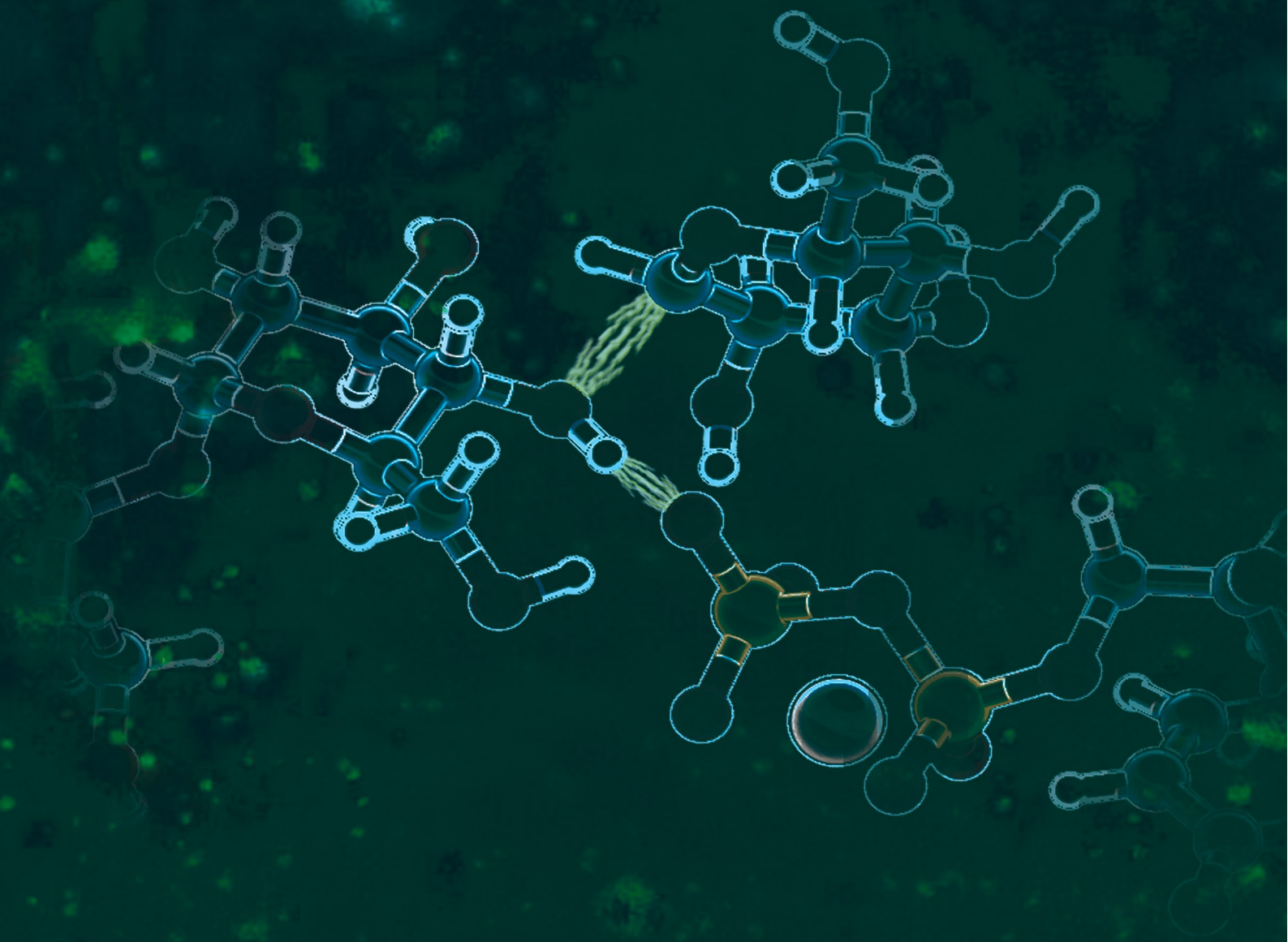
Unveiling Protein-Substrate Interactions and Conformations that Influence Catalysis in Carbohydrate-Active Enzymes

Lluís Adrià Raich Armendáriz

UNIVERSITAT DE BARCELONA

Unveiling Protein-Substrate Interactions and Conformations that Influence Catalysis in Carbohydrate-Active Enzymes



PhD Thesis

Lluís Raich Armendáriz

UNIVERSITAT DE BARCELONA
FACULTAT DE QUÍMICA

PROGRAMA DE DOCTORAT EN QUÍMICA TEÒRICA
I MODELITZACIÓ COMPUTACIONAL

Unveiling Protein-Substrate Interactions and Conformations that Influence Catalysis in Carbohydrate-Active Enzymes

Memòria de recerca presentada per en *Lluís Adrià Raich Armendáriz*

per tal d'optar al títol de Doctor per la Universitat de Barcelona.

Maig de 2018

Noms, cognoms i signatura del doctorand:

Lluís Adrià Raich Armendáriz

Departament de Química Inorgànica i Orgànica (Secció de Química Orgànica) i
Institut de Química Teòrica i Computacional (IQTC) de la Universitat de Barcelona.

Noms, cognoms i signatura de la directora i tutora:

Dra. Carme Rovira Virgili

Departament de Química Inorgànica i Orgànica (Secció de Química Orgànica) i
Institut de Química Teòrica i Computacional (IQTC) de la Universitat de Barcelona.
Institució Catalana de Recerca i Estudis Avançats (ICREA).

Prologue

Si può? Si può? Signore! Signori!
Scusatemi, se da sol me presento:
Iooooo soooooono iiiil Prroologo...

Enzymes have attracted the attention of chemists and biologists since long ago due to their molecular complexity and tremendous efficiency. They are highly specific catalysts that make chemical reactions possible at mild conditions with an astonishing rate enhancement. For this reason there have been many efforts to understand how these macromolecules work, trying to find out crucial factors that influence their activity. Whereas it is difficult to settle on how enzymes work in general, with hot debates over many years, significant progress has been made in elucidating specific catalytic mechanisms by means of experimental and computational approaches. The combination of structural data at atomic resolution, enzymology assays and cutting-edge QM/MM techniques has been necessary to identify subtle details that explain their great catalytic performance. Nonetheless, this interdisciplinary field is very young –the first crystallographic structure of an enzyme was solved in 1965 and the first QM/MM simulation was performed in 1976– and many intriguing questions remain unsolved.

Carbohydrate-active enzymes are particularly interesting because they deal with the huge diversity of structures that carbohydrates can adopt. Just a single change in the orientation of one exocyclic group in a carbohydrate substrate can make it completely inactive for enzyme catalysis, and therefore it is very important to understand how carbohydrate-active enzymes adapt to these –apparently irrelevant– molecular changes. Moreover, carbohydrates may have many functional groups and are extremely flexible molecules, which adds a degree of complexity in their study and comprehension. Several factors are known or presumed to enhance the reaction rate of carbohydrate-active enzymes, such as certain sugar conformations, enzyme-substrate interactions or the flexibility of the enzyme fold, but many of them remain poorly understood due to the lack of atomistic insights.

This thesis is aimed to unveil some of the essential enzymatic interactions and conformations that influence catalysis in carbohydrate-active enzymes, trying to provide computational proofs for general concepts that are usually assumed, as well as insights for the specific enzymes that have been studied. The manuscript is divided in seven chapters, including a general introduction, a theoretical framework, four chapters of results and a final chapter of general conclusions.

Chapter 1- General Introduction

We introduce the basic chemistry of carbohydrates, carbohydrate-active enzymes (glycoside hydrolases and transferases) and their essential features that contribute to catalysis. We highlight a set of open questions in the field and we list the objectives of the thesis.

Chapter 2- Theoretical Framework

We provide a detailed description of all methods that have been used along this work, paying special attention to the advantages and disadvantages of each one, along with a broad vision about the state of the art in computational chemistry.

Chapter 3- How Do Sugar Conformations Enhance Catalysis?

We perform an extensive conformational study of β-xylose in different environments and we estimate the reaction free energy contribution of sugar conformations to catalysis in a β-xylanase.

Chapter 4- The Contribution of 2-OH Interactions in the Reaction Rate of a Retaining *β*-glucosidase

We evaluate the influence of a crucial interaction (2-OH···Nucleophile) in the reaction mechanism and the free energy barrier of a retaining β-glucosidase.

Chapter 5- The Role of Water Binding Residues in the Active Site of an Inverting β-mannanase

We analyze the interactions of enzymatic residues that bind a catalytic water in the active site of an inverting β-mannanase and we disentangle its enzymatic reaction mechanism.

Chapter 6- Enzymatic Flexibility: Insights Into the Initial Steps of Glycogenesis

We decipher the reaction mechanism and the structural flexibility of a retaining α-glucosyl transferse, particularly its ability to adapt for different substrates and the conformational transitions that facilitate the release of the reaction product.

Chapter 7- General Conclusions

We summarize the main findings of the thesis and we provide a global concluding message.

Collaborations

Most of the works included in this thesis have been done in close collaboration with the following experimental groups:

- **Prof. Ramón Hurtado-Guerrero** (Chapter 4)

  Institute of Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, BIFI-IQFR (CSIC) Joint Unit, Mariano Esquillor s/n, Campus Rio Ebro, Edificio I+D, 50018 Zaragoza, Spain.

- **Prof. Gideon Davies** (Chapter 5)

  York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, YO10 5DD, U.K.

- **Prof. Spencer Williams** (Chapter 5)

  School of Chemistry and Bio21 Molecular Science and Biotechnology Institute and Department of Medical Biology, University of Melbourne, Parkville, Victoria 3010, Australia.

- **Prof. Ben Davis** (Chapter 6)

  Department of Chemistry, University of Oxford, Chemistry Research Laboratory, Mansfield Road, Oxford, OX1 3TA, U.K.

Additionally, thanks to the support of a PhD fellowship (APIF 2013-2014), I have been able to perform a doctoral stay of four months in a foreign research group of computational physics:

- **Prof. Mauro Boero** (Chapter 3)

  Institut de Physique et Chimie des Matériaux de Strasbourg (IPCMS), Département de Chimie et des Matériaux Inorganiques (DCMI), Centre National de la Recherche Scientifique (CNRS), 23 Rue du Loess, Strasbourg 67034 France.

My most sincere gratitude for all their insights, hints and support.

List of Abbreviations

All along this thesis we have used several abbreviations that are commonly employed in the field of theoretical Chemistry and molecular Biochemistry, here we list the most frequent ones:

Abbreviation	Full word
• A/B	*Acid/Base residue*
• C1	*Anomeric carbon*
• CAZyme	*Carbohydrate-active enzyme*
• CPMD	*Car-Parrinello molecular dynamics*
• CVs	*Collective variables*
• DFT	*Density functional theory*
• ESP	*Electrostatic potential derived atomic charges*
• FEL	*Free energy landscape*
• GEI	*Glycosyl-enzyme intermediate*
• GH	*Glycoside hydrolase*
• Glc	*Glucose*
• GS	*Glycosynthase*
• GT	*Glycosyl transferase*
• IP	*Ion-pair*
• MC	*Michaelis complex*
• MD	*Molecular dynamics*
• MM	*Molecular mechanics*
• MTD	*Metadynamics*
• Nuc	*Nucleophile residue*
• $O5/O_{Pyr}$	*Pyranic oxygen*
• O_{Gly}	*Glycosidic oxygen*
• O_{Nuc}	*Oxygen of the nucleophile residue*
• PBE	*Perdew-Burke-Ernzerhof exchange and correlation functional*
• PW	*Plane wave basis set*

- QM *Quantum mechanics*
- QM/MM *Quantum mechanics / molecular mechanics*
- RMSD *Root-mean square deviation*
- TG *Transglycosylase*
- TS *Transition state*
- UDP *Uridine diphosphate*
- US *Umbrella sampling*
- WT *Wild-type*

Contents

Chapter 1
General Introduction

1.1 A Sugar World

Why is there so much interest in the study of carbohydrates? This is a question that one may wonder when noticing that years of research are devoted to the study of these molecules. It is quite common to think about carbohydrates as a merely source of energy (*e.g.* white sugar) or as structural components of plants (*e.g.* cellulose), but recent advances in the field of glycobiology show that this simplistic view is far from reality. Nowadays it is well known that carbohydrates are not only one of the most important energy sources for living organisms nor the "architectural" units of most of their components, but they also play crucial roles in several cutting-edge applications.

In the field of biotechnology, for instance, carbohydrates are either used as sensors in industrial bioreactors, structural scaffolds in tissue engineering or heavy-metal ion-binders for the treatment of wastewater.[1–3] Carbohydrates are also at the forefront in biomedicine, where they are involved in drug delivery strategies, in the composition of wound dressing materials, vaccines and new therapeutics for treating many types of diseases.[4–6] This surprising range of applications leads to the natural question *why carbohydrates?* The answer is clear: because of their molecular complexity. Carbohydrates are formed by simple and abundant atoms –principally H, C and O– and yet they have a huge structural diversity that arises from several aspects of their arrangement (see Figure 1.1):

A) Carbohydrates are majoritarily assembled in cycles, forming rings that can have different sizes, with the most common being 5 and 6 membered rings (furanoses and pyranoses).

B) Carbohydrates are decorated with a variety of substituents or functional groups, including -OH, -NAc or $-OSO_3^-$ moieties among others.

C) These substituents are attached to chiral centers and can be found in different configurations (R and S), giving rise to multiple stereoisomers, such as α-glucose, β-glucose or β-mannose.

D) Carbohydrates can be linked together through several types of glycosidic bonds (*e.g.* 1-1, 1-4 or 1-6), forming oligosaccharides and polysaccharides.

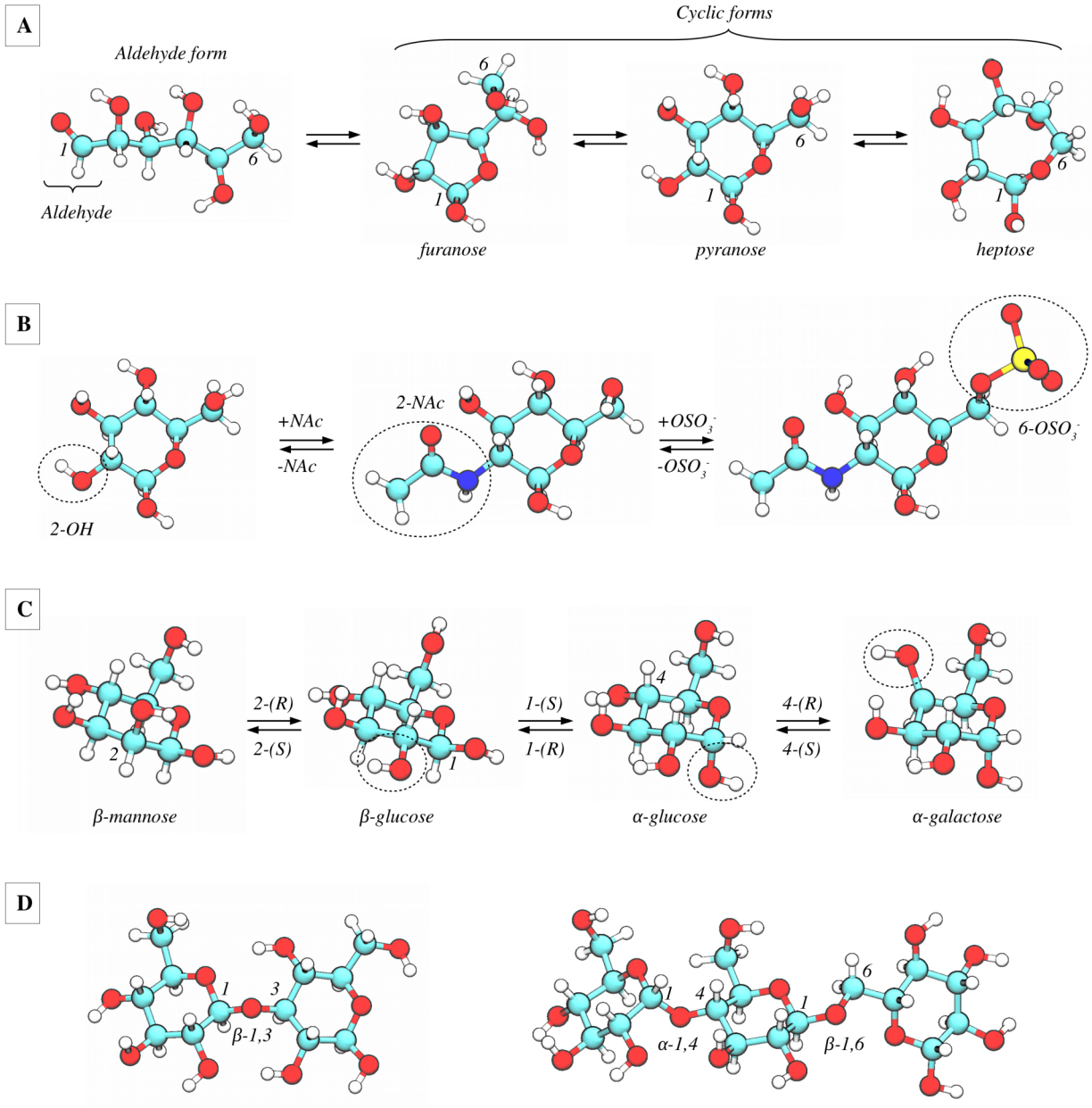


Figure 1.1- The structural diversity of carbohydrates arises from: (A) the ring size of its cyclic forms, (B) the functional groups attached to the cycle, (C) the spatial orientation of these groups, and (D) the glycosydic links between the monosaccharide units. The numbers next to the atoms refer to the carbon positions, and the arrows do not indicate chemical equilibrium.

The amount of possible structures that can be generated taking into account the above four aspects is enormous. To give an example, while the nucleotides that compose DNA can join forming *4.096* different hexanucleotides, the most frequent mammalian monosaccharides can join forming *192.780.943.360* hexasaccharides![7] Moreover, it is well known that carbohydrate rings possess a vast conformational space, which increases even more their complexity. They can adopt different shapes –up to 38 in the case of pyranoses– by the internal rotation of the bonds that compose their cyclic structure, leading to different orientations of the exocyclic groups and conferring them different properties (see Figure 1.2).[8]

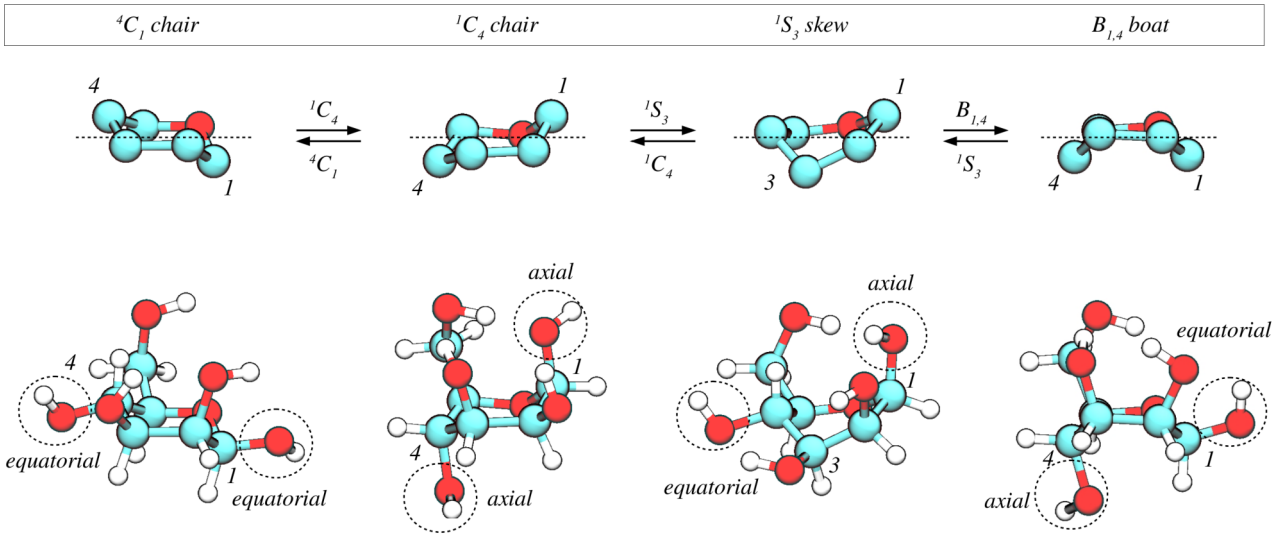


Figure 1.2- Four different ring conformations for a β-mannose monosaccharide. Notice that the conformational change does not only imply a rearrangement of the atoms that compose the cycle (top), but also –and most importantly– the ones of the exocylcic groups that are attached to it (bottom). According to the IUPAC,[9] pyranoses can be classified into *chair* (C), *half-chair* (H), *boat* (B), *skew-boat* (S), and *envelope* (E) conformations, each one displaying four atoms in a plane and two out-of-plane (except E conformers, which have only one out-of-plane atom). The out-of-plane atoms are indicated by superscript and subscript indexes.

All these structural particularities are the basis of an entire *world of sugars*, explaining their huge diversity and versatility for being used in many real-life applications. This aspect, at the same time, opens several interesting questions: how biological systems manage the incredible amount of existing sugars? Which are the most likely conformations adopted by a saccharide in different environments? How carbohydrates can be degraded and formed with precision?

1.2 The Biological Toolbox: Glycoside Hydrolases and Transferases

Enzymes are the most powerful catalysts in nature. They are able to enhance the rate of chemical reactions by several to many orders of magnitude. For instance, a glycosidic bond between two sugar monomers has an estimated half-life of about 5 million years,[10] whereas inside an enzyme the same bond cleaves in just a few milliseconds! Moreover, enzymes are very selective –both regio and stereochemically– and can work at mild conditions, being the perfect catalysts for cellular systems. In this regard, enzymes are essential for the processing and remodeling of carbohydrates in living organisms, carrying out manifold tasks in an efficient and precise manner. These enzymes are known as *carbohydrate-active enzymes* (CAZymes) and can be divided, majoritarily, in two big modules: glycoside hydrolases and glycosyl transferases.

1.2.1 Glycoside hydrolases

Glycoside hydrolases (GHs) or glycosidases are the enzymes in charge of the degradation of carbohydrates.[11] They are classified by their sequence, by the region where they cleave the carbohydrate and by their mechanism.[12] The sequence-based classification leads to the concept of GH *family* (enzymes that are related by sequence and, hence, by fold) and GH *clan* (group of families with high structural similarity).

Enzymes from the same family usually act on a similar substrate and share the same catalytic mechanism. The cleaving region differentiates *endo-* and *exo-* acting enzymes, with the former being able to cleave non-terminal sugars and the latest terminal sugars of a carbohydrate chain. Within the mechanistic classification we can find *retaining* or *inverting* enzymes depending if the outcome of the reaction retains or inverts the configuration at the anomeric center (see Figure 1.3). This classification is perhaps the most important one as it gives information about the reaction details. Despite the large number of GH families known (up to 153 by April 2018),[13,14] most of them follow a common reaction mechanism involving either one or two classical S_N2 displacement reactions, assisted by two essential residues: a *general acid* (or acid/base residue) and a *general ba*se (or nucleophile residue).[15] The former is usually a glutamic or aspartic acid, whereas the latest is a glutamate or aspartate conjugate base. Remarkable exceptions are sialidases, in which the nucleophile is an activated tyrosine,[16,17] or chitinases, in which the nucleophile is the 2-N-acetyl substituent of the -1 saccharide.[18]

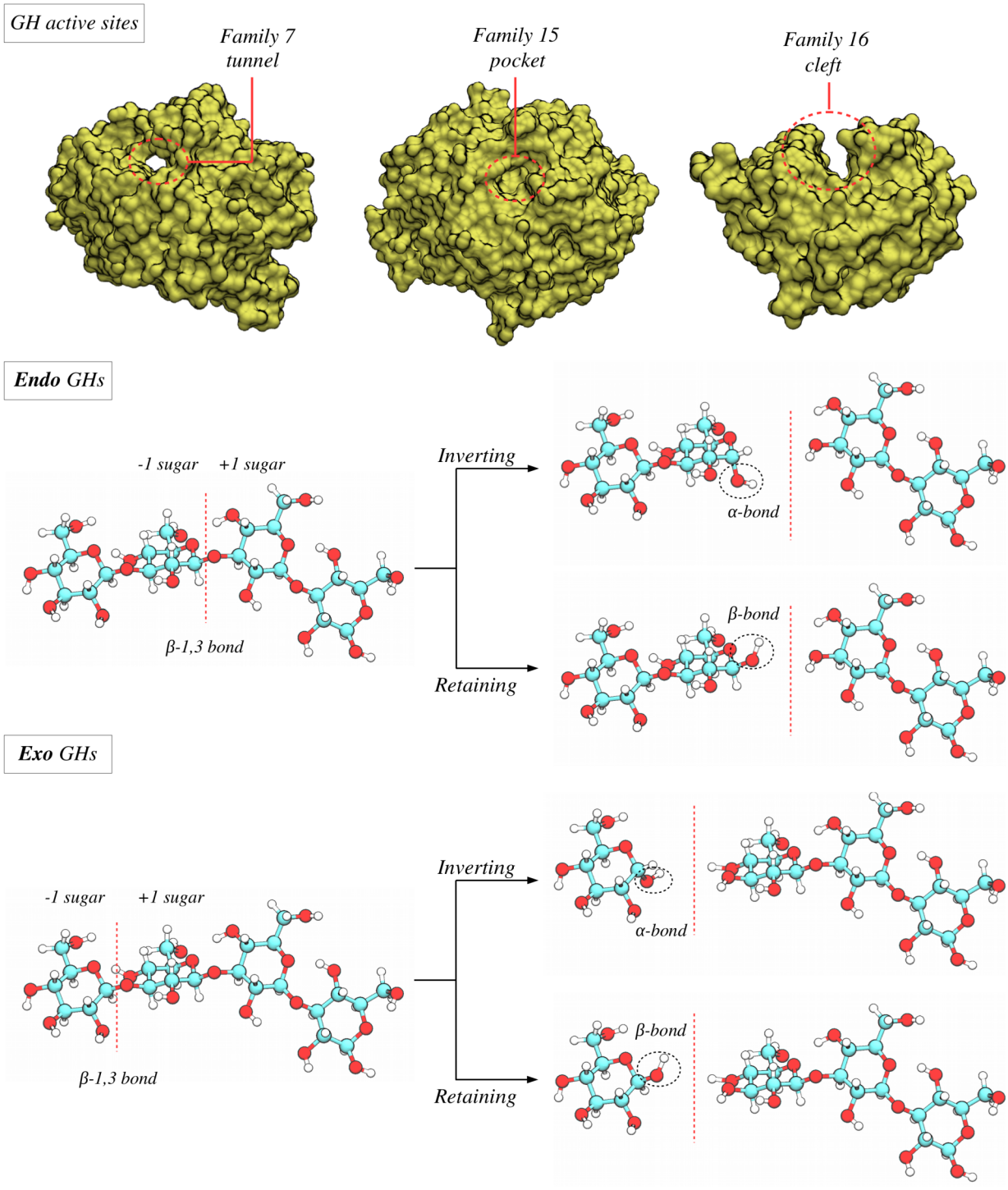
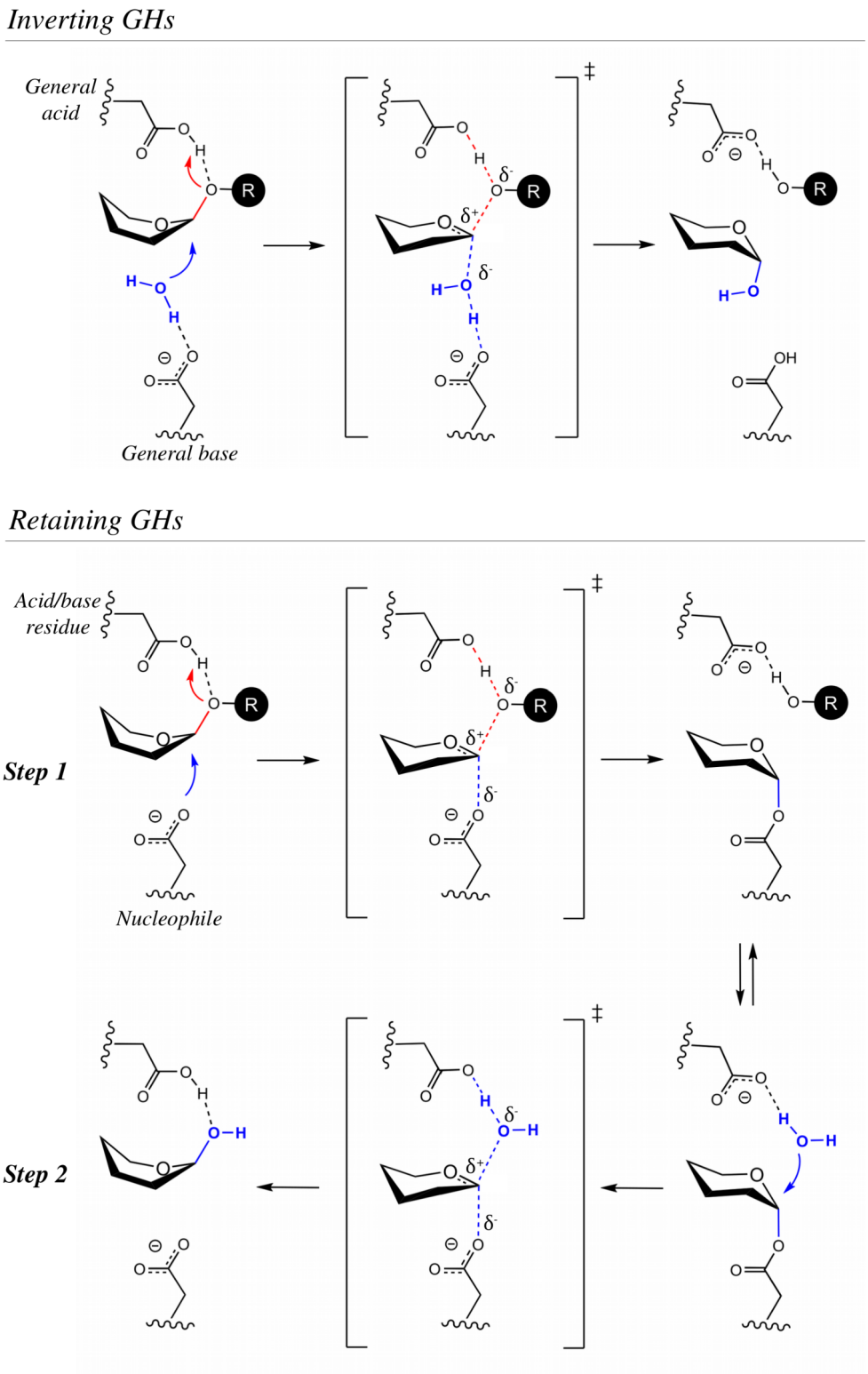


Figure 1.3- Classification of GHs. Three GH families with different sequences and different types of active sites: tunnel, pocket and cleft (PDBs 1GAI, 1GBG and 1EGN; top panel). Inverting and retaining substrates and products of *endo* and *exo* GHs (middle and bottom panels). Notice that a water molecule, not represented here, is needed to hydrolyze the substrates. The sugar at the -1 sub-site (or -1 sugar) is the one that reacts with the nucleophilic water either with inversion or retention of the anomeric configuration.



Scheme 1.1- Classical mechanisms of GHs: inverting and retaining mechanisms (top and bottom). The "R" moiety denotes a sugar molecule or chain. Notice that between step 1 and step 2 in the retaining mechanism a water molecule replaces the leaving group. Binding and unbinding molecules in the equilibrium have been omitted for clarity, as well as sugar exocylic groups.

Experimental and theoretical evidence support that inverting enzymes follow a single S_N2-type displacement with the nucleophilic attack of a water molecule.[19,20] In this case, the general acid activates the cleavage of the glycosidic bond and the general base facilitates the deprotonation of the nucleophilic water (see Scheme 1.1, top). The classical retaining mechanism, instead, consists on a two-step double S_N2-type displacement with the formation of a stable covalent intermediate (Scheme 1.1, bottom). In a first step, the acid/base residue acts as a general acid to activate the glycosidic bond at the same time as the nucleophilic residue attacks the anomeric carbon, forming a *glycosyl-enzyme intermediate* (GEI). In a second step, the GEI breaks by the nucleophilic attack of a water molecule, which is activated by the acid/base residue, now acting as a general base. The resulting product retains the configuration at the anomeric center –two inversions lead to net retention– and the whole mechanism restores the protonation states of the enzymatic residues.

It is important to highlight that in all the transition states (TS's), either for inverting or retaining mechanisms, the sugar that is transferred acquires a substantial oxocarbenium ion character.[21] The charge developed at the anomeric carbon can be stabilized by the internal delocalization of the pyranic oxygen lone pairs, which is maximal for certain sugar conformations that display the C2, C1, O5, and C5 atoms coplanar (for more details see section 1.4 below).

1.2.2 Glycosyl transferases

Glycosyl transferases (GTs) are the antithesis of GHs, as they are in charge of the synthesis of carbohydrates. These enzymes use activated substrates to synthesize new glycosidic bonds given that the glycosydic linkage is thermodynamically unfavorable (by *c.a.* 3 kcal·mol^{-1} for cellobiose[22]). Most GTs can transfer glycosyl units from a *donor substrate* –the activated substrate– to an *acceptor substrate* –the one that receives the sugar– with inversion or retention of configuration at the anomeric center (see Figure 1.4).

Two structural folds, GT-A and GT-B, have been so far identified for GTs that use activated sugars composed by nucleotide phosphates.[23] The departure of the leaving group in GT-A fold enzymes is typically promoted by a divalent cation (*e.g.* Mg^{2+} or Mn^{2+}), while in GT-B folds the role of the cation is substituted by positively charged residues. Experimental and theoretical evidence support that the inverting mechanism follows a S_N2-type displacement with the assistance of a general base residue that facilitates the deprotonation of the acceptor (see Scheme 1.2, top).[24]

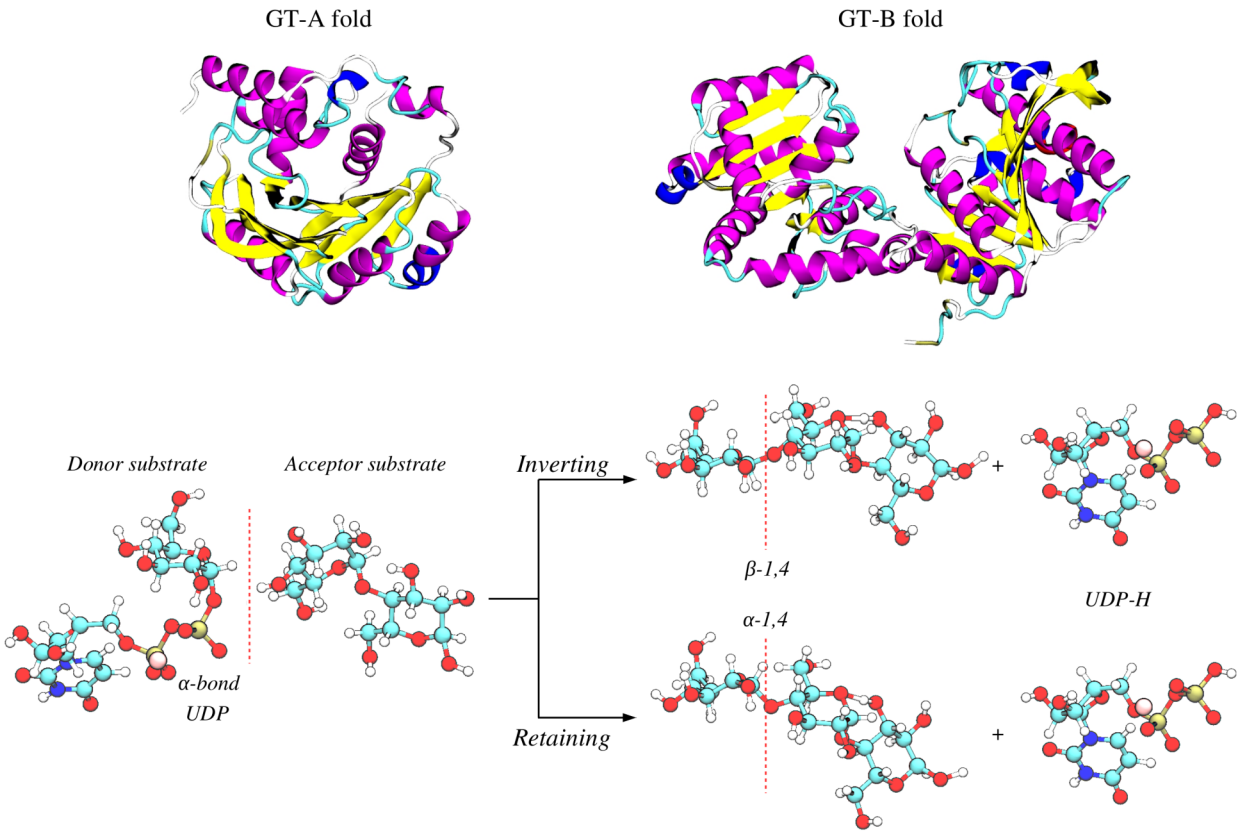
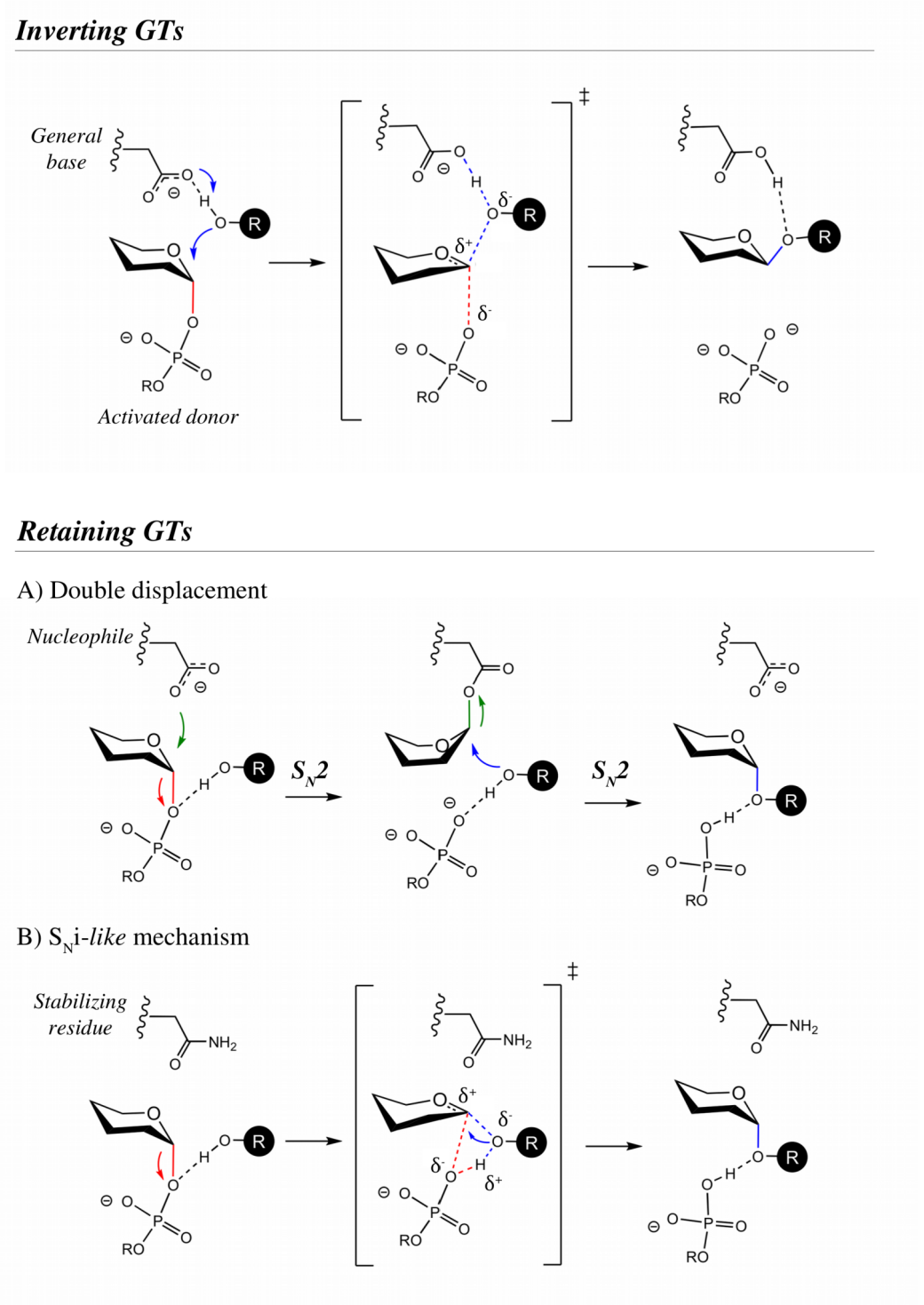


Figure 1.4- Fold-types A and B of glycosyl transferases families 2 and 5 (PDBs 1QGQ and 3FRO; top). Reaction between a donor and an acceptor substrates leading to inverting and retaining products (bottom).

The retaining mechanism is not as clear as the inverting one. There are at least two possibilities: (i) the classical mechanism with a double S_N2 displacement –just like in GHs– and (ii) the internal nucleophilic substitution (S_Ni; Scheme 1.2, bottom). In the S_Ni mechanism, the leaving group of the reaction acts as a base that abstracts the proton of the acceptor while it is approaching by the same face from which the leaving group has departed. Strictly speaking, we should refer it as an S_Ni-*like* mechanism, as the pure S_Ni mechanism is intramolecular (there is only one molecule), while in this case we refer to an intermolecular reaction (two molecules are involved, the donor and the acceptor sugars). The S_Ni-*like* mechanism is most likely to occur when there is no suitable nucleophilic residue near the catalytic center, such as glutamine or asparagine, thus making a double displacement improbable. Our group was pioneering in deciphering the catalytic S_Ni-like mechanism for retaining GTs,[25] which was later also associated to other family enzymes.[26]



Scheme 1.2- Classical mechanisms of GTs: (top) inverting mechanism and (bottom) the two possible retaining mechanisms: (A) double displacement and (B) S_Ni-*like*. The "R" moiety denotes a sugar molecule or chain. Notice that we have omitted the TS's in the double displacement mechanism of retaining GTs. In the S_Ni-*like* mechanism we have represented a non-nucleophilic glutamine that is presumed to stabilize the TS through electrostatic interactions.

1.3 Do not Stop Moving: Enzymatic Dynamics and Function

Enzymes have been classically viewed as rigid macromolecules that catalyze the reaction of specific substrates. The early *key-and-lock* model of Emil Fischer,[27] in which a single substrate –the key– binds to a rigid enzyme –the lock–, was the paradigm of this static point of view. While this model was able to explain the high specificity of enzymes, it failed to describe many other factors that arose over years of research, such as regulation, cooperativity or promiscuity.[28,29] This prompted Koshland to postulate the *induced fit* model,[30] which basically states that a substrate can induce appreciable changes in the structure of an enzyme upon binding, and that these changes are required for the catalytic activity (see Figure 1.5). The consideration of enzymatic flexibility allowed to rationalize the aforementioned factors and opened a novel dynamic vision of molecular recognition, nowadays proved by plenty of evidences.[31,32] Another accepted model states that the dynamical changes of enzymes do not occur after ligand binding (*i.e.* induced fit), but rather when the enzyme is still free of ligand.[33] This alternative view, called *conformational selection*, postulates that there is an ensemble of pre-existing enzyme conformations available for the ligand, which binds selectively to the most favored ones (Figure 1.5).

The difference between the two most accepted models lay on whether the conformational changes of the enzyme occur after or before substrate binding, but there is no doubt that such molecular changes take place before the chemical event. Although conformational flexibility can be seen as a drawback for catalytic efficiency (at the end only a small number of conformations can be optimally active), it confers versatility to the enzyme for acquiring new functions and structures that could be essential for adaptation in front of selection pressures.[34]

Among the huge amount of possible conformations that an enzyme can adopt (up to 10^{70} for a 100-aminoacid sequence), just a few that are similar to the "native" conformation are available at a given temperature.[35] Transitions between these conformations display a timescale that range from picoseconds to seconds depending if they involve local residue motions or entire domain rearrangements.[36] For this reason, several techniques are needed to capture the large amount of dynamical transitions, including NMR spectroscopy, X-ray diffraction and molecular simulations. As we will see all along this thesis, CAZymes are not an exception in terms of dynamism, both for local and global motions.

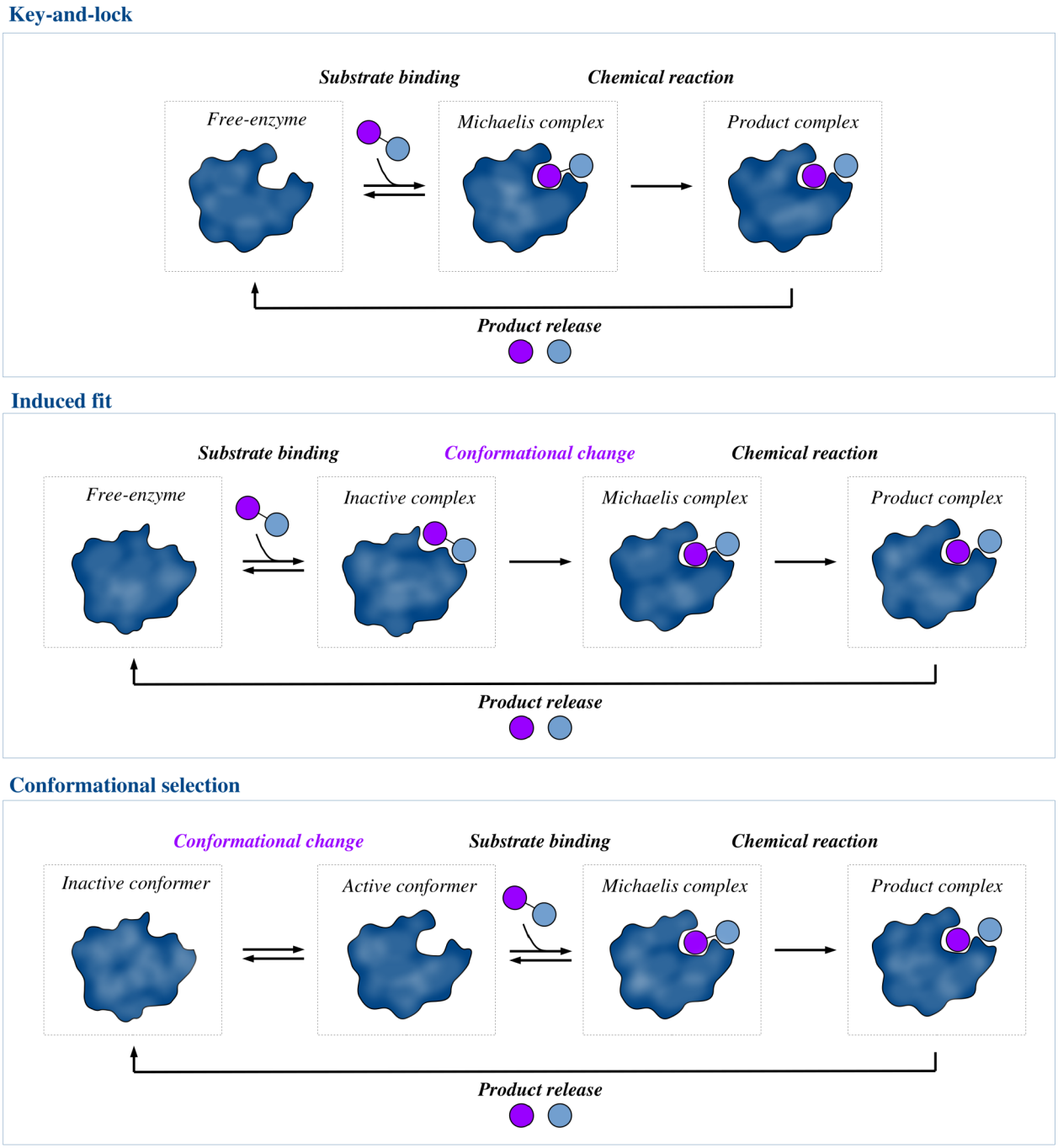


Figure 1.5- Different models of substrate-enzyme recognition: the *key-and-lock* (top), the *induced fit* (middle) and the *conformational selection* (bottom). Notice the subtle changes of the enzymatic cavity after the "conformational change" steps in the last two models, something that does not happen in the first given that the enzyme is considered as a rigid body. The *key-and-lock* model is a particular case of *conformational selection* in which there is a single conformation for the free-enzyme (the active one). For clarity, the substrate in the binding step has only been drawn once in the arrows that indicate equilibrium.

1.4 On the Importance of Sugar Conformations

Although sugar conformations may seem subtle details that lack importance, they are useful both for applications and scientific knowledge. The fact that different conformations lead the exocyclic groups in different orientations makes some of them fit better in the active site of a given CAZyme, which has a direct impact in the design of inhibitors.[37] Moreover, as mentioned before, only few conformations are able to –optimally– stabilize the oxocarbenium ion TS through the internal delocalization of the pyranic oxygen lone pairs, which represents a "restriction" that is informative of catalytic mechanisms. In particular, favored mechanisms are expected to connect the conformations of two stable states (*e.g.* the Michaelis complex and the glycosyl-enzyme intermediate) passing through one of the optimal TS conformations. The evolution of these conformations during catalysis is known as *the catalytic itinerary* of a given CAZyme (see Figure 1.6), and can taken as a reference for the goodness of a predicted mechanism.[8,20,38]

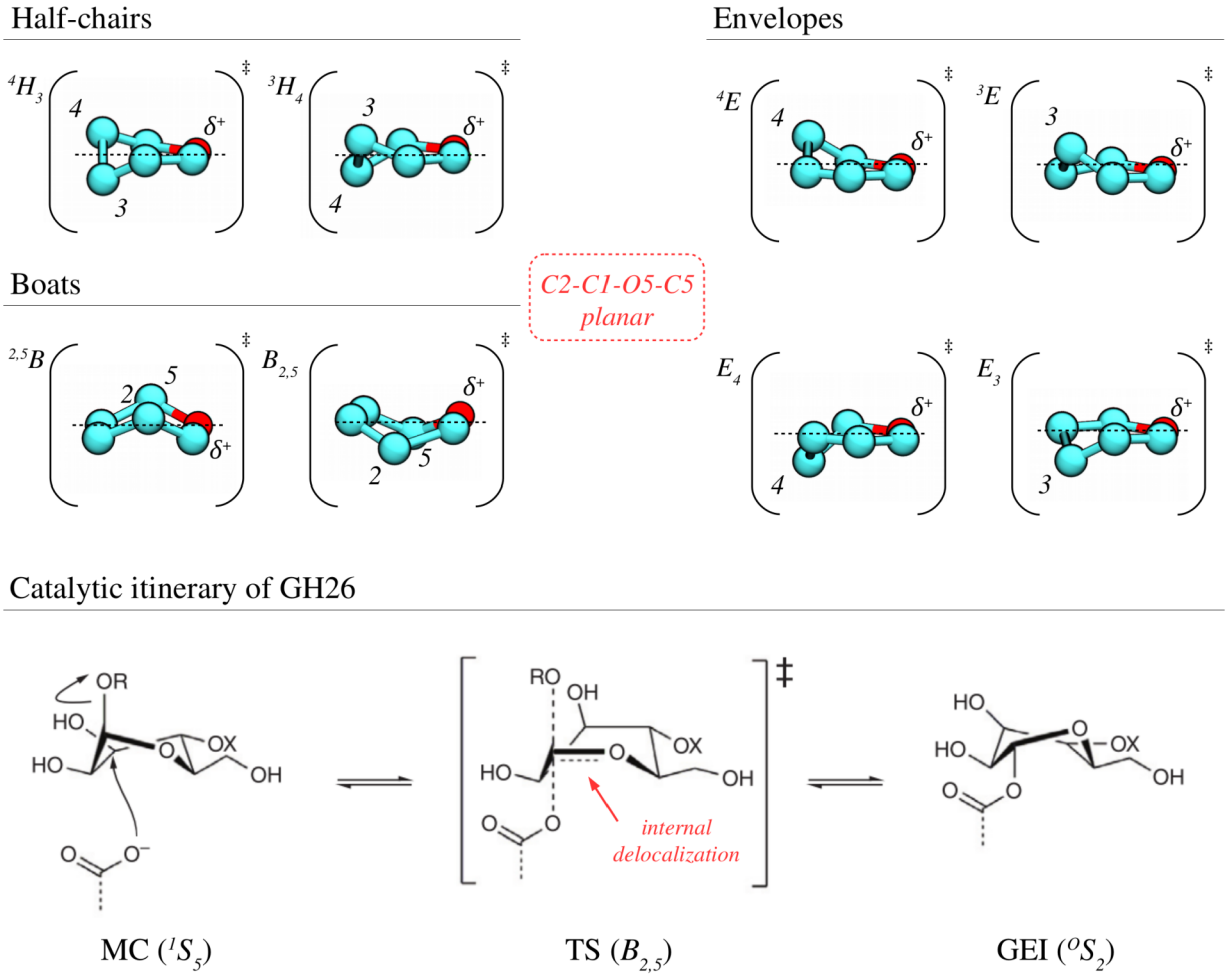


Figure 1.6- Favored transition state conformations for a pyranose ring and example of a catalytic itinerary (denoted as $^1S_5 \rightarrow [B_{2,5}]^{\ddagger} \rightarrow {}^OS_2$) proposed for a family 26 β-mannanase.[39] Notice that the C2-C1-O5-C5 atoms of the ring are planar to favor the internal delocalization of electrons.

Disclosing catalytic itineraries is therefore a topic of relevance in glycobiology, and several experimental and computational efforts are being devoted to fully resolve them.

1.4.1 Drawing catalytic itineraries on the world map of sugars

Ring conformations can be quantitatively classified by the use of a set of *puckering coordinates* that enclose their whole mathematical space (see Chapter 2 for details).[40] For a six-membered ring, these coordinates define a sphere in which one can map all the 38 existent conformations. This "world globe" of sugar conformations can be projected into two dimensional maps –Stoddart and Mercator representations; see Figure 1.7– that are very useful and popular for the discussion of catalytic itineraries.

Experimental techniques such as X-ray crystallography aim to trap the Michaelis complex and the product complex of a reaction, from which one can infer the possible catalytic itineraries by following pathways through the most likely TS's, while computational simulations aim to obtain this information by computing the lowest free energy pathway on these maps.

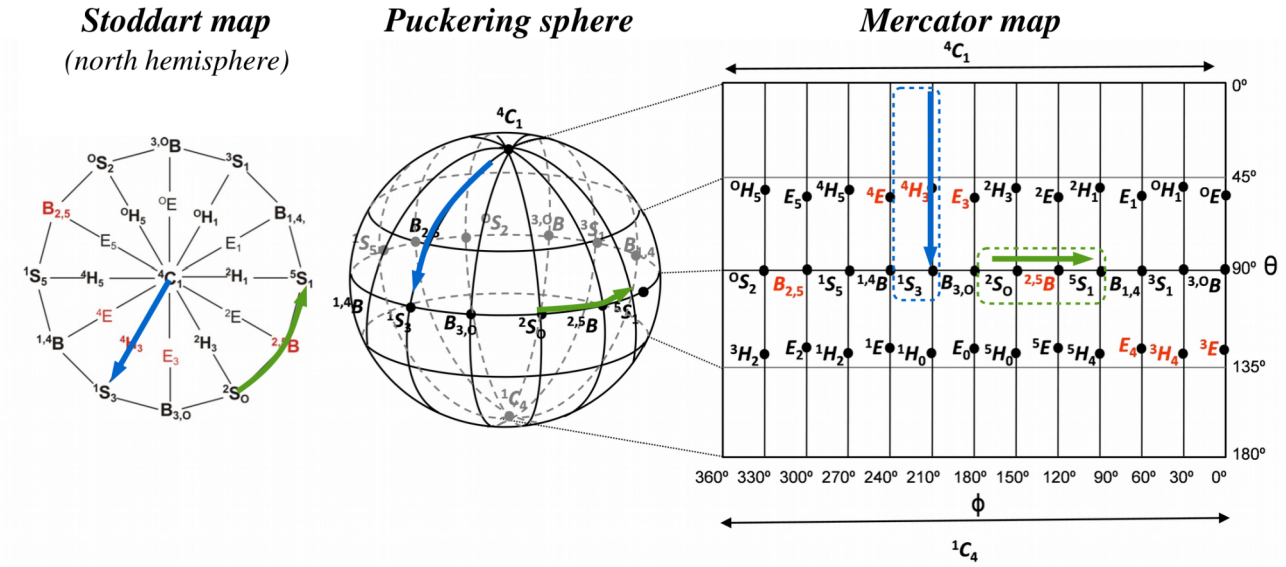


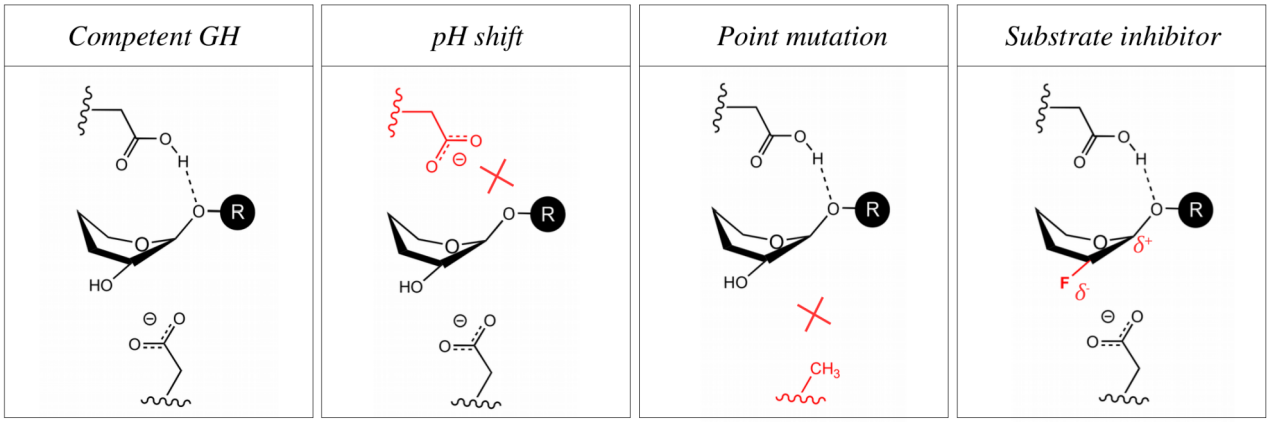
Figure 1.7- Stoddart and Mercator representations of the puckering sphere for a six-membered ring. The optimal TS conformations are highlighted in red. Two catalytic itineraries are drawn on the different representations, one that is equatorial ($^{2}S_{O} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$, in green, followed by β-xylosidases) and another that is northern ($^{4}C_{1} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{1}S_{3}$, in blue, followed by α-glucosidases).

1.4.2 Experimental traps for elusive complexes

One of the main challenges from experimental techniques lays on the fact that Michaelis complexes usually react in a timescale that is much below the time resolution of the experiment, and product complexes normally dissociate or their substrates relax towards conformations that are not informative.[39] This is the case of X-ray crystallography, for instance, in which measures are collected in a time window of several seconds, whereas CAZyme reactions occur at the millisecond time scale. Therefore, if one pursues to trap one of these states, the reaction should be somehow slowed down. There are several strategies for trapping elusive complexes, which involve modifications in the experimental conditions,[41] the enzyme[42] and/or the substrate[43] (see Scheme 1.3). These include:

- Shift the pH at ranges where the enzyme is inactive.
- Knockout the catalytic residues through site directed mutagenesis.
- Use sugar-based inhibitors, such as fluorinated or thioderivative substrates.

All these strategies lead to long-lived complexes that can be experimentally observed, but at the cost of perturbing the natural conditions at which the enzyme is competent. This leads to the question on whether the modifications that inactivate the enzyme could affect the conformation of the substrate, and for this reason it is mandatory to double-check the results in order to discard artifacts and verify that the modified system is a good mimic of the natural one.



Scheme 1.3- Experimental strategies for trapping elusive complexes, illustrated for a retaining GH. From left to right: (i) natural enzyme and natural substrate at optimal conditions; (ii) natural enzyme and natural substrate at a high pH, leading the general acid deprotonated; (iii) modified enzyme and natural substrate, with the nucleophilic residue mutated to alanine; (iv) natural enzyme and 2-fluoro modified substrate, which destabilizes the transition state through inductive effects.

1.4.3 Theoretical insights from computer simulations

In the last years, the astonishing growth of computational facilities has made theoretical simulations be at the vanguard of enzymatic research. Furthermore, the development of smart techniques such as the quantum mechanics/molecular mechanics approach (QM/MM), pioneered by Arieh Warshel and Michael Levitt,[44] has allowed to treat large enzymes with an all-atom viewpoint. The first application of QM/MM methods was precisely done on a GH enzyme, the egg white lysozyme, and sugar conformations were already discussed as an important issue.[45]

Predicting sugar conformations and catalytic itineraries from a theoretical perspective, however, is not a trivial task. The main challenge of computer simulations is to verify if the model is reliable enough to make accurate predictions. The goodness of a model involves the type of method used to describe the molecular interactions, the similarity of the modeled system with respect to the experimental system and the time that is devoted to explore the huge number of mechanistic pathways between two states of interest (see Chapter 2 for more details). All these features need to be approximated to some extent in order to have a reasonable balance between accuracy and computational cost, leading to the "Achilles heel" of computer simulations. While crude approximations can result in wrong –or not even wrong[46,47]– theoretical predictions, proper approximations can drive –and even correct[48]– experimental proposals. It is for this reason that both computer simulations and experiments are necessary and complementary tools for deciphering catalytic itineraries and enzymatic mechanisms. In the case of CAZymes, and GHs in particular, great advances have been achieved through the computation of reaction mechanisms and free energy landscapes (FELs).[49–51]

Recently, it has been shown that one of the most clever strategies in terms of accuracy and computational cost is to compute the conformational FEL of isolated sugars.[52,53] These predictions are based on the topology of the landscape –*i.e.* the number of minima and their energies– and the stereoelectronical particularities of each conformation, such as the glycosidic bond distance, its axiality and the charge at the anomeric carbon. Conformations that fall in low-energy regions and that display stereoelectronical features resembling the ones of TS-like oxocarbenium ions are considered as *preactivated* for catalysis. Remarkably, these preactivated conformations correlate with the ones found in X-ray structures of GH Michaelis complexes, reaction intermediates and TS-analogue complexes (see Figure 1.8),[8,54,55] representing a valuable mechanistic tool. The FEL of an isolated sugar, therefore, can be considered as a *fingerprint* that give information about the conformations available inside GHs.

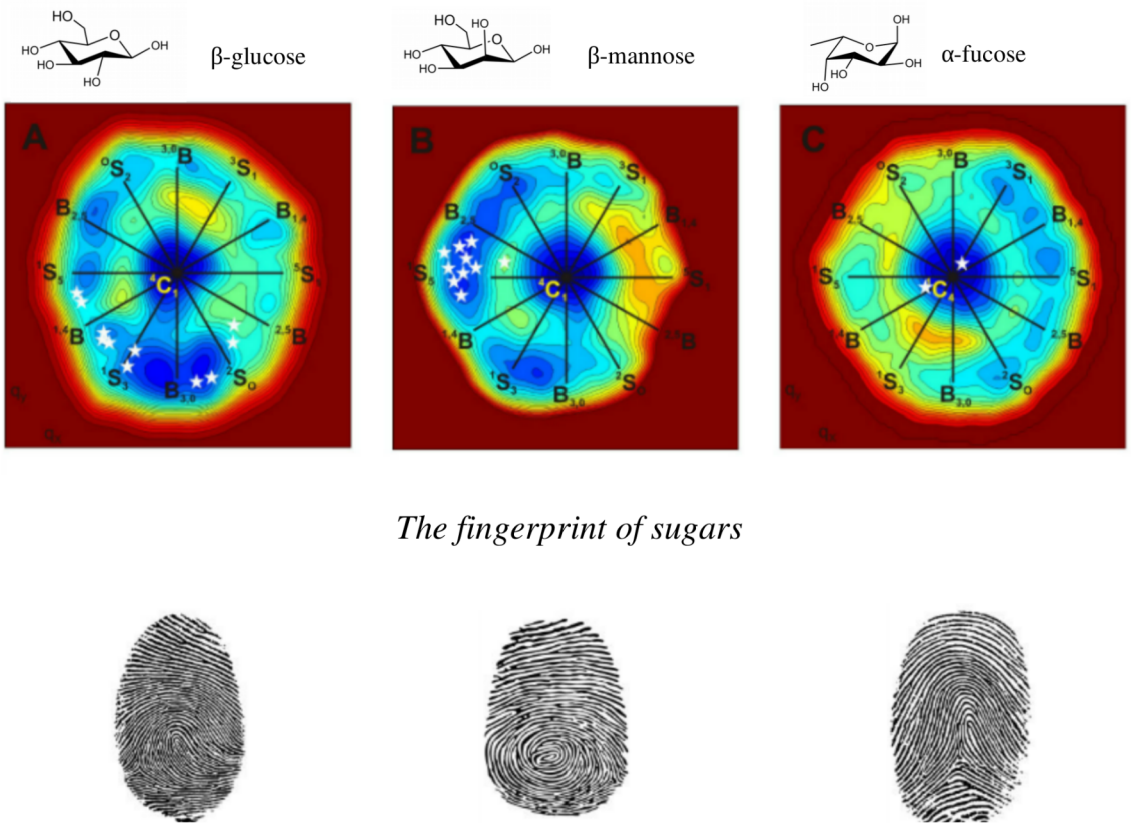


Figure 1.8- Conformational free energy landscapes (FELs) of β-glucose, β-mannose and α-fucose projected into a northern Stoddart representation.[20] Star symbols indicate the observed conformations of different Michaelis complexes structures.

1.5 Essential Features of a Fruitful Biocatalyst

The utopia of any researcher in enzymatic catalysis is to be able to create, in a rational manner, new enzymes from scratch.[56–58] This goal, however, is far from being a reality for two reasons: (i) first we should be able to determine which are the crucial factors of enzymatic catalysis and relate them with specific molecular distributions and motions, and (ii) we should be able create a precise fold of the new enzyme fulfilling those molecular conditions. Whereas it can be straightforward to determine the macroscopic factors that influence the catalysis of a particular enzyme (*e.g.* dependence on temperature, pH or the presence or absence of certain residues), finding an atomistic explanation for those observations is usually the critical point. The huge complexity of enzymatic systems makes their understanding arduous, and even more, in the case of CAZymes, if we add the inherent complexity of their substrates. Even though these intrinsic difficulties, significant progress has been made in the elucidation of the intrinsic details that make enzymes powerful catalysts, highlighting that conformational, electrostatic and entropic effects –among others– could be responsible

of their rate enhancements.[59–63] To illustrate the essential features of a fruitful CAZyme, based on experimental and theoretical evidences discovered so far, we will focus on GHs, which are able to enhance the cleavage of glycosidic bonds as much as ~10^{17} fold.[21] They achieve these astonishing rate enhancements, as well as their selectivity and specificity, by the use of:

- **General acid/base catalysis:** one of the major sources of catalytic power in GHs comes from the aforementioned assistance of general acid and general base residues. Subtle modifications in one of these residues leads to dramatic drops in the catalytic activity of GHs, rendering very inefficient or even inactive enzymes.[64,65] These residues participate directly in the chemical reaction and modify the catalytic mechanism with respect to the one that is expected in solution. For instance, the double displacement mechanism of retaining GHs is unlikely to occur in the absence of such catalytic residues, and thus the hydrolysis of carbohydrates in solution proceeds through alternative and less efficient mechanisms (*e.g.* specific acid/base catalysis to form an acyclic hemiacetal prior to hydrolysis[66]).

- **Conserved hydrogen bonds:** another major source of catalytic power in GHs are highly conserved enzyme-substrate hydrogen bonds, and particularly the ones involved with the OH groups of the -1 sugar. These interactions are able to change the stereoelectronic properties of both the substrate and the enzymatic residues,[67,68] leading to more reactive complexes. Rigorous experimental studies have been conducted to estimate the contribution of particular OH groups in the catalysis of β-glucosides by GHs, showing that they range from ~2 to 11 kcal·mol^{-1} depending on the position of the OH in the pyranic ring,[69] with the 2-OH interactions being the most important.

- **Optimal electrostatic environment:** long-range electrostatic interactions are also essential for GHs. They are involved in the modulation of the pK_a of the catalytic residues,[70–72] the stabilization of highly charged polysaccharides such as heparan sulfate[73] or the change in the stereoelectronic properties of the substrate (like hydrogen bonds do, but without the necessity to be near in the space).[74,75] Among the last, there are GHs whose activity crucially depend on the presence of metal ion cofactors, such as Ca^{2+} or Zn^{2+}, which are involved in the polarization of the substrate and in the stabilization of the nascent charges that develop at the transition state.[76,77]

- **Well-defined binding subsites:** the proper orientation of substrates in the active site of GHs is determined by well-defined binding subsites. Changes in the molecular interactions that comprise these subsites –not only of polar and electrostatic nature, but also of hydrophobic– have effects both in the selectivity and the activity of GHs.[78,79] For instance, mutations in the -2 and -4 subsites of a GH12 have revealed the importance of π-stacking interactions in the activity of the enzyme, highlighting that even distant regions from the active site can play a role in catalysis.[80] Similarly, modifications in both positive and negative subsites have been proposed to change the enzymatic activity from hydrolysis to synthesis,[81] leading to a type of GHs that are called transglycosylases (TGs). The importance of binding in GHs is also reflected in the fact that, nature, has destined efforts in the generation of entire domains called *carbohydrate-binding modules* that are crucial substrate recognition and catalysis.[82]

- **Enzyme and substrate flexibility:** as pointed above, flexibility of both the enzyme and the substrate is fundamental for the catalytic activity of CAZymes. In particular, GHs from families 5 and 12 bear flexible loops that undergo conformational transitions approaching to the bound substrates, which increases favorable enzyme-substrate interactions and improves the catalytic efficiency of the enzymes.[83,84] Substrate flexibility is also very important for GHs. The different properties of sugar conformations have been exploited by these enzymes, as evidenced in numerous X-ray structures with substrates distorted in unusual conformations.[39] These conformations are presumed to enhance catalysis by their specific stereoelectronical effects, leading to pre-activated conformations that are on the pathway towards the TS. This hypothesis is indirectly supported by experiments on sugar-based inhibitors locked in different conformations,[37,85,86] suggesting that certain conformations are better adapted for catalysis.

In summary, there are many features that are important for GHs and CAZymes in general, with several residues involved in such functions (*e.g.* catalytic, TS-stabilizing, binding or structural residues), highlighting that in an enzyme *every atom counts*.

1.6 Open Questions: The Veil of Mystery is Still Hiding Answers

In spite of the enormous advances in the understanding of how CAZymes work during the last twenty years,[8,21,87] there are still several questions in the field that remain without answers. For instance, there is an increasing amount of crystallographic structures that lack a general acid or base residue nearby the catalytic center,[41,88] opening the question on how the enzymes can be efficient without them. Do conformational changes place a proper residue before the reaction? Or do these enzymes employ different reaction mechanisms away from the classical ones? In the line of the last question, recent studies suggest that GH catalytic mechanisms can be more exotic than previously thought, with evidences pointing towards the formation of epoxide[89] or 1,2-unsaturated[90] intermediates during the reaction.

Other open questions are how some GHs, called transglycosylases (TGs), are able to synthesize carbohydrates in a "hostile" 55 M waterworld,[81] or how mannosidases, which degrade mannose substrates that have the 2-OH substituent in an axial position, deal with the destabilizing effects associated with those 2-configuration.[91,92] In addition, and as briefly explained in section 1.4, there is always the doubt whether some sugar distortions observed by X-ray experiments from modified systems are reliable enough, and while this issue can be properly addressed for families that are well characterized, there are serious concerns about the specific itineraries of newly discovered families.

The realm of GTs is also full of open questions, particularly those ones that operate through a retaining mechanism and display suspicious active site residues that could act as nucleophiles.[93,94] There is also a long debate on the existence of short-lived oxocarbenium intermediates in the S_Ni-*like* mechanisms, with studies supporting both possibilities.[25,95] Furthermore, mechanistic details about the conformational flexibility of GTs also remain to be characterized at the atomistic level, being nowadays poorly understood. Here we have addressed some of these questions by focusing in four specific CAZymes (3 GHs and 1 GT) that are interesting from both theoretical and practical points of view:

- β-xylanases, analyzed in Chapter 3, are GHs responsible for the hydrolysis of glycosidic bonds in β-xylans, a group of hemicelluloses of high biotechnological interest that are found in plant cell walls. The precise conformations followed by the substrate during catalysis, however, have not been unambiguously resolved, with three different catalytic itineraries being proposed from structural analyses, two of them for a single family (GH11). Can a sin-

gle family can harness different catalytic itineraries? How do different β-xylanases adopt different itineraries?

- *Sc*Gas2, addressed in Chapter 4, is a GH72 retaining β-glucosidase that is able to synthesize carbohydrates through transglycosylation. The conversion of GHs into TGs, *i.e.* from enzymes that hydrolyze carbohydrates to enzymes that synthesize them, represents a promising solution for the large-scale synthesis of complex carbohydrates for biotechnological purposes, but the lack of knowledge about the molecular mechanisms of TGs hampers their understanding and rational design. Recent studies in family GH1 show that mutations affecting the 2-OH interactions lead to high transglycosylation yields.[96] Which is the role of those interactions in *Sc*Gas2? What are the subtle mechanistic details that make the difference between hydrolysis and transglycosylation?

- *Ss*GH134, investigated in Chapter 5, is a newly discovered inverting β-mannanase of family 134 GHs with an unexpected fold and an unusual conformational itinerary ($^1C_4 \rightarrow [^3H_4]^{\ddagger} \rightarrow {}^3S_1$) predicted from crystallographic evidence. Whereas it is clear that the mechanism of the reaction is the classical single S_N2 displacement, both by kinetic and crystallographic data, the catalytic itinerary is incongruous: the -1 sugar of the product complex displays a $^1C_4$ conformation, instead of the $^3S_1$ that has been proposed. Does the product relax from $^3S_1$ to $^1C_4$ after the reaction? Or does the mutation made to trap the Michaelis complex has perturbed the observation?

- Glycogenin (GYG), studied in Chapter 6, is a GT of family 8 that is involved in the first steps of glycogen formation, a fundamental molecule for human life. GYG is able to catalyze sequentially the synthesis of a polysaccharide chain that is anchored to the enzyme, enlarging it up to more than 10 glucose units. This opens the question on how a single enzyme can adapt its active site to accept different lengths of acceptor substrates without loosing activity. Moreover, it is unknown whether the reaction mechanism of GYG consist on a double displacement or an S_Ni-like mechanism, with experimental proposals in favor of both possibilities.

Objectives

In this thesis we have used computational techniques with the object of unveiling protein-substrate interactions and conformations that influence catalysis in carbohydrate-active enzymes. To that end, we have uncovered the molecular reaction mechanism of the four different CAZymes highlighted in the previous section, pursuing the following specific objectives:

- Disclose the catalytic itineraries of β-xylanases and quantify the contribution of sugar conformations in catalysis for a prototypical retaining β-xylanase. These objectives are addressed in Chapter 3.

- Evaluate the contribution of non-covalent 2-OH interactions in the reaction catalyzed by *Sc*-Gas2 retaining β-glucosidase, deciphering the mechanism of transglycosylation and the interactions that could affect it. These objectives are pursued in Chapter 4.

- Unravel the reaction mechanism of the newly discovered *Ss*GH134 inverting β-mannanase, including the catalytic itinerary and the contribution of water binding residues in the orientation and stabilization of the catalytic water. These objectives are pursued in Chapter 5.

- Determine the ability of glycogenin (GYG) to adapt to different lengths of acceptor substrates and uncover the enzymatic reaction mechanism. In particular, find out whether a short-lived oxocarbenium intermediate is compatible with the reaction. These objectives are addressed in Chapter 6.

Chapter 2
Theoretical Framework

CONSPECTUS: one of the most important aspects of a scientific work is the methodology. It determines the quality of the evidences that arise from experimentation and, therefore, it conditions the validity of the interpretations. For this reason it is very important to emphasize all the theoretical details of the methods, their fundamentals and their advantages and disadvantages. This is particularly relevant in the field of theoretical chemistry, were one can find an enormous quantity of methods with exotic acronyms (*e.g.* QM/MM-B3LYP/6-311++G**, QM-CCSD(T)/STO-3G or MM-FF99SB) that can be fuzzy for non-experts. Can one reach the same conclusions using two different methods? Is the same a potential energy profile than a free energy profile? What is better, implicit or explicit solvent simulations? In this chapter we depict the methods that we have used along the thesis, paying special attention not only to little details but also giving a broad vision to make it understandable in general.

2.1 The Relationship Between Simulations and Experiments

The aim of any computational simulation is to obtain, with a minimum computational cost, a good estimation of a physical observable that could be ideally measured by experiments. Among the observables one can find, for instance, the density of a system, the absorbance of solute or the thermodynamic parameters of a chemical reaction. It is important to remark the point of *minimum computational cost*, as in simulations one can make many approximations according to the degree of accuracy that is desired, and it should not be applied unnecessary expensive methods if one can obtain the same results, in less time, with a cheaper one. The other important point is that, ideally, the computational prediction should be *compared with experiments*, to validate the approximations and give insights into other observables not captured by the experimental techniques (assuming that the validation stands for those other observables). Given this, it is relevant to distinguish between what is usually measured in experiments and what is measured in computer simulations.

In the field of enzymatic research, most experimental measures are microscopic averages of a macroscopic state,[a] made over a large number of molecules –or *ensemble of molecules*– and also over time. This is the case, for instance, of X-ray crystallography, which gives a static vision –time average– of the atomic positions of several enzymes that are enclosed in a crystal.[97] At a given time, the enzymes of the crystal will be in slightly different conformations according to a particular distribution, but the diffraction of X-rays will integrate all those conformations into a single diffraction pattern, leading to an average structure. This average structure, then, will be also averaged over the time in which the X-ray experiment collects the diffraction data.

In computer simulations, on the contrary, the system is commonly composed by a single "molecule" (*e.g.* one enzyme, not an ensemble), and thus a single measure on a given structure is not enough to obtain meaningful results. Rather, it is necessary to explore the conformational space of the molecule in order to make averages and compare with experiments. This means that, instead of making a single measure that integrate N-states, it is necessary to make N-measures (one for each state) and then combine them to obtain the average result. How to combine these single measures and how many one needs to obtain a good estimation of the experimental result is the critical issue.

a) A macroscopic state, or macrostate, is a physical state that is thermodynamically defined by certain macroscopic variables, such as the total number of particles (N), the volume (V) or the temperature of the system (T).

2.2 Statistical Mechanics: the Bridge Between Macro and Micro

An entire branch of physics, called statistical mechanics, is devoted to the connection between macroscopic and microscopic properties. It is based on a series of postulates that define their mathematical construction, one of which states the fact that a macroscopic observable is equal to the *ensemble average* of the *M* microstates that are compatible with it:[98]

$$A_{macro} = \sum_{i=1}^{M} p_i A_{i,micro} \tag{1}$$

where the A_{macro} is the macroscopic value, p_i is the probability weight of microstate i and $A_{i,micro}$ is the microscopic value of microstate i.[b] Therefore, there are two challenges for obtaining a good estimation of a macroscopic observable: (i) find out the "p_i" probability weights that connect both worlds, and (ii) compute accurate $A_{i,micro}$ values for each microstate. The first point depends entirely on statistical mechanics, and can be approached with different *statistical ensembles* that are equivalent between them, but that depending on the type of simulation ones are more convenient than others, as we will see in the following sections. The second depends on the expression used to determine the interactions between particles, and will be addressed in section 2.4.

2.2.1 One, two, three, four… the microcanonical ensemble

The easiest way to introduce statistical mechanics is starting with the microcanonical ensemble. In this ensemble, a macrostate is defined by a constant number of particles (N), a constant volume (V) and a constant energy (E), meaning that the system is isolated. According to these conditions, it is postulated that each of the *M* microstates is equally probable (*i.e.* $p_i = 1/M$), which is equal to say that transitions between microstates are purely random, and equation 1 becomes:

$$A_{macro} = \frac{1}{M} \sum_{i=1}^{M} A_{i,micro} \tag{2}$$

Hence, in the microcanonical ensemble it is enough to *count the total number of states*, *M*, in order to compute any macroscopic observable. This is easily seen in the typical example of N particles enclosed in an isolated box that has two compartments, say A and B (see Figure 2.1). This isolated system is, at a macroscopic level, defined by the volume of the box (V), the number of particles (N) and the energy of the system (E), but at a microscopic level there are several *microstates* that match with the macroscopic state. For instance, in the case of N=4, there are two microstates

b) This is only applies for mechanical properties, thermodynamic properties are indirectly derived from the classical equations of thermodynamics.

having all particles in one of the two compartments ($M_{4,0}$ = 1 for all in A, and $M_{0,4}$ = 1 for all in B), eight degenerate microstates having three particles in one compartment and one in the other ($M_{3,1}$ = 4 and $M_{1,3}$ = 4), and six degenerate microstates having two particles in each compartment ($M_{2,2}$ = 6). Therefore, the total number of states is M = 16.

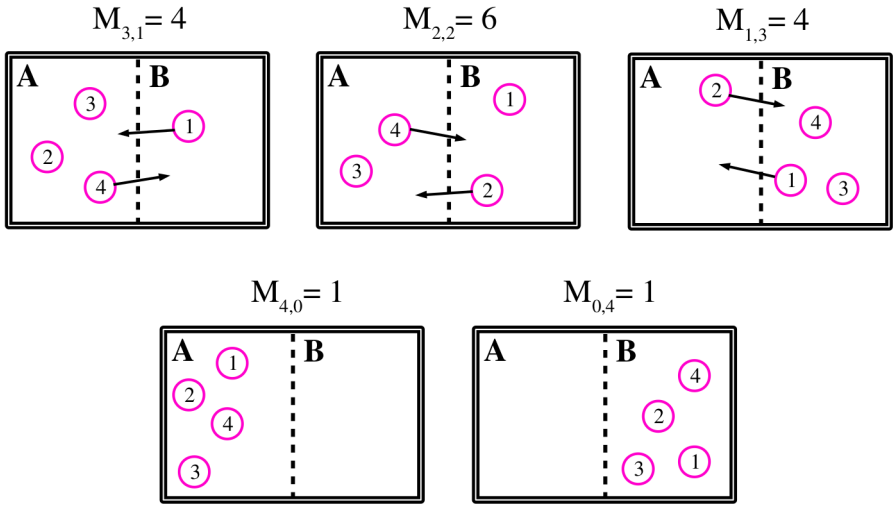


Figure 2.1- Five different microstates ($M_{4,0}$, $M_{3,1}$, $M_{2,2}$, $M_{1,3}$ and $M_{0,4}$) of a macrostate that is defined by four particles enclosed in an isolated box. Notice that, following a binomial distribution, $M_{3,1}$ and $M_{1,3}$ have four degenerate states each (the four particles alone in B and A, respectively) and $M_{2,2}$ has six degenerate states (all the possible pairs of particles alone in B, *i.e.* 1-2, 1-3, 1-4, 2-3, 2-4 and 3-4). The degeneracy of states is represented by arrows.

With the values derived above, then, what should one obtain when making a macroscopic measure of the density of particles in compartment A? According to equation 2, the result of such macroscopic measure will be the weighted average of all those possibilities:

$$\rho_{macro}=\frac{1}{16}\left[\left(M_{4,0}\cdot 4\right)+\left(M_{3,1}\cdot 3\right)+\left(M_{2,2}\cdot 2\right)+\left(M_{1,3}\cdot 1\right)+\left(M_{0,4}\cdot 0\right)\right]=2 \qquad (3)$$

Indeed, the most likely result is to find two particles in each compartment, following a maximum disorder principle.[c] This leads to one of the most famous laws in statistical mechanics, the Boltzmann equation, which relates the total number of microscopic states with a macroscopic thermodynamic quantity, the entropy:

$$S\left(N,V,E\right)\equiv k_B \ln\Omega\left(N,V,E\right) \qquad (4)$$

where k_B is the Boltzmann constant and $\Omega\left(N,V,E\right)$ is known as *the partition function*, which is basically the *M* defined above, *i.e.* the total number of states, but rewritten in a formal manner that specifies the compatibility of these microstates with the macroscopic restrictions on N, V and E.

c) It is important to note that a single measure could lead to any value, and most likely it will be different than $M_{2,2}$, so ensemble averages are crucial in order to compare with macroscopic experiments.

This equation states that the entropy increases with the increase of the number of available states, relating it with the ordinary view of "order and disorder", and is the fundamental equation from which one can derive any other thermodynamic property.

2.2.2 Turn on the heater: the canonical ensemble

While the microcanonical ensemble is useful for introducing statistical mechanics, other statistical ensembles that substitute the restriction on E by intensive variables –such as P or T– are more convenient for practical applications.[98] This is the case of the *canonical ensemble*, which restricts N, V and T, meaning that the isolated system is transformed into a system connected to a heat bath (also referred as a thermostat). Within this ensemble, the microstates can have different energies and therefore are not equally probable, but rather they follow a Boltzmann distribution:

$$p_i = \frac{\sum_{j=1}^{N_j} e^{-\beta E_{i,j}}}{\sum_{i=1}^{N_i}\sum_{j=1}^{N_j} e^{-\beta E_{i,j}}} = \frac{\sum_{j=1}^{N_j} e^{-\beta E_{i,j}}}{Z(N,V,T)} \tag{5}$$

where β is equal to $1/k_B \cdot T$, $E_{i,j}$ is the energy of a microstate i with j degenerate states, and $Z(N,V,T)$ is the canonical partition function. This equation highlights that states with a high energy will have a very low probability weight, and states with low energies will dominate the ensemble average of equation (1). In other words, in the canonical ensemble, the microscopic values of the most stable states account for the majority of the macroscopic observations. It also allows one to compare the likelihood of a state with respect to another, given as the difference of energies in the exponential function. For instance, imagine a system with two states, A and B, separated by an energy difference $\Delta E_{A \to B}$ =1 kcal·mol^{-1} (see Figure 2.2). According to this energy difference, is it straightforward to determine the population of each state by taking into account the ratio between the two probabilities –using equation (4)– and the number of N_j degenerate states:

$$\frac{p_B}{p_A} = \frac{\sum_{j=1}^{N_j(B)} e^{-\beta E_{B,j}}}{\sum_{j=1}^{N_j(A)} e^{-\beta E_{A,j}}} \tag{6}$$

which leads to a population of 86% for A and 14% for B considering that both states have the same number of N_j degenerate states. Nonetheless, the picture can change completely if the two states have a different number of degenerate states. For instance, in the case that B has 10 times more degenerate states than A, even with the energy difference disfavoring it, the resulting populations are

35% for A and 65% for B (see Figure 2.2). Therefore, *sampling is crucial* for exploring all those states, and the mere calculation of "single structures", even if they are the stabler minima, is not always enough for obtaining meaningful results.[d]

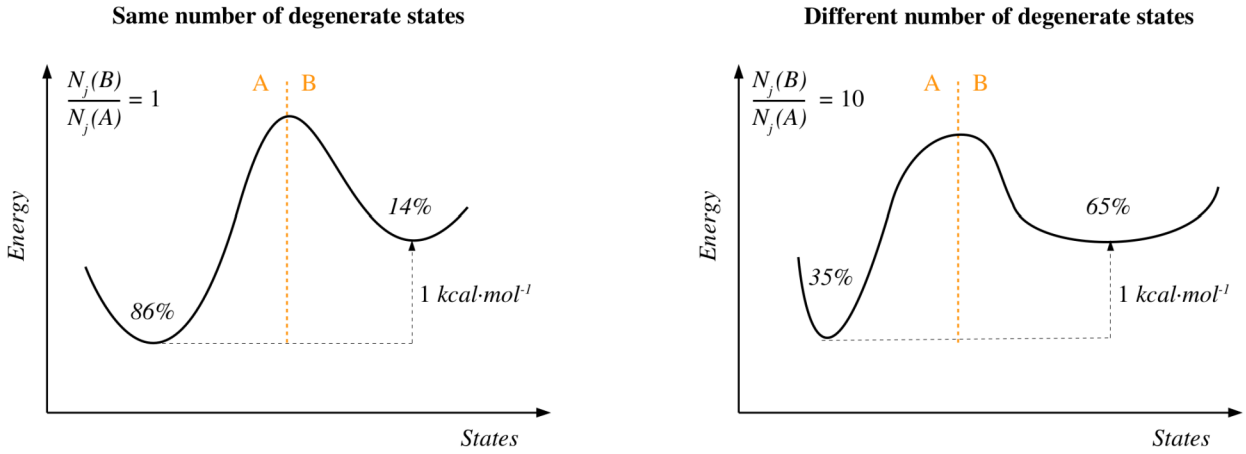


Figure 2.2- Schematic representation of two different states, A and B, that have an energy difference of 1 kcal·mol^{-1} in the canonical ensemble. At the left is represented a situation in which both states have the same number of N_j degenerate states, leading to a population of 86% for A and 14% B (at 300 K, *i.e.* $\beta = 0.6$ kcal·mol^{-1}), and at the right a situation in which basin B has 10 times more degenerate states than A, compensating the energy difference and leading to a population of 35% for A and 65% for B. Notice that state A in the right profile is narrower and state B is wider.

All the differences in populations can be translated into free energy terms with the master equation of the canonical ensemble:

$$F(N,V,T) \equiv -k_B T \ln Z(N,V,T) \tag{7}$$

which relates the canonical partition function $Z(N,V,T)$ with the Helmoltz free energy $F(N,V,T)$. Dividing the Z partition function in local Z_i partitions belonging to two states allows their comparison in terms of free energy. For instance, taking the above example, with state B having 10 times more degenerate states than A, it results that while the energy difference $\Delta E_{A\rightarrow B}$ is 1 kcal·mol^{-1}, the free energy difference $\Delta F_{A\rightarrow B}$ is -0.4 kcal·mol^{-1}. This means that the A→B transition is a favorable thermodynamic process, despite the fact that it is energetically unfavorable in terms of potential energy, highlighting *the importance of entropy*.

Free energy changes between different states along a chemical reaction is one of the main topics in computational simulations, and the principal challenge in computing them resides in the exploration of all the states of the system to construct the local partition functions.

d) Notice that for larger energy differences, *e.g.* $\Delta E_{A\rightarrow B}$ = 5 kcal·mol^{-1}, a considerable amount of degenerate states is necessary to compensate the effect of energy in the populations, and thus the "single structure" approach would be a reasonable approximation in such cases.

2.2.3 Cutting the partition function into pieces

One of the most common approaches in computational chemistry to evaluate partition functions is the *Rigid-Rotor Harmonic-Oscillator* approximation (RRHO).[99] This approximation basically "cuts" the partition function of one basin into translational, rotational, vibrational and electronic motions that are considered decoupled among them (*i.e.* they do not exchange energy), and so the partition function can be represented as a product of independent contributions:

$$Z = Z_{trans} Z_{rot} Z_{vib} Z_{elec} \tag{8}$$

These independent contributions are evaluated by solving ideal models such as a particle in a box (translational), a rigid rotor (rotational) and an harmonic oscillator (vibrational), according to the following expressions:

$$Z_{trans} = V \left(\frac{2\pi m}{\beta h^2} \right)^{3/2} \tag{9}$$

$$Z_{rot} = 8\pi^2 \left(\frac{2\pi}{\beta h^2} \right)^{3/2} \left(I_1 I_2 I_3 \right)^{1/2} \tag{10}$$

$$Z_{vib} = \prod_{i=1}^{3N-6} \frac{1}{1 - e^{-\frac{h\upsilon_i}{k_B T}}} \tag{11}$$

with m being the mass of the molecule, V the volume of the "box", I_i the principal moments of inertia and v_i the harmonic vibrational frequencies. Electronic degrees of freedom are usually neglected by considering a ground state approximation ($Z_{elec} = 1$). This approach allows to estimate the contribution of states *in a single basin* (*e.g.* a minima or a transition state) according to approximate solutions based on ideal models, with the advantage that it is very fast to evaluate such analytical formulas, but with two main disadvantages: (i) the ideal models are not always good enough for describing real systems, *e.g.* the harmonic approximation does not perform well for anharmonic or wide basins; and (ii) it does not consider the contributions of different *conformational states*.

Even assuming that the first disadvantage is not a problematic issue, that is usually not true for enzymatic systems, the second one is usually critical for any system (see Figure 2.3). In order to address it, one should explore the conformational space of the system, compute the RRHO approximation for each conformational basin, and then combine all the results together. While this approach can be somehow straightforward for very simple systems, such as small organic molecules where the exploration can be done "manually" (*e.g.* the different conformational states of *n*-butane with respect to its C-C-C-C dihedral angle), for complex systems such as enzymes, with a large number of degrees of freedom, it is a difficult task that requires sophisticated techniques to directly explore the space of interest.

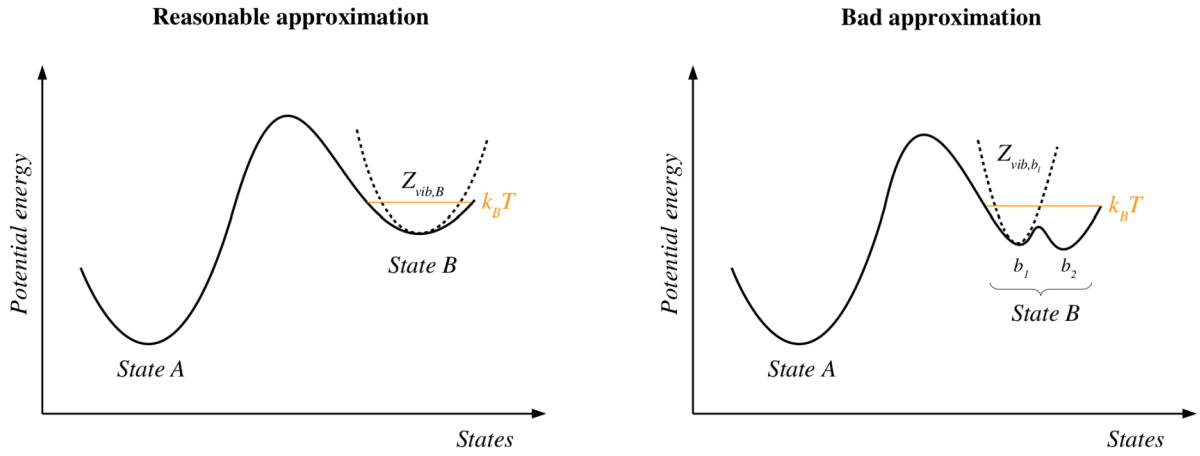


Figure 2.3- Schematic representation of the harmonic approximation in a state B for two different situations: (left) simple basin that is nearly harmonic, leading to a reasonable approximation for a given k_BT, and (right) basin with two low-lying minima, b_1 and b_2, for which the harmonic approximation leads to a severe underestimation of the available states if state b_2 is not explored and considered in the RRHO. The orange line indicates the energy "available" at a given temperature.

2.2.4 Direct exploration of the chemical space

An alternative approach to the RRHO approximation is to explore directly the chemical space, *i.e.* instead of estimating the available states of a basin from a single structure and model solutions, the available states are calculated one by one. At this point it is more convenient to leave the quantum state discretization that has been used up to now, and express the Boltzmann distribution as a continuous function of the phase-space $\vec{x}=(\vec{r},\vec{p})$ (*i.e.* classical approximation):

$$p(\vec{x})=\frac{e^{-\beta E(\vec{x})}}{\int d\vec{x}\, e^{-\beta E(\vec{x})}} \tag{12}$$

which emphasizes that the partition function, found in the denominator, represents a *hypervolume* of phase and space that has to be explored.[100] This requires the use of *sampling techniques* to visit all the energetically available states, which is of course much more expensive than the RRHO approximation in terms of computational time, but has also the advantage that it does not assume any ideal behavior (*e.g.* the shape of the basins is not assumed to be harmonic and the different terms are not assumed to be decoupled).[e]

Among the sampling techniques we find Monte Carlo (MC) and Molecular Dynamics (MD) approaches. The first one is based on the generation of new atomic configurations according to random values that follow a particular distribution –*e.g.* Boltzmann– and that are restricted by an en-

e) The only assumption that is taken is that the particles must be governed by classical statistical mechanics, which means that quantum effects are not included (*e.g.* zero-point energy contributions), but that can be added *a posteriori* with specific corrections.

ergy threshold. It is, therefore, an ensemble averaging approach. The second, instead, is based on the use of thermal energy to explore the chemical space *along time*, obtaining the new atomic configurations by solving a set of equations of motion. In this case, MD can be seen as a time averaging approach. This leads to the ergodic hypothesis, which states that a macroscopic observable can be obtained by the time average of all microscopic values of the property:

$$A_{macro} = \sum_{i=1}^{M} p_i A_{i,micro} = \lim_{t_m \to \infty} \frac{1}{t_m} \int_{t_0}^{t_m} dt\, A_{i,micro}(t) \tag{13}$$

In other words, the ergodic hypothesis states that by running a long enough simulation ($t_m \to \infty$), it is possible to recover the ensemble average with a time average. This inherently assumes that within the time of the simulation, the system is able to explore all the regions of the space (*i.e.* the system is "ergodic"), which is a matter of debate depending on the system and also on the simulation conditions.[100,101] An alternative option to time averaging is to run a number of short-time MD simulations and collect them in *ensembles of trajectories*, defining a kind of time and ensemble average that is more appropriate to explore the chemical space.[102]

2.3 Moving the World: Molecular Dynamics

As introduced above, Molecular dynamics (MD) is a powerful technique that allows to explore the phase-space of a system and obtain relevant statistical information. Given a set of initial atomic coordinates, all one needs is a mathematical expression to describe the *energy of the system* and then integrate a set of *equations of motion* for obtaining the trajectory of each particle along time. The first point depends on how particle interactions are handled, and will be addressed in Section 2.4. The last point depends on how particles are treated, leading to different types of classical and quantum MD approaches depending if they are considered as point particles or as waves of probability. Here we first outline the classical propagation of particles, and subsequently we derive the equations of *ab initio* MD from the beginning.

2.3.1 Classical propagation of particles

Point particles can be properly described by classical mechanics, which allow to define their trajectory $\vec{r}(t)$ using their position, momentum and acceleration:

$$\vec{r}(t_1) = \vec{r}(t_0) + \int_{t_0}^{t_1} dt\, \frac{\vec{p}(t)}{m} \quad \text{and} \quad \vec{p}(t_1) = \vec{p}(t_0) + m \int_{t_0}^{t_1} dt\, \vec{a}(t) \tag{14}$$

These equations of motion, in practice, have to be solved approximately using numerical algorithms –*e.g.* Verlet[103], velocity Verlet or leapfrog[99]– leading to expressions like:

$$\vec{r}(t+\Delta t)=2\vec{r}(t)-\vec{r}(t-\Delta t)+\vec{a}(t)\Delta t^2 \tag{15}$$

where Δt is the *time step* parameter that is used to integrate the equations of motion, and is key for the goodness of the numerical approximation. As a rule of thumb, to ensure that the system is adiabatic the time step is chosen as one tenth of the fastest molecular motion, which is usually dictated by the O-H bond stretching. The acceleration can be updated according to the second Newton law, which can be derived as a force that arises from the potential energy of the system:

$$m\frac{\vec{\partial}^2}{\partial t^2}\vec{r}(t)=-\frac{\partial}{\partial \vec{r}}E_{pot}(\vec{r}) \tag{16}$$

Therefore, using equations (15) and (16) one only needs two ingredients in order to propagate classical particles: (i) define a set of initial conditions, and (ii) find a mathematical expression for the energy of the system. While these equations can be applied satisfactorily to describe the evolution of nuclei treated as point particles –which is an excellent approximation for most systems of interest–, they fail to describe the evolution of light particles such as electrons, for which more general theories have to be considered.

2.3.2 From the beginning: quantum propagation of particles

If one aims to study systems characterized by significant electronic reorganizations, such as is the case of any chemical reaction, electrons have to be considered explicitly in the simulations. This can be approached in several manners, and different types of MD techniques arise depending on how do they describe the motion of electrons and nuclei. These techniques are based on Quantum Mechanics (QM), for which the *time dependent Schrödinger equation* has to be solved in order to follow the evolution of the system:[104]

$$i\hbar\frac{\partial}{\partial t}\psi(\{\vec{r}\}_n,\{\vec{R}\}_N;t)=\hat{H}\psi(\{\vec{r}\}_n,\{\vec{R}\}_N;t) \tag{17}$$

where "*i*" is the imaginary number, $\hbar$ is the Planck constant, ψ is the time dependent wave function describing "*n*" electrons and "*N*" nuclei, and $\hat{H}$ is the Hamiltonian of the system. The wave function is a mathematical entity from which one can derive –among other properties– the probability of finding a particle at time "*t*" in a particular region of the space, so it can be seen as a function of *where particles are*. Notice that here both electrons and nuclei are explicitly treated inside the wave function, *i.e.* nuclei are not anymore point particles. The Hamiltonian is an operator of the wave function that outputs the energy of the system. Depending on the property of interest it can be

simple or complex. If relativistic effects are not important, as it is the case of most chemical systems, the Hamiltonian can be written as follows:

$$\hat{H} = -\frac{1}{2}\sum_{I=1}^{N}\frac{\nabla_I^2}{M_I} - \frac{1}{2}\sum_{i=1}^{n}\nabla_i^2 - \sum_{i=1}^{n}\sum_{I=1}^{N}\frac{Z_I}{|\vec{r}_i - \vec{R}_I|} + \sum_{I=1}^{N-1}\sum_{J=I+1}^{N}\frac{Z_I Z_J}{|\vec{R}_I - \vec{R}_J|} + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{1}{|\vec{r}_i - \vec{r}_j|} \quad (18)$$

where the first and second terms correspond, respectively, to the kinetic energy of nuclei and electrons, the third and fourth terms to the nuclei-electron and nuclei-nuclei electrostatic potential, and the fifth term to the electron-electron repulsion.

At this point we aim to simplify equation (17) by dividing the nuclear and electronic wave function into two independent components: one depending uniquely on nuclei (nuclear wave function) and another depending uniquely on electrons (electronic wave function). This can be achieved under the Born-Oppenheimer approximation, which assumes that electrons and nuclei wave functions can be decoupled.[f] Within this approximation, the wave function is factorized in electronic and nuclear terms together with a phase-factor, leading to:

$$\psi(\{\vec{r}\}_n, \{\vec{R}\}_N; t) \approx \psi(\{\vec{r}\}_n; t)\,\chi(\{\vec{R}\}_N; t)\exp\left[\frac{i}{\hbar}\int_{t_0}^{t}\partial t'\,\widetilde{E}_e(t')\right] \quad (19)$$

This approximation, therefore, makes both electronic and nuclear wave functions independent one from the other, which facilitates the resolution of the Schrödinger equation. At this point, if we introduce the equations (19) and (18) in (17), splitting the nuclear and electronic degrees of freedom and integrating with χ^* and ψ^*, we obtain the following equations:

$$i\hbar\frac{\partial}{\partial t}\psi(\{\vec{r}\}_n; t) = -\frac{1}{2}\sum_{i=1}^{n}\nabla_i^2\psi(\{\vec{r}_i\}_n; t) + \left\{\int \partial\vec{R}\,\chi^*(\{\vec{R}\}_N; t)\hat{V}_{N-e}\chi(\{\vec{R}\}_N; t)\right\}\cdot\psi(\{\vec{r}\}_n; t) \quad (20)$$

$$i\hbar\frac{\partial}{\partial t}\chi(\{\vec{R}\}_N; t) = -\frac{1}{2}\sum_{I=1}^{N}\frac{\nabla_I^2}{M_I}\chi(\{\vec{R}_I\}_N; t) + \left\{\int \partial\vec{r}\,\psi^*(\{\vec{r}\}_n; t)\hat{H}_{elec}\psi(\{\vec{r}\}_n; t)\right\}\cdot\chi(\{\vec{R}\}_N; t) \quad (21)$$

These pair of coupled equations are the heart of the time-dependent self-consistent field (TD-SCF) method developed by Paul Dirac,[105] and they allow to describe the time evolution of both electron and nuclei wave functions according to a mean-field description of quantum dynamics.[104] The application of these equations, however, is limited to systems with few degrees of freedom, and further simplifications on both of them are necessary for treating larger systems.

f) It also assumes the adiabatic approximation between different electronic states, both in diagonal and non-diagonal terms, which is only acceptable in regions of the chemical space where there is enough energy gap between the states.

2.3.3 Ehrenfest and Born-Oppenheimer molecular dynamics

One of the obvious simplifications of the TDSCF method is to treat nuclei as point particles. This approximation gives rise to what are known as *ab initio molecular dynamics* techniques (AIMD). In these techniques, therefore, the nuclear wave functions of equations (20) and (21) are substituted by Dirac delta functions –*i.e.* points–, allowing to treat them according to their classical position and momentum. This simplification was first addressed by Paul Ehrenfest,[106] who obtained the following equations that define the *Ehrenfest molecular dynamics* (EMD) technique:

$$i\hbar\frac{\partial}{\partial t}\psi(\{\vec{r}\}_n;t)=-\frac{1}{2}\sum_{i=1}^{n}\nabla_i^2\psi(\{\vec{r}_i\}_n;t)+V_{N-e}(\{\vec{r}\}_n,\{\vec{R}\}_N)\cdot\psi(\{\vec{r}\}_n;t) \tag{22}$$

$$M_I\frac{\partial^2}{\partial t^2}\vec{R}_I(t)=-\nabla_I\int\partial\vec{r}\,\psi^*(\{\vec{r}\}_n;t)\hat{H}_{elec}\psi(\{\vec{r}\}_n;t) \tag{23}$$

Notice that equation (23) is essentially the same as the second Newton law –equation (16)– with the difference that in this case the potential energy is defined *ab initio* from the electronic wave function. The EMD method starts its dynamic cycle converging the electronic wave function to obtain its initial form, and then it propagates it *unitarily* –*i.e.* preserving its norm and the orthonormality of its orbitals– together with the classical nuclei that move under the effective field of electrons.[104]

This approach has the computational advantage that it does not require to converge the wave function at each step of the simulation (it evolves dynamically), which can be an arduous task depending on the system. At the same time, however, it has the disadvantage that it requires a very short time step –of the order of 1 a.u; 0.0241888 fs– in order to describe adiabatically the electronic motions.[g] An alternative approach to the EMD method is the *Born-Oppenheimer molecular dynamics* (BOMD) method, which does not consider the dynamic evolution of electrons given in equation (22), but rather it employs the stationary Schrödinger equation to obtain the electronic information, converging the wave function after each nuclear movement under Lagrangian constraints:

$$0=-\{\int\partial\vec{r}\,\psi^*(\{\vec{r}\}_n;t)\hat{H}_{elec}\psi(\{\vec{r}\}_n;t)\}+\sum_{i,j}^{n}\Lambda_{ij}(\langle\phi_i|\phi_j\rangle-\delta_{ij}) \tag{24}$$

$$M_I\frac{\partial^2}{\partial t^2}\vec{R}_I(t)=-\nabla_I\,min_\psi\{\int\partial\vec{r}\,\psi^*(\{\vec{r}\}_n;t)\hat{H}_{elec}\psi(\{\vec{r}\}_n;t)\} \tag{25}$$

where ϕ_i denote the orbitals used to construct the wave function and Λ_{ij} are the Lagrange multipliers. The disadvantage of BOMD is that it needs to converge continuously the wave function, but

g) The EMD method has also the advantage that it can include non-adiabatic transitions between different electronic states, although here we only consider ground-state wave functions.

has the advantage that the time step of the simulation can be taken much larger than in EMD (*c.a.* 100 times). This is possible given that, as the dynamics of electrons is not explicitly considered, the fastest movement is dictated by the nuclear O-H bond stretching.

2.3.4 Stay cold: Car-Parrinello molecular dynamics

The Car-Parrinello molecular dynamics (CPMD)[107] is a method that aims to combine the advantages of the two AIMD methods outlined above, *i.e.* to use a dynamic propagation of electrons (EMD) and at the same time use a large enough time step (BOMD). This is achieved by the use of a *fictitious mass* that is assigned to the electronic degrees of freedom, which can be tuned to make electrons "heavier" and, therefore, make possible to use a larger time step. With this approach, electrons acquire a classical kinetic energy and their dynamics, together with the one of nuclei, can be propagated using the following equations of motion:

$$\mu_{elec}\frac{\partial^2\phi_i(t)}{\partial t^2}=-\frac{\partial}{\partial\phi_i^*}\left(\langle\psi|\hat{H}_{elec}|\psi\rangle-\sum_{i,j}^{n}\Lambda_{ij}\left(\langle\phi_i|\phi_j\rangle-\delta_{ij}\right)\right) \tag{26}$$

$$M_I\frac{\partial^2}{\partial t^2}\vec{R}_I(t)=-\nabla_I\int\partial\vec{r}\,\psi^*(\{\vec{r}\}_n;t)\,\hat{H}_{elec}\,\psi(\{\vec{r}\}_n;t) \tag{27}$$

Notice that the difference between equation (27) and the equivalent equation in BOMD –equation (25)– lays on the fact that in the last one it is necessary to minimize the wave function, which is certainly a disadvantage compared to CPMD. This can be written altogether in a Lagrangian form as:

$$L_{CP}=\frac{1}{2}\sum_{I=1}^{N}M_I\frac{\partial^2\vec{R}_I(t)}{\partial t^2}+\frac{1}{2}\mu_{elec}\sum_{i=1}^{n}\langle\frac{\partial}{\partial t}\phi_i(\vec{r},t)|\frac{\partial}{\partial t}\phi_i(\vec{r},t)\rangle-\langle\psi_0|\hat{H}_{elec}|\psi_0\rangle+\sum_{j=1}^{n}\Lambda_{ij}\phi_j \tag{28}$$

The CPMD method relies on the assumption that there is minimal energy transfer between electrons and nuclei. In other words, electrons must remain "cold", they should not heat up with time, otherwise it means that their dynamics is not adiabatic. The degree of adiabaticity depends on the minimum and maximum electronic frequencies:

$$\omega_{elec}^{min}\propto\left(\frac{E_{gap}}{\mu_{elec}}\right)^{1/2} \tag{29}$$

$$\omega_{elec}^{max}\propto\left(\frac{E_{cut}}{\mu_{elec}}\right)^{1/2} \tag{30}$$

with E_{gap} being the energy gap between the HOMO and LUMO orbitals and E_{cut} the cutoff energy for a plane wave basis set expansion (see Section 2.4.3 below). The maximum frequency, in turn, can

be related with the maximum time step of the simulation, which is proportional to the inverse of the fastest frequency:

$$\Delta t^{max} \propto \left(\frac{\mu_{elec}}{E_{cut}} \right)^{1/2} \tag{31}$$

Therefore, the CPMD method can be seen as a type of "classical" approximation to EMD in which electronic degrees of freedom are weighted by a fictitious mass, allowing to lengthen the time step of the simulation according to a desired degree of adiabaticity. It can also be seen as a type of BOMD in which electrons fluctuate around the ground-state surface according to their temperature, which at the same time is related with the fictitious mass (notice that in the limit of $\mu_{elec} \rightarrow 0$, equation (26) reduces to equation (24), the expression of BOMD). The advantage of the CPMD method is that, for systems with a large E_{gap}, the time step can be chosen 5-10 times larger than in EMD when 500-1000 a.u fictitious masses are assigned to electrons.[104] The disadvantage is that it can not be applied to systems with a small E_{gap} , such as it is the case of metallic systems.

2.4 On the Definition of Energy Functions

In the previous section we have seen different variants of MD to evolve particles in time, which can be used to explore the phase-space of a system. These techniques require the definition of an energy function for obtaining the forces that are used to compute particle accelerations and propagate the equations of motion. Different degrees of approximations exist for such energy functions, and we will address here three general schemes.

In the simplest scheme, called Molecular Mechanics (MM), nuclei are treated as point particles and electrons are not explicitly considered in the simulation, but rather they are taken into account as a force field of *predefined potentials* that are used to determine the interactions between the nuclei. In more sophisticated schemes, based on Quantum Mechanics (QM), both nuclei and electrons *are explicitly considered* in the simulations, and thus there is no need to predefine any parameter from the beginning.

Finally, combination of both QM and MM schemes are also possible, leading to hybrid QM/MM methods. In this section we will review the three principal schemes, and for the QM schemes we will limit the discussion on the method used along this thesis, Density Functional Theory (DFT), albeit there is a myriad of other methods that can be used to calculate the energy of a polyelectronic system.

2.4.1 Cat's allergy: molecular mechanics

In molecular mechanics (MM) schemes nuclei are treated as point particles and electrons are not explicitly considered in the simulation. Electrons, however, are the fundamental particles that determine the structural features of a chemical compound (*e.g.* bond distances, angles and dihedrals), and thus they are crucial for understanding any type of chemical process. Therefore, they have to be somehow included in the description of the system. This is achieved by the use of simple force fields that include two types of interactions: (i) bond, angle and torsion *bonded interactions*, with the two first described by spring constants and the last by periodic functions; and (ii) van der Waals (VdW) and electrostatic *non-bonded interactions*, described respectively by Lennard-Jones (L-J) and Coulomb potentials. A typical force field has the following simplified expression for describing the potential energy of a system:[108]

$$E_{pot}^{FF} = \sum_{bonds} k_d (d - d_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2 + \sum_{torsions} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}^{atoms} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}} + \frac{q_i q_j}{\epsilon d_{ij}} \right) \quad (32)$$

were the first three terms account for the bonded interactions and the last term for the non-bonded ones. The most important aspect of a force field, apart from the mathematical expression, are the parameters that define it: the force constants and equilibrium values for bonds and angles (k_d, d_{eq}, k_θ and θ_{eq}, respectively), the number of "*n*" torsional energy barriers and their height (V_n), the parameters A_{ij} and B_{ij} of the L-J potential and the q_i charges of all atoms.[109] These parameters are defined for different *atom types*, which are "labels" on atoms necessary to take into account the different electronic properties of a given atom in different chemical states (*e.g.* a carbon can be *sp*, sp^2 or sp^3, and so the parameters for a C-C bond, for instance, should not be the same than the ones of a C=C bond, albeit both are bonds between two carbon atoms).

Typical force fields for enzymatic systems are FF99SB[110], CHARMM27[111] or OPLS,[112] and for water molecules the most used force field is TIP3P[113]. The main advantage of MM simulations is that the force evaluation of equation (32) can be done analytically, and so it is very fast to compute them even for systems with a large number of atoms, reaching time scales of hundreds of microseconds by *brute force* MD.[114–116] Some of the main disadvantages of MM are: (i) the accuracy of the results strongly depends on the force field parameters, which can be good or not depending on how they have been generated and for which purpose;[117] (ii) it is not possible to carry out simulations when parameters are not available for a given molecule; (iii) it does not allow to describe chemical reactions nor charge transfer reorganizations, given that electrons are not explicitly considered in the simulation. Therefore, if one is interested in exploring the conformational motions of large sys-

tems for which reliable parameters are available and that do not involve electronic effects, MM is the most appropriate method.

2.4.2 Cloud of electrons: density functional theory

In case of having to describe electrons, as also explained in the MD section above, QM based techniques are necessary. Given a set of fixed nuclear coordinates, the energy of a polyatomic system can be obtained by solving the stationary Schrödinger equation, which require the definition of a Hamiltonian and a wave function:

$$\hat{H}\,\psi(\vec{r}_1,\vec{r}_2...\vec{r}_n)=E\,\psi(\vec{r}_1,\vec{r}_2...\vec{r}_n) \tag{33}$$

with $\psi(\vec{r}_1,\vec{r}_2...\vec{r}_n)$ depending on the spatial coordinates of the "*n*" electrons that compose the system (neglecting spin). The solution to this equation can be approached by what are known as *wave function methods*, such as the ones derived from Hartree-Fock (HF) and post Hartree-Fock theories,[98] based on different simplifications of either the Hamiltonian and/or the wave function. These methods are purely derived from first principles and can be very accurate depending on their ability to include electron correlation, such as in the case of Møller-Plesset perturbational methods (*e.g.* MP2 and MP4) or coupled cluster methods (*e.g.* CCSD and CCSD(T)). The disadvantage of these methods is that they scale from N^4 to N^{10} depending on their degree of accuracy,[99] with *N* being the number of basis functions used to expand the wave function. This has been a bottleneck in quantum chemistry for many years.

Density functional theory (DFT)[118,119] is a method that is aimed to overcome such bottleneck with a compromise on the accuracy of the results. Walter Kohn, one of his fathers, said in his Nobel lecture that DFT "*has been most useful for systems of very many electrons where wave function methods encounter and are stopped by the exponential wall*",[120] highlighting its notorious importance. Following the Hohenberg and Kohn theorems,[118] any property of a stationary state can be described –in an *exact* form– by means of an observable that contains less information than the wave function: *the electronic density*. Indeed, the electron density does not depend on 3*n* spatial coordinates as the wave function, but rather only on three:

$$\rho(\vec{r}_1)=N\int d\vec{r}_2...\int d\vec{r}_n \quad \psi^*(\vec{r}_1,\vec{r}_2...\vec{r}_n)\psi(\vec{r}_1,\vec{r}_2...\vec{r}_n) \tag{34}$$

with *N* being a normalization factor. In the framework of DFT, the energy of a polyatomic system is a functional of the density (*i.e.* a function that, for a given density, returns an energetic value):

$$E[\rho(\vec{r})]=T_S[\rho(\vec{r})]+V_{N-e}[\rho(\vec{r})]+V_{e-e}[\rho(\vec{r})]+E_{XC}[\rho(\vec{r})] \tag{35}$$

where the first term accounts for the kinetic energy of a fictitious system of non-interacting electrons,[h] the second to the nuclear-electron attraction, the third to the classical[i] electron-electron repulsion, and the last term accounts for *corrections* on (i) the assumption of the non-interacting nature in T_s, and (ii) the exchange, the correlation and the self-interaction energy of electrons that arise from the classical electrostatic potential $V_{e\text{-}e}$. The total energy of the system can be obtained by finding the orbitals that minimize it according to the Kohn-Sham secular equation[119]:

$$\left[-\frac{1}{2}\nabla^2+v_{eff}(\vec{r})\right]\phi_i(\vec{r})=\varepsilon_i\phi_i(\vec{r}) \tag{36}$$

where the orbitals are defined as a linear combination of known $\chi_j(\vec{r})$ functions that can represent the electron density:

$$\phi_i(\vec{r})=\sum_{j=1}^{n} c_{ij}\chi_j(\vec{r}) \tag{37}$$

$$\rho(\vec{r})=\sum_{i=1}^{n_{occ}}\langle\phi_i(\vec{r})|\phi_i(\vec{r})\rangle \tag{38}$$

and $v_{eff}(\vec{r})$ is the effective potential that include the three last terms of equation (34):

$$v_{eff}(\vec{r})=\sum_{I=1}^{N}\frac{Z_I}{|\vec{r}-\vec{R}_I|}+\int d\vec{r}'\frac{\rho(\vec{r}')}{|\vec{r}-\vec{r}'|}+v_{XC}(\vec{r}) \tag{39}$$

The $v_{XC}(\vec{r})$ exchange and correlation potential is the most important aspect of any DFT method, as it accounts for all the corrections exposed above. Nonetheless, the exact expression of its functional form *is unknown*, and has to be constructed assuming different degrees of approximations. The most simple one is called local density approximation (LDA), which only takes into account the local density to handle with the corrections, *i.e.* it assumes that exchange and correlation effects are local and only depend on the electron density at each point.[98] The expression for the exchange functional is:

$$E_X^{LDA}[\rho(\vec{r})]=-\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int\rho(\vec{r})^{4/3}d\vec{r} \tag{40}$$

and the expression for the correlation part was derived from Monte Carlo simulations.[121] While this type of approximation can lead to reasonable good geometries, it does not have a good energetic performance.[120] A step up in accuracy is to consider gradient corrections, which take into account the density at each point and how it varies around it, leading to the *generalized gradient approximations* (GGA). In this case, the exchange functional has the following form:

h) A necessary step to derive the Kohn-Sham equations is to assume a fictitious system of non-interacting electrons under the field of an external potential.

i) Classical here refers to the Coulomb potential, which does not include electron exchange nor correlation.

$$E_X^{GGA}[\rho(\vec{r}),x]=\int \rho(\vec{r})^{4/3}F(x)d\vec{r} \tag{41}$$

where $x=|\nabla\rho(\vec{r})|/\rho(\vec{r})^{4/3}$ and *F(x)* is a function that is chosen to obey a gradient expansion of the density.[120] The Pedrew-Erzenhof-Bruke "PBE" functional[122] used along this thesis is a type of GGA exchange and correlation functional that adopts the following expression:

$$E_X^{PBE}[\rho(\vec{r})]=-\int \rho(\vec{r})^{4/3}\left[\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}+\frac{\mu s^2}{1+\mu s^2/\kappa}\right]d\vec{r} \tag{42}$$

where μ and κ are parameters derived *ab intio* (other functionals, such as Becke,[123] use empirical coefficients). These type of functionals lead to better geometries and charge densities with respect to LDA, as well as they improve thermochemical predictions and they are able to deal with systems involving hydrogen bonds.[98] Nonetheless, it is well known that GGA functionals do not perform well in the description of VdW forces and also that they underestimate HOMO-LUMO energy gaps, with the inherent effects in reactivity (*e.g.* S_N2 reactions[124]). More accurate approximations include the kinetic energy density (meta-GGA) or certain amount of exact Hartree-Fock exchange (hybrid functionals).

The main advantage of DFT is that it allows to overcome the computational bottleneck associated to wave function methods, as it formally scales cubically with respect to the number of basis functions (N^3_{basis}), an order of magnitude lower than HF (N^4_{basis}) and two with respect to MP2 (N^5_{basis}).[99] This makes possible to treat *ab initio* systems with large number of atoms and electrons. The drawback of DFT is that its results strongly depend on the proper choice of the exchange and correlation functional, and so it is necessary to test whether a given functional allows or not to describe properly the interactions of a particular system. In this regard, the PBE functional used in this thesis has been previously tested in our group, with successful results in the study of conformations and reactions in carbohydrate-active enzymes.[93,125–128]

2.4.3 Plane waves and atomic basis sets

As explained in the previous section, the electronic density of a system can be expressed in terms of orbitals that are linear combinations of known functions. This allows to solve numerically the Kohn-Sham equations. These functions can be of different types, with the most common ones centered on nuclei, such as the simple three Gaussian Slater-type basis set (STO-3G):[129]

$$\phi_{STO-3G}(\vec{r}-\vec{R})=c_1f_1e^{-\alpha_1(\vec{r}-\vec{R})^2}+c_2f_2e^{-\alpha_2(\vec{r}-\vec{R})^2}+c_3f_3e^{-\alpha_3(\vec{r}-\vec{R})^2} \tag{43}$$

where $\vec{R}$ indicates the position of the nucleus and the α and f_i are factors of the Gaussian functions parametrized to fit a Slater function.

An alternative option to atomic-centered basis sets is to expand the orbitals through a defined space using *plane waves* (PW):[104]

$$f_{\vec{G}}^{PW}(\vec{r}) = \frac{1}{\sqrt{\Omega}} \exp[i\vec{G}\vec{r}] \tag{44}$$

where Ω represents the volume of the cell that limits the space, and $\vec{G}$ is a vector in the reciprocal space that is defined as:

$$\vec{G} = i \cdot \frac{2\pi}{L_x} \cdot \vec{x} + j \cdot \frac{2\pi}{L_y} \cdot \vec{y} + k \cdot \frac{2\pi}{L_z} \cdot \vec{z} \tag{45}$$

The use of PW as a linear combination, therefore, takes the form:

$$\phi_i(\vec{r}, \vec{k}) = \frac{1}{\sqrt{\Omega}} \sum_{\vec{G}} c_i(\vec{G}, \vec{k}) \exp[i(\vec{G} + \vec{k})\vec{r}] \tag{46}$$

where $\vec{k}$ are vectors in the Brioullin zone. The size of the basis set in these linear combinations depends on an energy cutoff E_{cut} and the size of the cell:

$$N_{PW} = \frac{1}{2\pi^2} \Omega E_{cut}^{3/2} \tag{47}$$

The advantage of PW in front of the ones centered on nuclei are: (i) given that they are expanded through the whole space, they are perfectly unbiased, and thus there is no basis set superposition error; (ii) the Pulay forces vanish even for finite basis; (iii) they allow to describe the density of the system with a linear dependence with respect to the size of the system, and not a quadratic dependence; (iv) the improvement of the basis set is straightforward, it is only necessary to increment E_{cut}; (v) the derivatives in the real space are simple multiplications in the reciprocal space, allowing to optimize certain calculations together with the use of fast Fourier transforms (FFTs) to connect both spaces.

The main disadvantages are: (i) molecules have to be confined in the cell, and simulations become inestable if they approach the cell boundaries. This, in practice, can be addressed easily by using large enough cells, but at the cost of increasing the number of PW in the simulation, which means to increase the computational cost of the simulation; (ii) the use of hybrid functionals is terribly inefficient, as the evaluation of the Hartree-Fock exchange potential has to be performed in the Fourier space, while forces are computed in the real space. This makes necessary to perform several FFTs that hamper the efficiency of hybrid functionals, being >30 times more expensive than the GGA ones;[104] (iii) the use of *pseudopotentials* is necessary for the description of core electrons, as their explicit consideration would require to use a so large basis set that the calculations would become inefficient. The last point, however, is not a problematic issue as core electrons usually do

not play significative roles in chemistry, and moreover these pseudopotentials can include relativistic and other types of corrections.

2.4.4 Not Quantum, not Classical: hybrid QM/MM methods

Up to now we have seen that molecular mechanics (MM) schemes allow to describe a large number of atoms –hundreds of thousands– but rely on simple energy potentials and do not explicitly consider electrons, and thus they do not allow to describe electronic reorganizations. In contrast, we have seen that quantum based methods (QM) such as DFT are able to describe the energy of a system *ab initio*, but they can not handle more than –typically– few hundreds of atoms. These limitations highlight the problem of studying enzymatic reactions, which are very large systems (unfeasible with QM) in which the electronic structure has to be taken into account (impossible with MM).

Fortunately, nowadays it is possible to take the advantage of the strenghts of both QM and MM methods with the so-called *hybrid QM/MM methods*, originally developed by Warshel and Levitt.[45,130] These methods allow to describe one part of the system at a QM level –defined in the region where the chemical reaction takes place– and the rest of the system is treated at a MM level (see Figure 2.4). The atoms belonging to the first region are usually referred as QM atoms, and the ones of the second as MM atoms. Different degrees of QM/MM approximations exist, apart of the inherent approximations of QM and MM methods by themselves, according to the way in which the two subsystems interact between them.

In additive schemes, for instance, the energy of the system is the sum of exclusively QM (E^{QM}) and MM (E^{MM}) particles plus the coupling between the two ($E^{QM\text{-}MM}$):

$$E^{QM/MM} = E^{QM}\left(\{\vec{R}_q\}\right) + E^{MM}\left(\{\vec{R}_c\}\right) + E^{QM-MM}\left(\{\vec{R}_q\},\{\vec{R}_c\}\right) \tag{48}$$

The last term of this equation is the heart of the QM/MM approach. It takes into account the non-bonded VDW and electrostatic interactions between the two subsystems, but also bonded interactions in case that there are covalent bonds in between QM and MM atoms (*e.g.* enzymatic residues have to be split into QM and MM parts). The fact of "cutting" bonds raises the question of how to saturate the electronic requirements of the QM region, given than the electrons of the MM atom involved in the bond are not explicitly considered. This is usually addressed with two main strategies: (i) add monovalent "link" atoms such as hydrogens, also known as capping hydrogens; or (ii) use specific monovalent pseudopotentials.[131]

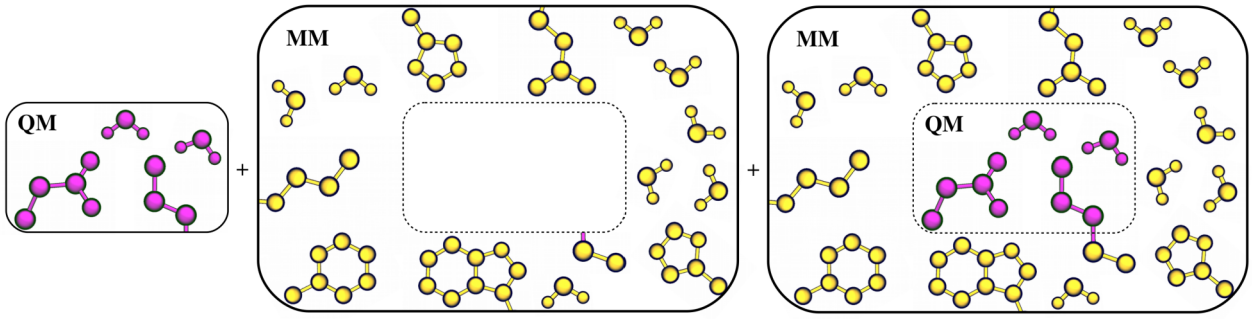


Figure 2.4- Schematic representation of the an additive QM/MM partition: (left) evaluation of the QM energy, which can be done either in vacuum or under the field of the MM charges; (middle) evaluation of the MM energy using common force fields; (right) evaluation of the QM-MM interactions, which in this representation does not only include non-bonded interactions, but also bonded interactions due to the presence of a bond that is in the interface between the QM and the MM region. QM atoms are represented in violet and MM atoms in yellow.

The addition of capping hydrogens has the disadvantage that includes an additional atom in the system, and its degrees of freedom and its interactions with the rest of the system should be minimized as much as possible. On the other hand, the use of monovalent pseudopotentials has the disadvantage that it requires specific pseudopotentials for each DFT functional and for each type of atom. In any case, both strategies perform well if the cut bonds are non-polar, such as carbon-carbon single bonds. It is also possible to cut other type of bonds, such as carbon-oxygen or carbon-nitrogen, but special care has to be taken in these cases to prevent strange behaviours (*e.g.* cutting a O_{QM}-C_{MM} bond with a capping hydrogen will lead a hydroxyl moiety in the QM region, which can establish fictitious hydrogen bonds with the environment). Given all the problems that can arise from those technical details, it is convenient to select a large enough QM region to avoid any spurious influence of the interface in the reaction center.[132]

Non-bonded QM-MM interactions of equation (48) are the most important ones, as their number is much larger than the one of bonded interactions. The most simple approach to deal with non-bonded QM-MM interactions is called *mechanical embedding*, which does not include the influence of the MM enviornment in the QM energy evaluation –*i.e.* the QM region is *like in the gas phase*– and treats all the non-bonded interactions at an MM level. This is a crude approximation given that MM polarization effects on the electron density are not considered, albeit they may be crucial for many reactions, but has the advantage that it is very fast to evaluate all the interactions. A more sophisticated approach, called *electrostatic embedding*, includes polarization effects in the QM region, which is achieved by introducing an additional term of MM charges in the QM Hamiltonian.

This makes necessary to couple the QM and MM softwares as they have to communicate between them. Moreover, the electrostatic interactions in the $E^{QM\text{-}MM}$ term can be evaluated using directly the electronic density:

$$E_{elec}^{QM\text{-}MM} = \sum_{I=1}^{MMatm} q_I \int d\vec{r}\, \frac{\rho(\vec{r})}{|\vec{r}-\vec{R}_I|} \tag{49}$$

The evaluation of this term, however, can be prohibitively expensive for large systems, as it requires to compute "*N·G*" electrostatic pairs, with *N* being the total number of MM atoms (typically ~10^5) and *G* the number of grid points to represent the electron density in the real space (typically ~10^6).[104] Nonetheless, there are smart schemes to handle this problem. In particular, the CPMD program allows to define three regions of interaction depending on the proximity of the MM atoms to the QM region.[133,134] For atoms that are near to the region (NN atoms), equation (49) is explicitly solved, and for atoms that are above a distance threshold (ESP atoms), the electron density is used to generate D-RESP charges and these fitted charges are employed in the electrostatic calculations.[j] This is a good approximation given that the distances involved in the definition of such regions can be selected to be arbitrarily accurate.

2.5 Making Rare Events not that Rare: Enhanced Sampling Methods

We have already outlined the basic techniques to carry out MD simulations with different schemes (EMD, BOMD and CPMD) and different levels of theory (MM, QM and QM/MM), so now it is time to explore the phase-space of the system and, for instance, unveil a hypothetical $A \rightarrow B$ chemical reaction and its associated free energy change. This means that, in the canonical ensemble, we have to evaluate:

$$\Delta F_{A\rightarrow B} = -k_B T \ln \frac{p(\vec{x}_B)}{p(\vec{x}_A)} \tag{50}$$

where $p(\vec{x}_A)$ and $p(\vec{x}_B)$ are the weighted probabilities associated to basins A and B:

$$\frac{p(\vec{x}_B)}{p(\vec{x}_A)} = \frac{\int_{\vec{x}\in B} d\vec{x}\; e^{-\beta E(\vec{x})}}{\int_{\vec{x}\in A} d\vec{x}\; e^{-\beta E(\vec{x})}} \tag{51}$$

The direct evaluation of this ratio using the energy values from the explored states is, however, not straightforward for two reasons: (i) these energies are large numbers that involve a huge amount of interactions, and thus for a finite sampling they lead to large statistical uncertainties;[99] and (ii)

j) There is an additional region in between the NN and the ESP, called MIX region, that takes into account the explicit form of the electron density for the calculation of electrostatic interactions with MM atoms above a threshold charge.

even assuming that uncertainty could be handled, the exponential average is dominated by low energy regions, which makes difficult its convergence.[135]

An alternative approach is to evaluate the ratio by *counting* the number of times that a MD simulation has spent in A and B, *i.e.* evaluate directly the probability, and not construct it from the energies. This requires the observation of many transitions from one basin to the other, so that the probability converges up to a given value. However, such transitions involve the crossing of high-energy regions (transition states) that are hampered by the exponential of the energy difference, making them unlikely to occur. For instance, a typical chemical reaction involves energy barriers of the order of 20 kcal·mol^{-1}, and the relative probability of being at the transition state is ~10^{-15} at 300 K, meaning that one should need *one thousand million of millions* of MD steps to just explore once the transition state!

This last point is known as the "*rare event*" challenge and makes necessary the use of *enhanced sampling techniques*, which allow to explore the phase-space of a system in an efficient manner. There are many of such techniques available (*e.g.* thermodynamic integration[100], string method[136], replica exchange[137] or transition path sampling[138]), but here we will briefly introduce the two that we have used along this thesis work: metadynamics[139] and umbrella sampling[140,141].

2.5.1 The space of collective variables

Most sampling techniques, including metadynamics and umbrella sampling, rely on a dimensionality reduction of the space, *i.e.* on the choice of a small set of collective variables (CVs) that describe a particular transition of interest. These CVs are functions of the coordinates of the system, *i.e.* $\xi(\vec{r})$, such as simple distances, angles or dihedrals, but they can also be more complex functions, such as the root-mean square deviation (RMSD) of a protein backbone, the radius of gyration, the number of hydrogen bonds or even the potential energy itself.[142] The expression of the probability function with respect to a CV adopts the following expression:

$$p(\xi(\vec{r}))=\frac{\int \exp[-\beta(E(\vec{r}))]\partial[\xi(\vec{r})-\xi]\partial\vec{r}}{\int \exp[-\beta(E(\vec{r}))]\partial\vec{r}} \tag{52}$$

where at the numerator the integral is made for all possible values except ξ. The choice of CVs is crucial for obtaining meaningful results, and they must fulfill –ideally– the following requirements:

- Be complete enough to distinguish between different states of interest along a particular transition, such as reactants, products, transition states and possible intermediates. If CVs do

not allow to do so, states may be mixed and it will not be possible to separate their free energy values (in the worst case they will not be even sampled).

- Be small enough in number to gain human insight. The inclusion of too many CVs makes their analysis complex and counterintuitive for deducing conclusions. Depending on the number of CVs, simple free energy profiles (1 dimensional), manageable free energy landscapes (2 dimensional) or complex free energy volumes and hypervolumes (3 and more dimensions) can be obtained, leading to increasing intricate analyses.[143] Moreover, the computational cost of certain methods, such as metadynamics or umbrella sampling, depends on the number of CVs, and so it is desirable to minimize that cost.

- Be selected so that they include the *slow motions* of the system, *i.e.* motions that have large free energy barriers along the transition and that will not be properly sampled within the time scale of common simulations. The lack of such slow motions in the CVs can lead to problems in the convergence of free energies and in the detection of the most likely pathway for the transition.

In this thesis we have used different types of CVs, including plain distances, differences of distances and dihedral angles. These metrics are usually enough for the proper description of enzymatic reactions, where the slow motions are related with the bonds that break and form during the reaction. Additionally, to study the conformational landscape of sugar rings we have used a set of *puckering coordinates*[144] that are not as trivial as simple metrics, but that follow the same philosophy than any other types of CVs. These coordinates are briefly outlined below.

2.5.2 The hills method: metadynamics

One way to overcome free energy barriers is to make the system be "uncomfortable" in the reactant basins. Metadynamics[139] (MTD) is an enhanced sampling technique that allows to do so, adding repulsive energy functions in the space defined by a limited number of $\xi(\vec{r})$ CVs. This pushes the system away from a particular basin and drives it to explore other regions of the space (see Figure 2.5). The repulsive functions are periodically deposited along the simulation, leading to a time-dependent *bias potential* that balances the effective potential of the system:

$$V_T(\xi(\vec{r}),t)=V_{eff}(\xi(\vec{r}))+V_G(\xi(\vec{r}),t) \tag{53}$$

Normally, Gaussian functions are used as repulsive potentials, and the time dependent potential can be written as the sum of all the N_G deposited Gaussians with a τ_G frequency in the d dimensional space:

$$V_G(\xi(\vec{r}),t)=\sum_{i=1}^{N_G} w(i\,\tau_G)\exp\left(-\sum_{j=1}^{d}\frac{(\xi_j(\vec{r})-\xi_j(\vec{r}(i\,\tau_G)))^2}{2\,\delta\xi_j^2}\right) \tag{54}$$

where $\xi_j(\vec{r})$ is the jth CV, and w and $\delta\xi$ are, respectively, the height and the width of the Gaussians. This external potential reduces the free energy barrier with every potential added, until the barrier becomes so small that it can be surmounted by thermal fluctuations (Figure 2.5). Moreover, at infinite time the bias potential converges towards:

$$\lim_{t\to\infty} V_G(S(\vec{R}),t)=-F(S(\vec{R}))+C \tag{55}$$

where $F(S(\vec{R}))$ is the free energy along the CV and C is an additive constant. Therefore, MTD does not only allow to uncover reaction mechanisms, but also their underlying free energy changes. This relation can be surprising at first glance, as it states that an equilibrium property –the free energy– can be estimated by means of a non-equilibrium simulation in which the total potential changes every time that a new Gaussian function is deposited. However, it can be viewed as a dynamic way of constructing the static free energy potential (the one that makes the probablity function to be equal to unity everywhere, *i.e.* all the states having the same probability of being explored). In fact, it is easy to see that in the limit of a Gaussian height tending to zero ($w\to 0$) or a deposition time tending to infinite ($\tau_G \to\infty$) plain MD is recovered. This means that if we choose a long-enough deposition time, the system can reach equilibrium between the addition of new Gaussian terms, and if we choose a very small Gaussian height, the system will be minimally perturbed after each addition. The error of a MTD estimate, therefore, is associated with these two parameters, and has been proben to be:

$$\varepsilon \propto \sqrt{\frac{w\,k_B T}{D\,\tau_G}} \tag{56}$$

where D is a difussion coeficient associated with the time that the system needs, once the free energy surface has been filled, to explore the CV space.

The advantages of MTD are several: (i) it allows to explore rare events, providing accurate free energy estimates as long as the CV space is complete and the orthogonal degrees of freedom –assumed to be fast motions– are properly sampled; (ii) it does not require any *a priori* assumption about the most likely reaction patwhay on the free energy landscape. This is a crucial point as other

methods, such as the umbrella sampling discussed below, rely on the definition of initial pathways, which can bias the results if such pathways are arbitrarily selected. (iii) it can be parallelizable, and many "walkers"[145] can explore different regions of the CV space at the same time. This leads to a linear scaling algorithm that is very efficient in terms of computational efforts; (iv) it can be used to obtain direct reaction rates.[146]

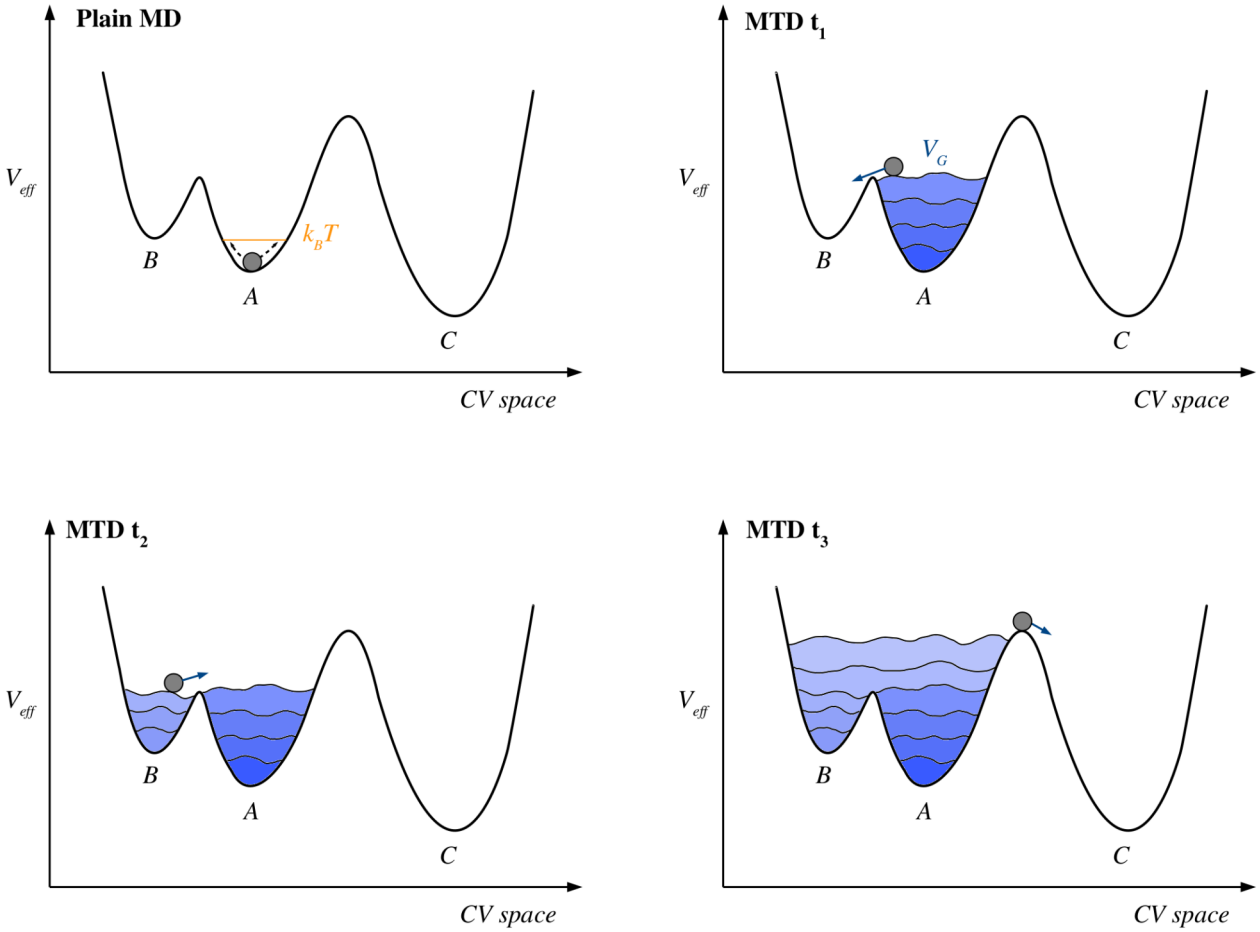


Figure 2.5- Schematic representation of the rare event challenge and the solution provided by metadynamics: (top left) a plain MD simulation is stuck in basin A and within k_BT it will take a long time to explore B and C; (top right) A metadynamics V_G bias potential deposited over time fills the A basin and allows to observe a transition from A to B; (bottom left) basin B has been already filled and the system can diffuse freely in between A and B, but there is still a notable barrier to cross to C; (bottom right) after a period of time, the V_G bias potential allows to observe the transition to basin C, and the accumulated bias potential –with opposite sign– is an estimate of the underlying free energy barrier. The time dependence of V_G has been represented with a blue color gradient scale.

Among the disadvantages of MTD we can find: (i) its computational cost scales with the number of CVs. This makes necessary to use the minimum number of variables that lead to meaningful results, which is difficult to find out; (ii) converging free energy estimates is complex in practice, particularly in the field of enzymatic reactions. It is indispensable to tune carefully the parameters that govern metadynamics, and also to test how they affect to the final outcome; (iii) transition state regions are poorly explored if short-time simulations are used; (iv) the lack of a slow motion CV leads to strong hysteresis problems in the free energies. This last point, however, is at the same an advantage, as it allows to identify slow motions that may be important to include in the CV space.

2.5.3 Using umbrellas on sunny days: umbrella sampling

A widely used method to explore rare events is umbrella sampling (US).[140,141] This method receives its name for using umbrella shaped potentials –harmonic functions– along a predefined CV space, with the aim of increasing the sampling at the region where the potential is centered (see Figure 2.6). The addition of such potentials makes energies to be biased:

$$E^b(\vec{r}) = E^u(\vec{r}) + w_i(\xi) \tag{57}$$

and so also the bias affects the probability distributions:

$$P_i^b(\xi) = \frac{\int \exp[-\beta(E^u(\vec{r}) + w_i(\xi))]\partial[\xi'(\vec{r}) - \xi]\partial\vec{r}}{\int \exp[-\beta(E^u(\vec{r}) + w_i(\xi))]\partial\vec{r}} \tag{58}$$

These biased probabilities can be properly evaluated given that they are defined in a "i" local region of the CV space, and then can be used to estimate the unbiased probability according to:

$$P_i^u(\xi) = P_i^b(\xi)\exp[\beta w_i(\xi)]\langle\exp[-\beta w_i(\xi)\rangle \tag{59}$$

which, ultimately, can be used to obtain free energy estimates following:

$$F_i(\xi) = -\frac{1}{\beta}\ln(P_i^u(\xi)) \tag{60}$$

$$F_i(\xi) = -\frac{1}{\beta}\ln(P_i^b(\xi)) - w_i(\xi) - \frac{1}{\beta}\ln(\langle\exp[-\beta w_i(\xi)]\rangle) \tag{61}$$

Therefore, one can sample a region of the space and obtain unbiased probabilities, but these regions have to overlap in order to obtain good statistical data (*i.e.* in the limit of $w_i \to 0$, plain MD is recovered). This makes necessary to use several "windows" to collect data along the CV space, whose probability functions can be combined afterwards to obtain the global free energy profile.

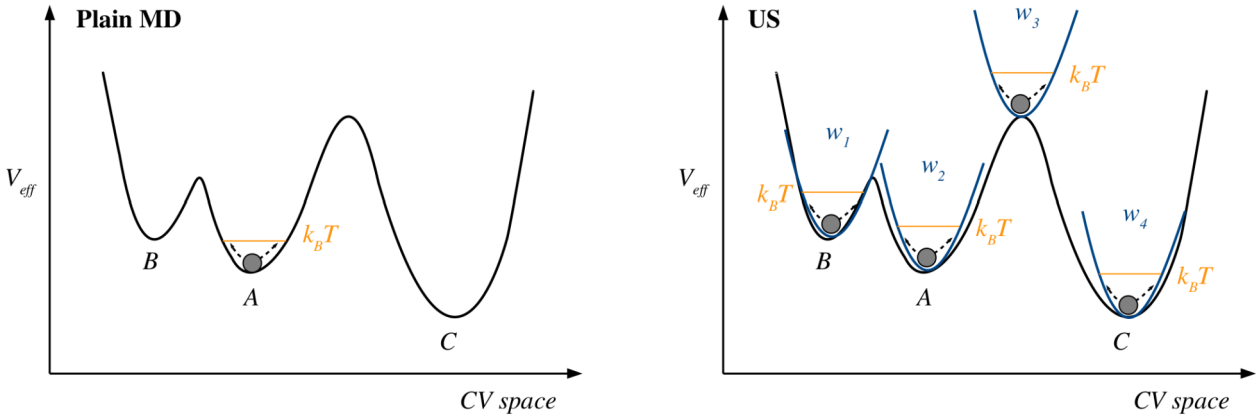


Figure 2.6- Schematic representation of the rare event challenge and the solution provided by umbrella sampling: (left) a plain MD simulation is stuck in basin A and within k_BT it will take a long time to explore B and C; (right) four independent umbrella potentials –or windows– are settled in different regions of the space that one aims at exploring. After the local convergence of the biased probability in each window, the unbiased probability can be recovered and combined to obtain the whole free energy profile.

There are different ways to address the combination of windows, although the weighted histogram analysis method (WHAM)[147,148] is one of the most used ones due to its good performance. This method finds the global probability distribution by means of a linear –weighted– combination of the local probabilities:

$$P^u(\xi) = \sum_{i=1}^{windows} p_i P_i^u(\xi) \tag{62}$$

where the weights of each window are chosen to minimze the variance on the global probability and having them normalized:

$$\frac{\partial \sigma^2(P^u)}{\partial p_i} = 0 \tag{63}$$

$$\sum_{i=1}^{windows} p_i = 1 \tag{64}$$

This is achieved by defining the probability weights as:

$$p_i = \frac{a_i}{\sum_{j=1}^{windows} a_j} \tag{65}$$

$$a_i = n_i \exp[-\beta(w_i(\xi) - Z_i)] \tag{66}$$

with $Z_i = \frac{1}{\beta} \ln(\langle \exp[-\beta w_i(\xi)] \rangle)$ and n_i being the total number of steps sampled in window i.[149]

The Z_i is a constant related with the unbiased probability distribution by:

$$\exp[-\beta Z_i] = \int P^u(\xi)\exp[-\beta w_i(\xi)]d\xi \tag{67}$$

and given that both Z_i and P^u appear in equations (62)[k] and (67), it is necessary to iterate these equations to obtain the final result. This approach requires a good overlap of distributions between consecutive windows, otherwise the construction of the global energy profile can result erroneous.

The main advantages of US are: (i) it is an equilibrium technique, so the error associated to the free energy estimates does not depend on external parameters such as the Gaussian heigh or the deposition time in MTD; (ii) the simulations of each window can be run in parallel without any communication between them; (iii) its computational cost is only related with the amount of sampling done in each window, and it does not strictly scale with the number of CVs considered in the simulation; (iv) all the points along the CV are properly sampled, including transition state regions.

The main disadvantages are: (i) it is necessary to define the initial pathway, otherwise one needs to explore a grid in the CV space, which can be very expensive depending on the number of grid points; (ii) it is usually limited to 2 CVs for post-processing reasons, although higher dimensional schemes have been developed to deal with more variables; (iii) it is not straightforward to determine possible pitfalls in the free energy estimates, as a bad choice of CVs can lead either to higher or lower energies than the ones expected.

2.5.4 Cremer and Pople puckering coordinates

Any of the possible conformations of an *N*-atom ring can be unequivocally assigned using the Cremer and Pople *puckering coordinates*.[144] These coordinates are defined by the displacements (z_j) of each ring atom from a mean plane that is centered at z = 0 and fixed by two conditions:

$$\sum_{j=1}^{N} z_j \cos[2\pi(j-1)/N] = 0 \tag{68}$$

$$\sum_{j=1}^{N} z_j \sin[2\pi(j-1)/N] = 0 \tag{69}$$

The orientation of the mean plane can be determined by the following vectors:

$$\vec{R}' = \sum_{j=1}^{N} \vec{R}_j \sin[2\pi(j-1)/N] = 0 \tag{70}$$

$$\vec{R}'' = \sum_{j=1}^{N} \vec{R}_j \cos[2\pi(j-1)/N] = 0 \tag{71}$$

k) Notice that Z_i enters in equation (62) via equations (65) and (66).

where $\vec{R}_j$ are the nuclear positions. The atomic displacements from the mean plane are given by the dot product:

$$z_j = \vec{R}_j \cdot \vec{n} \tag{72}$$

with $\vec{n}$ being the unit vector perpendicular to $\vec{R}'$ and $\vec{R}''$:

$$\vec{n} = \frac{\vec{R}' \times \vec{R}''}{|\vec{R}' \times \vec{R}''|} \tag{73}$$

This unit vector is taken as the molecular z-axis. At this point, Cremer and Pople defined the following generalized ring-puckering coordinates:

$$q_m \cos\phi_m = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} z_j \cos\left[\frac{2\pi m}{N}(j-1)\right] \tag{74}$$

$$q_m \sin\phi_m = -\sqrt{\frac{2}{N}} \sum_{j=1}^{N} z_j \sin\left[\frac{2\pi m}{N}(j-1)\right] \tag{75}$$

that apply for an odd number of $N > 3$ atoms, where q_m and ϕ_m are a set of puckering amplitudes and phase angles with m = 2, 3, … $(N - 1) / 2$. If the number of atoms is even, the coordinates in equations (73) and (74) apply up to m = $(N / 2) - 1$ and there is an additional coordinate:

$$q_{N/2} = \sqrt{\frac{1}{N}} \sum_{j=1}^{N} (-1)^{j-1} z_j \tag{76}$$

In the case of pyranoses, therefore, $N = 6$ and there are three puckering coordinates:

$$q_2 \cos\phi_2 = \sqrt{\frac{1}{3}} \sum_{j=1}^{6} z_j \cos\left[\frac{2\pi}{3}(j-1)\right] \tag{77}$$

$$q_2 \sin\phi_2 = -\sqrt{\frac{1}{3}} \sum_{j=1}^{6} z_j \sin\left[\frac{2\pi}{3}(j-1)\right] \tag{78}$$

$$q_3 = \sqrt{\frac{1}{6}} \sum_{j=1}^{6} (-1)^{j-1} z_j \tag{79}$$

These coordinates are usually replaced by a "*Q, θ, φ*" polar set according to:

$$q_2 = Q \sin\theta \tag{80}$$

$$q_3 = Q \cos\theta \tag{81}$$

with the Q coordinate being the total puckering amplitude:

$$Q^2 = \sum_{j=1}^{6} z_j^2 \tag{82}$$

Since these are essentially polar coordinates, any ring conformation falls within the puckering sphere-like volume (Figure 2.7). While Q may differ among different conformations, θ and φ are sufficient to differentiate between all the conformers. On the poles (θ = 0º and θ = 180º) are located

the two chair conformers ($^4C_1$ and $^1C_4$, respectively); on the equatorial region (θ = 90º) the 6 boat and 6 skew-boat structures are sequentially placed in steps of φ = 30º.

Two representations have been historically used to map the Cremer and Pople 3D sphere into simpler two dimensional plots (Figure 2.7). The Stoddart diagram corresponds to the projection of the polar coordinates onto the equatorial plane, and can be denoted by the Cartesian coordinates q_x and q_y. On the other hand, the so-called *plate carrée* or Mercator representation is an equidistant cylindrical projection that results in a rectangular map with respect to θ and φ. Which representation to use is essentially a matter of choice and interconversion between them are doable with Jacobian transformations.

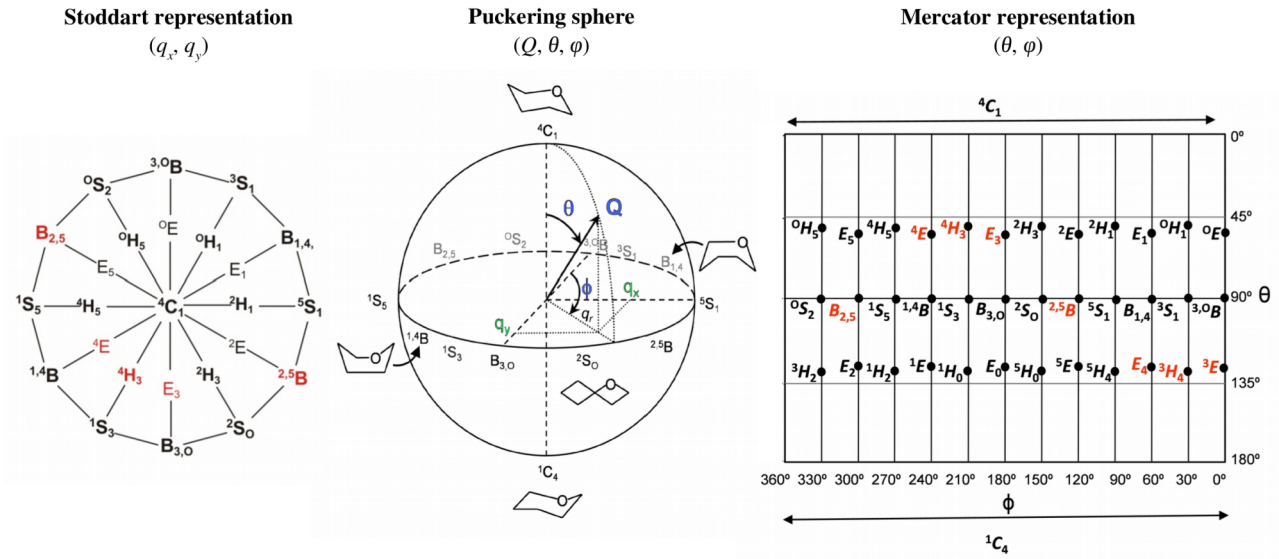


Figure 2.7- Stoddart and Mercator representations of the puckering sphere for a six-membered ring. The optimal TS conformations are highlighted in red. The "*Q, θ, φ*" and the "q_x, q_y" coordinates are represented in the puckering sphere. Notice that two Stoddart maps are possible (northern and southern representations), but in this case we only represent one of them (the northern).

2.6 The Three Pillars of Computational Predictions

In this section we summarize the important points that have to be taken into account for obtaining meaningful computational predictions. As exposed in the above sections, matching macroscopic experimental observables with computational tools is a difficult task, as it requires to average the microscopic values obtained in the simulations. This means that the accuracy of a computational prediction depends on (i) the statistical significance of the average, and (ii) the accuracy of the microscopic observables. The first point, in the particular case of MD simulations, depends entirely on

the *time* devoted to explore the phase-space. The second point depends on the expression used to describe the *energy* of the system, which at the same time depends on the *size* of the system itself (understanding by "size" the number of atoms that are described, *i.e.* if the system is completely considered in the simulation or if it is reduced to a minimal model that includes just the few molecules involved in the chemical problem). Therefore, a computational prediction is sustained by three "pillars" –*energy*, *time* and *size*– that have to be robust enough to maintain it (see Figure 2.8).

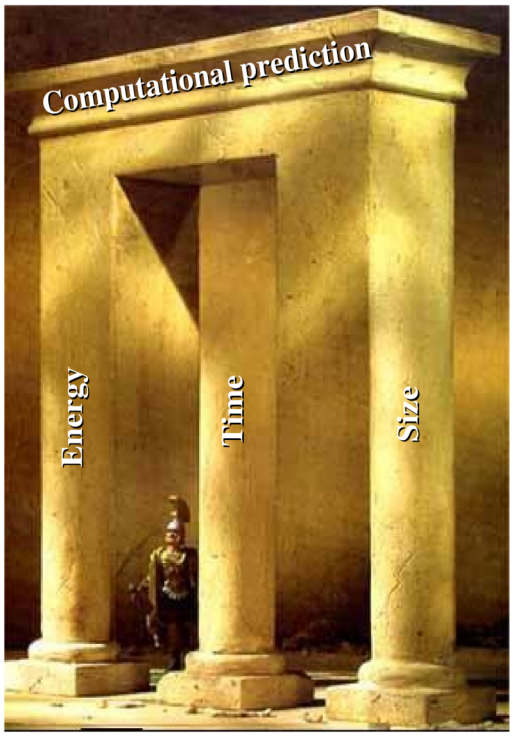


Figure 2.8- This optical view exemplifies the three pillars that sustain computational predictions, related with the *energy*, *time* and *size* challenges. If one looks at the bottom it seems that there are three cylindrical columns, while if one looks at the top, it seems that there are only two rectangular columns. The time challenge has been placed in the center –the column that can be seen or not– as it is usually the challenge that is commonly neglected.

Most computational predictions, however, neglect one of the three pillars in favor of computing efficiency, adopting different types of approximations that compromise the predictions. For instance, a very accurate description of the energy is only possible for systems with a reduced size and for short –or even non– times (*e.g.* cluster models of enzyme active sites at a B3LYP/MP2 level of theory[150]), and long time simulations are only possible hampering the energy accuracy and or the size of the system (*e.g.* force field based approaches that do not consider explicitly the solvent[151]). A good balance between the three pillars is necessary for obtaining proper results, and this requires either to make use of brute force approaches (*i.e.* take advantage of the astonishing growth of techno-

logical facilities, such as supercomputers, cloud computing or GPU graphics cards), or use smart approaches that permit to approximate the result within a compromise in accuracy. Among these last techniques we can find the hybrid QM/MM methods, that allow to alleviate the *energy-size* challenge, and enhanced sampling techniques such as metadynamics or umbrella sampling, that allow to alleviate the *time* challenge.

It is worth noting, however, that there are many types of QM/MM approaches depending on their treatment at the QM region (*e.g.* HF, post-HF, DFT or semiempirical methods) and at the QM-MM interface (*e.g.* mechanical and electrostatic embedding), and so the mere label of "QM/MM" in a simulation is not a guarantee for the goodness of the results. Moreover, QM/MM methods can be static or dynamic (*i.e.* include or not MD), and inside the dynamical ones one can further include the use of enhanced sampling techniques (*e.g.* QM/MM metadynamics). Similarly, enhanced sampling techniques can be applied with different descriptions of the energy, and it is not the same to make a MM metadynamics than a QM metadynamics or a QM/MM metadynamics.

In this thesis we have made use of several techniques to take into account the three pillars: we have used QM/MM schemes to consider the whole enzymatic environment, circumventing the size challenge; we have used DFT with a GGA functional for the QM part, allowing to achieve a good compromise in the energy challenge; and finally we have used MD simulations together with enhanced sampling techniques, alleviating the time challenge. In summary, we have used a QM(DFT)/MM metadynamics approach for the study of enzymatic reactions, which are cutting edge simulations for the current standards.

Chapter 3
How Do Sugar Conformations Enhance Catalysis?

Parts of this chapter have been published:
J. Iglesias-Fernández, L. Raich, A. Ardèvol, C. Rovira. "The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases" *Chemical Science*, **6**, 1167-1177 (2015).

ABSTRACT: sugar distortions are presumed to enhance GH catalysis by their crucial stereoelectronic advantages, but up to date there is not kinetic evidence, either experimental or computational, to proof this statement. In the present chapter we address this important issue by performing a complete conformational study of β-xylose inside a β-xylanase. Our results reveal that two conformations are available on-enzyme, a non-distorted and a distorted one, with the latest being ~42 kcal·mol^{-1} more reactive than the former. This enormous difference is principally attributed to the inability of the non-distorted conformation for reaching one of the favored conformations for an oxocarbenium ion transition state. Interestingly, although the non-distorted conformation exhibits low reactivity, it is the most stable conformation in the enzymatic cavity (by ~2 kcal·mol^{-1}), highlighting that GHs do not necessarily bind the substrate in its most reactive conformation. Finally, these results have allowed us to disentangle all the possible catalytic itineraries for β-xylanases, solving an intriguing controversy that arose from an experimental prediction made on the basis of an unreliable enzymatic mutant.

3.1 Introduction

3.1.1 Being distorted: evidences and presumptions

In the field of glycobiology it is generally accepted that GH catalysis is enhanced by certain sugar conformations. This conviction is principally sustained by strong –but indirect– experimental evidences: there are plenty of crystallographic structures showing Michaelis complexes with -1 sugars distorted in unusual conformations, *e.g. boats* or *skew-boats*, instead of the relaxed *chairs* that are mostly found in aqueous solution (see Figure 3.1).[43,152,153] These distortions, by simple visual analysis, place the scissile glycosidic bond in an axial position, which in principle could facilitate the prototypical S_N2 attack of a nucleophile and favor catalysis.

Although these evidences are certainly compelling, one has to consider that the crystal structures are usually not obtained for the natural systems, but rather for enzymatic mutants and/or substrate mimics that knock-out the enzymatic activity (otherwise the reaction would take place during the timescale of X-ray experiments and it would not be possible to trap the Michaelis complex). This fact brings up the question of whether the observed distortions are also present in the natural systems or if they are just an artifact resulting from the structural modifications made to prevent the reaction. Fortunately, the lack of precise experimental techniques to address this intriguing issue has not been an obstacle thanks to the power of computer simulations.

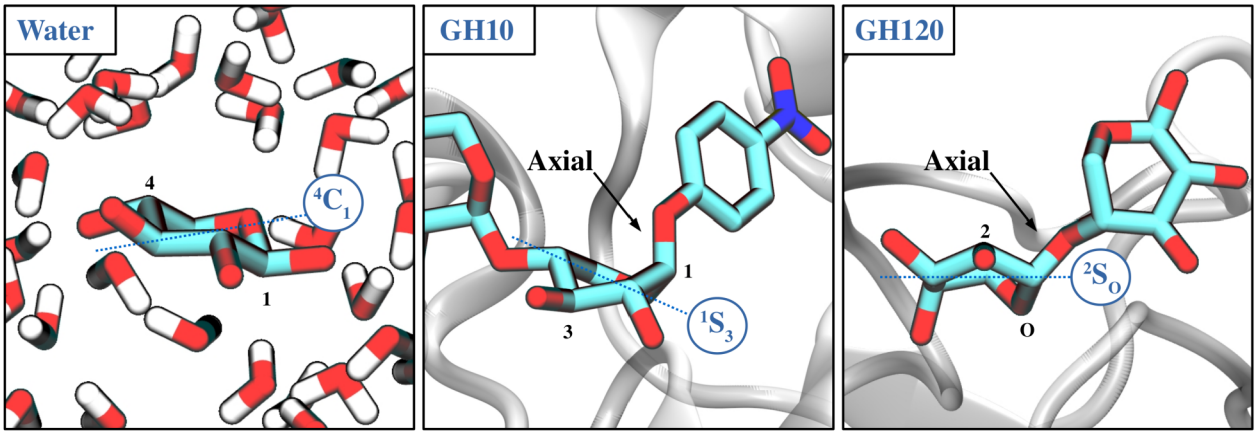
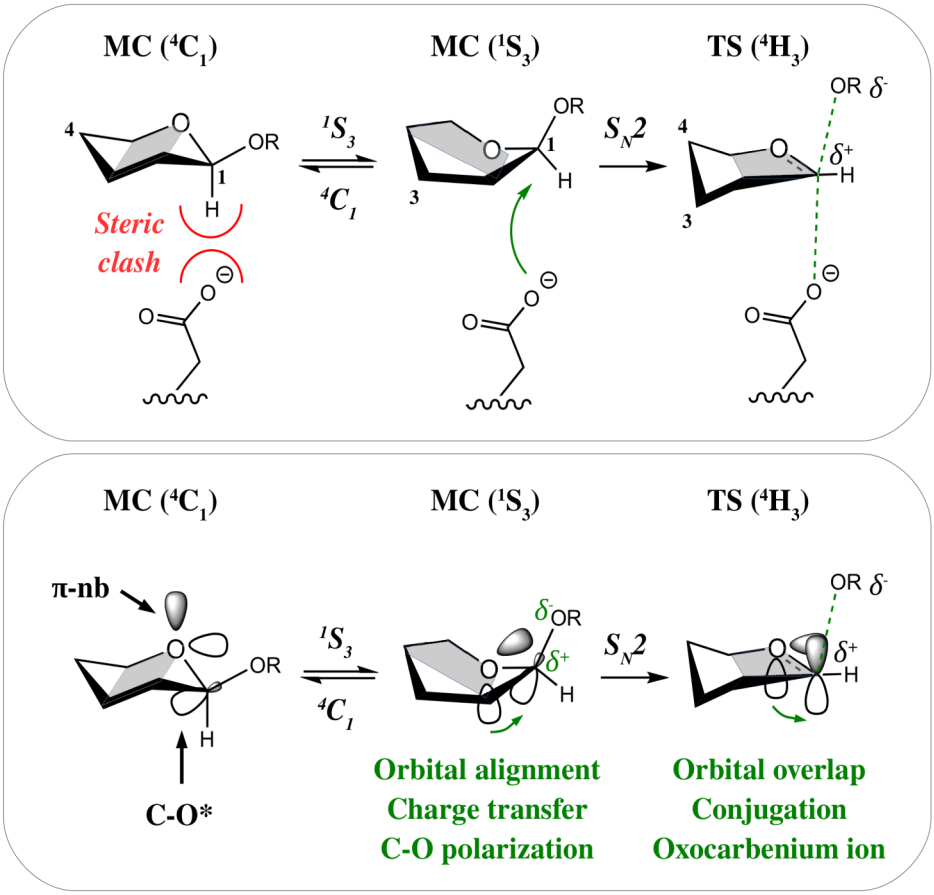


Figure 3.1- Conformation of a β-xylose sugar in different environments: in water solution, displaying a $^4C_1$ chair conformation (left); in the active site of a GH10 β-xylosidase, displaying a distorted $^1S_3$ skew-boat conformation (middle, PDB 2D20); in the active site of a GH120 β-xylosidase, displaying a distorted ^2S_O skew-boat conformation (right, PDB 3VSU). The dotted blue line indicates the four atoms that are nearly in a plane. Substrate hydrogen atoms have been omitted for clarity.

Ab initio QM/MM methods have been extensively employed to study the conformational preferences of sugars in the active site of GHs. The pioneering work of Biarnés *et al.* was the first to give insights about the structural and electronic advantages of sugar distortions on-enzyme.[154] By comparing the sugar in a distorted conformation with respect to a non-distorted one, it was found that the former displayed a slightly longer glycosidic bond distance, a slightly shorter C1-O5 intra-ring distance and an increase of the charge at the anomeric carbon, approaching to the stereoelectronic features of an oxocarbenium ion-like transition state (see Scheme 3.1). All these subtle differences conduced to the concept of "preactivated" conformations for catalysis, which would require less energy than non-preactivated conformations to reach the transition state.



Scheme 3.1- Sugar distortions are presumed to enhance GH catalysis by at least two factors: (top) they place the leaving group of the reaction in an optimal orientation for an S_N2 displacement, diminishing steric clashes, and (bottom) they induce polarization of the scissile bond by the alignment of the π-nb and the σ^* frontier orbitals of the pyranic oxygen and the glycosydic bond, reducing its dissociation energy. The gray shadow surface indicates the four atoms that are nearly in a plane.

Since the introduction of the preactivated concept, many computational works have confirmed the existence of substrate distortion in GH Michaelis complexes, reinforcing the crystallographic evidence from modified systems.[50,155–159] It is important to highlight, however, that not all the crystal structures are good mimics of the natural systems, but rather some of them, according to MD simulations, affect the conformation of the substrate.[160] Altogether, both by experimental and computational evidences, nowadays there is no doubt that *sugar distortions occur inside the active site of GHs*, but their role in enhancing catalysis remains as a presumption given that, although everything points towards that direction, there are no kinetic experiments demonstrating that one conformation is catalytically faster than another.

The problem is that it is not possible to restrict active sugars in a particular conformation, let is say, one can not have the -1 sugar of a polysaccharide in a *chair* and a *skew-boat* conformations, separate in two pots, and test their respective catalytic activities upon hydrolysis. Computational models, again, can give valuable insights about it. With computer simulations we can differentiate between different –short lived– conformational states and we can study their reaction pathways separately. Consequently, in this chapter we had the aim to quantify the contribution of different substrate conformations to the reaction rate of GHs.

3.1.2 The catalytic itineraries of β-xylosidases are not unambiguously resolved

We have centered our study in β-xylosidases, enzymes that are particularly interesting for their wide range of industrial applications, such as in biofuel, bread, pulp and feedstock industries.[161–164] They are responsible for the hydrolysis of glycosidic bonds in β-xylans, polymers that are present in plant-cell walls (see Figure 3.2 a). Albeit the general mechanism of these enzymes is well known, their *catalytic itineraries –i.e.* how the -1 sugar changes conformation during catalysis– are not completely resolved. These itineraries, according to what is exposed above, should start from a preactivated conformation and cross one of the energetically favored TS for an oxocarbenium ion (see Figure 3.2 b).

On the basis of X-ray experiments, one can trap a crystallographic structure of the Michaelis complex and, ideally, one of the products, with which throw a straight line connecting both conformations to define the catalytic itinerary. If a favored TS-like conformation is crossed with this approach, the catalytic itinerary earn points to be believed. In the case of β-xylosidases, three different catalytic itineraries have been predicted from crystallographic evidence: (i) $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ (ii) $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$ and (iii) $^{o}E \rightarrow [^{o}S_{2}]^{\ddagger} \rightarrow B_{2,5}$ (see Figure 3.2 c).[165] While the two first itineraries

propose plausible conformations both for the Michaelis complex ($^{1}S_{3}$ and $^{2}S_{O}$) and the transition state ($^{4}H_{3}$ and $^{2,5}B$), the third one starts from a non-distorted conformer ($^{4}C_{1}$) and crosses through a non-favored transition state ($^{O}S_{2}$). Moreover, to add to the confusion, the two last catalytic itineraries have been assigned to a single family (GH11), while enzymes of a given family usually follow just one pathway. This discrepancy is controversial and opens an intriguing question: is this an exception to the rule or a wrong experimental prediction?

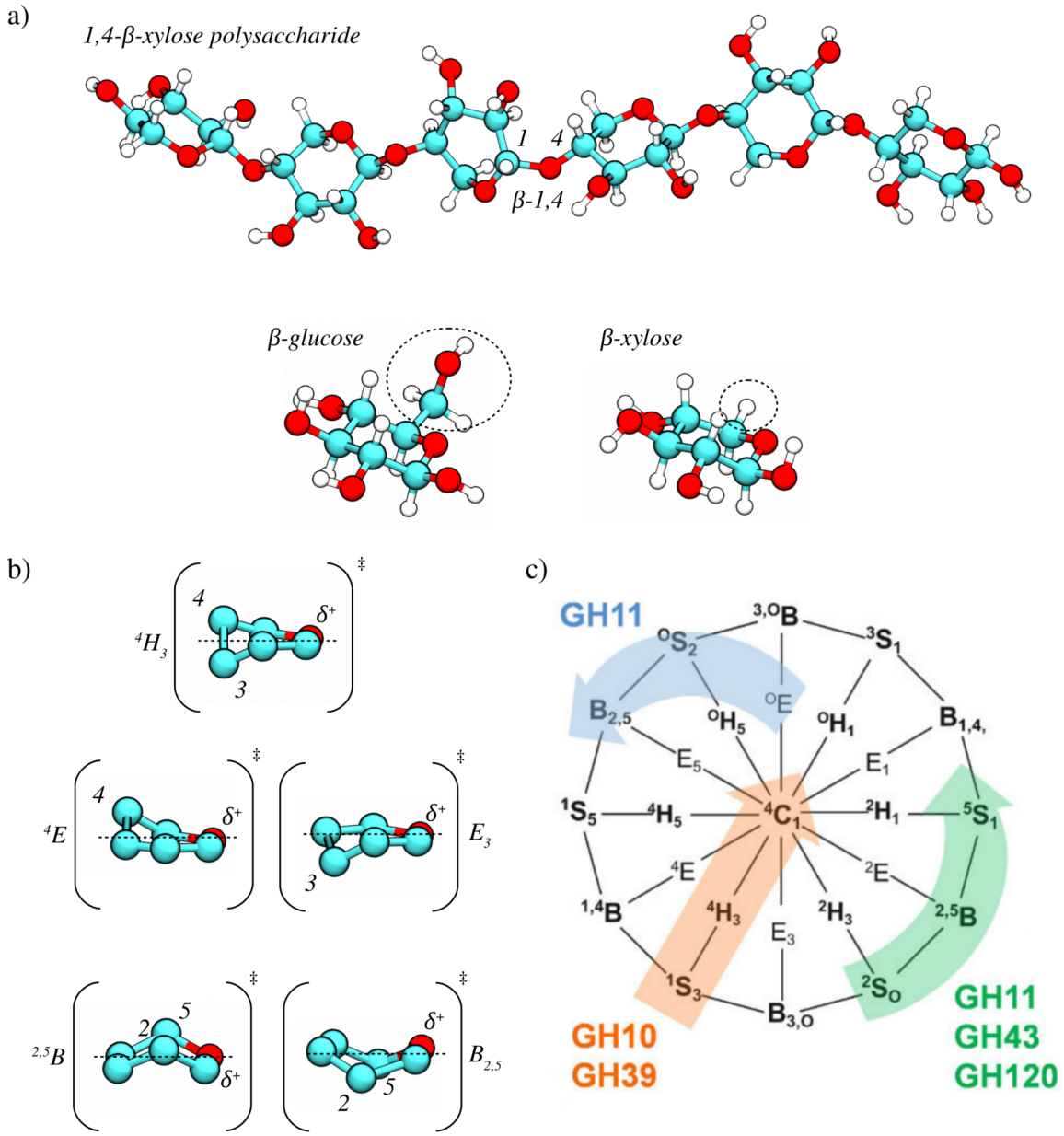


Figure 3.2- (a) Molecular structure of a 1,4-β-xylose polysaccharide. Notice that β-xylose differs from β-glucose in the substituent at position 5, with xylose bearing a hydrogen atom and glucose a hydroxymethyl group. (b) Energetically favored TS conformations of the north-pole. (c) Stoddart representation (north-pole) with the three proposed catalytic itineraries for β-xylosidases: $^{1}S_{3}$ → $[^{4}H_{3}]^{‡}$ → $^{4}C_{1}$ (orange), $^{2}S_{o}$ → $[^{2,5}B]^{‡}$ → $^{5}S_{1}$ (green) and ^{o}E → $[^{o}S_{2}]^{‡}$ → $B_{2,5}$ (blue).

To shade light into the main topic of this chapter –if sugar conformations enhance catalysis– and also into this last question, we have undertaken a three-fold strategy of increased technical complexity: (i) we have computed the free energy landscape of an isolated β-xylose, which gives information about all possible itineraries that β-xylosidases can take, showing that only two of the three proposed pathways fit with the low-energy regions of the landscape; (ii) we have obtained the same landscape but inside the active site of the controversial GH11 β-xylosidase, revealing how the enzyme restricts the access to certain regions of the landscape and favors only two conformations ($^{4}C_{1}$ and $^{2}S_{O}$) that match with one catalytic itinerary; (iii) finally, from this last result, we have studied the reaction mechanism starting from each of the two conformations to quantify their contribution to the reaction free energy barrier and verify the predicted catalytic itinerary, showing that only the $^{2}S_{O}$ conformation is kinetically competent and that the catalytic itinerary of the GH11 β-xylosidase should be $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$.

3.2 Results and Discussion

3.2.1 The fingerprint of β-xylose discards one catalytic itinerary

The free energy landscape (FEL) of an isolated β-xylose, shown in Figure 3.3, has been obtained by DFT-based metadynamics using the *theta* and *phi* puckering coordinates as collective variables (see section 3.4 Computational Details). It contains four local minima, two of them at the poles, corresponding to the $^{4}C_{1}$ and $^{1}C_{4}$ chair conformers (in blue, at theta 0º and 180º, respectively), as well as a two local minima in the equator (theta 90º): a wide minimum centered at $^{2}S_{O}/B_{3,O}$ (in green, phi 150º–180º) and a small minimum centered at $^{3}S_{1}/{}^{3,O}B$ (in yellow, phi 0º–30º). It is clear that the $^{4}C_{1}$ chair is the global minimum, the inverted $^{1}C_{4}$ chair is 4 kcal·mol^{-1} higher in energy, closely followed by the mixed $^{2}S_{O}/B_{3,O}$ conformation, which is 5 kcal·mol^{-1} above $^{4}C_{1}$. The remaining minima ($^{3,O}B/{}^{3}S_{1}$, $^{1}S_{5}$, $^{5}S_{1}$) are much higher in energy (7 kcal·mol^{-1}; see Figure 3.3).

A noticeable feature of the FEL is the presence of a wide valley on the equator, covering conformations $^{1}S_{3}$–$B_{3,O}$–$^{2}S_{O}$–$^{2,5}B$–$^{5}S_{1}$, which is connected with the $^{4}C_{1}$ global minimum by a low energy region at theta 75º, containing $^{4}H_{3}$ and E_{3} conformations. These results support both $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ and $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$ as possible itineraries for hydrolysis reactions catalyzed by β-xylosidases. The $^{o}E \rightarrow [^{o}S_{2}]^{\ddagger} \rightarrow B_{2,5}$ itinerary, however, is unusual because it starts from a non-preactivated conformation and crosses through a high-energy region of the landscape. This pathway was recently proposed for *Trichoderma ressei* GH11 xylanase (*Tr*GH11) in the basis of an atomic-reso-

lution (1.15 Å) X-ray structure, showing a Michaelis complex with the -1 sugar in a $^{4}C_{1}/^{O}E$ conformation.[166] In the light of this and previous studies on isolated sugars made in our group,[52,53,167] this is unprecedented and it might indicate that trends derived from the computed FEL break down in this case. However, the fact that the proposed $^{O}S_{2}$ conformation as TS does not fulfill the stereochemical requirements of a TS-like conformation suggests that the experimental prediction might not be correct. Moreover, the complex was obtained with a structural modification of the enzyme, which could affect the conformation of the substrate as we have observed in other GHs.

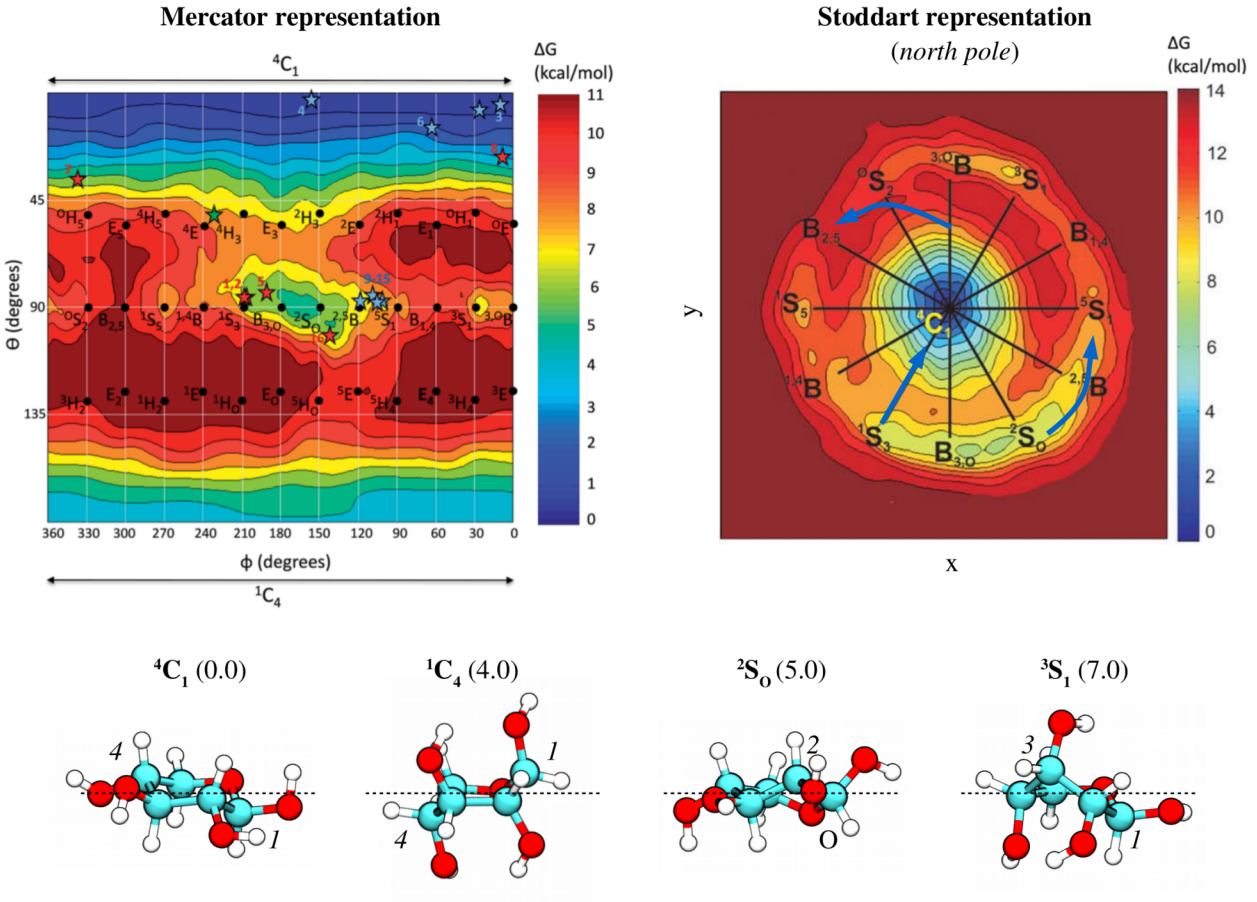


Figure 3.3- Computed free energy landscape of β-xylose with respect to ring distortion. Both Mercator and Stoddart-*like* ($x = \theta \cdot cos(\varphi)$ and $y = \theta \cdot sin(\varphi)$) representations of the same landscape are provided. Each contour line of the diagrams correspond to 1 kcal·mol^{-1}. The conformations found in experimental structures of retaining β-xylosidases are represented in the Mercator map by red stars (Michaelis complex structures) and blue stars (covalent intermediate structures).[165] The conformation of the TS-like inhibitor xylobio-imidazole in complex of family 10 xylanase Cex from *Cellulomonas fimi* has also been indicated (green star).[168] For the sake of clarity, only one star is displayed for several structures with nearly identical conformations. Notice in the Stoddart representation that from the proposed itineraries only the $^{1}S_{3} \rightarrow [^{4}H_{3}]^{‡} \rightarrow {}^{4}C_{1}$ and $^{2}S_{o} \rightarrow [^{2,5}B]^{‡} \rightarrow {}^{5}S_{1}$ cover low energy regions, but not the $^{o}E \rightarrow [^{o}S_{2}]^{‡} \rightarrow B_{2,5}$ one.

3.2.2 Simulations on-enzyme suggest the existence of two conformations

To verify that the FEL of the isolated β-xylose reflects the behavior of the β-xyloside on-enzyme, we have performed classical and *ab initio* QM/MM molecular dynamics simulations on the *Tr*GH11 structure reported by Wan *et al.*,[166] which corresponds to the complex of the Glu177Gln enzyme mutant with xylohexaose (Glu177 is the catalytic acid/base residue, see Figure 3.4). As we suspected that the unusual $^{4}C_{1}/^{O}E$ conformation may be due to the mutation of the acid/base residue, we have assessed its effect on the substrate conformation by performing simulations on both the mutant and the WT enzyme (in the last case, we manually reverted the Glu177Gln mutation; see details in section 3.4). Afterwards we have calculated the free energy landscape of β-xyloside in the active site of the WT enzyme.

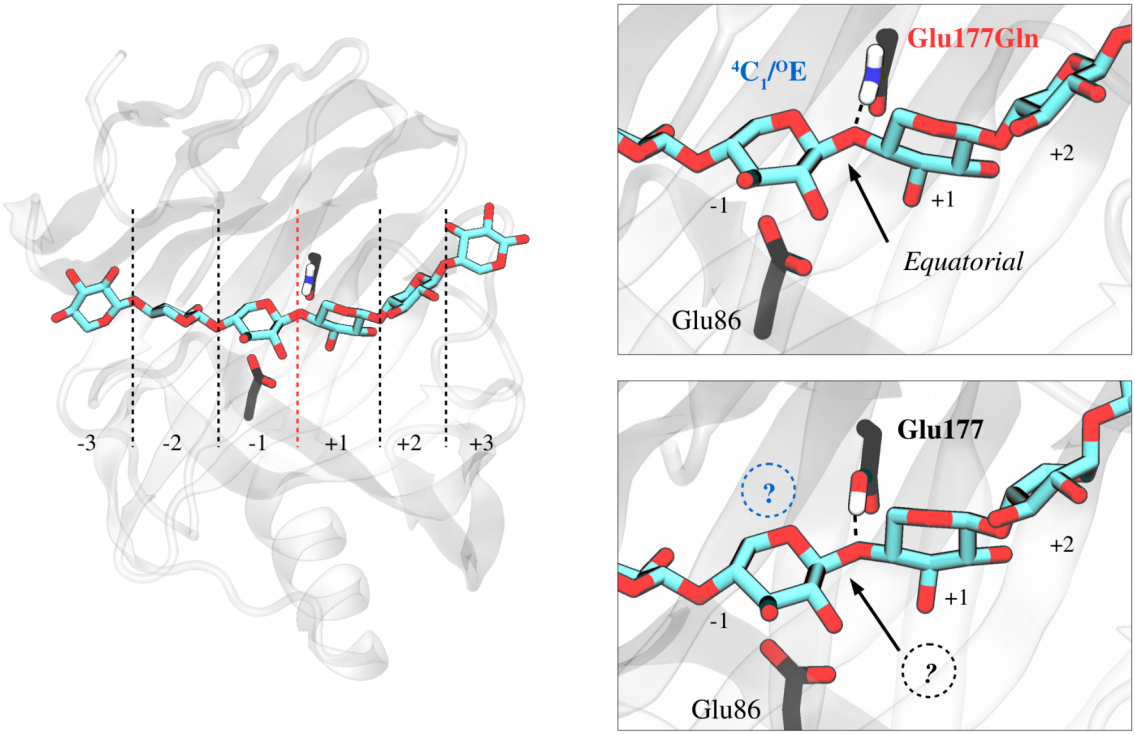


Figure 3.4- Michaelis complex of *Tr*GH11 β-xylanase in complex with a xylohexaose substrate (PDB 4HK8). The β-xyloside is known to exhibit a $^{4}C_{1}/^{O}E$ conformation in the Glu177Gln acid/base mutant (top right), but the conformation in the wild-type system (bottom right), modeled in this work, is still unknown. Glu86 is the catalytic nucleophile residue.

The results of the classical simulations for the enzyme mutant complex show that it fairly reproduced the experimental conformation of the -1 xylose: the sugar ring adopts an intermediate $^{4}C_{1}/^{O}E$ conformation (see Figure 3.5). However, once the mutation was reverted, the conformation of the -1 xylose ring started oscillating between $^{4}C_{1}/^{O}E$ and $^{2}S_{O}$, which is precisely the conformation that has been proposed for GH11 and it is supported by our computed FEL for β-xylose (Figure 3.5).

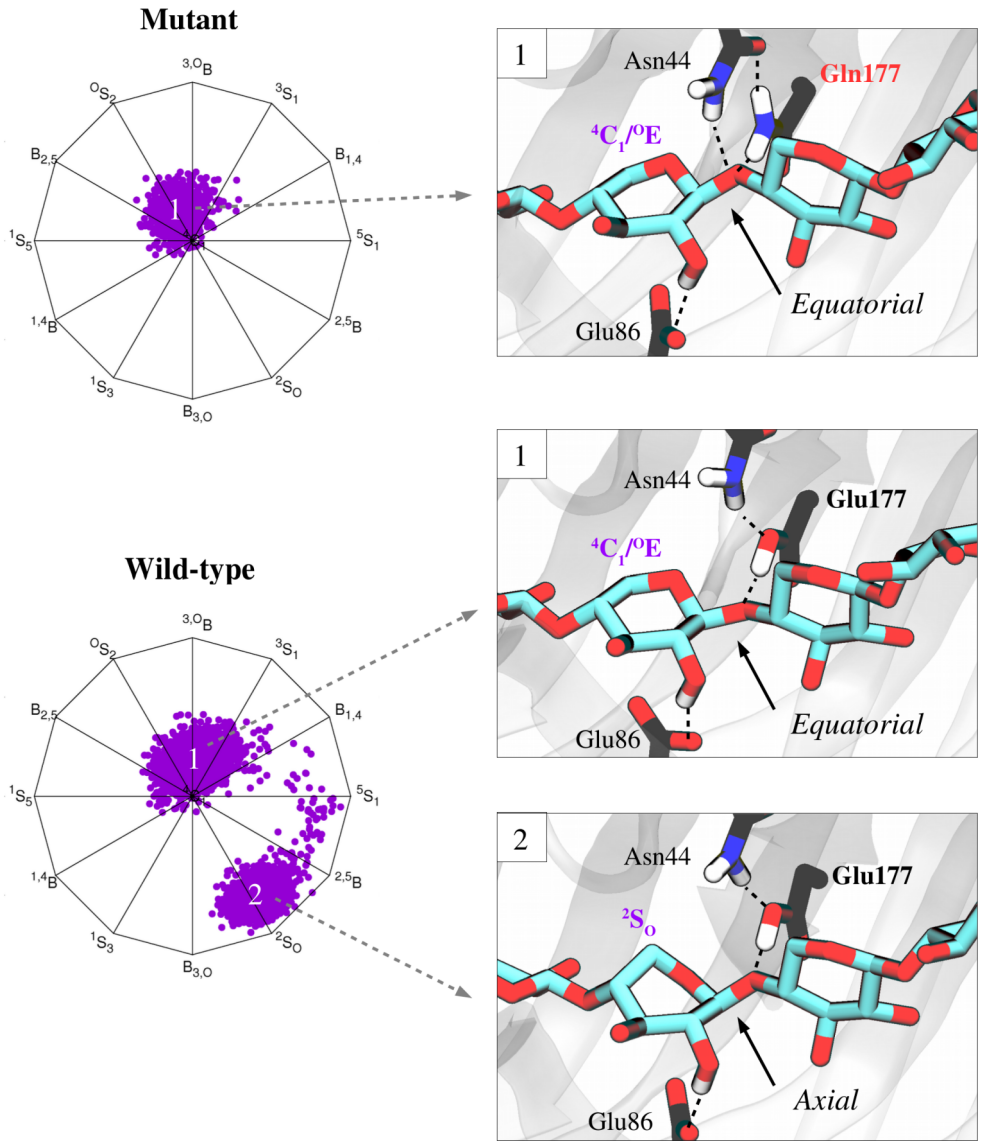


Figure 3.5- Stoddart graphs representing the distribution of conformations for (top) the complex of the β-hexasaccharide with the Glu177Gln mutant of *Tr*GH11 β-xylanase and (bottom) the same complex but with the wild-type enzyme. Notice that the Glu177Gln mutant makes Asn44 change its hydrogen bond network, which results in new interactions with the glycosidic bond that restrict its movement.

This surprising result can be attributed to the hydrogen bond pattern of each system. While in the WT enzyme a neighboring Asn44 interacts with the acid/base residue acting as a hydrogen bond donor, the opposite is found in the mutant enzyme (*i.e.* Asn44 acts as a hydrogen bond acceptor). At the same time, Asn44 acts as a hydrogen bond donor with the glycosidic bond, restricting its movement. In other words, the Asn44 and Gln177 combination acts as a kind of "molecular nippers" that grab the glycosidic bond, keeping the -1 sugar in the $^4C_1/^OE$ conformation by reducing the equato-

rial-axial C-O motion required to evolve from $^{4}C_{1}/^{O}E$ to $^{2}S_{O}$. Therefore, these results demonstrate that *the enzyme mutation affects the sugar conformation* and that the $^{4}C_{1}/^{O}E$ conformation observed in this complex, *a priori*, cannot be taken as informative for the conformational itinerary.

As a control calculation, we have considered an opposite case in which the complex obtained with a modified enzyme conforms with our computed FEL. We have selected a family 10 retaining xylanase from *Streptomyces olivaceoviridis* (*So*GH10, PDB code 2D24),[169] which was crystallized with a natural substrate and a double mutation (Asn127Ser and Glu128His, with Glu128 being the acid/base residue). In this structure, the -1 xylose displays a $^{1}S_{3}$ conformation that agrees with the $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ itinerary, covering a low-energy region of the FEL. However, can we trust this conformation taking into account that –as we have just seen– mutations can affect it?

To check for this concern, we have followed the same strategy as in the previous case, carrying out simulations for both the mutant and WT systems. Strikingly, in this case, the results show that there are two stable substrate conformations in the active site ($^{1}S_{3}/B_{3,O}$ and $^{4}C_{1}/^{4}E$) and these are independent of the mutations, *i.e.* both the modified and the wild type enzyme behave similarly with respect to the conformation of the xyloside at the -1 subsite (see Figure 3.6). Therefore, in this case the complex of the modified enzyme with its natural substrate is a good mimic of the WT enzyme complex (the "true" Michaelis complex), being in the pathway towards the transition state. In contrast, the complex of modified *Tr*GH11 xylanase with its natural substrate is not a good mimic of the WT Michaelis complex, and the $^{4}C_{1}/^{O}E$ conformation observed in this complex should not be taken as informative for the conformational itinerary .

At this point, we know that two conformations ($^{4}C_{1}/^{O}E$ and $^{2}S_{O}$) of the -1 xylose can accommodate in the GH11 enzyme active site, but we do not know if the result has been influenced by the force-field employed in the MD simulations –they usually overestimate the stability of $^{4}C_{1}$ conformations, so perhaps the $^{4}C_{1}/^{O}E$ conformation should not be present in the WT enzyme– and, in case that both conformations exist, we do not know which one is more stable. To exclude artifacts coming from force-field limitations, we have re-evaluated these results by *ab initio* QM/MM molecular dynamics simulations. The QM region included a total of 66/67 QM atoms (for the WT/mutated forms of the enzyme), including the xylose rings at the -1 and +1 subsites, half rings of the saccharides at the -2 and +2 subsites and the acid/base residue.

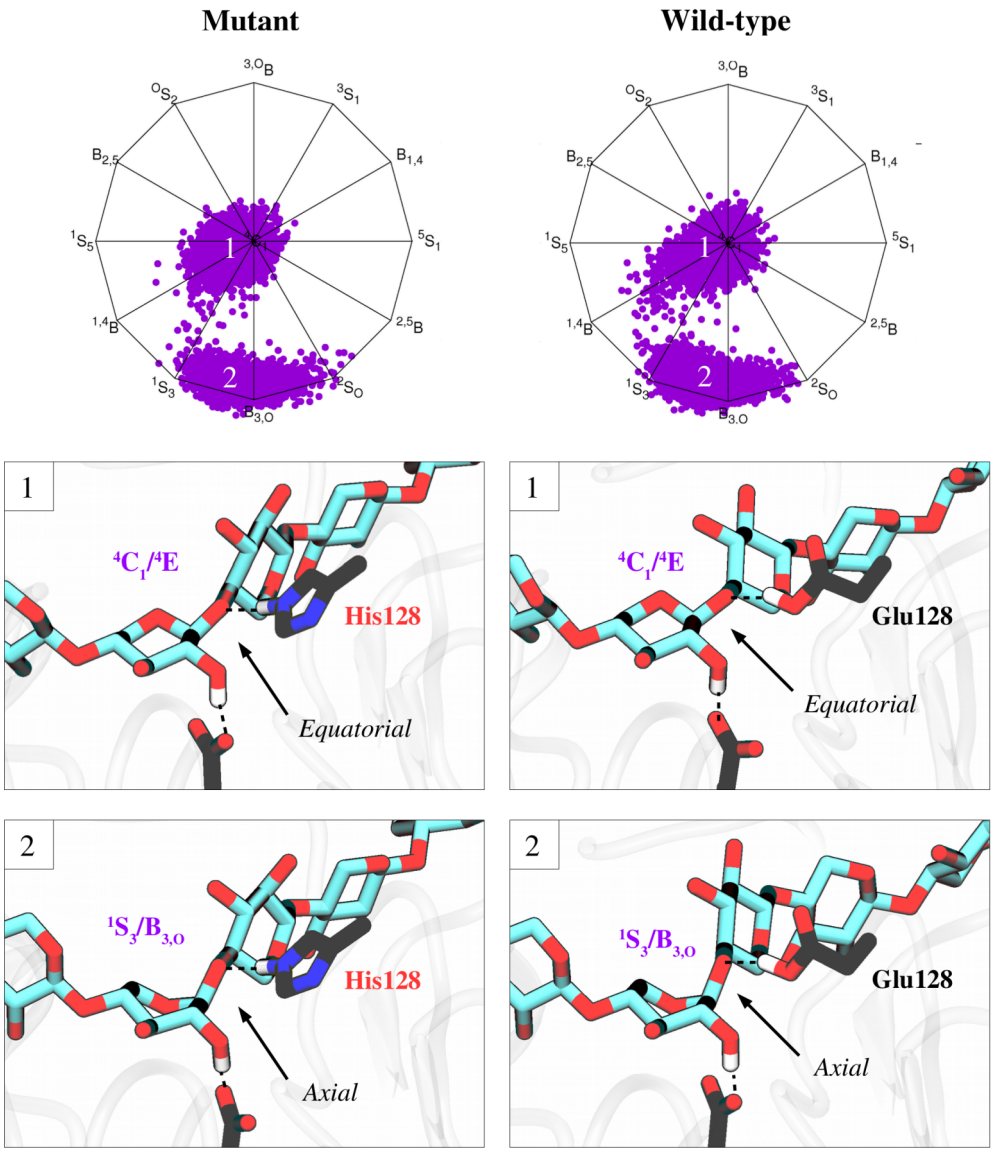


Figure 3.6- Stoddart graphs representing the distribution of conformations for (left) the complex of the β-pentasaccharide with the Asn127Ser and Glu128His double mutant of *So*GH10 enzyme and (right) the same complex but with the wild-type enzyme. The Asn127 residue is not shown for clarity.

Analysis of the QM/MM MD trajectories show that both conformations are stable during the equilibration (>2 ps), reinforcing the results obtained by the classical simulations. Subsequently, we have determined the stability of each conformation by computing the conformational free energy landscape on-enzyme, using QM/MM metadynamics and q_x and q_y "Stoddart" coordinates as collective variables. The FEL shows two available minima that agree with the classical pattern (see Figure 3.7). Both $^4C_1/^OE$ and ^2S_O conformations have similar energies, with ^2S_O being slightly higher with respect to $^4C_1/^OE$ (by ~2 kcal·mol^{-1}).

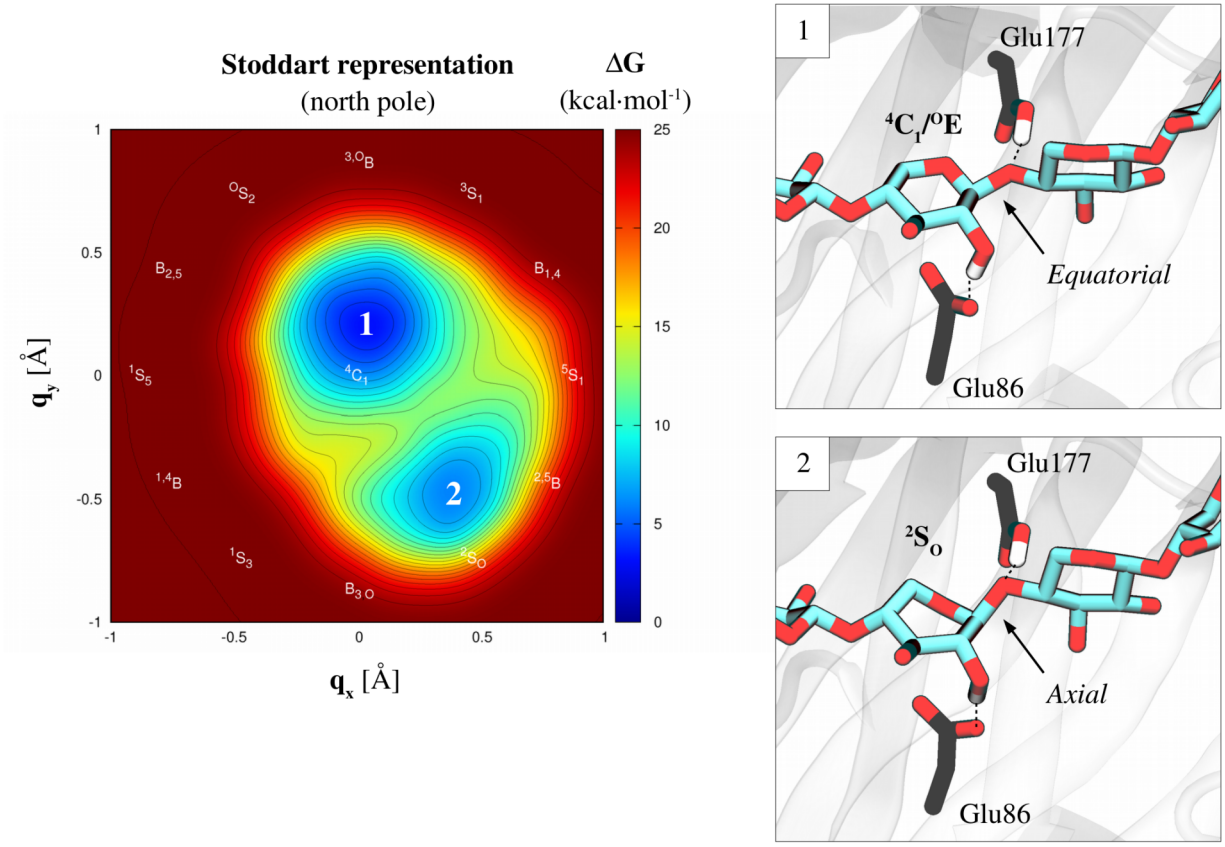


Figure 3.7- Computed QM/MM free energy landscape of β-xylose (Stoddart representation) in the active site of *Tr*GH11 β-xylanase (WT). Energy values are given in kcal·mol^{-1} and each contour line of the diagram corresponds to 1 kcal·mol^{-1}.

The fact that a practically undistorted conformation ($^{4}C_{1}/^{O}E$) is the most stable in the active site of a GH is unprecedented. It is usually assumed that the most populated conformation should be preactivated for catalysis, which is not the case for $^{4}C_{1}/^{O}E$. First, it displays an equatorial orientation of the glycosidic bond (see Figure 3.7). This is reflected in the O_{Nuc}-C1-O_{Gly} angle, which is notably more "closed" –with respect to the optimal 180º– for the $^{4}C_{1}/^{O}E$ conformation than for the $^{2}S_{O}$ one (by 20.1º, see Table 3.1), meaning that the former is less in-line for a proper S_N2 displacement. Second, the $^{2}S_{O}$ conformation shows the typical shortening of the intra-pyranic C1-O5 distance (by 0.05 Å) and lengthening of the glycosidic bond (by 0.07 Å), which makes the substrate to be "on the pathway" to the oxocarbenium ion-like transition state. Finally, the $^{2}S_{O}$ conformation exhibits a significant charge transfer from the pyranic oxygen to the "anomeric center" (0.16 e), reflecting the polarization of the glycosidic bond induced by hyperconjugation.

Table 3.1- Structural and electronic properties of the two possible conformations of β-xylose in the active site of *Tr*GH11 β-xylanase (WT). Atomic ESP charges are provided.

Properties	$^4C_1/^OE$	^2S_O	$\Delta(^2S_O-^4C_1/^OE)$
O_{Nuc}-C1-O_{Gly} angle (º)	135.8 ± 7.4	155.9 ± 6.6	20.1
C1-O5 distance (Å)	1.44 ± 0.04	1.39 ± 0.03	-0.05
C1-O_{Gly} distance (Å)	1.45 ± 0.03	1.52 ± 0.04	0.07
O5 charge (e)	-0.52 ± 0.02	-0.36 ± 0.04	0.16

Altogether, these stereoelectronic properties suggest that the ^2S_O conformation would be more reactive than the $^4C_1/^OE$ one, but this contrasts with the more stable conformation inside the enzyme, which is $^4C_1/^OE$. Does it mean that the $^4C_1/^OE$ conformation is catalytically faster than the ^2S_O? Do the ~2 kcal·mol^{-1} of binding stability in $^4C_1/^OE$ compensate for the preactivation advantages of ^2S_O? The answer to these questions is enclosed in the reaction mechanism of the enzyme.

3.2.3 Fast and distorted: evidence for a canonical $^2S_O \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^5S_1$ itinerary

We have computed the glycosylation reaction mechanism starting from both $^4C_1/^OE$ and ^2S_O conformations, using QM/MM metadynamics and two collective variables (CV1 and CV2) that include all bonds that are formed and cleaved during the chemical event. CV1 involves the nucleophilic attack of Glu86 and the glycosidic bond breakage, and CV2 the proton transfer between the acid/base residue (Glu 177) and the leaving group (see further details in section 3.4). The corresponding reaction mechanisms and FELs are shown in Figures 3.8 and 3.9.

The two reaction mechanisms consist in a one-step concerted S_N2 displacement, with the transition state characterized by having the glycosidic bond already broken, the glycosyl-enzyme bond not yet formed and the proton of the acid/base transferred to the leaving group (see Figures 3.8 and 3.9). The most important aspect to emphasize is that the two reaction pathways lead to the same glycosyl-enzyme intermediate (GEI), displaying a $^{2,5}B$ conformation. This result has two consequences: (i) it completely excludes the unusual $^OE \rightarrow [^OS_2]^{\ddagger} \rightarrow B_{2,5}$ catalytic itinerary proposed on the basis of the X-ray structure of the mutant enzyme; and (ii) it connects the two independent simulations as they both explore the same region of the space, so we can relate their free energy landscapes by a thermodynamic cycle (see Figure 3.9).

Another relevant aspect to point out is that the two pathways select very different transition state conformations, with the reaction starting from ^2S_O reaching a $^2S_O/^{2,5}B$ transition state (TS in Figure 3.8) and the $^4C_1/^OE$ reaching a $^2H_3$ transition state (TS').

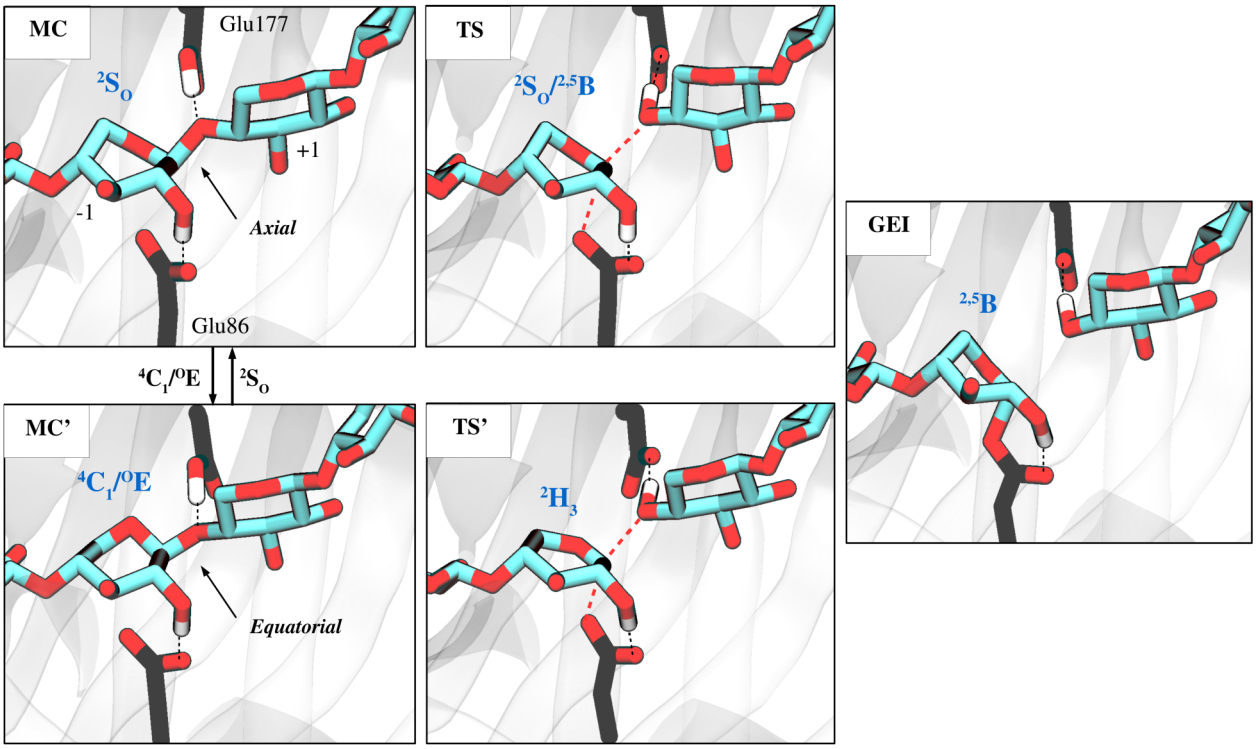


Figure 3.8- Close view of the catalytic center –subsites -1 and +1– along the two reaction mechanisms in *Tr*GH11: (top) starting from the $^{2}S_{O}$ conformation and (bottom) from the $^{4}C_{1}/^{O}E$ conformation. Sugar conformations are highlighted in blue. Red dashed lines indicate bonds that are being formed or broken. Most hydrogen atoms have been omitted for clarity.

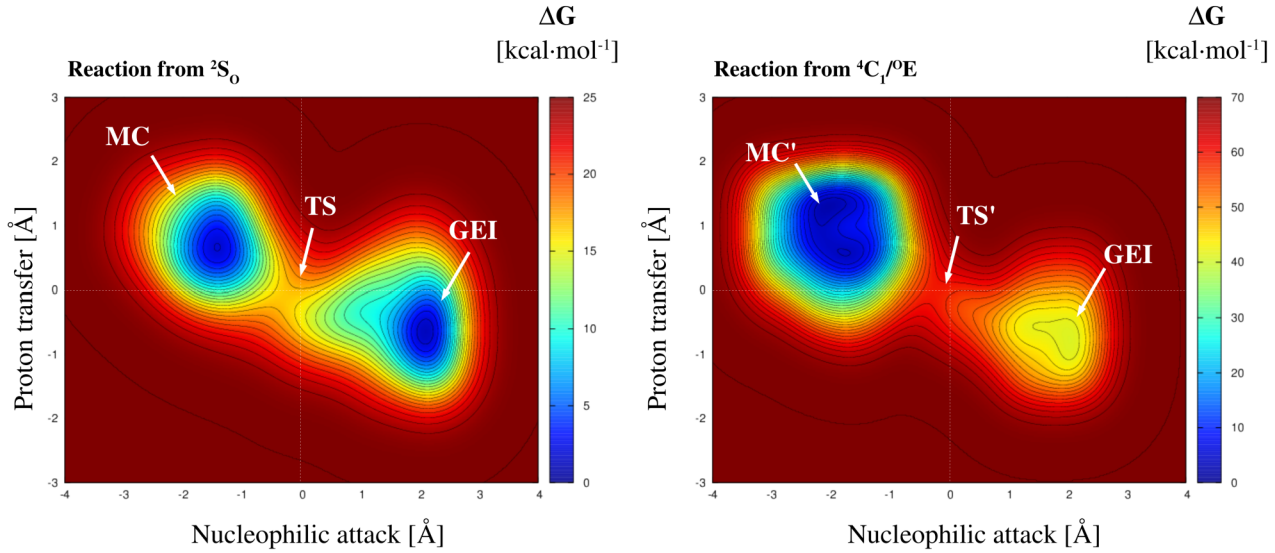


Figure 3.9- Reaction free energy landscapes for the two reaction mechanisms in *Tr*GH11: (left) starting from the $^{2}S_{O}$ conformation and (right) from the $^{4}C_{1}/^{O}E$ conformation. Contour lines correspond to 1 kcal·mol^{-1} and 2 kcal·mol^{-1} respectively (note the difference in energetic scales at the side bar). Reaction free energy barriers are provided in Table 3.2. Exploration of the GEI in the reaction starting from $^{4}C_{1}/^{O}E$ was considered not necessary because the reaction starting from $^{2}S_{O}$ already explored it.

The conformational difference between the two TS is crucial, as the $^{2}H_{3}$ conformation is non-planar around the anomeric center (C2-C1-O5-C5 atoms) and it can not optimally stabilize the oxocarbenium ion. This is evidenced by the comparison of the charge at the anomeric carbon between the two transition states, being notably more positive for the $^{2}H_{3}$ TS (see Table 3.2). Moreover, this conformation is not stable even at the Michaelis complex (by ~10 kcal·mol^{-1} according to the FEL in Figure 3.7), and it is expected that it would be less stable at the TS. Therefore, two reasons are behind the high energy observed for the reaction pathway that starts from the $^{4}C_{1}/^{O}E$ conformation: (i) the steric hindrance at the Michaelis complex, in which the glycosidic bond is in an equatorial position and difficults the nucleophilic attack; and (ii) the $^{2}H_{3}$ TS conformation that it has to take along the catalytic pathway, which is unfavored both by its shape and by the inability to stabilize properly the oxocarbenium ion. These two reasons make the $^{4}C_{1}/^{O}E$ conformation be 42 kcal·mol^{-1} less reactive than the preactivated $^{2}S_{O}$ one (see Table 3.2).

Table 3.2- Computed free energy barriers (kcal·mol^{-1}), transition state conformations and their values for the C2-C1-O5-C5 dihedral angle (measure of the planarity around the anomeric carbon) and the ESP charges of the anomeric carbon.

Properties	$^{4}C_{1}/^{O}E$	$^{2}S_{O}$	$\Delta(^{2}S_{O}-^{4}C_{1}/^{O}E)$
$\Delta G^{\ddagger}_{MC\rightarrow GEI}$	56.6[a]	14.6	-42.0
$\Delta G^{\ddagger}_{GEI\rightarrow MC}$	55.2[b]	15.2	-40.0
TS conformation	$^{2}H_{3}$	$^{2}S_{O}/^{2,5}B$	-
C2-C1-O5-C5 angle (º)	-31.2 ± 4.1	3.9 ± 3.7	35.2
Anomeric charge (e)	0.48 ± 0.05	0.29 ± 0.05	-0.18

a) An additional simulation with 3 collective variables, splitting the nucleophilic attack in two different CVs, shows that this barrier is >40 kcal·mol^{-1} b) This energy barrier is estimated by a thermodynamic cycle, taking into account the conformational FEL that is shown in Figure 3.7 and the reaction FEL of the $^{2}S_{O}$ conformation shown in Figure 3.9.

To further verify the lowest free energy pathway, we have performed an additional simulation starting from the glycosyl-enzyme intermediate (GEI). From this point, the system will evolve freely either to $^{4}C_{1}/^{O}E$ or $^{2}S_{O}$ as we do not impose any particular conformation for the Michaelis complex. Remarkably, during the metadynamics simulation we observe that the system leaves the GEI basin through the $^{2}S_{O}/^{2,5}B$ transition state, reaching the $^{2}S_{O}$ conformation at the Michaelis complex. This result, consistent with the reaction pathway that we have previously found, clearly demonstrates that the most favored pathway involves a $^{2}S_{O}/^{2,5}B$ transition state, which matches with a canonical $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$ catalytic itinerary. Altogether, we can conclude that the mechanism of *Tr*GH11 β-xylanase starts from a pre-Michaelis complex displaying a non-reactive $^{4}C_{1}/^{O}E$ con-

formation, changes to a preactivated $^{2}S_{O}$ conformation that is 2.0 kcal·mol^{-1} higher in energy and subsequently reacts via a $^{2}S_{O}/^{2,5}B$ transition state with a free energy barrier of 14.6 kcal·mol^{-1}, reaching a $^{2,5}B$ glycosyl-enzyme intermediate. The global free energy barrier, 16.6 kcal·mol^{-1}, is commensurate with the kinetic values of β-xylosidases.[170]

3.3 Summary and Conclusions

In this chapter we have addressed two principal objectives: (i) decipher the catalytic itineraries of β-xylosidases and (ii) obtain a computational proof for the kinetic advantage of sugar distortions in GHs. These objectives have been achieved in different steps. First, we have computed the free energy landscape of an isolated β-xylose, showing that only two of the three proposed itineraries ($^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ and $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$) fit with the low-energy regions of the landscape. A third itinerary, $^{O}E \rightarrow [^{O}S_{2}]^{\ddagger} \rightarrow B_{2,5}$, proposed on the basis of a GH11 X-ray structure, has been found to be unfavorable as it crosses a high energy region of the landscape.

In a second step, we have further analyzed the GH11 structure that led to the assignment of the anomalous itinerary, performing classical MD simulations. These simulations have showed that the mutation needed to trap the X-ray structure perturbs the conformation of the -1 sugar, which probably caused the wrong experimental prediction of the itinerary, adding a word of caution on using modified enzymes to inform on catalytic itineraries. Moreover, these simulations have suggested that two conformations ($^{4}C_{1}/^{O}E$ and $^{2}S_{O}$) of the -1 xylose can be accommodated in the active site of the enzyme, which has been further verified by QM/MM techniques.

Finally, we have computed the reaction mechanisms starting from the two conformations, showing that even though the $^{4}C_{1}/^{O}E$ conformation is 2 kcal·mol^{-1} more stable than the $^{2}S_{O}$, it is catalytically incompetent $-\Delta\Delta G^{\ddagger} = 42$ kcal·mol^{-1}– due to the steric hindrance of the equatorial scissile bond and the fact that it involves an unfavorable TS conformation ($^{2}H_{3}$). An enzyme complex displaying this conformation, hence, has to be considered as a pre-Michaelis complex. The "true" Michaelis complex displays a $^{2}S_{O}$ conformation and reacts through a $^{2}S_{O}/^{2,5}B$ transition state with a barrier of 14.6 kcal·mol^{-1}. Altogether, the following conclusions can be drawn from this chapter:

- The free energy landscape of β-xylose suggest that β-xylosidases follow only two catalytic itineraries: $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ or $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$. The unusual $^{O}E \rightarrow [^{O}S_{2}]^{\ddagger} \rightarrow B_{2,5}$ itinerary, predicted on the basis of X-ray experiments on a modified structure, is not compatible with the

computed FEL and arises from a conformation ($^{4}C_{1}/^{O}E$) that is a consequence of the Glu177Gln acid/base mutation.

- GHs restrict the conformational space of the isolated sugar to be more selective. As we have seen for GH11, the enzyme completely blocks the low-energy regions that cover the $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ itinerary (employed by GH10), defining a unique $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$ itinerary. Even though the restrictions, still two conformations ($^{4}C_{1}/^{O}E$ and $^{2}S_{O}$) can accommodate inside the enzyme, highlighting a local flexibility that can be attributed to the lack of the hydroxymethil moiety in β-xylose, which limits the possibilities of the enzyme for stabilizing a particular conformation by the use of specific interactions.

- GHs distort their -1 sugars to enhance catalysis. In the particular case of the *Tr*GH11 β-xylosidase, the catalytic enhancement between $^{4}C_{1}/^{O}E$ and $^{2}S_{O}$ conformations corresponds to ~42 kcal·mol^{-1} in terms of reaction free energy barrier. The kinetic advantage of distorted conformations is mainly due to low steric clashes with the nucleophile, prepared for an in-line $S_{N}2$ attack, and their "stereoelectronic similarity" to conformations that are favored for an oxocarbenium ion-like TS.

- GHs do not necessarily need to bind the -1 sugar in its most reactive conformation, but a conformational change to a distorted conformation is mandatory for the reaction. In the case of *Tr*GH11 β-xylosidase, the $^{4}C_{1}/^{O}E$ conformation is less reactive than the $^{2}S_{O}$ one, but at the same time it is ~2 kcal·mol^{-1} more stable.

- The FEL of isolated sugars provide a simple and effective approach to unveil catalytic itineraries. The FEL of the isolated β-xylose computed here does not only cost ~28.000 cpu·hours less than the one inside the GH11 β-xylanase, but also gives mechanistic information on β-xylosidases in general.

3.4 Computational Details

3.4.1 Modeling of the isolated β-xylose

The conformational free energy landcape (FEL) of the isolated β-xylose has been obtained by *ab initio* metadynamics simulations,[139,171] performed at room temperature within the Car–Parrinello approach,[107] as implemented in the CPMD 3.15.1 program.[172] The electronic structure has been computed within the density functional theory (DFT), using the Perdew, Burke, and Ernzerhoff generalized gradient-corrected approximation (PBE).[173] This functional gave reliable results in previous Car–Parrinello simulations of isolated carbohydrates and GHs.[52,53,126,157] In particular, the error on relative energies due to the DFT functional employed is ±0.6 kcal·mol^{-1} for β-glucose.[52] Kohn–Sham orbitals have been expanded in a plane wave (PW) basis set with a kinetic energy cutoff of 70 Ry. Norm-conserving pseudopotentials have been employed, generated within the Troullier–Martins scheme.[174] The isolated β-xylose has been enclosed in an orthorhombic box of size 12.5 Å x 13.5 Å 11.6 Å. The fictitious mass for the electronic degrees of freedom has been set to 850 au.

The puckering θ and φ polar coordinates have been used as collective variables.[144] The height of the Gaussian terms has been set to 0.18 kcal·mol^{-1}, which ensures sufficient accuracy for the reconstruction of the FEL. The width of the Gaussian terms has been set to 0.1 CV units, according to the oscillations of the selected collective variables observed in the free dynamics. A direct implementation of the MTD algorithm has been used to obtain the Mercator FEL. A new Gaussian potential has been added every 200–400 MD steps at the first stage of the simulation, increasing it up to 1000 MD steps to ensure a proper convergence of the simulation. A total number of 4500 Gaussian functions have been added to completely explore the free energy landscape. The convergence of the metadynamics simulations has been assessed by checking the invariance of energy differences and the free energy landscape with the progression of the simulation, following the work of Tiwary 2015.[175] The convergence error in the energies is found to be lower than 0.5 kcal·mol^{-1}. These simulations have been performed by Dr. Javier Iglesias, co-author of the manuscript in which we have published parts of the results of this chapter.

3.4.2 Modeling of the Michaelis complex of *Tr*GH11 β-xylanase

The initial structure for the simulations has been taken from the recently reported structure of *Tr*GH11 β-xylanase in complex with a xyloglucan hexasaccharide (PDB: 4HK8). In this structure, the acid/base Glu residue is mutated to glutamine (Glu177Gln) and the xylose saccharide at the -1 subsite adopts a $^{4}C_{1}/^{O}E$ conformation. Two separate systems have been modeled, with and without

the mutation of the acid/base residue (in the last case, the mutation has been manually reverted). For each system, two simulations have been performed, one with the sugar initially in a $^4C_1/^OE$ conformation and the other with the sugar in the ^2S_O conformation (the last one has been obtained by a restrained relaxation). The protonation states and hydrogen atom positions of all amino acid residues have been taken from the crystal structure, except His155, which has been changed from double to single protonation due to the close contact with Ser139. All crystallographic water molecules have been retained and extra water molecules have been added to form a 15 Å water box around the protein surface. Five chloride ions have been also added to neutralize the enzyme charge.

Molecular dynamics (MD) simulations using Amber11 software[176] have been performed. The protein has been modeled with the FF99SB force-field,[177] the carbohydrate substrate with the GLYCAM06 force-field[178] and water molecules were described with the TIP3P force -field.[113] The MD simulations have been carried out in several steps. First, the systems has been minimized, keeping the protein and substrate fixed. Then, the entire systems have been allowed to relax. To gradually reach the desired temperature of 300 K in the MD simulations, weak spatial constraints have been initially added to the protein and substrate, while the water molecules and chloride ions have been allowed to move freely at 100K. The constraints have been then removed and the working temperature of 300 K has been reached after two more 100K heatings in the NVT ensemble. Afterwards, densities have been converged up to water density at 300K in the NPT ensemble. The simulations have been further extended to 18 ns, when equilibration has been reached. In the case of the simulations of the wild type (WT) enzyme, the acid/base residue (Glu177) has been restrained (after reverting the Glu→Gln mutation) for the first 15 ns. Analysis of the trajectories has been carried out using standard tools of AMBER and VMD.[179]

A snapshot of the $^4C_1/^OE$ and ^2S_O conformations have been taken from the WT MD-equilibrated structures for subsequent QM/MM calculations. These calculations have been performed using the method developed by Laio *et al.*,[133] which combines Car–Parrinello MD with force-field MD methodology. The QM/MM interface has been modeled by the use of link-atoms that saturate the QM region. The electrostatic interactions between the QM and MM regions have been handled via a fully Hamiltonian coupling scheme, where the short-range electrostatic interactions between the QM and the MM regions are explicitly taken into account for all atoms. An appropriately modified Coulomb potential has been used to ensure that no unphysical escape of the electronic density from the QM to the MM region occurs. The electrostatic interactions with the more distant MM atoms have been treated via a multipole expansion. Bonded and van der Waals interactions between the

QM and the MM regions have been treated with the standard AMBER force field. Long-range electrostatic interactions between MM atoms have been described with the P3M implementation,[180] using a 64 Å x 64 Å x 64 Å mesh.

The QM region for studying the reaction pathways has included the xylose sugars at the -1 and +1 subsites, half sugars of the saccharides at the -2 and +2 subsites and the acid/base and nucleophile residues, leading to a total number of 76 QM atoms and 40.337 MM atoms. This QM region has been enclosed in an isolated supercell of size 21.2 Å x 18.5 Å x 18.5 Å. For the puckering study, the region has been reduced to the xylose sugar at -1 and half sugars of -2 and +1 subsites, leading to a total number of 38 QM atoms and 40.375 MM atoms. In this case, the QM region has been enclosed in an isolated supercell of size 20.1 Å x 14.1 Å x 13.4 Å. Kohn–Sham orbitals have been expanded in a plane wave basis set with a kinetic energy cutoff of 70 Ry. The fictitious mass for the electronic degrees of freedom was set to 700 au. Norm-conserving Troullier–Martins ab initio pseudopotentials were used for all elements. The calculations were performed using the Perdew, Burke and Ernzerhoff generalized gradient-corrected approximation (PBE). This functional form has been proven to give a good performance in the description of hydrogen bonds and was already used with success in previous works on glycoside hydrolases and transferases.[50,93,181] All systems have been equilibrated between 2.5 and 3.8 ps until they became stable according to the RMSD of the QM atoms.

The conformational free energy landscape given in Figure 3.7 has been explored using the well-tempered metadynamics –a variant of metadynamics in which the height of the Gaussians diminish during the simulation– approach with two puckering collective variables (q_x and q_y).[182] We have used the metadynamics driver provided by the Plumed2 plugin.[183] A hill height of 1 kcal·mol^{-1} (bias-factor of 10 kcal·mol^{-1}, T=300 K) and a hill width of 0.1 rad for each variable have been used to define the repulsive potentials. The deposition time has been set to 24 fs (200 MD steps). The FES has been completed after 1279 deposited Gaussians (~36 ps of simulation). The error between the two explored basins has been found to be of ±0.6 kcal·mol^{-1} according to the standard deviation from the last 18 ps.

The two reaction free energy landscapes given in Figure 3.9, the one starting from the $^4C_1/^OE$ conformation and the one from ^2S_O, have been explored using the lagrangian version of metadynamics (20.0 amu for the mass of the fictitious particle and 0.2 au for the force constant) with two collective variables. The first one (CV1) has been defined as the difference between the O_{Nuc}-C1 and the C1-O_{Gly} distances. This variable accounts for the nucleophilic attack of Glu86 and the cleavage

of the glycosidic bond. The second collective variable (CV2) has been defined as the distance difference between the $O_{A/B}$-H and H-O_{Gly} distances. This variable accounts for proton transfer between Glu177 and the glycosidic oxygen (leaving group). A hill height of 0.63 kcal·mol^{-1} (0.001 a.u) and a deposition time of 30 fs (250 MD steps) have been used in both cases. Spherical hills of 0.34 Å and 0.42 Å width have been used for the $^4C_1/^OE$ and the ^2S_O simulations, respectively, according to the fluctuations of the variables along a non-biased dynamics.

The simulation starting from the $^4C_1/^OE$ conformation have been stopped after the deposition of 912 Gaussians, spanning a time window of ~32 ps. The one from the ^2S_O conformation has been stopped according to the first-crossing criterion,[124] after the deposition of 238 Gaussians and a total simulation time of ~10 ps. The additional simulation starting from the glycosyl-enzyme intermediate has been done taking a snapshot obtained from the $^4C_1/^OE$ simulation. The same parameters as the ones given previously have been taken in this case. The simulation has been stopped after the reaction event, with 90 deposited Gaussians and ~3 ps of simulation. To further verify the high energy barrier obtained for the pathway that starts from the $^4C_1/^OE$ conformation, another simulation has been done by splitting the "nucleophilic attack" collective variable in two independent variables: CV1 defined as the distance between O_{Nuc}-C1 (nucleophilic attack) and CV2 the distance between C1-O_{Gly} (leaving group departure). CV3 has been defined as the proton transfer shown above. The hill height and deposition time has been also the same as in the previous case, but the hill widths have been adapted to the new oscillations: 0.36 Å for CV1, 0.08 Å for CV2 (the bond is initially formed, so its oscillations are very small) and 0.28 Å for CV3. Given the computational cost, the simulation has been stopped after the deposition of 780 hills and a total simulation time of ~30 ps, showing a basin of 42 kcal·mol^{-1} deep.

3.4.3 Modeling of the Michaelis complex of *So*GH10 β-xylanase

The initial structure for the simulations has been taken from the reported structure of *So*GH10 β-xylanase in complex with a xylose pentasaccharide (PDB: 2D24). The structure is a double mutant of residues 127 and 128 (N127S and E128H, being 128 the acid/base residue) and the xylose saccharide at the -1 subsite adopts a $^1S_3$ conformation. A similar protocol as in section 3.4.2 has been followed, with two separate systems (with and without mutations) being modeled. The mutated and the missing residues (304 to 312) have been taken from the structure of PDB entry 1ISV. The protonation states of all amino acid residues have been taken according to protein environment. All crystallographic water molecules have been retained and extra water molecules have been added to form

a 15 Å water box around the protein surface. One chloride ion has been also added to neutralize the enzyme charge. MD simulations using the Amber11 software have been performed 20-30 ns (until equilibration of the protein backbone RMSD has been reached).

3.5 Supplementary Figures

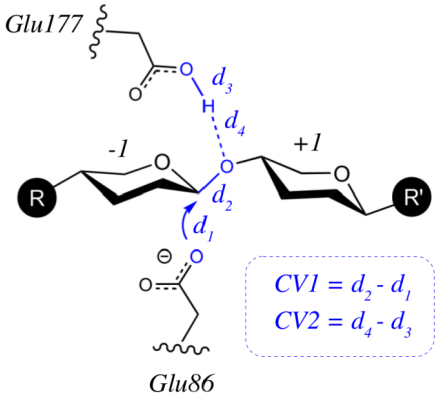


Figure S3.1- Collective variables used to study the reaction mechanisms: the nucleophilic attack (CV1) and the proton transfer (CV2).

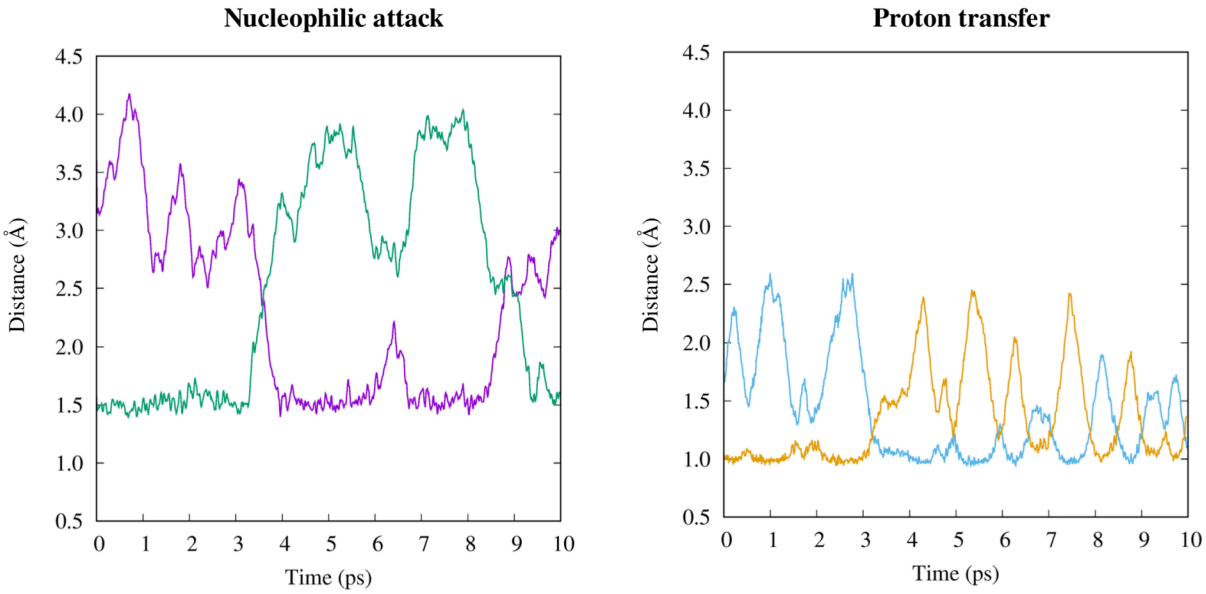


Figure S3.2- Evolution of the distances involved in the nucleophilic attack (left) and proton transfer (right) CVs for the *Tr*GH11 reaction mechanism starting from the ^2S_O conformation. The violet line corresponds to the O_{nuc}-C1 distance (d_1), the green to the C1-O_{Gly} (d_2), the orange to the $O_{A/B}$-H (d_3) and the blue to the H-O_{Gly} distances (d_4).

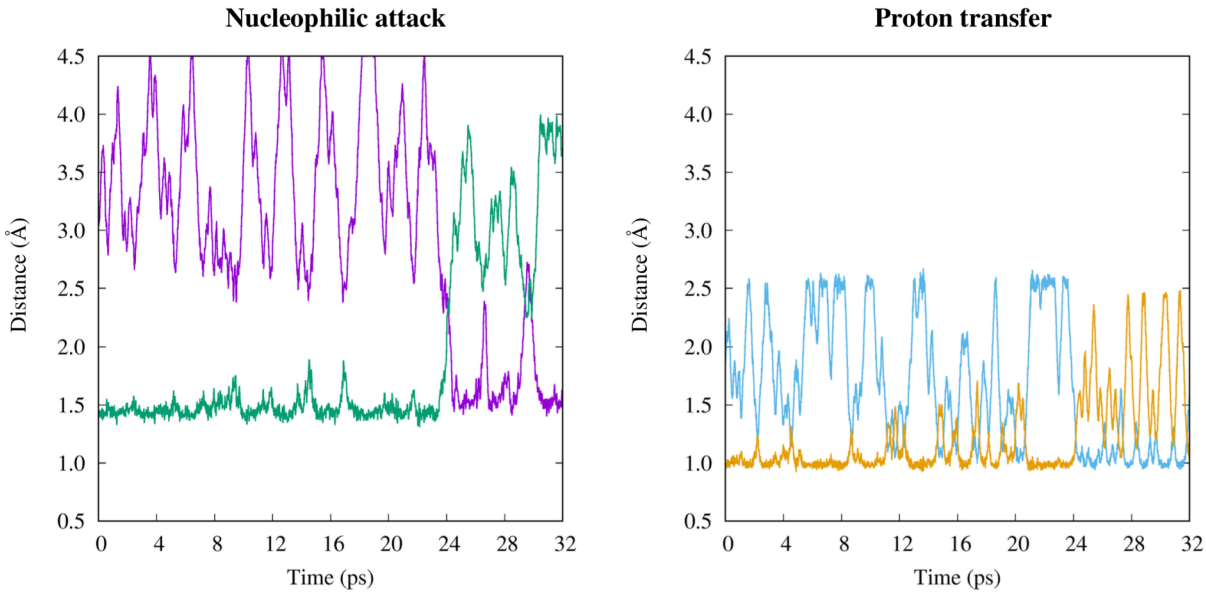


Figure S3.3- Evolution of the distances involved in the nucleophilic attack (left) and proton transfer (right) CVs for the *Tr*GH11 reaction mechanism starting from the $^{4}C_{1}/^{O}E$ conformation. The violet line corresponds to the O_{Nuc}-C1 distance (d_1), the green to the C1-O_{Gly} (d_2), the orange to the $O_{A/B}$-H (d_3) and the blue to the H-O_{Gly} distances (d_4).

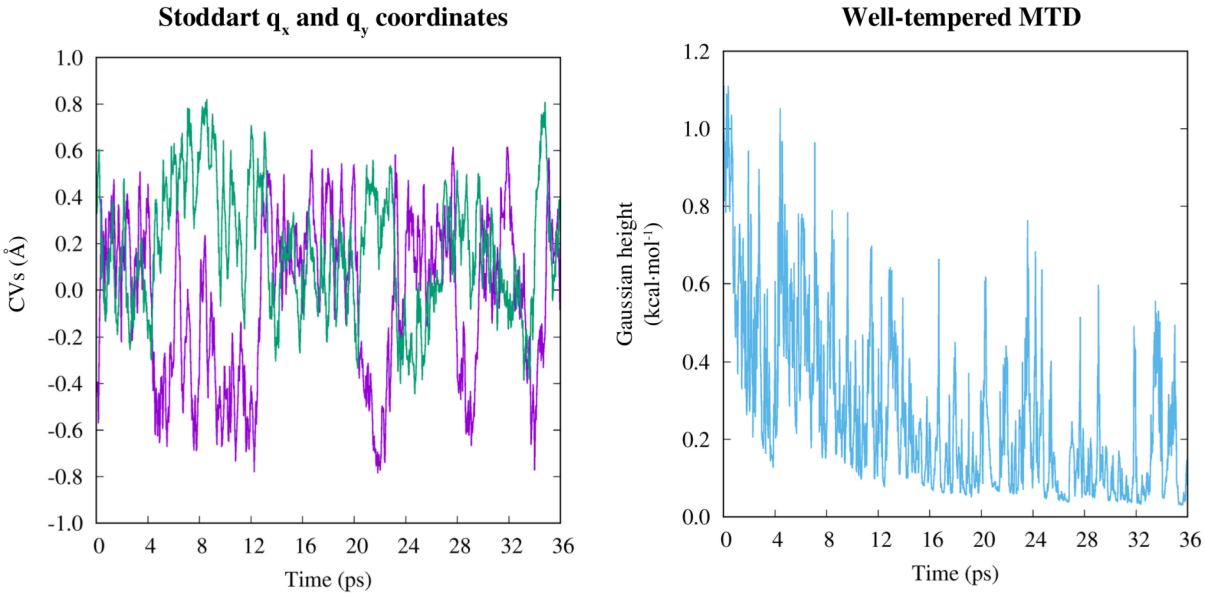


Figure S3.4- Evolution of the Stoddart q_x and q_y puckering coordinates (in green and violet, respectively) for the conformational FEL of β-xylose in the active site of *Tr*GH11 β-xylanase (left) and Gaussian height according to the algorithm of well-tempered MTD.

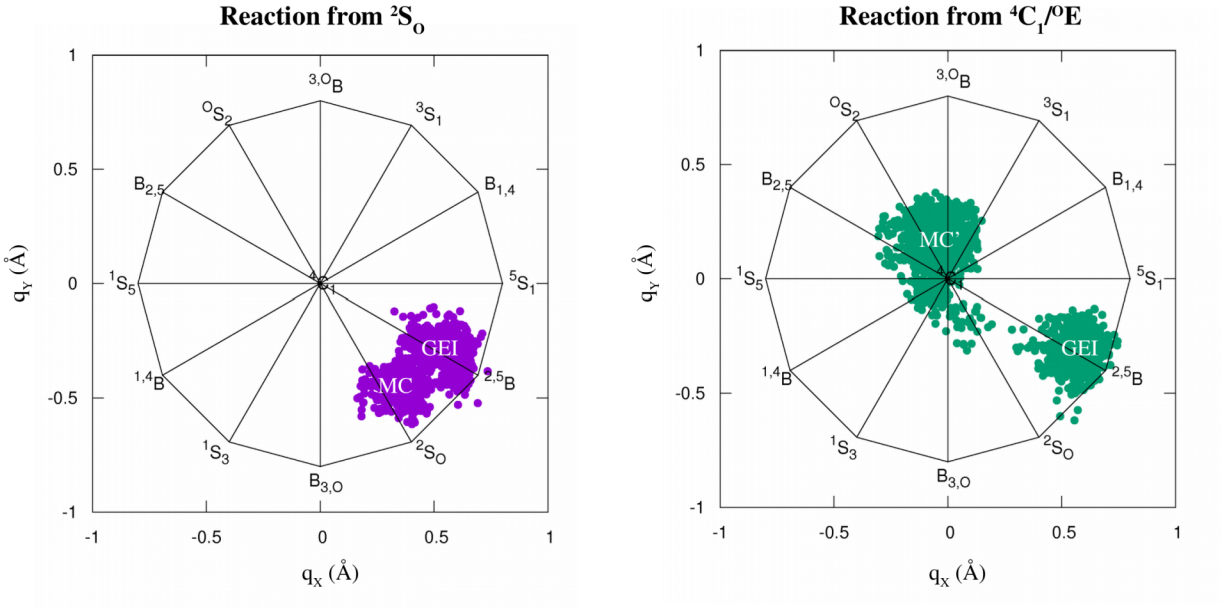


Figure S3.5- Evolution of the q_x and q_y puckering coordinates along the *Tr*GH11 reaction mechanism starting from the $^{2}S_{O}$ (violet) and the $^{4}C_{1}/^{O}E$ (green) conformations. Notice that they start from very different conformations (MC and MC') but they end both exploring the $^{2,5}B$ GEI basin.

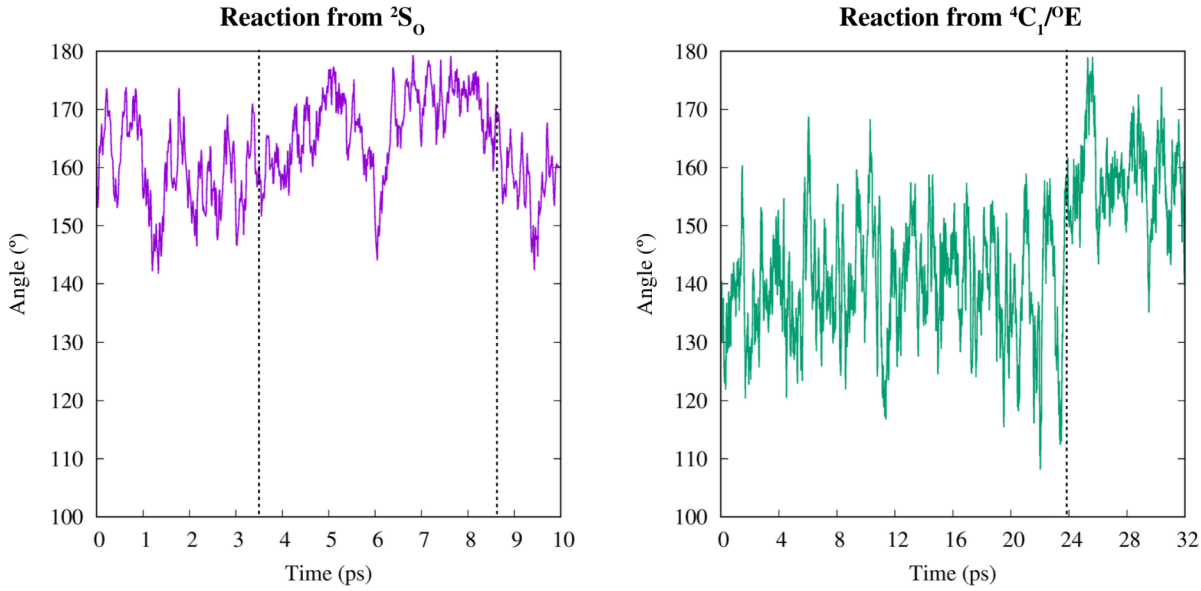


Figure S3.6- Evolution of the O_{Nuc}-C1-O_{Gly} angle along the *Tr*GH11 reaction mechanism starting from the $^{2}S_{O}$ (violet) and the $^{4}C_{1}/^{O}E$ (green) conformations. This angle should be 180º for an ideal S_N2 nucleophilic displacement. Notice that the reaction starting from $^{2}S_{O}$ (violet) is more in-line than the one starting from $^{4}C_{1}/^{O}E$ (green). Moreover, the former does not change along the reaction, while the latest suddenly changes from a mean of 135º to 160º after crossing the transition state (indicated by vertical dashed lines).

Chapter 4
The Contribution of 2-OH Interactions in the Reaction Rate of a β-glucosidase

Parts of this chapter have been published:
L. Raich, V. Borodkin, W. Fang, J. Castro-López, D. M. F. van Aalten, R. Hurtado-Guerrero, C. Rovira. "A trapped covalent intermediate of a glycoside hydrolase on the pathway to transglycosylation. Insights from experiments and QM/MM simulations" *Journal of the American Chemical Society*, **138**, 3325-3332 (2016).

ABSTRACT: in this chapter we study the importance of a crucial hydrogen bond interaction (2-OH···Nucleophile) in a β-glucosidase that displays a high synthetic activity. We demonstrate the existence of two states regarding the 2-OH···Nucleophile interaction (formed or broken) and we reveal its net contribution to the reaction rate and mechanistic outcome, showing that its absence raises free energy barriers up to 16 kcal·mol^{-1} and changes the catalytic itinerary of the substrate, from the expected $^{4}C_{1} \rightarrow [^{4}E]^{\ddagger} \rightarrow {}^{1,4}B/{}^{4}E$ to an unusual $^{4}C_{1} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ cyclic itinerary. Finally, we show how the lack of another interaction involving the 2-OH group (Asn175···2-OH), which is highly conserved in GHs, affects the glycosylation barrier without altering deglycosylation, which led us to propose mutations that could be used to convert GHs into TGs (transglycosylases).

4.1 Introduction

4.1.1 The molecular ties of sugars

Non-covalent interactions are of utmost importance in chemistry and biology. They determine the macroscopic properties of chemical species, the assembly and stability of supramolecular complexes or the three-dimensional fold of proteins and enzymes.[184] One of the most relevant non-covalent interaction is the hydrogen bond, which is characterized by its directional nature and its strong dissociation energies, ranging from 0.2 to 40 kcal·mol^{-1}.[185] These interactions are particularly important in carbohydrates given that they are polyhydroxylated and can form several hydrogen bonds with an enzyme upon binding. For instance, in a β-glucosidase of family 1 GHs,[186] each -1 sugar OH groups presumably form hydrogen bonds with the following residues: Asn165 and Glu352(2-OH); Gln20, His121 and Trp407 (3-OH); Gln20 and Glu406 (4-OH); and Arg325 and Glu406 (6-OH; see Figure 4.1). We write presumably because proton positions are rarely determined by X-ray crystallography (it requires sub-atomic resolutions, *i.e.* below 1 Å, and does not work for highly polarized hydrogen atoms and protons),[187] but they can be intuitively localized from their chemical environment.

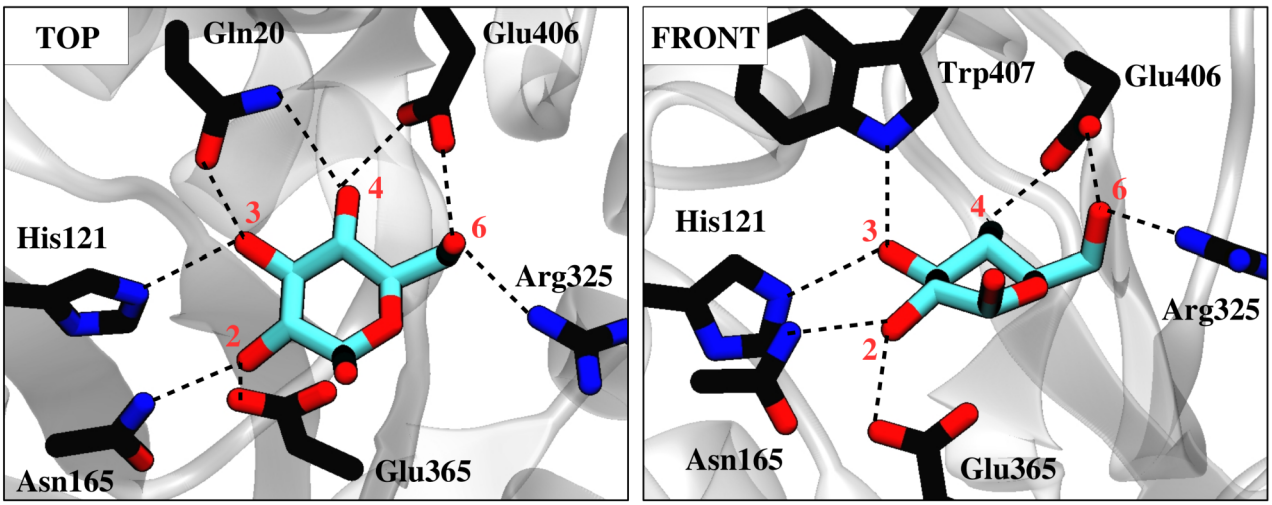
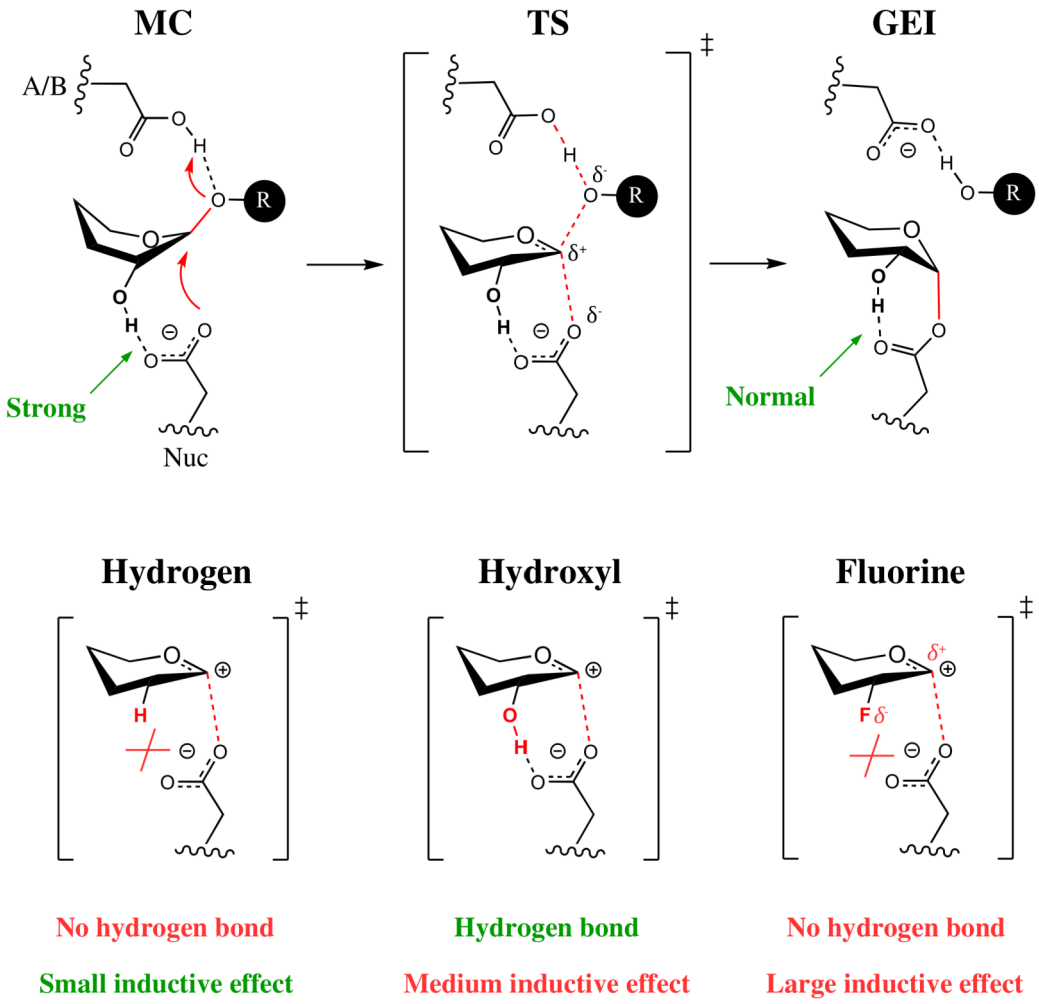


Figure 4.1- Enzyme-substrate hydrogen bond interactions at the -1 subsite of a GH1 β-glucosidase (PDB 3WH6, 1.6 Å resolution). Top and front views of the active site are displayed. Residues Gln20 and Trp407 have been omitted from the front and top views, respectively, for clarity purposes. The glucose substrate is shown in cyan and the enzymatic residues in black. Dashed lines represent hydrogen bond interactions. Substrate hydroxyl groups are numbered in red.

Kinetic experiments on *Agrobacterium faecalis* GH1 β-glucosidase, using modified substrates in which single hydroxyl groups of the -1 sugar were substituted by hydrogen or fluorine, showed that the rate of substrate hydrolysis was notoriously reduced in all cases, evidencing the crucial role of these interactions.[188] Interestingly, while 3-OH, 4-OH and 6-OH interactions were found to contribute up to ~2.4 kcal·mol^{-1} to the reaction free energy barrier, 2-OH interactions raised it up to ~10.8 kcal·mol^{-1}. This substantial difference may be attributed to the fact that the 2-OH hydroxyl interacts with the catalytic nucleophile residue (from now 2-OH···Nucleophile interaction), whose electronic properties change considerably during the reaction: it is negatively charged in the Michaelis complex but neutral at the glycosyl-enzyme intermediate (GEI; see Scheme 4.1 top).

The experimental free energy changes, however, are not direct estimates of the hydrogen bond contribution of each hydroxyl group, but rather indirect because the substitution of a hydroxyl by another functional group has a dual effect: it not only removes the hydrogen bond network of the hydroxyl, which is expected to decrease the reaction rate, but also changes the electronic features of the substrate, which can either increase or decrease the rate depending whether the substituent stabilizes or destabilizes the transition state by inductive effects. A hydrogen substituent, for example, is less electronegative than a OH substituent, and the decrease of the reaction rate by the lack of hydrogen bonds at position 2 is partially compensated by a decrease in the inductive effect. On the other hand, a fluorine substituent is more electronegative than OH, and the decrease of the reaction rate will be the sum of a double contribution: (i) the loss of hydrogen bonds at position 2 and (ii) transition state instability by the inductive effect (Scheme 4.1 bottom). Therefore, the observed increase in activation energies must be taken as lower and upper estimates of the real contribution involving the hydroxyl interactions.

In this chapter we have quantified the net effect of the interactions involving the 2-OH group in the mechanism and free energy barriers of *Saccharomyces cerevisiae* Gas2 β-glucosidase (*Sc*Gas2). This enzyme is a GPI-anchored glycoside hydrolase (GPI = glycophosphatidylinositol) that is attached to the plasma membrane. It belongs to family 72 (GH72) and its catalytic domain is found in the cell wall of fungi.[189,190] Members of this family, whose function is to regulate the assembly and rearrangement of the β-1,3-glucan that forms part of the fungal cell wall, are known to exhibit high synthetic activities through a mechanism called *transglycosylation.*[191] This mechanism, as we will detail below, is in direct competition with hydrolysis, and many efforts are devoted to favor one mechanism over the other by perturbing enzyme-substrate interactions.



Scheme 4.1- (Top) First step of the retaining double displacement mechanism of GHs. The 2-OH hydroxyl interacts strongly with the nucleophile at the Michaelis complex (MC) than at the glycosyl-enzyme intermediate (GEI). The "R" moiety represents a leaving group. Hydrogen bonds are represented by black dashed lines between heteroatoms. Red dashed lines represent bonds breaking or forming. (Bottom) Effects of the substitution at position 2. Statements in red and green refer, respectively, to negative and positive contributions to the reaction rate.

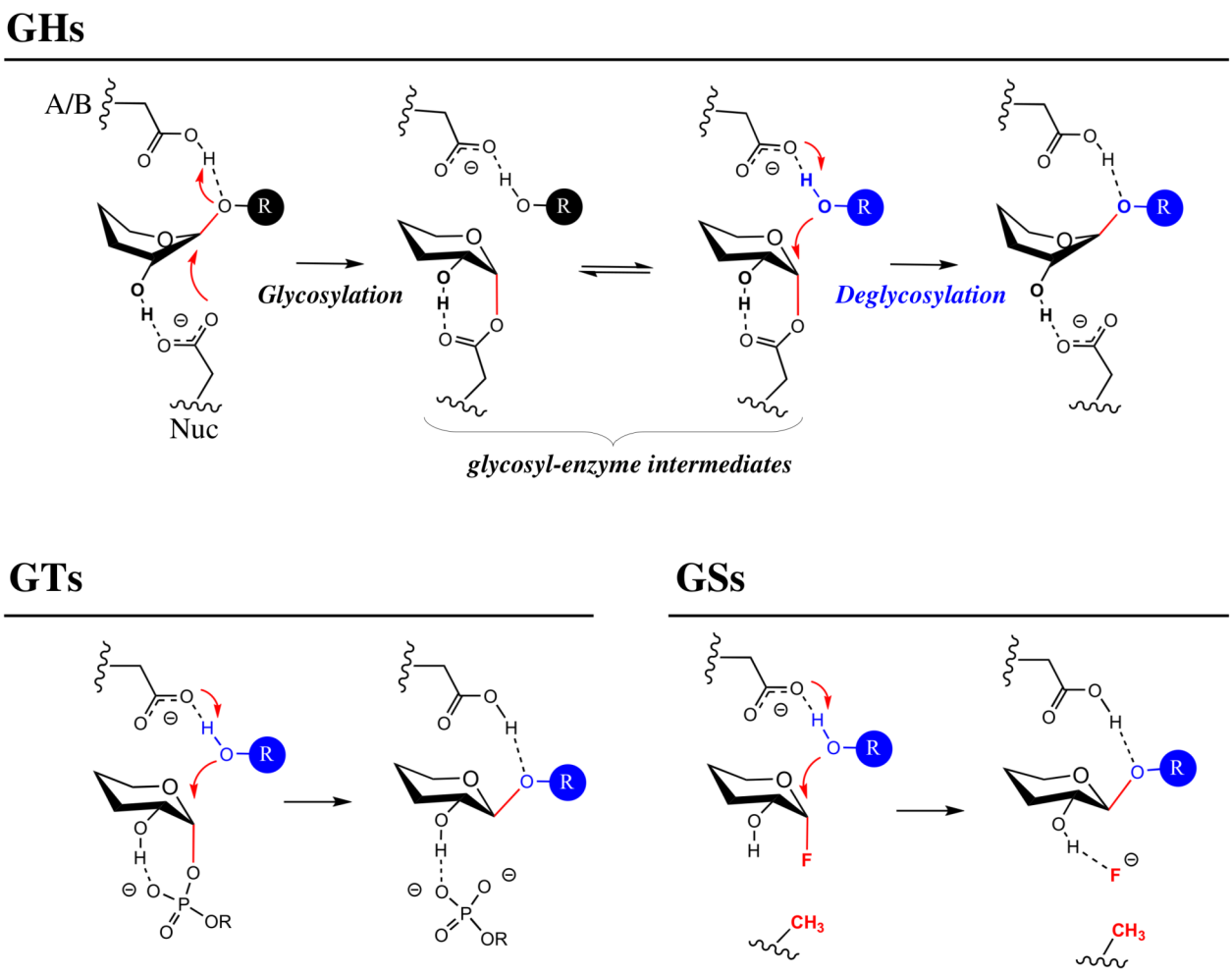
4.1.2 Transglycosylation: constructing sugars with deconstructing enzymes

The growth of glycomics and the development of diagnostic tests, vaccines, and new therapeutics based on carbohydrates depend on the availability of effective tools for their production.[4,192] Carbohydrate synthesis can be achieved by conventional chemistry approaches, but this strategy is generally harsh because it requires –given the similar reactivity of all sugar hydroxyls– several protecting and deprotecting steps, so reaction yields for the synthesis of a regioselective saccharide are usually very low.[193] Enzymatic synthesis, instead, does not suffer from selectivity problems and can be done in a fast and easy fashion with one single step. Glycosyl transferases (GTs),

for instance, are able to synthesize carbohydrates with a high precision, and are employed by cells for the construction of complex glycans such as glycogen or the antigens that decorate blood cells.[194] These enzymes use activated substrates –sugars attached to nucleotides– to drive the reaction towards glycosidic bond formation.

Surprisingly, retaining GHs can also be used to synthesize carbohydrates but without the need of activated substrates. This represents an economical advantage with respect to GTs as nucleotide-activated substrates are expensive in the market.[195] The two-step reaction mechanism of retaining GHs can explain this –apparently contradictory– activity for a hydrolase (see Scheme 4.2, top): once the glycosidic bond is broken and the glycosyl–enzyme intermediate (GEI) forms, the leaving group can be substituted by different acceptor molecules leading to distinct GEI complexes, which can subsequently react generating different products (*e.g.* hydrolysis product when R=H, and transglycosylation product when R=sugar). Although hydrolysis is thermodynamically favorable (by almost 3 kcal·mol^{-1} for cellobiose[22]), a few GHs known as transglycosylases (TGs), such as xyloglucan endo-TGs,[196] sucrase-type enzymes,[197] or trans-sialidases,[198] display significant transglycosylation activities and lead to high yields on reasonable time scales. This means that TGs take profit of kinetic advantages to struggle against thermodynamics, and that the ratio between hydrolysis and transglycosylation products –thermodynamic and kinetic products, respectively– can be modulated by affecting their reaction free energy barriers. However, despite the overall mechanism of transglycosylation is well known, it is still unclear how TGs can favor transglycosylation over hydrolysis in a 55 M "waterworld".[199]

One of the main problems of TGs is that their products are also substrates for the enzyme, so they can enter again into the catalytic cycle and be hydrolyzed with the time. To prevent such secondary hydrolysis, several approaches such as directed evolution,[200–202] site-directed mutagenesis,[96,203] or the use of endo/exo glycosynthases (GS)[195,204] have appeared in the last few years. GSs are engineered GHs that "emulate" the second step of retaining GHs (or the single step of inverting GTs; see Scheme 4.2 bottom). They use inactive variants of the enzyme in which the catalytic nucleophile is mutated, and together with 1-fluoro activated substrates –with *opposite* configuration to that of the natural substrate– and natural acceptor sugars stable synthetic products can be achieved (given that they can not be hydrolyzed by secondary hydrolysis due to the absence of a competent nucleophile). While this approach has been successful for many GHs, there are cases in which it does not work, and current research is focusing in understanding natural TGs to find new methods for the conversion of GHs into TGs.



Scheme 4.2- (Top) Retaining GH double displacement mechanism. The first step, glycosylation, involves the formation of the glycosyl-enzyme intermediate and the departure of the leaving group, which can be substituted by another molecule (shown in blue) and proceed with the second step, deglycosylation, either by hydrolysis (R=H; thermodynamic product) or transglycosylation (R=sugar; kinetic product). The main drawback of the transglycosylation reaction is that the product is also a substrate for the enzyme. (Bottom) Inverting GT and GS reaction mechanisms. Both involve a single displacement of a good leaving group –phosphate and fluorine, respectively– by a nucleophilic molecule, which can be a sugar, a peptide or a protein, among others. Notice that GSs, in comparison to GHs, have the catalytic nucleophile knocked-out by site directed mutagenesis (usually to alanine) to lead space for the fluorine of the donor and to avoid secondary hydrolysis.

Experiments on natural and engineered TGs show that they usually have lower catalytic efficiencies –*i.e.* higher reaction free energy barriers– in comparison to their purely hydrolytic relatives,[199,205] leading to long-lived species before the thermodynamic equilibrium is reached. In fact, GEI lifetimes as high as 30 min have been reported for wild-type TGs,[206] whereas the intermediate breaks down quickly in purely hydrolytic GHs. Factors such as substrate acceptor binding, water migration into the active site, and transition state (TS) interactions are known to influence the acti-

vation energy; thus, enzyme mutations affecting these factors can modify the transglycosylation/hydrolysis ratio.[199] Notwithstanding, there is not a straightforward, easy, and rational approach for generating such efficient enzyme variants. Moreover, the limited knowledge of the molecular basis of transglycosylation –particularly TS interactions– is hindering research in this field.

Recently, our collaborator Prof. Ramón Hurtado-Guerrero has trapped a unique crystal structure of *Sc*Gas2 that consists in a GEI (having a "G4" donor substrate attached to the catalytic nucleophile) together with a "G5" acceptor substrate bound into the active site (see Figure 4.2).[207] The observation of such ternary complexes is particularly difficult, as usually acceptors diffuse out of the active site when the GEI is formed to leave space for water molecules for deglycosylation,[208] precluding its characterization. This unprecedented structure represents the best departing point for studying transglycosylation from a computational point of view, given that no modeling approximation is needed (GEI structures without an acceptor substrate would require the use of docking approaches, which can be quite unreliable for large molecules such as G5).

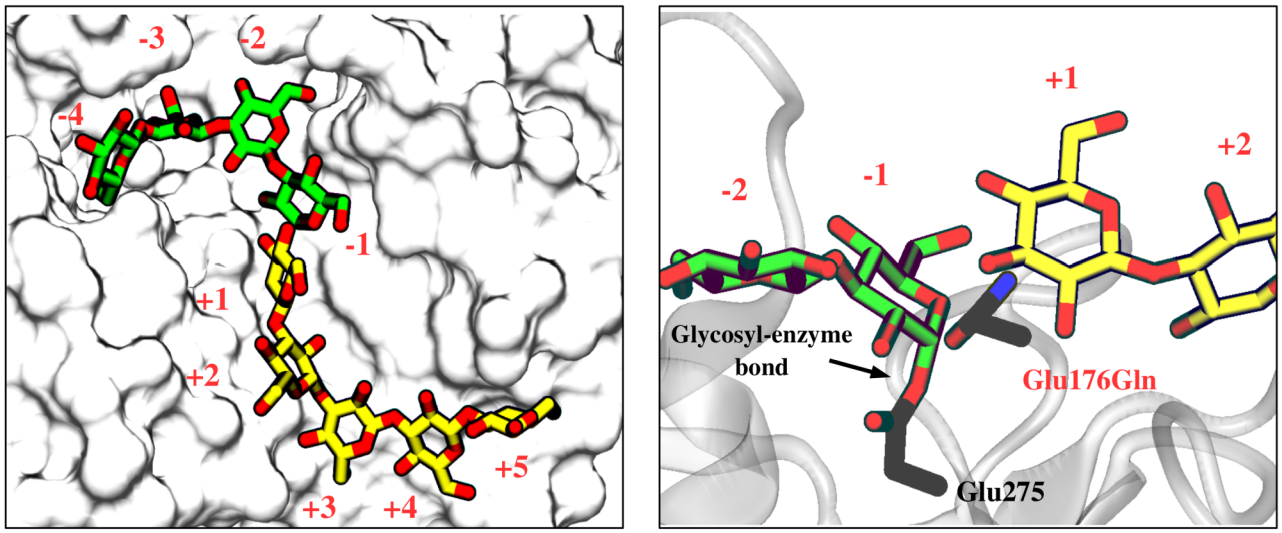


Figure 4.2- *Sc*Gas2 GEI with a bound acceptor substrate (PDB 5FIH). The G4 donor substrate is shown in green (subsites -1 to -4) and the G5 acceptor substrate in yellow (subsites +1 to +5). Notice the glycosyl-enzyme bond between the -1 sugar and Glu275 (nucleophile) as well as the Glu176Gln mutation (acid/base) used experimentally to prevent deglycosylation.

In this chapter we have dissected the transglycosylation mechanism of *Sc*Gas2 estimating the contribution of the 2-OH···Nucleophile interaction and using this information to improve the synthetic activity of the enzyme. This has been carried out in a three-fold scheme: (i) First, we have unveiled the hydrogen bond network of *Sc*Gas2 within molecular dynamics, finding two states for the 2-OH···Nucleophile interaction: one when it is formed and another when it is lost in favor of a 2-

OH···H_2O interaction. Moreover, we have quantified the energetics of these two states by means of QM/MM metadynamics, showing that the 2-OH···Nucleophile interaction is more stable by ~3 kcal·mol^{-1}. (ii) Next, we have computed the reaction free energy landscape for both states, revealing the importance of the 2-OH···Nucleophile interaction both in free energy barriers and catalytic itinerary: its removal raises free energy barriers by 11-16 kcal·mol^{-1} and changes the catalytic itinerary of the substrate, from $^4C_1 \rightarrow [^4E]^{\ddagger} \rightarrow {}^{1,4}B/{}^4E$ to $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$. (iii) Finally, we have tested the effect of replacing a high-conserved 2-OH interacting residue (Asn175···2-OH) by alanine, finding that it only affects the glycosylation barrier (by ~6.5 kcal·mol^{-1}), suggesting that this mutation can be used as a systematic target for the rational conversion of GHs into TGs.

4.2 Results and Discussion

4.2.1 Molecular dynamics reveal new enzyme-substrate interactions

We have unraveled the network of hydrogen bond interactions around the 2-OH of the -1 sugar by performing molecular dynamics (MD) simulations on the wild-type enzyme, reverting the mutation of the acid/base residue (Gln176→Glu176; see section 4.4 Computational Details). A detailed analysis of all active center interactions reveals two features that could not be observed in the crystallographic structure, presumably due to the enzymatic mutation: first of all, the 2-OH substituent often changes hydrogen bond partner, from the nucleophile (2-OH···Nucleophile interaction, hereafter named as *on configuration*) to a solvent water molecule (2-OH···H_2O interaction, named as *off configuration*; see Figure 4.3 top panel), with populations 45.4% and 54.6% respectively. Second, the 2-OH accepts a hydrogen bond from the amino group of Asn175 (average distance of 2.07 Å; Figure 4.3 middle panel).

To further characterize the dynamics of the C2-OH bond and obtain energetic information, we have performed a QM/MM metadynamics simulation using the H2−C2−O2−H dihedral angle as CV. Consistent with the above results, the free energy profile displays two minima that correspond to the *on* and *off* configurations, with the *on* configuration being ~3 kcal·mol^{-1} more stable (Figure 4.3 bottom panel). The free energy barriers for the interconversion from one configuration to the other are ~10 and ~7 kcal·mol^{-1}. To find out the net effect of the 2-OH···Nucleophile interaction in the free energy barriers, a snapshot of each state has been taken to initiate the modeling of the chemical reaction.

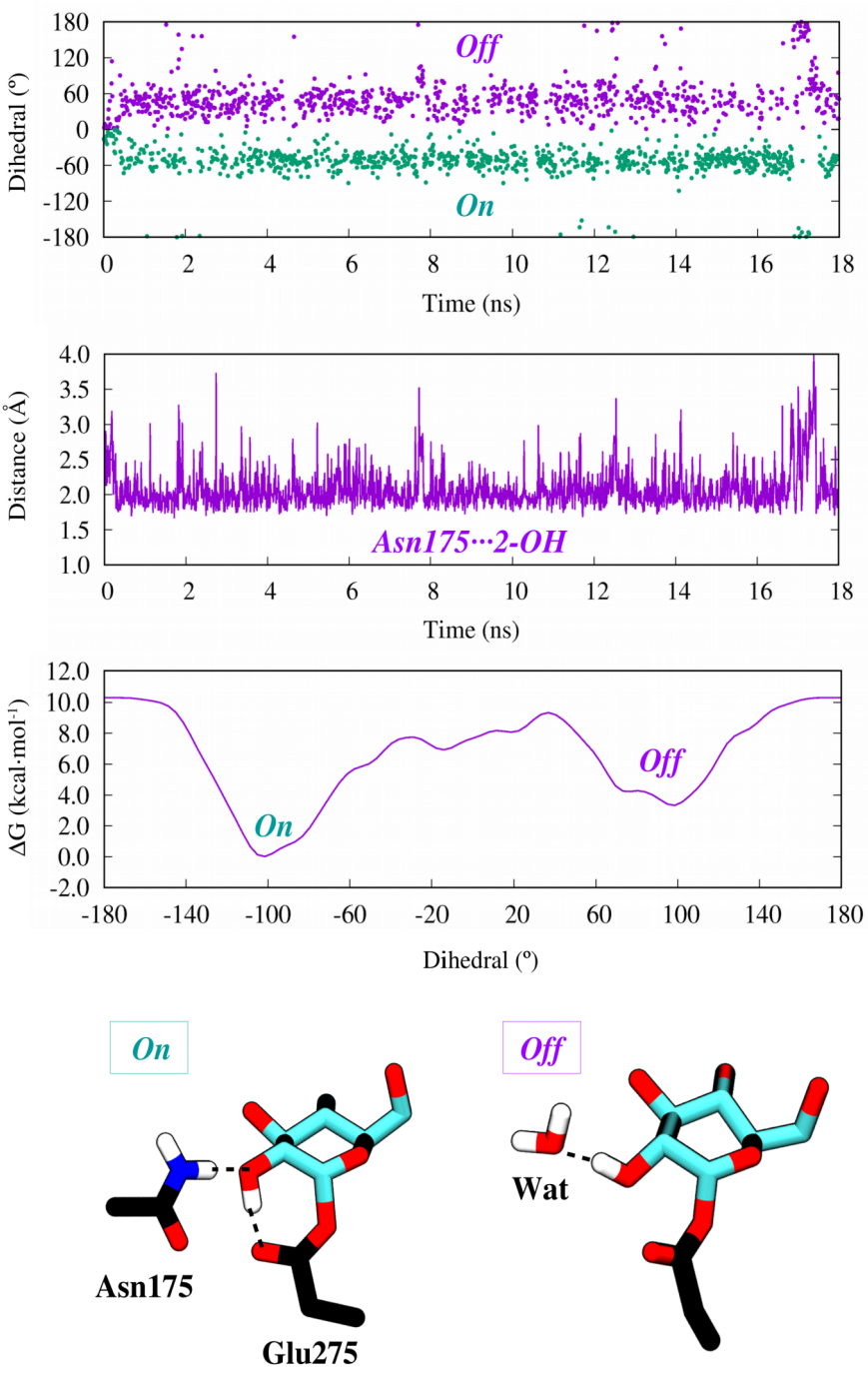


Figure 4.3- Evolution of the H2–C2–O2–H dihedral angle (top panel) and Asn175···2-OH hydrogen bond (middle panel) along the classical MD simulation; QM/MM free energy profile for the H2–C2–O2–H dihedral angle along a metadynamics simulation (bottom panel). *On* and *Off* configurations are shown below, with the -1 sugar covalently bound to the nucleophile (Glu275). Shifts of 130º and 75º have been applied to the dihedral angles of top and bottom panels in order to center the graphs.

4.2.2 The 2-OH···Nucleophile interaction acts as a molecular switch

The transglycosylation mechanism has been modeled using QM/MM metadynamics. A set of three collective variables, including all bonds that are formed or cleaved during the reaction, has been used to drive the reactants towards the transglycosylation products. CV1 measures the cleavage of the GEI bond, CV2 quantifies the degree of formation of the new donor–acceptor glycosidic bond, and CV3 takes into account the proton transfer between the acceptor and the Glu176 acid/base residue.

The two reaction mechanisms are shown in Figure 4.4. They are very similar in a general context, consisting in a single S_N2 displacement that is very dissociative, with the departure of the leaving group (Glu275) occurring before the new glycosidic linkage is formed and the proton transfer taking place after it (see Figures 4.4 and 4.5). However, despite its general similarity, there are three subtle but significant features to highlight: (i) the most relevant one is the difference in energetic barriers, with the *on state* notably more favored both forward and backward (transglycosylation and glycosylation), by 16 kcal·mol^{-1} and 11 kcal·mol^{-1} respectively (see Figure 4.5 and Table 4.1). These values, quantified for the first time by *ab initio* methods in the native enzyme, emphasize the importance of the 2-OH···Nucleophile interaction, which can be considered as a kind of *molecular switch*: it can activate or deactivate the enzymatic activity. Moreover, these results reinforce previous experimental estimations that concluded that this interaction contributes >10 kcal·mol^{-1} to the TS stabilization in retaining GHs.[188] (ii) The second relevant feature is the change in the catalytic itinerary of the substrate, from an expected $^4C_1 \rightarrow [^4E]^{\ddagger} \rightarrow {}^{1,4}B/^4E$ (similar to other β-glucosidases[8]) to an unusual $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$ cyclic itinerary, leading to completely different Michaelis complexes: one displaying a preactivated conformation ($^{1,4}B/^4E$) and the other a non-activated one ($^4C_1$). This highlights the fact that sugar conformations crucially depend on hydroxyl interactions. (iii) Finally, the third feature is related with the *on state*, and is the change in hydrogen bond partner of the 2-OH from one to another oxygen atom of the catalytic nucleophile. This feature appears to be common in various GHs –but not in all, see for instance the results in the preceding chapter– as it was previously observed in the study of a β-endoglucanase.[50]

It is also important to remark that both catalytic pathways bear transition state conformations that are very similar (^4E and $^4H_3$) and compatible with the requirement of a stable oxocarbenium ion, so that the crucial 2-OH···Nucleophile hydrogen bond does not affect the TS conformations. Likewise, in both cases the GEI is high in energy with respect to the MC (by 9 kcal·mol^{-1} for the *on state* and

1 kcal·mol^{-1} for the *off state*), something that has been recently observed by Piens *et al.* for a xyloglucan endo-TG.[206]

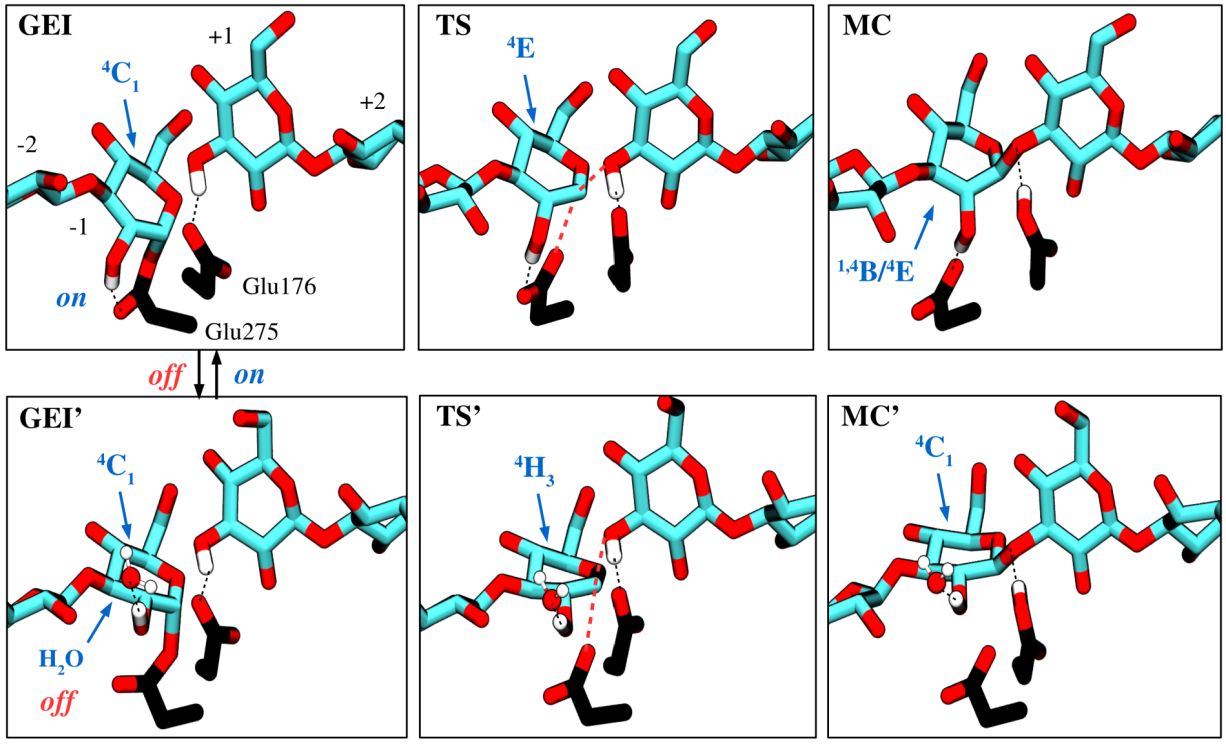


Figure 4.4- Close view of the catalytic center –subsites -1 and +1– along the two reaction mechanisms: (top) starting from the *On configuration* and (bottom) from the *Off configuration*. Notice that the 2-OH in the *Off configuration* is interacting with a water molecule. The interconversion between GEI (on) and GEI' (off) is shown in the bottom panel of Figure 4.3. Sugar conformations are highlighted in blue. Red dashed lines indicate the bonds that are being formed or broken. Most of the hydrogens have been omitted for clarity.

Globally, if we take into account the free energy profile for the *On*/*Off* transition and the corresponding reaction FELs, we can compare the energies of the two Michaelis complexes (MC and MC' in Figure 4.4). Interestingly, the greater stability of MC with respect to MC*'* (8 kcal·mol^{-1}, Figure 4.5 bottom) can be attributed to the strong 2-OH···Nucleophile hydrogen bond –which should be >3 kcal·mol^{-1}, the value obtained at the GEI– as well as the contribution of the -1 sugar conformation, which can account up to 6 kcal·mol^{-1} based on previous results of Biarnés *et al.*[50]

Table 4.1- Computed reaction free energy values (kcal·mol^{-1}) for *On* and *Off* configurations. The transgycosylation barriers are the ones from GEI to MC and glycosylation barriers form MC to GEI. The $\Delta\Delta G^{\ddagger}_{On\text{-}Off}$ difference has been taken as an absolute value.

	$\Delta G^{\ddagger}_{On}$	$\Delta G^{\ddagger}_{Off}$	$\Delta\Delta G^{\ddagger}_{On\text{-}Off}$
Transglycosylation	12	28	16
Glycosylation	21	32	11

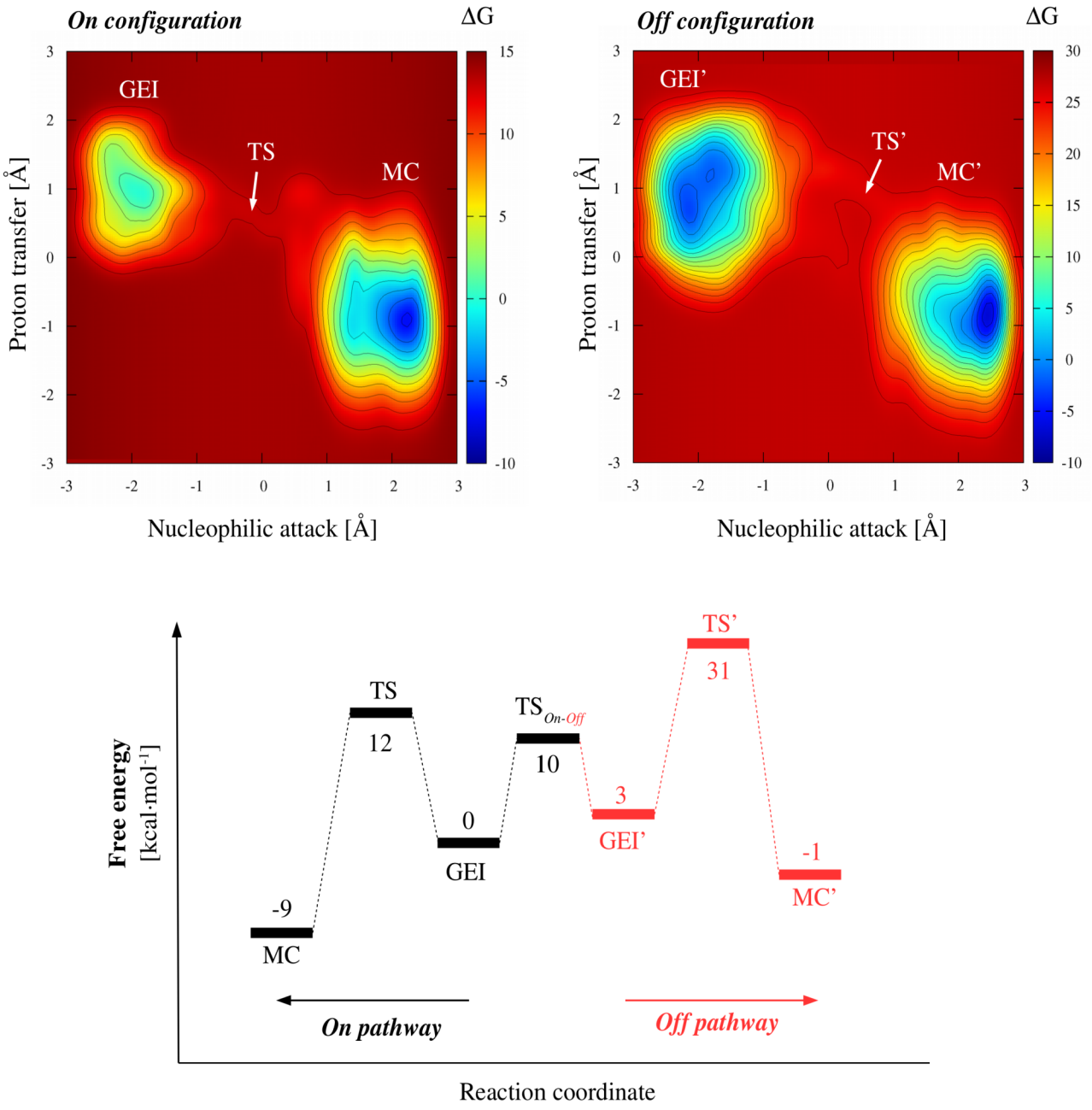


Figure 4.5- (Top) Computed free energy landscapes of the transglycosylation reaction for the *On* and *Off* configurations (left and right, respectively). To facilitate the analysis, the three dimensional FELs have been projected onto a 2D space defined by the nucleophilic attack (CV1-CV2) and the proton transfer (CV3). Energies are given in kcal·mol^{-1}. Contour lines are at 2 kcal·mol^{-1}. (Bottom) Schematic free energy profile for the two reaction pathways taking into account the above FELs and the free energy profile for the *On*/*Off* transition (Figure 4.3, bottom panel).

Altogether, it is clear that the 2-OH···Nucleophile interaction (*on configuration*) is crucial for catalysis, as its removal (*off configuration*) raises free energy barriers >10 kcal·mol^{-1} and alters the conformational itinerary of the substrate.

4.2.3 Worst is not always bad: biotechnological applications

The above results reveal that the 2-OH···Nucleophile interaction is essential for catalysis; thus, any interaction affecting it is expected to have an impact on the enzymatic activity. This is important since the transglycosylation product can be considered *metastable* in front of hydrolysis, and it will be long-lived if, once formed, free energy barriers are high enough to make it kinetically relevant (low free energy barriers would lead to a fast equilibrium). This means that a "worst" enzyme in terms of activity could make the transglycosylation product be stable enough for being observed and isolated experimentally. With this in mind, we have tried to find how could we indirectly affect the 2-OH···Nucleophile interaction. According to our simulations, the only residue that can directly influence it is Asn175, which forms a hydrogen bond with 2-OH.

To test how Asn175Ala mutation affects the energy barriers, we have performed QM/MM potential energy calculations in the GEI, TS, and MC for both wild-type and mutant systems (Figure 4.6), in the spirit of the work of Bueren-Calabuig.[209]

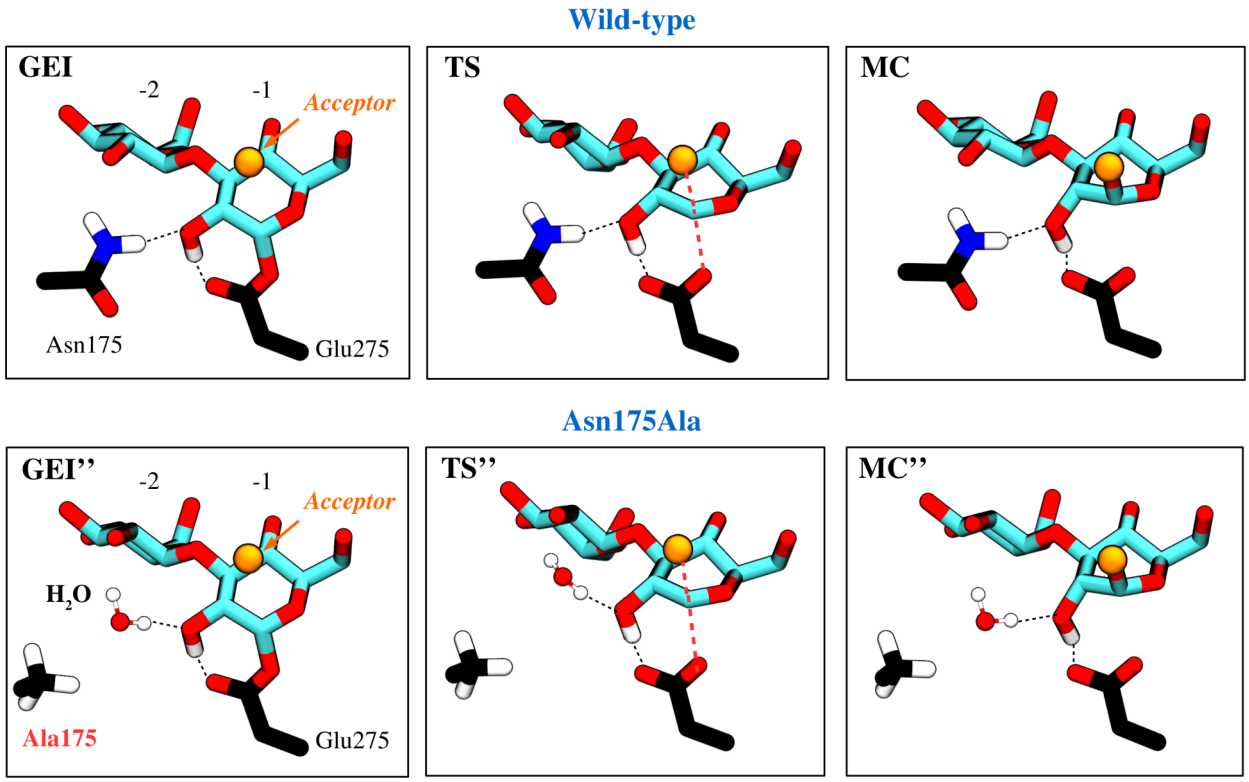


Figure 4.6- Stationary points for the WT and Asn175Ala potential energy optimizations (top and bottom, respectively). The point of view has been selected to highlight the differences in 2-OH interactions between both systems (Asn175···2-OH in the WT and H_2O···2-OH in the mutant enzyme). The acceptor molecule (a G5 sugar) is represented as an orange ball.

The results show that while the transglycosylation energy barrier is practically unaffected ($\Delta\Delta E^{\ddagger}$ = 0.2 ± 1.0 kcal·mol^{-1}), the glycosylation barrier increases by 6.5 ± 2.2 kcal·mol^{-1}. Structural analyses of the stationary points suggest that the high energy barrier observed for the Asn175Ala variant is related with the hydrogen bond that a water molecule performs with the 2-OH (up to 0.16 Å shorter than the Asn175···2-OH; see Table 4.2). The stronger hydrogen bond of water may increase the polarization of the C2-OH bond, which will enhance the inductive effect of the OH substituent and, ultimately, destabilize the transition state.

Table 4.2- Structural parameters (Å) for the WT and the Asn175Ala potential energy optimizations. Differences between the WT system and the Asn175Ala variant are listed, with the X in "X···2-OH" being Asn175 and H_2O, respectively. Standard deviations, taken from three different replicates, are ≤ 0.01 Å.

Wild-type	**GEI**	**TS**	**MC**
2-OH···Glu275	1.59	1.50	1.56
Asn175···2-OH	1.93	1.91	1.93
Asn175Ala			
2-OH···Glu275	1.61	1.51	1.55
H_2O···2-OH	1.84	1.79	1.77
Difference			
2-OH···Glu275	-0.02	-0.01	0.01
X···2-OH	0.09	0.12	0.16

Interestingly, in both cases the 2-OH···Nucleophile (Glu275) interaction reaches its shortest value at the TS (~1.50 Å), followed by the MC (~1.55 Å) and the GEI (~1.60 Å). This tendency indicates that the strength of the hydrogen bond interaction decreases in the order TS > MC > GEI, which is in agreement with the energy differences previously obtained: the loss of the 2-OH···Nucleophile interaction affects more deglycosylation (16 kcal·mol^{-1}) than glycosylation (11 kcal·mol^{-1}).

Enzymatic assays on the Asn175Ala mutant demonstrate the crucial role of the Asn175···2-OH interaction, as the catalytic activity of the enzyme is abolished.[210] These results suggest that mutations affecting the 2-position are expected to increase energy barriers. Given this scenario, one can envisage that the use of these mutants with –cheap– activated substrates (aryl- or fluoro-substituted donors) to generate the GEI, followed by addition of suitable acceptors to intercept the intermediate, could result in high yields of transglycosylation products, diminishing secondary hydrolysis by the increase of glycosylation barriers. This reasoning explains the observed enhancement in the hydrol-

ysis/transglycosylation ratio recently found for a GH1 enzyme:[96] the Asn163Ala variant –with Asn163 having an equivalent role to Asn175 in *Sc*Gas2– leads to higher transglycosylation yields, from 30% in the wild-type to 80% for the variant. Strikingly, this interaction pattern is conserved among several GHs:[153] Asn126 in Cex (PDB 2HIS), His108 in CtLic26A (PDB 2CIP), Asn163 in Ttβ-gly (PDB 1UG6), Asn175 in Cel7A (PDB 4C4C), Asn175 in TxAbf (PDB 2VRQ), or Asn127 in E-82 xylanase (PDB 2D24). Thus, targeting the 2-OH interacting residue may be a promising strategy to rationally convert a GH into a TG.

4.3 Summary and Conclusions

In this chapter we have studied the importance of the 2-OH···Nucleophile interaction in *Sc*Gas2, using this information to extract clues for improving its synthetic activity. First, we have demonstrated the existence of two states for 2-OH, namely the *On state* (2-OH···Nucleophile) and the *Off state* (2-OH···H_2O), with the former being favored by ~3 kcal·mol^{-1}. In a second step we have revealed the importance of the 2-OH···Nucleophile interaction both in reaction free energy barriers and the catalytic itinerary, showing that it contributes up to 16 kcal·mol^{-1} to the barrier and changes the catalytic itinerary of the substrate, from an expected $^4C_1 \rightarrow [^4E]^{\ddagger} \rightarrow {}^{1,4}B/{}^4E$ to an unusual $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$ cyclic itinerary. Lastly, we have shown how the suppression of a highly-conserved interaction involving the 2-OH (Asn175···2-OH) affects the glycosylation barrier without altering deglycosylation, suggesting that similar mutations can be used together with activated substrates for the rational conversion of GHs into TGs. Altogether, the following conclusions can be drawn from the present chapter:

- The 2-OH···Nucleophile interaction is crucial for the catalytic activity of retaining β-glucosidases. In *Sc*Gas2, its removal raises both glycosylation and deglycosylation free energy barriers and changes the catalytic itinerary of the substrate, from $^4C_1 \rightarrow [^4E]^{\ddagger} \rightarrow {}^{1,4}B/{}^4E$ to $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$, emphasizing that this interaction is necessary for substrate distortion at the Michaelis complex.

- The strength of the 2-OH···Nucleophile interaction decreases in the order TS (~1.50 Å) > MC (~1.55 Å) > GEI (~1.60 Å). This explains why the suppression of this interaction affects

more deglycosylation (16 kcal·mol^{-1}) than glycosylation (11 kcal·mol^{-1}), for which TS destabilization is partially compensated by MC destabilization.

- Substitution of Asn175 by alanine increases the glycosylation barrier by 6.5 kcal·mol^{-1} but leads the deglycosylation barrier practically unaffected. These changes may be attributed to the stronger hydrogen bond that water –replacing the amido group of Asn175– establishes with 2-OH.

- Mutation of conserved residues interacting with the 2-OH together with the use of activated substrates may be a promising strategy to rationally convert GHs into TGs. The expected increase in glycosylation barriers should diminish secondary hydrolysis, leading to long-lived transglycosylation species before the thermodynamic equilibrium is reached.

4.4 Computational Details

4.4.1 Modeling of the glycosyl-enzyme intermediate of *Sc*Gas2

The initial structure for the simulations has been taken from the structure of *Sc*Gas2 covalently bound with laminaritetraose donor (G4) and in complex with an incoming laminaripentaose acceptor (G5; PDB 5FIH). To simulate the wild type enzyme, the mutation of the acid/base residue (Glu176Gln) has been manually reverted (changing atom N by O without modifying its orientation) and the missing residues have been completed with available fragments of other *Sc*Gas2 structures (PDBs 2W61, 2W62 and 2W63) and by homology model according to *Sc*Gas2 sequence. The protonation states and hydrogen atom positions of all amino acid residues have been taken according to protein environment. A total number of 56.214 water molecules have been added to form a box of 15 Å around the protein surface and 27 sodium ions have been also added to neutralize the enzyme charge.

Molecular dynamics (MD) simulations using Amber11 software[176] have been performed. The protein has been modeled with the FF99SB force field,[110] and the carbohydrate substrate and water molecules were described with the GLYCAM06[178] and TIP3P[113] force fields, respectively. The parameters for the glycosylated glutamate have been taken from Parm99, taking as a reference structure a protonated glutamic acid residue and redistributing the charge of its proton over the whole molecule. The MD simulation has been carried out in several steps. First, the system has been minimized, holding the protein and substrate fixed. Then, the entire system has been allowed to relax.

To gradually reach the desired temperature, weak spatial constraints have been applied to the protein and substrate, while water molecules and sodium ions have been allowed to move freely at 100 K. Then, the constraints have been removed and the working temperature of 300K has been reached after two more 100 K heatings in the NVT ensemble. Afterwards, the density has been converged up to water density at 300 K in the NPT ensemble and the simulation has been extended to 18 ns in the NVT ensemble, when the system has reached equilibrium according to the root mean squared deviation of enzyme backbone. Analysis of the trajectory has been carried out using standard tools of AMBER and VMD.[179] The populations for the two states (*On* and *Off*), 45.4% and 54.6%, have been obtained considering -150º< *Off* <150º and *On* otherwise. Notice that, given the wide range of angles that encompasses the *Off* state, it also includes interactions with other residues rather than water, such as Asn175, Asn242 or Glu176.

The QM/MM calculations have been performed using the method developed by Laio et al.,[133] which combines Car–Parrinello MD,[107] based on Density Functional Theory (DFT), with force-field MD methodology. The QM region has included the glucose rings at the -1 and +1 subsites, half rings of the saccharides at the -2 and +2 subsites and the catalytic residues (Glu176 and Glu275), leading a total number of 88 QM atoms (including capping hydrogens) and 91.779 MM atoms for the system. The QM region has been enclosed in an isolated supercell of size 18.5 x 17.9 x 21.6 $Å^3$. Kohn–Sham orbitals have been expanded in a plane wave basis set with a kinetic energy cutoff of 70 Ry. Norm-conserving Troullier–Martins ab initio pseudopotentials[174] have been used for all elements. The calculations have been performed using the Perdew, Burke and Ernzerhoff generalized gradient-corrected approximation (PBE).[122] A fictitious electronic mass of 700 au and a timestep of 5 au has been used to ensure an adiabaticity of $4.73 \cdot 10^{-5}$ a.u·ps^{-1}·$atom^{-1}$ for the fictitious kinetic energy.

4.4.2 QM/MM metadynamics simulations of 2-OH rotation

A metadynamics[139] simulation has been performed to evaluate the energy barrier for rotation of the 2-OH substituent around the C2-O2 bond. The H2-C2-O2-HO2 dihedral angle (Ω) has been used as collective variable within the direct version of the metadynamics algorithm. The hill width and height have been set to 0.1 radiant and 0.00025 au (0.16 kcal·mol^{-1}), respectively, and the deposition time has been set to 30 fs (250 MD steps). First crossing criterion has been taken to determine the simulation end.[124] The free energy profile, after 592 deposited Gaussians, shows two principal minima, one centered at Ω= -106º and the other at Ω= 93º, representing *On* and *Off* configurations,

respectively. The *On* configuration turns out to be favored ~ 3 kcal·mol^{-1}, being 99.4% more populated according to Maxwell-Boltzmann probability weights at 300 K. This is different from the relative populations obtained by classical MD (45.4% and 54.6% for configurations *On*/*Off*, respectively), which we attribute to the limitations of the force-field parameters used to describe the glycosylated glutamate. Interconversion free energy barriers of 10 kcal·mol^{-1} (from *On* to *Off*) and 7 kcal·mol^{-1} (from *Off* to *On*) separate the two minimum wells, with a low stable intermediate ($\Omega \approx$ -20º) in which 2-OH is interacting with the nitrogen atom of Asn175.

4.4.3 QM/MM metadynamics simulations of the chemical reaction

The reaction free enery landscape (FEL) of the transglycosylation reaction has been explored for both states using the metadynamics approach with three collective variables (CVs). Two of them, CV1 and CV2 , have been taken as the C1−O_{Nuc} and C1−O_{Gly} distances, respectively. The third, CV3 , has been taken as the difference between the $O_{A/B}$−H and O_{Gly}−H distances. The parameters for the two simulations have been exactly the same with exception of the hill height, which for the *On* state has been set to 0.6 kcal·mol^{-1} and for the *Off* state to 1 kcal·mol^{-1} (given that, for this state, a higher barrier was expected). The deposition time has been set to 24 fs (200 MD steps). A fictitious harmonic coupling has been used to diminish the perturbation of the time-dependent potential, according to Lagrangian metadynamics,[211] using mass of 100 amu and constants of 1 au for CV1 and CV2, and mass of 100 amu and a constant of 3.5 au for CV3. Walls at 4 Å for each distance and +2 Å and −1.5 Å for the difference of distances have been used to reduce the FEL space to the chemical event. First crossing criterion has been taken to determine the simulation end. The three-dimensional free energy landscape (3D FEL) has been projected to 2D using a Jacobian procedure.

4.4.4 Metadynamics convergence tests

To estimate the statistical error of the free energies, we have carried out convergence tests to estimate the error in two of the metadynamics simulations (2-OH rotation and reaction mechanism for configuration *On*). To do so, we have launched several metadynamics simulations with different Gaussian heights, leading all other parameters (deposition time and Gaussian width) constant and starting from the half of the basin of our previous simulations (4 kcal·mol^{-1} for the 2-OH rotation simulation and 6 kcal·mol^{-1} for the reaction simulation). The molecular mechanism remains the same independently of the Gaussian height. The free energy results show that the error, taken as the

standard deviation, is of 1.2 kcal·mol^{-1} in both cases: $\Delta G^{\ddagger}_{reaction}$= 13.3±1.2 kcal·mol^{-1} and $\Delta G^{\ddagger}_{rotation}$= 8.1±1.2 kcal·mol^{-1} (see Table 4.3).

Table 4.3- Convergence tests for the free energy simulations of the reaction (left) and rotation (right) with respect to the Gaussian height. All values are given in kcal·mol^{-1}.

Gaussian height	$\Delta G^{\ddagger}$
0.31	15.0
0.63	13.0
1.00	11.5
1.51	13.0
2.01	14.0
Mean	13.3
SD	1.2

Gaussian height	$\Delta G^{\ddagger}$
0.08	6.7
0.16	10.0
0.31	8.0
0.47	7.8
Mean	8.1
SD	1.2

The relatively low standard deviation obtained indicate that the chosen collective variables account for all the relevant events during the process. Moreover, the standard deviation obtained should be taken as an upper bound, as it would decrease substantially if the change of the Gaussian height is coupled with a proper deposition time. In our test, we used a constant deposition time to see the net effect of the Gaussian height, but it should be taken into account that the greater the height, the larger must be the deposition time (otherwise "hill surfing" problems can be notorious).

4.4.5 Effect of the Asn175Ala mutation in the reaction energy barrier

Additional QM/MM simulations have been performed to test the impact of the Asn175Ala mutation in the reaction energy barriers. The Asn175Ala variant has been generated manually on the wild type system in order to have minimal differences in the environment (a classical MD equilibration has not been performed in view of the rigidity of the active site observed in different *Sc*Gas2 structures). Two water molecules have been included in the space that Asn175 was fulfilling, substituting the positions of its amido and carbonyl groups. This is a good modeling approximation based on the correspondence of the position of hydroxyl oxygens of a saccharide and water molecules of the solvent.[212] To save computer time, and because we are only interested in energy differences upon Asn175 mutation rather than absolute values, a potential energy analysis –rather than free energy– has been performed for both cases, considering three different configurations of the TS (snapshots of the previous metadynamics simulation). Structural optimization through an annealing pro-

cedure has been done for each TS structure, constraining the reactive bonds at their initial values to avoid scape from the TS. The convergence criterion has been set to 10^{-3} au·bohr^{-1} for the nuclear gradients and 10^{-7} au for the Kohn-Sham orbitals at the end of the optimization. From these optimized TS structures, we have performed forward and reverse trajectories along the nucleophilic attack to obtain the corresponding Michaelis complex (MC) and glycosyl-enzyme intermediate (GEI). The latter structures have been further optimized following the same protocol outlined above.

4.5 Supplementary Figures

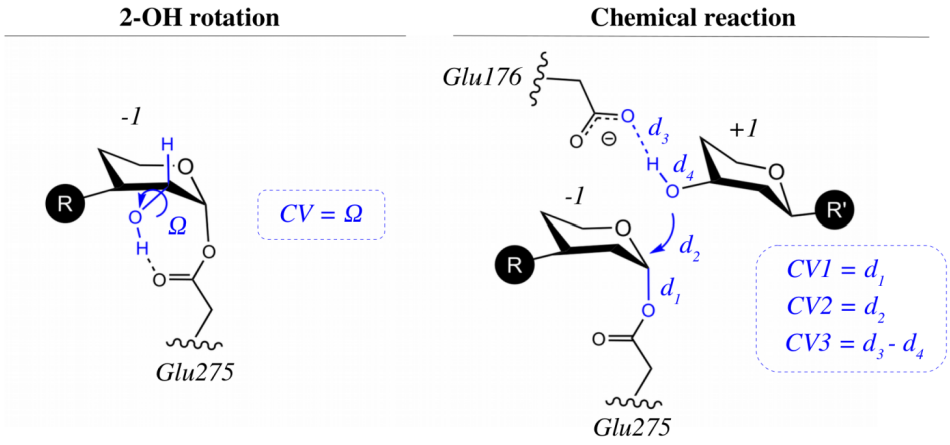


Figure S4.1- Collective variables used to study the rotation of the C2-O2 bond (left) and for the reaction mechanisms (right).

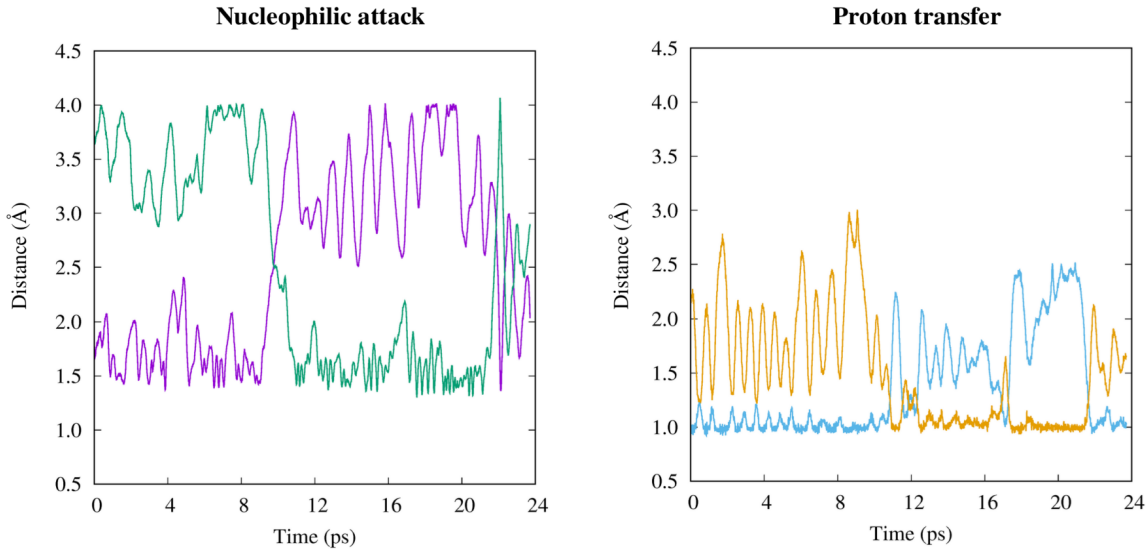


Figure S4.2- Evolution of the distances involved in the nucleophilic attack (left) and proton transfer (right) CVs along the reaction starting from the *On* configuration. The violet line corresponds to the O_{Nuc}-C1 distance (d_1), the green to the C1-O_{Gly} (d_2), the orange to the $O_{A/B}$-H (d_3) and the blue to the H-O_{Gly} (d_4) distances.

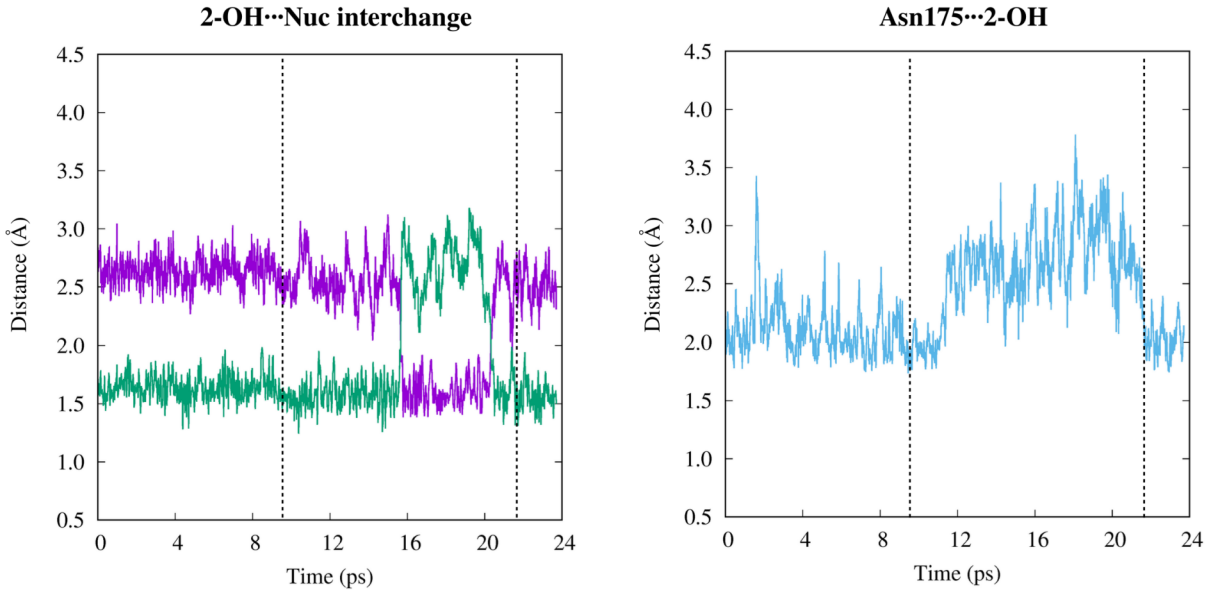


Figure S4.3- Evolution of the 2-OH···Nucleophile and Asn175···2-OH interactions (left and right) along the reaction starting from the *On* configuration. The green and violet lines correspond to the two oxygens of the nucleophile (left), and the blue one to the hydrogen bond between Asn175 and the 2-OH group. The vertical dashed lines indicate the transitions between GEI and MC. Notice that the Asn175···2-OH interaction is lost at the MC (>2.5 Å), which may facilitate the hydrogen bond interchange between the 2-OH and the oxygens of the nucleophile.

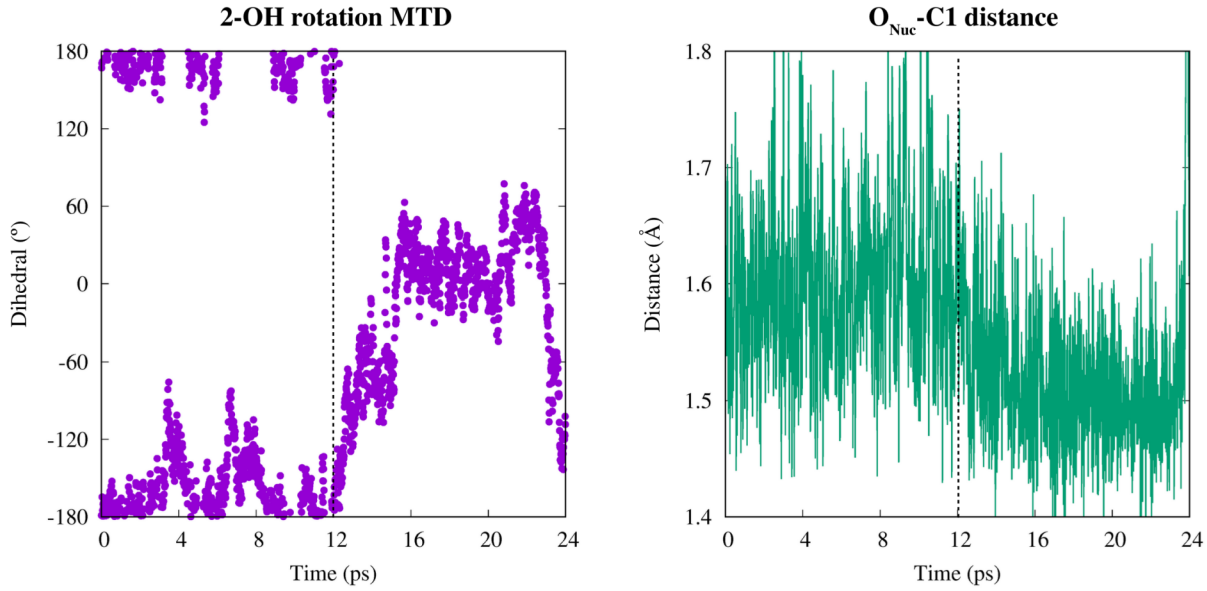


Figure S4.4- Evolution of the 2-OH rotation and the O_{Nuc}-C1 glycosyl-enzyme distance along the metadynamics simulation activating the H2-C2-O2-OH dihedral angle. The vertical dashed lines indicate the transitions from *On* to *Off* configurations. Notice that with this transition the O_{Nuc}-C1 distance (d_1) reduces from ~1.6 to ~1.5 Å, clearly indicating that the 2-OH···Nucleophile hydrogen bond facilitates the chemical reaction.

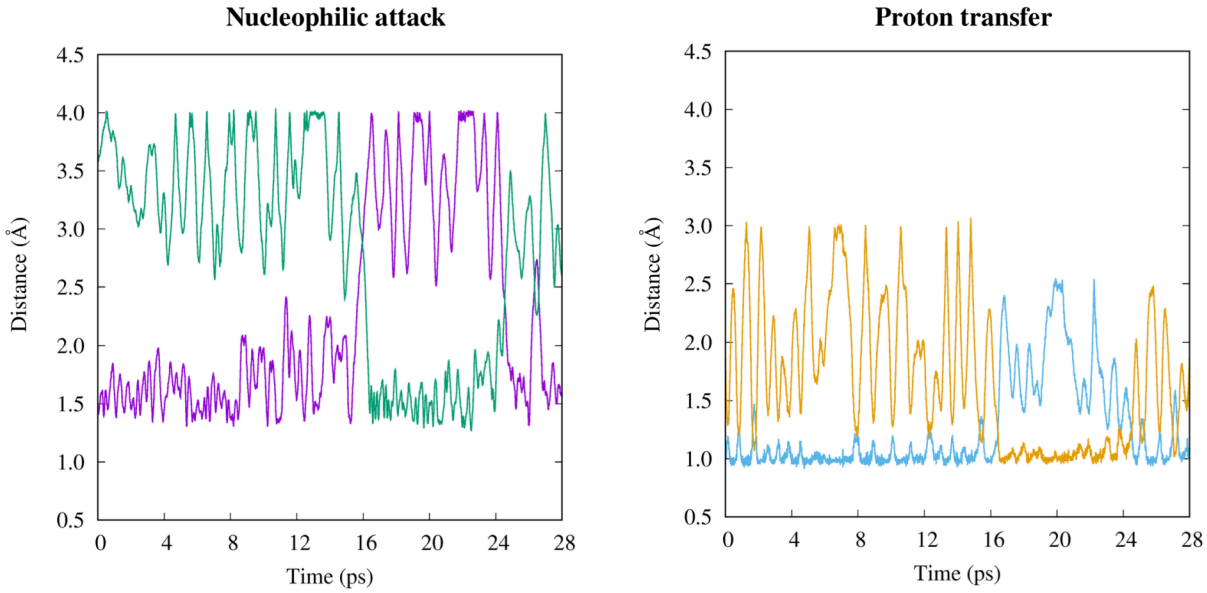


Figure S4.5- Evolution of the distances involved in the nucleophilic attack (left) and proton transfer (right) CVs along the reaction starting from the *Off* configuration. The violet line corresponds to the O_{Nuc}-C1 distance (d_1), the green to the C1-O_{Gly} (d_2), the orange to the $O_{A/B}$-H (d_3) and the blue to the H-O_{Gly} (d_4) distances.

Chapter 5
The Role of Water Binding Residues in the Active Site of an Inverting β-mannanase

Parts of this chapter have been published:
Y. Jin, M. Petricevic, A. John, L. Raich, H. Jenkins, L. Portela De Souza, F. Cuskin, H. Gilbert, C. Rovira, E. Goddard-Borger, S. Williams, G. Davies "A β-Mannanase with a Lysozyme-like Fold and a Novel Molecular Catalytic Mechanism" *ACS Central Science*, **2**, 896-903 (2016).

ABSTRACT: in this chapter we pay attention to non-catalytic residues that bind water in the active site of inverting GHs. We focus on a newly discovered β-mannanase with an inverting mechanism and an unusual fold, showing that its general structure is very rigid but its active site is highly dynamic. In particular, we reveal that three active site residues (Asp57, Lys59 and Asn65) display two different conformations and that waters are able to move *in and out* of the enzymatic cavity in the nanosecond timescale. We connect *end-to-end* two crystallographic structures (Michaelis complex and products) with QM/MM metadynamics, finding an unprecedented conformational itinerary and a subtle movement of Lys59 along the reaction coordinate. Finally, we estimate the contribution of Lys59 and Asn65 to the binding energy of water by means of alanine mutants, determining their crucial role in the binding and orientation of the catalytic water.

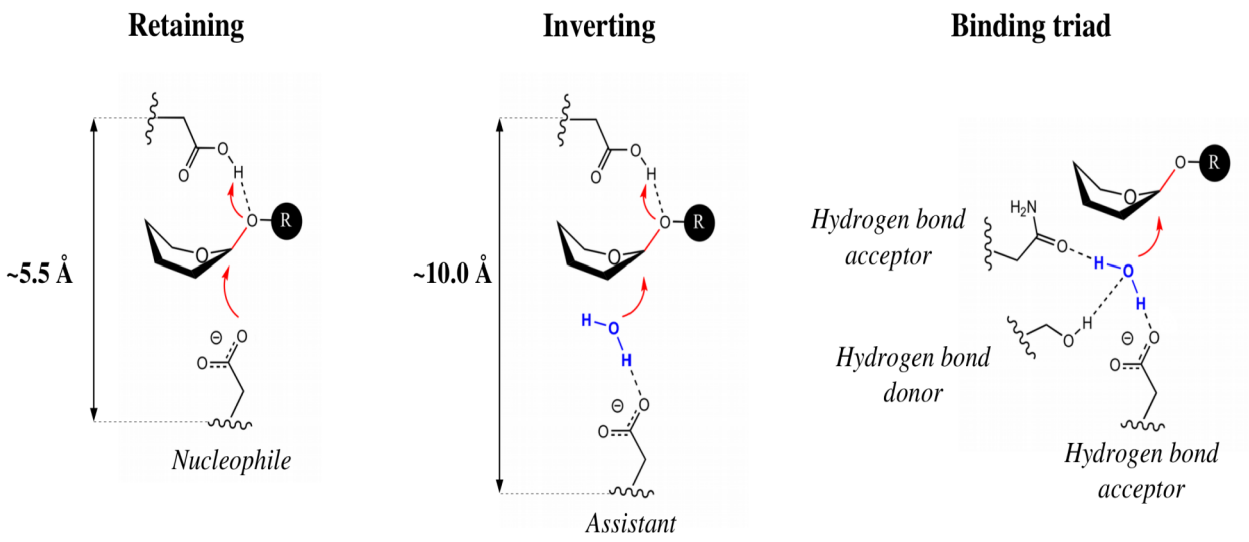
5.1 Introduction

5.1.1 From retention to inversion: tuning the machinery

One of the main differences between retaining and inverting GHs is the role of the general base. In retaining enzymes, the general base acts as a nucleophile, collapsing with the anomeric carbon to form the glycosyl-enzyme intermediate, while in inverting enzymes the nucleophile is a water molecule and the general base acts as an assistant residue to deprotonate the water molecule (see Scheme 5.1). This crucial difference of roles causes that, for inverting enzymes, two structural requirements need to be fulfilled: (i) there should be enough space for a water molecule between the anomeric carbon and the general base and (ii) the water molecule should have favorable interactions in that region for its binding and the stabilization of the transition state.

The first requirement is evidenced in a reference distance between the general acid and the general base: crystallographic structures of retaining GHs display an average distance of ~5.5 Å, while this value can increase almost twice (~10 Å) for inverting GHs.[213] The second requirement highlights that it is not enough to lead space for water molecules, but also that enzymatic non-covalent interactions are important.

Inverting GHs usually have hydrogen bond donor and acceptor residues next to the nucleophilic water, such as a glutamine, tyrosine or arginine (see Figure 5.1). These hydrogen bonds may "capture" the water molecule in the active site, orient it for a proper S_N2 displacement and stabilize the nascent charge on the water oxygen when the transition state is being formed.[214,215]



Scheme 5.1- Structural differences between retaining and inverting GHs. Distances represent an average of the four oxygen-oxygen distances between the carboxylates of the general acid and the general base.[216] Water-enzyme stabilizing interactions of a binding triad in an inverting GH.

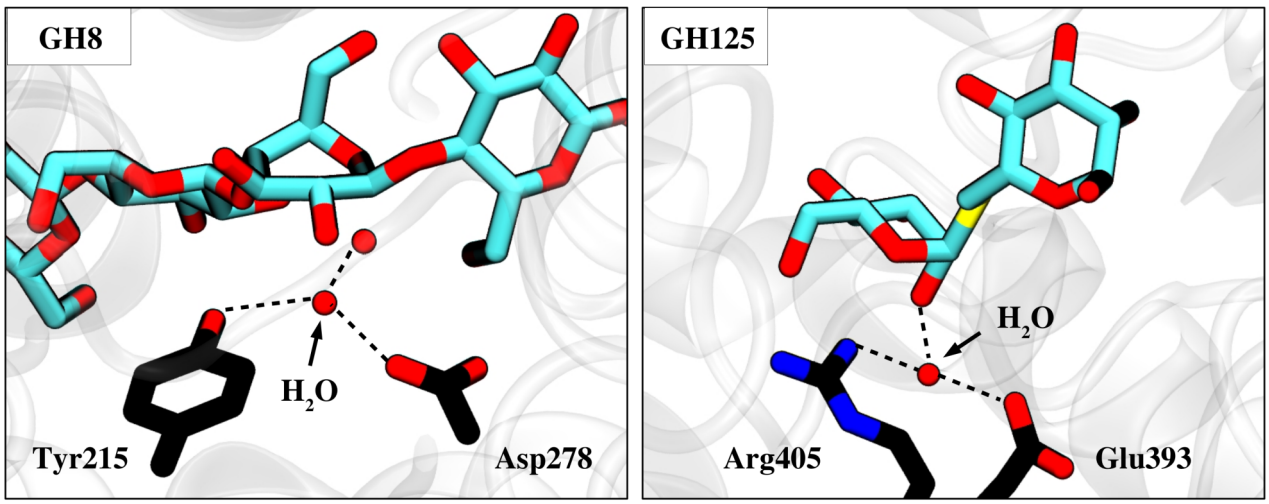


Figure 5.1- Tyrosine 215 and Arginine 405 stabilizing residues in a GH8 β-glucanase (PDB 1KW8) and GH125 α-mannosidase (PDB 3QT9). Asp278 and Glu393 are the assistant residues. Notice that these two enzymes do not have an additional residue to stabilize the catalytic water, but rather the catalytic water interacts with another water molecule (left) and with the 2-OH of the substrate (right).

According to what is exposed above, it is clear that the conversion of retaining GHs to their inverting analogues is a non-trivial engineering challenge. Nonetheless, nature has provided elegant examples of how enzymes with similar fold can bear different reactivity, such as sialidases from families GH34 and GH33 or chitinases from families GH22 and GH19 (both pairs retaining and inverting, respectively),[217] highlighting that it is a feasible goal. There are even members of the same family that have both types of mechanism, such as GH97 α-galactosidase and GH97 α-glucosidase enzymes (again, retaining and inverting).[218]

At an experimental level, conversion from one mechanism to the other has been achieved at least once for a β-glucosidase,[216] although with little differences with respect to the features of natural inverting enzymes. In this work, a first attempt was made by changing the nucleophile (Glu358) by a shorter aminoacid with similar properties (Asp358), which increased the separation of the acid/base and the general base by ~1 Å but did not changed the mechanism. Apparently, this separation is not long enough to allow the entrance of a water molecule, so the only possible mechanism is still the retaining one but in a less efficient manner (2500-fold slower than the wild-type). In a second attempt, the nucleophile was further shortened to alanine, leading enough space for a water molecule but making hydrolysis extremely inefficient (by 10^7-fold) due to the lack of a competent base residue. This made necessary the use of highly active nucleophiles such as azides or formates –instead of waters; here is the difference with respect to natural enzymes– to observe an inverting

mechanism.[216] These results emphasize the importance of the two structural requirements highlighted before, as (i) the small separation of the Glu358Asp mutant did not lead waters enter into the active site, and (ii) the Glu358Ala mutant did not had proper stabilizing interactions to make the reaction efficient with water. Hence, subtle variations in the arrangement of the key catalytic residues are enough to alter the enzymatic outcome, independently of the tertiary structure, but these variations need to be compensated with proper stabilizing interactions.

In this chapter, to understand the essence of these crucial interactions, we have studied the contribution of non-catalytic residues to the binding and stability of water in the active site of a novel β-mannanase with an inverting mechanism and an unusual fold.

5.1.2 The first β-mannanase with a lysozyme-like fold

The enzymatic hydrolysis of β-1,4-mannans is achieved by endo-β-1,4-mannanases,[219] enzymes involved in germination of seeds and microbial hemicellulose degradation, and which have increasing industrial and consumer product applications.[220,221] Mannanases occur in families 5, 26, and 113 of GHs, all with retaining mechanisms and similar $(\beta,\alpha)_8$-barrel folds (see Figure 5.2). However, our collaborators Prof. Gideon Davies and Prof. Spencer J. Williams had recently discovered that β-mannanases of a newly described family, GH134, differ from previously known β-mannanase families in both their mechanism and their tertiary structure.[222] In particular, they trapped a crystallographic structure of a *Streptomyces sp.* GH134 acid/base mutant (Glu45Gln) in complex with a mannopentaose substrate, showing that it displays a fold closely related to that of *hen egg white lysozyme* –a retaining enzyme– but acts with inversion of stereochemistry (see Figure 5.2).

The electron density, solved at 0.96 Å resolution, revealed a water molecule at the opposite face of the scissile glycosidic bond, well prepared for an inverting nucleophilic displacement assisted by the Asp57 general base (Figure 5.3 top). Furthermore, residues Lys59 and Asn65 appeared to be involved in direct hydrogen bonds with the putative catalytic water, being important for its binding and stabilization. These residues, together with Asp57, form a *binding triad* that is typical of inverting GHs. Finally, our collaborators noticed that while all other β-mannanases are known to operate through $^{1}S_{5} \rightarrow [B_{2,5}]^{\ddagger} \rightarrow {}^{O}S_{2}$ conformational itineraries,[223,224] with $^{1}S_{5}$ distorted substrates at the Michaelis complex, the -1 sugar of the present structure exhibits a $^{1}C_{4}$ inverted chair conformation. This suggests that GH134 operates through an unprecedented "southern hemisphere" catalytic itinerary, presumably $^{1}C_{4} \rightarrow [^{3}H_{4}]^{\ddagger} \rightarrow {}^{3}S_{1}$. Nonetheless, an additional crystal structure of a product complex –with the WT enzyme– lacking the acceptor substrate shows the -1 sugar distorted in a $^{1}C_{4}$ in-

verted chair, instead of the expected $^{3}S_{1}$ skew-boat (Figure 5.3 bottom). This evidence prompted up the question on whether the mutation made on the Michaelis complex (Glu45Gln) was perturbing the conformation of the substrate or whether the product relaxed from $^{3}S_{1}$ to $^{1}C_{4}$ after the reaction and the unbinding of the acceptor.

To shed light into this last question, we have unraveled the molecular mechanism and the catalytic itinerary of the GH134 endo-β-1,4-mannanase, paying special attention to the non-covalent interactions around the nucleophilic water. First, with molecular dynamics simulations we show that waters are highly dynamic in the active site and that the binding triad (Asp57, Lys59 and Asn65) is crucial for maintaining water in a proper orientation for catalysis. Second, we have obtained the reaction mechanism of the enzyme, connecting the two crystal structures and explaining the reason why the unexpected $^{1}C_{4}$ conformation is found at products. Lastly, we have performed a mutagenesis study substituting Lys59 and Asn65 by alanine, quantifying the importance of their interactions and their potential to convert the enzyme from inverting to retaining.

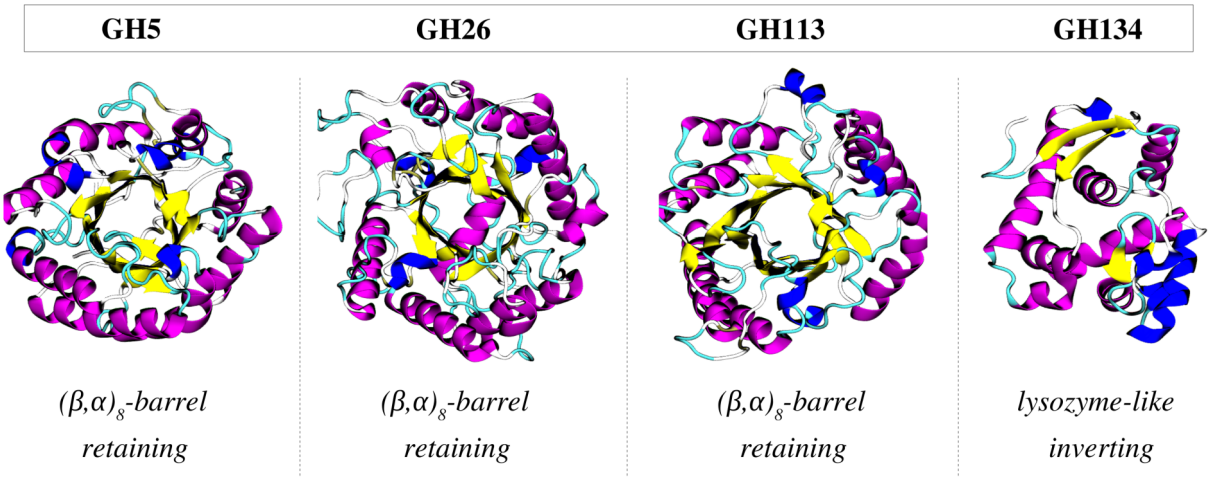


Figure 5.2- Three dimensional structures of families 5, 26 and 113 retaining β-mannanases displaying a $(\beta,\alpha)_8$-barrel fold (PDBs 3JUG, 4YN5 and 4CD8) as well as the newly described family, 134, with an inverting mechanism and a lysozyme-like fold (PDB 5JUG).

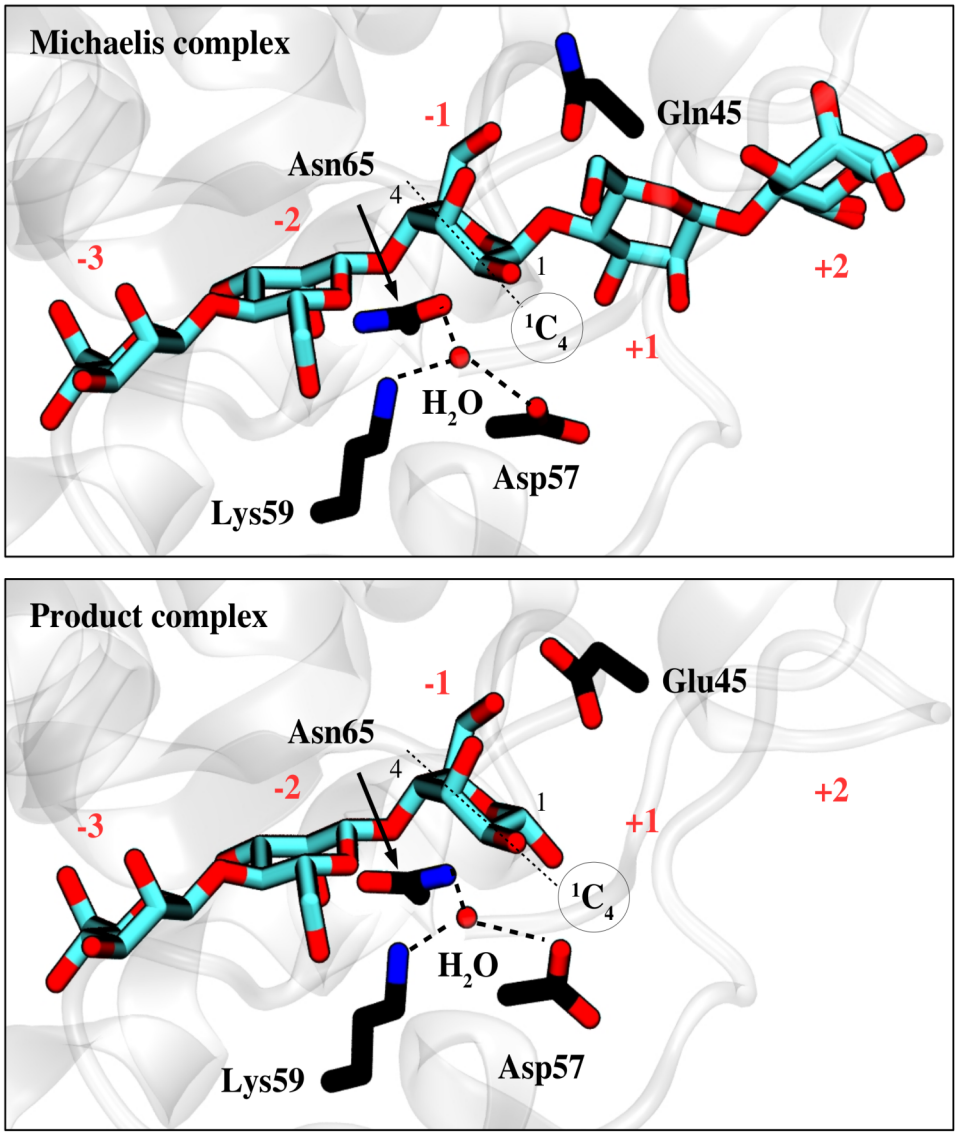


Figure 5.3- (Top) *Ss*GH134 acid/base mutant (Glu45Gln) in complex with a mannopentaose substrate bound in subsites -3 to +2. Notice that the -1 sugar displays a $^1C_4$ inverted chair conformation and that the catalytic water is fixed by a binding triad: Asp57 (the assistant residue), Lys59 and Asn65. (Bottom) Product complex of the WT enzyme in which there is a mannotriose substrate bound in subsites -3 to -1. In this case, the -1 sugar also displays a $^1C_4$ conformation and another water molecule, different from the previous one, is located in between the binding triad, ready to react with a new substrate. Interestingly, this structure was obtained from the incubation of *Ss*GH134 with a mannopentaose substrate, which suggests that after the hydrolytic reaction the leaving group leaves the acceptor subsites (+1 and +2) and thereafter the -1 sugar relaxes its conformation from an expected $^3S_1$ to the observed $^1C_4$.

5.2. Results and Discussion

5.2.1. Water can flow or water can react

We have checked the stability and flexibility of the natural Michaelis complex (reverting the acid/base mutation, *i.e.* Gln45→Glu45) via classical MD simulations. Three independent replicas of 40 ns each have been performed to equilibrate the system and explore the conformational space. Analysis of the trajectories show three noticeable features to highlight:

The first one is that although the enzymatic fold is very rigid, with few fluctuations along the simulation, the Asp57 and the Asn65 residues of the binding triad exhibit two different conformations. In particular, Asp57 has a conformation of 93% population that is exactly the one reflected in the crystallographic structure, with the residue interacting with the catalytic water and prepared to assist its deprotonation through an inverting mechanism (Figure 5.4 A). The other conformation, accounting for the remaining 7% of population, corresponds to a situation in which Asp57 has displaced the catalytic water and is near the anomeric carbon of the sugar, adopting the role of a possible nucleophile, resembling a retaining machinery (Figure 5.4 B). Similarly, in the case of Asn65, the most populated conformation –92%– is the one found by experiments, with the water molecule interacting either with the NH2 or C=O groups, while the other conformation corresponds to a rotation of the side chain in which the C=O approaches to the water molecule to interact exclusively with it (Figure 5.4 C). Whereas there is no experimental evidence for the conformational flexibility of Asp57, neither by means of X-ray crystallography nor kinetics studies (through the detection of retention products), the one of Asn65 is well characterized by the crystal structure of products, PDB 5JU9, which shows the same two conformations bearing 30% and 70% population each. Interestingly, the other residue in the binding triad, Lys59, is very rigid and has only one conformation in which it interacts with the catalytic water and also with the carboxylate group of the C-terminal region through a strong salt-bridge interaction (Figure 5.4 D).

The second relevant feature is that the -1 sugar of the substrate displays a stable $^{1}C_{4}$ chair conformation (Figure 5.5 top). This confirms that the structural modification of the acid/base residue used to trap the Michaelis complex (Glu45Gln) does not affect the conformation, reinforcing the experimental proposal of a southern hemisphere itinerary.

Finally, the third noticeable feature is that the nucleophilic water is highly dynamic, as it can flow *in and out* of the active site, being replaced by another water molecule, in the nanosecond timescale. For one of the simulations we have observed up to three water replacements, after 4, 28 and 32 ns (Figure 5.5 bottom). This dynamic behavior is in agreement with the “tunnel-type” topol-

ogy of the enzyme, reminiscent of processive GHs,[225] which do not unbind the substrate after the reaction –only the leaving group– and because of that they need water channels into the active site to bind another water molecule at the -1 position.

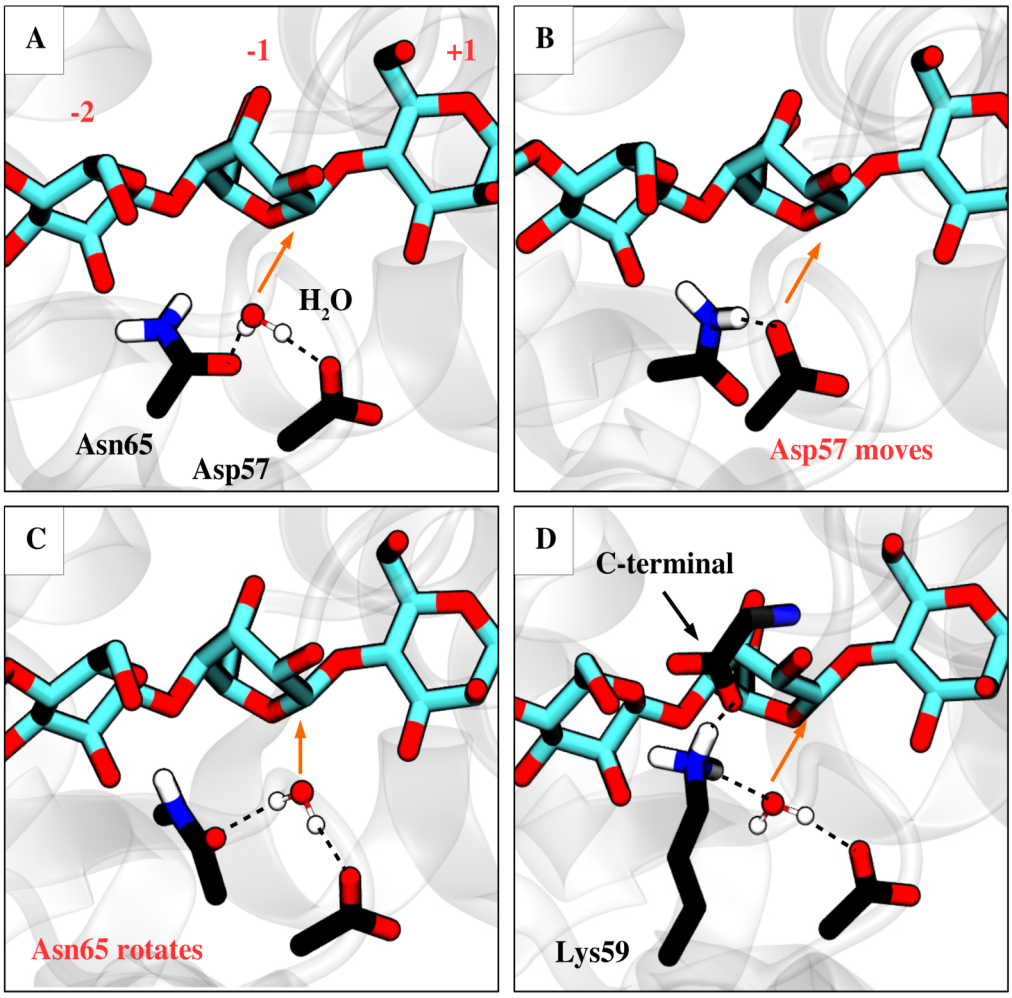


Figure 5.4- Active site conformations at different stages of the MD simulation: (A) the most populated conformation of Asp57 and Asn65, matching the crystallographic structure; (B) a low populated conformation in which Asp57 has displaced the catalytic water, adopting the role of a nucleophile in a retaining enzyme; (C) a low populated conformation in which Asn65 rotates, pointing with its carbonyl group to the catalytic water; and (D) strong salt-bridge interaction between Lys59 and the C-terminal carboxylate. Notice that Lys59 has been omitted from A to C in order to highlight Asn65, but it is always present interacting with the C-terminal carboxylate and the catalytic water. Orange arrows represent possible nucleophilic attacks on the anomeric carbon of the -1 sugar.

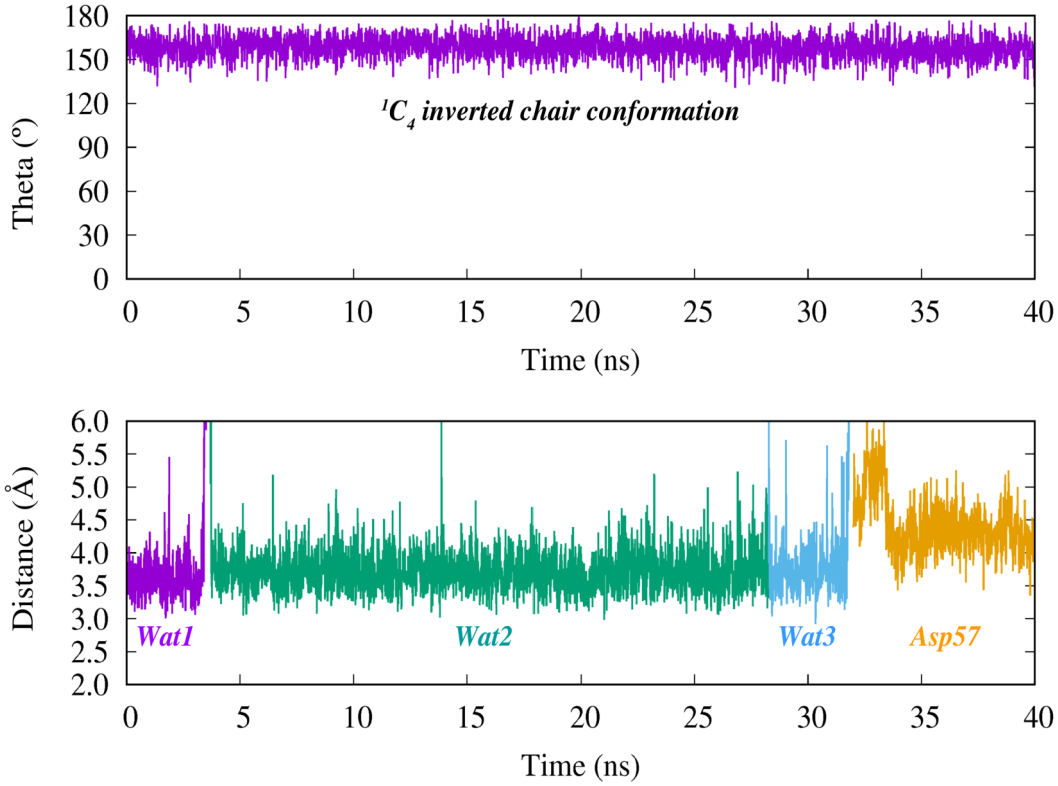


Figure 5.5- Evolution of the theta puckering variable (top panel) and distances between the oxygen atoms of the nucleophilic molecules and the anomeric carbon (bottom panel) along one of the 40 ns MD simulations. The value of 158.4 ± 7.5º for theta corresponds to a $^1C_4/^3E$ inverted chair conformation, which can be regarded as a canonical $^1C_4$. Water molecules –named as "Wat"– display an average distance of 3.7 ± 0.3 Å from the anomeric carbon. Notice that at after 32 ns of simulation Wat3 leaves the active site and at 34 ns Asp57 adopts the role of a possible nucleophile resembling a retaining enzyme.

Additionally, we have characterized the structural and thermodynamic properties of the nucleophilic water by means of a method called Grid Inhomogeneous Solvation Theory (GIST).[226] This method uses the information enclosed in the MD trajectory to create maps of water density and derive solute-water and water-water energetic terms as well as entropic contributions from translational and orientational restrictions, which can be combined to obtain binding ΔG values. The analysis shows that the catalytic water is stabilized in the active site by -3.1 ± 0.8 kcal·mol^{-1} in terms of potential energy and destabilized by 0.6 ± 0.2 kcal·mol^{-1} in terms of entropic penalty (-T·ΔS; T=300K), leading to a favorable binding free energy of -2.5 ± 0.6 kcal·mol^{-1}. The density of oxygen and hydrogen atoms shows that the catalytic water is perfectly oriented for catalysis, with the two hydrogens pointing towards Asp57 and Asn65 –establishing hydrogen bonds– and the two lone pairs pointing towards Lys59 and the anomeric carbon (the former establishing a hydrogen bond; see Figure 5.6). Moreover, we have found that there is a narrow water channel in between Asp57

and Lys59, suggesting that this pair of residues are crucial for the entrance of water molecules into the active site and, hence, for the hydrolytic activity of the enzyme. Therefore, it is clear that in this enzyme *water can flow*, now we will see if it can react.

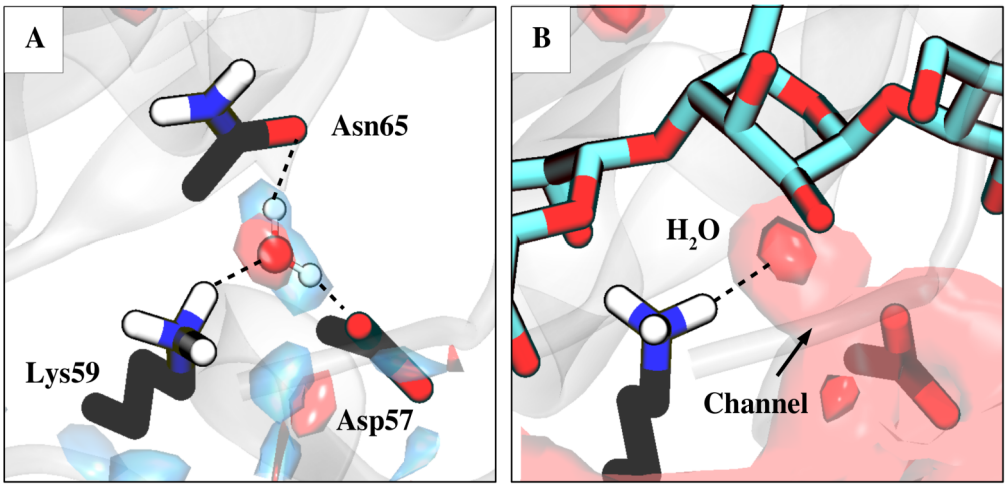


Figure 5.6- Contour plots of water density divided in their oxygen (red) and hydrogen (blue) contributions. (A) Top view of the -1 subsite, with the three binding residues surrounding the catalytic water. (B) Front view of the -1 subsite highlighting the water channel as a light red surface. Contour levels are, respectively, at 6 and 3 times the value of the bulk. The water channel is represented at 0.5 times the value of the bulk. The substrate and Asn65 have been omitted from the A and B, respectively, for the sake of clarity. A water molecule fitting the water density is displayed in A but not in B for clarity.

5.2.2. Computational proof for a southern hemisphere itinerary

The unusual $^1C_4$ inverted chair conformation observed in the X-ray structures of the Michaelis and product complexes, reproduced by our classical MD simulations, have prompted us study the catalytic mechanism and the conformational itinerary of the enzyme by QM/MM metadynamics. Three collective variables, including all bonds to be formed and broken during the reaction, have been considered. CV1 accounts for the nucleophilic attack of the catalytic water molecule; CV2 accounts for proton transfer between Asp57 and the water molecule; and CV3 accounts for the transfer of the Glu45 proton to the glycosidic oxygen atom (see section 5.4 Computational Details). The shape of the reconstructed free energy surface, projected onto CV1/CV2 , is indicative of a concerted one-step reaction (Figure 5.7). The reaction free energy barrier (17 kcal·mol^{-1}) is commensurate with the measured reaction rate.

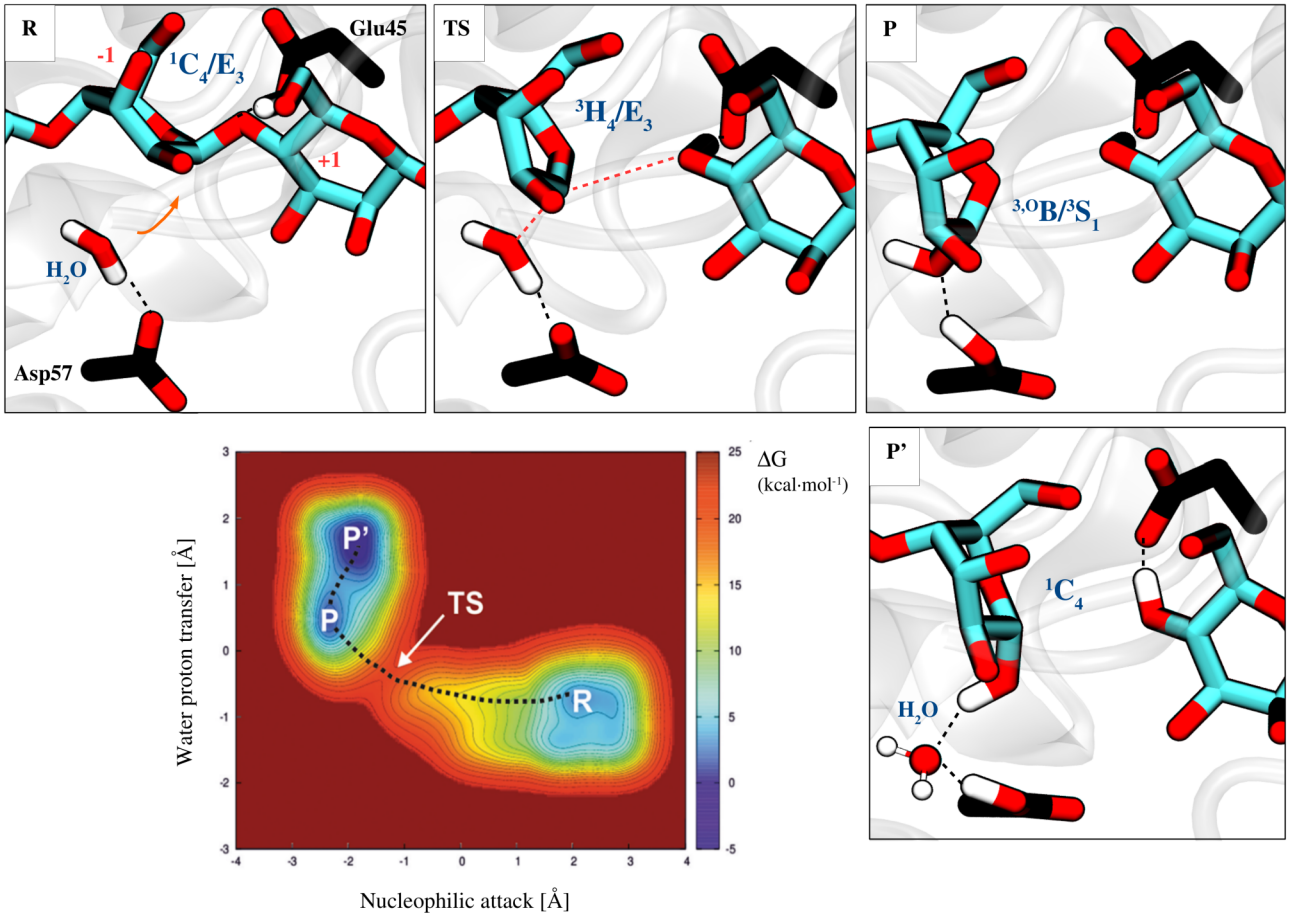


Figure 5.7- Free energy surface and representative states along the lowest free energy pathway. R denotes reactants (Michaelis complex), TS the transition state, and P the products. From P to P′, a solvent water molecule enters the active site and fills the space previously occupied by the catalytic water. Contour lines are given at 1 kcal·mol^{-1}. Hydrogen atoms have been omitted for clarity, except those of the catalytic water and the carboxylate group of Glu45.

The −1 sugar at the reactants state (R) adopts a conformation intermediate between $^1C_4$ and ^3E. The reaction starts with the elongation of the glycosidic bond simultaneous with the transfer of the carboxylic hydrogen atom of the acid residue to the glycosidic oxygen. At the transition state (TS), the −1 mannopyranose ring distorts from $^1C_4/^3E$ to $^3H_4/^3E$, a conformation that is compatible with the requirement of an oxocarbenium ion like TS. At this stage, the Glu45 proton is already transferred, the glycosidic bond is completely broken (3.3 Å; Table 5.1) and the bond between the nucleophilic water and the anomeric carbon is partially formed (2.0 Å). Proton transfer from the water to Asp57 then takes place, while the −1 sugar changes to a $^{3,O}B/^3S_1$ conformation (P), defining a canonical $^1C_4 \rightarrow [^3H_4]^{\ddagger} \rightarrow {}^3S_1$ "southern hemisphere" conformational itinerary. Thereafter, the anomeric OH loses its interaction with Asp57 (transition from P to P′, $\Delta G^{\ddagger}$= 2 kcal·mol^{-1}), and the −1 mannopyranose spontaneously undergoes relaxation to a $^1C_4$ conformation (Figure 5.7), matching the conformation observed in the product complex of the enzyme with mannotriose.

Remarkably, a new water molecule from the solvent enters into the active site at the same time that the conformational change occurs, emphasizing again the importance of water dynamics in this enzyme. It is interesting to notice that while Glu45, Asp57 and Asn65 residues are almost rigid during the reaction, Lys59 moves apart from reactants to products, leading space for the reactive sugar (Figure 5.8). This highlights the importance of the Lys59···Wat hydrogen bond interaction at the reactants and transition state, but not at the products.

Globally, the computed mechanism can be considered an electrophilic migration of the anomeric carbon from the departing sugar to the nucleophilic water, reacting with its oxygen, and assisted by Glu45 as general acid and Asp57 as general base.

Table 5.1- Change of the main distances (in Angstrom) involving the active site residues of *Ss*GH134 for each characteristic point along the reaction coordinate.

	O_{Wat}-C1	C1-O_{Gly}	O_{Wat}-H_{Wat}	H_{Wat}-O_{Asp57}	O_{Glu45}-H_{Glu45}	H_{Glu45}-O_{Gly}
R	3.68±0.18	1.55±0.15	1.01±0.03	1.74±0.06	1.20±0.30	1.78±0.57
TS	1.99±0.09	3.33±0.14	1.09±0.02	1.50±0.01	1.53±0.06	1.04±0.00
P	1.51±0.07	3.87±0.05	1.50±0.07	1.06±0.03	1.93±0.73	1.12±0.20
P'	1.43±0.03	3.28±0.03	2.60±0.02	0.98±0.04	1.62±0.01	0.99±0.00

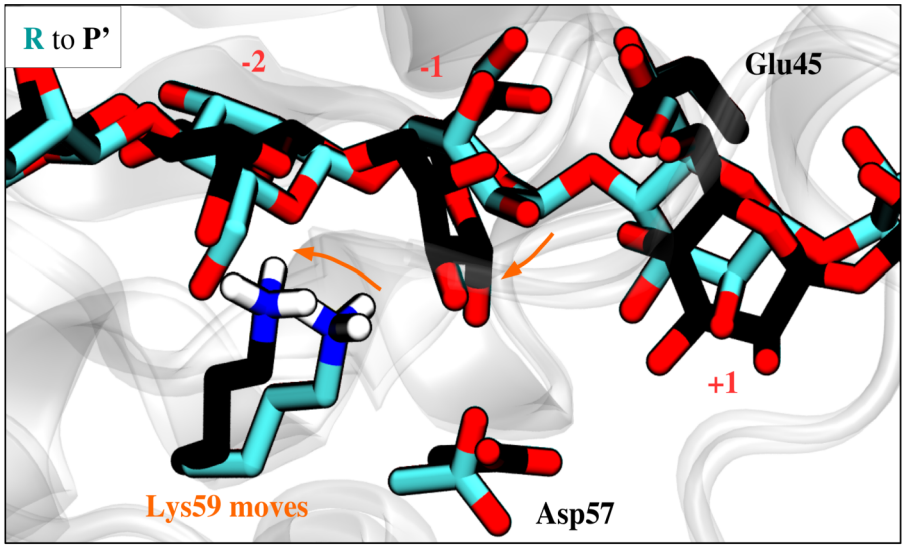


Figure 5.8- Structural differences between the Michaelis complex (R; cyan) and the product state (P'; black). Notice that the sugar migration makes Lys59 move apart to leave space for it. Additionally, Asp57 rotates to interact with the water molecule that enters after the reaction (not shown here for clarity). Only polar hydrogens of Lys59 are represented.

5.2.3. Breaking the chains: effect of Lys59Ala and Asn65Ala mutations

Given the importance of Lys59 and Asn65 residues in the binding of the catalytic water, we have studied the individual effect of alanine mutation in these two crucial positions. We have performed three independent 40 ns simulations of classical molecular dynamics for each system (Lys59Ala and Asn65Ala). Afterwards we have analyzed the properties of water with the GIST method outlined above, comparing the results with ones obtained for the WT form (section 5.2.1).

The mutation of Lys59 to alanine has a dramatic effect both in the enzymatic structure and the configuration of water molecules in the active site. First of all, this mutant lacks the strong interaction of Lys59 with the carboxylate of the C-terminal region, present in the WT enzyme, causing a dramatic movement of the terminal loop that opens a "big door" for the entrance of water molecules in the active site (Figure 5.9).

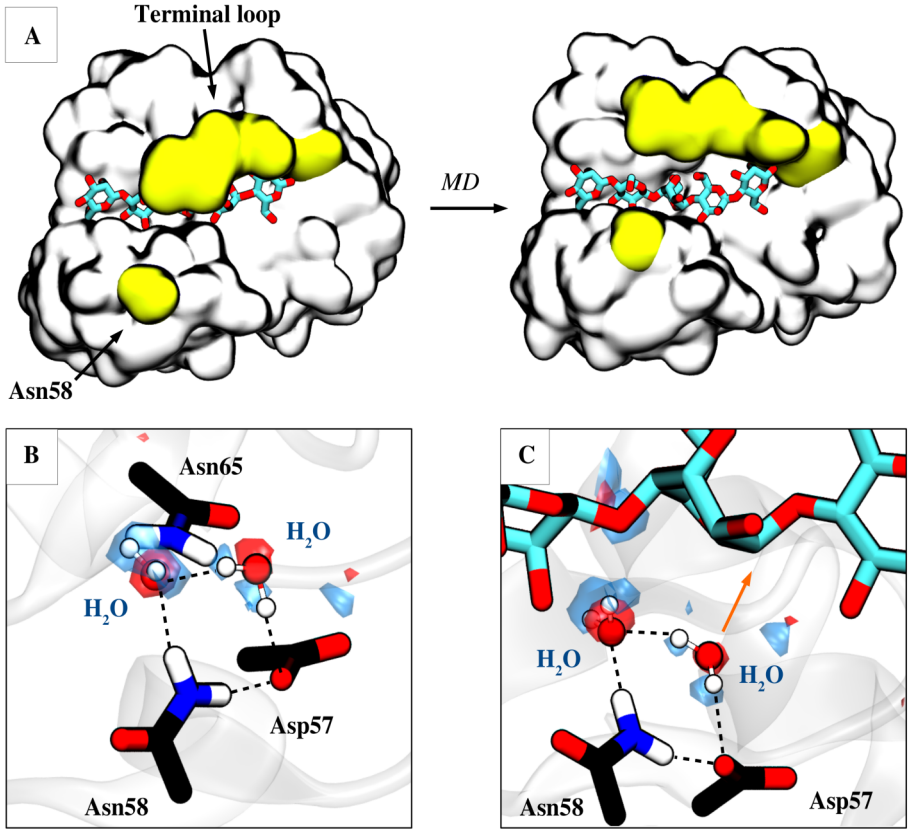


Figure 5.9- Effect of the Lys59Ala mutation: (A) it opens the terminal loop that closes the active site, leading more space for water molecules, and at the same time it induces a conformational change on Asn58, which is initially pointing to the solvent and ends up interacting with Asp57. (B) Top view of the -1 subsite, with the two binding residues (Asp57 and Asn65) surrounding the catalytic water and Asn58 and an additional water molecule fulfilling the space leaded by the mutation. (C) Front view of the -1 subsite. Contour plots of water density divided in their oxygen (red) and hydrogen (blue) contributions. Contour levels are, respectively, at 6 and 3 times the value of the bulk. The substrate and Asn65 have been omitted from B and C, respectively, for the sake of clarity. Water molecules fitting the water density are displayed in both views.

Moreover, it makes Asn58 (a residue that in the WT is in between Asp57 and Lys59, pointing to the solvent) change its conformation to interact with Asp57 through direct hydrogen bonds, partially substituting the role of Lys59. The larger space available for water molecules, however, is not expected to increase hydrolysis for two reasons: (i) the water population near the anomeric carbon lowers, as is reflected in the drop of free energy stabilization, from -2.4 ± 0.6 kcal·mol^{-1} to -0.8 ± 0.1 kcal·mol^{-1} ($\Delta G_{WT\text{-}Lys59Ala}$ = 1.6 kcal·mol^{-1}); and (ii) the catalytic water is worst oriented, since it prefers to point to the additional water molecule that fulfills the space leaded by the lack of Lys59 instead of interacting with Asn65 (Figure 5.9). This reorients the nucleophilic lone pairs of water around 90º from the perfect orientation for catalysis, which is expected to affect reaction rates. Thus, we conclude that Lys59 is important both for the stability of the enzymatic fold and for the binding and orientation of the catalytic water, with a binding free energy contribution of 1.6 kcal·-mol^{-1}.

The mutation of Asn65 to alanine has a more subtle effect than in the case of Lys59Ala, but it leads to more interesting changes. This mutant does not perturb the global structure of the enzyme nor the local environment of the active site, but it suppresses the directional interactions that waters perform with the NH2 and C=O groups of Asn65 present in the WT form. As a result, waters try to compensate the lack of these interactions by establishing new hydrogen bonds with the pyranic oxygen of the -1 sugar, causing the catalytic water to be badly oriented for catalysis (Figure 5.10). The worst orientation of the water molecule, however, should not affect notoriously its binding free energy, as it keep the same number of hydrogen bonds as in the WT form (Wat···O5 instead of Wat···Asn65). Nonetheless, the results show that there is a drop of 1.2 kcal·mol^{-1} of free energy ($\Delta G_{WT\text{-}Asn65Ala}$), from -2.4 ± 0.6 to -1.2 ± 0.5 kcal·mol^{-1}. This is because the interaction energy is averaged along all the simulation and, here is the interesting point, in this case the catalytic Asp57 residue fills the space previously occupied by the water molecule during most of the time (from 7% in the WT to 62% in the mutant; Figure 5.10 D). In other words, these results suggests that making the active site slightly more hydrophobic could affect the outcome of the reaction, from inversion to retention of configuration.

Finally, it is important to point out that none of the *in silico* mutations have affected the conformation of the -1 sugar, which display in all cases the inverted $^{1}C_{4}$ conformation that was initially found by X-ray experiments, consistent with the fact that these residues are not involved in the binding of the substrate, but only in the binding of the catalytic water.

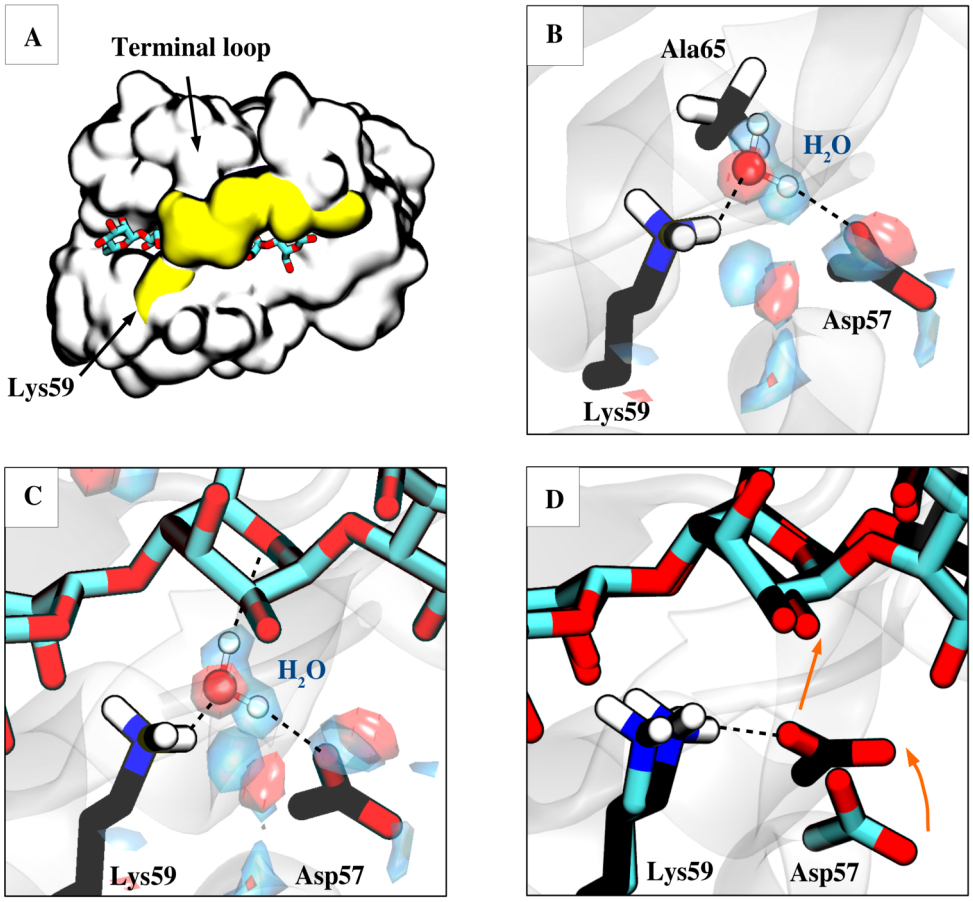


Figure 5.10- Effect of the Asn65Ala mutation: (A) The mutation does not affect the global structure of the enzyme nor the terminal loop that interacts with Lys59. (B) Top view of the -1 subsite, with the catalytic water bridged by Asp57 and Lys59. (C) Front view of the -1 subsite, showing the hydrogen bond of water with the pyranic oxygen. (D) Conformational change of Asp57, from a position in which is expected to act as an assistant residue (cyan) and as a possible nucleophile (black). Contour plots of water density are represented in red (oxygen) and blue (hydrogen), with levels being at 6 and 3 times the value of the bulk. The substrate and Ala65 residues have been omitted from A, B and C, respectively, for the sake of clarity. Water molecules fitting the water density are displayed.

5.3. Summary and Conclusions

In this chapter we have studied the relevance of the residues interacting with the catalytic water molecule in a novel GH134 inverting β-mannanase with an unusual fold. First, we have showed that although the enzymatic structure is very rigid, the catalytic water and its binding residues are highly dynamic, adopting conformations that may be important for water binding and entrance into the active site. Moreover, we have confirmed that the experimental mutation used to trap the Michaelis complex is not affecting the conformation of the -1 sugar, displaying a stable $^{1}C_{4}$ inverted chair conformation along the whole simulations. Second, we have computed the reaction mechanism by

means of QM/MM metadynamics, obtaining a computational proof for a $^1C_4 \rightarrow [^3H_4]^{\ddagger} \rightarrow {}^3S_1$ "southern hemisphere" catalytic itinerary. This allowed us to trace the conformational relaxation from $^3S_1$ to $^1C_4$ that takes place after the reaction, explaining the unexpected conformation present in the X-ray structure of the product complex. In addition, we have observed that the Lys59 binding residue moves apart after the transition state to lead space for the -1 sugar, highlighting that it is an important interaction for the stability of both reactant and transition state, but not for products. Finally, we have completed our study by analyzing the effect of mutation of Lys59 and Asn65 residues to alanine, finding that they contribute more than 1.6 and 1.2 kcal·mol^{-1} to the binding free energy of the catalytic water and that they are crucial for its proper orientation. Additionally, we have observed that the Asn65Ala mutant makes the general base (Asp57) to adopt a conformation that could be suitable for a retaining mechanism. Altogether, the following conclusions can be deduced from the present chapter:

- The *Ss*GH134 β-mannanase has a highly dynamic active center, with residues Asp57 and Asn65 bearing different low-populated conformations (<10%) and waters being able to enter *in and out* of the active site in the nanosecond time scale.

- The catalytic waters are translationally and orientationally restricted by strong hydrogen bonds with Asp57, Lys59 and Asn65 residues. The interactions with the two last resides account for 1-2 kcal·mol^{-1} of binding free energy and are crucial for the proper orientation of the catalytic water.

- The enzyme follows a $^1C_4 \rightarrow [^3H_4]^{\ddagger} \rightarrow {}^3S_1$ "southern hemisphere" itinerary. The unexpected $^1C_4$ conformation at the products, observed by X-ray crystallography, results from a spontaneous conformational change after the reaction.

- The Lys59 residue moves apart from reactants to products to lead space for the hydrolyzed substrate. Its role is to fix and orient the catalytic water at the same time that it stabilizes the nascent charge on the water oxygen as it approaches the TS.

- The Lys59Ala mutant perturbs both the enzymatic structure and the properties of water inside the enzymatic cavity, while the Asn65Ala mutant maintains the native structure but al-

ters the conformational preferences of Asp57, approaching it to the anomeric carbon. We suggest that this mutant could result in a mechanistic change, from inversion to retention of configuration.

5.4. Computational Details

5.4.1 Modeling of the *Ss*GH134 complexes with mannopentaose

The initial structure for the simulations has been taken from the present reported structure of *Ss*GH134 in complex with mannopentaose (PDB 5JUG).[222] To simulate the WT enzyme, the mutation of the acid residue (Glu45Gln) has been manually reverted (changing atom N by O without modifying its orientation). The protonation states and hydrogen atom positions of all amino acid residues have been taken according to protein environment. A total number of 12.102 water molecules have been added to within a radius of 15 Å from the protein and one sodium ion has been added to neutralize the enzyme charge.

Molecular dynamics (MD) simulations have been performed using Amber11 software.[176] The protein has been modeled using the FF99SB[110] force field. The carbohydrate substrate and water molecules have been described with the GLYCAM06[178] and TIP3P[113] force fields, respectively. The MD simulations have been carried out in several steps. First, the system has been minimized, holding the protein and substrate fixed, followed by energy minimization on the entire system. To gradually reach the desired temperature, weak spatial constraints have been initially added to the protein and substrate, while water molecules and the sodium ion have been allowed to move freely at 100 K. The constraints have been then removed and the working temperature of 300 K has been reached after two more 100 K heating steps in the NVT ensemble. Afterwards, the density has been converged up to water density at 300 K in the NPT ensemble and the simulation has been extended to 40 ns in the NVT ensemble, when the system has reached equilibrium according to the RMSD of the backbone. Two additional replicas of 40 ns each have been launched after the equilibration phase to enhance the conformational sampling. Lys59Ala and Asn65Ala mutant systems have been generated manually and have been equilibrated according the same procedure as in the WT enzyme.

Analysis of the trajectories has been carried out using standard tools of AMBER and VMD.[179] A water molecule has been considered to be inside the active site if the O_{Wat}-C1 distance was below 4.5 Å, otherwise it has been considered that Asp57 was fulfilling its space. To determine the population of Asn65 conformers, the distance between the carbonylic oxygen of Asn65 and the terminal nitrogen of Lys59 has been monitored, and the rotated conformation has been considered to be

present if the distance was below 4.5 Å. The percentages obtained for the different conformations do not significantly change upon variation of these parameters.

5.4.2 Modeling of the chemical reaction

QM/MM MD simulations have been performed using the method developed by Laio *et al.*,[133] which combines Car–Parrinello MD,[107] based on Density Functional Theory (DFT), with force-field MD methodology. The QM/MM interface has been modeled by the use of link-atom that saturates the QM region. The electrostatic interactions between the QM and MM regions have been handled via a fully Hamiltonian coupling scheme, where the short-range electrostatic interactions between the QM and the MM regions are explicitly taken into account for all atoms. An appropriately modified Coulomb potential has been used to ensure that no unphysical escape of the electronic density from the QM to the MM region occurs. The electrostatic interactions with the more distant MM atoms have been treated via a multipole expansion. Bonded and van der Waals interactions between the QM and the MM regions have been treated with the standard AMBER force-field. Long-range electrostatic interactions between MM atoms have been described with the P3M implementation,[180] using a 64 x 64 x 64 mesh.

The QM region included the mannose rings at the –1, +1 and +2 subsites, as well as the half ring of the saccharide at the -2 subsite and the catalytic residues (Glu45 and Asp57), leading a total number of 98 QM atoms (including capping hydrogens) and 38.693 MM atoms for the system. The QM region has been enclosed in an isolated supercell of size 20.1 x 17.7 x 20.9 $Å^3$. Kohn–Sham orbitals have been expanded in a plane wave basis set with a kinetic energy cutoff of 70 Ry. Norm-conserving Troullier–Martins ab initio pseudopotentials have been used for all elements.[174] The calculations have been performed using the Perdew, Burke and Ernzerhoff generalized gradient-corrected approximation (PBE).[122] A fictitious electronic mass of 700 au and a timestep of 5 au has been used to ensure an adiabaticity of $4.12 \cdot 10$ a.u$\cdot ps^{-1} \cdot atom^{-1}$ for the fictitious kinetic energy.

The free energy landscape (FEL) of the reaction has been explored using the metadynamics approach[139,171] with three collective variables (CVs). We have used the metadynamics driver provided by the Plumed2 plugin.[183] The first collective variable (CV1) has been defined as the difference between the O_{Wat}-C1 and the C1-O_{Gly} distances. This variable accounts for the nucleophilic attack of the water molecule and the cleavage of the glycosidic bond. The second collective variable (CV2) has been defined as the distance difference between the O_{Wat}-H and H-O_{Asp57}, accounting for the proton transfer between Asp57 and the water molecule. Finally, CV3 has been defined as the distance

difference of O_{Glu45}-H and H-O_{Gly}, which thus accounts for the transfer of the Glu45 proton to the glycosidic oxygen atom.

A hill height of 1 kcal·mol^{-1} and a deposition time of 30 fs (250 MD steps) have been used to explore the FEL. The shape of the Gaussian terms has been selected according to the fluctuations in the reactant basin, with 0.35, 0.29 and 0.35 Å for CV1, CV2 and CV3, respectively. Walls at 3 Å for CV1 and at -1.5 Å for CV2 and CV3 have been used to reduce the FEL space to the chemical event. The three-dimensional landscape has been completed after 552 deposited Gaussians. Enlarging the simulation leads to a relaxation of the –1 subsite mannose into a $^1C_4$ conformation, as observed experimentally.

5.4.3 Analysis of densities and binding free energies of the catalytic water

We have used the GIST[226] (grid inhomogeneous solvation theory) analysis tool to extract the local properties of water from the MD simulations. With this tool, a three dimensional grid is defined to discretize the space of interest and calculate different properties at each point of the grid. Among these properties we can find the number density of water oxygens, the solute-water and water-water potential energies (Esw and Eww, including Lennard-Jones and electrostatic terms) or the translational and orientational two-particle contributions to the entropy (Strans and Sorient). All these values are referred to the bulk and are averaged over the simulation frames. The total energy of the system can be obtained by summing Esw and Eww (Etot = Esw + Eww), and the total entropy by summing Strans and Sorient (Stot = Strans + Sorient). Finally, combining the two last values lead to the binding free energies. We have defined a grid of dimension 20x20x20 Å^3 centered at the anomeric carbon of the -1 sugar, just next to the catalytic water, using a fine grid spacing of 0.5 Å. A clustering of waters has been done to characterize their most concurrent positions, with a cutoff criterion higher than 3 for the number density of oxygens (three times more dense than in the bulk) and a total energy lower than 0 kcal·mol^{-1} (stable regions). From the clustering results we have selected the catalytic water and we have obtained the properties of the most stable point of the grid, including densities and energetic terms. Each of the three 40 ns replicas have been analyzed separately, integrating every 5 ns of simulation to obtain standard deviations and ensure convergence. Ions have been excluded to avoid spurious interactions with the solvent, as recommended in the literature.[227] The position of the solute has been allowed to move freely, and therefore the obtained values do not represent static interactions, but rather conformational averages. While this approach increments the noise of the results, the fluctuations along the different simulations that we have per-

formed are very small. To further diminish the noise, particularly the one coming from translation and rotational motions, we have centered each trajectory to a reference structure according to the root mean square of the protein backbone. The energetic values are summarized in Table 5.2.

Table 5.2- Binding free energies, energies and entropies (-T·ΔS, with T=300 K) of water inside the -1 subsite for the three different systems (WT, Asn65Ala and Lys59Ala) and the three MD replicas. All values are given in kcal·mol^{-1} and correspond to time averages for each simulation.

ΔG	**Replica 1**	**Replica 2**	**Replica 3**	**Mean**	**SD**
WT	-1.8	-2.7	-2.8	-2.4	0.6
Asn65Ala	-1.4	-0.6	-1.5	-1.2	0.5
Lys59Ala	-0.8	-0.7	-0.8	-0.8	0.1
ΔE					
WT	-2.2	-3.4	-3.6	-3.1	0.8
Asn65Ala	-1.8	-0.7	-1.9	-1.5	0.7
Lys59Ala	-0.9	-0.8	-0.9	-0.9	0.1
-T·ΔS					
WT	0.4	0.7	0.8	0.6	0.2
Asn65Ala	0.4	0.1	0.4	0.3	0.2
Lys59Ala	0.1	0.1	0.1	0.1	0.0

5.5 Supplementary Figures

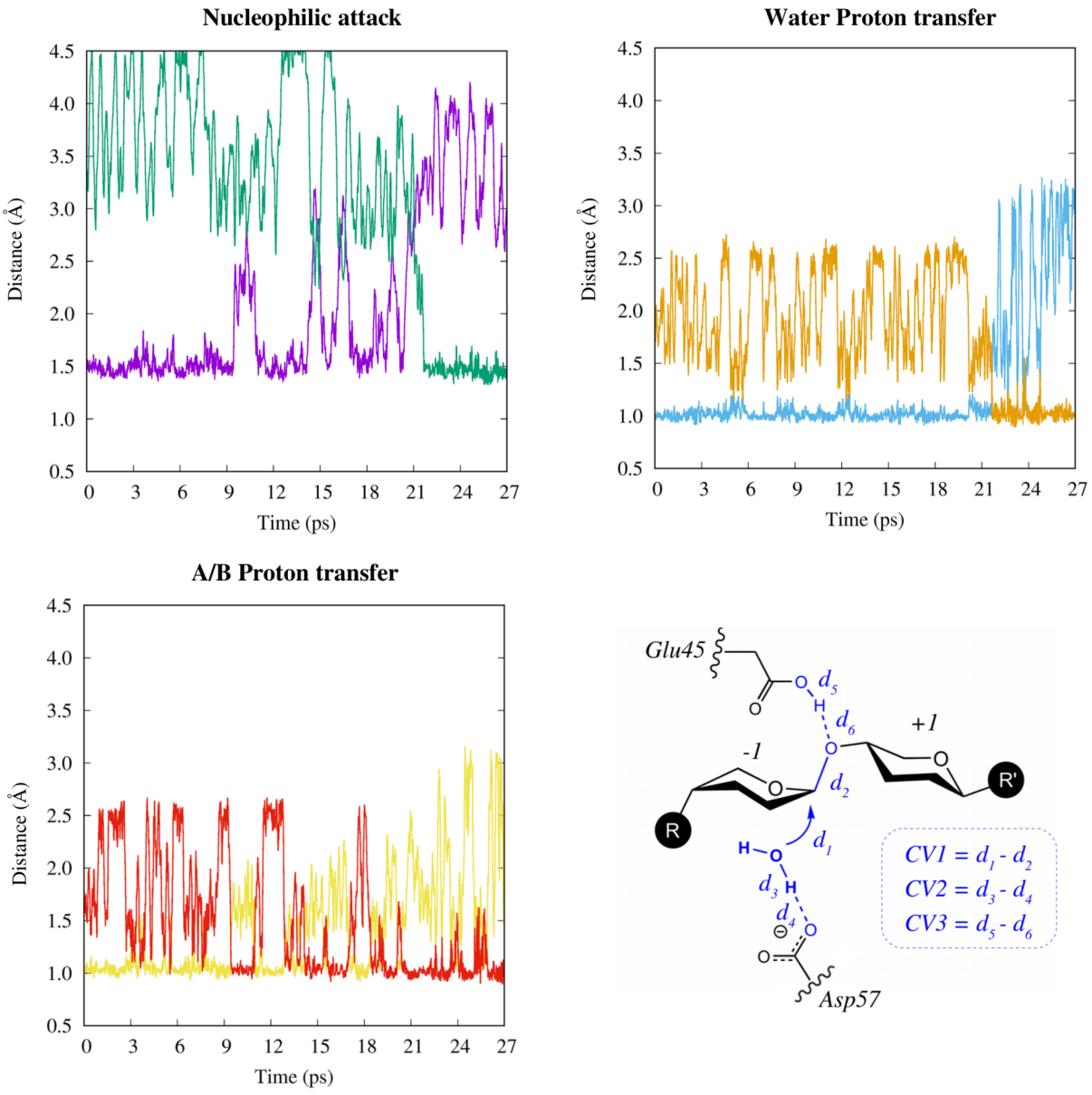


Figure S5.1- Evolution of the distances involved in the nucleophilic attack (top left), water proton transfer (top right) and acid/base proton transfer CVs along the metadynamics simulation. The green line corresponds to the O_{Wat}-C1 distance (d_1), the violet to the C1-O_{Gly} (d_2), the blue to the O_{Wat}-H (d_3), the orange to the H-O_{Asp57} (d_4), the yellow line to the O_{Glu45}-H (d_5) and the red line to the O_{Gly}-H distances (d_6).

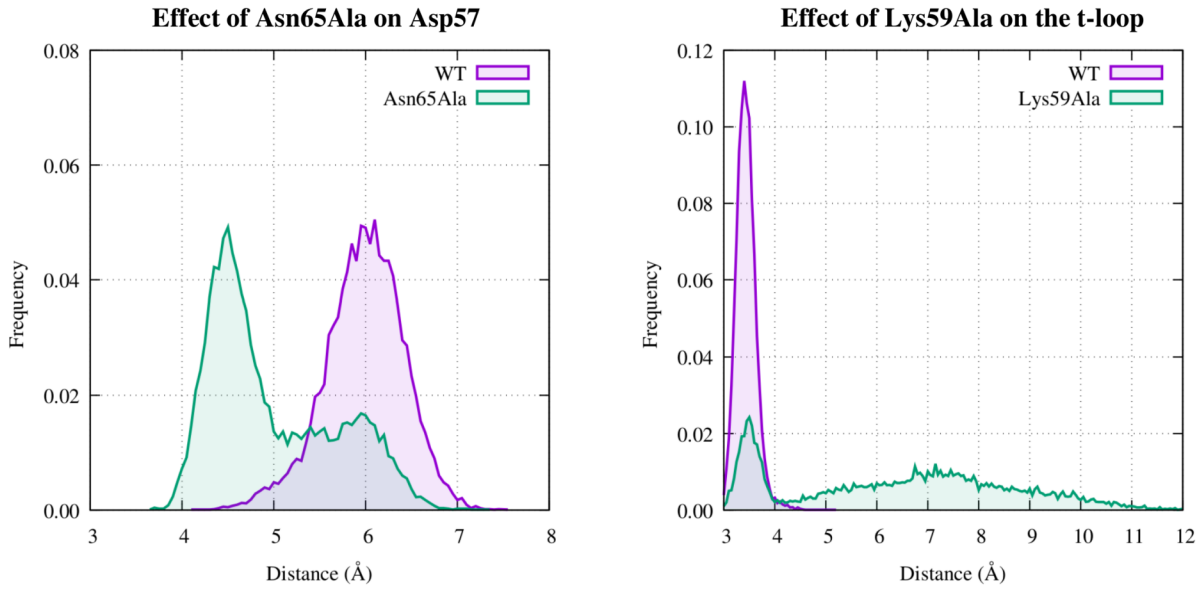


Figure S5.2- (Left) Effect of the Asn65Ala mutation on the conformational preferences of Asp57, monitored with the distance of the Asp57 oxygens to the anomeric carbon of the -1 sugar. Notice that the population shifts from 6.0 at the wild-type (WT) to 4.5 Å at the mutant, a distance that could be suitable for a nucleophilic attack. (Right) Effect of the Lys59Ala mutation on the terminal loop that covers the active center, monitored with a reference distance of the terminal carboxylate (the carbon attached to the oxygens) to the hydroxymethyl group of the -2 sugar. Notice that in the WT the terminal loop is fixed by strong interactions with Lys59, but the loss of these interactions makes the loop open. Frequencies are normalized over 12.000 distances computed for each system, taking structures at regular intervals of 0.01 ns.

Chapter 6
Enzymatic Flexibility: Insights Into the Initial Steps of Glycogenesis

Parts of this chapter are in the process of publication:
M. K. Bilyard, H. Bailey, L. Raich, J. Iglesias-Fernández, S. Seo Lee, C. D. Spicer, C. Rovira, W. W. Yue and B. G. Davis "Palladium-mediated enzyme activation suggests multiphase initiation of glycogenesis" *Under revision* (2018).

ABSTRACT: enzymes are highly dynamic and flexible entities that usually undergo subtle or large conformational changes that are crucial for their activity. In this chapter we study the flexibility and versatility of a polyvalent enzyme called glycogenin (GYG), a glycosyl transferase of family 8 that is able to attach glucose units to one of its own residues (Tyr195). Using classical MD, we analyze the conformational flexibility of Tyr195 and the loop that contains it, finding out how they adapt to different lengths and conformations of acceptor substrates. Furthermore, we unravel the catalytic mechanism of the reaction by means of QM/MM metadynamics, unveiling a prototypical S_Ni-*like* mechanism that evolves through a short-lived oxocarbenium intermediate. The predicted reaction free energy barrier (~10 kcal·mol^{-1}) is very low in comparison to other GTs, which we mainly attribute to an optimum donor/acceptor interaction and the presence of several charged moieties in the active site. This prompted us to explore the unbinding mechanism of the product, revealing that it could be the rate determining step of catalysis.

6.1 Introduction

6.1.1 Do not forget to close the lid for the reaction

Enzymes are not only impressive catalysts for their ability to enhance reaction rates, but also for the versatility of their folds to adapt for acquiring new functions.[31,34] The old view of enzymes as rigid entities has been replaced by models that consider enzymatic flexibility as a fundamental feature to explain regulation, cooperativity or promiscuity.[28,33] Conformational transitions between different states of an enzyme are associated to changes in the solvation of active sites, variations of the dielectric constant or shifts in the pK_a of crucial residues,[228] modifications that may be important both for substrate recognition and catalysis.

Glycosyl transferases (GTs), the enzymes that catalyze the synthesis of sugars, are not an exception in terms of dynamism. Several structural studies have revealed "open" and "closed" conformations of flexible loops that act as "lids" that cover the active center, protecting the nucleotide substrate and defining the acceptor binding site (see Figure 6.1).[229,230] For instance, in the case of a β-1,4-galactosyltransferase (Gal-T1), X-ray experiments have unveiled that the enzyme is in an open conformation in its *apo* form, but in a closed conformation when it is complexed with the donor substrate (UDP-Gal).[231] Similarly, the mammalian α-1,3-galactosyltransferase (α3GalT)[93,94] or the bacterial glycogen synthase (GS)[232,233] and maltodextrin phosphorylase (MalP)[23] are also known to display open and closed conformations of flexible loops and domains. In other words, there is compelling evidence suggesting that conformational changes such as the lid closure must precede the chemical reaction.

All these conformational reorganizations are usually highly dynamic motions, and many techniques are needed to capture their dynamical features. A good example of this complexity is reflected in the work of Lira-Navarrete *et al.* about polypeptide GalNAc-transferases (GalNAc-Ts),[234] enzymes that control protein O-glycosylation. Using a diverse set of techniques, comprising X-ray crystallography, atomic force microscopy, computational simulations and small-angle X-ray scattering, the authors revealed a complex conformational landscape involving compact and extended structures of the enzyme, demonstrating a dynamic and cooperative mechanism by which one of the two enzyme domains enables the binding of the substrate into a distant domain (see Figure 6.2). Nowadays there are so many experimental and computational evidences about the conformational diversity of enzymes that it is universally accepted that they are highly flexible entities and that their dynamical features are essential for substrate recognition and catalysis.

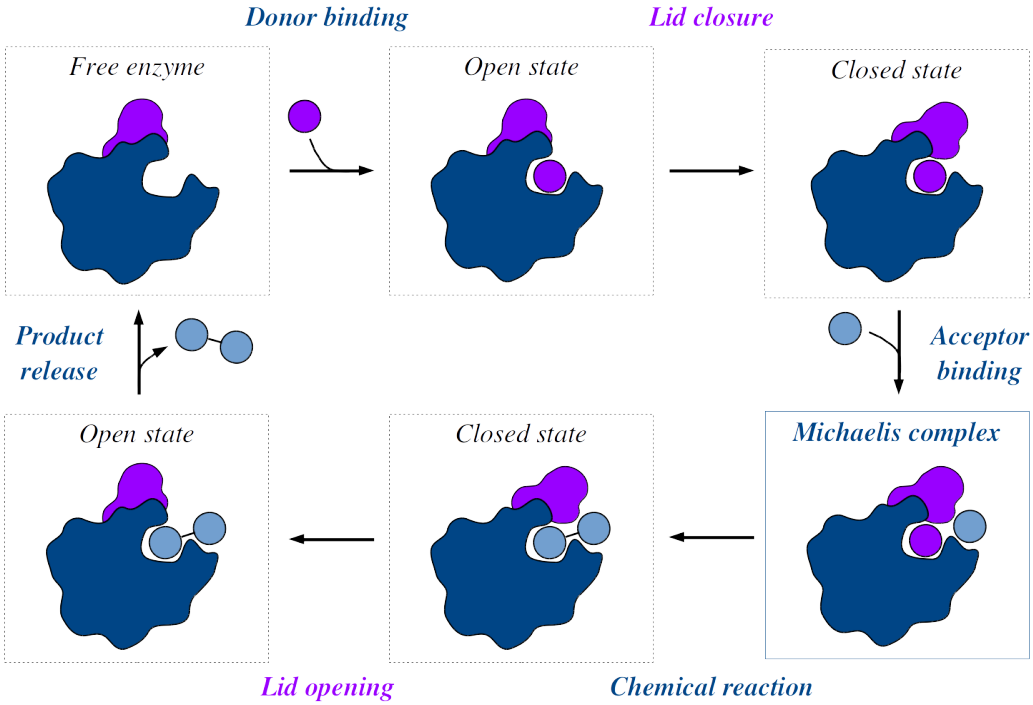


Figure 6.1- The proposed catalytic cycle of GTs possessing a lid segment (shown in violet). After the binding of the donor substrate, the lid closes defining the acceptor binding site, the acceptor binds forming the Michaelis complex, the chemical reaction takes place and subsequently the lid opens to release the product.

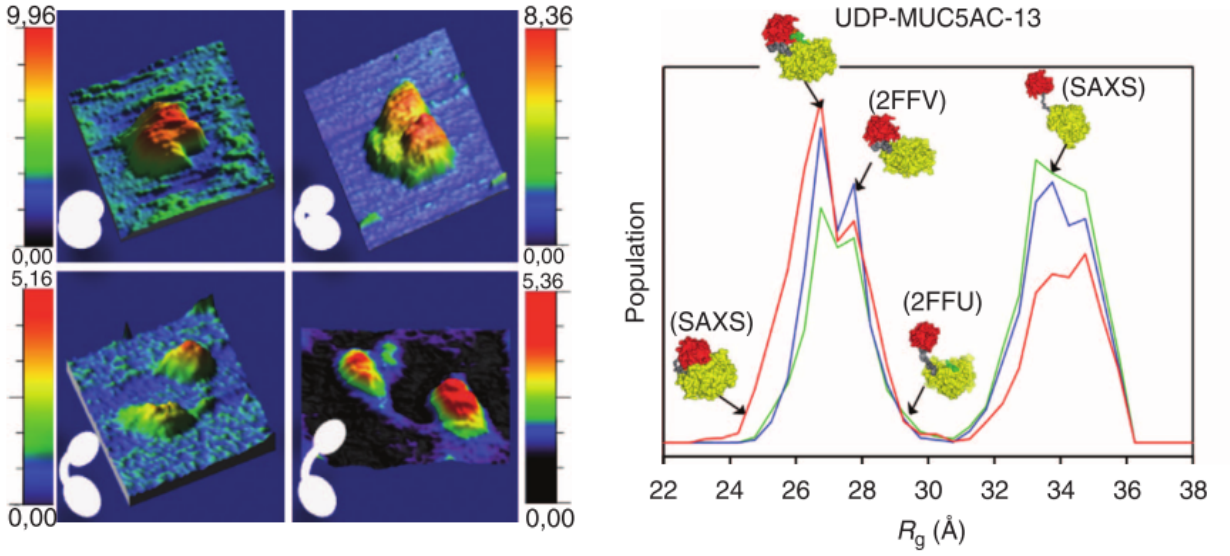


Figure 6.2- Conformational flexibility of GalNAc-T2. (Left) Three-dimensional topography AFM images of single molecules showing different conformations of the enzyme, from compact to extended. (Right) Radius of gyration distributions derived from ensemble analysis of solution scattering curves. Red, blue and green lines represent three different concentrations of GalNAc-T2 in its apo form, which allowed to identify ensembles of compact and extended structures. The two different domains of the enzyme are shown in red and yellow, connected by a flexible linker that is shown in gray. Figure adapted from the work of Lira-Navarrete *et al.*[234]

6.1.2 Glycogenin: the "Swiss army knife" of glycosyl transferases

One of the main energy storage molecules in the human body is glycogen. It is a complex polysaccharide of glucose units that are arranged in linear chains of α-1,4 linkages and branched modifications at certain intervals by α-1,6 glycosidic bonds.[235] The combination of linear and branched chains makes glycogen particles adopt spherical shapes of 10 to 290 nm size that can be detected by electron microscopy experiments.[236] At the core of these spheres there is a protein called *glycogenin* (GYG) that serves as a covalent anchor for the entire polysaccharide. Moreover, GYG is involved in the first steps of glycogen synthesis –also known as glycogenesis– acting as an enzyme that, surprisingly, attaches up to 12 glucose units to itself.[237,238] This multiple auto-glucosylation character is a tangible proof of its clear versatility for catalyzing related –but strictly different– enzymatic reactions.

Several experimental studies have revealed that GYG exists majoritarily as a dimer in solution,[239,240] and crystal structures have shown that the two monomers bind occluding a – mostly– hydrophobic interface (see Figure 6.3). The catalytic process of auto-glucosylation involves the binding of activated UDP-glucose molecules as donor substrates (UDP-Glc) together with a divalent Mn^{2+} cofactor that is properly coordinated to UDP and enzymatic residues. The first glucose addition reaction involves an acceptor tyrosine (Tyr195) located in a helix-turn-helix motif –the *acceptor arm*– that is far from the catalytic center (between 12 and 14 Å; see Figure 6.3).[241] Up to date it is not clear how does the enzyme attach a glucose unit in such a distant glucosylation point, but it is presumed that sizable conformational changes must precede the chemical reaction.[242] After the first step, the attached glucose acts as an acceptor and the α-1,4 polysaccharide grows progressively in a stepwise manner.

Furthermore, GYG also exhibits a flexible lid segment covering the active site, such as the ones depicted in Figure 6.1. This lid segment is shown to undergo major conformational changes depending on the presence of the substrates in the active site, as observed by means of X-ray crystallography.[241] In particular, the lid adopts an *open* conformation in its apo form, and a *closed* conformation when it is complexed with the substrates. These open to closed transitions allow to shield the active site from water entrance and to complete the essential catalytic machinery of GYG by establishing favorable interactions with the substrate.

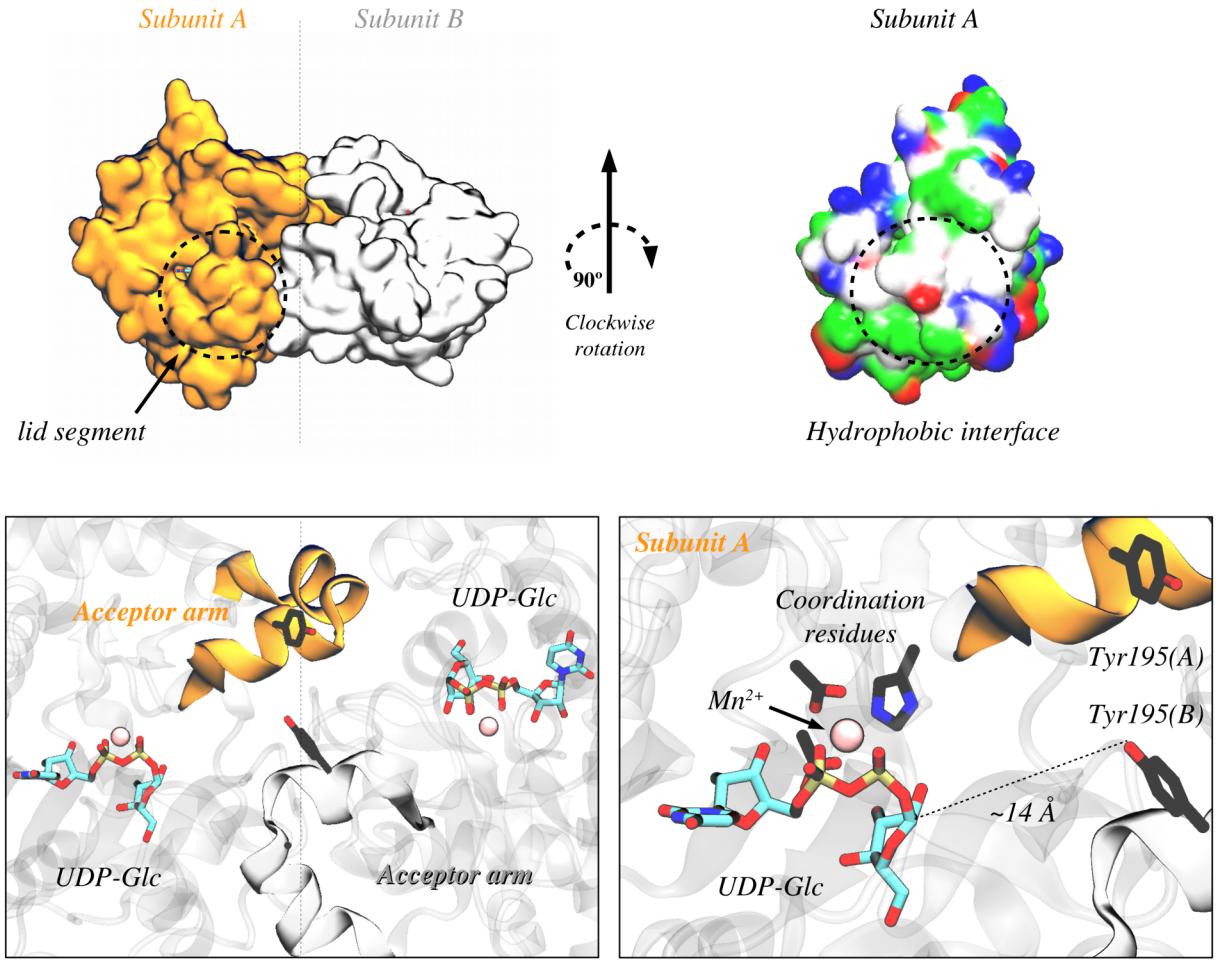


Figure 6.3- (Top) Three dimensional shape of the dimeric GYG and subunit A showing the hydrophobic interface between the two subunits (PDB 3T7O). The residues of the interface have been colored depending whether they are non-polar (white), polar (green), basic (blue) or acidic (red) according to the VMD program.[179] (Bottom) Active site amplification showing the acceptor arms that contain the catalytic Tyr195 and UDP-glucose donor with the Mn^{2+} ion and its coordination residues. This crystal structure also contains a free-glucose monosaccharide nearby the UDP-Glc of subunit B that has not been shown for clarity.

To add more complexity into the dynamical features of GYG, it has been proposed that intra- and intermonomeric transitions of the sugar chains can occur at different steps of the polysaccharide elongation. In particular, for short lengths of 3-4 glucose units the reaction is supposed to be intramonomeric (see Figure 6.4), while for larger lengths (5-12) it is assumed to be intermonomeric due to spatial constraints. This is supported by structural evidences in which product complexes of Glc_4 and Glc_6 chains were trapped, respectively, in intra- and intermonomeric conformations (Figure 6.4).[241] For lengths comprising 0-2 glucose units it is not known whether the reaction is *intra* or *inter*, but it is said that both could be possible after a proper conformational change.

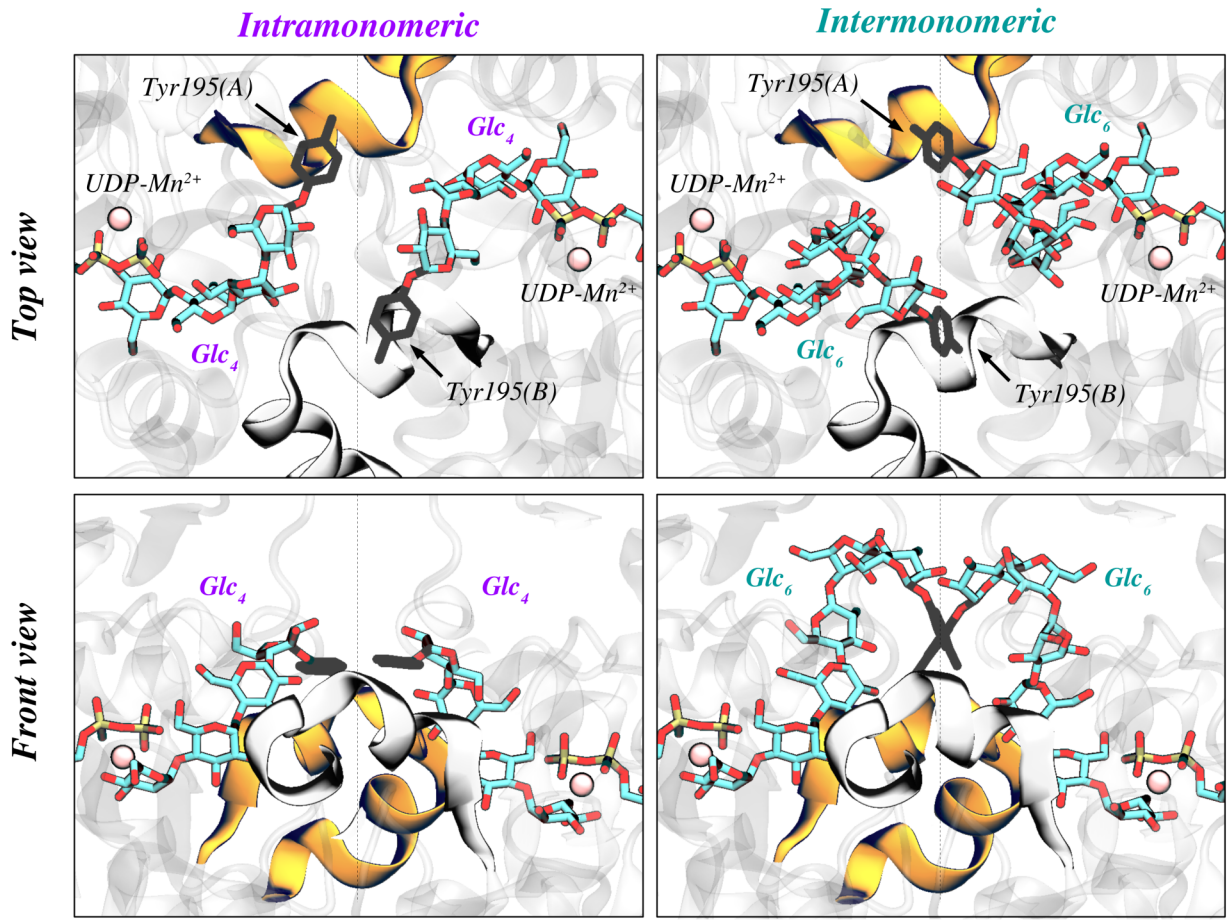


Figure 6.4- Product complexes of maltotetraose (Glc_4, left, PDB 3U2U) and maltohexaose (Glc_6, right, PDB 3U2V) chains attached to Tyr195. They are, respectively, in *intramonomeric* and *intermonomeric* conformations. The orange loop represents the acceptor arm of subunit A, and the white loop the one of subunit B.

Regarding the chemical step, GYG attaches glucose units with retention of configuration at the anomeric carbon. Retaining mechanisms are a topic of discussion in GTs as two possible reaction pathways have been proposed to explain the stereochemical outcome: (i) the S_N2-type double displacement mechanism and (ii) the S_Ni-*like* mechanism (see section 1.2.2 of the General Introduction). The former is expected to take place when there is a suitable nucleophilic residue near the anomeric carbon, prepared for a S_N2 displacement, while the latest is assumed to occur in the lack of such residue. According to the crystal structures of GYG, the only residue that could act as a nucleophile is Gln164, but it is located at a distance of 4.5 Å and its carbonyl is not expected to be a good enough base for a nucleophilic displacement. Thus, it has been proposed that GYG employs an S_Ni mechanism for the synthesis of the α-1,4 maltossacharide. This contrasts with the S_N2-type mechanism that has been tentatively suggested for the GYG extracted from rabbit muscle, which is

slightly different from the human GYG. Concretely, it displays a negatively charged aspartate that is apparently positioned for a nucleophilic displacement, and its substitution to serine or asparagine makes the activity drop by 190-fold.[239] While this catalytic drop is not very pronounced compared to what is observed upon nucleophile mutation in retaining GHs (6 orders of magnitude reduction!)[243], it is undoubtedly a source of controversy concerning the enzymatic mechanism.

In this chapter, to understand the high versatility of GYG, how the different lengths of sugar chains adapt to the enzyme and which is the catalytic mechanism of the reaction, we have performed a three-fold investigation: (i) we have analyzed the stability of different sugar chains in both *intra* and *inter* conformations, revealing an impressive flexibility of the acceptor arm that contains the anchoring Tyr195 and suggesting that for short lengths the reaction could preferably be inter-monomeric; (ii) we have studied the chemical step using QM/MM metadynamics, showing that an S_Ni-*like* mechanism is feasible with a very low reaction free energy barrier; (iii) we have studied the release of the reaction product, which involves the opening of the lid segment prior to its unbinding.

6.2. Results and Discussion

6.2.1. There is always room for one more

In order to study the versatility of GYG for accommodating acceptors of different sugar lengths, we have performed MD simulations with 0 to 3 "n" α-1,4-glucose molecules attached to Tyr195, *i.e.* "Tyr-$(Glc)_n$", both in their *intra* and *inter* conformations. Five independent 50 ns replicas have been carried out to converge the results, obtaining a total of ~0.5 μs of statistical data for each system. To identify reactive conformations, we have analyzed the distance between the nucleophilic hydroxyl of the acceptor molecule (O_{Tyr} for n = 0, O4 for n > 0) and the anomeric carbon of the donor sugar (C1), *i.e.* the "O_{Tyr}-C1" and "O4-C1" distances, using histogram plots (see Figure 6.5). The nearest the distance, the more reactive the conformation has been assumed to be.

Interestingly, sugar chains displaying low glucosylation patterns (n = 0 and n = 1) clearly prefer *inter* conformations, as their peaks appear at 3.8 and 3.2 Å, much below than the peaks observed for *intra* (6.4 and 3.6 Å; see Figure 6.5). In contrast, longer chains of glucose (n = 2 and n = 3) display almost no difference between both conformations, with the peaks appearing at the same distance (both at 3.3 Å for n = 2 and at 3.2 Å for n = 3) and practically equal populations. To explain the observed differences, we have extracted representative snapshots from the probability distributions for their analysis.

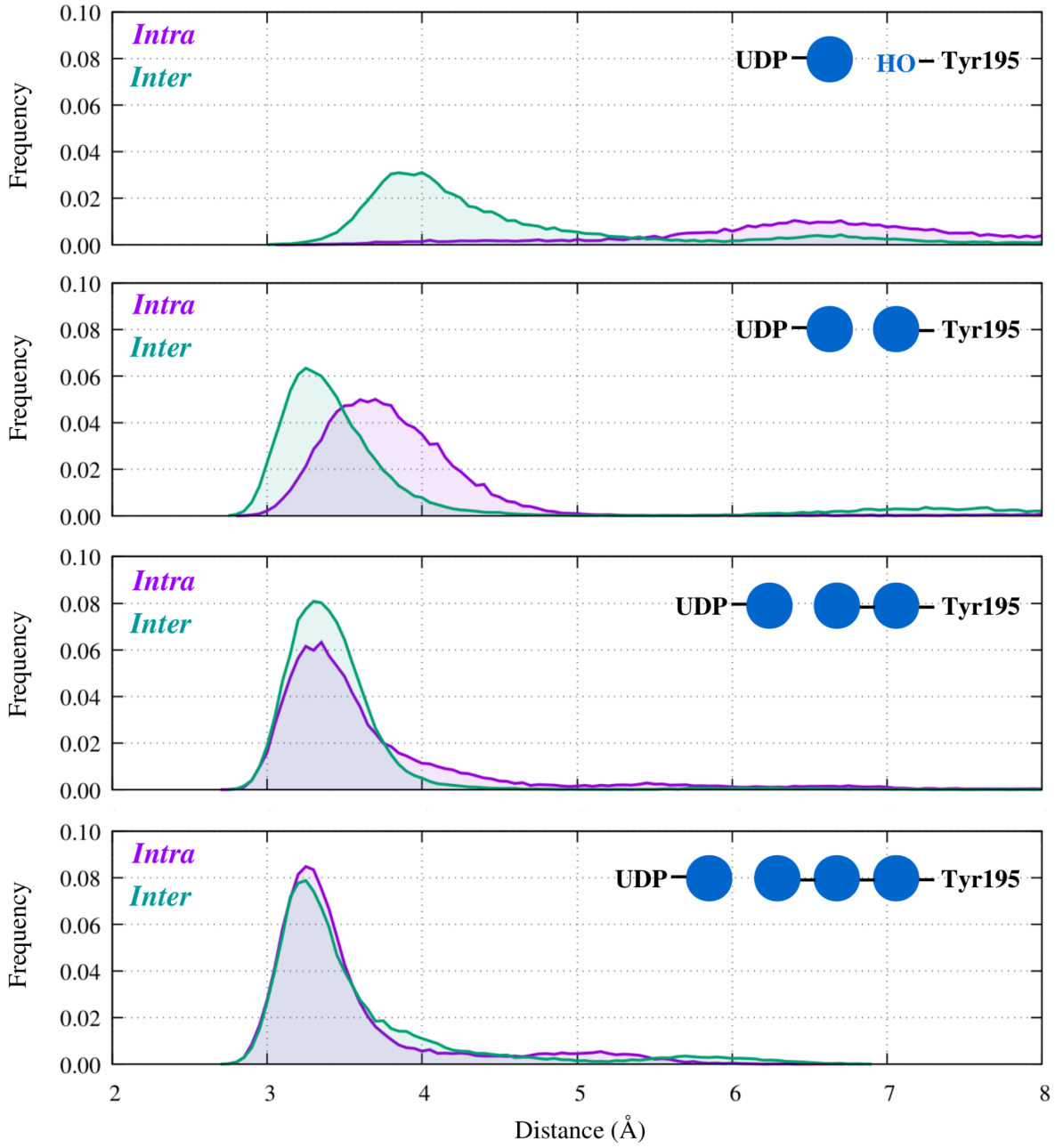


Figure 6.5- Histograms of the nucleophilic distance for n = 0 to n = 3 glucose units (represented as blue circles) attached to Tyr195. Both *intra* and *inter* conformations are represented in purple and cyan surfaces. Frequencies are normalized over 40.000 distances computed for each system, taking structures at regular intervals of 0.01 ns. For n = 0, the peaks of *intra* and *inter* conformations are found at 6.4 and 3.8 Å, at 3.6 and 3.2 Å for n = 1, at 3.3 Å both for n = 2 and at 3.2 Å both for n = 3.

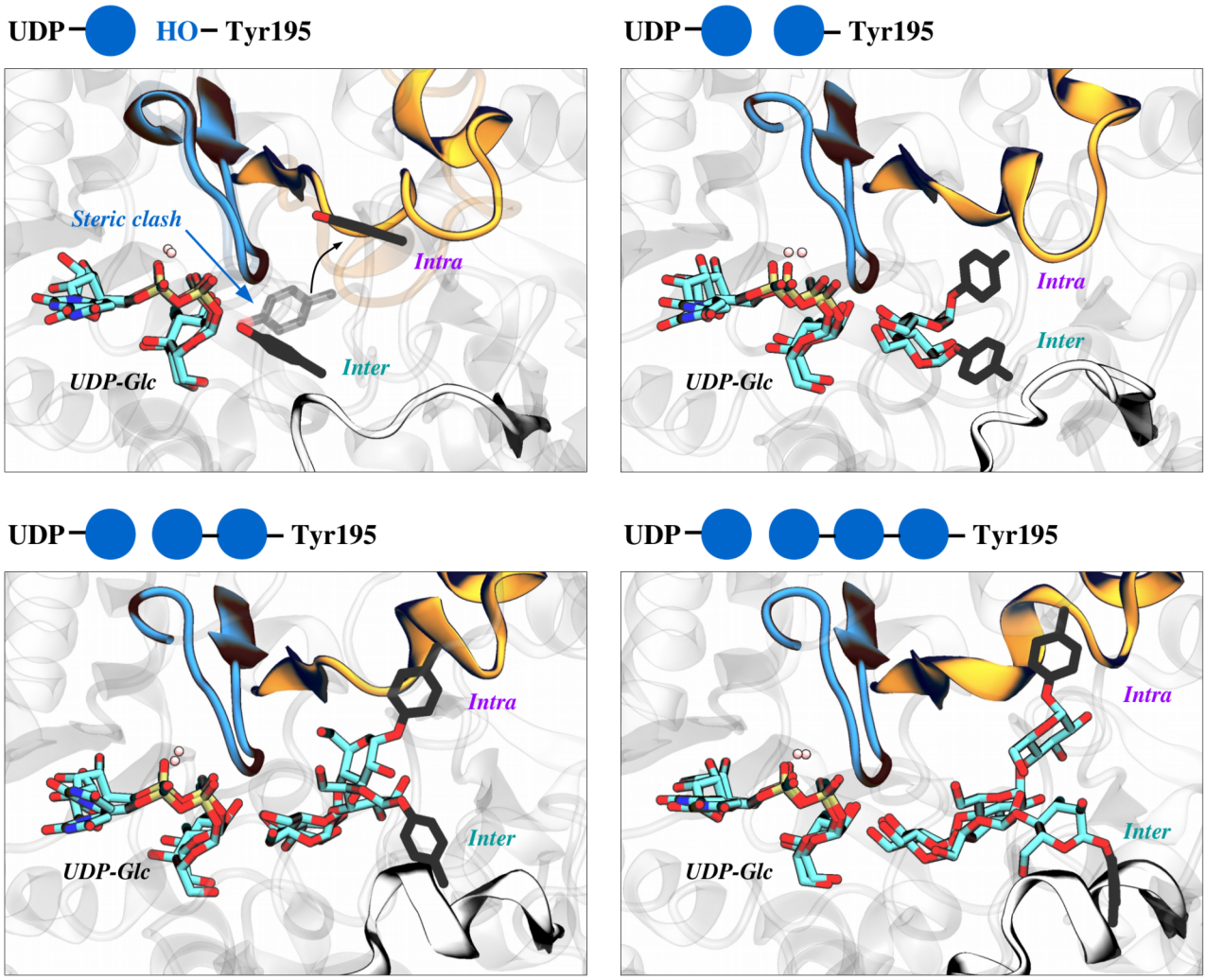


Figure 6.6- Structural snapshots of superposed *intra* and *inter* conformations for n = 0 to n = 3 glucose units (represented as blue circles) attached to Tyr195. The orange loop represents the acceptor arm of subunit A (*i.e. intra*), and the white loop the one of subunit B (*inter*). The blue loop represents an obstacle for the short *intra* conformations, particularly for n = 0, as it clashes with the transparent tyrosine (a non-stable "forced" conformation) and makes it evolve towards the opaque representation upon molecular dynamics. Hydrogens have been omitted for clarity.

A superposition of both *intra* and *inter* conformations revealed three important features to highlight: (i) first, there is a voluminous loop in between the *intra* tyrosine and the UDP-Glc substrate that makes difficult the approach of short chains –principally n = 0, but also n = 1– into the active site, explaining the observed shifts in the histograms (see Figure 6.6). This is not a problem for the *inter* tyrosine, which approaches the active site from "below" and does not have any structural obstacle in the pathway. Moreover, it is neither a problem for long chains (n = 2 and n = 3), which are more flexible and adapt to the environment, circumventing the obstacle. (ii) Second, the loop containing the anchoring tyrosines –either *intra* or *inter*– is more disordered for short chains than for the long ones. This is very relevant if one takes into account that the X-ray structure of the "naked"

GYG (n = 0) displays a well ordered loop and the tyrosine is far from the active site, meaning that it needs to distort the helicoidal shape of the loop when approaching, which could require a non-negligible energetic cost. (iii) Third, the anchoring tyrosine recoils approximately one sugar position after each glucosylation step (see Figure 6.7), remarking its high flexibility and its well-defined pathway for entering in the active site.

While these results clearly suggest that GYG would preferably start its first glucosylation steps (n = 0 and n = 1) from an *inter* conformation, they do not clarify whether it would be *intra* or *inter* for subsequent glucosylation steps. What it is clear is that as long as the polysaccharide chain elongates, the *intra* conformation becomes more prominent, and thus one could expect a conformational transition from *inter* to *intra* for chains with two or more glucose molecules attached to Tyr195. For longer chains (*e.g.* n = 5), based on the results of X-ray crystallography,[241] it is expected that *inter* conformations would be again preferred.

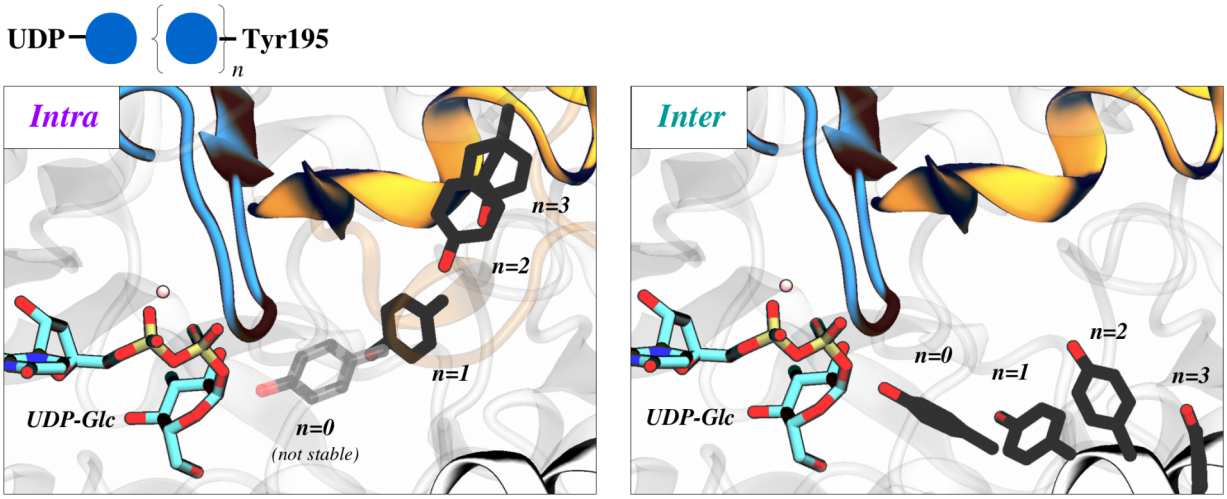


Figure 6.7- Recoil of Tyr195 for different "n" number of glucose units (represented as blue circles) attached to it. Both *intra* and *inter* conformations are displayed separately. The orange loop represents the acceptor arm of subunit A (*i.e. intra*), and the white loop the one of subunit B (*inter*). The blue loop represents an obstacle for short *intra* conformations (see also Figure 6.6). The transparent tyrosine indicates a non-stable "forced" conformation. Hydrogens and acceptor glucose units have been omitted for clarity.

6.2.2. Evidence for an S_Ni reaction mechanism with a short-lived oxocarbenium intermediate

The catalytic mechanism of GYG has been modeled using QM/MM metadynamics. The system with three glucose molecules attached to Tyr195 (n = 3) in an *intra* conformation has been selected in order to model the formation of the products trapped by X-ray experiments (PDB 3U2U, see computational details). Two collective variables have been used to drive the reactants towards the products. The first, named as *nucleophilic attack*, has been defined as a difference of distances be-

tween the C-O bond that breaks (C1-O_P) and the one that forms (C1-O4). Similarly, the second – named as *proton transfer*– is defined as a difference of distances between the H-O bond that breaks (H-O4) and the one that forms (H-O_P).

Structural snapshots along the reaction pathway are shown in Figure 6.8 and the free energy surface is shown in Figure 6.9. These results reveal a stepwise S_Ni-*like* catalytic mechanism in which the C-O bonds dissociate and form on the same face of the donor sugar, retaining the configuration at the anomeric carbon. In the reactants well, two different states with regard to the 4-OH of the acceptor appear to be stable, one in which it is hydrogen bonded to the 2-OH of the donor (R) and another one in which it is hydrogen bonded to O_P (R'). From the last one, the C1-O_P bond elongates reaching a transition state (TS, ~10 kcal·mol^{-1} higher in energy) that leads to a short-lived ion-pair oxocarbenium intermediate (IP, ~1 kcal·mol^{-1} lower than the TS). Subsequently, the O_4 of the acceptor substrate approaches the anomeric carbon (IP') and collapses with it at the same time that the proton of the acceptor hydroxyl is transferred to the UDP molecule (P).

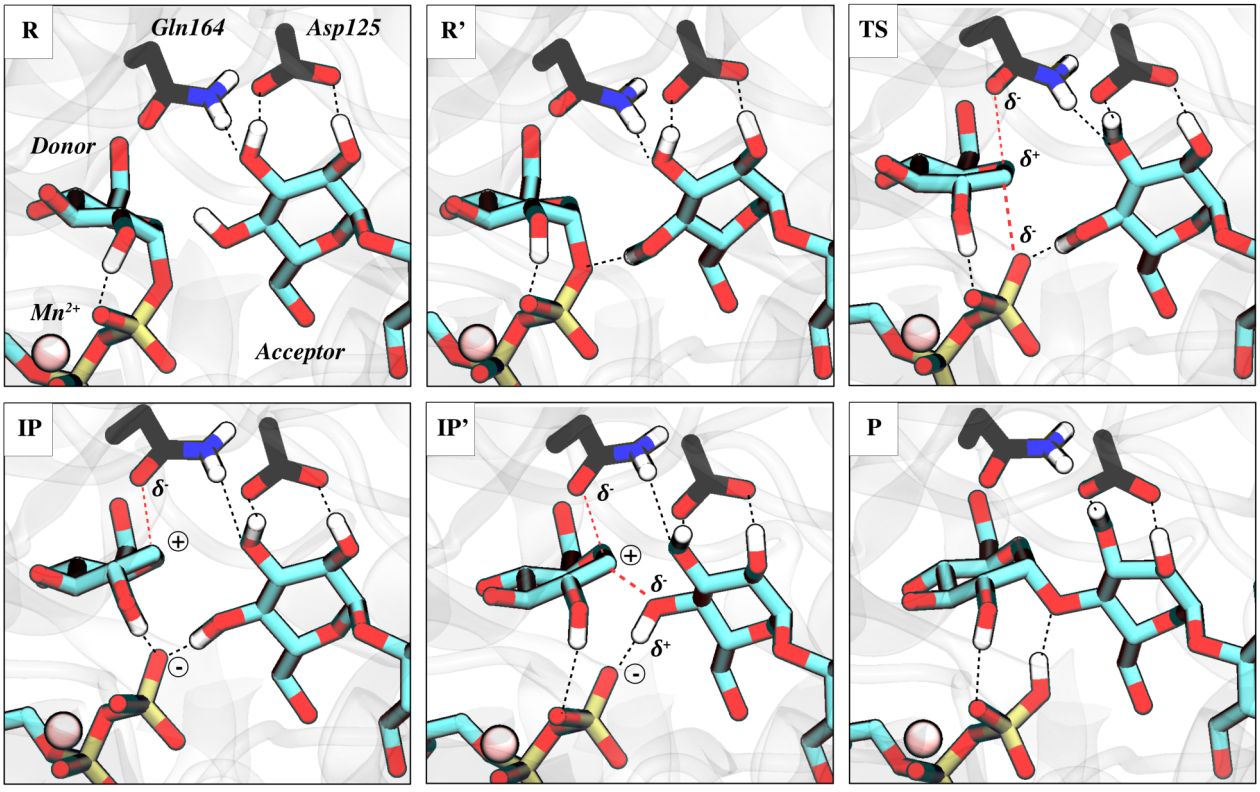


Figure 6.8- Catalytic mechanism of glycosidic bond formation in GYG. R and R' denotes reactants (Michaelis complex), TS the transition state, IP and IP' the ion-pair intermediates, and P denotes the products. Hydrogen atoms have been omitted for clarity, except the 2-OH of the donor sugar, the 2-OH, 3-OH and 4-OH of the acceptor, and the ones of Gln164. The last one is part of the MM region and its charges are not polarized, so the δ^- symbol just indicates that the negative charge of the oxygen stabilizes the oxocarbenium ion.

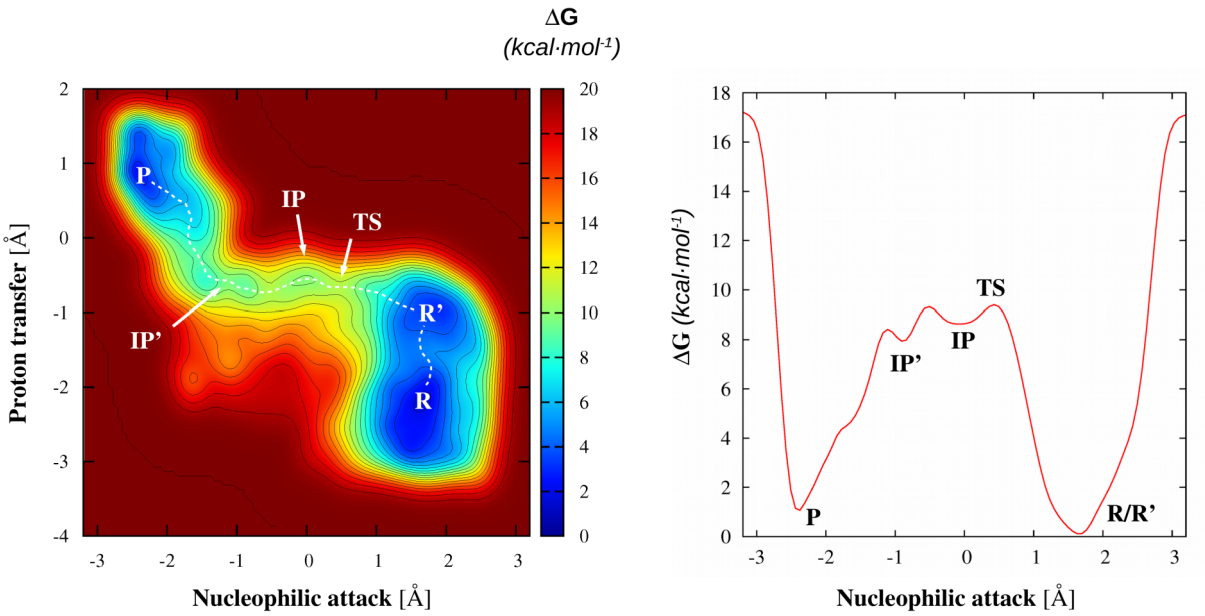


Figure 6.9- Computed free energy landscape of the GYG reaction (left) and monodimensional projection over the nucleophilic attack (right). Contour lines are given at 1 kcal·mol^{-1}. The labels are referred to the ones provided in Figure 6.8.

At this point, several mechanistic details are worth noting: (i) the Gln164 residue just acts as an electrostatic stabilizer for the TS and IP/IP' states, with a minimum distance of ~2.6 Å from the anomeric carbon. The formation of a possible glycosyl-enzyme with Gln164 in its imidic form is unlikely for a clear reason: there is not a good base nearby to deprotonate the amido group. At most, it could be deprotonated by a tyrosine residue that is next to Gln164, but before doing so it should release its proton to another aspartate, a process that is expected to involve higher barriers. Moreover, the modest reduction of activity upon the mutation of the "analogue" residue in rabbit GYG (190-fold[239]) is in agreement with its role as an electrostatic stabilizer. (ii) Once the IP forms, the 2-OH of the donor can change its hydrogen bond partner from one O_P to another O_P of UDP, a local flexibility that has been also observed for some retaining GHs.[50,244] (iii) The ion-pairs are stable upon optimization, but they are short-lived (~1 ps) upon MD. This is in agreement with the fact that the free energy surface is very flat at those regions, and is consistent with results that have been obtained previously in the group.[25,126] (iv) The cleavage of the C1-O_P bond (from R' to TS) is the rate-determining chemical step with a barrier of ~10 kcal·mol^{-1}. This value is very low in comparison with other GTs that have been previously studied, which range from 15 to 20 kcal·mol^{-1}.[20,26,95] This difference of energies can be related with two structural features of GYG: first, the acceptor OH in the reactant state (R') is hydrogen bonded directly to O_P (*i.e.* the bond that has to break), and not to

the other oxygens of UDP as it was found in *e.g.* OtsA, GalNac-T2 or α-3GalT GTs.[20] This stable and strong hydrogen bond could facilitate the dissociation of the C1-O_P bond. Second, while metal-dependent GTs have usually a Mn^{2+} ion interacting with UDP, non-metal GTs have positively charged residues to compensate for the lack of that ion (*e.g.* ToxB has a Mn^{2+} but no charged residues interacting with UDP, while OtsA has no metal but an arginine and a lysine, see Figure 6.10). In the case of GYG, interestingly, there is not only a Mn^{2+} metal in the active site, but also an arginine and a lysine (Arg77 and Lys218). This extremely charged environment could facilitate even more the dissociation of the C1-O_P bond.

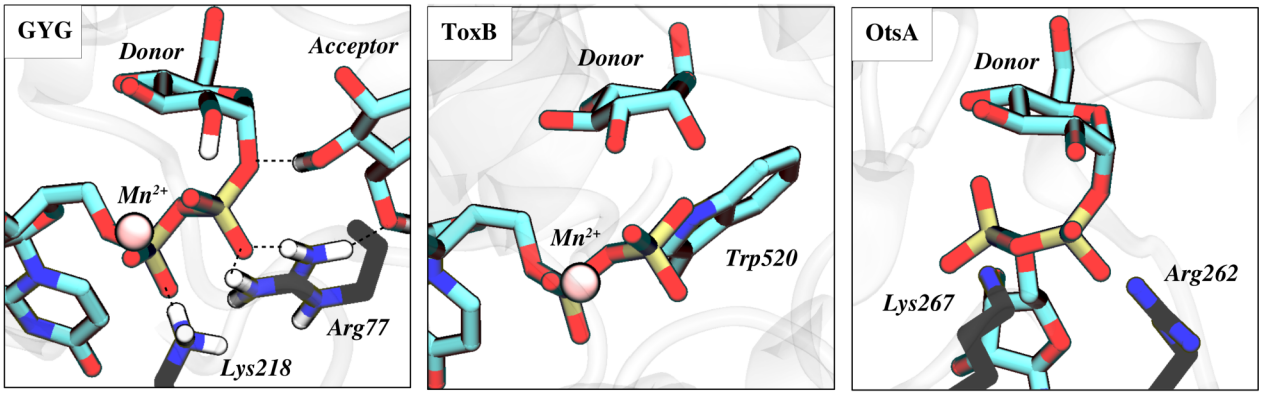


Figure 6.10- Active site comparison between GYG (R'), ToxB (PDB 2BVL) and OtsA (PDB 1UQU). Notice that while GYG has both a Mn^{2+} ion and two charged Arg77 and Lys218 residues interacting with UDP, ToxB has only a Mn^{2+} and a non-charged Trp520, and OtsA has only two charged Arg262 and Lys267 residues but no metal. The structures of ToxB and OtsA do not show hydrogens because they come from X-ray crystallography. The donor substrate show in ToxB is the retaining hydrolysis product of a UDP-Glc molecule.

Altogether, in view of the structural differences pointed above, it seems reasonable to expect a low free energy barrier for the GYG catalyzed reaction. Nonetheless, the experimental free energy value derived from transition state theory is ~20 kcal·mol^{-1},[245] suggesting that either our computed barrier is severely underestimated –*e.g.* by the chosen CVs or the DFT functional– or that there is another rate-determining step (*e.g.* substrate binding, product release or *inter* and *intra* conformational transitions). In order to assess the dependency of the CVs in the mechanism and the free energy barriers, we have repeated the simulations using two coordination numbers that are equivalent to the difference of distances used before, and the results show that there is no quantitative nor qualitative difference in the free energy barriers.[l] Furthermore, we have estimated the performance of the DFT functional used in this work (PBE) by computing the energies of the R and IP states with a

l) Data not shown. Results obtained by Dr. Iglésias-Fernández, co-author of the paper that is under revision.

more sophisticated functional (PBE0, which includes 25% of Hartree exchange), finding an upper bound error of 4.5 kcal·mol^{-1} in the potential energy. Even considering this error, the computed free energy barrier is still far from the ~20 kcal·mol^{-1} that are inferred by experiments, suggesting that the chemical reaction is not rate-limiting.

6.2.3. Product release: *intra* and *inter* conformational transitions could be rate-determining

In view of the results obtained in the previous section, we have studied the catalytic step that follows the chemical reaction: the product release. For that we have taken the product obtained by QM/MM metadynamics and we have performed a re-equilibration of 20 ns with classical MD. From that point, we have explored two distinct product release pathways that seem plausible from structural analyses: (i) the product chain recoils in a stepwise manner without perturbing the protein environment, in the spirit of the "*intra* pathway" envisaged in Figure 6.7; and (ii) the lid segment that closes the active center (see Figure 6.11) opens and subsequently the product chain exits the enzymatic cavity through the space previously occupied by the lid. This last pathway was already discussed on the basis of crystallographic evidence.[241]

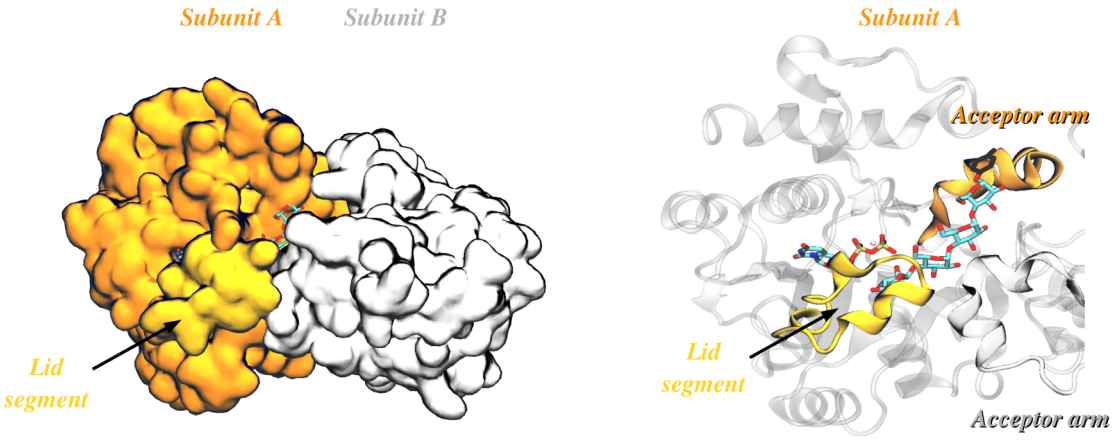


Figure 6.11- Three dimensional shape of the dimeric GYG showing in yellow the "lid" segment that closes the active center of subunit A. The sugar chain and the UDP-Glc molecule of subunit B is not shown for clarity.

We have used a combination of steered molecular dynamics (SMD) and umbrella sampling (US) simulations to generate the initial pathways for the product release and sample their conformational space (see section 5.4 Computational Details). The results for the first pathway, in which the sugar chain recoils in a stepwise manner, show that it is energetically unfeasible, with more than 30 kcal·mol^{-1} of free energy demand for moving the terminal glucose one position back (from "A" to "B" in Figure 6.12). Such high energy cost can be related with the number of h-bonds that the sugar

chain has to break during its movement. It starts with 8 h-bonds that come majoritarily from the interactions that Glc 0 and Glc -1 establish with several residues, including Asp125, Asp160 and Asp163, but also with Arg77 and Gln164 among others. The movement from "A" to "B" causes Glc 0 and Glc -1 loose their 4 h-bonds each (8 in total), but at the same time Glc 0 recovers the 4 h-bonds that were initially belonging to Glc -1 (see graphs in Figure 6.12). This explains why it is so costly to unbind the sugar chain along this pathway: at each step, all sugars need to break again the interactions of their "predecessor" adjacent unit.

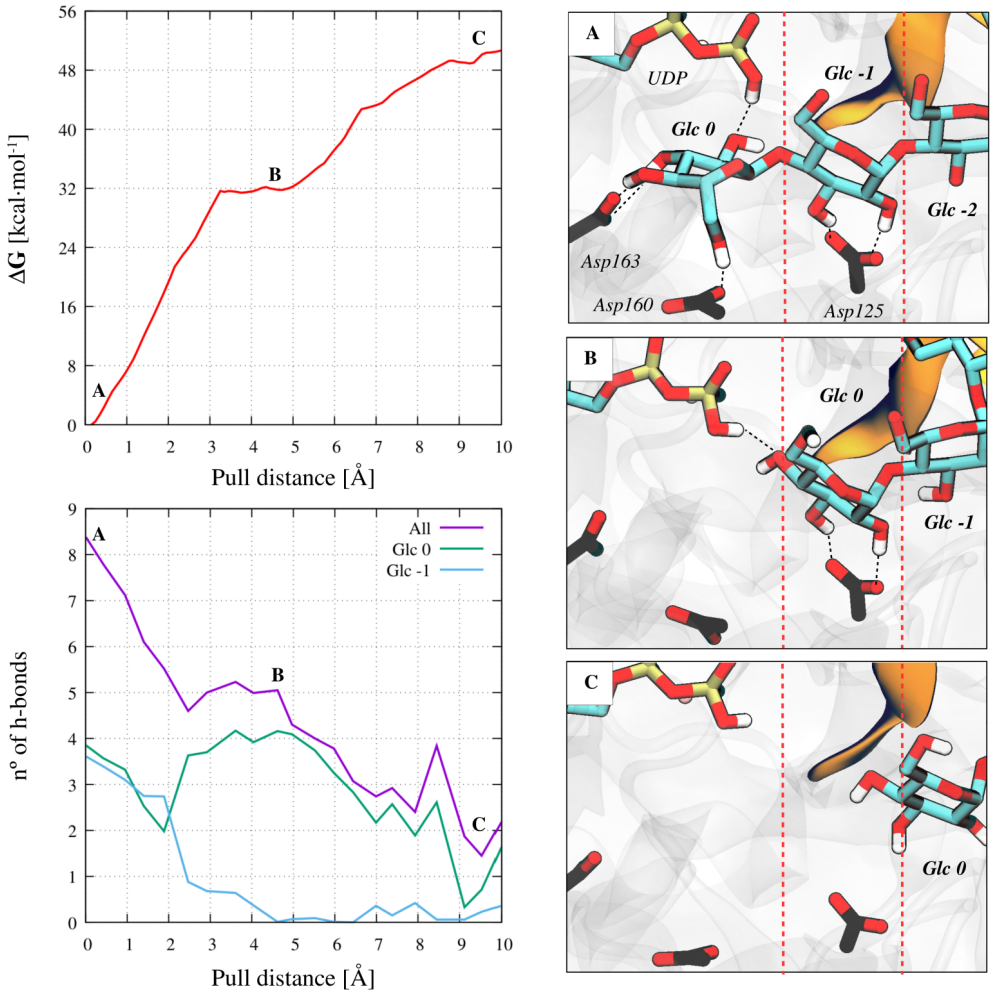


Figure 6.12- Structural snapshots, free energy profile and mean number of h-bonds that the sugar chain establishes with the receptor –enzyme and UDP– along the stepwise "recoil" pathway. The snapshot A corresponds to the product complex just after the chemical reaction, snapshot B corresponds to the chain shifted 1 position (Glc 0 to the position of Glc -1), and snapshot C corresponds to the chain shifted 2 positions (Glc 0 to the position of Glc -2). The Glc 0 and Glc -1 glucose molecules of the chain are equivalent to the donor and acceptor sugars defined previously for the reactant state. The pullout distance has been defined between the α-carbon of Asp163 and the anomeric carbon of the Glc 0 terminal sugar. Notice that in the h-bond graph, after 2 Å there is a "transfer" of h-bonds between Glc 0 and Glc -1. The space left by the sugar chain is filled by water molecules from the solvent, which are not shown for clarity.

Although the "recoil" pathway is unlikely to occur, the results are very informative. In particular, it is remarkable to see how well defined the binding sites are, as from "A" to "B" Glc 0 rotates ending up exactly in the same orientation that Glc -1 adopted at the starting state, and the same happens with Glc -1 moving back one position (it ends up in the same orientation as Glc -2). This highlights the pivotal role of several residues for the stabilization and the orientation of the substrate, particularly Asp125, which could bee seen as a "hook" for an incoming acceptor molecule. Moreover, it is interesting to see that none of the binding residues move after the departure of the substrate, highlighting a local rigidity that is necessary for enzyme specificity.

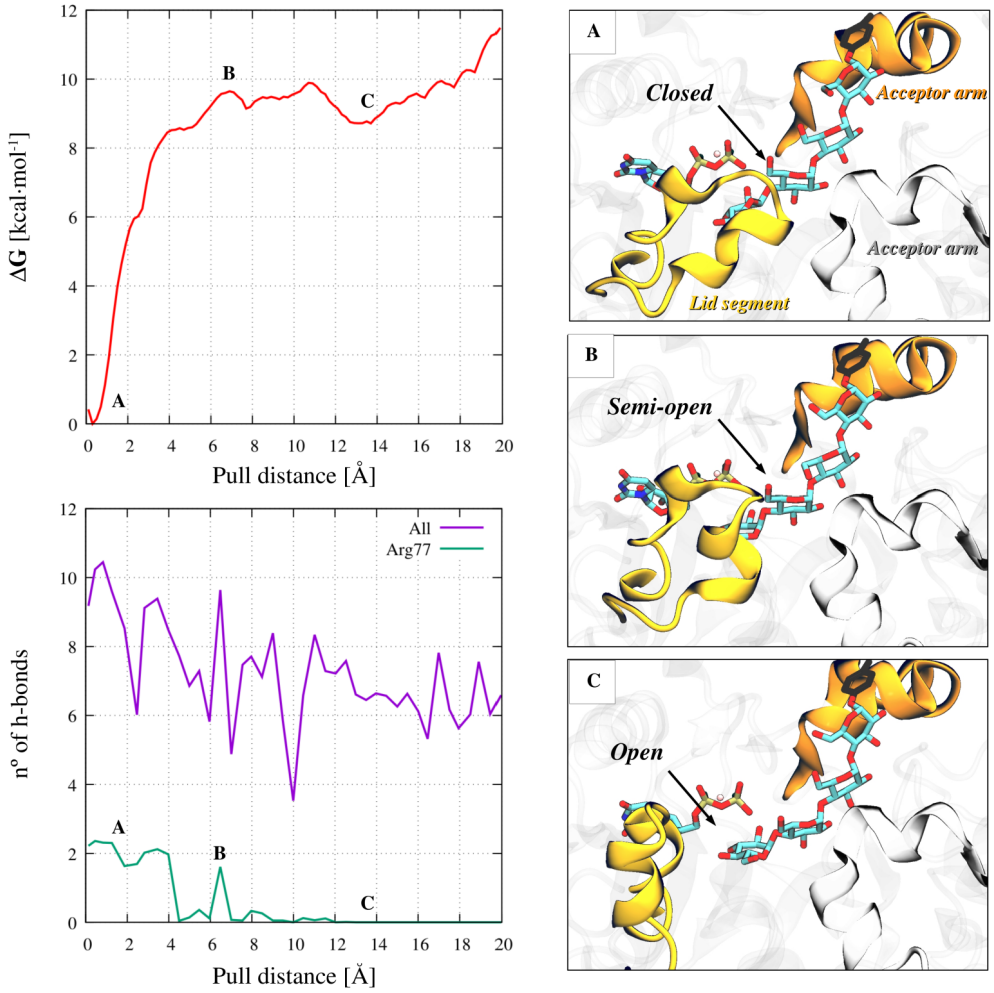


Figure 6.13- Structural snapshots, free energy profile and mean number of h-bonds that the lid segment establishes with the receptor –enzyme, UDP and sugar chains– along the lid opening. The snapshot "A" corresponds to the product complex just after the chemical reaction, snapshot "B" corresponds to the lid in the process of opening, and the snapshot "C" corresponds to the lid opened. The pulling distance has been defined between the α-carbon of the Met74 (located in the lid) and the α-carbon of the Tyr198 (located in the acceptor arm of subunit B). Notice that in the h-bond graph, the whole lid just looses ~2 h-bonds (from 9 to 7), which correspond to the ones of Arg77 (see Figure 6.14).

The second product release pathway has been modeled in two steps: (i) opening of the lid segment, and (ii) product release. Figure 6.13 shows the snapshots and the free energy profile for the first step. It turns out that the lid opens easily with an energy barrier of ~10 kcal·mol^{-1}. This is due to the fact that the interactions that it forms with the acceptor arm of subunit B –colored in white– are very weak, being mainly hydrophobic (see Figure 6.14). The only h-bonds that break during the opening are those involving Arg77, a residue that we have previously seen that is interacting with the UDP-Glc substrate (see h-bond graph in Figure 6.13 and snapshot in Figure 6.14).

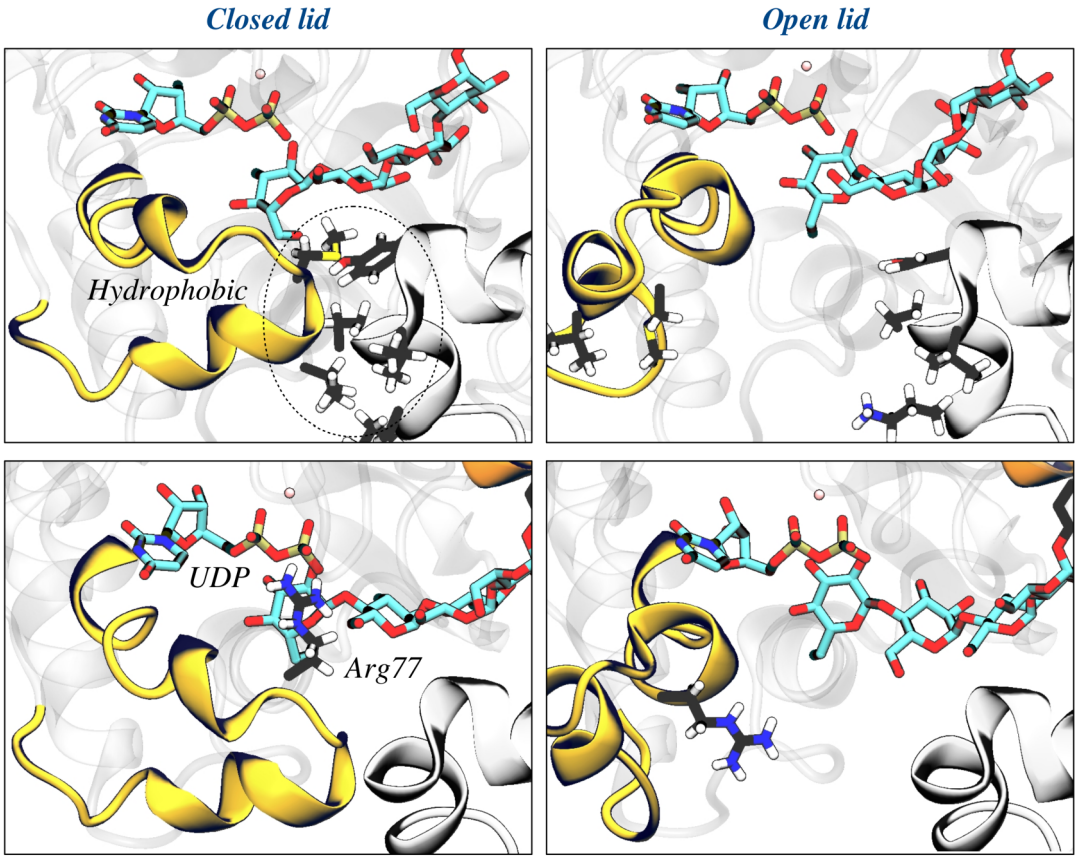


Figure 6.14- Interactions that are lost during the opening of the lid: (top) hydrophobic contacts with the acceptor arm of subunit B and (bottom) h-bond interactions of Arg77 with UDP and Glc -1.

After the lid opening step, we have explored the release of the product by pulling the sugar chain out from the enzymatic cavity. Figure 6.15 shows the snapshots and the free energy profile for this step. Two unbound states are observed in the free energy profile: one in which the sugar chain is interacting with both the acceptor arm and the sugar chain of subunit B (snapshot "B"), and another one (snapshot "C") in which the sugar chain is pointing "up" and establishes few interactions with its own subunit, specifically it interacts with a very movable loop including residues Asn221 to Pro238. The free energy barriers for reaching the first and the second state are 15 and 17 kcal·mol^{-1}

respectively. If the sugar chain is further pulled up to a distance of 30 Å, the free energy barrier increases up to 20 kcal·mol^{-1}, matching the experimental value of the overall catalytic process.

Remarkably, the last energy increase is related with a subtle conformational change of Tyr195, which reorients pointing towards the other subunit (see Figure 6.16). These results, therefore, suggest that the rate-determining step of the whole catalytic process could be not the product release by itself (as it can partially unbind to state "B" and then bind again after the replacement of UDP-Glc), but the transition between intra- and intermonomeric conformations of the substrate.

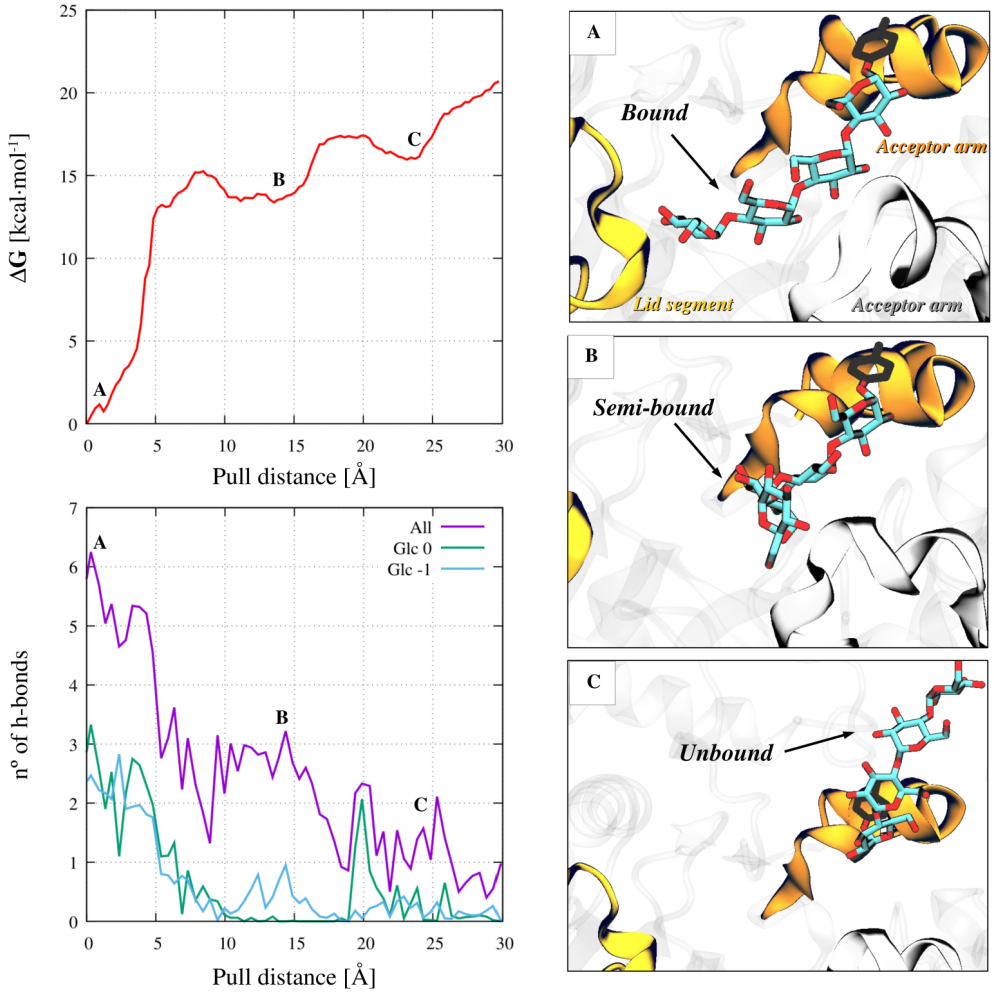


Figure 6.15- Structural snapshots, free energy profile and number of h-bonds that the sugar chain establishes with the receptor –enzyme, UDP and sugar chain of subunit B– along the product release. The snapshot "A" corresponds to the product complex just after the lid opening, snapshot "B" corresponds to the chain outside the enzymatic cavity, interacting with the acceptor arm of subunit B, and the snapshot "C" corresponds to the chain outside the enzymatic cavity but pointing "up" and interacting with a terminal loop of its own subunit. The pulling distance has been defined between the main chain carbon of the Ser131 of subunit A (buried in a stable β-sheet motif) and the anomeric carbon of the terminal sugar. Notice that in the h-bond graph, Glc 0 and Glc -1 account for almost all h-bonds but from 10 Å they vanish and Glc -2 and Glc -3 (not shown in the graph) account for the 3 h-bonds that are maintained until 15 Å, interacting with the sugar chain of subunit B.

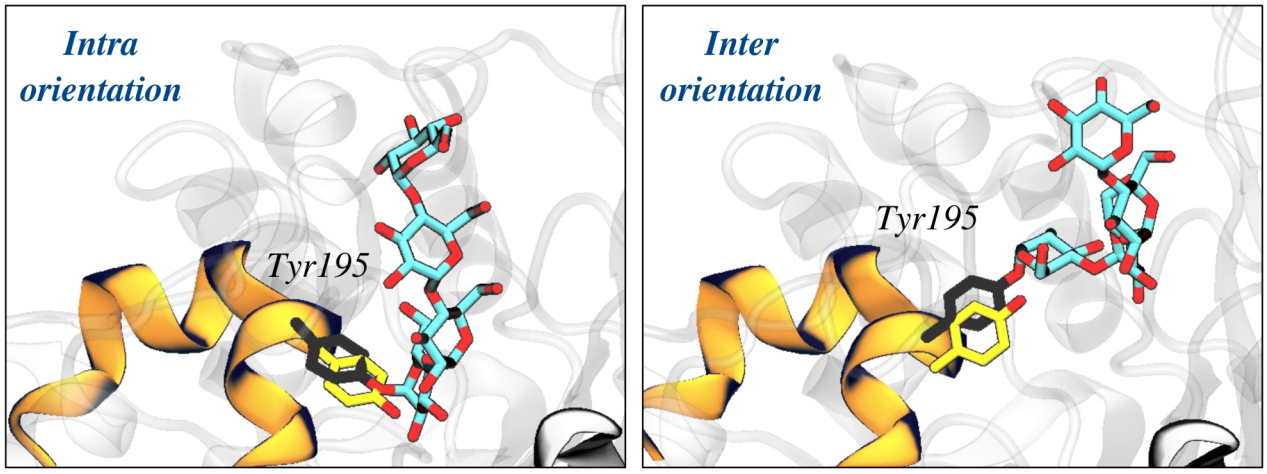


Figure 6.16- Change of Tyr195 orientation from *intra* to *inter* at long pullout distances of the sugar chain. The left image represents the "C" snapshot of figure 6.15, while the right image represents the Tyr195 conformation at a pullout distance of ~30 Å. The tyrosines shown in yellow are the ones of *intra* and *inter* conformations found by X-ray experiments, PDB 3U2U and 3U2V respectively (their sugar substrates, maltotetraose and maltohexaose, are not shown for clarity).

6.3. Summary and Conclusions

In this chapter we have studied the conformational flexibility and versatility of GYG at different stages of its catalytic cycle. First, we have showed how the enzyme adapts to the length of the acceptor substrate (*i.e.* the Tyr195-attached sugars), modifying the shape of the acceptor arms in order to approach short chains (n = 0 and n = 1) into the active center, and letting sugars adapt to the environment in the case of long chains (n = 2 and n = 3). This has allowed us to detect a loop that hampers the accommodation of short chains in an *intra* conformation, suggesting that the first steps of glycogenesis are intermonomeric. Second, we have modeled the chemical reaction using QM/MM metadynamics, finding that it proceeds via a prototypical S_Ni-*like* mechanism that involves a short-lived oxocarbenium intermediate. The Gln164 residue is found to stabilize the transition states and ion-pair intermediates, consistent with its moderate contribution to the experimental reaction rate. The reaction free energy barrier (~10 kcal·mol^{-1}) is very low in comparison to other GTs, something that we relate to an optimal donor/acceptor approach and a highly charged environment around the UDP. Finally, in view of the low energy barrier obtained for the chemical step, we have studied the product release. We have found that it requires the opening of the lid segment with a barrier that is comparable to the one of the reaction, followed by the exit of the product chain from the active site. We have characterized two states for the unbound chain, one in which it is interacting with the acceptor arm and the sugar chain of the contiguous subunit, and another in which it points "up", with

Tyr195 reorienting towards an *inter* conformation. This last state is associated with a free energy barrier that is nearly the one inferred from experiments (~20 kcal·mol^{-1}), suggesting that the RDS of the whole catalytic process could be the conformational transitions from intramonomeric to intermonomeric states. Altogether, the following conclusions can be drawn from the present chapter:

- The first glucosylation steps of GYG (n = 0 and n = 1) occur through intermonomeric conformations. The presence of a "blocking loop" in between the acceptor arm and the active site of the same subunit hampers the formation of intramonomeric conformations. These glucosylation steps require a substantial distortion of the acceptor arm, which could be associated with a high energy cost.

- Sugar chains of n = 2 and n = 3 glucose units adapt to the enzyme active site circumventing the blocking loop. This makes intramonomeric conformations be more likely and competitive compared to the intermonomeric ones, suggesting that intra- to intermonomeric transitions could take place at different glucosylation steps.

- GYG catalyzes the formation of retaining α-1,4 bonds through an S_Ni-*like* mechanism with a short-lived oxocarbenium intermediate. The Gln164 residue acts as an electrostatic stabilizer of the transition states and ion-pair intermediates, approaching its carbonyl to the –charged – anomeric carbon of the donor. The reaction free energy barrier (~10 kcal·mol^{-1}) is very low in comparison to other GTs. This can be attributed to: (i) the reactant state of GYG is hydrogen bonded directly to the O_P of the donor scissile bond, and not to the other oxygens of UDP, as it is the case in several other GTs; and (ii) while most GTs contain either a Mn^{2+} ion *or* two charged residues interacting with UDP, GYG has both a Mn^{2+} ion *and* two charged Arg77 and Lys218 residues.

- Intra- and intermonomeric conformational transitions could be the rate-determining step of the whole catalytic process. These transitions require the opening of the lid segment prior to the unbinding of the product chain –a step that takes place easily given that the interactions of the lid with the enzyme are mostly hydrophobic– and a conformational change of Tyr195, which needs to reorient in order to point towards the active site of the other subunit.

6.4. Computational Details

6.4.1 Modeling of the intra- and intermonomeric Michaelis complexes

Michaelis complex structures with acceptors of different lengths have been constructed from the available product complexes, PDBs 3U2U and 3U2V, which correspond to the dimeric form of the catalytic domain of human glycogenin-1 (GYG) in complex with manganese, UDP and maltotetraose/maltohexaose, respectively. Both protein subunits have been used for the calculations. The protonation states and hydrogen atom positions of all ionizable amino acid residues have been selected based on their hydrogen bond environment and their most favorable state at pH 7, with histidine residues modeled in their neutral state. All crystallographic water molecules have been retained and extra water molecules have been added to form a 15 Å water box around the protein surface. Fourteen sodium ions have been added to achieve neutrality.

In order to construct the Michaelis complexes, for each subunit the terminal sugar of the substrate has been manually attached to UDP, obtaining the *intra* UDP-Glc + $(Glc)_3$-Tyr195 complex (from PDB 3U2U) and the *inter* UDP-Glc + $(Glc)_5$-Tyr195 complex (from PDB 3U2V; this system is not discussed in the chapter). Starting with the former, we have constructed all the shorter *intra* variants by removing the terminal sugar, thus obtaining the UDP-Glc + $(Glc)_2$-Tyr195, UDP-Glc + Glc-Tyr195 and UDP-Glc + HO-Tyr195 complexes. With the latest, we have constructed the shorter variants in the same manner, obtaining the equivalent complexes in an *inter* conformation.

Molecular dynamics (MD) simulations of the GYG enzyme complexes have been performed with the Amber11 software package.[176] The protein has been modeled with the FF99SB[110] force field, whereas carbohydrates and water molecules have been modeled with the GLYCAM06[178] and TIP3P[113] force fields, respectively. The MD simulations have been carried out in several steps. First, the system has been energy minimized with a four-step minimization procedure, allowing to relax sequentially (i) the sugar donor, (ii) both the donor and the acceptor substrates, (iii) the substrates and all water molecules and (iv) the whole system. Structural restraints have been applied to specific hydrogen bonds during this step, the equilibration and the first 10 ns of MD in order to obtain productive complexes. These hydrogen bonds have been selected according to the interaction patterns that are present in PDBs 3T7O, 3U2V and 3U2U. To gradually reach the desired temperature, weak spatial constraints have been initially added to the protein and substrate, while water molecules and the sodium ions have been allowed to move freely at 100 K. The constraints have been then removed and the working temperature of 300 K has been reached after two more 100 K heating steps in the NVT ensemble. Afterwards, the density has been converged up to water density at

300 K in the NPT ensemble and the simulation has been extended to 50 ns in the NVT ensemble. Four additional replicas of 50 ns each have been launched after the equilibration phase to enhance the conformational sampling, leading to 250 ns for each system (0.5 μs of simulation data considering the two subunits). From each of the 50 ns simulations, the last 40 ns –free of any restraint– have been used for the histogram analyses showed in Figure 6.5. Convergence of these histograms has been checked by the cumulative addition of more replicas. Analysis of the trajectories has been carried out using standard tools of Amber and VMD.[179] Particularly, the hydrogen bond analyses has been carried out using the *cpptraj* utility from Amber14,[246] taking into account all the interactions between each substrate and receptor with a distance cutoff of 3.0 Å between heteroatoms and 135º for the angle that defines the hydrogen bond.

6.4.2 Modeling of the chemical reaction

A snapshot of the classical MD simulation of the *intra* UDP-Glc + $(Glc)_3$-Tyr195 complex has been taken for the QM/MM MD simulations. We have used the method developed by Laio *et al.*,[133] which combines Car–Parrinello MD,[107] based on Density Functional Theory (DFT), with force-field MD methodology. The QM region has been chosen as to include the two terminal sugar units of the maltotetraose molecule, the UDP phosphate groups, the active site Mn^{2+} ion and its coordination shell formed by residues Asp102, Asp104 and His212 (a total of 83 QM atoms). The QM region has been enclosed in a 18.1 x 18.6 x 23.0 $Å^3$ supercell. Kohn–Sham orbitals have been expanded in a plane wave basis set with a kinetic energy cutoff of 70 Ry. The QM/MM interface has been modeled by the use of a carbon monovalent pseudopotential that saturates the QM region. Troullier-Martins *ab initio* pseudopotentials have been used for all elements.[174] The PBE functional[173] in the generalized gradient-corrected approximation of DFT has been used, in consistency with previous works on glycosyltransferase reaction mechanisms.[25,93,126] A constant temperature of 300 K has been reached by coupling the system to a Nosé-Hoover thermostat.[247] A time step of 0.12 fs and a fictitious electron mass of 500 a.u. for the Car-Parrinello Lagrangian have been used. The electrostatic interactions between the QM and MM regions have been handled via a fully Hamiltonian coupling scheme, where the short-range electrostatic interactions between the QM and the MM regions are explicitly taken into account for all atoms. An appropriately modified Coulomb potential has been used to ensure that no unphysical escape of the electronic density from the QM to the MM region occurs. The electrostatic interactions with the more distant MM atoms have been treated via a multipole expansion. Bonded and van der Waals interactions between the QM and the MM re-

gions have been treated with the standard AMBER force-field. Long-range electrostatic interactions between MM atoms have been described with the P3M implementation,[180] using a 64 x 64 x 64 mesh.

The QM/MM MD simulations have been coupled with the metadynamics algorithm[139] to model the enzymatic reaction and reconstruct the free energy landscape (FEL) of the glucosylation reaction. Two collective variables (CVs) have been selected according to the bonds that break and form during the reaction. We have used the metadynamics driver provided by the Plumed2 plugin.[183] The first collective variable (CV1) has been defined as the difference of distances between the O_P-C1 of the donor and the C1-O_{Gly} of the acceptor sugars. This variable accounts for the break of the UDP-Glc bond and the *nucleophilic attack* of the acceptor. The second collective variable (CV2) has been defined as the difference of distances between the O_{Gly}-H of the acceptor and the H-O_P of UDP. This variable, thus, accounts for *proton transfer* between the two molecules. A Gaussian height of 1 kcal·mol^{-1} and a deposition time of 30 fs (250 MD steps) have been used to explore the FEL. The Gaussian width for each CV has been set to 0.2 Å. Walls at 4.0 Å for each distance and 1.5 Å for the proton transfer variable has been used to reduce the FEL space to the chemical event. Additionally, a wall at 2.5 Å has been set to the 2-OH···O_P' hydrogen bond that the donor substrate establishes with UDP, as we have found that it is lost in the product state and its lack makes recrossing unlikely. This distance is large enough to allow the 2-OH be flexible and to discard artifacts in the results. First crossing criterion has been taken to determine the simulation end. The FEL has been completed after 869 deposited Gaussians. Two additional replicas with random velocities have been simulated from the reactant state (bearing 80 deposited Gaussians) to estimate the error in the free energy barrier, which has been found to be of 9.6 ± 1.5 kcal·mol^{-1}.

6.4.3 Modeling of the product release

We have taken a snapshot of the products obtained by QM/MM metadynamics and we have performed a re-equilibration of 20 ns with classical MD. Subsequently, steered molecular dynamics[248,249] (SMD) and umbrella sampling[140,141] (US) simulations have been performed to explore the product release step. The first method has been used to generate the initial pathways from which the last method explored the phase-space. For each simulation, one collective variable (CVs) has been defined to differentiate between open/closed or bound/unbound states. The CV to pullout the product chain through the "recoil" pathway has been defined as the distance between the α-carbon of the Asp163 of subunit A (buried in a stable α-helix motif) and the anomeric carbon of the terminal

sugar. The CV to open the lid segment has been defined as the distance between the α-carbon of the Met74 of subunit A (located in the lid) and the α-carbon of the Tyr198 of subunit B (located in the acceptor arm). The CV to pullout the product chain after the lid opening has been defined as the distance between the main chain carbon of the Ser131 of subunit A (buried in a stable β-sheet motif) and the anomeric carbon of the terminal sugar. These distance-based CVs are commonly used to investigate enzyme-ligand binding pathways,[250–252] and although it is usually desirable to include additional metrics –such as RMSD– that take into account possible slow conformational motions,[253] in this case the use of distances are justified given that the ligand is always bound to the enzyme and its conformational space outside the enzymatic cavity is limited.

To generate the initial pathways, a total of 30 trajectories –10 for each pull– with random velocities have been launched from a reference structure taken from the equilibration MD, leading them relax for 1 ns. Subsequently, a movable harmonic potential of 50 $kcal \cdot mol^{-1} \cdot A^{-2}$ has been used to drive the CVs from bound to unbound states –closed to open in the case of the lid – during 2 ns, with pulling velocities between 5-15 $Å \cdot ns^{-1}$. The trajectories with lower energies have been taken as the initial points for the US simulations. Each pathway has been divided in windows spaced at regular intervals of 0.5 Å, thus obtaining 21 windows for the "recoil" pathway, 41 for the lid opening and 61 for the subsequent product release. Force constants of 10 $kcal \cdot mol^{-1} \cdot A^{-2}$ have been used for the harmonic potentials, increasing it to 30 $kcal \cdot mol^{-1} \cdot A^{-2}$ for windows in which the distributions have been displaced from their centers due to steep slopes. Every window has been sampled during 10 ns, leading to a total of ~1.2 μs of simulation data. The first two nanoseconds of each window have been considered as an equilibration step, and thus only the last 8 ns have been used for statistical purposes. Free energy profiles have been reconstructed using the weighted histogram analysis method (WHAM),[147,148] defining 100 bins for the histogram and a convergence threshold of 10^{-5} for the iterative optimization of the free energy values. The convergence of the profiles has been checked by looking at the cumulative evolution of the profiles over time, considering them converged when errors –taken as the standard deviation of the three last nanoseconds of simulation– have been below 1 $kcal \cdot mol^{-1}$. Additionally, bootstrap analyses with 20 data sets have been performed for each profile, with errors below 0.3 $kcal \cdot mol^{-1}$.

6.5 Supplementary Figures

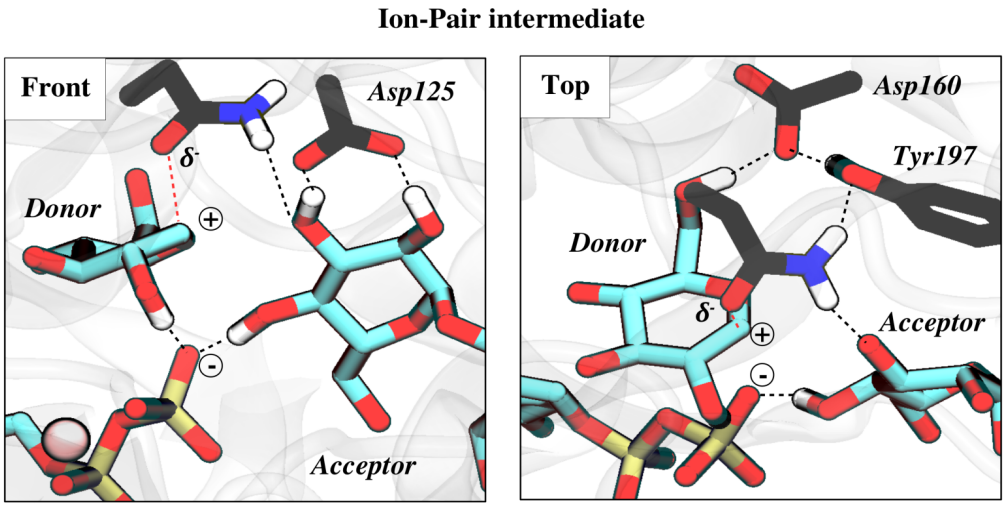


Figure S6.1- Top and front views of the hydrogen bond network of Gln164 at the ion-pair intermediate. Notice that in order to form a glycosyl-enzyme intermediate (GEI), Gln164 must release its proton to Tyr197, and at the same time this last one must release its proton to Asp160, whose basicity is decreased by the hydrogen bond that it establishes with the hydroxymethyl group of the donor. In other words, the formation of a GEI looks *a priori* unlikely.

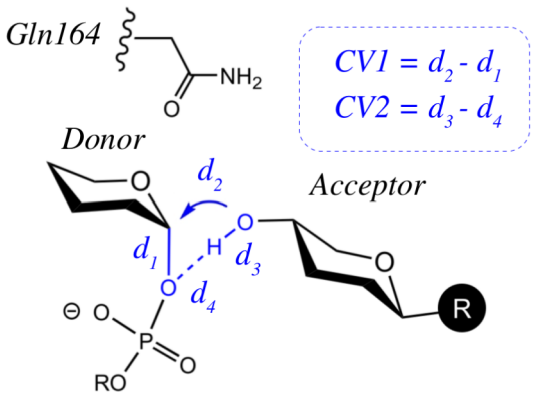


Figure S6.2- Collective variables used to study the reaction mechanism: nucleophilic attack (CV1) and proton transfer (CV2).

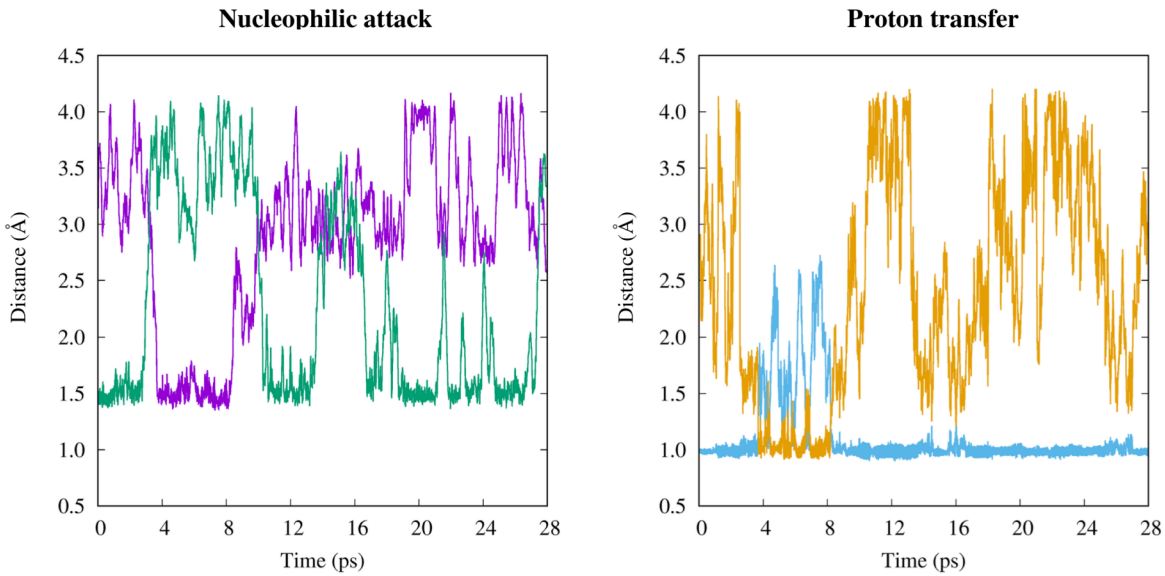


Figure S6.3- Evolution of the distances involved in the nucleophilic attack (left) and proton transfer (right) CVs along the metadynamics simulation of the reaction. The green line corresponds to the C1-O_p distance (d_1), the violet to the O_{Gly}-C1 (d_2), the blue to the O_{Gly}-H (d_3) and the orange to the O_p-H (d_4) distances.

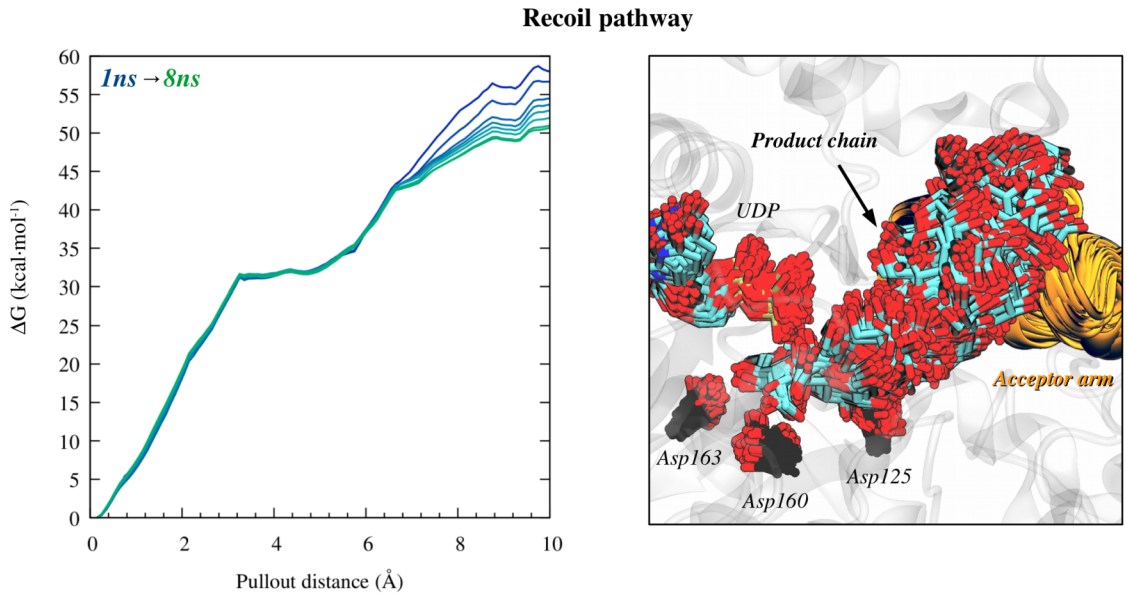


Figure S6.4- Time-convergence of the US free energy profiles of the product release (left) and structural superposition of the product chain conformations along the recoil pathway (right). Time evolution is represented in a gradient of colors, from blue to green, with a total of eight cumulative profiles. Only the last nanosecond of simulation of each US window is used for the superposition, taking structures at intervals of 250 ps. Notice that the free energy profiles only differ from > 7 Å given that the product chain has more mobility (see right panel) as it gets out from the enzymatic cavity.

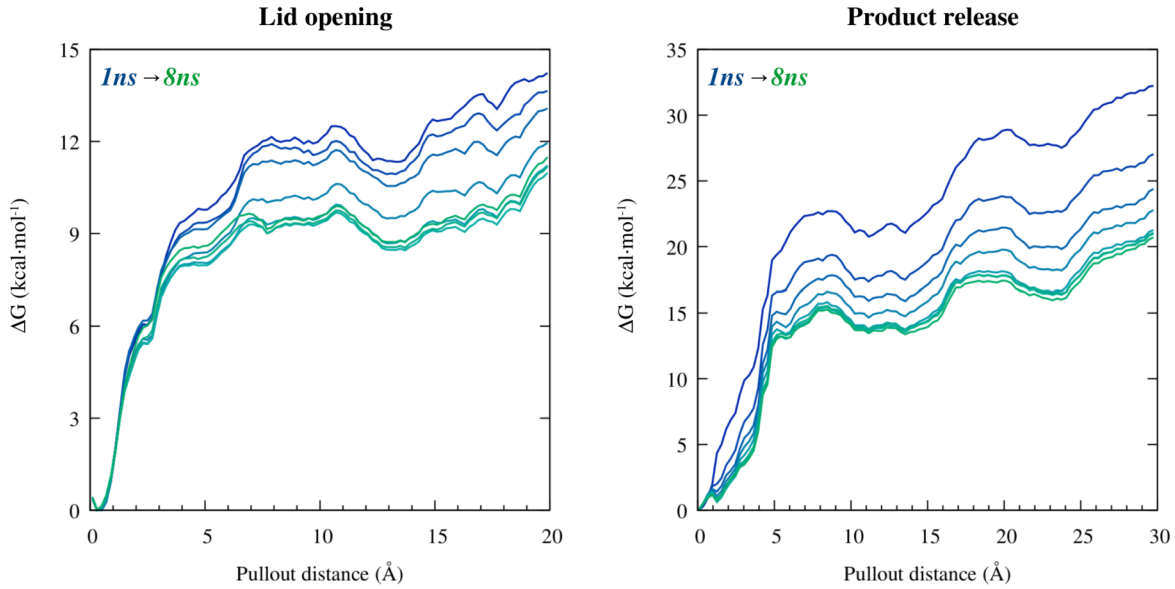


Figure S6.5- Time-convergence of the US free energy profiles of the lid opening (left) and the subsequent product release (right). Time evolution is represented in a gradient of colors, from blue to green, with a total of eight cumulative profiles. Notice that the profiles from the last 4 ns of simulation do not change over time.

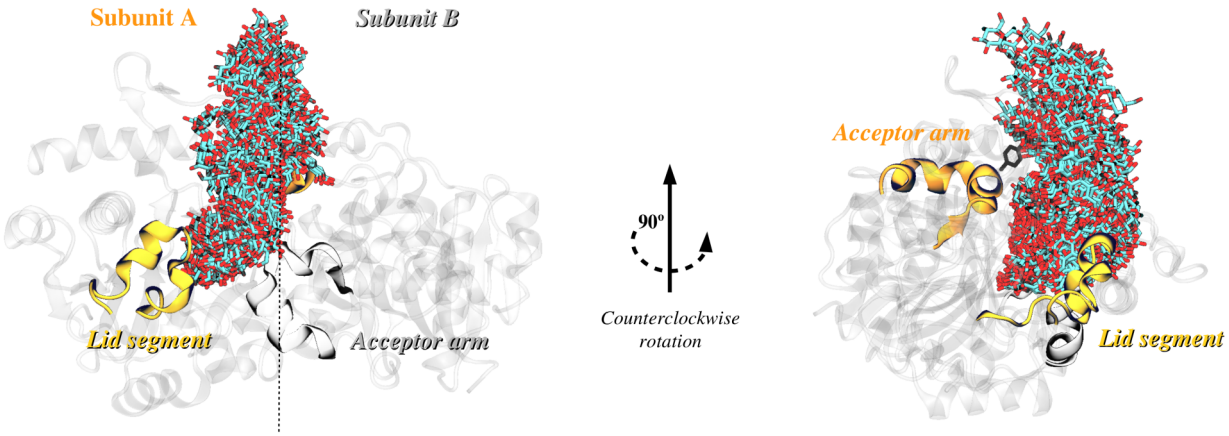


Figure S6.6- Structural superposition of the product chain conformations along the product release step that takes place after the lid opening. Only the last nanosecond of simulation of each US window is used for the superposition, taking structures at intervals of 250 ps. Notice that the substrate exits the enzymatic cavity interacting with the acceptor arm of subunit B and that at longer distances it mostly interacts with its own subunit.

Chapter 7
General Conclusions

In this thesis we have explored the fascinating world of carbohydrates and carbohydrate-active enzymes (CAZymes) using cutting-edge computational techniques, aiming to contribute to the general understanding of these complex systems. We have introduced the main topics and challenges of computational glycobiology and we have emphasized that, despite the great advances in the field, several open questions remain unsolved. Our work has focused on two types of questions that are related either with mechanistic details of *specific* CAZymes (GH11 β-xylanases, GH72 β-glucosidases, GH134 β-mannanases and GT8 α-glucosyltransferases) or with *general* concepts regarding interactions and conformations that seem to be common among these enzymes (sugar distortions, substrate hydroxyl interactions, water binding residues and enzymatic flexibility). The most important conclusions that have been obtained from the present work are the following:

- β-xylosidases can follow two different catalytic itineraries, $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ or $^{2}S_{o} \rightarrow [^{2,5}B]^{\ddagger} \rightarrow {}^{5}S_{1}$, and not three as it was experimentally proposed. We have demonstrated that the third catalytic itinerary, $^{O}E \rightarrow [^{O}S_{2}]^{\ddagger} \rightarrow B_{2,5}$, results from an over-interpretation of a GH11 X-ray structure of a mutant of the acid/base residue (Glu177Gln), which restricts the conformational preferences of the substrate. Furthermore, we have shown that the wild-type (WT) form of the enzyme is able to accommodate two different sugar conformations, $^{4}C_{1}$ and $^{2}S_{O}$, with the former being slightly more stable than the latest. This highlights that GHs do not necessarily bind the -1 sugar in a preactivated conformation, although a conformational change to preactivate the substrate must take place before the reaction.

- GHs distort their -1 sugars to enhance catalysis. We have obtained a computational proof for such rate enhancement, quantifying it in ~42 kcal·mol^{-1} in terms of reaction free energy barrier for a GH11 β-xylosidase. The kinetic advantage of distorted conformations is mainly due to reduced steric clashes with the nucleophile and the "stereoelectronic similarity" of these conformations to the ones that are favored for an oxocarbenium ion-like TS. The high energy barrier of non-distorted conformations, however, may change from enzyme to enzyme, and lower energy barriers can be expected for conformations that, although not being preactivated, are next to a TS-like conformation (*e.g.* in the *Sc*Gas2 enzyme, the $^{4}C_{1}$ conformation at the MC is still able to reach a $^{4}H_{3}$ TS).

- Interactions involving the 2-OH group are crucial for the catalytic activity of retaining GHs. We have confirmed the importance of these interactions in *Sc*Gas2 GH72 β-glucosidase, finding that the lack of the 2-OH···Nucleophile interaction raises reaction free energy barriers up to ~16 kcal·mol^{-1} and changes the catalytic itinerary of the substrate. The strength of this interaction decreases in the order TS (~1.50 Å) > MC (~1.55 Å) > GEI (~1.60 Å), explaining why it affects deglycosylation (GEI → TS) more than glycosylation (MC → TS), for which TS destabilization is partially compensated by MC destabilization. Furthermore, we have shown that the suppression of the Asn175···2-OH interaction by the Asn175Ala mutation increases the glycosylation barrier by 6.5 kcal·mol^{-1}, but leads deglycosylation practically unaffected. This suggests that the mutation of conserved residues interacting with the 2-OH could be used, together with activated substrates, for the rational conversion of GHs into TGs (*i.e.* from hydrolytic to synthetic enzymes).

- *Ss*GH134 β-mannanase follows a $^{1}C_{4} \rightarrow [^{3}H_{4}]^{\ddagger} \rightarrow {}^{3}S_{1}$ "southern hemisphere" catalytic itinerary, followed by spontaneous relaxation of the $^{3}S_{1}$ product to $^{1}C_{4}$. Our results reinforce the experimental proposal of the itinerary based on X-ray crystallography by connecting *end-to-end* the structures of MC and products. We also provide an atomistic explanation for the origin of the unexpected $^{1}C_{4}$ conformation of the substrate at the reaction products, which can be associated with the highly dynamic active center of the enzyme, where water molecules are able to enter *in and out* in the nanosecond time scale.

- Residues that confine water in the active site of inverting GHs are essential for the orientation and activation of the catalytic water. We have unveiled that in *Ss*GH134 β-mannanase the water molecule is translationally and orientationally restricted by strong hydrogen bonds with a binding triad (Asp57, Lys59 and Asn65), whose interactions with water account for 1-2 kcal·mol^{-1} in terms of binding free energy. The Lys59 residue stabilizes the MC and changes its conformation after the TS, leading space for the hydrolyzed product. Mutation of this residue (Lys59Ala) perturbs both the enzymatic structure and the properties of water inside the enzymatic cavity, which is expected to dramatically affect catalysis. On the contrary, the Asn65Ala mutant maintains the native structure of the enzyme, but alters the conformational preferences of Asp57, approaching it to the anomeric carbon. We suggest that this mutant could result in a mechanistic change, from inversion to retention of configuration.

- Human glycogenin (GYG) catalyzes the formation of α-1,4 bonds through a retaining S_Ni-*like* mechanism with a short-lived oxocarbenium intermediate. We have proven the feasibility of this mechanism using QM/MM metadynamics, finding a very low reaction free energy barrier (~10 kcal·mol^{-1}) that can be attributed to: (i) the ability of the acceptor to establish a direct hydrogen bond with the O_P of the scissile bond; and (ii) the highly charged environment at the donor region, containing both a Mn^{2+} ion and two charged Arg77 and Lys218 residues. In view of this result, and on the basis of enhanced sampling simulations, we suggest that intra- and intermonomeric conformational transitions of the substrate could be the rate-determining step of the whole catalytic process. These transitions require the opening of a flexible lid segment of the enzyme prior to the unbinding of the product chain, a step that takes place easily given that the interactions of the lid with its environment are mostly hydrophobic.

- The first two glucosylation steps of GYG (n = 0 and n = 1) occur through intermonomeric conformations of the substrate chains, while subsequent glucosylation steps can be either intra- or intermonomeric. We have shown that the presence of a "blocking loop" in between the acceptor arm and the active site of the same subunit hampers the formation of short-chain intramonomeric conformations. These early glucosylation steps, however, require a

substantial distortion of the acceptor arm, which could be associated with a high energy demand. In contrast, longer substrate chains (n = 2 and n = 3) are able to adapt to the enzymatic cavity circumventing the blocking loop, making intramonomeric conformations be more prominent compared to the intermonomeric ones. This suggests that intra- to intermonomeric transitions could take place at different glucosylation steps.

Whereas the contributions of this work have allowed to pave the understanding of certain CAZymes and have provided further evidence for concepts that are often presumed, many interesting questions remain open in the exciting field of glycobiology. In the coming years, it is expected that the increasing power of computer resources and the development of new techniques will allow to explore –with more precision and detail– the incredible amount of CAZymes that nature has created, deepening in their comprehension. Nonetheless, after each step in the lands of knowledge, new horizons of curiosities will emerge, and research will keep walking towards worlds that we still do not imagine, as it is perfectly reflected in Hamlet's famous quote: *there are more things in heaven and earth, Horatio, than are dreamt of in our philosophy*.

Publications in Journals

The work presented in this thesis has given rise to the following publications:

- "The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases" *Chemical Science*, **6**, 1167-1177 (2015).
- "A trapped covalent intermediate of a glycoside hydrolase on the pathway to transglycosylation. Insights from experiments and QM/MM simulations" *Journal of the American Chemical Society*, **138**, 3325-3332 (2016) .
- "A β-Mannanase with a Lysozyme-like Fold and a Novel Molecular Catalytic Mechanism" *ACS Central Science*, **2**, 896-903, (2016).
- "Palladium-mediated enzyme activation suggests multiphase initiation of glycogenesis" *Under revision.*

Besides, related works have resulted in other publications that are not included in the thesis:

- "Enzymatic Cleavage of Glycosidic Bonds: Strategies on How to Set Up and Control a QM/MM Metadynamics Simulation" *Methods in Enzymology*, **577**, 159-183 (2016).
- "Contribution of Shape and Charge to the Inhibition of a Family GH99 endo-α-1,2-mannanase" *Journal of the American Chemical Society*, **139**, 1089-1097 (2017).
- "Carba-cyclophellitols Are Neutral Retaining-Glucosidase Inhibitors" *Journal of the American Chemical Society*, **139**, 6534-6537 (2017).
- "Conformational Analysis of the Mannosidase Inhibitor Kifunensine: A Quantum Mechanical and Structural Approach" *ChemBioChem*, **18**, 1496-1501 (2017).
- "1,6-Cyclophellitol Cyclosulfates: A New Class of Irreversible Glycosidase Inhibitor" *ACS Central Science*, **3**, 784-793, (2017).
- "The Molecular Mechanism of Substrate Recognition and Catalysis of the Membrane Acyltransferase PatA from Mycobacteria" *ACS Chemical Biology*, **13**, 131−140, (2018).

Oral and Poster Communications

The work enclosed in this thesis has been presented as oral and poster communications in workshops and congresses:

- **February 2015-** Poster presentation "Transglycosylation: How Glycosyl Hydrolases can Overcome Hydrolysis?" at the CECAM (Centre Européen de Calcul Atomique et Moléculaire) tutorial "Hybrid Quantum Mechanics / Molecular Mechanics (QM/MM) Approaches to Biochemistry and Beyond", Lausanne, Switzerland.

- **May 2015-** Oral communication "Sugar Conformations that Enhance Cleavage of Glycosidic Bonds in Carbohydrate-Active Enzymes" at the 2nd BSC (Barcelona Supercomputing Center) International Doctoral Symposium, Barcelona, Spain.

- **May 2015-** Poster presentation "The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases" at the 11th Carbohydrate Bioengineering Meeting (CBM11), Helsinki, Finland.

- **June 2015-** Oral communication "The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases" at the XXXI Reference Network of R+D+I on Theoretical and Computational Chemistry (XRQTC) Annual Meeting, Girona, Spain.

- **October 2015-** Poster presentation " The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases" and award at the 4th edition of the New Trends of Computational Chemistry for Industry Applications organized by the XXXI Reference Network of R+D+I on Theoretical and Computational Chemistry (XRQTC), Barcelona, Spain.

- **April 2016-** Oral communication "Sugar Conformations that Enhance Cleavage of Glycosidic Bonds in Carbohydrate-Active Enzymes" at Institut de Physique et Chimie des Matériaux de Strasbourg.

- **July 2016-** Oral communication "Joining carbohydrates with modified enzymes. Insights from QM/MM simulations" at the XXXII Reference Network of R+D+I on Theoretical and Computational Chemistry (XRQTC) Annual Meeting, Bellaterra, Spain.

- **April 2017-** Flash presentation "A Trapped Covalent Intermediate of a Glycoside Hydrolase on the Pathway to Transglycosylation" and poster comunication "A β-Mannanase with a Lysozyme-like Fold and a Novel Molecular Catalytic Mechanism" at the 12th Carbohydrate Bioengineering Meeting (CBM12), Vienna, Austria.

- **July 2017-** Flash presentation "A β-Mannanase with a Lysozyme-like Fold and a Novel Molecular Catalytic Mechanism" and poster comunication "A Trapped Covalent Intermediate of a Glycoside Hydrolase on the Pathway to Transglycosylation" at the 19th European Carbohydrate Symposium (EuroCarb), Barcelona, Spain.

Acknowledgments

Per a dur a terme un projecte, un necessita una idea, però sobretot anar escàs de temps. Això darrer és potser el que més m'ha caracteritzat en els últims quatre anys, i bé ho saben totes les persones que han hagut de suportar la meva prolongada absència. Aquest aquelarre idiomàtic va per vosaltres, així com també pels que han fet més plaent la meva afanyosa immersió en el món paral·lel de la "veritat única":

A la meva mentora, la *Carme*, per haver-me introduït en la inhòspita selva de la recerca. He après tantes coses al llarg d'aquest període formatiu que trigaré anys en pair l'experiència. Moltes gràcies per tot el que m'has ensenyat, pels consells personals i professionals, els ànims, les innocentades –me les he empassat totes!– i per tota la confiança que has dipositat en mi *ab initio*.

Als companys de ciència: al *Javi*, per ser incombustible tot i que et cosim nit i dia amb les bales biodegradables de la glicogenina. Al *Víctor*, perquè jo sí que sé qui ha après més de qui, no en tinc cap dubte. A la *Meme*, per ser literalment increïble, i no pas per la teva habilitat d'esquivar fotons i evitar ser immortalitzada en plata (visca l'era digital), sinó per la teva insuperable meticulositat i perfeccionisme. *Alba*, sé que he estat un mal pare, però deixa'm dir-te que la meva poca virtut educativa t'ha fet una *self-made woman*, si treballes de valent estic convençut que seràs la pròxima Skłodowska Curie. *Joan*, has demostrat ser un bacteri termòfil, anaeròbic, endòspor i quimiolitoartòtrof –perdona, volia dir quimioteòricotòtrof, amb dos accents, al tanto!–, t'has guanyat tot el meu respecte científic (que no és massa, però tampoc és nul). *Binju*, 謝謝 for all your kindness and patience with our continuous *blablabla*. Als nadons, molt breument: *Beatriz*, "te asciendo a hijo", *Almacellas*, a partir d'ara seràs "la niña lista de la casa". Moltes gràcies a tots per suportar les meves queixes constants i les meves dissertacions para-filosòfiques sobre la teoria del dolor i la màquina del moviment perpetu.

Als del burg dels carrers bàvars (Strasbourg): à *Professeur Boero*, non seulement pour tout ce que j'ai appris de lui, mais aussi pour être une personne d'une qualité humaine incomparable. Merci beaucoup pour toute l'expérience avec vous. I also want to thank *Professor Guido*, *Professor Burak* and *Professor "rain"* for their energy, space and time devoted to me. También aprovecho para

recordar a toda la gente con la que pasé entrañables momentos, dígase *Edmond, Gabriella, Andrea, Farouk, Agostino*… pero sobretodo a *Ainara*, *Donata* y a mi querido *Pablo,* por la maravilla de una extraña amistad que surgió tal y como deben surgir las amistades: sin que nadie las entienda.

Als malalts d'òpera: *Pau,* anava a declarar-te el meu amor etern, però vist que no vas fer menció alguna sobre la meva honorable persona en la teva ignominiosa dissertació retòrica, només et diré una cosa: *non più andrai, farfallone amoroso, notte e giorno d'intorno girando…* mala peça!!! A l'*Albert*, per ser un referent intel·lectual i moral, espero seguir sent el teu *Traianus Hadrianus Augustus* tot i estar sempre al front comandant les legions, lluny dels plaers libidinosos de la capital. Moltes gràcies als dos per guardar-me un lloc en el vostre cap i cor, aquests quatre anys de xocolata, òpera i tapes Basques –¡todo en uno!– no s'esborraran fàcilment de la pell de la memòria.

A los poetas y a la poesía: a *Don Ramón*, otrora Archiduque de Despeñaperros, por ser el abanderado de la cultura decadente y el adalid del surrealismo ravalesco. Confío en que tras mi prolongada ausencia podamos volver a pintar cuadros poéticos inspirados en la venerable cópula de los quelonios. A mi Venezolano, *Luis*, Virrey de las Bahamas, por ser hombre y poema encarnados en un mismo saco de sombras y huesos. Todavía nos quedan muchos desastres por escribir bajo el cobijo de una cochambrosa y lúgubre taberna. También quisiera recordar a todos los cronopios y las cronopias, a *Joan*, *Pablowsky*, *Constanza, Tófoles*, *Anike, Saioa*, *Lirio*, *Llorenç*, *Dimarco...* y muchos más que me dejo en el tintero. ¡Todos sois cultura en carne viva!

A l'eternitat dels amics d'infància: senyors *Ruperti*, *Salla* i *Argüelles*, per ser epítoms de voluntat libèrrima en la talaia de la terra irreverent. Espero que em disculpeu per haver defugit els meus deures amicals. Tenim tantes anècdotes plegats que costa quedar-se amb alguna, estic segur que tornarem a reunir-nos per continuar tacant el ja ennegrit paper de la nostra infame història.

A les meves "*B*s", que no són Brahms, Bach, Beethoven ni Berlioz, sinó *Beatriz* y *Blanca*. A una per plantar la llavor de la curiositat acadèmica en aquest cap del desastre, y a la otra por ser mi *Liebestod,* mi *Léucade* y también mi primavera de turquesas pensamientos acuchillados por las notas de Sibelius. Te debo tanta felicidad que me será imposible pagártela con poemas, probaré con la alternativa más sensata: toneladas de tiramisú y litros y litros de orchata rancia (un remedio expeditivo tal seguro que funciona).

I per últim, a aquells que han configurat el que s'amaga sota l'iceberg de la meva consciència: a la meva família. *Papa*, mai deixes de sorprendre'm amb les teves anècdotes i el teu vast coneixement, t'estic eternament agraït per tot el suport i l'estimació que em professes. *Vasca*, eres una madraza que no merezco, si no fuera por tí no sería yo nada, gracias por tu santa paciencia y por estar al pié del cañón aguantando el duro envite de nuestros días. *Damià*, *Blanca* i *Mia*, tots tres el futur de la casa Raich, per la nova distinció que m'heu confiat (¡el padrino!), espero estar a l'alçada recordant, com a mínim, el dia de la mona i el del seu aniversari; que Déu m'agafi confessat... i acabo amb tu, *Leti*, raó de què aquest misantrop sigui alhora filantrop, per tot el que m'has ensenyat al llarg d'aquest ombrívol teatre que és la vida, i també per l'admirable coratge amb què afrontes l'indefugible destí que ens empresona. Ets tota la filosofia de la Humanitat col·lapsada en un sol ésser, la profunditat d'allò que ningú compren. Mai em perdonaré haver passat tan poc temps al teu cantó. La teva existència és tot el meu univers.

Bibliography

1. Hu, J., Seeberger, P. H. & Yin, J. Using carbohydrate-based biomaterials as scaffolds to control human stem cell fate. *Org. Biomol. Chem.* **14,** 8648–8658 (2016).

2. Reuel, N. F., Mu, B., Zhang, J., Hinckley, A. & Strano, M. S. Nanoengineered glycan sensors enabling native glycoprofiling for medicinal applications: towards profiling glycoproteins without labeling or liberation steps. *Chemical Society Reviews* **41,** (2012).

3. Krajewska, B. Membrane-based processes performed with use of chitin/chitosan materials. *Sep. Purif. Technol.* **41,** 305–312 (2005).

4. Seeberger, P. H. & Werz, D. B. Synthesis and medical applications of oligosaccharides. *Nature* **446,** 1046–1051 (2007).

5. Hart, G. W. & Copeland, R. J. Glycomics hits the big time. *Cell* **143,** 672–676 (2010).

6. Purcell, B. P. *et al.* Incorporation of sulfated hyaluronic acid macromers into degradable hydrogel scaffolds for sustained molecule delivery. *Biomater. Sci.* **2,** 693–702 (2014).

7. Werz, D. B. *et al.* Exploring the structural diversity of mammalian carbohydrates ('glycospace') by statistical databank analysis. *ACS Chem. Biol.* **2,** 685–691 (2007).

8. Davies, G. J., Planas, A. & Rovira, C. Conformational analyses of the reaction coordinate of glycosidases. *Acc. Chem. Res.* **45,** 308–316 (2012).

9. IUPAC. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Conformational nomenclature for five and six-membered ring forms of monosaccharides and their derivatives: recommendations 1980. *Eur J Biochem* **111,** 295–298 (1980).

10. Wolfenden, R. & Snider, M. J. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* **34,** 938–945 (2001).

11. Davies, G. & Henrissat, B. Structures and mechanisms of glycosyl hydrolases. *Structure* **3,** 853–859 (1995).

12. Henrissat, B. & Bairoch, A. Updating the sequence-based classification of glycosyl hydrolases. *Biochem. J.* **316,** 695–696 (1996).

13. http://www.cazy.org/. consulted on 20 of April 2018.

14. Cantarel, B. I. *et al.* The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Res.* **37,** 233–238 (2009).

15. Rose, A. W. and D. R. Mechanism of catalysis by retaining β-glycosyl hydrolases.

16. Pierdominici-Sottile, G., Horenstein, N. A. & Roitberg, A. E. Free energy study of the catalytic mechanism of Trypanosoma cruzi trans -sialidase. from the Michaelis complex to the covalent intermediate. *Biochemistry* **50,** 10150–10158 (2011).

17. Amaya, M. F. *et al.* Structural insights into the catalytic mechanism of Trypanosoma cruzi trans-sialidase. *Structure* **12,** 775–784 (2004).

18. van Aalten, D. M. *et al.* Structural insights into the catalytic mechanism of a family 18 exo-chitinase. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8979–84 (2001).

19. Carl S Rye and Stephen G Withers. Glycosidase mechanisms. *Curr. Opin. Chem. Biol.* **4,** 573–580 (2000).

20. Ardèvol, A. & Rovira, C. Reaction Mechanisms in Carbohydrate-Active Enzymes: Glycoside Hydrolases and Glycosyltransferases. Insights from ab Initio Quantum Mechanics/Molecular Mechanics Dynamic Simulations. *J. Am. Chem. Soc.* **137,** 7528–7547 (2015).

21. Zechel, D. L. & Withers, S. G. Glycosidase mechanisms: anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33,** 11–18 (2000).

22. Goldberg, Y. B. T. and R. N. Thermodynamics of hydrolysis of Disaccharides. *J. Biol. Chem.* **264,** 3966–3971 (1988).

23. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* **77,** 521–555 (2008).

24. Withers, S. G. Mechanisms of glycosyl transferases and hydrolases. *Carbohydr. Polym.* **44,** 325–337 (2001).

25. Ardèvol, A. & Rovira, C. The molecular mechanism of enzymatic glycosyl transfer with retention of configuration: evidence for a short-lived oxocarbenium-like species. *Angew. Chem. Int. Ed. Engl.* **50,** 10897–10901 (2011).

26. Ardèvol, A., Iglesias-Fernandez, J., Rojas-Cervellera, V. & Rovira, C. The reaction mechanism of retaining glycosyltransferases. *Biochem. Soc. Trans.* **44,** 51–60 (2016).

27. Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Ber. Dtsch. Chem. Ges.* **27,** 2984–2993 (1894).

28. Koshland, D. E. The Key–Lock Theory and the Induced Fit Theory. *Angew. Chemie Int. Ed. English* **33,** 2375–2378 (1995).

29. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5,** 789–796 (2009).

30. Koshland, D. E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci. USA* **44,** 98–104 (1958).

31. Pabis, A., Risso, V. A., Sanchez-Ruiz, J. M. & Kamerlin, S. C. Cooperativity and flexibility in enzyme evolution. *Curr. Opin. Struct. Biol.* **48,** 83–92 (2018).

32. Teague, S. J. Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2,** 527–541 (2003).

33. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: An extended view of binding events. *Trends Biochem. Sci.* **35,** 539–546 (2010).

34. Mathews, F. S. *et al.* Protein Dynamism and Evolvability. *Science (80-.).* **324,** 203–208 (2009).

35. Karplus, M. The Levinthal paradox: yesterday and today. *Fold. Des.* **2,** S69–S75 (1997).

36. Narayanan, C., Bernard, D. & Doucet, N. Role of Conformational Motions in Enzyme Function: Selected Methodologies and Case Studies. *Catalysts* **6,** 81 (2016).

37. Petricevic, M. *et al.* Contribution of Shape and Charge to the Inhibition of a Family GH99 endo-alpha-1,2-Mannanase. *J. Am. Chem. Soc.* **139,** 1089–1097 (2017).

38. Vocadlo, D. J. & Davies, G. J. Mechanistic insights into glycosidase chemistry. *Curr. Opin. Chem. Biol.* **12,** 539–555 (2008).

39. Speciale, G., Thompson, A. J., Davies, G. J. & Williams, S. J. Dissecting conformational contributions to glycosidase catalysis and inhibition. *Curr. Opin. Struct. Biol.* **28C,** 1–13 (2014).

40. Cremer, D. & Pople, J. A. General definition of ring puckering coordinates . *J. Am. Chem. Soc.* **97,** 1354–1358 (1975).

41. Hill, C. H., Graham, S. C., Read, R. J. & Deane, J. E. Structural snapshots illustrate the catalytic cycle of beta-galactocerebrosidase, the defective enzyme in Krabbe disease. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20479–20484 (2013).

42. Guérin, D. M. A. *et al.* Atomic (0.94 Å) resolution structure of an inverting glycosidase in complex with substrate. *J. Mol. Biol.* **316,** 1061–1069 (2002).

43. Sulzenbacher, G., Driguez, H., Henrissat, B., Schulein, M. & Davies, G. J. Structure of the Fusarium oxysporum endoglucanase I with a nonhydrolyzable substrate analogue: substrate distortion gives rise to the preferred axial orientation for the leaving group. *Biochemistry* **35,** 15280–15287 (1996).

44. Thiel, W. & Hummer, G. Nobel 2013 Chemistry: Methods for computational chemistry. *Nature* **504,** 96–97 (2013).

45. 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.* **103,** 227–249 (1976).

46. Plata, R. E. & Singleton, D. A. A case study of the mechanism of alcohol-mediated morita baylis-hillman reactions. the importance of experimental observations. *J. Am. Chem. Soc.* **137,** 3811–3826 (2015).

47. Winter, A. Computational chemistry: Making a bad calculation. *Nat. Chem.* **7,** 473–475 (2015).

48. Alonso-Gil, S. *et al.* Computational Design of Experiment Unveils the Conformational Reaction Coordinate of GH125 alpha-Mannosidases. *J. Am. Chem. Soc.* **139,** 1085–1088 (2017).

49. Biarnés, X., Nieto, J., Planas, A. & Rovira, C. Substrate distortion in the Michaelis complex of Bacillus 1,3-1,4-beta-glucanase. Insight from first principles molecular dynamics simulations. *J. Biol. Chem.* **281,** 1432–1441 (2006).

50. Biarnés, X., Ardèvol, A., Iglesias-Fernández, J., Planas, A. & Rovira, C. Catalytic itinerary in 1,3-1,4-beta-glucanase unraveled by QM/MM metadynamics. Charge is not yet fully developed at the oxocarbenium ion-like transition state. *J. Am. Chem. Soc.* **133,** 20301–20309 (2011).

51. Petersen, L., Ardèvol, A., Rovira, C. & Reilly, P. J. Mechanism of cellulose hydrolysis by inverting GH8 endoglucanases: a QM/MM metadynamics study. *J. Phys. Chem. B* **113,** 7331–7339 (2009).

52. Biarnés, X. *et al.* The conformational free energy landscape of beta-D-glucopyranose. Implications for substrate preactivation in beta-glucoside hydrolases. *J. Am. Chem. Soc.* **129,** 10686–10693 (2007).

53. Ardèvol, A., Biarnés, X., Planas, A. & Rovira, C. The Conformational Free-Energy Landscape of β-D-Mannopyranose: Evidence for a 1S5 → B2,5 → OS2 Catalytic Itinerary in β-Mannosidases. *J Am Chem Soc* **132,** 16058–16065 (2010).

54. Petricevic, M. *et al.* Contribution of shape and charge to the inhibition of a family GH99 endo-α-1,2-mannanase. *J. Am. Chem. Soc.* **139,** 1089–1097 (2017).

55. Beenakker, T. J. M. *et al.* Carba-cyclophellitols Are Neutral Retaining-Glucosidase Inhibitors. *J. Am. Chem. Soc.* **139,** 6534–6537 (2017).

56. Privett, H. K. *et al.* Iterative approach to computational enzyme design. *Proc. Natl. Acad. Sci.* **109,** 3790–3795 (2012).

57. Frushicheva, M. P., Cao, J., Chu, Z. T. & Warshel, A. Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. *Proc. Natl. Acad. Sci.* **107,** 16869–16874 (2010).

58. Jiang, L. *et al.* De novo computational design of retro- aldol enzymes. *Science (80-.).* **319,** 1387–1391 (2008).

59. Olsson, M. H., Parson, W. W. & Warshel, A. Dynamical contributions to enzyme catalysis: critical tests of a popular hypothesis. *Chem Rev* **106,** 1737–1756 (2006).

60. Moliner, V. 'Eppur si muove' (Yet it moves). *Proc. Natl. Acad. Sci. USA* **108,** 15013–15014 (2011).

61. Schowen, R. L. How an enzyme surmounts the activation energy barrier. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 11931–11932 (2003).

62. Hammes-Schiffer, S. Impact of enzyme motion on activity. *Biochemistry* **41,** 13335–13343 (2002).

63. Mireia Garcia-Viloca, Jiali Gao, Martin Karplus, D. G. T. How Enzymes Work: Analysis by Modern Rate Theory and Computer Simulations. **303,** 186–196 (2004).

64. Lawson, S. L., Wakarchuk, W. W. & Withers, S. G. Effects of both shortening and lengthening the active site nucleophile of Bacillus circulans xylanase on catalytic activity. *Biochemistry* **35,** 10110–10118 (1996).

65. Lawson, S. L., Wakarchuk, W. W. & Withers, S. G. Positioning the acid/base catalyst in a glycosidase: Studies with bacillus circulans xylanase. *Biochemistry* **36,** 2257–2265 (1997).

66. Capon, B. Mechanism in carbohydrate chemistry. *Chem. Rev.* **69,** 407–498 (1969).

67. Alabugin, I. V., Manoharan, M., Peabody, S. & Weinhold, F. Electronic basis of improper hydrogen bonding: A subtle balance of hyperconjugation and rehybridization. *J. Am. Chem. Soc.* **125,** 5973–5987 (2003).

68. Devereux, M. & Popelier, P. L. A. The effects of hydrogen-bonding environment on the polarization and electronic properties of water molecules. *J. Phys. Chem. A* **111,** 1536–1544 (2007).

69. Namchuk, M. N. & Withers, S. G. Mechanism of Agrobacterium β-Glucosidase: Kinetic Analysis of the Role of Noncovalent Enzyme/Substrate Interactions. *Biochemistry* **34,** 16194–16202 (1995).

70. McIntosh, L. P. *et al.* The pKa of the general acid/base carboxyl group of a glycosidase cycles during catalysis: A 13C-NMR study of Bacillus circulans xylanase. *Biochemistry* **35,** 9958–9966 (1996).

71. Nielsen, J. E. *et al.* Electrostatics in the active site of an alpha-amylase. *Eur J Biochem* **264,** 816–824 (1999).

72. Joshi, M. D. *et al.* Dissecting the electrostatic interactions and pH-dependent activity of a family 11 glycosidase. *Biochemistry* **40,** 10115–10139 (2001).

73. Bishop, J. R., Schuksz, M. & Esko, J. D. Heparan sulphate proteoglycans fine-tune mammalian physiology. *Nature* **446,** 1030–1037 (2007).

74. Davies, G. J. *et al.* Snapshots along an enzymatic reaction coordinate: Analysis of a retaining β-glycoside hydrolase. *Biochemistry* **37,** 11707–11713 (1998).

75. Warshel, A. Energetics of enzyme catalysis. *Proc. Natl. Acad. Sci.* **75,** 5250–5254 (1978).

76. Lipari, F. & Herscovics, A. Calcium binding to the class I α-1,2-mannosidase from Saccharomyces cerevisiae occurs outside the EF hand motif. *Biochemistry* **38,** 1111–1118 (1999).

77. Petersen, L., Ardévol, A., Rovira, C. & Reilly, P. J. Molecular mechanism of the glycosylation step catalyzed by Golgi α-mannosidase II: A QM/MM metadynamics investigation. *J. Am. Chem. Soc.* **132,** 8291–8300 (2010).

78. Zolotnitsky, G. *et al.* Mapping glycoside hydrolase substrate subsites by isothermal titration calorimetry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 11275–11280 (2004).

79. Hamre, A. G., Jana, S., Reppert, N. K., Payne, C. M. & Sørlie, M. Processivity, Substrate Positioning, and Binding: The Role of Polar Residues in a Family 18 Glycoside Hydrolase. *Biochemistry* **54,** 7292–7306 (2015).

80. Zhang, X. *et al.* Subsite-specific contributions of different aromatic residues in the active site architecture of glycoside hydrolase family 12. *Sci. Rep.* **5,** 1–12 (2015).

81. Bissaro, B., Monsan, P., Faure, R. & O'Donohue, M. J. Glycosynthesis in a waterworld: new insight into the molecular basis of transglycosylation in retaining glycoside hydrolases. *Biochem. J.* **467,** 17–35 (2015).

82. Boraston, A. B., Bolam, D. N., Gilbert, H. J. & Davies, G. J. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochem. J.* **382,** 769–781 (2004).

83. Yang, H. *et al.* Loop 3 of fungal endoglucanases of glycoside hydrolase family 12 modulates catalytic efficiency. *Appl. Environ. Microbiol.* **83,** 1–11 (2017).

84. Liang, P. H. *et al.* A flexible loop for mannan recognition and activity enhancement in a bifunctional glycoside hydrolase family 5. *Biochim. Biophys. Acta - Gen. Subj.* **1862,** 513–521 (2018).

85. Beenakker, T. J. M. *et al.* Carba-cyclophellitols Are Neutral Retaining-Glucosidase Inhibitors. *J. Am. Chem. Soc.* **139,** (2017).

86. Artola, M. *et al.* 1,6-Cyclophellitol Cyclosulfates: A New Class of Irreversible Glycosidase Inhibitor. *ACS Cent. Sci.* **3,** 784–793 (2017).

87. Sinnott, M. L. Catalytic mechanism of enzymic glycosyl transfer. *Chem. Rev.* **90,** 1171–1202 (1990).

88. Koivula, A. *et al.* The active site of cellobiohydrolase Cel6A from Trichoderma reesei: The roles of aspartic acids D221 and D175. *J. Am. Chem. Soc.* **124,** 10015–10024 (2002).

89. Fernandes, P. Z., Petricevic, M., Sobala, L., Davies, G. J. & Williams, S. J. Exploration of strategies for mechanism-based inhibitor design for family GH99 endo-α-1,2-mannanases. *Chem. - A Eur. J.* just accepted (2018). doi:10.1002/chem.201800435

90. Yip, V. L. *et al.* An unusual mechanism of glycoside hydrolysis involving redox and elimination steps by a family 4 beta-glycosidase from Thermotoga maritima. *J. Am. Chem. Soc.* **126,** 8354–8355 (2004).

91. Speciale, G., Thompson, A. J., Davies, G. J. & Williams, S. J. Dissecting conformational contributions to glycosidase catalysis and inhibition. *Curr. Opin. Struct. Biol.* **28,** 1–13 (2014).

92. Reeves, R. E. The Shape of Pyranoside Rings. *J. Am. Chem. Soc.* **72,** 1499–1506 (1950).

93. Rojas-Cervellera, V., Ardèvol, A., Boero, M., Planas, A. & Rovira, C. Formation of a covalent glycosyl-enzyme species in a retaining glycosyltransferase. *Chemistry (Easton).* **19,** 14018–14023 (2013).

94. Albesa-Jové, D., Sainz-Polo, M. Á., Marina, A. & Guerin, M. E. Structural Snapshots of α-1,3-Galactosyltransferase with Native Substrates: Insight into the Catalytic Mechanism of Retaining Glycosyltransferases. *Angew. Chemie - Int. Ed.* **56,** 14853–14857 (2017).

95. Gomez, H., Polyak, I., Thiel, W., Lluch, J. M. & Masgrau, L. Retaining glycosyltransferase mechanism studied by QM/MM methods: lipopolysaccharyl-alpha-1,4-galactosyltransferase C transfers alpha-galactose via an oxocarbenium ion-like transition state. *J. Am. Chem. Soc.* **134,** 4743–4752 (2012).

96. Teze, D. *et al.* Semi-rational approach for converting a GH1 β-glycosidase into a β-transglycosidase. *Protein Eng. Des. Sel.* **27,** 13–19 (2014).

97. Shi, Y. A glimpse of structural biology through X-ray crystallography. *Cell* **159,** 995–1014 (2014).

98. Ignacio Nebot-Gil, Fernando Martín, Rosa Caballol, Miquel Solà, J. J. N. *Theoretical and Computational Chemistry: Foundations, Methods and Techniques*. (Universitat Jaume I, 2007).

99. Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*. (Wiley, 2004).

100. Smit, D. F. and B. *Understanding Molecular Simulation: From Algorithms to Applications*. (Elsevier, 2002).

101. Cooke, B. & Schmidler, S. C. Preserving the Boltzmann ensemble in replica-exchange molecular dynamics. *J. Chem. Phys.* **129,** (2008).

102. Gordiz, K., Singh, D. J. & Henry, A. Ensemble averaging vs. time averaging in molecular dynamics simulations of thermal conductivity. *J. Appl. Phys.* **117,** (2015).

103. Verlet, L. Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159,** 98–103 (1967).

104. Dominik Marx and Jürg Hutter. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*. (Cambride University Press, 2009).

105. P. A. M. Dirac. Note on exchange phenomena in the Thomas atom. *Proc. Camb. Philol. Soc.* **26,** 376–385 (1930).

106. P. Ehrenfest. Bemerkung über die angenäherte Glütigkeit der klassischen Mechanik innerhalb der Quantenmechanik. *Zeitschrift für Phys.* **45,** 455–457 (1927).

107. Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55,** 2471–2474 (1985).

108. Cornell, W. D. *et al.* A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).

109. González, M. A. Force fields and molecular dynamics simulations. *Collect. SFN* **12,** 169–200 (2011).

110. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65,** 712–725 (2006).

111. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614 (2009).

112. W. L. Jorgensen, D. S. Maxwell, and J. T.-R. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118,** 11225–11236 (1996).

113. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935 (1983).

114. Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9,** 1005–1011 (2017).

115. Shaw, C. F. Gold-based therapeutic agents. *Chem. Rev.* **99,** 2589–2600 (1999).

116. Noe, F. Beating the Millisecond Barrier in Molecular Dynamics Simulations. *Biophys. J.* **108,** 228–229 (2015).

117. Martín-García, F., Papaleo, E., Gomez-Puertas, P., Boomsma, W. & Lindorff-Larsen, K. Comparing molecular dynamics force fields in the essential subspace. *PLoS One* **10,** 1–16 (2015).

118. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev. B* **136,** 1912–1919 (1964).

119. Kohn, W. & Sham, L. J. Self consistent equations including exchange and correlation effects. *Phys. Rev.* **140,** 1133–1138 (1965).

120. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for density functional theory. *Chem. Rev.* **112,** 289–320 (2012).

121. Ceperley, D. M. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **45,** 566–569 (1980).

122. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77,** 3865–3868 (1996).

123. Becke, A. D. Density functional calculations of molecular-bond energies. *J. Chem. Phys.* **84,** 4524–4529 (1986).

124. 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* **109,** 6676–6687 (2005).

125. Biarnés, X. *et al.* The conformational free energy landscape of β-D-glucopyranose. Implications for substrate preactivation in β-glucoside hydrolases. *J. Am. Chem. Soc.* **129,** 10686–10693 (2007).

126. Lira-Navarrete, E. *et al.* Substrate-Guided Front-Face Reaction Revealed by Combined Structural Snapshots and Metadynamics for the Polypeptide N-Acetylgalactosaminyltransferase 2. *Angew. Chem. Int. Ed. Engl.* **53,** 8206–8210 (2014).

127. Ardèvol, A. & Rovira, C. The molecular mechanism of enzymatic glycosyl transfer with retention of configuration: Evidence for a short-lived oxocarbenium-like species. *Angew. Chemie - Int. Ed.* **50,** 10897–10901 (2011).

128. Tersa, M. *et al.* The Molecular Mechanism of Substrate Recognition and Catalysis of the Membrane Acyltransferase PatA from Mycobacteria. *ACS Chem. Biol.* **13,** (2018).

129. W. J. Hehre, R. F. Stewart, and J. A. P. Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *J. Chem. Phys.* **51,** 2657 (1969).

130. Nobel Prizes 2013 M. Karplus, M. Levitt, A. Warshel. *Angew. Chem. Int. Ed.* **52,** 11972 (2013).

131. U. C. Singh and P. A. Kollman. 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.* **7,** 718–730 (1986).

132. Hu, L., Söderhjelm, P. & Ryde, U. On the convergence of QM/MM energies. *J. Chem. Theory Comput.* **7,** 761–777 (2011).

133. Laio, A., VandeVondele, J. & Rothlisberger, U. A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. *J. Chem. Phys.* **116,** 6941–6947 (2002).

134. Laio, A., VandeVondele, J. & Rothlisberger, U. D-RESP: Dynamically generated electrostatic potential derived charges from quantum mechanics/molecular mechanics simulations. *J. Phys. Chem. B* **106,** 7300–7307 (2002).

135. Sun, S. X. Equilibrium free energies from path sampling of nonequilibrium trajectories. *J. Chem. Phys.* **5769,** 5769–5775 (2012).

136. Ren, W. & Vanden-eijnden, E. String method for the study of rare events. *Phys. Rev. B* **66,** 052301 (2002).

137. Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314,** 141–151 (1999).

138. Dellago, C., Bolhuis, P. G., Csajka, F. & Chandler, D. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.* **108,** 1964–1977 (1998).

139. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99,** 12562–12566 (2002).

140. Torrie GM, V. J. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the subcritical Lennard-Jones fluid. *Chem Phys Lett* **28,** 578–581 (1974).

141. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. . *J. Comput. Phys.* **23,** 187–199 (1977).

142. Barducci, A., Bonomi, M. & Parrinello, M. Metadynamics. *WIREs Comput. Mol. Sci.* **1,** 826–843 (2011).

143. Ensing, B. & Klein, M. L. Perspective on the reactions between F- and CH3CH2F: the free energy landscape of the E2 and SN2 reaction channels. *Proc Natl Acad Sci U S A* **102,** 6755–6759 (2005).

144. Cremer, D. & Pople, J. A. A General Definition of Ring Puckering Coordinates. *J. Am. Chem. Soc.* **97,** 1354–1358 (1975).

145. Raiteri, P., Laio, A., Gervasio, F. L., Micheletti, C. & Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* **110,** 3533–3539 (2006).

146. Tiwary, P. & Parrinello, M. From metadynamics to dynamics. *Phys. Rev. Lett.* **111,** 1–5 (2013).

147. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, K. P. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem* **13,** 1011–1021 (1992).

148. Souaille M, R. B. Extension to the weighted histogram analysis method: Combining umbrella sampling with free energy calculations. *Comput Phys Commun* **135,** 40–57 (2001).

149. Kästner, J. Umbrella sampling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1,** 932–942 (2011).

150. Siegbahn, P. E. & Himo, F. Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *J. Biol. Inorg. Chem.* **14,** 643–651 (2009).

151. Bechor, D.; Ben-Tal, N. Implicit Solvent Model Studies of the Interactions of the Influenza Hemagglutinin Fusion Peptide with Lipid Bilayers. *Biophys. J.* **80,** 643–655 (2001).

152. Tews, I. *et al.* Bacterial chitobiase structure provides insight into catalytic mechanism and the basis of Tay-Sachs disease. *Nat. Struct. Biol.* **3,** 638–648 (1996).

153. Money, V. A. *et al.* Substrate distortion by a lichenase highlights the different conformational itineraries harnessed by related glycoside hydrolases. *Angew. Chem. Int. Ed. Engl.* **45,** 5136–5140 (2006).

154. Biarnés, X., Ardèvol, A., Planas, A. & Rovira, C. Substrate conformational changes in glycoside hydrolase catalysis. A first-principles molecular dynamics study. *Biocatal. Biotransfor.* **28,** 33–40 (2010).

155. Thompson, A. J. *et al.* The reaction coordinate of a bacterial GH47 α-mannosidase: A combined quantum mechanical and structural approach. *Angew. Chem. Int. Ed.* **51,** 10997–11001 (2012).

156. Thompson, A. J. *et al.* Evidence for a boat conformation at the transition state of GH76 alpha-1,6-Mannanases-Key enzymes in bacterial and fungal mannoprotein metabolism. *Angew. Chem. Int. Ed. Engl.* **54,** 5378–5382 (2015).

157. Petersen, L., Ardèvol, A., Rovira, C. & Reilly, P. J. Molecular mechanism of the glycosylation step catalyzed by Golgi alpha-mannosidase II: a QM/MM metadynamics investigation. *J. Am. Chem. Soc.* **132,** 8291–8300 (2010).

158. Knott, B. C. *et al.* The mechanism of cellulose hydrolysis by a two-step, retaining cellobiohydrolase elucidated by structural and transition path sampling studies. *J. Am. Chem. Soc.* **136,** 321–329 (2014).

159. Barnett, C. B., Wilkinson, K. A. & Naidoo, K. J. Molecular details from computational reaction dynamics for the cellobiohydrolase I glycosylation reaction. *J Am Chem Soc* **133,** 19474–19482 (2011).

160. Ardèvol, A. & Rovira, C. Reaction mechanisms in carbohydrate-active enzymes: glycoside hydrolases and glycosyltransferases. Insights from ab initio quantum mechanics/molecular mechanics dynamic simulations. *J. Am. Chem. Soc.* **137,** 7528–7547 (2015).

161. Kulkarni, N., Shendye, A. & Rao, M. Molecular and biotechnological aspects of xylanases. *FEMS Microbiol. Rev.* **23,** 411–456 (1999).

162. Deutschmann, R. & Dekker, R. F. From plant biomass to bio-based chemicals: latest developments in xylan research. *Biotechnol. Adv.* **30,** 1627–1640 (2012).

163. Sharma, M. & Kumar, A. Xylanases: An overview. *Br. Biotechnol. J.* **3,** 1–28 (2013).

164. Rubin, E. M. Genomics of cellulosic biofuels. *Nature* **454,** 841–845 (2008).

165. Iglesias-Fernández, J., Raich, L., Ardèvol, A. & Rovira, C. The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases. *Chem. Sci.* **6,** 1167–1177 (2015).

166. Wan, Q. *et al.* X-ray crystallographic studies of family 11 xylanase Michaelis and product complexes: implications for the catalytic mechanism. *Acta Crystallogr. D Biol. Crystallogr.* **70,** 11–23 (2014).

167. Lammerts van Bueren, A. *et al.* Analysis of the reaction coordinate of alpha-L-fucosidases: a combined structural and quantum mechanical approach. *J. Am. Chem. Soc.* **132,** 1804–1806 (2010).

168. Notenboom, V., Williams, S. J., Hoos, R., Withers, S. G. & Rose, D. R. Detailed structural analysis of glycosidase/inhibitor interactions: complexes of Cex from Cellulomonas fimi with xylobiose-derived aza-sugars. *Biochemistry* **39,** 11553–11563 (2000).

169. Suzuki, R. *et al.* Crystallographic snapshots of an entire reaction cycle for a retaining xylanase from Streptomyces olivaceoviridis E-86. *J. Biochem.* **146,** 61–70 (2009).

170. Wicki, J., Schloegl, J., Tarling, C. A. & Withers, S. G. Recruitment of both uniform and differential binding energy in enzymatic catalysis: Xylanases from families 10 and 11. *Biochemistry* **46,** 6996–7005 (2007).

171. Bussi, G., Laio, A. & Parrinello, M. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* **96,** 90601 (2006).

172. CPMD program, Copyright IBM Corp. 1990-2003, Copyright MPI für Festkörperforschung, Stuttgart 1997-2001. URL: http://www.cpmd.org.

173. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77,** 3865–3868 (1996).

174. Troullier, N. & Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **43,** 1993–2006 (1991).

175. Tiwary, P. & Parrinello, M. A time-independent free energy estimator for metadynamics. *J. Phys. Chem. B* **119,** 736–742 (2015).

176. Case, D. A. *et al. AMBER 11*. (2010). doi:citeulike-article-id:5692441

177. Cornell, W. D. *et al.* A 2nd generation force-field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117,** 5179–5197 (1995).

178. Kirschner, K. N. *et al.* GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J. Comput. Chem.* **29,** 622–655 (2008).

179. Humphrey, W., Dalke, A. & Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **14,** 33–38 (1996).

180. Hunenberger, P. H. Optimal charge-shaping functions for the particle-particle-particle-mesh (P3M) method for computing electrostatic interactions in molecular simulations. *J. Chem. Phys.* **113,** 10464–10476 (2000).

181. Ireta, J., Neugebauer, J. & Sheffler, M. On the accuracy of DFT for describing hydrogen bonds: dependence on the bond directionality. *J. Phys. Chem. A* **108,** 5692–5698 (2004).

182. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100,** 20603 (2008).

183. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: New feathers for an old bird. *Comp. Phys. Commun.* **185,** 604–613 (2014).

184. Desiraju, G. R. Chemistry beyond the molecule. *Nature* **412,** 397–400 (2001).

185. Steiner, T. The hydrogen bond in the solid state. *Angew. Chem. Int. Ed.* **41,** 49–76 (2002).

186. Matsuzawa, T. *et al.* Crystal structure and identification of a key amino acid for glucose tolerance, substrate specificity, and transglycosylation activity of metagenomic β-glucosidase Td2F2. *FEBS J.* **283,** 2340–2353 (2016).

187. Blakeley, M. P., Hasnain, S. S. & Antonyuk, S. V. Sub-atomic resolution X-ray crystallography and neutron crystallography: Promise, challenges and potential. *IUCrJ* **2,** 464–474 (2015).

188. Namchuk, M. N. & Withers, S. G. Mechanism of Agrobacterium beta-glucosidase: kinetic analysis of the role of noncovalent enzyme/substrate interactions. *Biochemistry* **34,** 16194–16202 (1995).

189. Latge, J. P. The cell wall: a carbohydrate armour for the fungal cell. *Mol. Microbiol.* **66,** 279–290 (2007).

190. Mazan, M., Ragni, E., Popolo, L. & Farkas, V. Catalytic properties of the Gas family beta-(1,3)-glucanosyltransferases active in fungal cell-wall biogenesis as determined by a novel fluorescent assay. *Biochem. J.* **438,** 275–282 (2011).

191. Hartland, R. P. *et al.* A novel beta-(1-3)-glucanosyltransferase from the cell wall of Aspergillus fumigatus. *J. Biol. Chem.* **271,** 26843–26849 (1996).

192. Sears, P. & Wong, C. H. Toward automated synthesis of oligosaccharides and glycoproteins. *Science (80-.).* **291,** 2344–2350 (2001).

193. Wang, C.-C. *et al.* Regioselective one-pot protection of carbohydrates. *Nature* **446,** 10–13 (2007).

194. Jöud, M., Möller, M. & Olsson, M. L. Identification of human glycosyltransferase genes expressed in erythroid cells predicts potential carbohydrate blood group loci. *Sci. Rep.* **8,** 6040 (2018).

195. Mackenzie, L. F., Wang, Q., Warren, R. A. J. & Withers, S. G. Glycosynthases: Mutant Glycosidases for Oligosaccharide Synthesis. *J. Am. Chem. Soc.* **120,** 5583–5584 (1998).

196. Baumann, M. J. *et al.* Structural evidence for the evolution of xyloglucanase activity from xyloglucan endo-transglycosylases: biological implications for cell wall metabolism. *Plant Cell* **19,** 1947–1963 (2007).

197. Monsan, P., Remaud-Simeon, M. & Andre, I. Transglucosidases as efficient tools for oligosaccharide and glucoconjugate synthesis. *Curr. Opin. Microbiol.* **13,** 293–300 (2010).

198. Montagna, G. *et al.* The trans-sialidase from the african trypanosome Trypanosoma brucei. *Eur. J. Biochem.* **269,** 2941–2950 (2002).

199. Bissaro, B., Monsan, P., Fauré, R. & O'Donohue, M. J. Glycosynthesis in a waterworld: new insight into the molecular basis of transglycosylation in retaining glycoside hydrolases. *Biochem. J.* **467,** 17–35 (2015).

200. Arnold, F. H. Combinatorial and computational challenges for biocatalyst design. *Nature* **409,** 253–257 (2001).

201. Feng, H. Y. *et al.* Converting a {beta}-glycosidase into a {beta}-transglycosidase by directed evolution. *J. Biol. Chem.* **280,** 37088–37097 (2005).

202. Osanjo, G. *et al.* Directed evolution of the alpha-L-fucosidase from Thermotoga maritima into an alpha-L-transfucosidase. *Biochemistry* **46,** 1022–1033 (2007).

203. Frutuoso, M. A. & Marana, S. R. A single amino acid residue determines the ratio of hydrolysis to transglycosylation catalyzed by beta-glucosidases. *Protein Pept. Lett.* **20,** 102–106 (2013).

204. Malet, C. & Planas, A. From beta-glucanase to beta-glucansynthase: glycosyl transfer to alpha-glycosyl fluorides catalyzed by a mutant endoglucanase lacking its catalytic nucleophile. *FEBS Lett* **440,** 208–212 (1998).

205. Bissaro, B. *et al.* Molecular Design of Non-Leloir Furanose-Transferring Enzymes from an α-l-Arabinofuranosidase: A Rationale for the Engineering of Evolved Transglycosylases. *ACS Catal.* **5,** 4598–4611 (2015).

206. Piens, K. *et al.* Mechanism-based labeling defines the free energy change for formation of the covalent glycosyl-enzyme intermediate in a xyloglucan endo-transglycosylase. *J. Biol. Chem.* **283,** 21864–21872 (2008).

207. Raich, L. *et al.* A Trapped Covalent Intermediate of a Glycoside Hydrolase on the Pathway to Transglycosylation. Insights from Experiments and Quantum Mechanics/Molecular Mechanics Simulations. *J. Am. Chem. Soc.* **138,** 3325–3332 (2016).

208. White, A., Tull, D., Johns, K., Withers, S. G. & Rose, D. R. Crystallographic observation of a covalent catalytic intermediate in a beta-glycosidase. *Nat. Struct. Biol.* **3,** 149–154 (1996).

209. Bueren-Calabuig, J. A., Pierdominici-Sottile, G. & Roitberg, A. E. Unraveling the differences of the hydrolytic activity of Trypanosoma cruzi trans-sialidase and Trypanosoma rangeli sialidase: a quantum mechanics-molecular mechanics modeling study. *J. Phys. Chem. B* **118,** 5807–5816 (2014).

210. Hurtado-Guerrero, R. *et al.* Molecular mechanisms of yeast cell wall glucan remodeling. *J. Biol. Chem.* **284,** 8461–8469 (2009).

211. Iannuzzi, M., Laio, A. & Parrinello, M. Efficient exploration of reactive potential energy surfaces using Car-Parrinello molecular dynamics. *Phys. Rev. Lett.* **90,** 238302 (2003).

212. Gauto, D. F., Di Lella, S., Guardia, C. M. A., Estrin, D. A. & Martí, M. A. Carbohydrate-Binding Proteins: Dissecting Ligand Structures through Solvent Environment Occupancy. *J. Phys. Chem. B* **113,** 8717–8724 (2009).

213. Zechel, D. L. & Withers, S. G. Glycosidase mechanisms:Anatomy of a finely tuned catalyst. *Acc. Chem. Res.* **33,** 11–18 (2000).

214. Adachi, W. *et al.* Crystal structure of family GH-8 chitosanase with subclass II specificity from Bacillus sp. K17. *J. Mol. Biol.* **343,** 785–795 (2004).

215. Ohnuma, T. *et al.* Crystal structure of a 'loopless' GH19 chitinase in complex with chitin tetrasaccharide spanning the catalytic center. *Biochim. Biophys. Acta - Proteins Proteomics* **1844,** 793–802 (2014).

216. Wang, Q., Withers, S. G., Graham, R. W., Trimbur, D. & Warren, R. A. J. Changing Enzymic Reaction Mechanisms by Mutagenesis: Conversion of a Retaining Glucosidase to an Inverting Enzyme. *J. Am. Chem. Soc.* **116,** 11594–11595 (1994).

217. Guo, X., Laver, W. G., Vimr, E. & Sinnott, M. L. Catalysis by Two Sialidases with the Same Protein Fold but Different Stereochemical Courses: A Mechanistic Comparison of the Enzymes from Influenza A Virus and Salmonella typhimurium. *J. Am. Chem. Soc.* **116,** 5572–5578 (1994).

218. Gloster, T. M., Turkenburg, J. P., Potts, J. R., Henrissat, B. & Davies, G. J. Divergence of Catalytic Mechanism within a Glycosidase Family Provides Insight into Evolution of Carbohydrate Metabolism by Human Gut Flora. *Chem. Biol.* **15,** 1058–1067 (2008).

219. Malgas, S.; van Dyk, J. S.; Pletschke, B. I. A review of the enzymatic hydrolysis of mannans and synergistic interactions between β-mannanase, β-mannosidase and α-galactosidase. *World J. Microbiol. Biotechnol.* **31,** 1167−1175 (2015).

220. Dhawan, S.; Kaur, J. Microbial mannanases: an overview of production and applications. *Crit. Rev. Biotechnol.* **27,** 197−216 (2007).

221. van Zyl, W. H.; Rose, S. H.; Trollope, K.; Gö rgens, J. F. Fungal β-mannanases: Mannan hydrolysis, heterologous production and biotechnological applications. *Process Biochem.* **45,** 1203−1213 (2010).

222. Jin, Y. *et al.* A β-mannanase with a lysozyme-like fold and a novel molecular catalytic mechanism. *ACS Cent. Sci.* **2,** (2016).

223. Tailford, L. E. *et al.* Structural and biochemical evidence for a boat-like transition state in β-mannosidases. *Nat. Chem. Biol.* **4,** 306–312 (2008).

224. Tankrathok, A. *et al.* A Single Glycosidase Harnesses Different Pyranoside Ring Transition State Conformations for Hydrolysis of Mannosides and Glucosides. *ACS Catal.* **5,** 6041–6051 (2015).

225. Davies, G. & Henrissat, B. Structures and mechanisms of glycosyl hydrolases. *Structure* **3,** 853–859 (1995).

226. Nguyen, C. N., Cruz, A., Gilson, M. K. & Kurtzman, T. Thermodynamics of water in an enzyme active site: Grid-based hydration analysis of coagulation factor xa. *J. Chem. Theory Comput.* **10,** 2769–2780 (2014).

227. Balius, T. E. *et al.* Testing inhomogeneous solvation theory in structure-based ligand discovery. *Proc. Natl. Acad. Sci.* 201703287 (2017). doi:10.1073/pnas.1703287114

228. Hammes, G. G., Benkovic, S. J. & Hammes-Schiffer, S. Flexibility, diversity, and cooperativity: pillars of enzyme catalysis. *Biochemistry* **50,** 10422–10430 (2011).

229. Qasba, P. K., Ramakrishnan, B. & Boeggeman, E. Substrate-induced conformational changes in glycosyltransferases. **30,** (2005).

230. Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J. & Imberty, A. Structures and mechanisms of glycosyltransferases. *Glycobiology* **16,** 29–37 (2006).

231. Ramakrishnan, B., Ramasamy, V. & Qasba, P. K. Structural snapshots of β-1,4-galactosyltransferase-1 along the kinetic pathway. *J. Mol. Biol.* **357,** 1619–1633 (2006).

232. Sheng, F., Jia, X., Yep, A., Preiss, J. & Geiger, J. H. The crystal structures of the open and catalytically competent closed conformation of Escherichia coli glycogen synthase. *J. Biol. Chem.* **284,** 17796–17807 (2009).

233. Baskaran, S., Roach, P. J., DePaoli-Roach, A. A. & Hurley, T. D. Structural basis for glucose-6-phosphate activation of glycogen synthase. *Proc. Natl. Acad. Sci.* **107,** 17563–17568 (2010).

234. Lira-Navarrete, E. *et al.* Dynamic interplay between catalytic and lectin domains of GalNAc-transferases modulates protein O-glycosylation. *Nat. Commun.* **6,** 6937 (2015).

235. V.M. Chikwana,M. Khanna, S. Baskaran, et al. Structural basis for 2′-phosphate in- corporation into glycogen by glycogen synthase. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20976–20981 (2013).

236. Adeva-Andany, M. M., González-Lucán, M., Donapetry-García, C., Fernández-Fernández, C. & Ameneiros-Rodríguez, E. Glycogen metabolism in humans. *BBA Clin.* **5,** 85–100 (2016).

237. Lomako J, Lomako WM, W. W. A self-glucosylating protein is the primer for rabbit muscle glycogen biosynthesis. *FASEB J* **2,** 3097–3103 (1988).

238. Smythe C, C. P. The discovery of glycogenin and the priming mechanism for glycogen biogenesis. *Eur J Biochem* **200,** 625–631 (1991).

239. Hurley, T. D., Stout, S., Miner, E., Zhou, J. & Roach, P. J. Requirements for catalysis in mammalian glycogenin. *J. Biol. Chem.* **280,** 23892–23899 (2005).

240. Issoglio, F. M., Carrizo, M. E., Romero, J. M. & Curtino, J. A. Mechanisms of monomeric and dimeric glycogenin autoglucosylation. *J. Biol. Chem.* **287,** 1955–1961 (2012).

241. Chaikuad, A. *et al.* Conformational plasticity of glycogenin and its maltosaccharide substrate during glycogen biogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 21028–21033 (2011).

242. Bazán, S., Issoglio, F. M., Carrizo, M. E. & Curtino, J. A. The intramolecular autoglucosylation of monomeric glycogenin. *Biochem. Biophys. Res. Commun.* **371,** 328–332 (2008).

243. Zhang, Y. *et al.* Roles of Individual Enzyme - Substrate Interactions by α -1,3-Galactosyltransferase in Catalysis and Specificity. *Biochemistry* **42,** 13512–13521 (2003).

244. Raich, L. *et al.* A Trapped Covalent Intermediate of a Glycoside Hydrolase on the Pathway to Transglycosylation. Insights from Experiments and Quantum Mechanics/Molecular Mechanics Simulations. *J. Am. Chem. Soc.* **138,** (2016).

245. Hurley, T. D., Walls, C., Bennett, J. R., Roach, P. J. & Wang, M. Direct detection of glycogenin reaction products during glycogen initiation. *Biochem. Biophys. Res. Commun.* **348,** 374–378 (2006).

246. D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R. E., Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. K., T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. S., R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. & Kollman, W. and P. A. AMBER 14. *University of California, San Francisco.* (2014).

247. Nose, S. Á molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **52,** 255–268 (1984).

248. Isralewitz, B., Gao, M. & Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **11,** 224–230 (2001).

249. Park, S., Khalili-Araghi, F., Tajkhorshid, E. & Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **119,** 3559 (2003).

250. Chen, P.-C. & Kuyucak, S. Accurate Determination of the Binding Free Energy for KcsA-Charybdotoxin Complex from the Potential of Mean Force Calculations with Restraints. *Biophys. J.* **100,** 2466–2474 (2011).

251. Ni, T. *et al.* Structure and lipid-binding properties of the kindlin-3 pleckstrin homology domain. *Biochem. J.* **474,** 539–556 (2017).

252. Sun, H. *et al.* Revealing the favorable dissociation pathway of type II kinase inhibitors via enhanced sampling simulations and two-end-state calculations. *Sci. Rep.* **5,** (2015).

253. Domański, J., Hedger, G., Best, R. B., Stansfeld, P. J. & Sansom, M. S. P. Convergence and Sampling in Determining Free Energy Landscapes for Membrane Protein Association. *J. Phys. Chem. B* **121,** 3364–3375 (2017).