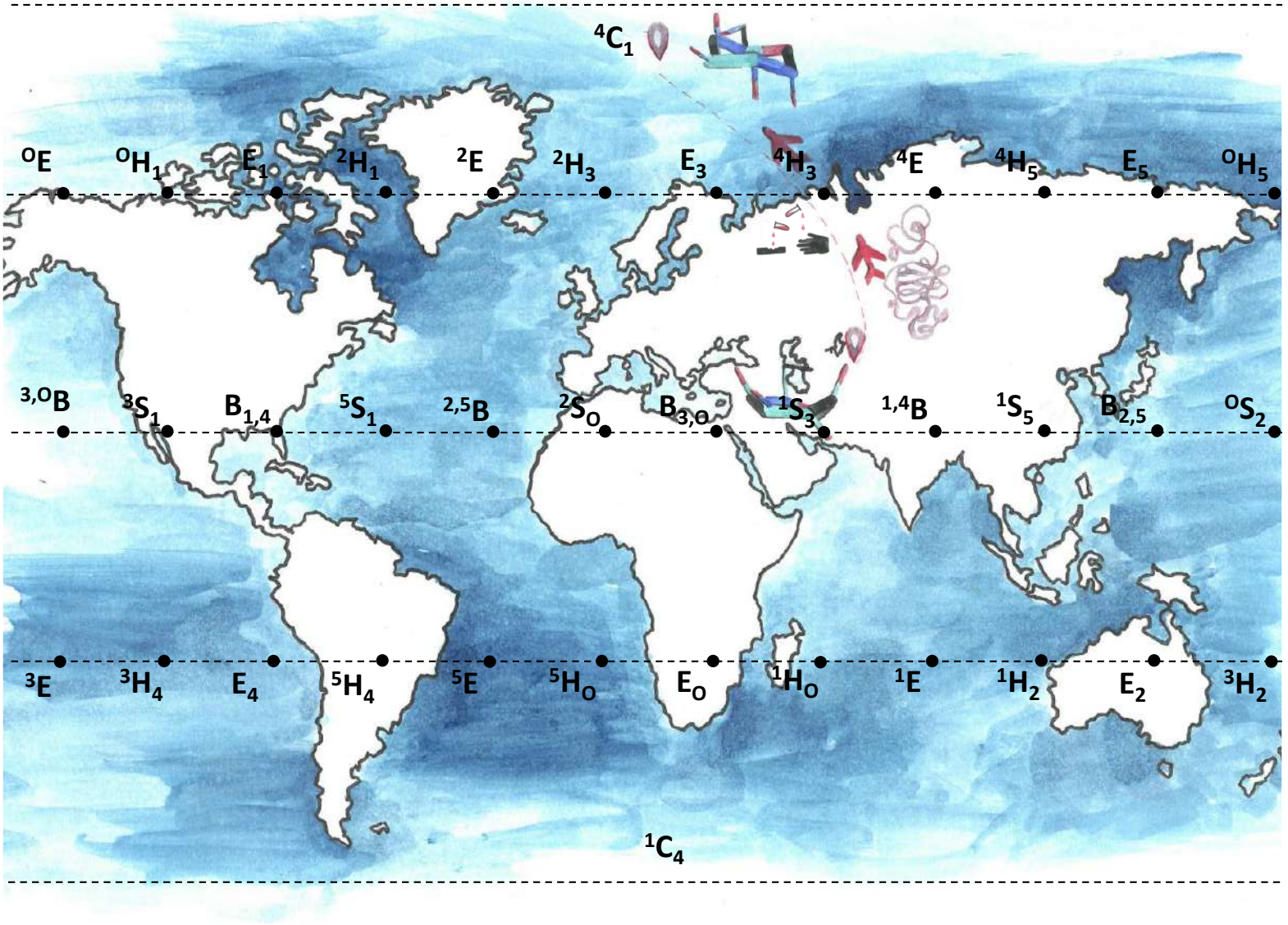
Conformational catalytic itineraries of five- and six-membered sugar rings in enzymatic and superacid media

Alba Nin Hill

UNIVERSITAT DE BARCELONA

Conformational catalytic itineraries of five- and six-membered sugar rings in enzymatic and superacid media



Alba Nin-Hill

Universitat de Barcelona
Facultat de Química

Programa de Doctorat en Química Teòrica i
Modelització Computacional

Conformational catalytic itineraries of five- and six-membered sugar rings in enzymatic and superacid media

Memòria de recerca presentada per n'**Alba Nin Hill** per tal d'optar al títol de Doctora per la Universitat de Barcelona.

Setembre de 2020

Noms, cognoms i signatura de la doctoranda:
Alba Nin Hill

Noms, cognoms i signatura de la directora i tutora:
Dra. Carme Rovira Virgili

Acknowledgements

Haig d'agrair a moltíssima gent el haver arribat fins aquí i no se ni com començar, ho farem per ordre temporal...

El primer lloc l'ocupen la meva família; pares, per ser els iniciadors de tot plegat i per sempre donar-me ànims per tirar endavant; avis, per acollir-me un temps a Barcelona mentre estudiava la carrera i el màster; també als altres avis, que tot i que fa un temps que no és el mateix, em feu recordar una infància molt maca; i a ma germana, que tot i no fer-li massa cas al principi, ara és de les meves millors amigues.

Si segueixo anant per ordre cronològic, haig d'agrair també tots els meus amics de Vilafranca, amb bastants dels quals ens coneixem des dels 3 anys i amb altres ens vam fer amics durant el batxillerat: Míriam, Judith, Raimon, Aleix, Andrea, Dani, Ferran, Raquel, Borja, Mireia (que m'has anat acompanyant per tot el periple de la universitat!).

Després va venir la carrera a Barcelona, una gran llista d'amistats amb els que hem creat la nostra pròpia tradició de Nadal i que espero que no acabi mai! Més tard va venir el màster i amb ell uns grans dos anys que no acabem mai de rememorar amb la "família" que vam crear; en particular Mireia (altre cop!), Alex, Toni (company de pis comunista durant un temps) i Efrem (tindràs dedicatòria personal, no pateixis!).

També haig d'agrair tot el grup SQPBIO, que òbviament sense ells aquesta Tesi no hagués estat possible. A la Carme per confiar amb mi des del TFG i motivar-me a seguir fent un màster i un doctorat i guiar-me durant tots aquests anys. També a la meva família científica; Mertxe, per sempre contestar-me tots els e-mails plens de dubtes i durant un temps fer-ho en persona també; Lluís, per compartir tants moments i també guiar-me i ajudar-me sempre tot i estar tan enfeinat (tot i que et pensis que vas ser un mal pare, no et canviaria pas); Joan, company de copets a l'esquena quan ho necessitem, aquests més de 4 anys haguessin estat més tristos sinó

t'haguessis passat a la dark side de la química computacional. Bea (t'has convertit en la veterana del grup, ets millor successora del que vam ser nosaltres!), Oriol, Irene, Daniil, Mariana (thanks for all the experimental literature and taking a look at it!), Binju... gràcies a tots!

Also, a special thanks to the people of YSBL for making me feel a little bit like home in the northern lands of England. Specially to Nick for helping a newbie at the lab, Eleanor for welcoming me to her home and Jon, Saioa and Iraia for all the films, pizzas, Eurovision and bbq which made me very happy.

I a tu, Efrem, per estar sempre al meu costat i aguantar-me durant bàsicament els mateixos anys que ha durat aquesta Tesi. Amb tots els alts i baixos que puguem tenir, ets el millor company de viatge que podria somiar.

I bàsicament a tothom (Grup Fem Pinya de Carnaval, els Bandarres, Irene en especial per ajudar-me a corregir una petita part d'aquesta tesi, Setem...) que ha anat formant part de la meva vida.

Full list of publications

This Thesis has resulted in the following publications:

- Nin-Hill, A.; Rovira, C.; "The catalytic reaction mechanism of β-galactocerebrosidase, the enzyme deficient in Krabbe disease" *ACS Catalysis* Published ahead of print (https://doi.org/10.1021/acscatal.0c02609) (2020) Selected as ACS Editors Choice.

- McGregor, N.[#]; Artola, M.[#]; Nin-Hill, A.[#]; Linzel, D.; Haon, M.; Reijngous, J.; Ram, A.; Rosso, M-N.; van der Marel, G.A.; Codée, J.D.C.; van Wezel, G.P.; Berrin, J-G.; Rovira, C.; Overkleeft, H.S.; Davies, G.J.; "Rational Design of Mechanism-Based Inhibitors and Activity-Based Probes for the Identification of Retaining α-L-Arabinofuranosidases" *Journal of the American Chemical Society*, 142, 10, 4648-4662 (2020) [#]Equal contribution.

- Raich, L.; Nin-Hill, A.; Ardèvol, A.; & Rovira, C. "Enzymatic cleavage of glycosidic bonds: strategies on how to set up and control a QM/MM metadynamics simulation" *In Methods in enzymology*, 577, 159-183. (2016)

Publications not included in this Thesis and unrelated to it:

- Gomila, A.; Rustler, K.; Maleeva, G.; Nin-Hill, A.; Wutz, D., Bautista-Barrufet, A.; Rovira, X.; Bosch, M.; Mukhametova, E.; Petukhova, E.; Ponomareva, D.; Mukhamedyarov, M.; Peiretti, F.; Alfonso-Prieto, M.; Rovira, C.; König, B.; Bregestovski, P.; Gorostiza, P.; "Photocontrol of endogenous glycine receptors in vivo" *Cell Chemical Biology*, 27, 1-9. (2020)

- López-Martínez, M.; López-Ortiz, M.; Antinori, M. E.; Wientjes, M.; Nin-Hill, A.; Rovira, C.; Croce, R.; Díez-Pérez, I.; Gorostiza, P.; "Electrochemically Gated Long-Distance Charge Transport in Photosystem I" *Angewandte Chemie International Edition*, 58(38), 13280-13284. (2019)

- Maleeva, G.; Wutz, D.; Rustler, K.; Nin-Hill, A.; Rovira, C.; Petukhova, E.; Bautista-Barrufet, A.; Gomila-Juaneda, A.; Scholze, P.; Peiretti, F.; Alfonso-Prieto, M.; König, B.; Gorostiza, P.; Bregestovski, P.; "A photoswitchable

GABA receptor channel blocker" *British journal of pharmacology*, 176(15), 2661-2677. (2019)

- Lagunas, A.; Guerra-Castellano, A.; Nin-Hill, A.; Díaz-Moreno, I.; Miguel, A.; Samitier, J.; Rovira, C.; Gorostiza, P.; "Long distance electron transfer through the aqueous solution between redox partner proteins" *Nature communications*, 9(1), 1-7. (2018)

- Gandioso, A.; Bresolí-Obach, R.; Nin-Hill, A.; Bosch, M.; Palau, M.; Galindo, A.; Contreras, S.; Rovira, A.; Rovira, C.; Nonell, S.; Marchán, V.; "Redesigning the coumarin scaffold into small bright fluorophores with far-red to near-infrared emission and large Stokes shifts useful for cell imaging" *The Journal of organic chemistry*, 83(3), 1185-1195. (2018)

- Gandioso, A.; Palau, M.; Nin-Hill, A.; Melnyk, I.; Rovira, C.; Nonell, S.; Velasco, D.; García-Amorós, J.; Marchán, V.; "Sequential Uncaging with Green Light can be Achieved by Fine-Tuning the Structure of a Dicyanocoumarin Chromophore" *ChemistryOpen*, 6(3), 375-384. (2017)

Collaborations

A three months secondment (research included in Chapter 4) was performed in the research groups led by:

- **Prof. Gideon Davies and Dr. Jon Agirre**. York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, YO10 5DD, United Kingdom.

Moreover, part of the research included in this Thesis (Chapter 5) was done in collaboration with the experimental research group led by:

- **Profs. Sebastien Thibaudeau and Ives Blériot**. Université de Poitiers, UMR-CNRS 7285, IC2MP, Equipe 'Synthèse organique', 4 rue Michel Brunet, TSA 51106, Poitiers Cedex 9 86073, France

Other experimental collaborations (work not related to this Thesis) were done with research groups led by:

- **Prof. Pau Gorostiza**. Institute for Bioengineering of Catalonia (IBEC), The Barcelona Institute of Science and Technology, Barcelona 08028, Spain. Networking Biomedical Research Center (CIBER), Madrid 28029, Spain. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Spain.

- **Prof. Burkhard König**. Institute of Organic Chemistry, University of Regensburg, Regensburg, Germany

- **Prof. Piotr Bregestovski**. INSERM, INS, Institut de Neurosciences des Systèmes, Aix-Marseille University, Marseille, France. Department of Normal Physiology, Kazan State Medical University, Kazan, Russia. Institute of Neurosciences, Kazan State Medical University, Kazan, Russia.

- **Prof. Vicente Marchán**. Secció de Química Orgànica, Departament de Química Inorgànica i Orgànica, Universitat de Barcelona, Martí i Franquès 1-11, E-08028 Barcelona, Spain. Institut de Biomedicina de la Universitat de Barcelona (IBUB), E-08028 Barcelona, Spain

List of abbreviations

Abbreviation	Description
CAZymes	Carbohydrate-active enzymes
GH	Glycosyl Hydrolase
GT	Glycosyl Transferase
GEI	Glycosyl-enzyme intermediate
TS	Transition state
MC	Michaelis Complex
PDB	Protein Data Bank
MD	Molecular Dynamics
QM/MM	Quantum Mechanics / Molecular Mechanics
FT	Fourier Transform
F	Structural factors
MR	Molecular Replacement
MIR/SIR	Multiple/Single Isomorphous replacement
MAD/SID	Multiple/Single wavelength Anomalous diffraction
FF	Force Field
FEL	Free Energy Landscape
CP	Car-Parrinello
DFT	Density Functional Theory
a.u.	atomic units

GALC	β-galactocerebrosidase
Gal-β-pNP	4-nitrophenyl β-D-galactopyranose
SapA	Saposin A
GalCer	Galactocerebroside
PUL	Polysaccharide utilization loci
YPD	Yeast extract peptone dextrose
BMGY	Buffered Glycerol-complex Medium
BMMY	Buffered Methanol-complex Medium
mPES	Modified Polyethersulfone
MWCO	Molecular Weight Cut-off
CV	Column Volumes (in section 4.4.1) or Collective Variables (everywhere else)
U	Unit (defined as the amount of enzyme required to remove more than 95% of the carbohydrate from 10 μg of denatured RNase B in 1 hour at 37°C in a total reaction volume of 10 μl)
PEG	Polyethylene glycol
RT	Room Temperature
MS	Mass Spectrum
AkAbfB	*Aspergillus Kawachii* abfB α-L-arabinofuranosidase
PNP-Araf	4-nitrophenyl α-L-arabinofuranoside
GH51	Glycoside Hydrolase family 51
GH54	Glycoside Hydrolase family 54
ABP	Activity-based probe
*Sc*Gas2	*Saccharomyces cerevisiae* Gas2

Summary

Carbohydrates were early considered important only for energy storage and structural support for living organisms. It was not until a few decades ago that scientists discovered that carbohydrates are also involved in much more complex mechanisms, such as modulation of protein structures, signalling in multicellular systems and cell-cell recognition processes, which are relevant in several serious diseases including bacterial and virus infections and cancer. Consequently, carbohydrates are getting more recognition, being nowadays used in drug delivery strategies, vaccines and new therapeutics for treating many types of diseases.

Carbohydrates come in many stereochemistries, configurations and conformations and this makes them very complex molecules to study. The vast amount of carbohydrate-based structures in nature needs a larger number of enzymes responsible of their degradation, synthesis and possible modifications. These are named as carbohydrate-active enzymes (CAZymes), which include glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases (PLs) and carbohydrate esterases (CEs). This Thesis focuses on GHs, which are the enzymes responsible for the hydrolysis of glycosidic bonds.

The possible sugar ring conformations that carbohydrates can adopt in the enzyme active sites during catalysis depends on distinct active site topologies observed for the more than 150 GH families. Each GH (or each family of GHs) hosts a particular catalytic conformational itinerary of the corresponding carbohydrate substrate. Understanding these itineraries is of utmost importance when designing selective inhibitors, as some of the most powerful inhibitors mimic the structure and properties of the transition state (TS) of the hydrolysis reaction.

Molecular mechanisms and catalytic conformational itineraries of GHs can be investigated with both experimental and computational approaches. Typically, a structure of the GH enzyme is needed to further perform computational analyses to elucidate the electronic and conformational changes of the carbohydrate substrate along the catalytic itinerary.

In recent years, our group has investigated the catalytic mechanisms for several GHs, uncovering the influence of substrate sugar conformations in catalysis.

In this Thesis we will focus on two GHs that are of biochemical and biotechnological relevance, for which the catalytic conformational itineraries were still unknown: human β-D-galactocerebrosidase (GALC) and *Aspergillus kawachii* abfB α-L-arabinofuranosidase (AkAbfB). These GHs act on carbohydrates formed by five- and six-membered sugar rings (derivatives of β-D-galactose and α-L-arabinose, respectively). Additionally, and motivated by recent NMR experiments of isolated glycosyl cations, we analyse the properties of five- and six-membered glycosyl cations in different environments: in vacuum, in superacid media and on-enzyme.

In particular, this Thesis is organized as follows:

- **Chapter 1. Introduction.** It contains a general introduction of carbohydrates, carbohydrate-active enzymes and their general catalytic mechanisms. The main objectives of this work are listed in this chapter.

- **Chapter 2. Methods.** It provides a description of all the experimental and computational methods used along this work.

- **Chapter 3. Unravelling the origin of brain diseases: catalytic study of β-Galactocerebrosidase.** We uncover the catalytic molecular mechanism of GALC with both a synthetic substrate analogue (4-nitrophenyl β-D-galactopyranose) and its natural β-galactocerebroside substrate. We also study the influence of the binding of the attached lipid-transfer saposin A protein (SapA) to GALC in the catalytic reaction mechanism.

- **Chapter 4. Retaining α-L-arabinofuranosidases: structure, substrate conformations and catalysis.** We will investigate the conformational properties of potential α-L-arabinofuranosidase inhibitors and compute the catalytic molecular mechanism of the AkAbfB enzyme with the synthetic 4-nitrophenyl α-L-arabinofuranoside substrate using a Michaelis complex structure determined in this Thesis.

- **Chapter 5. Mimicking the transition state of glycosylation reactions using superacids.** We study the geometrical and electronic properties of two different glycosyl cations in various environments (in vacuum, in superacid media and in the active site of a retaining GH enzyme) and determine to which extend a superacid media is able to mimic the enzymatic environment.

- **Chapter 6: General Conclusions.** The main conclusions of this work are summarized.

Contents

Chapter 1

Introduction

1.1 Carbohydrate chemistry

Which are the first images that you see when someone says carbohydrates? Probably, in your brain you will see flashes of bread, pizza, spaghetti, sweet desserts dressed with icing sugar, etc. Thus, all the food needed as energy storage for most of the living organisms. After all the food flashes, our brain probably will also consider the carbohydrates as responsible for structure and strength, such as cellulose. But they are much more important in much more complex mechanisms in living organisms such as modulating protein structures and locating and signalling in multicellular systems and cell-cell recognition, including bacterial and virus infections and some aspects of cancer [1–3]. We can put as an example the worldwide known COVID-19 virus which entry into host cells is mediated by a transmembrane spike glycoprotein, a protein with glycans attached covalently to the aminoacid side chains [4]. Thus, carbohydrates are getting more recognition and being nowadays used in drug delivery strategies, vaccines and new therapeutics for treating many types of diseases [5–9].

But, which is exactly the composition of a carbohydrate? They have C, H and O as their main atoms with the molecular formula of $C_5H_{10}O_5$ and $C_6H_{12}O_6$ being the most common. They come in many stereochemistries, configurations and conformations which converts them into a very complex world to study who had challenged chemists and biochemists for many years, from Emil Fischer (1852-1919), considered the father of carbohydrate chemistry, to nowadays. The simplest way to represent them is using the Fischer projection in which we have a linear chain of a polyhydroxylated aldehydes/ketones, termed as *aldoses* and *ketoses* being the *aldopentoses/aldohexoses* and *ketopentasoses/ketohexoses* the most common ones. If we place the most oxidized functional group (aldehyde or ketone) at the top and

we look at the pointing direction of the OH group in the bottom chiral centre[1] we can discern between D- and L- sugars. When the hydroxyl group is pointing to the right, the sugar is referred to as D-; whereas when the hydroxyl group is pointing to the left, it is referred to as a L-sugar. This notation is the easiest way to distinguish between sugar enantiomers, the first of the many forms that a carbohydrate can adopt. Also, all the possible orientations of the OH groups creates different kind of sugars, such as between glucose and galactose, in which the only difference is the position of the hydroxyl at carbon 4, see Figure 1.1. Besides, there is a great amount of functional groups that can also be attached to the carbon chain (-NAc, -OSO_3^-, among others) and increases exponentially the world of carbohydrates. The linear form of the sugars has multiple orientations when rotating around carbon-carbon bonds, one of these orientations places the hydroxyl group of C5 very close to the aldehyde group (C1, afterwards called as anomeric carbon) in where a chemical reaction can take place, creating a more stable cyclic carbohydrate. Now, there is an extra stereogenic centre (C1) and two isomeric forms can exist; called α or β. When, in the linear form, C4 is adjacent to C1, instead of C5, it can also react and form a 5-membered cyclic sugar. Their corresponding names are pyranoses (6-membered ring) and furanoses (5-membered rings) [1]. These cyclic monosaccharides can also be linked together in a variety of glycosidic bonds; between both anomeric carbons of each monosaccharide (1-1) or between the anomeric carbon of the first monosaccharide and C4 of the second carbohydrate (1-4), etc.; forming disaccharides, oligosaccharides and polysaccharides.

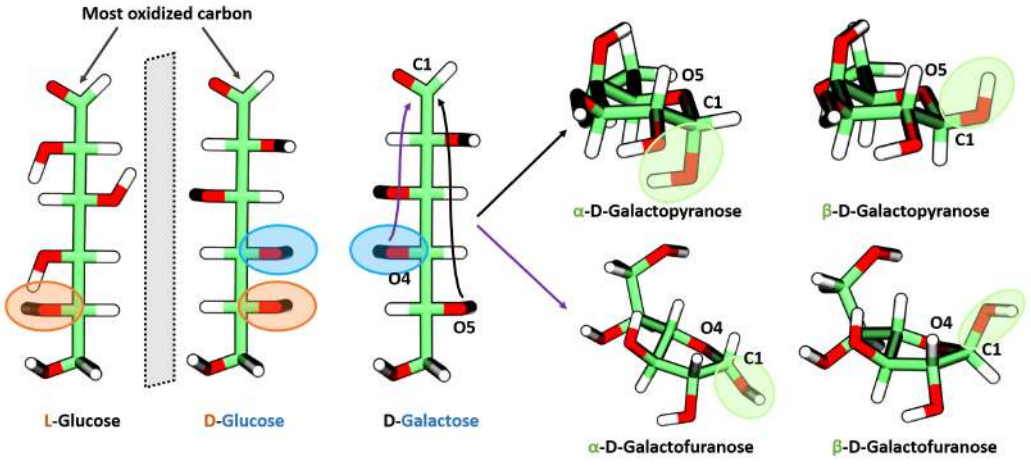


Figure 1.1: One of the infinite examples of the diversity obtained in carbohydrates from subtle changes

[1]defined as an atom in a molecule that is bonded to four different chemical species, allowing for optical isomerism.

1.1.1 Ring puckering coordinates

Besides the vast amount of possible forms that carbohydrates can take, seen in the previous section, we have to sum up the wide range of conformations that the sugar ring can explore; 38 in the case of pyranoses and 20 for furanoses. The terms used to describe them derive from a mathematical study made by Cremer and Pople [10], who built on the treatment of Altona *et. al.* for furanoses [11].

In 5-membered sugar rings the conformations can be classified in two groups: envelope (E) and twist (T) and a defining mean plane in which one, for E, or two, for T, atoms are placed below (denoted by a subscript) or above (denoted by a superscript) to this plane. For better understanding, the ^{3}E conformation has C1, C2, C4 and the ring oxygen in the same plane and C3 is placed above it, see Figure 1.2. To be able to define each conformation only one variable is needed, the so-called pseudorotational angle, ϕ or P. All the conformations do not differ much in terms of free energy, as seen in Lightner *et. al.* [12] and also in this work, in section 4.2.1.

The concept of pseudorotation was first introduced by Kilpatrick *et. al.* [13] and further improved by Altona *et. al.* [11]. In the latest work they propose to use the five torsion angles, θ, that are directly derived from the atomic coordinates and are all zero in the ring planar form. They proposed a relationship of the form;

$$\theta_j = \theta_m cos(P + 4\pi(j-1)/5) \tag{1.1}$$

for these torsion angles (j=0-4), where θ_m is an amplitude and P is a phase angle. Given values for the five θ_j, the phase P can be obtained from;

$$tanP = \frac{(\theta_2 + \theta_4) - (\theta_1 + \theta_3)}{2\theta_0(sin1/5\pi + sin2/5\pi)} \tag{1.2}$$

The standard conformation (P=0°) is defined as the one with the maximal positive C1-C2-C3-C4 torsion angle corresponding to the conformational name of $^{3}T_2$. The main problem from these derivations of the pseudorotational phase of a 5-membered sugar ring is the fact that the relations in equation 1.1 are only approximate and the full set of torsion angles cannot be expressed exactly in terms of two parameters in this way. Consequently, the puckering amplitude calculated by the above procedure will depend somewhat on which atom is chosen as number one. There have been many attempts to develop the puckering description [10, 14–16], being the one implemented in the Plumed software [17] the procedure chosen for this Thesis [18]. In Huang *et. al.* work, they used two proper torsions to describe the puckering of the five-membered rings and its pseudorotation behaviour, as seen in the following equations;

$$Z_x = \frac{\theta_4 + \theta_1}{2cos(4\pi/5)} \tag{1.3}$$

$$Z_x = \frac{\theta_4 - \theta_1}{2sin(4\pi/5)} \tag{1.4}$$

$$tanP = \frac{Z_y}{Z_x} \tag{1.5}$$

$$\theta_m = \sqrt{Z_x^2 + Z_y^2} \tag{1.6}$$

Pyranoses work in the same manner but with a higher number of possible conformations classified in five groups: chairs (C), boats (B), skew-boats (S), half-chairs (H) and envelopes (E). For the 6-membered rings is necessary to have two variables to be able to describe all the possible conformations, θ and ϕ angles. These polar coordinates are represented in a sphere of amplitude Q; which can be represented in a three-dimensional way or just with two-dimensions making use of the so-called plate carrée or Mercator representation, in a similar way as Earth is represented in maps. In contrast to furanoses, pyranose rings do have clear conformational preferences, being the chairs the most preferred one in gas phase and in water solution [19–22].

The puckering coordinates for the 6-membered rings, developed by Cremer and Pople [10], 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:

$$\sum_{j=1}^{6} z_j cos[2\pi(j-1)/6] = 0 \tag{1.7}$$

$$\sum_{j=1}^{6} z_j sin[2\pi(j-1)/6] = 0 \tag{1.8}$$

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

$$\overrightarrow{R'} = \sum_{j=1}^{6} \overrightarrow{R_j} sin[2\pi(j-1)/6] = 0 \tag{1.9}$$

$$\overrightarrow{R''} = \sum_{j=1}^{6} \overrightarrow{R_j} cos[2\pi(j-1)/6] = 0 \tag{1.10}$$

where $\overrightarrow{R_j}$ are the atomic positions. Then, the atomic displacements from the mean plane are given by the dot product:

$$z_j = \overrightarrow{R_j}\,\overrightarrow{n} \tag{1.11}$$

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

$$\overrightarrow{n} = \frac{\overrightarrow{R'} x \overrightarrow{R''}}{|\overrightarrow{R'} x \overrightarrow{R''}|} \tag{1.12}$$

This unit vector is taken as the molecular z-axis. The following system of equations must be solved to obtain the puckering coordinates for a 6-membered ring:

$$Qsin\theta cos\phi = \sqrt{\frac{1}{3}}\sum_{j=1}^{6} z_j cos[\frac{2\pi}{3}(j-1)]$$

$$Qsin\theta sin\phi = -\sqrt{\frac{1}{3}}\sum_{j=1}^{6} z_j sin[\frac{2\pi}{3}(j-1)] \tag{1.13}$$

$$Qcos\theta = \sqrt{\frac{1}{6}}\sum_{j=1}^{6}(-1)^{(j-1)} z_j$$

the Q coordinate being the total puckering amplitude and θ and ϕ the polar coordinates needed to describe the whole puckering sphere, see Figure 1.2.

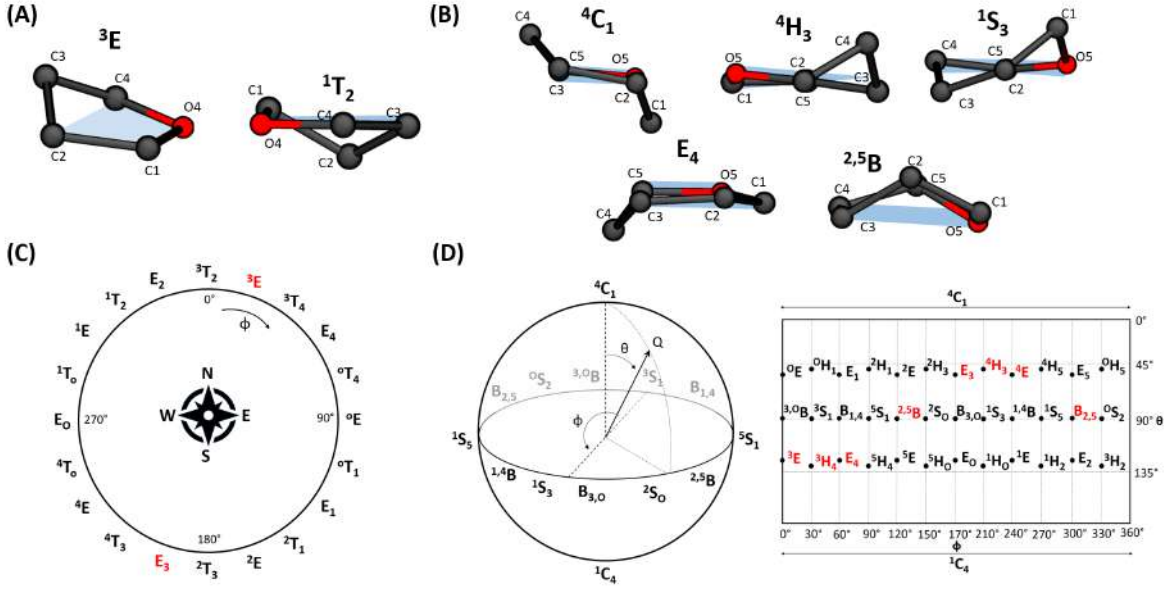


Figure 1.2: **(A)** One conformation example for the two possible groups, E and T, of 5-membered sugar rings. The mean plane is represented in light blue. **(B)** One conformation example for the five possible groups, C, H, S, E and B, of 6-membered sugar rings. The mean plane is represented in light blue. **(C)** Pseudorotational pathway of the furanose ring. **(D)** Puckering sphere and Mercator representation of pyranose rings. In red are marked all the conformations that stabilize an oxocarbenium ion as explained in section 1.2.4.

1.2 Glycoside Hydrolases

The vast amount of carbohydrate-based structures in nature needs a larger group of enzymes responsible for their synthesis, degradation and modification. Their

name is carbohydrate-active enzymes (CAZymes) which can be classified in glycoside hydrolases (GHs) or glycosiltransferases (GTs), among others; depending on their specific function. These enzymes are not only important to be essential for live organisms but also for food, detergent, oil, gas and biotechnology industries. Their myriad of applications goes from biofuel production to drug design. In this work, we focus on GHs, which are responsible to hydrolyse glycosidic bonds. These bonds are among the strongest ones in nature, with a half-life of 4.7 million years [1]. In the presence of GHs, glycosidic bonds are hydrolysed in just a few milliseconds.

1.2.1 Classification

GHs can be classified according to:

i. The stereochemistry of the product. GHs can be either **retaining** or **inverting** with respect to the configuration of the anomeric carbon in the reactant and product state.

ii. The location of the glycosidic bond to be cleaved. We refer either to **endo-** or **exo-**glycoside hydrolases. In the former, the cleaved bond is in the middle of the carbohydrate chain whereas in the latter, the broken bond is in the non-reducing extreme, i.e., at the end of the chain. Whether a GH is exo- or endo- has a relation with the enzyme active site topology. In this sense, exo-glycosidases usually have a pocket binding site topology while endo-glycosidases adopt cleft or tunnel topologies, see Figure 1.3.

iii. The similarity in the amino acid sequence. This is the most useful classification to help predicting the function of new sequences of enzymes. The online database CAZy [23] (*www.cazy.org*, see also *www.cazypedia.org*) contains an updated classification of GHs based in their sequence [24]. When this work is written, GHs are grouped in 162 families. Enzymes that belong to the same family usually share the same enzymatic mechanism.

Even though there is an increasing number of X-ray structures available in the Protein Data Bank (PDB), a database with more than 150 000 structures (*www.rcbs.org*, 2020) [25], the lack of structures of CAZyme complexes with their corresponding carbohydrate substrates hampers the study of their molecular catalytic mechanisms.

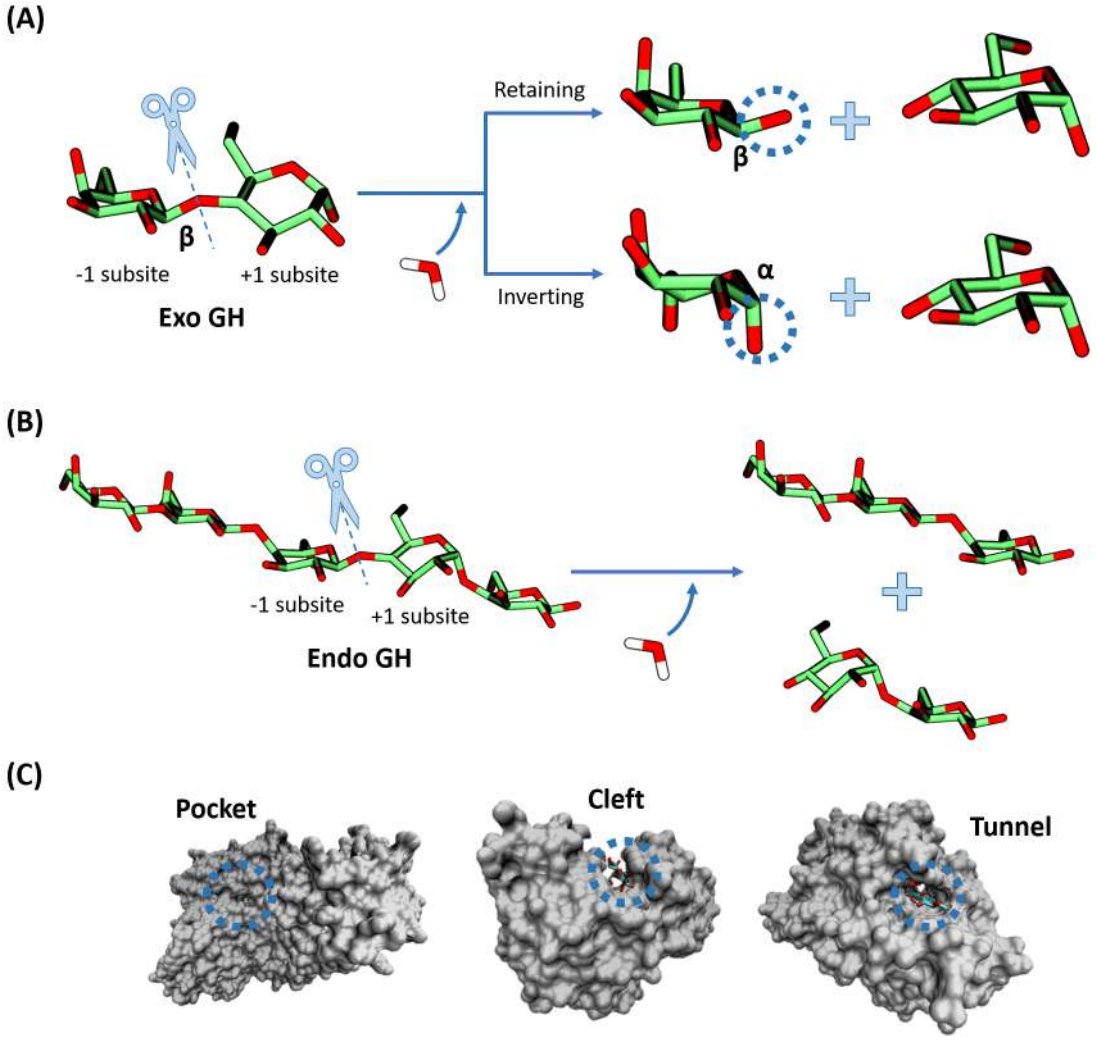


Figure 1.3: **(A)** Classification according to the stereochemistry of the product, retaining or inverting the configuration of the anomeric carbon. As an example, we use an exo-GH in which the location of the glycosidic bond is placed at the end of the sugar chain. **(B)** The location of the cleaved glycosidic bond is placed in the middle of the chain, an endo-GHs. They can also retain or invert the configuration of the anomeric carbon, not being represented here. **(C)** Active site topologies in GHs from left to right; pocket, cleft and tunnel topologies, the active site is marked with a blue dashed circle.

1.2.2 Reaction mechanism

Despite their differences, most of GHs share a common mechanism where two residues, placed in the so-called -1 subsite, assist the reaction, acting as acid (acid/base residue) and base (nucleophile residue) [26].

The **inverting mechanism** is achieved via a one step, single-displacement mechanism with the assistance of the two catalytic residues. The classical Koshland **retaining mechanism** is a two-step double S_N2-type displacement mechanism, in which the first step consists on the acid/base residue activating the glycosidic bond while the nucleophile attacks the anomeric carbon to form a stable covalent glycosyl-enzyme intermediate (GEI). The reaction ends when a water activated by the acid/base residue, now acting as a general base, attacks as a nucleophile to the GEI to finally reach the product of a cleaved sugar with the same anomeric center configuration as the initial compound, see Figure 1.4. This mechanism is the one found throughout this Thesis.

Regardless of the type of mechanism each reaction step involves an oxocarbenium ion-like character of the sugar as a transition state (TS). They are very elusive due to its extremely short life time, and they have been studied by quantum chemical calculations [27] and mass spectrometry [28]. Also, kinetic measurements of the isotope effect [27] and the use of a cationic cyclization reaction [29] indirectly support the existence of oxocarbenium ion-like species. Recently, pure oxocarbenium ions have been trapped by nuclear magnetic resonance (NMR) using superacid media [30].

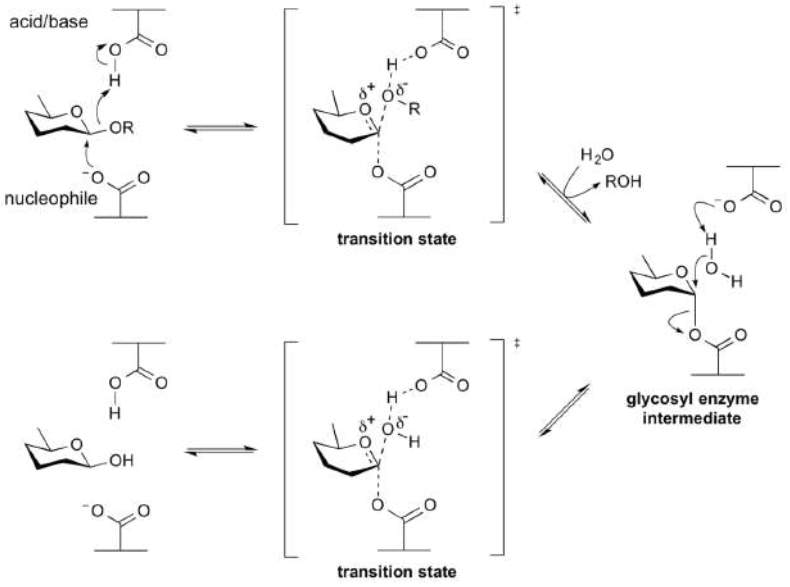


Figure 1.4: Catalytic mechanisms of retaining GHs. *Figure modified from Cazypedia (Spencer Williams, 17/3/2020 via Wikimedia Commons, Creative Commons Attribution)*

1.2.3 Superacid chemistry, strategies to trap oxocarbenium ions

Superacids are defined classically as an acid more acidic than pure sulfuric acid [31] or, according to the modern definition, a superacid is a medium in which the chemical potential of a proton is higher than in pure sulfuric acid [32]. They are mainly composed by mixing a Brønsted acid with a Lewis acid and some examples of them are fluoroantimonic acid ($HF:SbF_5$), magic acid ($HSO_3F:SbF_5$) or fluoroboric acid ($HF:BF_3$); being the former the most acidic one, having a protonating ability over a billion times greater than 100% sulfuric acid [33].

Martin *et. al.* [30] were able to observe the elusive glycosyl cation thanks to using a superacid media (fluoroantimonic acid) which stabilized the cation long enough to be able to detect it by NMR techniques. Even though this represented a great advance in the field of glycosyl cations, there was an important question to be solved; can the observed glycosyl cation in the extreme superacid media be related to the one formed in the active site of a GH during the hydrolysis reaction? In other words, is the superacid media a good mimic of a GH active site? This question will be answered in Chapter 5.

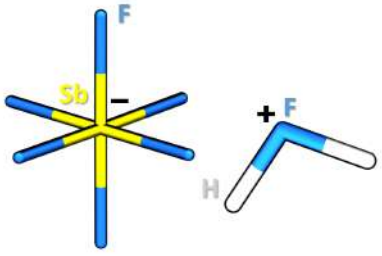


Figure 1.5: Fluoroantimonic acid ($HF_2^+:SbF_6^-$)

1.2.4 Conformational catalytic itinerary

The myriad of conformations that sugars can adopt (section 1.1.1) is very important in GH catalysis. A vast amount of Michaelis (enzyme-substrate) complexes found in X-ray crystallography, specially for β-GHs, exhibit a distorted substrate conformation. In particular, the sugar ring located at the -1 enzyme subsite (i.e. the one bearing the C-O bond to be hydrolysed by the enzyme), is distorted away from the ground-state $^4C_1$ conformation [34–37]. Why is this distortion induced by the enzyme?

Three major reasons are invoked to justify the need for a substrate to change from a ground-state to a distorted and typically higher energy conformation.

i. The anomeric carbon and the nucleophile need to be "in line", however there is a steric hindrance between both due to the axial orientation of the anomeric carbon hydrogen in the ground-state of pyranose rings, $^4C_1$, whereas a distorted conformation removes this steric hindrance and also places the glycosidic oxygen in a good orientation for catalysis (see Figure 1.6).

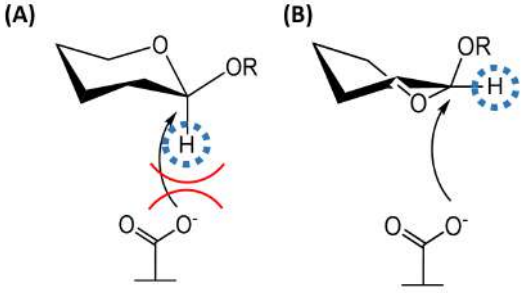


Figure 1.6: **A)** Steric hindrance for a $^4C_1$ conformation **B)** Alleviation of the hindrance upon ring distortion.

ii. Since the reaction is typically dissociative, the TS exhibits a high oxocarbenium ion-like character. The partial charge (δ^+) developed on the anomeric carbon is stabilized by a partial double bond formation with the pyranic oxygen (O5/O4 in six-/five-membered rings). This particular stabilization requires planarity of the C4/C5, O5/O4, C1, C2 atoms that not all the possible conformations can achieve. Actually, only 8 conformers for pyranoses and 2 conformers for furanoses are able to optimally stabilize an oxocarbenium ion-like sugar. These are $^4H_3$, $^3H_4$, $^{2,5}B$, $B_{2,5}$, ^4E, E_4, ^3E and E_3 for pyranoses and ^3E and E_3 for furanoses, see the previous Figure 1.2.

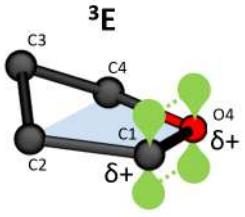


Figure 1.7: One of the possible furanose ring conformations at the oxocarbenium ion-like TS of the reaction catalyzed by GHs.

iii. It can also be explained either by the antiperiplanar lone-pair hypothesis (ALPH) or the principle of least nuclear motion (PLNM). The ALPH proposes

that, for instance in sugars, the hydrolysis reaction is favoured when the non-bonding (i.e. lone pair) electrons of O5 (for six-membered sugar rings) or O4 (for five-membered sugar rings) lie antiperiplanar to the leaving group bond. Thus, in this way the overlapping between the O5/O4 lone pair orbital and the antibonding σ^* orbital of the leaving group is maximized [38, 39]. The leaving group of β-glycosides is placed in an equatorial position in their ground-state ($^4C_1$) and thus, their hydrolysis is unfavoured according to ALPH. Instead, when they adopt a distorted conformation, the lone pair of O5/O4 is disposed in an antiperiplanar way to the leaving group bond. The same conclusion is reached from the PLNM, which stated that those reactions involving the least nuclei motion are favoured [40, 41]. In this case, the equatorial orientation of the leaving group in β-glycosides in its ground-state is unfavoured for the hydrolysis reaction as the C1-H1 bond has to rotate nearly 90° from the MC to the TS to achieve the planarity of the oxocarbenium ion. When the sugar adopts a distorted conformation, less nuclear motion needs to be done to arrive to the oxocarbenium ion configuration.

Which distorted conformation a sugar will adopt at different GHs also depends on their distinct active site topologies as different conformations place the exocyclic groups in different orientations which will fit better for particular enzymes. Thus, depending on these restrictions (fitting in the active site and adopting a MC conformation which will eager the path to the restricted oxocarbenium ion-like TS conformation) we can draw specific catalytic conformational itineraries for each enzyme (or family of enzymes) into the puckering coordinates map (see Figure 1.8). Understanding this is of great importance when designing selective inhibitors as the usually more powerful inhibitors are the ones mimicking the properties of the TS [42]. But, how can we study the reaction mechanisms and conformational catalytic itineraries? The first step is to obtain a high resolution MC structure to start deciphering the conformation adopted by the -1 sugar, which it is not always easy (as discussed below). It is nearly impossible to trap the TS experimentally due to its extremely short life-time, thus, this is when computational methods are needed.

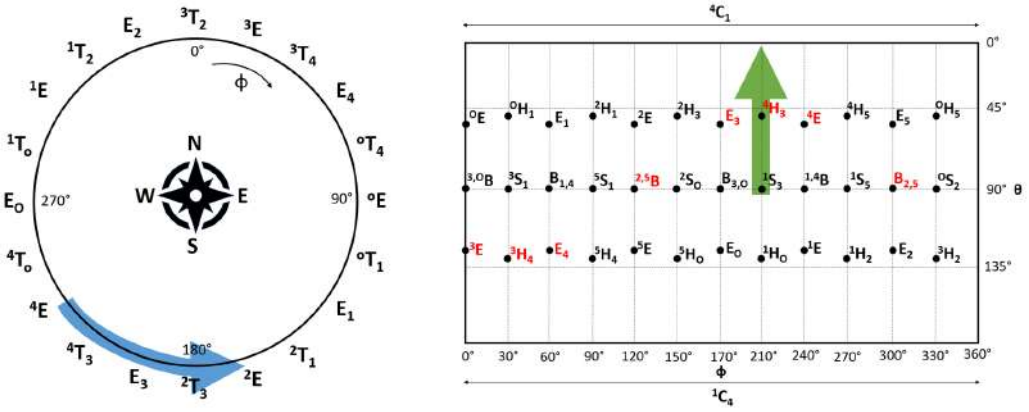


Figure 1.8: Conformational catalytic itinerary from the MC to the GEI for a five-membered sugar ring (α-L-arabinofuranosidase of GH54 family, this Thesis) marked with a blue arrow and for a six-membered sugar ring (β-D-glucosidase of GH5 family [34]) marked with a green arrow.

1.3 Crystal structures and QM/MM simulations: deciphering catalytic mechanisms

1.3.1 First step, obtaining a high resolution MC structure: a X-ray crystallography job

Ideally, to be able to study catalytic mechanisms, one would like to start with the Michaelis complex of the wild-type enzyme with its corresponding natural glycoside substrate. This kind of structure is rarely available as the reaction timescale (ms) is below the time resolution of the X-ray crystallography experiment (s). To be able to trap a structure as similar as possible to the ideal one, crystallographers use various techniques which slows or knock-out the enzymatic reaction. They can modify different aspects of the experiment, such as:

i. **the substrate.** Use sugar-base inhibitors or substrates that are not able to react with the catalytic residues.

ii. **the catalytic residues.** Mutate either the nucleophile, the acid/base or even some assistant residues.

iii. **the protein.** Work in non-optimal pHs in where the whole enzyme is inactive.

Using these modifications can also lead to non-competent structures, thus the study of reaction mechanisms needs to be performed with care.

1.3.2 Second step, uncovering the reaction pathway: a computational job.

To assure that the modifications done to be able to trap a MC result in a competent structure, or to transform the non-competent crystal structures into competent ones; we would need a technique which has less limitations than experiments, and in here is where computer simulations enter into the game. After verifying the first step of the reaction, one can follow the whole reaction mechanism, characterizing the nature of the TS and what is more, provide further information on electronic rearrangements not able to be observed experimentally. Using classical molecular dynamics (MD), quantum mechanics/molecular mechanics (QM/MM) and metadynamics methods (Chapter 2) we are able to simulate catalytic itineraries of GHs in an accurate and affordable computational cost. Summarizing all, in this particular field, and in many others, computational and experimental approaches complement each other to reach the final objective; understanding the complex world of carbohydrates and of the enzymes acting on them and consequently, how they influence in all the living organisms.

1.3.3 Open questions in GH catalysis

Despite the recent advances done in the field of CAZymes mechanisms, there are still many unsolved mechanistic aspects in various GHs families. In this Thesis we broad the knowledge of GH catalysis by focusing in two GHs whose catalytic mechanisms were unknown or controversial.

Specifically, we addressed the question why an undistorted chair conformation of the synthetic 4-nitrophenyl β-D-galactopyranose substrate is observed in the crystal structure of the MC of β-galactocerebrosidase (GALC). GALC is a GH that catalyzes the cleavage of β-galactosidic bonds in glycosphingolipids and its misfunction causes Krabbe disease, a neurodegenerative rare disorder. We investigated the origin of the unusual substrate conformation and uncovered the conformational catalytic itinerary. Additionally, a recently solved crystallographic structure of GALC bound to the lipid-transfer saposin A (SapA) protein made us wonder which is the role of SapA in the molecular mechanism of GALC. We tried to answer these questions in Chapter 3.

The second GH studied in this Thesis is an α-L-arabinofuranosidase from GH54 family (AkAbfB). We investigated whether furanoside-active GHs work similarly to the better studied pyranoside-active GHs. In Chapter 4 we used both experimental and computational techniques to solve the first MC structure of AkAbfB in complex with 4-nitrophenyl α-L-arabinofuranoside substrate and computed its molecu-

lar mechanism. We also analysed the ability of potential α-L-arabinofuranosidase inhibitors to mimic the putative MC/TS conformation in the enzyme active site.

Finally, in Chapter 5 we studied the molecular properties of two types of glycosyl cations, either with five- or six-membered sugar rings, in superacid media and assess whether these cations are similar to the ones formed during GH catalysis.

Objectives

In this Thesis we have used computational techniques (MD, QM/MM MD and metadynamics) with the aim of unravelling the molecular mechanism of two GHs and the properties of glycosyl cations in different environments. Experimental techniques (protein production and X-ray crystallography) have been also used in Chapter 4. The following specific objectives have been pursued:

- Clarify whether the unusual chair conformation of the synthetic Gal-β-pNP substrate observed in the Michaelis complex of GALC is due to the non-optimal experimental conditions or it is a functional feature of the GALC enzyme (Chapter 3).
- Find out whether the conformations adopted by the Gal-β-pNP substrate can be extended to the GALC natural substrate, GalCer, and understand the role of the SapA protein in complex with GALC in molecular mechanism catalyzed by GALC *in vivo* (Chapter 3).
- Assess the conformational mimicry of some α-L-arabinofuranose-like inhibitors with respect to the preferred conformations adopted by α-L-arabinofuranose (Chapter 4).
- Obtain a Michaelis complex structure of a family GH54 α-L-arabinofuranosidase (Chapter 4).
- Determine the molecular mechanism of catalysis by α-L-arabinofuranosidase (Chapter 4).
- Parametrize a force field for the SbF_6^-/H_2F^+ mixture to investigate the conformational free energy landscape of glycosyl cations in superacid media (Chapter 5).
- Determine whether a glycosyl cation trapped in superacid media is a *bona-fide* mimic of the short-lived oxocarbenium ion-like species formed during the enzymatic glycosylation reaction (Chapter 5).

Chapter 2

Methods

As stated in Chapter 1, the study of GHs is obtained by combining the experimental and computational results which let us understand in detail their reaction mechanisms. Thus, it is important to understand the experimental and computational basis that we will need to use to be able to know its limitations and the results validity. Firstly, the experimental procedure done to be able to obtain a reliable MC structure of any GHs is explained. Secondly, the theory behind the computational tools used along this Thesis is also clarified.

2.1 Protein structure determination in a nutshell

Obtaining a protein structure is a long process with multiple experimental steps. In this section we will try to explain it as simple as possible, as most of the projects presented in this work are not experimental. But as we also touched some experimental techniques in Chapter 4, we considered necessary to explain the basics about these methods.

Protein crystallography has evolved considerably during the more than 50 years since the first protein structure was released (myoglobin structure by Max Perutz [43], Nobel Prize in 1962); from the advance of X-ray crystallography (Dorothy Hodgkin, Nobel Prize in 1964) to the automation of protein production and crystallization, as well as new algorithms and computer software facilitating the data collection and evaluation.

2.1.1 Previous steps; obtaining the protein to crystallize

The first step in protein production consists on obtaining a plasmid[1] that carries the genes of the protein of interest. This plasmid is obtained via **molecular cloning** [45]; this technique consists on taking the gene of interest (usually previously amplified by PCR) and separated (cut) from the rest of the DNA by the use of a restriction enzyme[2] which is also used to cut a commercial circular plasmid, called expression vector. By being treated with same restriction enzymes, the end from both the gene of interest and the expression vector can be attached to each other via base pairing forming weak hydrogen bonds, a process called **ligation**. A DNA ligase facilitates the bond between the plasmid and the gene containing the information of the protein of interest by catalysing the formation of a phosphodiester bond. After ligation, the recombinant plasmids (containing the expression vector with our gene of interest) are inserted into the host cells (which can be bacteria, yeast, mammalian...) by a process called **transformation**[3]. Expression vectors feature different elements that enhance the gene expression and optimize the production of the protein of interest, such as an origin of replication, elements that allow them to be maintained in other organisms (plasmids usually come from bacteria but transformations can also be done in fungal cells, for instance), inducible promoters (inducing gene expression by feeding the host cells with a specific component, such as methanol) and also selectable markers (for instance, antibiotic resistance), to allow the separation of cells that contain the vector from those that do not. Once done the transformation into the host of interest, the cells containing recombinant plasmids (expression vector with our gene of interest) are selected by the use of necessary nutrients to grow (some examples are medias, YPD[4], BMGY[5] and BMMY[6]) together with the antibiotic which the vector-containing cells are resistant to. Afterwards, cells are grown in large-scale in the corresponding media supplemented with the necessary components to promote gene expression.

The second step is **protein purification**. Generally speaking, consists on flowing the solution containing the protein of interest through a column packed with

[1] an extrachromosomal DNA molecule which can replicate independently from chromosomal DNA. Commonly found as a double stranded DNA molecule in a small circle form. They only carry a small number of genes and often, some of them consist on genes that benefit their survival such as antibiotic resistance [44].

[2] an enzyme that cleaves DNA into fragments close to the so-called restriction sites, a specific sequence of nucleotides [46].

[3] Alteration of a cell resulting from the direct uptake and incorporation of DNA coming from outside the organism through the cell membrane [47].

[4] Yeast Extract-Peptone-Dextrose

[5] Buffered Glycerol-complex Medium

[6] Buffered Methanol-complex Medium

different materials; then the various compounds present in the solution will interact differently with the column being able to be separated by the different rates going through it or by different conditions to elute[7]. Before starting the protein purification, it is needed to extract the protein from the cells and obtain the protein solution. The disruption of the cells can be done using lysozyme, or any other substance that breaks the cells walls and optionally, before or after, homogenize it by high-pressure with a French press which disrupt the cells by making them go through a small valve at a high pressure. Subsequently, to separate proteins from the disrupted cells ultracentrifugation is used in which the pellet (with the unwanted cells parts) and the supernatant (with the protein and other soluble substances)[8] are obtained. Also, by means of centrifugation, bigger compounds using a filter of a specific MWCO size can be separated.

The usual protocol for protein purification consists on numerous steps using different properties of the protein. We will enumerate the ones used in Chapter 4 which consist also in the most common protocol:

1. **Ion exchange chromatography.** A process which separates substances based on their charges using a charged resin, negatively or positively charged depending on the isoelectric point of the protein of interest. For instance, a column packed with a positively charged resin will interact strongly with a protein negatively charged in its surface while all the positive or neutral substance will go through. To be able to elute the protein there are two different ways; the first one is increasing salt concentration, such as NaCl, so that the Cl^- ions will compete with the protein. The second way is decreasing the pH of the solution resulting in a more positive charge on the protein, making it interact weaker with the column resin. The most common ion exchangers are listed in Table 2.1 [49].

2. **Size exclusion chromatography.** As its name indicates, this technique consists on separating the substances by their size. Depending on their ability to go through the gel beads, the substances will have to traverse more column volume and thus elute later in time. The columns are usually composed of dextran polymers (Sephadex), agarose (Sepharose) or polyacrylamide (Sephacryl or Biogel P) with different pore sizes [49].

3. **Reversed-phase chromatography or hydrophobic interaction chromatography.** In this case the column contains hydrophobic ligands attached

[7] to wash out (a substance) by the action of a solvent [48]

[8] After sedimentation, specially when using a centrifuge, the precipitate is referred as pellet and the solution above the solid is called supernatant

to the stationary phase, such as phenyl, that retains the hydrophobic substances with increasing retention times with higher polar mobile phases. For instance, you can decrease the concentration of salt (thus, decrease the polarity of the mobile phase) to decrease the retention time and elute faster the hydrophobic substance [50].

Name	**Type**	**Functional Group**
DEAE cellulose (Anion exchanger)	Weakly basic	(DEAE) Diethylaminoethyl
QAE Sephadex (Anion exchanger)	Strongly basic	QAE (Quaternary aminoethyl)
Q Sepharose (Anion exchanger)	Strongly basic	Q (Quaternary ammonium)
CM-Cellulose (Cation exchanger)	Weakly acidic	CM (Carboxymethyl)
SP Sepharose (Cation exchanger)	Strongly acidic	SP (Sulfopropyl)
SOURCE S (Cation exchanger)	Strongly acidic	S (Methyl sulfate)

Table 2.1: Most common ion exchangers columns.

Another purification process not enumerated above as it was not used in this work, but worth mentioning as it is one of the most popular purification techniques, is the addition of a poly-histidine tag (His-tag) at one extreme of the protein aminoacid chain (using a specific gene in the expression vector) and then use a metal-ion affinity chromatography to separate the His-tagged protein of other proteins. Thanks to the relatively small size of the His-tag is not essential to be eliminated before the crystaliization process starts.

Finally, the purified protein must be concentrated and it is usually done by ultrafiltration, which by means of centrifugation and a membrane with a specific MWCO, separates the solution and small molecules of the protein. Also, it is important to evaluate all the purification process and this is usually done by running a SDS-PAGE, an electrophoresis method[9] that allows protein separation by mass [51], see Figure 2.1. In addition, it is necessary to quantify the amount of protein obtained before trying to crystallize it. Approximately, a concentration of 10 $mg{\cdot}ml^{-1}$ of protein is needed as a rule of thumb, but the true appropriated concentration for crystallization depends a lot on the sample (i.e. the protein characteristics, specially solubility). The quantification is done by looking at the absorbance of light at 280 nm.

[9]the motion of dispersed particles relative to a fluid under the influence of a spatially uniform electric field

Our purified and concentrated protein usually contains glycans attached to it. These glycans difficult the task of obtaining a high-quality crystal and sometimes the protein needs to be deglycosylated prior the crystallization procedure. This can be done with different enzymes such as endoglycosidase H or Peptide-N-Glycosidase F. Then, a second round of purification process needs to be done in order to get rid of these new enzymes placed in our protein solution besides the small amount of protein not deglycosylated and also, the removed glycans.

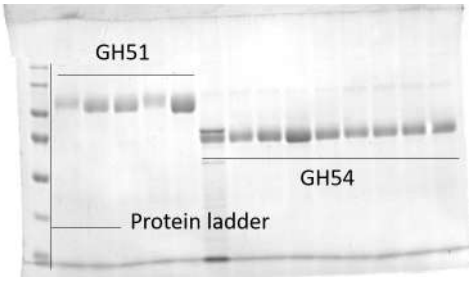


Figure 2.1: SDS-PAGE done for two different weighted enzymes, GH51 and GH54, the later being studied in Chapter 4.

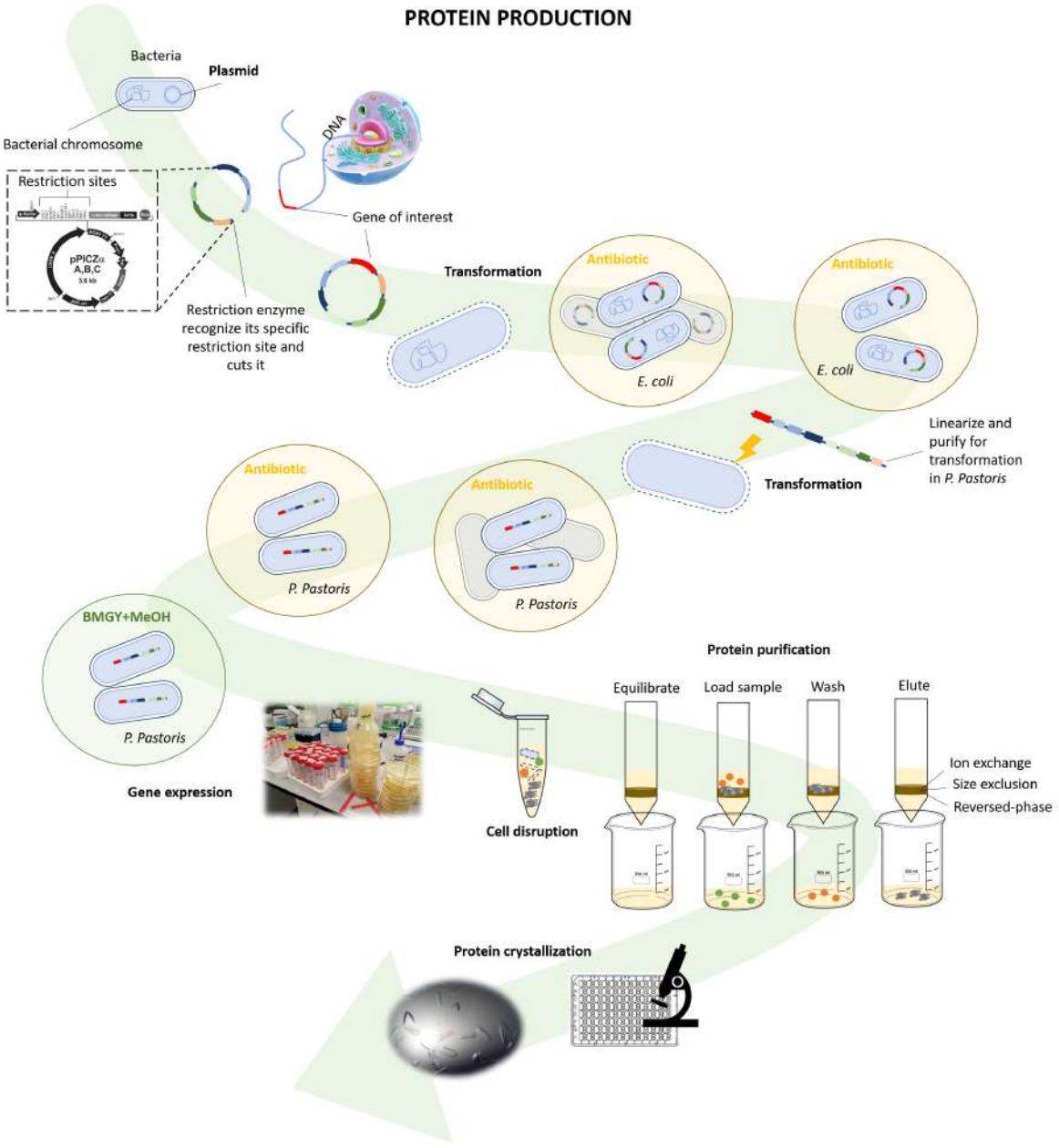


Figure 2.2: Schematic representation of molecular cloning, protein expression, purification and crystallization in *P. Pastoris*. The crystals shown correspond to EndoH-deglycosylated AkAbfB, from GH54 family. The vector pPICZα, used in Chapter 4, is shown in black and white from ref. [52]

2.1.2 X-ray crystallography

Firstly, there is the arduous task of finding an optimal crystallization condition. Nowadays, there are at our disposal a variety of commercial screens optimized for the crystallization of proteins [53,54] in which various ingredients mixes are preassembled and thanks to liquid-handling robots the process of dropping a very small amount of our protein to each different conditions well is faster and easier to manipulate. After this, the crystallization screens are usually placed in a robot machine which controls using a camera the crystal growth. Once there is some crystal growth (which can vary from some days to weeks or months), starts the process of crystal optimization usually performed by designing screens based on the initial conditions where crystals were grown [55], pretending in this way to obtain a higher number of crystals but also more ordered and bigger.

Secondly, with a big and ordered crystal, the process called cryocooling is done. This technique consists on placing the crystal in a cryosolution, a solution which contains different types of alcohols, salts or oils to prevent ice formation. Then these crystals are freezed by using liquid nitrogen which maintain them at 100K. All these steps are needed as the beam lines used in the synchrotrons to diffract the crystal are so intense that an unprotected crystal would evaporate before the diffraction experiment would finish.

Finally, the loop containing the cryocooled crystal is placed in-line with the X-ray beam and rotates in all its possible angles so that all the reflections across the crystal are collected.

2.1.3 Solving and refining crystal structures

The last step consists on recollecting the data from the X-ray experiment. But it is not as easy as it seems. What we recover from a crystal after a X-ray beam it is what is called the **diffraction pattern**; a set of spots (or more technically, the reflections at different intensities) corresponding to the resulting diffraction[10] of the waves that go through the crystal and are scattered and interfered with each other deviating from the central beam. We obtain a large number of diffraction patterns to recover all the possible orientations of the crystal. X-ray is the beamline of choice due to its wavelength, in the Å range, $\approx$ of 10^{-10} m, which can be scattered by the electrons present in the chemical bonds of the molecules. The interference of the diffracted X-rays depend on all the atoms present in the crystal, i.e., the shape and symmetry of the unit cell define the directions of the diffracted beams and the location of all

[10]Bragg diffraction: Consequence of the interference between waves reflecting from different crystal planes.

atoms in the cell define their intensities. When there is a larger unit cell, we obtain more diffracted beams (reflections or spots in the diffraction pattern) which help in increasing the accuracy of the final estimation of the averaged reflection intensity. What is more, the position of each atom is responsible for the intensities of all the reflections or in other words, the intensity of each reflection depends on the positions of all atoms in the unit cell. This fact makes impossible to solve only a small region of the crystal as it is needed to model the whole system to be able to obtain it [56]. Each reflection has two main properties, its amplitude and its phase. Thus, once the amplitude and phase of all the reflections are known, as they are scattered by the electrons of all the atoms of the unit cell, we would be able to reconstruct the electron density of the protein present in the crystal.

To obtain the distribution of the electrons present in our crystal it is necessary to calculate the Fourier transform (FT) of the structural factors or F, which correspond to the amplitudes and phases of the scattered waves. The FT let us going from the reciprocal space, the diffraction pattern, to the direct space; the electron density of the atoms present in our protein. Amplitude can be recovered from the intensities of the reflections but there is no direct way to obtain the phases of the reflections and we face what is called the phase problem. There are several methods to obtain the phases in an indirect way, listed below:

1. **Molecular replacement.** Also abbreviated as MR, it is based on finding a protein model resembling the protein of interest (more than 20 % of sequence similarity) which it is translated and rotated to fit the new unit cell, then obtain the phases of the protein model and finally, build the first electron density map of our protein. Thanks to the increasing number of proteins diposited in PDB [25] the probability of finding an adequate model has also been increased. It is the most common method to solve the phase problem, in 2013 almost 60% of all the structures deposited in PDB where solved by MR [57].

2. **Multiple/Single Isomorphus replacement.** Abbreviated as MIR/SIR, it is based on comparing two isomorphic[11] structures, a native one and another of the same protein but soaked or co-crystallized with heavy-atoms such as Hg. Diffraction patterns are measured from both the native and the heavy-atom crystals, and the structure factors for the two crystals are compared. If a structure factor from the heavy-atom crystal is significantly stronger than the corresponding structure factor from the native crystal, then the scattering from the heavy atoms must be interfering constructively with the scattering

[11] Crystals are described as isomorphous if they are closely similar in shape [58]

from the rest of the crystal. In this case the (unknown) phase of the native structure factor must be close to the (known) phase of the scattering from the heavy atom alone. Similarly, if a structure factor from the heavy-atom crystal is significantly weaker than the corresponding structure factor from the native crystal, then the scattering from the heavy atoms must be interfering destructively and the phases must be offset by nearly 180°. These two cases would be the ideal ones and usually, the scattering lies between both extremes. To resolve this problem, it is necessary to use multiple crystals with different heavy-atoms bonded to different places on the protein [59].

3. **Multiple/Single wavelength Anomalous diffraction.** MAD/SAD is nowadays the method of choice for novel protein structures. It can be seen as a similar method compared to MIR, but in this case compares a single crystal in which the scattering is measured at the absorption edge[12] and at a wavelength distant from it, or a set of several wavelengths to provide better phase estimation. The wavelength tuning can be done thanks to the use of synchrotrons and the atoms anomalous scattered are usually Se placed instead of S in methionine residues, even though other heavy atoms can be used. The only drawback of this method is that the experiment requires a very careful measurement of small differences in the diffraction pattern and therefore, it must be performed in a high resolution crystal [59].

After solving the waves phases by either experimentally derived or MR-based methods only approximate models are build and they require further **refinement**. It is an iterative process which goes on until the best fit is achieved between the observed diffraction pattern and the one calculated by FT from the model structure. The iterative process includes rounds of automated optimization and also manual corrections to avoid the former one to get stuck in a local minimum instead of the global one. The parameters optimized are the x, y, z coordinates of the atoms and a parameter reflecting their smearing in space known as atomic displacement parameter (ADP) or B-factor. Refinement is a tremendous task as, for example, a molecule of 20 KDa would take approximately 6000 parameters to refine [60]. Frequently, the number of reflections obtained in the diffraction pattern are not enough and thus, some stereochemical restraints are applied, being careful to do not use very tight restraints which could suppress the information coming from the experiment.

[12] each atomic type has a few absorption edges around which the scattering varies rapidly (in amplitude and phase) with wavelength. By varying the wavelength around the absorption edge for an atomic type, the contribution from those atoms to the total scattering can be varied.

The final step would be structure **validation**, which consist on checking the interpretation of the data into the atomic model (model-to-data validation) and also to check the consistency of the model with known physical and chemical properties; such as making use of Ramachandran plots, side-chain rotamers, etc. Also it is important to validate the ligands present in the structure as at lower resolutions it is very difficult to solve them properly, specially carbohydrates [61, 62].

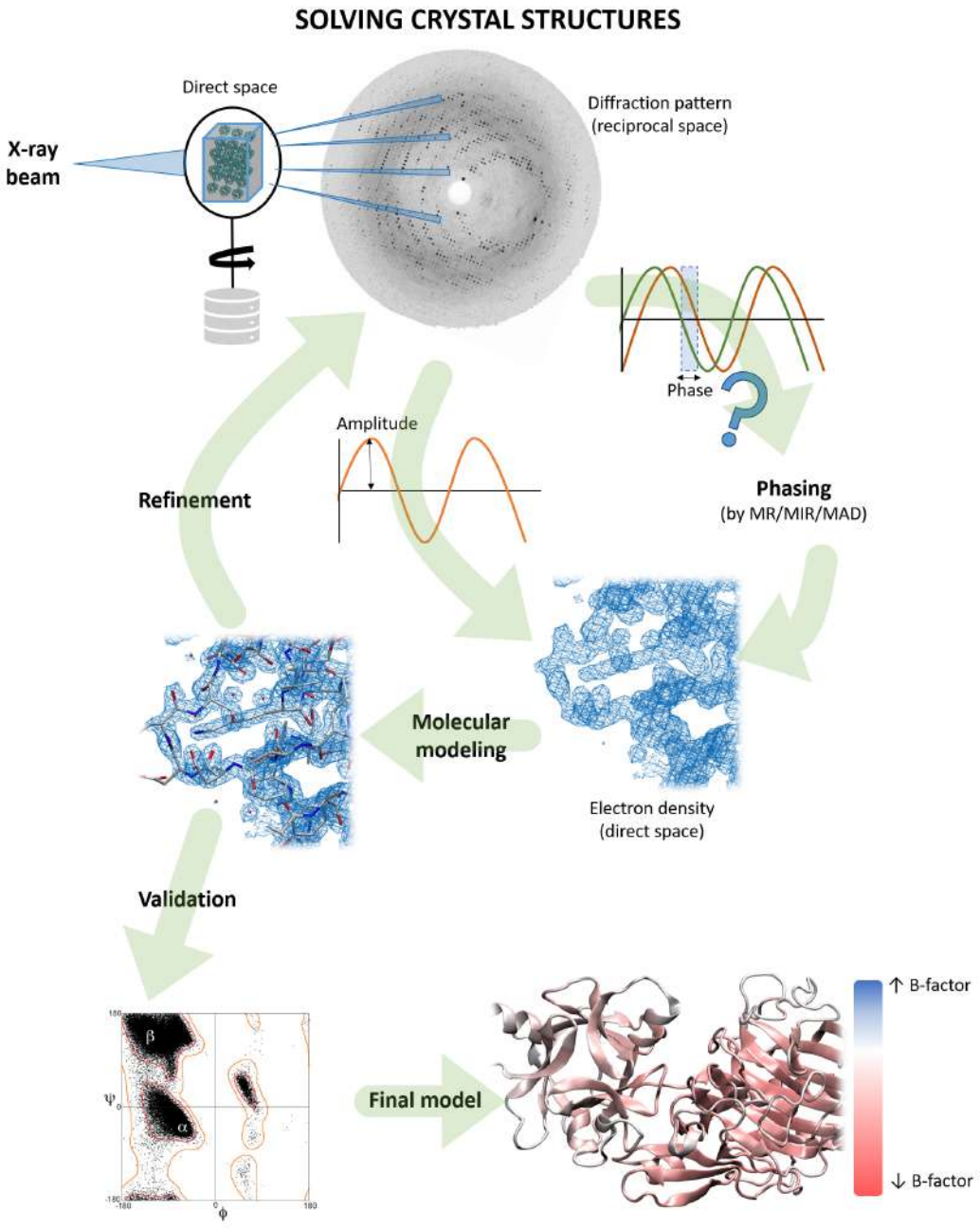


Figure 2.3: Schematic representation of solving a macromolecular structure such as a protein by a X-ray experiment.

2.1.4 Protein ensembles and single proteins; experimental vs computational protein structures

It is important to bear in mind the fact that when solving a protein from a crystal structure, it is not a single molecule but instead a space and time average over the whole crystal volume and the duration of the X-ray experiment. Thus, the diffraction pattern is an average of possible **static disorders**: different conformations adopted by some regions of the protein in various unit cells; and **dynamic disorders**: the atomic thermal motion, mobility of whole molecular fragments and crystal lattice vibrations. The error of this averaged structure coming from an ensemble of structures is enclosed in the above mentioned B-factor or ADP.

On the other hand, when we simulate proteins we use a single one and not an ensemble averaged over space and time. To obtain significant results it is needed to run dynamic simulations to be able to recover all the possible conformations of the system by making use of the ergodic hypothesis[13]. In this case, the factor which enclose the "smearing" of atoms would be the root mean square fluctuation (r.m.s.f.) of the atoms.

The experimental and computational error parameters can be related by:

$$rmsf_{exp} = \sqrt{\frac{3B}{8\pi^2}} \tag{2.1}$$

$$rmsf_{calc} = \sqrt{\langle [r_i(t) - \langle r_i(t) \rangle]^2 \rangle} \tag{2.2}$$

2.2 Computational tools

From experiments we obtain an averaged but static picture of a specific protein and usually, as explained in Chapter 1 (section 1.3.1, page 12), in a non-competent picture due to the modifications necessarily done to be able to crystallize the MC; or with some missing loops which were not able to observe due to its high mobility. Then, if we want to study its competent state (reverting the mutations done) and its full completeness (modeling missing loops) we need to make use of the computational tools available nowadays. Also, we want to see the dynamics of that protein to be able to discern how this protein actually works. Molecular dynamics is based on describing the movement of the atoms thanks to mathematical expressions with equations of motion and the corresponding energy at each step to be able to integrate a trajectory of each atom along time. There are different ways to treat the movement

[13] It states that by running an infinite time simulation it would be possible to recover the ensemble average by a time average.

and energy of these particles; basically using quantum or classical methods. The particles motion can be described as point particles (classical description) or as probability waves (quantum description). The system energy depends on the atom interactions and they can be described classically, by means of potential force fields or quantically, in which some approximations are needed as the computational cost needed is extremely huge.

2.2.1 Classical Molecular Dynamics

The propagation of the particles using this method is based on the Newton's second law of motion, which is;

$$\overrightarrow{F_i}(t) = m_i \overrightarrow{a_i}(t) \Rightarrow -\frac{\delta V(\overrightarrow{r})}{\delta r_i} = m_i \frac{\delta^2 \overrightarrow{r_i}(t)}{\delta t^2} \tag{2.3}$$

The equation of motion is solved approximately using numerical algorithms such as the Verlet integrator [63] and the atomic propagation takes the following form;

$$\overrightarrow{r}(t + \Delta t) = 2\overrightarrow{r}(t) - \overrightarrow{r}(t - \Delta t) + \overrightarrow{a}(t)\Delta t^2 \tag{2.4}$$

where Δt is the time step parameter, its choice being a counterbalance between accuracy and computational cost. A small time step gives a better integration quality, but more integration steps will be required for the same dynamics length. A good rule of thumb is choosing a time step one tenth of the fastest molecular motion which is usually the O-H bond stretching.

2.2.1.1 Energy calculation, use of force fields

To solve the potential energy parameter of equation 2.3 which corresponds to the interatomic interactions we make use of the so-called force-fields, a set of energy functions and parameters that give the potential energy for a given atomic position. One possible function is the one used in the Amber program and has the following form [64];

$$V^{FF} = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} 1/2V_n[1 + cos(n\phi - \gamma)] + \sum_{i<j}^{atoms} (\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{+}} + \frac{q_i q_j}{\epsilon R_{ij}}) \tag{2.5}$$

There are two main drawbacks of classical MD. Firstly, FF use parameters developed from data on small molecules into much larger entities, thus, the effect of the molecular environment is neglected. These parameters are obtained by fitting

a potential function with experimental data and high level quantum chemical calculations, which have the inconvenient to do not be always accessible. Secondly, electrons are not considered, but instead point charges, localized in each nuclear center, are used to reproduce the electrostatic properties of each molecule. Simple functions (e. g. Hooke's law) are used to describe processes such as the stretching of the bonds; due to these approximations electron density is not explicitly considered and so, chemical reactions can not be performed.

2.2.2 Quantum Molecular Dynamics

To overcome the drawback of not being able to study chemical reactions, we would need to include explicitly the electrons in the molecular dynamics formalism. The equation that follows the evolution of the system in a quantum way, is the time dependent Schrödinger equation [65];

$$i\hbar\frac{\delta}{\delta t}\psi(\overrightarrow{r_i},\overrightarrow{R_I},t) = \widehat{H}\psi(\overrightarrow{r_i},\overrightarrow{R_I},t) \tag{2.6}$$

in where i is the imaginary number, the constant Planck is $\hbar$, ψ corresponds to the time dependent wave function for i electrons and I nuclei from which a probability of finding a particle in a particular space region at time t can be derived from; and $\widehat{H}$ is the Hamiltonian[14] of the system. $\widehat{H}$ outputs the energy of the system and in a non-relativistic way can be expressed as;

$$\widehat{H} = -\frac{\hbar}{2}(\sum_{I=1}^{N}\frac{\nabla_I^2}{m_I} + \sum_{i=1}^{n}\frac{\nabla_i^2}{m_i}) + V_{n-e}(\overrightarrow{r_i},\overrightarrow{R_I}) \tag{2.7}$$

in where the first term corresponds to the kinetic energy of the nuclei (N) and electrons (n) and the second term to the potential energy which can be divided in;

$$V_{n-e}(\overrightarrow{r_i},\overrightarrow{R_I}) = \frac{1}{4\pi\epsilon_0}(\sum_{i<j}\frac{e^2}{|\overrightarrow{r_i}-\overrightarrow{r_j}|} - \sum_{I,i}\frac{e^2 Z_I}{|\overrightarrow{R_i}-\overrightarrow{r_i}|} + \sum_{I<J}\frac{e^2 Z_I Z_J}{|\overrightarrow{R_I}-\overrightarrow{R_J}|}) \tag{2.8}$$

where the first term corresponds to the electron-electron electrostatic repulsion, the second term to the nuclei-electron electrostatic potential and the last term to the nuclei-nuclei electrostatic potential. Using the Born-Oppenheimer approximation we can assum that the motion of the nuclei and electrons in a molecule can be treated separately. Unfortunately, these equations are limited to small systems and some approximations need to be done for larger systems. The following methods can be approached:

[14] an operator corresponding to the sum of the kinetic energies plus the potential energies for all the particles in the system (this addition is the total energy of the system in most of the cases under analysis)

1. **Ehrenfest MD (EMD).** The first simplification that can be done in the Schrödinger time-dependent equation is to assume the nuclei as point particles and then treat them in a classical way, then the equations to solve are the followings;

$$M_I \frac{\delta^2 \overrightarrow{R_I}(t)}{\delta t^2} = -\nabla_I \langle \psi_0(\overrightarrow{r_n}, t) | \widehat{H}_e | \psi_0(\overrightarrow{r_n}, t) \rangle \tag{2.9}$$

$$i\hbar \frac{\delta \psi_o(\overrightarrow{r_n}, t)}{\delta t} = \widehat{H_e} \psi_o(\overrightarrow{r_n}, t) \tag{2.10}$$

   As it can be seen, equation 2.9 is basically the Newton's equation of motion (found in equation 2.3), but now calculating the potential energy in an *ab initio* way, taking into account the electrons [65].

2. **Born-Oppenheimer MD (BOMD).** A second approach is the BOMD in which the nuclei are also treated as point particles moving in a classical way but now, in an effective potential which is given by the Born-Oppenheimer potential energy surface obtained by solving the time-independent electronic Schrödinger equation at the given nuclear configuration $\overrightarrow{R_I}$(t). Contrary to the EMD where the wavefunction is minimized initially and the electronic system will stay in its respective ground state minimum and propagated as the nuclei move; in BOMD the minimum of the electronic system has to be reached at each time step. It has the following form [65];

$$M_I \frac{\delta^2 \overrightarrow{R_I}(t)}{\delta t^2} = -\nabla_I min_{\psi_0(\overrightarrow{r_n}, t)} \langle \psi_0(\overrightarrow{r_n}, t) | \widehat{H}_e | \psi_0(\overrightarrow{r_n}, t) \rangle \tag{2.11}$$

$$E_0 \psi_0(\overrightarrow{r_n}, t) = \widehat{H_e} \psi_o(\overrightarrow{r_n}, t) \tag{2.12}$$

3. **Car-Parrinello MD (CPMD).** This technique combines the advantages of both Ehrenfest and Born-Oppenheimer molecular dynamics in an optimal way. The basic idea of Car-Parrinello Molecular Dynamics (CPMD) is taking advantage of the quantum-mechanical adiabatic time scale separation of fast electronic (quantum) and slow nuclear motion (classical) [66]. Using this idea, a Lagrangian was proposed where both electrons and nuclei are included;

$$L_{CP} = \sum_{I=1}^{N} \frac{1}{2} M_I \frac{\delta^2 \overrightarrow{R_I}}{\delta t^2} + \frac{1}{2} \mu_e \sum_{i=1}^{n} \langle \frac{\delta \phi_i(\overrightarrow{r}, t)}{\delta t} | \frac{\delta \phi_i(\overrightarrow{r}, t)}{\delta t} \rangle - \langle \psi_0 | \widehat{H_e} | \psi_0 \rangle + \sum_{j=1}^{n} \Lambda_{ij} \phi_j(t) \tag{2.13}$$

   where the first two terms correspond to the kinetic energy, the third term corresponds to the potential energy and the last one is a constraint imposed to assure the orthonormality of the orbitals. The equations of motion are derived as;

$$M_I \frac{\delta^2 \overrightarrow{R_I}(t)}{\delta t^2} = -\nabla_I \langle \psi_0(\overrightarrow{r_n}, t) | \widehat{H}_e | \psi_0(\overrightarrow{r_n}, t) \rangle \tag{2.14}$$

$$\mu_e \frac{\delta^2 \phi_i(t)}{\delta t^2} = -\widehat{H_e}\phi_i(t) + \sum_{j=1}^{n} \Lambda_{ij}\phi_j(t) \tag{2.15}$$

where μ_e is the fictitious mass assigned to the orbitals degrees of freedom. Thus, in a CPMD simulation both electrons and nuclei evolve simultaneously, where the electrons are following adiabatically the nuclear motion very close to their instantaneous ground state [67]. Basically, the main issue is how to achieve adiabatic separation of the electronic and nuclear subsystems and how to control it to guarantee a correct description of the system. This can be done if the frequency difference, ω_{min} - ω_{max}, between both subsystems is relatively large. The lowest possible electronic frequency is directly proportional to the E_{gap} (energy difference between the lowest unoccupied, LUMO, and the highest occupied, HOMO, orbital) and inversely proportional to the fictitious mass, μ_e.

$$\omega_e^{min} \alpha \{\frac{E_{gap}}{\mu_e}\}^{1/2} \tag{2.16}$$

According to equation 2.16 the only parameter to control adiabatic separation is the fictitious mass, μ_e. Decreasing μ_e increases the minimum frequency of the electronic subsystem (as the energy gap depends only on the physics of the system and thus cannot be changed) and then adiabiacity is increased. However, a decrease of μ_e also leads to an increase of the maximum electronic frequency according to;

$$\omega_e^{max} \alpha \{\frac{E_{cut}}{\mu_e}\}^{1/2} \tag{2.17}$$

where E_{cut} is the largest kinetic energy in an expansion of the wave function in terms of a plane wave basis set (explained in section 2.2.2.1). Also, the time step is inversely proportional to the highest frequency in the system, and thus limits the largest time step that is possible.

$$\Delta t^{max} \alpha \{\frac{\mu_e}{E_{cut}}\}^{1/2} \tag{2.18}$$

Then, a compromise have to be achieved in order to control parameter μ_e. For large-gap systems (most molecules), values of μ_e = 500-1000 a.u. with time steps of about 5-10 a.u. (0.12-0.24 fs) ensure a good control of adiabaticity, but some tests need to be done to each different system to ensure the optimal values [68].

4. **Benefits and drawbacks of each *ab initio* MD method.** The main reason to choose between different *ab initio* molecular dynamics methods is basically the efficiency in terms of computer time for a specific problem. This can reside in the time step one has to choose for the dynamics which depends

on the fastest motion considered as degree of freedom. In the EMD the time step is dictated by the intrinsic dynamics of the electrons as described by the time-dependent Schrödinger equation, thus, the largest possible time step is that which allows us to integrate properly the electronic equations of motion. On the contrary, in the BOMD there is no electron dynamics whatsover involved in the equations of motion as the electronic subsystem is treated as time-independent. In this case, the time step is given by the nuclear motions which is much slower than the electronic one allowing larger time steps. The only drawback is the fact that the electronic structure has to be solved self-consistently at each molecular dynamics step whereas this is avoided in EMD. Then, EMD and CPMD have a smaller time step than BOMD. Between the two methods which treat electrons in a time-dependent manner, EMD and CPMD, the latter has a time step of one order of magnitude higher due to the time step dependence on the fastest electronic motion. In the former corresponds to equation 2.17 whereas in the latter can be estimated within a plane wave expansion as $\omega_e^E \alpha \, \mathrm{E}_{cut}$. When the electronic fictitious mass is introduced, the time step can be increased. But the time step size it is not the only parameter that will increase or decrease the used computer time. In the BOMD the iterative optimization of the wave function is done at each time step but only once in CPMD, this process becoming harder the larger the system is. Depending on the system, the computational effort can be the same for the two methods (CPMD and BOMD) or even smaller for CPMD even though the electrons are treated dynamically.

2.2.2.1 Energy calculation

To calculate the energy needed to propagate the equation of motions in a quantum way we "only" need to solve the stationary Schrödinger equation which has the following form;

$$\widehat{H}\psi(\overrightarrow{r_n}) = E\psi(\overrightarrow{r_n}) \tag{2.19}$$

The solution to this equation can be approached by different techniques some of them called wave function methods which derive from first principles to solve either the Hamiltonian or the wave function, from lower to higher level of accuracy there is Hartree-Fock (HF) theory, Møller-Plesset perturbational theory or coupled cluster methods. The main disadvantage of this kind of wave function methods is its dependency on a high number of variables for a n electron system, corresponding to the three spatial and one spin coordinates for each electron. Thus, the computational cost increases exponentially with the number of electrons and making it virtually impossible to solve for large systems [69].

Density Functional Theory (DFT) is based on the proof by Hohenberg and Kohn [70] that the ground state electronic energy is determined completely by the electron density ($\rho(\overrightarrow{r})$) which only depends on three variables for the whole system, the spatial x, y and z coordinates, overcoming the bottleneck of the exponential "wall" found in wave function approaches. So, in theory, we would only need to find a functional[15] which takes us from $\rho(\overrightarrow{r})$ to E. We will further develop this technique as it is the one used along this work.

$$E = F[\rho(\overrightarrow{r})] \tag{2.20}$$

The DFT theorem was further developed by Kohn and Sham (KS) in which the intractable many-body problem of interacting electrons in a static external potential is reduced to a tractable problem of non-interacting electrons moving in an effective potential [71]. Then the energy functional is approximated to;

$$E[\rho(\overrightarrow{r})] = Ts[\rho(\overrightarrow{r})] + V_{N-e}[\rho(\overrightarrow{r})] + V_{e-e}[\rho(\overrightarrow{r})] + E_{XC}[\rho(\overrightarrow{r})] \tag{2.21}$$

where the first term corresponds to the kinetic energy assuming non-interacting electrons, the second term corresponds to the nuclei-electron attractions, the third to the Coulomb interactions between electrons and the last one to a correction term which contains the contributions from exchange and correlation. The density of the system is written with a technical construct called the KS-orbitals which ensures that the true electron density can be derived even though the kinetic energy and Coulomb energies were derived from an approximated non-interacting system.

$$\rho(\overrightarrow{r}) = \sum_{i=1}^{N} |\psi_i(\overrightarrow{r})|^2 \tag{2.22}$$

The drawback in DFT is the unknown $\mathrm{E}_{XC}[\rho(\overrightarrow{r})]$ which is derived with different approximations. A popular way to order them is by the *Jacob's ladder* [72] seen in Figure 2.4 which goes from the less accurate (Hartree world) to the most accurate (Chemical accuracy heaven) E_{XC} approximations.

As we stated previously, we use the KS-orbitals to solve numerically the energy functional from the electron density. These orbitals are functions that can be described in various ways, using **basis sets** or **plane waves**. The former is a function centred on the nuclei, while the latter is based on expanding the orbitals through a defined space using plane waves (PW);

$$\psi_i(\overrightarrow{r}) = \frac{1}{\omega^{1/2}} \sum_{G}^{G_{max}} C_G e^{i\overrightarrow{G_r}} \tag{2.23}$$

[15] A function whose argument itself is a function. It is a device which assigns a number, e.g. energy, to a function, e.g. electron density.

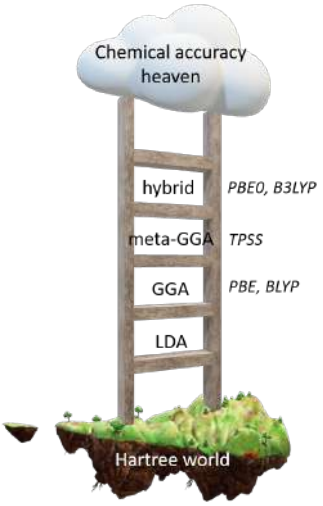


Figure 2.4: DFT Jacob's ladder. In cursive in the right side of the picture there are some examples for GGA, meta-GGa and hybrid functionals.

where ω is the volume of the cell and G is the plane wave momentum. The corresponding basis sets are denoted by an energy value, E_{cut}, which is related to the maximum G value in the PW expansion, G_{max}. One advantage is that the calculations are not affected by the basis set superposition error (BSSE, found in the basis sets approach), but otherwise, the whole space is equally described, and this means paying a price for the calculations of non-extended systems (e.g. molecules), in which the electron density is largely concentrated around the atoms. Because describing core electrons with a PW basis set would require even a higher number of PWs, the effect of the core electrons is described with pseudopotentials (an effective potential constructed to replace the atomic all-electron potential and valence electrons are described by pseudo-wave functions with significantly fewer nodes than the all-electron wave function) [73].

2.2.3 A mix in between quantum and classical, QM/MM

The main problem of using quantum MD methods is its size restriction, which is around a few hundred atoms; impossible to be able to simulate most of biological systems. Classical MD, even though capable of simulating huge systems, it describes them less accurately and with the impossibility of electron rearrangement, such as chemical reactions. Thus, combining both methods, we can overcome the size and electron rearrangement restrictions. In 2013, Karplus, Levitt and Warshel were awarded with a Nobel Prize for "the development of multiscale models for complex chemical systems" [74]. Thus, the award was for the invention of the method called

QM/MM, a combination between quantum mechanics and molecular mechanics.

In this method, there is a quantum region, in which there is an interest of its electronic rearrangement, and a classical region. Both have a myriad of different quantum methods and force fields to choose which will obviously affect on the results obtained, but also you can choose different ways on how to treat the interaction between both regions. There are mainly two schemes to calculate the whole QM/MM energy of the system; one called the **subtractive** scheme in which the whole system energy is calculated in a classical way, then the quantum region is added calculated with a quantum method and finally the calculated classically energy of the quantum region is subtracted;

$$E = E^{MM}(QM + MM) + E^{QM}(QM) - E^{MM}(QM) \tag{2.24}$$

But the most common one, and the one used in this work, is using the **additive** scheme which the total energy system is the sum between the quantum region (using quantum methods), the classical region (using force fields) and the energy coupling both regions;

$$E = E^{QM} + E^{MM} + E^{QM-MM} \tag{2.25}$$

The interactions coupling QM and MM degrees of freedom can be grouped in bonded and non-bonded. When the QM-MM boundary divides a covalent bond there are two suitable options to saturate the QM region within a MD formalism: either "cap" the QM site with a link atom (frequently a hydrogen atom), mostly known as **capping hydrogens** or parametrize a boundary atom described by a **pseudopotential**. The electrostatic interactions are dealt at different levels of sophistication called as **mechanical embedding**, **electrostatic embedding** and **polarized embedding** [75]. In this work we are using the QM/MM scheme implemented in the CPMD [76] code which corresponds to an electrostatic embedding based on a multilayer approach [77]. There are three different layers called NN, MIX and ESP regions. The Coulombic electrostatic interaction between the QM atoms and the closest MM atoms (names as NN atoms), and the atoms in the MIX region with a charge larger than $0.1e_0$, are treated exactly;

$$\widehat{H_{NN}} = \sum_{i=1}^{NN_{atoms}} q_i \int_0^{\inf} d\overrightarrow{r} \frac{\rho(\overrightarrow{r})}{|\overrightarrow{r} - \overrightarrow{r_i}|} \tag{2.26}$$

For classical atoms in the previous mentioned region with a charge smaller than $0.1e_0$ and all MM-atoms included between the MIX and ESP boundary regions are coupled to the QM system treating them as point charges;

$$\widehat{H_{ESP}} = \sum_{i=1}^{ESP_{atoms}} \sum_{j=1}^{QM_{atoms}} \frac{q_i Q_j^{RESP}}{|\overrightarrow{r_i} - \overrightarrow{r_j}|} \tag{2.27}$$

The longer range electrostatic interactions are treated via a multipole expansion. It is important to choose the size of this layers between accuracy and computational cost as a big NN or ESP layer can increase exponentially the time employed for a MD step.

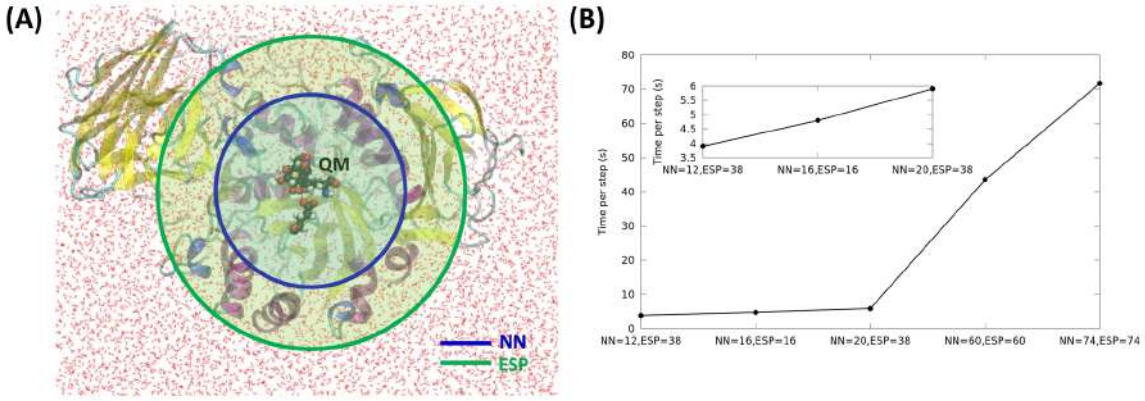


Figure 2.5: **(A)** Electrostatic regions illustrated for a retaining GH. Values correspond to R_{NN}=20 a.u. and R_{ESP}=38 a.u. The water box has been reduced for visualization purposes. **(B)** Computational cost for various values of R_{NN} and R_{ESP} (calculations performed using 80 Intel SandyBridge processors). Image modified from Raich *et. al.* [78]

.

2.2.4 Exploring more regions of the system space, metadynamics

When one performs molecular dynamics simulations, the main aim is to sample all the phase space of the system. To be able to achieve it, a very large timescale has to be computed due to many relevant systems are characterized by several metastable states separated by high barriers that lead to kinetic bottleneck [79]. The transition between different states, thus, become rare events taking place often at much longer timescales that one can simulate in practice. Then, these kinetic bottlenecks slow down or even make impossible the sampling of the whole phase space. The equation of probability of finding a particular system in any state is [80];

$$p(\overrightarrow{S'}) = 1/Q \int d\overrightarrow{R} d\frac{\delta \overrightarrow{R}}{\delta t} exp(-\widehat{H}(\overrightarrow{R}, \frac{\overrightarrow{\delta \overrightarrow{R}}}{\delta t})/K_B T)\delta(\overrightarrow{S}(\overrightarrow{R}) - \overrightarrow{S'}) \tag{2.28}$$

What is showing equation 2.28 is the fact that the probability to observe the system in a TS during a simulation decreases exponentially with the height of the TS barrier. Taking into account that chemical barriers are typically one or two order of

magnitudes higher than the average kinetic energy ($K_BT \approx 0.6$ kcal/mol) it results in a very low probability. Advanced sampling methods have been developed to be able to enhance sampling of the phase space based on the definition of some crucial parameters or collective variables. One branch of these methods is based on an introduction of a bias potential to be able to cross the high barriers between these metastable states. In this work, the method of choice has been metadynamics [81]. **Metadynamics** lowers the probability of revisiting regions in the low-energy reactant and product states such that it would be easier for the system to cross high energy barriers between metastable states. This can be achieved by adding a bias potential to the effective potential coming from the electron system;

$$V_{total}(t) = V_e + V_{bias}(t) \tag{2.29}$$

The bias potential has the form of Gaussian-shaped functions which are added after a specified time interval, δt.

$$V_{bias}(s,t) = \sum_{t_i} W exp(\frac{-|s - s(t_i)|^2}{2\Delta S_i}) \tag{2.30}$$

where W is the gaussian height, Δs is its width and s is the position of the corresponding collective variable. This bias potential discourages the system from revisiting a specific point. The history-dependent potential builds up until it counterbalances the underlying free energy well, thus, being the system able to cross the high barrier and visit other metastable states of the phase space. This procedure is repeated until all the minima are filled with Gaussian potential "hills" and the system can move barrier-free among the different states. The free energy surface can be obtained by

$$V_{bias}[t \to \inf](s,t) \approx -\Delta G(s) \tag{2.31}$$

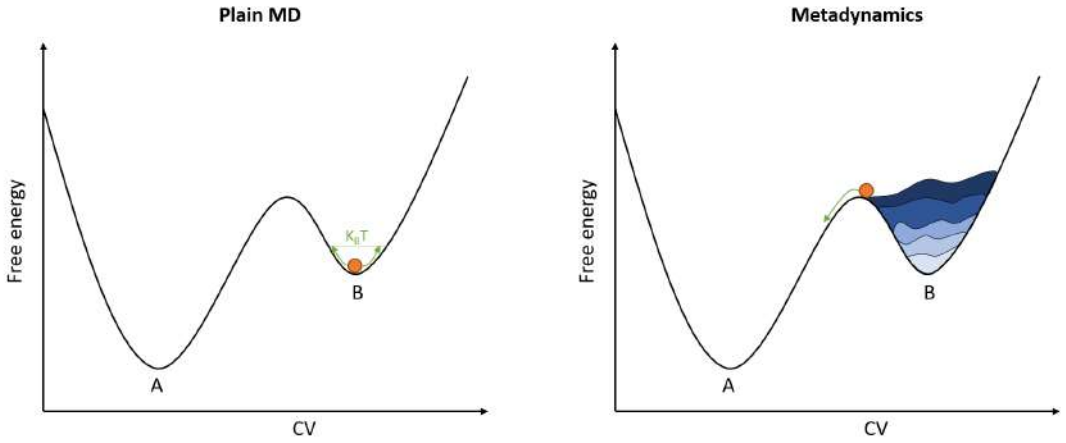


Figure 2.6: Schematic representation of a fictitious particle in plain MD (left) and metadynamics (right). The fictitious particle cannot scape from state B when using plain MD while when metadynamics is applied it will move through the "hills" (in blue) added with bias potential, V_{bias}, crossing high barriers between metastable states.

There are two issues to take into account in a metadynamics simulation (or any other advances sampling technique). The first one is the fact that is not trivial to decide when to stop a simulation. As a general rule, it should be stopped when the motion of the CVs becomes diffusive in the region of interest, which means that local convergence has been achieved. Some authors have argued that one recrossing over the transition state (TS) is enough to converge a metadynamics simulation of a chemical reaction [80]. In our experience, the free energy error associated to lack of recrossing is less than 1.5 kcal/mol and in general the molecular mechanism remains unchanged. The second issue is the the choice of an appropriate set of CVs for the system of interest. A rational choice has to be made in order to obtain a good physical description of the process, otherwise the computational time to convergence can be prohibitively large. Optimal CVs should fulfill the following requirements [82]:

1. They are explicit functions of the atomic positions.
2. They are able to distinguish the different states of the system.
3. They should include all the slow modes of the system. Slow modes are referred to those variables that cannot be sampled in the timescale of a simulation. It is expected that all the other fast variables will adjust rapidly to the evolution of the slow variables.
4. There should be a limited number of CV as the computational cost increases exponentially with the number of CVs.

2.2.5 QM/MM metadynamics simulation protocol

To be able to study the chemical reaction mechanisms of enzymes we would need to do a QM/MM metadynamics simulation. QM/MM because we want to simulate a big biological system such as an enzyme solvated by water molecules which would be impossible to calculate by quantum methods alone but also, we want to study the electron rearrangement, some bonds are being broken and formed and, thus, we cannot use classical MD alone either. Metadynamics is used to enhance the sampling of the cleavage and formation of the bonds which are in a chemical time scale not possible to simulate with QM/MM alone with the nowadays computer power. It is important to carefully set all the previous steps to finally do not reach to a simulation artifact. All the consideration that have to be taken into account are explained below.

First, a reliable crystallographic structure has to be chosen. This was enough discussed in Chapter 1 and section 2.1. Then, the proper modifications need to be done to recover a competent state of the enzyme. Afterwards, the protein needs to adapt at these modifications and explore different possible states found in room temperature. Nowadays, these steps need to be performed with a classical MD method, as the time scales needed to reach thermal equilibrium (nanoseconds) are far from the time scales that are affordable with *ab initio* and QM/MM MD. The steps performed in a classical MD are minimization, heating and equilibration to proceed afterwards with the production phase from which we will take the snapshots to further develop our study. At least one frame is selected from the MM simulation, it is important to start with a suitable snapshot since, for example, ring conformational changes might involve sizeable energy barriers ($\geq$5 kcal/mol) and thus are unlikely in the picosecond time scale at room temperature. Afterwards, you need to set the quantum described atoms, mainly all the atoms involved in the reaction, decide where would be the best part to cut between the QM and MM regions and include the corresponding capping hydrogens or monovalent pseudopotentials. Once the QM region is determined, there are some parameters which need to be decided; the fictitious electronic mass of the QM region (if using CPMD), the dynamics time step and also, the electrostatic QM-MM interaction radius must be carefully set before QM/MM MD equilibration is initiated, as important features of the active site might be influenced by non-bonded interactions with the environment. Once we have set up the QM/MM system we also equilibrate it to the new method; by a minimization step of the wave function and the geometry of the system; afterwards we set the thermostats to the corresponding temperature and now we are able to start the metadynamics simulation. First, the collective variables need to be chosen, which is a crucial process for obtaining a good physical description of the process.

In general, a selected small number of CVs are chosen and by failures attempts you improve them until reaching the ones which lets you achieve convergence of the metadynamics simulation; finally obtaining the proper data to be evaluated to obtain the reaction mechanisms results.

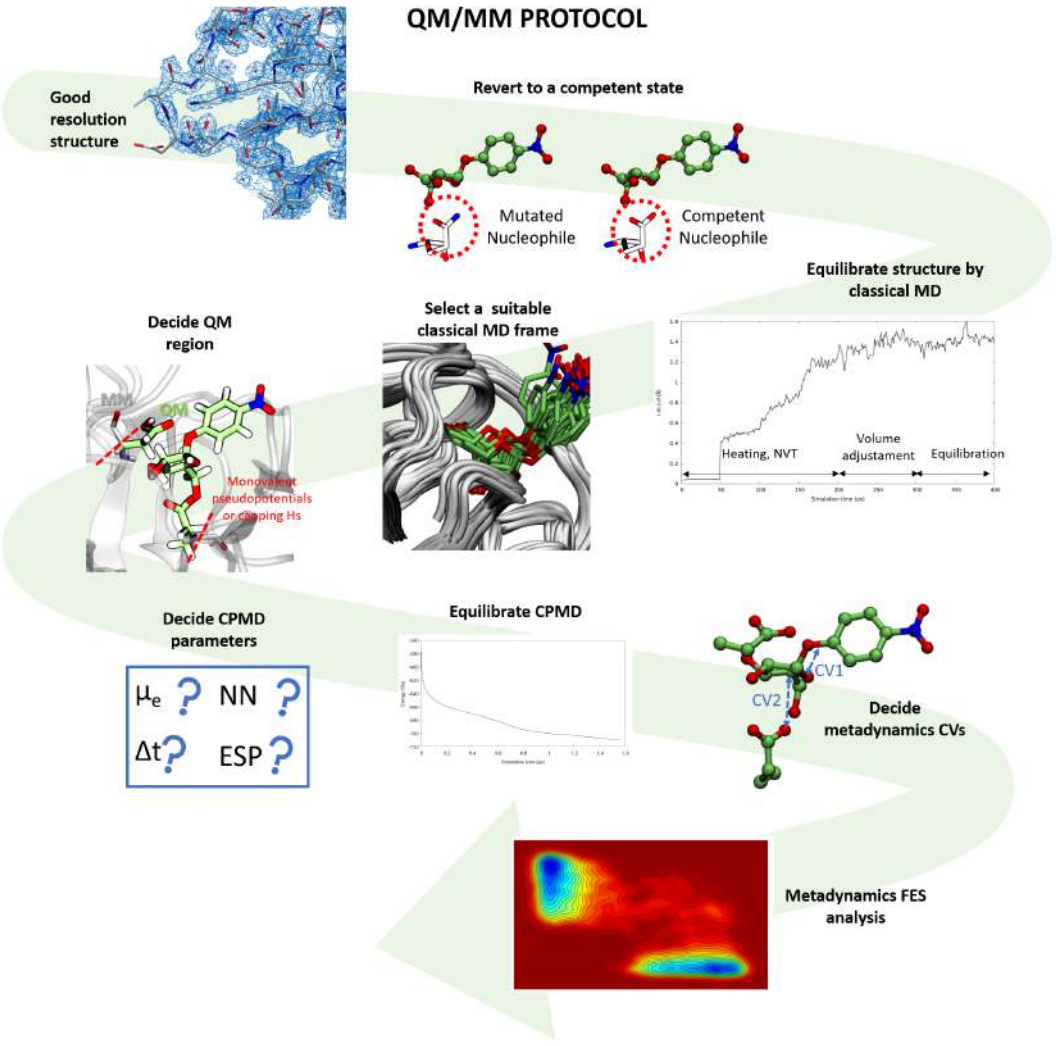


Figure 2.7: Schematic representation of the protocol used for a QM/MM metadynamics simulation.

Chapter 3

Unravelling the origin of brain diseases: catalytic study of β-Galactocerebrosidase

Parts of this chapter have been published:
Nin-Hill, A.; Rovira, C.; "The catalytic reaction mechanism of β- galactocerebrosidase, the enzyme deficient in Krabbe disease." *ACS Catalysis* Published ahead of print (https://doi.org/10.1021/acscatal.0c02609). (2020)

ABSTRACT: Krabbe disease is a neurodegenerative disorder related to misfunction of β-galactocerebrosidase (GALC), a glycosidase that catalyzes the cleavage of β-galactosidic bonds in glycosphingolipids. Here we uncover the catalytic molecular mechanism of GALC with both a synthetic (Gal-β-pNP) and its natural (GalCer) substrate. Our results clarify the unusual chair conformation of the Gal-β-pNP substrate observed in the crystal structure and show that catalysis can take place via two distinct conformational pathways $^{1}S_3 \rightarrow [^{4}H_3]^{\ddagger} \rightarrow {}^{4}C_1$ and $^{4}C_1 \rightarrow [^{4}H_3]^{\ddagger} \rightarrow {}^{4}C_1$ with similar free energy barriers, due to leaving group flexibility. In contrast, only the $^{4}C_1 \rightarrow [^{4}H_3]^{\ddagger} \rightarrow {}^{4}C_1$ itinerary turns out to be feasible when the natural GalCer substrate is hydrolysed. Our results also show that binding of the lipid-transfer protein saposin A (SapA) is necessary for the reaction to proceed with a low reaction energy barrier. This mechanistic insight could aid in the design of Krabbe diagnosis probes and GALC conformational chaperones.

3.1 Introduction

The first GH enzyme studied in this Thesis is β-galactocerebrosidase (GALC) (enzyme comission 3.2.1.46) which, according to the classification discussed in the Introduction section, corresponds to an enzyme operating via a retaining mechanism, i.e., retaining the anomeric carbon configuration of the scissile glycosidic bond at the end of the carbohydrate chain, since GALC is an exo-GH. Regarding its sequence similarity, GALC is classified in family 59. This enzyme is essential for the catabolism of glycosphingolipids, besides its function in cancer cell metabolism, primary open-angle glaucoma and the maintenance of a hematopoietic stem cell niche. GALC is responsible of catalyzing the hydrolysis of β-galactocerebroside (GalCer) to β-D-galactose and ceramide, as well as the cleavage of psychosine to β-D-galactose and sphingosine (Figure 3.1**A**). The ability of GALC to cleave two different sphingolipids is due to the fact that the lipid tail is projecting from the surface of the enzyme. This suggests that the sphingolipid leaving group contributes little to substrate recognition, which is governed by the interactions of the enzyme with the β-galactose sugar at the active site [83].

GALC works in the lysosome which has a slightly acidic environment, pH=4.8, compared to the cytosol, pH=7.2. Lysosomal degradative enzymes are very sensitive to pH, which is very convenient for the cell, as they become inactive if there is some leak in the lysosomal membrane. Hill *et. al.* [84], measured the activity of GALC in a range of different pHs, determining that the optimum pH of GALC activity is between 4.5 and 5, consistent with the lysosomal pH. Malfunction of GALC leads to Krabbe disease which is a rare and usually deadly disorder of the nervous system, being one of the most prevalent lysosomal storage disorders [85, 86]. Malfunction of GALC leads to accumulation of the cytotoxic metabolites galactocerebroside and psychosine, leading to apoptosis of myelin-forming cells. This causes severe neurodegenerative disorders in Krabbe patients, especially small children [87]. Several therapies are available to fight Krabbe disease, such as hematopoietic stem cell transplantation, gene therapy and enzyme replacement, or a combination of them [86]. More recently, pharmacological chaperone therapy has emerged as a promising alternative or complementary approach [8, 9]. By binding of a small stabilizer molecule, degradation of GALC before reaches the lysosome is prevented. Some of these molecules are small molecule inhibitors designed to mimic the conformation of the substrate at the Michaelis complex or transition state of the enzymatic reaction. Therefore, understanding the catalytic mechanism of GALC at atomic detail is important to boost the development of efficient conformational chaperones for Krabbe disease therapy.

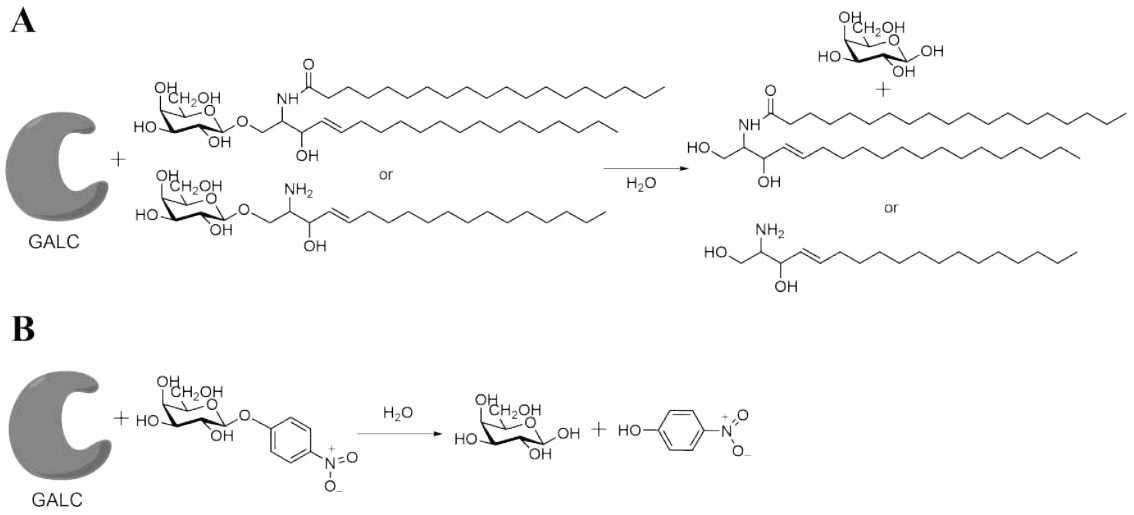


Figure 3.1: **(A)** Hydrolysis reaction of galactocerebroside (top) and psychosine (bottom) and the subsequent products, galactose and ceramide or sphingosine. **(B)** Hydrolysis reaction of the analogue Gal-β-pNP substrate by GALC and the subsequent products, galactose and 4-nitrophenol.

As an exo-GH with a classical retaining mechanism, GALC has two active site residues, identified by the work of Deane *et. al.* [83]. The nucleophile and acid/base residues correspond to glutamates, Glu258 and Glu182. Currently, there are 12 X-ray structures of GALC available in the Protein Data Bank (PDB). Most of these structures correspond to the unligated form of the enzyme [83], complexes with non-hydrolysable inhibitors [88] and the unligated enzyme in complex with Saposin A [89]. There are also structures in complex with hydrolyzable substrate analogues, which will be discussed in the next section [84].

3.1.1 Unusual sugar conformation in a MC structure

The most interesting GALC crystallographic structure for mechanistic analysis is the Michaelis complex (MC) with the non-natural substrate 4-nitrophenyl-β- D-galactopyranoside (Gal-β-pNP), PDB code 4CCC (Figure 3.1**B**). Interestingly, this complex was trapped without using a catalytically inactive mutant form of the enzyme or a non-hydrolysable substrate analogue, as usually done in GHs [34]. Instead, the authors could trap the Michaelis complex by working at pH 6.8, in which GALC shows reduced catalytic activity, as the catalytic acid/base residue is unprotonated, and it is not able to perform its function. Some issues have been raised in the literature about the reliability of this structure being a true Michaelis complex in view of the non-catalytic orientation of the acid/base residue, which points towards the solvent [34], see Supplemental Figure 3.14, page 73. Another issue about this structure is the conformation of the galactose of the Gal-β-pNP

substrate. Surprisingly, the galactose ring is in a relaxed chair conformation, instead of being distorted to a boat or skew-boat conformation as it is observed in most MCs of GHs [35].

As discussed in Chapter 1, a distorted conformation is more preactivated for hydrolysis than the chair conformation. The lack of sugar distortion observed in the GALC Michaelis complex could be a consequence of the misorientation of the acid/base residue (Supplemental Figure 3.14), or could be a functional feature of the enzyme. In this regard, the solvent-exposed leaving group, which adopts an equatorial orientation in the crystal structure, could easily reorient during the catalytic reaction, facilitating the sugar to adopt a distorted conformation. QM/MM MD metadynamics simulations of the GALC reaction mechanism will be performed in this chapter to clarify this issue.

3.1.2 Natural conditions: GALC-Saposin A complex with β-D-galactocerebroside

An important agent for GALC activity *in vivo* is a lipid-transfer protein called Saposin A (SapA). SapA belongs to the family of saposin domains which serve as activators of various lysosomal lipid-degrading enzymes. Their main function consists on making the lipid substrate, placed in the cell membrane, more accessible to the water soluble hydrolases. There are two hypothesis in the literature describing the saposins mechanism of action [90–93]: the "solubiliser" model and the "liftase" model. The solubiliser model assumes that the saposin extracts the substrates from the membrane so that the hydrolising activity of GHs can take place further away from it. In the liftase model, the saposin helps catalysis distorting the membrane in a way that the substrate is placed close to the GH active site, so that the hydrolysis reaction takes place very close to the membrane, if not just in the membrane itself. SapA is thought to work via the "solubiliser" model, as discussed in various works [89, 94], see Figure 3.2.

Several structurally distinct states, open or closed, of saposins have been previously crystallized [94–100], being the open state the one that is believed to be mostly present in complex with GALC. This is the state determined in the recent apo structure of GALC-SapA dimer [89]. The correct GalCer degradation depends on the presence of SapA, as demonstrated by mutagenesis studies. In particular, mutations present only in SapA affect the GALC function, while GALC is maintained intact. Likewise, mutations of important residues of GALC interacting directly with SapA are among the ones causing Krabbe disease [101–105].

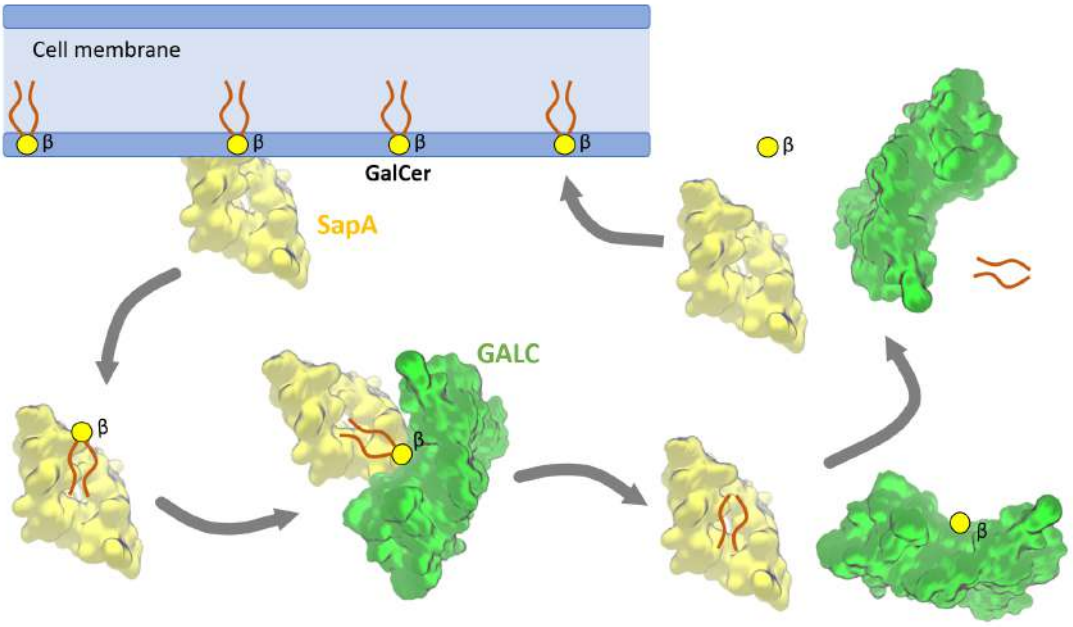


Figure 3.2: Schematic diagram for the saposin-mediated mechanism of GALC by the "solubiliser" model. The SapA protein (in yellow) extracts the GalCer molecule (illustrated as a yellow circle and two brown lines corresponding to β-galactose and cerebrosidase, respectively) from the cell membrane and carries it to the GALC enzyme (in green) where the hydrolysis reaction takes place.

3.2 Results and Discussion

3.2.1 Conformations of the substrate on-enzyme

To answer the question whether the non-catalytically competent pH would affect substrate conformation in GALC, the conformational free energy landscape of Gal-β-pNP substrate in the active site of GALC was computed with the catalytic protonated Glu182. The starting structure for the simulations was taken from the reported GALC—Gal-β-pNP complex structure, after reverting the orientation of the acid/base residue to the catalytic form observed in the WT enzyme (Supplemental Figure 3.14, page 73). The MD simulations show that the system is stable, with the two catalytic residues correctly oriented for catalysis, while the β-galactosyl ring remains in the $^4C_1$ conformation during the entire simulation (130 ns, Supplemental Figure 3.15, page 74).

The FEL reconstructed from the metadynamics simulation (Figure 3.3) shows that not only the $^4C_1$ conformation corresponds to a free energy minimum; there is another relevant minimum close to the $^1S_3$ conformation. Interestingly, this is the conformation expected for β-galactosidases [34], compatible with NMR [106, 107] and QM/MM studies on GH2 β-galactosidase [108]. It is also the conformation

that is often observed for the closely-related β-glucosidases [34–36]. This conformation, hereafter classified for simplicity as $^1S_3$, is stabilized by interactions of the enzyme with the galactose unit (see Figure 3.3). In fact, calculations on an isolate β-galactose molecule (i.e. in the absence of the enzyme) show a completely different landscape, dominated by the $^4C_1$ conformation (Supplemental Figure 3.13, page 72). As found also in endo-GH complexes, this illustrates how the enzyme confines the -1 subsite sugar in the active site so that only a small set of conformations (in principle the catalytically relevant ones) "survive". The two conformations observed on-enzyme are practically isoenergetic (the small energy difference of ≈ 0.5 kcal·mol^{-1}, in favour of the chair conformer, is within the limits of the accuracy of the method used). This means that the substrate can adopt two conformations in the active site of GALC, with $\approx 50\%$ probability each. It is unclear so far, though, whether the two conformations are catalytically competent.

The Gal-β-pNP substrate in both conformations is only able to adopt one of the three possible hydroxymethyl rotameric states, the gt one (C4-C5-C6-O6 dihedral angle has $120° \leq \omega \leq 180°$ or $-120° \leq \omega \leq -180°$) [109, 110]. This rotameric state is fixed due to the interaction with Arg380 and Ser261 conserved along all the classical and QM/MM *ab initio* molecular dynamics simulations. Interestingly, the sugar conformer calculated in vacuum with the highest gt population ($\sim$25%) is the undistorted $^4C_1$, compared to populations of $\sim$10% for all the other conformations (Table 3.6, further discussed in section 3.5 of the Supplemental Information, page 72). This can favour $^4C_1$ conformer in the enzyme active site with respect to other conformations.

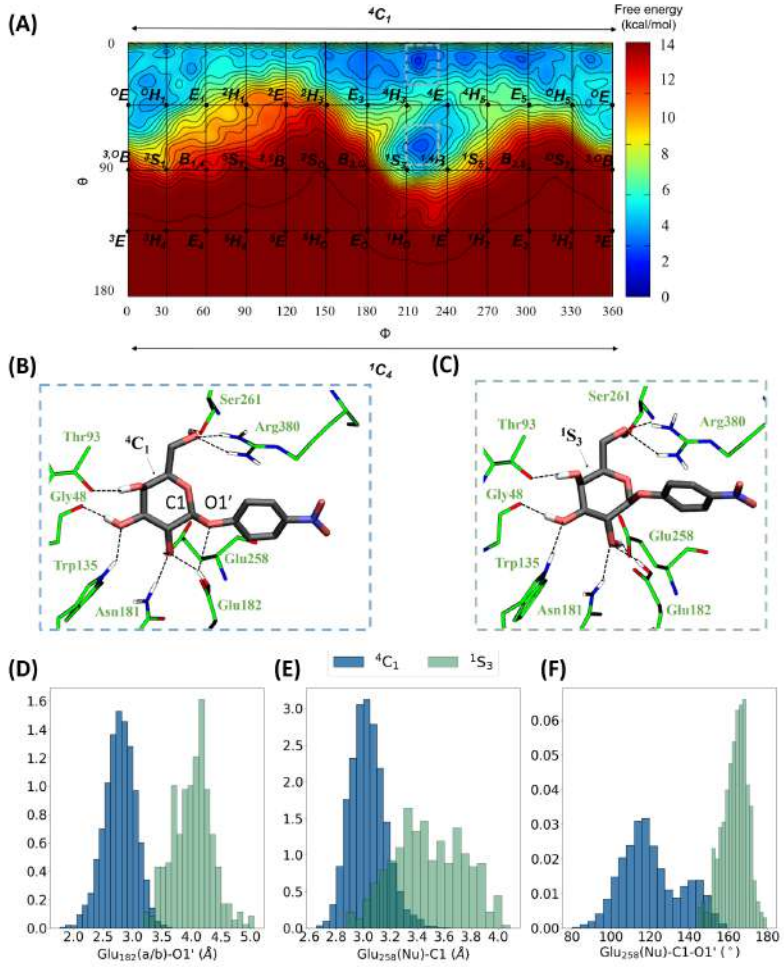


Figure 3.3: **(A)** Conformational FEL of Gal-β-pNP residue at the -1 enzyme subsite of GALC. Energy values are given in kcal·mol^{-1} and each contour line corresponds to 1 kcal·mol^{-1}. **(B)** Analysis of substrate-enzyme interactions of the two conformations of the β-galactose unit in the active site of GALC. Carbon atoms in the substrate are represented in black whereas the carbon atoms of the enzyme residues are represented in green; oxygen and nitrogen atoms are coloured red and blue, respectively. Only the residues with direct contact with the substrate are shown. Black dashed lines are used for hydrogen bond interactions. Representative snapshot of the active site with a $^4C_1$ substrate conformation. **(C)** Representative snapshot of the active site with a $^1S_3$ substrate conformation. **(D)** Normalized distribution of the distance between the acid/base catalytic residue (Glu182) and the glycosidic oxygen of the two most stable conformers, $^4C_1$ and $^1S_3$, in blue and green, respectively. **(E)** Normalized distribution of the distance between the nucleophile catalytic residue (Glu258) and the anomeric carbon of the two most stable conformers, $^4C_1$ and $^1S_3$, in blue and green, respectively. **(F)** Normalized distribution of the O_{Glu258} – C1 – O1' angle. The undistorted $^4C_1$ conformation exhibits a wider distribution at low angle values, indicating a less favourable S_N2 attack compared with the distorted conformer.

3.2.2 Reactivity of Gal-β-pNP from two different conformations

To investigate how the substrate conformation evolves during catalysis, we modelled the first step of the enzymatic reaction (the formation of the glycosyl-enzyme intermediate). Two collective variables (CVs), corresponding to the main covalent bonds that are formed or cleaved during the reaction, were used to drive the system from reactants (the Michaelis complex, MC) to products (the glycosyl-enzyme intermediate, GEI). The first collective variable (CV1), named "glycosidic bond cleavage", quantifies the cleavage of the glycosidic bond. The second one (CV2), named as "nucleophilic attack" accounts for the formation of the covalent bond between the anomeric carbon of the β-galactose and the closest oxygen atom of the catalytic nucleophile (Glu258). Two simulations were performed, starting from snapshots corresponding to the two minima of the conformational FEL, i.e. starting either from the $^4C_1$ and $^1S_3$ conformations of the β-galactoside substrate. The free energy landscape of the reaction starting from the $^1S_3$ conformation (Figure 3.4) shows that the system evolves towards the GEI in a concerted pathway with only one transition state (TS). The reaction free energy barrier (13.5 ± 1.4 kcal·mol^{-1}) agrees with the one estimated from the experimental reaction rate (≈ 15 kcal·mol^{-1}) [84], indicating that the reaction is feasible.

The reaction begin with the elongation of the glycosidic bond (C1-O1') while the sugar ring changes conformation from $^1S_3$ to $^4H_3$ as it reaches the reaction TS (Table 3.1 and Figure 3.4). As previously observed, the nucleophile practically does not move along the first stages of the reaction (from R to TS) indicative of a dissociative DN*AN mechanism, where the glycosidic bond breaks before the nucleophilic attack takes place [111]. The TS state exhibits a shorter C1-O5 bond distance (1.29 Å) than the MC (1.40 Å) and an almost coplanar arrangement of C2, C1, O5 and H1 atoms, consistent with a sp2-like hybridization of the anomeric carbon and an oxocarbenium ion-like character. Protonation of the leaving group is not seen, this is probably due to the interaction of the deprotonated p-nitrophenyl group (negatively charged) and Arg380 (positively charged) creating a strong ion-pair electrostatic interaction that adds to the typical cation-π interaction [112, 113] for these two groups, stabilizing this aglycon moiety even when it is deprotonated. The conformation of the substrate at the GEI ($^4C_1$) is in agreement with the one observed in the corresponding X-ray structure [84].

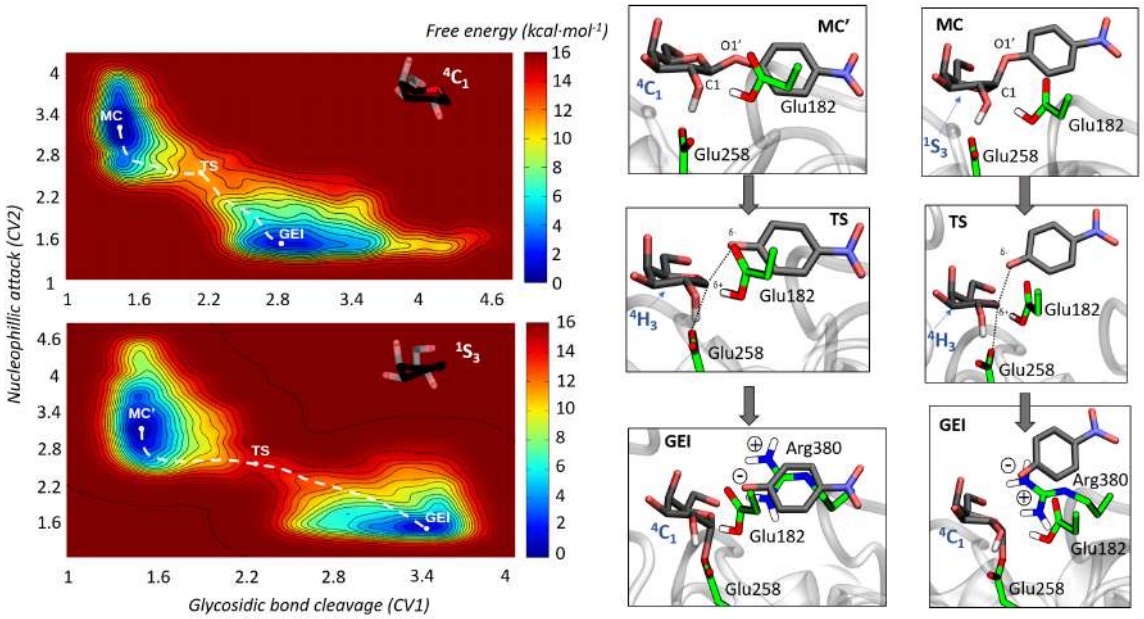


Figure 3.4: Reaction free energy landscape and representative structures along the catalytic itinerary for the QM/MM metadynamics simulation starting with the substrate in the $^4C_1$ and $^1S_3$ conformations. For the sake of clarity, most substrate hydrogen atoms have been omitted. Isolines are at 1 kcal·mol^{-1}. The dashed line indicates the minimum free energy pathway. Bonds being broken/formed are shown as dashed black lines.

Interestingly, the reaction departing from the $^4C_1$ minimum of the substrate conformational FEL (Figure 3.4); i.e. from the undistorted substrate conformation; turns out to be feasible as well. It involves a similar dissociative transition state and product complex (with conformations $^4H_3$ and $^4C_1$, respectively) as the reaction starting from the distorted substrate conformation. The computed free energy barrier (15.4 $\pm$ 2.4 kcal·mol^{-1}) is not much above the one obtained for the reaction starting from the $^1S_3$ conformation (13.5 $\pm$ 1.4 kcal·mol^{-1}), both values being compatible with experiments.

	Cyclic itinerary			**Classic itinerary**		
	MC' ($^4C_1$)	**TS** ($^4H_3$)	**GEI** ($^4C_1$)	**MC** ($^1S_3$)	**TS** ($^4H_3$)	**GEI** ($^4C_1$)
C1-O1' (Å)	1.44 ± 0.06	2.12 ±0.03	2.82 ± 0.07	1.49 ± 0.03	2.32 ±0.05	3.50 ± 0.03
C1-O_{Glu258} (Å)	3.17±0.06	2.63±0.04	1.56±0.09	3.13 ± 0.02	2.65 ± 0.04	1.57 ± 0.03
O_{Glu182}-H_{Glu182} (Å)	1.04 ± 0.04	1.01 ± 0.02	1.00 ± 0.03	1.03 ± 0.03	1.04 ± 0.02	1.01 ± 0.03
H_{Glu182}-O1' (Å)	3.1 ± 0.2	3.71 ± 0.07	3.7 ± 0.3	4.0 ± 0.3	4.2 ± 0.6	4.8 ± 0.3
C1-O5 (Å)	1.44 ± 0.04	1.30 ± 0.01	1.38 ± 0.04	1.40 ± 0.05	1.29 ± 0.03	1.37 ± 0.03
δ_{C1}	0.7 ± 0.1	0.878 ± 0.008	0.7 ± 0.1	0.9 ± 0.06	1.0 ± 0.1	0.8 ± 0.1
δ_{O5}	-0.73 ± 0.08	-0.554 ± 0.004	-0.67 ± 0.09	-0.75 ± 0.05	-0.54 ± 0.06	-0.6 ± 0.1
θ (°)	19 ± 3	45 ± 12	15 ± 8	74 ± 7	49 ± 10	20 ± 8
ϕ (°)	277 ± 20	208 ± 25	176 ± 56	206 ± 6	208 ± 8	150 ± 23

Table 3.1: Structural and electronic parameters of the characteristic points along the reaction pathways.

That the enzyme stabilizes an undistorted substrate conformation might seem at first sight surprising, as β-GHs are known to act via distorted conformations. One example is family GH16 1,3-1,4-β-glucanase, in which the $^4C_1$ conformation was 10-11 kcal·mol^{-1} higher than the most stable distorted $^1S_3$ conformation [114], thus not populated at room temperature. However, 1,3-1,4-β-glucanase is an endo-β-GHs, i.e. it cleaves the substrate in the middle of the oligosaccharide chain. The enzyme can accommodate a leaving group with up to 3 sugar units (expanding from +1 to +3 subsites) and can control the distortion of the -1 sugar by evolving an appropriate shape of the binding cavity [36], which forces the leaving group to adopt an axial orientation. Instead, GALC is an exo-acting enzyme and does not have positive subsites. It is worth noting that a recent docking and MD study on clan GH-A β-galactosidases [115], which are all exo-GHs, predicts a cloud of different closely-related interchanging conformers for the β-galactoside substrate, in agreement with the results here found for GALC. The leaving group of the GALC reaction is solvent exposed and can move freely. This lack of steric determinants probably is the reason that the β-galactose moiety can exchange between the distorted and undistorted conformers both of which are catalytically relevant.

Of course, the nucleophilic substitution is more favoured when the glycosidic bond is in a pseudo-axial orientation in a way to avoid the steric hindrance between the hydrogen of the anomeric carbon and the nucleophile. This is only observed in the $^1S_3$ conformer (Figure 3.3). However, the acid/base catalytic residue (Glu182) of the $^4C_1$ conformer is closer to the glycosidic oxygen by more than 1 Å compared to the $^1S_3$ conformer, which favours the hydrolysis reaction. Similarly, the nucleophile (Glu258) is closer to the anomeric carbon in the undistorted substrate compared to the distorted one. As a result, both itineraries (not only the "classic" β-glucosidase itinerary, also expected for β-galactosidase, that start from the $^1S_3$ conformation but also a novel cyclic itinerary that start from a $^4C_1$ conformation) are competitive.

In summary, the active site of GALC can accommodate the β-galactose substrate in two alternative conformations ($^1S_3$ and $^4C_1$) that can interconvert. The hydrolysis reaction can take place via either of the two conformations, leading to two distinct catalytic itineraries, $^1S_3 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$ and $^4C_1 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$. Although the first itinerary is slightly favoured, both itineraries are expected to contribute to GALC catalysis. This scenario can possibly be extended to other exo-GHs that, similarly to GALC, exhibit solvent-exposed active sites.

3.2.3 The natural conditions: considering the GalCer substrate and the GALC-SapA dimer

3.2.3.1 GALC in complex with GalCer

Using a hydrolysable analogue such as Gal-β-pNP two different catalytic itineraries were found to be accessible. However, there is still the question if the same results would be obtained with the natural substrate GalCer in which the aglycon moiety differs both in size and polarity. Thus, GalCer substrate was considered (see Section 3.4.1.1 for the system setup, page 65). A classical molecular dynamics was performed to equilibrate the new system and find the most probable position of the ceramide tail, shown in Supplemental Figure 3.18. The conformational free energy landscape of GalCer inside the protein cavity, computed by QM/MM metadynamics, indicates that the ceramide tail restricts even more the conformational space that the galactose moiety is able to explore, compared to the pNP leaving group (Figure 3.3). In contrast to the simulation with the synthetic Gal-β-pNP substrate, there is only one main minimum which corresponds to the undistorted $^4C_1$ conformer (Figure 3.5).

Even though GalCer only adopts the undistorted conformation, the angle between the nucleophile oxygen, C1 and O1' ($\sim 140°$) is placed in-between the angles previously observed for the $^4C_1$ ($\sim 120°$) and $^1S_3$ ($\sim 170°$) conformations of Gal-β-pNP (Figure 3.5**C**). This indicates that the leaving group is pseudo-axial, which facilitates the nucleophillic attack. All the active site interactions are maintained.

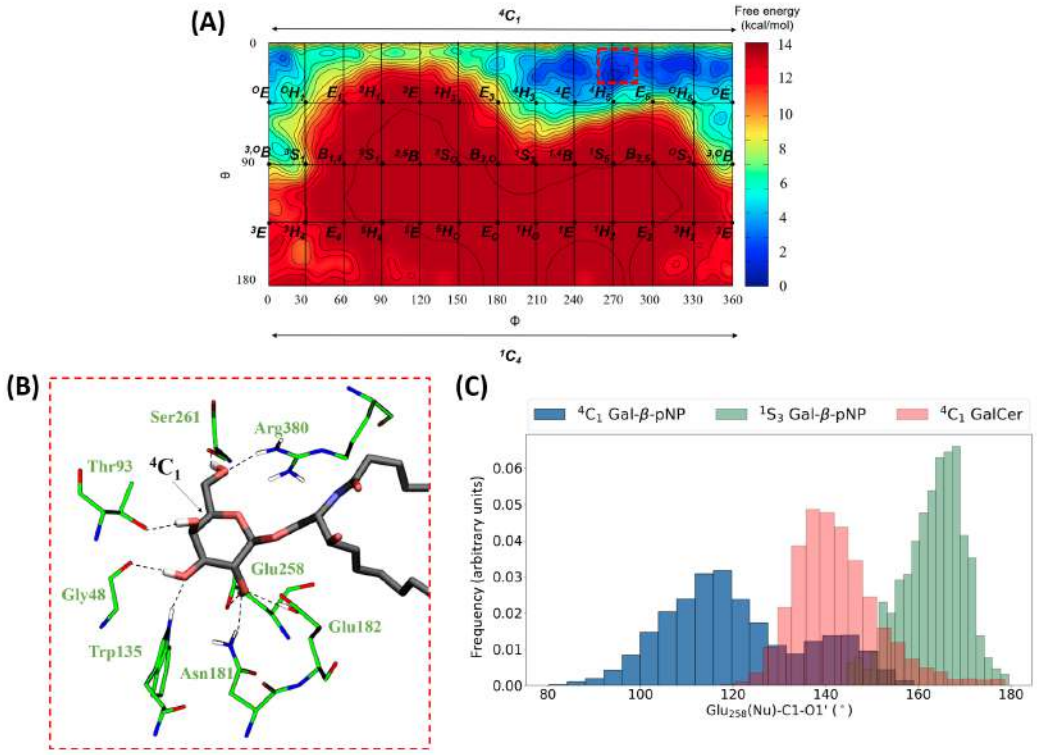


Figure 3.5: **(A)** Conformational FEL of the galactose moiety of the GalCer substrate at the -1 enzyme subsite of GALC. Energy values are given in kcal·mol^{-1} and each contour line corresponds to 1 kcal·mol^{-1}. **(B)** Representative snapshot of the most stable conformer of GalCer in GALC. Carbon atoms in the substrate are represented in black whereas the carbon atoms of the enzyme residues are represented in green, oxygen and nitrogen atoms are colored red and blue, respectively. Only the residues with direct contact with the substrate are shown. Black dashed lines are used for hydrogen bond interactions. **(C)** Normalized distribution of the O_{Glu258}-C1-O1' angle of GalCer in its most stable conformation, $^4C_1$, and the same angle for the two possible conformations of Gal-β-pNP, $^4C_1$ and $^1S_3$, in red, blue and green, respectively.

The reaction simulation of GalCer by GALC was initiated from the only minima present in the conformational FEL, $^4C_1$. A different set of CVs were used for GalCer as opposed to Gal-β-pNP, the acid/base residue was now included, as the lipid tail is a bad leaving group and the reaction is not expected to proceed without strong acid/base assistance. The first CV called "nucleophillic attack" quantifies both the glycosidic cleavage and the formation of the covalent bond between the anomeric carbon of the β-galactose and the closest oxygen of the nucleophile (Glu258) and was defined as a difference of these two distances. The second CV, named as "proton transfer" accounts for the protonation of O1' of the ceramide tail from the acid/base residue (Glu182) and it was described also as a distance difference between O1'-H_{Glu182} and H_{Glu182}-O_{Glu182}.

The simulation show that the reaction initiates by the approach of the acid/base

residue, Glu182, to the glycosidic oxygen so that the protonation of the leaving group can occur. At the same time, the glycosidic bond (C1-O1') is elongated while the nucleophile practically does not move along the first stages of the reaction, as also observed in the simulations with Gal-β-pNP, see Table 3.2 in page 60, and Supplemental Figure 3.23 in page 81. The sugar ring changes conformation from $^4C_1$ (MC) to $^4H_3$ as it reaches the reaction TS to finally return to $^4C_1$ in the GEI state. The reaction free energy barrier is substantially higher, 32.3 kcal·mol^{-1}, than the previous ones found for Gal-β-pNP (15.4 or 13.5 kcal·mol^{-1} depending on the initial substrate conformation).

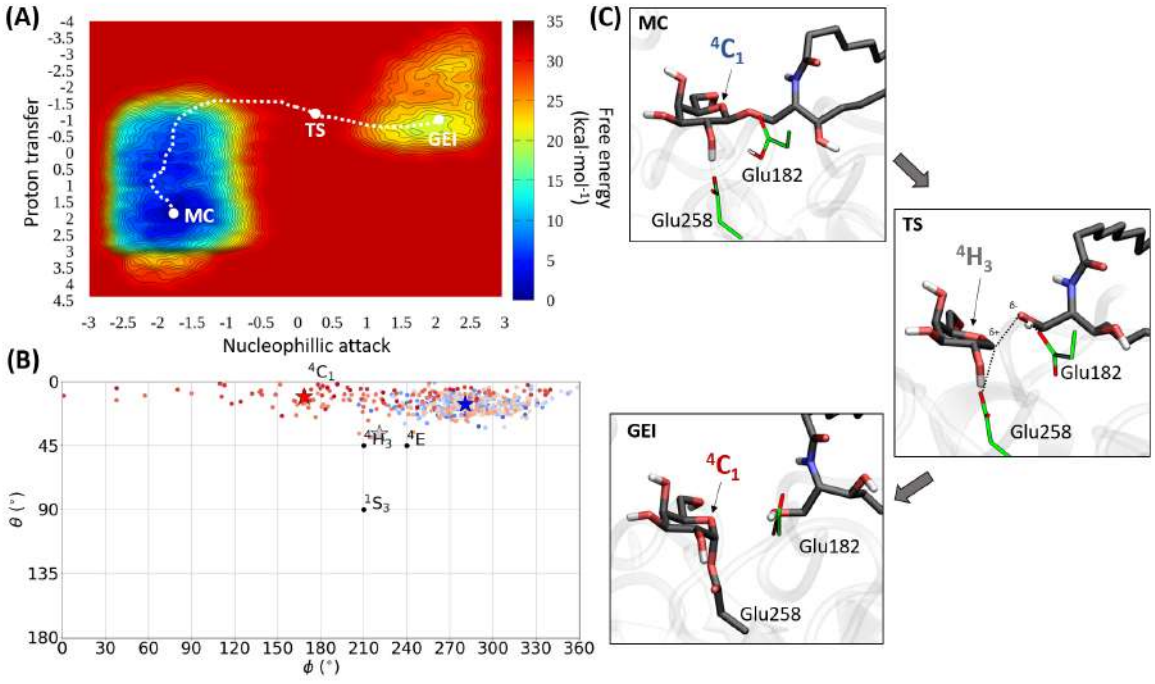


Figure 3.6: **(A)** Free energy landscape of the hydrolysis reaction of GalCer in GALC. The white dashed line indicates the lowest free energy path (LFEP) calculated with MEPSAnd [116]. **(B)** GalCer ring conformation along the catalytic itinerary of the LFEP from the MC, in blue, to GEI, in red; going through the TS, in gray. The averaged structures of the main states are marked with a star in their corresponding colours. **(C)** Representative structures of the catalytic itinerary along the reaction mechanism. The two catalytic residues are also shown in green. For the sake of clarity, most hydrogen atoms have been omitted.

In summary, the natural GalCer substrate follows the same cyclic itinerary previously found for Gal-β-pNP; $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$, but the ceramide tail increases the reaction barrier by approximately 15 kcal·mol^{-1}, making the reaction unfeasible from a kinetic point of view.

3.2.3.2 GALC-SapA dimer in complex with GalCer

The lipid-transfer protein SapA is essential for the function of GALC *in vivo* helping to bring the amphiphilic substrate GalCer close to the water soluble GALC protein. Very recently, Hill *et. al.* were able to crystallize the complex of GALC with SapA [89]. To fully understand how GALC works *in vivo*, and to be able to compare it with the results obtained for Gal-β-pNP and GalCer substrates with only GALC, new simulations were designed considering the recently available GALC-SapA dimer. The GalCer substrate was placed manually in the active site of GALC-SapA structure by superimposing the previously equilibrated structure of GALC in complex with GalCer with the crystallographic structure of the unligated form of GALC-SapA dimer. Classical MD simulations were done in the unligated and GalCer-bound forms of GALC-SapA dimer. The good behaviour of the GALC-SapA:GalCer complex during the MD (Supplemental Figure 3.20), as observed in the previous systems analyzed, made us confident of the quality of such large model. We can suggest from the MD simulations of GALC-SapA dimer with GalCer that SapA stabilizes the lipid tail acting as a polar solvent shield and doubling the hydrophobic interactions between GalCer and GALC-SapA dimer compared to GALC alone (Supplemental Figure 3.21).

The conformational free energy landscape of GalCer in the active site of GALC-SapA dimer changes respect the conformational FEL of GalCer with GALC-only. Besides the $^4C_1$ minima, there is also another minimum corresponding to $^1S_3$ conformation, which is 3 kcal·mol^{-1} higher in energy. Therefore, the conformational FEL of GalCer in GALC-SapA ressembles that of Gal-β-pNP in GALC, with the $^4C_1$ conformation being favoured this time. $^4C_1$ is more stable due to the presence of two hydrogen bonds that the distorted conformation is not able to form; a hydrogen bond interaction between the acid/base residue, Glu182, and the glycosidic oxygen, O1', which seems to be facilitated by the higher flexibility of the glycosidic bond; and also a hydrogen bond interaction between the hydroxyl group at C2 of the galactose moiety and the nucleophile, Glu258, an interaction found to be essential for catalysis in other β-GHs [117], see Figure 3.7. The $^1S_3$ conformation in the GALC:Gal-β-pNP complex was able to form this hydrogen bond interaction between the hydroxyl group at C2 of the galactose moiety and the nucleophile, Glu258, so that was nearly isoenergetic to the $^4C_1$ conformation, see Figure 3.3.

The hydrolysis of GalCer by GALC-SapA was investigated with the same CVs set used for GalCer in GALC-only, starting the reaction from the most stable $^4C_1$ conformation. The chemical reaction displays a similar mechanism as previously observed with only the GALC subunit; the protonation and cleavage of O1' happens almost simultaneously while the bond between C1 and Glu258 is formed at the last

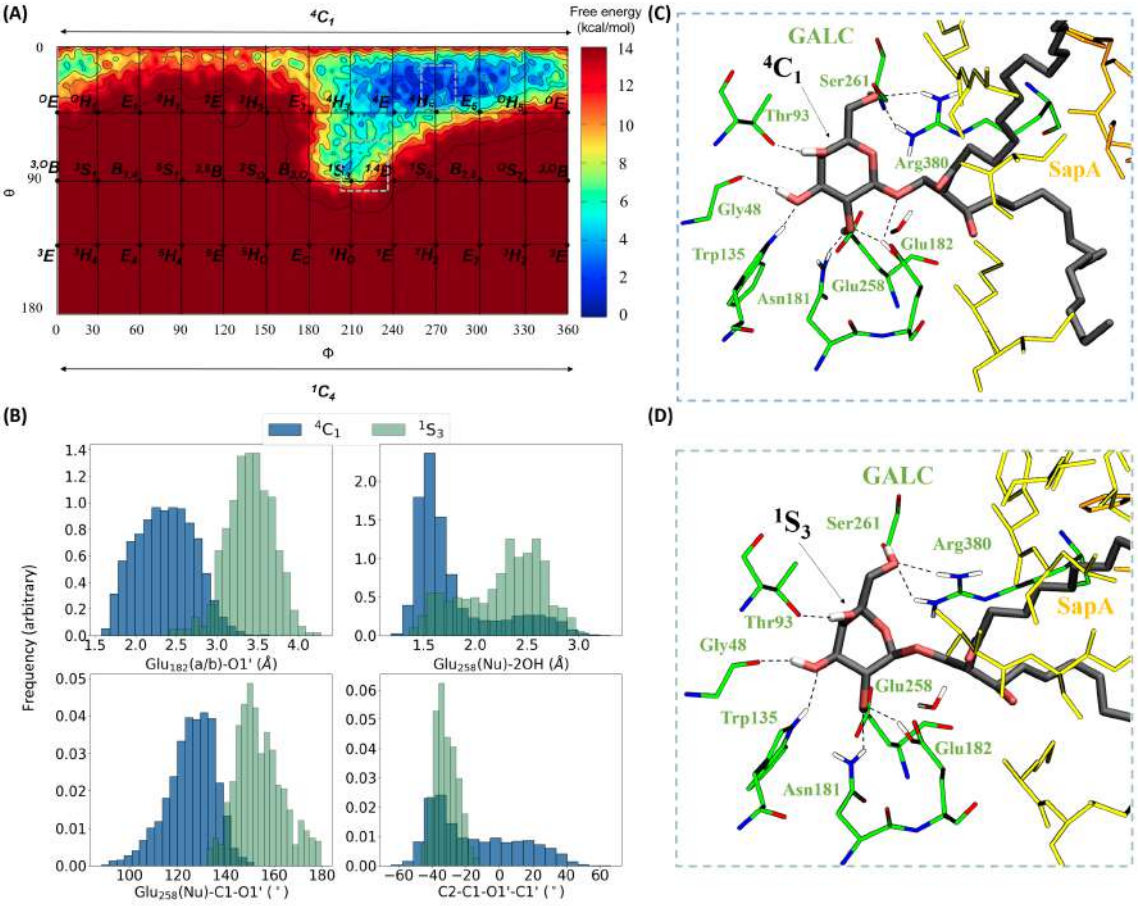


Figure 3.7: **(A)** Conformational FEL of GalCer residue at the -1 enzyme subsite of GALC-SapA dimer. Energy values are given in kcal·mol^{-1} and each contour line corresponds to 2 kcal·mol^{-1}. **(B)** Normalized distribution of the main interactions that differentiate between the most stable conformers; $^4C_1$ in blue and $^1S_3$ in green. **(C)** Analysis of the most stable conformers with the catalytic residues interactions of GalCer in GALC-SapA dimer. Carbon atoms in the substrate are represented in black whereas the carbon atoms of the enzyme residues are represented in green, oxygen and nitrogen atoms are colored red and blue, respectively; for the residues in GALC subunit. The residues of SapA are colored yellow and orange. Only the residues with direct contact with the substrate are shown. Black dashed lines are used for hydrogen bond interactions. Representative snapshot of the active site with a $^4C_1$ substrate conformation. **(D)** Representative snapshot of the active site with a $^1S_3$ substrate conformation.

stages of the reaction, see Supplemental Figure 3.25 (page 82) and Table 3.2 (page 60). The sugar ring conformation evolves from a $^4C_1$/ $^4H_5$ conformation to a $^4C_1$/^4E conformation at the reaction TS, to finally reach $^4C_1$ for the GEI state, thus the catalytic itinerary is slightly different from the previous cyclic itineraries found for GalCer and Gal-β-pNP in a GALC subunit. Strikingly, the free energy barrier is much lower (by almost 10 kcal/mol) with respect to the one obtained for the same substrate with GALC-only. The computed free energy barrier (24.7 kcal·mol^{-1}) is

in good agreement with the estimated experimental value, 23.4 ± 0.3 kcal·mol^{-1} (an average from different works of GALC-SapA dimer *in vivo* from studies in different cells types [118, 119], such as human fibroblast, leukocytes and brain).

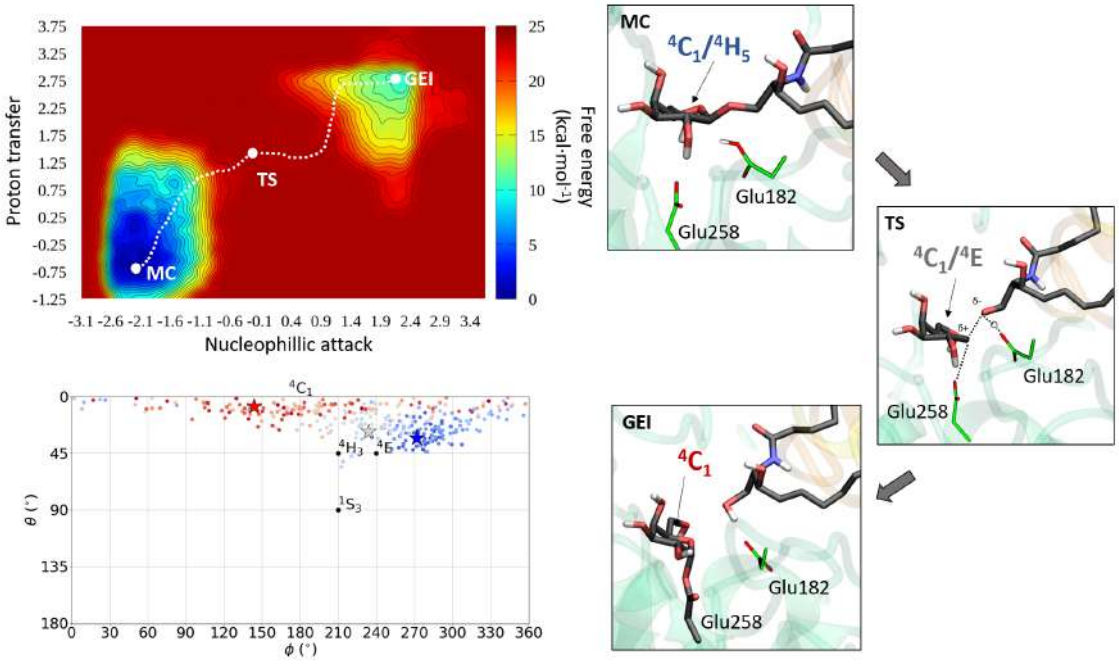


Figure 3.8: **(A)** Free energy landscape of the hydrolysis reaction of GalCer in GALC-SapA dimer. The white dashed line indicates the lowest free energy path (LFEP) calculated with MEPSAnd [116]. **(B)** GalCer ring conformation along the catalytic itinerary of the LFEP from the MC, in blue, to GEI, in red; going through the TS, in gray. The averaged structures of the main states are marked with a star in their corresponding colours. **(C)** Representative structures of the active site along the reaction mechanism. The two catalytic residues are also shown in green. For the sake of clarity, most hydrogen atoms have been omitted.

The observed significant decrease in the reaction free energy barrier compared to the model in the absence of SapA is likely due to the TS stabilization by a higher number of hydrophobic interactions of the lipid tail, with the SapA protein, and an extra hydrogen bond between Arg183 from GALC and the hydroxyl group from the lipid tail. This interaction is only possible thanks to the lipid tail orientation when SapA is present, see Figure 3.9.

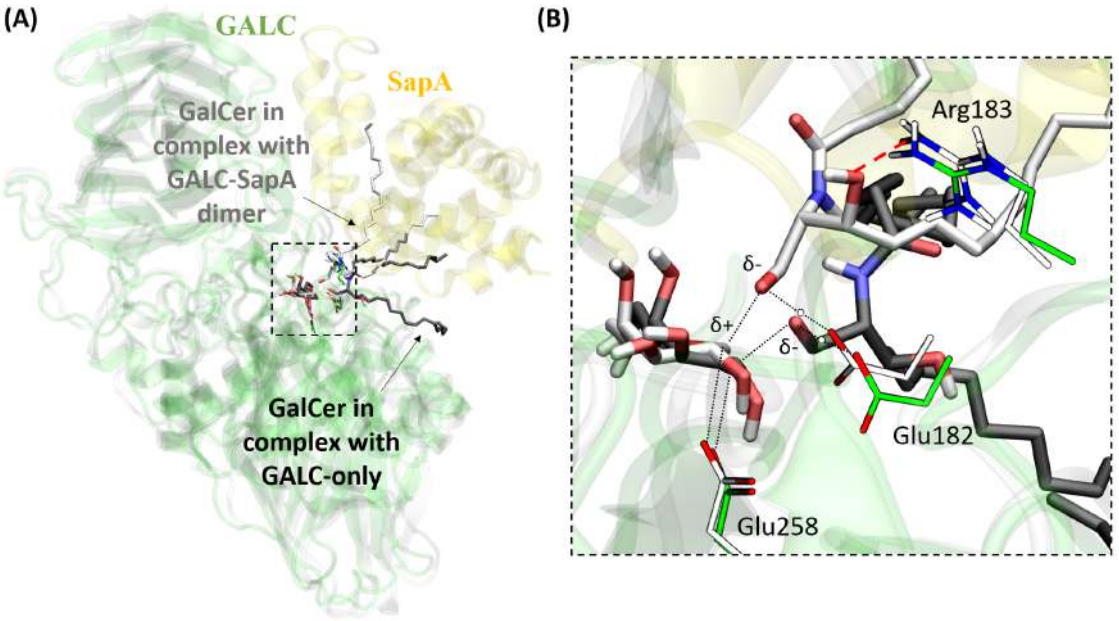


Figure 3.9: **(A)** Zoom out of the hydrolysis reaction TS snapshot of GalCer substrate in complex with GALC-only (shown in black) and with GALC-SapA dimer (shown in white). **(B)** Zoom in of the hydrolysis reaction TS snapshot of GalCer substrate. GalCer in complex with GALC-only is illustrated in black and GalCer in complex with GALC-SapA dimer is illustrated in white. The hydrogen bond interaction from the hydroxyl group of the lipid tail to Arg183, only present in the GALC-SapA dimer, is shown as a red dashed line.

	GALC-only			GALC-SapA dimer		
	MC ($^4C_1$)	**TS** ($^4H_3$)	**GEI** ($^4C_1$)	**MC** ($^4C_1/^4H_5$)	**TS** ($^4C_1/^4E$)	**GEI** ($^4C_1$)
C1-O1' (Å)	1.43 ± 0.04	2.1	3.31 ± 0.06	1.43 ± 0.04	2.4	3.7 ± 0.04
C1-O_{Glu258} (Å)	3.2 ± 0.04	2.36	1.54 ± 0.05	3.6 ± 0.04	2.6	1.51 ± 0.03
O_{Glu182}-H_{Glu182} (Å)	1.03 ± 0.03	1.34	1.63 ± 0.04	0.98 ± 0.02	1.35	1.7 ± 0.1
H_{Glu182}-O1' (Å)	2.87 ± 0.04	1.12	1.03 ± 0.04	2.98 ± 0.07	1.14	1.01 ± 0.03
C1-O5 (Å)	1.46 ± 0.03	1.32	1.39 ± 0.03	1.48 ± 0.04	1.28	1.4 ± 0.03
δ_{C1}	1.06 ± 0.06	1.12	1.01 ± 0.1	1.34 ± 0.2	1.19	0.96 ± 0.08
δ_{O5}	-0.66 ± 0.03	-0.25	-0.66 ± 0.06	-0.75 ± 0.06	-0.24	-0.54 ± 0.08
θ (°)	16 ± 3	37	11 ± 6	33 ± 5	28	8 ± 5
ϕ (°)	281 ± 17	221	168 ± 100	272 ± 11	234	144 ± 75

Table 3.2: Structural and electronic parameters of the characteristic points along the reaction pathway of GalCer in GALC subunit and in GALC-SapA dimer. δ_{C1} is a charge sum of C1,C2, H_{C1} and H_{C2}. No standard deviations are presented for the TS structure as it corresponds to the only frame found by committor analysis [120].

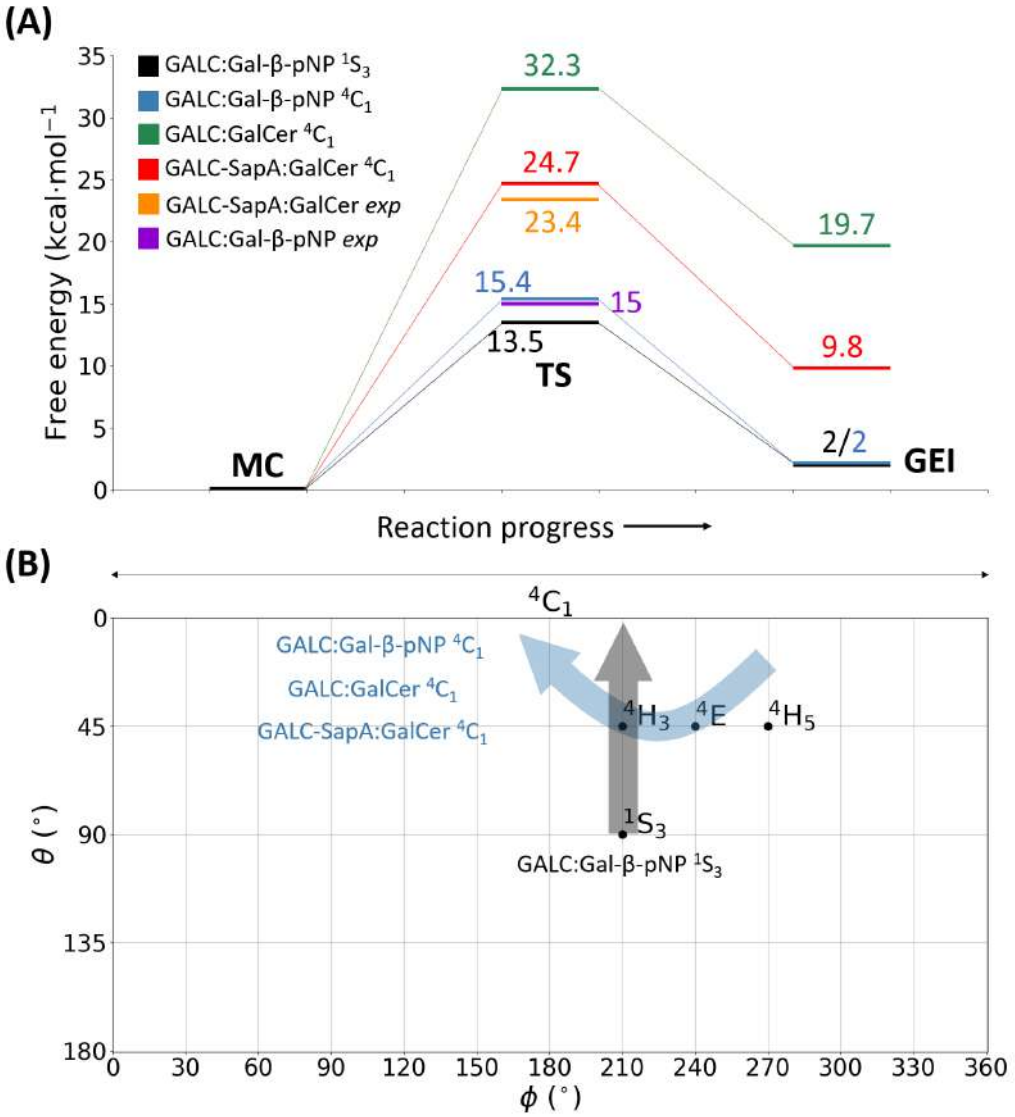


Figure 3.10: **(A)** Free energy barriers for the 4 simulated systems. **(B)** Conformational catalytic itinerary for all the simulated systems.

3.2.3.3 Analysis of water accessibility to the active site

Once the covalent glycosyl-enzyme intermediate is formed, the enzyme should hydrolyse it via an appropriately positioned active site water. This reaction step was not modelled since it is not rate-limiting. However, the accessibility of water molecules to the substrate anomeric carbon was analysed in order to check the feasibility of the reaction in all the systems analysed. The hydrolysis of the glycosyl-enzyme intermediate should not be difficult to achieve for Gal-β-pNP and GalCer in GALC-only as the active site is surrounded by water molecules which can easily

approach the acid/base residue to be deprotonated and attack C1. The situation completely changes in GALC-SapA dimer as the active site is now covered by the SapA protein. However, water is still able to reach the active site thanks to a water tunnel close to the polar groups of the ceramide tail placing the possible catalytic water close to the galactose moiety and the acid/base residue, see Figure 3.11.

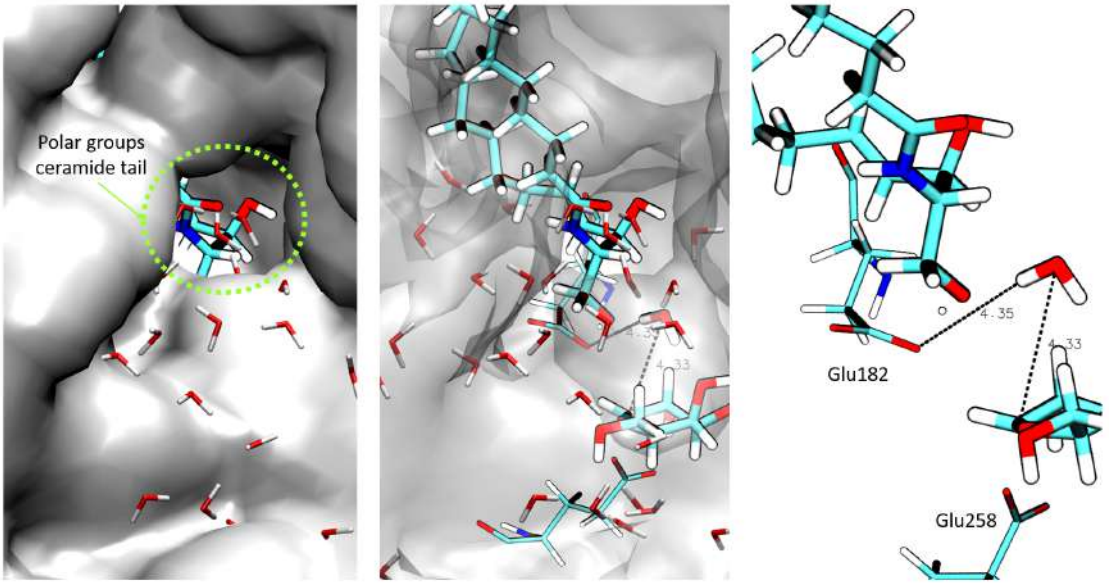


Figure 3.11: Different views of the water tunnel reaching the polar groups of the ceramide tail.

3.3 Summary and Conclusions

In this chapter we have studied the catalytic mechanism of β-galactocerebrosidase, GALC with two different substrates; Gal-β-pNP, an analogous substrate, and Gal-Cer, the natural substrate. Additionally, the effect of SapA, the protein that forms a complex with GALC *in vivo*, was investigated and compared to the GALC-only case. A crystallographic structure of the GALC:Gal-β-pNP complex with an unusual undistorted sugar conformation was the departure point in where we asked ourselves if such undistorted conformation was due to the non-optimal experimental conditions (a higher pH which caused a misorientation of the acid/base residue) or a functional feature induced by the GALC enzyme. To answer this question we computed the conformational free energy landscape of the Gal-β-pNP substrate in the active site of the enzyme in catalytic conditions; two conformational minima were found, the undistorted $^4C_1$ and a distorted one, $^1S_3$. To know if both conformations were catalytically competent we calculated the rate limiting step of the hydrolysis reaction departing from both MC conformations. The results obtained ($\Delta G = 15.4$

and 13.5 kcal·mol^{-1} for the reactions starting from $^4C_1$ and $^1S_3$, respectively) indicate that both conformational itineraries ($^1S_3 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$ and $^4C_1 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$) contribute to GALC catalysis. The cyclic $^4C_1 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$ itinerary, which is unprecedented in GHs, is probably observed because of the lack of steric determinants that keep the leaving group in place. Being an exo-GH, the leaving group is pointing towards the solvent with minimal contacts to the enzyme surface. After these striking results we wanted to understand if the same conclusions hold for the GALC natural substrate, GalCer, which has a large hydrophobic lipid tail in contrast to the smaller and polar 4-nitrophenol group of the Gal-β-pNP analogue. The conformational FEL of GalCer on the GALC active site shows that only the $^4C_1$ undistorted conformation "survive" in this case. However, the cyclic catalytic itinerary has a high energy barrier, $\approx$ 30 kcal·mol^{-1}. In views of these results, we thought that perhaps SapA, a protein domain complexed with GALC *in vivo*, would not only act as a substrate transport agent but also would help in stabilizing the large lipid tail and thus, could decrease the energy barrier needed for the first step of the hydrolysis reaction. Thus, taking advantage that a GALC-SapA structure was recently obtained, we designed a new simulation considering the complex of GalCer with GALC-SapA. The conformational FEL of GalCer on the active site of GALC-SapA dimer displays two minima, $^4C_1$ and $^1S_3$, recovering the results previously obtained for the complex of GALC with Gal-β-pNP. However, the undistorted $^4C_1$ conformation has a clear preference. The hydrolysis reaction follows a very similar mechanism as the one obtained for GalCer with only GALC, but now the reaction has a lower energy barrier, $\approx$ 24 kcal·mol^{-1} in good agreement with the experimental reaction rate ($\approx$ 23 kcal·mol^{-1}), thanks to the TS stabilization done by the SapA protein, by increasing the number of hydrophobic interactions and orientating the lipid tail so that it can create an extra hydrogen bond interaction with GALC. In summary, a cyclic conformational itinerary, in which the β-galactose starts from an undistorted conformation in the MC to reach again the $^4C_1$ conformation in the GEI via a $^4H_3$ TS is a functional feature of GALC, for Gal-β-pNP and GalCer substrates. The hydrolysis of GalCer is only possible in the presence of the SapA domain.

The conclusions reached in this chapter are the following:

- The protonation state and orientation of the acid/base residue do not change the conformation of Gal-β-pNP in GALC, however, the conformational FEL of Gal-β-pNP indicates that two substrate conformations, $^4C_1$ and $^1S_3$, are stable.

- Two catalytic itineraries, $^{1}S_{3} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$ and $^{4}C_{1} \rightarrow [^{4}H_{3}]^{\ddagger} \rightarrow {}^{4}C_{1}$, contribute to the reaction rate of GALC with Gal-β-pNP substrate, see Figure 3.10.

- The natural GalCer substrate on the active site of GALC can only adopt the undistorted $^{4}C_{1}$ conformation and the hydrolysis reaction exhibits a high energy barrier.

- The natural GalCer substrate can adopt two conformations ($^{4}C_{1}$ and $^{1}S_{3}$) in the active site of GALC-SapA, with $^{4}C_{1}$ being favoured.

- The SapA protein not only acts as a transport agent of the GalCer substrate from the cell membrane to the water soluble GALC active site, but also it helps decreasing the hydrolysis energy barrier to a catalytically competent value by stabilizing the reaction TS with a higher number of interactions between the GALC-SapA dimer and the GalCer substrate.

3.4 Computational details

3.4.1 Classical MD simulations

The 4 systems simulated are:

i. GALC in complex with Gal-β-pNP

ii. GALC in complex with GalCer substrate

iii. GALC-SapA dimer in the unligated form

iv. The GALC-SapA dimer in complex with GalCer

The same protocol was used in all classical MD simulations. First, a minimization step was performed stabilizing only the solvent composed by the water and different ion molecules, followed by a minimization of the substrate and then a final minimization step by allowing the whole system to relax. Afterwards, heating was performed gradually; keeping restrained the protein and substrate while the water and ions were allowed to move freely at 100K for a few ps, then the whole system can move, and the temperature was being increased bit by bit until the desired temperature of 300K was achieved. Later, the system density was converged up to the water density in the NPT ensemble. Finally, the simulation was extended in the NVT ensemble to achieve equilibrium and then the production phase was elongated between 50 to 150 nanoseconds. Analysis of the trajectories were carried out using VMD tools [121],

cpptraj (an AMBER package) [122] and the BINANA algorithm [123]. Images were processed with VMD (version 1.9.2) [121], Gnuplot [124] and matplotlib (module found in Python3) [125]. The specific parameters for each simulation are detailed below.

3.4.1.1 GALC in complex with Gal-β-pNP and GalCer

The initial structure for the simulations with Gal-β-pNP and GalCer substrates was obtained from the reported structure of GALC with PDB code 4CCC [84] in which a missing loop was inserted (residues 416-419) via the Chimera [126] program and FASTA [127] and the N-acetyl-glucosamine (NAG) molecules were removed. Finally, the protonation state of the protein residues was decided using the PropKa [128] program and manually inspecting the aminoacid environment taking into account the optimal pH for GALC which is around 5. The catalytically inactive position of the acid/base (Glu182) was exchanged by the one found in the glycosyl enzyme intermediate (GEI) structure (PDB code 4CCD) which seems in a better position for the reaction. For the simulation with GalCer substrate placed in the active site; the galactose moiety of GalCer was superimposed with the galactose moiety of Gal-β-pNP substrate. The coordinates of GalCer were obtained using the CHARMM-GUI web service [129,130]. The 2 systems were prepared using LeaP (a code integrated in the Amber simulation program [131]) adding the appropriate ions to counterbalance the total protein and calcium charges and solvating the whole systems with a cubic box of water molecules, a total number of 33354 water molecules and 7 sodiums ions to neutralize the protein were added. The simulations were performed using AMBER 11 [131] software and the force fields used for the protein, the galactose sugar and waters were FF99SB [132], Glycam06 [133] and TIP3P [134], respectively. The parametrization of the aglycon moieties, the 4-nitrophenol and the ceramide tail, were done using the antechamber package [135] and the GAFF force field [136].

3.4.1.2 GALC-SapA dimer unligated and in complex with GalCer

The initial structure for the simulations in the unligated form and bounded to GalCer was taken from the reported structure of the two subunits of the GALC-SapA dimer with PDB code 5N8K [89] in the apo form in which the N-acetyl-glucosamines (NAG), the detergent molecules (LDA) and a β-D-mannose were removed. The protonation states for the GALC subunits were taken as the ones used in the previous GALC subunit simulations and the titrable residues of the SapA protein were visually inspected. Molprobity web service [137] was also used to reassure the structure and the recommended flips were applied. The GalCer substrate was placed manually

in the active site of GALC-SapA dimer by superimposing the previously equilibrated structure of GALC-only bound to GalCer with the crystallographic structure of the unligated form of GALC-SapA dimer, and the ceramide tail was manually moved to be accommodated in the SapA hydrophobic cavity avoiding possible close contacts. The catalytic GALC subunit (the one with GalCer substrate placed in its active site) was chosen to be chain B (from the two subunits of GALC-SapA dimer) due to a crystallographic water placed in the active site interacting with Glu258 which was seen in previous simulations to be essential to stabilize and place the nucleophile in a reactive orientation. The 2 systems were also prepared using LeaP [131] adding the appropriate ions to counterbalance the total protein and calcium charges and solvating the whole systems with a cubic box of water molecules; in total, 52220 water molecules and 4 sodium ions were added. The simulations were performed using AMBER 14 [138] software and the force fields used for the protein, the galactose sugar and waters were FF99SB [132], Glycam06 [133] and TIP3P [134], respectively. The same ceramide tail parameters used in the previous simulations were also used.

3.4.2 *Ab initio* MD and metadynamics

3.4.2.1 β-D-galactose in vacuum

The isolated galactose (23 atoms) was computed with the Car-Parrinello approach [66], based on DFT, using the CPMD 3.15.1 program [76]. The galactose unit was enclosed in an orthorombic box of size 14.1 Å x 13.3 Å x 11.5 Å and the electronic structure was computed within the density functional approach (DFT) with the Perdew, Burke and Ernzerhoff generalized gradient-corrected approximation (PBE) [139] which has been proven to give a good performance in previous works on GHs and glycosyltransferases [140], besides being proved that it is the cheapest computational method which reaches chemical accuracy [141]. Kohn-Sham orbitals were expanded in a plane wave basis set with a kinetic energy cutoff of 70 Ry. Norm-conserving Trouiller-Martins pseudopotentials [142] were employed. The fictitious electronic mass and time step were set to 850 and 5 a.u. to ensure an adiabacity smaller than 10^{-5} a.u./atom for the fictitious kinetic energy. The conformational free energy landscape was calculated via the enhanced sampling method called metadynamics [81]; using the CPMD program and the Plumed driver [17]. The collective variables (CVs) used are the Cremer-Pople puckering coordinates θ and ϕ. The statistical error calculated with the free energy estimator developed by Tiwary [143] was below 0.5 kcal·mol^{-1}. The specific details are listed in Table 3.3 and Table 3.4.

3.4.2.2 GALC in complex with Gal-β-pNP and GalCer

The QM/MM calculations were performed using the method developed by Laio *et. al.* [77] in which the system is divided into a quantum region, where the atoms are treated within the Car-Parrinello MD and PBE functional approach as done previously for the isolated galactose; and a molecular mechanics region where the atoms are treated with the AMBER force field, as previously done in the classical MD simulations. The QM-MM boundary which can divide a covalent bond is treated by saturating the QM region within a MD formalism parametrizing a boundary atom described by a pseudopotential. The QM region included the Gal-β-pNP substrate and the catalytic residues (Glu182 and Glu258) until their beta carbon for the reaction simulation of Gal-β-pNP substrate; the galactose moiety and a part of the ceramide tail for the puckering simulation of GalCer substrate or the galactose moiety with a smaller part of the ceramide tail and the catalytic residues (Glu182 and Glu258) until their beta carbon for the reaction simulation of GalCer substrate (see Figure 3.12).

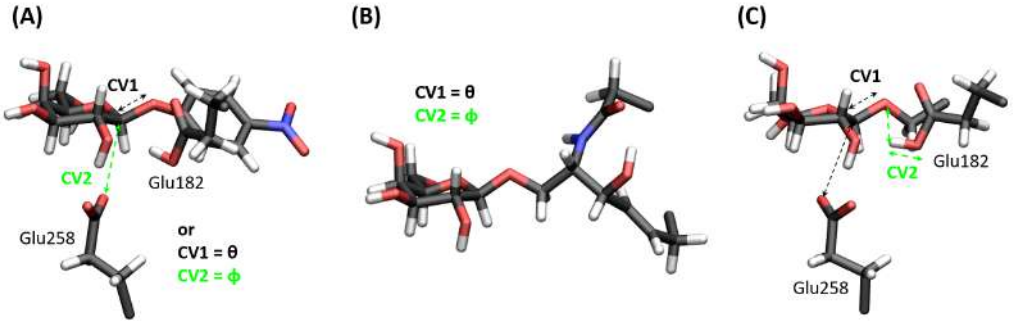


Figure 3.12: QM regions considered for the GALC subunit in the QM/MM MD simulations. **(A)** Gal-β-pNP substrate in the active site. Two separate sets of metadynamics simulations were done; one using the distance of the glycosidic bond as CV1 and the distance between the nucleophile Glu258 and the anomeric carbon of galactose as CV2, the other is using the polar puckering coordinates θ and ϕ. **(B)** GalCer substrate in the active site for the calculation of the conformational free energy landscape. The CVs are the puckering coordinates θ and ϕ. **(C)** GalCer substrate in the active site for the reaction metadynamics simulation. CV1 involves the bonds responsible for the nucleophillic attack, a distance difference of C1-O1' and O_{Glu258}-C1. CV2 involves the bonds responsible for the proton transfer, the distance difference between H_{Glu182}-O1' and H_{Glu182}-O_{Glu182}.

The conformational free energy landscapes of the studied substrates placed in the active site of GALC, Gal-β-pNP and GalCer, were calculated using the same method and software as with galactose in vacuum. The error of the principal minima of the reconstructed free energy surfaces [143] (standard deviation from the last 46 ps) is below 0.6 kcal·mol^{-1}. The specific values are listed in Table 3.3 and Table 3.4.

The rate limiting step of the hydrolysis reaction was also calculated for the two different substrates. The Gal-β-pNP hydrolysis reaction can depart from two different conformations, $^4C_1$ and $^1S_3$, and both itineraries were investigated. The CVs used in all simulations were chosen such as to take into account the main bonds to be broken or formed, Figure 3.12. The proton transfer from the acid/base residue to the leaving group was not directly included in the CV specification of Gal-β-pNP hydrolysis due to the relatively good leaving group character that the aglycon moiety has; a p-nitrophenol group (pKa = 7.15) [144], which made us think that the reaction would take place without leaving group protonation, oppositely to GalCer in which proton transfer needed to be included in the CVs set due to the bad leaving group character of ceramide. The energy barrier was further refined by launching several metadynamics simulations with different Gaussian heights (from 0.31 to 1.5 kcal·mol^{-1}) following Nair et. al. [145], with all other parameters (deposition time and Gaussian width) constant and starting from the half of the reactant basin of our previous simulations. The molecular mechanism remains the same independently of the Gaussian height. The values for each simulation are listed in Table 3.3 and 3.4.

3.4.2.3 GALC-SapA dimer in complex with GalCer

The same methodology was used for the dimer complex of GALC-SapA with GalCer in the active site. The conformational free energy landscape was calculated including in the quantum region only the galactose moiety and the minimal atoms needed of the lipid tail and using as collective variables the puckering coordinates θ and ϕ.

The rate limiting step of the hydrolysis reaction was also calculated departing from the two most stable conformers, $^4C_1$ and $^1S_3$. The quantum region used included the galactose moiety of the GalCer substrate, the minimal atoms of the lipid leaving group and the catalytic residues until their beta carbon (see Figure 3.12**C**). The same CVs for the hydrolysis reaction of GALC-GalCer were used. The statistical error of the free energy barrier was estimated by running three additional replicas with random velocities starting from half of the reactants basin and the same metadynamics parameters were applied. The specific values for each simulation are listed in Table 3.5. The reaction from the distorted conformer, $^1S_3$, was unsuccessful and thus not reported here.

	Type of simulation		
Parameters	**Vacuum**[*1]	**Gal-pNP puck**[*2]	**Gal-pNP $^4C_1$**[*3]
Quantum atoms	23	57	57
Functional	PBE	PBE	PBE
PW E_{cut} (Ry)	70	70	70
Pseudopotential	Norm-conserving Martin-Trouillers		
Fictitious electronic mass (a.u.)	850	700	700
Time step (a.u.)	5.0	5.0	5.0
Electrostatic regions (NN/ESP in a.u.)	-	20/38	20/38
Supercell size (Å)	14.1,13.3,11.5	18.7,17.6,18.4	18.7,17.6,18.4
CV1	θ	θ	C1-O1'
CV2	ϕ	ϕ	O_{Glu258}-C1
Gaussian width (a.u.)	0.1	0.1	0.1
Gaussian height (kcal·mol^{-1})	0.6/0.2	1/0.2	1/0.5
Deposition time (MD steps)	250	300	250
n$^{\underline{o}}$ deposited gaussians	9192	3680	567
Simulation time (ps)	276	133	17

Table 3.3: Parameters of the metadynamics QM/MM MD simulations of GALC in complex with Gal-β-pNP and GalCer.
[*1]Conformational FEL simulation of the isolated β-galactose
[*2]Conformational FEL simulation of Gal-β-pNP in the active site of GALC
[*3]Hydrolysis reaction simulation of Gal-β-pNP departing from the $^4C_1$ conformation

	Type of simulation		
Parameters	**Gal-pNP $^1S_3$*1**	**GalCer puck*2**	**GalCer reaction*3**
Quantum atoms	57	48	48
Functional	PBE	PBE	PBE
PW E_{cut} (Ry)	70	70	70
Pseudopotential	Norm-conserving Martin-Trouillers		
Fictitious electronic mass (a.u.)	700	700	700
Time step (a.u.)	5.0	5.0	5.0
Electrostatic regions (NN/ESP in a.u.)	20/38	20/38	16/24
Supercell size (Å)	18.7,17.6,18.4	15.4,17.4,18.1	17.8,19.6,14.2
CV1	C1-O1'	θ	(C1-O1')-(C1-O_{Glu258})
CV2	O_{Glu258}-C1	ϕ	(H_{Glu182}-O_{Glu182})-(H_{Glu182}-O1')
Gaussian width (a.u.)	0.1	0.1	0.1
Gaussian height (kcal·mol^{-1})	1/0.5	1	1(0.31/0.5/0.63)
Deposition time (MD steps)	250	300	250
n$^{\underline{o}}$ deposited gaussians	561	1144	4797
Simulation time (ps)	17	41	144

Table 3.4: Parameters of the QM/MM MD metadynamics simulations of GALC in complex with Gal-β-pNP and GalCer (continuation).
*1Hydrolysis reaction simulation of Gal-β-pNP departing from the $^1S_3$ conformation.
*2Conformational FEL simulation of GalCer in the active site of GALC
*3Hydrolysis reaction simulation of GalCer

Parameters	**GalCer puck**[*1]	**GalCer reaction**[*2]
Quantum atoms	27	48
Functional	PBE	PBE
PW E_{cut} (Ry)	70	70
Pseudopotential	Norm-conserving Martin-Trouillers	
Fictitious electronic mass (a.u.)	700	700
Time step (a.u.)	5.0	5.0
Electrostatic regions (NN/ESP in a.u.)	20/26	16/24
Supercell size (Å)	13.3,13.3,13.3	16.6,15.4,17.1
CV1	θ	(C1-O1')-(C1-O_{Glu258})
CV2	ϕ	(H_{Glu182}-O_{Glu182})-(H_{Glu182}-O1')
Gaussian width (a.u.)	0.04	0.1
Gaussian height (kcal·mol^{-1})	1	1/0.31
Deposition time (MD steps)	250	250
n$^{\underline{o}}$ deposited gaussians	3340	4066
Simulation time (ps)	110	122

Table 3.5: Parameters of the QM/MM MD metadynamics simulations of the GALC-SapA dimer in complex with GalCer.
[*1]Conformational FEL simulation of GalCer in the active site of the GALC-SapA dimer
[*2] Hydrolysis reaction simulation of GalCer

3.5 Supplemental Information

The isolated galactose conformational FEL was reconstructed obtaining that the most stable conformer is the $^4C_1$ chair conformation ($\theta = 0°$) followed by the inverted chair, $^1C_4$ ($\theta = 180°$), and two distorted conformations corresponding to $B_{3,O}$ and $^1S_3/^{1,4}B$ ($\theta = 90°$ and ϕ= 180° and 220°, respectively). The energy difference between the global minima $^4C_1$ and the other three minima are 3, 5.7 and 5.8 kcal·mol^{-1}, according to their stability. The convergence error in the energies was found to be lower than 0.5 kcal·mol^{-1}. The $^4C_1$ main minima is connected to the two distorted minima via a wide valley covering the $^2H_3$, E_3, $^4H_3$, ^4E and $^4H_5$ conformations in $\theta = 75°$. The second main minima, $^1C_4$ conformer, is connected to the distorted regions via the $^1H_2$ conformer in $\theta = 160°$. All rotations of the exocyclic groups, including the hidroxymethyl group, are explored for all the sugar configurations (Table 3.6). For the hydroxymethyl exocyclic group, the gg conformer is the most explored one in most of the sugar conformations, except $^4C_1$ and $^4H_3$ which have less than 50% of the population. Interestingly, all of the experimental structures fall around in the most stable regions of the conformational FEL ($^4C_1$) and the region between $B_{3,O}$, $^1S_3$ and $^{1,4}B$ (see Supplemental Table 3.7).

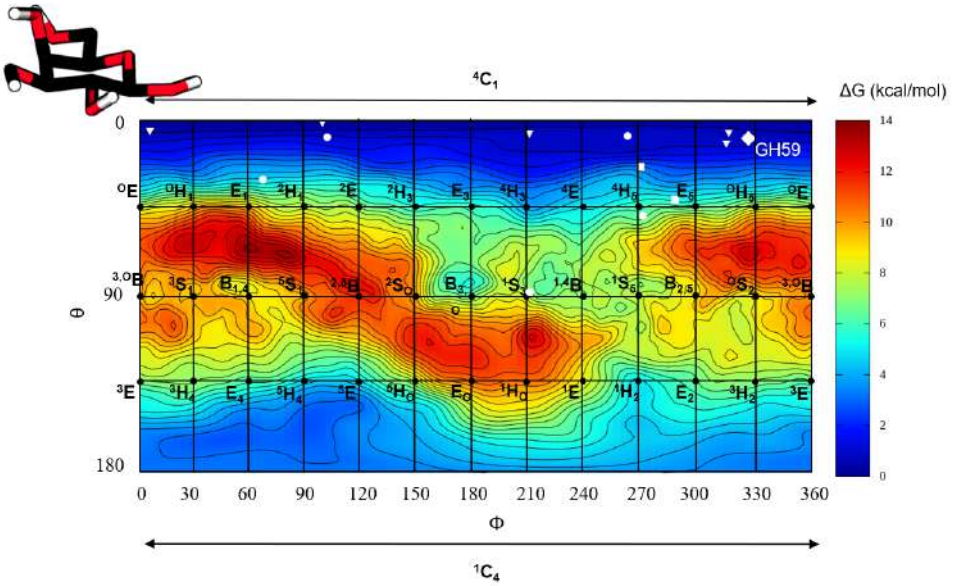


Figure 3.13: Isolated β-D-galactose conformational FEL (in a Mercator representation). Energy values are given in kcal·mol^{-1} and each contour line corresponds to 0.5 kcal·mol^{-1}. The MC conformations found in crystallographic structure of different galactosidases families are represented as an inverted triangle for GH2 family, as a square for GH30 family, as a rhombus for GH59 family (studied in this work) and as a circle for the inhibitors of GH59 family (Supplemental Table 3.7).

Conformation	**tg**	**gt**	**gg**
$^4C_1$	46.8	26.4	26.8
$^1C_4$	21.0	3.4	75.6
$^{1,4}B$	27.8	11.1	61.1
$B_{2,5}$	25.3	8.4	66.3
$^1S_5$	32.1	14.1	53.8
$^1S_3$	22.0	11.5	66.5
$B_{3,O}$	15.8	4.7	79.5
$^4H_3$	45.7	17.9	36.4
Total	24.9	8.1	67.0

Table 3.6: Populations (%) of the three possible orientations of the exocyclic group CH_2OH in the most stable conformers of the sugar ring obtained from the metadynamics simulation.

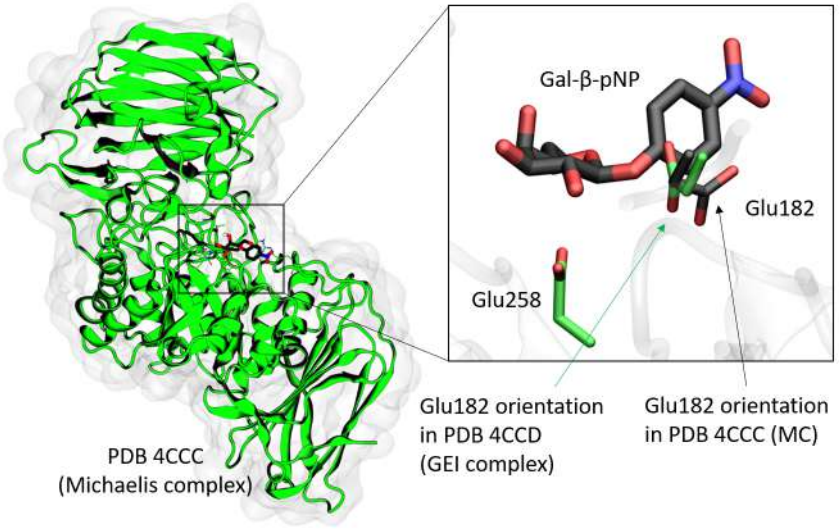


Figure 3.14: Structure of GALC from the Michaelis complex at pH 6.8 (PDB 4CCC) [84]. The acid/base residue (Glu182) points towards the solvent. The corresponding orientation in the structure of the GEI complex (trapped by enzyme soaking with D-galactal) is shown for comparison. The same orientation is observed in the structure of the unligated enzyme (PDB 3ZR5) [83].

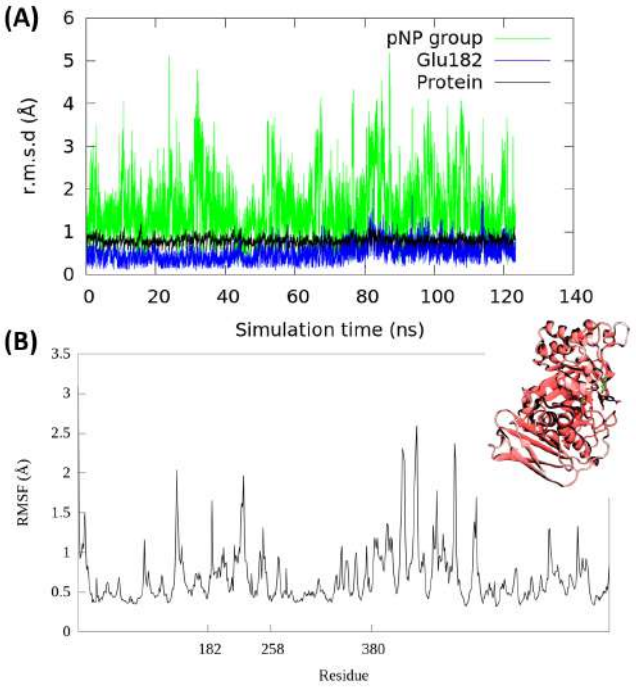


Figure 3.15: **(A)** r.m.s.d. of the protein backbone and the pNP leaving group (ignoring the hydrogen atoms) during the whole simulation. The protein is maintained stable (black) along the whole simulation, while the pNP leaving group (green) is more flexible as it is pointing towards the solvent. The acid/base catalytic residue is also maintained stable pointing towards the glycosidic oxygen the whole simulation (blue). **(B)** Cα r.m.s.f. graph and mapped on GALC surface. The color scale goes from the blue, representing the more mobile residues, to red, which are the one with the lowest r.m.s.f. value.

Family	**Proposed itinerary**	**MC**	**Inhibitors or Chaperones**	**GEI**	**Product**
GH1	$^{1,4}B \rightarrow [^4H_3/^4E]^{\ddagger} \rightarrow {}^4C_1$	-	$^4H_3$ (1uwt)	$^4C_1$ (1uwr)	-
GH2	$^1S_3 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$	$^4C_1$(1jyv, 4v44, 1jyx, 1jyw,1jyn)	$^4H_3/^4E$ (1jz6,3vd7,3vdb,1jz5)	-	$^4C_1$ (5ldr)
GH35	$^{1,4}B \rightarrow [^4H_3/^4E]^{\ddagger} \rightarrow {}^4C_1$	-	-	-	$^4C_1$ (3ogr, 1xc6, 4e8c, 4iug, 3thc)
GH42	$^1S_5 \rightarrow [^4E/^4H_5]^{\ddagger} \rightarrow {}^4C_1$	-	$^4C_1$ (4ucf)	-	$^4C_1$ (4uni,3tty,1kwk)
GH59	Unknown	$^4C_1$ (4ccc)	$^4C_1$ (4ufh, 4ufi, 4ufk); $^1S_3$ (4ufm); $^4H_5$ (4ufj); E_1 (4ufl)	$^4C_1$ (4ccd)	$^4C_1$ (4cce, 3zr6)

Table 3.7: Reported structures of exo-β-D-galactosidases, adapted from Speciale *et. al.* [34] and Kumar *et. al.* [115]

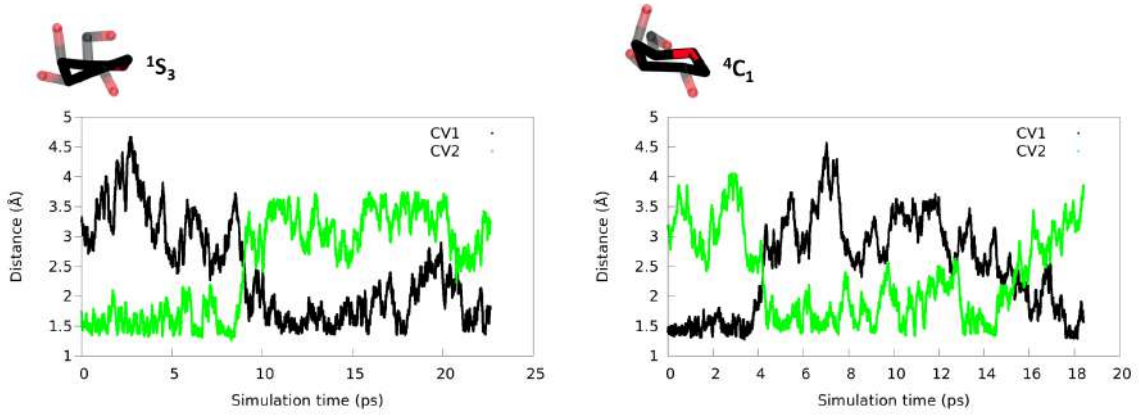


Figure 3.16: Evolution of the two collective variables in the metadynamics simulation for the rate limiting step of the hydrolysis reaction departing from both $^4C_1$ and $^1S_3$ conformers.

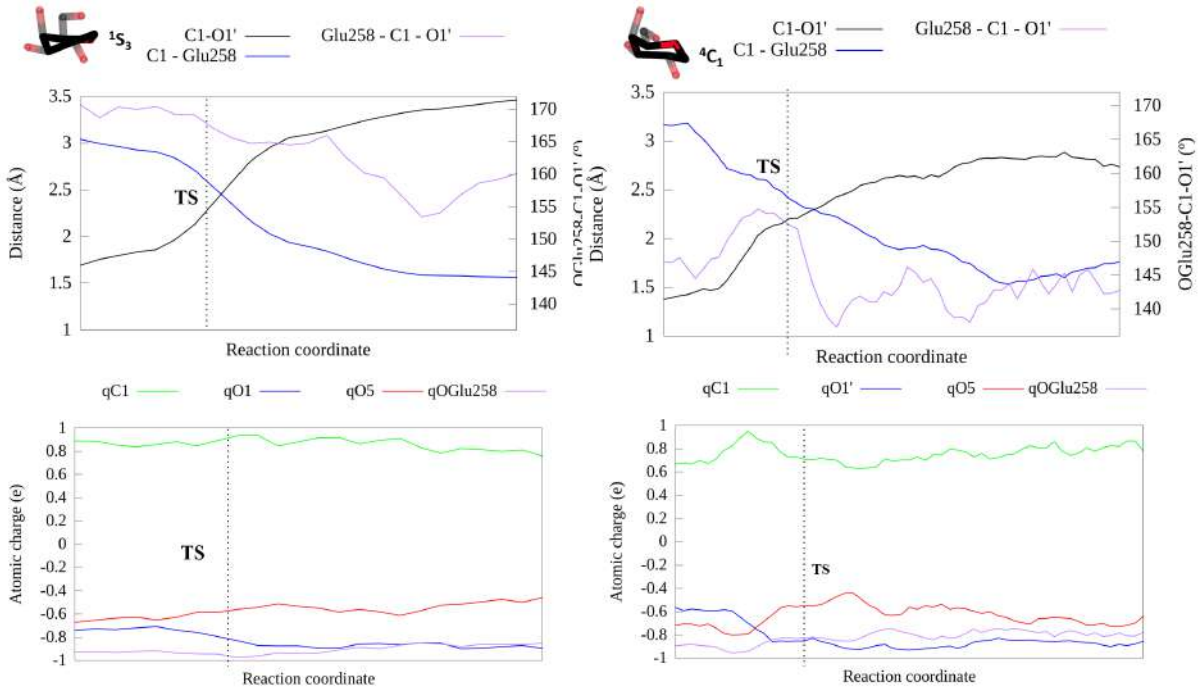


Figure 3.17: Variation of the relevant distances (top) and charges (bottom) along the reaction pathway departing from both $^1S_3$ (left) and $^4C_1$ (right) conformers. The TS is shown in the graph as a black dashed line. A running average over five data values is taken.

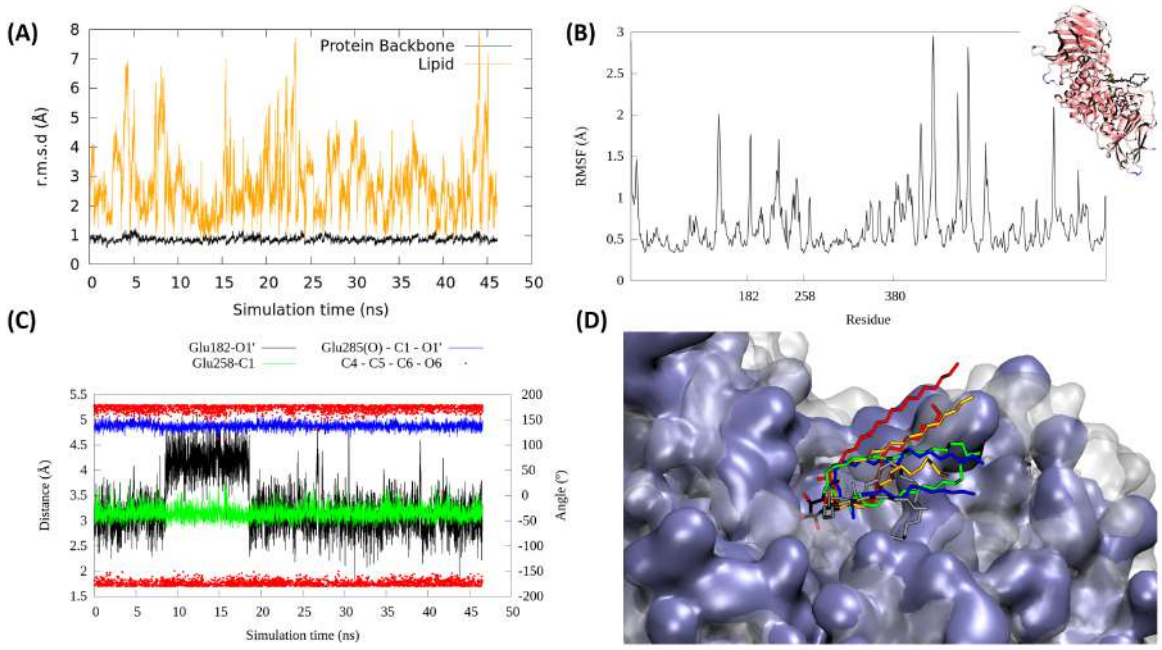


Figure 3.18: **(A)** r.m.s.d. of the protein backbone (black) and the ceramide tail (yellow, ignoring the hydrogen atoms) during the whole simulation. The protein is maintained stable along the whole simulation, while the ceramide tail is more flexible as it is pointing towards the solvent. **(B)** Cα r.m.s.f. graph and mapped on the GALC surface. The color scale goes from the blue, representing the more mobile residues, to red, which are the one with the lowest r.m.s.f. value. **(C)** Evolution of distances, angle and dihedral (placed in the y label in the right side) of some relevant interactions of the galactose moiety along the simulation. The hydroxymethyl group can only explore the gt conformer as seen for Gal-β-pNP, and the angle between the Glu258, the anomeric carbon and the oxygen of the glycosidic bond shows a similar value, approximately 140°, of Gal-β-pNP in $^4C_1$ conformation. **(D)** Representative structures of a cluster analysis (r.m.s.d. cutoff of 2Å which corresponds to the 18% of the total frames) of the ceramide moiety in red, orange, green, blue and grey. The protein surface is represented in transparent white and the apolar regions are represented as opaque iceblue.

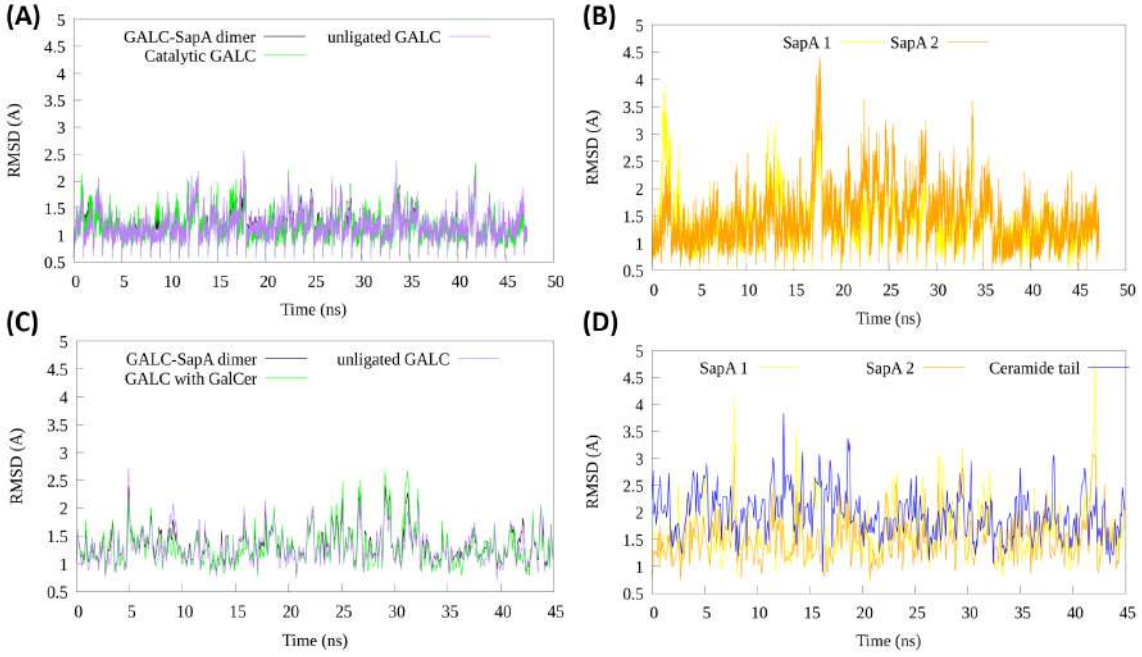


Figure 3.19: **(A)** Apo form system, r.m.s.d. of the two GALC subunits (green and purple) and the whole complex (black). In this case, the catalytic GALC (green) is referred to the subunit which should have the GalCer substrate bound to it. It can be seen that the system is maintained stable oscillating between 1-1.5 Å along the whole MD simulation. **(B)** Apo form system, r.m.s.d. of the two SapA subunits (orange and yellow) and the whole complex (black). Both SapA proteins are more flexible than the GALC subunits, oscillating around 2 Å . **(C)** GalCer bound to the active site system, r.m.s.d. of the two GALC subunits (green and purple). GalCer substrate in the active site seems to not affect the movement of the whole protein and it is also maintaned stable oscillating around 1-1.5 Å . **(D)** GalCer bound to the active site system, r.m.s.d. of the two SapA subunits (orange and yellow). Both SapA proteins are a bit more flexible than the GALC subunits, oscillating around 1.5 Å , thus being less mobile than in the apo form. In blue it can also be seen the r.m.s.d. of the ceramide tail which is also maintained stable along the whole MD simulation and oscillating around 2 Å , in contrast to the r.m.s.d. of the ceramide tail when only one subunit of GALC is present (Figure 3.18**(A)**) in where it was oscillating between 2 and 7 Å .

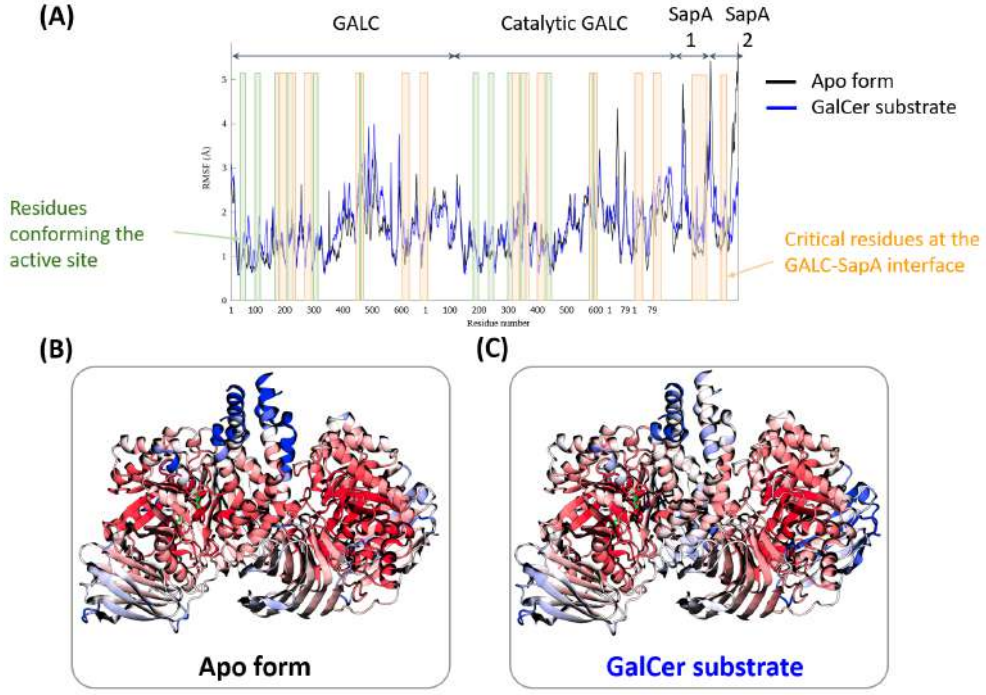


Figure 3.20: **(A)** Cα root mean square fluctuations (r.m.s.f.) of the GALC and SapA complex in the apo form (black) and bound to GalCer (blue). On the top of the graph, with arrows, are indicated the residues that correspond to the different subunits of GALC and SapA. The residues conforming the active site are marked with transparent green boxes and the residues involved in the GALC and SapA interaction are marked with transparent orange boxes. Generally, the apo form has residues which move more than when GalCer substrate is bound. **(B)** r.m.s.f. mapped on the apo form surface of GALC and SapA complex. The color scale goes from the blue, representing the more mobile residues, to red, which are the one with the lowest r.m.s.f. value. **(C)** r.m.s.f. mapped on the GALC and SapA complex with GalCer bound. The same color scale used previously is used.

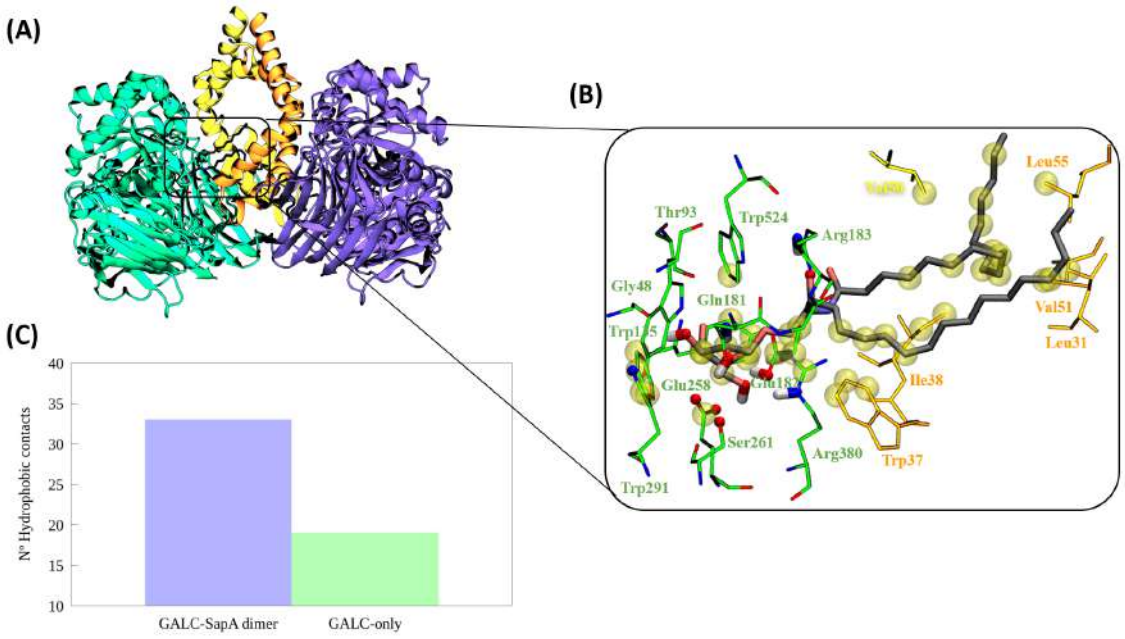


Figure 3.21: **(A)** The protein complex represented with its corresponding helixes and β-sheets colored in green (catalytic GALC), purple (the other GALC subunit), yellow and orange (the two SapA subunits). **(B)** Enlarged view of the active site and the GalCer substrate. Carbon atoms of the substrate are represented in black whereas the carbon atoms of GALC residues are represented in green and carbon atoms of SapA are represented in yellow and orange, oxygen and nitrogen atoms are colored red and blue, respectively. Only the residues with direct contact with the substrate are shown. Thicker atoms represent hydrogen bond interactions and transparent yellow surfaces represent hydrophobic contacts. **(C)** Comparison of protein-substrate hydrophobic contacts between GALC-only (green) and the GALC-SapA dimer (blue).

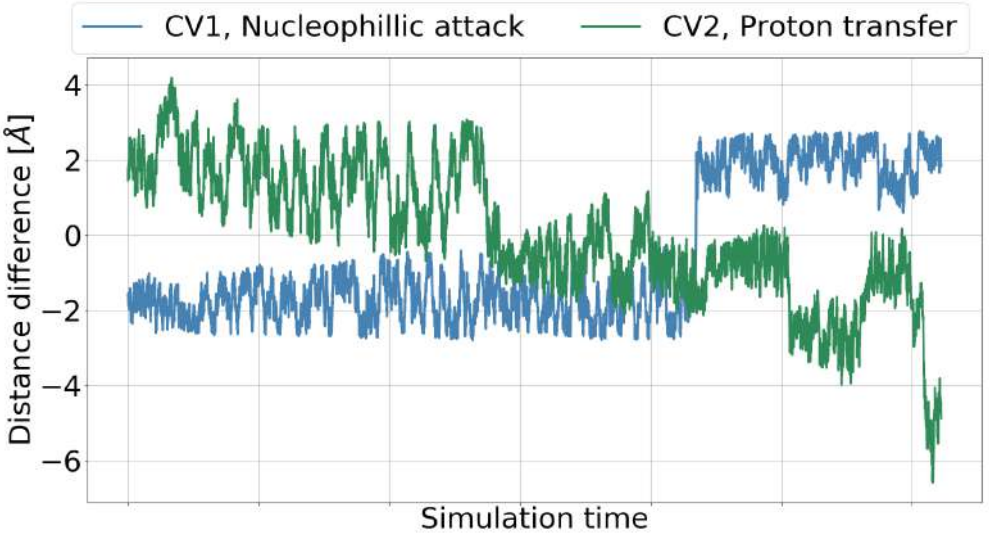


Figure 3.22: CVs evolution in the metadynamics simulation of the hydrolysis reaction of GalCer bound to the GALC-only.

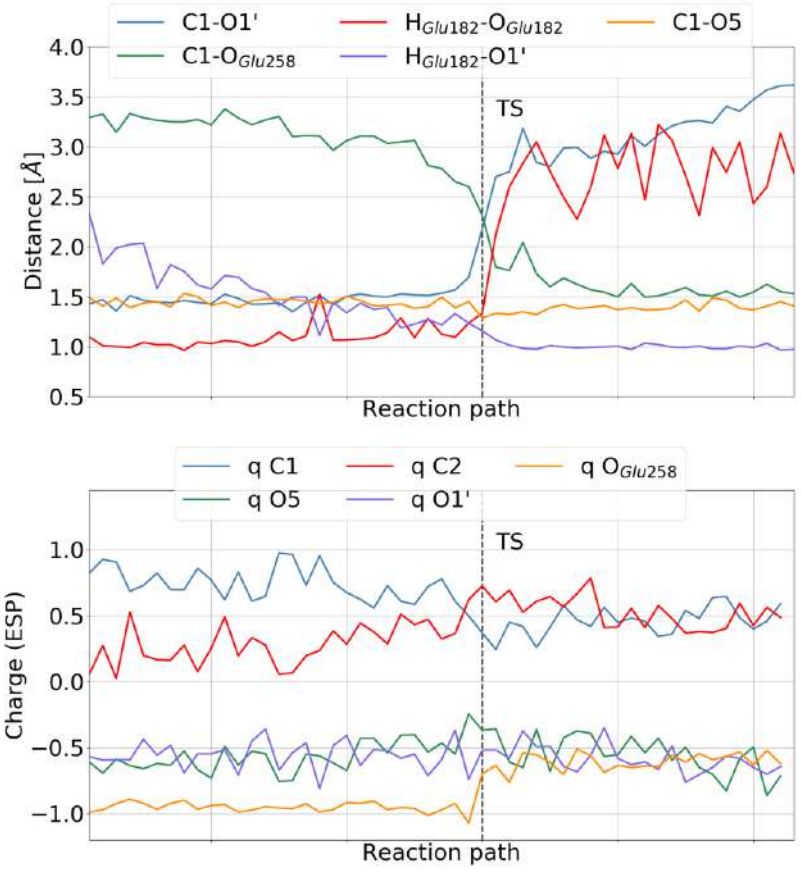


Figure 3.23: Variation of the relevant distance (top) and charges (bottom) along the lowest free energy path of the hydrolysis reaction of GalCer bound to the GALC-only.

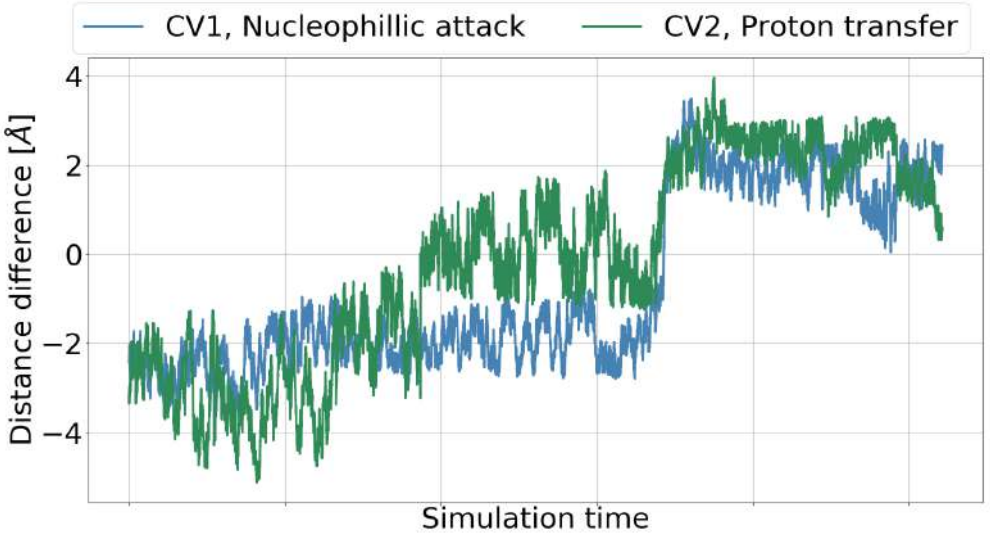


Figure 3.24: CVs evolution in the metadynamics simulation of the hydrolysis reaction of GalCer bound to the GALC-SapA dimer.

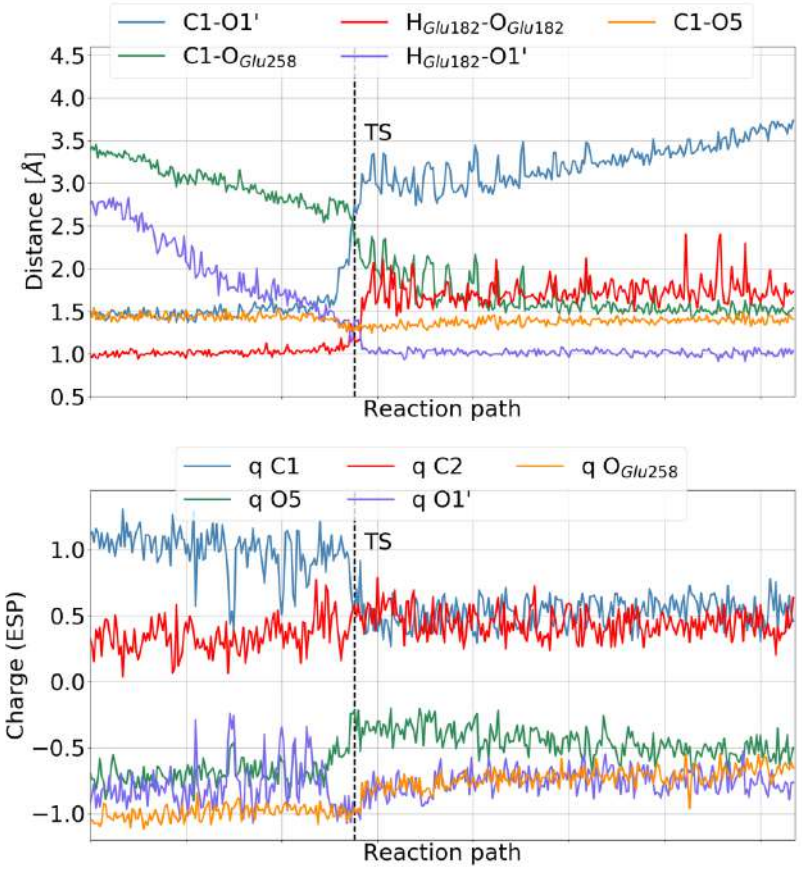


Figure 3.25: Variation of the relevant distance (top) and charges (bottom) along the lowest free energy path of the hydrolysis reaction of GalCer bound to the GALC-SapA dimer.

Chapter 4

Retaining α-L-arabinofuranosidases: structure, substrate conformations and catalysis

Parts of this chapter have been published in:
McGregor, N.[#]; Artola, M.[#]; Nin-Hill, A.[#]; Linzel, D.; Haon, M.; Reijngous, J.; Ram, A.; Rosso, M-N.; van der Marel, G.A.; Codée, J.D.C.; van Wezel, G.P.; Berrin, J-G.; Rovira, C.; Overkleeft, H.S.; Davies, G.J.; "Rational Design of Mechanism-Based Inhibitors and Activity-Based Probes for the Identification of Retaining α-L-Arabinofuranosidases" *JACS*, 142, 10, 4648-4662 (2020) [#]Equal contribution

ABSTRACT: The conformational free energy landscapes of isolated α-L- arabinofuranose and three covalent α-L-arabinofuranosidases inhibitors were analysed. The crystallographic structures of α-L-arabinofuranosidase, a glycoside hydrolase from family GH54, in complex with these inhibitors were obtained demonstrating that they are able to label efficiently the catalytic nucleophile. In addition, the first Michaelis complex crystallographic structure of a GH54 family member was obtained, using 4-nitrophenyl α-L-arabinofuranoside (PNP-Araf) as substrate. Finally, the enzyme catalytic reaction mechanism was studied using QM/MM metadynamics, revealing a $^{4}E/^{4}T_{o} \rightarrow [E_{3}]^{\ddagger} \rightarrow {}^{2}T_{3}$ conformational catalytic itinerary of the α-L-arabinofuranoside substrate. These results are in good concordance with the conformations observed in the structure of the Glu221Gln mutant in complex with PNP-Araf, as well as with inhibitor complex structures, which mimic the glycosyl-enzyme intermediate.

4.1 Introduction

As mentioned earlier in this Thesis, the carbohydrate-degrading machinery is a fundamentally important component of the metabolic systems that underpin the global carbon cycle. α-L-arabinofuranosidase is one of the enzymes involved in this machinery. Specifically, α-L-arabinofuranoside "side-chains" are usually found on both hemicellulosic and pectinaceous polysaccharides which are present in almost all plant cell walls. The efficient removal of α-L-arabinofuranose branches enhances the break-down of xylan-rich biomass [146]. Furthermore, α-L- arabinofuranosidases are an essential part of the polysaccharide utilization loci (PUL)[1], which ferment arabinan chains in dietary rhamnogalacturonan I and arabinogalactan within the human gut [147].

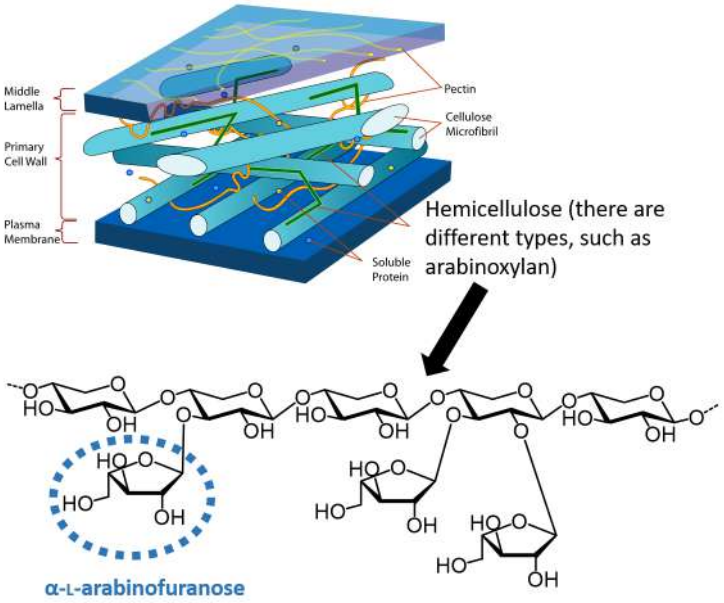


Figure 4.1: Plant cell wall diagram (Figure taken from Wikipedia, LadyofHats, 05/08/2020 via Wikimedia Commons, Creative Commons Attribution) in which one type of hemicellulose is drawn (arabinoxylan) and the sidechain of α-L-arabinofuranose is indicated with a blue circle.

Our understanding of these systems is dependent on an ability to identify the capacities of the carbohydrate-active enzymes produced by an organism. Inspired by the work of Withers [148–150] and Wright [151], our collaborators developed cyclophellitol-derived activity-based inhibitors and probes (ABPs) for the rapid detection and identification of specific biomass-degrading GHs within complex systems [152–154]. Previous reports on these ABPs have demonstrated their potential

[1] a set of physically-linked genes that orchestrates the breakdown of complex glycans.

as tools for the detection and identification of retaining GHs [155]. It has also been demonstrated that mimicking the half chair conformation of the enzymatic transition state, cyclophellitol and cyclophellitol aziridine derivatives react specifically with the catalytic nucleophile of a retaining GH, forming a non-hydrolyzable ester linkage through a ring-opening addition [156], see Figure 4.2**A**.

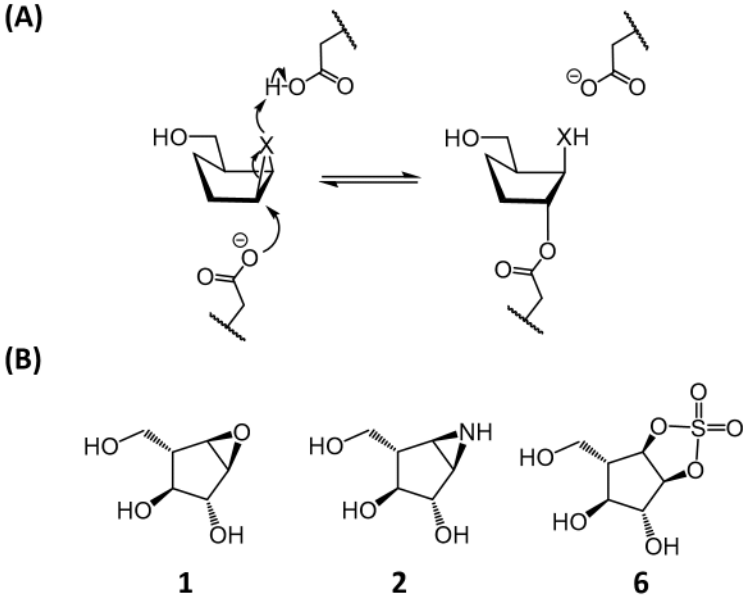


Figure 4.2: **(A)** General proposed mechanism for covalent intermediate trapping by cyclophellitol derivative. **(B)** Chemical structures of α-L-arabinofuranosidase inhibitors, hereafter named compounds **1** (cyclophellitol), **2** (aziridine) and **6** (cyclicsulphate), respectively; to be consistent with the naming in ref. [157]

This general strategy has been exploited to inhibit and label glycosidases displaying a variety of specificities including α- and β-D-glucosidases [158–160], β-D-glucuronidases [161], α- and β-D-galactosidases [162, 163], β-D-xylanases and β-D-xylosidases [164], among others. Thus, cyclophellitol-derived ABPs and inhibitors for α-L-arabinofuranosidases could be used to identify the enzymes responsible for the breakdown of a variety of complex polysaccharides, but it still remains unknown whether cyclophellitol derivatives can be effectively extended to target furanosidases.

Specific α-L-arabinofuranosidases have been identified within GH families 43, 51, 54, and 62, of which only families 51 and 54 follow the anomeric stereochemistry-retaining Koshland double-displacement mechanism, see Figure 1.4. The most detailed studies of α-L-arabinofuranosidase mechanisms have been performed using bacterial GH51 enzymes. Paes *et. al.* obtained the structure of an intact branched pentasaccharide substrate bound to the active site of TxAbf, a thermostable GH51 from *Thermobacillus xylanilyticus* (PDB ID 2VRQ) [165]. Hövel *et. al.* reported

the crystal structure of *Geobacillus stearothermophilus* AbfA (hereafter referred to as GsGH51) bound to 4-nitrophenyl α-L-arabinofuranoside (PNP-Araf) (PDB ID 1QW9) [166]. In both of these Michaelis complexes, the α-L-arabinofuranose ring was found in the ^{4}E conformation, see Figure 4.3. Therefore, we would like to know if, similar to most of GHs acting on pyranose sugars, furanosidases distort the -1 sugar to a conformation that is pre-activated for catalysis and thus, in a conformation close to the one found in the TS. In pyranose-acting GHs, it is well known that there are eight possible conformers that stabilize an oxocarbenium ion and thus, each of these conformations correspond to a possible TS for the hydrolysis reaction. The selected one depends on the GHs family and the substrate bound in the active site. In furanoses, the number of possible conformations that stabilize an oxocarbenium ion is reduced to only two, ^{3}E and E_3. Therefore, all the catalytic mechanisms of GHs acting on furanoses will go through one of them; which one depending on the active site architecture of the furanosidase enzyme.

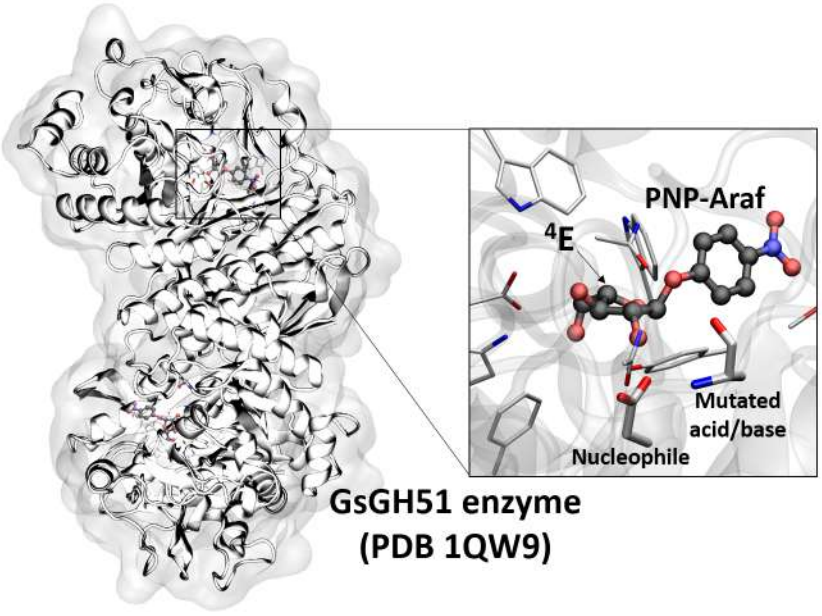


Figure 4.3: GsGH51 Michaelis complex crystallographic structure bound to PNP-Araf (PDB ID 1QW9) [166].

Beyond the bacterial GH51 enzymes, there is only one retaining α-L- arabinofuranosidase for which structural information of a substrate complex is available. The structure of *Aspergillus kawachii* AbfB (a member of GH54 hereafter referred to as AkAbfB) with arabinose in the active site (PDB ID 1WD4), displays a product complex ring conformation of ^{4}E [167]. Unfortunately, no Michaelis complex of this enzyme had been reported by the time this study was performed.

In this chapter, we will determine the ability of potential α-L-arabinofuranosidase inhibitors to mimic the putative MC/TS conformation in the enzyme active site in the two major families of retaining α-L-arabinofuranosidases, GH51 and GH54. With this aim, our collaborators designed a collection of putative α-L-arabinofuranosidase inhibitors (Figure 4.2**B**) and ABPs with different electrophilic traps and detection tags, which in turn we analized *in silico* for their ability to mimic the natural 5-membered ring structure, stereochemistry, and conformational itinerary of retaining α-L-arabinofuranosides. Inhibition kinetics measured with α-L-arabinofuranosidases from glycoside hydrolase families 51 and 54 were measured by our collaborators to validate our predictions. Then, the crystallographic structures of AkAbfB bound to the best inhibitors were obtained, validating the reactivity with the catalytic nucleophile. Also, the first Michaelis complex of the family GH54 was crystallized and the corresponding reaction mechanism was simulated and analysed. To the best of our knowledge, this is the first computed reaction mechanism of a furanoside-active GH.

My contribution to this study was to perform all theoretical simulations, fungal enzyme expression for α-L-arabinofuranosidases of GH51 and GH54 families, crystallizing AkAbfB enzyme and solving the X-ray structure of AkAbfB bound to **2** and **6** inhibitors and to PNP-Araf, together with Dr. McGregor (during a PhD secondment in the group of Prof. G. J. Davies).

4.2 Results and discussion

4.2.1 Conformational FEL of α-L-arabinofuranose, α-L- arabinofuranosidase inhibitors and α-L-arabinofuranosyl cation in vacuum

To gain insight into the ability of our potential inhibitors to mimic the natural conformational preferences of α-L-arabinofuranoside, we computed the relative energy of all ring conformations of α-L-arabinofuranose-configured inhibitors **1** (cyclophellitol), **2** (aziridine) and **6** (cyclicsulphate), as well as α-L-arabinofuranose. The conformational free energy landscape (FEL) of each molecule was calculated

using *ab initio* metadynamics and the Cremer-Pople puckering coordinates. This approach has recently been successful in predicting the performance of pyranose-like inhibitors [160, 164].

In contrast to GHs which act on pyranosides (e.g. β-galactosidases, found in the previous chapter), little is known about the catalytic conformational itineraries of α-L-arabinofuranosidases. The computed conformational FEL of α-L-arabinofuranose (Figure 4.5**A**) shows that all conformations lie in an energy window of ≈ 5 kcal·mol^{-1}. This window is significantly narrower than what is typical for pyranose compounds (e.g. a ≈ 15 kcal·mol^{-1} window energy was observed for β-galactose in Figure 3.13, page 72) [19, 21], and shows that most α-L-arabinofuranose conformations are thermally accessible. The most stable conformation is $^{1}T_{2}$. Its stability come from the ability to create a higher number of internal hydrogen bond interactions (between hydroxyls in C1 and C3 and the hydroxyl of C2 with the hydroxymethyl group) and a smaller steric hindrance between the ring hydrogens compared to the second minima found in the ^{2}E-^{4}E region, Figure 4.4. However, this conformation is not catalytically competent since the axial 2-OH group creates steric hindrance with the nucleophile residue located on the "beta" face of the sugar, see Figure 4.4**B**. Conformations between ^{2}E and ^{4}E, being only ≈ 2 kcal·mol^{-1} higher in energy, feature an equatorial 2-OH, eliminating such steric hindrance. Thus, the ideal Michaelis complex conformation for an α-L-arabinofuranosidase should be between ^{2}E and ^{4}E (shaded region in Figure 4.5**A**).

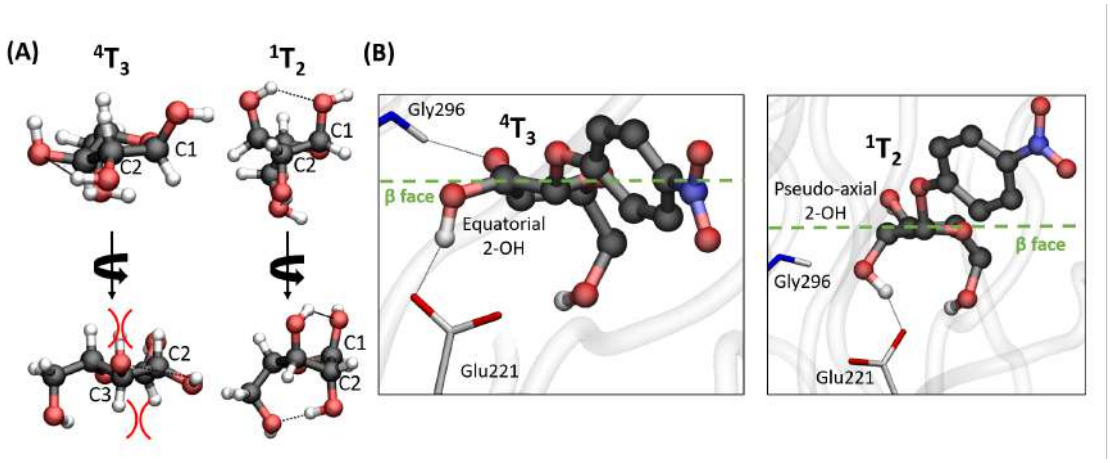


Figure 4.4: **(A)** α-L-arabinofuranose in $^4T_3$, a conformation within the ^2E-^4E region, and $^1T_2$ in vacuum. Hydrogen bond interactions are marked with black dashed lines and steric hindrance as a red inverse parenthesis symbol. **(B)** Comparison of PNP-Araf in $^4T_3$ and $^1T_2$ conformations in the active site of AkAbfB α-L-arabinofuranosidase. The pseudo-axial 2-OH group in $^1T_2$ creates such steric hindrance that does not permit the nucleophile to be in-line for the catalytic attack, whereas in $^4T_3$ this steric hindrance is minimized due to the equatorial 2-OH (see also panel A). Moreover, the hydrogen bond binding interaction between 3-OH and Gly296 disappears when the sugar adopts the $^1T_2$ conformation. These structures come from the QM/MM MD metadynamics simulation computing the substrate conformational FEL on-enzyme, Supplemental Figure 4.21 (page 114).

To determine where on this landscape the observed conformations of enzyme-bound species lie, we surveyed all the conformations of α-L-arabinofuranosidase complexes of GH51 and GH54 enzymes. As discussed above, only two MC structures exist for the GH51 family and in these structures the substrate is found in a ^4E conformation, marked in Figure 4.5**A** as red stars. The new GH54 MC structure also shows the same ^4E conformation. Thus, adopting also a distorted conformation in the enzyme active site, as previously observed in pyranose-acting GHs. Therefore, in order to be a suitable covalent α-L-arabinofuranosidase inhibitor, the molecule should readily adopt a ^4E conformation, in which the atom that mimics the anomeric carbon is accessible for nucleophilic attack from the β face of the sugar ring. Computed conformational FELs for compounds **1**, **2**, and **6** (Figure 4.5**B**) show that these conformations are energetically favoured for **6**, whereas **1** and **2** instead prefer conformations in which the 2-OH is axial (in the $^1T_2 - E_2 - {}^3T_2$ region). Thus, cyclic sulphate **6** was anticipated to be a potentially stronger inhibitor than the epoxide (**1**) or aziridine (**2**) for both GH51 and GH54 α-L-arabinofuranosidases.

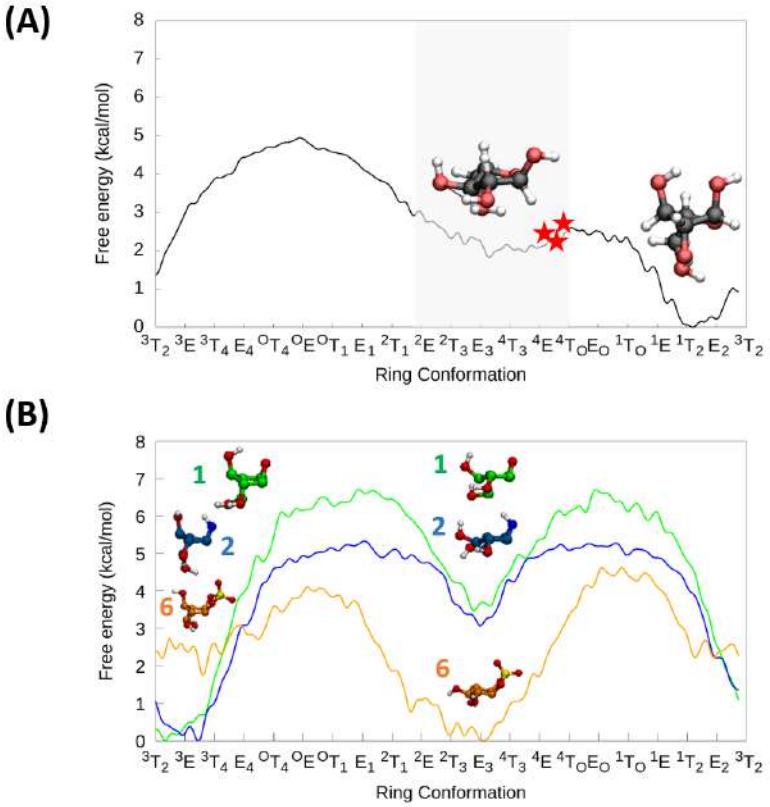


Figure 4.5: **(A)** Conformational FEL of isolated α-L-arabinofuranose. Conformations observed in Michaelis complexes of α-L-arabinofuranosidases are represented with a red star (PDB 2VRQ and 1QW9 for GH51 and PDB 6SXR, this work, for GH54). The conformational region having an equatorial 2-OH is shaded. **(B)** Conformational FEL of α-L-arabinofuranose-configured cyclophellitol (**1**), aziridine (**2**) and cyclic sulphate (**6**).

The conformational FEL of the α-L-arabinofuranosyl cation was also computed in order to determine the ideal conformation for the transition state of the glycosylation reaction of α-L-arabinofuranosides, which is expected to have oxocarbenium ion character. There are only two main minima, located around E_3 and ^3E, as previously predicted, see Figure 4.6**B**. The ^3E conformer is more stable than E_3 by $\sim$ 6 kcal·mol^{-1} due to a higher number of intramolecular interactions and smaller steric hindrance between ring hydrogens, Figure 4.6**A**. A more detailed analysis via calculation of a 2 dimensional FEL, ϕ vs the hydroxymethyl group dihedral angle, shows that the less stable E_3 conformer is the most stable one for the hydroxymethyl in a gg orientation. The gg orientation is the one observed on-enzyme due to the specific active site architecture. Together with the fact that in the computed conformational FEL of PNP-Araf in the active site of AkAbfB all the conformers around the ^3E conformer are destabilized in energy (see Supplemental Figure 4.21, page 114), because they lack the important hydrogen bond interactions observed in the

crystal structure, it can be concluded that even though globally ^{3}E in vacuum is more stable than E_3, the TS of α-L-arabinofuranosidase AkAbfB is more likely to adopt an E_3 conformation.

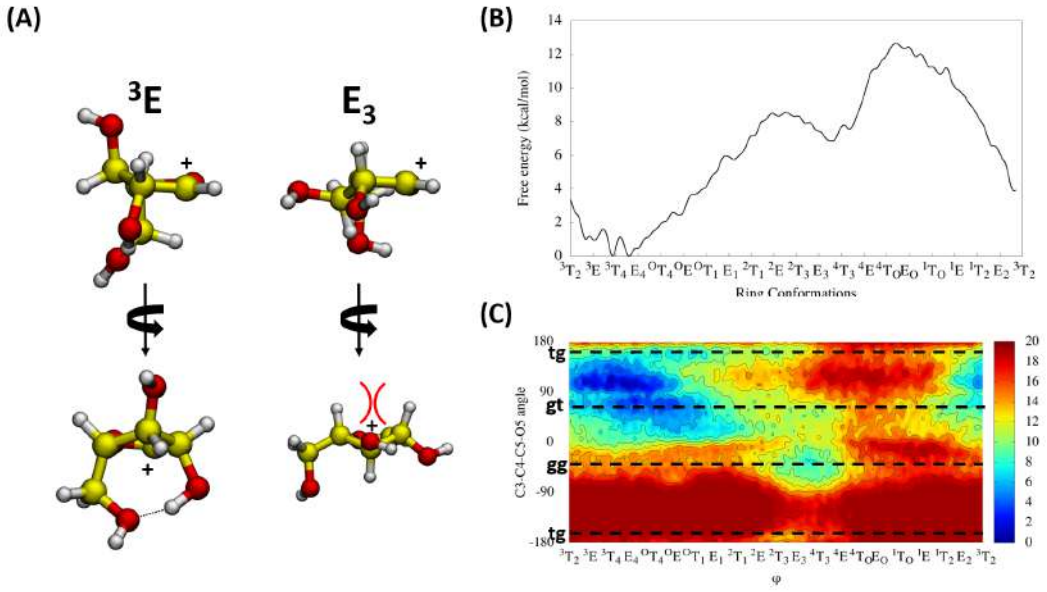


Figure 4.6: **(A)** α-L-arabinofuranosyl cation in ^{3}E and E_3 conformations in vacuum. Hydrogen bond interactions are marked with black dashed lines and steric hindrance as as a red inverse parenthesis symbol. **(B)** Conformational FEL of the α-L-arabinofuranosyl cation projected in one dimension (ϕ). **(C)** Conformational FEL of the α-L-arabinofuranosyl cation projected in two dimension (ϕ and the hydroxymethyl group dihedral angle). Isolines at 2 kcal·mol^{-1}.

4.2.2 Structural analysis of AkAbfB α-L-arabinofuranosidase bound to PNP-Araf and 2 and 6 inhibitors

To obtain the Michaelis complex, crystals of deglycosylated AkAbfB Glu221Gln were soaked in a saturated solution of PNP-Araf in mother liquor. The resulting 1.64 Å crystal structure contained 3 PNP-Araf molecules: two full occupancy molecules bound to the carbohydrate binding module and a partial occupancy molecule bound in the active site (Figure 4.7).

Overall, the Michaelis complex shows that O2 forms hydrogen bonds with the carbonyl oxygen of Gln221 and the backbone amide of Asp297. O3 forms hydrogen bonds with the backbone amide of Gly296 and the carboxylate of Asp219. The ring oxygen forms a hydrogen bond with the backbone amide of Asn222 and O5 formed hydrogen bonds with the carboxylate of Asp219 and the backbone amide of Asn223 displaying a gg hydroxymethyl orientation. The furanose ring is found in a ^{4}E conformation (ϕ=225.9°), stacked against a hydrophobic surface formed

by Trp206 and the Cys176-Cys177 disulfide linkage. The axial nitrophenyl leaving group points out of the active site into a solvent channel. The electrophilic carbon (C1) is positioned 3 Å away from the amide nitrogen of Gln221.

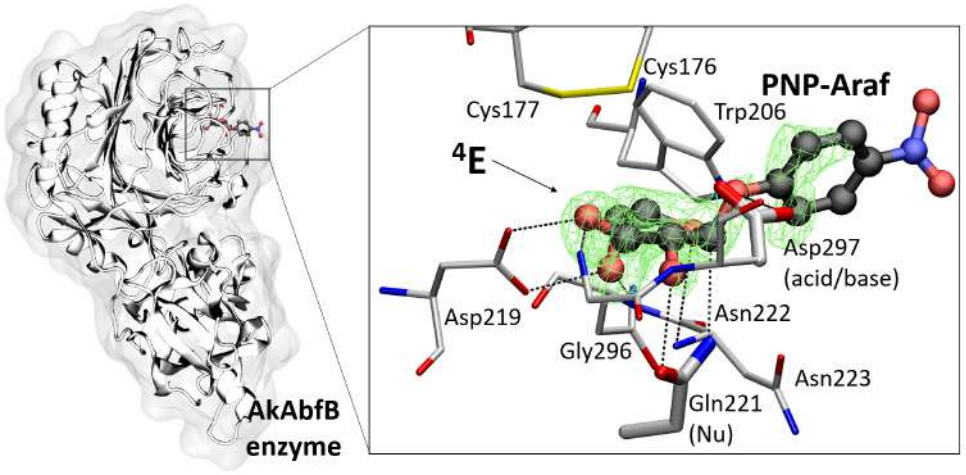


Figure 4.7: Michaelis complex structure of AkAbfB (Glu221Gln) bound to PNP-Araf substrate. The electron density of the PNP-Araf molecule (black) within the AkAbfB (Glu221Gln) active site (grey) is contoured to 1σ (green mesh). The catalytic residues (Asp297 and Glu221, mutated to Gln) are shown with thicker lines.

To generate covalent complexes, crystals of AkAbfB were soaked with 0.2 mM of inhibitors **1**, **2** and **6** for 1 h. A complex structure of AkAbfB with inhibitor **1** was not able to solve but it was possible for both inhibitors, **2** and **6**, bound to Glu221 in almost identical positions and conformations (Figure 4.8), forming hydrogen bonds from O2 to Gly296 and the sulphur of Met195, from O3 to Asn297 and Asp219, and from O5 to Asp219 and Asn223. The interaction between the ring oxygen and Asn222 found in the product complex (PDB ID 1WD4) cannot be formed, but the axial amine presents an additional hydrogen bond with Asp297, the general acid/base. The interactions between both inhibitors **2** and **6** and the active site of AkAbfB cause no significant change in the protein structure. The active site appears to be sufficiently open to accommodate the sulphate of inhibitor **6** without any steric clashes. Thus, these complexes could be good representations of the glycosyl-enzyme intermediate (GEI) structure, further confirmed once the first step of the hydrolysis reaction is simulated, explained below. The ring in each covalent complex is found in the ^{2}E conformation (**2**, ϕ=166.0° and **6**, ϕ=168.5°). This consensus conformation represents a 1.2 Å migration of C1 from its position in the AkAbfB Michaelis complex toward Glu221 coupled with a ∼15° axial rotation of the ring around C3.

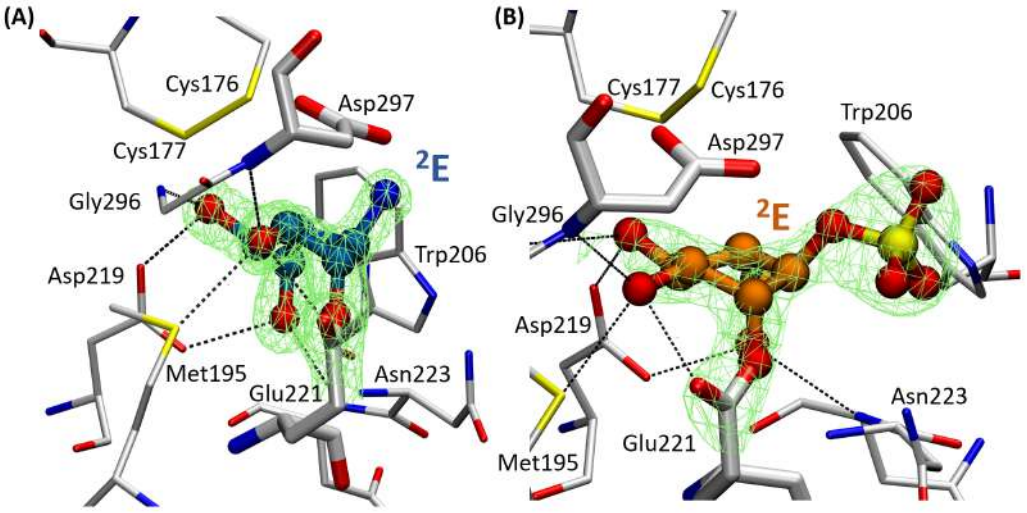


Figure 4.8: Crystal structures of the covalently-bound inhibitors **2** (panel A, coloured in blue) and **6** (panel B, coloured in orange) in AkAbfB enzyme. 2F0 - Fc electron density is shown for both the ligand and the catalytic nucleophile as a green mesh contoured at 1σ. Active site residues are shown as gray sticks. Apparent hydrogen bonding interactions are shown as dotted black lines.

4.2.3 Inhibition of 1, 2 and 6 compounds in AkAbfB enzyme

Inhibition kinetics were measured by our collaborators for the potential **1**, **2** and **6** inhibitors in AkAbfB. Initial overnight incubations of each enzyme with compounds **2** and **6** resulted in the complete loss of activity, while no loss of activity was observed with compound **1** explaining the lack of a complexed crystal structure. As predicted by our conformational analysis, inhibitor **6** reacts rapidly with the catalytic nucleophile of AkAbfB with a k_{inact} well above 1 min^{-1} (estimated from the limited speed of their assay). However, the lack of any apparent nonlinearity in the k_{app} vs [I] curve suggests poor initial binding, see ref. [157]. In spite of this, it has a performance constant of 240 M^{-1} s^{-1}. Contrary to our prediction, compound **2** also proved to be a potent inhibitor of AkAbfB, having performance constants only 8-fold lower than inhibitor **6** with AkAbfB (28 M^{-1} s^{-1}). In contrast to inhibitor **6**, inhibition kinetics with inhibitor **2** provided evidence of stronger initial binding in the enzyme active site, having KI values of 0.3 mM. As predicted from our conformational analysis, compound **1** is not an inhibitor of retaining α-L-arabinofuranosidases at concentrations up to 0.25 mM.

The inhibition experiment results are supported by the trends observed in the binding energies calculated for the *in silico* built Michaelis complexes (methodology explained in section 4.4.3.1 in page 104) of inhibitor **1**, **2**, and **6** in AkAbfB active site (Figure 4.9). The binding energy of inhibitor **2** was calculated in 3 different situations in each active site: deprotonated inhibitor **2** with protonated acid/base

residue, protonated inhibitor **2** with protonated acid/base residue, and protonated inhibitor **2** with deprotonated acid/base residue. Protonated compound **2** (with the optimal acid/base residue protonation) binds better than all the other compounds (Figure 4.9**F**).

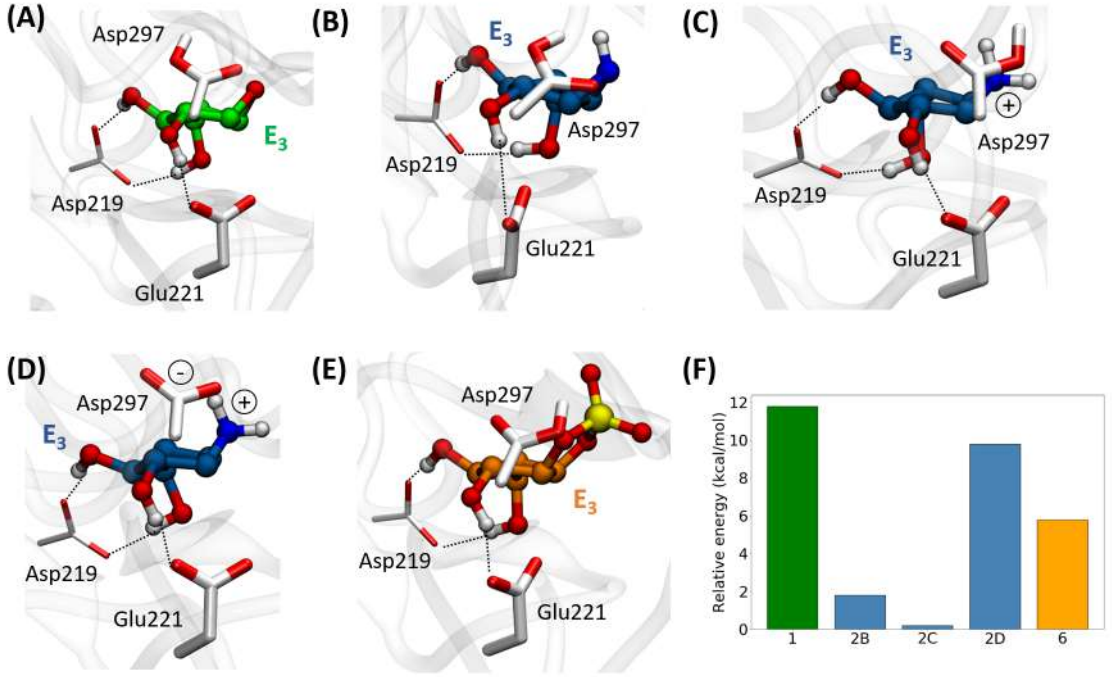


Figure 4.9: Simulated Michaelis complex configurations of inhibitors **1 (panel A)**, **2 (panels B, C and D)**, and **6 (E)** used for the calculation of the binding energy within the active site of AkAbfB. Panel **B** shows the protonated general acid/base with deprotonated aziridine nitrogen, **C** shows the protonated general acid/base with protonated aziridine nitrogen, **D** shows the deprotonated general acid/base with protonated aziridine nitrogen. **(F)** Relative binding energies of the inhibitors into the active site of AkAbfB. Compounds **2B**, **2C** and **2D** corresponds to the different tested protonation states.

4.2.4 Reaction mechanism of AkAbfB α-L-arabinofuranosidase

In order to decipher the reaction mechanism of AkAbfB enzyme and in particular analyze the conformational itinerary of the arabinofuranoside substrate, we reverted the nucleophile mutation present in the MC crystallographic structure bound to PNP-Araf; from a Gln to Glu, *in silico*. After performing a classical MD simulation to equilibrate the new system, the chemical reaction was computed by means of a QM/MM MD metadynamics simulation in where two collective variables (CVs), corresponding to the main covalent bonds that are formed or cleaved during the reaction, were used to drive the system from reactants (the MC) to products (the

glycosyl-enzyme intermediate, GEI). The first collective variable (CV1), named "glycosidic bond cleavage", quantifies the cleavage of the glycosidic bond. The second one (CV2), named as "nucleophilic attack" accounts for the formation of the covalent bond between the anomeric carbon of the α-L-arabinofuranose and the closest oxygen atom of the catalytic nucleophile (Glu221).

There are two main minima in the chemical reaction FEL corresponding to an ensemble of MC structures (in the left-hand side of Figure 4.10**A**) and an ensemble of GEI structures (in the right hand-side of the same figure). The TS (confirmed with a committor analysis [120]) is found in-between both minima, displaying a distance of 2.92 Å between the anomeric carbon of PNP-Araf and Glu221, corresponding to CV1, and a distance of 3.29 Å between the anomeric carbon and the glycosidic oxygen, corresponding to CV2. The experimental kinetics measured for PNP-Araf in optimal conditions (pH 5 at 37°C), data from our collaborators, gives a K_M of 1.4 mM and a Kcat corresponding to 5700 min^{-1}, which would give an energy barrier of $\approx$ 15 kcal·mol^{-1}. The energy barrier found computationally gives a value of 17.8 $\pm$ 1.1 kcal·mol^{-1}, which it is a little bit higher compared to the experimental value, but still in the range of predictions of free energy barriers in DFT.

The most relevant bond distances that characterize the reaction mechanism are listed in Table 4.1, in which the properties for each state were computed from all the configurations falling into a small region ($\pm$0.05 and $\pm$0.05 in terms of the CV) around the corresponding point in the FEL.

	MC	**TS**	**GEI**
C1 - O1' (Å)	1.53 ± 0.09	3.29	3.80 ± 0.09
C1 - O_{Glu221} (Å)	3.94 ± 0.08	2.92	1.47 ± 0.08
O_{Asp297} - H_{Asp297} (Å)	1.02 ± 0.03	0.98	1.05 ± 0.04
H_{Asp297} - O1' (Å)	3.29 ± 0.14	3.09	1.98 ± 0.65
C1 - O4 (Å)	1.40 ± 0.04	1.29	1.41 ± 0.04
C1 (RESP charge)	0.52 ± 0.06	0.36	0.92 ± 0.13
O4 (RESP charge)	-0.66 ± 0.04	-0.26	-0.68 ± 0.12
ϕ (°)	244 ± 11	206.4	184 ± 7

Table 4.1: Structural and electronic parameters of the characteristic points along the reaction pathway. There are no standard deviations for the TS structure as it corresponds to the only frame found by committor analysis [120].

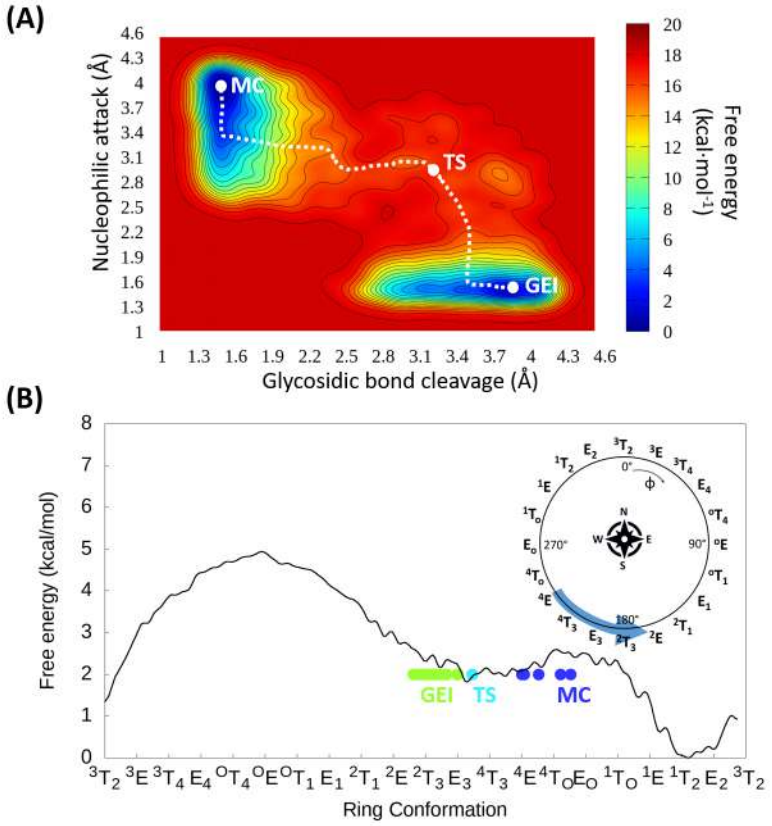


Figure 4.10: **(A)** Free energy landscape of the computed first step of the hydrolysis reaction in where the dashed line corresponds to the lowest free energy path, computed using MEPSAnd [116]. **(B)** Conformational catalytic itinerary of the lowest free energy path projected onto the conformational FEL of the isolated α-L-arabinofuranose.

The reaction mechanism of AkAbfB α-L-arabinofuranosidase corresponds to a dissociative mechanism, in which the glycosidic bond is cleaved before the nucleophile approaches the anomeric carbon (see Table 4.1, C1-O1' is 3.29 Å and C1-O_{Glu221} is 2.92 Å at the TS). As observed for β-galactocerebrosidase in the previous chapter, the p-nitrophenol group does not need proton transfer from the acid/base catalytic residue to be able to react. The conformation of the arabinofuranose ring of the PNP-Araf substrate in the MC changes very little compared to the mutated Glu221Gln crystal structure, from a ^{4}E conformation in the crystal structure to a conformation in-between ^{4}E and ^4T_O in the computed structure (see Figure 4.10**B**). This slight difference can probably be attributed to the Glu221Gln mutation of the crystal structure. Thus, in this case the nucleophile mutation does not change substantially the MC conformation of the substrate. This is in contrast with GH complexes with pyranose-type substrates, in which mutation of the nucleophile, a negatively charged residue that strongly interacts with the sugar hydroxyl residues,

can change the sugar conformation significantly [168]. As in pyranose-acting GHs, the reaction goes through a TS with significant oxocarbenium ion-like character in an E_3 conformer, as previously hypothesized. Finally, the GEI intermediate forms, in which the arabinofuranose ring adopts a $^2T_3$ conformation. Therefore, the conformational catalytic itinerary of AkAbfB α-L-arabinofuranosidase is ^4E / $^4T_O \rightarrow [E_3]^{\ddagger} \rightarrow {}^2T_3$.

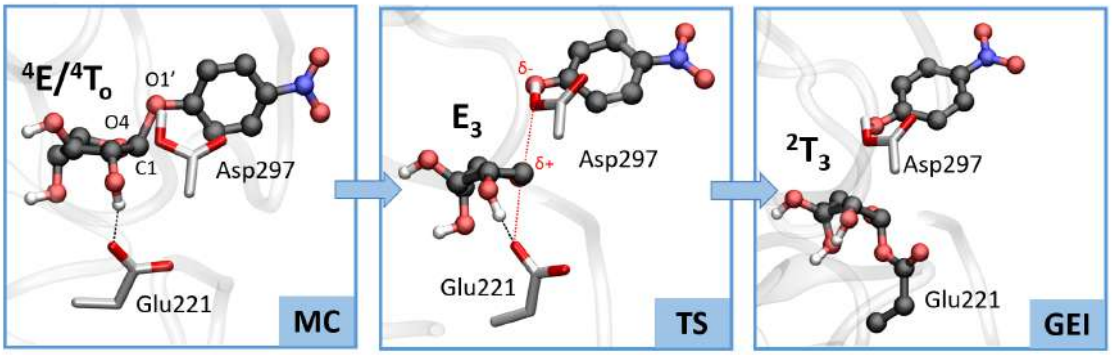


Figure 4.11: Representative structures along the catalytic itinerary of AkAbfB α-L-arabinofuranosidase.

The $^2T_3$ conformation of the GEI intermediate is very close to the ^2E conformer found in the crystal structures of inhibitors **2** and **6** (Figure 4.12), which also react with the nucleophile residue, forming a covalent complex. Practically all the protein interactions are the same between the inhibitor bound complexes and the computed GEI with PNP-Araf. Comparing the crystal structures corresponding to the mutated MC and the protein-bound inhibitors, it can be observed that one main difference between them is the hydrogen bond between Met195 and 2-OH present in the latter but not in the former (Figures 4.7 and 4.8). Along the reaction mechanism one can observe how this hydrogen bond is formed just after crossing the TS and when the GEI complex is fully formed. We can conclude that the structures of AkAbfB complexes with inhibitors **2** and **6** are good representations of the GEI structure of GH54 α-L-arabinofuranosidase.

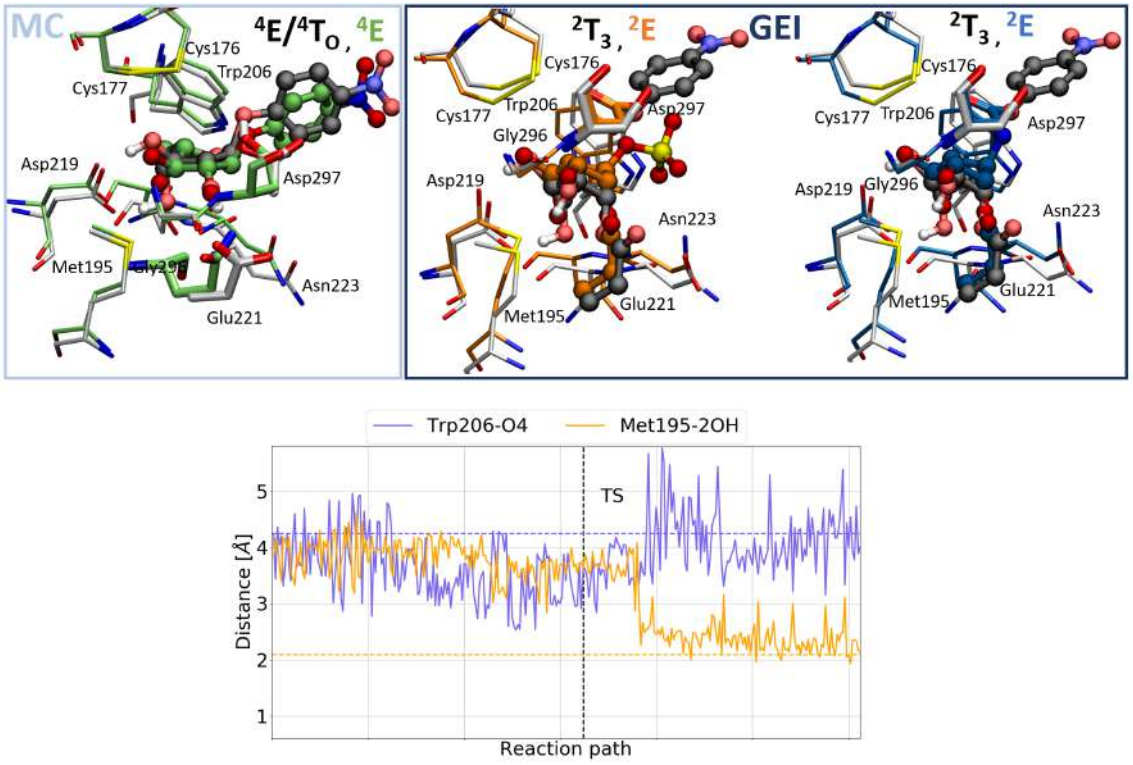


Figure 4.12: **(A)** Superimposed structures of the AkAbfB Glu221Gln crystal structure (green) with the MC found *in silico* (substrate in black and protein residues in gray). All the protein interactions are conserved. **(B)** Superimposed crystal structures of covalently-bound inhibitors **6** (orange) and **2** (blue) bounded crystal structures with the GEI structure found *in silico* (substrate in black and protein residues in gray). All protein interactions are conserved. **(C)** Distance evolution of the protein interactions which significantly change (besides the ones involved in the chemical reaction) along the reaction pathway from MC to GEI. Dashed lines show the distance value found in the crystal structures of the covalent-bound inhibitors. It is observed that when the GEI is reached in the chemical reaction simulation the distance values from the crystal structures are restored.

4.3 Summary and Conclusions

In this chapter we have tackled different goals in a multidisciplinary approach, experimental and computational. First, we assessed the conformational mimicry of some α-L-arabinofuranose-like inhibitors with respect to the preferred conformations adopted by α-L-arabinofuranose. By means of computing the conformational FEL, we predicted which of the inhibitors are able to mimic the putative MC/TS conformation in the enzyme active site of retaining α-L-arabinofuranosidases, we concluded that the ideal MC conformation should be between ^{2}E and ^{4}E region. Therefore, compound **6** was hypothesized to be potentially a better inhibitor than compounds **1** and **2** as the conformations energetically favoured for **6** corresponded to the ones found in the region of ideal MC conformation whereas **1** and **2** pre-

ferred conformations in the $^1T_2$-$^3T_2$ region. Additionally, calculations of the conformational FEL of the isolated α-L-arabinofuranosyl cation indicated that probably the TS conformation of the retaining α-L-arabinofuranosidases would correspond to E_3. Once the inhibitors were synthesized by our collaborators, the best inhibitors turned out to be **6** and **2**, corroborating our prediction for **6** but not so much for **2**. To characterize the mode of inhibition, we determined the crystallographic structure of the best inhibitors, **6** and **2**, covalently bound to a fungal enzyme of family GH54 (AkAbfB) in which both compounds exhibit a ^2E ring conformation. At the same time, a Michaelis complex of the enzyme mutant (Glu221Gln) with a hydrolysable substrate (PNP-Araf) was determined by X-ray crystallography. The structure showed that PNP-Araf binds to the active site with the α-arabinose ring in a ^4E conformation, as predicted from calculations of the isolated conformational FEL of α-L-arabinofuranose. Finally, the reaction mechanism of AkAbfB α-L-arabinofuranosidase was modelled, using as starting structure the MC structure obtained by X-ray crystallography. The computed free energy barrier (17.8 ± 1.1 kcal·mol^{-1}) is in reasonable agreement with experiments and the predicted catalytic itinerary of the α-L-arabinose ring is $^4E/^4T_O \rightarrow [E_3]^{\ddagger} \rightarrow {}^2T_3$. This itinerary is consistent with the results obtained from the conformational FEL of the isolated α-L-arabinofuranose, in where we predicted that the ideal MC conformation would lie in-between the ^2E-^4E region; also, it is consistent with the predicted ring conformation found in the TS from the conformational FEL of the isolated α-L-arabinofuranosyl cation and finally; the GEI substrate conformation is very close to the one found in the covalently-bound crystal structures of inhibitors **2** and **6**.

The main conclusions reached in this chapter are the following:

- By computing the conformational FEL of α-L-arabinofuranose we predicted that the most preactivated conformation of α-L-arabinofuranose in α-L- arabinofuranosidases Michaelis complex (MC) should be in between the conformations ^2E and ^4E, as these conformations feature an equatorial 2-OH, minimizing the steric hindrance with the nucleophile residue along the chemical reaction.

- Compound **6** shows preference in its conformational FEL, for the ideal MC conformation predicted above, anticipating its potential as an α-L- arabinofuranosidases inhibitor. Our prediction was correct as it was proved to react rapidly with the catalytic nucleophile of AkAbfB, but with poor initial binding (experiments performed by our collaborators).

- Compound **2**, contrary to our prediction, was also found to be a potent inhibitor, with a performance 8-fold lower than **6** but with stronger initial bind-

ing. These results are supported by the trends observed in the computed binding energies.

- Compound **1**, as computationally predicted by its conformational FEL in vacuum, was found not to be an inhibitor of α-L-arabinofuranosidases.

- The computed conformational FEL of the isolated α-L-arabinofuranosyl cation indicates that the most stable conformation for the hydroxymethyl orientation found in the active site of AkAbfB enzyme is E_3.

- The crystallographic structures of the covalently bound inhibitors **2** and **6** were obtained with high resolution (1.47 Å and 1.86 Å). The arabinose ring exhibits in both structures a ^2E conformation.

- The Glu221Gln MC structure of AkAbfB in complex with PNP-Araf was crystallized with high resolution (1.64 Å). The substrate ring adopts a ^4E conformation, as predicted by the computed conformational FEL of the isolated α-L-arabinofuranose.

- The catalytic conformational itinerary of AkAbfB α-L-arabinofuranosidase (GH54 family) is predicted to be $^4E/^4T_O \rightarrow [E_3]^{\ddagger} \rightarrow {}^2T_3$ from QM/MM MD metadynamics simulations. This demonstrates that the crystallographic structures of AkAbfB in complex with PNP-Araf and inhibitors **2** and **6** are good pictures of a MC and GEI structures, respectively.

4.4 Experimental and computational methods

4.4.1 Enzyme production and crystallization

4.4.1.1 Recombinant enzyme production

For α-L-arabinofuranosidase of family GH51, the sequence came from an eukaryotic organism; specifically a fungi called *Aspergillus niger* abfA (AnAbfA, GenBank: CAK43424). For family GH54 the sequence also came from a fungi organism which was previously synthesized (*Aspergillus kawachii* abfB, AkAbfB, GenBank: BAB96816), the synthesis procedure is further explained in ref. [157]. The AkAbfB (Glu221Gln) mutant was generated using the Q5 site-directed mutagenesis kit (New England Biolabs) with primers designed by the NEBaseChanger tool.

The genes were produced in *P. Pastoris* X-33. Plasmid DNA for transformation into *P. Pastoris* was linearized with SacI and purified using a PCR clean-up kit

(Qiagen) using ultrapure water as eluent. 100 ng of linearized DNA was electroporated into 80 μL of X-33 electrocompetent cells prepared following the protocol of Wu and Letchworth [169]. Nine colonies from each transformation were purified on YPD-Zeocin plates, then grown in 5 mL of BMGY medium. At saturation (OD600 $\sim$ 20) cells were collected by centrifugation and re-suspended in 5 mL of BMMY medium for expression screening at 20^{o}C. The transformant which gave the highest titer of the target protein with minimal detectable contamination after 3 daily 0.5 % MeOH feedings was grown in 500 mL of BMGY in a 2.5L baffled shaking flask at 30^{o}C overnight. The culture was then cooled to 20^{o}C and supplemented with 2.5 mL of 100% MeOH each day for 3 days. The culture supernatant was clarified by centrifugation followed by 0.45 μm-filtration. 500 mL of medium was concentrated using a KrosFlo tangential flow system fitted with a 30 KDa MWCO mPES filter, then diluted with 9 volumes of 10 mM pH 5 sodium acetate buffer and concentrated again. Protein was then collected onto a 5 mL Q sepharose HP column (GE Healthcare), washed with 3 CV of 50 mM pH 5 sodium acetate buffer, then eluted with 25 CV gradient from 0 to 0.5 M NaCl in the same buffer. Fractions from the largest UV-active peak were pooled, concentrated to 10-30 mg/mL using a 30 kDa MWCO centrifugal concentrator (Amicon) and purified over Superdex 200 (GE Healthcare) into 50 mM sodium acetate pH 5. Protein-containing fractions were pooled and concentrated to give a colourless 15-25 mg/mL protein solution. Approximately, 5 mg of protein was then treated with 1000 U of EndoHf (New England Biolabs) overnight at RT. This was purified using a 5 mL Q sepharose HP column as above. To prepare the sample for crystallization, the eluent from Q sepharose was mixed 1:1 with saturated ammonium sulphate and purified over a 1 mL phenyl sepharose HP column with a 25 CV gradient from 2M to 0M of ammonium sulphate in 50 mM pH 5 sodium acetate buffer. Protein-containing fractions were pooled, desalted into 20 mM sodium acetate pH 5, concentrated to 10-30 mg/mL and frozen to -80°C.

4.4.1.2 Enzyme crystallization and diffraction

The AnAbfA emzyme from GH51 family was not able to crystallize, being the crystals from family GH54 (wild-type and Glu221Gln), AkAbfB, the only ones able to grown. They were grown essentially as described by Miyanaga *et. al.* [170]. Optimized crystals grew from 0.5 μL of 10mg/mL AkAbfB in 50 mM pH 5 sodium acetate mixed with 0.5 μL of 100 mM Tris-HCl pH 8.0, 200mg $MgCl_2$, 400 mM NaCl, 20% PEG6000, 2.5% DMF at 279 K. However, preferential formation of poor-quality needle clusters and poor diffraction of these crystals led us to explore other crystallization conditions. EndoH-deglycosylated AkAbfB or AkAbfB (Glu221Gln) (12 mg/mL in 50 mM sodium acetate pH 5.0) formed slow-growing isolated crys-

tals when mixed 2:1 with 0.2 M lithium sulphate, 0.1 M sodium acetate pH 4.5, 50% PEG400. Supplementation with 0.2-0.5 M NaCl resulted in more rapid crystal growth. To generate inhibitor-bound complexes, crystals were transferred to mother liquor supplemented with inhibitor **2** or **6** to a final concentration of 0.2 mM, or saturated with PNP-Araf (for AkAbfB [Glu221Gln]). Crystals were soaked for 1 hour at RT prior to freezing. Diffraction data were collected at Diamond Light Source (Harwell, UK) on beamline I04 and automatically processed using autoPROC [171] (AkAbfB-2 and AkAbfB-6), or Xia2 [172] (AkAbfB-PNP-Araf) pipelines. Computation was carried out using programs from the CCP4 suite [173]. Data collection and processing statistics for all structures are given in Supplemental Table 4.3.

4.4.1.3 Structure solution and refinement

Solved by Dr. N. McGregor:
Data for AkAbfB bound to inhibitors **2** and **6** were collected to 1.47 and 1.86 Å , respectively. Each structure was solved by molecular replacement using Phaser [174] with the known structure (PDB ID 1WD3) as the search model. The resulting solution showed clear density for the bound ligand within the enzyme active site. Ligand coordinates and dictionaries were generated using jLigand [175] and built into the model using Coot [176], followed by alternating rounds of manual model building and refinement using Coot and REFMAC5 [177].

Data for AkAbfB (Glu221Gln) bound to PNP-Araf were collected to 1.64 Å. The structure was solved by molecular replacement using Phaser with the AkAbfB-**2** complex as the search model. The resulting structure showed clear density for two PNP-Araf (ligand ID: KHP) molecules bound to the carbohydrate-binding module. Following several rounds of manual model building and refinement, partial density for an additional PNP-Araf molecule, which was modelled at 60% occupancy, became apparent in the active site.

4.4.2 Conformational FELs of isolated molecules

Conformational free energy landscapes (FELs) were computed for α-L- arabinofuranose, compounds **1**, **2**, **6** (Figure 4.2**B**, page 86) and α-L-arabinofuranosyl cation using DFT-based molecular dynamics, according to the Car-Parrinello method [66]. Each molecule was enclosed in an isolated cubic box of 12.5 $Å^3$. A fictitious electron mass of 500 atomic units was used for the CP Lagrangian and a time step of 0.12 fs was used in all CPMD simulations to ensure that the adiabacity of the fictitious kinetic energy of the electrons was smaller than 10^{-5} a.u./atom. The Kohn-Sham orbitals were expanded in a plane wave basis set with a kinetic energy cut-off of 70

Ry. Ab initio pseudopotentials, generated within the Troullier-Martins scheme, were employed [142]. The Perdew, Burke, and Ernzerhoff (PBE) generalized gradient-corrected approximation [139] was selected in view of its good performance [141] in previous work on isolated sugars [19], glycosidases, and glycosyltransferases [140]. The metadynamics algorithm [81], provided by the Plumed 2 plugin [17], was used to explore the conformational free energy landscape of the systems, taking as collective variables the pseudorotational phase (ϕ) puckering coordinate [18], as well as a dihedral angle accounting for the rotation of the sugar hydroxymethyl group. The energy was projected into the ϕ coordinate for representation purposes. Initially, the height of these Gaussian terms was set at 0.6 kcal·mol^{-1} and a new Gaussian-like potential was added every 500 MD steps. Once the whole free energy space was explored, the height of the Gaussian terms was reduced to 0.2 kcal·mol^{-1} to facilitate convergence of the FEL. The width of the collective variables was set according to their oscillations in the free dynamics which corresponded to 0.035 and 0.1 rad for ϕ and the hydroxymethyl dihedral angle, respectively. The simulations were stopped when energy differences among wells remain constant, which was further confirmed by a time-independent free energy estimator [143]. The exploration of the phase space was extended up to 380, 360, 324, 474 and 960 ps for α-L-arabinofuranose, compound **1**, compound **2**, compound **6** and α-L-arabinofuranosyl cation, respectively. The errors in the principal minima, taken as a standard deviation from the last 200 ps, are below 0.6 kcal mol^{-1}. Conformational FELs computed using only ϕ as CV gave very similar results, see Supplemental Figure 4.14.

4.4.3 MD simulations of AkAbfB enzyme in complex with compounds 1, 2, 6; as well as with the PNP-Araf substrate mimic

4.4.3.1 Classical MD simulations

The Michaelis complexes of compounds **1**, **2**, and **6** were built using the crystal structures of the covalently-bound inhibitors obtained for AkAbfB as a reference. In the case of compounds **1** and **2**, the Michaelis complexes were reconstructed by removing the covalent bond between the inhibitor and the nucleophile in the protein structure bound to inhibitor **2**. The amine group was reverted to an aziridine, for compound **2** (Figure 4.13), which was replaced with an oxygen atom to give compound **1**. The same strategy was followed for compound **6**. Molecular dynamics (MD) simulations were set up employing the program LEaP included in the Amber suite [178] and the ff14SB protein force field [179] was used. The compounds were parametrized using gaff2 [136]. The systems were solvated with explicit TIP3P

water molecules [134]. They were neutralized with 21 sodium atoms for all neutral compounds in AkAbfB. The systems with protonated compound **2** were neutralized with one fewer sodium atom (20 atoms). MD simulations were performed using Amber16 [178]. A thermal equilibration to 300 K was done prior to the equilibration of dynamics in the NPT ensemble with a production phase of 51 ns for each system. The SHAKE algorithm, with an integration time step of 2 fs, was used. The binding free energy of the compounds were obtained by using the MMPBSA method [180] integrated in the Amber suite.

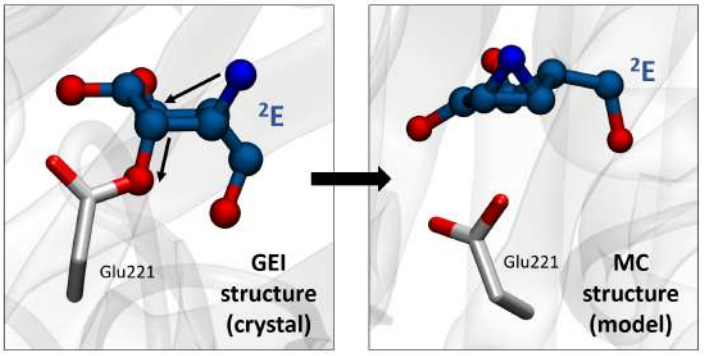


Figure 4.13: Building process of the MC structure of compound **2** from the crystal structure of the covalently-bound inhibitor (GEI structure) obtained for AkAbfB as a reference. The same process was followed for compounds **6** with its corresponding GEI structure and for compound **1** the nitrogen atom of compound **2** was replaced with an oxygen atom.

For AkGH54 (Glu221Gln) bound to PNP-Araf, the mutation was reversed *in silico* and the classical MD simulation was set up with the same procedure as above (with 21 sodium atoms) and a production phase of 48 ns. Analysis of the trajectories were carried out using VMD tools [121] and images were processed with VMD (version 1.9.2), Gnuplot [124] and matplotlib (module found in Python3) [125].

4.4.3.2 QM/MM MD metadynamics simulations

Two different QM/MM MD simulations were done using the method developed by Laio *et. al.* [77] (further explained in Chapter 2); the substrate conformational FEL simulation and the chemical reaction simulation. In both, the quantum region is treated using the PBE functional and Car-Parrinello MD whereas the molecular mechanics region is treated using the AMBER force field; when a covalent bond is placed in the QM-MM boundary, this is treated by saturating the QM region within a MD formalism parametrizing a boundary atom described by a pseudopotential. The electrostatic interactions between the QM and MM region are handled with two layers, NN and ESP regions. Both systems were simulated by means of metady-

namics [81] with the CPMD program (version 3.15.1) [76] and the Plumed package (version 2.3.3) [17].

The substrate conformational FEL simulation in the active site of AkAbfB only included the PNP-Araf substrate in the active site as the QM region and the ϕ puckering coordinate was considered as collective variable (CV).

The chemical reaction simulation included the PNP-Araf substrate and the catalytic residues (Glu221 and Asp297) until their beta carbon in the QM region. The CVs chosen were the bonds to be broken and formed; the first CV corresponds to the formation of the covalent bond between the anomeric carbon of the PNP-Araf substrate and the nucleophile whereas the second CV quantifies the cleavage of the glycosidic bond. The proton transfer between the aglycon moiety and Asp297 (acid/base) was not included due to the relatively good leaving group character that the aglycon moiety has; a p-nitrophenol group (pKa = 7.15), which made us think that the reaction would take place without leaving group protonation, as seen previously for Gal-β-pNP which had the same aglycon moiety (Chapter 3). The statistical error of the reaction free energy barrier was computed by launching several metadynamics simulations changing the Gaussian height and keeping all the other parameters constant [145]. This should be considered as an approximation to the statistical error of the metadynamics simulation, since only the Gaussian height (not the width) was varied. Each simulation was restarted from the point along the metadynamics run in which the reactants basin is half filled and its height was changed to 0.3, 0.5 and 0.63 kcal·mol^{-1}. The results showed that the reaction mechanism was not affected by the change of the Gaussian height.

Parameters	**Simulation of the substrate conformational FEL**	**Simulation of the chemical reaction**
Number of quantum atoms	32	50
Functional	PBE	PBE
PW E_{cut} (Ry)	70	70
Pseudopotential	Norm-conserving Martin-Trouillers	Norm-conserving Martin-Trouillers
Fictitious electronic mass (a.u.)	500	700
Time step (a.u.)	5.0	5.0
Size of the electrostatic regions (NN/ESP in a.u.)	10/18	14/22
Supercell size (Å)	18, 14.4, 12.6	19, 15.6, 16.9
CV1	ϕ	dist(O_{Glu221}-C1)
CV2	-	dist(C1-O1')
Gaussian width (a.u.)	0.035	0.1/0.1
Gaussian height (kcal·mol^{-1})	0.6	1 and 0.31/0.5/0.63
Deposition time (MD steps)	500	300
Deposited gaussians	3776	747
Simulation time (ps)	227	27

Table 4.2: Parameters of the QM/MM MD metadynamics simulations.

4.5 Supplemental Figures and Tables

	AkAbfB		**AkAbfB (Glu221Gln)**
	Complex with **2**	Complex with **6**	PNP-Araf complex
Data Collection			
Space Group	$H3_2$	$H3_2$	$H3_2$
a,b,c (Å)	112.2, 112.2, 342.7	111.14, 111.14, 342.7	112.0, 112.0, 341.4
α, β, γ (°)	90.0, 90.0, 120.0	90.0, 90.0, 120.0	90.0, 90.0, 120.0
Resolution (Å)	84.54-1.47	83.92-1.86	93.28-1.64
	(1.49-1.47)	(1.89-1.86)	(1.67-1.64)
R_{meas}	0.109 (3.128)	0.133 (5.537)	0.129 (2.319)
I / σI	13.0 (0.7)	10.2 (0.5)	12.5 (1.0)
Completeness (%)	99.3 (98.5)	100 (100)	98.2 (97.0)
Redundancy	12.4 (10.7)	12.1 (12.8)	12.6 (12.2)
Refinement			
Resolution (Å)	84.69-1.47	84.06-1.86	93.45-1.64
No. Reflections	141273	68786	99210
R_{work}/R_{free}	0.157/0.179	0.181/0.215	0.159/0.181
No. atoms			
Protein	3628	3568	3572
Ligand/ion	202	131	260
Water	456	236	352
B-factors			
Protein	20.9	43.7	26.8
Ligand/ion	42.1	58.5	50.4
Water	34.3	49.4	37.6

R.m.s. deviations			
Bond lengths (Å)	0.016	0.012	0.014
Bond angles (°)	1.94	1.77	1.87
PDB Code	6SXT	6SXS	6SXR

Table 4.3: Data Collection and refinement statistics (molecular replacement)

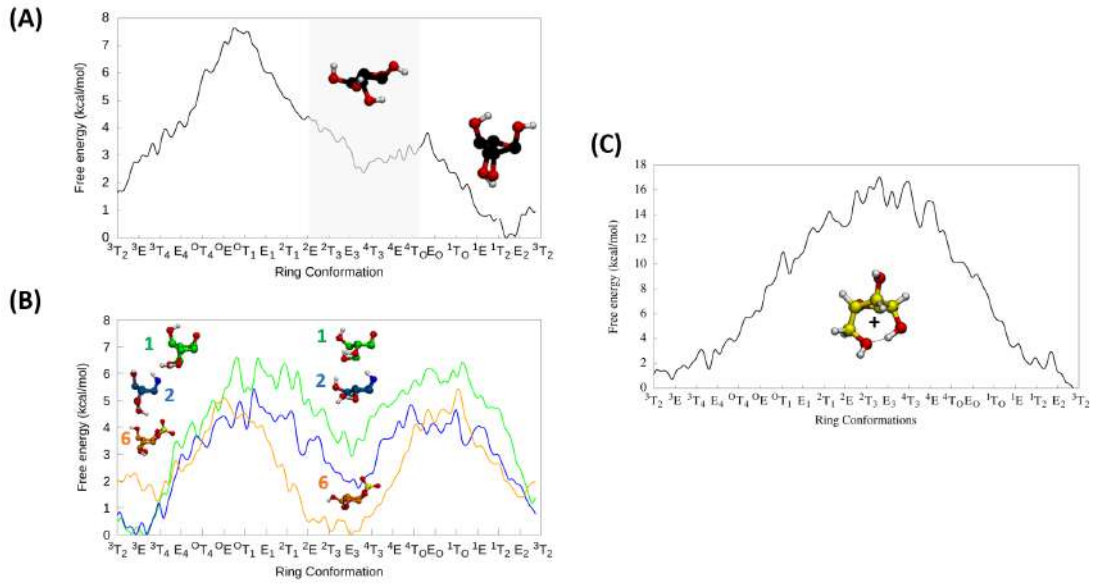


Figure 4.14: **(A)** Conformational FEL of isolated α-L-arabinofuranose computed with 1 CV (ϕ). The conformational region having an equatorial 2-OH is shaded. **(B)** Conformational FEL of α-L-arabinofuranose-configured cyclophellitol(**1**), aziridine (**2**) and cyclic sulfate (**6**) computed with only one CV (ϕ). **(C)** Conformational FEL of isolated α-L-arabinofuranosyl cation computed with 1 CV (ϕ).

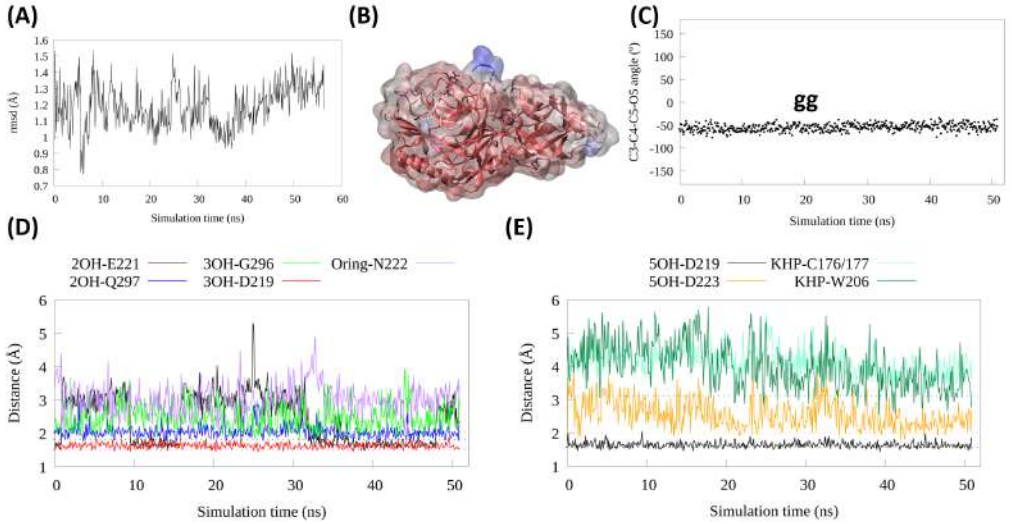


Figure 4.15: Analysis of the classical MD simulation of PNP-Araf. **(A)** Backbone protein r.m.s.d. **(B)** Protein backbone r.m.s.f. from red, for the less mobile residues, to blue for the more mobile ones. **(C)** Hydroxymethil rotation angle value, in a gg orientation along the whole simulation **(D)** Active site bond distances with PNP-Araf, the dashed lines in their respective colours correspond to the value found in the crystallographic structure **(E)** Active site bond distances with PNP-Araf, continuation.

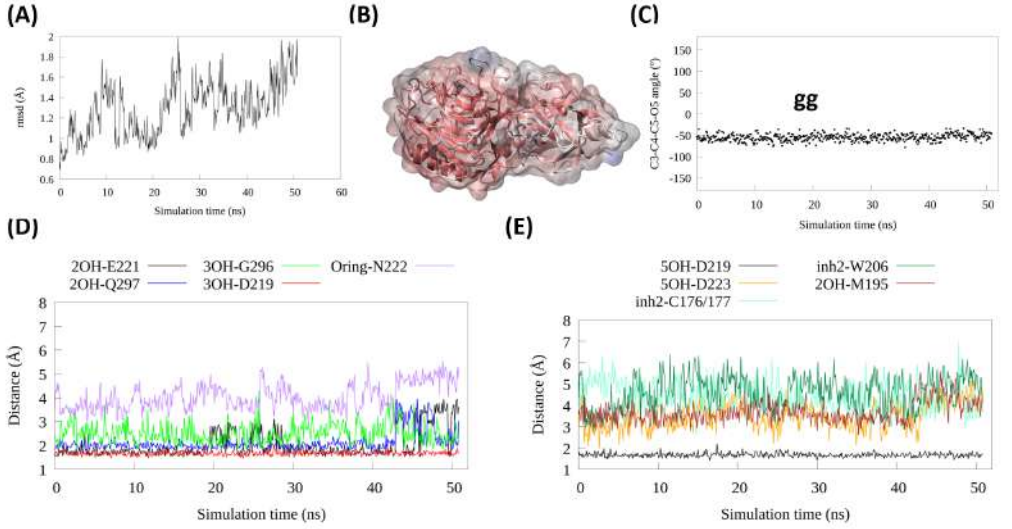


Figure 4.16: Analysis of the classical MD simuation of compound **1**. **(A)** Backbone protein r.m.s.d. **(B)** Protein backbone r.m.s.f. from red, for the less mobile residues, to blue for the more mobile ones. **(C)** Hydroxymethil rotation angle value, in a gg orientation along the whole simulation **(D)** Active site bond distances with inhibitor **1**, the dashed lines in their respective colours correspond to the value found in the crystallographic structure of the covalently bound inhibitor **2**, used as a reference **(E)** Active site bond distances with inhibitor **1**, continuation.

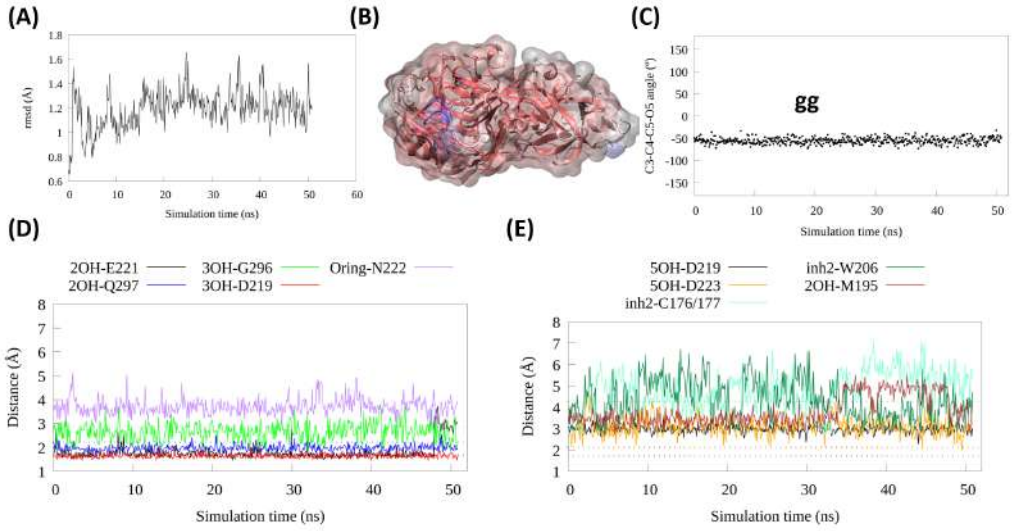


Figure 4.17: Analysis of the classical MD simulation of compound **2** deprotonated and protonated acid/base residue. **(A)** Backbone protein r.m.s.d. **(B)** Protein backbone r.m.s.f. from red, for the less mobile residues, to blue for the more mobile ones. **(C)** Hydroxymethil rotation angle value, in a gg orientation along the whole simulation **(D)** Active site bond distances with inhibitor **2** with deprotonated aziridine nitrogen and protonated general acid/base, the dashed lines in their respective colours correspond to the value found in the crystallographic structure of the covalently bound inhibitor **2** **(E)** Active site bond distances with inhibitor **2** with deprotonated aziridine nitrogen and protonated general acid/base, continuation.

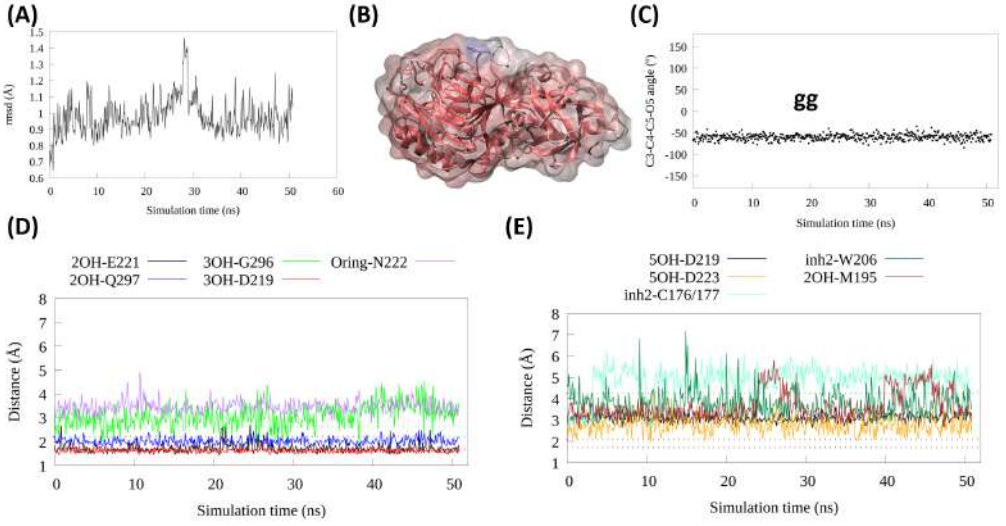


Figure 4.18: Analysis of the classical MD simulation of compound **2** protonated and protonated acid/base residue. **(A)** Backbone protein r.m.s.d. **(B)** Protein backbone r.m.s.f. from red, for the less mobile residues, to blue for the more mobile ones. **(C)** Hydroxymethil rotation angle value, in a gg orientation along the whole simulation **(D)** Active site bond distances with inhibitor **2** with protonated aziridine nitrogen and protonated general acid/base, the dashed lines in their respective colours correspond to the value found in the crystallographic structure of the covalently bound inhibitor **2** **(E)** Active site bond distances with inhibitor **2** with protonated aziridine nitrogen and protonated general acid/base, continuation.

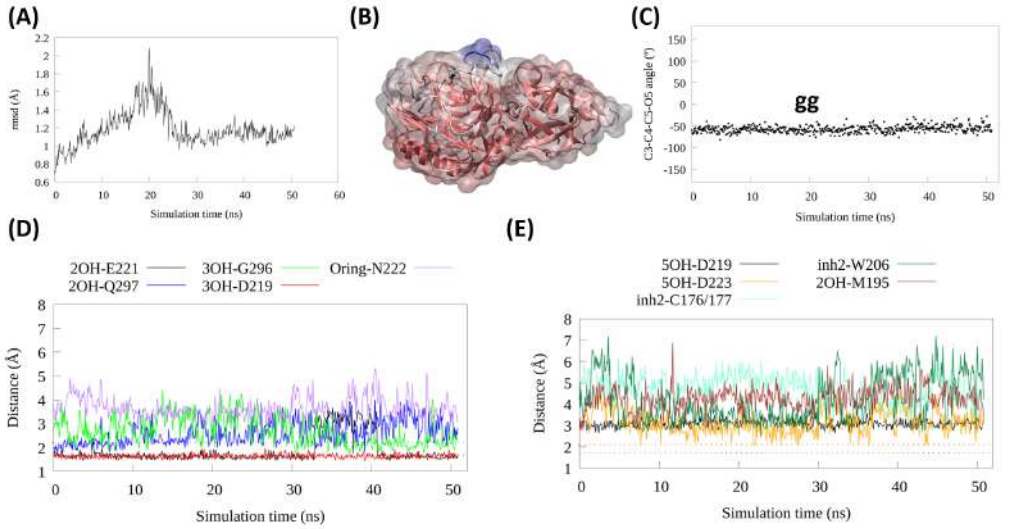


Figure 4.19: Analysis of the classical MD simulation of compound **2** protonated and deprotonated acid/base residue. **(A)** Backbone protein r.m.s.d. **(B)** Protein backbone r.m.s.f. from red, for the less mobile residues, to blue for the more mobile ones. **(C)** Hydroxymethil rotation angle value, in a gg orientation along the whole simulation **(D)** Active site bond distances with inhibitor **2** with protonated aziridine nitrogen and deprotonated general acid/base, the dashed lines in their respective colours correspond to the value found in the crystallographic structure of the covalently bound inhibitor **2** **(E)** Active site bond distances with inhibitor **2** with protonated aziridine nitrogen and deprotonated general acid/base, continuation.

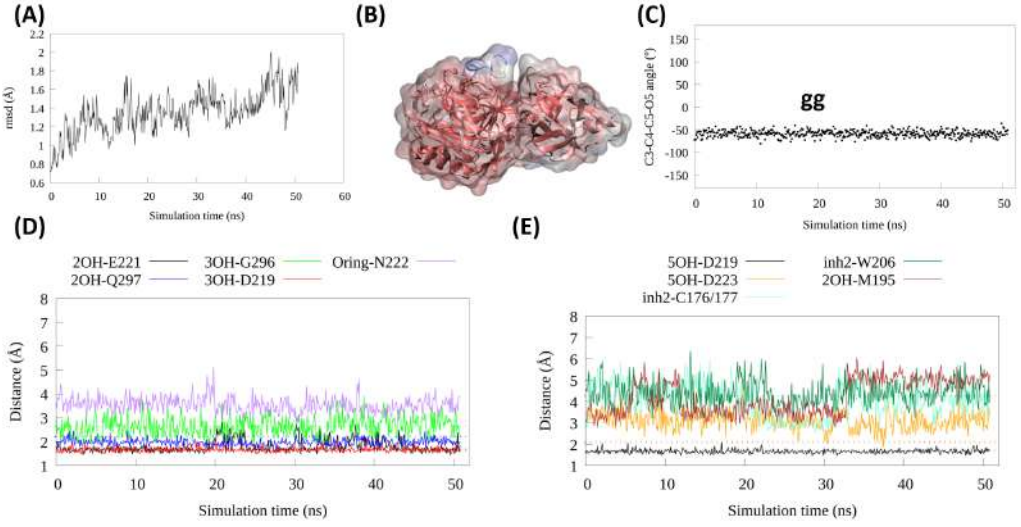


Figure 4.20: Analysis of the classical MD simulation of compound **6**. **(A)** Backbone protein r.m.s.d. **(B)** Protein backbone r.m.s.f. from red, for the less mobile residues, to blue for the more mobile ones. **(C)** Hydroxymethil rotation angle value, in a gg orientation along the whole simulation **(D)** Active site bond distances with inhibitor **6**, the dashed lines in their respective colours correspond to the value found in the crystallographic structure of the covalently bound inhibitor **6** **(E)** Active site bond distances with inhibitor **6**, continuation.

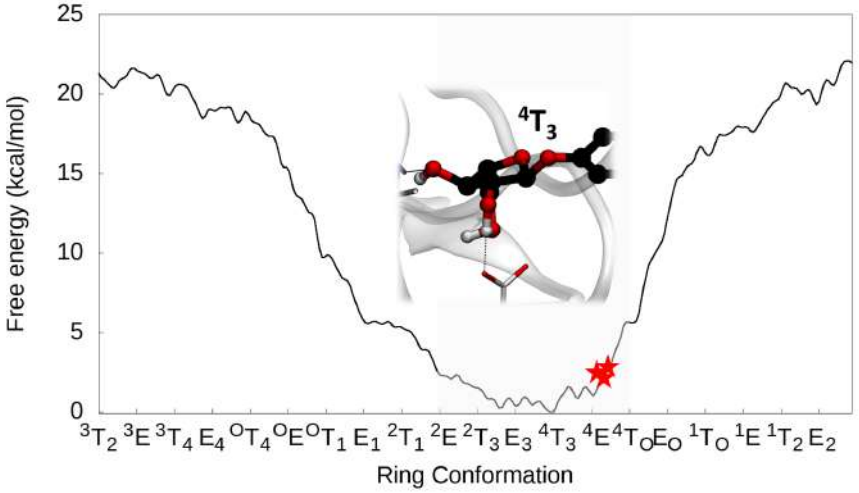


Figure 4.21: Arabinose ring conformational FEL of PNP-Araf on the active site of WT-AkAbfB. Conformations observed in Michaelis complexes of α-L-arabinofuranosidases are represented with a red star (PDB 2VRQ and 1QW9 for GH51 and PDB 6SXR, this work, for GH54). The conformational region having an equatorial 2-OH is shaded and a representative structure of the most stable conformation, $^4T_3$ is shown.

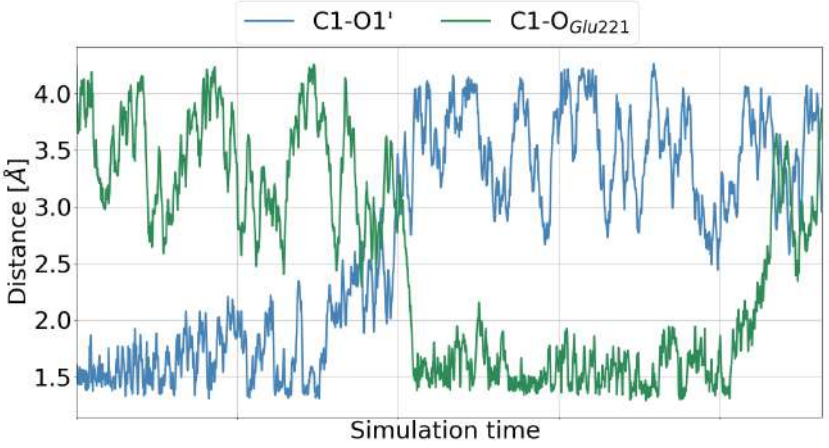


Figure 4.22: Collective variables evolution along the reaction simulation.

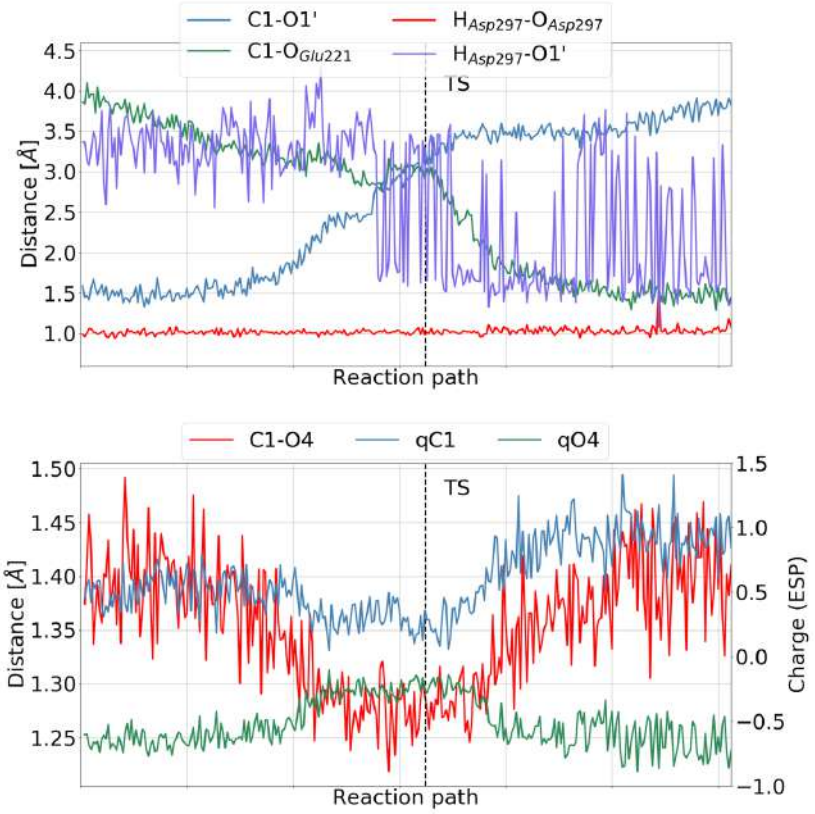


Figure 4.23: Structural and electronic parameters along the lowest free energy reaction path. The big variation of Asp297-O1' distance indicates that for PNP-Araf, the acid/base residue is not needed for the reaction to take place and the GEI complex it is stable either with a hydrogen bond between the p-nitrophenol moiety and Asp297 or with no interaction between them.

Chapter 5

Mimicking the transition state of glycosylation reactions using superacids

Parts of this chapter have been published:
Manuscript in preparation

ABSTRACT: Recently, the experimentally elusive glycosyl cation was trapped for the first time by NMR using liquid hydrofluoric acid-antimony pentafluoride (HF/SbF5) superacid. It was found that 2-deoxy-β-D-glucosyl cation adopts a ^{4}E conformation, which in principle mimic the transition state of glycosylation reactions. Nevertheless, whether the superacid media can be an useful tool to study the glycosyl cation properties present on the active site of a GH enzyme remains controversial. To solve this issue, here we study the molecular properties of 2-deoxy-β-D-glucosyl and 2-deoxy-α-L-arabinofuranosyl cations in different environments; in vacuum, in superacid media and on-enzyme, using QM/MM methods. A superacid force field was parametrized for this purpose. We found that the geometrical properties, such as the sugar ring conformation, bond distance and torsion angles are conserved along all the studied environments in both glycosyl cations. The superacid media enhances slightly the cationic character of the glycosyl cations compared to the vacumm and on-enzyme environments. Our results clarify that a superacid can be considered as a good mimic of the active site of GHs and thus they constitute a useful tool to study the properties of the transition state of glycosylation reactions.

5.1 Introduction

As discussed in previous chapters of this Thesis, the reaction mechanism of GHs go through a TS structure where the substrate acquires oxocarbenium ion character, which it is important for the design of new specific drugs. Even though the glycosyl oxocarbenium ion (glycosyl cation) plays a key role in glycosylation reactions (e.g. hydrolysis, transglycosylation and synthesis of glycans), its observation in the condensed phase has been elusive to experiments due to its extremely short life time. In seminal work, Martin *et. al.* [30] and Lebedel *et. al.* [181] were recently able to trap glycosyl cations in liquid hydrofluoric acid–antimony pentafluoride (HF/SbF_5) superacid by nuclear magnetic ressonance (NMR). This work, leaded by Sébastien Thibadeau, Ives Blériot (University of Poitiers, France) and Jesús Jiménez Barbero (CIC Biogune, Bilbao), was inspired in the experiments by George Olah (1994 Nobel Prize), in which he used superacidity to stabilize and study carbocations using low-temperature NMR. Interestingly, G. Olah wrote in his autobiography that there could be a parallelism between the superacid media and some enzyme active sites [182].

Superacids are defined as acids that are more acidic than pure sulphuric acid (the Hammet acidity function[1], H_0, of sulphuric acid corresponds to -12 while HF/SbF_5 superacid has a value of -23 [183]). In this highly acidic environment, organic molecules can be (poly)protonated, shifting their reactivity from "classical" to superelectrophilic. Martin *et. al.* used peracetylated 2-deoxy-β-D-glucopyranose to avoid potential problems associated with the presence of a substituent at C2; their results indicated a ^{4}E conformation and an α face stereoselectivity for nucleophillic reactions, see Figure 5.1**A**. More recently, exploring the idea that (poly)protonation of multifunctional substrates can also acts as a (poly)protection of the molecule, the same experimental groups applied a similar strategy to trap glycosyl cations of both five- and six-membered sugar rings (2-deoxy-β-D-glucosyl and 2-deoxy-α-L-arabinofuranosyl cations) with no protecting groups (unpublished work), see Figure 5.1 panels **B** and **C**. They found that the 2-deoxy-β-D-glucosyl cation conserves the same ^{4}E conformation while the conformation of 2-deoxy-α-L-arabinofuranosyl cation has not yet been analised. In collaboration with the above experimental groups, we asked ourselves if the superacid media is a good mimic of the active site environment of carbohydrate-active enzymes such as GHs; i.e., to which degree the glycosyl cation properties observed in superacid media are similar to the properties of the TS in the enzymatic hydrolysis glycosylation reaction?

[1] A measure of acidity used for very concentrated solutions of strong acids as the Henderson-Hasselbach equation is no longer valid in these conditions.

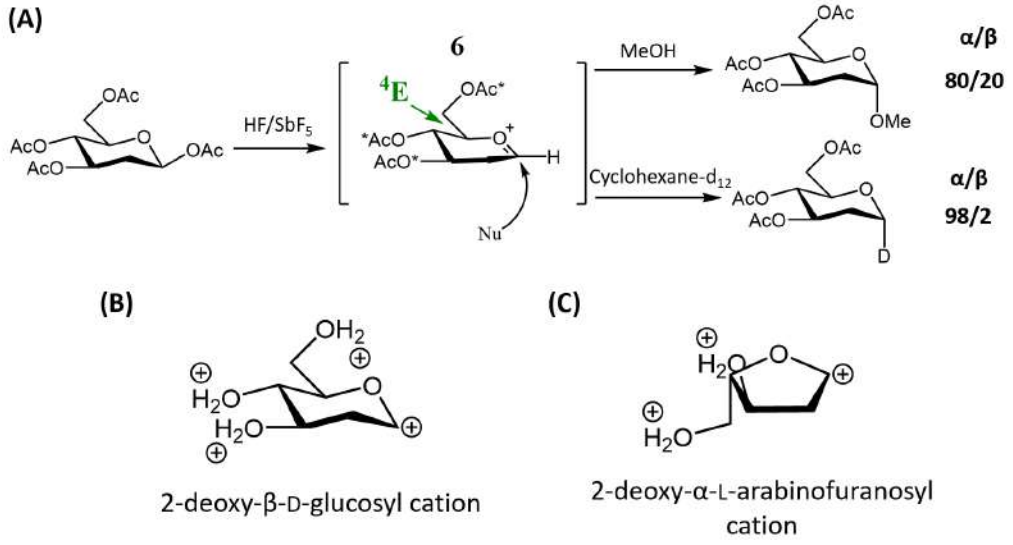


Figure 5.1: **(A)** Generation of 2-deoxy-glucosyl oxocarbenium ion 6 in HF/SbF_5 followed by its trapping with methanol (40 equiv.) or cyclohexane-d12 (2 equiv.). Image modified from Martin *et. al.* [30]. **(B)** The polyprotonated 2-deoxy-β-D-glucosyl cation trapped in superacid. **(C)** The polyprotonated 2-deoxy-α-L-arabinofuranosyl cation trapped in superacid.

To answer these questions, we designed QM/MM simulations of the glycosyl cations analysed experimentally, in comparison to the TS of representative GH. First, the intrinsic properties of both molecules (2-deoxy-β-D-glucosyl and 2-deoxy-α-L-arabinofuranosyl cations) in gas-phase are studied; the geometrical and electronic intramolecular properties such as bond distances, torsion angles, sugar puckering conformation and atomic charges when the sugar is in its most stable state (neutrally charged) are compared with these properties when the molecule is in the form of an oxocarbenium ion (when the hydroxyl group in C1 is not present and thus, the molecule is positively charged). Secondly, the polyprotonated glycosyl cations are simulated in superacid media, i.e., surrounded by SbF_6^-, H_2F^+ and HF molecules. Our simulations were performed with the QM/MM approach, after developing a suitable force field for the superacid media (see methods section 5.4, page 143). Finally, we analysed the properties of these glycosyl cations in the active site of two representative GHs, one acting on hexoses (the *Saccharomyces cerevisiae* Gas2, ScGas2, previously investigated in the group [117]) and one acting on furanoses (the *Aspergillus kawachii* abfB, AkAbfB, investigated in Chapter 4). The structural and electronic properties of these glycosyl cations in an enzymatic environment are discussed and compared to the ones found in superacid media.

5.2 Results

5.2.1 2-deoxy-β-D-glucosyl cation in different environments

5.2.1.1 The isolated 2-deoxy-β-D-glucose

In order to study the intrinsic properties of 2-deoxy-β-D-glucosyl cation we first computed the conformational FEL of the isolated neutral form by means of CPMD metadynamics using the Cremer-Pople cartesian coordinates [10] as CVs. The most stable conformation was found to be the same one obtained for its 2-oxy counterpart, β-D-glucose [19], the relaxed $^4C_1$ chair (Figure 5.2). Even though in both cases (β-D-glucose and 2-deoxy-β-D-glucose) the distorted conformations region in which the most stable minima are corresponds to one half of the diagram (from ^2S_O to $^OS_2$), the lack of the hydroxyl at C2 changes the stability order of the distorted conformations; $^1S_5$ is now the second minima at ~ 4 kcal·mol^{-1} above $^4C_1$, compared to the second most stable minima of β-D-glucose which is $B_{3,O}$ at ~ 3 kcal·mol^{-1}, with $^1S_5$ being ~ 6 kcal·mol^{-1} above it. The third minima of 2-deoxy-β-D-glucose corresponds to a region between $B_{3,O}$ and $^1S_3$ at 5 kcal·mol^{-1} (see Table 5.1). This interchange of stability of the $B_{3,O}$ and $^1S_5$ conformations between the natural and 2-deoxy molecules is probably due to the lack of one intramolecular hydrogen bond interaction in the $^1S_5$ conformation for β-D-glucose (between the hydroxyls of C2 and C3) while in the 2-deoxy case, both distorted conformers have only one hydrogen bond, Figure 5.2. The $^4C_1$ main minima is connected to the distorted conformations via a valley covering $^2H_3$, E_3, $^4H_3$ and ^4E.

	Conformations		
Molecule	$^4\mathbf{C}_1$	$\mathbf{B}_{3,O}$	$^1\mathbf{S}_5$
β-D-glucose	0	3	6
2-deoxy-β-D-glucose	0	5	4

Table 5.1: Relative energies, in kcal·mol^{-1}, of the main minima in the conformational FEL of β-D-glucose [19] and 2-deoxy-β-D-glucose.

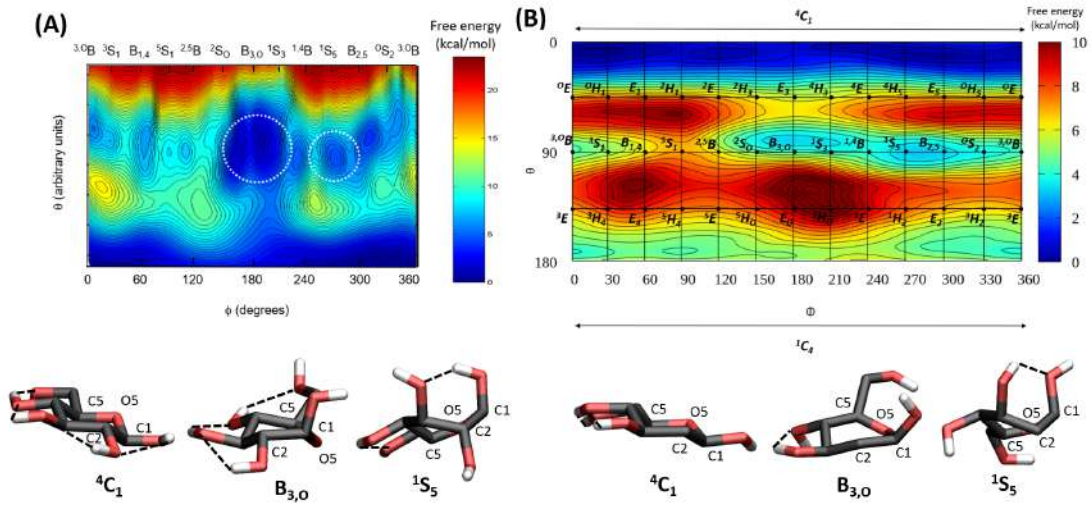


Figure 5.2: **(A)** Conformational FEL of β-D-glucose, figure modified from Biarnés *et. al.* [19]. Energy values are given in kcal·mol^{-1} and each contour line corresponds to 0.5 kcal·mol^{-1}. Representative structures of $^4C_1$, $B_{3,O}$ and $^1S_5$ conformations are shown with their corresponding intramolecular hydrogen bond interactions, represented as dashed lines. The discussed distorted conformations, $B_{3,O}$ and $^1S_5$, are marked with a dashed circle. **(B)** Conformational FEL of 2-deoxy-β-D-glucose. Energy values are given in kcal·mol^{-1} and each contour line corresponds to 0.5 kcal·mol^{-1}. Representative structures of $^4C_1$, $B_{3,O}$ and $^1S_5$ conformations are shown with their corresponding intramolecular hydrogen bond interactions represented as dashed lines.

5.2.1.2 The isolated 2-deoxy-β-D-glucosyl cation

As a second step in our investigation, we calculated the conformational FEL of the corresponding isolated oxocarbenium ion (2-deoxy-β-D-glucosyl cation). As shown in Figure 5.3**B**, the FEL is completely different. We have a completely different scenario as now the lowest energy region corresponds to the conformers $^4H_3$, ^4E, $^3H_4$ and E_3. In fact, these are some of the conformers in which the C5, O5, C1 and C2 atoms are on a plane, stabilizing the sp2-hybridization and anomeric charge of the oxocarbenium ion. The main minima (see Figure 5.3**B**) correspond to the region between $^4H_3$ and ^4E (θ=45° and ϕ=210-240°); a region placed in the connecting valley previously found for 2-deoxy-β-D-glucose; followed by a minima in the E_4 region (θ=110° and ϕ=80°) at 11 kcal·mol^{-1}. There is a high energy barrier (22 kcal·mol^{-1}) connecting both minima. The conformational preferences of the 2-deoxy-β-D-glucosyl cation are also different with those of the plain glucosyl cation (i.e. with the 2-OH substituent), shown in Figure 5.3**A**, which was computed by J. Iglesias-Fernández early on [184]. In that case, the $E_4/^3H_4$ conformation is the most stable one, followed by $^4H_3$, and they are both connected via $^{2,5}B$ and $B_{2,5}$ conformers. Iglesias-Fernández *et. al.* (unpublished work) argued that the stabil-

ity of $^3H_4$ is due to the axial orientation of the hydroxyls at C3 and C4 and the hydrogen bond interaction between C2 and C4 hydroxyl groups, in contrast to $^4H_3$ which has C3 and C4 hydroxyl groups in the less stable equatorial orientation but with a stabilizing "clockwise" hydrogen bond pattern between the hydroxyls of C2, C3 and C4. These conclusions were based on the work of Woods *et. al.* [185] in where they describe by molecular mechanical (MM2) and *ab initio* calculations that oxocarbenium ions are stabilized when C3 and C4 hydroxyl groups have an axial orientation, then, being closer to C1, as in this way there is a stabilizing interaction between the partial negative charge in the oxygen of the hydroxyl group and the partial positive charge of C1. For the 2-deoxy-β-D-glucosyl cation molecule, of interest here, the same orientation of the hydroxyls groups at C3 and C4 is found in $^4H_3$ and $^3H_4$ conformations (axial for $^3H_4$ and equatorial for $^4H_3$), however, $^3H_4$ is not any more a free energy minimum as the C2-C4 hydrogen bond interaction is not present while $^4H_3/^4E$ is most stable due to the higher number of intramolecular hydrogen bonds, one hydrogen bond interaction between the hydroxyls of C3 and C4 and another one between the hydroxyls of C4 and C6 (Figure 5.3**D**).

Besides the sugar ring conformational stabilities, the main difference between the neutral form and the glycosyl cation form of 2-deoxy-β-D-glucose is the C1-O5 bond distance, which shrinks from 1.44 Å (neutral) to 1.27 Å. Additionally, the bond distance between C1-C2 gets shorter, from 1.53 Å to 1.45 Å. Interestingly, C2-C1-O5-C5 dihedral angle, which needs to be close to zero for an oxocarbenium ion molecule, changes from -33° in the neutral 2-deoxy-β-D-glucose to 1.7°, almost planar, for the glycosyl cation form. Moreover, the anomeric charge switches from a negative value (-0.19) to a positive one (0.29), see Table 5.2 in page 130.

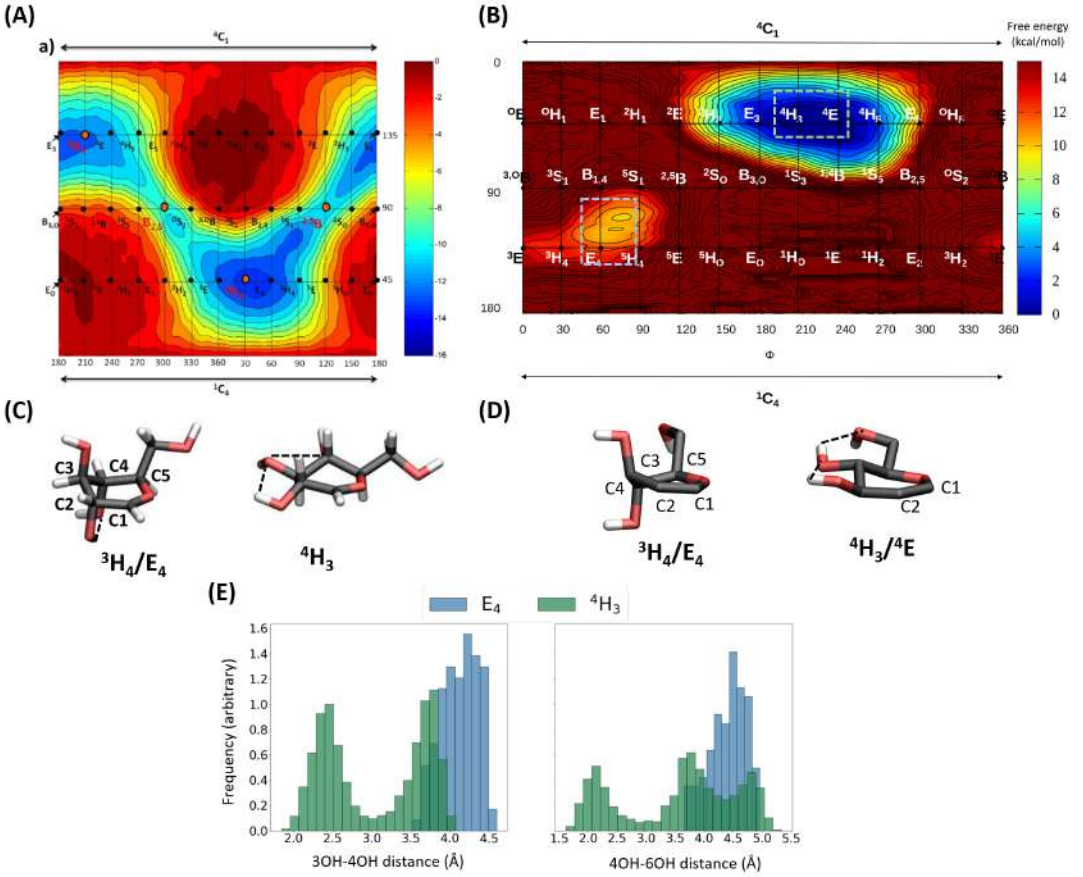


Figure 5.3: **(A)** Conformational free energy landscape of an isolated β-D-glucosyl cation. Energy values are in kcal·mol^{-1} and each contour line corresponds to 1 kcal·mol^{-1}. Image taken from J. Iglesias-Fernández thesis [184]. **(B)** Conformational FEL of an isolated 2-deoxy-β-D-glucosyl cation. Energy values are given in kcal·mol^{-1} and each contour line corresponds to 1 kcal·mol^{-1}. The two most stable regions of the FEL are indicated with dashed squares. **(C)** β-D-glucosyl cation representative structures of $^3H_4$ and $^4H_3$ conformations are shown with their corresponding intramolecular hydrogen bond interactions, represented as dashed lines. **(D)** 2-deoxy-β-D-glucosyl cation representative structures of $^3H_4$ and $^4H_3$ conformations are shown with their corresponding intramolecular hydrogen bond interactions, represented as dashed lines. **(E)** Histogram of 3OH-4OH and 4OH-6OH distances (Å) corresponding to $^4H_3$ and E_4 conformations, indicating a higher probability of $^4H_3$ to stablish hydrogen bond interactions.

5.2.1.3 The 2-deoxy-β-D-glucosyl cation in superacid media

The second step of this project was to assure that our superacid force field was able to reproduce the results obtained experimentally and, afterwards, use this force-field to study properties for which experimental information is lacking. The computed properties (structure, energy and electron density distribution) will be compared with the ones found in vacuum to understand how the superacid media affects the

electronic and conformational preferences of the glycosyl cation.

A classical MD simulation of the 2-deoxy-β-D-glucosyl cation in superacid with all the hydroxyl groups protonated (see Supplemental section 5.5.1) was initially performed to analyse the distribution of SbF_6^- and HF^- solvent molecules (Figure 5.4). A larger concentration of SbF_6^- in the β face of the sugar cation is observed, compared to the α face. This means that nucleophiles (i.e. negatively charges) will have a preference to attack the anomeric carbon from the α face, where there is less negative charge density and thus, less electrostatically hindered. The increased selectivity for the α face for bigger nucleophiles, experimentally observed for peracetylated 2-deoxy-β-D-glucopyranosyl cation (so-called oxocarbenium ion **6**) studied in the work of Martin *et. al.* [30], is here justified due to the steric and electrostatic hindrance of the big SbF_6^- molecules in the β face.

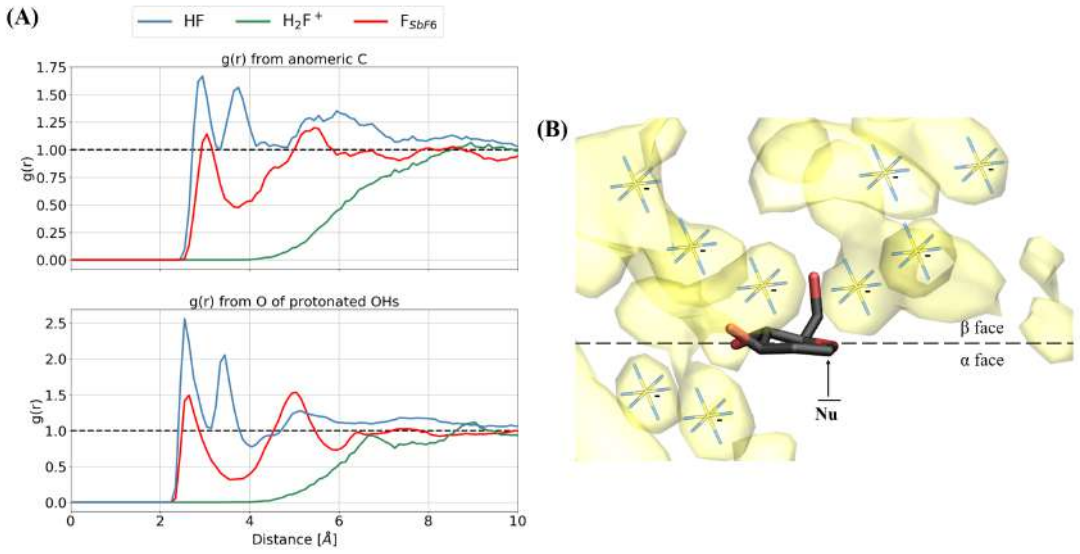


Figure 5.4: **(A)** Radial pair distribution functions for the anomeric carbon and the oxygens of protonated hydroxyls of 2-deoxy-β-D-glucosyl cation with the solvent molecules, HF (blue line), H_2F^+ (green line) and SbF_6^- (red line). The two peaks at distances smaller than 4 Å in g(r) for the anomeric carbon and the protonated hydroxyls with HF solvent molecule correspond to the radial distribution of F_{HF} and H_{HF}, respectively. The distance difference between these two peaks correspond to the bond distance of HF, $\sim$ 0.9 Å. Also, the first solvation shell of HF with the glycosyl cation hydroxyls is $\sim$ 1.5 Å closer compared to the anomeric carbon due to their ability to create hydrogen bonds besides the electrostatic interactions. **(B)** Averaged occupation of SbF_6^- molecules along all the classical MD simulation shown as transparent yellow surfaces. The SbF_6^- molecules are placed only for representation purposes. The negative charged is accumulated in the β face facilitating the nucleophilic attack on the α face, as seen experimentally for a similar glycosyl cation [30].

An enhanced QM/MM MD simulation of the conformational FEL was run to obtain the conformational free energy landscape of the oxocarbenium ion surrounded by the superacid molecules. An unique minimum was observed covering the region between E_3 and ^4E conformations (θ=45° and ϕ=170-230°). Even though the minimum is displaced a little bit compared to the isolated molecule, it is still placed in the same region and including ^4E as the most stable conformer, as observed experimentally.

The computed conformational properties of the 2-deoxy-β-D-glucosyl cation in superacid media and the distribution of the superacid solvent molecules around it fit perfectly with the experimental results and thus, our superacid force field satisfactorily reproduce and explain the experiments.

Comparing the FEL of the isolated glycosyl cation in superacid media with the one obtained for the isolated cation, we can see that the E_4 region is less stable in superacid due to the lower pseudo-axiality of the exocyclic groups at C3 and C4 and the lower number of possible hydrogen bond interactions with the solvent (see Figure 5.5). The bond distance between the anomeric carbon and the ring oxygen is maintained to the value found for the isolated case, 1.27 Å. The value of the C2-C1-O5-C5 angle is slightly different, from 1.7° to -2.5°, but still indicates the almost planar arrangement between these four atoms. Finally, the anomeric charge increases in the superacid environment compared to the isolated molecule, with values of 0.57 and 0.29, respectively (Table 5.2, page 130). Thus, the superacid media does not significantly affect the conformational free energy landscape of the glycosyl cation, neither its intramolecular properties such as bond distances and dihedral angles, but it enhances its cationic character.

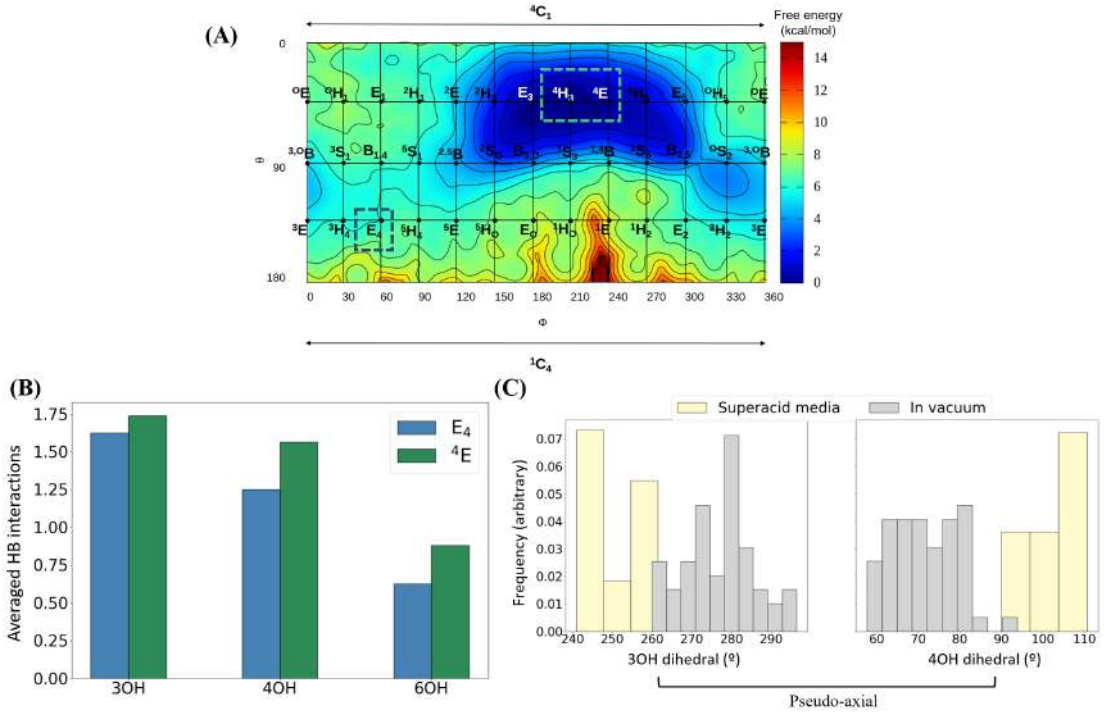


Figure 5.5: **(A)** Conformational FEL of polyprotonated 2-deoxy-β-D-glucosyl cation in superacid media, energy values are given in kcal·mol^{-1} and each contour line corresponds to 1 kcal·mol^{-1}. **(B)** Averaged hydrogen bond interactions from the protonated hydroxyls in C3, C4 and C6 of the glycosyl cation to the superacid solvent molecules in the ^{4}E and E$_4$ conformers. **(C)** Orientation of the hydroxyl group in C3 and C4 for the E4 conformer, in gas-phase (grey) and in superacid media (yellow)

5.2.1.4 The 2-deoxy-β-D-glucosyl cation in the active site of a GH enzyme

To discern the ability of the superacid media to mimic the active site of glycosidases we computed the conformational FEL of the glycosyl cation in the active site of the β-D-glucanase *Saccharomyces cerevisiae* Gas2, ScGas2, discussed in reference [117]. We chose this specific enzyme because in that previous work of the group, classical MD simulations were performed to unravel the hydrogen bond interactions involving the C2 hydroxyl group. They observed that the 2-OH group changes hydrogen bond partner, from the nucleophile (so-called the *on* configuration) to a solvent water molecule (named as the *off* configuration) with populations around 50%. The transglycosylation reaction was modelled using QM/MM metadynamics from both *on* and *off* configurations demonstrating that the transglycosylation reaction from the *on* configuration was feasible while the removal of the 2-OH···nucleophile interaction (*off* configuration) raised the free energy barrier significantly. The same TS was found for the reverse reaction, corresponding to a glycosylation reaction.

The TS found in the *off* configuration was chosen due to a higher similarity in the active site interactions with the 2-deoxy molecule (see Figure 5.14 in the methods section, page 147). We removed *in silico* the hydroxyl in position 2 to simulate the 2-deoxy-β-D-glucosyl cation studied in this project. Only the glycosyl cation was included in the quantum region of the QM/MM simulation. Using this strategy, we were able to avoid the progress of the reaction and obtaining the FEL of the corresponding glycosyl cation, which mimics very well the TS of the reaction.

The computed conformational FEL shows that the sugar ring is even more restricted in the enzyme active site than in superacid media. The 2-deoxy-β-D-glucosyl cation is only able to explore the ^{4}E region (θ=45° and ϕ=210-270°, Figure 5.6). The E_4 conformer is not possible into the active site as it will probably disrupt the hydrogen bond interaction between 4-OH and Tyr307 and the stacking between the saccharide in the -2 subsite and Tyr107, both important for the substrate chain stabilization [117] (Figure 5.6**B**). Regarding the charge distribution around the oxocarbenium ion, a similar pattern compared to the superacid media is found; a much more negative electrostatic potential in the β face of the sugar is observed (Figure 5.6**C**) in the enzyme active site compared to the α face of the sugar.

With respect to the intramolecular properties, we can see that the bond distance between the ring oxygen and the anomeric carbon has the same value in the enzyme as in superacid; 1.27 Å. The same distance was observed in the isolated molecule (Table 5.2). Similarly, the C2-C1-O5-C5 angle has a very similar value to the two previous systems, 0.52°. Therefore, intramolecular bond distances are not affected by the environment. On the contrary, the anomeric charge decreases in the enzyme compared to the one found in the superacid media, from 0.57 to 0.40, but it is still higher than the one obtained for the isolated glycosyl cation (0.29).

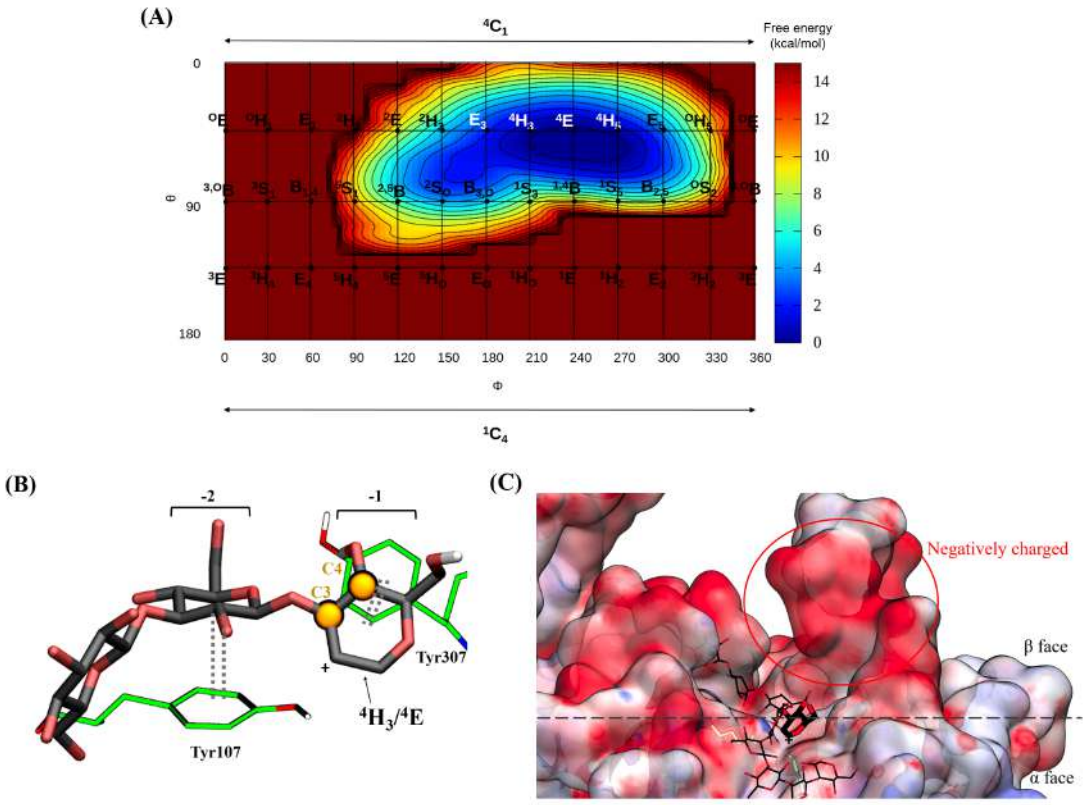


Figure 5.6: **(A)** Conformational FEL of 2-deoxy-β-D-glucosyl cation on the active site of *Sc*Gas2, energy values are given in kcal·mol^{-1} and each contour line corresponds to 1 kcal·mol^{-1}. **(B)** Active site interactions for the substrate with *Sc*Gas2 enzyme. The 2-deoxy-β-D-glucosyl cation placed in the -1 subsite displays a hydrogen bond (thin dashed line) and a stacking (thick dashed line) interactions with Tyr307 and the carbohydrate placed in -2 subsite has also a stacking interaction with Tyr107. To change the sugar ring conformation close to the E_4 region, C3 and C4 should move up and down, respectively; this would disrupt all the showed interactions which are very important to stabilize the sugar chain in the active site of the enzyme. **(C)** Calculated electrostatic potential [186] using default values for *Sc*Gas2 enzyme. The color scheme goes from red, for negative values, to blue, for positive values.

	Enviroment			
Property	**In vacuum**	**In vacuum**	**Superacid**	**On-enzyme**
	(2-deoxy-β-glucose)	(2-deoxy-β-glucosyl cation)		
C1-O5 (Å)	1.44 ± 0.01	1.27 ± 0.02	1.27 ± 0.01	1.27 ± 0.02
C1-C2 (Å)	1.53 ± 0.05	1.45 ± 0.02	1.46 ± 0.03	1.47 ± 0.03
C2-C1-O5-C5 (°)	-33 ± 47	1.7 ± 1.0	-2 ± 11	0.5 ± 15
3OH orientation (°)	-170 ± 6	-165 ± 8	-153 ± 8	-148 ± 9
4OH orientation (°)	-173 ± 5	168 ± 7	161 ± 8	184 ± 8
C1 charge (ESP)	-0.19 ± 0.04	0.29 ± 0.02	0.57 ± 0.04	0.40 ± 0.04

Table 5.2: Properties of 2-deoxy-β-D-glucose, the neutral molecule (only in vacuum) and 2-deoxy-β-D-glucosyl cation in vacuum, superacid media and on-enzyme. The anomeric carbon charge listed in this table, C1, is a sum of the charges in C1, H1 and O5 (including O1 and HO1 for the neutral 2-deoxy-β-D-glucose).

5.2.2 2-deoxy-α-L-arabinofuranosyl cation in different environments

To extend our knowledge of how the superacid media affects glycosyl cations, a similar study to the one carried out for 2-deoxy-β-D-glucosyl cation, a pyranosyl cation, was carried out for 2-deoxy-α-L-arabinofuranosyl cation, a furanosyl cation, whose NMR spectrum was also being recorded by our experimental collaborators.

5.2.2.1 The isolated 2-deoxy-α-L-arabinofuranose

To study the intrisic properties of a five-membered sugar ring (furanose derivative) we first computed the conformational FEL of the isolated 2-deoxy-α-L-arabinofuranose by means of CPMD metadynamics. The conformational FEL obtained is very similar to the one found for α-L-arabinofuranose, analysed in Chapter 4 (Figure 4.5 in page 91), where all conformations lie in an energy window of $\sim$ 5 kcal·mol^{-1}, narrower than the one obtained for pyranose compounds studied along this work, see Figure 5.7. Despite the lack of the hydrogen bond interaction between the hydroxyls at C2 and C5, in the case of the 2-deoxy sugar, $^1T_2$ is still the most stable conformation as it conserves the hydrogen bond interaction between the hydroxyl at

C1 and the hydroxyl at C3 (Figure 5.7, bottom-right) and has less steric hindrance among ring hydrogens compared to the second most stable conformation, which fall in the region E_3-$^4T_3$-^4E-^4T_O-E_O, 2 kcal·mol^{-1} higher in energy. The conformations in this second local minimum are stabilized by hydrogen bond interactions between the hydroxyls of C3 and C5 (Figure 5.7, bottom-left), even though this interaction was only observed 12% of the simulated time.

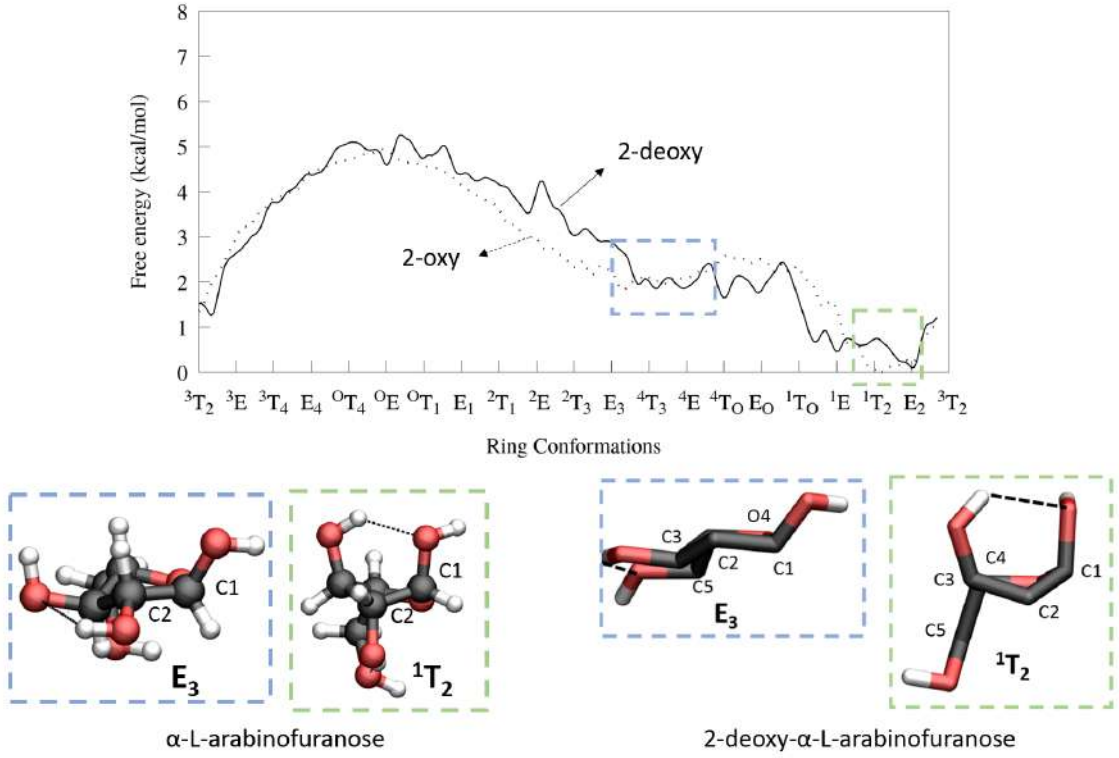


Figure 5.7: Conformational FEL of 2-deoxy-α-L-arabinofuranose as a continuous line and conformational FEL of α-L-arabinofuranose represented as a dashed line, also shown in Figure 4.5**A**. α-L-arabinofuranose and 2-deoxy-α-L-arabinofuranose representative structures of E_3 and $^1T_2$ conformations with their corresponding intramolecular hydrogen bond interactions represented as dashed lines.

5.2.2.2 The isolated 2-deoxy-α-L-arabinofuranosyl cation

Using the same methodology we calculated the conformational FEL of the corresponding isolated glycosyl cation (2-deoxy-α-L-arabinofuranosyl cation). The results show that it exhibits a different conformational FEL compared to the corresponding neutral molecule (explained above), but also, to the α-L-arabinofuranosyl cation (i.e. the same cation with the 2-OH group, Figure 4.6 in page 92 and Figure 5.8).

The main minima of the furanosyl cation corresponds to the region between ^{2}E-^{4}T$_{O}$ conformations, centred in E_3, as this conformation has the ability to create a hydrogen bond between the hydroxyls at C3 and C5 (Figure 5.8, bottom-right; $\sim$ 12% of the simulation time). $^{1}T_2$ loses its stability compared to the neutral 2-deoxy-α-L-arabinofuranose molecule due to the lack of the hydrogen bond between C1 and C3 hydroxyls. The second minima corresponds to the region around ^{3}E at $\sim$ 6 kcal·mol^{-1} above in energy. In the 2-oxy counterpart, i.e., the α-L-arabinofuranosyl cation, the minima order is inverted, as discussed in Chapter 4, page 88. The ^{3}E destabilization in the 2-deoxy-α-L-arabinofuranosyl cation is due to the lack of the hydrogen bond interaction between the hydroxyls at C2 and C5. Besides the sugar ring conformational stabilities, the main difference between the neutral molecule and the furanosyl cation of 2-deoxy-α-L-arabinofuranose is the C1-O5 bond distance, which shrinks from 1.45 Å (neutral molecule) to 1.31 Å (cation) for the E_3 conformation and from 1.43 Å (neutral molecule) to 1.27 Å (cation) for the ^{3}E conformation. These values are very similar to the ones obtained for the TS of the hydrolysis reaction of 4-nitrophenyl α-L-arabinofuranose in *Aspergillus kawachii* α-L-arabinosidase (AkAbfB) in Chapter 4 (page 95). It was found that the arabinose ring adopts an E_3 TS conformation, with a C1-O5 bond distance of 1.29 Å, a value in-between the ones found here for the arabinofuranosyl cation.

Our results also show that the bond distance between C1-C2 gets shorter in the furanosyl cation compared to the neutral 2-deoxy-arabinofuranose molecule (from 1.54 Å to 1.46 Å) in both ^{3}E and E_3 conformations. The required planarity arrangement of C2-C1-O4-C4 for an oxocarbenium ion is here observed that it does not change much in the furanosyl cation compared to the neutral molecule (from 13° and 0.45° to -0.58° and -0.95° for the E_3 and ^{3}E conformations, respectively). Moreover, the anomeric charge switches from a negative value (-0.53, E_3, or -0.52, ^{3}E) to a positive one (0.13, E_3, or 0.16, ^{3}E), see Table 5.3 in page 139.

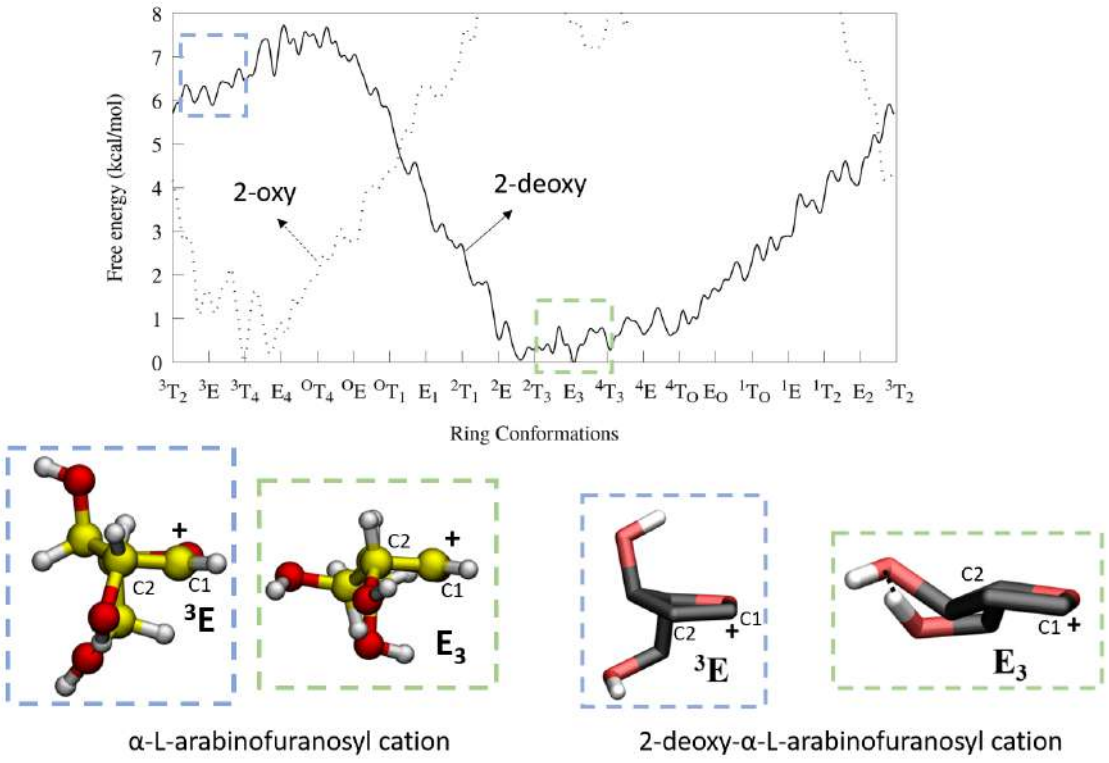


Figure 5.8: Conformational FEL of 2-deoxy-α-L-arabinofuranosyl cation as a continuous line and conformational FEL of α-L-arabinofuranosyl cation represented as a dashed line, also shown in Figure 4.6 in page 92. α-L-arabinofuranosyl cation (from Figure 4.6) and 2-deoxy-α-L-arabinofuranosyl cation representative structures of E_3 and ^3E conformations with their corresponding intramolecular hydrogen bond interactions represented as dashed lines.

5.2.2.3 The 2-deoxy-α-L-arabinofuranosyl cation in superacid media

The same strategy used for the polyprotonated 2-deoxy-β-D-glucosyl cation in superacid media was used to study the polyprotonated 2-deoxy-α-L-arabinofuranosyl cation in superacid media. From the classical MD simulation of 2-deoxy-α-L- arabinofuranosyl cation in superacid with all the hydroxyl groups protonated, it can be observed that the F atoms in HF and SbF_6^- molecules distribute around C1 at $\sim$ 3 Å due to the electrostatic interactions; whereas they distribute closer around the hydroxyls as they are also able to create hydrogen bond interactions, at $\sim$ 1.5 Å of the hydroxyl hydrogens, see Figure 5.9. Additionally, a larger concentration of SbF_6^- in the β face of the sugar cation is also observed, as seen for the β-D-glucosyl cation, compared to the α face. As stated for the glucosyl cation, this means that the nucleophiles will have a preference to attack the anomeric carbon from the α face where there is less electrostically hindered.

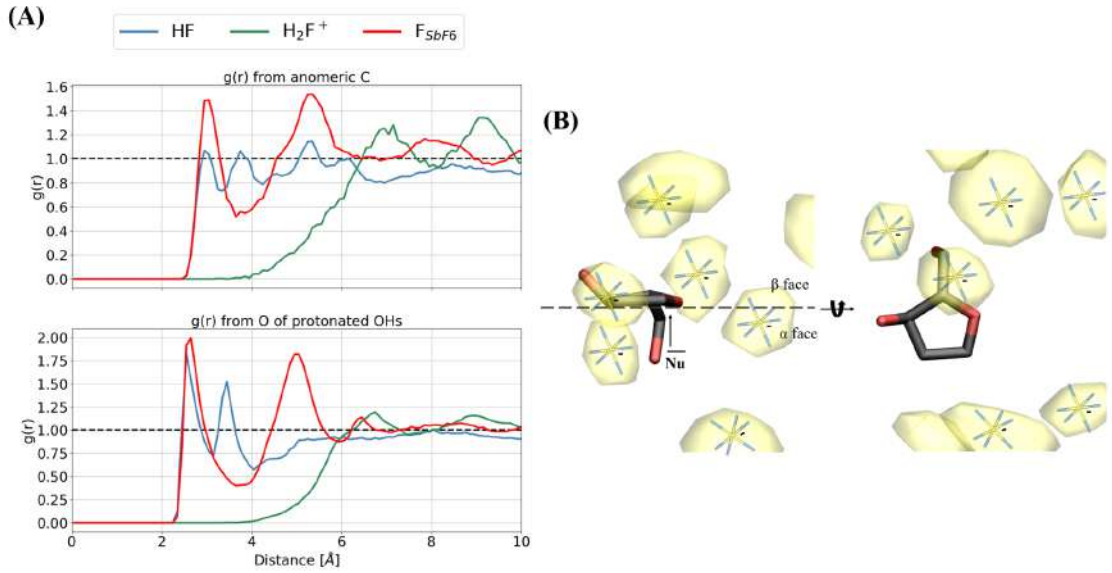


Figure 5.9: **(A)** Radial pair distribution functions for the anomeric carbon and the oxygens of protonated hydroxyls of 2-deoxy-α-L-arabinofuranosyl cation with the solvent molecules, HF (blue line), H_2F^+ (green line) and SbF_6^- (red line). The two peaks at distances smaller than 4 Å in g(r) for the anomeric carbon and the protonated hydroxyls and HF solvent molecule correspond to the radial distribution of F_{HF} and H_{HF}, respectively. The distance difference between these two peaks correspond to the bond distance of HF, $\sim$ 0.9 Å. Also, the first solvation shell of the HF solvent molecule with the hydroxyls is $\sim$ 1.5 Å closer compared to the anomeric carbon due to their ability to create hydrogen bonds besides the electrostatic interactions. **(B)** Averaged occupation of SbF_6^- along all the classical MD simulation shown as transparent yellow surfaces. The SbF_6^- molecules are shown only for representation purposes. The negative charged is accumulated in the beta face facilitating the nucleophilic attack on the alpha face.

The conformational FEL of the polyprotonated 2-deoxy-α-L-arabinofuranosyl cation in superacid media changes considerably compared to the isolated cation. All the conformations lie in an energy window of 2 kcal·mol^{-1}, being ^{3}E (ϕ=18°) and E_3 (ϕ=198°) the most stable conformations with an energy difference of 1 kcal·mol^{-1}, and thus, being practically isoenergetic. The 3OH orientation for the ^{3}E conformation in superacid media gets close to the value found for the E_3 conformation either in the isolated case and surrounded by superacid molecules. Additionally, the small energy difference between the two main minima is due to the almost same number of hydrogen bond interactions between the furanosyl cation and the superacid media in both conformations (Figure 5.10). This is different from what was observed for the to 2-deoxy-β-D-glucosyl cation where there was a clear conformational preference due to a higher number of solvent interactions for the most stable conformation.

This indicates that the surroundings of the furanosyl cation, which has a smaller number of protonated hydroxyl groups and a smaller size compared to the glucosyl cation, barely varies when changing conformations, from E_3 to ^3E.

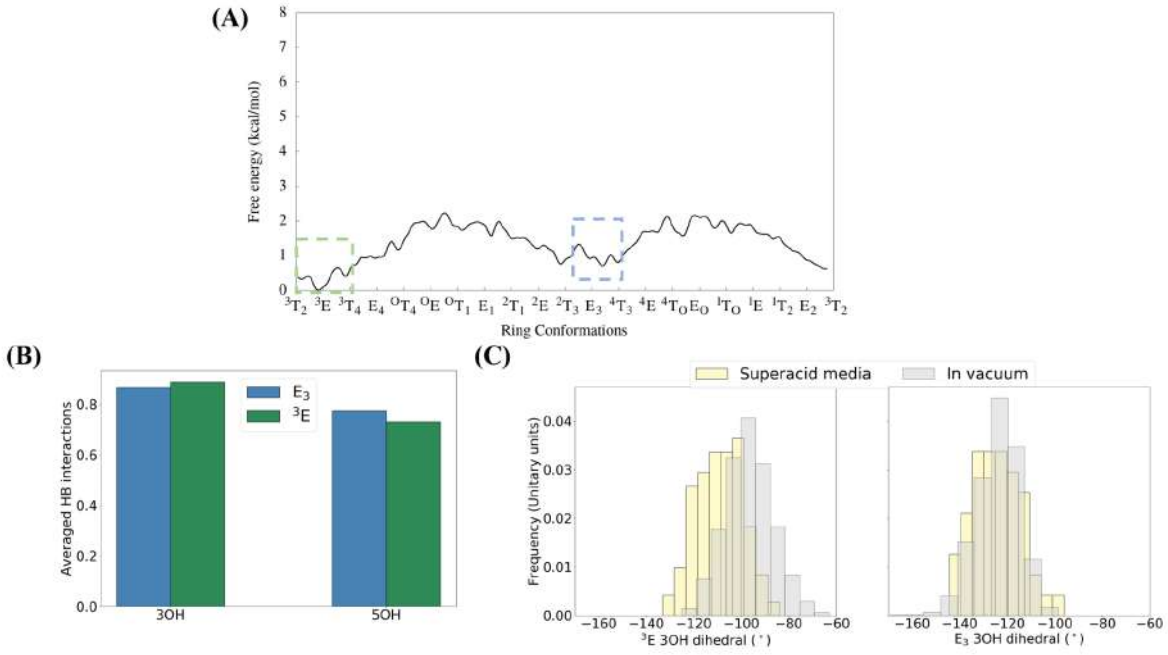


Figure 5.10: **(A)** Conformational FEL of polyprotonated 2-deoxy-α-L-arabinofuranosyl cation in superacid media, energy values are given in kcal·mol^{-1}. **(B)** Averaged hydrogen bond interactions with the superacid solvent in ^3E and E_3 conformations for the protonated hydroxyls placed at C3 and C5. **(C)** Orientation of the hydroxyl group in C3 for ^3E and E_3 conformations, in gas-phase (grey) and in superacid media (yellow).

The C1-O4 bond distance of the furanosyl cation in superacid has a value of 1.27 Å in both E_3 and ^3E conformations, contrary, to what was found for the isolated furanosyl cation. The other relevant geometrical properties, C1-C2 distance and the C2-C1-O4-C4 angle, are conserved in both conformations with respect to the isolated cation. However, as it was previously observed for the glucosyl cation, the anomeric charge increases in the superacid media, for both the E_3 conformation, from 0.13 (in vacuum) to 0.26; or the ^3E conformation, from 0.16 to 0.24 (see Table 5.3, page 139).

Therefore, the calculations of the furanosyl cation in superacid show that there is no clear prefernce between E_3 and ^3E due to the same solvent-sugar interactions in both conformations. Additionally, the classical MD simulation of the furanosyl cation in superacid has the same SbF_6^- distribution, concentrated in the β face, observed for the glucosyl cation.

5.2.2.4 The 2-deoxy-α-L-arabinofuranosyl cation in the active site of a GH enzyme

To fully understand the superacid media ability to mimic the active site of glycosidases, the conformational FEL of 2-deoxy-α-L-arabinofuranosyl cation in the active site of *Aspergillus kawachii* α-L-arabinofuranosidase (AkAbfB) was calculated. The TS snapshot was taken from the previously calculated hydrolysis reaction in Chapter 4, see Figure 4.11 in page 98. As done for the pyranose case, the C2 hydroxyl was removed *in silico* and the reaction progress was avoided by only including the sugar into the QM region, excluding the catalytic residues.

Contrary to what is found for the 2-deoxy-β-D-glucosyl cation on-enzyme, the conformational FEL of 2-deoxy-α-L-arabinofuranosyl cation is not restricted to a unique conformation, but instead, all the possible furanose conformers lie in a $\sim$ 3 kcal·mol^{-1} energy window, much similar to what was found for the furanosyl cation in superacid media. The $^{3}T_{2}$-E_{4} region, where the ^{3}E conformer lies, and the ^{2}E-$^{4}T_{3}$ region, where E_{3} is placed, are the most stable ones with practically no energy difference between them. Both E_{3} and ^{3}E conformations are possible in the active site of AkAbfB enzyme as the main substrate-protein interactions are maintained in both conformations; forming hydrogen bonds from C3 hydroxyl to Gly296 and Asp219 and from the hydroxylmethyl exocyclic group to Asp219. As we are studying the 2-deoxy-α-L-arabinofuranosyl cation, the hydrogen bonds from 2OH to the carbonyl oxygen of Glu221 and to the backbone amide of Asp297 seen in Chapter 4 (Figure 4.11) for α-L-arabinofuranose are lost. This gives more mobility to the substrate placed in the active site and, thus, is able to accommodate a wide range of conformations maintaining the other aforementioned hydrogen bonds. In order to stabilize the ^{3}E conformation the substrate is displaced a bit above in the active site, so that C1-Glu221 and O4-Asn222 distances are far-off compared to the E_{3} conformation (see Figure 5.11). The 2-deoxy-arabinofuranosyl cation in the E_{3} conformation resembles the TS of the reaction found for AkAbfB (Figure 4.11 in Chapter 4, page 98) in terms of sugar ring conformation and distance between the anomeric carbon and the carbonyl oxygen of Glu221, the attacking nucleophile. The p-nitrophenol moiety, the leaving group of the hydrolysis reaction, leaves completely the active site to go into the solvent for the whole QM/MM MD metadynamics simulation due to the high-solvent exposure of the active site. Interestingly, the more mobile furanosyl cation in the active site of AkAbfB seems to be mimicked in superacid, with a smaller number of solvent-substrate interactions compared to the pyranosyl cation, which it is also tighter placed in the active site of *Sc*Gas2 enzyme, see Figure 5.12.

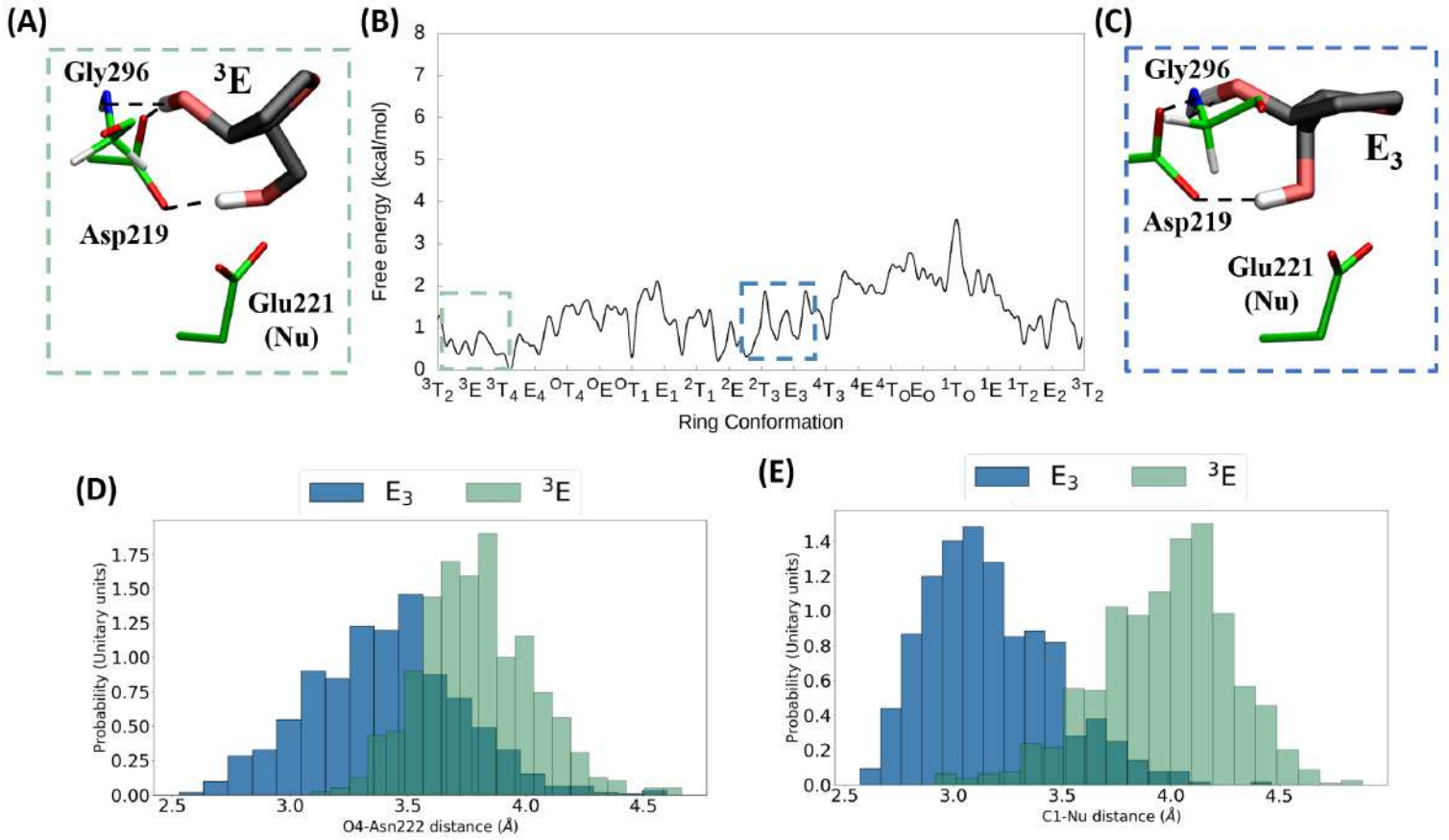


Figure 5.11: **(A)** Representative structure of 2-deoxy-α-L-arabinofuranosyl cation in a ^{3}E conformation in the active site of the AkAbfB enzyme. The main protein-substrate interactions are represented as black dashed lines. **(B)** Conformational FEL of 2-deoxy-α-L-arabinofuranosyl cation on the active site of AkAbfB α-L-arabinofuranosidase. Energy values are given in kcal·mol^{-1}. **(C)** Representative structure of 2-deoxy-α-L-arabinofuranosyl cation in a E_3 conformation in the active site of the AkAbfB enzyme. The main protein-substrate interactions are represented as black dashed lines. **(D)** O4-Asn222 distance distribution of E_3 and ^{3}E conformations of 2-deoxy-α-L-arabinofuranosyl cation on-enzyme. **(E)** C1-Glu221 (Nu) distance distribution for the same conformations.

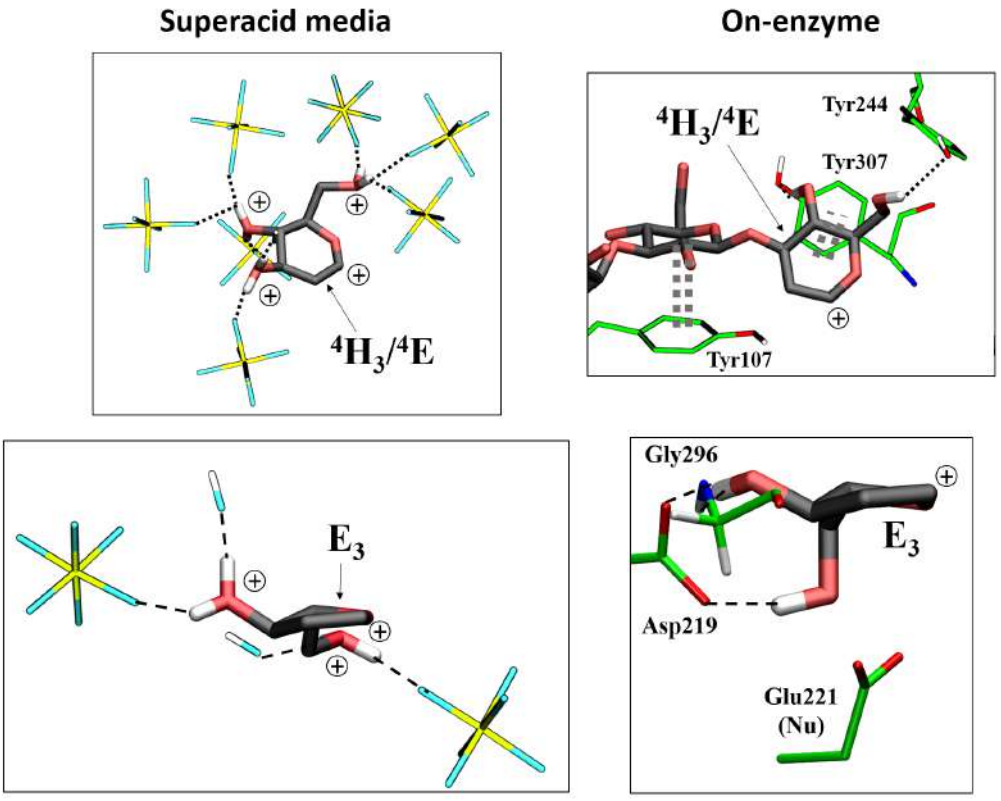


Figure 5.12: Interactions present in superacid media (left) and on the active site of an enzyme (right), glucanase and α-L-arabinofuranosidase, with 2-deoxy-β-D-glucosyl (top) and 2-deoxy-α-L-arabinofuranosyl (bottom) cations. Hydrogen bond/salt bridges interactions are represented as black dashed lines and π-stacking interactions are represented as double grey dashed lines.

The intrinsic furanosyl cation properties are very similar to the ones found when it is surrounded by superacid molecules in both E_3 and ^3E conformations, besides the small decrease of the anomeric charge when surrounded by AkAbfB residues (from 0.26 to 0.22 for the E_3 conformation and from 0.24 to 0.19 for the ^3E conformation); the same tendency seen for the pyranosyl cation, see Table 5.3. In Figure 5.13 all the studied properties for the pyranosyl and furanosyl cations are shown.

		Environment			
Property		**In vacuum**	**In vacuum**	**Superacid**	**On-enzyme**
		(2-deoxy-α-arabinofuranose)	(2-deoxy-α-arabinofuranosyl cation)		
C1-O4 (Å)	$\mathbf{E}_3$	1.45 ± 0.01	1.31 ± 0.02	1.27 ± 0.02	1.27 ± 0.02
	$^3\mathbf{E}$	1.43 ± 0.04	1.27 ± 0.002	1.27 ± 0.02	1.27 ± 0.02
C1-C2 (Å)	$\mathbf{E}_3$	1.54 ± 0.03	1.46 ± 0.03	1.47 ± 0.03	1.46 ± 0.03
	$^3\mathbf{E}$	1.54 ± 0.04	1.46 ± 0.03	1.46 ± 0.03	1.46 ± 0.03
C2-C1-O4-C4 (°)	$\mathbf{E}_3$	13 ± 9	-0.58 ± 7	-0.51 ± 8	1.4 ± 10
	$^3\mathbf{E}$	0.45 ± 8	-0.95 ± 1.7	-1.5 ± 8	-2 ± 9
3OH orientation (°)	$\mathbf{E}_3$	-156 ± 9	-124 ± 9	-124 ± 10	-134 ± 9
	$^3\mathbf{E}$	-84 ± 9	-97 ± 10	-109 ± 10	-106 ± 8
C1 charge (ESP)	$\mathbf{E}_3$	-0.53 ± 0.09	0.13 ± 0.06	0.26 ± 0.08	0.22 ± 0.06
	$^3\mathbf{E}$	-0.52 ± 0.1	0.16 ± 0.07	0.24 ± 0.08	0.19 ± 0.06

Table 5.3: Properties of 2-deoxy-α-L-arabinofuranose, the neutral molecule (only in vacuum) and 2-deoxy-α-L-arabinofuranosyl cation in vacuum, superacid media and on-enzyme for the E_3 and ^{3}E conformations. The anomeric carbon charge listed in this table, C1, is a sum of the charges in C1, H1, O5 and C2 (including O1 and HO1 for the neutral molecule).

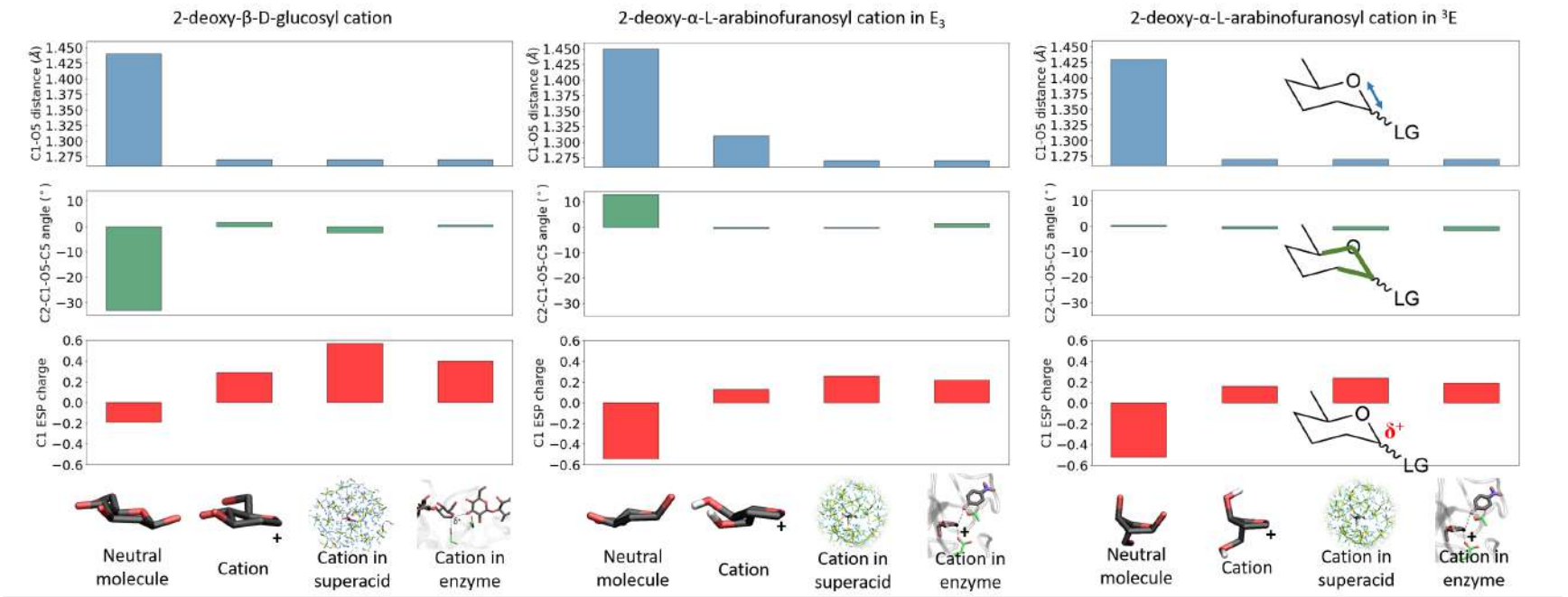


Figure 5.13: Comparison of the intramolecular properties of 2-deoxy-β-D-glucosyl (in ^{4}E conformation) and 2-deoxy-α-L-arabinofuranosyl (in E_3 and ^{3}E conformations) in its neutral form in vacuum and as a glycosyl cation in vacuum, in superacid and enzymatic environments, from left to right. From top to bottom, the oxygen ring and anomeric carbon bond distance (blue), the C2-C1-O5/O4-C5/C4 dihedral angle indicating the molecule planarity (green) and the anomeric charge (red).

5.3 Discussion and Conclusions

The objective of this chapter was to quantify the intrinsic properties of glycosyl cations in different environments (in vacuum, superacid media and on the active site of a GH enzyme) and analyze how these properties are affected by the superacid environment. These glycosyl cations are being isolated experimentally, thus our results will serve to answer the question whether the conditions used experimentally (the superacid media) truly reflect the environment of the glycosyl cation in a GH enzyme. First, the 2-deoxy-β-D-glucosyl cation, a 6 membered sugar ring, was systematically studied as isolated molecule, in superacid media and in the active site of the *Sc*Gas2 glucosidase enzyme. The results obtained indicate a clear preference for the ^{4}E conformation in all environments, with the geometrical properties little affected by the superacid environment. However, the electronic properties slightly differ with the environment of the cation. The superacid media enhances the cationic character by a factor of 1.4 compared to the enzymatic environment and almost doubling the anomeric charge compared to the isolated molecule. This is most probably due to a higher number of negative molecules (SbF_6^-) interacting with the glycosyl cation compared to the enzymatic environment, which polarizes the sugar. Additionally, our results also show that the density of negative charges around the glycosyl cation in the superacid media is higher in the β face of the sugar compared with the α face. This nicely explains the stereoselectivity of nucleophillic reactions observed experimentally [30].

To find out whether our findings can be extended to glycosyl cations of any size, we have also studied the 2-deoxy-α-L-arabinofuranosyl cation, a 5 membered sugar ring, which was recently isolated in superacid by our collaborators. As observed for the pyranosyl cation, all the geometrical properties previously mentioned are also conserved along all the studied environments. Contrary from the conformational FELs of pyranose-like molecules, the FELs of furanose-like molecules (either neutral or cationic and independently of the environment) do not show a clear preference for any specific conformation as all them lie in a 3 kcal·mol^{-1} energy window from the two most stable conformers, E_3 and ^{3}E, which are almost isoenergetic. The anomeric charge shows the same trend seen previously for pyranoses, the superacid media enhances the cationic character of the glycosyl cation compared to the enzymatic or vacuum environments. The anomeric charge of the furanosyl cation is systematically smaller than the anomeric charge in the glucopyranosyl cation probably due to a decrease of interactions either in the superacid media or in the enzymatic environment. Moreover, the highest density of negative charges around the furanosyl cation in superacid media is found in the β face of the sugar, as observed previously for the

glucosyl cation. Thus, our results predict that nucleophillic attack will preferentially take place on the α face of the furanosyl cation.

To recapitulate, the superacid media do not alter the geometrical intramolecular properties of the glycosyl cations; the conformations explored, the bond distances and dihedral angles are similar in superacid and in enzyme, for both pyranose and furanose-type of cations. On the contrary, the cationic character is slightly enhanced in superacid compared to the enzymatic environment. Our results confirm the predictions of George Olah [182] that the multidentate interactions of the active site of an enzyme with the carbohydrate stabilizing the glycosyl cation can be mimicked in superacid media, as we can see salt bridges between SbF_6^- and the protonated hydroxyl groups in all simulations.

The main conclusions reached in this chapter are the following:

- The density of the SbF_6^- molecules in the superacid media, analysed by classical MD simulations, is much higher in the β face of the glycosyl cation compared to the α face, either in pyranoses or furanoses. This explains the stereoselectivity seen experimentally in nucleophillic reactions.

- The conformational FEL of the pyranosyl cation is not significantly affected by the environment (isolated molecule, superacid media or on-enzyme) all displaying ^{4}E as the most stable conformation.

- The conformational FEL of the furanosyl cation changes significantly from the isolated case (in which E_3 is the most stable conformation) to the superacid or enzymatic environments (in where the E_3 and ^{3}E conformations are almost isoenergetic).

- The structural properties of glycosyl cations in superacid media (conformational preferences, bond distances and ring planarity) are very similar to the ones found in the enzyme active site.

- The cationic character of the glucosyl and furanosyl cations is slightly enhanced by the superacid media compared to the enzymatic environment due to the higher polarization effect of the surrounding SbF_6^- molecules.

- The effect of the multidentate interactions formed between enzyme active site residues and the glycosyl cation is well reproduced in superacid media as we can see salt bridges between SbF_6^- and the protonated hydroxyl groups along the whole classical MD simulations.

5.4 Computational details

5.4.1 Conformational FELs in vacuum

The 2-deoxy-β-D-glucose, 2-deoxy-β-D-glucosyl cation (hereafter also called the pyranose cases), 2-deoxy-α-L-arabinofuranose and 2-deoxy-α-L-arabinofuranosyl cation (hereafter called the furanose cases) conformational FELs in vacuum were computed according to the Car-Parrinello (CP) method [66] using the CPMD code [76]. They were enclosed in isolated cubic boxes of 12.94 x 10.39 x 13.65 $Å^3$, 14.8 x 13.5 x 10.3 $Å^3$, 12.5 x 12.5 x 12.5 $Å^3$ and 12.5 x 12.5 x 12.5 $Å^3$, respectively. A fictitious electron mass of 850 and 500 atomic units (a.u.) was used for the CP Lagrangian for the pyranose and furanose cases, respectively, and a time step of 0.12 fs was used to ensure that the adiabacity of the fictitious kinetic energy of the electrons was smaller than 10^{-5} a.u./atom. The Kohn-Sham orbitals were expanded in a plane wave basis set with a kinetic energy cut-off of 70 Ry. Ab initio pseudopotentials, generated within the Troullier-Martins scheme, were employed. The Perdew, Burke, and Ernzerhoff generalized gradient-corrected approximation [139] was selected in view of its good performance [140, 141]. The metadynamics algorithm [81], provided by the Plumed 2 plugin [17], was used to explore the conformational free energy landscape taking as collective variables (CVs) the Cremer-Pople puckering coordinates [10] qx, qy and qz divided by the amplitude (Q) for the pyranose cases and taking as collective variables the pseudorotational phase (ϕ) puckering coordinate [18], as well as a dihedral angle accounting for the rotation of the sugar hydroxymethyl group for the furanose cases. Initially, the height of the gaussian-like potential was set to 0.6 kcal·mol^{-1} and added every 250 (pyranoses) and 500 (furanoses) molecular dynamics steps. Once the whole free energy space was explored, the height of the Gaussian terms was reduced to 0.2 kcal·mol^{-1} to facilitate convergence of the FEL. The width of the collective variables was set according to their oscillations in the free dynamics which corresponded to 0.035, 0.03, 0.02 rad (2-deoxy-β-D-glucose); 0.05, 0.05 and 0.015 rad (2-deoxy-β-D-glucosyl cation) and 0.035 and 0.1 rad (2-deoxy-arabinofuranose and 2-deoxy-arabinofuranosyl cation). The simulations were stopped when energy differences among wells remain constant and the standard deviation of the oscillations were below 0.5 kcal·mol^{-1}, which corresponded to 9900, 16500, 12800 and 35171 deposited gaussians for 2-deoxy-β-D-glucose, 2-deoxy-β-D-glucosyl cation, 2-deoxy-α-L-arabinofuranose and 2-deoxy-α-L-arabinofuranosyl cation, respectively. The FEL was projected against the polar coordinates θ and ϕ for representation purposes using the reweighting method described in Branduardi *et. al.* [187] for the pyranose cases and projected into the ϕ coordinate for representation purposes for the furanose cases.

5.4.2 Conformational FELs in superacid media

5.4.2.1 Parametrization of a superacid force-field

The anionic species present in the superacid solutions were identified by our collaborators by ^{19}F NMR spectroscopy (see Supplemental Figure 5.18). A mixture of SbF_6^- and $Sb_2F_{11}^-$ species, with a major concentration of SbF_6^-, was determined. The superacid solvent was decided to be simplified by only modelling SbF_6^- and H_2F^+ species. To be able to simulate the required neutral system for the classical MD simulations and also reach the molar percentage of SbF_5 used experimentally (21.6%), HF was also modelled in the superacid solvent. The parametrization approach used was the following:

1. **SbF_6^- molecule**. MCPB.py [188] from AmberTools16 [178] was used, with no differentiation between the small and large models. The first step consisted on preparing the input files for Gaussian (version g09 [189]) which result from geometry optimization, force constant calculation and the Merz-Kollman RESP charge calculation. We omitted the geometry optimization step as we used the optimized structure obtained by Kim *et. al.* [190] by CPMD simulations in which SbF_6^- has an octahedral structure with an average bond length for Sb-F of 0.925 Å when surrounded by HF molecules. For the force constant and RESP charge calculations, the level of theory of HF/LANL2DZ for Sb and HF/6-31G* for F was used. The MCPB.py second step generated the force-fields parameters using the Seminario method [191]. Afterwards, we used the ChgModB to perform the RESP charge fitting and generate the mol2 files. Finally, we generated the Amber tleap input file specifying all the parameter files for SbF_6^-.

2. **H_2F^+ and HF molecules**. The ffTK [192] plugin in VMD [121] was used. ffTK is designed to obtain a complete set of CHARMM-compatible parameters, thus, our last step of this parametrization consisted in changing it to Amber format. The first ffTK step consisted on assigning the LJ/VdW parameters by extracting them from the par_all36_cgenff.prm file by analogy to a similar system. Afterwards, a geometry optimization was done using Gaussian (g09 [189]) in the MP2/6-31G* level of theory. In the ffTK workflow, the next step consists on deriving the partial charges on atoms, following the CHARMM convention of reproducing QM interactions with TIP3P water molecules [134]. In contrast, in the AMBER force field, charges result from fitting to the electrostatic potential surrounding the molecule [193]. Therefore, as we were following the AMBER convention, we derived the RESP charges

at the HF/6-31G* level of theory and, afterwards, we used the antechamber plugin [135] to obtain the prep file with all the corresponding partial charges. Finally, a Hessian calculation was done at the MP2/6-31G* level of theory to later optimize the bond and angles parameters using the ffTK plugin.

5.4.2.2 Classical MD simulations

Classical MD simulations were set up employing the program LEaP included in the Amber suite and using our superacid media force field parameters. The 2-deoxy-β-D-glucosyl and 2-deoxy-α-L-arabinofuranosyl cations with protonated hydroxyl groups were parametrized using antechamber [135] and gaff [136]. As Amber is not designed for small periodic systems, the glycosyl cations were separately solvated by 861 SbF_6^-, 857 or 858 H_2F^+ and 2270 or 2269 HF molecules; for pyranose and furanose cases, respectively. The MD simulations were performed using Amber14 [138]. A thermal equilibration to 300 K was done prior to the MD equilibration in the NPT ensemble, followed by a production phase of 51 ns. The SHAKE algorithm, with an integration step of 2 fs, was used. One frame from the production MD simulation was taken to set up the forthcoming QM/MM MD metadynamics simulations.

5.4.2.3 QM/MM MD metadynamics simulations

The method developed by Laio *et. al.* [77] was used in each system. The quantum (QM) region is composed of the corresponding polyprotonated glycosyl cation (24 atoms, for pyranose and 19 atoms, for furanose), where the atoms are treated using the Car-Parrinello MD [66] method with the PBE functional [139], as done previously for the isolated molecules. The molecular mechanics (MM) region is composed of the superacid solvent (3998 atoms for pyranose and 3999 atoms for furanose) and it is treated with the Amber force-field parameters previously used in the classical MD simulations. The electrostatic interactions between the QM and MM regions were handled via a fully Hamiltonian scheme base on a multilayer approach. The radii of the three different layers, called NN, MIX and ESP, were chosen as 12, 12 and 22 a.u., respectively, after testing according to reference [78]. The 2-deoxy-β-D-glucosyl and 2-deoxy-α-L-arabinofuranosyl cations were enclosed in an isolated cubic box of 13.07 x 12.5 x 12.4 $Å^3$ and 12.6 x 11.6 x 12.9 $Å^3$, respectively. A fictitious electron mass of 500 atomic units (a.u.) and a timestep of 3 a.u. were used for the CP Lagrangian. The Kohn-Sham orbitals were expanded in a plane wave basis set with a kinetic energy cut-off of 70 Ry. Ab initio pseudopotentials, generated within the Troullier-Martins scheme, were employed. After equilibrating the systems in the QM/MM scheme we performed the metadynamics simulations [81], provided by the

Plumed 2 plugin [17], to explore the conformational free energy landscapes taking as collective variables (CVs) the Cremer-Pople puckering coordinates qx, qy and qz divided by the amplitude (Q) and the hydroxymethyl dihedral angle for the pyranosyl cation and the pseudorotational phase and the hydroxymethyl dihedral angle for the furanosyl cation. The height of the gaussian-like potential was set to 0.6 (pyranose) and 0.3 (furanose) kcal·mol^{-1} and added every 250 molecular dynamics steps. The width of the collective variables was set according to their oscillations in the free dynamics which corresponded to 0.05, 0.05, 0.015 and 0.1 rad (pyranose) and 0.035 and 0.1 rad (furanose). The simulations were stopped when energy differences among wells remain constant and the standard deviation of the oscillations were below 0.5 kcal·mol^{-1}, which corresponded to 12824 and 20848 deposited Gaussians. The pyranose conformational FEL was projected against the polar coordinates θ and ϕ for representation purposes using the reweighting method described in Branduardi *et. al.* [187] and projected into the ϕ coordinate for representation purposes for the furanose case.

5.4.3 Conformational FELs on-enzyme

The 2-deoxy-β-D-glucosyl cation conformational FEL was computed in the active site of the β-D-glucanase *Saccharomyces cerevisiae* Gas2, ScGas2, discussed in reference [117]. In that study, MD simulations were performed to unravel the hydrogen bond interactions involving the C2 hydroxyl group. They observed that the 2-OH group changes hydrogen bond partner, from the nucleophile (so-called the *on* configuration) to a solvent water molecule (named as the *off* configuration) with populations around 50%. The transglycosylation reaction was modelled using QM/MM metadynamics from both *on* and *off* configurations demonstrating that the transglycosylation reaction from the *on* configuration was feasible while the removal of the 2-OH···nucleophile interaction (*off* configuration) raised the free energy barrier significantly. The same TS was found for the reverse reaction, the glycosylation. We chose the TS found in the *off* configuration due to a higher similarity in the active site interactions with the 2-deoxy molecule, see Figure 5.14**A**. For the 2-deoxy-α-L-arabinofuranosyl cation conformational FEL on-enzyme, the TS of the hydrolysis reaction in the active site of *Aspergillus Kawachii* abfB (AkAbfB) computed in Chapter 4 was taken (see Section 4.2.4 in page 95), see Figure 5.14**B**.

The hydroxyl in C2 was replaced manually with a hydrogen in both TS snapshots and a QM/MM MD equilibration was performed in each system. The values of the electrostatic layers NN, MIX and ESP were set to 12, 12 and 16 a.u. for the pyranose case and 10, 10 and 18 a.u. for the furanose case. In the pyranose simulation, the QM region included the glycosyl cation in -1 and half ring of the saccharide placed

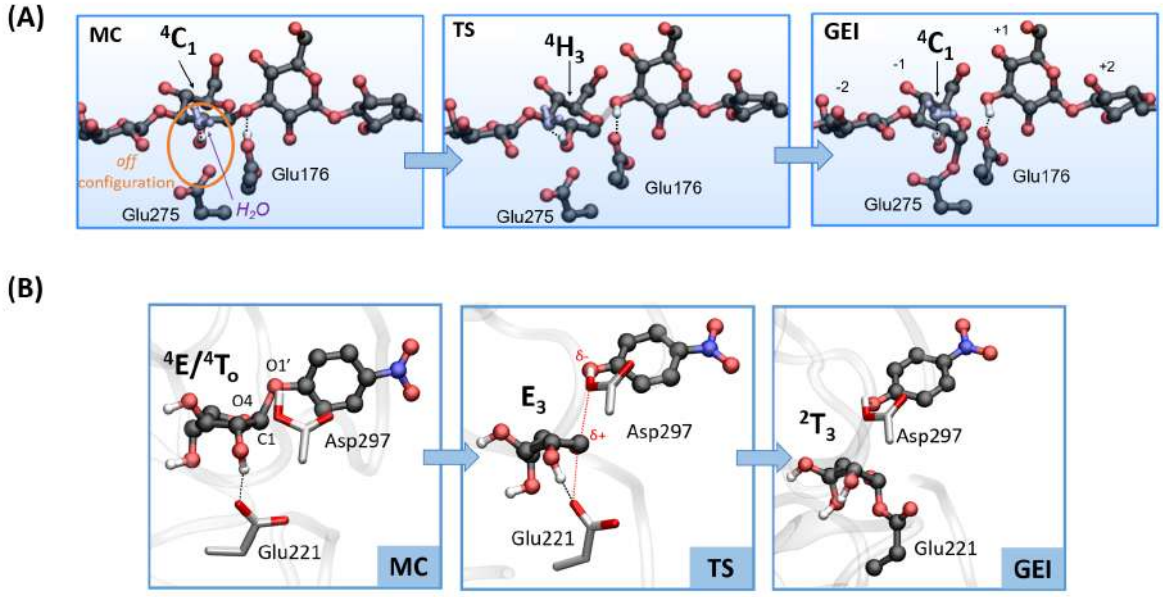


Figure 5.14: **(A)** Representative structures along the catalytic itinerary of ScGas2 β-D-glucanase from the *off* configuration [117]. **(B)** Representative structures along the catalytic itinerary of AkAbfB α-L-arabinofuranosidase.

in the -2 subsite leading to a total number of 32 quantum atoms (including capping hydrogens) and 91830 MM atoms; in the furanose simulation, the QM region only included the carbocation itself leading to 17 quantum atoms and 78810 MM atoms. The QM atoms were enclosed in an isolated supercell of size 15.3 x 15.3 x 15.3 $Å^3$ and 12.1 x 12.1 x 11.9 $Å^3$ and a fictitious electron mass of 700 and 500 atomic units (a.u.), for the pyranose and furanose cases, respectively; and a time step of 5 a.u. were used for the QM/MM MD simulations. The CP MD setup done in all the previous systems was used. The cartesian Cremer and Pople puckering coordinates were used as CVs for the pyranose case and the pseudorotational phase was used as CV for the furanose case. The height of the gaussian-like potential was set to 0.6 kcal·mol^{-1} and added every 250 (pyranose) or 500 (furanose) molecular dynamics steps and once the whole free energy spaces were explored, the height of the Gaussian terms were reduced to 0.2 kcal·mol^{-1} and the deposition time was increased to 1000 molecular dynamics steps to facilitate convergence of the FEL. The width of the collective variables was set to 0.05, 0.05, 0.015 rad for the pyranose case and 0.035 rad for the furanose case. The simulation was stopped after 1807 and 4998 deposited gaussians.

5.5 Supplemental Information

5.5.1 Superacid force field validation

To assess the quality of our in-house parametrized superacid force field we compared the properties calculated from the classical MD simulations with reference data obtained from *ab initio* MD simulations. First, the bond distances of the superacid molecules are monitored along the whole simulation, see Figure 5.15, where H-F bond distance in HF molecule has a value of 0.93 Å comparable to the values found in the work of Kim *et. al.* [190] using CPMD (0.93 Å) and found experimentally (0.92 Å) [194]. The F-H covalent bond distance for H_2F^+ has a higher value (0.987 Å), comparable to the value found by *ab initio* methods in Sophy *et. al.* work ($\sim$ 1.02 Å) [195]. The Sb-F bond distance has the average value obtained in the Kim *et. al.* work [190] (0.92 Å) as in classical MD is not possible to distinguish between the two Sb-F bonds, one coordinated to HF molecules by a hydrogen bond and the other that it is not (their corresponding values being 1.95 Å and 1.90 Å, respectively).

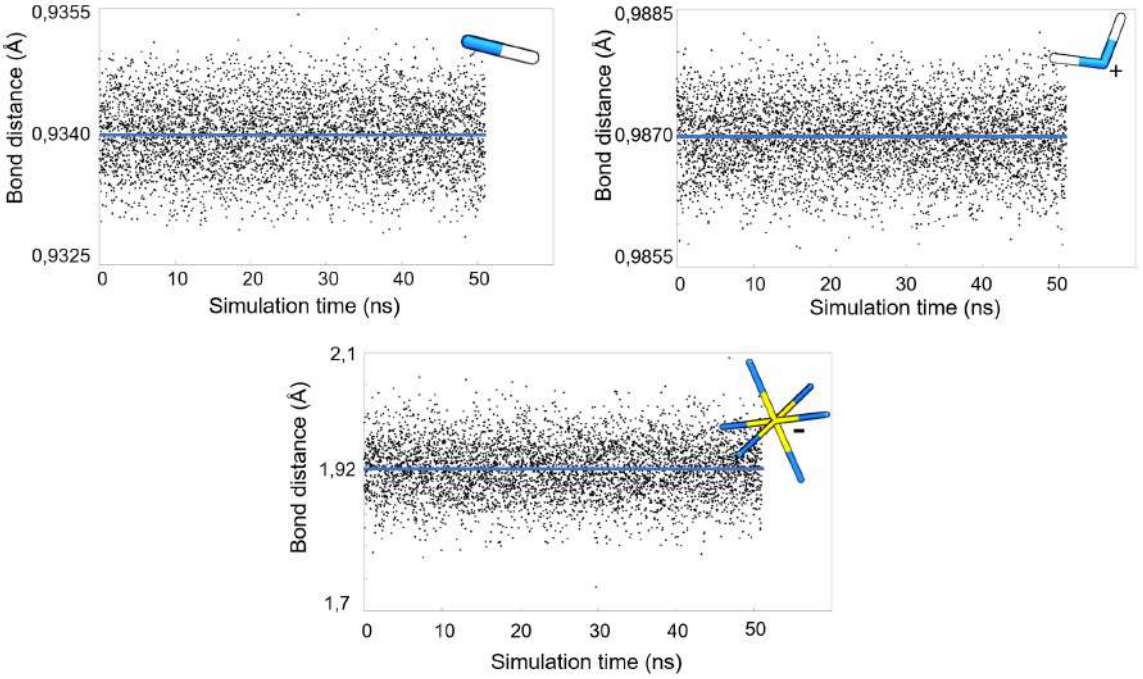


Figure 5.15: Bond lengths (in Å) of the parametrized superacid molecules; HF, H_2F^+ and SbF_6^- along the classical MD simulation, the average value is indicated by a blue line.

The classical MD energies, volume and density of the simulations are also monitored, see Figure 5.16. Equilibration is reached when the system volume has a value of $2{\cdot}10^5 Å^3$ and a density of 2.4 g·cm^{-3}, in-between the experimentally density values of the liquid SbF_5 (2.99 g·cm^{-3}) and HF (0.987 g·cm^{-3}) at ambient conditions [196].

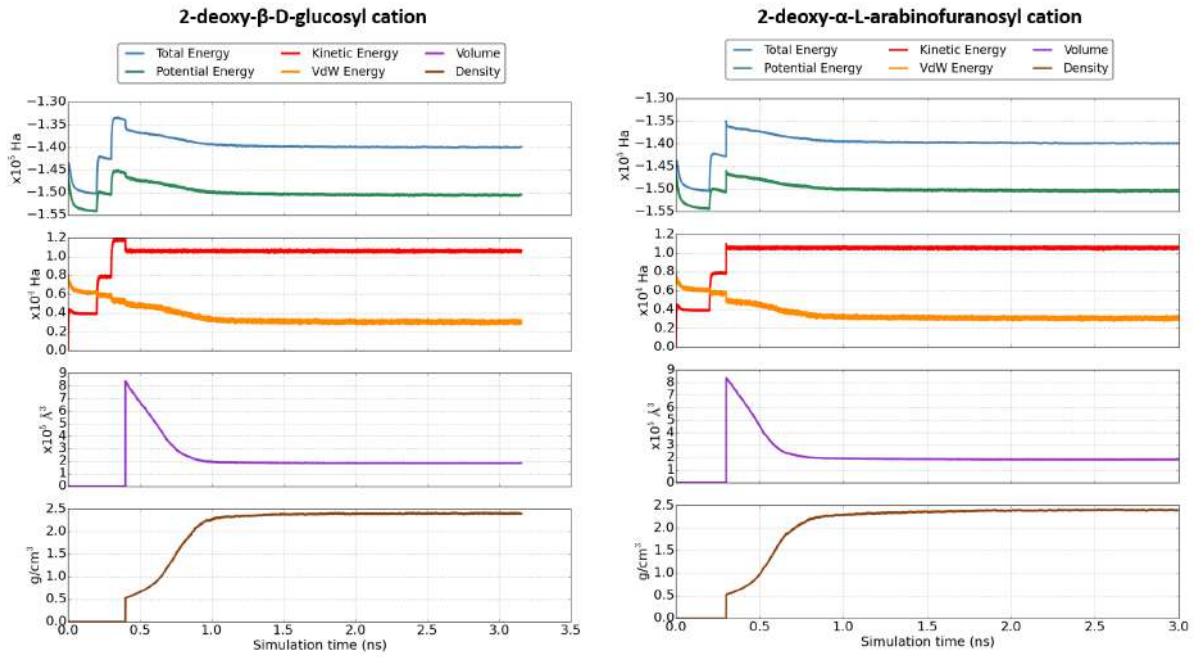


Figure 5.16: Total, potential, kinetic and Van der Waals energies and volume and density of the equilibration phase of the classical molecular dynamics simulations of the polyprotonated 2-deoxy-β-D-glucosyl and 2-deoxy-α-L-arabinofuranosyl cations in superacid media.

The structural properties of the solvent can be seen with the radial pair distribution function (Figure 5.17) where the radial distribution function of H (from HF molecules) atoms and F (from SbF_6^- molecules) is compared with g(r) of H (from HF molecules) and F from HF solvent molecules. The big peak at approximately 0.9 Å corresponds to the covalent bond distance between H and F in HF molecules. The peak at $\sim$ 1.6 Å corresponds to the hydrogen-bonded F atoms of the SbF_6^- anion coordinated to nearby HF molecules. It coincides with the second peak of g(H-F_{HF}) which it also corresponds to hydrogen bond interactions, in this case between F and H of different HF molecules. These results agree with the ones found in Kim *et. al.* [190] using *ab initio* MD with the only difference being the disappearance of a shoulder peak for g(H-F_{HF}) and a small peak for g(H-$F_{SbF_6^-}$), both at 1.2Å. These two unseen peaks corresponded to the interaction between an excess proton with HF or SbF_6^- molecules, therefore, being the manifestation of the existence of a contact ion pair. These contact ion pairs cannot be simulated by classical MD.

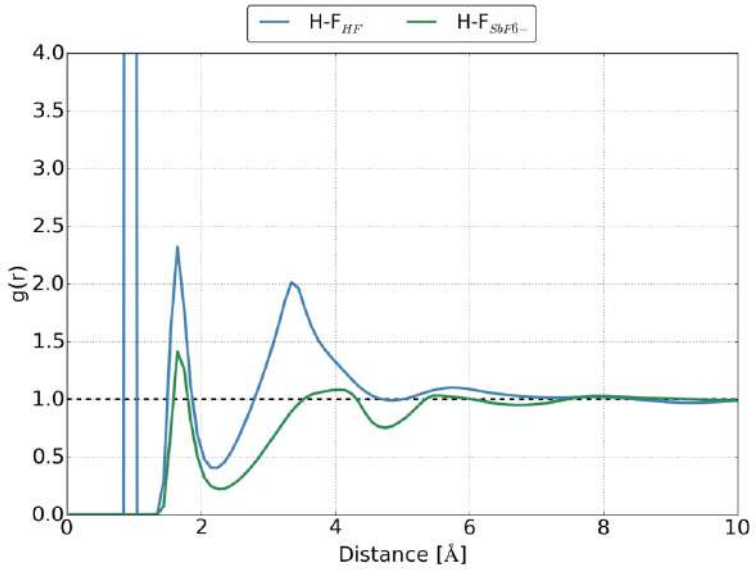


Figure 5.17: Radial pair distribution function, g(r), for H (HF) and F (HF) solvent molecules (black line) and for H (HF) and F (SbF_6^-) solvent molecules (blue line).

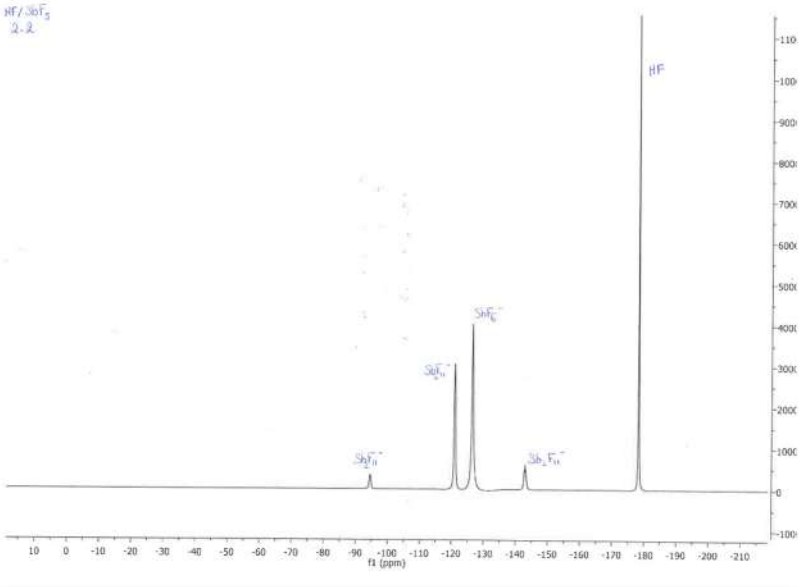


Figure 5.18: ^{19}F NMR of the superacid media used to characterize the glycosyl cation.

Chapter 6

Conclusions

By means of computational techniques, in particular, QM/MM MD metadynamics simulations (in all the chapters of the present work) and experimental techniques, such as protein production and X-ray crystallography (in Chapter 4) we investigated the conformational preferences and reactivity of carbohydrates in the active site of retaining GHs. We focused on two enzymes of biomedical and biotechnological interest: β-galactocerebrosidase (GALC), the defective enzyme in Krabbe disease, and the biomass-degrading enzyme α-L-arabinofuranosidase (AkAbfB). In the former, we uncovered the reaction mechanism of the enzyme alone and in complex with the lipid-transfer Saposin A protein (SapA). In the latter, we investigated the ability of potential inhibitors to reproduce the conformational properties of the natural substrate by comparing their conformational free energy landscapes. Finally, we investigated the conformational flexibility and electrostatic properties of five- and six-membered sugar ring glycosyl cations in the gas phase, superacid and enzymatic media.

The main conclusions that have been obtained in this Thesis are the following:

- The computed conformational free energy landscape of Gal-β-pNP in the active site of GALC in catalytic conditions indicates that two conformational minima are possible, the undistorted chair conformation ($^4C_1$) found in the non-catalytically competent MC structure; and the distorted $^1S_3$ conformation. The calculated first step of the hydrolysis reaction departing from both conformations demonstrates that two catalytic itineraries, $^1S_3 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$ and $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$, contribute to the hydrolysis of the Gal-β-pNP substrate by GALC. The cyclic $^4C_1 \rightarrow [^4H_3]^{\ddagger} \rightarrow {}^4C_1$ itinerary, which is unprecedented in GHs, is probably observed because of the lack of steric determinants that keep the leaving group in place. Being an exo-GH, the leaving group is pointing towards the solvent with minimal contacts to the enzyme surface.

- The conformational free energy landscape of the natural substrate of GALC, GalCer in the active site of GALC, indicates that only the undistorted conformation is stable. GalCer follows one of the two itineraries previously found for Gal-β-pNP; $^4C_1 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$, but the ceramide leaving group tail increases the reaction barrier by approximately 15 kcal·mol^{-1}, making the reaction unfeasible from a kinetic point of view.

- The conformational free energy landscape of GalCer in the active site of the GALC-SapA dimer displays two minima, $^4C_1$ and $^1S_3$, recovering the results previously obtained for the complex of GALC with Gal-β-pNP. However, the undistorted $^4C_1$ conformation this time has a clear preference and the reaction can only take place from this conformation. The reaction conformational catalytic itinerary is very similar to the one obtained for GalCer in complex with only GALC, i.e. the cyclic $^4C_1 \rightarrow [^4H_3]^\ddagger \rightarrow {}^4C_1$ itinerary, but now the reaction has a lower energy barrier, which is in excellent agreement with the experimental reaction rate. This is because of the stabilization of the TS by the SapA protein, which increases the number of interactions between the GalCer substrate and the GALC-SapA dimer. Therefore, the SapA protein not only acts as a transport agent but also helps stabilizing the TS of the reaction.

- The conformational free energy landscape of α-L-arabinofuranose shows that the most preactivated conformation of α-L-arabinofuranose in α-L- arabinofuranosidases are in the region ^2E-$^2T_3$-E_3-$^4T_3$-^4E, as these conformations feature an equatorial 2-OH, minimizing its steric interaction with the nucleophile residue along the chemical reaction. The conformational free energy landscape of three α-L-arabinofuranose-configured inhibitors can serve to predict their inhibition potency.

- The structure of the Glu221Gln AkAbfB mutant in complex with PNP-Araf was crystallized with high resolution (1.64 Å). The substrate ring adopts a ^4E conformation, as predicted by the computed conformational free energy landscape of the isolated α-L-arabinofuranose. The computed catalytic conformational itinerary of AkAbfB α-L-arabinofuranosidase turns out to be $^4E/^4T_O \rightarrow [E_3]^\ddagger \rightarrow {}^2T_3$, demonstrating that the crystallographic structures of AkAbfB in complex with PNP-Araf and α-L-arabinofuranose-configured aziridine and cyclicsulphate are good mimics of the Michaelis complex and glycosyl-enzyme intermediate structures, respectively.

- The superacid media do not alter the intrinsic structural properties of glycosyl cations. The accessible sugar ring conformations, the intramolecular

bond distances and dihedral angles are conserved among at least, the two glycosyl cations analysed, the 2-deoxy-β-D-glucosyl cation and the 2-deoxy-α-L-arabinofuranosyl cation. The cationic character is slightly enhanced in superacid compared to both the isolated cation and the cation in the enzyme active site due to the polarization effect of the high number of negative molecules surrounding the cation (SbF_6^- molecules).

- The superacid media mimics the multidentate interactions that stabilize the glycosyl cation in the enzyme active site.
- The density of SbF_6^- molecules is much higher in the β face of the glycosyl cation, either in pyranose- and furanose-based cations, which explains the stereoselectivity of nucleophillic reactions of these cations in superacid media that is observed experimentally.

Bibliography

[1] R. V. Stick and S. Williams, *Carbohydrates: the essential molecules of life*. Elsevier, 2010.

[2] G. W. Hart and R. J. Copeland, "Glycomics hits the big time," *Cell*, vol. 143, no. 5, pp. 672–676, 2010.

[3] J. Y. Zhou, D. M. Oswald, K. D. Oliva, L. S. Kreisman, and B. A. Cobb, "The glycoscience of immunity," *Trends in immunology*, vol. 39, no. 7, pp. 523–535, 2018.

[4] A. C. Walls, X. Xiong, Y.-J. Park, M. A. Tortorici, J. Snijder, J. Quispe, E. Cameroni, R. Gopal, M. Dai, A. Lanzavecchia, *et al.*, "Unexpected receptor functional mimicry elucidates activation of coronavirus fusion," *Cell*, vol. 176, no. 5, pp. 1026–1039, 2019.

[5] K. M. Koeller and C.-H. Wong, "Emerging themes in medicinal glycoscience," *Nature biotechnology*, vol. 18, no. 8, pp. 835–841, 2000.

[6] H. Ghazarian, B. Idoni, and S. B. Oppenheimer, "A glycobiology review: carbohydrates, lectins and implications in cancer therapeutics," *Acta histochemica*, vol. 113, no. 3, pp. 236–247, 2011.

[7] P. H. Seeberger and D. B. Werz, "Synthesis and medical applications of oligosaccharides," *Nature*, vol. 446, no. 7139, pp. 1046–1051, 2007.

[8] A. C. E. Graziano, G. Pannuzzo, R. Avola, and V. Cardile, "Chaperones as potential therapeutics for krabbe disease," *Journal of neuroscience research*, vol. 94, no. 11, pp. 1220–1230, 2016.

[9] S. J. Spratley and J. E. Deane, "New therapeutic approaches for krabbe disease: the potential of pharmacological chaperones," *Journal of neuroscience research*, vol. 94, no. 11, pp. 1203–1219, 2016.

[10] D. t. Cremer and J. Pople, "General definition of ring puckering coordinates," *Journal of the American Chemical Society*, vol. 97, no. 6, pp. 1354–1358, 1975.

[11] C. t. Altona and M. Sundaralingam, "Conformational analysis of the sugar ring in nucleosides and nucleotides. new description using the concept of pseudorotation," *Journal of the American Chemical Society*, vol. 94, no. 23, pp. 8205–8212, 1972.

[12] D. A. Lightner and J. E. Gurst, *Organic conformational analysis and stereochemistry from circular dichroism spectroscopy*, vol. 23. John Wiley & Sons, 2000.

[13] J. E. Kilpatrick, K. S. Pitzer, and R. Spitzer, "The thermodynamics and molecular structure of cyclopentane1," *Journal of the American Chemical Society*, vol. 69, no. 10, pp. 2483–2488, 1947.

[14] S. T. Rao, E. t. Westhof, and M. Sundaralingam, "Exact method for the calculation of pseudorotation parameters p, τm and their errors. a comparison of the altona–sundaralingam and cremer–pople treatment of puckering of five-membered rings," *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, vol. 37, no. 3, pp. 421–425, 1981.

[15] A. D. Hill and P. J. Reilly, "Puckering coordinates of monocyclic rings by triangular decomposition," *Journal of chemical information and modeling*, vol. 47, no. 3, pp. 1031–1035, 2007.

[16] T. Sato, "Another method for specifying furanose ring puckering," *Nucleic acids research*, vol. 11, no. 14, pp. 4933–4938, 1983.

[17] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "Plumed 2: New feathers for an old bird," *Computer Physics Communications*, vol. 185, no. 2, pp. 604–613, 2014.

[18] M. Huang, T. J. Giese, T.-S. Lee, and D. M. York, "Improvement of dna and rna sugar pucker profiles from semiempirical quantum methods," *Journal of chemical theory and computation*, vol. 10, no. 4, pp. 1538–1545, 2014.

[19] X. Biarnés, A. Ardevol, A. Planas, C. Rovira, A. Laio, and M. Parrinello, "The conformational free energy landscape of β-d-glucopyranose. implications for substrate preactivation in β-glucoside hydrolases," *Journal of the American Chemical Society*, vol. 129, no. 35, pp. 10686–10693, 2007.

[20] H. B. Mayes, L. J. Broadbelt, and G. T. Beckham, "How sugars pucker: electronic structure calculations map the kinetic landscape of five biologically paramount monosaccharides and their implications for enzymatic catalysis," *Journal of the American Chemical Society*, vol. 136, no. 3, pp. 1008–1022, 2014.

[21] A. Ardévol, X. Biarnés, A. Planas, and C. Rovira, "The conformational free-energy landscape of β-d-mannopyranose: evidence for a 1s5 $\rightarrow$ b2,5 $\rightarrow$ os2 catalytic itinerary in β-mannosidases," *Journal of the American Chemical Society*, vol. 132, no. 45, pp. 16058–16065, 2010.

[22] J. Iglesias-Fernández, L. Raich, A. Ardèvol, and C. Rovira, "The complete conformational free energy landscape of β-xylose reveals a two-fold catalytic itinerary for β-xylanases," *Chemical science*, vol. 6, no. 2, pp. 1167–1177, 2015.

[23] V. Lombard, H. Golaconda Ramulu, E. Drula, P. M. Coutinho, and B. Henrissat, "The carbohydrate-active enzymes database (cazy) in 2013," *Nucleic acids research*, vol. 42, no. D1, pp. D490–D495, 2013.

[24] B. Henrissat, "A classification of glycosyl hydrolases based on amino acid sequence similarities," *Biochemical journal*, vol. 280, no. 2, pp. 309–316, 1991.

[25] T. N. Bhat, P. Bourne, Z. Feng, G. Gilliland, S. Jain, V. Ravichandran, B. Schneider, K. Schneider, N. Thanki, H. Weissig, *et al.*, "The pdb data uniformity project," *Nucleic Acids Research*, vol. 29, no. 1, pp. 214–218, 2001.

[26] D. Koshland Jr, "Stereochemistry and the mechanism of enzymatic reactions," *Biological reviews*, vol. 28, no. 4, pp. 416–436, 1953.

[27] D. L. Zechel and S. G. Withers, "Glycosidase mechanisms: anatomy of a finely tuned catalyst," *Accounts of chemical research*, vol. 33, no. 1, pp. 11–18, 2000.

[28] C. Denekamp and Y. Sandlers, "Anomeric distinction and oxonium ion formation in acetylated glycosides," *Journal of mass spectrometry*, vol. 40, no. 6, pp. 765–771, 2005.

[29] M. Huang, P. Retailleau, L. Bohé, and D. Crich, "Cation clock permits distinction between the mechanisms of α-and β-o-and β-c-glycosylation in the mannopyranose series: evidence for the existence of a mannopyranosyl oxocarbenium ion," *Journal of the American Chemical Society*, vol. 134, no. 36, pp. 14746–14749, 2012.

[30] A. Martin, A. Arda, J. Désiré, A. Martin-Mingot, N. Probst, P. Sinaÿ, J. Jiménez-Barbero, S. Thibaudeau, and Y. Blériot, "Catching elusive glycosyl cations in a condensed phase with hf/sbf 5 superacid," *Nature chemistry*, vol. 8, no. 2, p. 186, 2016.

[31] N. F. Hall and J. B. Conant, "A study of superacid solutions. i. the use of the chloranil electrode in glacial acetic acid and the strength of certain weak bases," *Journal of the American Chemical Society*, vol. 49, no. 12, pp. 3047–3061, 1927.

[32] D. Himmel, S. K. Goll, I. Leito, and I. Krossing, "A unified ph scale for all phases," *Angewandte Chemie International Edition*, vol. 49, no. 38, pp. 6885–6888, 2010.

[33] G. A. Olah, "Crossing conventional boundaries in half a century of research," *The Journal of organic chemistry*, vol. 70, no. 7, pp. 2413–2429, 2005.

[34] G. Speciale, A. J. Thompson, G. J. Davies, and S. J. Williams, "Dissecting conformational contributions to glycosidase catalysis and inhibition," *Current opinion in structural biology*, vol. 28, pp. 1–13, 2014.

[35] G. J. Davies, A. Planas, and C. Rovira, "Conformational analyses of the reaction coordinate of glycosidases," *Accounts of chemical research*, vol. 45, no. 2, pp. 308–316, 2012.

[36] X. Biarnés, J. Nieto, A. Planas, and C. Rovira, "Substrate distortion in the michaelis complex of bacillus 1, 3–1, 4-β-glucanase insight from first principles molecular dynamics simulations," *Journal of Biological Chemistry*, vol. 281, no. 3, pp. 1432–1441, 2006.

[37] A. Vasella, G. J. Davies, and M. Böhm, "Glycosidase mechanisms," *Current opinion in chemical biology*, vol. 6, no. 5, pp. 619–629, 2002.

[38] A. J. Kirby, "Stereoelectronic effects on acetal hydrolysis," *Accounts of chemical research*, vol. 17, no. 9, pp. 305–311, 1984.

[39] P. Deslongchamps, *Stereoelectronic effects in organic chemistry*, vol. 1. Pergamon, 1983.

[40] F. Rice and E. Teller, "The role of free radicals in elementary organic reactions," *The Journal of Chemical Physics*, vol. 6, no. 8, pp. 489–496, 1938.

[41] M. L. Sinnott, "The principle of least nuclear motion and the theory of stereoelectronic control," in *Advances in physical organic chemistry*, vol. 24, pp. 113–204, Elsevier, 1988.

[42] M. E. Caines, S. M. Hancock, C. A. Tarling, T. M. Wrodnigg, R. V. Stick, A. E. Stütz, A. Vasella, S. G. Withers, and N. C. Strynadka, "The structural basis of glycosidase inhibition by five-membered iminocyclitols: The clan a glycoside hydrolase endoglycoceramidase as a model system," *Angewandte Chemie International Edition*, vol. 46, no. 24, pp. 4474–4476, 2007.

[43] J. C. Kendrew, G. Bodo, H. M. Dintzis, R. Parrish, H. Wyckoff, and D. C. Phillips, "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis," *Nature*, vol. 181, no. 4610, pp. 662–666, 1958.

[44] F. Hayes, *The function and organization of plasmids*. Springer, 2003.

[45] B. R. Glick and C. L. Patten, *Molecular biotechnology: principles and applications of recombinant DNA*, vol. 34. John Wiley & Sons, 2017.

[46] R. J. Roberts and K. Murray, "Restriction endonuclease," *CRC critical reviews in biochemistry*, vol. 4, no. 2, pp. 123–164, 1976.

[47] C. Johnston, B. Martin, G. Fichant, P. Polard, and J.-P. Claverys, "Bacterial transformation: distribution, shared mechanisms and divergent control," *Nature Reviews Microbiology*, vol. 12, no. 3, pp. 181–196, 2014.

[48] J. M. Sinclair, *Collins English dictionary and thesaurus*. Collins, 1993.

[49] P. Kumar, *Fundamentals and Techniques of Biophysics and Molecular Biology*. Pathfinder Publication unit of PAPL, 2016.

[50] A. Mehta, "Principle of reversed-phase chromatography hplc/uplc," 2012.

[51] U. K. Laemmli, "Cleavage of structural proteins during the assembly of the head of bacteriophage t4," *Nature*, vol. 227, no. 5259, pp. 680–685, 1970.

[52] InVitrogen, "B, and c: Pichia expression vectors for selection on zeocin™ and purification of secreted, recombinant proteins, man0000035," Invitrogen, Corporate Headquarters, 2010.

[53] M. Chruszcz, A. Wlodawer, and W. Minor, "Determination of protein structures—a series of fortunate events," *Biophysical Journal*, vol. 95, no. 1, pp. 1–9, 2008.

[54] J. Newman, D. Egan, T. S. Walter, R. Meged, I. Berry, M. Ben Jelloul, J. L. Sussman, D. I. Stuart, and A. Perrakis, "Towards rationalization of crystallization screening for small-to medium-sized academic laboratories: the pact/jcsg+ strategy," *Acta Crystallographica Section D: Biological Crystallography*, vol. 61, no. 10, pp. 1426–1431, 2005.

[55] C. Carter and Y. Yin, "Quantitative analysis in the characterization and optimization of protein crystal growth," *Acta Crystallographica Section D: Biological Crystallography*, vol. 50, no. 4, pp. 572–590, 1994.

[56] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures," *The FEBS journal*, vol. 275, no. 1, pp. 1–21, 2008.

[57] G. Scapin, "Molecular replacement then and now," *Acta Crystallographica Section D: Biological Crystallography*, vol. 69, no. 11, pp. 2266–2275, 2013.

[58] A. F. Wells, *Structural inorganic chemistry*. Oxford university press, 2012.

[59] K. Cowtan, "Phase problem in x-ray crystallography, and its solution," *e LS*, 2001.

[60] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski, "Protein crystallography for aspiring crystallographers or how to avoid pitfalls and traps in macromolecular structure determination," *The FEBS journal*, vol. 280, no. 22, pp. 5705–5736, 2013.

[61] J. Agirre, G. Davies, K. Wilson, and K. Cowtan, "Carbohydrate anomalies in the pdb," *Nature chemical biology*, vol. 11, no. 5, p. 303, 2015.

[62] J. Agirre, J. Iglesias-Fernández, C. Rovira, G. J. Davies, K. S. Wilson, and K. D. Cowtan, "Privateer: software for the conformational validation of carbohydrate structures," *Nature structural & molecular biology*, vol. 22, no. 11, p. 833, 2015.

[63] O. M. Becker, A. D. MacKerell Jr, B. Roux, and M. Watanabe, *Computational biochemistry and biophysics*. Crc Press, 2001.

[64] D. A. Pearlman, D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham III, S. DeBolt, D. Ferguson, G. Seibel, and P. Kollman, "Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural

and energetic properties of molecules," *Computer Physics Communications*, vol. 91, no. 1-3, pp. 1–41, 1995.

[65] D. Marx and J. Hutter, *Ab initio molecular dynamics: basic theory and advanced methods*. Cambridge University Press, 2009.

[66] R. Car and M. Parrinello, "Unified approach for molecular dynamics and density-functional theory," *Physical review letters*, vol. 55, no. 22, p. 2471, 1985.

[67] R. Car, M. Parrinello, and M. Payne, "Error cancellation in the molecular dynamics method for total energy calculations," *Journal of Physics: Condensed Matter*, vol. 3, no. 47, p. 9539, 1991.

[68] C. Rovira, *Study of ligand-protein interactions by means of density functional theory and first-principles molecular dynamics*. Springer, 2005.

[69] F. Jensen, *Introduction to computational chemistry*. John wiley & sons, 2017.

[70] P. Hohenberg and W. Kohn, "Density functional theory (dft)," *Phys. Rev*, vol. 136, p. B864, 1964.

[71] W. Kohn, "Time-dependent kohn–sham density-functional theory," *Physical Review A*, vol. 1133, pp. 140–148, 1965.

[72] J. P. Perdew and K. Schmidt, "Jacob's ladder of density functional approximations for the exchange-correlation energy," in *AIP Conference Proceedings*, vol. 577, pp. 1–20, American Institute of Physics, 2001.

[73] H. Hellmann, "A new approximation method in the problem of many electrons," *The Journal of Chemical Physics*, vol. 3, no. 1, pp. 61–61, 1935.

[74] M. Karplus, M. Levitt, and A. Warshel, "The nobel prize in chemistry 2013," *Nobel Media AB 2014*, 2013.

[75] H. Lin and D. G. Truhlar, "Qm/mm: what have we learned, where are we, and where do we go from here?," *Theoretical Chemistry Accounts*, vol. 117, no. 2, p. 185, 2007.

[76] I. CPMD, "Corp 1990–2006," *MPI für Festkörperforschung Stuttgart*, vol. 2001, 1997.

[77] A. Laio, J. VandeVondele, and U. Rothlisberger, "A hamiltonian electrostatic coupling scheme for hybrid car–parrinello molecular dynamics simulations," *The Journal of chemical physics*, vol. 116, no. 16, pp. 6941–6947, 2002.

[78] L. Raich, A. Nin-Hill, A. Ardèvol, and C. Rovira, "Enzymatic cleavage of glycosidic bonds: strategies on how to set up and control a qm/mm metadynamics simulation," in *Methods in enzymology*, vol. 577, pp. 159–183, Elsevier, 2016.

[79] O. Valsson, P. Tiwary, and M. Parrinello, "Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint," *Annual review of physical chemistry*, vol. 67, pp. 159–184, 2016.

[80] B. Ensing, M. De Vivo, Z. Liu, P. Moore, and M. L. Klein, "Metadynamics as a tool for exploring free energy landscapes of chemical reactions," *Accounts of chemical research*, vol. 39, no. 2, pp. 73–81, 2006.

[81] A. Laio and M. Parrinello, "Escaping free-energy minima," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, pp. 12562–12566, 2002.

[82] B. Ensing, A. Laio, M. Parrinello, and M. L. Klein, "A recipe for the computation of the free energy barrier and the lowest free energy path of concerted reactions," *The journal of physical chemistry B*, vol. 109, no. 14, pp. 6676–6687, 2005.

[83] J. E. Deane, S. C. Graham, N. N. Kim, P. E. Stein, R. McNair, M. B. Cachón-González, T. M. Cox, and R. J. Read, "Insights into krabbe disease from structures of galactocerebrosidase," *Proceedings of the National Academy of Sciences*, vol. 108, no. 37, pp. 15169–15173, 2011.

[84] C. H. Hill, S. C. Graham, R. J. Read, and J. E. Deane, "Structural snapshots illustrate the catalytic cycle of β-galactocerebrosidase, the defective enzyme in krabbe disease," *Proceedings of the National Academy of Sciences*, vol. 110, no. 51, pp. 20479–20484, 2013.

[85] E. F. Neufeld, "Lysosomal storage diseases," *Annual review of biochemistry*, vol. 60, no. 1, pp. 257–280, 1991.

[86] A. C. E. Graziano and V. Cardile, "History, genetic, and recent advances on krabbe disease," *Gene*, vol. 555, no. 1, pp. 2–13, 2015.

[87] M. L. Beltran-Quintero, N. A. Bascou, M. D. Poe, D. A. Wenger, C. A. Saavedra-Matiz, M. J. Nichols, and M. L. Escolar, "Early progression of krabbe disease in patients with symptom onset between 0 and 5 months," *Orphanet journal of rare diseases*, vol. 14, no. 1, p. 46, 2019.

[88] C. H. Hill, A. H. Viuff, S. J. Spratley, S. Salamone, S. H. Christensen, R. J. Read, N. W. Moriarty, H. H. Jensen, and J. E. Deane, "Azasugar inhibitors as pharmacological chaperones for krabbe disease," *Chemical science*, vol. 6, no. 5, pp. 3075–3086, 2015.

[89] C. H. Hill, G. M. Cook, S. J. Spratley, S. Fawke, S. C. Graham, and J. E. Deane, "The mechanism of glycosphingolipid degradation revealed by a galc-sapa complex structure," *Nature communications*, vol. 9, no. 1, pp. 1–9, 2018.

[90] S. Locatelli-Hoops, N. Remmel, R. Klingenstein, B. Breiden, M. Rossocha, M. Schoeniger, C. Koenigs, W. Saenger, and K. Sandhoff, "Saposin a mobilizes lipids from low cholesterol and high bis (monoacylglycerol) phosphate-containing membranes patient variant saposin a lacks lipid extraction capacity," *Journal of Biological Chemistry*, vol. 281, no. 43, pp. 32451–32460, 2006.

[91] N. Remmel, S. Locatelli-Hoops, B. Breiden, G. Schwarzmann, and K. Sandhoff, "Saposin b mobilizes lipids from cholesterol-poor and bis (monoacylglycero) phosphate-rich membranes at acidic ph: Unglycosylated patient variant saposin b lacks lipid-extraction capacity," *The FEBS journal*, vol. 274, no. 13, pp. 3405–3420, 2007.

[92] J.-R. Alattia, J. E. Shaw, C. M. Yip, and G. G. Privé, "Direct visualization of saposin remodelling of lipid bilayers," *Journal of molecular biology*, vol. 362, no. 5, pp. 943–953, 2006.

[93] J.-R. Alattia, J. E. Shaw, C. M. Yip, and G. G. Privé, "Molecular imaging of membrane interfaces reveals mode of β-glucosidase activation by saposin c," *Proceedings of the National Academy of Sciences*, vol. 104, no. 44, pp. 17394–17399, 2007.

[94] K. Popovic, J. Holyoake, R. Pomès, and G. G. Privé, "Structure of saposin a lipoprotein discs," *Proceedings of the National Academy of Sciences*, vol. 109, no. 8, pp. 2908–2912, 2012.

[95] K. Popovic and G. G. Privé, "Structures of the human ceramide activator protein saposin d," *Acta Crystallographica Section D: Biological Crystallography*, vol. 64, no. 5, pp. 589–594, 2008.

[96] V. E. Ahn, P. Leyko, J.-R. Alattia, L. Chen, and G. G. Privé, "Crystal structures of saposins a and c," *Protein science*, vol. 15, no. 8, pp. 1849–1857, 2006.

[97] V. E. Ahn, K. F. Faull, J. P. Whitelegge, A. L. Fluharty, and G. G. Privé, "Crystal structure of saposin b reveals a dimeric shell for lipid binding," *Proceedings of the National Academy of Sciences*, vol. 100, no. 1, pp. 38–43, 2003.

[98] E. de Alba, S. Weiler, and N. Tjandra, "Solution structure of human saposin c: ph-dependent interaction with phospholipid vesicles," *Biochemistry*, vol. 42, no. 50, pp. 14729–14740, 2003.

[99] C. H. Hill, R. J. Read, and J. E. Deane, "Structure of human saposin a at lysosomal ph," *Acta Crystallographica Section F: Structural Biology Communications*, vol. 71, no. 7, pp. 895–900, 2015.

[100] J. Li, M. R. Richards, D. Bagal, I. D. Campuzano, E. N. Kitova, Z. J. Xiong, G. G. Privé, and J. S. Klassen, "Characterizing the size and composition of saposin a lipoprotein picodiscs," *Analytical chemistry*, vol. 88, no. 19, pp. 9524–9531, 2016.

[101] J. Matsuda, M. T. Vanier, Y. Saito, J. Tohyama, K. Suzuki, and K. Suzuki, "A mutation in the saposin a domain of the sphingolipid activator protein (prosaposin) gene results in a late-onset, chronic form of globoid cell leukodystrophy in the mouse," *Human molecular genetics*, vol. 10, no. 11, pp. 1191–1199, 2001.

[102] S. J. Spratley, C. H. Hill, A. H. Viuff, J. R. Edgar, K. Skjødt, and J. E. Deane, "Molecular mechanisms of disease pathogenesis differ in krabbe disease variants," *Traffic*, vol. 17, no. 8, pp. 908–922, 2016.

[103] B. Tappino, R. Biancheri, M. Mort, S. Regis, F. Corsolini, A. Rossi, M. Stroppiano, S. Lualdi, A. Fiumara, B. Bembi, *et al.*, "Identification and characterization of 15 novel galc gene mutations causing krabbe disease," *Human mutation*, vol. 31, no. 12, pp. E1894–E1914, 2010.

[104] R. De Gasperi, M. G. Sosa, E. L. Sartorato, S. Battistini, H. MacFarlane, J. F. Gusella, W. Krivit, and E. H. Kolodny, "Molecular heterogeneity of late-onset forms of globoid-cell leukodystrophy.," *American journal of human genetics*, vol. 59, no. 6, p. 1233, 1996.

[105] E. H. Kolodny, S. Raghavan, and W. Krivit, "Late-onset krabbe disease (globoid cell leukodystrophy): clinical and biochemical features of 15 cases," *Developmental neuroscience*, vol. 13, no. 4-5, pp. 232–239, 1991.

[106] J. Espinosa, E. Montero, A. Vian, J. Garcia, H. Dietrich, R. Schmidt, M. Martin-Lomas, A. Imberty, F. Canada, and J. Jimenez-Barbero, "Es-

cherichia coli β-galactosidase recognizes a high-energy conformation of c-lactose, a nonhydrolizable substrate analogue. nmr and modeling studies of the molecular complex," *Journal of the American Chemical Society*, vol. 120, no. 6, pp. 1309–1318, 1998.

[107] A. García-Herrero, E. Montero, J. L. Muñoz, J. F. Espinosa, A. Vián, J. L. García, J. L. Asensio, F. J. Cañada, and J. Jiménez-Barbero, "Conformational selection of glycomimetics at enzyme catalytic sites: Experimental demonstration of the binding of distinct high-energy distorted conformations of c-, s-, and o-glycosides by e. c oli β-galactosidases," *Journal of the American Chemical Society*, vol. 124, no. 17, pp. 4804–4810, 2002.

[108] N. F. Brás, P. A. Fernandes, and M. J. Ramos, "Qm/mm studies on the β-galactosidase catalytic mechanism: Hydrolysis and transglycosylation reactions," *Journal of chemical theory and computation*, vol. 6, no. 2, pp. 421–433, 2010.

[109] C. B. Barnett and K. J. Naidoo, "Stereoelectronic and solvation effects determine hydroxymethyl conformational preferences in monosaccharides," *The Journal of Physical Chemistry B*, vol. 112, no. 48, pp. 15450–15459, 2008.

[110] M. Rahal-Sekkal, N. Sekkal, D. C. Kleb, and P. Bleckmann, "Structures and energies of d-galactose and galabiose conformers as calculated by ab initio and semiempirical methods," *Journal of computational chemistry*, vol. 24, no. 7, pp. 806–818, 2003.

[111] V. L. Schramm and W. Shi, "Atomic motion in enzymatic reaction coordinates," *Current opinion in structural biology*, vol. 11, no. 6, pp. 657–665, 2001.

[112] J. L. Asensio, A. Ardá, F. J. Cañada, and J. Jiménez-Barbero, "Carbohydrate–aromatic interactions," *Accounts of chemical research*, vol. 46, no. 4, pp. 946–954, 2013.

[113] L. Montalvillo-Jimenez, A. G. Santana, F. Corzana, G. Jiménez-Osés, J. Jiménez-Barbero, A. M. Gomez, and J. L. Asensio, "Impact of aromatic stacking on glycoside reactivity: Balancing ch/π and cation/π interactions for the stabilization of glycosyl-oxocarbenium ions," *Journal of the American Chemical Society*, vol. 141, no. 34, pp. 13372–13384, 2019.

[114] X. Biarnés, A. Ardevol, J. Iglesias-Fernández, A. Planas, and C. Rovira, "Catalytic itinerary in 1, 3-1, 4-β-glucanase unraveled by qm/mm metadynamics.

charge is not yet fully developed at the oxocarbenium ion-like transition state," *Journal of the American Chemical Society*, vol. 133, no. 50, pp. 20301–20309, 2011.

[115] R. Kumar, B. Henrissat, and P. M. Coutinho, "Intrinsic dynamic behavior of enzyme: substrate complexes govern the catalytic action of β-galactosidases across clan gh-a," *Scientific reports*, vol. 9, no. 1, pp. 1–14, 2019.

[116] I. Marcos-Alcalde, E. López-Viñas, and P. Gómez-Puertas, "Mepsand: minimum energy path surface analysis over n-dimensional surfaces," *Bioinformatics*, vol. 36, no. 3, pp. 956–958, 2020.

[117] L. Raich, V. Borodkin, W. Fang, J. Castro-López, D. M. van Aalten, R. Hurtado-Guerrero, and C. Rovira, "A trapped covalent intermediate of a glycoside hydrolase on the pathway to transglycosylation. insights from experiments and quantum mechanics/molecular mechanics simulations," *Journal of the American Chemical Society*, vol. 138, no. 10, pp. 3325–3332, 2016.

[118] G. Wiederschain, R. Srinivasa, and E. Kolodny, "Characterization of 6-hexadecanoylamino-4-methylumbelliferyl-β-d-galactopyranoside as fluorogenic substrate of galactocerebrosidase for the diagnosis of krabbe disease," *Clinica chimica acta*, vol. 205, no. 1-2, pp. 87–96, 1992.

[119] K. Harzer, B. C. Paton, H. Christomanou, M. Chatelut, T. Levade, M. Hiraiwa, and J. S. O'Brien, "Saposins (sap) a and c activate the degradation of galactosylceramide in living cells," *FEBS letters*, vol. 417, no. 3, pp. 270–274, 1997.

[120] P. G. Bolhuis, C. Dellago, and D. Chandler, "Reaction coordinates of biomolecular isomerization," *Proceedings of the National Academy of Sciences*, vol. 97, no. 11, pp. 5877–5882, 2000.

[121] W. Humphrey, A. Dalke, K. Schulten, *et al.*, "Vmd: visual molecular dynamics," *Journal of molecular graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[122] D. R. Roe and T. E. Cheatham III, "Ptraj and cpptraj: software for processing and analysis of molecular dynamics trajectory data," *Journal of chemical theory and computation*, vol. 9, no. 7, pp. 3084–3095, 2013.

[123] J. D. Durrant and J. A. McCammon, "Binana: a novel algorithm for ligand-binding characterization," *Journal of Molecular Graphics and Modelling*, vol. 29, no. 6, pp. 888–893, 2011.

[124] T. Williams, C. Kelley, H.-B. Bröker, J. Campbell, R. Cunningham, D. Denholm, G. Elber, R. Fearick, C. Grammes, L. Hart, *et al.*, "gnuplot 4.6," *An Interactive Plotting Program*, 2012.

[125] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[126] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "Ucsf chimera—a visualization system for exploratory research and analysis," *Journal of computational chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[127] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences*, vol. 85, no. 8, pp. 2444–2448, 1988.

[128] M. Rostkowski, M. H. Olsson, C. R. Søndergaard, and J. H. Jensen, "Graphical analysis of ph-dependent properties of proteins predicted using propka," *BMC structural biology*, vol. 11, no. 1, p. 6, 2011.

[129] S. Jo, T. Kim, V. G. Iyer, and W. Im, "Charmm-gui: a web-based graphical user interface for charmm," *Journal of computational chemistry*, vol. 29, no. 11, pp. 1859–1865, 2008.

[130] J. Lee, D. S. Patel, J. Ståhle, S.-J. Park, N. R. Kern, S. Kim, J. Lee, X. Cheng, M. A. Valvano, O. Holst, *et al.*, "Charmm-gui membrane builder for complex biological membrane simulations with glycolipids and lipoglycans," *Journal of chemical theory and computation*, vol. 15, no. 1, pp. 775–786, 2018.

[131] D. A. Case, T. A. Darden, T. r. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, *et al.*, "Amber 11," tech. rep., University of California, 2010.

[132] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters," *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 3, pp. 712–725, 2006.

[133] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley, and R. J. Woods, "Glycam06: a generalizable biomolecular force field. carbohydrates," *Journal of computational chemistry*, vol. 29, no. 4, pp. 622–655, 2008.

[134] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of chemical physics*, vol. 79, no. 2, pp. 926–935, 1983.

[135] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," *Journal of molecular graphics and modelling*, vol. 25, no. 2, pp. 247–260, 2006.

[136] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," *Journal of computational chemistry*, vol. 25, no. 9, pp. 1157–1174, 2004.

[137] I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall III, J. Snoeyink, J. S. Richardson, *et al.*, "Molprobity: all-atom contacts and structure validation for proteins and nucleic acids," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W375–W383, 2007.

[138] D. Case, V. Babin, J. Berryman, R. Betz, Q. Cai, D. Cerutti, T. Cheatham III, T. Darden, R. Duke, H. Gohlke, *et al.*, "Amber 14, 2014," *University of California, San Francisco*, 2014.

[139] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," *Physical review letters*, vol. 77, no. 18, p. 3865, 1996.

[140] A. Ardèvol and C. Rovira, "Reaction mechanisms in carbohydrate-active enzymes: glycoside hydrolases and glycosyltransferases. insights from ab initio quantum mechanics/molecular mechanics dynamic simulations," *Journal of the American Chemical Society*, vol. 137, no. 24, pp. 7528–7547, 2015.

[141] M. Marianski, A. Supady, T. Ingram, M. Schneider, and C. Baldauf, "Assessing the accuracy of across-the-scale methods for predicting carbohydrate conformational energies for the examples of glucose and α-maltose," *Journal of chemical theory and computation*, vol. 12, no. 12, pp. 6157–6168, 2016.

[142] N. Troullier and J. L. Martins, "Efficient pseudopotentials for plane-wave calculations," *Physical review B*, vol. 43, no. 3, p. 1993, 1991.

[143] P. Tiwary and M. Parrinello, "A time-independent free energy estimator for metadynamics," *The Journal of Physical Chemistry B*, vol. 119, no. 3, pp. 736–742, 2015.

[144] M. L. Sinnott and I. J. Souchard, "The mechanism of action of β-galactosidase. effect of aglycone nature and α-deuterium substitution on the hydrolysis of aryl galactosides," *Biochemical Journal*, vol. 133, no. 1, pp. 89–98, 1973.

[145] N. N. Nair, E. Schreiner, and D. Marx, "Peptide synthesis in aqueous environments: The role of extreme conditions on amino acid activation," *Journal of the American Chemical Society*, vol. 130, no. 43, pp. 14148–14160, 2008.

[146] L. Bezalel, Y. Shoham, and E. Rosenberg, "Characterization and delignification activity of a thermostable α-l-arabinofuranosidase from bacillus stearothermophilus," *Applied microbiology and biotechnology*, vol. 40, no. 1, pp. 57–62, 1993.

[147] A. Cartmell, J. Muñoz-Muñoz, J. A. Briggs, D. A. Ndeh, E. C. Lowe, A. Baslé, N. Terrapon, K. Stott, T. Heunis, J. Gray, *et al.*, "A surface endogalactanase in bacteroides thetaiotaomicron confers keystone status for arabinogalactan degradation," *Nature microbiology*, vol. 3, no. 11, pp. 1314–1326, 2018.

[148] S. J. Williams, O. Hekmat, and S. G. Withers, "Synthesis and testing of mechanism-based protein-profiling probes for retaining endo-glycosidases," *ChemBioChem*, vol. 7, no. 1, pp. 116–124, 2006.

[149] O. Hekmat, Y.-W. Kim, S. J. Williams, S. He, and S. G. Withers, "Active-site peptide "fingerprinting" of glycosidases in complex mixtures by mass spectrometry discovery of a novel retaining β-1, 4-glycanase in cellulomonas fimi," *Journal of Biological Chemistry*, vol. 280, no. 42, pp. 35126–35135, 2005.

[150] O. Hekmat, S. He, R. A. J. Warren, and S. G. Withers, "A mechanism-based icat strategy for comparing relative expression and activity levels of glycosidases in biological systems," *Journal of proteome research*, vol. 7, no. 8, pp. 3282–3292, 2008.

[151] L. M. Chauvigné-Hines, L. N. Anderson, H. M. Weaver, J. N. Brown, P. K. Koech, C. D. Nicora, B. A. Hofstad, R. D. Smith, M. J. Wilkins, S. J. Callister, *et al.*, "Suite of activity-based probes for cellulose-degrading enzymes," *Journal of the American Chemical Society*, vol. 134, no. 50, pp. 20521–20532, 2012.

[152] Y. Liu, J. K. Fredrickson, N. C. Sadler, P. Nandhikonda, R. D. Smith, and A. T. Wright, "Advancing understanding of microbial bioenergy conversion processes by activity-based protein profiling," *Biotechnology for biofuels*, vol. 8, no. 1, p. 156, 2015.

[153] L. Wu, Z. Armstrong, S. P. Schröder, C. de Boer, M. Artola, J. M. Aerts, H. S. Overkleeft, and G. J. Davies, "An overview of activity-based probes for glycosidases," *Current opinion in chemical biology*, vol. 53, pp. 25–36, 2019.

[154] J. J. Rosnow, L. N. Anderson, R. N. Nair, E. S. Baker, and A. T. Wright, "Profiling microbial lignocellulose degradation and utilization by emergent omics technologies," *Critical reviews in biotechnology*, vol. 37, no. 5, pp. 626–640, 2017.

[155] C.-L. Kuo, E. van Meel, K. Kytidou, W. W. Kallemeijn, M. Witte, H. S. Overkleeft, M. E. Artola, and J. M. Aerts, "Activity-based probes for glycosidases: profiling and other applications," in *Methods in enzymology*, vol. 598, pp. 217–235, Elsevier, 2018.

[156] T. M. Gloster, R. Madsen, and G. J. Davies, "Structural basis for cyclophellitol inhibition of a β-glucosidase," *Organic & biomolecular chemistry*, vol. 5, no. 3, pp. 444–446, 2007.

[157] N. McGregor, M. Artola, A. Nin-Hill, D. Linzel, M. Haon, J. Reijngoud, A. F. Ram, M.-N. Rosso, G. A. van der Marel, J. D. Codée, *et al.*, "Rational design of mechanism-based inhibitors and activity-based probes for the identification of retaining α-l-arabinofuranosidases," *Journal of the American Chemical Society*, vol. 142, pp. 4648–4662, 2020.

[158] J. Jiang, C.-L. Kuo, L. Wu, C. Franke, W. W. Kallemeijn, B. I. Florea, E. van Meel, G. A. van der Marel, J. D. Codée, R. G. Boot, *et al.*, "Detection of active mammalian gh31 α-glucosidases in health and disease using in-class, broad-spectrum activity-based probes," *ACS central science*, vol. 2, no. 5, pp. 351–358, 2016.

[159] M. D. Witte, W. W. Kallemeijn, J. Aten, K.-Y. Li, A. Strijland, W. E. Donker-Koopman, A. M. Van Den Nieuwendijk, B. Bleijlevens, G. Kramer, B. I. Florea, *et al.*, "Ultrasensitive in situ visualization of active glucocerebrosidase molecules," *Nature chemical biology*, vol. 6, no. 12, p. 907, 2010.

[160] M. Artola, L. Wu, M. J. Ferraz, C.-L. Kuo, L. Raich, I. Z. Breen, W. A. Offen, J. D. Codée, G. A. Van Der Marel, C. Rovira, *et al.*, "1, 6-cyclophellitol cyclosulfates: a new class of irreversible glycosidase inhibitor," *ACS central science*, vol. 3, no. 7, pp. 784–793, 2017.

[161] L. Wu, J. Jiang, Y. Jin, W. W. Kallemeijn, C.-L. Kuo, M. Artola, W. Dai, C. Van Elk, M. Van Eijk, G. A. Van Der Marel, *et al.*, "Activity-based probes

for functional interrogation of retaining β-glucuronidases," *Nature chemical biology*, vol. 13, no. 8, pp. 867–873, 2017.

[162] L. I. Willems, T. J. Beenakker, B. Murray, S. Scheij, W. W. Kallemeijn, R. G. Boot, M. Verhoek, W. E. Donker-Koopman, M. J. Ferraz, E. R. van Rijssel, *et al.*, "Potent and selective activity-based probes for gh27 human retaining α-galactosidases," *Journal of the American Chemical Society*, vol. 136, no. 33, pp. 11622–11625, 2014.

[163] A. R. Marques, L. I. Willems, D. Herrera Moro, B. I. Florea, S. Scheij, R. Ottenhoff, C. P. van Roomen, M. Verhoek, J. K. Nelson, W. W. Kallemeijn, *et al.*, "A specific activity-based probe to monitor family gh59 galactosylceramidase, the enzyme deficient in krabbe disease," *ChemBioChem*, vol. 18, no. 4, pp. 402–412, 2017.

[164] S. P. Schröder, C. De Boer, N. G. McGregor, R. J. Rowland, O. Moroz, E. Blagova, J. Reijngoud, M. Arentshorst, D. Osborn, M. D. Morant, *et al.*, "Dynamic and functional profiling of xylan-degrading enzymes in aspergillus secretomes using activity-based probes," *ACS central science*, vol. 5, no. 6, pp. 1067–1078, 2019.

[165] G. Paës, L. K. Skov, M. J. O'Donohue, C. Rémond, J. S. Kastrup, M. Gajhede, and O. Mirza, "The structure of the complex between a branched pentasaccharide and thermobacillus xylanilyticus gh-51 arabinofuranosidase reveals xylan-binding determinants and induced fit," *Biochemistry*, vol. 47, no. 28, pp. 7441–7451, 2008.

[166] K. Hövel, D. Shallom, K. Niefind, V. Belakhov, G. Shoham, T. Baasov, Y. Shoham, and D. Schomburg, "Crystal structure and snapshots along the reaction pathway of a family 51 α-l-arabinofuranosidase," *The EMBO Journal*, vol. 22, no. 19, pp. 4922–4932, 2003.

[167] A. Miyanaga, T. Koseki, H. Matsuzawa, T. Wakagi, H. Shoun, and S. Fushinobu, "Crystal structure of a family 54 α-l-arabinofuranosidase reveals a novel carbohydrate-binding module that can bind arabinose," *Journal of Biological Chemistry*, vol. 279, no. 43, pp. 44907–44914, 2004.

[168] L. Petersen, A. Ardèvol, C. Rovira, and P. J. Reilly, "Molecular mechanism of the glycosylation step catalyzed by golgi α-mannosidase ii: A qm/mm metadynamics investigation," *Journal of the American Chemical Society*, vol. 132, no. 24, pp. 8291–8300, 2010.

[169] S. Wu and G. J. Letchworth, "High efficiency transformation by electroporation of pichia pastoris pretreated with lithium acetate and dithiothreitol," *Biotechniques*, vol. 36, no. 1, pp. 152–154, 2004.

[170] A. Miyanaga, T. Koseki, H. Matsuzawa, T. Wakagi, H. Shoun, and S. Fushinobu, "Expression, purification, crystallization and preliminary x-ray analysis of α-l-arabinofuranosidase b from aspergillus kawachii," *Acta Crystallographica Section D: Biological Crystallography*, vol. 60, no. 7, pp. 1286–1288, 2004.

[171] C. Vonrhein, C. Flensburg, P. Keller, A. Sharff, O. Smart, W. Paciorek, T. Womack, and G. Bricogne, "Data processing and analysis with the autoproc toolbox," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 293–302, 2011.

[172] G. Winter, C. M. Lobley, and S. M. Prince, "Decision making in xia2," *Acta Crystallographica Section D: Biological Crystallography*, vol. 69, no. 7, pp. 1260–1273, 2013.

[173] M. D. Winn, C. C. Ballard, K. D. Cowtan, E. J. Dodson, P. Emsley, P. R. Evans, R. M. Keegan, E. B. Krissinel, A. G. Leslie, A. McCoy, *et al.*, "Overview of the ccp4 suite and current developments," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 235–242, 2011.

[174] A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, and R. J. Read, "Phaser crystallographic software," *Journal of applied crystallography*, vol. 40, no. 4, pp. 658–674, 2007.

[175] A. A. Lebedev, P. Young, M. N. Isupov, O. V. Moroz, A. A. Vagin, and G. N. Murshudov, "Jligand: a graphical tool for the ccp4 template-restraint library," *Acta Crystallographica Section D: Biological Crystallography*, vol. 68, no. 4, pp. 431–440, 2012.

[176] P. Emsley, B. Lohkamp, W. G. Scott, and K. Cowtan, "Features and development of coot," *Acta Crystallographica Section D - Biological Crystallography*, vol. 66, p. 486, 2010.

[177] G. N. Murshudov, P. Skubák, A. A. Lebedev, N. S. Pannu, R. A. Steiner, R. A. Nicholls, M. D. Winn, F. Long, and A. A. Vagin, "Refmac5 for the refinement of macromolecular crystal structures," *Acta Crystallographica Section D: Biological Crystallography*, vol. 67, no. 4, pp. 355–367, 2011.

[178] D. Case, R. Botello-Smith, D. Cerutti, T. Cheatham, T. Darden III, R. Duke, *et al.*, "Amber 16," *San Franc. Univ. Calif*, 2016.

[179] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb," *Journal of chemical theory and computation*, vol. 11, no. 8, pp. 3696–3713, 2015.

[180] B. R. Miller III, T. D. McGee Jr, J. M. Swails, N. Homeyer, H. Gohlke, and A. E. Roitberg, "Mmpbsa. py: an efficient program for end-state free energy calculations," *Journal of chemical theory and computation*, vol. 8, no. 9, pp. 3314–3321, 2012.

[181] L. Lebedel, A. Arda, A. Martin, J. Désiré, A. Mingot, M. Aufiero, N. Aiguabella Font, R. Gilmour, J. Jiménez-Barbero, Y. Blériot, *et al.*, "Structural and computational analysis of 2-halogeno-glycosyl cations in the presence of a superacid: An expansive platform," *Angewandte Chemie International Edition*, vol. 58, no. 39, pp. 13758–13762, 2019.

[182] G. A. Olah, *A Life of Magic Chemistry: Autobiographical Reflections of a Nobel Prize Winner*. John Wiley & Sons, 2002.

[183] G. A. Olah, Á. Molnár, and G. S. Prakash, *Hydrocarbon Chemistry, 2 Volume Set*. John Wiley & Sons, 2017.

[184] J. Iglesias Fernández, "Doctoral thesis: Elucidating catalytic mechanisms of glycoside hydrolases and transferases by means of ab initio molecular dynamics simulations," *Universitat de Barcelona*, 2014.

[185] R. J. Woods, C. W. Andrews, and J. P. Bowen, "Molecular mechanical investigations of the properties of oxocarbenium ions. 2. application to glycoside hydrolysis," *Journal of the American Chemical Society*, vol. 114, no. 3, pp. 859–864, 1992.

[186] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson, J. Chen, K. Liles, *et al.*, "Improvements to the apbs biomolecular solvation software suite," *Protein Science*, vol. 27, no. 1, pp. 112–128, 2018.

[187] D. Branduardi, G. Bussi, and M. Parrinello, "Metadynamics with adaptive gaussians," *Journal of chemical theory and computation*, vol. 8, no. 7, pp. 2247–2254, 2012.

[188] P. Li and K. M. Merz Jr, "Mcpb. py: A python based metal center parameter builder," *J. Chem. Inf. Model.*, vol. 56, no. 4, pp. 599–604, 2016.

[189] M. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. e. Petersson, *et al.*, "Gaussian 09 revision d. 01," 2014.

[190] D. Kim and M. L. Klein, "Ab initio molecular dynamics study of the superacid system sbf5/hf solution," *The Journal of Physical Chemistry B*, vol. 104, no. 43, pp. 10074–10079, 2000.

[191] J. M. Seminario, "Calculation of intramolecular force fields from second-derivative tensors," *International journal of quantum chemistry*, vol. 60, no. 7, pp. 1271–1277, 1996.

[192] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J. C. Gumbart, "Rapid parameterization of small molecules using the force field toolkit," *Journal of computational chemistry*, vol. 34, no. 32, pp. 2757–2770, 2013.

[193] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules," *Journal of the American Chemical Society*, vol. 117, no. 19, pp. 5179–5197, 1995.

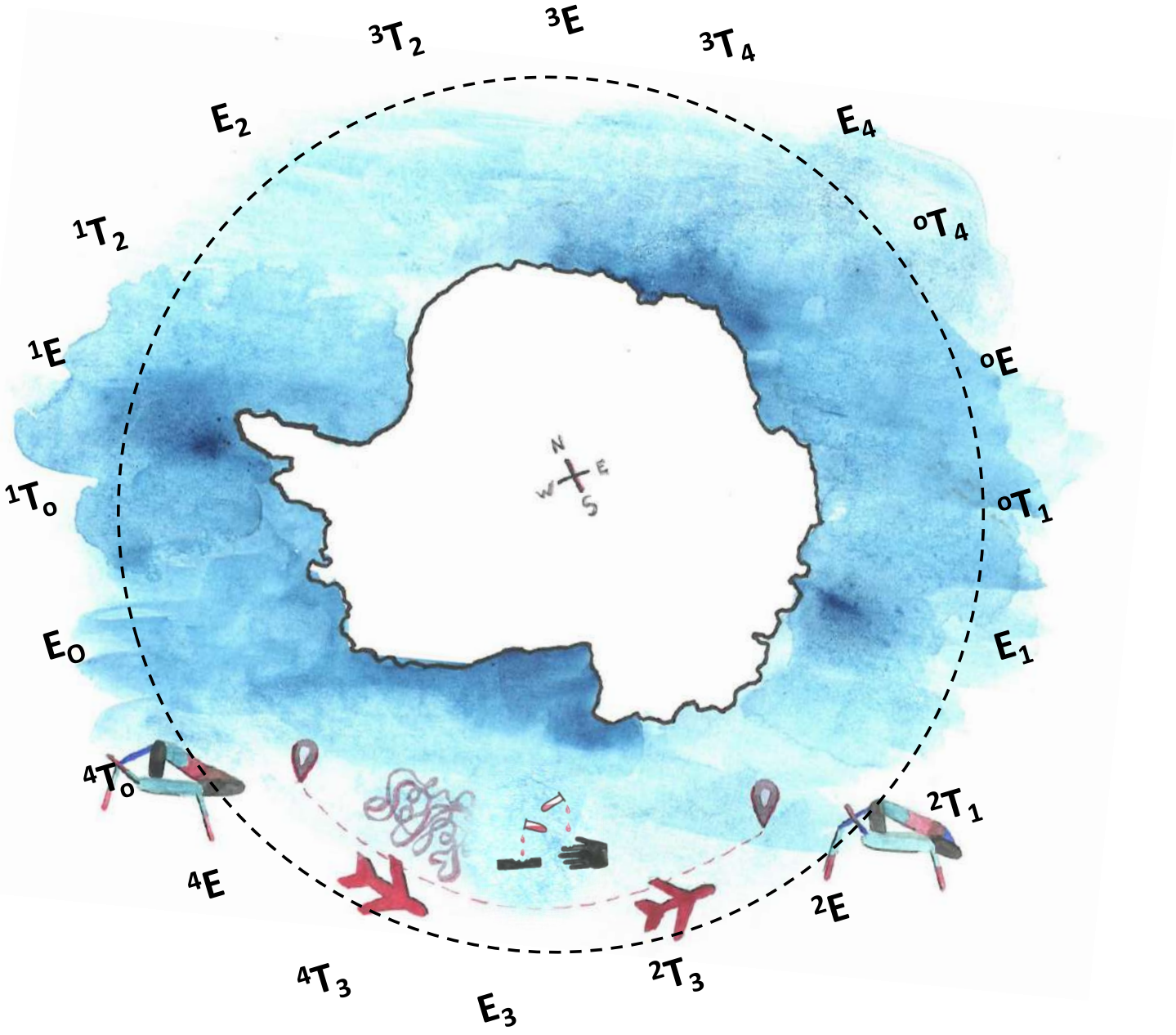
[194] Z. Latajka and Y. Bouteiller, "Application of density functional methods for the study of hydrogen-bonded systems: The hydrogen fluoride dimer," *The Journal of chemical physics*, vol. 101, no. 11, pp. 9793–9799, 1994.

[195] K. Sophy and J.-L. Kuo, "Structure and vibrational spectra of h+(hf) n (n= 2–9) clusters: An ab initio study," *The Journal of chemical physics*, vol. 131, no. 22, p. 224307, 2009.

[196] D. R. Lide, *CRC handbook of chemistry and physics*, vol. 85. CRC press, 2004.

List of Figures

List of Tables



UNIVERSITAT DE BARCELONA