UNIVERSITA' DEGLI STUDI DI SALERNO

Dipartimento di Chimica e Biologia "A. Zambelli"

DOTTORATO DI RICERCA IN CHIMICA
Ph. D. IN CHEMISTRY
XXXIV CYCLE

A computational approach to investigate the interactions between potential pharmacological chaperones and GALT enzyme

Un approccio computazionale per studiare le interazioni tra potenziali chaperoni farmacologici e l'enzima GALT

Tutor:
Prof.ssa
Anna MARABOTTI

Ph.D. student:
Anna Verdino
Matr. 8800100055

Ph.D. coordinator:
Prof.
Claudio PELLECCHIA

A. A. 2021-2022

ABSTRACT

Classic galactosemia is an inborn error of metabolism associated with mutations that impair the activity and the stability of the dimeric enzyme galactose-1-phosphate uridylyltransferase (GALT), which catalyzes the third step in galactose metabolism. Out of more than 300 known mutations, p.Gln188Arg, a missense mutation located at the active site and in the dimer interface, is the most frequently found for GALT. It causes the almost total inactivation of the enzyme and impairs its stability, resulting in the most severe phenotype of the disease. In the past, and more recently, the structural effects of this mutation were deduced on the static structure of the wild-type human enzyme; however, we feel that a dynamic view of the protein is necessary to deeply understand their behavior and obtain tips for possible therapeutic interventions.

We performed molecular dynamics simulations of both wild type and p.Gln188Arg GALT proteins in the absence or in the presence of the substrates in different conditions of temperature. Our results suggest the importance of the intersubunit interactions for the correct activity of this enzyme and can be used as a starting point for the search of drugs able to rescue the activity of this enzyme in galactosemic patients.

Since no treatments, including the current one (the removal of galactose from the diet), are adequate to solve lifelong physical and cognitive disability, some research groups started searching for pharmaceutical chaperones towards GALT. Pharmaceutical chaperones are small molecules able to bind specific target proteins and to stabilize their native conformation or to correct misfolding in proteins affected by mutations thus rescuing their original function. In particular, it has been found that arginine was able to rescue the activity of several mutant GALT enzymes including p.Gln188Arg in a bacterial model of the disease. However, more recently, this rescue was not confirmed testing Arg directly on four galactosemic patients affected by p.Gln188Arg mutation. Given that no molecular characterization of the possible effects of arginine on GALT has been performed, and given that the number of patients treated with arginine is extremely limited for drawing definitive conclusions at the clinical level, we performed

computational simulations to predict the interactions (if any) between this amino acid and the enzyme. Our results do not support the possibility that arginine could function as a pharmacochaperone for GALT, but information obtained by this study could be useful for identifying, in the future, possible pharmacochaperones for this enzyme.

Simultaneously, we wondered if there might be an allosteric site in the GALT enzyme and if it could be used as a target to develop new pharmacochaperones for this enzyme. Through a computational predictor and considering our previous results, we identified a potential allosteric site corresponding also the to portion of the enzyme to which arginine interacts. This potential allosteric site can be a target for new candidate pharmachochaperones for human GALT.

A possible interaction between putative pharmacochaperones was simulated by molecular docking of both wild type and p.Gln188Arg GALT proteins. Starting from the best conformation of docking, the next step was to proceed with the search for pharmacophores, using the method of receptor-based pharmacomodelling. This led to the identification of five new ligands, which were selected for further docking on the allosteric site. All ligands selected showed promising results. These results were used to set up further molecular dynamics studies that are currently ongoing.

Preliminary tests of these ligands on fibroblasts from galactosemic patients showed their ability to lower galactose-1-phosphate concentration when fibroblasts are stressed by galactose. These preliminary data obviously need to be confirmed, but they are promising for the development of pharmacochaperon therapy for galactosemia.

RIASSUNTO

La galattosemia classica è un disturbo metabolico genetico raro causato da mutazioni che compromettono l'attività e la stabilità dell'enzima dimerico galattosio-1-fosfato uridiltransferasi (GALT), che catalizza la terza fase del metabolismo del galattosio. Tra le oltre 300 mutazioni note, p.Gln188Arg, una mutazione missenso situata nel sito attivo e all'interfaccia del dimero, è la più frequentemente riscontrata per l'enzima GALT. Essa causa l'inattivazione quasi totale dell'enzima e ne compromette la stabilità, determinando il fenotipo più grave della malattia. In passato, e più recentemente, gli effetti strutturali di questa mutazione sono stati dedotti dalla struttura statica dell'enzima umano wild-type; tuttavia, abbiamo ritenuto che una visione dinamica della proteina fosse necessaria per comprenderne a fondo il comportamento e ottenere suggerimenti per possibili interventi terapeutici. Abbiamo, perciò, eseguito simulazioni di dinamica molecolare della proteina GALT wild-type e del mutante p.Gln188Arg, in assenza o in presenza dei substrati, in diverse condizioni di temperatura. I nostri risultati suggeriscono l'importanza delle interazioni intersubunitarie per una corretta attività di questo enzima e possono essere utilizzati come punto di partenza per la ricerca di farmaci in grado di ripristinare l'attività di questo enzima nei pazienti galattosemici.

Poiché l'attuale trattamento della malattia (l'eliminazione del galattosio dalla dieta) non è adeguato a risolvere la disabilità fisica e cognitiva dei pazienti galattosemici, che può durare tutta la vita, alcuni gruppi di ricerca hanno iniziato a cercare farmacochaperoni per GALT. I farmacochaperoni sono piccole molecole, in grado di legare specifiche proteine bersaglio, che possono stabilizzare la loro conformazione nativa o addirittura correggere il misfolding di proteine affette da mutazioni, ripristinando così la loro funzione originale. In particolare, si è scoperto che l'arginina era in grado di ripristinare l'attività di diversi enzimi mutanti di GALT, tra cui p.Gln188Arg, in un modello batterico della malattia. Tuttavia, recentemente, testando l'arginina direttamente su quattro pazienti galattosemici affetti dalla mutazione p.Gln188Arg, questo ripristino

funzionale non è stato confermato. Dato che non sono state effettuate caratterizzazioni molecolari dei possibili effetti dell'arginina nei confronti di GALT e dato che il numero di pazienti trattati con arginina è estremamente limitato per trarre conclusioni definitive a livello clinico, abbiamo effettuato simulazioni computazionali per prevedere le interazioni (se esistono) tra questo aminoacido e l'enzima. I nostri risultati non supportano la possibilità che l'arginina possa funzionare come farmacochaperone per GALT, ma le informazioni ottenute da questo studio potrebbero essere utili per identificare, in futuro, possibili farmacochaperoni per questo enzima.

Allo stesso tempo, ci siamo chiesti se potesse esistere un sito allosterico nell'enzima GALT e se potesse essere usato come bersaglio per sviluppare nuovi farmacochaperoni per questo enzima. Attraverso un predittore computazionale e considerando i nostri precedenti risultati, abbiamo identificato un potenziale sito allosterico corrispondente anche alla porzione di interazione dell'enzima con l'arginina. Questo potenziale sito allosterico può essere un bersaglio per nuovi farmacochaperoni, che abbiamo identificato come candidati per l'enzima GALT umano.

Mediante docking molecolare, è stata simulata una possibile interazione tra i nuovi farmacochaperoni e le proteine GALT wild type e p.Gln188Arg. Partendo dalla migliore conformazione del docking, il passo successivo è stato quello di procedere con la ricerca di farmacofori, utilizzando un medoto chiamato “receptor-based”. Ciò ha portato all'identificazione di cinque nuovi ligandi, che sono stati selezionati per un ulteriore docking sul sito allosterico. Una prima analisi rivela che tutti i ligandi selezionati hanno dato risultati promettenti. Questi risultati sono stati utilizzati per impostare ulteriori studi di dinamica molecolare, che sono attualmente in corso. Inoltre, risultati preliminari eseguiti su fibroblasti di pazienti galattosemici hanno suggerito che questi composti siano in grado di migliorare l'attività dell'enzima.

Test preliminari di questi ligandi su fibroblasti di pazienti galattosemici hanno mostrato la capacità di tutti di abbassare la concentrazione di galattosio-1-fosfato quando i fibroblasti sono stressati dal galattosio. Questi dati preliminari devono ovviamente

essere confermati, ma sono dati promettenti per lo sviluppo di una terapia basata su farmacochaperoni per la galattosemia.

ABBREVIATIONS

AGAL (deficit of acid alpha-galactosidase A enzyme)

3D (three-dimensional)

C1q (complement receptor)

CFTR (anion channel)

FcγR (Fc gamma receptor)

FSPF (focal sclerosis permeability factor)

FDA (Food and Drug Administration)

G1P (glucose 6-phosphate)

G6PDH (glucose-6-phosphate dehydrogenase)

GALE (UDP-galactose-4'-epimerase)

GALK1 (galactokinase)

GALM (galactose mutorotase)

GALT (galactose-1-phosphate uridylyltransferase)

GALNS (N-acetylgalactosamine-6-sulfate sulfatase)

GCase (deficiency of glucocerebrosidase)

H2U (5,6-dihydrouridine-5-monophosphate)

HexA (human hexosaminidase)

hGALT (the human enzyme GALT)

IgG (gamma immunoglobulins)

MD (molecular dynamics)

PCs (pharmacochaperones)

PGM (phosphoglucomutase)

PMM2 (phosphomanno-mutase 2 enzyme)

RMSD (root mean square deviation)

RMSF (root mean square fluctuation)

RBC (red blood cells)

SASA (predicted solvent accessible surface area)

SRNS (idiopathic steroid-resistant nephrotic syndrome)

UDP (uridine diphosphate)

UMP (uridine monophosphate)

UPR (unfolded protein response)

UROS (uroporphyrinogen III synthase)

INDEX

1. INTRODUCTION

1.1 Galactose

Louis Pasteur discovered galactose in milk in 1856 and called it "lactose". It was only later that the name "galactose", derived from the Greek word "*γαλακτο*", which means "milk", was attributed to this sugar [Coelho et al., 2015a].

Galactose is a monosaccharide, aldohexose sugar, C4 epimer of glucose. It is more common in nature in its D-configuration, it is ubiquitous in all bacteria, plants and animals [Bell et al., 2012]. In humans, it is produced endogenously in small amounts, but it is introduced in large amounts from the diet. In fact, it is an energy-providing nutrient present mainly in milk and dairy products, bound to glucose to form the disaccharide sugar lactose, a vital source of energy for infants [Bell et al., 2012]. Moreover, it is also present as a free monosaccharide in many food plants or legumes, nuts, and cereals [Acosta, et al., 1995].

The biologically active isomer, D-galactose, plays a fundamental role for our organism not only at the energetic level, but also in many other processes (Figure 1.1). At the respiratory level, for example, galactose is one of the main components of mucin, the glycoprotein that synthesizes mucus, a viscous colloid that forms a physical barrier in all epithelial surfaces of the human body, including the gastrointestinal, respiratory, reproductive, and urinary tracts [Bansil and Turner, 2018].

Galactose is also involved in galactosylation of the most abundant immunoglobulins in plasma (IgG). In particular, this galactosylation is an essential step to achieve immune activation by autoantibodies either through complement (C1q) or Fc gamma receptors (FcγR). In fact, the agalactosylation decreases affinity for FcγR and also C1q binding, leading to immune diseases [Kemna et al., 2017].

In kidneys, the role of galactose relates to a condition called idiopathic steroid-resistant nephrotic syndrome (SRNS). In particular, SRNS has been associated with the

presence of a circulating focal sclerosis permeability factor (FSPF), which is believed to damage the renal glomerular barrier. In vitro, galactose has been shown to bind to FSG, inactivating it. Galactose may have also an effect in vivo on glomerular permeability but its role has yet to be clarified by further studies [Sgambat et al., 2013]. Moreover, galactose is used in the biosynthesis of several macromolecules in the human body, including glycolipids and glycoproteins. Glycoproteins are cell membrane components that are important for cellular signal transduction [Alberts et al., 2002]. The galactosylation of proteins protects and stabilize surface proteins, improving their structural stability. Galactose plays also an essential role in the central nervous system, because most glycolipids, such as gangliosides, cerebrosides, and sphingolipids contains it. For example, galactosylceramide is the primary sphingolipid found in the myelin sheath [Zöller et al., 2005].

Finally, at the gastrointestinal level, galactose has a prebiotic role, maintaining the microbiota, and at the reproductive level it contributes to fertility, facilitating sperm penetration into the zona pellucida [Kotb et al., 2019].

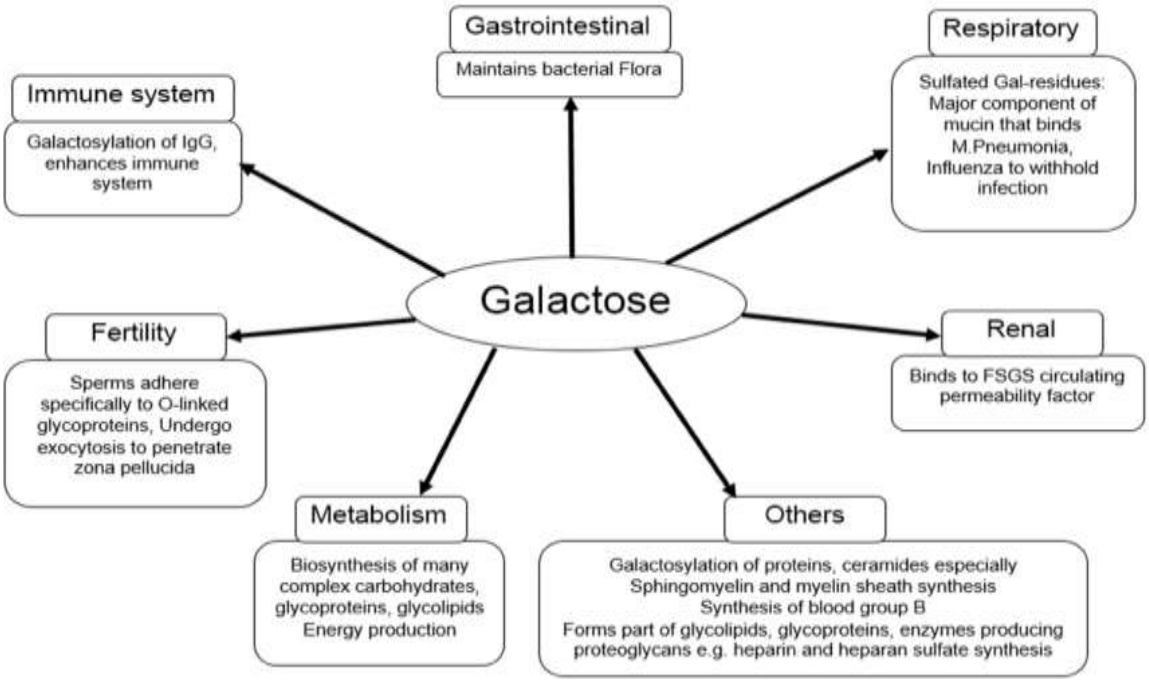


Figure 1.1: The roles of galactose *in vivo* [Kotb et al., 2019]

1.2 The catabolism of D-galactose through the Leloir pathway

After being released from lactose and hydrolyzed into its monosaccharide components by the disaccharidase lactase (β-galactosidase) in the enterocytes of intestinal villi [Coelho et al., 2015a], β-galactose is converted into glucose-1-phosphate (G1P), through a metabolic pathway called "Leloir pathway" in the honour of the researcher who first discovered it [Leloir, 1951].

The Leloir metabolic pathway is highly conserved, from bacteria to yeast to humans, confirming the importance of galactose in living organisms [Chai et al., 2013].

This pathway is divided into 4 steps, each of which catalyzed by a specific enzyme: galactose mutarotase (GALM, E.C. 5.1.3.3), galactokinase (GALK1, E.C. 2.7.1.6), galactose-1-phosphate uridylyltransferase (GALT, E.C. 2.7.7.12) and UDP-galactose-4'-epimerase (GALE, E.C. 5.1.3.2) (Figure 1.2).

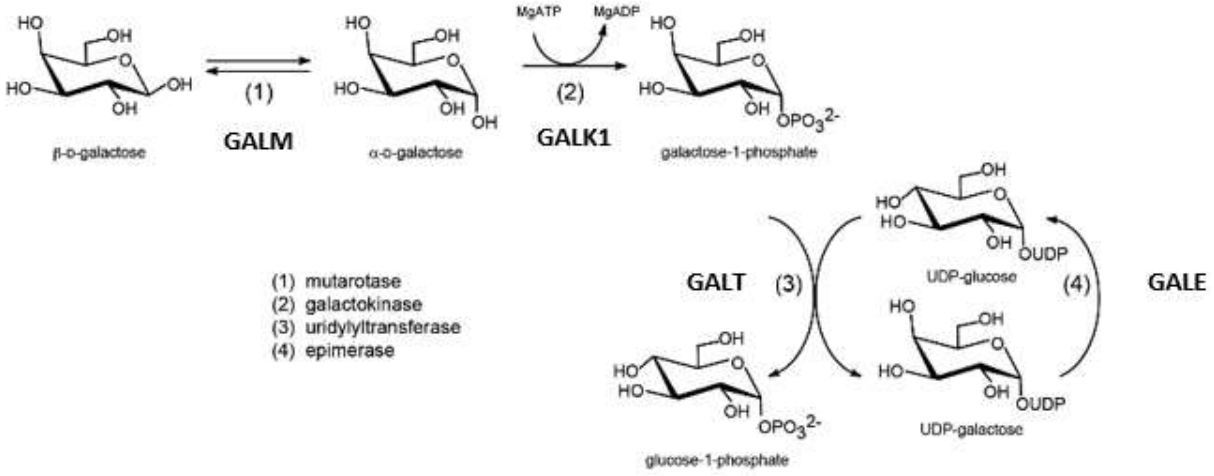


Figure 1.2: The Leloir pathway [Reinhardt et al., 2013]

STEP 1: Conversion of β-D-galactose to the α-anomer by the GALM enzyme (Figure 1.3). The enzymatic catalysis follows an acid-base mechanism, involving two crucial residues*:* glutamic acid at position 304 (Glu304) and histidine at position 170 (His170). Glu304 accepts a proton of the hydroxyl group in the C-1 position and His170 gives a

proton to the oxygen in the C-5 position. This causes the opening of the galactose ring, followed by a rotation of 180º of the C1 - C2 bond [Thoden et al., 2003].

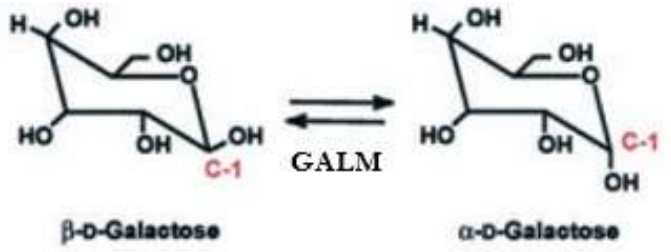


Figure 1.3: Step 1 of Leloir's pathway [Holden et al., 2003]

STEP 2: conversion of α-D-galactose to galactose-1-phosphate (Figure 1.4). The enzyme GALK1 catalyzes the phosphorylation of the hydroxyl group bound to the anomeric C-1 carbon by a reaction with a well-defined temporal sequence: ATP binds to the enzyme after the sugar [Holden et al., 2003].

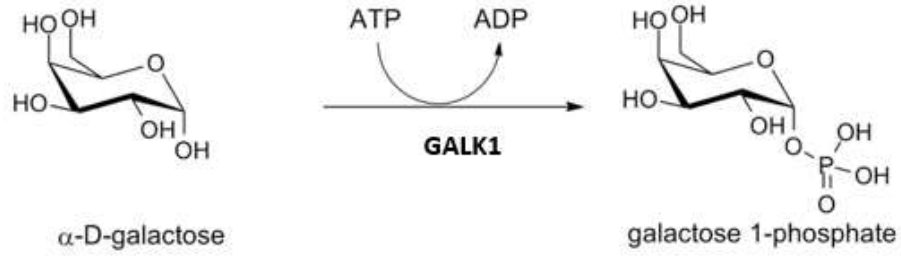


Figure 1.4: Step 2 of the Leloir's pathway [Holden et al., 2003]

The active site of the human enzyme has correctly positioned negative and positive side chains, namely aspartic acid at position 186 (Asp186) and arginine at position 37 (Arg37), both of which influence the overall activity of GALK1. The proposed mechanism is divided into two steps: the first reaction between enzyme and galactose and the second reaction between galactose and ATP (Figure 1.5). The anionic form of Asp186 is stabilized by the neighboring Arg37. In fact, one of the roles of Arg37 is to increase the pKa of Asp186. The positive charge of Arg37 also helps bind negatively charged species in the active site. The aspartate residue accepts a proton from the C1-

OH of the sugar. The resulting extremely nucleophilic alkoxide ion attacks the γ-phosphorus of ATP, transferring the phosphate group to the sugar. The protonation state of Asp186 can be restored by an interaction with water once the products have diffused away from the active site [Megarity et al., 2011].

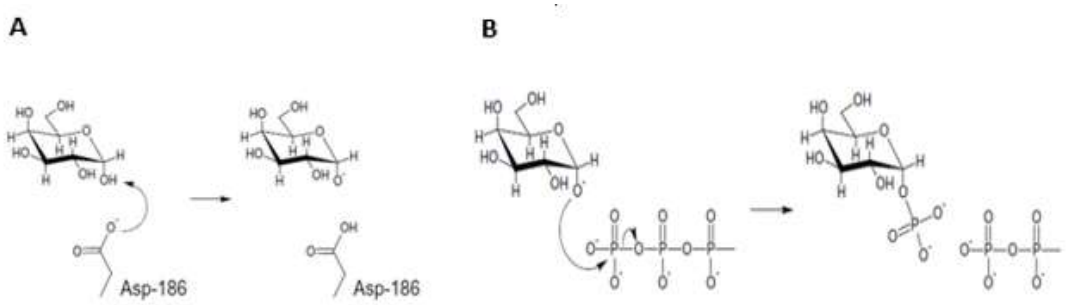


Figure 1.5 (A) A potential active site mechanism for galactokinase: attack of Asp186 to galactose (B): attack of galactose to ATP [Megarity et al., 2011]

STEP 3: conversion of galactose-1-phosphate into glucose-1-phosphate (G1P) (Figure 1.6). This reaction is catalyzed by GALT enzyme and takes place in the presence of the complex uridine diphosphate (UDP)-glucose and using as a substrate galactose-1-phosphate produced in the previous step. In particular, there is the transfer of the uridine monophosphate (UMP) group from UDP-glucose to galactose-1-phosphate, resulting in the formation of UDP-galactose and G1P, which is subsequently transformed into glucose-6-phosphate that enters as an intermediate in the glycolytic pathway [Holden et al., 2003].

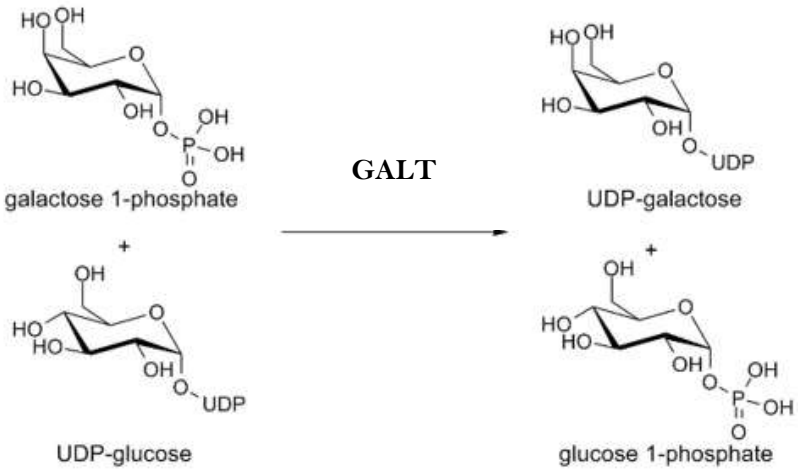


Figure 1.6: Step 3 of Leloir's pathway [Holden et al., 2003]

STEP 4: The enzyme GALE catalyzes the final step of this metabolic pathway by regenerating the UDP-glucose molecule used in the previous step (Figure 1.7).

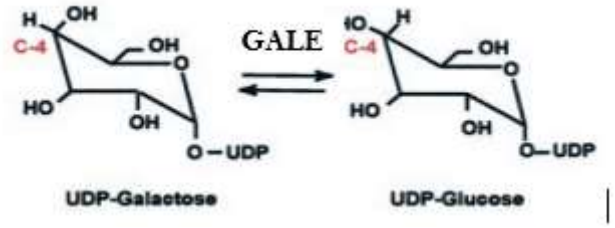


Figure 1.7: Step 4 of Leloir's pathway [Holden et al., 2003].

The enzymatic mechanism of this epimerase probably proceeds in three steps (Figure 1.8): i. a tyrosine residue extracts a proton from the 4'-hydroxyl of UDP-galactose, and the 4'-hydride is added to NAD^+, producing NADH and a 4-ketopyranose intermediate; ii. the 4-ketopyranose intermediate rotates 180°, showing its opposite side to NADH; iii. the hydride is transferred from NADH to C-4 of the sugar, reversing the stereochemistry of the 4'-center. At the same time, a tyrosine residue donates its proton and regenerates the 4'-hydroxyl group [Nam et al., 2019].

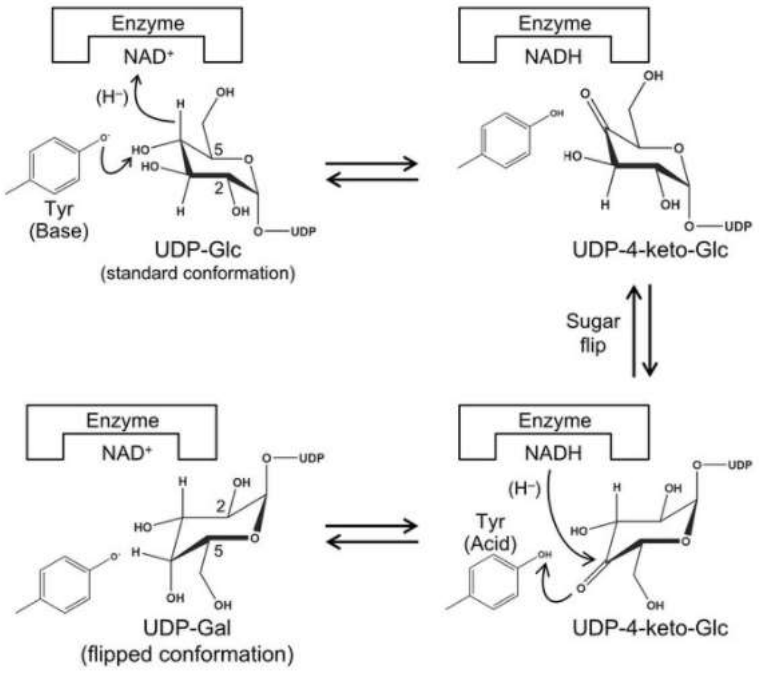


Figure 1.8: Enzymatic mechanism of GALE [Nam et al., 2019]

1.3 GALT enzyme

GALT enzyme belongs to the family of histidine triad transferases characterized by the sequence motif **HhHhHhh** (where **H** stands for histidine and **h** for a hydrophobic amino acid) [Brenner, 2002]. Since the present Ph.D. project is focused on this enzyme, its mechanism of action, structure, and models are described in details below.

1.3.1 The mechanism of action of GALT

The mechanism of action of GALT occurs via a ping-pong reaction kinetics in two steps: uridylation and deuridylation. In the first step (Figure 1.9), a nucleophilic histidine residue (His186 in human GALT and His166 in *Escherichia coli (E. coli)*) attacks the α-phosphate of UDP-glucose to form a covalent adduct, 5,6-dihydrouridine-5-monophosphate (H2U) bound to GALT. Simultaneously, this reaction releases G1P [McCorvie et al., 2016].

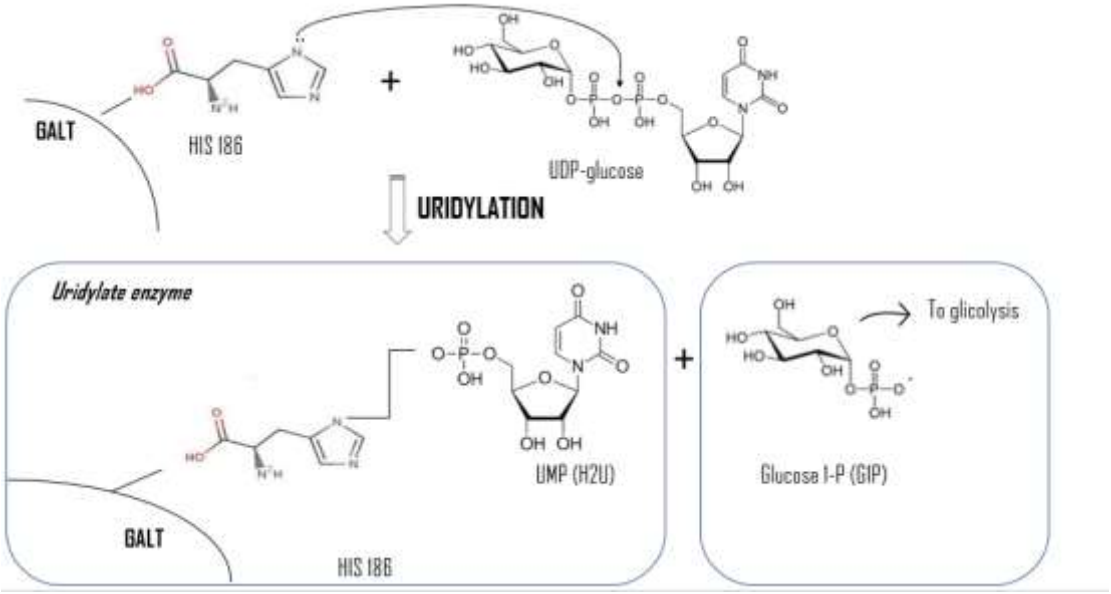


Figure 1.9: Uridylation mechanism of GALT

In the second step (Figure 1.10), the uridylate enzyme interacts with the second substrate of the reaction, the α-galactose-1-phosphate molecule resulting from the previous step catalyzed by the enzyme GALK1. This reaction produces UDP-galactose and regenerates the enzyme GALT, which can then participate in another round of

catalysis [McCorvie et al., 2016]. As reported above, the enzyme GALE catalyzes later the conversion of the UDP-galactose produced by GALT into UDP-glucose so that GALT can use it again as a substrate in a subsequent reaction.

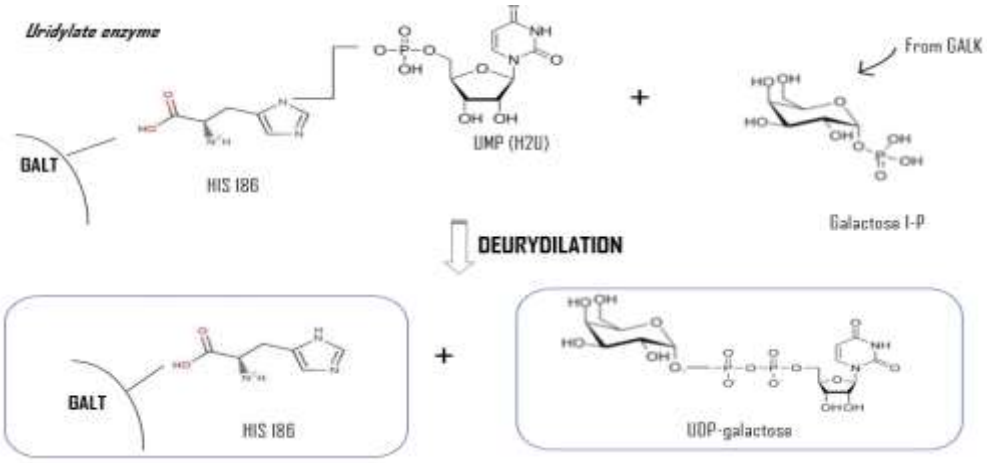


Figure 1.10: Deuridylation mechanism of GALT

1.3.2 GALT: structures and models

Before an experimental structure of the human enzyme GALT (hGALT) was made available, three-dimensional (3D) structures of GALT from *E. coli* in complex with different ligands were solved.

The first structure from *E. coli* was solved by multiple isomorphous replacement and electron density modification techniques and refined to a resolution of 1.8 Å (PDB code 1HXP) [Wedekind et al., 1995]. Thanks to this structure, it was possible to visualize the quaternary assembly typical of GALT: it consists of two identical polypeptide chains with a central cavity between the subunits. This cavity readily accepts water molecules bound by hydrogen bonds inside the cavity. In addition, the protein interface could be visualized, which also contains a considerable number of hydrophobic residues. Each subunit of the enzyme consists of a single domain with a *"half a barrel"* topology. The barrel "staves" are formed by an antiparallel β-sheet made by nine strands [Wedekind et al., 1995].

In that structure, it has been shown that an iron atom is important for stabilizing the structure of the protein. The iron atom is located outside the barrel, at the center of the subunit interface. The coordination within the subunit resembles a distorted square pyramid, which is formed by the equatorial bonding of two histidines and a bidentate carboxylate group, as well as a single axial histidine [Wedekind et al., 1995].

Subsequently, the second structure was solved from *E. coli* with a resolution of 1.86 Å by X-ray diffraction (PDB code 1HXQ) [Wedekind et al., 1996].

In particular, this study was carried out to gain a better understanding of the existing structural and mechanistic studies of this enzyme. This structure has revealed the covalent attachment of the H2U α-phosphorus to His166, representing a genuine reaction intermediate in the double-displacement mechanism [Wedekind et al., 1996].

Next, Thoden and his co-authors determined the structures of *E. coli* enzyme/UDP-glucose (PDB code 1GUQ) and enzyme/UDP-galactose (PDB code 1GUP) complexes, in which the catalytic nucleophile His166 was replaced by a glycine residue. The structures were refined to 1.8 Å resolution by single-crystal X-ray diffraction analysis. These models have provided an important key to understanding the composition and properties of the active site, showing that it is formed by amino acid residues derived from both subunits of the dimer. For example, these models have highlighted the importance of the side chains of Glu317 and Gln323, able to accommodate both UDP-galactose and UDP-glucose substrates. In addition, these models have shown that Gln168 plays an essential role by binding to the phosphate of the substrate, and that three residues (Leu54, Val61, Phe151) provide important hydrophobic surfaces for the active site.

Considering that hGALT shares 46% sequence identity with the bacterial enzyme and that the two sequences are very similar starting from the 20th amino acid of the human sequence, the homology modeling strategy was successfully applied to build a suitable model of the hGALT enzyme, starting from these bacterial structures. This theoretical model allowed to study the catalytic mechanism of the human enzyme and to lay the

foundations for the study of the mutations of this enzyme (as will be explained below) [Marabotti and Facchiano, 2005; Facchiano and Marabotti, 2010].

The first crystallographic structure of hGALT (in particular, of the variant p.Asn314Asp) was obtained in 2016 and deposited in the Protein Data Bank with the PDB code 5IN3 [McCorvie et al., 2016]. hGALT was obtained by molecular replacement using bacterial GALT (PDB code 1HXP) as a template [Wedekind et al., 1995]. Unlike GALT from *E. coli*, which had an iron atom, hGALT is a metalloprotein in which two zinc ions are bound to a site approximately 20 Å far from the active site, formed by residues Glu202, His301, His319, and His321. As expected, hGALT exhibits a homodimeric quaternary arrangement and consists of two identical chains (A and B) of 379 amino acids each [Brenner, 2002]. Each protomer is arranged in a central nine-stranded β-sheet flanked on either side by five α-helices and a small three-stranded β-sheet (Figure 1.11). The interface between the two subunits is stabilized by 17 hydrogen bonds and 2 salt bridges between residues Asp 113 (chain B)-Arg228 (chain A) and His114 (chain B)-Glu 220 (chain A). In particular, these salt bridges are located at the end of a dimerization loop (residues 106-122) and stabilize the interface.

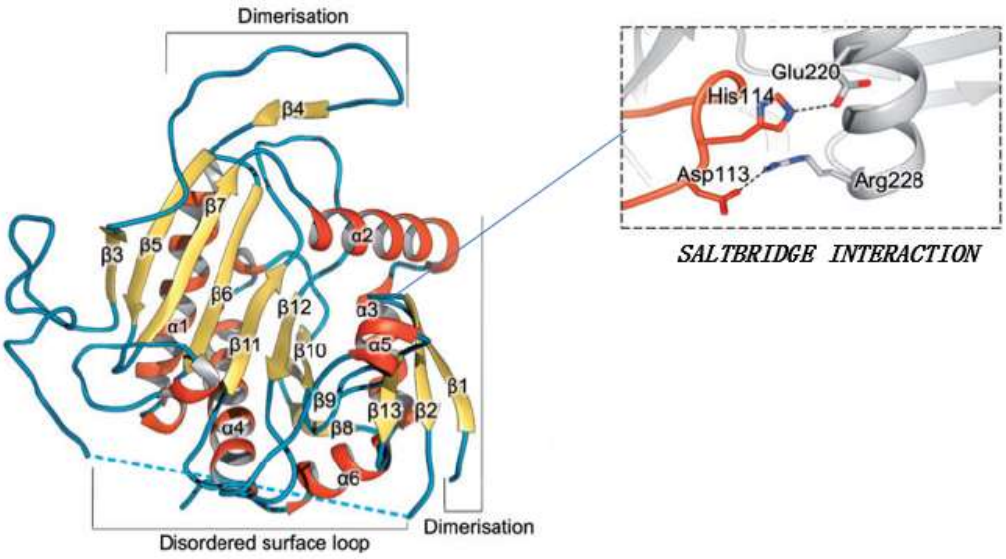


Figure 1.11 Cartoon representation of hGALT structure (chain A) showing secondary structure elements and showing the two salt bridges Asp113B-Arg228A and His114B-Glu220A [McCorvie et al., 2016].

.

The two chains (A and B) are arranged in a mirror-inverted manner with respect to each other, forming a central cavity and two active sites at the interfaces between them (Figure 1.12). The association of the two monomers leads to the catalytically active form of the protein, as both active sites consist of residues belonging to both chains (table 1.1).

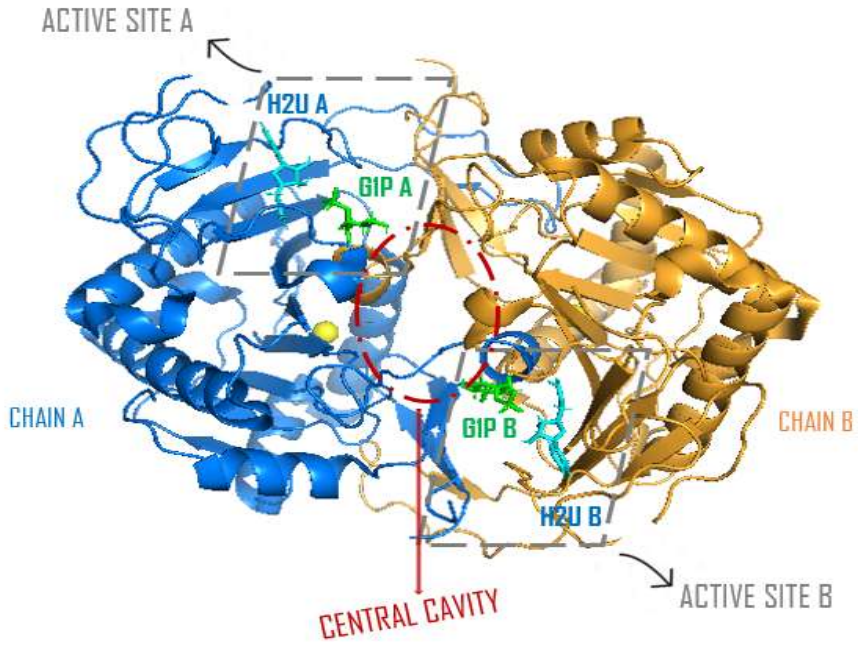


Figure 1.12: Structure of hGALT: Chain A is in blue, chain B in orange; the substrates G1P are represented in green; the substrates H2U are represented in cyan. Figure obtained with Pymol.

SUBSTRATE	INTERACTING RESIDUES	
ACTIVE SITE A	**CHAIN A**	**CHAIN B**
G1P A	N173 and Q188	K334, F335, V337, Y339, E340, Q346
H2U A	G179, S181, H186	R48, R51
ACTIVE SITE B	**CHAIN A**	**CHAIN B**
G1P B	K334, F335, V337, Y339, E340, Q346	N173 and Q188
H2U B	R48, R51	G179, S181, H186

Table 1.1: Residues interacting with G1P and H2U in hGALT [McCorvie et al., 2016].

The uridylation induces structural changes to the protein and reduces its flexibility. The conformational changes are shown synthetically below:

- *loop 49-63*, corresponding to a surface-exposed region in the H2U-binding site. It was not possible to determine an electron density map for this loop, but it is supposed to become ordered when bound to UDP sugar, which makes it more compact, less flexible, less susceptible to degradation, and more thermostable;
- *loop 76-90*, which is involved in H2U binding. It loses flexibility during uridylylation; H2U-hGALT is more compact than apo-hGALT, as evidenced by the fact that uridylylation reduces the radius of gyration and intrinsic hydrophobicity [McCorvie et al., 2016].

Recently, this crystallographic structure of hGALT was used as a template to re-model the structure of this enzyme. This made it possible to remove non-canonical residue conformations, to add the missing residues, to remove the mutation p.Asn314Asp in the crystallographic structure, and to model the loop at position 48-62, which is not present in the crystallographic structure. This final model is deposited in a web-accessible resource called Galactosemia Proteins Database 2.0 [d'Acierno et al., 2018] (https://proteinvariants.eu/galactosemia). The superposition of the new model with the crystallographic structure revealed that the backbone and the conformations of the side chains are generally preserved. This final model was used as a reference for many analyses, including those presented in this work.

In 2018, a second hGALT structure in a complex with H2U was deposited in PDB with the code 6GQD [Fairhead et al., 2018]. This structure consists only of the A chain (from Tyr 21 to Tyr 366) and was solved by X-ray diffraction at a resolution of 1.52 Å, with crystallization of some epitope mutations. More precisely, it is an hGALT artificial variant, with substitutions at the protein surface to a crystallization-prone epitope motif (A21Y:A22T:T23P:R25L), aimed at enhancing protein crystallizability.

1.4 Galactosemia: different forms

Galactosemia is an ensemble of rare genetic metabolic disorders characterized by the impairment of galactose metabolism. The discovery of galactosemia dates back to 1908 when von Ruess, in the publication *"Sugar Excretion in infancy"*, classified it as a disorder belonging to the class of carbohydrate metabolism [Bray et al., 1951]. The different galactosemia types are caused by mutations in the genes coding for the four enzymes essential in the Leloir galactose degradation metabolic pathway. All these diseases follow an autosomal recessive pattern of inheritance [Wada et al., 2020].

Galactosemia type I (OMIM: #230400) is caused by homozygous or compound heterozygous mutations in the gene encoding for the enzyme GALT, located on chromosome 9p13. This disease is also called "classic galactosemia" [Kotb et al., 2019]. The incidence of this disease varies when comparing different nations: a higher incidence is found in Irish ancestry (1:24,000), whereas the lowest incidence is found in Swedish ancestry (1:100,000) [Kotb et al., 2019], with an annual incidence of 1 per 30,000/60,000 births worldwide and 1 per 47,000 in the Caucasian population [Moammar et al., 1996]. This form of galactosemia will be discussed in detail in the following paragraph.

Galactosemia type II (OMIM: #230200), first described by Gitzelmann [Gitzelmann, 1965], is the mildest form of galactosemia and is caused by mutations in the gene on chromosome 17q24 that encodes GALK1. It is estimated that the incidence of galactosemia type II is less than 1 in 100,000 births [Hennermann et al., 2011]. The prevalence of galactosemia type II is estimated to be approximately 1 in 1,000,000 in Japan and 1 in 60,000 in the United States [Sneha et al., 2018]. The only recurrent clinical sign is an early cataract due to the accumulation of galactitol in the lens. Cataract occurs in the neonatal period and can lead to significant visual impairment, but galactose-reduced diet leads to its regression. Other clinical signs, such as hypoglycemia, hepatomegaly, and hypercholesterolemia, are more difficult to detect, while others, such as symptomatic mental retardation, microcephaly, and failure to

thrive, are sporadic and difficult to diagnose. Patients with GALK1 deficiency generally have increased plasma galactose concentration and galactitol excretion in the urine [Hennermann et al., 2011].

Galactosemia type III (OMIM: #230350) is caused by mutations in the gene located on chromosome 1p36 encoding the enzyme UDP-galactose 4-epimerase (GALE). GALE deficiency was demonstrated to exist in a rare but clinically severe "generalized" form [Openo et al., 2006]. It has been first described by Gitzelmann [Gitzelmann, 1965] and confirmed by Holton when he reported the case of a child who had similar symptoms to patients with classical galactosemia, but with normal GALT activity and diminished GALE activity [Holton et al., 1981].

In 1990, two forms of type III galactosemia were identified: the peripheral and the generalized form [Endres, 1990]. The first form is a benign disease in which only the galactose level is altered in patients, as in type II galactosemia, whereas the second form is severe and resembles type I galactosemia. Later, several mutations were found to be associated with an "intermediate" form of the disease, i.e., the patients and/or their cells with intermediate GALE impairment may have abnormally high galactose 1-phosphate blood levels, referred to the cells, in the presence of galactose, as well as abnormally high UDP-galactose levels and low UDP-glucose levels, even in the absence of galactose in the diet [Openo et al., 2006]. Kalckar hypothesized that patients with GALE deficiency, unlike individuals with classical galactosemia, require at least a small amount of dietary galactose to maintain homeostasis [Kalckar, 1961].

GALE catalyses also the conversion of N-acetylgalactosamine and N-acetylglucosamine, a reaction important in maintaining the pools of UDP-sugars, and the loss of its activity may explain the abnormal glycosylation patterns seen in some cell cultures and animal models of type III galactosemia [Kingsley et al., 1986].

Galactosemia type IV (OMIM: #618881) is caused by mutations in the gene located on chromosome 2p22 encoding GALM and is a recently identified form of galactosemia [Wada et al., 2020]. Iwasawa and coauthors estimated the incidence of

GALM deficiency to be almost 1:10,000 in African populations, almost 1:80,000 in the Japanese population, and much lower in many other populations [Iwasawa et al., 2019].

This new type of genetic galactosemia was detected during newborn screening for classical galactosemia in Japan [Kikuchi et al., 2021]. To date, despite the diagnosis and genotyping of many thousands of galactosemic patients worldwide, the incidence and long-term consequences of GALM deficiency are unknown. However, patients with reduced GALM activity may be asymptomatic for many years. These symptoms are similar to those of II type galactosemia: increased blood galactose concentrations and cataracts at a young age [Timson, 2019].

Patients with GALM deficiency are identified by elevated galactose levels, while GALT activity is normal and galactose 1-phosphate levels are usually below the threshold. A substantial number of cases are likely to go undiagnosed, especially in countries and regions where newborn screening for galactosemia is not available or where blood galactose is not measured [Kikuchi et al., 2021].

1.4.1 Diagnosis of galactosemia

To date, the Beutler test is the most common test for the diagnosis of galactosemia (Figure 1.13). It is based on the detection of the conversion of galactose-1-phosphate to gluconate-6-phosphate, a process involving the activities of GALT, phosphoglucomutase (PGM) and glucose-6-phosphate dehydrogenase (G6PDH). The activity of these three enzymes is revealed by the fluorescence produced by NADPH. A GALT deficit results in a true positive test for galactosemia type I, but a positive result could indicate also PGM or G6PDH deficiency, resulting in a false positive test for galactosemia type I [Banford et al., 2021].

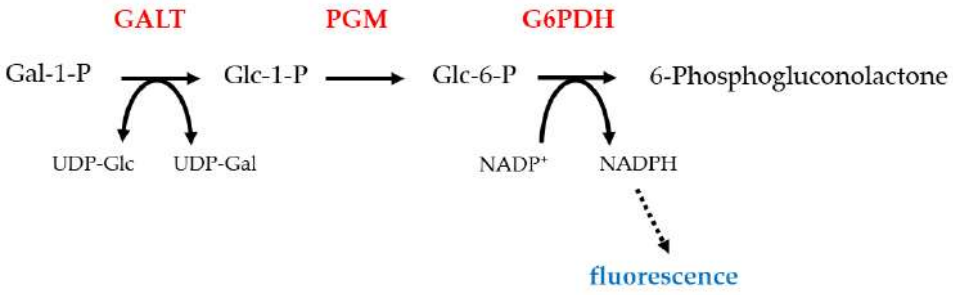


Figure 1.13: Beutler test [Banford et al., 2021]

An alternative test is based on fluorescence produced by NAD^+, the cofactor required for galactose dehydrogenase, a non-human enzyme. Galactose, previously obtained from galactose-1-phosphate by a nonspecific phosphatase enzyme, is then detected by oxidation catalyzed by galactose dehydrogenase [Banford et al., 2021]. With this last test, a false positive result could occur in patients with galactosemia type II because the galactose concentration is not as elevated as in patients with galactosemia type I or III.

Another useful test measures the concentration of galactitol in urines by gas chromatography [Allen et al., 1988].

Complete sequencing of the suspect gene is recommended to obtain a correct diagnosis. In fact, molecular biology techniques are usually used to confirm the diagnosis. Previously, site-specific probes were used to detect common mutations, but these methods inevitably missed unusual and unexpected mutations. To ensure a tailored medical approach, full sequencing of the putative causative gene is now recommended [Viggiano et al., 2018]. Fortunately, galactosemia is included in newborn screening programs in several countries, for example in the European Union, the screening of galactosemia is performed in about one-third of countries. Even in countries where galactosemia is included in screening programs, such as Japan and the United States, screening is primarily for classical galactosemia, the most severe form of galactosemia [Kikuchi et al., 2021].

1.5 Classical galactosemia: clinical picture

Classic galactosemia is the first discovered [Timson, 2016] and the most studied form of galactosemia, due to its frequency and complications [Kotb et al., 2019]. It can manifest with both acute and long-term consequences. The acute form mainly affects newborns, as they feed almost exclusively on breast milk, which is rich in lactose. Once galactosemic newborns come into contact with milk, they show typical symptoms of galactose poisoning such as jaundice, feeding difficulties, failure to thrive, liver cell damage, hemorrhage, and possibly death from *E. coli* sepsis. The acute disease can usually be resolved with a life-long galactose-restricted diet, which unfortunately does not prevent most galactosemic patients from developing a late complication during childhood and adolescence [McCorvie 2011]. Very common clinical signs that may be present include those related to the nervous system. This is expected since, as described in paragraph 1.1, one of the most important roles of galactose in human body is the formation of the myelin sheath of nerve fibers [Kotb et al., 2019]. The symptoms related to the nervous system are mental dysfunction, dysarthria, anxiety, ataxia, attention deficit, hyperactivity, apraxia disorder, and autistic behavior [Lynch et al., 2015]. Other symptoms that are often seen in patients with this disease include abnormalities of the reproductive system, especially in females who may suffer from oligomenorrhea, premature ovarian insufficiency, delayed puberty and decreased fertility. Additionally, many patients also experience hepatic failure, hepatomegaly, and elevated hepatic transaminase levels. Other symptoms regard encephalopathy, feeding difficulties, gait disturbance and imbalance, hypoglycemia, low levels of vitamin D, and osteoporosis [Kotb et al., 2019].

In addition to the different GALT variants (see below), it is worth mentioning the Duarte polymorphism, identified by Reichard and Woo in 1991. In the same year, Elsas and coauthors discovered that one of two biochemical phenotypes of the GALT enzyme, called Duarte 2 (D2), was caused by a single nucleotide variant (c.940A→G

substitution in GALT exon 10) leading to the mutation replacing the original aspartate in position 314 with asparagine (p.Asn314Asp) [Elsas et al., 1994].

The characteristic Duarte isoform is also associated with a variant allele (c.652C→T substitution in GALT exon 7) in cis with p.Asn314Asp leading to a rare and neutral polymorphism for leucine at amino acid 218 (p.Leu181Leu) [Podskarbi et al., 1996]. This last is called Los Angeles phenotype (D1). D1 and D2 variants differ in GALT activity, with D1 showing 110% to 130% of normal red blood cell activity, and D2 only showing 40% to 50%. Podskarbi and coworkers in 1996 suggested that the decrease in GALT activity in D2 could be the result of regulation of GALT gene expression by the intronic mutations causing an aberrant splice processing, possibly inducing the formation of a low level of correctly spliced mRNA [Podskarbi et al., 1996].

In 1997, Langley et al. postulated a favorable codon bias suggesting that the increased GALT activity in D1 may be due to increasing GALT protein abundance without increasing transcription or decreasing thermolability [Langley et al., 1997].

In pediatric subjects with Duarte polymorphism, cognitive abilities (memory, executive function, and auditory processing), communication processes (speech and language), physical development (including motor skills, coordination, and occurrence of tremor), and social-emotional development are only occasionally impaired (https://clinicaltrials.gov/ct2/show/NCT02519504). Typically, patients do not show any symptom, although GALT activity is low. Women are at an increased risk of ovarian cancer and premature ovarian insufficiency [Fung et al., 2003]. However, a large amount of conflicting literature exists: some reports consider this biochemical phenotype to be benign and do not recommend life-long galactose restrictions, while others consider it a pathological condition [Kotb et al., 2018].

1.5.1 Classical galactosemia: the most common associated variants

Galactosemia type I is characterized by high allelic heterogeneity. To date, according to ARUP database (http://arup.utah.edu/database/GALT/GALT_welcome.ph),

[Calderon et al., 2007] more than 330 variants of hGALT are known, most of which are missense mutations (Figure 1.14).

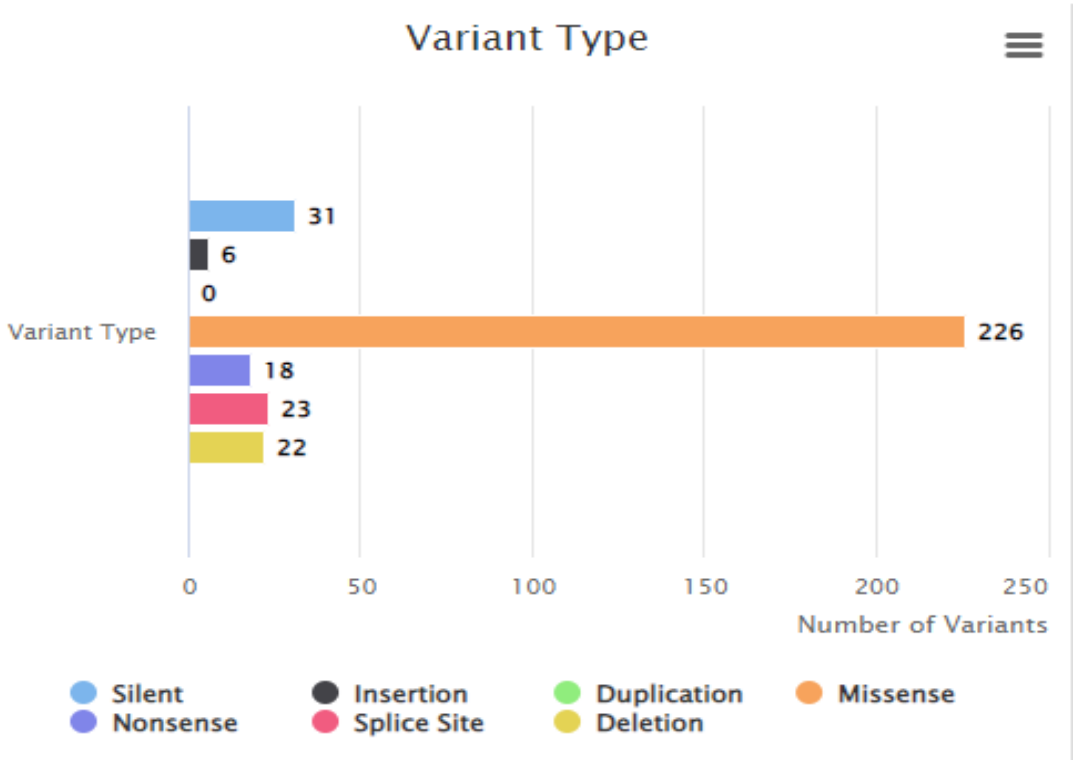


Figure 1.14: Variant types of hGALT [Calderon et al., 2007]

The prevalence of the mutations is different in different geographical and ethnic groups. We report only the most common variants below; for a full list of variants, please refer to http://arup.utah.edu/database/GALT/GALT_welcome.ph.

- ***p.Gln188Arg***: it is the most frequent mutation found for GALT in the Caucasian population and represents up to 70% of classic galactosemia-associated variants. It is a missense mutation replacing the original glutamine, located in the active site and at the dimer interface of GALT enzyme, with arginine. It causes the almost total inactivation of the enzyme and impairs its stability. This has been confirmed by biochemical studies in yeast, which have demonstrated that the homodimeric mutant is characterized by almost total loss of function, while the heterodimer shows a residual activity varying between 15 and 45% compared to the wild-type, instead of the expected 50% [Elsevier and Fridovich-Keil, 1996];

- ***p.Lys285Asn:*** it is a missense mutation replacing the original lysine with asparagine, and it is the second most common European mutation, especially in the East Europeans. It represents up to 34% of galactosemic alleles [Flanagan et al., 2010] and is rare in individuals with non-European ancestry [Manga et al., 2007]. A study of Czech, Slovak, Polish and Austrian galactosemic patients [Kozak et al., 2000] found that the frequency of this mutation was higher in these populations than in other European populations. Particularly, it is one of the most frequent in healthy Slovenian population [Lukac-Bajalo et al., 2005]. Although the individuals who are homoallelic for this mutation have a severe phenotype with complete loss of enzyme activity [Podskarbi et al., 1996; Shin et al., 1999], the heterozygotes have about 50% of normal GALT activity and are asymptomatic at birth. There is some evidence that the risk of developing certain diseases, such as cataracts, is increased later in life for heterozygotes [Karas and Goldberg, 2003];
- ***p.Ser135Leu:*** it is found almost exclusively in African Americans, suggesting that it occurred more than 100,000 years ago, after the first wave of migration of *Homo sapiens* out of Africa [Tyfield et al., 1999]. It is a missense mutation replacing the original serine with leucine and it accounts for almost 50% of mutant alleles in this ethnic group [Lai et al., 1996; Wang et al., 1998]. This mutation was first reported by Reichardt and colleagues [Reichardt et al., 1992]. However, they concluded that it was a polymorphism, using a fibroblast-like cell line (COS cell) transient expression system. The studies of haemolysates, lymphocytes, and lymphoblasts derived from patients homozygous for p.Ser135Leu also revealed levels of activity ranging from undetectable to ~5% wild-type [Lai et al., 1996; Wang et al., 1998]. Subsequent studies, in contrast with the previous study, using a null-background yeast expression system, have demonstrated that there were approximately 5% wild-type levels of activity associated with this variant [Fridovich-Keil et al., 1995]. The explanation for this apparent disparity remains unknown but it may be due either to differences in the

properties of the host cells or to differences between transient- and stable-expression assays [Fridovich-Keil et al., 1995].

1.5.2 p.Gln188Arg: the importance of this variant

Among all variants of hGALT, p.Gln188Arg is a very clinically relevant mutation since it is the most widespread mutation in Caucasian population, as reported above.

The high frequency of p.Gln188Arg mutation in European populations, compared to the very low frequency in Asian populations, suggests that this mutation arose after immigration from Africa and the subsequent divergence of Caucasian and Asian populations [Singh et al., 2012]. It is interesting to know that the frequency increases moving from East to West across the globe [Tyfield et al., 1999].

As told before, this mutation is located at the active site and at the dimer interface of the quaternary GALT assembly. Early biochemical experiments in yeast revealed that the homodimeric mutant has practically complete loss of function, but the heterodimer has activity in the range of 15-45% of wild-type, rather than the expected 50%. Therefore, a partial dominant negative effect on the functionality of the mutant enzyme was observed in heterozygosity, indicating that the negative effect of this mutation is present not only in homozygosity but also when a healthy copy of the monomer is present [Elsevier and Fridovich-Keil, 1996].

Both homodimers and heterodimers with the p.Gln188Arg mutation were among the first theoretical models created for hGALT [Marabotti and Facchiano, 2005]. These models also allowed researchers to study the impact of the mutation on both enzyme-substrate and interchain interactions based on comparisons with the wild-type enzyme. The mutant Arg188 residue establishes different interactions with the substrate than the wild-type Gln188 residue, according to the examination of enzyme-substrate contacts (Figure 1.15):

- Gln188 forms two hydrogen bonds with the UDP-glucose substrate's phosphate component (with the O2 and O1 of the phosphate group). This creates an electron

density dispersion on the α-phosphate, favoring its subsequent nucleophilic attack by the galactose 1-phosphate.

- Arg188 forms only one hydrogen bond with the phosphate portion of the UDP-sugar (with the O2 of the phosphate group) and an additional hydrogen bond with the O5 that connects the uridyl portion to the phosphate chain; this bond is unable to disperse the negative charge in the same way as the other hydrogen bonds [Marabotti and Facchiano, 2005]. The reaction at this point cannot proceed to the second step of the ping-pong mechanism, and there is an accumulation of G1P at the cellular level.

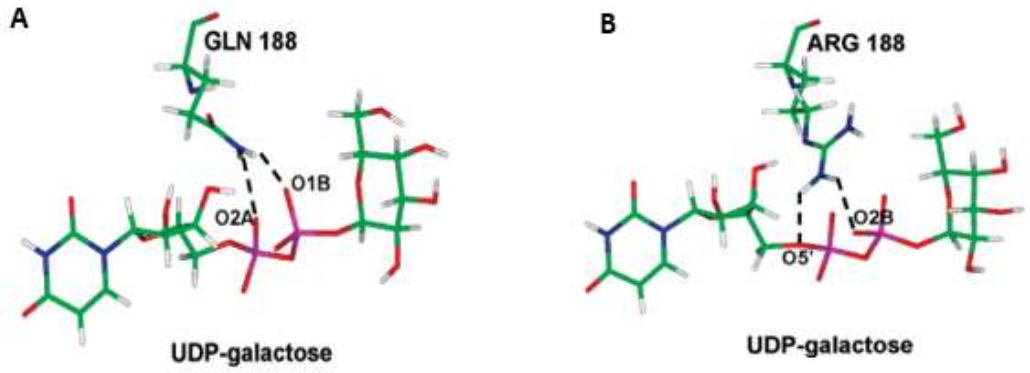
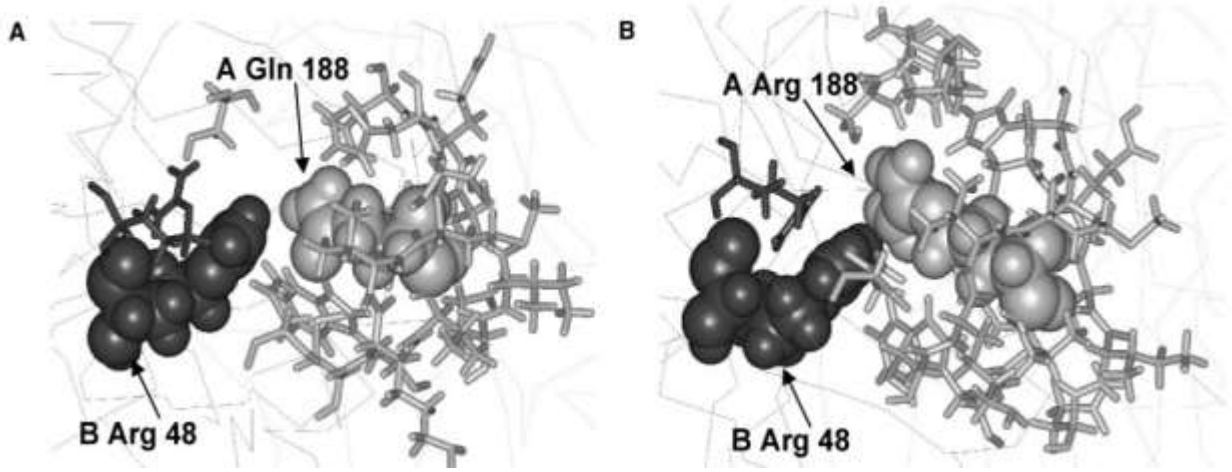


Figure 1.15: A. Interactions of Gln188 with UDP-galactose B. Interactions of Arg188 with UDP-galactose [Marabotti and Facchiano, 2005].

An interchain contact analysis showed that there were significantly fewer hydrogen bonds in the mutant heterodimer and homodimer protein than in the wild-type. This predicted that the mutation not only had functional effects within the enzyme, but also structural ones. Since the active site is part of the dimer interface, being formed by residues from both subunits, this means that a mutation in the active site could disrupt not only the enzymatic catalytic function but also the interaction between the subunits [Marabotti and Facchiano, 2005]. Marabotti and Facchiano confirmed that amino acids included within a distance of 5 Å from residue 188 belong to both chains of the dimer. Among these amino acids, the nearest one is Arg48 of the other chain to which residue 188 belongs (Figure 1.16). They evidenced that, when residue 188 is mutated to

arginine, it is in strict contact with Arg 48. Furthermore, within a radius of 5 Å from residue 188, there are three further positively charged residues (Arg48, Arg51 and Lys127), while only one residue (Glu172) is negatively charged. This suggests that the causes of unfavourable dimerisation are both steric hindrance, due to the high proximity of two bulky residues, and repulsion due to an unfavourable electrostatic interaction [Marabotti and Facchiano, 2005].



Figur 1.16: (A) Close-up of the contact between Gln188 (light gray) in chain A and Arg48 (dark gray) in chain B of hGALT. (B) Close-up of the contact between Arg188 (light gray) in the chain A and Arg48 (dark gray) in chain B of Q188R-hGALT. Amino acids within a distance of 5 Å from residue 188 are represented in stick mode [Marabotti and Facchiano, 2005].

McCorvie and coworkers also investigated the effects of uridylation and zinc binding on p.Gln188Arg aggregation. They discovered that p.Gln188Arg is more prone to aggregation than H2U-hGALT but has comparable aggregation rates to apo-hGALT. This indicates that uridylylation decreases the kinetics of hGALT aggregation, and that p.Gln188Arg aggregation is likely related to the lower degree of uridylylation. They observed that zinc binding causes aggregation of all mutant proteins; however, this aggregation is severely increased in the p.Gln188Arg protein. Unexpectedly, high Zn^{2+} content prevented aggregation of all hGALT species during the experiment. Using a method available at "*The Aggrescan server"* (http://bioinf.uab.es/aggrescan/), they also predicted that the active site of hGALT, which has a β-strand-rich structure, is a hot region for aggregation that can be altered by uridylylation at the His186 active site.

Regardless, there seems to be a correlation between the aggregation of the p.Gln188Arg protein, its solubility, and the level of protein in the cell. Higher levels of protein are more likely to be aggregated. This suggests that, although the protein is lowering its solubility significantly, the protein may become dysfunctional in the presence of higher levels of Zn. These results indicate that p.Gln188Arg has a stronger aggregation ability due to its lower ability to be uridylylated and that Zn is a structurally important ion which influences the stability and aggregation tendency of hGALT [McCorvie et al., 2016].

1.5.3 Classical galactosemia: a misfolding disease?

Although the effects of mutations of hGALT are difficult to predict, in most cases they impact on the correct folding of the protein, and the resulting enzyme is unstable from a structural point of view, therefore causing dysfunction [McCorvie et al., 2016].

It is appropriate to describe here previous and parallel studies that put an effort to clarify whether galactosemia is a misfolding disease.

In literature, there are several pieces of evidence that suggest the unfolded protein response (UPR) is activated in galactosemia. Accumulated unfolded protein during the UPR is either correctly refolded, or unsuccessfully refolded and degraded by the ubiquitin-proteasome pathway [Kim et al., 2006].

Research on UPR activation in classic galactosemia has been conducted using a yeast model and a human cell line. Slepak and coworkers showed that in a yeast model of classic galactosemia (gal7Δ mutant, human GALT equivalent) [Slepak et al., 2005] and in a human cell line model of classic galactosemia [Slepak et al., 2007], genes controlled by the UPR are also induced by galactose.

A study conducted on two yeast models of classic galactosemia suggests that galactose-1-phosphate synthesis is essential to causing endoplasmic reticulum stress (ER). ER stress, caused by UPR activation, has been shown to play a protective role against the cytotoxic effect of galactose [De-Souza et al., 2014]. This study provides evidence that

molecules that interfere with ER stress may be effective in treating classic galactosemia [Kraskiewicz and FitzGerald, 2012].

In a 2013 work, McCorvie and co-authors discovered that five hGALT mutants, p.Asp28Tyr, p.Leu74Pro, p.Phe171Ser, p.Phe194Leu, and p.Arg333Gly, lack key interactions required for protein structure stability and substrate binding. However, each one produces distinct modifications in various features of the protein at the same time. Each of the five mutants was tested *in vitro* to examine the stability, substrate binding, capacity to dimerize, and enzyme kinetics. p.Asp28Tyr is the only substitution among the five mutations that does not cause significant changes in the substrate binding, which is reflected by different characteristics from the other four. The fluorescence results suggest that all variants but p.Asp28Tyr have an altered conformation compared to the wild-type enzyme. Additionally, p.Asp28Tyr was only slightly more resistant to thermal denaturation.

Although p.Phe194Leu and p.Arg333Gly still have the ability to bind substrates, their binding is not as strong as that of the wild-type protein.

Their results suggested that only p.Phe171Ser and p.Leu74Pro, which are both positioned in the active site like p.Gln188Arg, severely impair enzyme activity and appear to produce a significant decrease in substrate binding ability. Those variants that are impaired in the formation of the intermediates may be more prone to protease degradation. This suggests that these mutations, especially those occurring at the active site, cause protein misfolding, and that there is compelling evidence that this is a common molecular mechanism that causes hGALT deficiency in patients [McCorvie et al., 2013]. In the same study, the molecular causes of the disability of the mutants to bind substrates and proceed with the reaction were investigated. In this context, modifying the protein sequence affects the amount of active protein by causing conformational changes that alter the active site. This compromises the substrate binding, uridylylated intermediate generation and changes the protein's overall stability.

For example, Leu74 forms some crucial hydrogen bonds with Cys120 and Tyr89 (via the backbone oxygen) and with Asn72 (via the backbone nitrogen). The mutation of leucine to proline in position 74 results in the removal of the hydrogen bond with Asn72 on the same chain. Moreover, Leu74 is flanked by Pro73 and Cys75 and results to be close to uracil moiety of UDP-galactose. This leads to a van der Waals contact between Leu74 and part of the substrate; the mutation of leucine to proline removes this contact, causing a conformation change of the active site and disability to bind substrate for enzyme.

p.Phe171Ser showed similar effects on enzyme function to p.Lys74Pro, considering its location at the active site. This residue is important because it forms a hydrogen bond with Gln188 of the same subunit, a crucial residue for the ping-pong mechanism [Mc Corvie et al., 2013]

The results of this study show that hGALT requires some degree of flexibility to function optimally, and that the impaired function of some variants is due to altered folding. Since the majority of disease-associated mutations of hGALT are not found at the active site, it has been hypothesized that protein misfolding is the most likely cause of their effects [Mc Corvie et al., 2013].

In the same study, using FTMap server (*http://ftmap.bu.edu*), McCorvie and coauthors identified a cavity between the two subunits of hGALT, and hypothesized that it probably has an allosteric function; however, no literature exists describing which residues are involved [Mc Corvie et al., 2013].

1.6 The pharmacochaperones

The term “Pharmacological chaperones”, also known as "pharmacochaperones" (PCs) was introduced to describe chemical compounds that bind specifically and stabilize target proteins. Morello and coworkers first defined with this term the action of a specific antagonist that stabilizes some mutants of the vasopressin receptor [Morello et

al., 2000]. PCs are small molecules able to bind specific target protein, which can stabilize their native conformation or even correct misfolding in proteins affected by mutations, thus rescuing their original function.

The main class of proteins target of PCs is represented by transferases, followed by transporters and receptors. These proteins can have different subcellular localisation but in the majority of cases they are situated in plasma membrane and lysosomes [Liguori et al., 2020].

There are different types of PCs, each with specific characteristics, and they can be classified into different types: competitive inhibitors, activating compounds and allosteric ligands. Despite their diversity, all of them are low molecular weight chemical molecules that have entered clinical practice for many diseases caused by protein instability [Liguori et al., 2020].

PCs are different from chemical chaperones, because they interact with proteins in a specific way. On the contrary, chemical chaperones interact with proteins in a non-specific way, binding them near the interfaces between protein domains. This induces the formation of a network of favorable weak interactions that eventually allows the different parts of a protein to be stably linked [Scafuri et al., 2022].

Since their discovery in the early 2000s, PC have been considered as candidate treatments for an increasing number of rare genetic diseases, first of all Fabry disease [Fan et al., 1999], lysosomal storage disorders [Thomas et al., 2019], cystic fibrosis [Hanrahan et al., 2017], and also phenylketonuria [Pampalone et al., 2021].

Despite the high expectations and the continuously increasing amount of studies on PC, only few drugs belonging to this class have been entered in the clinics [Scafuri et al., 2022]. However, PCs remain a promising therapeutic approach for the treatment of rare inborn errors of metabolism, caused by genetic mutations that often can destabilize the structure of the wild-type proteins expressed by that gene [Matalonga et al., 2017].

Starting from these considerations, McCorvie and co-workers proposed the use of PC as drug candidates for classic galactosemia based on the protein structural instability

caused by disease-associated mutations in the human GALT gene [McCorvie et al., 2013].

1.6.1 "First generation" PCs

The majority of PCs developed early (called "first-generation" PCs) compete with real substrates of their target proteins. In fact, these PCs are competitive inhibitors that bind to the active site of the enzyme in the folded state preferentially, stabilizing the protein [Scafuri et al., 2022].

The following examples of first generation PCs illustrate the wide diffusion and pharmacological efficacy of this type of drugs.

Migalastat, an iminosugar analog of galactose, is the first representative example of a competitive inhibitor used as a PC. It was approved by Food and Drug Administration (FDA) as the first oral therapy for Fabry disease, a X-linked lysosomal disorder with a deficit of acid alpha-galactosidase A enzyme (AGAL). In 1999, it was demonstrated that Migalastat raised the residual activity of responsive AGAL mutants (p.Arg301Gln and p.Gln279Glu) [Fan et al., 1999] and, subsequently, many others experiments have confirmed that migalastat is a potent inhibitor of AGAL. Meanwhile, migalastat has been a good starting point for the synthesis of similar glycomimetic molecules [Liguori et al., 2020]. Isofagomine, for example, is a modified iminosugar, and appeared to be very promising in stabilizing mutant of glucosylceramidase [Sun et al., 2012].

A second representative example of competitive inhibitors is *Ambroxol*, a mucolitic agent used as to treat hypersecretion and hyaline membrane disease in newborns [Maegawa et al., 2009]. Ambroxol has also been shown to be effective in treating Gaucher disease, the most frequent lysosomal storage disease caused by a deficiency of glucocerebrosidase (GCase) [Ivanova et al., 2018]. This is due to its inhibitory activity, as well as to its ability to bind and stabilize the enzyme. In particular, ambroxol has been shown to increase the activity of some mutants of GCase (p.Asn370Ser and p.Phe231Ile) [Maegawa et al., 2009]. Additionally, the usage of ambroxol has been linked to other diseases. For example, increasing the levels of GCase can lower the

level of alpha-synuclein, the protein that causes Parkinson's disease dementia [Silveira et al., 2019].

Pyrimethamine is another example of a competitive inhibitor that has been discovered to inhibit one of the three isoforms of human hexosaminidase (HexA). GM2 gangliosidosis is a rare genetic illness that causes the gradual destruction of nerve cells in the brain and spinal cord due to a genetic mutation in the gene coding for HexA [Yamanaka et al., 1994]. Pyrimethamine, a drug already approved by the FDA against malaria and toxoplasmosis [Leport et al., 1996, Weiss et al., 1992], has been identified as a chaperone in the late-onset variant of GM2 ganglosidosis [Tropak et al., 2007, Beck et al., 1998].

There are many other examples of competitive inhibitors, such as lumacaftor for cystic fibrosis. Lumacaftor binds to an anion channel (CFTR) expressed at the apical surface of secretory epithelia in the pancreas, intestine, exocrine glands, and lungs. Mutations in CFTR have a major and dangerous effect on all of these organs [Carlile et. al., 2018]. Ezetimibe and pranlukast bind to and increase the activity of N-acetylgalactosamine-6-sulfate sulfatase (GALNS), which is mutated in a rare disease called Morquio A syndrome (a mucopolysaccharidosis IVA) [Alméciga Diaz et al., 2019].

1.6.2 "Second generation" PCs

Although the study and use of first-generation PCs is increasing, they present several problems that prevent them from becoming widely diffused drugs. The first real problem lies in their nature: they are competitive inhibitors that, by binding to the target enzyme, do not allow the true substrate to bind, thereby compromising the biological mechanism [Scafuri et al., 2022]. The second issue is that each genetic disease is defined by a variety of distinct mutations (many of which are private), but only a small percentage of these variants respond to PCs.

For these reasons, a second generation of PCs has more recently been developed, molecules that specifically bind to a different position than the active site of the target enzyme [Scafuri et al., 2022]. Interacting with non-catalytic domains allows them to

avoid competition with the substrate and to expand the spectrum of responsive mutations [Parenti et al., 2015].

The second generation PCs are considered activators compounds and below are described some representative examples.

The first example is *glucose-1,6-bisphosphate,* which activates the phosphomanno-mutase 2 enzyme (PMM2). The main activity of PMM2 *in vivo* consists of the isomerization of mannose-6-phosphate into mannose-1-phosphate, which is activated and eventually introduced into glycans [Citro et al., 2018]. A mutation in the gene coding for this enzyme causes PMM2-CDG disease, the most common form of congenital N-glycosylation pathology. The disease, characterized by cerebellar dysfunction, abnormal fat distribution, strabismus and hypotonia, has a highly variable clinical picture: some adults only have a mild form of the disease, while some die in the first year of life, and other are asymptomatic carriers [Lam et al., 2021].

Another example of PCs that binds to the active site and stabilizes mutants in cells is *11-cis-retinal*, which is a natural cofactor of rhodopsin. This activator promotes the cellular folding of this protein, which is associated with the autosomal dominant disease retinitis pigmentosa, a slowly progressive and bilateral degeneration of the retina and retinal pigment epithelium characterized by choroidal neovascularization and macular edema, which eventually leads to irreversible loss of central vision [Noorwez et al., 2004].

Tetrahydrobiopterin, the natural cofactors of phenylalanine hydroxylase, is a further example of second generation PCs in the treatment of a rare form of phenylketonuria, called mild hyperphenylalaninemia, an inborn error of amino acid metabolism, with or without clinical manifestations of impaired cognitive function, and behavioral and developmental disorders [de Baulny et al., 2007]. In particular, tetrahydrobiopterin treatment led to normal or nearly normal blood phenylalanine concentrations in most patients with residual phenylalanine hydroxylase activity, suggesting that

responsiveness to tetrahydrobiopterin is a common feature of mild hyperphenylalaninemia phenotypes [Muntau et al., 2002].

The examples given above are meant to show how much second generation PCs are developing and how promising they are as potential therapies for pathologies that currently do not have a cure.

As reported above, the second generation PCs bind specifically to a different position than the active site in the target enzymes, and when this different position is in an allosteric region, these PCs are called allosteric non-inhibitory PCs. There are several molecules that act as allosteric non-inhibitory PCs, such as 2,6-dithiopurine which stabilizes AGAL in Fabry disease, as extensively described above. This molecule preferentially binds to an AGAL region identified as an allosteric hot spot for ligand binding [Citro et al., 2016].

Erythropoietic porphyria is characterized by severe skin photosensitivity that may lead to scarring, blistering, and increased hair growth on the face and back of the hands. It is a condition caused by an overproduction of porphyrins as a result of a deficiency in the enzyme uroporphyrinogen III synthase (UROS). In silico docking was used to identify an allosteric binding site on the surface of the enzyme. 2500 diverse chemical fragments were looked for among ligands. The authors discovered that the antifungal ciclopirox could help to stabilize UROS. In particular, ciclopirox binds to UROS in an allosteric location, away from the active site, with no effect on the enzyme's catalytic activity [Urquiza et al., 2019].

1.6.3 Computational strategies for the identification of PCs

Computational technologies (computer-aided drug design) have revolutionized the drug discovery process by reducing time and costs [Feinstein, 2016]. In a general workflow for PCs development, it is possible to identify the different stages of drug discovery that a particular compound may undergo, which can benefit from a computational approach (Figure 1.17) [Scafuri et al., 2022].

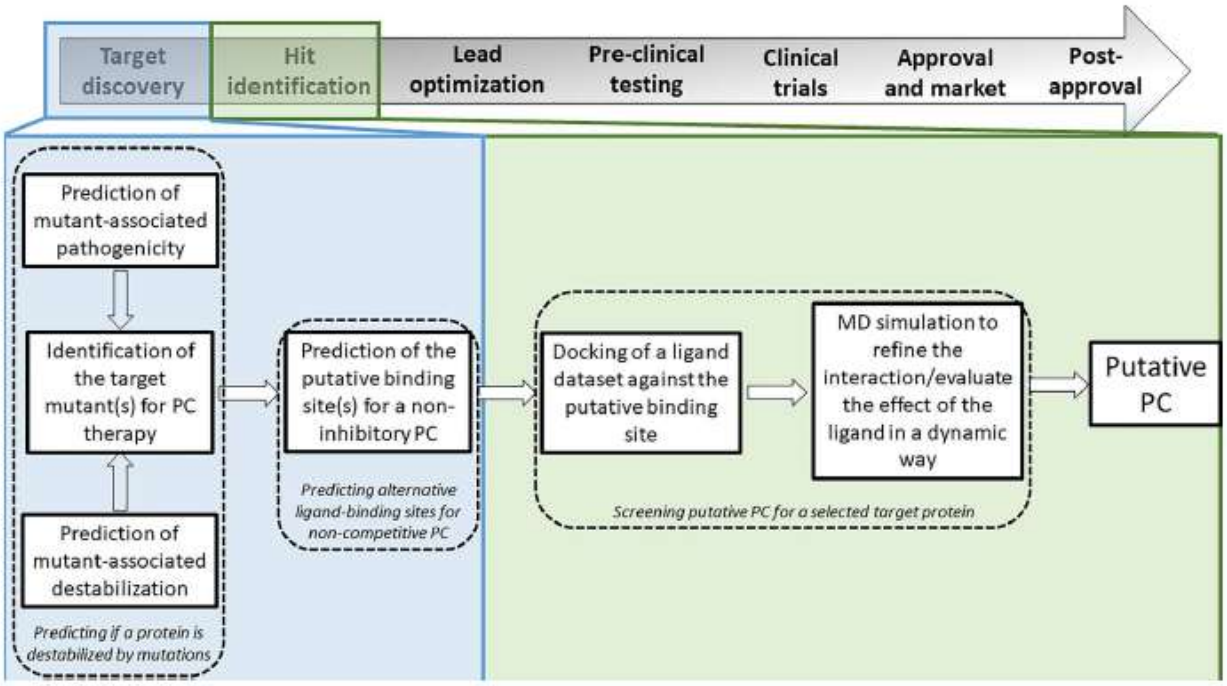


Figure 1.17: Workflow showing the general roadmap for PCs development, with a highlight on where computational solutions can be inserted to improve the discovery of a putative PC [Scafuri et al., 2022].

The first step is the target discovery, which consists of identifying mutant targets for PCs therapy. The protein responsible for the disease must have mutations that destabilize the protein itself. In most cases, the mutations are numerous and have multiple genotypes and phenotypes, which makes the selection more complex. The pathogenicity of amino acid substitutions can be predicted by studying the effect of the substitutions on the structure, function, and stability of proteins. This requires computational tools that are currently available [Scafuri et al., 2022].

The second step, which can also be done contemporaneously to the first, is to identify binding sites for the second generation PCs. As discussed above, the recent trend is to identify PCs that bind to alternative protein sites, different to the active site of the enzyme [Parenti et al., 2015]. However, even when the 3D structure of the protein of interest is available, it is not trivial to find those sites. A plethora of different computational tools and approaches have been developed in recent years for this scope, some of which are accessible from the Web (table 1.2).

Name	URL	Principle	Reference
CASTp	http://sts.bioe.uic.edu/castp/	Grid-based geometry	[Binkowski et al., 2003; Tian et al., 2018]
COACH-D	https://yanglab.nankai.edu.cn/COACH-D/	Consensus; Support Vector Machine	[Wu et al., 2018]
3DLigandSite	https://www.wass-michaelislab.org/3dlig/	Template-based method	[Wass et al., 2010]
FunFold2	https://www.reading.ac.uk/bioinf/FunFOLD/	Template-based method	[Roche et al., 2013]
DeepSite	https://playmolecule.com/deepsite/	Template-based methods, neural networks	[Jiménez et al., 2017]

Table 1.2: Web accessible predictors of protein cavities and ligand binding sites [Scafuri et al., 2022].

Different approaches to identify active sites in biomolecules exist, each based on different methods. The oldest of these approaches are geometry-based methods, either grid-based or probe-based. In the first case the molecule is positioned in a 3D Cartesian grid and some geometric conditions must be satisfied for a pocket to be identified. In the second case, the pockets are identified by the probe spheres that are tangent to the surfaces of two atoms of the biomolecule [Scafuri et al., 2022].

The next step is to screen for putative PCs at the possible site(s) previously identified. This step requires a virtual screening of large compound libraries. This screening serves two fundamental purposes: to reduce the number of compounds to be evaluated experimentally and to expand the chemical diversity of the compounds that are preliminaryly evaluated [Shaker et al., 2021]. To date, millions of chemical compounds are available in many databases, either public or owned by private companies (table 1.3) that can be freely accessed or used under request to perform virtual screening.

	Name	URL
freely available	ZINC	https://zinc20.docking.org/
	PubChem	https://pubchem.ncbi.nlm.nih.gov/
	DrugBank	https://go.drugbank.com/
	ChEMBL	https://www.ebi.ac.uk/chembl/
	e-Drug3D	https://chemoinfo.ipmc.cnrs.fr/MOLDB/index.php
	BindingDB	https://www.bindingdb.org/bind/index.jsp
private companies	Asinex	http://www.asinex.com/
	Selleckchem	https://www.selleckchem.com/
	Reaxys	https://www.reaxys.com
	KNApSAcK	http://www.knapsackfamily.com/KNApSAcK/

Table 1.3 Web accessible databases of chemical compounds [Scafuri et al., 2022].

In the simplest case where both the structure of the compounds and the target protein are available, the best method for predicting the binding affinity between ligand-protein is molecular docking [Kitchen et al., 2004]. The development of molecular docking programs dates back to 40 years ago [Kuntz et al., 1982], and today there are several programs: ICM [Totrov and Abagyan, 1997], FlexX [Kramer et al., 1999], Glide [Friesner et al., 2004; Halgren et al., 2004], GOLD [Verdonk et al., 2003], MDock [Huang et al., 2007], MOE [Vilar et al., 2008], AutoDock [Morris et al., 2009], AutoDockVina [Trott and Olson 2010], DOCK [Anderson et al., 2005], to mention only some of the most popular ones.

The use of pharmacophoric models is undoubtedly one of the most well-known approaches for drug research. The term "pharmacophore" is defined as *"An ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response"* [Balakumar et al., 2018]. Therefore, it is an abstract concept that explains the interaction capabilities of a compound towards its target structure rather than an actual molecule or association of functional groups [Wermuth et al., 1998]. These models simulate phenomena (physical, chemical, and biological) that occur during drug-receptor interactions [Drwal et al., 2011].

In drug design, two types of pharmacophore modeling techniques are used: ligand-based pharmacophore modeling and structure-based pharmacophore modeling. If the

target protein structure is not known, the ligand-based pharmacophore modeling approach is used to create new chemical entities; otherwise, structure-based pharmacophore modeling can be used when the 3D structure of the macromolecule of interest (typically a receptor or enzyme with or without a bound ligand) is available [Balakumar et al., 2018].

In a 3D pharmacophoric model with both chemical and functional interactions, the spatial arrangement of the chemical features represents the ligand's interactions with the receptor, and the pharmacophoric pattern generated represents the binding mode of those ligands that bind to the same target in the same way [Wolbere al., 2006]. Each pharmacophoric model has specific properties that correspond to the observed interactions in the drug-receptor complex: hydrogen bond acceptor, hydrogen bond donor, cation, anion, aromatic, and hydrophobic features The knowledge of these features helps in the screening of the databases and in the identification of target molecules for the following steps.

The last step to find putative PCs for a selected target protein consists in two phases: docking of a ligand dataset against the putative binding site and molecular dynamics (MD) simulation to refine the interaction and to evaluate the effect of the ligand in a dynamic way.

Standard docking protocols cannot handle the entire flexibility of the protein. However, there are two main ways to include protein flexibility; in both it is possible to study the best accommodation of the ligand into the cavity. The first way allows a limited conformational variability of the residues in close contact with the ligand on the binding site. The second way consists in creating a set of alternative receptor conformations to simulate the protein conformational changes [Wong et al., 2021].

Nevertheless, in both ways it is not possible to identify the effect induced by the binding of the ligand on the overall structure of the protein. For this reason, MD simulations are necessary to perform a complete study about the conformational flexibility of protein and to allow to perform an accurate prediction of the binding free

energy of the ligands [Graff et al., 2021]. There are many popular programs for biomolecular MD simulations, including GROMACS [Abraham et al., 2015a], NAMD [Phillips et al., 2021], AMBER [Case et al., 2005], CHARMM [Jo et al., 2008], Desmond [Bowers et al., 2006], among others.

In conclusion, to search any PC with a computational approach it is necessary to apply docking and MD simulations. In both cases, the choice of which program to use is highly problematic because it may depend on many factors, such as speed, accuracy, and free availability. However, both programs (for docking and MD simulations) must be set up on the system to be studied, obtaining the best experimental conditions in order to have results that are as accurate as possible [Scafuri et al., 2021].

1.6.4. Arginine as a possible PC for GALT: testing the hypothesis

It is known that the amino acid arginine is a stabilizing agent for poorly folded proteins, preventing their aggregation and cellular accumulation. As a result, it has already shown a beneficial effect in some hereditary metabolic disorders [Berendse et al., 2013; Silva et al., 2017]. In 2014, Coelho and coauthors showed that the increased tendency of several GALT mutants to aggregate, associated with protein misfolding, could be a pathogenetic mechanism in classic galactosemia [Coelho et al., 2015b].

In particular, they focalized their studies on functional and structural impact on the most frequent variations in classic galactosemia: p.Gln188Arg, p.Ser135Leu, p.Lys285Asn, and p.Asn314Asp. Based on this study, the most surprising and novel observation was that most of the variants, particularly relevant for p.Gln188Arg, show disturbed aggregation profiles, despite the absence of detectable structural effects on their secondary and tertiary structures. This is an important observation because, at the cellular level, the accumulation of aggregation-prone proteins can seriously interfere with cellular homeostasis [Coelho et al., 2015b]. Moreover, two studies, one based on increased ER stress in p.Gln188Arg homozygous patients [Slepak et al., 2007] and the other one based on increased unfolded protein response in GALT-null galactosemia model [De-Souza et al., 2014] also suggested that there is a basal level of protein

homeostasis disturbance associated with galactosemia. The results of Coelho and coauthors in 2014 suggested that GALT aggregation associated with protein misfolding could be an important pathogenetic mechanism in classical galactosemia, laying the groundwork for future studies on GALT aggregation *in vivo* [Coelho et al., 2015b]. This led the same group to successfully test the ability of arginine, an amino acid known for its activity as a protein stabilizer with an anti-aggregation effect [Coelho et al., 2015b], in improving the activity of GALT mutants, including p.Gln188Arg. In particular, galactose-sensitive prokaryotic models were developed with a twofold aim: to evaluate the negative effect of mutations on the one hand and the possible positive effect of arginine on the other. These models were deprived of their endogenous *GALT* gene, and added with the human *GALT* genes carrying different mutations, including p.Ser135Leu, p.Gly175Asp, p.Gln188Arg, p.Arg231Cys, p.Arg231His, p.Lys285Asn and p.Asn314Asp. After the addition of galactose, the p.Ser135Leu, p.Arg231His and p.Asn314Asp mutants showed no growth arrest (with a growth rate comparable to that observed in the presence of the control medium). On the other hand, p.Gln188Arg, p.Arg231Cys, and p.Lys285Asn cultures showed impaired growth, with different levels of galactose toxicity. Subsequently, the cultures were treated with arginine. In the p.Gly175Asp, p.Gln188Arg and p.Lys285Asn mutants, a significant improvement in activity was observed, underlined by an increased ability to cope with galactose-induced toxicity (Figure 1.18) [Coelho et al., 2015b].

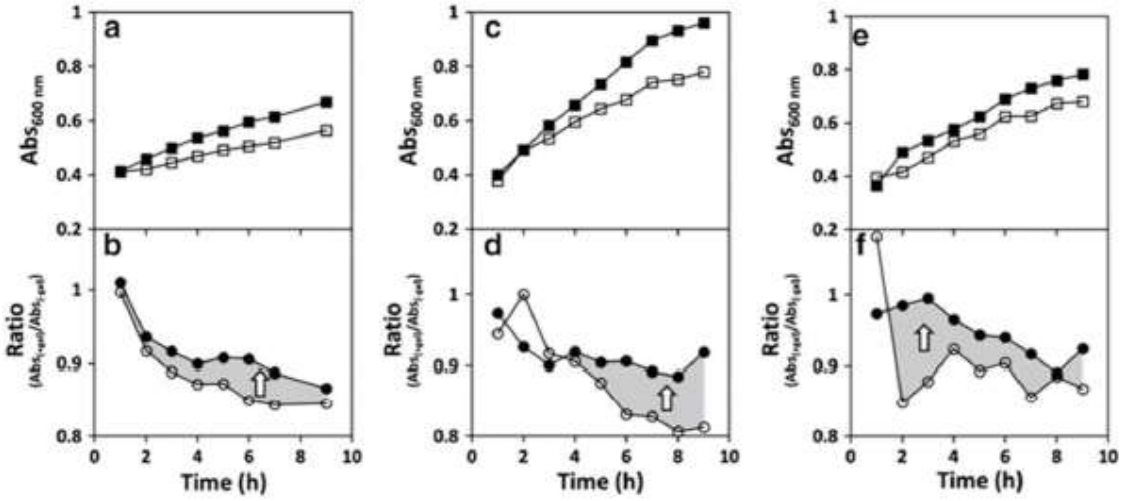


Figure 1.18: Arginine improves the function of p.Gln188Arg (panels a and b), p.Lys285Asn (panels c and d), and p.Gly175Asp (panels e and f) hGALT in prokaryotic models of the disease.
Panels a, c, and e depict the growth profiles of *Escherichia coli ΔgalT* expressing p.Gln188Arg, p.Lys285Asn and p.Gly175Asp, respectively, in the absence and in the presence of galactose. Panels b, d, and f depict the ratio curves for bacteria expressing, respectively, p.Gln188Arg, p.Lys285Asn, and p.Gly175Asp hGALT, in the presence or absence of galactose (black circles, in the presence of 25 mM arginine; hollow circles, absence of arginine). The gray-shaded areas and the white arrows depict the effect of arginin in improving the ability of these variants to alleviate galactose toxicity, highlighted by white arrows. [Coelho et al., 2015b]

This study revealed that arginine shows a mutation-specific beneficial effect, in particular on two of the most widespread pathogenic variants, which lays the foundations for further studies, since arginine could have a great therapeutic impact against classical galactosemia.

After this promising results, Haskovic and coworkers conducted in vivo studies on four galactosemic patients to evaluate the effect of arginine in galactose metabolism, conducted in vitro studies with three fibroblast cell lines derived from classic galactosemia patients as well as recombinant protein experiments. Patients were treated with arginine aspartate (in the commercially available form of Asparten®) in a dose of 15 g/day, for one month. Patients did not show a significant improvement in whole-body galactose oxidative capacity, which remained the same before and after arginine aspartate administration (Figure 1.19 A). GALT activity analysis in red blood cells (RBC) revealed no statistically significant difference after treatment compared to

baseline. Galactose metabolite concentrations did not significantly change (Figure 1.19 B) [Haskovic et al., 2018]. Thus, it was deduced that, at least for people carrying this mutation, arginine has no potential therapeutic role.

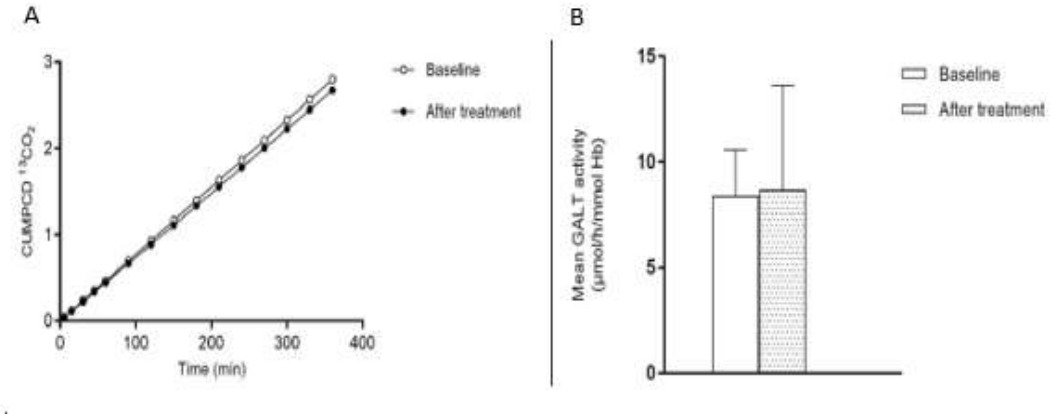


Figure 1.19. A. Mean galactose oxidative capacity before and after Asparten® supplementation; x axis: time in minute, y axis: CUMPC and CO_2 eliminated in air. B Mean GALT activity is expressed as µmol of UDP-Gal formed per hour per mmol hemoglobin [Haskovic et al., 2018]

However, neither study was an in-depth analysis of the molecular interactions between the GALT enzyme and arginine. Moreover, the number of patients recruited in the clinical trial was very small (only four), and the short duration of the study (1 month) was insufficient for evaluating the effects of arginine on long-term clinical outcomes. Furthermore, a single dose (15 g/day) was tested during the clinical trial, and the authors argued that the discrepancy with the results of the prokaryotic model could be due to the higher concentration of arginine used for those experiments. Therefore, the molecular aspects of the putative interaction between arginine and GALT enzyme remain unclear.

1.7 Objectives of the thesis

The goal of the project was to identify potential PCs for one of the most common and debilitating mutations found for GALT, p.Gln188Arg [Timson, 2016].

This project was structured in several stages. First of all, since in the past and more recently, the structural effects of this mutation were deduced from the static structure

of the wild-type human enzyme, we felt that a dynamic view of the protein was necessary to deeply understand its behavior and obtain tips for possible therapeutic interventions. Thus, we performed MD simulations of both wild-type and p.Gln188Arg proteins in the absence or the presence of the substrates in different conditions. On the other hand, the first evidence of a PC's potentially positive role against the p.Gln188Arg mutation was delivered through research on arginine amino acid [Coelho et al., 2015b], a known suppressor of protein aggregation [Baynes et al., 2005]. Arginine, however, showed a positive effect only in a prokaryotic model [Coelho et al., 2015b], but no therapeutic effect was found after one month of the administration, for patients affected by the p.Gln188Arg variant in homozygosity [Haskovic et al., 2018]. The lack of information about molecular interactions of this amino acid with respect to the protein prompted us to investigate its binding in the active site and central cavity of both wild type and p.Gln188Arg mutant, either in the presence or in the absence of the substrates G1P and H2U.

In this respect, the aim was to evaluate the efficacy of this amino acid in restoring the correct folding of the mutated enzyme and therefore its function, maintaining the bond with the enzyme over time. First of all, the binding of arginine to the enzyme was simulated by molecular docking. The representative conformations obtained by docking simulations were used as starting point for the following MD simulations at two different temperatures: 310 K, corresponding to the normal body temperature, and 334 K, a higher temperature to induce protein destabilization.

In the second part of the project, we wondered if there might be an allosteric site in GALT enzyme, as suggested by the literature [McCorvie et al., 2013], and if this allosteric site could be used as a target to develop new PCs for this enzyme.

Through a computational predictor, we identified an allosteric site, corresponding to the portion of the protein interacting with arginine.

The identification of the allosteric site occurred simultaneously with the search for new PC candidates for hGALT.

A possible interaction between PCs, selected from drugs already in therapeutic use for different diseases, and hGALT (both wild type and mutant), was simulated by molecular docking on both the active site and the central cavity. The next step was to proceed with the search for pharmacophores, starting from the best docking conformations. This led to the identification of new ligands, which were selected for further docking on the allosteric site. Different ligands have been shown to provide promising results both in "in silico" and in preliminary "wet" experiments.

Finally, the MD protocol was also improved to find the best experimental conditions to set up further MD studies focused on the interactions between hGALT system and these new ligands, and on the search of allosteric pathways in hGALT, starting from MD simulations of greater length.

2. MATERIALS AND METHODS

2.1 Databases

I will describe in the following paragraph the databases used for the present work, some of which have been developed *in house*.

2.1.1 Protein Data Bank

The Protein Data Bank *(https://www.wwpdb.org/)* was founded in 1971 as a free, public database containing the crystallographic coordinates of biological macromolecules (proteins, nucleic acids, carbohydrates, and a variety of complexes) [Berman et al., 2000]. It is one of the oldest scientific databases, having been established at the Cambridge Crystallographic Data Centre in 1965 [Attwood et al., 2011]. The Research Collaboratory for Structural Bioinformatics (RCSB) has been fully responsible for its management since 1 July 1999 [Bhat et al., 2001] and still provides a portal (*https://www.rcsb.org/*) for the data access.

There were only seven protein structures when it was founded, and the number has steadily increased since 1980, to around 200,000 today (last access: 31 October 2022). The structures are obtained by various experimental methods: X-ray crystallography for the majority (166,649), nuclear magnetic resonance (NMR) for more than ten thousands (13,653), and the remaining by other methods such as cryo-electron microscopy (cryoEM) which has considerably expanded in the last years [Callaway 2020].

Presently, in addition to experimentally-determined macromolecular structures, RCSB.org now offers access to ~1 million Computed Structure Models (CSMs) from AlphaFoldDB [Jumper et al., 2021 and Varadi et al., 2021] and RoseTTAFold (from Model Archive *(https://www.modelarchive.org/*).

Certainly, the number of macromolecular structures has increased since the beginning, but the rate of growth in the number of structures remains slow. Furthermore, there is

a high level of redundancy in the PDB that demonstrates the great difficulty in obtaining structural data on new proteins.

The resources in PDB can be accessed freely using the available tools, which are personalized to the various needs of the user community [Di Costanzo et al., 2016]. The 3D structure can be graphically represented, and there are links to other databases, such as UniProt (the main protein sequence database) [UniProt Consortium, 2021]. Alternatively, the structure coordinates file can be downloaded from the website in the PDB format. This is a universally recognized format used by all molecular graphics programs, in which the positions of all the macromolecular atoms in the space are represented by their cartesian coordinates. Identified by a unique four-character code (one number and three alphanumeric characters) and divided into lines (records) and 80 columns, the pdb file appears as a text file divided into two fundamental parts.

There is a series of useful information in the first part (HEADER), such as the type of macromolecule, the reference organism, various parameters used to determine the quality of the macromolecule structure (resolution, R-value, and R-free), experimental details, cell parameters, missing atoms or residues. The Cartesian coordinates of the atoms of the macromolecule and of any ligands/cofactors are marked with ATOM and HETATM, respectively. For crystallographic structures, these records include also the occupancy factor, which varies from 0.0 to 1.0 and can be lowered if the atom spends only a fraction of the time in the position identified by the Cartesian coordinate, and the B-factor, which represents the displacement of atoms from their average position identified by the Cartesian coordinates. If an anisotropic B-factor is present, the record ANISOU is added below each record ATOM.

This data bank was used to download the hGALT structure with PDB code 5IN3, solved by X-ray crystallography, selected for this project among the other structures of GALT enzyme available [McCorvie et al., 2016].

2.1.2 Galactosemia Proteins Database 2.0

Galactosemia Proteins Database 2.0 [d'Acierno et al., 2018], *(https://proteinvariants.eu/galactosemia)* is a database developed *in house* that collects information about the structural and functional effects of the variants of the four enzymes of the galactose metabolism. It is accessible to the entire scientific community and developed at the Institute of Food Sciences of the National Research Council of Avellino, in partnership with the Department of Chemistry and Biology "A. Zambelli" of the University of Salerno.

This database has been organized into two main areas: a common general part in which there is a description of the disease, a list of several useful resources (such as scientific and curated databases) and a list of associations or no-profit organizations devoted to galactosemia; and an enzyme-specific part that provides the access to the information about all the enzymes of the Leloir's pathway. It collects a set of theoretical models of the variants associated to the different forms of galactosemia, as well as tools for searching and visualizing the results of analyses performed on the models, allowing people to investigate the structural and functional effects of variations.

The model of the 3D structure of the human wild-type hGALT, which was generated from the 5IN3.pdb crystal structure, as well as the p.Gln188Arg model, have been downloaded from this database.

2.1.3 Databases of small molecules

PubChem, ZINC[15], and DrugBank were the databases of small molecules used for this project.

PubChem *(https://pubchem.ncbi.nlm.nih.gov/)* is a public and free database, hosted at the National Institutes of Health (NIH), which contains structures of small organic molecules, as well as information about their chemical and physical properties, biological activities, patents, health, safety, and toxicity data [Kim et al., 2019]. Small molecules predominate in PubChem, but larger molecules such as nucleotides, carbohydrates, lipids, peptides, and chemically modified macromolecules are also

present. *Substance* (compound information), *Compound* (structures), and *BioAssay* (bioactivity data) are the three main sections of PubChem. The structure of interest can be downloaded from the Compound section in a variety of formats, including sdf, json, xml, and asn.1, in both 2D and 3D formats.

ZINC[15] (*http://zinc.docking.org*), developed in 2004 by the Department of Pharmaceutical Chemistry at the University of California, San Francisco (UCSF), is a free database that contains the 3D structures of millions of commercially available compounds, in standard, ready-to-use formats (mol2 and sdf). In the *'Substances'* section, a quick search allows to enter the name of a known compound of interest and obtain its 3D structure in the most appropriate form for a specific purpose, such as molecular docking simulations, which is the database's original focus. Compounds can also be found using designed chemical structure, biological activity, physical properties, similarity to a starting compound, predictions or annotations to a specific target, and other features [Sterling and Irwin, 2015].

The structures of arginine and of the possible PCs for hGALT were retrieved, respectively, from ZINC[15] and PubChem.

Another Web-based database that contains detailed molecular information about drugs, their mechanisms, interactions, and targets is DrugBank (*https://go.drugbank.com/*). It also includes data on the effects of hundreds of drugs at metabolite levels (pharmacometabolomics), gene expression levels (pharmacotranscriptomics), and protein expression levels (pharmacoproteomics) [Wishart et al., 2018]. It has been first described in 2006, has evolved over the past 12 years in response to significant improvements in web standards and changing needs for drug research and development.

It is possible to write the name of the molecule in the search section and instantly get a result structured in several sections, the main ones being two:

- *a general section* in which the ligand is described in the details, including the summary, the generic name, the background (drug description), the type of molecule with structure, weight and chemical formula;
- *a section dedicated to pharmacology*, including the associated therapies, any contraindications, the mechanism of action, the metabolism, the eventual toxicity or pharmacogenomic effects.

The DrugBank database is extensively used by drug researchers and developers, pharmaceutical companies, the medical community, and the general public. It has been used here for the selection of the PCs for hGALT.

2.2 Programs and tools for molecular structures visualization and manipulation

There are numerous programs available, both commercial and free, for visualizing the structures of macromolecules such as proteins and possibly associated small ligands.

The goal of using these programs is always to gain a better understanding of the conformational characteristics of molecules of biological interest and their interactions.

All these programs read the coordinate portion of the pdb file and allow images to be manipulated, rotated and translated, zoomed in and out, and displayed in different colors and representations.

2.2.1 PyMOL

PyMOL (*http://www.pymol.org*) (Schrödinger, Inc.) is a molecular graphics program created in 1998 by the US biophysicist Warren Lyford DeLano and released to the scientific community in 2000. This program supports several functions, including 3D molecule visualization and molecule motion, such as translation, rotation, and magnification [Rigsby et al., 2016].

PyMOL was used to visualize the 3D structures of hGALT and its variants and manipulate them with appropriate modifications, as well as to visualize the results of MD simulations.

2.2.2 UCSF Chimera

Chimera *(https://www.cgl.ucsf.edu/chimera/)* is a free and highly extensible molecular graphics program [Pettersen et al., 2004]. Accepted file formats include the pdb and mol2 formats. This program supports a wide range of functions, including molecular structure analysis, electronic density map visualization, and sequence alignments. It is also possible to make structural changes to the various elements loaded within it.

UCSF Chimera has been used in the current project to minimize the structure of some ligands used for docking and to restore the normal phosphate group of H2U in the structure of hGALT and p.Gln188Arg, as it will be explained in the following chapter.

2.2.3 BIOVIA Discovery Studio

Discovery Studio (Dassault Systèmes BIOVIA, San Diego, 2015) is a comprehensive software for analyzing and modeling molecular structures and sequences. Its many features include input file editing and receptor-ligand interaction analysis. In this last case, the types of interactions that can be graphically represented are conventional hydrogen bond, van der Waals, π interactions and covalent bonds. In addition, it can be used to assess the receptor surface's hydrophobicity, charge, solvent-accessible and non-solvent-accessible area, and ionisability (acid-base).

In this project, Discovery Studio has been primarily used for protein-ligand interaction analysis and for the generation of pharmacophoric models (Discovery Studio-Interaction generation protocol) [Vuorinen and Schuster, 2015], by means of the HypoGen algorithm [Meslamani et al., 2012], which attempts to automatically identify a 3D spatial arrangement of chemical features common to the training molecules, using a variety of sources of information, including gene expression data, protein-protein interactions, and chemical databases. The HypoGen algorithm ranks potential drug targets based on their potential to treat the disease. HypoGen consists of 3 main stages:

construction, stabilization, and optimization. In the first step, pharmacophores with common characteristics to the identified bioactive molecules are generated, and then all pharmacophoric conformations with a maximum of five default characteristics are collected. Only those pharmacophores best suited to bioactive molecules will be available at the end of this phase. Pharmacophores mapping inactive molecules are removed during the stabilization phase. Finally, the pattern collection is optimized using the simulated annealing algorithm [Vuorinen and Schuster, 2015]. As result, only the pharmacophoric models with higher scores are shown as output [Khedkar et al., 2007].

Discovery Studio Interaction Generation protocol consists of two main steps. In the first, pharmacophoric features are identified: hydrogen bond acceptor (A), hydrogen bond donor (D), cation (P), anion (N), aromatic system (R), and hydrophobic (H); in the second, all features that do not correspond to protein-ligand interactions are eliminated. In this way, Discovery Studio creates all possible combinations of pharmacophoric models and classifies them basing on decreasing selectivity score. The selectivity of a pharmacological model depends on the number and types of features and their 3D arrangement.

Among the generated pharmacophore models, the best one (which generally corresponds to the first one) is selected and used for the last step, i.e. searching for pharmacophoric hits in a selected database. In particular, the ’Search 3D-database’ function provides a rapid filtering of the input database, selecting those hits that match the minimum amount of features defined or not by user. In the latter case, the default setting is one. A hit list is finally provided in the form of a table, presented in descending order of fit value, with associated information for each compound (see paragraph 2.6.4).

2.3 Programs for protein cavity identification.

Cavity identification programs allows to identify cavities formed because of the specific folding of the amino acid chain, corresponding to sites of the protein that may play important functional roles.

2.3.1 CASTp

The Computed Atlas of Surface Topography of Proteins 3.0 (CASTp 3.0) (http://sts.bioe.uic.edu/castp/index.html) is a Web-server that lets users to locate, delineate, and measure the cavities within the protein structure, by means of a geometric approach. By uploading the pdb file of the molecule of interest, the program returns a graphic representation of the cavities. Moreover, it displays volumes of cavities and channels, topographical features of specific assemblies found in PDBs, secondary structure information, functional sites, site variants, and other annotations on protein residues [Tian et al., 2018].

The two starting structures of hGALT and p.Gln188Arg were analyzed with the CASTp 3.0 Web server (*http://sts.bioe.uic.edu/castp/calculation.html; last accessed 5 October 2021*) [Tian et al., 2018], to identify the possible cavitie(s) to perform docking with the ligand arginine and other potential PC ligands.

2.3.2 FTMap

To identify potential allosteric site(s) of hGALT, a computational mapping server, FTMap (*http://ftmap.bu.edu*), was used.

FTMap is a computational mapping server that identifies binding regions on the surface of macromolecules, called *hotspots* [Brenke et al., 2009]. One key property of hotspots is their ability to bind with high affinity small molecules with size similar to drugs [Owens, 2007].

The server is based on a combination of three methods: i. Conformational analysis of the small molecule and the protein; ii. Electrostatic potential analysis; iii. Hydrogen bonding analysis [Kozakov et al., 2015]. The only input required by FTMap is the structure of either a protein, or DNA, or RNA, obtained by X-ray crystallography or

NMR techniques. The input can be provided by giving a PDB ID, or by loading a structure in PDB format [Kozakov et al., 2015].

In details, the FTMap algorithm consists of five steps: *i. rigid body docking of probe molecules.* The server distributes small organic probes of various sizes, shapes and polarity (ethanol, isopropanol, isobutanol, acetone, acetaldehyde, dimethyl ether, cyclohexane, ethane, acetonitrile, urea, methylamine, phenol, benzaldehyde, benzene, acetamide, and N,N-dimethylformamide) in the 3D structure of the macromolecule to be analyzed. For each probe, billions of docked conformations are sampled by a rigid body docking step. At the end, the 2000 best poses for each probe are retained for further processing; *ii. minimization and rescoring:* the free energy of each of the 2000 complexes, generated in Step 1, is minimized using the CHARMM potential with the Analytic Continuum Electrostatic (ACE) model [Brooks et al., 1983]; *iii. clustering and ranking:* the minimized probe conformations from step 2 are grouped into clusters using a simple greedy algorithm, excluding clusters with less than 10 members. The clusters are ranked on the basis of their Boltzmann averaged energies, and for each probe, 6 clusters with the lowest average free energies are retained [Ruvinsky and Kozintsev, 2006]; *iv: determination of consensus sites (CSs):* the groups of different probes are clustered using the distance between the centers of mass of the cluster centers as the distance measure; *v: characterization of the binding site:* first the largest CSs is selected because it is the most important subsite. Moreover, additional CSs are identified expanding the binding site by adding any CS (irrespective of its size) within 7 Å from any CS already in the binding site, and this procedure continues until no further expansion is possible [McDonald and Thornton, 1994].

At the end, a hotspot can be obtained and it is considered “druggable” if it fulfills two conditions:

- Consensus site strength (S), defined as the number of probes within the cluster. A consensus site with $S \geq 16$ will be druggable; a consensus site with $S \leq 13$ is not druggable due to very weak binding sites;

- Distance between consensus sites: the distance of a consensus from a secondary consensus site or another hotspot by a maximum of 8 Å [Kozakov et al., 2015].
Through this method, a "possible allosteric site" for hGALT was identified, also in agreement with data from MD simulations of arginine.

2.4 Molecular docking

Docking is a set of methodologies that simulate the interactions between two molecular entities (protein/protein or protein/ligand) in order to identify the best molecular complex in terms of energy [Morris et al., 2008]. These interactions may involve two similar or different molecules that play key roles in different biochemical and cellular pathways. This is of great relevance especially from a pharmaceutical point of view.

2.4.1 Basic concepts for molecular docking

To perform a docking, the interacting structures must first be complete, and all possible interactions between them must be identified and evaluated energetically. There are two main challenges: managing the possible number of interactions, as the size of the interacting entities increases, and taking into account the possible conformational variability of proteins in the absence or presence of a ligand.
There are several types of docking:
- *rigid-body docking*, in which the two interactors are kept in fixed conformations [Shoichet and Kuntz, 1991];
- *rigid-protein docking*, in which the conformation of the largest interactor (typically the protein) is kept fixed, while the conformation of the smaller interactor (typically the ligand) is varied in order to have as many conformations as possible [Banitt et al., 2011];
- *flexible docking*, where both interactors can explore the entire conformational space of their interaction, as they are both considered flexible bodies [Koshland, 1963];

- *covalent docking*: a possible method for simulating covalent ligand-protein interactions [Bianco et al., 2016].

Docking consists essentially of three steps:

i. ***Initial exploration***, in which the properties of the interactors are evaluated and compared with any known similar complexes. A docking simulation requires high quality atomic resolution structures of both interactors in order to produce reliable results, since the interactions that are observed involve single atoms. Furthermore, in some crystallographic structures, there may be alternative positions of observed atoms. In these cases, both alternatives must be tested during the docking simulations [Morris and Lim-Wilby 2008];

ii. ***Selection of the conformation of both interactors and docking***, this step refers to the simulations that enable to identify the most favorable conformations of the interactors. In order to study the conformational variability of the ligand relative to its receptor, algorithms have been developed that allow to explore comprehensively and efficiently the different "poses" that the ligand can assume;

iii. ***The refinement phase***, in which specific contacts in the complexes are analyzed, assessing the volume occupied by the ligand at the binding site and distinguishing similarities and differences with other ligands of the same class.

For docking, it is necessary to use an algorithm to perform the searching phase, and a scoring function to rank the results. The search methods can be divided into two main categories: systematic and stochastic. The outcome of a systematic search is deterministic, but the quality of the solution depends on the granularity of the sampling of the search space. These methods are commonly used in rigid protein docking. Stochastic methods that rely on an element of randomness are more suitable for higher-dimensional problems, such as flexible ligand–protein docking, in which the conformational space is sampled by performing random changes to a single ligand or a population of ligands [Sousa et al., 2006].

Another classification of search methods is based on the number of potential solutions that are explored. For example, local search methods tend to find the closest minimum energy to the current conformation, while global search methods seek the best (i.e. the global) minimum energy within the defined search space. Hybrid global-local research methods have been shown to work even better than global methods alone, being more efficient and able to find lower energies [Morris et al., 1998].

Genetic algorithms are effective to perform conformational search in docking. They are inspired by biology, mimicking the main characteristics of Darwinian evolution and Mendelian genetics. The main concept of these algorithms is that a ligand bound to a protein can be described by a set of state variables that define the ligand translation, orientation, and conformation with respect to the protein. In a genetic algorithm, each state variable corresponds to a "gene". The state of the ligand corresponds to the "genotype", and its atomic coordinates correspond to the "phenotype". In molecular docking, the "fitness" is the total interaction energy of the ligand with the protein, and it is evaluated using an energy function. Pairs of individuals are randomly mated and new individuals inherit genes from one or both parents in a process of crossover. Some offspring randomly mutates one gene, which changes their fitness. The selection of the offspring of the current generation is based on the fitness of the individual: thus, those best suited to the environment in which they are located reproduce, while those less suited are discarded [Morris et al., 1998].

The Lamarckian Genetic Algorithm (LGA) uses different starting conformations of the ligand, referred to as the "population of individuals", to find the conformation that best interacts with the protein environment. This is done by choosing the individual with the highest "fitness", which is determined by an energy score. However, this procedure is based on an inverse mapping function, which yields a genotype from a given phenotype. It is possible to finish a local search by replacing the individual with the result of the local search [Morris et al., 1998].

The scoring functions required for docking have two fundamental properties: selectivity (the ability to distinguish between correct and incorrect structures) and efficiency (the ability to find the solution in a reasonable amount of computing time).
Three types of functions can be used: classical force field-based functions, knowledge-based functions, and empirical functions. The classical force field-based functions compute the direct interactions between protein and ligand, including non-covalent protein-ligand interactions. Considering the intrinsic error of each individual energetic term, these methods often need empirical scaling parameters to fit their results to experimental binding data [Li et al., 2019]. Later, scoring functions were improved by solvation energy terms to take into account the free energy change in a protein-ligand binding process [Zou et al., 1999]. The energy terms were computed with either Poisson–Boltzmann (PB) or Generalized Born (GB) continuum solvation models [Liu and Wang 2015].
In knowledge-based potentials, the frequency of a pairwise contact can be assumed to be a measure of the energy it contributes to protein-ligand binding. If a specific pairwise contact occurs more frequently than in a reference state, it indicates a favorable interaction between the given atom pair. If it occurs less frequently, it indicates an unfavorable interaction. The standard approach for deriving desired pairwise potentials is to use a large set of protein-ligand complex structures from PDB as the training set [Evers et al., 2003].
In an empirical-based scoring function, the final score is determined by linear regression from experimental data collected for a particular category of ligands. A training set of protein-ligand complexes with known 3D structures and binding affinity data is required to perform the regression analysis. Empirical scoring functions must first be calibrated to reproduce protein-ligand binding affinities. The generalization of these scoring functions to other categories of ligands is dependent on the quality and quantity of experimental data collected [Liu and Wang, 2015].

2.4.2 AutoDock suite

For the docking simulations made in our project, we used AutoDock version 4.2, setting up the molecular system with the AutoDockTools (ADT) 1.5.6 software [Morris et al., 2008].

Since its release in 1990, AutoDock is an effective tool in predicting bound conformations and binding energies of ligands for macromolecular targets, accurately and quickly [Goodsell, 1990].

AutoDock is made up of two programs, AutoGrid and AutoDock, which work together within the ADT graphical interface. ADT is useful for the coordinate preparation, experiment design, and analysis. It is implemented in the object-oriented programming language Python and is built from reusable software components. ADT includes methods for formatting input molecule files, calculating charges, and specifying rotatable bonds in the ligand and the protein, as well as methods for clustering, displaying, and analyzing the results of docking experiments [Morris et al., 2009]. To prepare the input, the receptor and ligand pdb files are converted into an AutoDock proprietary format called pdbqt, which contains information on the partial charge, the position of the polar hydrogens, and the rotational degrees of freedom of the ligand bonds in addition to the coordinates. Then, a three-dimensional grid, either including the entire protein in the case of "blind" docking (i.e. without any indication of the location of the binding site) or focusing solely on the binding site in the case of "focused" docking is set up.

The grid coordinates and other information required to create the interaction maps are saved in a grid parameter file (gpf). AutoGrid assigns a pre-calculated interaction energy between each individual atom of the ligand and the protein, to each grid point [Morris et al., 2009]. AutoDock can apply a variety of search algorithms, including two local search methods [Solis and Wets 1981], two global search methods: Monte Carlo (MC) simulated annealing (SA) [Kirkpatrick et al., 1983], the genetic algorithm (GA) [Goldberg 1989] and one hybrid global–local search method, the Lamarckian GA

(LGA) [Morris et al., 1998]. The docking parameters are saved in a docking parameter file (dpf). The, AutoDock carries out the docking calculations and generates the output file (dlg) containing the simulation results based on the dpf file and the information generated by AutoGrid. The output file contains the histogram of the obtained solution clusters (called "poses"), the numerosity of each cluster, and the interaction energy. The best clusters are selected based on a compromise between the lowest (i.e. the best) interaction energy and the highest number of poses in each cluster, and are saved for later analysis.

The force field is based on a comprehensive thermodynamic model that allows the incorporation of intramolecular energies into the predicted free energy of binding. This is performed by evaluating energies for both the bound and unbound states. A new charge-based desolvation method has been incorporated which uses a typical set of atom types and charges. The method has been calibrated on a set of 188 different protein-ligand complexes of known structure and binding energy, showing a standard error of about 2-3 kcal/mol in prediction of binding free energy in cross-validation studies [Morris et al., 2009]

2.5 Molecular Dynamics (MD) Simulations

Proteins do not correspond to the immobile images derived from X-ray crystallographic analyses, but rather they continuously move their chains, and many of these movements are critical to the function they perform. MD studies are an important approach to understand the structural and functional properties of molecular systems, as well as their dynamic behavior. The MD method was first introduced by Berni Julian Alderin in the late 1950's when, together with his collaborators, he made a series of remarkable numerical simulations of a simple model system, a set of hard spheres [Ceperley and Libby, 2021]. The next major advance was in 1964, when Rahman carried out the first simulation using a realistic potential for liquid argon [Rahman,

1964]. The first MD simulation of a realistic system was done by Rahman and Stillinger in their simulation of liquid water in 1974 [Stillinger, 1974]. The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor [McCammon et al., 1977]. Currently, we can perform MD simulations of nearly every biological system [Hollingsworth e Dror, 2018].

2.5.1 The simplification of motion and energy calculations for macromolecules

MD must be based on molecular mechanics in order to measure the energy associated with a molecule and study the forces that cause its motion [Durrant and McCammon, 2011].

The energy of a molecule is determined by the number and the type of atoms and bonds in the molecule, as well as by the forces exerted by each atom on the others, i.e. the set of atomic interactions in the molecule [Abraham et al., 2015]. It is possible to parameterize the characteristics of the atoms and bonds present in a molecule through a force field. One can use various approximations to represent the atoms, such as the "united" approximation in which hydrogen atoms bound to the carbons are treated as a unique atom, or the "corse-grained" approximation which coarsely represents amino acids while still capturing characteristics of their side chains [Jamroz et al., 2013]; the more precise the details on the type of atom, the more complex the calculation to be made.

The properties of a covalent bond depend on the type of bond and the type of atoms involved in the bond. These characteristics include e.g. bond length, polarizability, and geometry. These properties can be calculated and tabulated from experiments conducted on real molecules.

After the parameterization of atoms and bonds has been carried out, the interaction energy can be calculated [Wang et al., 2006]. The classical force fields widely used today describe the total energy of a molecule as a sum of contributions of different

nature that may act on the molecule, with an equation that can be represented in a simplified form as:

$$E_{tot}=E_{bond}+ E_{non\text{-}bond}+ E_{other}$$

The term E_{bond} represents the contribution of covalent bonds to the energy of the molecule. This term can be further subdivided into contributions from the energy of stretching (vibrations), bending (rotations), and twisting (for dihedral angles), and other phenomena [Ponder and Case, 2003]. Instead, the term $E_{non\text{-}bond}$ represents the contribution of atoms not covalently bound. This term usually includes the van der Waals interactions, electrostatic interactions, and contributions from hydrogen bonds. E_{other} describes other energetic contributions, such as, for example, those of the solvent in which the molecule is immersed.

The classical force fields have a very complex mathematical treatment, which makes it computationally very demanding, considering that each energy calculation is repeated for each atom every time during the simulations.

Each type of force field has its advantages and disadvantages, each is specific for a certain type of molecule, based on experimental reference models used in the parameterization of atoms and bonds. As a result, MD programs have integrated various types of force fields to allow users the ability to choose the most suitable option for their purposes [Scheraga et al., 2007].

The dynamic behavior of a system can be described using Newton's equations in a MD simulation. This equation is solved by double integration, which introduces two arbitrary constants into the solution: one related to the initial velocity (v_i), the other to the initial position (r_i). The position vector can then be expressed as a function of time (i_t) (Figure 2.1).

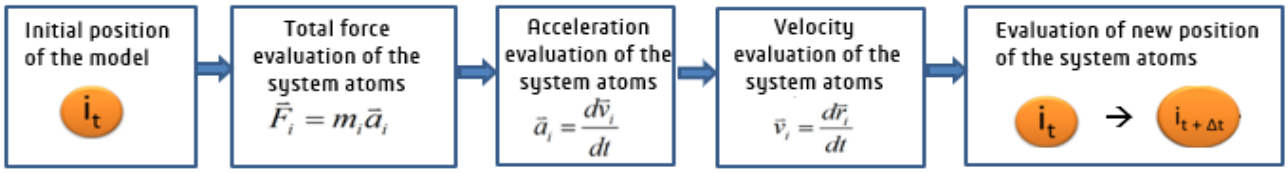


Figure 2.1: The dynamic behavior of a system of atoms a MD simulation.

However, Newton's equation represents a set of N second-order differential equations, and it is complicate to solve this set of equations precisely in order to obtain atomic trajectories. For this reason, the equation is numerically solved at discrete time steps to determine the trajectory of each atom. There are several algorithms available for the numerical integration of Newton's equations of motion and calculation of atomic trajectories in practical MD runs. Verlet Algorithm [Grubmüller et al., 1990] is an example and it consists in a Taylor expansion for forward and backward positions in time. Given the coordinates from the structure file, the initial velocities (attributed statistically), and the potential energy (calculated through the force field on the starting structure), it is possible to simulate the evolution of the system at a given time.

2.5.2 The simulation environment

The "life" of a biological macromolecule, such as a protein, occurs within an environment crowded of other molecules, atoms, or ions with which our protein is constantly in contact, at specific temperature and pressure conditions. As a result, in order to achieve a realistic prediction of the behavior of a protein during a MD simulation, the environment in which it is immersed must be represented, but since this system would be too complicated to simulate, the best compromise is to represent the protein as immersed in a box of water, which is viewed as a rigid molecule capable of short- and long-range interactions. Different types of water models are commonly used during simulations to recreate the aqueous environment inside the box [Scheraga et al., 2007]. The presence of water molecules in the system, of course, increases the computational complexity of the calculations [Scheraga et al., 2007]. Furthermore, in order to recreate a neutral system, the presence of counter-ions is required to neutralize

the net charge of the protein or its ligands, just as if a buffer solution was used in the laboratory [Scheraga et al., 2007].

When such a simulation environment is used, in order to avoid the presence of unnatural interactions at the edges of the box between the atoms inside the box and the vacuum environment outside the box, usually periodic boundary conditions are applied. The simulation box is placed in the center of a lattice of boxes identical to itself, with the same conditions reproduced [Scheraga et al., 2007]. This box lattice that surrounds the molecular simulation environment (Figure 2.2) is required to prevent the protein from "escaping" the environment in which it is contained and landing in a space that is different from the one in which it is immersed.

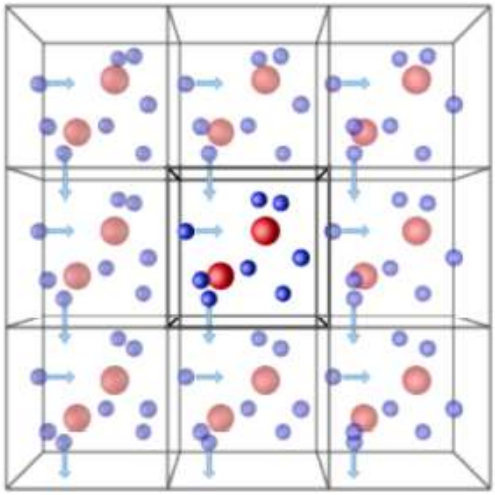


Figure 2.2: Representation of periodic boundary conditions [Ercolessi et al., 1997]

During the simulation, the mirror images of each protein must maintain a certain minimum distance from each other, avoiding interacting in any way [Abraham et al., 2015]. Each cell is identified by three vectors, the dimensions of which are listed at the bottom of the protein co-ordinate file. Different shapes (e.g. cubic, triclinic, octahedric, dodecahedric) can be selected for the box, and usually the protein is positioned in the central part, fixing a distance from the box edges that ensures the two periodic images of the protein are not coming into contact [Lemkul, 2018].

The environmental variables that a protein is exposed to must also be replicated within a MD simulation. Within the simulations, these parameters can be specified using thermostats and barostats to recreate environments at the desired temperature and pressure and observe the behavior of our system under these conditions [Scheraga et al., 2007].

The dynamic state of a system is defined solely by a small set of parameters. The canonical isothermal ensemble (NVT) and the canonical isobaric ensemble (NPT) are of particular interest in MD simulation. NVT ensemble is characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed temperature, T. NPT is distinguished by a constant number of atoms, N, a constant pressure, P, and a constant temperature, T [Scheraga et al., 2007]. As a result, by incorporating these latest values into simulations using a thermostat and a barostat, it is possible to observe the system's behavior in desired conditions [Lemkul, 2018].

In particular, we can use different algorithms to set a constant temperature of the system. The most popular algorithm used are Berendsen [Berendsen et al.,1984] and Nosè-Hoover [Nosè, 1984]. Berendsen thermostat is able to maintain the desired temperature of the system representing it coupled to an external bath [Eslami et al., 2010]. The only limit of Berendsen could be an incorrect distribution of kinetic energy [Abraham et al., 2015]. To solve this, a modified Berendsen thermostat has been developed, called V-rescale, in which an additional stochastic term ensures a correct kinetic energy distribution [Basconi et al., 2013]. Instead, Nosè-Hoover considers the heat bath as an integral part of the system [Nosè, 1984]. We can also use different algorithms to set a constant pressure of the system. The most popular algorithms used are Berendsen, with the same considerations previously described, but applied to the desired pressure [Lin et al., 2017], and Parrinello-Rahman, which is similar to the Nosé-Hoover temperature coupling, and in theory gives the true NPT ensemble [Bussi et al., 2009].

2.5.3 Programs used for MD simulations

2.5.3.1 GROMACS

MD simulations in our work have been performed using GROMACS (GROningen MAchine for Chemical Simulations), a package developed by the Department of Biophysical Chemistry at the University of Groningen in the Netherlands that is used to perform MD simulations on systems containing hundreds of thousands of atoms. This software was originally developed to examine biological molecules such as proteins, lipids, and nucleic acids, but it has been extended to the study of non-biological systems as well [Abraham et al., 2015].

GROMACS contains a number of utilities for preparing and executing the simulations, and for analyzing results. It includes 15 force fields for automatically generating topologies for proteins and multimeric structures. The libraries contained within them, in particular, allow to parameterize the 20 natural and some modified amino acids, the 4 nucleotides, various sugars and lipids, as well as special groups like heme and other small molecules. External tools compatible with the reference force fields can be used in the presence of small ligands that are not recognized by the internal force fields [Abraham et al., 2015]. Furthermore, GROMACS includes many tools for analyzing trajectories at the end of the simulation. The results are returned in the form of graphs complete with legends, labels, and so on, in a format that can be interpreted by a variety of external visualisation programs [Abraham et al., 2015].

2.5.3.2 ANTECHAMBER, ACPYPE

Since GROMACS is not able to write the topology of many ligands, it is necessary to use external tools. A fundamental aspect of this procedure is that the parameters used to define the topology of the ligands designed with these tools must be consistent with those used to define the topology of the protein using GROMACS' internal force fields. As a result, there are specific softwares for each force field present in GROMACS that attempt to provide parameters compatible with the available force fields [Abraham et al., 2015]. ANTECHAMBER, a program for creating the topology of generic organic

molecules and metal centers, is one of these. It is based on the General Amber Force Field (GAFF), which was specifically designed to parameterize many pharmaceutical compounds while remaining consistent with AMBER [Wang et al., 2006]. The ligand must be submitted in pdb format. It is then transformed into mol2 format, including coordinates and charges for each atom, by ANTECHAMBER. ANTECHAMBER does not provide a topology file and a coordinate file that can be used in GROMACS. To convert these files into a format compatible with GROMACS, it is necessary to use another tool, known as ACPYPE, which is based on the Python language [Sousa and Vranken, 2012]. In particular, the mol2 file is given as input to LEaP, a tool in the ANTECHAMBER package. This converts all the parameters of the ligand, specified in the mol2 file, to a format compatible with GAFF. At the end of the operation, a prmtop file and an inpcrd file for our ligand are returned, two intermediate files that will be used as input to obtain the final ones for our ligand [Wang et al., 2004].

In the present work, ANTECHAMBER and ACPYPE were used to generate the topology and coordinate file for the arginine, H2U and G1P ligand.

2.5.3.3. CHARMM-GUI

It is a web-based graphical user interface that generates input files for a variety of programs such as CHARMM, NAMD, GROMACS, AMBER, GENESIS, LAMMPS, Desmond, OpenMM, and CHARMM/OpenMM [Lee et al., 2016]. Since its original development in 2006, CHARMM-GUI has been widely adopted for various purposes, but the main one is to prepare complex biomolecular systems for molecular simulations. It is organized through a different number of modules, to read and modify molecules (PDB Reader & Manipulator, Glycan Reader, and Ligand Reader & Modeler); to build all-atom simulation systems in various environments (Quick MD Simulator, Membrane Builder, Nanodisc Builder, HMMM Builder, Monolayer Builder, Micelle Builder, and Hex Phase Builder). In particular, Solution Builder [Jo et al., 2008] has been used for this project to generate a series of input files for the simulation of MD in aqueous solvent environments. To use Solution Builder, one must

load the complex (protein and cofactors) and the ligand separately, in compatible formats. It is then possible to add water and ion to the system through the Solution Builder. In this step, the users can select the waterbox type and enter the edge distance from a drop-down menu. The users can choose the neutralisation action by adding counterions only, or by selecting a concentration (default is 0.15 M) of the buffer solution. At the end, the user chooses both a force field and a desired simulation package, and downloads the input files needed to perform the simulations with the selected package.

CHARMM-GUI was used to create the topology of the systems to be prepared for long MD simulations (see paragraph 3.4).

2.5.4 Workflow of the MD simulations using GROMACS

The following general steps are performed for a MD workflow when using GROMACS:

- ***Protein topology preparation***: the protein pdb file is processed to generate a topology file (top), a position restrains file (itp), and an atomic coordinates file (gro). The top file contains the parametrization of atoms, bonds, angles, and dihedrals according to the selected force field. The itp file contains the information used to 'constrain' the position of the heavy atoms and is useful for the subsequent equilibration steps. The gro file is the structure file in the proprietary Gromos87 format. In this step, the user is also prompted to choose which water model to use as a solvent in the system;

- ***Preparation of the ligand topology and creation of the files for the final system***: The same procedure that is performed for the protein is also necessary for the ligand(s), whenever present. If the ligand is recognized by GROMACS, its topology is automatically added to that of the protein; otherwise, the topology of the ligand must be created with the external tools described before (see paragraph 2.5.3.2). Only when ANTECHAMBER is used, the coordinates files of the protein and the ligand, as well as their topology files, must be merged manually. If CHARMM-GUI is used, the

complex (protein + ligand) is uploaded and the topology of the complex is automatically prepared (see paragraph 2.5.3.3);

- ***Box construction and solvation****:* the system's coordinate file is processed to allow the construction of a box around the protein. The shape of the box and type of water for the system are selected in this phase (see paragraph 2.5.2). After adding the solvent, it is critical to update the top file with the reference to the number and type of water molecules added to the system [Lemkul, 2018]. This step can be fully automated when performed using the CHARMM-GUI program, as explained in the previous section;

- ***Neutralisation:*** After the addition of water, the system must be neutralized by the addition of counterions. This is achieved by replacing water molecules with ions (usually, sodium or chlorine) in sufficient numbers to achieve neutrality. At the end of this step, the topology file will be added with the number and type of the counterions [Lemkul, 2018]. Also this step is fully automated when performed using the CHARMM-GUI, as explained in the previous section;

- ***Energy minimization***: Before starting the dynamics simulation, it is necessary to ensure that the system has no steric inconsistencies or incorrect geometry. Therefore, the structure must be relaxed by applying a minimization procedure. It is an iterative process that ends the search for the best conformation when the maximum force is lower than a preset value, or when a maximum number of steps is performed, parameters specified in the mdp file. To determine whether the minimisation is successful, the potential energy of the protein must be examined [Lemkul, 2018]. The algorithm used for minimisation is usually the steepest descent, which is very robust and easy to implement, but other algorithms are also available in GROMACS, such as the conjugate gradient algorithm [Lemkul, 2018];

- ***Equilibration:*** by minimizing, we obtain a geometrically relaxed structure of the protein. The solvent and ions surrounding the protein must also be equilibrated before the dynamics can begin. Moreover, it is necessary to bring the system up to the temperature at which we want to run the simulation, then, once the optimal temperature

is reached, a pressure must be applied so that the system can reach the proper density [Lemkul, 2018]. This phase is usually divided into two steps. A set of restraints is applied to the macromolecule, and a first equilibration is run with the system considered as NVT ensemble (fixed number of particles, volume and temperature). If the system is inserted in a thermostated bath at a desired temperature value, it can vary its energy until it equilibrates around the set temperature value. The mdp file specifies the temperature value in Kelvin that the user wants to set for the system. By analyzing the edr output file, it is possible to observe a curve describing the temperature trend during the equilibration and confirm that it has reached the desired value [Lemkul, 2018].

A secondo equilibration step is usually run by considering the system in a NPT ensemble (fixed number of particles, pressure, and temperature), always keeping the position of the atoms in the macromolecules restrained. The addition of a barostat allows the system to reach the applied pressure value and to find the right density [Lemkul, 2018]. In a similar fashion to the previous equilibration phase, the mdp file allows for specification of the desired pressure value and the barostat to be used. At this stage, no new initial speeds are generated, the speeds from the previous equilibration step are used. The pressure trend during this equilibration phase can be visualized to make sure it has reached the set value;

- ***Production dynamics***: After the equilibration steps, the system is well-balanced at the desired temperature and pressure values. The forces imposed by spatial restraints on the ligand-protein system during the previous equilibration can be removed, allowing the final MD simulation to run for the desired time;

- ***Post-processing and analysis of the results:*** As in any simulation conducted with periodic boundary conditions, molecules may appear "broken" or may "jump" back and forth across the box. To re-center the protein and rewrap the molecules within the unit cell to recover the desired shape of the box, at the end of the simulations, usually there

are post-processing commands that must be executed before the trajectory can be analyzed. After this post-processing, the trajectories can be analyzed to evaluate:

- *the stabilization of simulations:* the analysis of the energetic components (including total energy, kinetic energy component, potential energy component, pressure, temperature, volume, and density), the analysis of the minimum distance of periodic images, and of the root mean square deviation (RMSD) of atom distances;
- *the global structural features*: the analysis of the root mean square fluctuation (RMSF), of the radius of gyration, of the secondary structures, of the predicted solvent accessible surface area (SASA);
- *other analyses made to analyze the structural features of the enzyme:* for example, the quantitative and qualitative analysis of hydrogen bonds and salt bridges, in terms of percentage of existence.

The trajectories can also be visualized by means of graphic interfaces, such as Visual Molecular Dynamics (VMD), a molecular visualization program for displaying, animating, and analyzing large biomolecular systems [Humphrey et al., 1996].

2.5.5 Performing MD calculations on HPC systems

Because of the massive amount of calculations required, MD simulations are typically run on clusters or supercomputers with hundreds to thousands of processors running in parallel [Durrant and McCammon, 2011]. The MD simulations for this project were performed on advanced calculation platforms provided by CINECA, a non-profit inter-university consortium made of 102 members (including 69 Italian universities and 33 institutions) for automatic calculation, founded in 1969, which is one of Europe's large-scale facilities for high-performing computing and among the most powerful in the world.

CINECA's HPC environment is made up of general-purpose computers that are constantly kept at the cutting edge of technology. All CINECA HPC systems share a common environment to facilitate resource utilization and data and program portability. Despite the different charactesistics of the various systems, users can access

any system in similar ways, expect similar behavior, and have access to shared resources. There are areas available for both users and projects. Our simulations were run on MARCONI, built on the LENOVO NeXtScale platform and Intel Xeon Phi processors. Since June 2016, it has been gradually upgraded, and the current configuration is Marconi-A3 with SkyLake (in production since August 2017, upgraded in January 2018 and completed in November 2018). Marconi was ranked 12th in November 2016 and 19th in November 2018 on the Top500 list of the most powerful supercomputers.

To launch the simulations, users must typically prepare a shell script containing all operations to be executed in batch mode, once the necessary resources are available and allocated to the process. The process is then started and executed in the computing nodes of the cluster.

The structure of a typical job in which there is a request for computing resources is shown in the figure 2.3:

```
#SBATCH --job-name jobname                    # name of the job
#SBATCH -o job.out                            # output file
#SBATCH -e job.err                            # error file
#SBATCH --gres=gpu:2                          # generic resources
#SBATCH --time=1:00:00                        # d:hh:mm:ss
#SBATCH --partition=<partition>               # chosen partition
#SBATCH --qos=quality-of-service              # select a qos
#SBATCH --account=<my_account>                # name of the account
#SBATCH --nodes=1                             # nr. Of nodes
#SBATCH --ntasks-per-node=4                   # nr of MPI ranks
#SBATCH --cpus-per-tasks=   8                 # nr.of OpenMP threads
#SBATCH --ntasks-per-socket=2                 # nr.of ranks per socket
```

Figure 2.3: A typical shell script preparing containing all operations to be executed in batch mode. In red are described the meanings of each line composing the typical shell script.

The access to these computational resources was made available thanks to ELIXIR-IT project [Castrignanò et al., 2020]

2.6 Set up applied in the present Ph.D. project

2.6.1 Starting system

We have used as a starting point the models of wild type hGALT and of the mutant p.Gln188Arg, derived from the crystallographic structure of hGALT enzyme (PDB

code: 5IN3) [McCorvie et al., 2016] and modelled according to d'Acierno et al., 2018. Both models contain the ligands G1P and H2U that were visible in the crystallographic structure, as well as the Zn ions.

During the analysis of the crystallographic structure of wild type hGALT, we found a series of discrepancies (shown in the table 2.1) between some data present in the scientific article related to the crystallographic experiment and those released in the PDB file 5IN3. In this structure, the enzyme was captured during the first ping-pong step, having H2U and G1P in the active sites of both chains (A and B). Particularly, the terminal phosphate group of both H2U has a phosphorus atom which apparently binds 3 oxygen atoms instead of 4 (Figure 2.4). During the modelling process, the covalent bond between H2U and the residue His186 has not been modelled, and the structure of this ligand in both cases has been modified (Figure 2.4 B) in order to restore the normal phosphate group, by using Chimera. This covalent bond is atypical because it is a transient bond; the choice not to model it has also avoided problems in the parameterization of this anomalous bond in the following steps. Moreover, the phosphate groups of the two ligands (H2U and G1P) were considered in their charged (deprotonated) form throughout the simulations (Figure 2.4 B and C).

Discrepancy	PDB file	McCorvie et al., 2016
Resolution	1.73 Å	1.9 Å
Template used for the molecular replacement technique	1HXQ (structure from E. coli) [Wedekind et al.,1996]	1HXP (structure from E. coli) [Wedekind et al.,1996]
Temperature for the growth of crystal	12.85 °C	20 °C
Pixel detector	PILATUS 6M	PILATUS 2M

Table 2.1: Discrepancies between the experimental data contained in the PDB file 5IN3 with respect to the article associated to the file McCorvie et al., 2016.

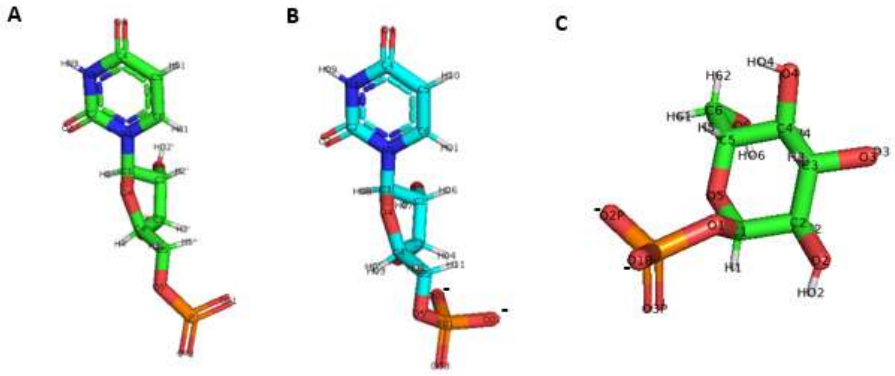


Figure 2.4: A. Crystallographic structure of H2U from 5IN3; B. Structure of H2U with two negative charges; C. Structure of G1P with two negative charges.

2.6.2 Set up of the docking simulations using AutoDock

2.6.2.1 Docking of Arginine

Arginine was considered in its zwitterionic form with the side chain protonated throughout the simulations. For all the docking simulations, polar hydrogens were added to the proteins and ligands (except for those groups in ligands that were considered deprotonated), and charges were assigned according to Gasteiger [Gasteiger, 1980]. For the docking with arginine into the active site of the protein, we used a grid map with a spacing of 0.375 Å and dimensions of 58x80x74 points, focused on the residues belonging to the active site of GALT formally identified as "A" (containing His 186 of the chain A), as reported in the PDB file. These simulations were performed either by alternatively keeping G1P and H2U in the active site A, or by removing both ligands from the active site A; instead, active site B was left without ligands. A grid map with the same spacing as above and dimensions of 92x112x102 points, centered on the central cavity of the enzyme, was used to set up the simulations of the binding of arginine in the central cavity of the enzyme. For each system, 100 docking runs were performed using the AutoDock Lamarckian genetic algorithm, treating the protein as rigid and the ligand as flexible. All the other parameters were kept as default (population size: 150; number of energy evaluations: 2,500,000; and

number of generations: 27,000), as is advisable for this molecule with 6 torsional degrees of freedom. The docking poses were clustered using an RMSD value of 2.0 Å. The conformations representative of the best energetic and/or the most populated cluster of poses were selected, saved in pdb format, and analyzed for their interactions with the enzyme by using Discovery Studio. Those identified as the best conformation for each system were used as a starting point for the following MD simulations.

2.6.2.2 Docking of PCs

A total of five PCs were selected (we will not provide their names to protect possible future patent applications, but we will indicate them as follows: PC1, PC2, PC3, PC4 and PC5). These PCs were simulated in the central cavity in the same condition of arginine. A grid map with the same spacing as above and dimensions 68x86x72 was used. At the end, 16 docking were made for each PCs (see paragraph 3.3.1). Once identified a potential allosteric site (see paragraph 2.7), all five PCs were simulated for the docking also on the putative allosteric sites of both chains. A grid map with dimensions 48x48x56 (for potential allosteric site of chain A) and with dimensions 50x53x43 (for potential allosteric site of chain B) were used.

2.6.2.3. Docking of pharmacophoric hits

The best conformations selected from each previous docking were used as the starting point for receptor-based pharmacophore modeling. The steps of receptor-based pharmacological modelling are explained in detail in paragraph 2.7.2. For now, it is sufficient to know that the result of this procedure provides as output a set of ligands (called a hit list). From this hit list, a total of 19 ligands, called pharmacophoric hits were selected (according to the parameters explained in paragraph 2.7.2). For each hit, two types of docking were simulated: on the potential allosteric site of chain A and on the potential allosteric site of chain B, identified by the FTMap server (see paragraph 2.7.1). For simulations on the potential allosteric site of chain A, a grid map with the same spacing and dimensions 48x48x56 was used. The best docking results concerned only 4 ligands among the 19. So, only for these 4 ligands docking were done also into

a putative allosteric site of chain B. For this, a grid map with the same spacing as above and dimensions 50x53x43 was used.

2.6.3. Set-up of MD simulation procedures

Both the dynamics of GALT enzyme and of its pathogenic mutant p.Gln188Arg under different experimental conditions and dynamics of arginine have the same protocol, described in paragraph 2.6.3.1.

During the last year of PhD, the MD simulation protocol has been modified in order to set-up long MD simulations for the study of the allosteric communications inside GALT protein. This protocol is described in section 3.4.

MD simulations concerned the following systems:

- no arginine: wild type hGALT; wild type hGALT + G1P + H2U; p.Gln188Arg; pGln188Arg + G1P + H2U at two different temperatures (310 K, corresponding to the normal body temperature, and 334 K, a temperature close to the Tm of the enzyme). We have used as a starting point the models of wild type hGALT and of the mutant p.Gln188Arg obtained as described previously;
- arginine in the active site A: wild type hGALT + arginine; p.Gln188Arg + arginine; wild type hGALT + G1P + arginine; p.Gln188Arg + G1P + arginine; wild type hGALT + H2U + arginine; p.Gln188Arg + H2U + arginine;
- arginine in the central cavity: wild type hGALT + arginine; p.Gln188Arg + arginine; wild type hGALT + G1P + arginine; p.Gln188Arg + G1P + arginine; wild type hGALT + H2U + arginine; p.Gln188Arg + H2U + arginine.

The starting points for the MD of arginine simulations were the best docking results.

The package used for the MD simulations was GROMACS 2018.3 [Abraham et al., 2015]. The force field used throughout the simulation of systems with arginine was Amber ff99SB-ILDN [Ponder, et al., 2003; Lindorff-Larsen, et al., 2010], and the packages ANTECHAMBER and ACPYPE [Wang et al., 2006; Sousa da Silva et al., 2012] were used according to their instructions to calculate the correct topology for the ligands. Each of the starting systems was included in a cubic box centered on the

protein, with a distance of 1.2 nm from it, filled with water (using the TIP4P model [Abascal et al., 2005] according to the suggestions in the GROMACS manual) and neutralized with sodium or chlorine counterions. The systems were first minimized by applying steepest descent minimization, setting the cut-off for short-range electrostatic and van der Waals interactions to 1.2 nm, and using the grid method to determine the neighbor list. Minimization stopped when the maximum force reached a value lower than 10.0 kJ/mol/nm.

Equilibration steps with position-restrained MD simulations were run first in NVT conditions for 100 ps and subsequently in NPT conditions for 1000 ps. For the NVT equilibration, the V-rescale thermostat [Bussi et al., 2007] was applied, fixing two temperatures: 310 K (the normal body temperature) and 334 K (a temperature close to the Tm of the enzyme); for NPT equilibration, the Berendsen barostat [Berendsen, et al., 1984] was added to keep the pressure constant at 1.0 bar. At the end of the equilibration, for each system, we performed 200 ns-long MD simulations in NPT conditions, at a temperature of 310 K or 334 K. For the production runs, the Berendsen barostat was replaced with the Parrinello–Rahman barostat [Parrinello et al., 1981].

The other parameters selected for the production simulations were: the leap-frog algorithm [van Gunsteren et al., 1988] for integrating Newton equations of motion; the LINear Constraint Solver (LINCS) algorithm [Hess et al., 1997] to constrain bonds; Verlet [Verlet et al., 1967] as cutoff scheme in the neighbor searching section; Particle Mesh Ewald (PME) method [Darden et al., 1993] to handle long-range electrostatic interactions. Two replicas have been made for each simulation.

At the end of the simulations, the trajectories were analyzed using GROMACS analysis tools. The obtained results were plotted by using XMGrace software (https://plasma-gate.weizmann.ac.il/Grace/; last accessed 24 September 2022).

2.6.4 Set up of the search of allosteric site and identification of pharmacophores for GALT

As already reported in paragraph 2.3.2, FTMap, a computational mapping server that identifies hotspots or binding regions in proteins [Brenke et al., 2009], has been used to identify putative allosteric sites in the structures of wtGALT and p.Gln188Arg. The search was made by using default parameters. As reported in paragraph 2.6.2.2, we performed docking of 5 selected PCs either on central cavity or on the putative allosteric site of GALT enzyme. Following, the best conformations of docking on the putative allosteric site were selected and used as a starting point for the following step, made by applying the Discovery Studio Interaction Generation protocol (see paragraph 2.2.3). In our set up, Discovery Studio takes as input only the docking conformations (best energy and most populated) and creates all possible combinations of pharmacophoric model. A representative result is shown in table 2.2, in which (as reported in the paragraph 2.2.3) is ranked on decreasing selectivity score, describing the pharmacophoric features (or feature set).

7 features match the receptor-ligand interactions: AAADDDP
10 pharmacophores generated.

Pharmacophore Summary			
Pharmacophore	Number of Features	Feature Set	Selectivity Score
Pharmacophore_01	5	ADDDP	11,536
Pharmacophore_02	5	AADDP	10,622
Pharmacophore_03	5	AADDP	10,622
Pharmacophore_04	4	DDDP	10,021
Pharmacophore_05	5	AAADP	9,7086
Pharmacophore_06	4	ADDP	9,1074
Pharmacophore_07	4	ADDP	9,1074
Pharmacophore_08	4	ADDP	9,1074
Pharmacophore_09	4	ADDP	9,1074
Pharmacophore_10	4	ADDP	9,1074

Table 2.2: Summary of pharmacophores generated by BIOVIA Discovery Studio, a representative result

Among the generated pharmacophore models, the best one (which generally corresponds to the first one having a better selectivity score) is selected, and this best model is used for the last step, i.e. the search for pharmacophoric hits. In our study, we uploaded DrugBank database (the last version on Febrary 2020) into Discovery Studio. In particular, the ’Search 3D-database’ function provides a rapid filtering of the input database, selecting those hits that match the minimum amount of features defined by user. A hit list is finally provided in the form of a table, presented in descending order of fit value, with associated information for each compound (e.g. the DrugBank code, the IUPAC and generic name, the status of the trial (drug approved, in trial or under investigation) and finally whether it meets Lipinski's "rule of 5” [Lipinski et al., 1997, 2001]. The next step involves the selection of hits based on two parameters: i. a fit value greater than or equal to 3; ii. only pharmacophoric hits found from the pharmacophore model generated by the docking conformations of the p.Gln188Arg mutant were reported.

In the last step, the selected hits (19 in total) were simulated on the potential allosteric site of chain A. Additionally, only for the best 4 among 19 hits, docking were performed also on the potential allosteric site of chain B.

3. RESULTS AND DISCUSSION

3.1 Analysis of the structure-function-dynamics relationships of GALT enzyme and of its pathogenic mutant p.Gln188Arg by means of MD simulations

We performed MD simulations of both wtGALT and p.Gln188Arg proteins in the absence or in the presence of the substrates (G1P and H2U) at different temperatures (310 K and 334 K). 310 K corresponds to the normal human body temperature, whereas 334 K is a temperature near the Tm of the enzyme [McCorvie et al., 2016; Coelho et al., 2014]. In this last case, we would like to simulate an environment that favors the destabilization of the enzyme, to predict the molecular effects (if any) of this destabilization, particularly at the level of intersubunit interactions.

3.1.1 Analysis of MD Simulations at 310 K

The simulations performed on both wtGALT and p.Gln188Arg at 310 K, in the presence and in the absence of the substrates, were analyzed to confirm that they were not significantly affected by perturbations.

The analysis of the energetic components (including total energy, kinetic energy component, potential energy component, pressure, temperature, volume, and density), of the minimum distance of periodic images, and of the RMSD of atom distances, showed that both systems quickly reached the stabilization *(see Supplementary Files 1, 2, 3, 4)*

A slight difference in the global structural features of these two systems, in the absence of substrates, can be detected with the analysis of the RMSF, which shows a slightly enhanced flexibility of the mutant in the zone of residues ~50–70 (including segment 50–60 formed by very conserved residues at the intersubunit interface, some of which are also involved in substrate interactions), ~300–320 (a long loop including the three conserved His residues His301, His319, His321 forming the Zn binding site), and marginally in the zone of mutation. The fluctuation, especially for segment 300–320,

is not identical in the two chains (Figure 3.1 a,b). For the mutant in the presence of the substrates, the RMSF graph shows for both systems an enhancement of the flexibility around residues 50–70 and a decrease in the flexibility of the segment between residues 300–320 with respect to the systems in the absence of the ligands (Figure 3.1 c, d), in particular for one of the two chains. Thus, it appears that the presence of the substrates has a different effect on the local flexibility of wt hGALT with respect to p.Gln188Arg.

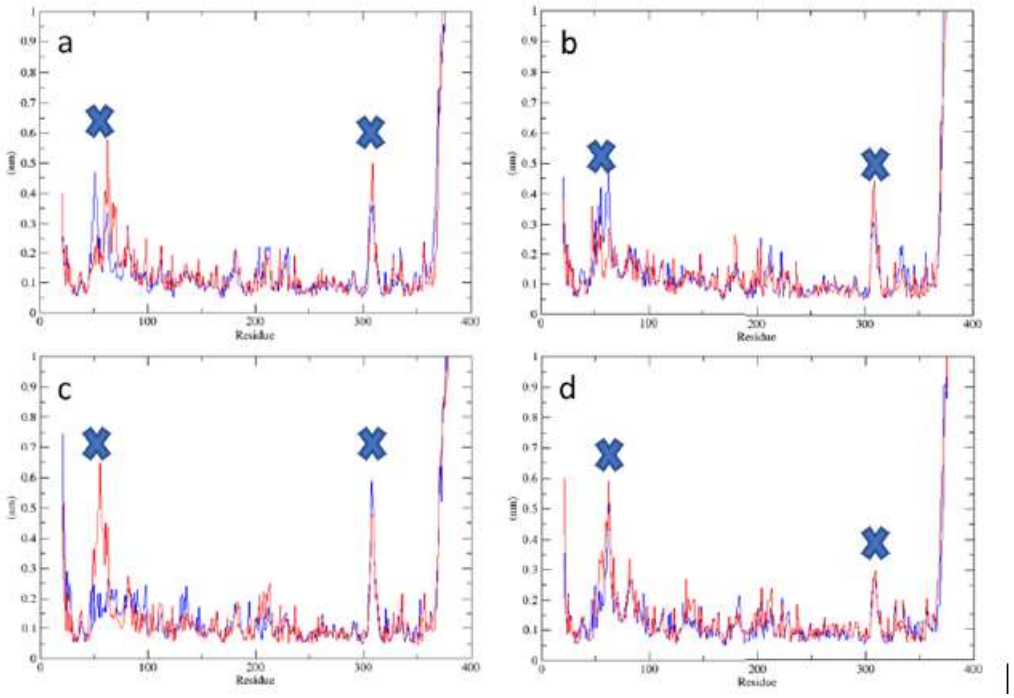


Figure 3.1 RMSF analysis for simulations at 310 K of: (**a**): wtGALT; (**b**): p.Gln188Arg; (**c**) wtGALT + G1P + H2U. (**d**) p.Gln188Arg+ G1P + H2U. Blue lines represent RMSF fluctuation of chain A, red lines RMSF fluctuation of chain B.

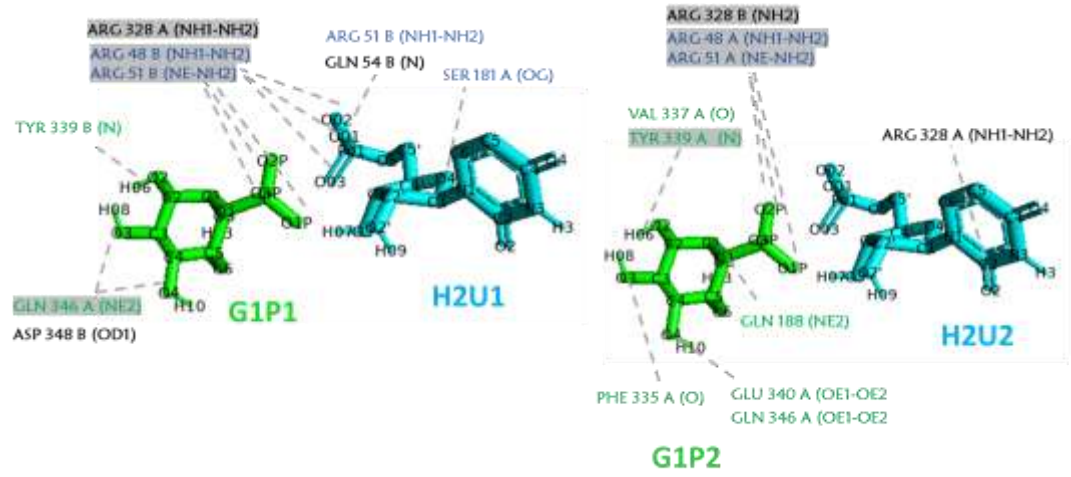
The results obtained are in agreement with the work of McCorvie et al., 2016, discussed in the introduction (see paragraph 1.3.2). McCorvie and coauthors identified loop 49-63 as one of the crucial loops, corresponding to a surface-exposed region in the H2U-binding site which is ordered only when H2U is bound to the active site. The enhancement of flexibility around residues 50-70 (observed in our MD simulations) includes the loops 49-63, which is more flexible in the mutant, where the binding of H2U is compromised and cannot impart order to this region.

The analysis of the content of secondary structures did not show main differences between the two systems in the absence of the ligands (*see Supplementary File 9 a, b).* The mutant enzyme with the substrates, instead, highlights a slightly increased content in less regular structures, such as the π-helix (indicated as 5-helix in the graph) *(see Supplementary File 9 c, d).* Thus, once again, in the presence of the substrates the mutant enzyme shows a perturbation that could be diagnostic of the alteration of its structural features due to the mutation.

During the simulations, the radius of gyration remained stable and practically identical for all systems *(see Supplementary File 10, panels a–d).* The SASA was similar for systems in the absence of substrates, with a small decrease (from 310 to 290 nm^2) in both cases and a slightly more accentuated decreasing trend for the mutant enzyme *(see Supplementary File 11, panels a, b).* In the simulation in the presence of the substrates, the SASA of wtGALT is fluctuating around an average value of 305 nm^2, whereas in p.Gln188Arg it shows a marked decreasing tendency along the trajectory from about 320 nm^2 to 300 nm^2 (*see Supplementary File 11, panels c, d).*

The analysis of the interactions between the enzymes and the ligands present in both active sites show that both ligands are stably bound to the proteins during the simulations *(see Supplementary File 12, panels a, d).* Both in wtGALT and in p.Gln188Arg, G1P interacts with residues belonging to both chains, mainly with residues Arg48 and Arg51 and less stably with residues belonging to the segment 330–340 of the protein. In wtGALT, H2U interacts with several residues, but does not show persistent interactions. On the contrary, in p.Gln188Arg, H2U shows persistent interactions with His186 and Arg188. In particular, with this last residue, both hydrogen bonds and salt bridges are predicted to occur. The interactions of the substrates are not exactly symmetrical in both active sites (Figure 3.2).

a)



b)

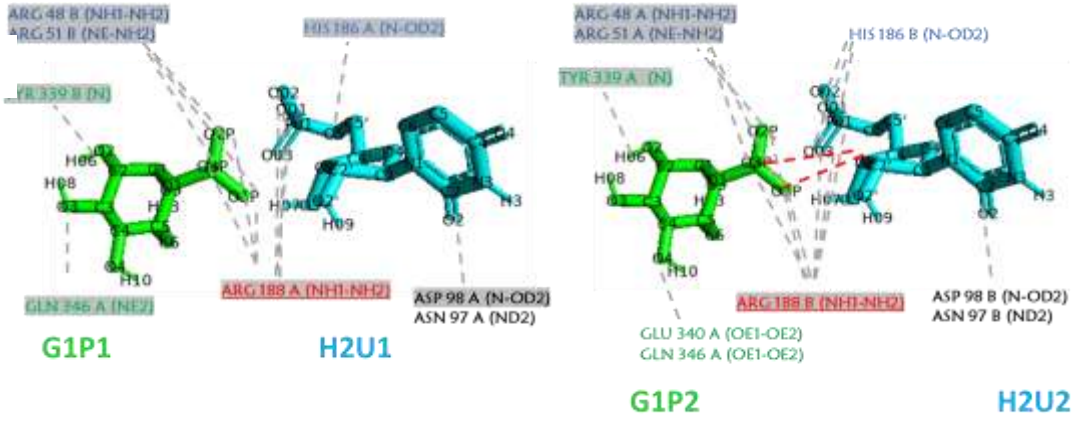


Figure 3.2: Interactions between enzyme and ligands in simulations at 310 K. (a) wtGALT; (b) p.Gln188Arg. Panels on the left indicate the ligands of the active site 1 (formally belonging to subunit A), panels on the right indicate the ligands of the active site 2 (formally belonging to subunit B). Grey background indicates interactions that persist for more than 50% of the simulation time. Grey dashed lines indicate H-bonds; residues underlined are also able to form salt bridges. Red dashed lines indicate interactions between the ligands.

To verify if the introduction of the mutation p.Gln188Arg in the protein could destabilize its quaternary assembly, we analysed the H-bonds present at the intersubunit interface. The results are shown in table 3.1, in which are listed many stable interactions that were not detected in the static models of the enzymes [d'Acierno et al., 2018; McCorvie et al., 2016], but that are conserved during the simulations and are present in both systems. The average number of intersubunit H-

bonds calculated per each timeframe is slightly higher in p.Gln188Arg than in wtGALT, but, interestingly, when the ligands are bound to the enzymes, the number of stable H-bonds is frequently higher in wtGALT than in p.Gln188Arg (table 3.1). This suggests that the ligands exert a stabilizing effect on the quaternary assembly of the wild-type enzyme and that this effect is lost in the mutant.

wtGALT	p.Gln188Arg	wtGALT + Ligands	p.Gln188Arg + Ligands
Average Number of H-Bonds per Timeframe: 27	Average Number of H-Bonds per Timeframe: 29	Average Number of H-Bonds per Timeframe: 30	Average Number of H-Bonds per Timeframe: 25
ILE32A-LYS120B	**ILE32A-LYS120B**	**ILE32A-LYS120B**	**ILE32A-LYS120B**
ILE32B-LYS120A	**ILE32B-LYS120A**	**ILE32B-LYS120A**	**ILE32B-LYS120A**
TYR34A-GLN118B	**TYR34A-GLN118B**	**TYR34A-GLN118B**	**TYR34A-GLN118B**
TYR34B-GLN118A	**TYR34B-GLN118A**	**TYR34B-GLN118A**	**TYR34B-GLN118A**
ILE198A-ALA343B	**ILE198A-ALA343B**	ARG48A-PHE99B	ARG48A-PHE99B
ILE198B-ALA343A	ARG48A-PRO100	**ILE198A-ALA343B**	ASP197A-GLN344B
HIS301A-LEU342B	HIS301A-LEU342B	ILE198B-ALA343A	SER45A-ALA101B
ASP197A-GLN334B	ASP197A-GLN334B	HIS47B-PRO100A	**ILE198A-ALA343B**
ARG48A-PHE99B	GLN30A-GLN103B	ASP197B-GLN344A	HIS47B-PRO100A
SER45A-ALA101B	GLN30A-ALA122B	**TRP41A-ASP197B**	**TRP41A-ASP197B**
GLN30A-GLN103B	**TRP41A-ASP197B**	TRP41B-ASP197A	TRP41B-ASP197A
GLN30A-ALA122B	HIS47A-PRO100B	GLN30B-GLN103A	GLN30B-GLN103A
TRP41A-ASP197B	**ARG228B-ASP113A**	GLN30B-ALA122A	GLN30B-ALA122A
HIS47A-PRO100B	ARG333B-GLU58A	ARG228A-ASP113B	ARG228A-ASP113B
ARG228B-ASP113A	ARG228A-ASP113B	**ARG228B-ASP113A**	**ARG228B-ASP113A**
ARG333B-GLU58A	ARG201A-ASP39B	GLY338A-SER297B	GLY338A-SER297B
SER45B-ALA101A	GLN103A-GLN30B	SER45B-ALA101A	SER45B-ALA101A
ARG51B-ASP98A	GLY338B-SER297A	GLY338B-SER297A	GLN224A-HIS114B
		ARG51B-ASP98A	ARG48B-PHE99A
		GLN30A-GLN103B	
		GLN30A-ALA122B	
		ARG48B-PHE99A	

Table 3.1. Pairs of residues involved in stable intersubunit H-bonds in simulations at 310 K.Bold: stable interactions present in all systems. Pairs of residues are considered to have a stable H-bond interaction if the sum of % of existence of the H-bonds between the two residues is >50.

We also calculated the intersubunit salt bridges present in the different simulations and found that their number is slightly higher in the mutant enzyme with respect to wtGALT in the absence of ligands, whereas the opposite is true in the presence of the ligands (table 3.2), as it happens for H-bonds (table 3.1). However, considering the low number of stable salt bridges detected during the simulations, it is not possible to deduce if this difference is significant. A salt bridge between Asp113 of chain A and

Arg228 of chain B visible in the static model, is conserved in the simulations. Instead, another intersubunit salt bridge that is visible in the starting structure (His114B-Glu 220A) [d'Acierno et al., 2018; McCorvie et al., 2016], is not preserved during all the simulations. Additional salt bridges between the Glu58 of chain A and the Arg333 of chain B and between the Asp98 of chain A and the Arg51 of chain B are stably present in p.Gln188Arg only in the absence of substrates.

wtGALT	p.Gln188Arg	wtGALT + Ligands	p.Gln188Arg + Ligands
GLU58A-ARG333B	GLU58A-ARG333B		
	ASP113B-ARG228A	ASP113B-ARG228A	ASP113B-ARG228A
ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B
ASP98A-ARG51B	ASP98A-ARG51B	ASP98A-ARG51B	

Table 3.2. Pairs of residues involved in stable intersubunit salt bridges in simulations at 310 K. Pairs of residues are considered to have a stable salt bridge interaction when the sum of % of existence of the salt bridges between the two residues is >50.

From these analyses it is possible to infer that the presence of the mutation is able to perturb the correct intersubunit interactions present in wtGALT, but this perturbation is more evident when the ligands are bound in the active site. Although the length of the simulation (200 ns) cannot allow for the detection of a complete destabilization of the enzyme, these effects appear to confirm what it was deduced from the analysis of the static structure, i.e., the replacement of Gln188 by Arg not only impairs the enzymatic activity, but also the stability of the quaternary assembly, and probably, the two effects are correlated.

3.1.2 Analysis of MD Simulations at 334 K

The quality checks of the trajectories obtained at higher temperature confirmed the good stabilization and the correct behavior of the systems during the simulations *(see Supplementary Files 5,6,7,8).*

Similarly to the equivalent systems at 310 K, the RMSF graphs of the simulations in the absence of the substrates (Figure 3.3 a, b) do not reveal a great enhancement in the flexibility of the protein structure, confirming that the most fluctuating parts of the

enzymes are the segments including residues ~50–70 and ~300–320 and again the fluctuation is not identical in the two subunits. In the mutant enzyme, differently from the simulations at 310 K, in the mutant enzyme, the presence of the ligands shows the fluctuation in these segments, with respect to the simulation at 334 K in the absence of ligands (Figure 3.3 c, d).

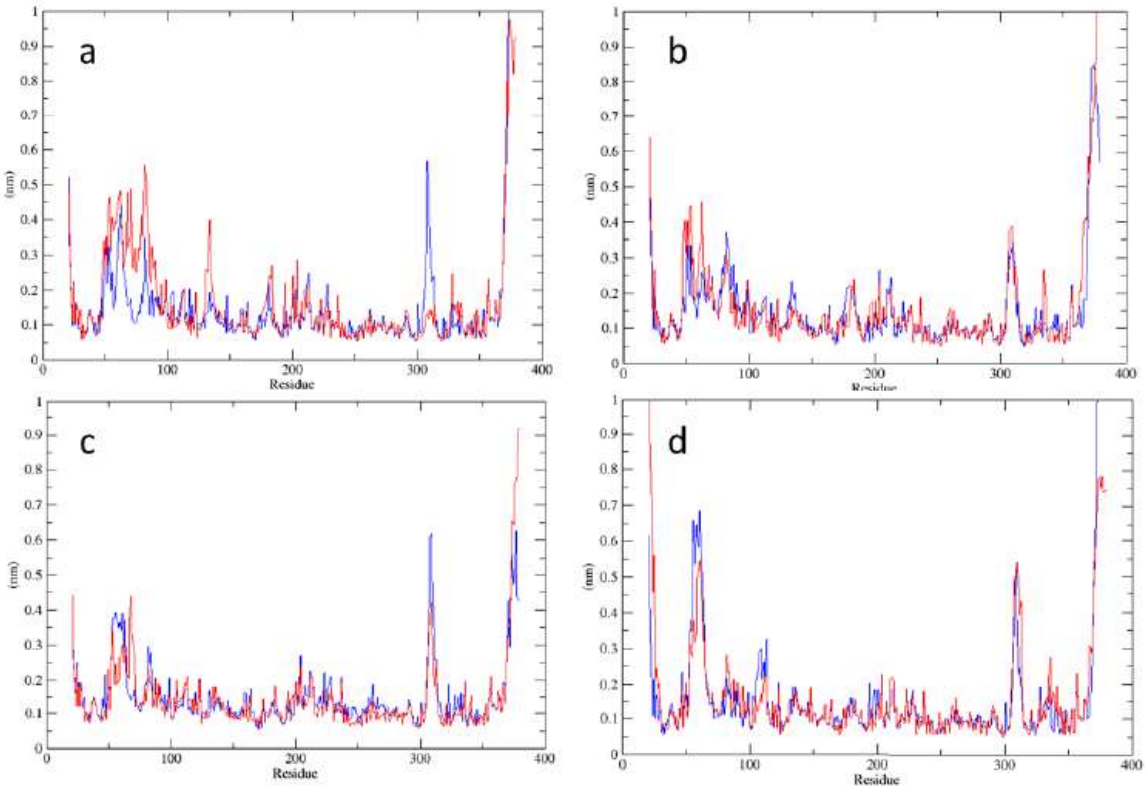


Figure 3.3 RMSF analysis for simulations at 334 K of: (a): wtGALT; (b): p.Gln188Arg; (c) wtGALT +G1P + H2U; (d) p.Gln188Arg+G1P + H2U. Blue lines represent RMSF fluctuation of chain A, red lines of chain B.

The analysis of secondary structures *(see Supplementary File 9, panels e–h)* reveals the formation of more irregular structures such as the π-helix (indicated as 5-helix in the graph produced by DSSP algorithm) in wtGALT, and a small increase in the number of disordered structures such as coils in p.Gln188Arg. The presence of the ligands seems to stabilize the secondary structures, since in wtGALT the irregular structures such as π helices are no longer detected and the number of residues in the coils is reduced in p.Gln188Arg; however, this last system is still more perturbed, given

that the presence of the π-helices is still detectable. Thus, as in the case of the simulations at lower temperature, in the mutant enzyme, the presence of the substrates is not able to rescue completely the destabilization.

Even at higher temperature, the ligands in the active site remain stably associated with the enzymes *(see Supplementary File 12 e–h).* The detailed interactions between the ligands and the residues of the active site are reported in Figure 3.4.

a)

b)

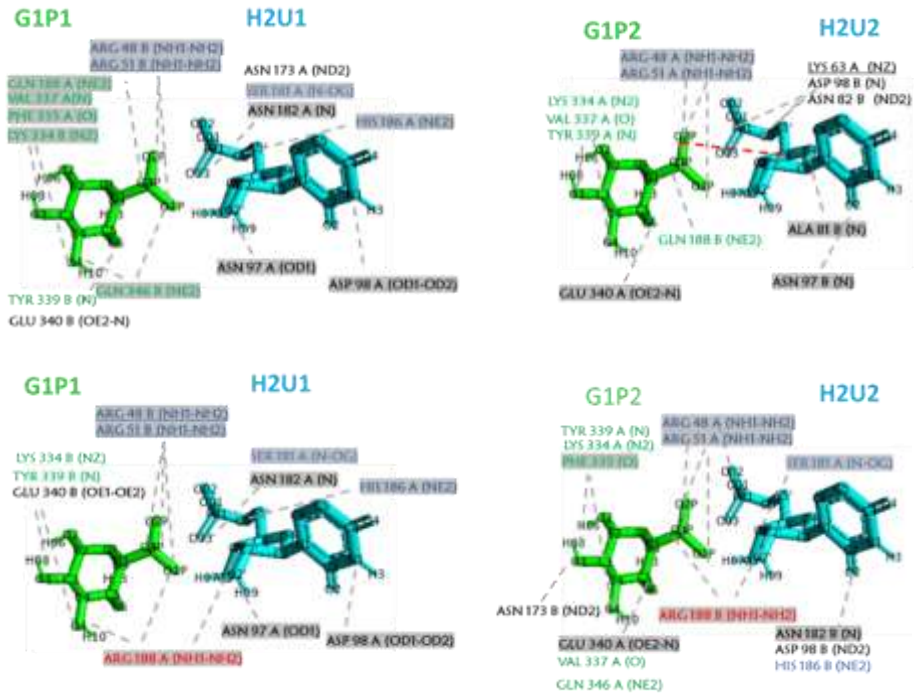


Figure 3.4 Interactions between enzyme and ligands in simulations at 334 K. (**a**) wtGALT; (**b**) p.Gln188Arg. Panels on the left indicate the ligands of the active site 1 (formally belonging to subunit A), panels on the right indicate the ligands of the active site 2 (formally belonging to subunit B). Grey background indicates interactions that persist for more than 50% of the simulation time. Grey dashed lines indicate H-bonds; residues underlined are also able to form salt bridges. Red dashed lines indicate interactions between the ligands

The persistent interactions between G1P and Arg48 or Arg51, predicted at 310 K both in wtGALT and p.Gln188Arg, are conserved even at this higher temperature, as well as the persistent interactions (hydrogen bonds and salt bridges) between H2U and Arg188. Therefore, also at high temperature it is possible to suppose that the mutant residue Arg188 is able to interfere with the correct enzymatic activity of GALT. Again, the interactions found in the two active sites are not identical. The analysis of the radius

of gyration *(see Supplementary File 10, panels e–h)* does not highlight the main differences between the two proteins, whereas the analysis of the SASA *(see Supplementary File 11, panels e–f)* shows that p.Gln188Arg has a higher value of this parameter (average value between 330 and 340 nm^2 compared to the average value between 290 and 300 nm^2 in wtGALT), indicating that the mutant protein tends to expose a higher area to the solvent, and this effect is clearly more visible at high temperature.

In the presence of the ligands, the SASA is slightly lower in wtGALT than in p.Gln188Arg and, for the mutant, is lower than the value predicted in the absence of the ligands *(see Supplementary File 11, panels g–h)*, reaching the same value of the simulations at 310 K.

The analysis of the interface interactions shows that, in the simulations in the absence of the ligands at higher temperatures, the average number of H-bonds per timeframe are practically unchanged in both systems with respect to the simulations at 310 K (Table 3.3). Contrarily to what happens in the systems at 310 K, the presence of the ligands in wtGALT does not increase this parameter, suggesting that at a higher temperature, they are no longer able to promote a further stabilization of the intersubunit interface. Instead, in p.Gln188Arg, this parameter is decreased, as it happens at body temperature, showing that the mutation still destabilizes this interface. In addition, at high temperature, many stable H-bond interactions, including some that are not detectable in the static structures, are formed during the simulations, and their number is higher in the absence than in the presence of the ligands, in both systems. The residues involved in these interactions are essentially the same predicted to interact at body temperature.

wtGALT	p.Gln188Arg	wtGALT + Ligands	p.Gln188Arg + Ligands
Average Number of H-Bonds per Timeframe: 27	Average Number of H-Bonds per Timeframe: 28	Average Number of H-Bonds per Timeframe: 27	Average Number of H-Bonds per Timeframe: 26
ILE32A-LYS120B	**ILE32A-LYS120B**	**ILE32A-LYS120B**	**ILE32A-LYS120B**
ILE32B-LYS120A	**ILE32B-LYS120A**	**ILE32B-LYS120A**	**ILE32B-LYS120A**
TYR34A-GLN118B	**TYR34A-GLN118B**	**TYR34A-GLN118B**	**TYR34A-GLN118B**
ASP197A-GLN344B	TYR34B-GLN118A	TYR34B-GLN118A	TYR34B-GLN118A
ILE198A-ALA343B	ASP197A-GLN344B	ARG48A-PHE99B	ARG48A-PHE99B
ARG48A-PHE99B	ILE198A-ALA343B	SER45B-ALA101A	SER45A-ALA101B
SER45A-ALA101B	GLN30A-GLN103B	ASP197A-GLN344B	ASP197B-GLN344A
HIS47B-PRO100A	GLN30A-ALA122B	ILE198A-ALA343B	ILE198B-ALA343A
HIS301B-LEU342A	**ARG228A-ASP113B**	HIS301A-LEU342B	HIS301B-LEU342A
TRP41B-ASP197A	**ARG228B-ASP113A**	GLY338A-SER297B	GLY338A-SER297B
GLY338B-SER297A	**TRP41B-ASP197A**	**GLY338B-SER297A**	**GLY338B-SER297A**
ARG333B-GLU58A	**GLY338B-SER297A**	ARG48B-PHE99A	**ARG228A-ASP113B**
GLN103B-GLN30A	ARG333A-GLU58B	GLN30B-GLN103A	**ARG228B-ASP113A**
GLN30A-ALA122B	SER297B-VAL337A	GLN30B-ALA122B	ARG48B-PHE99A
ARG228B-ASP113A	TRP41A-ASP197B	TRP41A-ASP197B	TRP41A-ASP197B
ARG228A-ASP113B	GLN169B-LEU342A	**TRP41B-ASP197A**	**TRP41B-ASP197A**
GLN30B-GLN103A	SER45B-ALA101A	**ARG228A-ASP113B**	ALA122A-GLN30B
GLN30B-ALA122A	GLN56A-PRO59B	**ARG228B-ASP113A**	
ARG201B-ASP39A	ARG48B-GLN103A		
HIS47A-PRO100B			

Table 3.3 Pairs of residues involved in stable intersubunit H-bonds in simulations at 334 K. Bold: stable interactionspresent in all systems. Pairs of residues are considered to have a stable H-bond interaction when the sum of % of existence of the H-bonds between the two residues is >50.

The number of salt bridges is identical both between the two systems in the absence of ligands and between the two systems in the presence of ligands and is lower in the latter case (Table 3.4). Again, the number of these interactions is so low that it is not possible to state if this difference is significant or not. In all systems, the salt bridge between Asp113 and Arg228 is conserved, whereas the interaction between the Glu58 of chain A and the Arg333 of chain B is lost in the presence of the substrates, as it was the case for simulations at 310 K (table 2). From these results, it appears that at 334 K, both systems are not dramatically perturbed, although some differences are present especially in the flexibility of the most mobile segments and in the secondary structures. The temperature seems not to perturb drastically the quaternary assembly of the enzymes, but the ability of the substrates to stabilize especially the intersubunit H-bonds seems to be lost in wtGALT at higher temperatures, and in p.Gln188Arg in all conditions.

wtGALT	p.Gln188Arg	wtGALT + Ligands	p.Gln188Arg + Ligands
GLU58A-ARG333B	GLU58A-ARG333B		
ASP113B-ARG228A	ASP113B-ARG228A	ASP113B-ARG228A	ASP113B-ARG228A
ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B

Table 3.4 Pairs of residues involved in stable intersubunit salt bridges in simulations at 334 K. Pairs of residues are considered to have a stable salt bridge interaction when the sum of % of existence of the salt bridges between the two residues is >50.

3.2 Arginine as a possible pharmacochaperone for GALT

As discussed in the paragraph 1.6.4, first the apparent ability, and then the apparent failure of arginine to act as a pharmacochaperone for GALT, prompted us to apply computational simulations in order to understand, at the molecular level, the possible interactions between the enzyme and this amino acid, in an effort to predict the putative effect (if any) of arginine on the GALT enzyme.

3.2.1 Docking simulations

First of all, a preliminary docking study, performed with a blind approach, showed that arginine tends to interact mainly with the active sites of both wtGALT and p.Gln188Arg mutant protein, with a preference for active site A (paragraph 1.3.2, figure 1.12). Additionally, the same docking study shows the interaction of arginine with other aspecific sites on the protein surface, without any clusterization (data not shown).

Therefore, we considered the active site as a first putative target for the binding of arginine. Targeting the active site of the enzyme could have an effect on the overall stability of the structure of GALT, given that the two active sites of the protein are at the interface between the two subunits forming the quaternary assembly and are formed by residues belonging to both monomeric chains. Additionally, the mutant p.Gln188Arg shows a dominant negative effect due to the perturbation of the intersubunit interface caused by the mutation [Marabotti et al., 2005, McCorvie et al., 2016]. Therefore, as a first approach, we docked arginine in the active site A of both wtGALT and p.Gln188Arg, in the presence or in the absence of the natural substrates,

to predict possible effects induced by this amino acid on the enzyme's quaternary assembly and stability. Additionally, given that the central cavity of the GALT enzyme represents a very interesting target for putative PCs [Timson et al., 2016], we also decided to simulate the possibility that arginine could bind to this portion of the enzyme. The central cavity is formed by residues belonging to both subunits, creating important networks of interactions [d'Acierno et al., 2018; McCorvie et al., 2016] and our goal was to detect if the presence of arginine in this position can influence, either favorably or unfavorably, the activity of the enzyme.

The two starting structures of wtGALT and p.Gln188Arg were analyzed with the CASTp 3.0 Web server (*http://sts.bioe.uic.edu/castp/calculation.html*; last accessed 5 October 2021) [Tian et al., 2018], which identified three main cavities: the two active sites and the central cavity, with an area of 1772.5 $Å^2$ and a volume of 2442.6 $Å^3$, formed by 67 residues, of which 32 belong, formally, to the subunit A and 35 to subunit B (Figure 3.5).

A

RESIDUES OF CENTRAL CAVITY	
CHAIN A	CHAIN B
ASN35, LEU37, GLN38, ASP39, GLU40, TRP41, VAL42, GLN56, GLU58, MET177, MET 178, ASP197, ILE198, ARG201, TRP246, ALA247, THR248, TRP249, GLN252, HIS301, HIS321, TYR323, PRO325, ARG333, PHE335, MET336, VAL337, GLU340, MET341, LEU 342, GLN344, ALA345.	ASN35, LEU37, GLN38, GLU40, TRP41, VAL42, GLN56, GLU58, MET177, MET 178, ASP197, ILE198, GLN200, ARG201, ARG204, PRO244, TRP246, ALA247, THR248, TRP249, GLN252, THR253, LEU254, HIS301, HIS321, TYR323, ARG333, PHE335, MET336, VAL337, GLU340, MET341, LEU342, GLN344, ALA345.

B

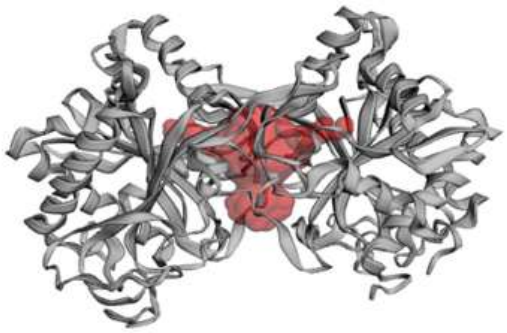


Figure 3.5: A. Table listing the residues of the central cavity of wtGALT detected by the CASTp 3.0 servers: the underlined residues are part of active site; B. representation of the central cavity of wtGALT with a front view. Image obtained by CASTp 3.0 [Tian et al., 2018].

Then, we performed docking simulations targeting the whole central cavity, and also in this case, we simulated the binding of arginine in the presence and in the absence of both natural substrates in the active site. The docking results for arginine in the different

conditions simulated are reported in table 3.5. They show the different poses corresponding to each result and the details of their interactions with the residues of the protein *(see Supplementary Files 13, 14).*

System	Binding Energy of the Representative Pose	Number of Poses in the Cluster	Predicted Interactions with Residues
wtGALT + arginine (active site)	−5.25	35	R48, R51, R333, K334, F335, V337, E340, D348
p.Gln188Arg + arginine (active site)	−5.26	36	R48, R51, R188, R333, K334, F335, V337, E340, D348
wtGALT + G1P + arginine (active site)	−5.55	87	R48, N97, D98, G1P
p.Gln188Arg + G1P + arginine (active site)	−5.22	84	R48, N97, D98, F99, R188, G1P
wtGALT + H2U + arginine (active site)	−6.96	72	R48, R51, G179, S181, K334, F335, V337, E340, Q346, H2U
p.Gln188Arg + H2U + arginine (active site)	−6.76	90	R48, R51, N173, G179, R188, K334, F335, V337, G338, Y339, E340, H2U
wtGALT + arginine (central cavity)	−4.73	26	Q38, E40, D197, R201
p.Gln188Arg + arginine (central cavity)	−5.00	15	Q38, E40, T248, Y323, M341
wtGALT + G1P + H2U + arginine (central cavity)	−5.19	20	Q38, E40, W41, D197, R201
p.Gln188Arg + G1P + H2U + arginine (central cavity)	−5.09	19	Q38, E40, M341, Q344, A345

Table 3.5. Docking results for arginine used as starting point for MD simulations

The energies of interactions predicted for arginine are negative (favorable) for all the systems, although their absolute values are not that high, indicative of the fact that arginine does not interact strongly with the protein. They are lower (more favorable) in those systems in which H2U alone is present in the active site. In these systems, arginine interacts with the residues of the active site (in particular, Arg48 and Lys334, which interact with the negatively charged part of the amino acid, and Glu340, Ser181, and Arg51, which form H-bonds with the polar groups of arginine) and with a strong, favorable interaction with the phosphate group of H2U. When the active site of the enzyme is partly occupied by G1P, the positions of arginine seem slightly different in the two systems. Indeed, in wtGALT, arginine interacts with the residues Glu172, Asn173, and Ser181, the catalytic residue His186, and Gln188; on the contrary, in the mutant p.Gln188Arg, there is an unfavorable interaction with the residue Arg188 that

probably displaces arginine towards Arg48, Asn 97, and Asp98. In both cases, there is also a favorable interaction with the phosphate group of G1P.

We also simulated the condition in which the active site of the GALT enzyme and of mutant p.Gln188Arg are occupied by both ligands, but as expected, arginine cannot enter in it and stays on the protein surface, contacting a portion of the external part of the enzyme, with a predicted binding energy significantly higher than that obtained in the other simulated conditions (data not shown).

The docking results for the central cavity of the enzymes gave less defined results than those in the active site, because the cavity is very large and, thus, arginine has a higher conformational freedom. However, all the simulations predicted a negative binding energy, suggesting the possibility that arginine could also bind to this cavity. In these systems, arginine frequently binds to Gln38 and Glu40, with occasional contacts with Asp197, Arg 201, Thr248, Met341, and Gln344. The predicted binding energies in all these conditions seem not to be significantly different, indicating that neither the mutation nor the presence of the substrate in the active site would affect the binding of arginine in the central cavity.

3.2.2 MD Simulations — Arginine in the active site

MD simulations were performed at 310 K and for 100 ns, a timescale in which it is possible to evaluate if arginine remains or not in the active site. The starting point for the MD simulations was the best docking results, reported in Table 3.5. The analyses of the energetic components, of the minimum distance of the periodic images, and of the RMSD of the atom distances for these simulations showed that the systems reached stabilization and that no major perturbation affected them *(see Supplementary Files 25, 26, 27,28,29,30).* From the data obtained by the two different replicas of the simulations, it appears that arginine is not bound stably to the active sites of both the wild type and the mutant enzyme, irrespective of the absence or the presence of either ligand *(see Supplementary File 17).* When arginine binds into the active site and ligands are not present, arginine occupies the cavity that hosts G1P (the same identified

with the docking simulation) and binds to residue Arg48 and to residues belonging to loop 334–340 in both wt GALT and p.Gln188Arg (Figure 3.6 a, b). When G1P alone is in the active site, arginine occupies the place that usually hosts H2U, but interacts only with Asp98 in wtGALT, in addition to the phosphate group of G1P itself (Figure 3.6 c). In p.Gln188Arg, the interactions are made with Gln54, His186, and Arg188 (Figure 3.6 d). When H2U alone is in the active site of wtGALT, arginine is hosted again in the cavity of G1P and interacts with the same residues listed above. Additionally, arginine also interacts with the phosphate group of H2U. We observed that the presence of a molecule of arginine in the active site of p.Gln188Arg determines the creation of a cluster of positive charges that perturbs not only the interactions that H2U can keep with the active site, but also the binding of arginine itself. In all these simulations, G1P and H2U remained stably bound to both wtGALT and p.Gln188Arg *(see Supplementary File 18).* G1P stably interacts with Arg48 and Arg51, and, additionally, with residue 188 and residues 339 and 340, in both wtGALT and p.Gln188Arg (Figure 3.6 c,d). H2U in wtGALT is in contact with Asn97, Asp98, and His186, whereas in p.Gln188Arg, it contacts Arg48, Arg51, and Arg188 (Figure 3.6 e,f). The replacement of Gln188 with Arg is able to perturb the pattern of interactions of the substrate [Marabotti et al., 2005]. Concerning the global state of the systems, the radius of gyration was constant during these simulations *(see Supplementary File 19).* The SASA appears to be affected by the binding of arginine in the active site: this parameter tends to increase when arginine moves away from the active site *(see Supplementary File 20).*

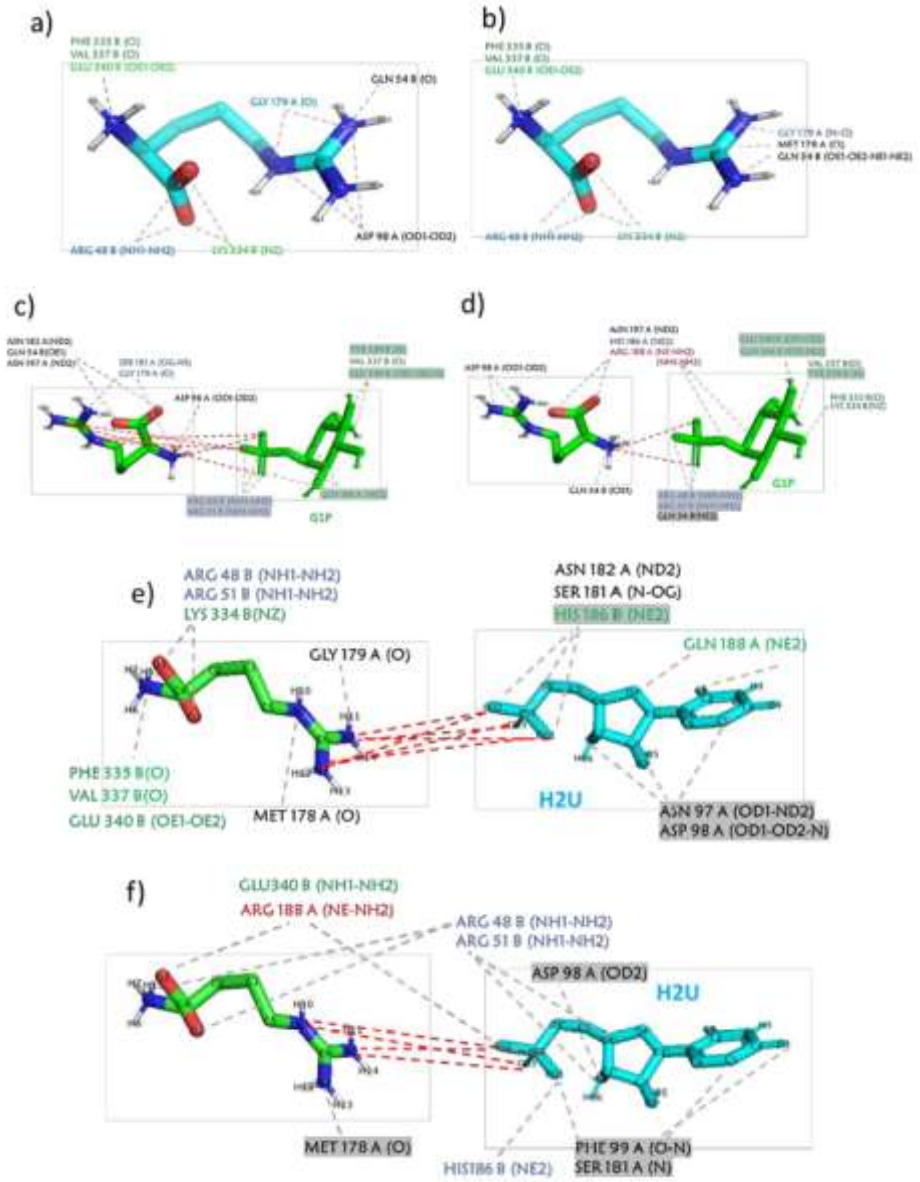


Figure 3.6. Interactions of arginine in the active site with enzyme and ligands. (a) wtGALT +arginine; (b) p.Gln188Arg + arginine; (c) wtGALT + G1P + arginine; (d) p.Gln188Arg + G1P + arginine; (e) wtGALT + H2U + arginine; (f) p.Gln188Arg +H2U + arginine. Gray background indicates interactions that persisted for more than 50% of the simulation time. Gray dashed lines indicate H-bonds. Red dashed lines indicate interactions between the ligands.

The analysis of the evolution of the secondary structures (Figure 3.7) shows that, in the presence of arginine alone, wtGALT shows a slightly higher presence of irregular structures such as the π-helix with respect to the mutant enzyme. No differences are detectable in both systems in the presence of G1P. In the presence of H2U, irregular structures are detected in the mutant enzyme when arginine is bound to the active site.

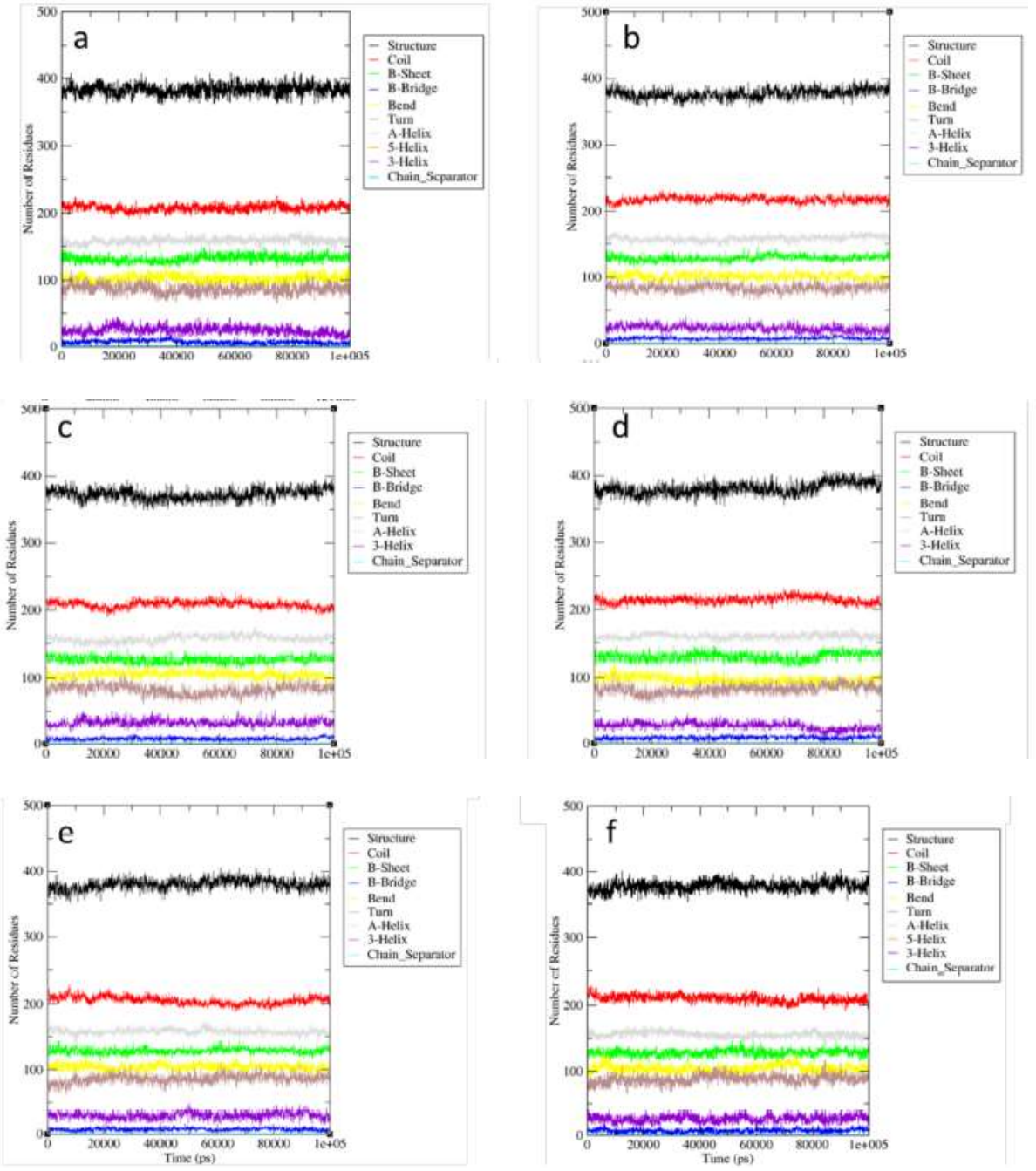


Figure 3.7. Representative results of DSSP analysis for simulations with arginine in the active site. (a) wtGALT + arginine; (b) p.Gln188Arg + arginine; (c) wtGALT + G1P + arginine; (d) p.Gln188Arg + G1P + arginine; (e) wtGALT + H2U + arginine; (f) p.Gln188Arg+ H2U + arginine.

Comparing the RMSF graphs (Figure 3.8), the main variability in all the simulations seems to be focused on the same segments already shown in simulations discussed in the paragraph 3.1.1 and 3.1.2., i.e., mainly segments 50–70 (including segment 50–60 formed by very conserved residues at the intersubunit interface) and 300–320 (a long

loop including the conserved residues of the Zn-binding site). For segment 300–320, there is an asymmetrical flexibility of the two chains, more evident in the simulation with G1P for wt hGALT and with H2U for p.Gln188Arg.

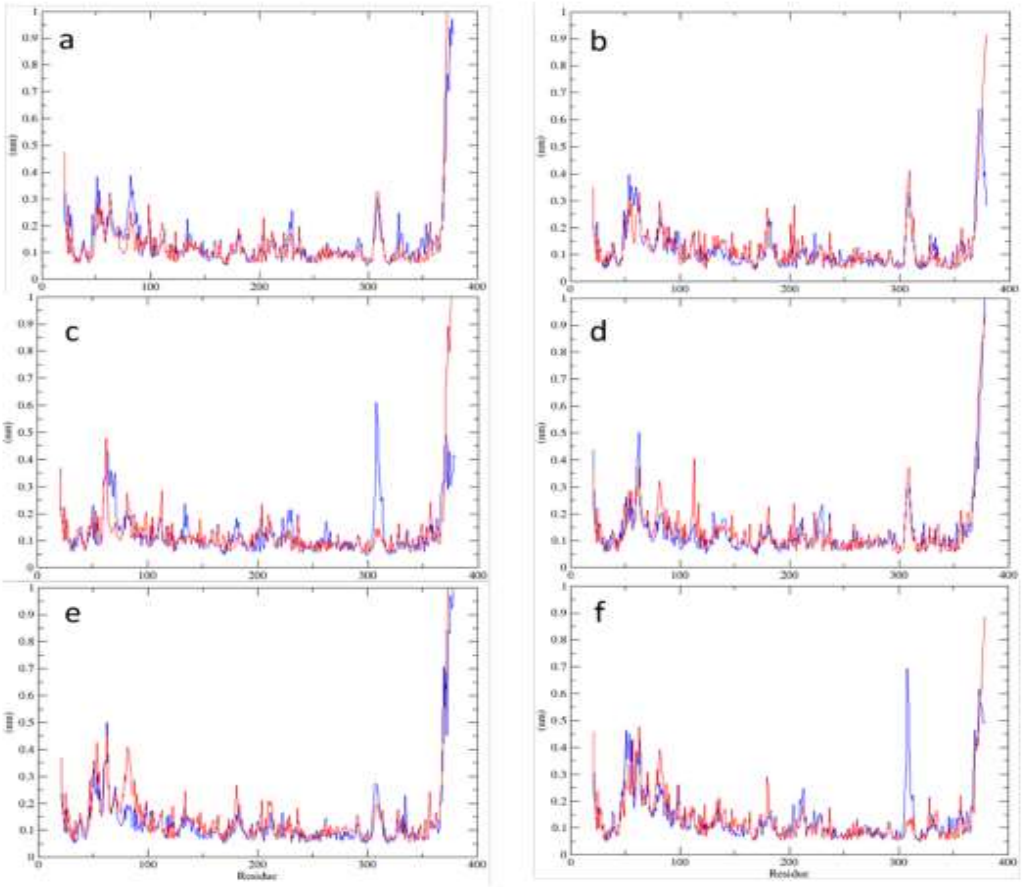


Figure 3.8. Representative results of RMSF analysis for simulations with arginine in the active site. (a) wtGALT + arginine;(b) p.Gln188Arg + arginine; (c) wtGALT + G1P + arginine; (d) p.Gln188Arg + G1P + arginine; (e) wtGALT + H2U + arginine;(f) p.Gln188Arg + H2U + arginine. Blue lines represent RMSF fluctuation of chain A; red lines, that of chain B.

The analysis of the stable intersubunit interactions is reported for H-bonds in Table 3.6. Almost half of the stable intersubunit H-bonds monitored were also present in the static models, whereas the others were formed during the simulations. The average number of intersubunit H-bonds per timeframe was very similar in the simulations involving wtGALT with respect to the equivalent simulations involving the mutant enzyme, also considering the variation between the two different replicas of each simulation.

This parameter is constantly lower than the same parameter obtained for simulations in the absence of arginine (see the paragraph 3.1). The most notable difference is visible in the simulations of arginine in the presence of H2U, where the average number of H-bonds per timeframe was higher in wtGALT than in p.Gln188Arg. These data show that arginine does not have a favorable effect on the intersubunit contacts in the mutant enzyme; rather, it seems to perturb the intersubunit interactions in these systems.

The analysis of the stable intersubunit salt bridges during the simulations is reported in Table 3.7. Only a few stable interactions of this type were present in the systems during the simulations, and most of them involved the residue Asp113 of one chain and Arg 228 of the other chain. The presence of arginine in the active site seems not to influence their existence. In contrast with the results obtained for H-bonds, the simulation of arginine in the active site of p.Gln188Arg bound to H2U is the one with the highest number of salt bridges (3), but given this low number of interactions, it is difficult to consider this difference as significant.

From the results of these simulations, it seems that, if arginine binds into the active site of the mutant enzyme, it is not able to counteract the loss of activity; rather, it could even worsen it, as in the case of the simulation of arginine in the active site of the mutant enzyme when H2U is also bound to the site.

wtGALT + Arginine	p.Gln188Arg + Arginine	wtGALT + G1P + Arginine
Average Number of H-Bonds per Timeframe: 26/26	Average Number of H-Bonds per Timeframe: 25/25	Average Number of H-Bonds per Timeframe: 26/22
ILE32A-LYS120B	*ILE32A-LYS120B*	*ILE32A-LYS120B*
ILE32B-LYS120A	*ILE32B-LYS120A*	*ILE32B-LYS120A*
TYR34A-GLN118B	**TYR34A-GLN118B**	*TYR34A-GLN118B*
TYR34B-GLN118A	*TYR34B-GLN118A*	*TYR34B-GLN118A*
ILE198A-ALA343B	*ILE198A-ALA343B*	*ILE198A-ALA343B*
ILE198B-ALA343A	*ILE198B-ALA343A*	*ILE198B-ALA343A*
ASP197A-GLN344B	*ASP197A-GLN344B*	*ASP197A-GLN344B*
ARG48A-PHE99B	*ASP197B-GLN344A*	*HIS301A-LEU342B*
TRP41A-ASP197B	ARG48A-PHE99B	**TRP41A-ASP197B**
TRP41B-ASP197A	*SER45A-ALA101B*	*TRP41B-ASP197A*
ARG228A-ASP113B	HIS47A-PRO100B	*ARG228B-ASP113A*
ARG228B-ASP113A	*HIS301B-LEU342A*	GLN30A-ALA101B
ARG204A-ASP39B	*TRP41A-ASP197B*	**GLY338A-SER297B**
ARG201A-ASP39B	**TRP41B-ASP197A**	*GLY338B-SER297A*
ARG333B-GLU58A	*ARG228A-ASP113B*	GLN56A-GLU58B
ARG48B-PHE99A	**ARG228B-ASP113A**	ARG51A-ASP98B
GLY338A-SER297B	ARG48B-PRO100A	ARG201A-ASP39B
ALA122A-GLN30B	GLN30A-ALA122B	ARG333A-GLU58B
	GLN30A-GLN103B	*ARG48B-PHE99A*
	GLY338A-SER297B	SER45B-ALA101A
	ARG204A-ASP39B	SER45A-ALA101B
	GLN169B-ILE32A	GLN103A-GLN30B
		TYR339A-SER192B
		GLN30B-ALA122A

p.Gln188Arg + G1P + Arginine	wtGALT + H2U + Arginine	p.Gln188Arg + H2U + Arginine
Average Number of H-Bonds per Timeframe: 24/27	Average Number of H-Bonds per Timeframe: 29/27	Average Number of H-Bonds per Timeframe: 23/26
ILE32A-LYS120B	*ILE32A-LYS120B*	*ILE32A-LYS120B*
ILE32B-LYS120A	*ILE32B-LYS120A*	*ILE32B-LYS120A*
TYR34A-GLN118B	*TYR34A-GLN118B*	*TYR34A-GLN118B*
TYR34B-GLN118A	*TYR34B-GLN118A*	*TYR34B-GLN118A*
ILE198A-ALA343B	*ILE198A-ALA343B*	*ILE198A-ALA343B*
ILE198B-ALA343A	**ILE198B-ALA343A**	*ILE198B-ALA343A*
ASP197B-GLN344A	*ARG48A-PHE99B*	*ASP197B-GLN344A*
HIS301A-LEU342B	HIS301B-LEU342A	*TRP41A-ASP197B*
HIS301B-LEU342A	*TRP41A-ASP197B*	*TRP41B-ASP197A*
TRP41A-ASP197B	*TRP41B-ASP197A*	*ARG228A-ASP113B*
TRP41B-ASP197A	*ARG228A-ASP113B*	*ARG228B-ASP113A*
ARG228B-ASP113A	**ARG228B-ASP113A**	*ARG201B-ASP39A*
GLN30A-GLN103B	*GLN30A-GLN103B*	*ARG204A-ASP39B*
GLN30A-ALA122B	GLN30A-ALA122B	GLY338B-SER297A
ARG48B-PRO100A	ARG201B-ASP39A	**GLY338A-SER297B**
TYR339A-SER192B	*GLN103A-GLN30B*	GLN30A-GLN103B
SER45B-ALA101A	ARG201A-ASP39B	SER45A-ALA101B
ARG204A-ASP39B	GLN30B-ALA122A	ARG48A-PHE99B
ARG228A-ASP113B	*ARG48B-PHE99A*	ARG201A-ASP39B
GLY338A-SER297B	ARG51B-ASP98A	GLN30B-GLN103A
HIS114B-GLU220A	*GLY338A-SER297B*	
ARG201B-GLN38A	SER45A-ALA101B	
	SER45B-ALA101A	
	ASP197A-GLN344B	
	ARG204A-ASP39B	
	HIS301A-LEU342B	

Table 3.6. Pairs of residues involved in stable intersubunit H-bonds in simulations with arginine in the active site. The average number of H-bonds per timeframe is reported for each replica. Dark gray background: intersubunit interactions identified in the initial models and conserved throughout the simulations; bold: stable interactions present in all systems; italics: interactions maintained in the two replicas of each system. Pairs of residues are considered to have a stable H-bond interaction when the sum of % of existence of the H-bonds between the two residues is >50.

wtGALT + Arginine	p.Gln188Arg + Arginine	wtGALT + G1P + Arginine	p.Gln188Arg + G1P + Arginine	wtGALT + H2U + Arginine	p.Gln188Arg + H2U + Arginine
GLU58B-ARG333A					
ASP113B-ARG228A	ASP113B-ARG228A	ASP113B-ARG228A		ASP113B-ARG228A	ASP113B-ARG228A
	ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B
ARG204A-ASP39B			ARG204A-ASP39B		ASP39A-ARG201B

Table 3.7. Pairs of residues involved in stable intersubunit salt bridges in simulations with arginine in the active site. Pairs of residues are considered to have a stable salt bridge interaction when the sum of % of existence of the salt bridges between the two residues is > 50.

3.2.3 MD Simulations—Arginine in the central cavity

As for the simulations with arginine in the active site, also for these systems the starting point for MD simulations was the best docking results of arginine in the central cavity (see *Supplementary Files 14, 15*) reported above in Table 1. The MD simulations were conducted at 310 K, but given the size of the central cavity, we decided for these systems to run 200 ns-long simulations, in order to allow arginine to perform a deeper exploration of the conformational space.

All the analyses of the energetic components (including the total energy, kinetic energy component, potential energy component, pressure, temperature, volume and density), of the minimum distance of the periodic images, of the RMSD, and of the atom distances for these simulations showed that the systems reached stabilization and that no major perturbation affected them *(see Supplementary Files 31, 32, 33, 34).*

In all the simulations, arginine stably interacted with the protein, both in the presence and in the absence of the substrates *(see Supplementary File 21).* Additionally, G1P and H2U, in turn, stably interacted with the enzyme for all the simulations *(see Supplementary Files 22).* The detailed interactions of arginine and the substrates in these simulations are represented in Figure 3.9. In the absence of the ligands, as for the docking simulations, arginine mainly interacted with two negatively charged residues belonging to the two protomers of the enzyme, i.e., Glu40 and Asp197, which mainly formed interactions with the guanidinium group of the amino acid. These residues were located in proximity to the Zn-binding site, in a cavity that was putatively identified as

an allosteric site for the enzyme [McCorvie et al., 2016]. The interaction with these residues seems to be more stable and persistent in wtGALT than in p.Gln188Arg (Figure 3.9 a, b). In the presence of the ligands, arginine remained in stable contact with Glu40 and occasionally interacted with the other residues identified in the docking simulations (Figure 3.9 c, d). G1P maintained H-bonds and salt bridges with the residues Arg48 and Arg51, also seen in the absence of the arginine (see the paragraph 3.1). Moreover, other interactions were maintained with residues belonging to the flexible loop 335-340. Additionally, it was possible to detect an interaction with Arg188 in the mutant p.Gln188Arg. H2U was mainly bound to Asn97 and Asp98 in wtGALT, and to His186 and Arg188 in p.Gln188Arg. Thus, as seen for the simulations of the GALT enzyme in the absence of arginine and in the previous simulation with arginine bound to the active site, the presence of the mutation is able to perturb the interactions of H2U with the active site residues (Figure 3.9 c,d), but the presence of arginine seems not to be able to modify this situation.

Concerning the global state of the systems, the radius of gyration was constant throughout the simulations, indicating that the protein did not change its shape *(see Supplementary File 23).* In the presence of arginine, the SASA of both wtGALT and p.Gln188Arg shows a decreasing trend, whereas in the presence of the substrates, wt GALT shows an increasing trend *(see Supplementary File 24).* The analysis of the secondary structures by means of DSSP (Figure 3.10) showed no significant differences in the presence of arginine only, whereas when ligands were also bound to the active site, there was a higher content of irregular structures such as the π-helix (indicated as a 5-helix in the graph) in the p.Gln188Arg system.

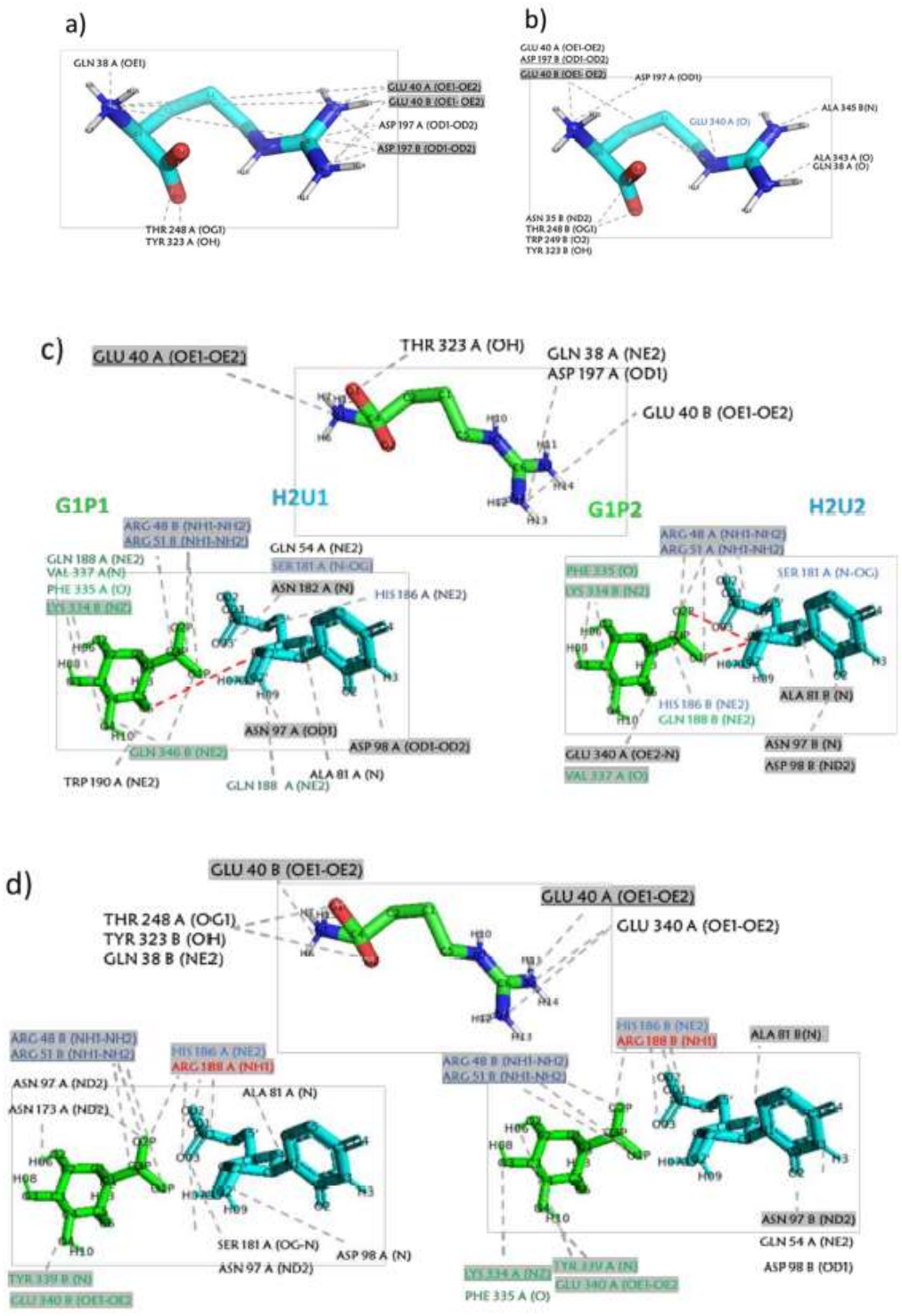


Figure 3.9. Interactions of arginine in the central cavity with enzyme and ligands. (a) Interactions of arginine in the simulation wtGALT + arginine; (b) Interactions of arginine in the simulation p.Gln188Arg + arginine; (c) Interactions of arginine (top) and of the substrates in the two active sites(bottom) in the simulation wtGALT + G1P + H2U + arginine; (d) Interactions of arginine (top) and ofthe substrates in the two active sites (bottom) in the simulation p.Gln188Arg + G1P + H2U + arginine. Gray background indicates interactions that persisted for more than 50% of the simulation time. Graydashed lines indicate H-bonds. Red dashed lines indicate interactions between the ligands.

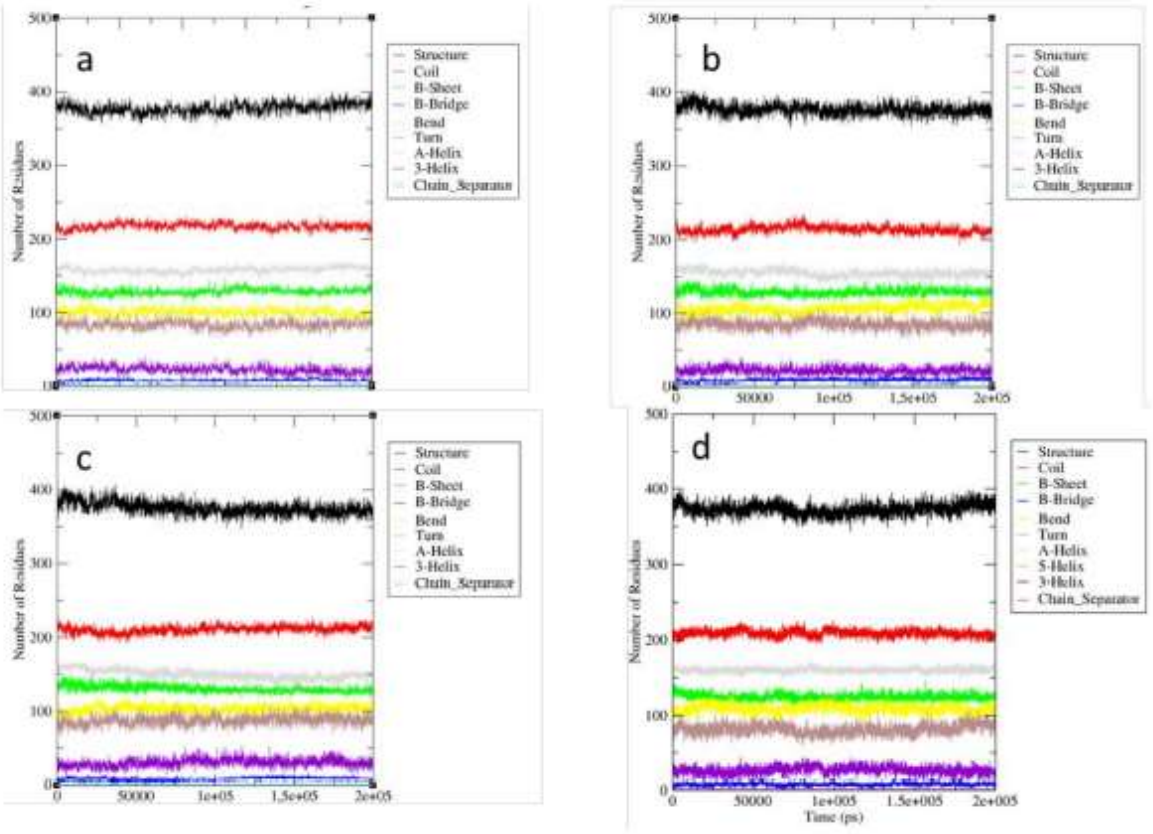


Figure 3.10. DSSP analysis for simulations with arginine in the central cavity. (a) wtGALT + arginine; (b) p.Gln188Arg +arginine; (c) wtGALT + G1P + H2U + arginine; (d) p.Gln188Arg + G1P + H2U + arginine.

Finally, the analysis of the RMSF (Figure 3.11) showed that, apart from N- and especially C-terminals, the more flexible segments of the protein were those around residues 40, 200 and 320, including portions of the active site. In the presence of arginine only, the mutant p.Gln188Arg showed higher flexibility in the segment around the position of the mutation, whereas, when ligands were also present, the flexibility of the mutant seemed to be decreased, contrarily to what happens to wtGALT. Similarly to previous simulations, in these graphs, it is also possible to detect an asymmetry in the flexibility of the two chains, especially concerning the segment 300–320.

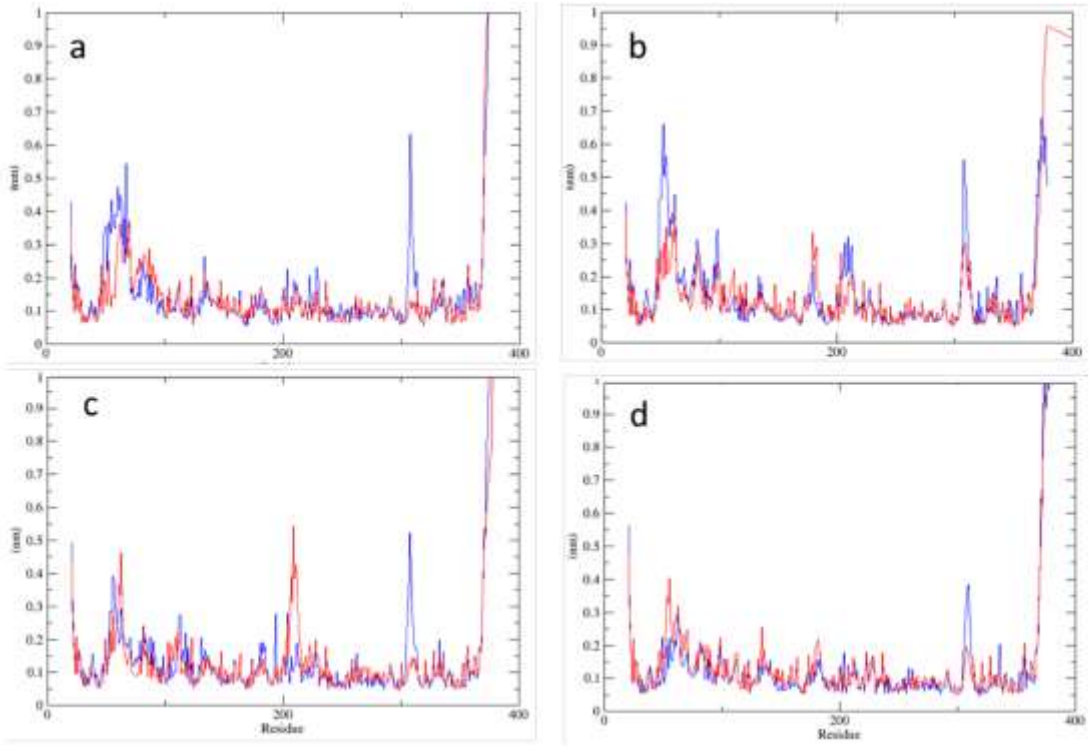


Figure 3.11. RMSF analysis. Simulations with arginine in the central cavity. (a) wtGALT + arginine; (b) p.Gln188Arg + arginine;(c) wtGALT + G1P + H2U + arginine; (d) p.Gln188Arg + G1P + H2U + arginine. Blue lines represent RMSF fluctuation of chain A; red lines, that of chain B.

We monitored the variation of stable intersubunit H-bonds in wtGALT and p.Gln188Arg during these simulations. The results are reported in Table 3.8. Several intersubunit H-bonds that were identified previously in the static models of wtGALT and of p.Gln188Arg [d'Acierno et al., 2018, McCorvie et al., 2016] appeared to be stable, and some of them were conserved in all the different systems. However, as we found in our results in the absence of arginine (see paragraph 3.1), several other persistent H-bonds, which were not detectable in the static models, were formed during the simulations and contributed to stabilizing the intersubunit interface. When arginine was present in the central cavity, the average number of H-bonds per timeframe was identical in wtGALT and in p.Gln188Arg, but when ligands were also in the active site, the mutant enzyme showed a notable increase in this parameter. Thus, it seems that arginine bound to the central cavity is able to tighten these interactions between the two subunits.

wtGALT + Arginine	p.Gln188Arg + Arginine	wtGALT+ Ligands + Arginine	p.Gln188Arg + Ligands + Arginine
Average Number of H-Bonds per Timeframe: 26	Average Number of H-Bonds per Timeframe: 26	Average Number of H-Bonds per timeframe: 25	Average Number of H-Bonds per Timeframe: 30
ILE32A-LYS120B	**ILE32A-LYS120B**	**ILE32A-LYS120B**	**ILE32A-LYS120B**
ILE32B-LYS120A	**ILE32B-LYS120A**	**ILE32B-LYS120A**	**ILE32B-LYS120A**
TYR34A-GLN118B	**TYR34A-GLN118B**	**TYR34A-GLN118B**	**TYR34A-GLN118B**
TYR34B-GLN118A	**TYR34B-GLN118A**	**TYR34B-GLN118A**	**TYR34B-GLN118A**
ILE198A-ALA343B	ILE198B-ALA343A	ILE198B-ALA343A	ILE198B-ALA343A
ASP197A-GLN344B	ASP197B-GLN344A	ILE198A-ALA343B	ILE198A-ALA343B
ARG48A-PHE99B	HIS301A-LEU342B	ASP197B-GLN344A	HIS301B-LEU342A
HIS301A-LEU342B	SER297A-VAL337B	ARG48A-PHE99B	TRP41B-ASP197A
HIS114B-GLU220A	SER45A-ALA101B	HIS301A-LEU342B	ARG201A-ASP39B
HIS47A-PRO100B	**TRP41A-ASP197B**	**TRP41A-ASP197B**	**TRP41A-ASP197B**
GLY338A-SER297B	GLY338A-ASN173B	TRP41B-ASP197A	**ARG228B-ASP113A**
GLN30A-ALA101B	GLN30A-ALA122B	ARG201A-ASP39B	GLN30A-GLN103B
GLN103A-GLN30B	ARG48A-PRO100B	**ARG228B-ASP113A**	GLN30A-ALA122B
GLN30A-GLN103B	ARG204A-ASP39B	SER45B-ALA101A	HIS47A-PHE99B
ARG333A-GLN58B	**ARG228B-ASP113A**	TRP167B-TYR339A	ARG204A-ASP39B
ARG201A-ASP39B	ARG201A-GLU40B	ARG228A-ASP113B	SER297A-VAL337B
ARG228B-ASP113A	GLN103A-GLN30B		GLN346A-ALA101B
TRP41A-ASP197B	GLN30A-GLN103B		GLN30B-ALA101A
TRP41B-ASP197A	ARG228A-ASP113B		ARG48B-PHE99A
GLY338B-SER297A	ARG201B-ASP39A		ARG51B-ASP98A
ARG48B-PHE99A			ARG201B-ASP39A
GLN224A-HIS114B			ARG228A-ASP113B
			ARG333B-GLU58A

Table 3.8. Pairs of residues involved in stable intersubunit H-bonds in simulations with arginine in the central cavity. Dark gray background: intersubunit interactions identified in the initial models and conserved throughout the simulations; bold: stable interactions present in all systems. Pairs of residues are considered to have a stable H-bond interaction when the sum of % of existence of the H-bonds between the two residues is >50.

As for the previous simulations, only a few stable intersubunit salt bridges were detected throughout the simulations (Table 3.9), and most of them involved the residue Asp113 of one chain and Arg228 of the other chain. In the presence of the ligands, the number of these stable interactions was increased; however, the numbers were so small that they did not allow the evaluation of the significance of these data.

wtGALT + Arginine	p.Gln188Arg + Arginine	wtGALT + Ligands + Arginine	p.Gln188Arg + Ligands + Arginine
GLU58B-ARG333A			
	ASP113B-ARG228A	ASP113B-ARG228A	ASP113B-ARG228A
	ASP113A-ARG228B	ASP113A-ARG228B	ASP113A-ARG228B
		ASP39B-ARG201A	ASP39B-ARG201A

Table 3.9. Pairs of residues involved in stable intersubunit salt bridges in simulations with arginine in the central cavity. Pairs of residues are considered to have a stable salt bridge interaction when the sum of % of existence of the salt bridges between the two residues is >50.

3.2.4 Comparison of the results of simulations in the presence of Arginine

When arginine was simulated in the active site, the simulations revealed that this interaction was unstable and that arginine tended to leave the active site during the simulations. When present, it did not favorably affect any structural feature of the enzymes; rather, sometimes, it seemed to perturb them, such as in the case of secondary structuresin wtGALT. The mutation-perturbed intersubunit interactions also did not appear to be improved by the presence of arginine in the active site; indeed, the number of both H-bonds and salt bridges was slightly lower in the presence of arginine when compared to those for the corresponding simulations in the absence of this putative pharmacochaperone. Moreover, the presence of arginine in the active site, as could be expected, seemed to perturb the interactions between the enzyme and the substrates. This is more evident in the mutant than in the wtGALT, probably because the accumulation of positive charges in the binding site of p.Gln188Arg determined the formation of a repulsive force that could even result in the expulsion of arginine outside the binding site.

The simulations with arginine in the central cavity showed that the amino acid found a favorable interaction with residues near a putative “allosteric site” [McCorvie et al., 2016] and maintained it constantly for the whole simulations. This is interesting, considering that the central cavity of the enzyme is quite big, and that our simulations (both docking and MD simulations) allowed arginine to move freely in this cavity. The presence of arginine in the central cavity did not perturb the secondary structures of the

enzyme and seemed to slightly enhance the flexibility of those segments that were in contact with the substrates. This effect, however, was more visible in the absence of the substrates rather than in their presence, and it is difficult to associate this with a functional meaning. We also analyzed the intersubunit interactions in the simulations in the presence of arginine in the central cavity. When ligands were not present, the average number of H-bonds per timeframe decreased slightly with respect to the simulations in the absence of arginine. When ligands were bound to the enzymes, instead, this parameter decreased in wtGALT and increased in p.Gln188Arg. The number of intersubunit salt bridges was so low that the variations recorded in the different simulations cannot be considered significant. Finally, looking for a possible effect that the arginine bound in the central cavity could exert on the binding of the substrates in the active site, we noticed that the most stable interactions in wt hGALT were maintained with Arg48, Arg51, and residues of the loop 330–340, in addition to the catalytic residue His186. Additionally, in the simulations with arginine in the central cavity, two stable interactions with Asn97, Asp98, and, less frequently, with Ala81, Ser181, and Asn182 were formed. Thus, arginine bound to the central cavity seemed to affect the pattern of interactions of the substrates with the wildtype enzyme. In p.Gln188Arg, however, the stable interactions between the substrates and the enzyme were the same either in the absence or in the presence of arginine in the central cavity. In particular, in all the systems, Arg188 created a strong interaction with H2U, both with H-bonds and salt bridges, and this strong interaction, which persisted for all the simulation time, impaired the mutant enzyme in performing the correct catalysis. In our simulations, the presence of arginine was not able to alter this strong interaction; therefore, we predict that the binding of this amino acid is not able to rescue the enzymatic activity of the mutant enzyme.

3.3 Search for possible pharmacochaperones for GALT

3.3.1 Docking simulations of putative PCs on central cavity

We started by searching the literature for PCs already in therapeutic use, selecting drugs approved for misfolding pathologies. At the end, five PCs were identified, which we will refer to as PC1-PC2-PC3-PC4-PC5. As reported in paragraph 2.6.2.2, we will not disclose here their molecular structures to protect possible patent opportunities.

First of all, we decided to simulate the possibility that all five PCs could bind to the central cavity, in the same condition of arginine. These docking have been performed in two conditions: in the presence (table 3.10) and in the absence (table 3.11) of both ligands in the active site of the enzyme.

The docking results for the central cavity of the enzymes gave less defined results, because the cavity is very large and, thus, all five PCs have a higher conformational freedom. For this, we report both the best energy and the most populated poses as representative results of docking. However, all the simulations predicted a negative binding energy, suggesting the possibility that all five PCs could also bind to this cavity. The predicted binding energies in all these conditions seem not to be significantly different, indicating that neither the mutation nor the presence of the substrate in the active site would affect the binding of all five PCs in the central cavity. However, the highest (less favorable) value of the interaction energy remains those of PC3 with both wtGALT and p.Gln188Arg mutant protein.

PCs	docking wtGALT+G1P +H2U	Predicted Interactions with Residues	docking p.Gln188Arg+G1P+H2U	Predicted Interactions with Residues
PC1	BE=MP Energy value: -7,5 kcal/mol N.P: 14	GLU38A,GLU40A,TRP249A,TYR323A, MET336A; ARG197B,ILE198B, MET341B,ALA343B, GLN344B	BE Energy value: -7,7 kcal/mol N.P.: 1	GLN58A, MET177A, MET178A, TRP249A, PRO325A, ARG333A, MET 336A; GLN56B, GLN58B, ARG333B, PHE335B, VAL337B, MET341B
			MP Energy value: -7.0 kcal/mol N.P: 14	GLN38A, GLU40A, THR248A, TRP249A, MET336A, GLU340A, MET341A, GLN344A, ALA345A; ILE198B, MET341B
PC2	BE Energy value: -5,7 kcal/mol N.P: 27	MET177A, MET178A, TRP249A, ARG333A, MET336A, MET341A; MET336 B, MET341 B	BE=MP E: -6,7 kcal/mol N.P:73	GLU40A, TRP41A; ASP197B, ILE198B, GLU340B, MET341B, ALA343B, GLN344B, ALA345B
	MP Energy value: -5,0 kcal/mol N.P: 37	MET336A, GLU340A, MET341A; ILE198B, MET336B		
PC 3	BE Energy value: -4,8 kcal/mol N.P: 30	ASP197A, ILE198A, ARG201A; GLU 40B	BE Energy value: -4,8 kcal/mol N.P: 19	ASP197A, ILE198A, ARG201A, GLU40B
	MP coincide con BE		MP E: -4,7 kcal/mol N.P: 20	ASP197A, ILE198A, ARG201A; GLU 40B, GLU340B, GLN344B, ALA345B
PC4	BE=MP Energy value:-7,1 kcal/mol N.P: 75	MET177A, MET178A, TRP249A, ARG333A; MET341B, ARG333B, PHE335B, MET336B, VAL337B	BE=MP Energy value:-7,1 kcal/mol N.P: 88	MET177A, MET178A, TRP249A, ARG333A; MET341B, ARG333 B, PHE335B, MET336B, VAL337B
PC5	BE E:-8,45kcal/mol N.P: 9	GLN56A, GLU58A, ARG333A, PHE335A, MET336A, VAL337A; GLN56B, GLU58B, MET178B, ARG333B, PHE335B, MET336B	BE Energy value:-8,4 kcal/mol N.P:11	GLN56A, GLU58A, ARG333A, PHE335A, MET336A, VAL337A; GLN56B, GLU58B, MET178B, ARG333B, PHE335B, MET336B
	MP E:-8,3 kcal/mol N.P: 82	GLU40A, GLU340A, MET341A, ALA343A, GLN344A, ALA345A; ASP197B, ILE198B, TRP249B	BE Energy value:-8,2 kcal/mol N.P: 83	GLU40A, GLU340A, MET341A, ALA343A, GLN344A, ALA345A; ASP197B, ILE198B, TRP249B

Table 3.10. Results of focused docking (central cavity) for all five PC (1,2,3,4,5) with wtGALT and p.Gln188Arg in the presence of natural substrates bound to the active site.
BE=The binding energy of the best energy pose; MP=The binding energy of the most populated pose; N.P=number of poses in the cluster.

PC	docking wtGALT	Predicted Interactions with Residues	docking p.Gln188Arg	Predicted Interactions with Residues
PC1	BE Energy value: -7,6 kcal/mol N.P: 9	MET341A; ASN35B,GLN38B, GLU40B,ILE198B, THR248B,GLN252B, TYR323B,GLU340B, GLN344B,ALA345B	BE Energy value: -7,6 kcal/mol N.P: 5	GLU340A, MET341A, GLN344A, ALA345A; ASN35B, GLU40B, TYR323B, ILE 398B, ASP197B, HIS301B
	MP Energy value: -7,4 kcal/mol N.P: 11	GLU340A, MET341A; THR248B, TRP249B, GLN252B, TYE323B, MET336B	MP Energy value: -7,2 kcal/mol N.P: 10	THR248B, TRP249B, GLN252B, TYR323B, MET336B, GLU340B, MET341B
PC2	BE=MP Energy value: -7.0 kcal/mol N.P 70	GLU40A, TRP41A; MET341A, ALA343A, GLN344A,ALA345A; ASP197B, ILE198 B, GLU340B	BE=MP Energy value: -6,8 kcal/mol N.P: 78	GLU40A, TRP41A, GLU340A, MET341A, ALA343A,GLN344A, ALA345A; ASP197B, ILE198B
PC3	BE Energy value: -5,5 kcal/mol N.P: 7	PHE171A, ASN173A; ARG48B,LYS334B, PHE335B, VAL337B, TRYR339B, GLU349B	BE Energy value:-5,1 kcal/mol N.P: 18	GLU40B, ASN35B, THR248B, GLN252B, TYR323B
	MP Energy value:: -5,1 kcal/mol N.P: 20	GLU40B, ASN35B, THR248B, GLN252B, TYR323B	MP Energy value:: -4,8 kcal/mol N.P: 25	ASP197A, ILE198A, ARG201; GLU40B, GLU340B, GLN344B, ALA345 B
PC4	BE=MP Energy value: -7,1 kcal/mol N.P: 74	MET177A,MET178A, TRP249A, ARG333A; TRP249B, ARG33B, PHE335B, VAL337B	BE=MP E: -7,1 kcal/mol N.P: 77	MET177A, MET178A, TRP249A, ARG333A; TRP249B, ARG33B, PHE335B, MET336B
PC5	BE Energy value: -8,45 kcal/mol N.P: 11	GLN56A, GLU58A, ARG333A, PHE335A, MET336A, VAL337A; GLN56B, GLU58B, MET178B, ARG333B, PHE335B, MET336B	BE Energy value: -8,4 kcal/mol N.P: 8	GLN56A, GLU58A, ARG333A, PHE335A, MET336A, VAL337A; GLN56B, GLU58B, MET178B, ARG333B, PHE335B, MET336B
	MP Energy value: -8,45 kcal/mol N.P: 83	GLU40A, GLU340A, MET341A, ALA343A, GLN344A, ALA345A; ASP197B, ILE198B, TRP249B	MP Energy value: -8,4 kcal/mol N.P: 88	GLU40A, GLU340A, MET341A, ALA343A, GLN344A, ALA345A; ASP197B, ILE198B, TRP249B

Table 3.11: Results of focused docking (central cavity) for all five PC (1,2,3,4,5) with wtGALT and p.Gln188Arg in the absence of natural substrates bound in the active site.
BE=The binding energy of the best energy pose; MP=The binding energy of the most populated pose; N.P=number of poses in the cluster.

3.3.2 Search of the allosteric site

The strategy used to search for the allosteric site of hGALT was based on (i) the literature [McCorvie et al., 2013], (ii) docking results of the five PCs (see paragraph 3.3) in the central cavity, and (iii) the results of both docking and MD of arginine (see paragraph 3.2).

Based on the literature, McCorvie and coauthors in 2013 identified, using FTMap server (*http://ftmap.bu.edu*), a possible allosteric site for hGALT. In particular, it was defined as present at the dimer interface and in the side opposite the binding site, on the old model of hGALT (PDB: 1R3A). However, in that and also in other following study [McCorvie et al., 2016], the residues involved are not described.

Therefore, we submitted the new model of hGALT enzyme (see paragraph 2.6.4) to the FTMap server, which fully automatically, generated 11 clusters. Among these, the cluster containing the largest number of probes per cluster was selected as primary hotspot (Figure 3.12).

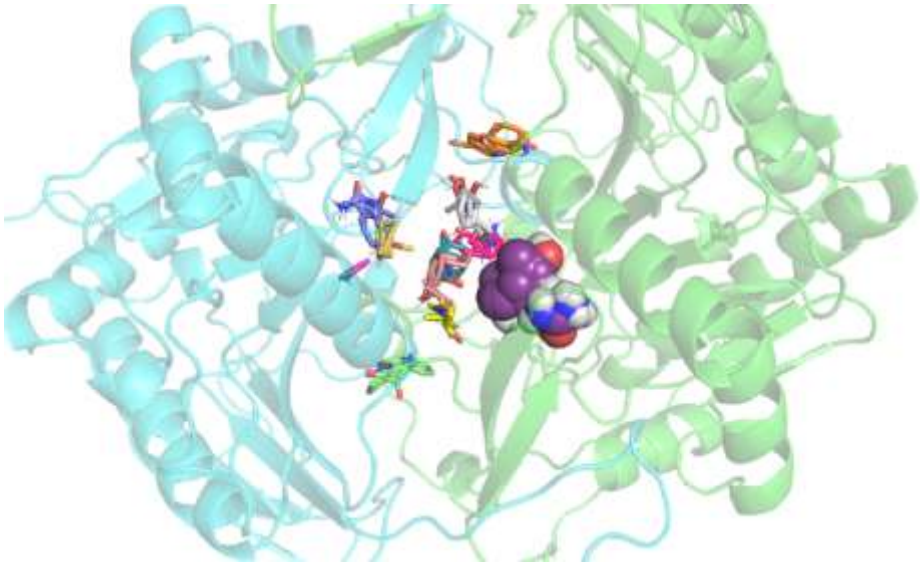


Figure 3.12 Output of FTMap. Purple spheres identify the primary hotspot. In sticks other probes forming the other CSs. Image obtained by PyMOL.

Once verified that the primary hotspot is “druggable", as per criteria described in Paragraph 2.3.2, the amino acids around 5 Å from the amino acids of the others CSs were selected using PyMOL.

We submitted to FTMap analysis also the theoretical model of hGALT [Marabotti et al., 2005] used by McCorvie and coworkers in 2013. The FTMap server identified a total of 13 clusters for this model. Among them, the primary hotspot consists of 15 probes. The residues 5 Å apart were also identified for this cluster.

Figure 3.13 shows these residues, which in both cases correspond to a portion of the central cavity.

A

RESIDUES OF CENTRAL CAVITY					
CHAIN A			CHAIN B		
ASN35,	LEU37,	GLN38,	ASN35,	LEU37,	GLN38,
ASP39,	GLU40,	TRP41,	GLU40,	TRP41,	VAL42,
VAL42,	GLN56,	GLU58,	GLN56, GLU58, MET177, MET		
MET177,	MET 178,	ASP197,	178, ASP197, ILE198, GLN200,		
ILE198,	ARG201,	TRP246,	ARG201,	ARG204,	PRO244,
ALA247,	THR248,	TRP249,	TRP246,	ALA247,	THR248,
GLN252,	HIS301,	HIS321,	TRP249,	GLN252,	THR253,
TYR323,	PRO325,	ARG333,	LEU254,	HIS301,	HIS321,
PHE335,	MET336,	VAL337,	TYR323,	ARG333,	PHE335,
GLU340,	MET341,	LEU 342,	MET336,	VAL337,	GLU340,
GLN344,	ALA345,		MET341,	LEU342,	GLN344,
			ALA345,		

B

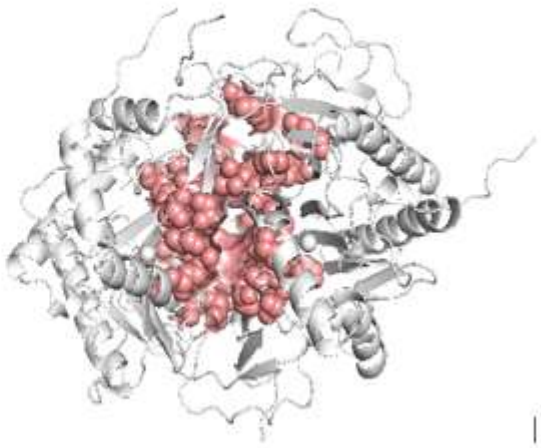


Figure 3.13. A. Residues of central cavity and residues identified by FTmap and Pymol circled in red. B. Structure of hGALT with residues identified by FTmap and Pymol in pink spheres

It is worth noting that the analysis of the interactions of all 5 PC docked into the central cavity showed that the most populated poses always identifies the same specific zone included in both hotspots found by FTMap Moreover, the arginine results from both docking and MD (see paragraph 3.2) reinforced the idea that the zone of interest includes the same residues.

On the basis of these results, we were able to identify these areas as the potential allosteric sites postulated by McCorvie et al. on chain A and on chain B of hGALT enzyme (Figure 3.14).

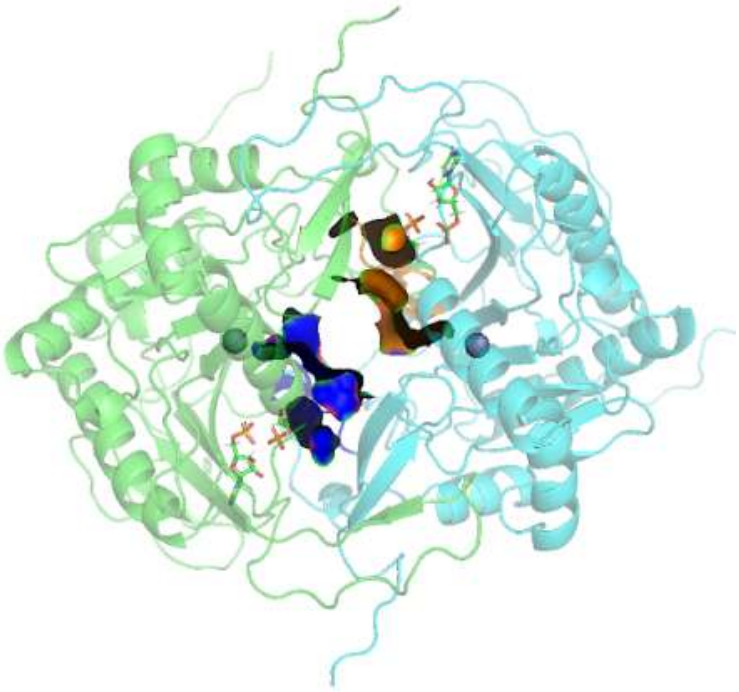


Figure 3.14. Representation of the possible allosteric site identified by comparing FTMap, docking results between hGALT/p.Gln188Arg and all five PCs on the central cavity and docking ad MD results of arginine. In orange the allosteric cavity portion on chain A, in blue the allosteric cavity portion on chain B. Note that the two cavities are at the dimer interface, as indicated in the article by [McCorvie et al., 2013]; note that in sticks are G1P and H2U of both active site (A and B).

3.3.2.1 Docking on the potential allosteric sites

Once the two potential allosteric sites have been identified, we decided to simulate the possibility that all five PCs could bind to both these sites.

These docking have been performed in the presence of both ligands for wtGALT and p.Gln188Arg bound to the active site of the enzyme, although their presence is not significant for these docking, as seen in tables 3.14 and 3.15.

The docking results for the simulations focused on the allosteric site A (table 3.12) and B (table 3.13) of the enzymes gave more defined results than those performed in the central cavity.

Again for this dockings, it has been chosen to report both the best energy and most populated poses as a representative result of docking.

PCs	docking wtGALT	Predicted Interactions with Residues	docking p.Gln188Arg	Predicted Interactions with Residues
PC1	BE/MP En: -7,6 kcal/mol N. in cluster:25	GLN38A, GLU40A, THR249A, TYR323A, MET336A, ALA343, GLN344A, ALA345A, ASP197B, MET341B	BE En:-7,5 kcal/mol N. in cluster:27	GLN38A, GLU40A, MET336A, GLU340A, ALA343A, GLN344A, ALA345A, ASP197B, ILE198B, MET341B
			MP En:-6,7 kcal/mol N. in cluster:30	GLN38A, GLU40A, MET336A, GLU340A, MET341A, GLN344A, ALA345A, ASP197B, MET341B
PC2	BE/MP En: -6,8 kcal/mol N. in cluster:100	GLU 40 A, TRP 41 A, GLU 340 A, MET 341 A, ALA 343 A, ALA 345 A, ASP 197B, ILE 198 B	BE/MP En:-6,7 kcal/mol N. in cluster:100	GLU 40 A, TRP 41 A, GLU 340 A, MET 341 A, ALA 343 A, GLN344 A, ALA 345 A, ASP 197B, ILE 198 B
PC3	BE En: -4,6 kcal/mol N. in cluster:27	GLN38 A, GLU 40 A, THR 248 A, ALA 345A	BE/MP En: -4,5 kcal/mol N. in cluster: 58	GLU40 A, GLU 340 A, GLN344 A, ASP197 A
	MP En: -4,3 kcal/mol N. in cluster:39	GLU 40 A, GLU340 A, GLN 344 A, ALA345 A, ASP197 B		
PC4	BE/MP En: -5,3 kcal/mol N. in cluster:52	TRP249 A, MET336 A, MET341 A, MET336 B,GLU340 B, MET341 B	BE En: -5,3 kcal/mol N. in cluster:22	THR248 A, TRP 249 A, TYR323 A, MET336 A, GLU340 A, ALA 345 A, MET341 B
			MP En: -5,3 kcal/mol N. in cluster:46	THR248 A, TRP 249 A, TYR323 A, MET336 A, GLU340 A, ALA 345 A, MET341 B
PC5	BE/MP En: -8,4 kcal/mol N. in cluster:77	GLU40A, TRP41A, MET336A, GLU340A, MET341A, ALA343A. GLN344A, ALA345A, ASP197B, ILE198B, TRP249B	BE/MP En:-8,2 kcal/mol N. in cluster:78	GLU40A, GLU340A, MET341A, ALA343A, GLN344A, ALA345A, ASP197B, ILE198B, TRP249B

Table 3.12. Results of focused docking (potential allosteric site of chain A) for all five PC (1,2,3,4,5) with wtGALT and p.Gln188Arg in the presence of natural substrates bound in the active site. BE=The binding energy of the best energy pose; MP=The binding energy of the most populated pose; N. =number of poses in the cluster

PCs	docking wtGALT	Predicted Interactions with Residues	docking p.Gln188Arg	Predicted Interactions with Residues
PC1	BE En:-7,2 kcal/mol N. in cluster:28	GLN38A, ASP197A, ILE198A, TRP249A, TYR323A, MET336A, GLU40B, GLU340B, MET341B, ALA345B	BE/MP/MP En: -7,1 kcal/mol N. in cluster: 34	GLN38A, ASP197A, ILE198A, ARG201A, TYR323A, GLU40B, GLU340B, MET341B
PC2	BE/MP En: -6,6 kcal/mol N. in cluster: 50	GLU 40 A, TRP 41 A, GLU 340 A, MET 341 A, ALA 343 A, GLN344 A, ALA 345 A, TRP249 B	BE/MP En:-6,1 kcal/mol N. in cluster:99	ASP197 A, ILE198 A, GLU 40 B, TRP 41 B, MET 336 B, GLU 340 B, MET 341 B, ALA345 B
	BE/MP En:-6,6 kcal/mol N. in cluster:50	ASP 197 A, ILE 198 A, GLU 40 B, TRP 41B, MET 336 B, GLU 340 B, MET 341 B, GLN 344 B, ALA 345 B		
PC3	BE/MP En:-7.0 kcal/mol N. in cluster:55	ASP197 A; ILE 198A, ARG201 A, GLU40 B, GLU340 B, GLN344 B, ALA345 B	BE/MP En:-4,8 kcal/mol N. in cluster:44	ASP197 A, ILE198 A, ARG201 A, GLU40 B
PC4	BE En:-5,9 kcal/mol N. in cluster:1	MET 178 A, ARG333 A, MET 336 A, ARG333 B, PHE 335 B, MET 336 B, MET 341 B	BE En:-5,9 kcal/mol N. in cluster:1	MET 178 A, TRP249 A, ARG 333 A, MET 336 A, ARG333 B, PHE 335 B, MET336B, VAL 337 B, MET 34 1 B
	MP En:-5,3 kcal/mol N. in cluster:51	GLN38 A, TRP249 A, TYR 323 A, GLU340 B, MET341 B, ALA345 B	BE En:-5,3 kcal/mol N. in cluster:62	GLN38 A, TRP 249 A, TYR 323 A, GLU 340 B, MET341 B, ALA 345 B
PC5	BE En:-7,3 kcal/mol N. in cluster:1	GLU40A, TRP41A, TRP249A, MET336A, GLN344A, ALA345A, ASP197B, MET341B	BE/MP En: -7,0 kcal/mol N. in cluster: 99	ASP197A, ILE198A, TRP249A, GLU40B, GLU340B, MET341B, GLN344B, ALA345B
	MP En:-7,2 kcal/mol N. in cluster:99	ASP197A, ILE198A, TRP249A, GLU40B, GLU340B, MET341B, GLN344B, ALA345B		

Table 3.13. Results of focused docking (potential allosteric site of chain B) for all five PC (1,2,3,4,5) with wtGALT and p.Gln188Arg in the presence of natural substrates bound in the active site.
BE=The binding energy of the best energy pose; MP=The binding energy of the most populated pose; N=number of poses in the cluster.

All the simulations predicted a negative binding energy, suggesting the possibility that all five PCs could indeed bind to this potential allosteric site. The predicted binding energies in all these conditions seem not to be significantly different, indicating that neither the mutation nor the presence of the substrate in the active site would affect the binding of all five PCs in the central cavity (Figure 3.15 and Figure 3.16).

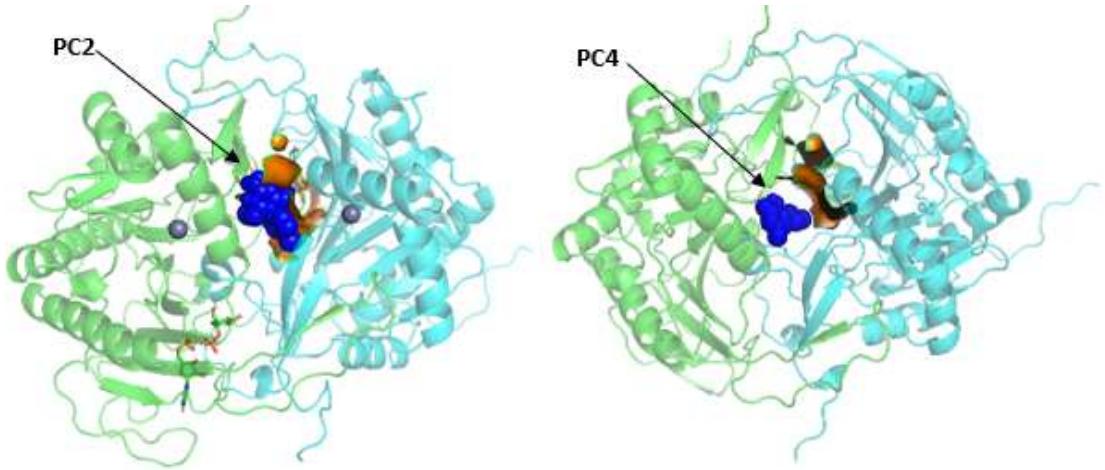


Figure 3.15. Left: representation of the best docking conformations of PC2 in blue within wtGALT. Right: representation of the best docking conformations of PC4 in blue within wtGALT. The putative allosteric site of chain A is shown in orange. The analysis of the interactions showed that PC4 is not bound inside the putative allosteric site, rather it accommodates in a cavity of the protein close to it.

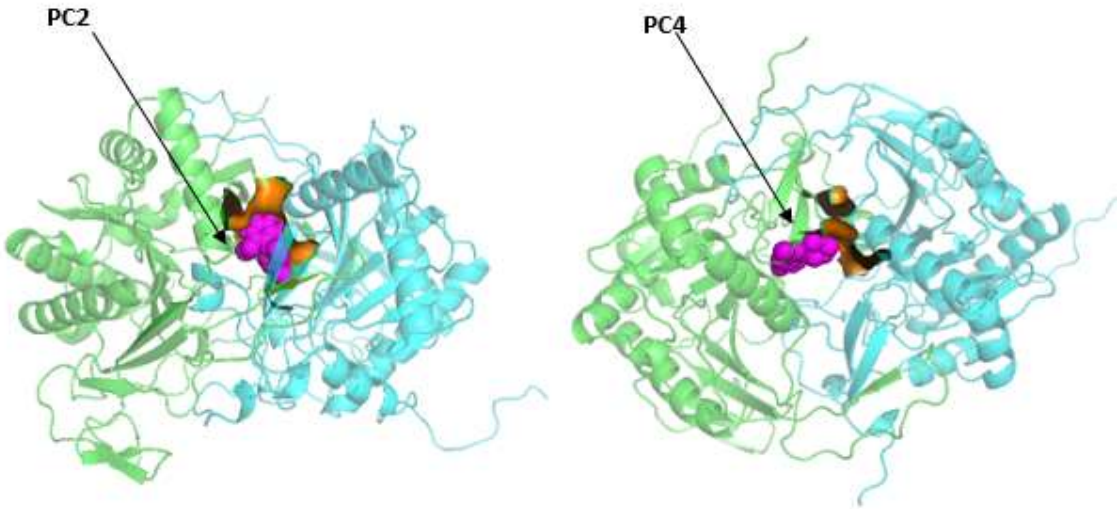


Figure 3.16. Left: representation of the best docking conformations of PC2 in magenta within p.Gln188Arg. Right: representation of the best docking conformations of PC4 in magenta within p.Gln188Arg. The putative allosteric site of chain A is shown in orange. The analysis of the interactions showed that PC4 is not bound inside the putative allosteric site, rather it accommodates in a cavity of the protein close to it.

3.3.3 Receptor-based pharmacophoric modelling for GALT

The best docking conformations of the PCs on both potential allosteric sites (A and B) were used as a starting point to generate the pharmacophoric models according to the receptor-based method (see paragraph 2.2.3).

Discovery Studio takes as input the best docking conformations and creates all possible combinations of pharmacophoric models. As reported in paragraph 2.2.3, among the generated pharmacophore models, the best one (which generally corresponds to the first one) is selected, with a better selectivity score. Notably, no pharmacophoric models were generated from the best docking conformations of PC4 and PC2. Instead, PC1, PC3, and PC5 generated pharmacophoric models from which we started the search for pharmacophoric hits.

3.3.4 Search of pharmacophoric hits and virtual screening

The best selected pharmacophoric models were used for the searching for pharmacophoric hits in DrugBank.

As reported in the paragraph 2.6, only pharmacophoric hits with a value ≥ 3 and only pharmacophoric hits found from the pharmacophore model generated by the docking conformations of the p.Gln188Arg mutant were selected.

As result, a total of 19 hits were selected (table 3.14) and all hits have the same pharmacophoric features identified as AHHHP [hydrogen bond acceptor (A), cation (P), and hydrophobic (H)].

Again, their names and molecular structures are not disclosed to protect possible future patent applications of these compounds.

PHARMACOPHORIC HIT	FITVALUE	GROUPS	LIPINSKI RULE	FEATURE
HIT 1	3,0731	experimental	SI	AHHHP
HIT 2	3,20849	experimental	SI	
HIT 3	3,18965	Approved/investigational vet_approved	SI	
HIT 4	3,122	experimental	SI	
HIT 5	3,38402	approved investigational	SI	
HIT 6	3,18336	investigational	SI	
HIT 7	3,15195	approved investigational	SI	
HIT 8	3,21927	experimental	SI	
HIT 9	3,06716	experimental	SI	
HIT 10	4,01559	investigational	SI	
HIT 11	3,00595	approved	SI	
HIT 12	3,18988	approved investigational	SI	
HIT 13	3,25559	approved investigational	SI	
HIT 14	3,51184	experimental investigational	SI	
HIT 15	3,36574	investigational	SI	
HIT 16	3,5515	vet_approved	SI	
HIT 17	3,20016	approved	SI	
HIT 18	3,62319	approved investigational	SI	
HIT 19	3,02465	experimental	SI	

Table 3.14: The 19 pharmacophoric hits with a value ≥ 3 and only found from the pharmacophore model generated by the docking conformations of the p.Gln188Arg mutant were selected.

3.3.4.1 Docking results of pharmacophoric hits on potential allosteric site A

We docked all 19 hits on the potential allosteric site of chain A of both wtGALT and p.Gln188Arg. Results are reported in table 3.15. By analyzing the interactions of the best conformations of hits 5, 7, 10 and 11 within the putative allosteric site of wtGALT and p.Gln188Arg, the residues with which most of the ligands interact are the same: Glu40A, Met336A, Glu340A, Met341A, Ala343A, Ala345A, Asp197B, Ile198B, Trp249B (Figure 3.17 and 3.18).

HIT	wtGALT (Kcal/mol)	p.Gln188Arg (Kcal/mol)
HIT 1	6,7	-6.6
HIT 2	-6.7	-6.8
HIT 3	-9.0	-9.0
HIT 4	-9.3	-8.5
HIT 5	**-9.5**	**-9.5**
HIT 6	-7.5	-7.7
HIT 7	**-10.9**	**-10.9**
HIT 8	-8.6	-8.7
HIT 9	-7.8	-7.8
HIT 10	**-8.0**	**-8.3**
HIT 11	**-9.6**	**-9.0**
HIT 12	-6.2	-6.4
HIT 13	-8.5	-7.1
HIT 14	-7.5	-7.6
HIT 15	-8.4	-8.0
HIT 16	-8.4	-8.3
HIT 17	-7.8	-7.8
HIT 18	-6.0	-6.1
HIT 19	-7.9	-7.8

Table 3.15: The results of the docking simulations focused on the possible allosteric site of chain A in p.Gln188Arg and wtGALT models with the pharmacophoric hits identified in the virtual screening. Results are displayed in terms of energy to allow for initial screening. The drugs selected for further analysis are highlighted in bold and underlined.

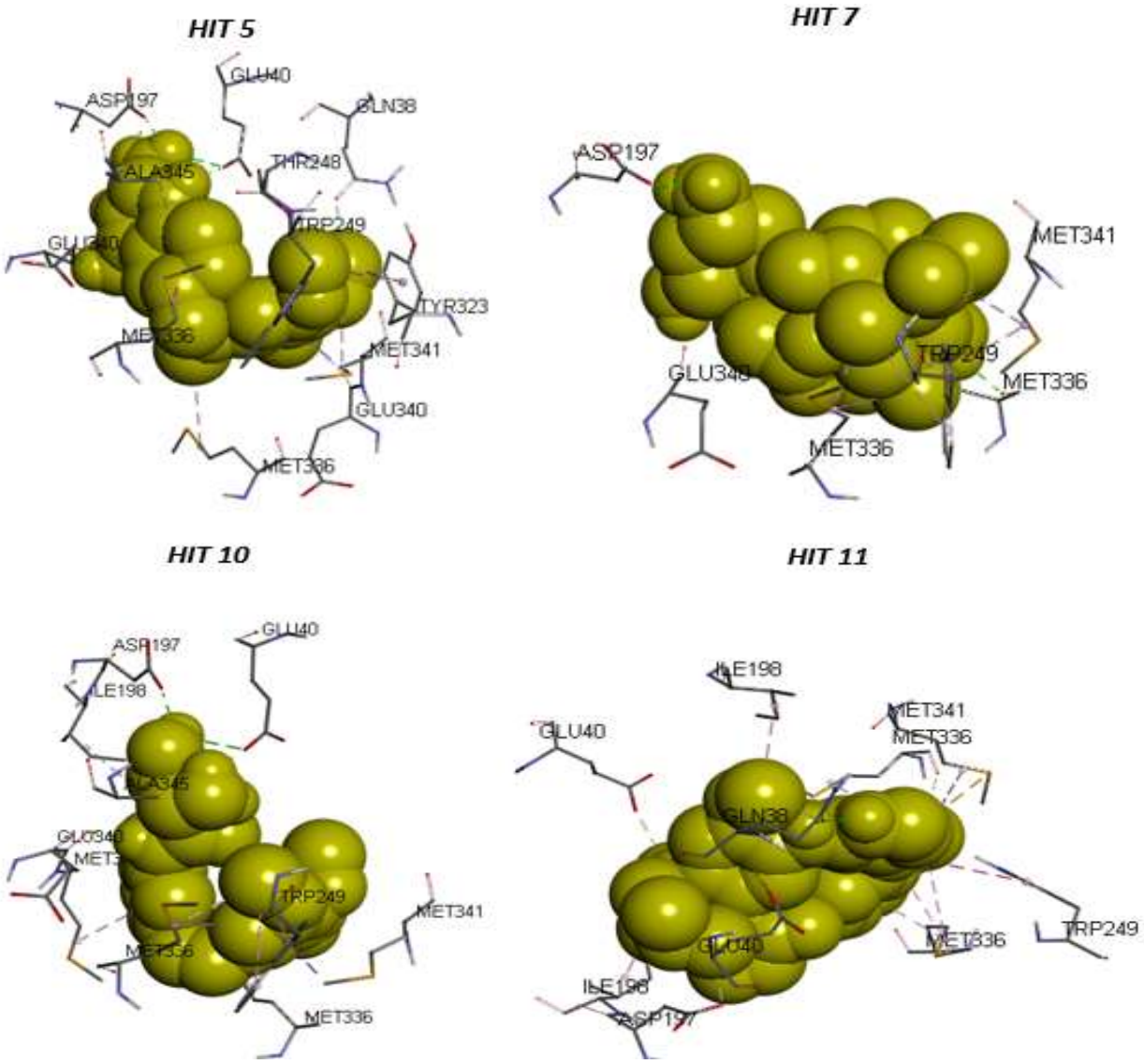


Figure 3.17. Representation of the interactions of the best docking conformations of the 4 hits selected within wtGALT models on the putative allosteric site of chain A. In sticks we visulize the interacting amino acids, in yellow spheres the above-mentioned ligands. Images obtained through BIOVIA Discovery Studio.

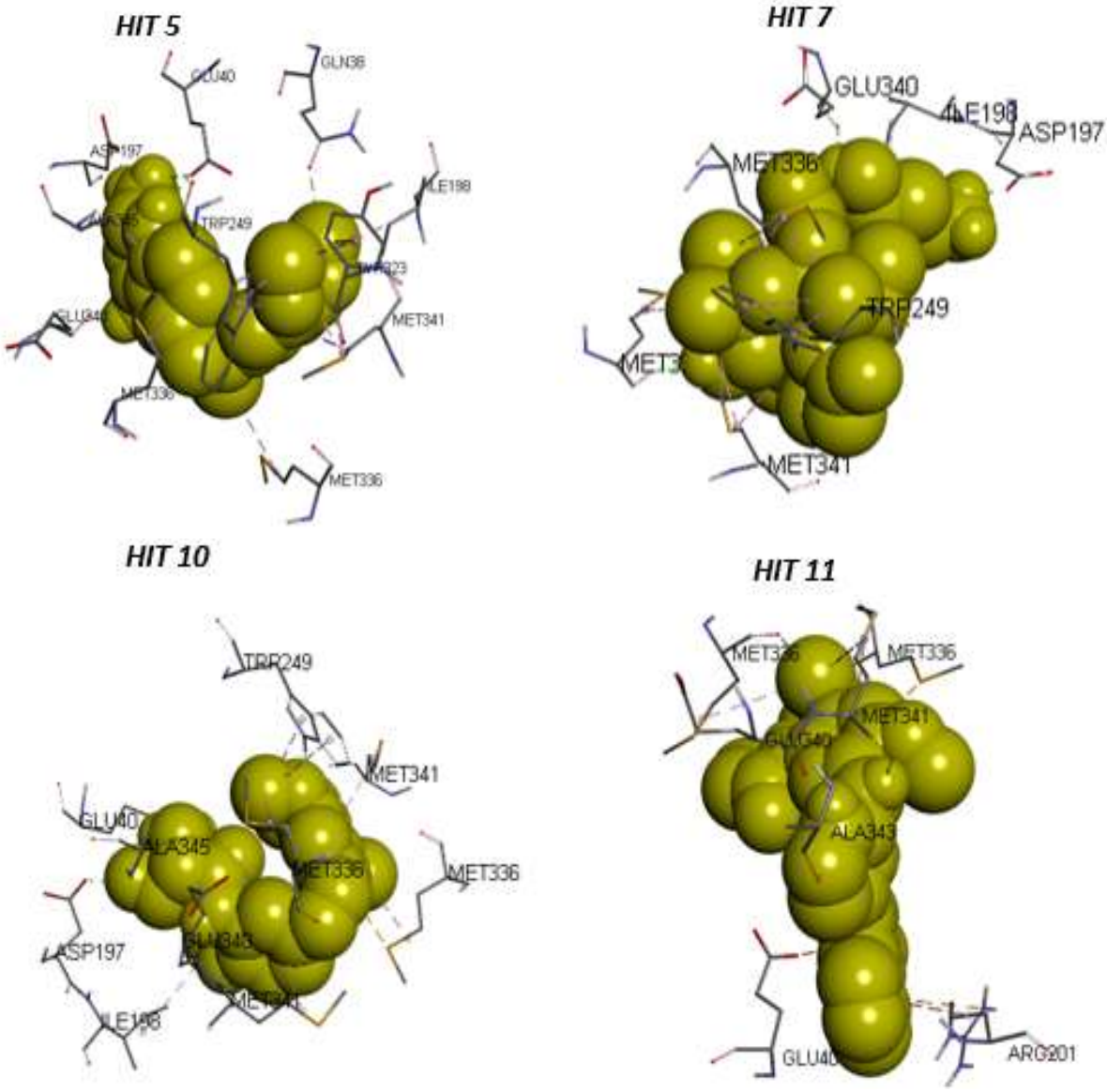


Figure 3.18: Representation of the interactions of the best docking conformations of the 4 hit selected within p.Gln188Arg models on the putative allosteric site of chain A. In sticks we visualize the interacting amino acids, in yellow spheres the above-mentioned ligands. Images obtained through BIOVIA Discovery Studio.

From the structural superpositions of the selected conformations, it can be seen that all the 4 hits fit well into the putative allosteric cavity of GALT enzyme, wtGALT (Figure 3.19) and p.Gln188Arg (Figure 3.20).

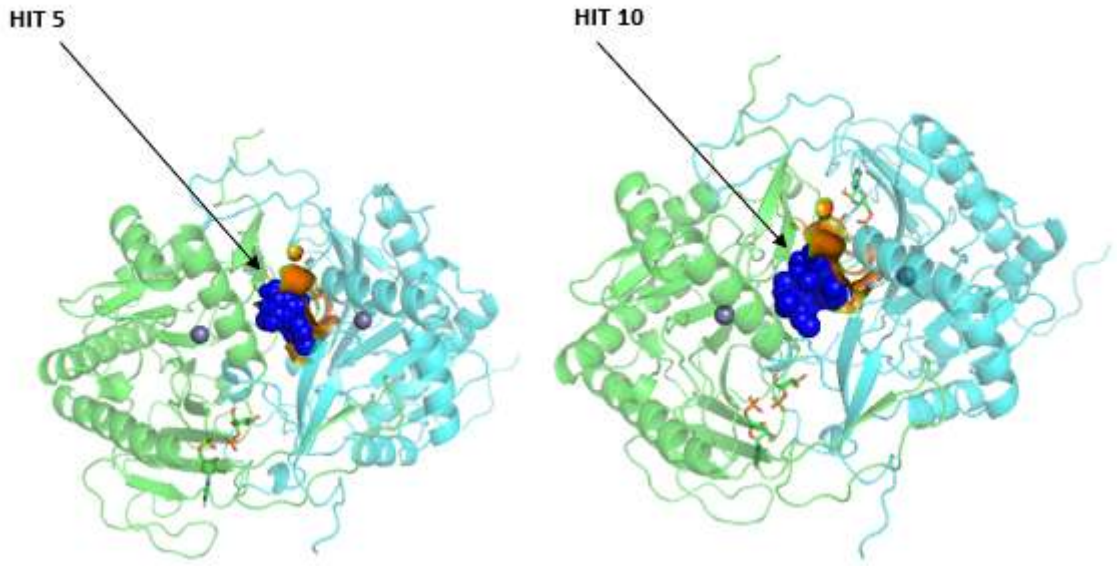


Figure 3.19: Left: representation of the best docking conformations of HIT5 in blue within wtGALT. Right: representation of the best docking conformations of HIT10 in blue within wtGALT. The putative allosteric site of chain A is shown in orange. The analysis of the interactions showed that both hits are bound inside the putative allosteric site.

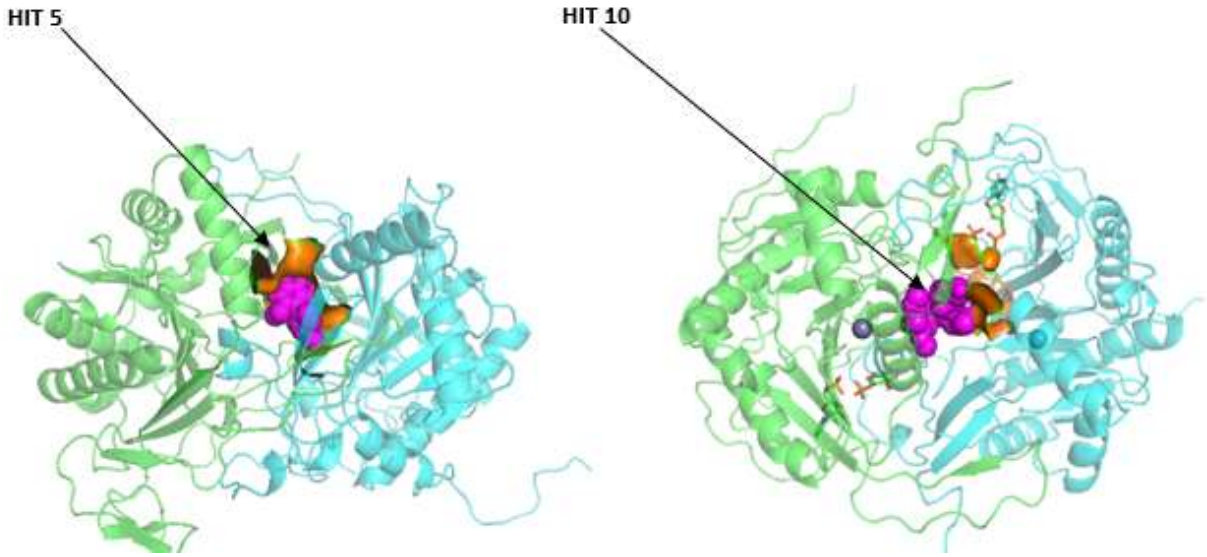


Figure 3.20: Left: representation of the best docking conformations of HIT5 in magenta within p.Gln188Arg. Right: representation of the best docking conformations of HIT10 in magenta within p.Gln188Arg. The putative allosteric site of chain A is shown in orange. The analysis of the interactions showed that both hits are bound inside the putative allosteric site.

3.3.4.2 Docking results of selected pharmacophoric hits on potential allosteric site B

The docking performed on the potential allosteric site of chain A was repeated on that of chain B. We choose to perform the docking only for the best 4 identified from previous docking simulation on chain A, namely hits 5, 7, 10 and 11. Also in this case, the docking simulations were performed on wtGALT and p.Gln188Arg in the presence of the G1P and H2U ligands bound into the active site. The result are reported in Table 3.16 for wtGALT and in Table 3.17 for p.Gln188Arg.

	BE		*MP*	
HIT 5	*RUN 87*	-7,6 (N.42)	*RUN 87*	-7,6 (N.42)
HIT 7	*RUN 10*	-10,5 (N.23)	*RUN 93*	-9,8(N.28)
HIT 10	*RUN 91*	-7,7 (N.1)	*RUN 59*	-7,0 (N.19)
HIT 11	*RUN 97*	-7,9 (N.4)	*RUN 82*	-7,4 (N.11)
	wtGALT			

Table 3.16: Docking results focused on the allosteric B-chain site for wtGALT systems.
BE=The binding energy of the best energy pose; MP=The binding energy of the most populated pose; N.=number of poses in the cluster.

	BE		*MP*	
HIT 5	*RUN 51*	-7,6 (N.35)	*RUN 51*	-7,6 (N.35)
HIT 7	*RUN 8*	-10,5 (N.28)	*RUN 3*	-9,6(N.31)
HIT 10	*RUN 62*	-7,8 (N.8)	*RUN 3*	-6,9(N.20)
HIT 11	*RUN 82*	-6,9 (N.9)	*RUN 17*	-5,6 (N.11)
	Q188R			

Table 3.17: Docking results focused on the allosteric B-chain site for p.Gln188Arg systems.
BE=The binding energy of the best energy pose; MP=The binding energy of the most populated pose; N.=number of poses in the cluster

The values obtained on chain B are similar to those obtained on chain A (see Table 3.15). By analyzing the best conformations of the hits, both for the wtGALT and p.Gln188Arg, the residues of interactions are again the same as those for chain A (Figure 3.21 and Figure 3.22).

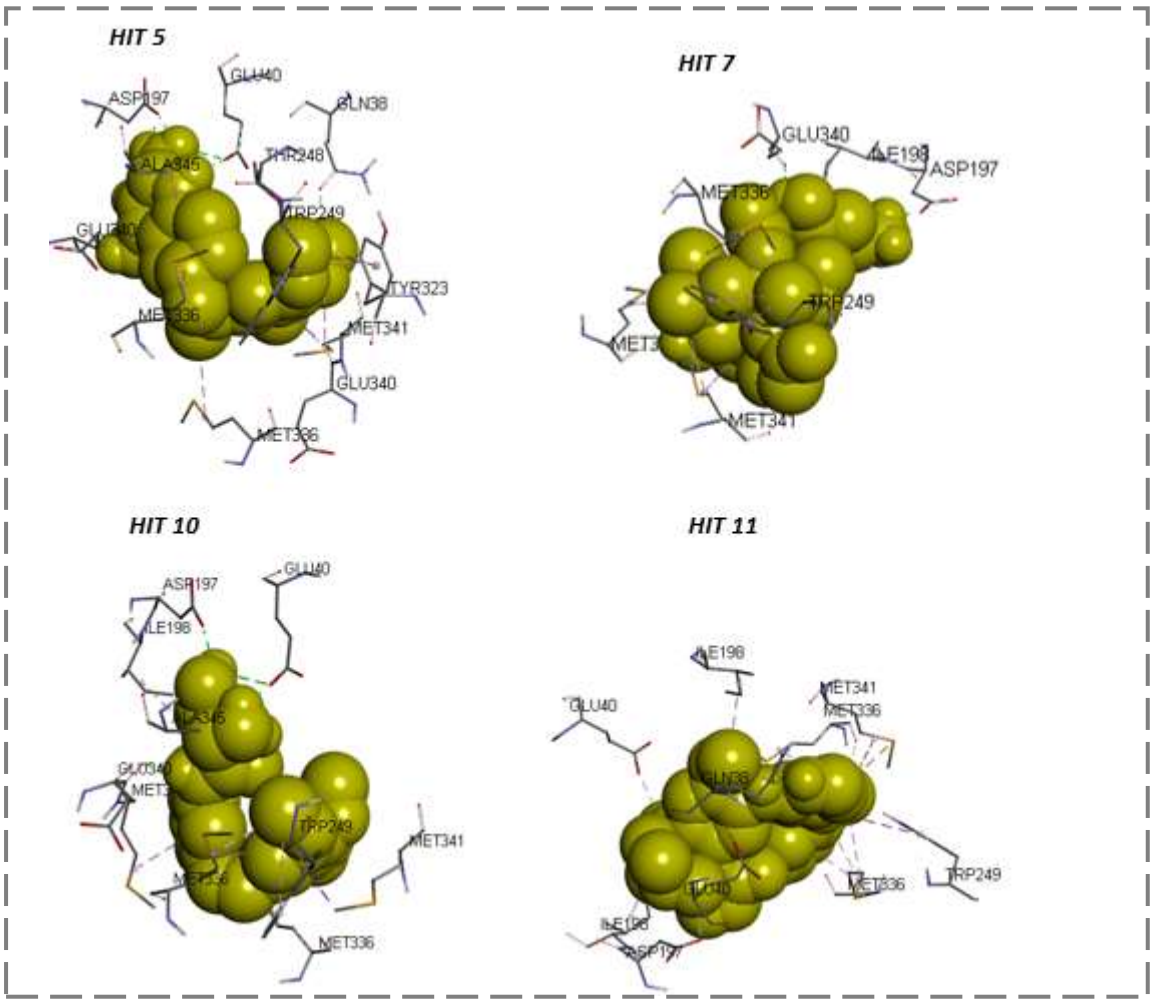


Figure 3.21: Representation of the interactions of the best docking conformations of the 4 hit selected within wtGALT models on the putative allosteric site of chain B. In sticks we visualize the interacting amino acids, in yellow spheres the above-mentioned ligands. Images obtained through BIOVIA Discovery Studio.

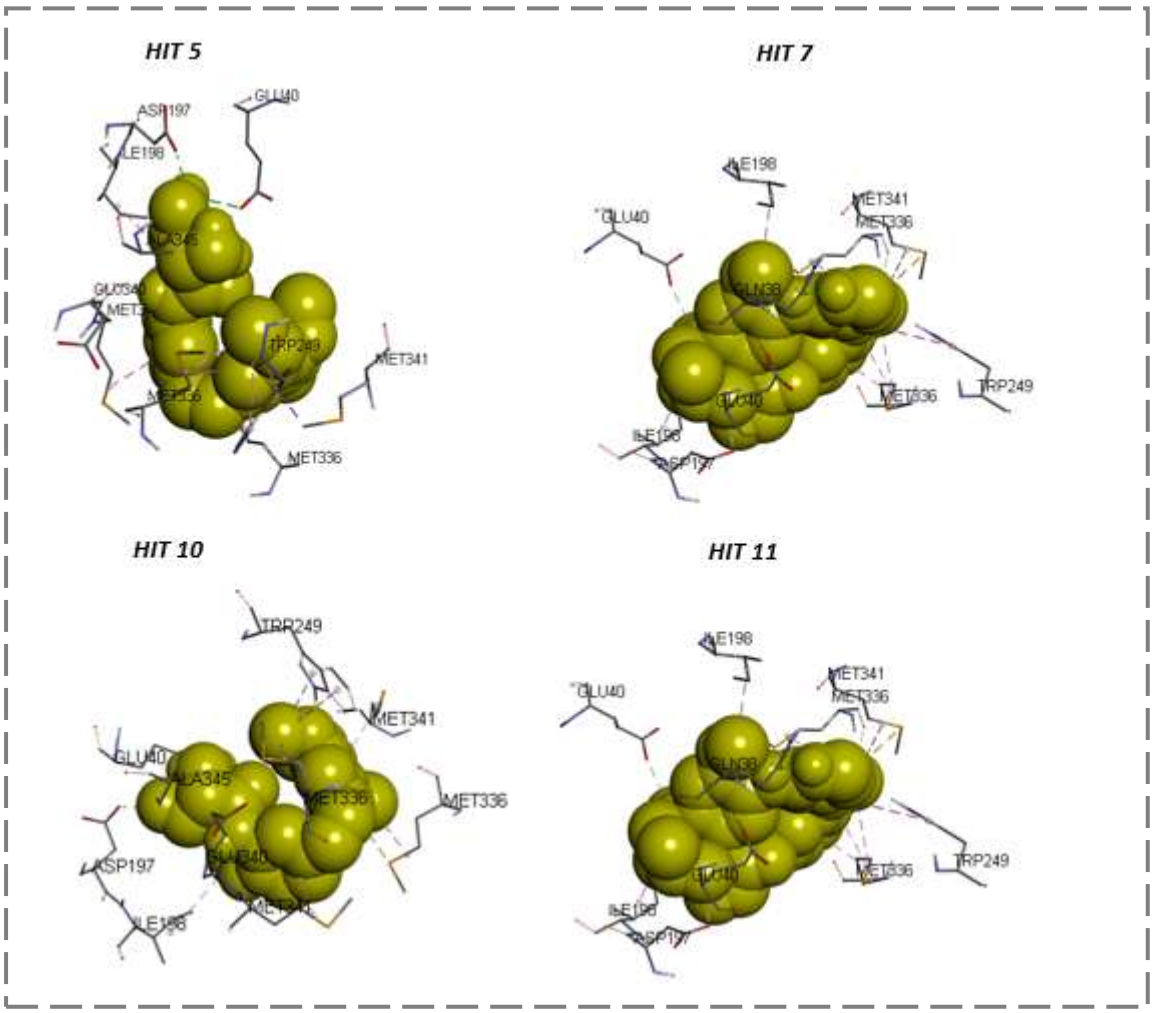


Figure 3.22: Representation of the interactions of the best docking conformations of the 4 hit selected within p.Gln188Arg models on the putative allosteric site of chain B. In sticks we visulize the interacting amino acids, in yellow spheres the above-mentioned ligands. Images obtained through BIOVIA Discovery Studio

Preliminary tests of PC1, PC2, PC4, and hits 7 and 10 on fibroblasts from galactosemic patients showed the ability of all 5 PCs to lower galactose-1-phosphate concentration when fibroblasts are stressed by galactose (personal communication). These preliminary data were made by Prof. Kent Lai of Utah University, USA, and obviously they need to be confirmed, but they are promising data towards the development of pharmacochaperone therapy for galactosemia.

3.4 Optimization of MD protocol for long simulations and for simulations with pharmacochaperones and selected pharmacophoric hits

Once PC1, PC2, PC4 and HIT 5, 7, 10, 11 were selected as promising ligands for hGALT, we decided to set up studies from a dynamic point of view at the level of the two potential allosteric sites.

To do this, we decided to set up a suitable protocol for long MD in order to study possible allosteric paths for hGALT. In fact, the search for allosteric pathways in proteins by means of MD simulations requires to perform long simulations in conditions that allow to detect the fine movements that occur in a macromolecule in the presence of these molecular phenomena. The setting of the new MD protocol resulted from the combination of a literature study, from which it was possible to identify the methods used to study communication in proteins in some works [Genoni et al., 2012; Sanchez-Martin et al., 2020] and from several experimental tests that differed in some conditions.

Long MD concerned the following systems: wtGALT; wtGALT + ligands; p.Gln188Arg; p.Gln188Arg + ligands at 310 K.

We have used as a starting point the models of wtGALT and of the mutant p.Gln188Arg obtained as described previously.

Unexpectedly, some of the possible PCs and HT selected in the previous steps could not be parameterized with the Amber force field. For this reason, we had to select another force field that was suitable for the correct parameterization of both the protein and different molecules. We selected CHARMM [Vanommeslaeghe et al., 2010], which, in addition to being a widely used force field for the study of proteins [Brooks et al., 2009], also has the advantage of making available to the scientific community CHARMM-GUI [Sousa da Silva et al., 2012], a Web interface that allows easy parameterization of molecules other than proteins.

The critical steps in order to perform the long MD simulations are the minimization of the starting structure, and the equilibration of the systems. The minimization protocol applied for our previous studies (see paragraph 2.6.3) is sufficient for general purposes, but to detect allosteric paths, it is necessary that the structure is deeply minimized [Moroni et al., 2018]. Therefore, in order to optimize this step, we performed several tests to assess the effect of different minimization protocols on the structure of the protein. As a reference, we analyzed the Ramachandran plot of GALT enzyme before and after the minimization process, and we selected the protocol that allowed to obtain a structure with no residues in the disallowed areas of this plot.

The analysis of the Ramachandran plot after the first minimization showed Asp90 as a residue with non-favorable dihedral angles. Therefore, we decided to carry out an additional minimization cycle. In this second cycle, the minimization has stopped when the maximum force reached a value lower than 1.0 kJ/mol/nm. In this way, the analysis of the Ramachandran plot after the second minimization showed no residues with non-favorable dihedral angles (Figure 3.23).

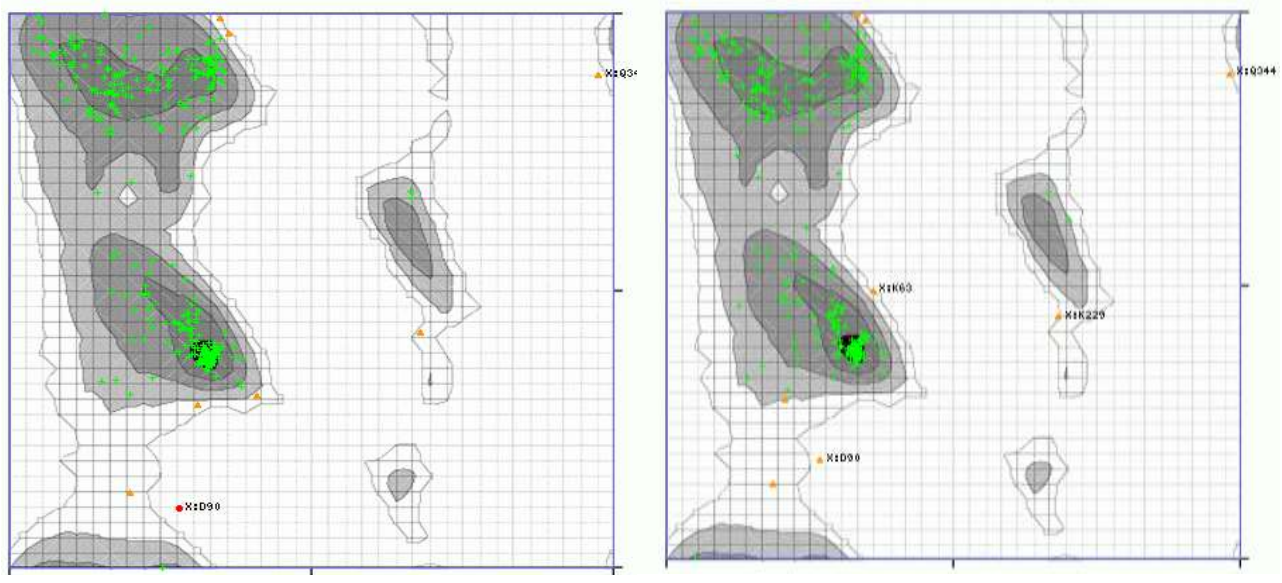


Figure 3.23. Left: Ramachandran plot analysis after minimization stopped when the maximum force reached a value lower than 10.0 kJ/mol/nm, with residues in disallowed regions shown as red circles: 1 (0.309%), is ASP90. Right: Ramachandran plot analysis after minimization stopped when the maximum force reached a value lower than 1.0 kJ/mol/nm. No residues in disallowed regions are visible.

Also, regarding minimization, a further test was carried out following the double minimization cycle (just discussed) with the conjugate gradient algorithm, stopped when the maximum force reached a value less than 10.0 kJ/mol/nm, and a further cycle stopped when the maximum force reached a value less than 4.0 kJ/mol/nm. These tests were made by running 100 ns simulations on wtGALT at 310 K, in the presence of the substrates. The analyses of these two simulations confirmed that the two protocols don't have any significant difference. Therefore, we concluded that our system had already achieved the best possible minimization after steepest descent algorithm. As representative of all the analyses performed, we report the RMSD of atom distances, showed that both minimization protocols allowed the systems to reach quickly the stabilization without significant differences (Figure 3.24).

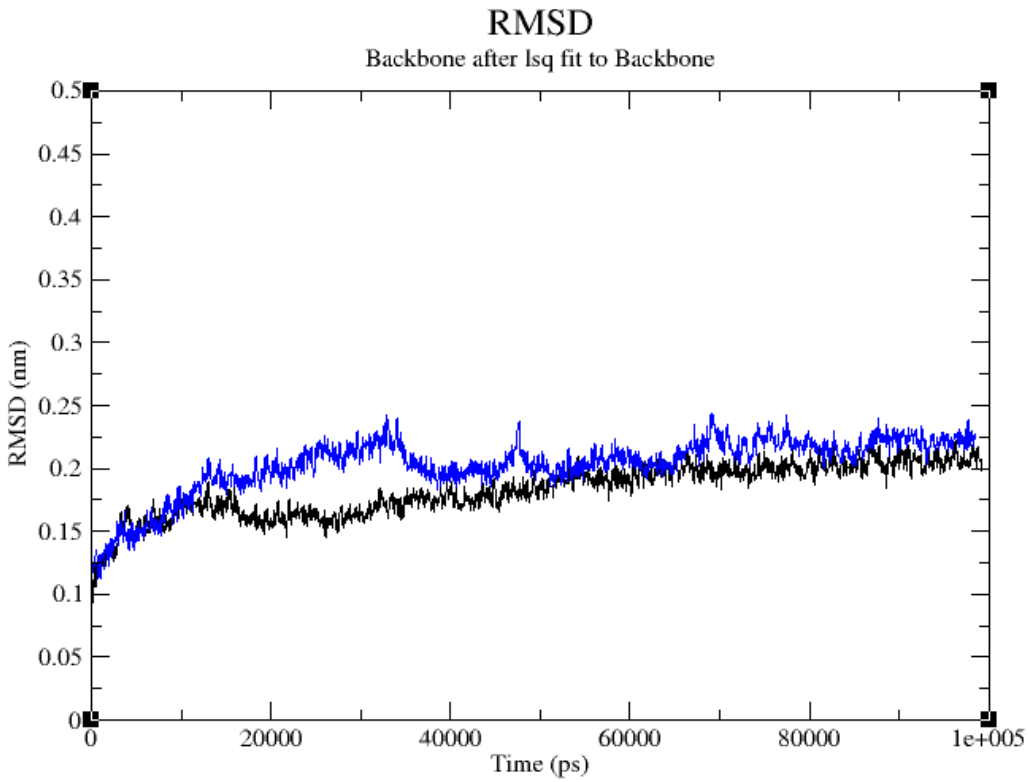


Figure 3.24: In blue, RMSD of atom distances analysis after minimization with steepest descent gradient; in black, RMSD of atom distances analysis after minimization with conjugate gradient algorithm.

In the second part of this study, the modified protocol for the equilibration phase was decided, considering either NVT and NPT mdp file. The new combination of thermostats in the NVT ensemble and thermostats and barostats in the NPT ensemble is particularly necessary in the case of long MD. In this respect, several tests were performed, each differing in a different combination of thermostats (Berendsen, V-rescale and Nosè-Hoover) and of barostats (Berendsen and Parrinello-Rahman). To evaluate a correct NPT equilibration, we based on two parameters. The first is the average pressure, to be never higher than 1.5 bar; the second is the value of the total drift (Tot-Drift). This last is calculated by performing a least-squares fit of the data to a straight line. The reported total drift is the difference of the fit at the first and last point. This value is considered acceptable when it does not exceed 2 [Abraham et al., 2015]. Then, to give the optimal condition for the pressure equilibration, we considered as variables:

1. different combinations of thermostats and barostats
2. several groups in the system can be coupled separately, as specified in the *tc-grps* of .mdp file
3. Length of nvt and npt equilibration

The different tests made are summarized in table 3.18. The best result was represented by test 9 in the table. This result highlighted how the groups coupled separately, the time, the right combination of thermostat/barostat affects a good equilibration.

We confirmed that, for our system, the best results were obtained with the V-rescale thermostat, in agreement with the MD protocol used in our previous protocol (paragraph 2.6.3). On the contrary, the use of Nosè-Hoover thermostat during NVT ensemble have shown the worst results (tests 1, 2, 3).

Concerning the different coupling groups, we observed the best result when the ligands and the Zn ions are grouped together with the enyme (wtGALT or p.Gln188Arg) as the first group, and water + ions are separated in the second group. In fact, if we compare test 8 with test 9, in which we operate under the same conditions of NVT and NPT but

with different groups, the result deteriorates dramatically when ligands are grouped together with water and ions, with an average pressure of -0.01.

TEST	Average of Energy pressure	Tot-Drift	NVT: Thermostat	NPT: Thermostat	NPT: Barostat	TC-GROUPS	TIME NVT/NPT
TEST 1	5.78	14.3	NOSE-HOOVER	NOSE-HOOVER	PARRINELLO	Protein and Non Protein (Ligand+cofactor+water+ions) i.e: GROUP 1: Protein; GROUP 2: G1P+H2U+ZN(2ATOMS)+WATER+Cl/Na	**100ps/ 1ns**
TEST 2	-0.16	10.5		V-RESCALE	BERENDSEN		
TEST 3	2.20	-10.6		V-RESCALE	PARRINELLO		
TEST 4	1.56	-13.6	V-RESCALE	NOSE-HOOVER	PARRINELLO		
TEST 5	0.99	5.4		V-RESCALE	BERENDSEN		
TEST 6	-6.89	0.6		V-RESCALE	PARRINELLO		
TEST 7	0.6	0.8		V-RESCALE	PARRINELLO		
TEST 8	**-0.01**	**7.2**		V-RESCALE	BERENDSEN	Protein+Ligand+cofactor and water+ions i.e: GROUP 1: Protein; GROUP G1P+H2U+ZN(2ATOMS)+ 2: WATER+Cl/Na	**500ps /3ns**
TEST 9	**0.9**	**0.5**		**V-RESCALE**	**PARRINELLO**		

Table 3.18: Test to evaluate the best combination of thermostat and barostat during NVT and NPT equilibration

To evaluate how the time of equilibration has contributed to achieve the best result, we compared test 6 with test 7. If we consider only the mean value of pressure, we went from a value of -6.89 to a value of 0.6, which is far close to 1.

We finally confirmed that the best combination of thermostat and barostat results in V-rescale and Parrinello-Rahman, respectively, in agreement with tutorial of A. Lemkul version 2018 (http://www.mdtutorials.com/gmx/lysozyme/index.html).

Following this very careful and detailed study, a new protocol of MD (Figure 3.25) was created, through which we performed 600 ns-long MD simulations with two replicas. The analysis of these simulations is presently (September 2022) still ongoing.

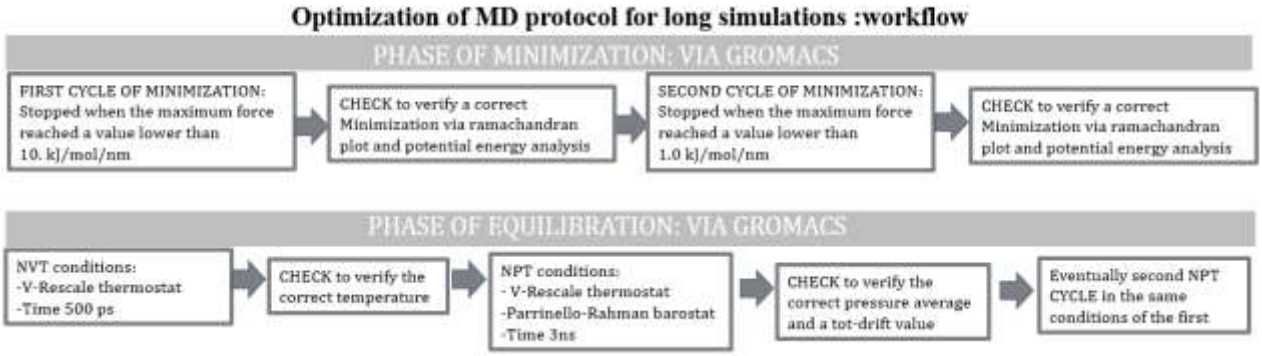


Figure 3.25: Worfkflow of the phase of new protocol for long MD simulations.

4. CONCLUSIONS

During this Ph.D. project, we focused on the deepening of the knowledge about wtGALT and p.Gln188Arg, the most common pathogenic mutant form of GALT enzyme [Timson, 2016], which is associated with the most severe phenotype and to a poor outcome of the classic galactosemia disease. This mutant enzyme has no or barely detectable enzymatic activity in the erythrocytes and liver of homozygous patients, and less than 50% activity in heterozygous individuals [Marabotti et al., 2005]. It has been proposed that this partial dominant effect could be related to the perturbation of the molecular interface between the two subunits forming the quaternary structure of the enzyme, considering that Gln188 is not only a residue of the active site but also a residue located at the interface between the two subunits [Marabotti et al., 2005, d'Acierno et al., 2018; McCorvie et al., 2016].

In the past, and more recently, the structural effects of this mutation were deduced on the static structure of the wild-type human enzyme; however, as first stage, we felt that a dynamic view of the proteins is necessary to deeply understand their behavior and obtain tips for possible therapeutic interventions. To carry out this study, we have performed MD of wtGALT enzyme and of its pathogenic mutant p.Gln188Arg under different experimental conditions, using as starting point the best conformation of docking. From our results, at body temperature (310 K) it appears that the negative effects of the mutation on the intersubunit interactions are more evident in the presence of the ligands (G1P and H2U). Indeed, the wild-type enzyme bound to the substrates shows an increased number of intersubunit H-bonds, most of which are not predicted in the static structure but are formed and persist during the MD simulations. On the contrary, the mutant p.Gln188Arg shows a marked decrease of the number of intersubunit H-bonds in the presence of the substrates. The number of intersubunit salt bridges is very small and it is not possible to infer if the variations detected are significant or not, but their trend in both wtGALT and pGln188Arg is analogous to that of H-bonds. It is also interesting to see that, despite this being an homodimeric enzyme,

the flexibility of the same segments in the two subunits is not of the same extent. This is intriguing, considering that the zones characterized by higher flexibility are either at the subunit interface or are involved in the stabilization of Zn, which is considered to have an important structural role for this enzyme [d'Acierno et al., 2018, McCorvie et al., 2016]. The higher temperature used to perturb our systems (334 K) seems to have few effects on the overall structure of the enzyme, but in these simulations, it is also possible to see that the mutation perturbs the quaternary assembly of the enzyme. Overall, our simulations confirm the importance of the intersubunit interactions of GALT for its correct functioning and suggest that their preservation in the mutant could improve the functioning of the enzyme, thereby rescuing, at least partially, its activity. Simultaneously, the lack of information about molecular interactions of arginine amino acid with respect to the protein prompted us to investigate its binding in the active site and central cavity of both wild type and p.Gln188Arg mutant. In this work, we did not find clear evidence about the ability of arginine to counteract the unfavorable effects of the mutation p.Gln188Arg in the mutant most often associated with classic galactosemia. In particular, the putative binding of arginine to the active site in the mutant enzyme is predicted to create a cluster of positive charges that further destabilizes the quaternary structure, and that, at last, can result in the expulsion of the arginine itself from the site. The putative binding of arginine to the central cavity is predicted to have more favorable effects on the overall structure and function of the enzyme, but also, in this case, we have no clear evidence of a stabilization of the enzymatic structure. Thus, the favorable effect (if any) of arginine on this enzyme is not predicted to be due to an activity similar to that of other pharmacochaperones. Notably, however, arginine is predicted to stably bind to some residues, one of which belongs to a cavity of the enzyme that was previously identified as an allosteric site. This cavity could be considered as a possible target for the development of true pharmacochaperones, also taking into account the interactions identified as crucial in

this study and in the other reported above that we conducted on this system [Verdino et al., 2021a].

These two parallel works prompted us to ask whether there really is an allosteric site in GALT enzyme, as also speculated in the literature [McCorvie et al., 2013], and if this allosteric site could be used as a target to develop new PCs for this enzyme. We were able to identify a potential allosteric site on chain A and one on chain B of GALT enzyme.

The identification of the potential allosteric sites occurred simultaneously with the search for new PCs already in therapeutic use, selecting drugs approved for pathologies due and not to misfolding. This search is resulted in the selection of five putative PCs (PC1, PC2, PC3, PC4, PC5). The next step, the search for pharmacophores starting from the best docking conformations of previous PCs, led to the identification of new hits, which were selected for further docking on the allosteric site. Other ligands, in particular HIT5, HIT7, HIT10 and HIT11, seem to interact with hGALT. Preliminary experimental tests performed at Utah University in collaboration with prof. Kent Lai seem to highlight the ability of some of these compounds to lower the levels of galactose-1-P in fibroblasts extracted from galactosemic patients (personal communication); further experiments will be needed to confirm this preliminary evidence.

The future prospects include the search for potential allosteric pathways in hGALT. To achieve this objective, the MD protocol was improved, testing the best experimental conditions for the hGALT system and ensuring the most reliable results starting from longer dynamics.

In details, the future analyses will consider the allosteric communication paths on sets of structures derived from these long MD simulations. To achieve this, we plan to perform principal component analysis (PCA) [David and Jacobs, 2014] to reveal the most important motions in proteins. Moreover, in order to capture the multi-modal behaviors of some atoms, which often play essential roles, particularly at the interfaces

of macromolecules, just like GALT enzyme, the dynamic cross correlation (DCC) analysis has been planned as subsequent essential analysis. Moreover, the protein motion could be represented as a linear combination of mutually independent normal mode vectors, throughout a normal mode analysis. This analysis gives results similar to those produced by PCA of a molecular dynamics simulation, but with only a fraction of the computational effort. It is also possible to represent the effect of external perturbations, e.g., ligand binding or tightly packed amino acid residues interacting with each other.

References

Abascal JL, Vega C. A general purpose model for the condensed phases of water: TIP4P/2005. J Chem Phys. 2005;123(23):234505.

Abraham MJ, Murtola T, Schulz R, et al. GROMACS: high perfor- mance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015a;1–2:19–25.

Abraham MJ, van der Spoel D, Lindahl E, Hess B., and the GROMACS development team. GROMACS User Manual version 5.0.7, 2015b, www.gromacs.org

Acosta PB, Gross KC. Hidden sources of galactose in the environment. Eur J Pediatr. 1995;154(7 Suppl 2):S87-92.

Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. General Principles of Cell Communication. Available from: https://www.ncbi.nlm.nih.gov/books/NBK26813/

Allen JT, Holton JB, Lennox AC, Hodges IC. Early morning urine galactitol levels in relation to galactose intake: a possible method of monitoring the diet in galactokinase deficiency. J Inherit Metab Dis. 1988;11 Suppl 2:243-5.

Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, Case DA, Kuntz ID, Rizzo RC. DOCK 6: Impact of new features and current docking performance. J Comput Chem. 2015;36(15):1132-56.

Alméciga-Diaz CJ, Hidalgo OA, Olarte-Avellaneda S, Rodríguez-López A, Guzman E, Garzón R, Pimentel-Vera LN, Puentes-Tellez MA, Rojas-Rodriguez AF, Gorshkov K, Li R, Zheng W. Identification of Ezetimibe and Pranlukast as Pharmacological Chaperones for the Treatment of the Rare Disease Mucopolysaccharidosis Type IVA. J Med Chem. 2019;62(13):6175-6189.

Anderson RJ, Weng Z, Campbell RK, Jiang X. Main-chain conformational tendencies of amino acids. Proteins. 2005 Sep 1;60(4):679-89. Andreotti G, Citro V, De Crescenzo A, et al. Therapy of Fabry disease with pharmacological chaperones: from in silico pre- dictions to in vitro tests. Orphanet J Rare Dis 2011;6:66.

Attwood TK, Gisel A, Eriksson NE, Bongcam-Rudloff E. Concepts, historical milestones and the central place of bioinformatics in modern biology: a European perspective. Bioinformatics - Trends and Methodologies, Dr. Mahmood A. Mahdavi (Ed.), InTech, 2011, 1-38.

Balakumar C, Ramesh M, Tham CL, Khathi SP, Kozielski F, Srinivasulu C, Hampannavar GA, Sayyad N, Soliman ME, Karpoormath R. Ligand- and structure-based in silico studies to identify kinesin spindle protein (KSP) inhibitors as potential anticancer agents. J Biomol Struct Dyn. 2018;36(14):3687-3704.

Banford S, McCorvie TJ, Pey AL, Timson DJ. Galactosemia: Towards Pharmacological Chaperones. J Pers Med. 2021;11(2):106.

Banitt I, Wolfson HJ. ParaDock: a flexible non-specific DNA--rigid protein docking algorithm. Nucleic Acids Res. 2011;39(20):e135.

Bansil R, Turner BS. Mucin structure, aggregation, physiological functions and biomedical applications. Current Opinion in Colloid & Interface Science. 2006;11:164–170.

Basconi JE, Shirts MR. Effects of Temperature Control Algorithms on Transport Properties and Kinetics in Molecular Dynamics Simulations. J Chem Theory Comput. 2013;9(7):2887-99.

Baynes BM, Wang DI, Trout BL. Role of arginine in the stabilization of proteins against aggregation. Biochemistry. 2005;44(12):4919-25.

Beck M, Sieber N, Goebel HH. Progressive cerebellar ataxia in juvenile GM2-gangliosidosis type Sandhoff. Eur J Pediatr. 1998;157(10):866-7.

Bell D, editor. Natural monosaccharides and oligosaccharides: their structures and occurrence. In Florkin M, editor. Comparative biochemistry: a comprehensive treatise. Vol. 3. Elsevier; 2012. pp. 287–354.

Berendse K, Ebberink MS, Ijlst L, Poll-The BT, Wanders RJ, Waterham HR. Arginine improves peroxisome functioning in cells from patients with a mild peroxisome biogenesis disorder. Orphanet J Rare Dis. 2013;8:138.

Berendsen H.J.C, Postma J.P.M, van Gunsteren W.F, DiNola, A.and Haak J.R Molecular dynamics with coupling to an external bath. 1984 Apr; J. Chem. Phys. 81, 3684

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42.

Berry GT. Classic Galactosemia and Clinical Variant Galactosemia. 2000 Feb 4 [Updated 2021 Mar 11]. In: Adam MP, Everman DB, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. Available from: https://www.ncbi.nlm.nih.gov/books/NBK1518/

Bhat TN, Bourne P, Feng Z, Gilliland G, Jain S, Ravichandran V, Schneider B, Schneider K, Thanki N, Weissig H, Westbrook J, Berman HM. The PDB data uniformity project. Nucleic Acids Res. 2001;29(1):214-8.

Bianco G, Forli S, Goodsell DS, Olson AJ. Covalent docking using autodock: Two-point attractor and flexible side chain methods. Protein Sci. 2016;25(1):295-301.

Binkowski TA, Naghibzadeh S, Liang J. CASTp: Computed Atlas of Surface Topography of proteins. Nucleic Acids Res. 2003;31(13):3352-5.

Bowers K.J, Chow E., Xu H.,O. Dror R., Eastwood M.P, Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *ACM/IEEE SC 2006 Conference (SC'06)* 2006: 43-43.

Bray PT, Isaac RJ, Watkins AG. Galactosaemia. Arch Dis Child. 1952;27(134):341-7.

Brenke R, Kozakov D, Chuang GY, Beglov D, Hall D, Landon MR, Mattos C, Vajda S. Fragment-based identification of druggable 'hot spots' of proteins using Fourier domain correlation techniques. Bioinformatics. 2009;25(5):621-7.

Brenner C. Hint, Fhit, and GalT: function, structure, evolution, and mechanism of three branches of the histidine triad superfamily of nucleotide hydrolases and transferases. Biochemistry. 2002;41(29):9003-14.

Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30(10):1545-614.

Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. J. Chem. Phys. 2007, 126, 014101.

Bussi G, Zykova-Timan T, Parrinello M. Isothermal-isobaric molecular dynamics using stochastic velocity rescaling. J Chem Phys. 2009;130(7):074101.

Calderon FR, Phansalkar AR, Crockett DK, Miller M, Mao R. Mutation database for the galactose-1-phosphate uridyltransferase (GALT) gene. Hum Mutat. 2007;28(10):939-43.

Callaway E. Revolutionary cryo-EM is taking over structural biology. Nature. 2020 Feb;578(7794):201.

Carlile GW, Yang Q, Matthes E, Liao J, Radinovic S, Miyamoto C, Robert R, Hanrahan JW, Thomas DY. A novel triple combination of pharmacological chaperones improves F508del-CFTR correction. Sci Rep. 2018;8(1):11404.

Case DA, Cheatham TE 3rd, Darden T, Gohlke H, Luo R, Merz KM Jr, Onufriev A, Simmerling C, Wang B, Woods RJ. The Amber biomolecular simulation programs. J Comput Chem. 2005;26(16):1668-88.

Castrignanò T, Gioiosa S, Flati T, Cestari M, Picardi E, Chiara M, Fratelli M, Amente S, Cirilli M, Tangaro MA, Chillemi G, Pesole G, Zambelli F. ELIXIR-IT HPC@CINECA: high performance computing resources for the bioinformatics community. BMC Bioinformatics. 2020 A;21(Suppl 10):352.

Ceperley D.M and Libby S.B Berni Julian Alder, theoretical physicist and inventor of molecular dynamics, 1925–2020. 2021; Vol.118 No.11

Chai Y, Beauregard PB, Vlamakis H, Losick R, Kolter R. Galactose metabolism plays a crucial role in biofilm formation by Bacillus subtilis. mBio. 2012;3(4):e00184-12.

Citro V, Cammisa M, Liguori L, Cimmaruta C, Lukas J, Cubellis MV, Andreotti G. The Large Phenotypic Spectrum of Fabry Disease Requires Graduated Diagnosis and Personalized Therapy: A Meta-Analysis Can Help to Differentiate Missense Mutations. Int J Mol Sci. 2016;17(12):2010.

Citro V, Cimmaruta C, Monticelli M, Riccio G, Hay Mele B, Cubellis MV, Andreotti G. The Analysis of Variants in the General Population Reveals That PMM2 Is Extremely Tolerant to Missense Mutations and That Diagnosis of PMM2-CDG Can Benefit from the Identification of Modifiers. Int J Mol Sci. 2018;19(8):2218.

Coelho AI, Berry GT, Rubio-Gozalbo ME. Galactose metabolism and health. Curr Opin Clin Nutr Metab Care. 2015a;18(4):422-7.

Coelho AI, Trabuco M, Silva MJ, de Almeida IT, Leandro P, Rivera I, Vicente JB. Arginine Functionally Improves Clinically Relevant Human Galactose-1-Phosphate Uridylyltransferase (GALT) Variants Expressed in a Prokaryotic Model. JIMD Rep. 2015b;23:1-6.

Coelho, A.I.; Trabuco, M.; Ramos, R.; Silva, M.J.; Tavares de Almeida, I.; Leandro, P.; Rivera, I.; Vicente, J.B. Functional andstructural impact of the most prevalent missense mutations in classic galactosemia. Mol. Genet. Genom. Med. 2014, 2, 484–496.

d'Acierno A, Facchiano A, Marabotti A. GALT protein database: querying structural and functional features of GALT enzyme. Human Mutation. 2014;35(9):1060-1067.

d'Acierno A, Scafuri B, Facchiano A, Marabotti A. The evolution of a Web resource: The Galactosemia Proteins Database 2.0. Hum Mutat. 2018;39(1):52-60.

Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An Nlog(N) method for Ewald sums in large systems. J. Chem. Phys.1993, 98, 10089–10092.

David CC, Jacobs DJ. Principal component analysis: a method for determining the essential dynamics of proteins. Methods Mol Biol. 2014;1084:193-226

de Baulny HO, Abadie V, Feillet F, de Parscau L. Management of phenylketonuria and hyperphenylalaninemia. J Nutr. 2007 Jun;137(6 Suppl 1):1561S-1563S; discussion 1573S-1575S.

Demain AL, Elander RP. The beta-lactam antibiotics: past, present, and future. Antonie Van Leeuwenhoek. 1999;75(1-2):5-19.

De-Souza EA, Pimentel FS, Machado CM, Martins LS, da-Silva WS, Montero-Lomelí M, Masuda CA. The unfolded protein response has a protective role in yeast models of classic galactosemia. Dis Model Mech. 2014 Jan;7(1):55-61.

Di Costanzo L, Ghosh S, Zardecki C, Burley S.K. Using the Tools and Resources of the RCSB Protein Data Bank, 2016;55 1.9.1 - 1.9.35

Drwal MN, Agama K, Wakelin LP, Pommier Y, Griffith R. Exploring DNA topoisomerase I ligand space in search of novel anticancer agents. PLoS One. 2011;6(9):e25150.

Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol. 2011;9:71.

Elsas LJ, Dembure PP, Langley S, Paulk EM, Hjelm LN, Fridovich-Keil J. A common mutation associated with the Duarte galactosemia allele. Am J Hum Genet. 1994;54(6):1030-6.

Elsevier JP, Fridovich-Keil JL. The Q188R mutation in human galactose-1-phosphate uridylyltransferase acts as a partial dominant negative. J Biol Chem. 1996;271(50):32002-7
.
Endres W, Shin YS. Cataract and metabolic disease. J Inherit Metab Dis. 1990;13(4):509-16.

Ercolessi F., A molecular dynamics primer. International School for Advanced Studies, 1997 Trieste, IT
.
Eslami H, Mojahedi F, Moghadasi J. Molecular dynamics simulation with weak coupling to heat and material baths. J Chem Phys. 2010;133(8):084105.

Evers A, Gohlke H, Klebe G. Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. J Mol Biol. 2003;334(2):327-45

Facchiano A, Marabotti A. Analysis of galactosemia-linked mutations of GALT enzyme using a computational biology approach. Protein Eng Des Sel. 2010;23(2):103-13.

Fairhead M., Strain-Damerell C., Kopec J., Bezerra G.A., Zhang M., Burgess-Brown N., von Delft F., Arrowsmith C., Edwards, A., Bountra C., Yue W.W. Structure of human galactose-1-phosphate uridylyltransferase (GALT), with crystallization epitope mutations A21Y:A22T:T23P:R25L (*To be published)*

Fan JQ, Ishii S, Asano N, Suzuki Y. Accelerated transport and maturation of lysosomal alpha-galactosidase A in Fabry lymphoblasts by an enzyme inhibitor. Nat Med. 1999;5(1):112-5.

Feinstein W, Brylinski M. Structure-Based Drug Discovery Accelerated by Many-Core Devices. Curr Drug Targets. 2016;17(14):1595-1609.

Flanagan JM, McMahon G, Brendan Chia SH, Fitzpatrick P, Tighe O, O'Neill C, Briones P, Gort L, Kozak L, Magee A, Naughten E, Radomyska B, Schwartz M, Shin JS, Strobl WM, Tyfield LA, Waterham HR, Russell H, Bertorelle G, Reichardt JK, Mayne PD, Croke DT. The role of human demographic history in determining the distribution and frequency of transferase-deficient galactosaemia mutations. Heredity (Edinb). 2010;104(2):148-54.

Fridovich-Keil JL, Quimby BB, Wells L, Mazur LA, Elsevier JP. Characterization of the N314D allele of human galactose-1-phosphate uridylyltransferase using a yeast expression system. Biochem Mol Med. 1995;56(2):121-30.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem. 2004;47(7):1739-49.

Fung WL, Risch H, McLaughlin J, Rosen B, Cole D, Vesprini D, Narod SA. The N314D polymorphism of galactose-1-phosphate uridyl transferase does not modify the risk of ovarian cancer. Cancer Epidemiol Biomarkers Prev. 2003;12(7):678-80

.

Gasteiger, J. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. Tetrahedron 1980, 36,3219–3228.

Genoni A, Morra G, Colombo G. Identification of domains in protein structures from the analysis of intramolecular interactions. J Phys Chem B. 2012;116(10):3331-43.

Gitzelman R, Auricchio S. The Handling of soya Apha-Galactosides by a normal and a galactosemin child. Pediatrics. 1965;36:231-5. PMID: 14320033.

Goldberg D. The Design Innovation.Lessons from and for Competent Genetic Algorithms. GENA; 2002;volume 7

Goodsell DS, Olson AJ. Automated docking of substrates to proteins by simulated annealing. Proteins. 1990;8(3):195-202.

Graff DE, Shakhnovich EI, Coley CW. Accelerating high-throughput virtual screening through molecular pool-basedactive learning. Chem Sci 2021;12:7866–81.

Grubmüller H, Heller H,Windemuth A, Schulten K. Generalized Verlet Algorithm for Efficient Molecular Dynamics Simulations with Long-range Interactions, Molecular Simulation,.1991;6:1-3, 121-142.

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem. 2004;47(7):1750-9.

Hanrahan JW, Sato Y, Carlile GW, Jansen G, Young JC, Thomas DY. Cystic Fibrosis: Proteostatic correctors of CFTR trafficking and alternative therapeutic targets. Expert Opin Ther Targets. 2019;23(8):711-724.

Haskovic M, Derks B, van der Ploeg L, Trommelen J, Nyakayiru J, van Loon LJC, Mackinnon S, Yue WW, Peake RWA, Zha L, Demirbas D, Qi W, Huang X, Berry GT, Achten J, Bierau J, Rubio-Gozalbo ME, Coelho AI. Arginine does not rescue p.Q188R mutation deleterious effect in classic galactosemia. Orphanet J Rare Dis. 2018;13(1):212.

Hennermann JB, Schadewaldt P, Vetter B, Shin YS, Mönch E, Klein J. Features and outcome of galactokinase deficiency in children diagnosed by newborn screening. J Inherit Metab Dis. 2011;34(2):399-407

Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comp Chem 1997;18:1463–72.

Holden HM, Rayment I, Thoden JB. Structure and function of enzymes of the Leloir pathway for galactose metabolism. J Biol Chem. 2003;278(45):43885-8.

Hollingsworth SA, Dror RO. Molecular Dynamics Simulation for All. Neuron. 2018 Sep 19;99(6):1129-1143.

Holton JB, Gillett MG, MacFaul R, Young R. Galactosaemia: a new severe variant due to uridine diphosphate galactose-4-epimerase deficiency. Arch Dis Child. 1981;56(11):885-7.

Huang SY, Zou X. Ensemble docking of multiple protein structures: considering protein structural variations in molecular docking. Proteins. 2007;66(2):399-421.

Humphrey W, Dalke A, Schulten K. VMD - Visual Molecular Dynamics. J Mol Graph 1996;14:33-8.

Ivanova MM, Changsila E, Turgut A, Goker-Alpan O. Individualized screening for chaperone activity in Gaucher disease using multiple patient derived primary cell lines. Am J Transl Res. 2018;10(11):3750-3761.

Iwasawa S, Kikuchi A, Wada Y, Arai-Ichinoi N, Sakamoto O, Tamiya G, Kure S. The prevalence of GALM mutations that cause galactosemia: A database of functionally evaluated variants. Mol Genet Metab. 2019;126(4):362-367.

Jamroz M, Orozco M, Kolinski A, Kmiecik S. Consistent View of Protein Fluctuations from All-Atom Molecular Dynamics and Coarse-Grained Dynamics with Knowledge-Based Force-Field. J Chem Theory Comput. 2013;9(1):119-25.

Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 2017;33(19):3036-3042.

Jo S, Cheng X, Lee J, Kim S, Park SJ, Patel DS, Beaven AH, Lee KI, Rui H, Park S, Lee HS, Roux B, MacKerell AD Jr, Klauda JB, Qi Y, Im W. CHARMM-GUI 10 years for biomolecular modeling and simulation. J Comput Chem. 2017;38(15):1114-1124.

Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem. 2008 (11):1859-65.

Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596, 583–589.

Kalckar HM. Biochemical genetics as illustrated by hereditary galactosemia. Am J Clin Nutr. 1961;9:676-82.

Karas AZ, Goldberg AM. Recognizing signs of pain. J Am Vet Med Assoc. 2003;223(3):298-9; author reply 299.

Kemna MJ, Plomp R, van Paassen P, Koeleman CAM, Jansen BC, Damoiseaux JGMC, Cohen Tervaert JW, Wuhrer M. Galactosylation and Sialylation Levels of IgG Predict Relapse in Patients With PR3-ANCA Associated Vasculitis. EBioMedicine. 2017;17:108-118.

Khedkar SA, Malde AK, Coutinho EC, Srivastava S. Pharmacophore modeling in drug discovery and development: an overview. Med Chem. 2007;3(2):187-97.

Kikuchi A, Wada Y, Ohura T, Kure S. The Discovery of GALM Deficiency (Type IV Galactosemia) and Newborn Screening System for Galactosemia in Japan. Int J Neonatal Screen. 2021;7(4):68.

Kim R, Emi M, Tanabe K, Murakami S. Role of the unfolded protein response in cell death.Apoptosis. 2006;11(1):5-13.

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, Bolton EE. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102-D1109.

Kingsley DM, Krieger M, Holton JB. Structure and function of low-density-lipoprotein receptors in epimerase-deficient galactosemia. N Engl J Med. 1986;314(19):1257-8.

Kirkpatrick, S. Optimization by simulated annealing: Quantitative studies. J Stat Phys. 1984; 34, 975–986.

Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004;3(11):935-49.

Koshland De Jr. Enzyme flexibility and enzyme action. J Cell Comp Physiol. 1959;54:245-58.

Kotb MA, Mansour L, William Shaker Basanti C, El Garf W, Ali GIZ, Mostafa El Sorogy ST, Kamel IEM, Kamal NM. Pilot study of classic galactosemia: Neurodevelopmental impact and other complications urge neonatal screening in Egypt. J Adv Res. 2018;12:39-45.

Kotb MA, Mansour L, Shamma RA. Screening for galactosemia: is there a place for it? Int J Gen Med. 2019;12:193-205

Kozák L, Francová H, Fajkusová L, Pijácková A, Macku J, Stastná S, Peskovová K, Martincová O, Krijt J, Bzdúch V. Mutation analysis of the GALT gene in Czech and Slovak galactosemia populations: identification of six novel mutations, including a stop codon mutation (X380R). Hum Mutat. 2000;15(2):206.

Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D, Vajda S. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat Protoc. 2015;10(5):733-55.

Kramer B, Rarey M, Lengauer T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins. 1999;37(2):228-41.

Kraskiewicz H, FitzGerald U. InterfERing with endoplasmic reticulum stress. Trends Pharmacol Sci. 2012 Feb;33(2):53-63.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol. 1982;161(2):269-88.

Lai K, Langley SD, Singh RH, Dembure PP, Hjelm LN, Elsas LJ 2nd. A prevalent mutation for galactosemia among black Americans. J Pediatr. 1996;128(1):89-95.

Lam C, Krasnewich DM. PMM2-CDG. 2005 Aug 15 [updated 2021 May 20]. In: Adam MP, Everman DB, Mirzaa GM, Pagon RA, Wallace SE, Bean LJH, Gripp KW, Amemiya A, editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993–2022..

Langley SD, Lai K, Dembure PP, Hjelm LN, Elsas LJ. Molecular basis for Duarte and Los Angeles variant galactosemia. Am J Hum Genet. 1997;60(2):366-72.

Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, Wei S, Buckner J, JeonJC, Qi Y, Jo S, Pande VS, Case DA, Brooks CL 3rd, MacKerell AD Jr, Klauda JB, Im W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. J Chem Theory Comput. 2016 Jan 12;12(1):405-13.

Leloir LF. The enzymatic transformation of uridine diphosphate glucose into a galactose derivative. Arch Biochem Biophys. 1951;33(2):186-90.

Lemkul JA. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package, v1.0. Living J. Comp. Mol. Sci. 2018,. 1 (1), 5068.

Leport C, Chêne G, Morlat P, Luft BJ, Rousseau F, Pueyo S, Hafner R, Miro J, Aubertin J, Salamon R, Vildé JL. Pyrimethamine for primary prophylaxis of toxoplasmic encephalitis in patients with human immunodeficiency virus infection: a double-blind, randomized trial. ANRS 005-ACTG 154 Group Members. Agence Nationale de Recherche sur le SIDA. AIDS Clinical Trial Group. J Infect Dis. 1996 Jan;173(1):91-7.

Liguori L, Monticelli M, Allocca M, Hay Mele B, Lukas J, Cubellis MV, Andreotti G. Pharmacological Chaperones: A Therapeutic Approach for Diseases Caused by Destabilizing Missense Mutations. Int J Mol Sci. 2020;21(2):489.

Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J.L.; Dror, R.O.; Shaw, D.E. Improved side-chain torsion potentialsfor the Amber ff99SB protein force field. Proteins 2010, 78, 1950–1958.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 1997;23(1–3):3-25

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46(1-3):3-26.

Liu, J. and Wang R. On Classification of Current Scoring Functions. Chem. Inf. Model. 2015 , 55, 3, 475–482

Lukac-Bajalo J, Mencej S, Karas N, Mlinar B, Zitnik IP, Gersak K. Q188R, K285N, and N314D mutation-associated alleles in the galactose-1-phosphate uridyltransferase gene and female infertility. Fertil Steril. 2005;83(3):776-8.

Lynch ME, Potter NL, Coles CD, Fridovich-Keil JL. Developmental Outcomes of School-Age Children with Duarte Galactosemia: A Pilot Study. JIMD Rep. 2015;19:75-84.

Maegawa GH, Tropak MB, Buttner JD, Rigat BA, Fuller M, Pandit D, Tang L, Kornhaber GJ, Hamuro Y, Clarke JT, Mahuran DJ. Identification and characterization of ambroxol as an enzyme enhancement agent for Gaucher disease. J Biol Chem. 2009;284(35):23502-16.

Manga N, Jenkins T, Jackson H, Whittaker DA, Lane AB. The molecular basis of transferase galactosaemia in South African negroids. J Inherit Metab Dis. 2007;22(1):37-42.

Moammar H, Ratard R, Cheriyan G, Mathew P. Incidence and features of galactosaemia in Saudi Arabs. J Inherit Metab Dis. 1996;19(3):331-4.

Marabotti A, Facchiano AM. Homology modeling studies on human galactose-1-phosphate uridylyltransferase and on its galactosemia-related mutant Q188R provide an explanation of the molecular effects of the mutation on homo- and heterodimers. J Med Chem. 2005;48(3):773-9

Matalonga L, Gort L, Ribes A. Small molecules as therapeutic agents for inborn errors of metabolism. J Inherit Metab Dis. 2017 Mar;40(2):177-193.

McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977 Jun 16;267(5612):585-90.

McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977 Jun 16;267(5612):585-90.

McCorvie TJ, Timson DJ. Structural and molecular biology of type I galactosemia: disease-associated mutations. IUBMB Life. 2011;63(11):949-54.

McCorvie TJ, Kopec J, Pey AL, Fitzpatrick F, Patel D, Chalk R, Shrestha L, Yue WW. Molecular basis of classic galactosemia from the structure of human galactose 1-phosphate uridylyltransferase. Hum Mol Genet. 2016;25(11):2234-2244.

McDonald I, Thornton,J. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238: 777–93.

Megarity CF, Huang M, Warnock C, Timson DJ. The role of the active site residues in human galactokinase: implications for the mechanisms of GHMP kinases. Bioorg Chem. 2011;39(3):120-6

Meslamani J, Li J, Sutter J, Stevens A, Bertrand HO, Rognan D. Protein-ligand-based pharmacophores: generation and utility assessment in computational ligand profiling. J Chem Inf Model. 2012;52(4):943-55.

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Žídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 2022 Jan 7;50(D1):

Moammar H, Ratard R, Cheriyan G, Mathew P. Incidence and features of galactosaemia in Saudi Arabs. J Inherit Metab Dis. 1996;19(3):331-4.

Morello JP, Salahpour A, Laperrière A, Bernier V, Arthus MF, Lonergan M, Petäjä-Repo U, Angers S, Morin D, Bichet DG, Bouvier M. Pharmacological chaperones rescue cell-surface expression and function of misfolded V2 vasopressin receptor mutants. J Clin Invest. 2000;105(7):887-95.

Moroni E, Agard DA, Colombo G. The Structural Asymmetry of Mitochondrial Hsp90 (Trap1) Determines Fine Tuning of Functional Dynamics. J Chem Theory Comput. 2018;14(2):1033-1044.

Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, Olson AJ. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30(16):2785-91.

Morris GM, Lim-Wilby M. Molecular docking. Methods Mol Biol. 2008;443:365-82.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem. 1998;19:1639–166

Muntau AC, Röschinger W, Habich M, Demmelmair H, Hoffmann B, Sommerhoff CP, Roscher AA. Tetrahydrobiopterin as an alternative treatment for mild phenylketonuria. N Engl J Med. 2002;347(26):2122-32.

Nam YW, Nishimoto M, Arakawa T, Kitaoka M, Fushinobu S. Structural basis for broad substrate specificity of UDP-glucose 4-epimerase in the human milk oligosaccharide catabolic pathway of Bifidobacterium longum. Sci Rep. 2019;9(1):11081.

Noorwez SM, Malhotra R, McDowell JH, Smith KA, Krebs MP, Kaushal S. Retinoids assist the cellular folding of the autosomal dominant retinitis pigmentosa opsin mutant P23H. J Biol Chem. 2004;279(16):16278-84.

Nosé S. A unified formulation of the constant temperature molecular dynamics methods. J. Chem. Phys 1984. 81, 511;

Openo KK, Schulz JM, Vargas CA, Orton CS, Epstein MP, Schnur RE, Scaglia F, Berry GT, Gottesman GS, Ficicioglu C, Slonim AE, Schroer RJ, Yu C, Rangel VE, Keenan J, Lamance K,

Fridovich-Keil JL. Epimerase-deficiency galactosemia is not a binary condition. Am J Hum Genet. 2006;78(1):89-102.

Owens J. Determining druggability. Nat Rev Drug Discov. 2007;6(3):187

Pampalone G, Grottelli S, Gatticchi L, Lombardi EM, Bellezza I, Cellini B. Role of misfolding in rare enzymatic deficits and use of pharmacological chaperones as therapeutic approach. Front Biosci (Landmark Ed). 2021;26(12):1627-1642.

Parenti G, Andria G, Valenzano KJ. Pharmacological Chaperone Therapy: Preclinical Development, Clinical Translation, and Prospects for the Treatment of Lysosomal Storage Disorders. Mol Ther. 2015;23(7):1138-1148.

Parrinello, M., and A. Rahman. 1981. Polymorphic transitions in single crystals: a new molecular dynamics method. J. Appl. Phys. 52:7182–7190.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25(13):1605-12.

Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, Buch R, Fiorin G, Hénin J, Jiang W, McGreevy R, Melo MCR, Radak BK, Skeel RD, Singharoy A, Wang Y, Roux B, Aksimentiev A, Luthey-Schulten Z, Kalé LV, Schulten K, Chipot C, Tajkhorshid E. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J Chem Phys. 2020;153(4):044130.

Podskarbi T, Kohlmetz T, Gathof BS, Kleinlein B, Bieger WP, Gresser U, Shin YS. Molecular characterization of Duarte-1 and Duarte-2 variants of galactose-1-phosphate uridyltransferase. J Inherit Metab Dis. 1996;19(5):638-44..

Ponder JW, Case DA. Force fields for protein simulations. Adv Prot Chem. 2003. 66, 27-85.

Rahman A. Correlations in the Motion of Atoms in Liquid Argon. Physical Review. 1964; 136:405-411

Reichardt JK, Woo SL. Molecular basis of galactosemia: mutations and polymorphisms in the gene encoding human galactose-1-phosphate uridylyltransferase. Proc Natl Acad Sci USA. 1991;88(7):2633-7.

Reichardt JK. Genetic basis of galactosemia. Hum Mutat. 1992;1(3):190-6.

Reinhardt LA, Thoden JB, Peters GS, Holden HM, Cleland WW. pH-rate profiles support a general base mechanism for galactokinase (Lactococcus lactis). FEBS Lett. 2013;587(17):2876-81

Rigsby RE, Parker AB. Using the PyMOL application to reinforce visual understanding of protein structure. Biochem Mol Biol Educ. 2016;44(5):433-7.

Roche DB, Buenavista MT, McGuffin LJ. The FunFOLD2 server for the prediction of protein-ligand interactions. Nucleic Acids Res. 2013;41(Web Server issue):W303-7.

Ruvinsky AM, Kozintsev AV. Novel statistical-thermodynamic methods to predict protein-ligand binding positions using probability distribution functions. Proteins. 2006;62(1):202-8.

Sanchez-Martin C, Moroni E, Ferraro M, Laquatra C, Cannino G, Masgras I, Negro A, Quadrelli P, Rasola A, Colombo G. Rational Design of Allosteric and Selective Inhibitors of the Molecular Chaperone TRAP1. Cell Rep. 2020 Apr 21;31(3):107531. doi: 10.1016/j.celrep.2020.107531. PMID: 32320652.

Scafuri B, Verdino A, D'Arminio N, Marabotti A. Computational methods to assist in the discovery of pharmacological chaperones for rare diseases. Brief Bioinform. 2022

Scheraga HA, Khalili M, Liwo A. Protein-Folding Dynamics: Overview of Molecular Simulation Techniques. Annu Rev Phys Chem. 2007. 58:57-83.

Sgambat K, Banks M, Moudgil A. Effect of galactose on glomerular permeability and proteinuria in steroid-resistant nephrotic syndrome. Pediatr Nephrol. 2013;28(11):2131-5.

Shaker B, Ahmad S, Lee J, Jung C, Na D. In silico methods and tools for drug discovery. Comput Biol Med. 2021;137:104851.

Shin YS, Zschocke J, Das AM, Podskarbi T. Molecular and biochemical basis for variants and deficiency forms of galactose-1-phosphate uridyltransferase. J Inherit Metab Dis. 1999;22(3):327-9.

Shoichet BK, Kuntz ID. Protein docking and complementarity. J Mol Biol. 1991;221(1):327-46.

Silva EP Jr, Borges LS, Mendes-da-Silva C, Hirabara SM, Lambertucci RH. l-Arginine supplementation improves rats' antioxidant system and exercise performance. Free Radic Res. 2017;51(3):281-293.

Silveira CRA, MacKinley J, Coleman K, Li Z, Finger E, Bartha R, Morrow SA, Wells J, Borrie M, Tirona RG, Rupar CA, Zou G, Hegele RA, Mahuran D, MacDonald P, Jenkins ME, Jog M, Pasternak SH. Ambroxol as a novel disease-modifying treatment for Parkinson's disease dementia: protocol for a single-centre, randomized, double-blind, placebo-controlled trial. BMC Neurol. 2019;19(1):20.

Singh R, Thapa BR, Kaur G, Prasad R. Frequency distribution of Q188R, N314D, Duarte 1, and Duarte 2 GALT variant alleles in an Indian galactosemia population. Biochem Genet. 2012;50(11-12):871-80

Slepak T, Tang M, Addo F, Lai K. Intracellular galactose-1-phosphate accumulation leads to environmental stress response in yeast model. Mol Genet Metab. 2005;86(3):360-71

Slepak TI, Tang M, Slepak VZ, Lai K. Involvement of endoplasmic reticulum stress in a novel Classic Galactosemia model. Mol Genet Metab. 2007;92(1-2):78-87.

Sneha P, Ebrahimi EA, Ghazala SA, D TK, R S, Priya Doss C G, Zayed H. Structural analysis of missense mutations in galactokinase 1 (GALK1) leading to galactosemia type-2. J Cell Biochem. 2018;119(9):7585-7598.

Solis FJ, Wets RJB. Minimization by random search techniques. Maths Operat Res 1981;6(1):19-30.

Sousa da Silva AW, Vranken WF. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes. 2012 23;5:367.

Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins. 2006;65(1):15-26.

Sterling T, Irwin JJ. ZINC 15--Ligand Discovery for Everyone. J Chem Inf Model. 2015;55(11):2324-37.

Stillinger F. Improved simulation of liquid water by molecular dynamics.1973 J. Chem. Phys. 60, 1545

Sun Y; Liou B; Xu YH; Quinn B; Zhang W; Hamler R; Setchell KD; Grabowski GA. Ex vivo and in vivo eects of isofagomine on acid beta-glucosidase variants and substrate levels in gaucher disease. J Biol Chem 2012;287:4275–87.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Res 2021;49(D1):D480–9.

Thoden JB, Kim J, Raushel FM, Holden HM. The catalytic mechanism of galactose mutarotase. Protein Sci. 2003;12(5):1051-9.

Thomas R, Kermode AR. Enzyme enhancement therapeutics for lysosomal storage diseases: Current status and perspective. Mol Genet Metab.;126(2):83-97.

Tian W, Chen C, Lei X, Zhao J, Liang J. CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. 2018;46(W1):W363-W367.

Timson DJ. The molecular basis of galactosemia - Past, present and future. Gene. 2016, 589, 133-141

Timson DJ. Type IV galactosemia. Gene. 2019;21(6):1283–5

Totrov M, Abagyan R. Flexible protein-ligand docking by global energy optimization in internal coordinates. Proteins 1997;Suppl 1:215-20.

Tropak MB, Yonekawa S, Karumuthil-Melethil S, Thompson P, Wakarchuk W, Gray SJ, Walia JS, Mark BL, Mahuran D. Construction of a hybrid β-hexosaminidase subunit capable of forming stable homodimers that hydrolyze GM2 ganglioside in vivo. Mol Ther Methods Clin Dev. 2016;3:15057.

Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-61.

Tyfield L, Reichardt J, Fridovich-Keil J, Croke DT, Elsas LJ 2nd, Strobl W, Kozak L, Coskun T, Novelli G, Okano Y, Zekanowski C, Shin Y, Boleda MD. Classical galactosemia and mutations at the galactose-1-phosphate uridyl transferase (GALT) gene. Hum Mutat. 1999;13(6):417-30..

Urquiza P, Laín A, Sanz-Parra A, Moreno J, Bernardo-Seisdedos G, Dubus P, González E, Gutiérrez-de-Juan V, García S, Eraña H, San Juan I, Macías I, Ben Bdira F, Pluta P, Ortega G, Oyarzábal J, González-Muñiz R, Rodríguez-Cuesta J, Anguita J, Díez E, Blouin JM, de Verneuil H, Mato JM, Richard E, Falcón-Pérez JM, Castilla J, Millet O. Repurposing ciclopirox as a pharmacological chaperone in a model of congenital erythropoietic porphyria. Sci Transl Med. 2018;10(459):eaat7467.

van Gunsteren,W.F.; Berendsen, H.J.C. A leap-frog algorithm for stochastic dynamics. Mol. Simul. 1988, 1, 173–185.

Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, Mackerell AD Jr. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010;31(4):671-90.

Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003;52(4):609-23.

Verlet L. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Phys Rev 1967:159:98–103.

Viggiano E, Marabotti A, Politano L, Burlina A. Galactose-1-phosphate uridyltransferasdeficiency: A literature review of the putative mechanisms of short and long-term complications and allelic variants. Clin Genet. 2018;93(2):206-215.

Vilar S, Cozza G, Moro S. Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. Curr Top Med Chem. 2008;8(18):1555-72.

Vuorinen A, Schuster D. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. Methods. 2015;71:113-34.

Wada Y, Kikuchi A, Arai-Ichinoi N, Sakamoto O, Takezawa Y, Iwasawa S, Niihori T, Nyuzuki H, Nakajima Y, Ogawa E, Ishige M, Hirai H, Sasai H, Fujiki R, Shirota M, Funayama R, Yamamoto M, Ito T, Ohara O, Nakayama K, Aoki Y, Koshiba S, Fukao T, Kure S. Correction: Biallelic GALM pathogenic variants cause a novel type of galactosemia. Genet Med. 2020;22(7):1281.

Wang BB, Xu YK, Ng WG, Wong LJ. Molecular and biochemical basis of galactosemia. Mol Genet Metab. 1998;63(4):263-9..

Wang J, Wang W, Kollman PA, Case DA. Automatic atom typeand bond type perception in molecular mechanical calculations. J Mol Graph Mod 2006;25:247-60.

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and Testing of a General Amber Force Field. J Comp Chem 2004;25:1157 - 73.

Wass MN, Kelley LA, Sternberg MJ. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucleic Acids Res. 2010;38(Web Server issue):W469-73.

Wedekind JE, Frey PA, Rayment I. Three-dimensional structure of galactose-1-phosphate uridylyltransferase from Escherichia coli at 1.8 A resolution. Biochemistry. 1995;34(35):11049-61.

Wedekind JE, Frey PA, Rayment I. The structure of nucleotidylated histidine-166 of galactose-1-phosphate uridylyltransferase provides insight into phosphoryl group transfer. Biochemistry. 1996;35(36):11560-9.

Weiss LM, Luft BJ, Tanowitz HB, Wittner M. Pyrimethamine concentrations in serum during treatment of acute murine experimental toxoplasmosis. Am J Trop Med Hyg. 1992;46(3):288-91.

Wermuth G., Ganellin C.R., Lindberg P., Mitscher L.A. Glossary of terms used in medicinal chemistry (iupac recommendations 1998). Pure Appl. Chem. 1998;70:1129–1143.

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46(D1):D1074-D1082.

Wolber G, Dornhofer AA, Langer T. Efficient overlay of small organic molecules using 3D pharmacophores. J Comput Aided Mol Des. 2006;20(12):773-88.

Wong KM, Tai HK, Siu SWI. GWOVina: A grey wolf optimization approach to rigid and flexible receptor docking. Chem Biol Drug Des. 2021;97(1):97-110.

Wu Q, Peng Z, Zhang Y, Yang J. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res. 2018 Jul 2;46(W1):W438-W442.

Yamanaka S, Johnson ON, Norflus F, Boles DJ, Proia RL. Structure and expression of the mouse beta-hexosaminidase genes, Hexa and Hexb. Genomics. 1994;21(3):588-96.

Zapun A, Contreras-Martel C, Vernet T. Penicillin-binding proteins and beta-lactam resistance. FEMS Microbiol Rev. 2008;32(2):361-85.

Zöller I, Büssow H, Gieselmann V, Eckhardt M. Oligodendrocyte-specific ceramide galactosyltransferase (CGT) expression phenotypically rescues CGT-deficient mice and demonstrates that CGT activity does not limit brain galactosylceramide level. Glia. 2005;52(3):190-8.

Zou X., Yaxiong, and Irwin D. Kuntz Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model J. Am. Chem. Soc. 1999 121, 35, 8033–8043

Appendix

Supplementary materials

Supplementary File 1:Quality check for wtGALT at 310K
Supplementary File 2: Quality check for p.Gln188Arg at 310K
Supplementary File 3:Quality check for wtGALT + ligands at 310K
Supplementary File 4:Quality check for p.Gln188Arg + ligands at 310K
Supplementary File 5:Quality check for wtGALT at 334K
Supplementary File 6:Quality check for p.Gln188Arg at 334K
Supplementary File 7:Quality check for wtGALT + ligands at 334K
Supplementary File 8:Quality check for p.Gln188Arg + ligands at 334K
Supplemenaty File 9:DSSP analysis for all systems of wtGALT p.Gln188Arg at 310K and 334K
Supplemenaty File 10 Radius of gyration analysis for all systems of wtGALT p.Gln188Arg at 310K and 334K
Supplemenaty File 11 SASA analysis for all systems of wtGALT p.Gln188Arg at 310K and 334K
Supplemenaty File 12 Pair distance of ligands analysis for all systems of wtGALT p.Gln188Arg at 310K and 334K
Supplementary File 13 Docking results Simulations with arginine in the active site Position of the selected pose
Supplementary File 14 Docking results Simulations with arginine in the active site – interactions
Supplementary File 15 Docking results Simulations with arginine in the central cavity Position of the selected pose
Supplementary File 16 Docking results Simulations with arginine in the central cavity – interactions
Supplementary File 17 Pair distance between enzyme and arginine, simulations with arginine in the active site
Supplementary File 18 Pair distance between enzyme and substrates, simulations with arginine in the active site
Supplementary File 19 Radius of gyration simulations with arginine in the active site
Supplementary File 20 SASA simulations with arginine in the active site
Supplementary File 21 Pair distance between enzyme and arginine simulations with arginine in the central cavity
Supplementary File 22 Pair distance between enzyme and substrates, simulations with arginine in the central cavity
Supplementary File 23 Radius of gyration simulations with arginine in central cavity
Supplementary File 24 SASA simulations with arginine in the central cavity
Supplementary File 24 Quality check for wtGALT + Arg (active site)
Supplementary File 26 Quality check for wtGALT + G1P + Arg (active site)
Supplementary File 27 Quality check for wtGALT + H2U + Arg (active site)
Supplementary File 28 Quality check for p.Gln188Arg + Arg (active site)
Supplementary File 29 Quality check for p.Gln188Arg + G1P + Arg (active site)
Supplementary File 30 Quality check for p.Gln188Arg + H2U + Arg (active site)
Supplementary File 31 Quality check for wtGALT + Arg (central cavity)
Supplementary File 32 Quality check for wtGALT + ligands + Arg (central cavity)
Supplementary File 33 Quality check for p.Gln188Arg + Arg (central cavity)
Supplementary File 34 Quality check for p.Gln188Arg + ligands + Arg (central cavity)