A Computational Pipeline to Study Donor Substrate Binding to the Wild-Type and Mutants of a Xanthan Gum Glycosyltransferase

Tova Alenfalk
Supervisors (NTNU): Gaston Courtade and Davide Luciano
Supervisor (LTH): Eva Nordberg Karlsson
Examinator: Carl Grey

Master Thesis in Biotechnology
Spring, 2024

Acknowledgments

This master's thesis was done as part of an ERASMUS+ exchange at the Norwegian University of Science and Technology (NTNU) in Trondheim, Norway.

First of all I would like to express my gratitude to Gaston Courtade for giving me the opportunity to write my master's thesis in his group and for all the guidance and support throughout the process. I also want to thank Davide Luciano for all the help and valuable inputs, as well as everyone else in the lab group. Then I would like to thank my supervisor at LTH, Eva Nordberg Karlsson, and my examiner, Carl Grey, for the valuable feedback on my thesis.

Lastly, I want to thank my family and friends for their love and support during my master's thesis and throughout the five years of study leading up to it.

Populärvetenskaplig artikel

I en allt varmare värld där resurserna börjar ta slut är behovet av mer miljövänliga material och bränslen större än någonsin. Kan kolhydrater, ett av naturens mest förekommande organiska material, vara en del av lösningen för ett mer hållbart samhälle?

Kolhydrater består av en eller flera sammansatta sockermolekyler. Det finns en mängd olika varianter av kolhydrater som används inom allt från mat till kläder och medicin (1). Bildandet av kolhydrater möjliggörs av enzymer, specifika proteiner som fungerar som katalysatorer och påskyndar kemiska reaktioner. Dessa verkar exempelvis genom att katalysera sammansättningen av två sockermolekyler under byggandet av en längre sockerkedja.

Om man lyckas modifiera de enzymer som är delaktiga i uppbyggnaden av kolhydrater så att de sätter ihop andra typer av sockermolekyler än vanligt, kan det fungera som ett verktyg för att skapa nya kolhydrater med önskade egenskaper. Detta skulle i sin tur kunna öppna upp dörrarna för ännu fler applikationer av denna biomolekyl.

Beräkningsbiologi använder datorer för att lösa biologiska problem och erbjuder verktyg för att både förstå och designa enzymer. I detta examensarbete har beräkningsbiologiska metoder använts för att studera ett enzym som kallas GumK. GumK är delaktigt i syntetiseringen av den kommersiellt viktiga kolhydraten xantan som används bland annat som förtjocknings- och stabiliseringsmedel i livsmedel. Mer specifikt sätter den på sockermolekylen glukuronsyra under uppbyggnaden av kolhydratkedjan.

Det är inte helt känt varför GumK endast kan använda just glukuronsyra och inte andra liknande sockermolekyler, som exempelvis glukos. I ett försök att svara på den frågan analyserades geometrin av både glukuronsyra och glukos när det var bundet till enzymet. Det observerades en skillnad i hur de två molekylerna var orienterade i relation till GumK. Dessa resultat kan eventuellt tillämpas för att leta efter varianter av GumK som kan använda andra sockermolekyler, såsom glukos, och på så sätt ändra uppsättningen av xantan för att potentiellt utöka dess användningsområden. Ett antal enzymvarianter analyserades, men ingen verkade ha en tydlig ändrad preferens för glukos. För att kunna dra en slutsats kring hur realistiska resultaten är och om metoden verkligen kan användas på detta sätt skulle de behöva jämföras med experimentell data.

Abstract

Polysaccharides are a versatile group of biopolymers widely utilized in a range of industries, from food and medicine to construction. Engineering of the enzymes involved in the synthesis of polysaccharides could provide a way to chemically modify the composition, and therefore be used as a tool to obtain polysaccharides with new desired properties. In this thesis, the substrate binding to the donor domain of GumK, an enzyme involved in the synthesis of the commercially important polysaccharide Xanthan, was studied using molecular docking. Both the native substrate, *uridine diphosphate glucuronic acid* (UDP-GlcA), and the analog *uridine diphosphate glucose* (UDP-Glc) were docked to GumK to gain insight to what causes the enzyme's specificity. When analyzing a larger number of poses, a difference in the distribution was seen, where the native substrate contained more poses with an interaction between the carboxylic group and K307. Generating and analyzing the distribution of the poses could, therefore, potentially be used as a method to screen for mutants with an altered specificity towards UDP-Glc. A number of mutants were tested, but none of them seemed to have an identical UDP-Glc distribution to that obtained from docking UDP-GlcA to the wild type. However, to really draw a conclusion regarding the activity for the mutants tested, and the accuracy of the method used, experimental data is needed.

Abbreviations

Abbreviation	Explanation
CNN	Convolutional neural network
GDP-Man	Guanosine diphosphate mannose
GT	Glycosyltransferase
MD simulations	Molecular dynamics simulations
ML	Machine learning
NMA	Normal mode analysis
RMSD	Root mean square deviation of atomic positions
UDP	Uridine diphosphate
UDP-Glc	Uridine diphosphate glucose
UDP-GlcA	Uridine diphosphate glucuronic acid
WT	Wild-type

Contents

1. Introduction

Polysaccharides are biologically abundant molecules consisting of long chains of monosaccharides connected via glycosidic bonds. They are widely utilized in a range of industries, from food and medicine to construction (1). This versatile group of biopolymers can offer a sustainable alternative to materials from non-renewable sources. Xanthan gum is a commercially important polysaccharide produced by bacteria of the genus *Xanthomonas*, where mainly *X. campestris* is used for industrial production. It consists of repeating cellobiose units, with a side chain composed of D-mannose (β-1,2), D-glucuronic acid (β-1,4), and D-mannose (2). Modifying the chemical composition of Xanthan could potentially be used as a way to change its properties, opening up for new possible applications.

One approach to modify the chemical composition of biomolecules is through enzyme engineering. GumK is one of multiple enzymes involved in the synthesis of Xanthan in *X. campestris* and consists of two domains, an acceptor domain and a donor domain. The donor domain binds uridine diphosphate glucuronic acid (UDP-GlcA) and transfers GlcA to the first mannose on the side chain of Xanthan, leaving UDP as a side product. (3) The composition of Xanthan can potentially be altered by mutating GumK to accept other types of monosaccharides than GlcA.

Computational methods provide powerful tools to investigate the protein-ligand interaction, facilitating the process of engineering enzymes to alter their specificity. These include for instance molecular docking, molecular dynamics simulations, and free energy calculations (4). The different methods have varying physical accuracy and speed. Molecular docking is a time-efficient method that aims to predict the structure of protein-ligand complexes. Although it involves many simplifications and assumptions, it is a popular tool especially when studying a larger library of compounds, and continued efforts are made to improve the accuracy of the results. In this thesis, the donor substrate binding to the wild type (WT) and mutants of GumK has been studied using primarily molecular docking.

1.1 Aim of the study

This thesis aims to study the donor substrate binding of GumK using computational methods, primarily molecular docking. The project is based on the assumption that the donor substrate binds the donor domain independently of the acceptor domain. Therefore, the studies have been performed mainly on the donor domain.

The docking results for both the native substrate, UDP-GlcA, and the analog UDP-Glc was analyzed to gain insight into what causes the substrate specificity towards UDP-GlcA. Furthermore, mutants of GumK were analyzed to find out how the mutations affect the binding, and if any of the mutants seemed to have an altered specificity towards UDP-Glc.

2. Background

2.1 Protein-ligand interaction

The binding of a ligand (L) to a protein (P) to form a protein-ligand complex (PL) can often be described by the reversible reaction:

$$P + L \Leftrightarrow PL \quad (1)$$

The dissociation constant, K_d, for the reaction described by Eq. 1 is defined as:

$$K_d = \frac{[P][L]}{[PL]} \quad (2)$$

where [PL], [P] and [L] and is the concentrations of protein-ligand complex, free protein and free ligand at equilibrium (5). K_d reflects the change in free energy, ΔG, between the free protein and ligand, and the bound complex. During standard conditions (i.e. 1 atm pressure, 298K and 1 M reactants concentrations), the Gibbs free energy of binding is denoted ΔG^0, and can be related to K_d in the following way:

$$\Delta G^0 = RT \ln(K_d) \quad (3)$$

where R is the universal gas constant. A lower K_d means a more negative binding free energy, indicating a stable protein-ligand complex. Therefore, K_d is often used as a measurement for the binding affinity, meaning the "strength" of the interaction between the protein and ligand (4).

The free energy of the binding process is composed of both entropic, ΔS, and enthalpic, ΔH, contributions as described in the following equation:

$$\Delta G = \Delta H - T\Delta S \quad (4)$$

During ligand binding, the change in enthalpy is a result of the loss and formation of interactions between the ligand, protein and solvents. This could for instance include the loss of hydrogen bonds between the ligand or protein and solvent, and the formation of new hydrogen bonds between the protein and ligand (4).

The entropy change is often divided into the three entropic contributions: solvent entropy change, ΔS_{solv}, conformational entropy change, ΔS_{conf}, and translational-rotational, $\Delta S_{r/t}$. First of all, the creation of a protein-ligand complex often increases the entropy of the solvents since the surface area that comes in contact with the solvent decreases, making ΔS_{solv} positive. The ligand and protein, however, will lose rotational/translational freedom when bound in a complex, contributing unfavorably to the entropy change. The conformational entropy can either increase or decrease, depending on the system studied. The total entropy change is the sum of the different contributions (4):

$$\Delta S = \Delta S_{solv} + \Delta S_{conf} + \Delta S_{r/t} \quad (5)$$

2.1.1 Binding models: lock-key, induced fit and conformational selection

There are currently three main models used to theoretically describe the mechanism for the binding of a ligand to a protein: *lock-and-key*, *induced fit,* and *conformational selection model*. The lock-and-key model describes the protein and ligand as rigid bodies whose shapes complement each other like a key in a lock, as illustrated in Figure 1 below. In the induced fit model it is instead proposed that the ligand induces a conformational change in the protein upon binding. This aligns with experimental results showing that a ligand can bind a protein that does not initially have a structure complementing it. The last model, conformational selection, is derived from the theory that the native state of a protein rather is an ensemble of different populated states/conformations. As ligands bind to the protein, the equilibrium is shifted towards the state(s) to which the ligand can bind well (4).

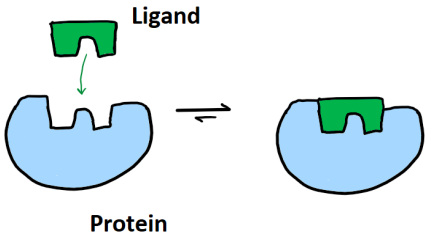

Figure 1: Illustration of the "lock-and-key" model describing the binding of a ligand to a protein.

2.1.2 Methods to study protein-ligand interactions

Multiple methods have been developed to study the interaction of ligands and proteins, both experimental and computational. Experimental techniques include methods to determine the structure of protein-ligand complexes and/or study the dynamics of binding events, such as X-ray crystallography and NMR, and methods to get information on the binding affinity, such as isothermal titration calorimetry, Surface Plasmon Resonance, and Fluorescence Polarization techniques. *In silico* studies of the interaction can be done using for instance molecular docking, molecular dynamics simulations, and free energy calculations (4). Some of these computational methods will be further discussed below.

2.2 Molecular Dynamic simulations

Molecular dynamics simulation is a computational method used to study the behavior and dynamics of a molecular system. A particle-based model of the system is first defined and the positions of the particles are then step-wise updated by integrating Newton's equations of motion, to generate a time dependent trajectory (6).

A so-called force-field with empirical equations and parameters is used to calculate the potential energy of the system. Commonly, the interactions between atoms/particles of the system are described by bonded interactions, such as bond stretching and bending, and nonbonded interactions, usually defined by Lennard-Jones and Coulomb interactions, see Figure 2 (7).

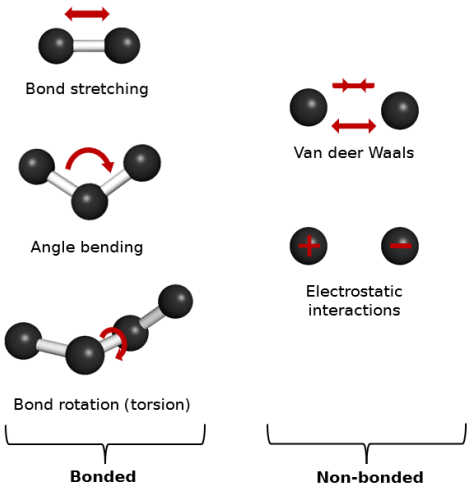

Figure 2: Interactions used to calculate the potential energy of the system during an MD simulation. The interactions are described by bonded and non-bonded interactions.

The simulation often includes multiple steps: preparation of the system, energy minimization, equilibration, and production MD simulation. In the case of simulating a protein in a solution, the simulation box is first defined with the protein placed in it, and then filled with solvent molecules. The energy minimization moves the system to a local minimum, to avoid too large forces on the atoms and structure distortion when running the simulation. Instead of integrating Newton's equation of motion, the positions of the atoms are moved in the direction of decreasing energy using, for example, the steepest descent algorithm. Energy minimization can also be used to refine low-resolution experimental structures and predicted structures (7). Equilibration of the system brings it to a representative equilibrium state for the conditions studied. Both the equilibration and the production MD is typically run in a particular thermodynamic ensemble, for example, NVT and NPT ensembles (6). In an NVT ensemble, the moles (N), volume (V), and temperature (T) are kept constant, while in an NPT ensemble, the pressure (P) is constant instead of the volume. After the equilibration, the system is hopefully at a stable and balanced state and a production simulation can be run to collect desired information about the system.

When simulating a microscopic system using MD simulations, it is more interesting to know how it would behave in a macroscopic solution rather than in vacuum (6). This is done by replicating the simulation box in all directions. All simulation boxes are exact images of each other, meaning that if a molecule leaves the box at one end, it will return to the box at the next end. This way, the simulation resembles that of a bulk system, enabling the prediction of macroscopic properties.

2.3 Molecular Docking

Unlike MD simulations, which provide a time-dependent trajectory that can be used to study the dynamics and calculate macroscopic properties of a system, molecular docking aims to predict the structure of a stable complex between two molecules, a receptor and a ligand. It is a more time efficient method, and is, therefore, commonly used for virtual screenings. The docking algorithm can generally be described in two steps; (1) sampling of different ligand-receptor structures (referred to as poses), and (2) scoring and ranking the poses to find the most likely one (8).

In so-called rigid docking, both molecules are represented as rigid bodies, meaning that the bond lengths and bond angles are set. The only degrees of freedom taken into account are the translational and rotational of the entire molecule, while the conformations are kept the same (9). A more accurate method, however, is to also include single-bond rotations to consider the conformational space of one or both of the molecules. This increases the search space, making it more computationally expensive. Most docking algorithms therefore only consider the ligand (partially or fully) flexible, while the receptor is kept rigid. This is called flexible-ligand docking (10).

2.3.1 Sampling of ligand conformations

It is practically impossible to explore all possible conformations of a ligand, so the search algorithm explores a limited amount of conformations (by rotation and translation) with a given threshold for how identical two conformations are allowed to be (11). Typically, the search space is limited by a user-defined search box to only include the part of the protein that is believed to bind the ligand (12). Three types of search algorithms have been developed for flexible-ligand docking; systematic, random/stochastic, and deterministic. A commonly employed systematic search algorithm is to fragmentize the ligand and individually dock the fragment before covalently linking them together again. Stochastic search algorithms include the Monte Carlo algorithm, where the conformation and/or orientation of the ligand is randomly changed slightly in each step. Molecular dynamics simulations are an example of deterministic search algorithms, however, due to the computational time required for these types of simulations they are rarely used for larger screenings (10).

2.3.2 Scoring functions

The poses produced during the search algorithm are scored using a scoring function. The scoring function should preferably be able to both distinguish binders from non-binders and correctly rank the poses to find the most "real-like" receptor-ligand complex (10). Furthermore, many scoring functions also aim to correctly estimate the binding free energy.

Traditionally, there are three types of scoring functions; force-field based, empirical and knowledge based. Some docking programs combine the results from different scoring functions, which have been shown to sometimes improve the results (10). Scoring functions based on machine learning (ML) have emerged in recent years, and can be considered a fourth category (13).

Force-field based scoring function

Physics, or force-field, based scoring functions estimate the interaction energy between the receptor and ligand by calculating the sum of various energy terms. The non-covalent interaction between the receptor and ligand atoms is usually described as the sum of the Van der Waals and electrostatic interactions represented as Lennard-Jones and coulomb potentials. Some scoring functions also include other terms to account for, for example, hydrogen bonding and solvation energy. A force-field based scoring function can therefore often be described by the following equation:

$$\Delta G_{binding} = \Delta E_{WdV} + \Delta E_{el} + \Delta E_{H\text{-}bond} + \Delta G_{sol} \quad (6)$$

The parameters for the energy terms are either determined based on experimental observations, or by *ab initio* quantum mechanical calculations (13). A major challenge with using force field based scoring functions in molecular docking is to estimate the solvation and the entropy terms in Eq. 5, and these are therefore often oversimplified or ignored (14).

Empirical scoring functions

Empirical, or regression-based, scoring functions predict the binding affinity by summing up a set of weighted scoring terms. Examples of scoring terms often included are Van der Waals interactions and hydrogen bonding. The coefficients, or weights, of these terms are determined using linear regression analysis on experimental binding affinity data. An empirical scoring function may look as the following, where *w* is the weight coefficient (13,14):

$$\Delta G_{binding} = w_0 + w_1 \Delta E_{WdV} + w_2 \Delta E_{el} + w_3 \Delta E_{H\text{-}bond} + w_4 \Delta G_{entropy} \quad (7)$$

Knowledge based scoring functions

Knowledge based scoring functions use atom pairwise statistical potentials derived from structurally determined protein-ligand complexes. The score is the sum of these pairwise potentials (13,14).

Machine learning based scoring functions

Machine learning (ML) refers to algorithms that learn how to perform a task without being explicitly programmed on how to do so. ML based scoring functions are trained on structure data labeled with experimentally determined binding affinities to learn patterns from the dataset. There are many different types of machine learning algorithms that have been employed for the task (13,14). Section 2.5.1 will introduce a docking program that enables the use of ML scoring functions based on convolutional neural networks (CNNs), a type of ML algorithm architecture.

2.3.3 Flexible docking

Although flexible-ligand docking is much more accurate than rigid docking, it still makes a major simplification of reality, that being that the receptor is fully rigid (15). If the substrate binds according to the induced fit or conformational selection model, see section 2.1.1, treating the protein as rigid does not give a representative view of the binding process. Including the flexibility of the receptor without greatly increasing the computationally cost, however, remains a challenge in molecular docking. Various methods have been developed to somewhat account for the receptor flexibility. This includes for instance ensemble docking and side chain flexibility. In ensemble docking, an ensemble of different rigid protein conformations are docked to instead of only one. Sometimes multiple crystal structures of the protein exist to make up the protein ensemble. Otherwise, the set of protein conformers are often generated through MD simulations, or using methods based on normal mode analysis. Docking with side chain flexibility instead means that different rotamers of the residue side chains are used, while the backbone is unchanged (15).

2.4 The enzyme studied

The enzyme studied in this thesis is GumK from *Xanthomonas campestris*. It is one of multiple glycosyltransferases involved in the synthesis of Xanthan. This section will give some background

information on xanthan gum and glycosyltransferases in general, as well as a review of what is currently known about the structure and function of GumK.

2.4.1 Glycosyltransferases

Glycosyltransferases (GT) are a group of enzymes that transfer the sugar moiety from a donor sugar substrate to another biomolecule by catalyzing the formation of a glycosidic bond. The nucleophile of the acceptor substrate is typically an oxygen, forming an O-glycosidic bond, but can also be a nitrogen, sulfur, or carbon. The donor sugar substrate contains a phosphate leaving group, most commonly a nucleoside diphosphate sugar such as UDP or GDP (16). In general, glycosyltransferases exhibit high specificity towards the donor and acceptor substrates used, although there exist examples of GTs with a broader specificity (17).

The enzymes are categorized into families based on the sequence. However, although there exist more than 100 glycosyltransferase families, almost all GTs have a structure that can be grouped into one of four different folds; GT-A, GT-B, GT-C, and lysozyme-type folds. Enzymes belonging to either GT-A or GT-B catalyze reactions with nucleotide sugars as the donor substrate. They both contain Rossmann-like folds, a tertiary fold commonly observed in proteins characterized by segments of alternating alpha helices and beta strands (18). The difference is that GT-A consists of one domain and binds a metal ion, while GT-B has two domains and are generally metal-ion independent. GT-C and lysozyme-type folds instead utilize lipid-linked sugar donors (17).

Glycosyltransferases are also classified into two groups depending on the stereochemistry of the product. If it is the same as for the donor substrate, it is a retaining GT. Inverting GTs, on the other hand, inverts the stereochemistry at the anomeric carbon atom (17). The reaction mechanism for these two types differs. For inverting GTs, the reaction follows an SN2-like reaction, with one of the active site residues serving as a base to deprotonate the nucleophile of the acceptor substrate. Figure 3 shows a schematic representation of the reaction mechanism for inverting GTs. The reaction mechanism for retaining GTs has instead been proposed to contain a covalently bound glycosyl-enzyme intermediate, following a double-displacement mechanism (16).

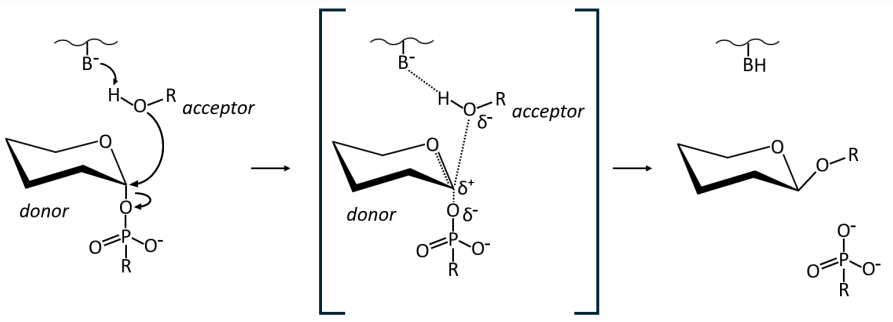

Figure 3: A schematic representation of the reaction mechanism for inverting glycosyltransferases. One of the active site residues, here denoted B^-, serves as a base and deprotonates the nucleophile of the acceptor substrate. The reaction then follows an SN2-like reaction.

2.4.2 Xanthan gum and the biosynthetic pathway of xanthan

Xanthan gum is an extracellular polysaccharide produced through fermentation by bacteria of the genus *Xanthomonas*, where mainly *X. campestris* is used for industrial production. Xanthan gum is hydrophilic and adds viscosity to liquids. Therefore, it is often used as a thickener and stabilizer (2). Other important properties of xanthan gum include a high thermostability and stability across a broad range of pH values. It has a wide range of applications in various industries, as a food additive, emulsion stabilizer in cosmetic products, biomedical applications, and more (19).

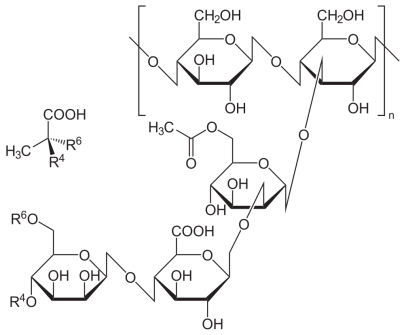

Figure 4: The chemical structure of the repeating unit that builds up Xanthan gum, retrieved from Wikipedia (20).

The chemical structure of xanthan consists of repeating cellobiose units, with a side chain composed of D-mannose (β-1,2), D-glucuronic acid (β-1,4), and D-mannose (2), see Figure 4. During the biosynthetic pathway, the sugars of one repeating unit are sequentially added by different GTs to a polyprenol phosphate carrier. Thereafter, acetylation and pyruvylation of the mannose residues occur to varying degrees, before the polymerization of the pentasaccharide subunits and secretion of the polymer happen (21). The polysaccharide chain arranges itself in a helical shape to form fibers, contributing to the stability of the structure (19).

2.4.3 Structure and function of GumK

GumK is a membrane-associated inverting glycosyltransferase responsible for transferring GlcA from UDP-GlcA to the first mannose on the side chain of Xanthan during the synthesis of the pentasaccharide repeating unit of the polymer, see Figure 5. The structure of GumK has been resolved by x-ray crystallography (PDB id: 2HY7). It has a fold typical to that of GT-Bs, with two domains and a catalytic cleft between them. The N-domain, consisting of 10 alpha helices and 8 beta-sheets, binds the acceptor substrate, while the C-domain, consisting of six alpha helices and six beta sheets, binds the donor substrate (UDP-GlcA). Both domains have a Rossmann-like fold and are connected via an interdomain linker of 7 residues. For some GT-Bs, the catalytic activity has been linked to interdomain motions that bring the domains together (3). This has also been suggested for GumK, where the binding of UDP-GlcA seems to cause an interdomain twisting motion, resulting in a more closed conformation that enables the reaction (22). The proposed reaction mechanism follows the one shown in Figure 3. An aspartic acid on the acceptor domain, D157 has been identified as the amino acid serving as the base catalyst that deprotonates the hydroxyl group on the C2 atom of the mannose on the acceptor substrate (3).

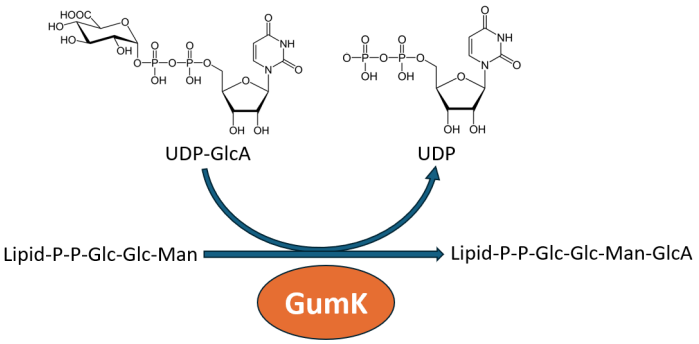

Figure 5: The reaction catalyzed by GumK. GumK binds uridine diphosphate glucuronic acid (UDP-GlcA) and transfers GlcA to the trisaccharide glucose-glucose-mannose (Glc-Glc-Man) connected to a polyprenol phosphate carrier (lipid-P-P), during the build-up of the pentasaccharide repeating unit of Xanthan. The reaction forms the product lipid-P-P-Glc-Glc-Man-GlcA, leaving UDP as a side product.

A crystal structure of the UDP-bound GumK is also available (PDB id: 2Q6V), see Figure 6. The binding pocket for the donor substrate is situated by the two helices named Cα3 and Cα4, and residues important for the binding of UDP include M231, E272, M273, Y292, M306, K307 and Q301. Mutations of M231, E272, Y292 and K307 have been shown to have a negative effect on the protein activity, although none of the mutations fully inactivated the enzyme (3).

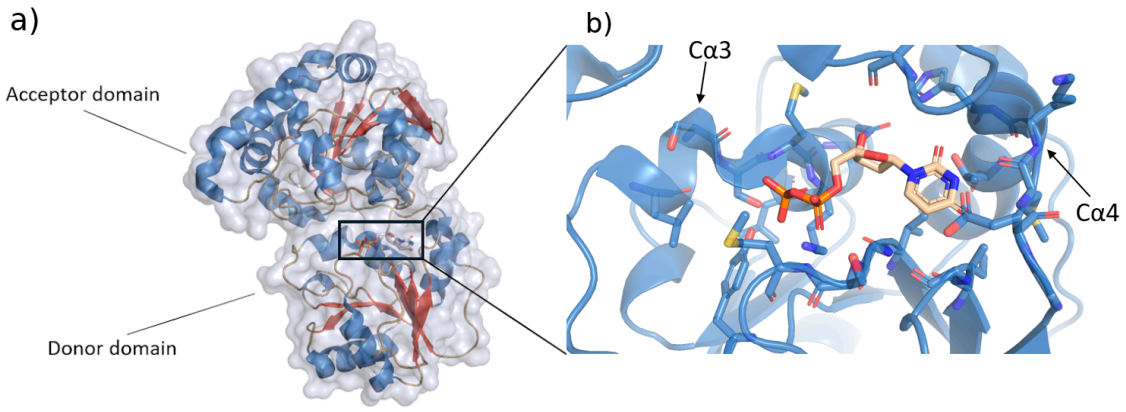

Figure 6: Crystal structure of GumK bound to UDP (PDB entry: 2Q6V) visualized using PyMol. a) shows the entire protein, and b) provides a more detailed picture of the binding site, where all residues within 5Å of the bound UDP are represented as sticks.

GumK has been shown to be able to hydrolyze UDP-GlcA even in the absence of the acceptor substrate. As a result, no crystal structure with UDP-GlcA bound to GumK exists (3). The D157A mutant of GumK cannot hydrolyze UDP-GlcA. Despite this, the sugar moiety of the donor substrate remained unresolved in co-crystallization attempts with the D157A mutant and UDP-GlcA. The UDP, however, was still resolved and positioned exactly as in the WT. The dynamics of the sugar part when the acceptor substrate is not bound was mentioned as a possible reason for not being able to co-crystallize the enzyme with the whole donor substrate in this case (3,22).

2.5 Programs used in the project

2.5.1 Docking program: GNINA

The molecular docking program used in this project is called GNINA. It is a fork of SMINA which is a fork of the popular docking software AutoDockVina (called Vina) (23). SMINA was developed by modifying the source code of Vina to support custom scoring functions (24). GNINA was then developed to further support scoring functions based on convolutional neural networks (CNN), a type of machine learning model (23).

Scoring functions in GNINA

GNINA has multiple built in scoring functions, both empirical and CNN based, but also provides the option of manually defining a scoring function. Different empirical scoring functions, as defined in section 2.3.2, are available in GNINA and include Vina and Vinardo scoring. For both of these, the score is given in units of a binding affinity (kcal/mol) although it should be taken as a docking score rather than an actual binding affinity. Furthermore, multiple CNN based scoring functions exist, with either varying architectures and/or trained on different data sets. The default CNN scoring function is trained to both predict a pose score that correlates to how probable it is that the pose has a low RMSD compared to the real structure, as well as the binding affinity in pK units. The pose score is given as a score between 0-1, where 1 means that the CNN model is very confident that it is a good pose and 0 is not confident at all (23).

GNINA docking pipeline

Aside from a 3D structure file of the ligand and protein, the user also needs to specify a so-called autobox before running a molecular docking using GNINA. The autobox defines the part of the protein where the ligand should bind. It can be provided either as the dimension of a box or a structure file of a molecule. If a structure file is provided, GNINA will create a box around the molecule and then add an additional 4Å in each direction. The option *autobox_extend* can be used to expand the autobox if it is smaller than the provided ligand (23).

The search algorithm used in GNINA is the Monte Carlo algorithm, which starts with a random configuration of the ligand, unbiased towards the 3D structure of the provided ligand file. In each step of the algorithm, the ligand is first randomly modified and then energy minimized. The modification is done either by translating or rotating the molecule, or by changing the torsional angles (23). Ring structures and bond lengths, however, are not changed. The new ligand structure is scored according to the chosen scoring function, and is only kept if the score passes the Metropolis acceptance criterion (23).

The number of Monte Carlo samplings that are run is determined by the setting *exhaustiveness*. After each sampling, the top scored ligand poses are saved and refined. During the refinement the ligand conformation is altered by following the scoring function gradients, to reach a local energy minima. The refined ligand poses are scored again and filtered to exclude poses that have a pairwise RMSD lower than *min_rmsd_filter*, and the top scored poses are provided to the user. The number of poses output by GNINA is given by *num_modes*, with the default being the 9 top poses (23).

The scoring functions used during the Monte Carlo sampling algorithm, for refinement and in the last scoring step do not necessarily need to be the same, and can be determined by the user. After testing different possibilities of scoring functions the authors set, with regards to both docking performance and run time, the default scoring to only use CNN scoring in the last ranking step and Vina scoring for the first steps. An illustration of the docking pipeline used in GNINA can be seen in Figure 7 below.

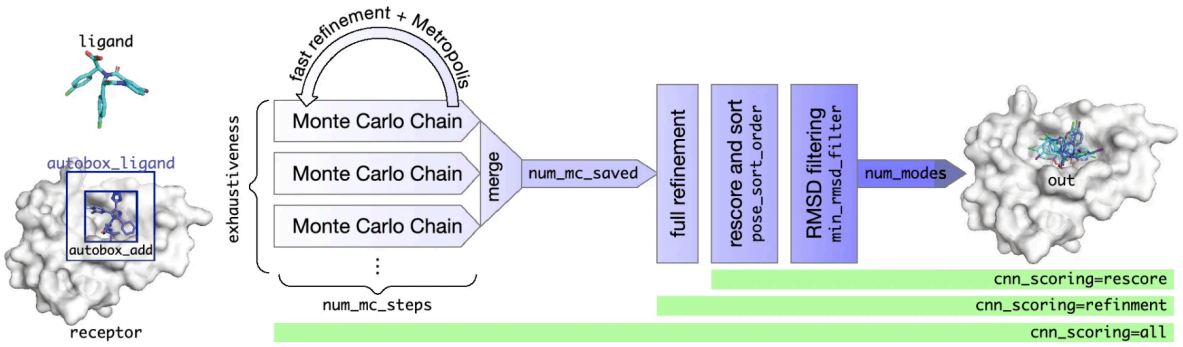

Figure 7: An overview of the docking process of GNINA with some of the different commandline parameters available, from McNutt et al., 2021 (23). The user provides coordinate files for the ligand and receptor, and the *autobox_ligand* defines within what region of the protein the ligand should be docked to. Sampling of different ligand conformations is done using the Monte Carlo algorithm, where the number of Monte Carlo chains is given by *exhaustiveness*, and the number of steps in each chain is given by *num_mc_steps*. After each sampling, the *num_mc_saved* number of top-scored ligand poses is saved and refined by following the scoring function gradients to reach a local energy minimum. The saved poses are then rescored and filtered based on their pairwise RMSD to exclude poses with a pairwise RMSD less than *min_rmsd_filter* Å. The top *num_modes* poses are given to the user. The *cnn_scoring* parameter determines to what extent CNN scoring is used as visualized by the green rectangles.

Flexible docking

GNINA has the option to make side chains of the receptor protein flexible during docking. In this case, the torsion angles of the selected side chains are also sampled. The backbone, however, is still kept rigid (23).

2.5.2 ProDy

ProDy is a Python package developed by the University of Pittsburgh that provides various methods for structure-based analysis of protein dynamics. Among others, they enable normal mode analysis (NMA) of proteins (25), a method used to get information about the collective motions of the structure. In NMA, the protein is often simplified to a network of alpha carbons connected with springs (7). From the analysis, so-called normal modes are obtained describing the collective motions of the protein. Each normal mode is associated with vectors, representing the direction and magnitude of the movement, and the frequency of the motion in an arbitrary unit of frequency. A higher frequency generally means a faster and smaller vibrational motion, whereas the normal modes with a lower frequency instead describe slower, large-scale motions (26). The slowest motions are typically assumed to be those with functional relevance (7). Aside from performing NMA, ProDy also offers a VMD plugin named Normal Mode Wizard (NMWiz) that enables visualization of normal mode data (25), and various other methods to study protein dynamics.

A module implemented within ProDy employed in the project is ClustENMD, which enables sampling of protein conformations to be used in, for example, ensemble docking. The ClustENMD algorithm iterates through three steps; conformer generation, clustering and MD simulation, for a determined number of cycles. The conformers are generated using anisotropic network model, a tool for NMA of proteins. The generated protein conformations are then clustered and the representative conformations from each cluster is relaxed with a short MD simulation (27). The workflow for ClustENMD is visualized in Figure 8.

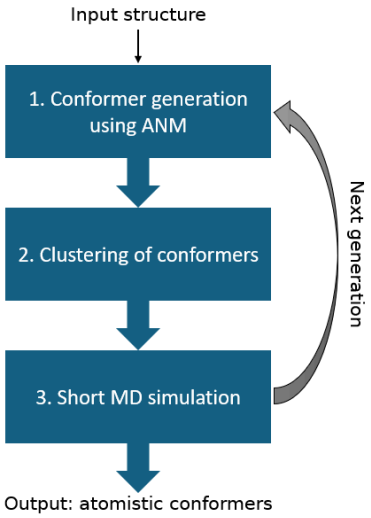

Figure 8: The workflow of ClustENMD, a module implemented within the Python package ProDy used to sample protein conformers using NMA (27).

3. Materials and methods

This section is described in two steps. First the performance of the docking program was evaluated by docking to the crystal structure as described in 3.1. Different docking settings were tested and the ability of the docking program to distinguish between an active substrate and inactive substrate was assessed. Thereafter, ensembles of ligand conformations were generated to study the binding of the native substrate, UDP-GlcA, and the analog UDP-Glc to both the WT GumK donor domain and a number of mutants, as described in 3.2. An overview of the workflow is depicted in Figure 9. The left part of Figure 9 shows the pipeline used to study the donor substrate binding as described in 3.2.

The scripts used to generate and analyze the poses can be found here: https://github.com/tovaalen/substrate_binding_GumK. All molecular dockings were performed using the GNINA (23). Preparation of the ligand and receptor, and analysis of the poses was done using RDkit, an open-source toolkit for cheminformatics (29); obabel, an open-source chemical toolkit (30); and MDanalysis, a python library to analyze trajectories from MD simulations (35,36).

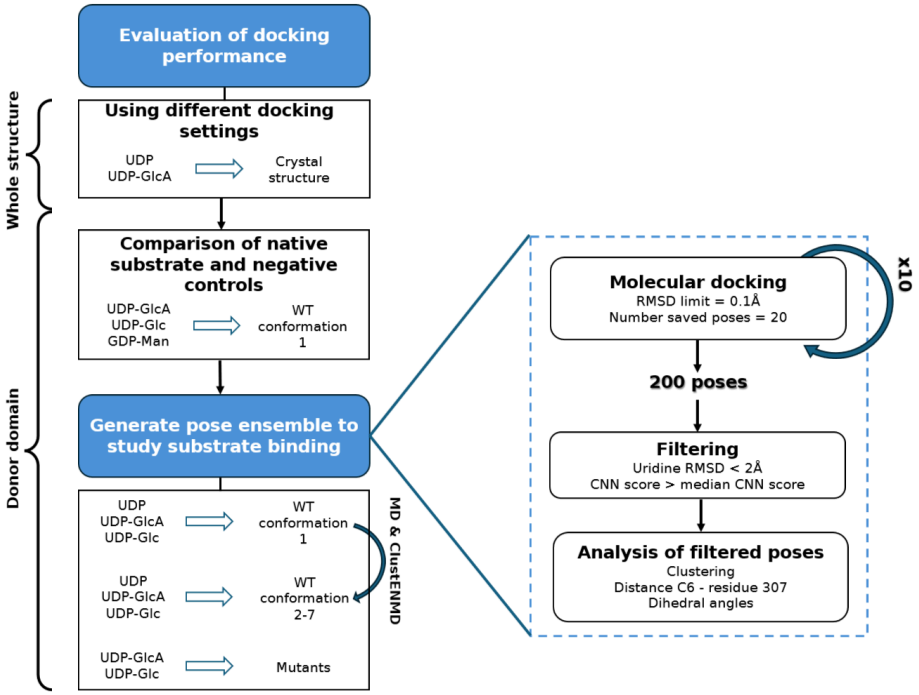

Figure 9: The left part of the figure shows a flowchart over the workflow in this thesis, with the different steps performed and what ligands were docked to which receptor in each step. The right part of the figure shows the pipeline used to study the substrate binding to WT and mutants of GumK.

3.1 Protein and ligand structures

The UDP bound GumK crystal structure was downloaded from the RCSB Protein Data Bank (id: 2Q6V). Using PyMol (28), the water molecules were removed from the structure and the ligand was extracted and saved in a separate pdb file. A conformation of the donor domain of the WT was provided by PhD student Davide Luciano (generated through MD simulations), and will in this thesis be denoted as conformation 1.

Four different ligands were docked throughout the project: UDP, UDP-GlcA, UDP-Glc and guanosine diphosphate mannose (GDP-Man). Docking of UDP was made using the pdb file with the extracted ligand from the crystal structure. A 3D conformer of UDP-GlcA and UDP-Glc was downloaded from PubChem as SDF files. As no 3D conformer was found for GDP-Mannose, it was instead generated from the SMILES code (retrieved from PubChem) using RDkit (29). The SDF files for the ligands were then converted to pdb files through PyMol (28).

3.2 Preparation of the receptor and ligands prior to docking

Independent of the receptor and ligand used for the docking, they were always prepared in the following way. The protein structure was aligned to the crystal structure using PyMol's *align* command (28). Both the receptor and ligand were protonated at pH 7 using the obabel (30), with the -p flag.

3.3 Evaluation of docking performance by docking to crystal structure

Docking of UDP and UDP-GlcA to the UDP-bound GumK crystal structure was made using different docking settings. Specifically, the two settings exhaustiveness and autobox were varied, meaning the number of Monte Carlo chains run during the sampling step and the area within which the ligand is docked, see Figure 7. Four different exhaustiveness levels were tested: 4, 8, 16, and 32. For each exhaustiveness level, two different autoboxes were used. First, the pdb file for the extracted UDP from the crystal structure was used to define the autobox. Thereafter, a larger autobox was manually picked to cover the whole binding site using MGLTools, see Figure A3. Independent of which autobox was used, the option *autobox_extend* was always on. This option ensures that the autobox would be expanded if the ligand is bigger than the box, so that it can always rotate freely. Except for the settings described, the default settings for GNINA docking were used. This includes for example that the top 9 poses are provided to the user after docking (*num_modes* = 9), which are filtered to have a maximum pairwise RMSD of 1Å (*min_rmsd_filter* = 1).

The docking was rerun 10 times for each ligand and setting. The default scoring was used (e.i., Vina scoring for the sampling and refinement steps and CNN scoring for ranking of the poses). However, in addition to this, the final poses were also scored with Vina and Vinardo. The heavy-atom RMSD of the docked UDP compared to the crystal structure UDP was calculated using obabel's *obrms* option (30). Similarly, the heavy-atom RMSD for the uridine part of the docked UDP-GlcA relative to the uridine part of the crystal structure UDP was calculated.

The residues on the donor domain within 5Å of the highest scored UDP-GlcA pose (in terms of CNN score) obtained from the redocking were selected and used as autobox for all future docking.

3.3.2 Docking of UDP-GlcA, UDP-Glc and UDP-Man to crystal structure

To evaluate if the scoring functions used can distinguish between an active and inactive substrate for the enzyme studied, the three compounds UDP-GlcA, UDP-Glc and UDP-Man, see Figure 10, were docked to

the crystal structure. The exhaustiveness was set to 16, and the autobox used was given by the residues within 5Å of the highest scored pose as previously described with the *autobox_extend* option on. The docking of each ligand was rerun 10 times, and all poses were scored using both CNN scoring (default), Vina and Vinardo.

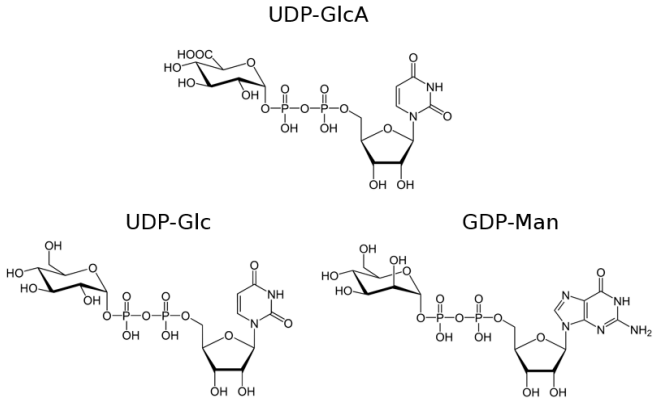

Figure 10: The chemical structure of UDP-GlcA, UDP-Glc, and GDP-Man. Retrieved from Wikipedia (31–33).

3.4 Study of donor substrate binding to GumK WT and mutants

To further study the donor substrate binding to the GumK donor domain, an ensemble of poses was generated. In order to do this, the number of saved poses after each docking, *num_modes*, was increased to 20 (default 9). Furthermore, the limit for how similar two poses are allowed to be was lowered to an RMSD of 0.1 Å (default is 1), using the *min_rmsd_filter* option. The exhaustiveness was set to 16, and the autobox used was given by the residues within 5Å of the highest scored pose from section 3.3 with the *autobox_extend* option on. The final poses were scored using both CNN scoring, Vina and Vinardo. For each receptor/ligand, the docking was rerun 10 times, resulting in a total of 200 saved poses. This was done both for the WT and a number of mutants. Docking to the WT donor domain included docking to conformation 1, retrieved as explained in section 3.1, as well as docking to an ensemble of rigid protein conformations generated based on conformation 1. A more detailed description of these different docking scenarios is found in section 3.4.1-3.4.2.

Prior to analysis of the poses, they were filtered based on two criterias: how high their score was and how well-docked the uridine part of the substrate was. Only poses with a CNN score higher than the median CNN score for the 200 generated poses and a heavy atom RMSD lower than 2Å when comparing the uridine part to that of a reference UDP were used for analysis. For the WT conformation 1 and the mutants, this reference UDP was given by the crystal structure UDP. For the other conformations of the WT donor domain, a new reference UDP was selected as described in 3.4.2.

3.4.1 WT conformation 1

Docking to conformation 1 of the WT donor domain was done for both UDP-GlcA and UDP-Glc.

3.4.2 Ensemble docking (WT)

Different conformations of the donor domain were generated by both MD simulations and ClustENMD. From these, a few were selected to which UDP, UDP-GlcA and UDP-Glc were docked to.

MD simulation

The starting structure for the simulation (conformation 1), was first prepared in PyMol by shortening the chain to contain residues S217-A352. This was done because the ends of the protein chain are likely very flexible. Excluding them prevents the protein from coming too close to the simulation box boundaries. Furthermore, it avoids interference of the termini with other parts of the protein. The end termini were then capped with an acetyl group at the N-termini (ACE) and a methanamine at the C-termini (NME) to avoid any artificial strong interactions. The system was set up in GROMACS by placing the protein in the center of a dodecahedron box with a distance of 1.2Å to the box boundary using the *editconf* tool, and filling the box with water molecules using GROMACs *solvate*. Potassium and chloride ions were added to the system using the command *genion* to neutralize the overall charge.

The prepared system was first energy minimized using a steepest descent integrator with a maximum of 5000 minimization steps, to resolve any steric clashes and unfavorable geometry in the system, with the *grompp* and *mdrun* commands. Thereafter an 100 ps NVT and a NPT equilibration was performed, before the 1000 ns production MD.

The distance between L232 and L301, and H275 and Q310 were measured for all frames of the trajectory in VMD (34) in order to, approximately, describe the distance between the Cα3 and Cα4 helices. The trajectory was grouped into time periods during which the distances were somewhat constant. A representative protein conformation for each of these time periods was given by the one with the lowest RMSD compared to the average structure (as calculated in VMD (34)). From these representative conformation, a few were selected for the ensemble docking.

ProDy

To obtain more different conformations, especially ones with a more open binding pocket, conformational sampling using ProDy's ClustENMD module was also made. The starting structure of the donor domain, conformation 1, was again shortened in PyMol to contain residue G222 - A352. This was done to exclude the flexible ends, so that the conformer sampling focuses on the dynamics of the rest of the protein instead. Data for the normal modes were first calculated using ProDy's Anisotropic Network Model (ANM) to gain insight of the major motions of the donor domain. An ensemble of protein conformation was then sampled using the 2 slowest normal modes.

Reference UDP

Since the structures in the protein ensemble have different conformations, evaluating how well the UDP part is docked by comparing to the crystal structure would likely give faulty results. Therefore, a new reference UDP was selected for each conformation from the UDP poses obtained when docking to the same conformation. This was done by aligning the uridine part of the UDP pose to the crystal structure UDP, and then calculating the RMSD of the protein using MDanalysis (35,36). The UDP pose giving the lowest protein RMSD was used as a new reference when calculating the uridine RMSD for the UDP, UDP-GlcA and UDP-Glc poses.

3.4.3 Mutants

UDP-GlcA and UDP-Glc were docked to the donor domain of six GumK mutants to study how the mutations affect the binding of these substrates. The mutants were provided by my supervisors and are listed in Table 1. All mutants with only one point mutation were mutated using the *Mutagenesis Wizard* in PyMol (28). ChimeraX (37) was used to alter the side chain rotamer to the most likely configuration that didn't interfere with the binding site and did not give any steric clashes with other side chains in the protein. Mutants with more than one point mutation were instead folded using AlphaFold2 (38). An visual inspection of the binding site for the AlphaFold generated mutants was conducted, and the rotamer of any side chains that pointed into the binding site was changed in ChimeraX (37). All new structures were energy minimized using GROMACS following the same procedure described under *MD simulations* in section 3.4.2. Mutant 1 was docked twice, once with the protein fully rigid and once with one side chain (R231) flexible.

Table 1: The mutations for the six mutations analyzed.

Mutant 1	S230A M231R L232G L301A M306G K307T
Mutant 2	K307A
Mutant 3	M306A
Mutant 4	L301A
Mutant 5	S305A
Mutant 6	K307A M231R

3.4.4 Analysis of the poses

After filtering the poses generated from one docking round, they were analyzed in different ways. First, the pairwise heavy-atom RMSD for all poses was calculated with obabel (30). Based on this RMSD matrix, the poses were clustered by utilizing the python package *scipy* (39). The clustering method employed was hierarchical clustering, where the distance between two clusters where calculated using the "average" approach, as described in Eq. 8:

$$d(u,v) = \sum_{ij} \frac{d(u[i], v[j])}{(|u| * |v|)} \tag{8}$$

where d is the distance between cluster u and v, and $|u|$ and $|v|$ is the number of poses in that cluster. A threshold distance of 7 was used to form the clusters. For the biggest clusters, the pose with the lowest pairwise RMSD to the other poses in the cluster was selected as a representative pose and visualized in PyMol (28).

A number of attributes of the poses were calculated in an attempt to describe the conformation of the poses with these. First, the distance between carbon 6 in the sugar ring and the nitrogen in the side chain of

residue K307 was calculated. The coordinates for the atoms were extracted using RDkit (29). For mutants with a mutated residue 307, i.e. mutant 1,2 and 6, a different atom in the side chain was used when calculating the distance. The oxygen of the threonine side chain was used for mutant 1, and the carbon of the alanine side chain was used for mutant 2 and 6. Furthermore, two dihedral angles denoted α and β were calculated using MDanalysis (35,36). α was defined as the dihedral angle given by the four atoms O^{β}-P^{β}-O^{1}-C^{1} and β was given by P^{β}-O^{1}-C^{1}-O^{5}, see Figure 11. To analyze if the dihedral angles were a good description of the orientation of the sugar molecule, the normal vector given by the sugar ring plane, and the projection of this vector on the x, y and z-axis, was calculated for the UDP-GlcA docked to the WT conformation 1. As this should uniformly represent a single orientation of the ring, comparing it to the dihedral angles can give an indication of the usefulness of the angles.

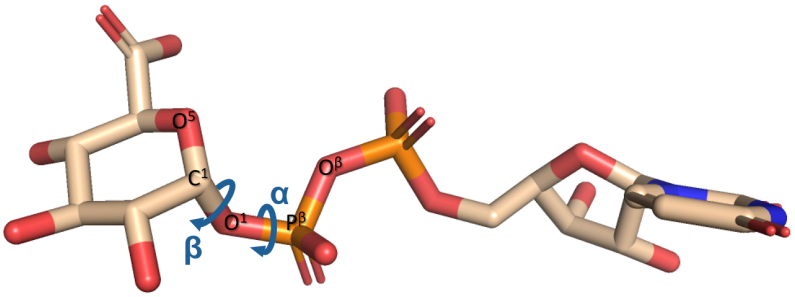

Figure 11: The two dihedral angles calculated for the docked poses, denoted α and β. α is the dihedral angle given by the four atoms O^{β}-P^{β}-O^{1}-C^{1} and β is given by P^{β}-O^{1}-C^{1}-O^{5}.

4. Results

4.1 Evaluation of docking performance and docking settings

Redocking of UDP and docking of UDP-GlcA to the crystal structure of GumK was done to evaluate the performance of the docking program using different settings. A list with the median and average scoring values and RMSD for the different docking scenarios can be found in the appendix, see Table A1-A4.

When utilizing the crystal structure UDP as an autobox, the top ranked pose for both UDP and UDP-GlcA almost always had an RMSD lower than 2Å compared to crystal structure UDP. These poses can be considered well-docked. A slight increase in the scoring could be seen at higher exhaustiveness levels in some cases. For example, for the median CNN score when docking UDP (0.49 at exhaustiveness 4 and 0.67 at exhaustiveness 32), see Table A1. Overall, however, the four exhaustiveness levels tested (4, 8, 16, and 32) gave similar scoring results.

When increasing the volume of the autobox to cover a larger portion of the protein, the exhaustiveness had a bigger impact on the performance of the docking, see Table A1-A4. The median Vina score increased from -5.4 kcal/mol at exhaustiveness 4 to -8.1 kcal/mol at exhaustiveness 32 for the docked UDP, and from -7.5 to -10.2 for UDP-GlcA. Although the top pose for the docked UDP still almost always had an RMSD < 2Å independent of the exhaustiveness, this was not the case for the docked UDP-GlcA. The average RMSD at exhaustiveness 4 was 10±10 Å, compared to 1.1±0.3 Å for exhaustiveness 32.

Figure 12 shows the average percentage of the top 3 poses with an RMSD less than 2Å, when docking with the different settings. It reflects what was described above, that the top poses are generally well-docked for the smaller autobox without being that affected by the exhaustiveness. When using a bigger autobox, however, it is more likely that the top poses are well-docked when increasing the exhaustiveness for both the docked UDP and UDP-GlcA.

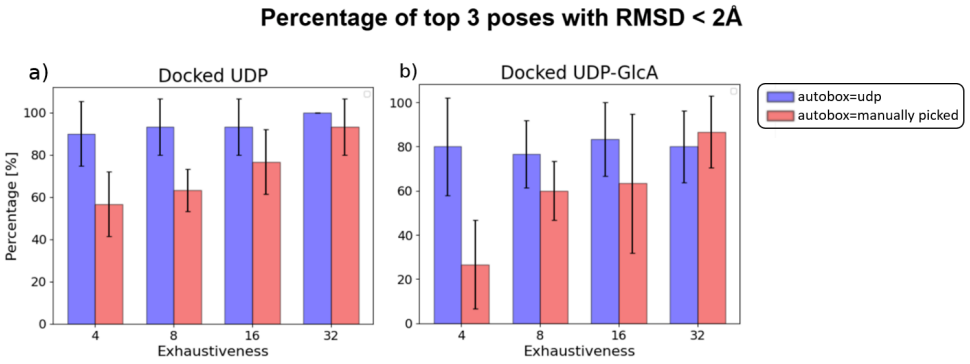

Figure 12: The average percentage of top 3 poses with an RMSD < 2Å from docking UDP and UDP-GlcA to the crystal structure using different exhaustiveness levels and autoboxes. Four different exhaustiveness levels 4, 8, 16 and 32, and two different autoboxes was tested. The autobox "udp" corresponds to the crystal structure UDP and the autobox "manually picked" was manually picked using MGLTools to cover the whole binding site. The docking was rerun 10 times for each exhaustiveness level and autobox size, and the error bar shows the standard deviation. a) shows the results for the docked UDP. The RMSD was calculated as the heavy-atom RMSD compared to the crystal structure UDP. b) shows the results for the docked UDP-GlcA, for which the RMSD represents the heavy-atom RMSD compared to the uridine part of the crystal structure.

To see how the different scoring functions manage to predict the quality of a docked pose, the score vs the RMSD have been plotted for all poses. Figure 13 shows these plots for the CNN scoring and Vina scoring. Both scoring functions have a rather large variation for how high scored a pose with a low RMSD is. For example, the CNN score for the docked UDP poses with an RMSD < 2Å ranges from 0.37 to 0.94, and the Vina score from -6.6 to -11.0 kcal/mol.

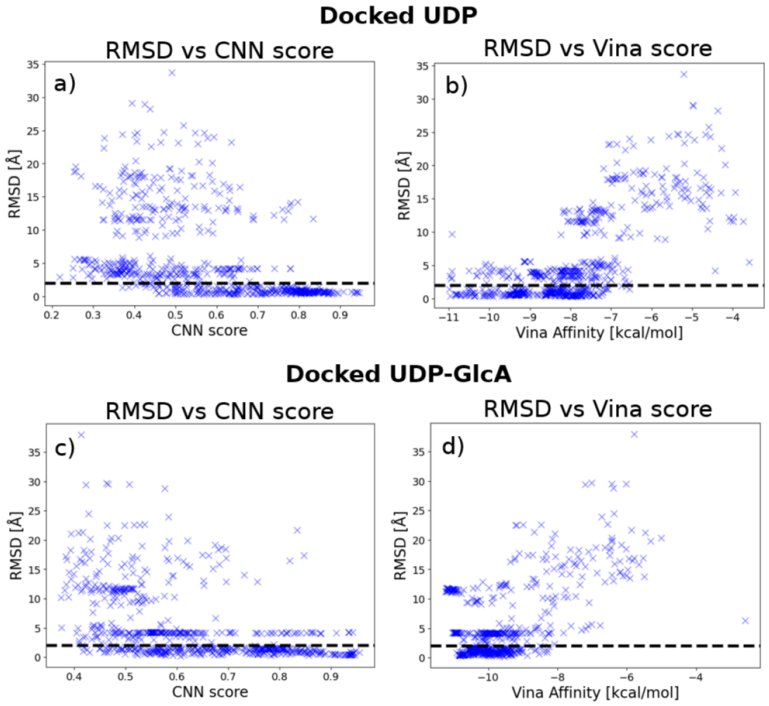

Figure 13: The score of the poses generated from docking UDP and UDP-GlcA to the crystal structure GumK plotted against the RMSD. The RMSD for the docked UDP was calculated as the heavy-atom RMSD compared to the crystal structure UDP, whereas the RMSD for the docked UDP-GlcA was calculated as the heavy-atom RMSD compared to the uridine part of the crystal structure. The dotted line shows where the RMSD is equal to 2Å.

Plots of the Vinardo score and CNN affinity vs RMSD can be found in the Appendix (Figures A1 and A2). However, also these scoring functions have a broad variation in how high a pose with a low RMSD is scored, and multiple poses with a high RMSD are ranked high. Therefore, they do not seem to give a better result than the CNN and Vina scoring.

The UDP-GlcA pose with the highest CNN score (0.96) had an RMSD of 0.85Å and is shown in Figure 14. The residues on the donor domain within 5Å of this pose were picked as the autobox for all future docking. These are: V228, G229, S230, M231, I253, G271, E272, M273, K274, H275, T278, Y292, L301, S304, S305, M306, K307, L308, Q310. A visualization of this autobox compared to the two previously used is shown in Figure A3. Furthermore, the exhaustiveness for the future dockings was always set to 16.

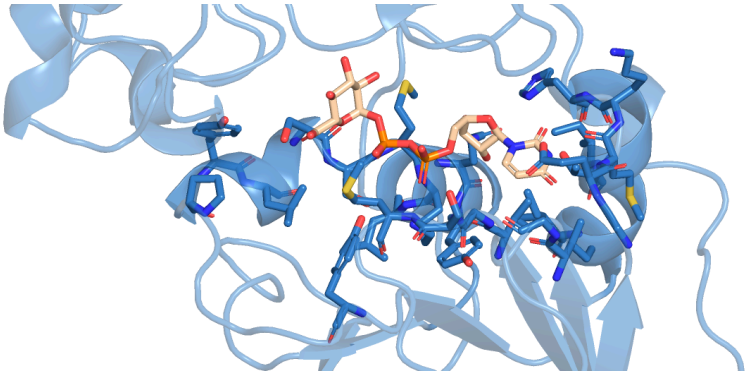

Figure 14: The docked UDP-GlcA with the highest CNN scoring (0.96) obtained from docking to the crystal structure. The residues on the donor domain within 5Å of the ligand are shown with a stick representation.

4.1.1 Docking of UDP-GlcA, UDP-Glc and GDP-Man to crystal structure

To further test if the scoring functions used were able to distinguish between an active compound, UDP-GlcA, and inactive compounds, UDP-Glc and GDP-Man, these three ligands were docked to the crystal structure. The average score for the highest ranked pose is given in Table 2. Although the Vina score is slightly worse for GDP-Man than the other compounds, the difference is not significant. The CNN scoring, however, is significantly lower for GDP-Man. Both scoring functions gave a similar score for UDP-GlcA and UDP-Glc.

Table 2: Average CNN and Vina scores with standard deviation for the top pose obtained when docking UDP-GlcA, UDP-Glc and GDP-Man to the crystal structure. The docking was rerun 10 times.

Substrate	CNN score	Vina score [kcal/mol]
UDP-GlcA	0.92±0.01	-7.6±0.5
UDP-Glc	0.93±0.01	-7.9±0.3
GDP-Man	0.58±0.05	-7.4±0.4

4.2 Study of donor substrate binding to GumK WT and mutants

4.2.1 WT conformation 1

After filtering the poses obtained when docking UDP-GlcA and UDP-Glc to conformation 1 of the WT donor domain as described in 3.4, a total of 84 poses were left for UDP-GlcA (42%) and 90 poses of UDP-Glc (45%). These had an average CNN score of 0.86±0.05 and 0.89±0.04 respectively, see Table A8. Clustering of the poses gave a total of 12 clusters each, where the four biggest clusters contained more than 60% of the poses. A heatmap, a type of graphic representation, of the pairwise RMSD between the poses can be found in Figures 15 and 16. The value of the RMSD is represented as a color, where blue corresponds to a low RMSD and red means a higher RMSD. A dendrogram showing the distance between the poses calculated as described by Eq. 8 is also displayed. Furthermore, the representative poses of the four biggest clusters are shown in the figures.

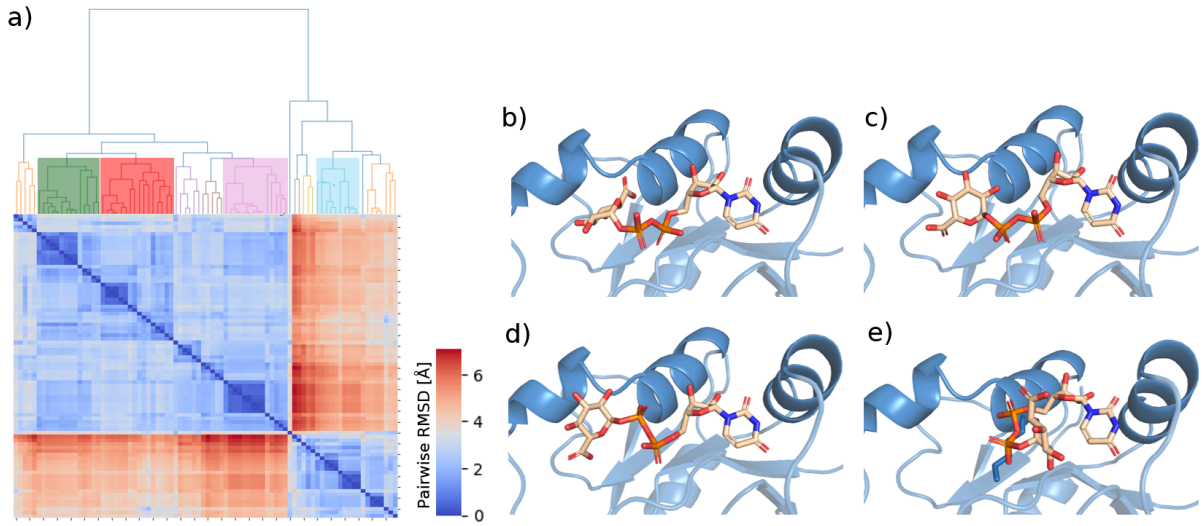

Figure 15: Overview of the results from clustering of UDP-GlcA poses after docking to conformation 1 of the WT donor domain. A heatmap of the pairwise RMSD and the corresponding dendrogram is shown in a). The four biggest clusters are highlighted and the representative pose for these are visualized in b)-e). The biggest cluster colored in red in the dendrogram consists of 19 % of the total poses, with the representative pose shown in b). The clusters represented by the poses in c and d both contain 17% of the total poses. These correspond to the clusters marked with green and pink in the dendrogram. The fourth biggest cluster, highlighted in blue, with the representative pose depicted in e) consists of 12% of the total poses.

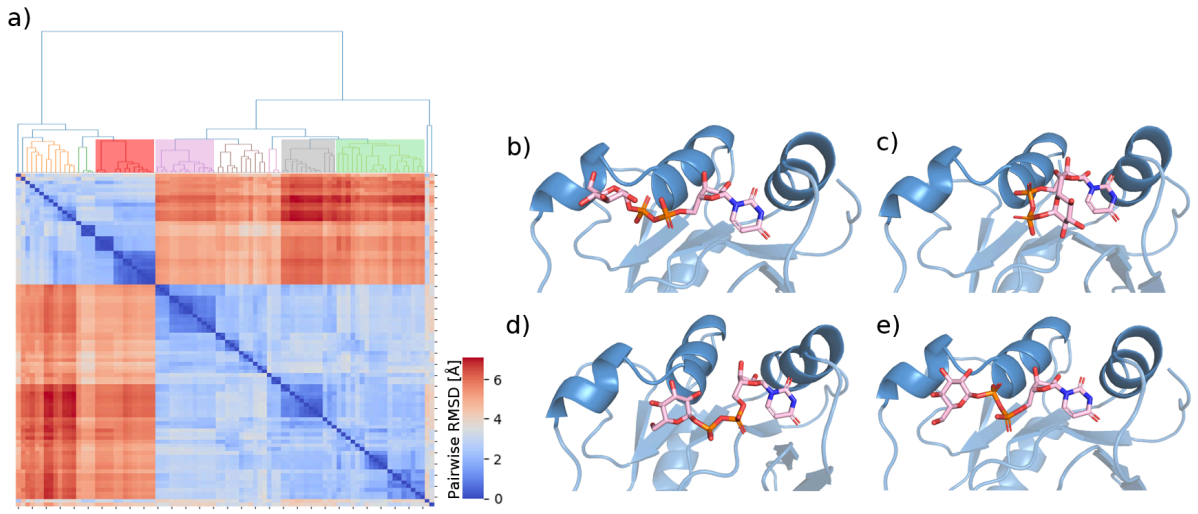

Figure 16: Overview of the results from clustering of UDP-Glc poses after docking to conformation 1 of the WT donor domain. A heatmap of the pairwise RMSD and the corresponding dendrogram is shown in a). The four biggest clusters are highlighted and the representative pose for these are visualized in b)-e). The biggest cluster consists of 21% with the representative pose shown in b). The clusters represented by the poses in c and d both contain 14% of the total poses. The fourth biggest cluster with the representative pose depicted in e) consists of 13% of the total poses. The poses in f and g both contain 12% of the poses.

The distance between C6 of the sugar ring and the nitrogen on the side chain of K307 was calculated for all the filtered poses to describe the position of the sugar ring. Density maps with this calculated distance vs the docking score are shown in Figure 17 and 18. The density is given by the frequency of poses, where areas in the plot with a high frequency are colored in dark red, and areas with few poses are colored in yellow. The poses seem to have a distribution similar to bimodal in terms of the distance. However, UDP-GlcA has more poses with a shorter distance of 4-7Å, as opposed to UDP-Glc where the majority of the poses have a distance longer than 7Å. The UDP-GlcA poses with a shorter distance are also the highest scored poses in regards of both the CNN scoring and the Vina scoring. For UDP-Glc, on the other hand, this trend is not seen. Although the UDP-Glc poses with a shorter distance have among the better Vina scores, the same can not be said for the CNN scoring.

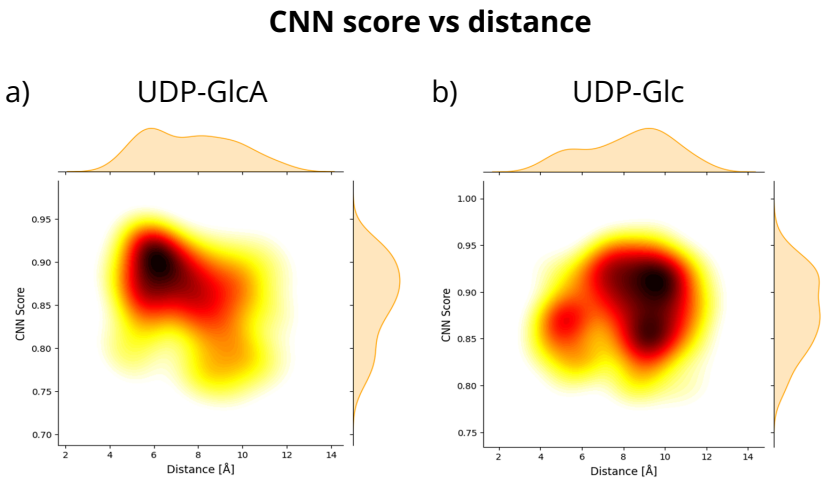

Figure 17: Density plots of the CNN score vs the distance K307-C6 for the filtered poses after docking to conformation 1 of WT donor domain. a) shows the docked UDP-GlcA and b) shows the docked UDP-Glc.

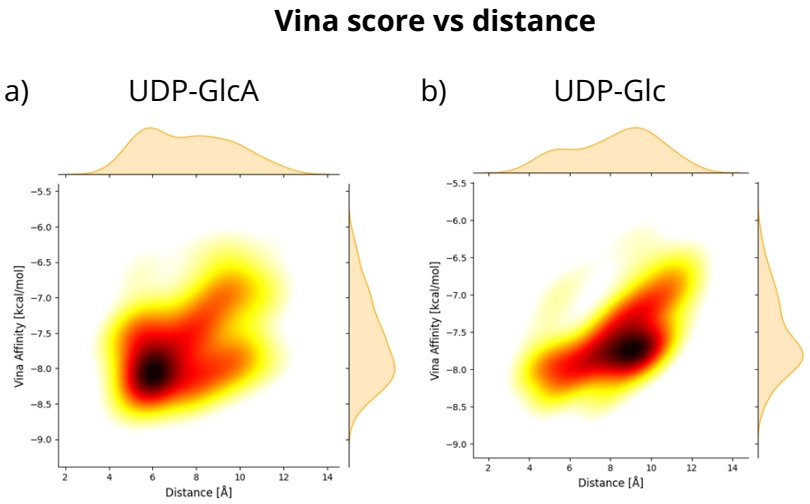

Figure 18: Density plots of the Vina score vs the distance K307-C6 for the filtered poses after docking to conformation 1 of WT donor domain. a) shows the docked UDP-GlcA and b) shows the docked UDP-Glc.

Furthermore, the dihedral angles denoted α and β given by the atoms O^{β}-P^{β}-O^{1}-C^{1} and O^{β}-P^{β}-O^{1}-C^{1} respectively, as shown in Figure 11, were calculated to describe the orientation of the sugar ring. The poses seem to be able to adapt about any value of α. However, the β angle is limited to 60-160 degrees. Figure 19a and 19b shows the dihedral angles plotted against each other for the docked UDP-GlcA and UDP-Glc, divided into two plots. The first plot contains the poses with a C6-K307 distance of less than 7Å and the second one contains the poses with a distance longer than 7Å. The poses corresponding to some data points are also visualized.

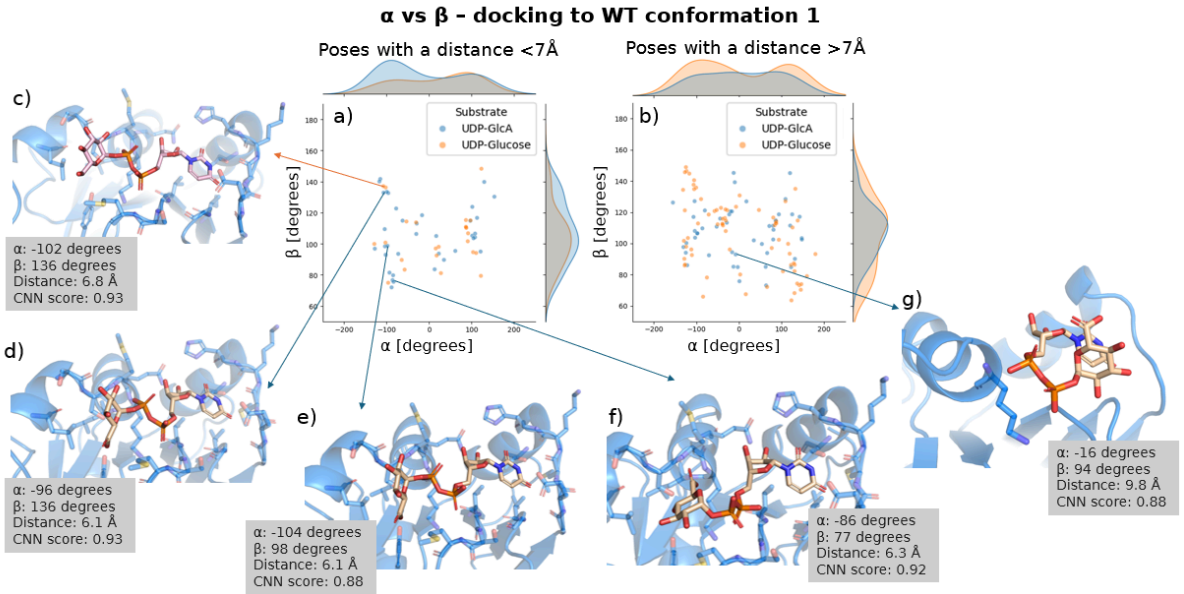

Figure 19: Results from docking of UDP-GlcA and UDP-Glc to conformation 1 of WT donor domain. Plot a) shows the dihedral angles, α vs β, for the poses with a distance between C6 of the sugar and NZ of K307 of less than 7.5Å (after filtering), while b) shows the same plot but for the poses with a distance longer than 7.5 Å. The docked UDP-GlcA are plotted in orange and UDP-Glc in blue. Five poses from the plots are visualized in c-g).

By comparing the projections of the normal vector given by the sugar plane for poses with a similar set of dihedral angles the suitability for using the angles to describe the sugar orientation was estimated. The dihedral angle and normal vectors for the filtered UDP-GlcA poses are listed in Table A9. From this, it was noted that two poses with the same normal vector, i.e. the same orientation of the sugar ring, sometimes gave different dihedral angles. For example, the poses in Figures 19d and 19e have a 34-degree difference in the β angle but are almost identical. Similarly, the same dihedral angles sometimes gave different sugar orientations if the C6-K307 distance differed.

Although the spread of the data points in Figure 19a and 19b is high, some angles are more populated than others. One area of the graph in Figure 19a that is more populated by the UDP-GlcA and contains very few UDP-Glc poses is the bottom left corner. This corresponds to $\alpha < -50$ and $\beta < 100$ degrees. Although the poses in this area have varying conformations, many have the conformation shown in Figure 19f, which is about the same as the representative pose for the biggest UDP-GlcA cluster. This represents a conformation where the carboxylic group interacts with the charged K307. Due to the positioning of the anomeric carbon, the reaction would likely be able to occur as described in 2.4.3 with the donor substrate oriented in this manner. For the poses with a longer distance, K307 instead seems to generally interact with the phosphate groups, as seen for example in Figure 19g, giving the ligand a conformation where the sugar is too far away from the acceptor substrate for the reaction to occur.

4.2.2 Ensemble docking (WT)

To get further information on the binding to the WT donor domain, an ensemble of protein conformations was generated using MD simulations and ClustENMD.

A 1000 ns MD simulation of the donor domain was done. From visually inspecting the trajectory, the Cα3 and Cα4 helices, which contribute to shaping the geometry of the binding pocket (see Figure 6), seemed to be closer together compared to conformation 1. This gave the conformations from the MD simulation a more closed binding site. After around 300 ns an hydrophobic interaction between Met301 and Met231, "blocking" the binding pocket, was seen and continued throughout the rest of the simulation. Therefore, only the first 300ns were used to select protein conformations for the ensemble docking. The distance between L232 and L301 and H275 and Q310 were calculated to approximately describe the distance between the Cα3 and Cα4 helices. Figure 20 shows the plot of these distances for the first 300ns of the simulation. The regions marked in the plot correspond to the time periods used to get three of the conformations for the ensemble docking. They were selected to give conformations with varying distances, one with a shorter distance for both L232-L301 and H275-Q310, one with a longer L232-L301 distance, and one with a longer H275-Q310 distance. Addition to these three conformations, the first frame was also picked to have a conformation similar to the original one (i.e., conformation 1).

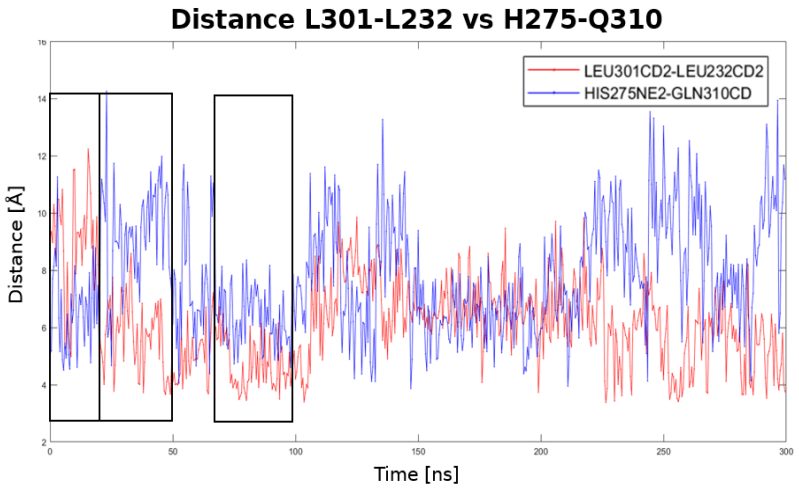

Figure 20: The distance L232-L301 vs H275-Q310 for the first 300ns of an MD simulation of the WT donor domain. The first distance was measured between the CD2 atoms, and H275-Q310 was measured between NE2 and CD.

To improve the sampling of protein conformations, a normal mode analysis approach was also employed. Using ProDy's ANM, data for the normal modes of the donor domain was obtained. The motions described by the two slowest normal modes mainly affect the Cα3 and Cα4 helices, see Figure 21. These modes were used for sampling of protein conformers using ClustENMD.

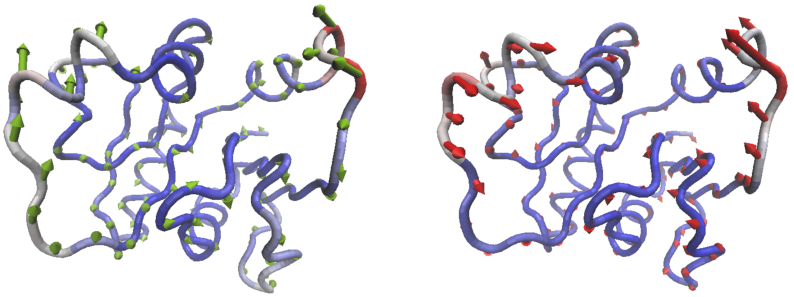

Figure 21: The two slowest normal modes of the donor domain**,** visualized using the NMWiz plugin in VMD (25,34). The slowest normal mode is shown to the right and the next slowest to the left. The vectors show the directions of the motion, and the protein is colored based on the magnitude of movement in each part. The red regions of the protein structure are the parts most affected by the motion, followed by the white regions, while the blue regions are the most rigid.

A lot of conformations generated from ClustENMD had a partial unfolding of one alpha helix (Cα3, see Figure 6b). However, two conformations with a more open binding pocket and an intact Cα3 helix were selected from the ensemble. A total of six conformations (Conformation 2-7) were therefore used for docking of UDP, UDP-GlcA and UDP-Glc, in addition to Conformation 1. The structures of the protein conformations used are visualized in Figure 22. Conformation 2-3 is the ones sampled with ClustENMD, where conformation 2 has a very open binding site, while conformation 3 is only slightly more open than the first conformation. Conformation 4 (from MD) is very similar to conformation 1, whereas conformation 5-7 (MD) have a more closed binding site. The L232-L301 and H275-Q310 distances for each conformation are listed in Table 3.

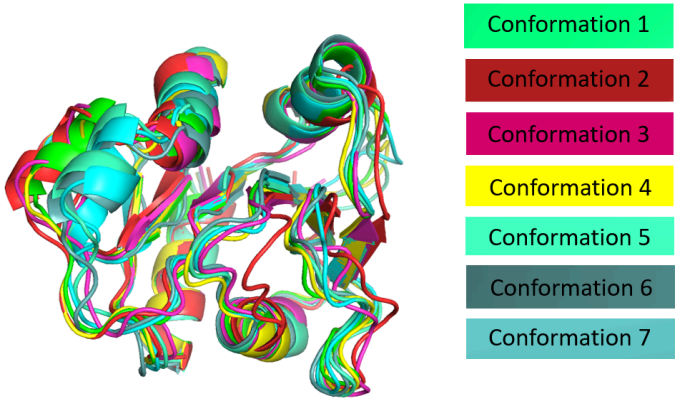

Figure 22: The protein conformations selected for ensemble docking. Conformation 1 (shown in green) is the input protein conformation. Conformation 2 and 3 (red and pink) were generated using ClustENMD, while the other conformations were selected from the MD simulation of the donor domain.

For each new conformation, a reference UDP was selected from the docked UDP poses as described in 3.4.2, see Figure A4. The main docking results from docking to all the WT conformation is found in Table 3. For the docked UDP, this includes the number of UDP poses with a low uridine RMSD compared to the reference UDP, and their average CNN scoring. For UDP-GlcA and UDP-Glc, the number of poses left after filtering (based on their uridine RMSD and CNN score), and the average CNN scoring for these are given.

Table 3: Data from docking of UDP, UDP-GlcA and UDP-Glc to different conformation of the WT donor domain. The distance between the CD2 atom of the residues L310 and L301, and the distance between NE2 of H275 and CD of Q310, for each conformation is given. Furthermore, the number of docked UDP poses with an RMSD < 2Å, and the number of filtered UDP-GlcA and UDP-Glc poses (i.e. the poses with RMSD < 2Å and CNN score > median score) is listed, as well as the average CNN score with standard deviation for these.

Conformation	**L232-L301 [Å]**	**H275-Q310 [Å]**	**UDP**		**UDP-GlcA**		**UDP-Glc**	
			Number poses	**CNN score**	**Number poses**	**CNN score**	**Number poses**	**CNN score**
1	9.1	5.1	83	0.79±0.05	84	0.86±0.05	90	0.89±0.04
2	16.9	9.4	5	0.51±0.1	0	NA	0	NA
3	10.7	5.9	24	0.67±0.1	19	0.65±0.08	14	0.64±0.08
4	9.3	5.2	26	0.57±0.1	3	0.53±0.04	5	0.59±0.10
5	6.0	8.7	78	0.53±0.2	41	0.60±0.2	39	0.63±0.1
6	9.4	6.8	50	0.55±0.1	22	0.62±0.1	36	0.58±0.09
7	3.7	5.5	69	0.54±0.1	53	0.56±0.08	65	0.56±0.07

None of the conformations gave as many filtered poses nor an as high average CNN score as conformation 1. Furthermore, all conformations except for conformation 3 gave poses with a relatively long distance between C6 and K307 compared to conformation 1, and no major difference in terms of this distance was seen between UDP-GlcA and UDP-Glc, see Figure A4. Conformation 3 is the only one containing poses with a distance lower than 6Å (for both UDP-GlcA and UDP-Glc).

The poses from docking to conformation 3 are more narrowly distributed, see Figure A6-A7, compared to the first conformation. When clustering the poses, both substrates ended up with only one cluster containing more than one pose. Most of the poses in this cluster, however, had a pairwise RMSD of 2-3Å. The representative poses for these clusters are shown in Figure 23.

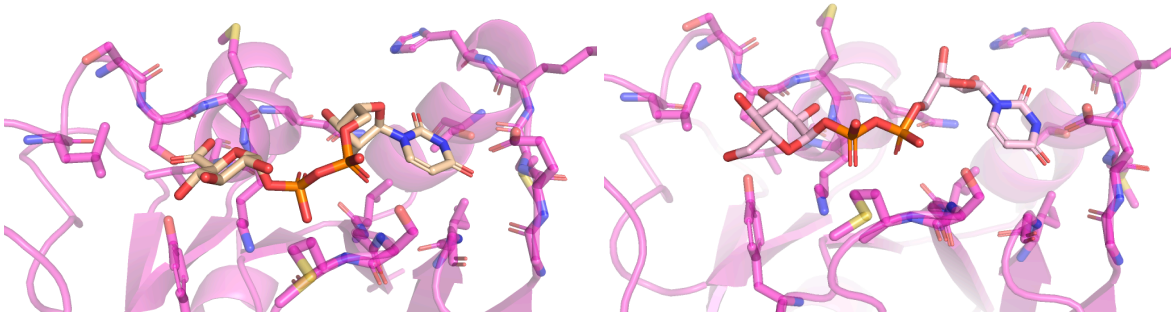

Figure 23: The representative poses for the largest clusters when docking to conformation 3 of the WT donor domain. a) shows the UDP-GlcA, and b) shows the UDP-Glc.

Conformations 5-7 gave more clusters, with the majority of the poses belonging to the largest 2-3 clusters. These are fairly similar for all the conformations, and between UDP-GlcA and UDP-Glc. The three biggest clusters of UDP-GlcA docked to conformation 7 is shown in Figure 24, and the other poses can be found in the Appendix (Figure A8-A12).

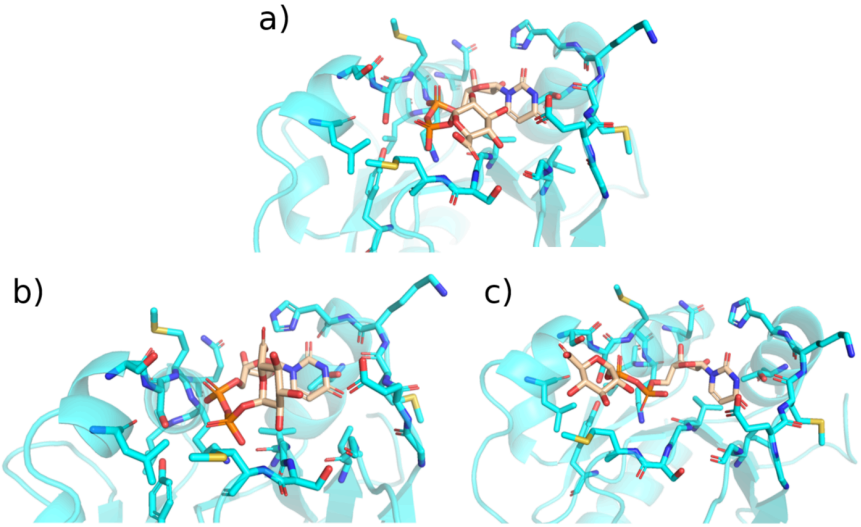

Figure 24: The representative poses for the three biggest clusters from docking of UDP-GlcA to conformation 7 of the WT donor domain. 19% of the poses belong to the cluster represented by the pose in a), 14 % belong to the poses represented in b) and c) respectively.

The dihedral angles for all the filtered poses are plotted in Figure 25, with the data points colored according to the distance between C6 and K307. Also here it can be seen that conformation 3 has mostly poses with a shorter distance. The UDP-Glc poses seem to mainly adopt an α of around -100 or 100 degrees, whereas UDP-GlcA also include poses with an α angle closer to 0 degrees. The plots for the three more closed conformations differ slightly among themselves, but no big difference is seen for the two substrates for each conformation. Conformation 2 and 4 are excluded from the figure since no, or only very few, poses remained after filtering.

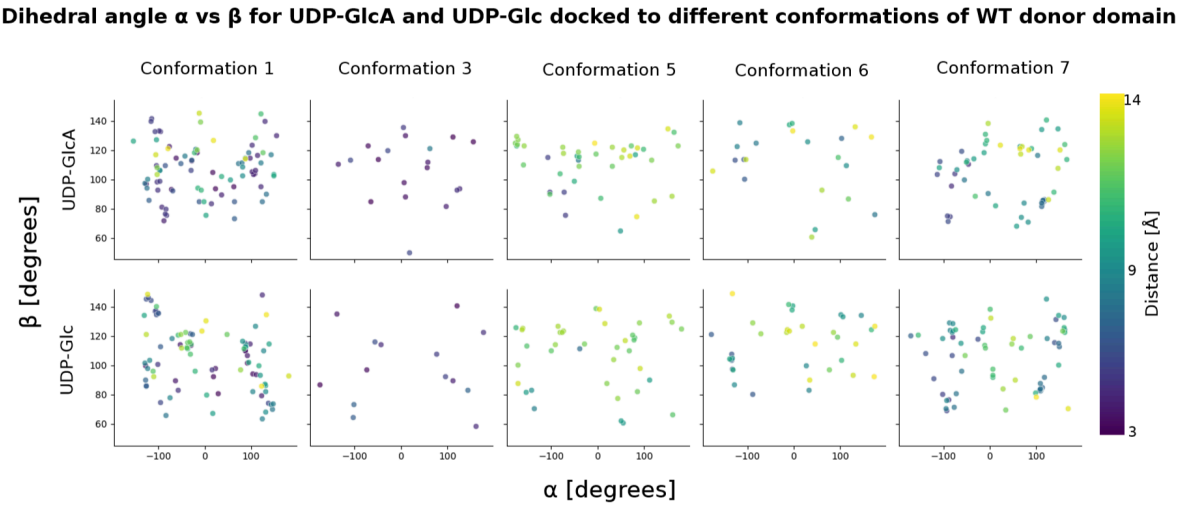

Figure 25: The dihedral angles, α vs β, for the UDP-GlcA and UDP-Glc poses remaining after filtering are plotted against each other for conformation 1,3,5,6 and 7 of the WT donor domain. The data points are colored after the distance between C6 of the sugar and NZ of K307.

4.2.3 Mutants

Furthermore, UDP-GlcA and UDP-Glc were docked to six different mutants. A list of the different mutations of the mutants is found in Table 1. The coordinates for mutant 1 and 6 were generated using AlphaFold2 (38). The predicted structures had a high–very high confidence score, and were very similar to that of the WT conformation 1, see Figure A13, with a 1.9 and 1.6 Å RMSD compared to the WT for mutant 1 and 6 respectively. After visual inspection of the binding site, though, it was noted that the side chain of R218 for both mutants pointed in towards the binding pocket. Therefore, the rotamer for this side chain was changed in ChimeraX (37) and the new structures were energy minimized as described in 3.4.3. However, the number of filtered poses and the CNN score for most of these were relatively low compared to the WT. It was therefore decided to redo the docking for mutant 1 with R218 as a flexible side chain instead to see if this would improve the result. When doing this, the average CNN score for the filtered poses increased slightly for the docked UDP-GlcA (0.67±0.11 vs 0.71±0.07), but the number of filtered poses decreased, so the results from the rigid docking were used instead. The results for the flexible side-chain docking can be found in Appendix (Table A14 and Figure A28-31). All other mutants were made in PyMol and then energy minimized as described in section 3.4.3.

After filtering the poses generated based on the CNN score and RMSD of the uridine part compared to the UDP-GumK crystal structure, a varying number of poses were left for the different mutants, as given in Table 4. The table also includes the average CNN score and Vina score for the filtered poses.

Table 4: Data from docking of UDP-GlcA and UDP-Glc to the six mutants. The number of poses remaining after filtering and their average CNN and Vina score with standard deviation is listed.

Mutant	UDP-GlcA			UDP-Glc		
	Number filtered poses	CNN score	Vina score [kcal/mol]	Number filtered poses	CNN score	Vina score [kcal/mol]
1	53	0.67±0.1	-8.04±0.7	54	0.72±0.1	-8.03±0.6
2	61	0.78±0.09	-7.64±1	59	0.77±0.09	-7.62±1
3	86	0.87±0.05	-7.33±0.6	81	0.87±0.05	-7.18±0.5
4	62	0.82±0.07	-7.65±0.8	80	0.85±0.05	-7.33±0.7
5	88	0.87±0.05	-7.30±0.6	80	0.87±0.04	-7.30±0.5
6	40	0.70±0.1	-7.53±0.5	52	0.73±0.1	-7.53±0.5

Density plots of the CNN score vs the distance between the C6 of the sugar and the side chain of residue 307 for all of the mutants is found in Figure 26. The distance for mutant 1 is measured from the oxygen of T307, for mutant 2 and 6 it is measured to CB of A307, and for the rest it is measured to NZ of K307. Due to the difference in atoms used when calculating the distance, they are not directly comparable to the WT for all of the mutants. Still, most mutants and substrates seem to obtain a bimodal distribution in regards to the distance. However, none of distributions have a majority of poses with a shorter distance, as was seen for the native substrate (UDP-GlcA) when docked to the WT conformation 1.

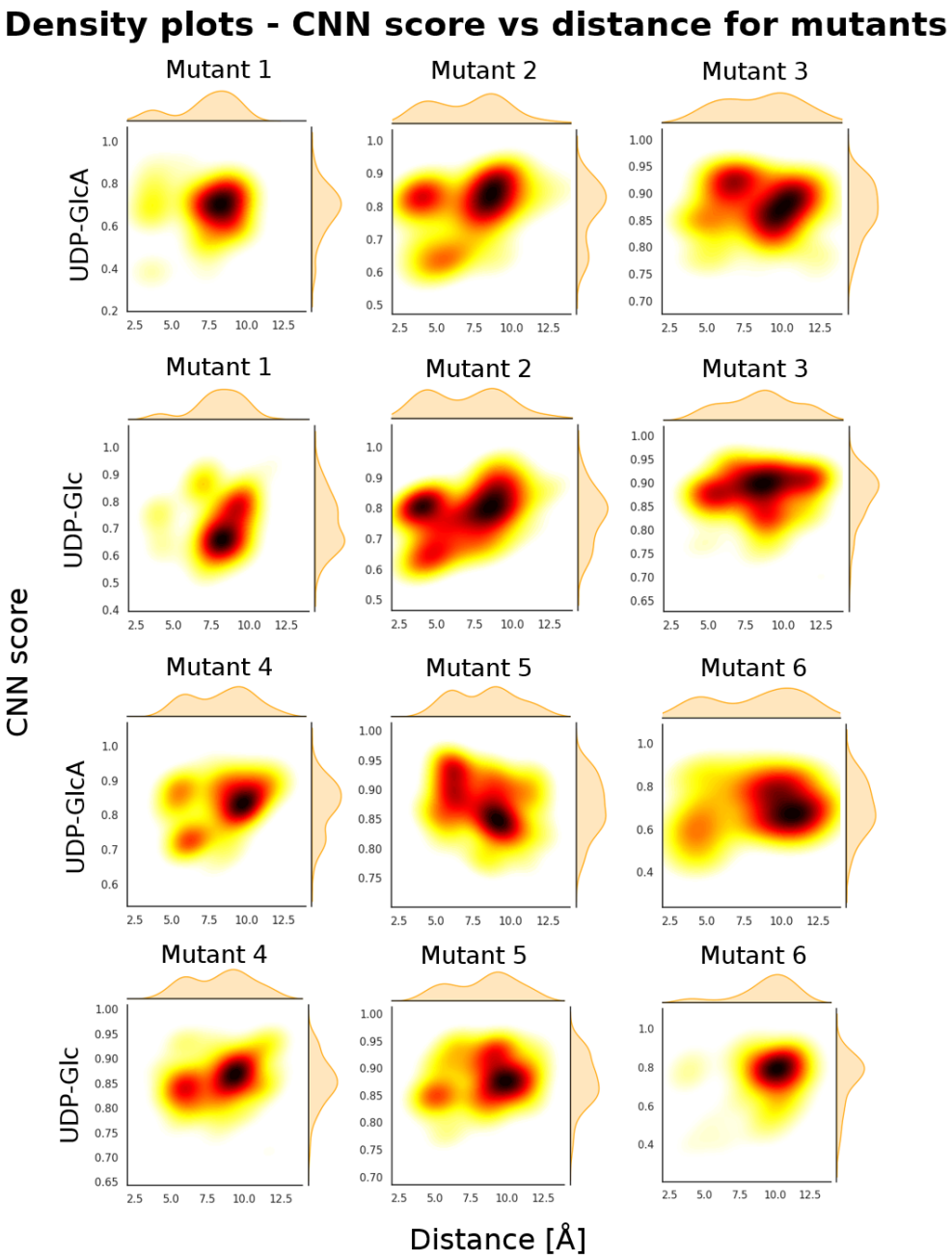

Figure 26: Density plots of CNN scoring vs distance between C6 and the side chain of residue 307 for UDP-GlcA and UDP-Glc docked to the six mutants. The distance for mutant 1 is measured from the oxygen of T307, for mutant 2 and 6 it is measured to CB of A307, and for the rest it is measured to NZ of K307.

Furthermore, the dihedral angles for the filtered poses, colored after the distance between C6 and the side chain of residue 307 can be seen in Figure 27.

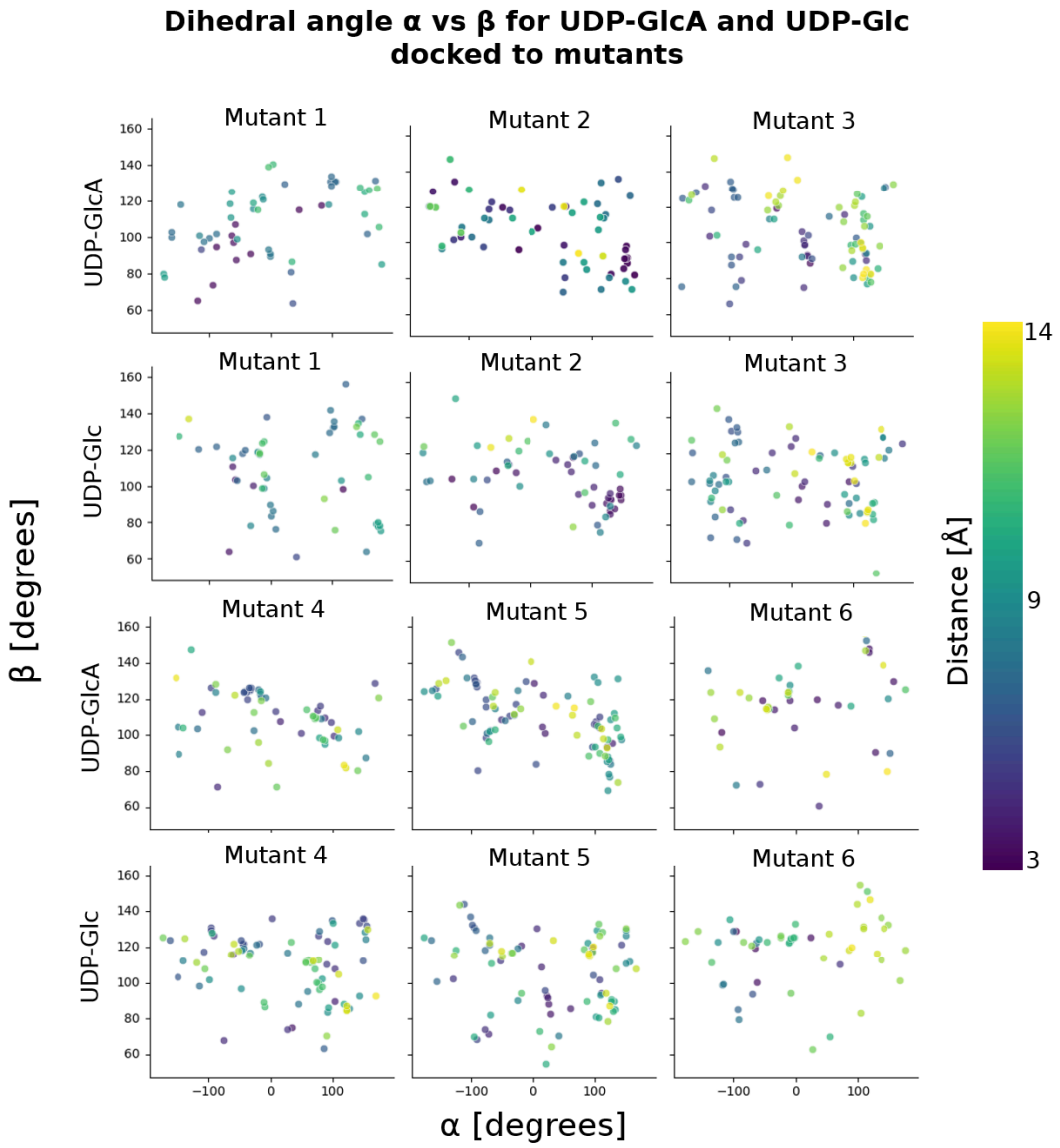

Figure 27: The dihedral angles, α vs β, for the docked UDP-GlcA and UDP-Glc poses remaining after filtering are plotted against each other for the six different mutants. The data points are colored after the distance between C6 of the sugar and the side chain of residue 307. The distance for mutant 1 is measured from the oxygen of T307, for mutant 2 and 6 it is measured to CB of A307, and for the rest it is measured to NZ of K307.

Clustering of the filtered poses from docking to mutant 1 resulted in 8 clusters for both UDP-GlcA and UDP-Glc. The representative poses for the biggest clusters are very similar for the two substrates, and include conformations where the sugar ring is in close contact to the mutated arginine at residue 231. The representative poses for the largest UDP-Glc clusters are shown in Figure 28.

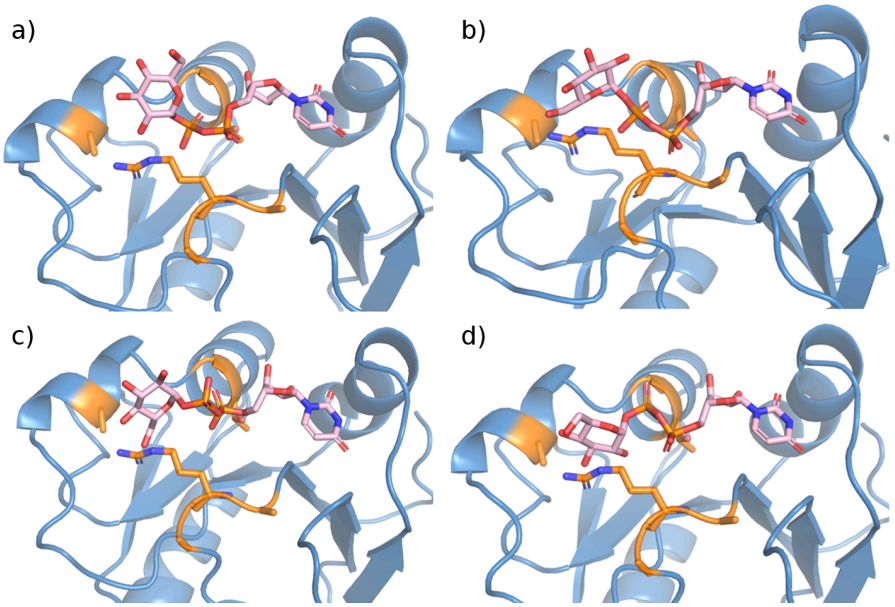

Figure 28: The representative poses for the four biggest clusters from docking UDP-Glc to mutant 1. 22% of the poses belong to the cluster given by a), 20% belong to the clusters given by the poses in b) and c) respectively, and 17% belong to the cluster given by d).

Due to the mutation on residue 307, the distance for mutant 2 is calculated from the CB of the alanine. The poses still appear to exhibit a bimodal distribution in regard to the C6-A307 distance, see Figure 26. The docked UDP-GlcA has a slightly shifted distribution of more poses with a longer distance than what was observed for conformation 1 of the WT. The UDP-Glc poses, on the other hand, have more poses with a shorter distance than when docking UDP-Glc to the WT conformation 1. However, from the plot of the dihedral angles for these poses given in Figure 27, they seem to differ from the WT. Many of them have an $\alpha > 100$ and $\beta < 100$ degrees, as opposed to WT which has few poses with a distance of less than 7Å in this region, see Figure 19a. When looking at the representative poses of the clusters, it becomes evident that some of the poses with a shorter distance for both UDP-GlcA and UDP-Glc correspond to a conformation not seen for the WT, as shown in Figure 29. These belong to the third biggest cluster containing several almost identical poses. The other larger clusters contain poses with a broader distribution.

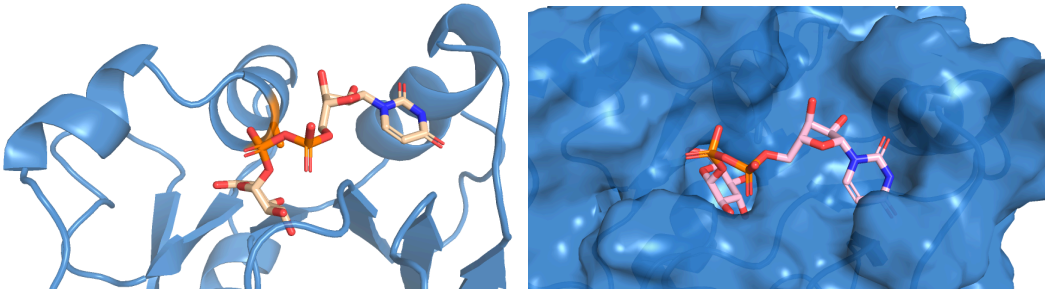

Figure 29: The representative poses for the third biggest cluster when docking to mutant 2. This cluster is very similar for both UDP-GlcA and UDP-Glc and contains 6% of the filtered poses, which all have an pairwise RMSD within 0.7Å. a) shows the representative pose for UDP-GlcA, with the mutated residues highlighted in orange. b) shows the representative pose for UDP-Glc with the protein surface visualized.

The docked UDP-GlcA to mutant 3 contains a majority of poses with a longer distance, and none of the representative poses for the four largest clusters (containing 68% of the poses) resemble that of the representative pose of the largest UDP-GlcA cluster when docking to WT conformation 1. Similarly, none of the three biggest clusters for the docked UDP-Glc (containing 66% of the poses) have a representative pose such as the one from the largest UDP-Glc clusters when docked to the WT conformation 1. However, other than this, they are similar to the WT UDP-Glc results.

Furthermore, clustering of the mutant 4 poses gave 10 clusters for both UDP-GlcA and UDP-Glc, with almost identical representative poses between the substrates for the largest clusters. Figure 30 shows the representative poses for the 3 biggest clusters of UDP-Glc.

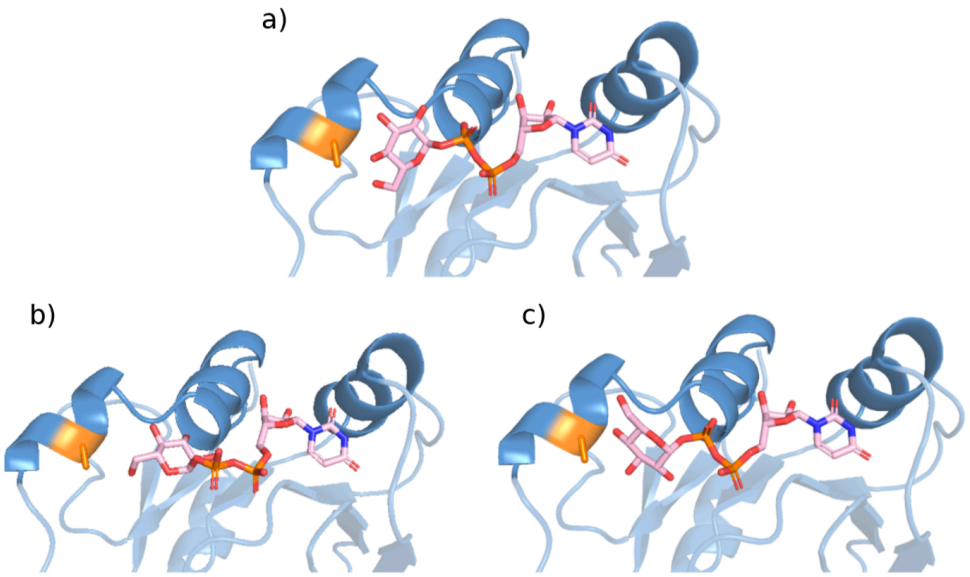

Figure 30: The representative poses for the third biggest cluster when docking UDP-Glc to mutant 4. 26% of the poses belong to the cluster with a representative pose as shown in a), 18% of the poses belong to the cluster given by b) and 13% belong to c).

Mutant 5 gave 14 clusters each for UDP-GlcA and UDP-Glc. The largest UDP-GlcA cluster has a representative pose similar to that of the largest UDP-Glc cluster when docked to WT conformation 1. The other three biggest clusters are more like the ones obtained from docking of UDP-GlcA to the WT. The four biggest clusters for UDP-Glc docked to mutant 5 can be seen in Figure 31.

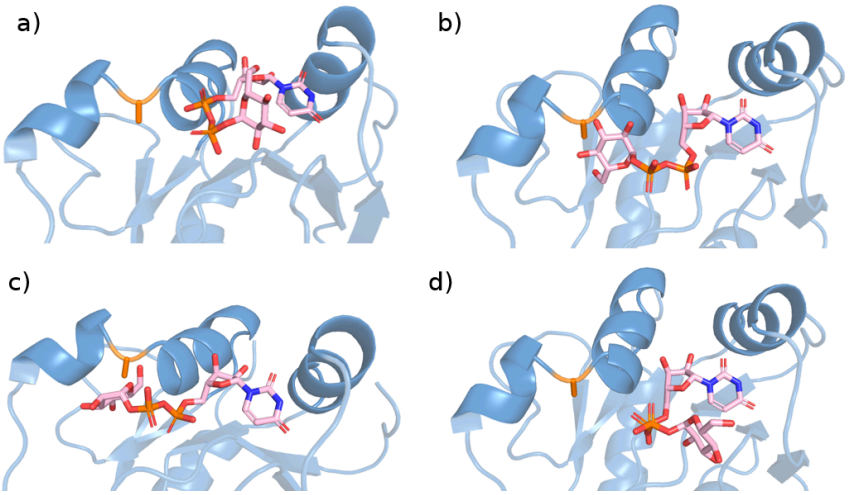

Figure 31: The representative poses for the four biggest clusters from docking UDP-Glc to mutant 5. The two largest clusters contain 18% of the poses each, with the representative poses shown in a) and b). 11% belong to the cluster with the representative pose in c) and another 11% to the one in d).

Clustering of the poses for mutant 6 resulted in quite few clusters, 4 for UDP-GlcA and 6 for UDP-Glc, with a very large portion of the poses belonging to the biggest cluster (60% and 50%). The representative pose for the biggest cluster is very similar for UDP-GlcA and UDP-Glc, see Figure 32.

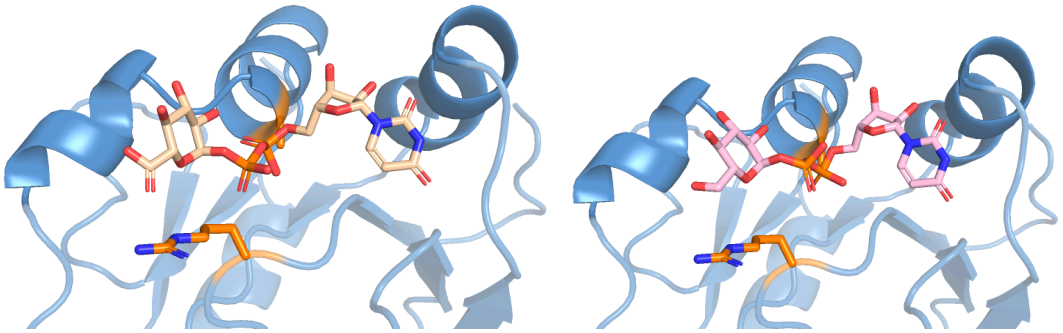

Figure 32: The representative poses for the largest UDP-GlcA and UDP-Glc clusters when docking to mutant 6. 60% of the UDP-GlcA belongs to the cluster represented by a) and 50% of the UDP-Glc poses belong to the cluster represented by b)

Heatmaps of the pairwise RMSD for the poses obtained from docking to the mutants, as well as the representative poses for the biggest clusters of the substrates/mutants that were not shown in the result section, can be found in the Appendix (Figure A14-A26).

5. Discussion

Tailoring the chemical composition and properties of Xanthan presents a promising approach for broadening its potential applications. One method for achieving this chemical modification involves engineering the glycosyltransferase enzymes in the Xanthan biosynthesis pathway. This requires a deep understanding of the enzyme's mode of action to allow for targeted mutations to alter their function. In the case of glycosyltransferases, this understanding revolves around how they bind to specific donor and acceptor substrates. Mutagenesis can then be used to modify their substrate specificity. A computationally fast and cost-effective way to study the interactions between biomolecules, such as glycosyltransferases and their substrates, is through molecular docking. The aim of this thesis was, therefore, to utilize molecular docking to study the substrate binding to the donor domain of the glycosyltransferase GumK and mutants of GumK. Furthermore, the possibility of using docking as a method to screen for mutants with an altered specificity was explored. To accomplish this, the performance of the docking program used and the impact of different docking settings were first tested, and the results from these tests are discussed in 5.1. Thereafter, further analysis of the binding of UDP-GlcA and UDP-Glc to the WT and mutants was performed, with the results discussed in section 5.2.

5.1 Evaluation of docking performance and docking settings

The docking performance when changing the two different settings exhaustiveness (number of Monte Caro chains in the sampling) and autobox (the search area) was tested by redocking of UDP and UDP-GlcA to the crystal structure. If the scoring function is assumed to perfectly be able to rank the poses, an increase in exhaustiveness should lead to an increased possibility of ending up with a well-docked pose. This is because increasing the number of Monte Carlo chains run in the sampling step means a more thorough exploration of the ligand conformational space. However, since the scoring functions involve many simplifications and assumptions, they are, in reality, likely not perfect. Due to this, increasing the exhaustiveness only means that the poses generated from the sampling step are more likely to have a high score. Considering the fact that the scoring functions used in this docking setup differ between the sampling step (vina scoring) and the final ranking of the poses (CNN scoring), and the possibility that these do not fully correlate, makes it difficult to predict anything regarding how the exhaustiveness will affect the scoring.

Although the scoring did increase with a higher exhaustiveness level for a few of the docking scenarios, this correlation was mostly not seen and nothing conclusive can be said regarding how the exhaustiveness affects the score of the poses, see Appendix (Table A1-A4). However, what is more interesting to know is whether or not the exhaustiveness affects how similar the top-ranked poses are to the crystal structure. This effect seems apparent, at least when using a larger autobox, see Figure 12. The percentage of the top 3 poses with an RMSD < 2Å is very high independent of the exhaustiveness when utilizing a smaller autobox (average was always over 70%), whereas for the larger autobox, a clear increase of well-docked top poses can be seen with higher exhaustiveness. This is not surprising, since a larger autobox increases the search space and more sampling would need to be done to find a good conformation.

Except for when utilizing a large autobox and low exhaustiveness, the top poses are generally very well-docked, see Figure 12, which is promising. However, the receptor used here is the UDP-bound crystal structure. This likely facilitates the discovery of favorable poses, since this conformation presumably binds the substrate well, with the amino acids positioned in a way that enables favorable protein-substrate interactions. The analysis of the donor substrate binding as described in section 4.2 (and discussed in 5.2) instead uses non-crystal protein structures, which may have an impact on the results.

Figure 13 shows the correlation between the scoring functions (Vina and CNN) and the RMSD compared to the crystal structure UDP. Generally, poses with an RMSD less than 2Å compared to the crystal structure are considered well-docked. However, no equivalent thresholds were found for the docking scores. Therefore, no statistical methods were employed to determine the number of false positives/negatives, as this would need a clear definition for what is considered a "good" and "bad" docking score. Still, the graphs can be used to get an idea of how well the scoring functions manage to predict the quality of a docked pose. Figure 13a and 13b shows this correlation for the docked UDP. The high scored poses generally have a low RMSD. However, not all well-docked poses have a high score. This means that filtering poses based on the scoring functions might exclude relevant/interesting poses in terms of RMSD to the ligand-bound crystal structure.

For the docked UDP-GlcA, a low RMSD of the UDP part is needed for the pose to have a realistic conformation. However, in contrast to when docking only the UDP, a low RMSD does not necessarily mean that the pose is well-docked since the sugar part still could have an inaccurate conformation. Due to this, Figure 13c and d, showing the docking scores vs the RMSD for the docked UDP-GlcA, can mainly be used to get an idea of to what extent the scoring function gives an incorrect pose a high score, rather than how well the scoring functions manage to predict a well-docked pose. Here, a bigger difference between the accuracy of the scoring functions can be seen, where the Vina scoring seems to give more poses with a large RMSD a high score, compared to the CNN scoring.

The highest scored UDP-GlcA from redocking to the crystal structure, see Figure 14, was used to define the autobox used for the future dockings. This was mainly done to have an autobox that would be replicable across different receptors and ensure that the whole binding site was covered. The autobox is similar to the one manually picked in MGLTools used when testing the docking settings, see Figure A3. However, it is possible that using UDP as the autobox would have been as appropriate since the option *autobox_extend* was always on.

Furthermore, as one of the aims of the project is to use docking as a method to test GumK mutants for an altered donor substrate specificity, the docking score for docking the three compounds UDP-GlcA, UDP-Glc and GDP-Man was compared, see Figure 10 for their chemical structure. The CNN scoring function was the only one that gave the negative control GDP-Man a significantly lower score than the native substrate UDP-GlcA, see Table 2. However, all scoring functions failed to distinguish between the inactive UDP-Glc and UDP-GlcA. It is possible that UDP-Glc actually binds to GumK although it is not reactive, or the scoring function can simply not differentiate between these two analogues. The lack of discrimination between UDP-GlcA and UDP-Glc served as a motivation to try an alternative approach, and

investigate if generating an ensemble of docked poses could indicate differences between the two substrates. The results from this are discussed in section 5.2.

Overall, the CNN scoring appears to perform slightly better than the empirical scoring functions tested, since it gave the negative control GDP-Man a significantly lower score, see Table 2, and seemed to give fewer poses with a high RMSD a good score, see Figure 13. Therefore, the CNN scoring was used when filtering the poses, in addition to the RMSD. The fact that the CNN scoring function gave better results than the more physics-based one shows the potential of these types of algorithms. The field of ML is rapidly evolving, so optimistically future ML scoring functions will have an even better accuracy than the CNN model used here. However, a drawback of using ML algorithms to study the protein-ligand interaction is the lack of physical interpretation, making it difficult to know exactly what causes the scoring functions to rank the poses the way it does.

5.2 Study of donor substrate binding to GumK WT and mutants

The donor substrate binding to the WT and mutants of GumK were further studied by docking to the donor domain as described in section 3.4. GumK is a flexible enzyme, with a proposed interdomain motion bringing the domains closer/further apart from each other. The reason for not using the whole protein during the study is based on the assumption that UDP-GlcA binds the donor domain independently of the acceptor domain when the enzyme has a more open conformation. By excluding the acceptor domain, the possibility that the protein structure used has a conformation unsuitable for substrate binding, due to the positioning of the two domains, is avoided. Furthermore, it enables the sampling of protein conformations for the ensemble docking to focus on the dynamics of the donor domain, instead of the protein as a whole. However, although only considering one of the domains has its advantages, it is important to keep in mind that it is an approximation that could have an impact on the results.

The decision to analyze a larger number of ligand conformations, instead of just looking at the top-ranked pose(s) served multiple purposes. As already mentioned, it could possibly be used as a way to distinguish between UDP-GlcA and UDP-Glc, since the scoring functions alone did not manage to do this. Furthermore, since the sugar part of the donor substrate likely is flexible, it should give a more realistic understanding of the binding than only considering one possible bound conformation. The outcome for the different docking scenarios is discussed below.

5.2.1 WT conformation 1

Docking of UDP-GlcA to the WT conformation 1 donor domain resulted in numerous diverse ligand conformations with a high score and low uridine RMSD value. This was expected since the sugar part likely is flexible and therefore can adapt different conformation while still being considered a stable ligand-protein complex. Although the poses display a diversity, some conformations are more commonly obtained, see Figure 15a displaying a heatmap of the pairwise RMSD between the poses. Clustering of the poses gives an insight into what conformations are most frequently populated. The representative pose for the largest cluster shows an interaction between the charged lysine (residue 307) and the carboxyl group of the sugar as well as the first phosphate group. For the other bigger clusters, this interaction only involves one of the two phosphates, see Figure 15b-15e. However, it should be taken into consideration that the poses belonging to

one cluster still differ slightly as seen in Figure 15a, and might not have the exact same conformation as the representative pose.

Even though there are similarities between the poses generated for UDP-Glc when compared to UDP-GlcA, the distribution of the poses differs. This difference in distribution could potentially be linked to the donor substrate specificity in GumK. If so, it could be used as a method to see how different mutations affect the activity. The idea would then be that mutants with a pose distribution similar to the one obtained for the native substrate docked to the WT should be able to perform the reaction with the substrate tested. Additionally, since the scoring functions should correlate, at least to an extent, with the possibility that the substrate binds the enzyme, a relatively high score among the poses is likely also necessary.

The main difference between the distribution of poses seen was that more UDP-GlcA poses had a shorter distance between the C6 of the sugar and K307 than UDP-Glc, see Figures 17 and 18. In an attempt to further distinguish the populations of poses, the dihedral angles for the poses were analyzed, see Figure 19. The most obvious area of the plot that seemed to be more populated by the UDP-GlcA was the region with $\alpha < -50$ and $\beta < 100$ for the poses with a distance less than 7Å, where many of the poses obtained a conformation like the one in Figure 19f. The pose in Figure 19f represents a conformation similar to the representative pose of the largest UDP-GlcA cluster, where the carboxylic group interacts with the charged K307. An interaction between two charged particles is typically stronger than between a charged particle and a dipole. So, it is not unreasonable to expect this conformation to be more favorable for UDP-GlcA, which contains a negatively charged carboxylic group, compared to the uncharged glucose in UDP-Glc. Due to the positioning of the anomeric carbon, the reaction would likely be able to occur as described in 2.4.3 with the donor substrate oriented in this manner. Therefore, this region could perhaps serve as an indicator of whether or not the protein can undergo the reaction with the tested substrate. However, the fact that it is possible to obtain quite different dihedral angles for two very similar conformations, as exemplified in Figure 19, reduces the reliability of this approach since conformations similar to Figure 19f may not consistently generate a data point in the lower left corner of the dihedral plot.

5.2.2 Ensemble docking (WT)

Ensemble docking, meaning docking to several different protein conformations, is one way to somewhat account for the dynamics of a protein in molecular docking without it being too computationally expensive. Although conformation 1 gave many poses with both a high score and a well docked UDP part, incorporating the protein dynamics can give a more comprehensive understanding of the binding if the substrate binds according to the induced fit or conformational selection model, see section 2.1.1.

The two slowest normal modes of the donor domain correspond to a motion of mainly the Cα3 and Cα4 helices, giving a more open or closed binding pocket, see Figure 16. To get a protein ensemble that somewhat represents these motions, the ensemble used contained two conformations with a more open binding site (compared to conformation 1), one very similar to the first conformation, and three more closed ones, see Figure 17. All protein conformations had a worse docking result compared to conformation 1 for all substrates, both in terms of the scoring and the number of poses with a well-docked uridine part. This suggests that they do not form as favorable ligand-protein complexes as the first conformation.

Docking of UDP to the different protein conformations was done partly to obtain a new reference UDP for estimating the quality of the docked uridine part, but also to see how the different structures affect the UDP binding without considering the sugar part of the substrate. Previous studies indicate that the UDP part of the donor substrate is less flexible than the sugar, at least when bound to GumK in absence of the acceptor substrate (3). It is, therefore, assumed that the UDP part should bind in a similar fashion to the protein conformations in the protein ensemble as in the UDP-bound crystal structure. Hence the use of a reference UDP that is based on the position of the crystal structure UDP. However, it is possible that the change in protein conformation causes the UDP to bind in a way that differs more than anticipated. In that case, the filtering process, where only poses with an RMSD less than 2Å compared to the reference structure were kept for analysis, might have resulted in relevant poses being excluded.

The most open conformation (conformation 2) only had 5 UDP poses out of 200 with an RMSD less than 2Å. This means that the UDP likely does not have a strong affinity for this conformation, perhaps because the residues are too far away from each other and the substrate for all of them to have a favorable interaction with UDP. Since the UDP likely can not bind this conformation well, it is not surprising that it did not have any UDP-GlcA nor UDP-Glc poses left after filtering. The next most open conformation, conformation 3, only has around a third as many well-docked UDP poses as the original conformation. However, these are scored relatively high compared to the other conformations. It is difficult to know how to interpret this, is it better with a high number of structurally well-docked poses, or is a few well-docked poses with a high score better? Few poses with a low RMSD means that the docking program ranks other ligand conformations high, reducing the reliability that the poses with a low uridine RMSD actually correspond to realistic ligand-protein structures. On the other hand, many filtered poses that are scored low implies that the docking program is uncertain about the poses. Therefore, either scenario likely indicates a reduced binding affinity towards that protein conformation, although it is possible that the substrate still can bind, just not as strongly compared to the case where the docking gives many high-scored poses.

The three more closed conformations (conformation 5-7) obtained mainly UDP-GlcA and UDP-Glc poses with a longer distance, see Figure A5, suggesting that the reaction does not occur when the protein has a conformation like these, since the sugar is too far away from the acceptor substrate. This was expected because there is not much space in the binding pocket for the sugar to fit. Docking to conformation 3, on the other hand, resulted in poses with a shorter C6-K307 distance.

Something surprising is the fact that the conformation most similar to the original conformation, conformation 4, has among the worst docking results for all substrates. The residues at the binding site differ minimally from the first conformation, see Figure A4, so it is difficult to tell the reason why the docking results are not better. However, it illustrates the point that docking is very sensitive to the structure given.

5.2.3 Mutants

All mutants have a structure very similar to that of the first WT conformation. Due to this, and the fact that the biggest difference between the UDP-GlcA and UDP-Glc poses was seen for this WT conformation, they will be compared to the WT conformation 1.

The number of poses left after filtering and their average CNN score was rather low for mutant 1 suggesting that it does not bind UDP-GlcA nor UDP-Glc that well. Interestingly though, both substrates have a better average Vina score for mutant 1 than for the WT, see Table 4. The fact that the scoring functions do not agree makes the estimation more unreliable. However, from looking at the representative poses for the largest clusters it seems like the distributions for both UDP-GlcA and UDP-Glc differ from that of the WT.

Mutants 2-5 all contain a point mutation to alanine. None of the mutants have a distribution of UDP-Glc poses exactly as the one one obtained when docking the native substrate to the first conformation of the WT, indicating that they probably are not active with UDP-Glc as substrate.

Mutant 4 (L301A), though, obtained a higher number of filtered poses for the docked UDP-Glc than UDP-GlcA (80 vs 62), with docking scores similar to that of the WT. The distribution of the poses based on the distance can be said to be in between that obtained for the two substrates when docking to the WT, see Figure 26. This could indicate that the enzyme is able to catalyze the reaction with UDP-Glc, but not as efficient as the WT. However, there is a lack of poses in the region where $\alpha < -50$ and $\beta < 100$, see Figure 27, which goes in line with the fact that none of the biggest UDP-Glc clusters for mutant 4 seem to represent the conformation where the sugar has a very close interaction with K307, see Figure 30. If this region of dihedral angles is important for the specificity, as speculated from the results of the WT docking, the possibility that mutant 4 has a changed specificity towards UDP-Glc is less likely.

Furthermore, the mutations of mutant 2-5 seem to have a negative impact on the enzyme's ability to undergo the reaction with the native substrate UDP-GlcA, based on the number of poses after filtering, their score and distribution. Mutant 2 (K307A) obtained a quite low number of UDP-GlcA poses after filtering (61), with a relatively low average CNN score. This implies that the mutation has a negative effect on the binding affinity, which aligns with previous studies showing that this particular mutation leads to an increase K_M and decrease of the V_{max} (3). As already mentioned, mutant 4 (L301A), also had relatively few filtered poses, with a distribution shifted toward more poses with a longer distance compared to the WT. Mutants 3 and 5 (M306A and S305A), on the other hand, had a number of high-scored filtered poses similar to the WT. Even though this might mean that the substrate binds well to the enzyme, the mutations seem to have a negative effect on the enzyme activity when comparing the distribution of the poses. The results suggest that the residues K307, M306, L301, and S305 are crucial for the substrate binding, since mutating these to alanine seem to have a negative effect on the enzyme activity. Therefore, targeting these residues could be a strategic method when searching for mutants with an alter substrate specificity.

Mutant 6 (K307A and M231R), is mutated so that residue 231 contains a positive charge instead of 307. Therefore, it is expected that UDP-GlcA should prefer an interaction between the carboxylic group and the mutated arginine at 231 rather than the mutated alanine at 307. This was also seen, both from the shifted distribution of the distance, see Figure 26, and from the clustering of the poses. Similar results were also obtained for UDP-Glc. Considering this and the fact that both the number of filtered poses and the average CNN score were quite low, this mutant is most likely not active with any of the substrates.

The RMSD for the uridine part when docking to the mutants was calculated compared to the crystal structure after aligning the proteins. It is possible that it would have been more accurate to obtain a new reference UDP as done for the ensemble docking to the WT. The reason for not doing this was because the mutants were docked to first, prior to the ensemble docking and before making the decision to use a new reference pose for the different conformations. However, it would likely not have any major impact on the results since all the mutants have a conformation very similar to that of conformation 1, with all residues involved in the UDP binding oriented in a similar way.

Most of the mutants have fewer filtered poses than the WT conformation 1, see Table 4. Since a lot of emphasis in this analysis is put on the distribution of the ligand conformations, a reduced number of analyzed poses impacts the analysis of the mutants. It is unknown how many poses are needed to get a reliable pose distribution, but fewer poses reduces the probability of getting a comprehensive view of the binding landscape. It is possible that removing the filtering based on the CNN score and just filtering the poses based on the uridine RMSD would increase the number of poses and improve the analysis. However, this goes back to the question discussed in section 5.1, if few higher scored poses or many low scored poses are better, which is difficult to know.

5.3 General discussion and future work

Generating and analyzing a large number of docked poses rather than only looking at one or a few poses can potentially give a better insight into the substrate binding, especially for an enzyme with a flexible substrate, such as GumK. Furthermore, looking at the distribution of the generated conformations could possibly be used as a way to screen for mutants of GumK with an altered specificity towards UDP-Glc (or another sugar). The workflow can by advantage be set up and run automatically with many mutants being screened in a short amount of time. Interesting mutants could this way be selected and further tested using experimental and/or more exact computational methods.

Although clustering and visualization of the representative poses give valuable insight into what conformations are mainly adopted, plotting attributes describing the conformation provides a faster way to easily get information about all the poses. Additionally, the clusters usually contain a smaller variation of poses. How similar the poses in a cluster are depends on the overall distribution. In some cases, the poses all represent about the same conformation and sometimes they differ more. Therefore, only looking at the representative pose might not always give the best overview of the distribution.

Ideally, the attributes selected for plotting the pose distribution should capture important structural differences, and be able to distinguish an active compound from an inactive. Here, the distance between the side chain of residue 307 and the carbon 6 of the sugar was calculated. Although a clear difference in terms of this distance was seen between UDP-GlcA and UDP-Glc when docking to the WT conformation 1, some mutants had a mutation at residue 307, making it difficult to use it for direct comparison. It was tested to use the distance to the α carbon of residue 307 instead (data not shown), but doing this did not display an as distinct difference between the two substrates. Furthermore, the dihedral angles given by the atoms

O^{β}-P^{β}-O^{1}-C^{1} and O^{β}-P^{β}-O^{1}-C^{1} were also used to describe the distribution of the poses. Although it seemed like a specific set of dihedral corresponded to a certain conformation for a given distance, a small change in the distance sometimes made the conformation very different, and in some cases very similar conformations could give significantly different dihedral angles, which complicated the analysis. Finding attributes that can describe the distribution of poses in a way that enables the observer to easily distinguish between an active and inactive compound is key for a fast analysis of the substrate binding of mutants. However, it might be difficult to find better parameters than the ones used, since the substrate can obtain many different conformations. An alternative is to utilize more parameters, but this of course also makes the analysis more complex and time consuming.

As already mentioned in section 5.2.2, another aspect reducing the possibility of comparing the results for the distribution of poses between the different structures is the fact that the number of poses often differs. For example, mutant 6 only has around half of UDP-GlcA poses compared to conformation 1 of the WT. A potentially improved way of the method would be to, instead of defining the number of docking reruns, decide the number of poses wanted for analysis. The docking could then be rerun until the desired amount of well-docked poses are reached, with a set number of maximum reruns to avoid time spent on docking to proteins that do not bind the substrate.

A downside with molecular docking that was also seen in this thesis is the fact that it is very sensitive to the structure of the receptor given. For example, utilizing a different but still very similar conformation during the ensemble docking did not only give lower docking scores, but also a major reduction in the number of poses with a well-docked uridine part. Since the mutants tested are folded via AlphaFold2, or mutated through PyMol, there is a degree of uncertainty regarding the structures that could impact the docking results. Furthermore, the conformation of the protein affects the distribution of the poses, which also needs to be considered when analyzing the mutants. Docking to an ensemble of conformations for the mutants as well would likely increase the possibility of getting reliable results.

Screening for mutants with the method used in this thesis builds on the assumption that the specificity, at least partially, depends on the conformations adopted by the donor substrate when bound to an open conformation of the protein. However, another possibility is that the specificity is associated with the substrate conformation acquired when the protein is in a more closed state. In this case, docking to only the donor domain will likely not be sufficient to draw a reliable conclusion about different mutants.

Considering the assumptions made and the fact that docking algorithms include many approximations that could affect the accuracy of the results, experimental data would be needed to conclude if this way of using molecular docking truly gives valuable insight and can be used to find interesting mutants.

6. Conclusion

In this thesis, the substrate binding to the wild type GumK donor domain and a number of mutants has been studied using molecular docking. Although the docking did not manage to distinguish between the active substrate UDP-GlcA and the inactive UDP-Glc in terms of the docking score (for the scoring functions tested), a difference in the distribution of conformations adapted was seen when generating an ensemble of docked poses. The main difference observed was that UDP-GlcA more often obtained poses with a shorter distance between the sugar and the residue K307.

Normal mode analysis revealed that the slowest motions of the donor domain involve the alpha helices around the binding site, giving a more open/closed binding pocket. Docking to more closed protein conformations resulted mainly in poses with a longer distance between the sugar and K307, likely due to the fact that the sugar could not fit that well into the binding pocket. A slightly more open conformation instead gave primarily poses with a shorter distance. Unexpectedly, the conformation most similar to the one initially tested did not dock the substrates well at all. This illustrates one of the major shortcomings of molecular docking, that the results depend a lot on the structure of the protein used during the docking.

From the results obtained for the WT, it was predicted that a mutant with a UDP-Glc pose distribution similar to that of the WT native substrate might have an altered substrate specificity towards UDP-Glc. If true, it could be used as a screening method to find interesting mutants that can catalyze the reaction with glucose instead. Although none of the mutants tested obtained this exact distribution, experimental data is needed to conclude the accuracy of the method. Furthermore, improvements, such as performing ensemble docking of the mutants would likely improve the reliability of the results.

7. References

1. Srivastava RK, Sushant P, Sathvik AS, Kolluru VC, Ahamad MI, Alharthi MA, et al. Sources and industrial applications of polysaccharides. In: Food, Medical, and Environmental Applications of Polysaccharides [Internet]. Elsevier; 2021 [cited 2024 May 1]. p. 511–30. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128192399000221

2. Petri DFS. Xanthan gum: A versatile biopolymer for biomedical and technological applications. J Appl Polym Sci. 2015 Jun 15;132(23):app.42035.

3. Barreras M, Salinas SR, Abdian PL, Kampel MA, Ielpi L. Structure and Mechanism of GumK, a Membrane-associated Glucuronosyltransferase. J Biol Chem. 2008 Sep;283(36):25027–35.

4. Du X, Li Y, Xia YL, Ai SM, Liang J, Sang P, et al. Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods. Int J Mol Sci. 2016 Jan 26;17(2):144.

5. Thomas E. Creighton, W. H. Freeman, New York, 1992. Proteins: Structures and molecular properties.

6. Braun E, Gilmer J, Mayes HB, Mobley DL, Monroe JI, Prasad S, et al. Best Practices for Foundations in Molecular Simulations [Article v1.0]. Living J Comput Mol Sci [Internet]. 2019 [cited 2024 Jan 22];1(1). Available from: https://www.livecomsjournal.org/article/5957-best-practices-for-foundations-in-molecular-simulations-article-v1-0

7. Kukol A, editor. Molecular Modeling of Proteins [Internet]. Totowa, NJ: Humana Press; 2008 [cited 2024 May 12]. (Walker JM, editor. Methods in Molecular Biology; vol. 443). Available from: http://link.springer.com/10.1007/978-1-59745-177-2

8. Meng XY, Zhang HX, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. Curr Comput Aided-Drug Des. 2011 Jun 1;7(2):146–57.

9. Molecular Docking - Recent Advances. [Elektronisk resurs] [Internet]. 2023. Available from: https://ludwig.lub.lu.se/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat07147a&AN=lub.7613381&site=eds-live&scope=site

10. Marjana Novič, Tjaša Tibaut, Marko Anderluh, Jure Borišek, Tihomir Tomašič. 2016. The Comparison of Docking Search Algorithms and Scoring Functions: An Overview and Case Studies. doi: 10.4018/978-1-5225-0115-2.ch004.

11. Vlachakis DP, editor. Molecular Docking [Internet]. InTech; 2018 [cited 2024 May 1]. Available from: http://www.intechopen.com/books/molecular-docking

12. Feinstein WP, Brylinski M. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. J Cheminformatics. 2015 Dec;7(1):18.

13. Meli R, Morris GM, Biggin PC. Scoring Functions for Protein-Ligand Binding Affinity Prediction Using Structure-based Deep Learning: A Review. Front Bioinforma. 2022 Jun 17;2:885983.

14. Yang C, Chen EA, Zhang Y. Protein–Ligand Docking in the Machine-Learning Era. Molecules. 2022 Jul 18;27(14):4568.

15. Lexa KW, Carlson HA. Protein flexibility in docking and surface mapping. Q Rev Biophys. 2012 Aug;45(3):301–43.

16. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: Structures, Functions, and Mechanisms. Annu Rev Biochem. 2008 Jun 1;77(1):521–55.

17. Rini JM, Moremen KW, Davis BG, et al. Glycosyltransferases and Glycan-Processing Enzymes.

18. Hanukoglu I. Proteopedia: Rossmann fold: A beta-alpha-beta fold at dinucleotide binding sites. Biochem Mol Biol Educ. 2015;43(3):206–9.

19. Chaturvedi S, Kulshrestha S, Bhardwaj K, Jangir R. A Review on Properties and Applications of Xanthan Gum. In: Vaishnav A, Choudhary DK, editors. Microbial Polymers [Internet]. Singapore: Springer Singapore; 2021 [cited 2024 May 1]. p. 87–107. Available from: https://link.springer.com/10.1007/978-981-16-0045-6_4

20. NEUROtiker. Structure of Xanthan gum [Internet]. 2008. Available from: https://en.wikipedia.org/wiki/Xanthan_gum#cite_note-carl-roth-1

21. Becker A, Katzen F, Pühler A, Ielpi L. Xanthan gum biosynthesis and application: a biochemical /genetic perspective. Appl Microbiol Biotechnol. 1998 Aug 27;50(2):145–52.

22. Salinas SR, Petruk AA, Brukman NG, Bianco MI, Jacobs M, Marti MA, et al. Binding of the substrate UDP-glucuronic acid induces conformational changes in the xanthan gum glucuronosyltransferase. Protein Eng Des Sel. 2016 Jun;29(6):197–207.

23. McNutt AT, Francoeur P, Aggarwal R, Masuda T, Meli R, Ragoza M, et al. GNINA 1.0: molecular docking with deep learning. J Cheminformatics. 2021 Dec;13(1):43.

24. Koes DR, Baumgartner MP, Camacho CJ. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. J Chem Inf Model. 2013 Aug 26;53(8):1893–904.

25. Bakan A, Meireles LM, Bahar I. *ProDy* : Protein Dynamics Inferred from Theory and Experiments. Bioinformatics. 2011 Jun 1;27(11):1575–7.

26. Jacob A. Bauer, Vladena Bauerová-Hlinková. Normal Mode Analysis: A Tool for Better Understanding Protein Flexibility and Dynamics with Application to Homology Models. In: Rafael Trindade Maia, Rômulo Maciel de Moraes Filho, Magnólia Campos, editors. Homology Molecular Modeling [Internet]. Rijeka: IntechOpen; 2020 [cited 2024 May 12]. p. Ch. 2. Available from: https://doi.org/10.5772/intechopen.94139

27. Kaynak BT, Zhang S, Bahar I, Doruker P. ClustENMD: efficient sampling of biomolecular conformational space at atomic resolution. Cowen L, editor. Bioinformatics. 2021 Nov 5;37(21):3956–8.

28. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.

29. RDKit: Open-source cheminformatics; http://www.rdkit.org.

30. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. J Cheminformatics. 2011 Dec;3(1):33.

31. Yikrazuul. Structure of UDP glucuronic acid [Internet]. 2014. Available from: https://en.wikipedia.org/wiki/Uridine_diphosphate_glucuronic_acid#/media/File:Uridine_diphosphate_glucuronic_acid.svg.

32. Yikrazuul. Structure of UDP-glucose [Internet]. 2014. Available from: https://en.wikipedia.org/wiki/Uridine_diphosphate_glucose#/media/File:UDP-Glucose.svg.

33. Yikrazuul. Structure of Guanosine Diphosphate Mannose [Internet]. 2014. Available from: https://en.wikipedia.org/wiki/Guanosine_diphosphate_mannose#/media/File:GDP-mannose.svg.

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

35. Gowers RJ, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In: Benthall S, Rostrup S, editors. Proceedings of the 15th Python in Science Conference. 2016. p. 98–105.

36. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. J Comput Chem. 2011;32(10):2319–27.

37. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, et al. UCSF ChimeraX : Tools for structure building and analysis. Protein Sci. 2023 Nov;32(11):e4792.

38. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.

39. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;17(3):261–72.

8. Appendix

Redocking of UDP and docking of UDP-GlcA to the crystal structure of GumK was done as described in section 4.1. The two settings *autobox* and *exhaustiveness* were varied. Four exhaustiveness levels, 4, 8, 16, and 32, and two different autobox sizes were tested. First, the UDP from the UDP bound GumK crystal structure was provided as an autobox, and then a larger manually selected autobox was used. The median and average scoring values from redocking of UDP are found in Table A1 and A2, where Table A1 includes all poses and Table A2 are calculated using only the top ranked poses. Furthermore, the median and average scoring values from docking of UDP-GlcA are found in Table A3 and A4. For each setting, the docking was rerun 10 times.

Table A1: Scoring data from docking UDP to the crystal structure using different exhaustiveness levels and two different sizes for the autobox. For each setting, the docking was rerun 10 times. The median and average score with standard deviation for all poses generated at the specified settings is listed, as well as the median and average RMSD compared to the crystal structure UDP.

Auto-box	**Exhaus-tiveness**	**CNN score**		**CNN affinity [pK]**		**Vina score [kcal/mol]**		**Vinardo score [kcal/mol]**		**RMSD [Å]**	
		Median	Average	Median	Average	Median	Average	Median	Average	Median	Average
UDP	4	0.49	0.52±0.18	4.5	4.58±0.27	−7.8	−7.84±0.94	−5.58	−5.54±1.12	2.02	2.48±1.7
UDP	8	0.52	0.56±0.18	4.57	4.63±0.29	−8.07	−8.16±0.63	−6.01	−5.92±0.97	1.48	2.37±2.15
UDP	16	0.68	0.69±0.13	5.09	5.11±0.2	−10.1	−10.13±0.49	−7.74	−7.89±0.95	1.02	1.93±1.74
UDP	32	0.67	0.67±0.15	4.77	4.78±0.25	−8.3	−8.43±0.59	−6.13	−6.16±0.95	0.86	1.6±1.64
Manual	4	0.54	0.58±0.16	4.04	4.12±0.58	−5.37	−6.08±1.69	−4.05	−4.47±1.41	14.7	13.42±8.2
Manual	8	0.47	0.51±0.18	4.44	4.42±0.45	−7.12	−7.33±1.15	−5.09	−5.27±1.09	7.38	8.82±7.26
Manual	16	0.48	0.54±0.18	4.59	4.56±0.38	−7.74	−7.85±0.82	−5.64	−5.83±0.88	5.24	8.01±6.95
Manual	32	0.55	0.57±0.18	4.66	4.68±0.28	−8.13	−8.19±0.61	−5.89	−5.94±0.94	3.36	4.66±4.84

Table A2: Scoring data for the highest ranked poses when docking UDP to the crystal structure using different exhaustiveness levels and two different sizes for the autobox. For each setting, the docking was rerun 10 times. The median and average score with standard deviation for the highest ranked poses generated at the specified settings is listed, as well as the median and average RMSD compared to the crystal structure UDP.

Auto-box	Exhaus-tiveness	CNN score		CNN affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]		RMSD [Å]	
		Median	Average	Median	Average	Median	Average	Median	Average	Median	Average
UDP	4	0.84	0.83±0.04	5	5.01±0.09	−9.14	−8.88±0.54	−6.84	−6.7±0.58	0.61	0.67±0.18
UDP	8	0.84	0.84±0.02	5.01	5.04±0.1	−9.34	−9.14±0.39	−7.34	−7.17±0.44	0.64	0.68±0.14
UDP	16	0.94	0.92±0.03	5.47	5.45±0.05	−10.78	−10.65±0.32	−9.38	−9.18±0.5	0.65	0.64±0.1
UDP	32	0.87	0.87±0.02	5.04	5.05±0.03	−9.49	−9.17±0.66	−7.46	−7.2±0.56	0.6	0.67±0.28
Manual	4	0.83	0.84±0.03	5.08	5.07±0.08	−8.8	−8.86±0.52	−6.93	−6.92±0.49	0.58	0.6±0.18
Manual	8	0.81	0.81±0.03	4.95	5.01±0.12	−9.13	−8.89±0.41	−6.85	−6.77±0.43	0.85	0.82±0.26
Manual	16	0.82	0.81±0.02	4.96	5.00±0.14	−9.01	−8.87±0.54	−7.17	−6.9±0.62	0.76	0.81±0.29
Manual	32	0.84	0.83±0.03	5.02	5.01±0.09	−9.28	−9.14±0.34	−7.38	−7.17±0.45	0.62	0.67±0.17

Table A3: Scoring data from docking UDP-GlcA to the crystal structure using different exhaustiveness levels and two different sizes for the autobox. For each setting, the docking was rerun 10 times. The median and average score with standard deviation for all poses generated at the specified settings is listed, as well as the median and average RMSD compared to the uridine part of the crystal structure UDP.

Autobox	Exhaus-tiveness	CNN score		CNN affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]		RMSD [Å]	
		Median	Average	Median	Average	Median	Average	Median	Average	Median	Average
UDP	4	0.64	0.67±0.14	5.05	5.08±0.19	−9.78	−9.6±1.18	−7.2	−7.19±1.61	1.95	2.85±2.16
UDP	8	0.67	0.69±0.14	5.15	5.13±0.19	−10.04	−10.02±0.57	−7.85	−7.87±0.95	1.59	2.32±1.5
UDP	16	0.68	0.69±0.13	5.09	5.11±0.2	−10.1	−10.13±0.49	-7.58	-7.83±1.03	1.42	2.04±1.64
UDP	32	0.76	0.74±0.13	5.18	5.17±0.18	−10.2	−10.22±0.42	−7.93	−8.03±0.83	1.33	1.76±1.45
Manual	4	0.5	0.54±0.13	4.63	4.64±0.37	−7.45	−7.87±1.81	−5.54	−5.96±1.36	12.98	12.86±8.13
Manual	8	0.51	0.54±0.12	4.91	4.91±0.24	−9.72	−9.39± 1.26	−7.12	−7.06±1.15	9.11	8.11±6.29
Manual	16	0.52	0.55±0.13	4.97	4.98±0.21	−10.12	−9.98±0.8	−7.85	−7.69±0.86	4.16	6.5±5.89
Manual	32	0.55	0.61±0.13	5.04	5.05±0.18	−10.2	−10.12±0.59	−7.88	−7.74±0.8	1.55	4.29±4.77

Table A4: Scoring data for the highest ranked poses when docking UDP-GlcA to the crystal structure using different exhaustiveness levels and two different sizes for the autobox. For each setting, the docking was rerun 10 times. The median and average score with standard deviation for the highest ranked poses generated at the specified settings is listed, as well as the median and average RMSD compared to the uridine part of the crystal structure UDP.

Auto-box	Exhaus-tiveness	CNN score		CNN affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]		RMSD [Å]	
		Median	Average	Median	Average	Median	Average	Median	Average	Median	Average
UDP	4	0.88	0.88±0.05	5.37	5.37±0.08	−10.18	−10.09±0.64	−7.98	−7.91±1.2	1.12	1.35±1.15
UDP	8	0.89	0.9±0.04	5.41	5.39±0.08	−10.48	−10.38±0.44	−8.87	−8.60±0.91	0.98	1.76±1.68
UDP	16	0.94	0.92±0.03	5.47	5.45±0.05	−10.78	−10.65±0.32	-8.42	−8.50±0.83	0.44	0.87±1.16
UDP	32	0.93	0.93±0.01	5.46	5.46±0.02	−10.77	−10.75±0.11	−9.44	−9.35±0.23	0.44	0.48±0.11
Manual	4	0.75	0.75±0.14	4.58	4.85±0.43	−6.96	−8.06±1.97	−6.06	−6.62±1.56	8.88	10.52±9.6
Manual	8	0.77	0.76±0.1	5.27	5.16±0.31	−10.07	−9.72±1.21	−7.98	−7.64±1.23	1.1	4.18±6.37
Manual	16	0.8	0.8±0.1	5.26	5.26±0.19	−10.26	−10.07±0.51	−8.18	−8.05±0.64	1.27	3.8±5.95
Manual	32	0.83	0.84±0.08	5.3	5.3±0.09	−10.07	−10.03±0.41	−7.96	−7.93±0.78	0.96	1.06±0.29

The CNN affinity and Vinardo score vs the RMSD for all poses generated when redocking UDP and docking UDP-GlcA to the crystal structure is plotted in Figure A1 and A2.

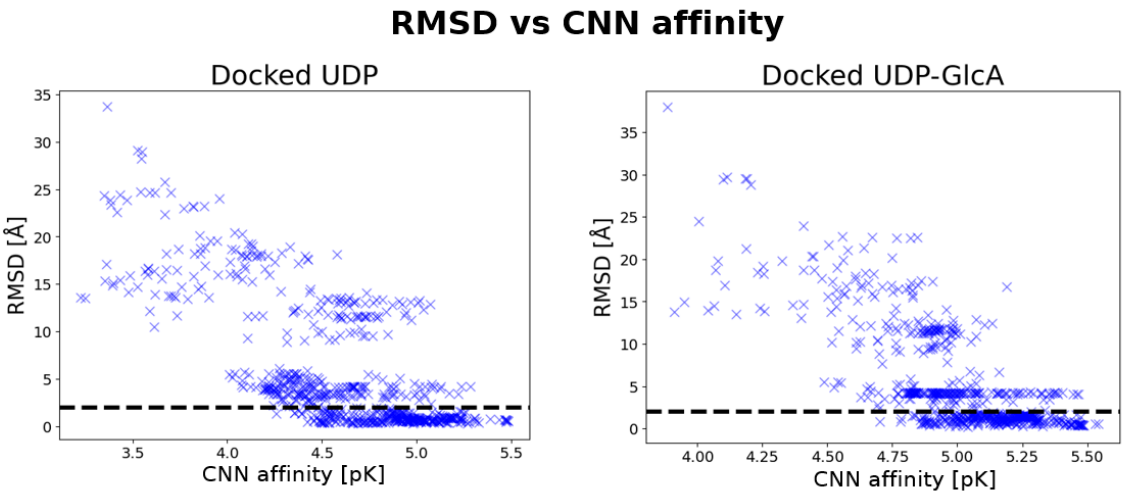

Figure A1: The CNN affinity of the poses generated from docking UDP and UDP-GlcA to the crystal structure GumK plotted against the RMSD. The RMSD for the docked UDP was calculated as the heavy-atom RMSD compared to the crystal structure UDP, whereas the RMSD for the docked UDP-GlcA was calculated as the heavy-atom RMSD compared to the uridine part of the crystal structure. The dotted line shows where the RMSD is equal to 2Å.

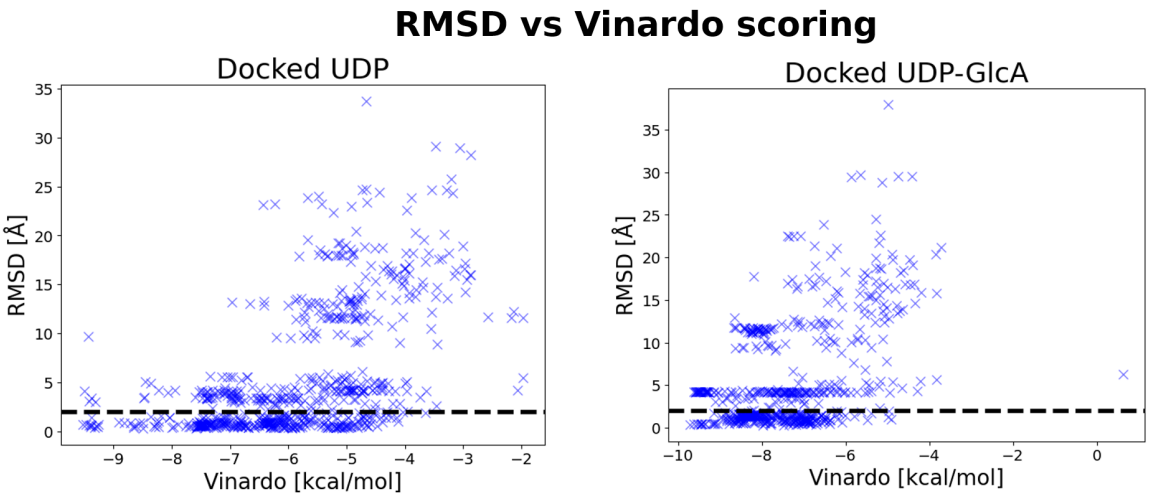

Figure A2: The Vinardo affinity score of the poses generated from docking UDP and UDP-GlcA to the crystal structure GumK plotted against the RMSD. The RMSD for the docked UDP was calculated as the heavy-atom RMSD compared to the crystal structure UDP, whereas the RMSD for the docked UDP-GlcA was calculated as the heavy-atom RMSD compared to the uridine part of the crystal structure. The dotted line shows where the RMSD is equal to 2Å.

Two different autobox sizes were used when redocking UDP and docking UDP-GlcA to the crystal structure. First the UDP from the crystal structure was provided as the autobox, then a larger autobox, manually picked in MGLTools was used. Thereafter, the autobox was given by the residues within 5Å of the UDP-GlcA pose with the highest CNN score from docking UDP-GlcA to the crystal structure. All of these autobox sizes are visualized in Figure A3.

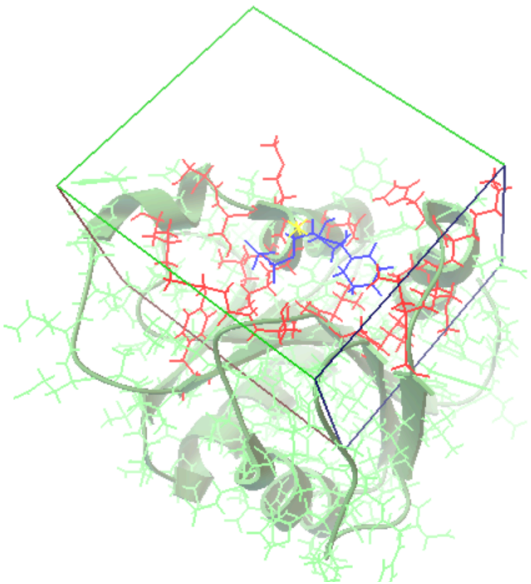

Figure A3: A visualization of the different autobox sizes used in the thesis. The rectangle box shows the autobox manually picked in MGLTools, and the UDP (colored blue) is the crystal structure UDP, both which was used as an autobox when docking to the crystal structure. The red colored amino acids show the residues used to define the autobox for all other dockings.

The three compounds UDP-GlcA, UDP-Glc and GDP-Man were docked to the donor domain of the WT GumK provided by PhD student Davide Luciano (generated through MD simulations), denoted conformation 1, as described in section 5.1. The average and median score for all poses are found in Table A5 and for the top ranked poses in Table A6.

Table A5: Scoring data for all poses from docking UDP-GlcA, UDP-Glc and GDP-Man to the crystal structure as described in section 4.1.

Substrate	CNN score		CNN affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]	
	Average	Median	Average	Median	Average	Median	Average	Median
UDP-GlcA	0.86±0.06	0.86	4.67±0.12	4.68	-7.57±0.56	-7.56	-6.73±0.69	-6.77
UDP-Glc	0.85±0.07	0.87	4.62±0.12	4.66	-7.51±0.57	-7.64	-6.79±0.87	-6.96
GDP-Man	0.47±0.07	0.46	4.42±0.11	4.43	-7.32±0.45	-7.38	-6.19±0.58	-6.21

Table A6: Scoring data for the top ranked poses from docking UDP-GlcA, UDP-Glc and GDP-Man to the crystal structure as described in section 4.1.

Substrate	CNN score		CNN affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]	
	Average	Median	Average	Median	Average	Median	Average	Median
UDP-GlcA	0.92±0.02	0.92	4.81±0.06	4.84	-7.59±0.53	-7.79	-7.15±0.39	-7.26
UDP-Glc	0.93±0.01	0.93	4.72±0.07	4.70	-7.90±0.27	-7.93	-7.17±0.74	-7.33
GDP-Man	0.58±0.05	0.57	4.48±0.09	4.48	-7.44±0.41	-7.49	-6.54±0.39	-6.67

As no big difference was seen between the substrates UDP-GlcA and UDP-Glc in regards to the docking scores, they were docked again to the donor domain WT conformation 1. However, this time with different docking settings, as described in section 4.2. First of all, the number of saved poses after each docking was increased to 20 instead of 9, and the limit for how similar two poses were allowed to be was lowered to an RMSD of 0.1 Å instead of 1Å. Table A7 lists the average and median scoring for the generated UDP-GlcA and UDP-Glc poses. After filtering the poses to only consider poses with a CNN score higher than the median CNN score and a uridine RMSD lower than 2Å, the average and median scores were calculated again and can be found in Table A8.

Table A7: Scoring data for all poses from docking UDP-GlcA and UDP-Glc to WT donor domain conformation 1 as described in section 4.2.

Substrate	CNN score		CNN affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]	
	Average	Median	Average	Median	Average	Median	Average	Median
UDP-GlcA	0.73±0.14	0.75	4.52±0.20	4.55	-7.50±0.63	-7.58	-6.57±0.84	-6.67
UDP-Glc	0.75±0.14	0.79	4.48±0.21	4.49	-7.38±0.61	-7.49	-6.53±0.86	-6.66
GDP-Man	0.39±0.11	0.37	4.33±0.16	4.34	-7.21±0.45	-7.23	-6.00±0.55	-5.96

Table A8: The number filtered poses and their average docking score (with standard deviation) from docking UDP-GlcA and UDP-Glc to WT donor domain conformation 1 as described in section 5.2.

Substrate	Number filtered poses	CNN score	CNN affinity [pK]	Vina score [kcal/mol]	Vinardo score [kcal/mol]
UDP-GlcA	84 (42%)	0.86±0.05	4.69±0.11	-7.62±0.58	-6.84±0.8
UDP-Glc	90 (45%)	0.89±0.04	4.66±0.11	-7.54±0.42	-6.87±0.75

The dihedral angles α and β, given by the atoms O^{β}-P^{β}-O^{1}-C^{1} and O^{β}-P^{β}-O^{1}-C^{1} respectively, were calculated for the docked poses to describe the orientation of the sugar ring. Furthermore, the normal vector given by the sugar ring plane, and the projection of this vector on the x, y and z-axis, was calculated for the UDP-GlcA docked to the WT conformation 1. As this should uniformly represent a single orientation of the ring, comparing it to the dihedral angles can give an indication of how well the dihedral angles represent the ring orientation. Table A9 lists the dihedral angles and normal vector projections for the filtered UDP-GlcA poses from docking to WT conformation 1.

Table A9: All the dihedral angles, α and β, for the filtered UDP-GlcA poses from docking to WT conformation 1 is listed in the two tables below. Furthermore, the distance between the C6 atom of the sugar and the nitrogen of K307, and the projection of the normal vector given by the sugar ring on the x, y, and z-axis is listed (under column "x", "y", and "z").

x	y	z	Distance [Å]	α [degrees]	β [degrees]
0.9	-0.3	-0.3	6.1	13	84
0.9	0.1	-0.4	5.5	35	90
0.9	-0.5	0.2	9.2	54	100
0.9	0.3	-0.4	5.3	-77	123
0.8	0.2	0.5	9.4	1	76
0.8	0.2	0.5	9.6	-3	85
0.8	0.2	-0.5	5.0	-62	115
0.8	0.2	0.5	9.8	-8	93
0.8	0.0	0.6	11.7	18	127
0.8	-0.2	-0.6	9.3	148	103
0.8	0.4	0.5	11.1	-12	145
0.8	0.0	0.7	9.8	-23	119
0.8	0.1	-0.7	4.5	-29	113
0.8	0.3	0.6	9.8	-16	94
0.7	0.0	-0.7	9.9	120	145
0.7	-0.1	-0.7	4.9	-17	118
0.7	-0.3	-0.7	9.3	66	94
0.6	0.3	-0.7	8.3	143	117
0.6	-0.3	-0.7	6.8	-30	106
0.6	-0.2	-0.8	5.2	-73	94
0.6	-0.6	-0.5	7.5	63	85
0.6	0.4	-0.7	4.8	62	95
0.6	-0.8	-0.1	8.8	80	103
0.6	-0.8	0.0	8.4	-44	116
0.6	-0.8	-0.1	8.8	81	109
0.5	-0.8	-0.1	9.1	86	115
0.5	0.5	0.7	11.7	86	114
0.5	-0.3	-0.8	6.2	-85	80
0.5	-0.3	-0.8	6.2	-84	76
0.5	-0.4	-0.8	6.3	-98	99
0.5	-0.3	-0.8	6.3	-86	77
0.5	-0.4	0.8	11.7	-80	121
0.5	-0.2	-0.9	5.8	-93	81
0.4	-0.8	-0.4	7.7	-98	106
0.4	-0.9	-0.1	10.0	109	128
0.3	0.1	0.9	11.2	-107	117
0.3	0.9	0.0	4.5	23	94
0.3	1.0	0.0	4.9	17	105
0.2	-1.0	0.2	10.5	-10	139
0.2	0.3	-0.9	5.5	-88	72
0.1	-1.0	0.1	10.5	-17	121
0.0	-1.0	0.2	9.1	-98	106
0.0	0.3	-1.0	5.1	-128	97

x	y	z	Distance [Å]	α [degrees]	β [degrees]
0.0	0.8	-0.5	5.6	97	115
0.0	0.9	-0.4	5.6	113	117
0.0	0.8	-0.6	5.5	104	104
0.0	-0.9	-0.4	7.9	94	119
0.0	0.9	-0.5	5.5	108	104
-0.1	0.2	-1.0	5.9	101	124
-0.1	0.8	-0.6	5.6	101	105
-0.1	0.8	-0.5	5.6	108	107
-0.1	-0.9	-0.4	10.5	-104	108
-0.1	0.8	-0.6	5.7	153	130
-0.1	0.8	-0.5	5.7	121	95
-0.1	0.8	-0.5	5.9	130	140
-0.1	-0.7	-0.7	10.1	125	119
-0.2	-0.8	-0.5	9.3	-104	127
-0.2	0.6	-0.8	6.1	-96	133
-0.2	-1.0	0.0	9.6	148	103
-0.2	0.6	-0.8	6.2	-99	133
-0.2	0.5	-0.9	6.3	-115	142
-0.2	0.5	-0.9	6.4	-117	140
-0.2	0.5	-0.8	6.3	-107	133
-0.2	0.6	-0.8	7.9	131	90
-0.2	0.6	-0.8	7.9	123	85
-0.2	0.4	-0.9	6.2	-104	109
-0.3	0.5	-0.8	6.1	-102	93
-0.3	0.5	-0.8	6.1	-104	98
-0.3	0.8	-0.6	8.1	71	111
-0.3	-0.5	-0.8	9.5	-155	126
-0.4	-0.7	-0.6	10.9	-103	104
-0.7	-0.2	0.7	7.7	-122	86
-0.7	0.5	-0.5	8.4	120	111
-0.7	-0.2	0.7	8.0	-125	94
-0.7	0.5	-0.5	8.1	80	111
-0.7	-0.3	0.7	7.9	-130	98
-0.7	0.0	0.7	7.7	64	73
-0.7	-0.6	0.4	9.7	35	102
-0.8	0.0	0.6	7.6	-13	85
-0.8	-0.4	0.4	8.5	-124	100
-0.9	0.1	0.5	7.8	-29	113
-0.9	0.3	0.3	7.5	-86	110
-0.9	-0.4	0.0	7.4	-71	92
-1.0	0.0	0.2	8.5	-20	121
-1.0	-0.1	-0.1	7.7	-52	111

An ensemble of six additional rigid protein conformations were generated using MD simulation and ClustENMD as described in section 3.4.2. UDP, UDP-GlcA and UDP-Glc were docked to the protein ensemble, see Table A10 for the average and median docking scores for all the poses obtained. A new

reference UDP from the docked UDP poses was selected for each conformation and is visualized in Figure A4. The average and median docking score of the docked UDP poses with a uridine RMSD < 2Å compared to the reference UDP for that conformation is found in Table A11, and the docking score for the UDP-GlcA and UDP-Glc poses filtered on both the uridine RMSD compared to the reference UDP and the CNN score is found in Table A12.

Table A10: Scoring data for all poses from docking UDP, UDP-GlcA and UDP-Glc to WT conformation 2-7 as described in section 5.2.

Substrate	Conformation	**CNN score**		**CNN affinity [pK]**		**Vina score [kcal/mol]**		**Vinardo score [kcal/mol]**	
		Average	Median	Average	Median	Average	Median	Average	Median
UDP	1	0.65±0.20	0.66	4.08±0.36	4.08	-6.21±0.84	-6.11	-4.09±1.03	-4.79
UDP-GlcA	2	0.45±0.10	0.45	4.22±0.25	4.28	-7.13±0.54	-7.17	-6.22±0.68	-6.21
UDP-Glc	2	0.44±0.12	0.42	4.16±0.25	4.19	-7.00±0.57	-7.05	-6.18±0.72	-6.20
UDP	2	0.47±0.10	0.48	3.76±0.21	3.77	-6.00±0.81	-6.00	-4.78±0.67	-4.87
UDP-GlcA	3	0.50±0.15	0.46	4.2±0.19	4.20	-6.75±0.56	-6.69	-5.73±0.77	-5.70
UDP-Glc	3	0.50±0.15	0.43	4.25±0.20	4.15	-6.64±0.52	-6.56	-5.62±0.74	-5.59
UDP	3	0.54±0.14	0.50	3.82±0.26	3.78	-5.59±0.53	-5.52	-4.42±0.69	-4.36
UDP-GlcA	4	0.52±0.11	0.48	4.27±0.15	4.27	-7.31±0.53	-7.30	-5.91±0.76	-5.94
UDP-Glc	4	0.50±0.13	0.45	4.20±0.17	4.20	-7.15±0.51	-7.14	-5.74±0.73	-5.73
UDP	4	0.51±0.11	0.51	3.87±0.19	3.89	-5.96±0.57	-5.94	-4.44±0.62	-4.43
UDP-GlcA	5	0.43±0.14	0.39	4.19±0.22	4.17	-7.54±0.59	-7.54	-6.38±0.76	-6.34
UDP-Glc	5	0.45±0.13	0.42	4.16±0.24	4.14	-7.50±0.57	-7.38	-6.37±0.78	-6.30
UDP	5	0.41±0.16	0.37	3.92±0.28	3.91	-6.32±0.71	-6.24	-4.80±0.68	-4.79
UDP-GlcA	6	0.51±0.12	0.48	4.23±0.20	4.22	7.31±0.48	-7.31	6.32±0.70	-6.33
UDP-Glc	6	0.46±0.12	0.44	4.14±0.20	4.13	-7.17±0.50	-7.12	6.18±0.70	-6.10
UDP	6	0.45±0.14	0.40	3.79±0.20	3.77	-6.14±0.61	-6.04	-4.84±0.75	-4.87
UDP-GlcA	7	0.46±0.11	0.45	4.24±0.21	4.23	-7.50±0.59	-7.52	-6.32±0.80	-6.28
UDP-Glc	7	0.45±0.11	0.42	4.18±0.21	4.17	-7.34±0.55	-7.34	-6.26±0.75	-6.23
UDP	7	0.45±0.13	0.42	3.94±0.25	3.95	-6.15±0.70	-5.98	-4.62±0.75	-4.59

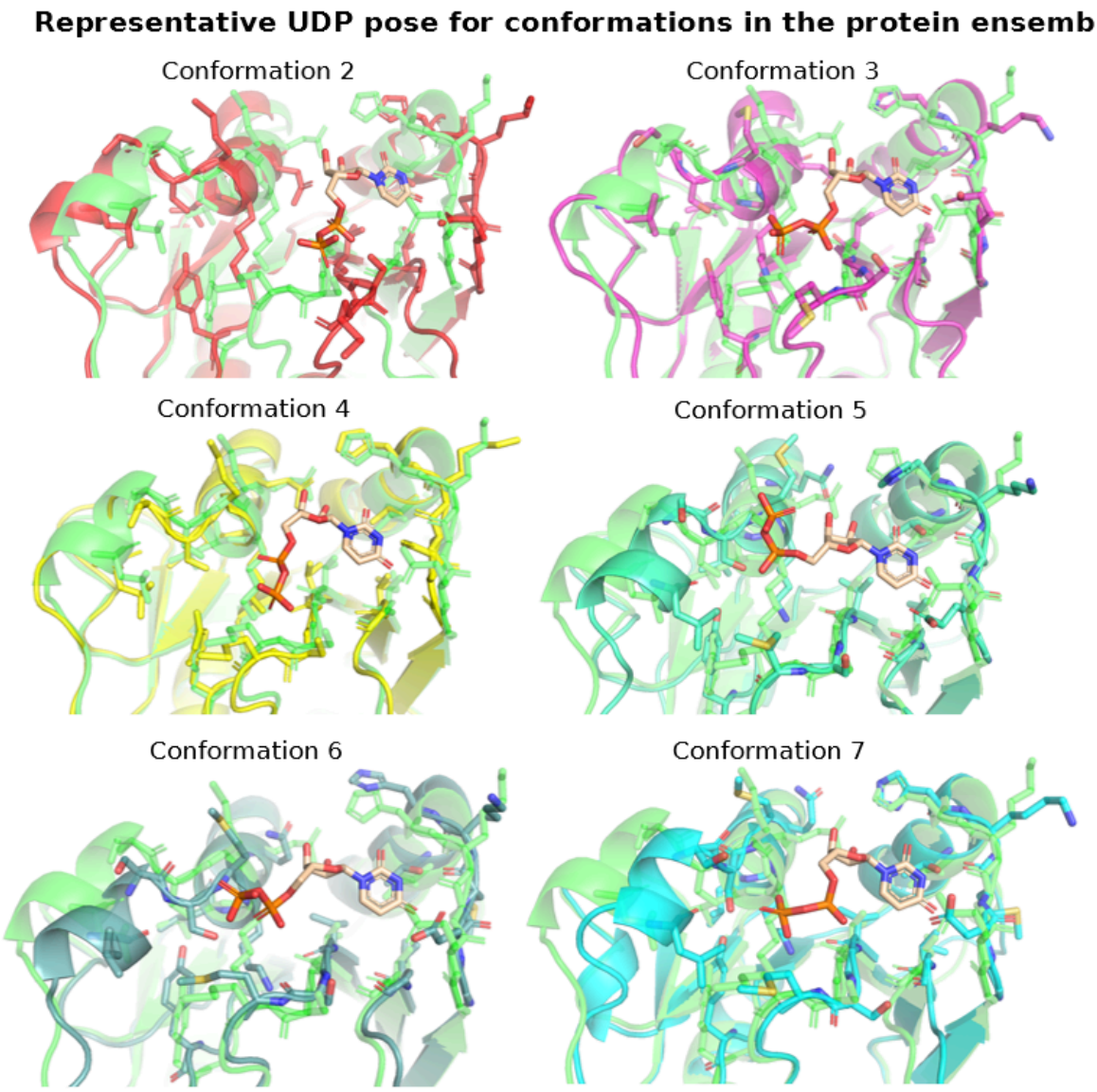

Figure A4: The representative UDP poses for the different protein conformations used to calculate the RMSD of the uridine part for the other poses. Conformation 1 is shown in green in all of the pictures.

Table A11: Average docking score with standard deviation for the UDP poses with an uridine RMSD < 2Å compared to the reference UDP for that conformation, from docking to WT conformation 2-7 as described in section 5.2.

Ligand	Receptor	CNN affinity	Vina score [kcal/mol]	Vinardo score [kcal/mol]
UDP	conformation 1	4.34±0.29	-6.60±0.84	-5.34±1.06
UDP	conformation 2	3.84±0.04	-5.23±0.17	-4.36±0.52
UDP	conformation 3	4.09±0.14	-5.38±0.31	-3.84±0.42
UDP	conformation 4	4.03±0.11	-6.18±0.54	-4.72±0.72
UDP	conformation 5	4.18±0.19	-6.92±0.60	-5.26±0.60
UDP	conformation 6	3.99±0.18	-6.62±0.67	-5.25±0.95
UDP	conformation 7	4.16±0.17	-6.86±0.64	-5.22±0.68

Table A12: Average docking score with standard deviation for the filtered poses from docking UDP-GlcA and UDP-Glc to WT conformation 2-7 as described in section 5.2.

Substrate	Conformation	CNN score	CNN affinity [pK]	Vina score [kcal/mol]	Vinardo score [kcal/mol]
UDP-GlcA	2	NA	NA	NA	NA
UDP-Glc	2	NA	NA	NA	NA
UDP-GlcA	3	0.65±0.08	4.44±0.14	-6.89±0.45	-6.89±0.45
UDP-Glc	3	0.64±0.08	4.44±0.10	-6.56±0.30	-5.06±0.50
UDP-GlcA	4	0.53±0.04	4.41±0.01	-7.31±0.55	-5.58±0.73
UDP-Glc	4	0.59±0.10	4.40±0.09	-6.92±0.46	-5.32±0.59
UDP-GlcA	5	0.60±0.16	4.48±0.19	-7.34±0.53	-6.29±0.75
UDP-Glc	5	0.63±0.13	4.49±0.15	-7.51±0.56	-6.54±0.72
UDP-GlcA	6	0.62±0.11	4.43±0.15	-7.31±0.40	-6.36±0.61
UDP-Glc	6	0.58±0.09	4.33±0.13	-7.24±0.35	-6.32±0.57
UDP-GlcA	7	0.56±0.08	4.46±0.12	-7.72±0.58	-6.84±0.86
UDP-Glc	7	0.56±0.07	4.38±0.13	-7.57±0.52	-6.63±0.77

The distance between the C6 atom of the sugar and the nitrogen on the side chain of K307 when docking UDP-GlcA and UDP-Glc to the WT protein ensemble was calculated. Figure A5 shows the density plots of this distance vs the CNN score for the filtered poses. Figure A6-A7 shows heatmaps of the pairwise RMSD for the filtered UDP-GlcA and UDP-Glc poses for conformation 3,5,6 and 7. Conformation 2 and 4 is excluded since they contained no or very few (3-5) poses. The heatmaps also include the corresponding dendrogram showing the distance between the poses calculated according to Eq. 8. Clustering of the filtered poses from docking to conformation 3,5,6 and 7 was done as described in 3.4.4. The representative poses for the largest clusters are shown in Figure A8-A12 (and Figure 23-24 in the result section).

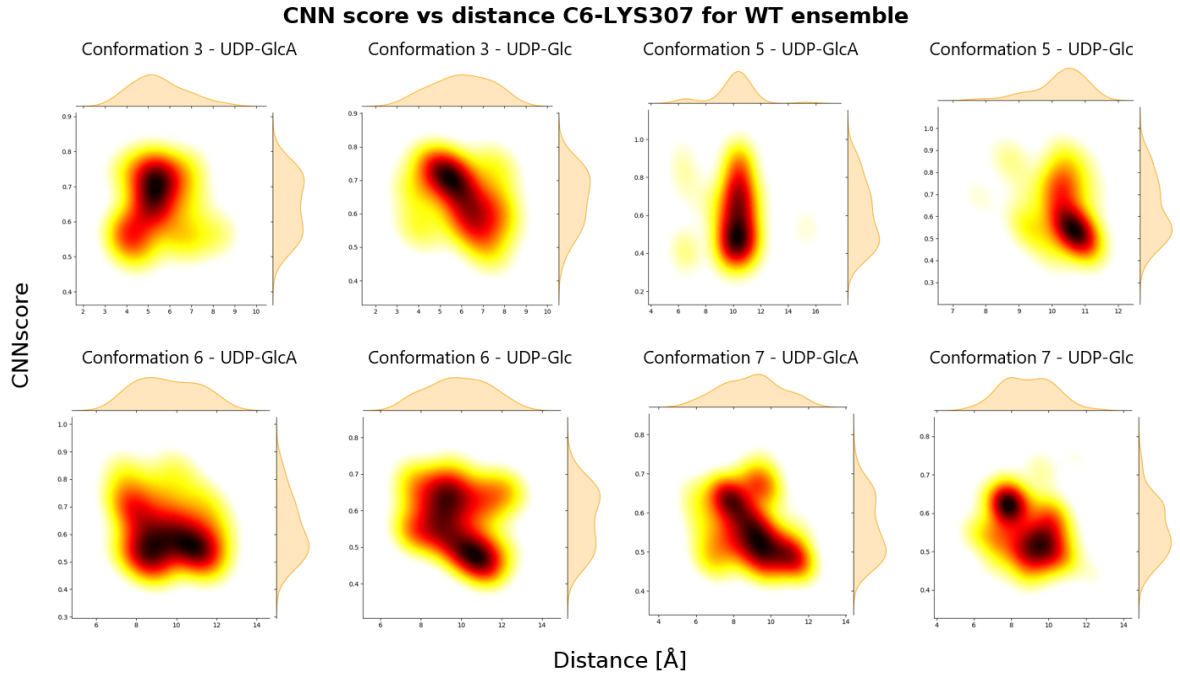

Figure A5: Density plots of the CNN score vs the K307-C6 distance for the filtered UDP-GlcA and UDP-Glc poses after docking to WT conformations 3, 5, 6 and 7.

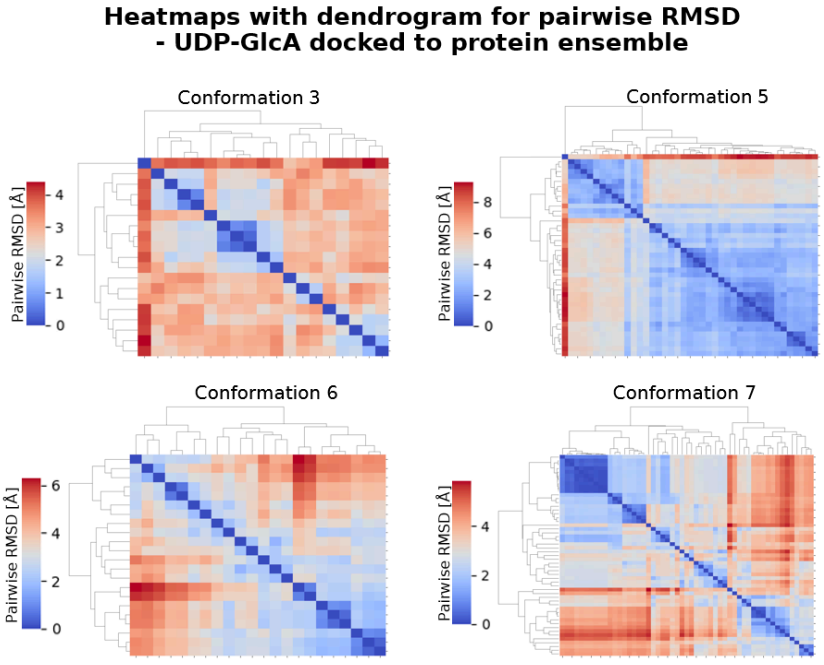

Figure A6: Heatmaps with the pairwise heavy-atom RMSD for the UDP-GlcA poses from docking to the WT conformations 3, 5, 6 and 7. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

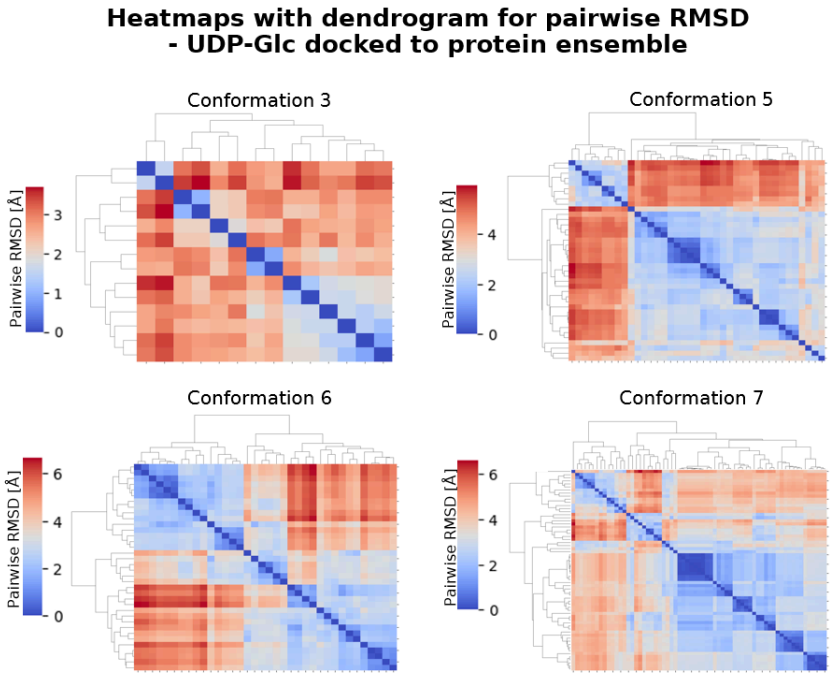

Figure A7: Heatmaps with the pairwise heavy-atom RMSD for the UDP-Glc poses from docking to the WT conformations 3, 5, 6 and 7. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

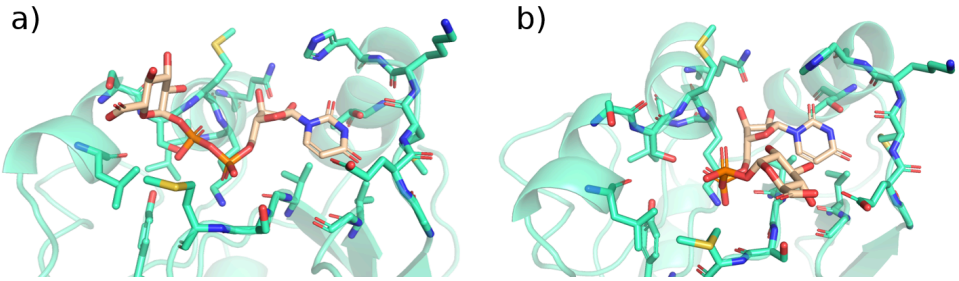

Figure A8: The representative poses for the two biggest clusters when docking UDP-GlcA to WT conformation 5. 54% of the poses belong to the cluster in a) and 23% belong to the cluster b).

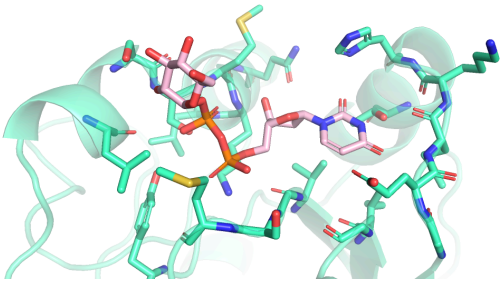

Figure A9: The representative poses for the largest clusters, containing 64% of the poses, from docking UDP-Glc to WT conformation 5.

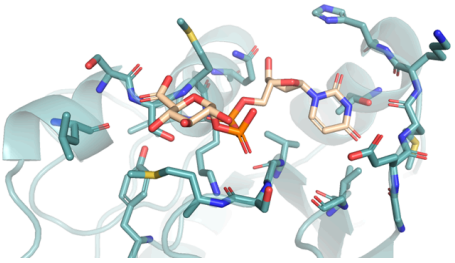

Figure A10: The representative poses for the largest clusters, containing 68% of the poses, from docking UDP-Glc to WT conformation 6.

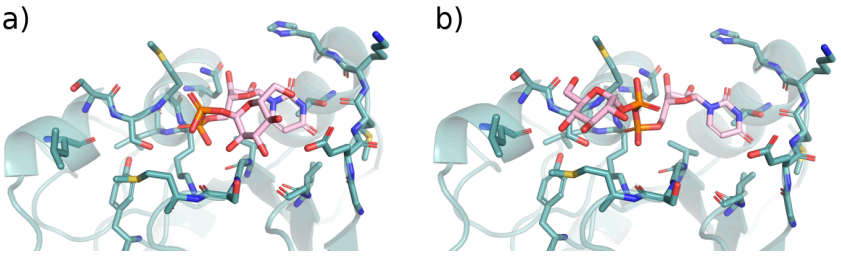

Figure A11: The representative poses for the two biggest clusters when docking UDP-Glc to WT conformation 6. Both clusters shown in a) and b) contain 42% of the poses each.

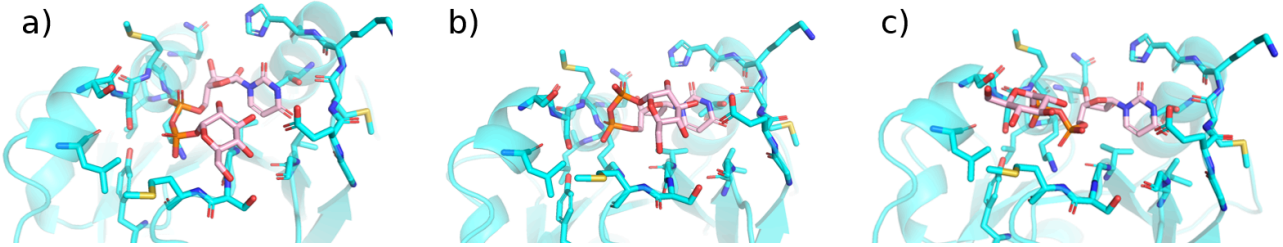

Figure A12: The representative poses for the three biggest clusters when docking UDP-Glc to WT conformation 7. 22% of the poses belong to a), 17% belong to b), and 12% belong to c).

Additionally to the WT conformations, six different mutants were docked to. A list of the different mutations of the mutants is found in Table 1. The coordinates for mutant 1 and 6 were generated using AlphaFold2, and a visualization of the conformation for these compared to the WT conformation 1 can be seen in Figure A13. The mutants are colored in blue, except for the mutated residues that are highlighted in orange, while the WT conformation 1 is colored in green.

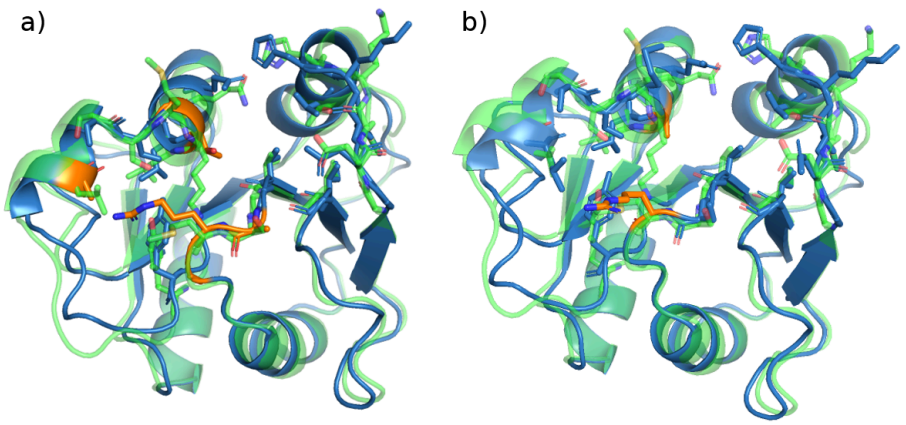

Figure A13: The structure of the mutants folded with AlphaFold2, with the mutant shown in blue and mutated residues colored in orange. The structure of conformation 1 of the WT is shown in green. The residues in the binding site are visualized as sticks. a) shows mutant 1 and b) shows mutant 6.

The median and average scores for all the UDP-GlcA and UDP-Glc poses from docking to the six mutants are found in Table A13. The poses were then filtered based on their uridine RMSD compared to the crystal structure UDP, and the median and average score for the filtered poses is found in Table A14. The pairwise RMSD for the filtered UDP-GlcA and UDP-Glc poses was calculated for each mutant and is visualized as heatmaps, see Figure A14-A18. Figure A14-A18 also includes the corresponding dendrogram showing the distance between the poses calculated according to Eq. 8. Clustering of the filtered poses was also done as described in 3.4.4. The representative poses for the largest clusters are shown in Figure A19-A26 (and Figure 23-27 in the result section).

Due to the low score obtained for mutant 1, and because the side chain of R218 was more oriented towards the binding pocket compared to WT conformation 1, see Figure A13, mutant 1 was docked again but with a flexible R218 side chain. The scores from this docking are found in Table A13 and A14. Figure A27 shows the heatmap and corresponding dendrogram of the pairwise RMSD for the filtered poses, and Figure A30-31 shows the representative poses from clustering of the filtered poses. Furthermore, density plots of the distance for C6 atom of the sugar and the oxygen of T307 vs the CNN score is found in Figure A28 and plots of the dihedral angles are shown in Figure A29.

Table A13: Scoring data for all poses from docking UDP-GlcA and UDP-Glc to the 6 mutants. Each substrate was docked to mutant 1 twice, once with the protein fully rigid and once with the side chain of R218 flexible.

Mutant	Ligand	Cnn score		Cnn affinity [pK]		Vina score [kcal/mol]		Vinardo score [kcal/mol]	
		Median	Average	Median	Average	Median	Average	Median	Average
1	UDP-GlcA	0.35	0.42±0.17	4.04	4.14±0.32	-7.57	-7.65±0.76	-6.73	-6.71±1.0
1	UDP-Glc	0.36	0.44±0.19	4.02	4.13±0.33	-7.58	-7.65±0.65	-6.67	-6.74±0.81
1 (flexible side chain)	UDP-GlcA	0.61	0.58±0.13	4.63	4.62±0.22	-8.67	-8.63±0.51	-	24.23±20.81
1 (flexible side chain)	UDP-Glc	0.59	0.59±0.14	4.62	4.58±0.23	-8.53	-8.51±0.5	-	24.24±20.55
2	UDP-GlcA	0.57	0.6±0.16	4.39	4.39±0.27	-7.34	-7.58±0.91	-6.26	-6.37±0.89
2	UDP-Glc	0.56	0.58±0.16	4.34	4.34±0.27	-7.29	-7.51±0.95	-6.24	-6.33±0.95
3	UDP-GlcA	0.74	0.73±0.14	4.5	4.48±0.22	-7.31	-7.29±0.63	-6.39	-6.28±0.82
3	UDP-Glc	0.7	0.73±0.14	4.42	4.43±0.22	-7.11	-7.09±0.56	-6.22	-6.17±0.79
4	UDP-GlcA	0.58	0.63±0.16	4.29	4.35±0.24	-7.47	-7.44±0.67	-6.02	-6.11±0.82
4	UDP-Glc	0.70	0.69±0.16	4.37	4.39±0.24	-7.3	-7.29±0.57	-6.12	-6.13±0.76
5	UDP-GlcA	0.76	0.75±0.13	4.53	4.52±0.2	-7.18	-7.26±0.59	-6.28	-6.21±0.76
5	UDP-Glc	0.75	0.73±0.15	4.46	4.44±0.22	-7.17	-7.16±0.51	-6.26	-6.22±0.75
6	UDP-GlcA	0.37	0.43±0.17	4.15	4.19±0.25	-7.83	-7.84±0.69	-6.5	-6.59±0.89
6	UDP-Glc	0.4	0.48±0.19	4.15	4.24±0.29	-7.68	-7.73±0.56	-6.53	-6.52±0.76

Table A14: The number of filtered poses and their average docking score with standard deviation from docking UDP-GlcA and UDP-Glc to the 6 mutants. Each substrate was docked to mutant 1 twice, once with the protein fully rigid and once with the side chain of R218 flexible.

Mutant	Number filtered poses	Substrate	Cnn score	Cnn affinity [pK]	Vina score [kcal/mol]	Vinardo score [kcal/mol]
1	53	UDP-GlcA	0.67±0.11	4.61±0.15	-8.04±0.72	-7.11±0.93
1	54	UDP-Glc	0.72±0.10	4.62±0.13	-8.03±0.55	-7.11±0.67
1 (flexible side chain)	40	UDP-GlcA	0.71±0.07	4.83±0.13	-8.64±0.42	-
1 (flexible side chain)	40	UDP-Glc	0.72±0.08	4.82±0.12	-8.72±0.43	-
2	61	UDP-GlcA	0.78±0.09	4.64±0.20	-7.64±0.99	-6.44±0.82
2	59	UDP-Glc	0.77±0.09	4.62±0.22	-7.62±1.06	-6.52±0.92
3	86	UDP-GlcA	0.87±0.05	4.67±0.11	-7.33±0.63	-6.47±0.76
3	81	UDP-Glc	0.87±0.05	4.63±0.12	-7.18±0.51	-6.42±0.69
4	62	UDP-GlcA	0.82±0.07	4.62±0.13	-7.65±0.77	-6.64±0.73
4	80	UDP-Glc	0.85±0.05	4.62±0.10	-7.33±0.68	-6.38±0.74
5	88	UDP-GlcA	0.87±0.05	4.68±0.11	-7.30±0.56	-6.37±0.65
5	80	UDP-Glc	0.87±0.04	4.64±0.11	-7.30±0.46	-6.52±0.60
6	40	UDP-GlcA	0.70±0.12	4.57±0.19	-7.53±0.48	-6.32±0.62
6	52	UDP-Glc	0.73±0.13	4.62±0.21	-7.53±0.46	-6.35±0.56

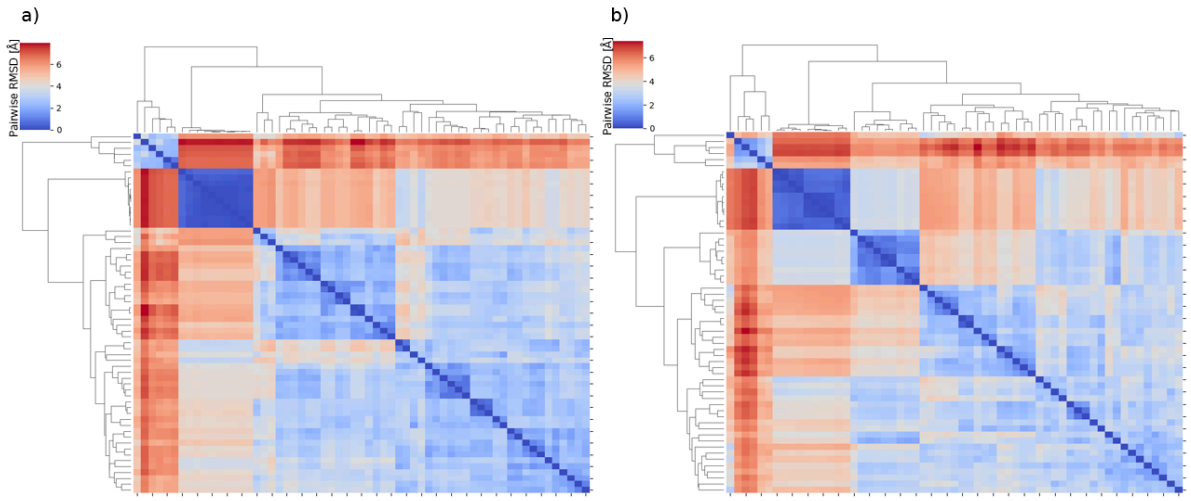

Figure A14: Heatmaps with the pairwise heavy-atom RMSD for the filtered UDP-GlcA and UDP-GlcA poses docked to mutant 2. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

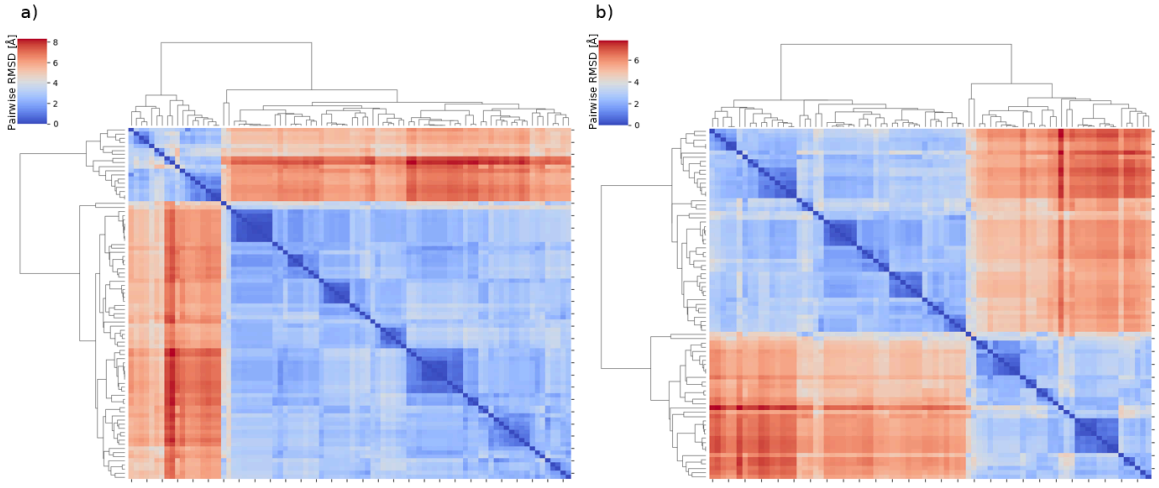

Figure A15: Heatmaps with the pairwise heavy-atom RMSD for the filtered UDP-GlcA and UDP-GlcA poses docked to mutant 3. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

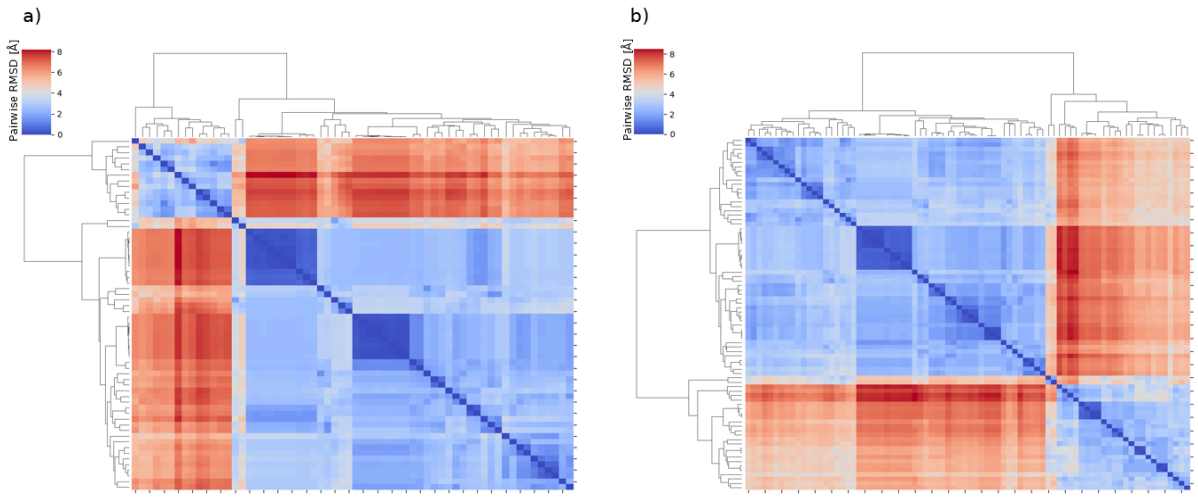

Figure A16: Heatmaps with the pairwise heavy-atom RMSD for the filtered UDP-GlcA and UDP-GlcA poses docked to mutant 4. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

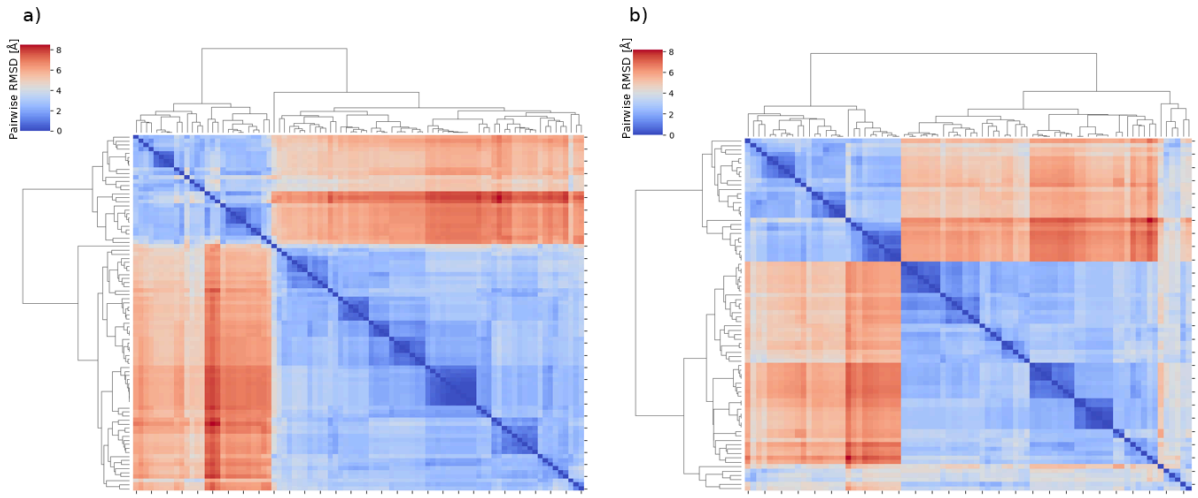

Figure A17: Heatmaps with the pairwise heavy-atom RMSD for the filtered UDP-GlcA and UDP-GlcA poses docked to mutant 5. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

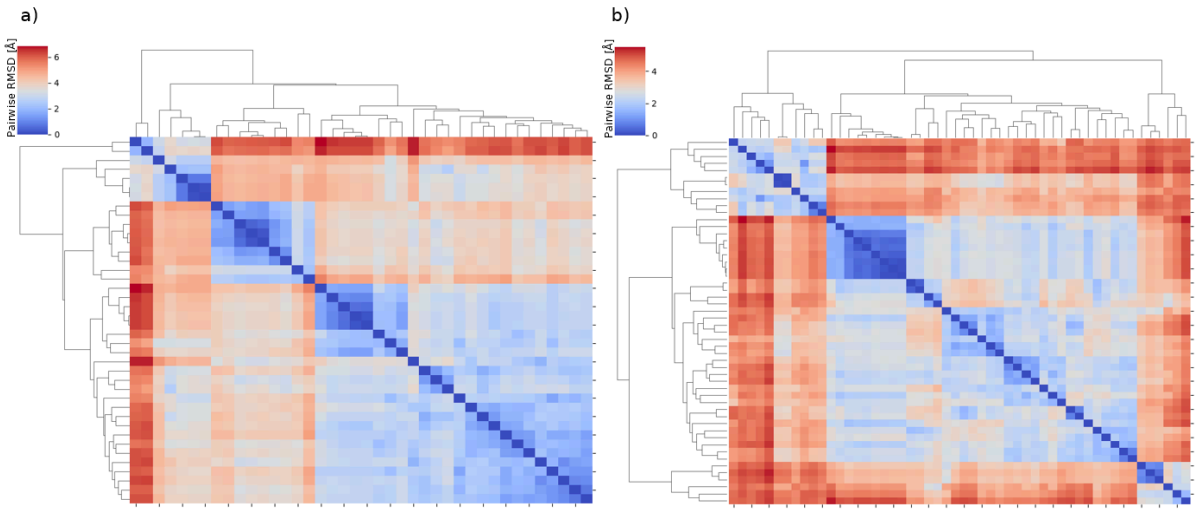

Figure A18: Heatmaps with the pairwise heavy-atom RMSD for the filtered UDP-GlcA and UDP-GlcA poses docked to mutant 6. The dendrogram beside the heatmap shows the distance between the poses calculated with *scipy* using the "average" linkage method.

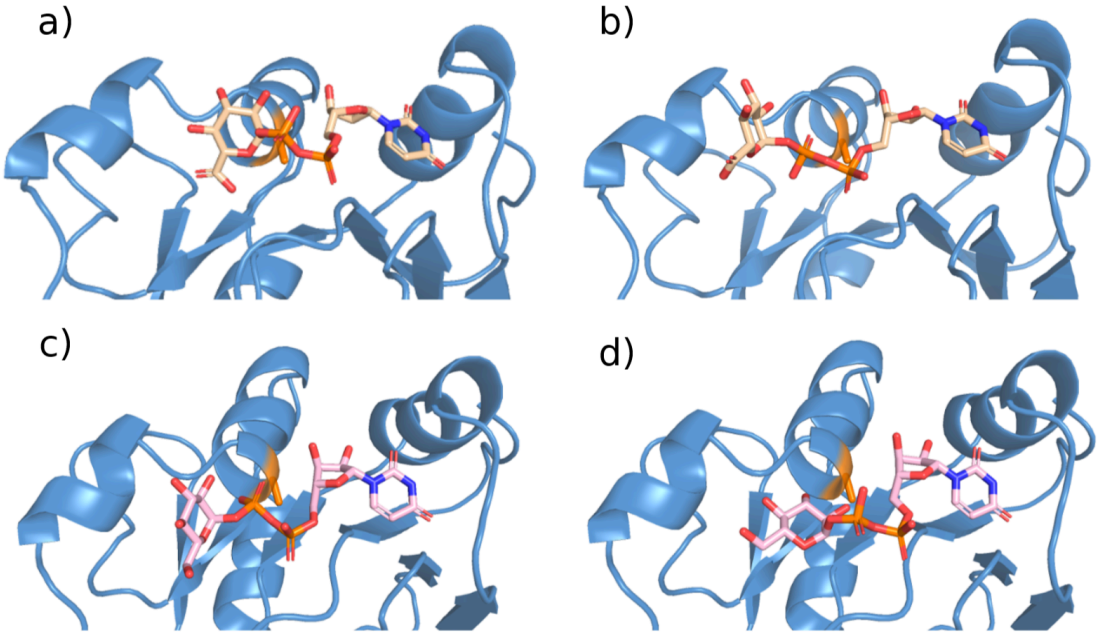

Figure A19: The representative poses for the two biggest clusters when docking UDP-GlcA and UDP-Glc to mutant 2. 36% of the docked UDP-GlcA belong to the cluster represented by a) and 26% belong to b). 25% of the docked UDP-Glc belong to the cluster represented in c) and another 25% belong to d). The mutated residue is highlighted in orange.

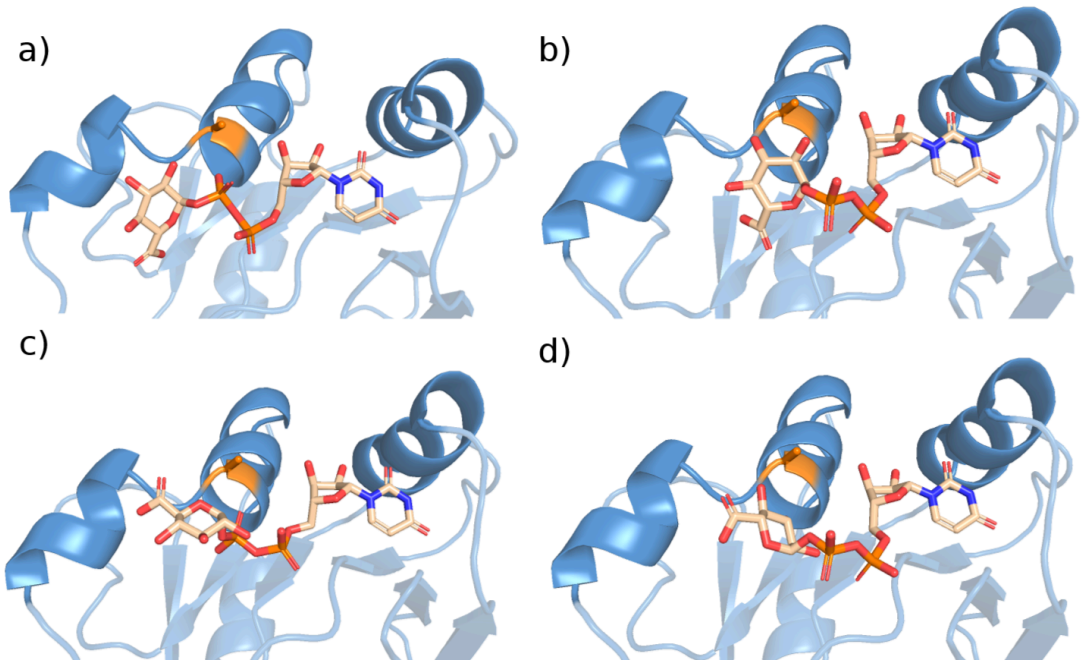

Figure A20: The four biggest clusters for the UDP-GlcA docked to mutant 3. 20% belong to the cluster represented by a), 20% belong to b), 16% belong to c) and 12 % belong to d). The mutated residue is highlighted in orange.

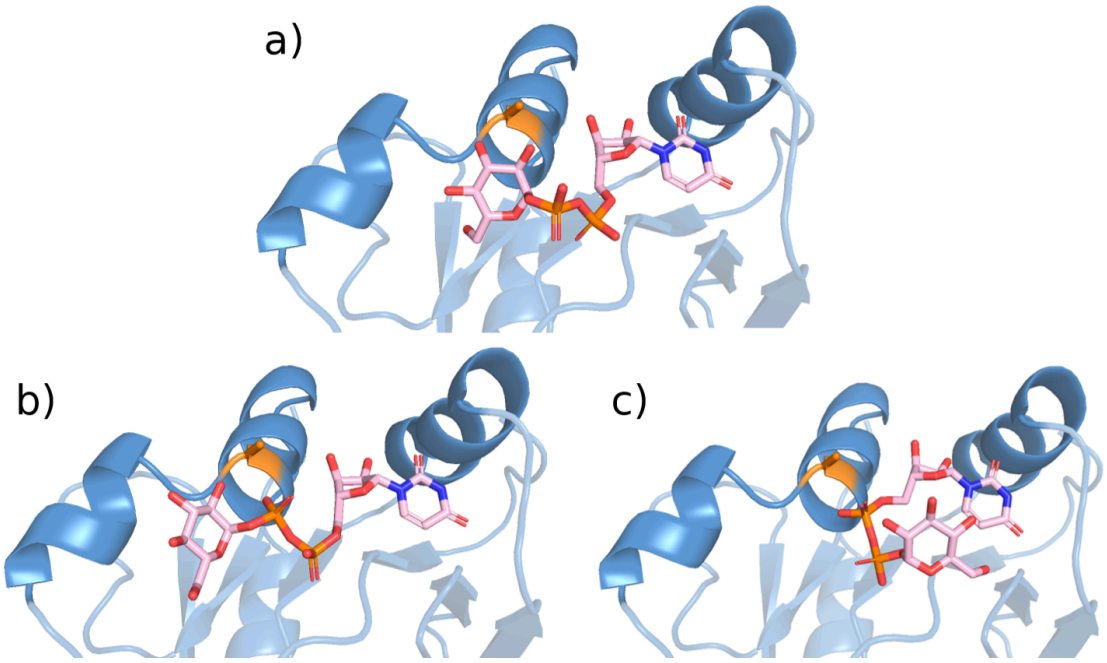

Figure A21: The three biggest clusters for the UDP-GlcA docked to mutant 3. 32% of the poses belong to the cluster represented by a), 20% belong to b) and 14% abelong to c). The mutated residue is highlighted in orange.

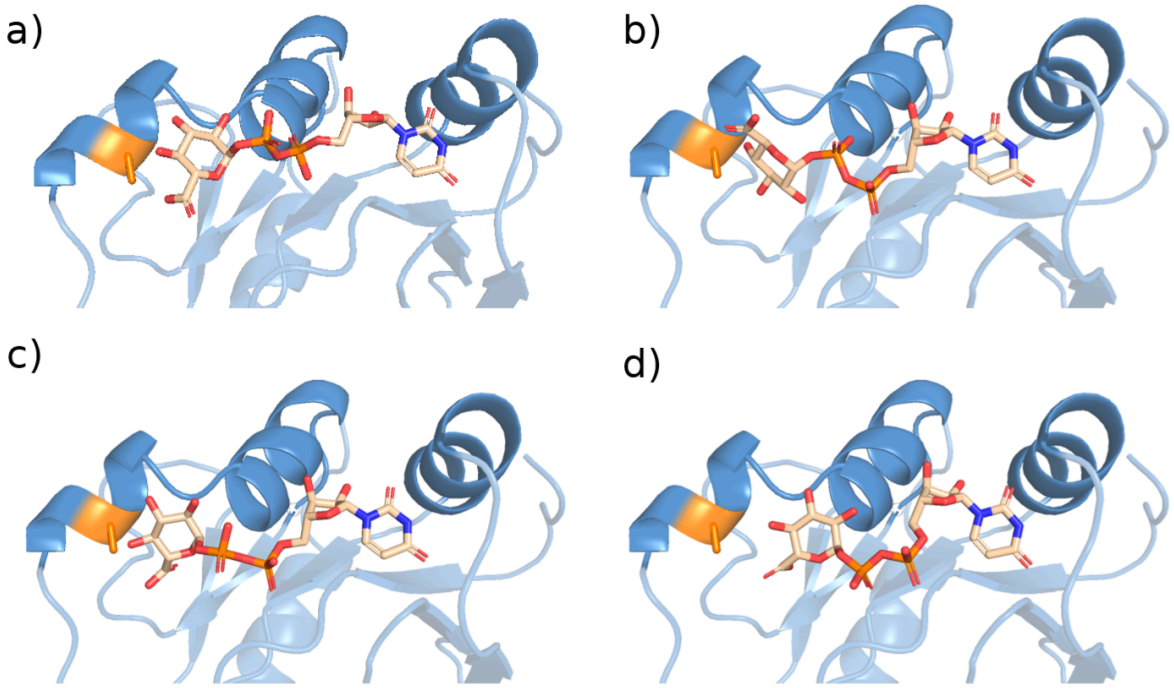

Figure A22: The four biggest clusters for the UDP-GlcA docked to mutant 4. 18% of the poses belong to the cluster represented by a) and 16% belong to b), c) and d) each. The mutated residue is highlighted in orange.

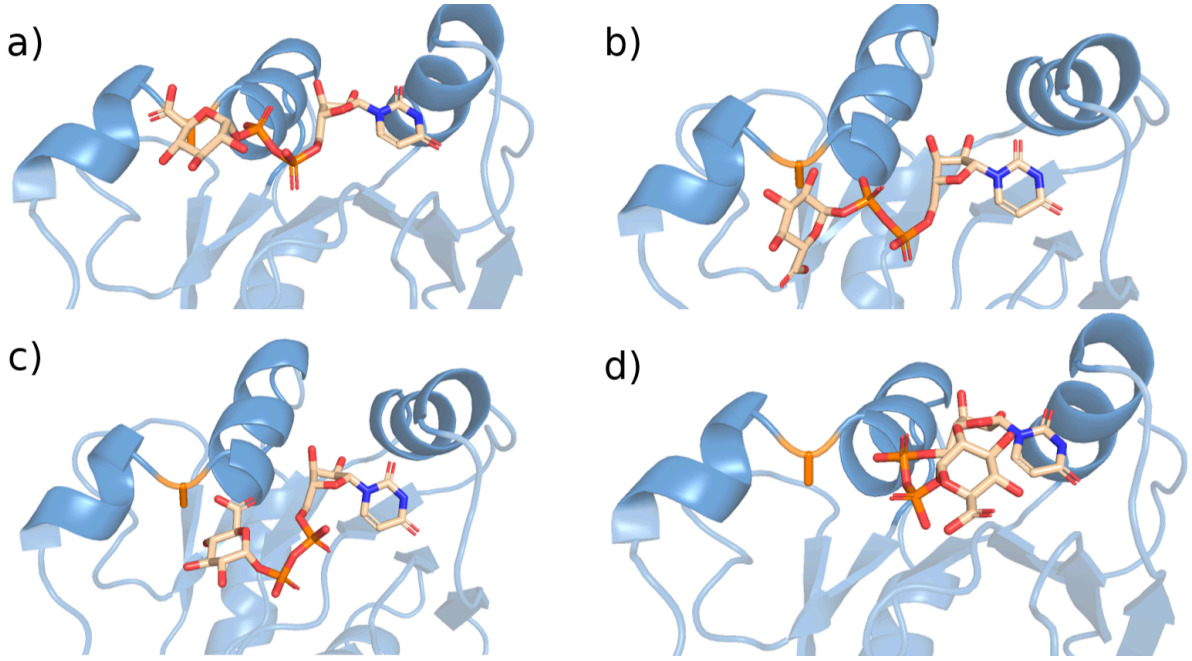

Figure A23: The four biggest clusters for the UDP-GlcA docked to mutant 5. 17% of the poses belong to the cluster represented by a), 15% belong to b), 14% belong to c) and 11% belong to d).

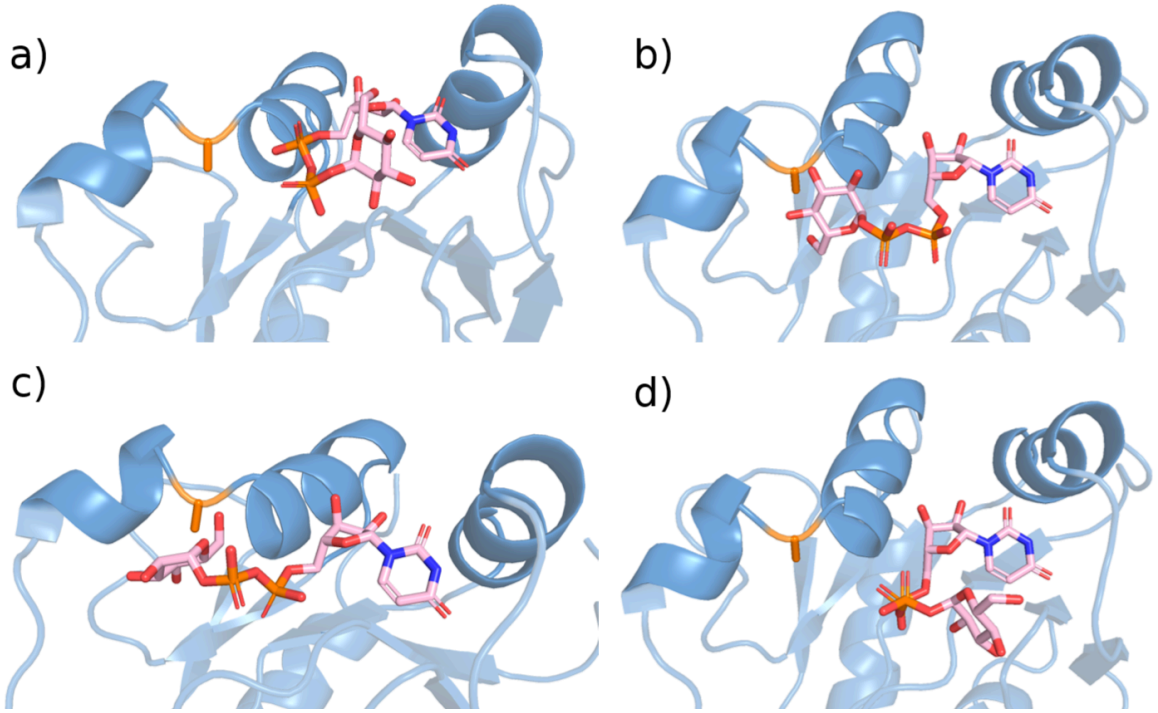

Figure A24: The four biggest clusters for the UDP-Glc docked to mutant 5. 18% of the poses belong to the cluster represented by a) and 18% belong to b), 11% belong to c) and 11% belong to d).

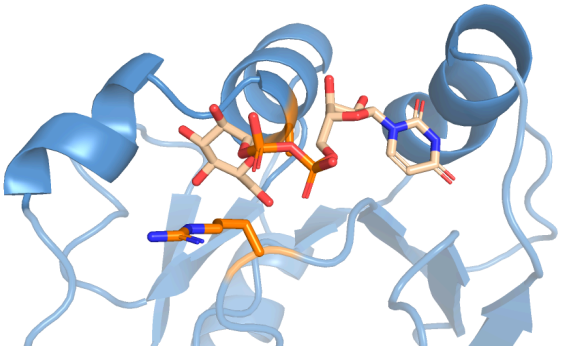

Figure A25: The representative pose for the next biggest cluster when docking UDP-GlcA to mutant 6, with 23% of the poses.

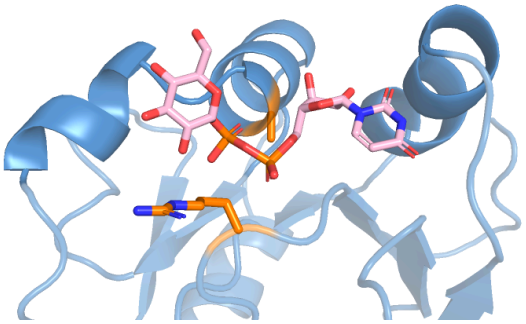

Figure A26: The representative pose for the next biggest cluster when docking UDP-Glc to mutant 6, with 17% of the poses.

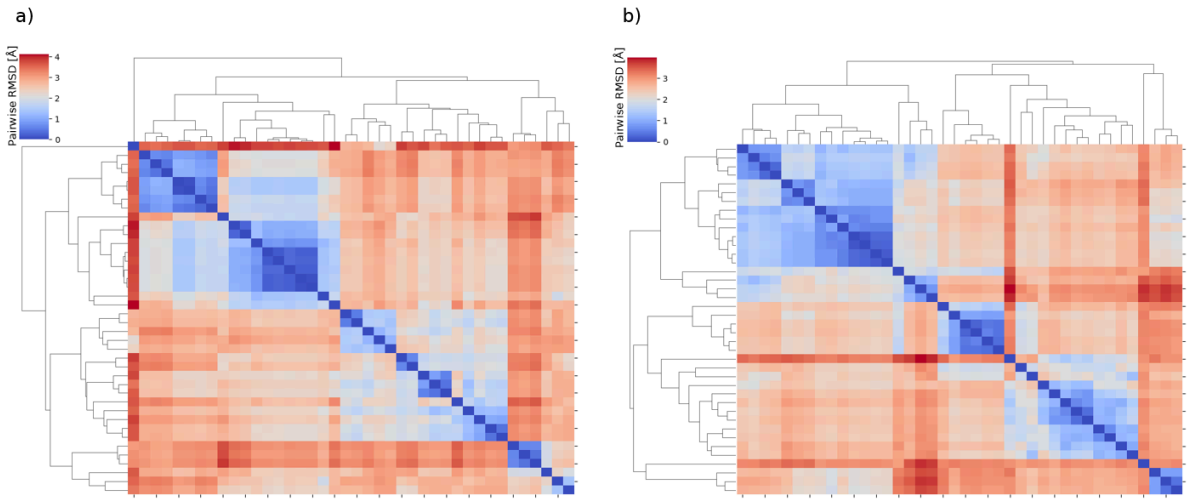

Figure A27: Heatmap of the pairwise heavy-atom RMSD for the poses docked to mutant 1 when docking with a flexible R218 side chain, with the corresponding dendrogram. a) shows the heat map for UDP-GlcA and b) for UDP-Glc.

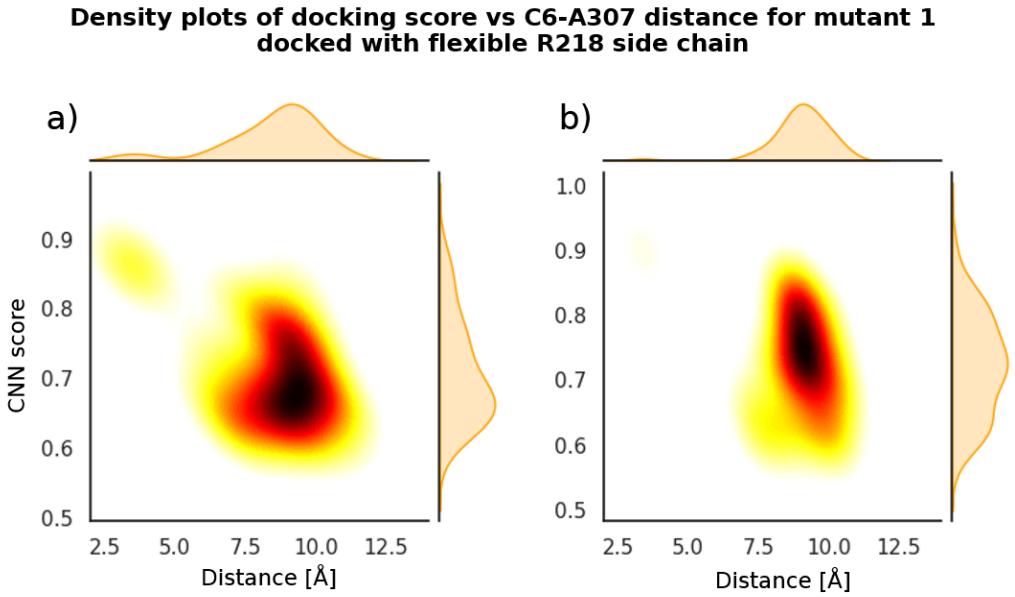

Figure A28: Density maps for the distance between the C6 atom of the sugar and A307 vs the CNN score for the filtered UDP-GlcA and UDP-Glc poses from docking to mutant 1 with a flexible R218 side chain.

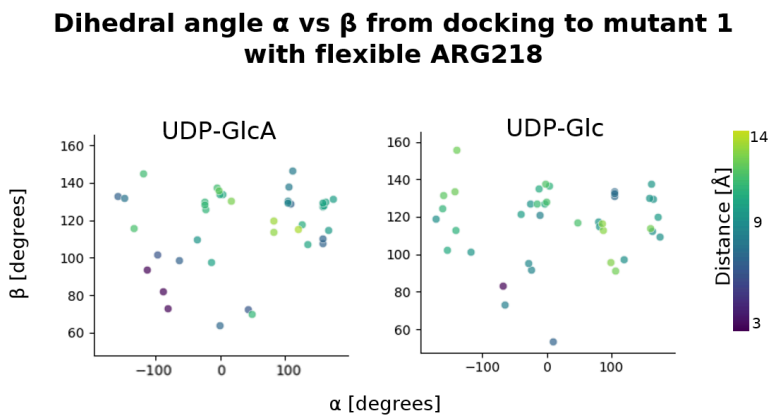

Figure A29: The dihedral angles plotted against each other for the poses from docking to mutant 1 with a flexible R218 side chain. The data points are colored after the distance between the C6 of the sugar and A307.

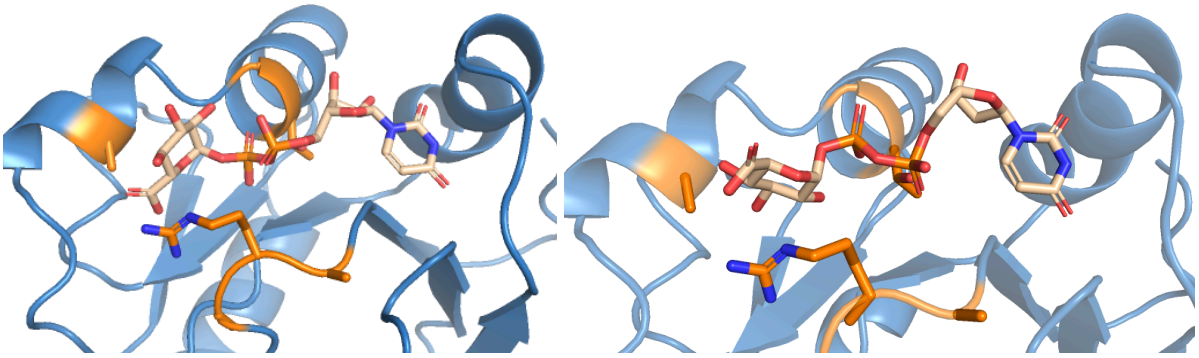

Figure A30: The two biggest clusters from docking UDP-GlcA to mutant 1 with flexible side chain. 45% of the poses belong to the cluster represented by the pose to the left and 38% belong to the cluster with the representative pose shown to the right.

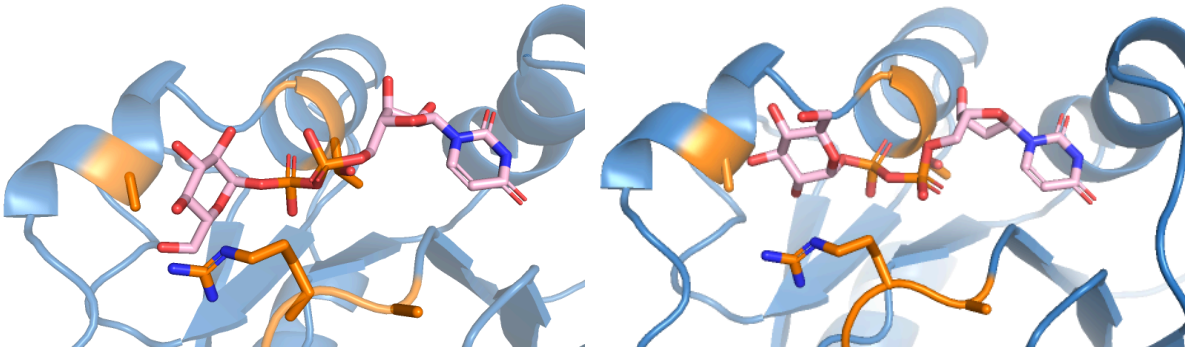

Figure A31: Two biggest clusters from docking UDP-Glc to mutant 1 with flexible side chain. Both clusters contain 45% of the poses.