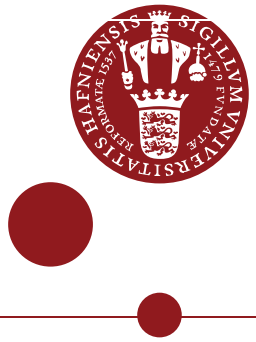
UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE
COMPUTATIONAL AND RNA BIOLOGY



Master thesis

Dovydas Vaitkus

Are molecular mechanics enough to predict glycosylation specificity by glycosyltransferases?

Supervisors: Amelie Stein, David Teze

Submitted on: 14 September 2022

Name of department:	Computational and RNA Biology; DTU Biosustain
Author(s):	Dovydas Vaitkus
Title and subtitle:	Are molecular mechanics enough to predict glycosylation specificity by glycosyltransferases?
Supervisor:	Amelie Stein
External supervisor	David Teze
Submitted on:	14 September 2022
Grade:	

Table of Contents

ABSTRACT

Polyphenol glycosides are bioactive compounds known for their health benefits, and are used as pharmaceuticals, food ingredients and nutraceuticals. However, chemical glycosylation glycosylation is challenging, and the enzymatic pathways are difficult to elucidate because of the lack of general knowledge about suitable enzymes, and there being a high number of potential glycosylation sites on any given polyphenol. Despite that, AlphaFold2 can now be used to predict high quality structures of family 1 glycosyltransferases (GT1s) – enzymes that can form different glycosidic bonds efficiently and selectively. This development thus enables a shift of focus to enzyme structure-function studies that could possibly exploit the copious amounts of structural information. Based on the GT1 research, it is often proposed that glycosyltransferase specificity can be rationalized by the lock-and-key principle on the formation of the Michaelis complex. As AlphaFold2 can output full protein structural models, we have decided to test this hypothesis by combining molecular docking and molecular dynamics (MD) simulations for site-specific *O*- and/or *C*-glycosylation predictions. We developed a full MD pipeline to produce ternary (enzyme:glycosyl donor:glycosyl acceptor) Michaelis complexes, in a computationally cheap and scalable fashion. We carried out simulations of phloretin, 5,7-dihydroxy-4-methylcoumarin, and 5,7-dihydroxy-4-phenylcoumarin with multiple (13–34) well-characterized plant GT1s. Importantly, the formation of Michaelis complex could be observed for most ternary complexes with the coumarin derivatives, independently of experimental reactivity. Overall, no correlation was found between experimental results and predicted regioselectivities. These results show that the formation or the stability of a Michaelis complex is poorly correlated to reactivity, and that transition state stabilization has to be considered to rationalize glycosyltransferase specificity.

ABBREVIATIONS

57DHMC – 5,7-dihydroxy-4-methylcoumarin

57DHPC – 5,7-dihydroxy-4-phenylcoumarin

AF2 – AlphaFold2

QM/MM – quantum mechanics/molecular mechanics

CV – collective variable

GT – glycosyltransferase

MD – molecular dynamics

ML – machine learning

UGT – UDP-glucose dependent glycosyltransferase

$TS^{\ddagger}$ – transition state

BACKGROUND AND INTRODUCTION

Biocatalysis and enzyme engineering

Biologically catalyzed reactions, also called bioconversions, are usually classified into two distinct groups[1]. The first group is conversion associated with cell growth, which is more generally known as fermentation. While historically it has been primarily utilized in food and drink industries, it is still widely used in the production of amino acids[2], and several different organic acids[3,4]. The second type of bioconversion is called biocatalysis. The main difference between traditional fermentation and biocatalysis is the separation of catalyst production and formulation (e.g., whole-cells displaying enzymes, cell extracts, purified enzymes, immobilized enzymes) and the actual substrate conversion in time and/or space[5]. Depending on the subtype of biocatalysis, it elucidates benefits of increased control over catalyst concentration, reaction mixture composition, as well as reduced cross-reactivity with cellular metabolic pathways, when compared to growth associated conversions[6,7].

Another great interest for biocatalysis comes from its stance as an environmentally friendly and efficient strategy for industrial synthesis. Current methods of fine chemical and pharmaceutical manufacturing often rely on the use of mostly inorganic compounds, e.g. reductions by metals and metal hydrides, oxidations by permanganate, manganese dioxide, and chromium(VI) reagents, all of which result in major amounts of waste[5]. Enzymes, on the contrary, operate under mild conditions (regarding pH, temperature and pressure) and can accelerate reactions up to 10^{17} times[8]. Additionally, enzymatic reactions are usually chemo-, regio-, and stereospecific, which nullifies the need of functional group activation, protection and deprotection[9].

Most natural enzymes are the result of billions of years of evolution, which turned them into highly specialized and efficient catalysts[10]. Unfortunately, high production cost and limited physico-chemical capabilities often restrict their large-scale application. Correspondingly, using the advances in molecular biology and bioinformatics, a field of enzyme engineering has been developed, with a general goal of creating process-specific biocatalysts[11]. To achieve this, existing proteins can be enriched with novel properties, such as selectivity for new substrates and/or isomers, or formation of unusual bonds (e.g., carbon-silicon[12], carbon-boron[13]). However, the improvement of existing characteristics is more widely practiced. This includes reaction productivity, natural substrate selectivity, thermostability, and tolerance for high concentrations of salts, organic solvents, substrates, products[14,15].

Two main strategies for enzyme engineering are directed evolution and rational design. Directed evolution is based on mimicking natural evolution in experimental setting, only decreasing the cycle time to weeks or even days[16,17]. It relies on an efficient and specific selection of improved enzyme mutants generated by random or saturation mutagenesis. Rational design, on the other hand, is a

strategy encompassing only one or several specific mutations, thus requiring much deeper knowledge about the protein structure and its relationship with catalytic function[18]. Therefore, this field is highly reliant on both the accessible data, including protein sequences, structures and their connection to function and computational methods used to analyze it, ranging from sequence and structural alignments to docking, molecular dynamics, and machine learning tools[19,20].

Enzyme regiospecificity

Apart from catalytic activity, one of the most important enzymatic properties is selectivity towards specific substrates and their isomers, and the ability to produce single products. It has been reported that around 57% of commercially available drugs and about 99% of purified natural products are chiral compounds[21]. That is expected, since enantiomers can have completely different chemical properties and effects in medicine[22]. Conventional chemistry is unable to produce pure chiral compounds without using chiral components, which must be obtained from natural sources. However, even on a chemo- and regioselectivity level it is still not completely understood how such precision is achieved in Nature, and multiple mechanisms have been proposed in over a hundred years now.

Lock-and-key model

Emil Fischer proposed in 1894 that the specificity of an enzyme towards its substrate is based on a complete complementarity of two geometric shapes, a perfect fit between ligand and protein, analogous to "lock and key"[23] (Fig. 1, A). Albeit simple, this model successfully explains a large number of specificity cases. As a notable example, many hydrolases, and lipases in particular, have successfully been engineered using the lock-and-key principle[24].

Induced-fit model

Proposed by David Koshland, Jr. in 1958, the induced-fit mechanism builds on the lock-and-key model with a substantial modification to the "lock" part. It takes into account the flexibility of protein, describing it as a dynamic entity that gets structurally altered by the binding of ligand[25] (Fig. 1, B). According to this theory, a "correct" substrate aligns with the active site residues in a way that allows conformational changes required for the reaction. Induced-fit explains why some molecules are not substrates of enzymes despite fitting perfectly into the active site.

Selected-fit model

Also known as “conformational selection”, this model describes an alternative origin of protein conformational changes[26]. Selected-fit model assumes that there exists an equilibrium of protein conformational states, and the ligand only binds the “most compatible” ones (Fig. 1, C). Following the initial binding, enzyme:substrate complex then undergoes an induced-fit process, during which the local structure gets optimized and stabilized. Finally, the ensemble of protein conformational states gets redistributed and equilibrated.

Keyhole-lock-key model

Building on top of the traditional models, Prokop and colleagues postulate that the importance of substrate tunnels (i.e. “keyholes”) in the enzymes are as important as the active sites themselves[27]. Size, dynamics, and physico-chemical properties of substrate tunnels are hypothesized to determine kinetics and equilibria of substrate entry and product exit (Fig. 1, D). This model shifts the focus of protein engineering from only the active site residues to also the ones surrounding the ligand entry path.

Combination lock-and-key model

All beforementioned models provide a simplistic view on an otherwise intricate phenomenon of molecular recognition/interaction processes. Based on the combination lock and key system, only a combination of complementary features provided by both enzyme and the ligand make the binding possible[28]. Additionally, both feature variables on protein and the ligand fine-tune and adapt for the best fit, somewhat similar to induced-fit systems. Features themselves can be geometric properties (e.g., size, volume, shape, surface area, etc.) and physico-chemical properties (e.g., electrostatics, hydropathic and van der Waals energetic components), that, together with other chemical features (e.g., hydrogen bond donors/acceptors, aromatic centers, etc.), would define a molecule interaction fingerprint. The implementation of this model is made possible by advances of machine learning methods, that make the feature extraction from large datasets possible.

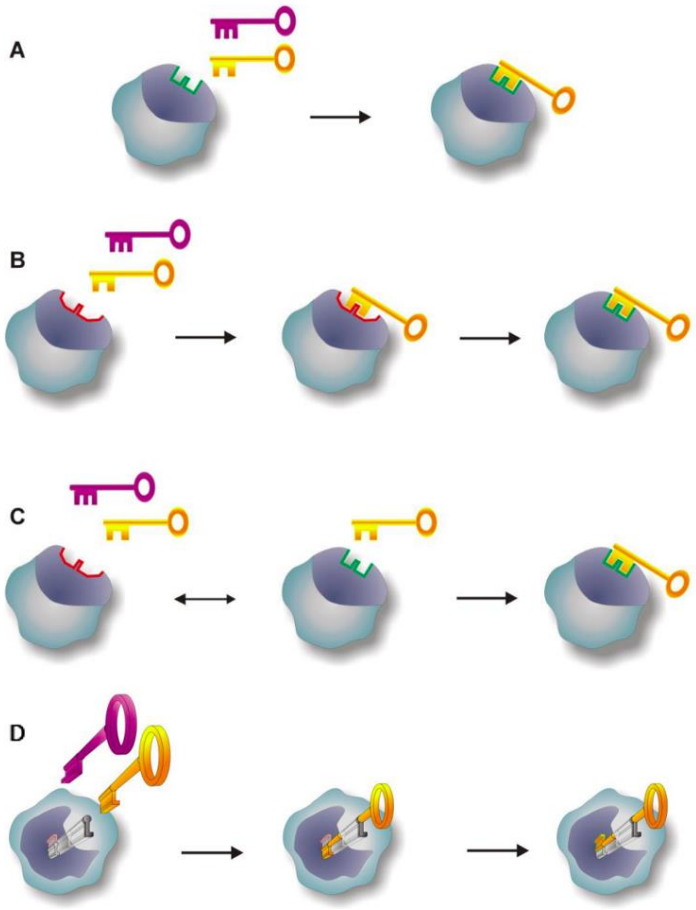


Figure 1. Different models of enzyme catalytic recognition and formation of a transition state: A) lock-and-key model; B) induced-fit model; C) selected-fit model; D) keyhole-lock-key model. Figure from Prokop et al., 2012.

Polyphenol glycosides and their production

Glycosides are organic molecules with one or more sugar molecules bound through glycosidic linkages. In a broad sense, glycosylated compounds carry out a myriad of functions in every living cell, including energy and information storage, structural maintenance, molecular and cell-cell recognition, cellular regulation[29]. However, most often this term is used to describe glycosylation of natural products, especially secondary metabolites that exert defensive functions in plants or other organisms. Polyphenols are a prominent family of such compounds. Characterized by having one or more phenol units, they are widely known due to their chronic disease (cardiovascular disease, diabetes, cancer, cognitive disorders) risk reducing effects in humans[30,31]. As an example, some coumarin-based drugs are used in medicine as anticoagulants and antineurodegenerative agents, and are considered as some of the most promising compounds in anticancer drug research[32,33]. Phloretin and

its derivatives exhibit antifungal, anticancer, antioxidant, antiosteoclastogenic, antiviral, anti-inflammatory, antibacterial properties[34]. Quercetin similarly show antioxidant, antibacterial, antiparasitic, antifungal, anti-immunosuppression properties[35]. However, low bioavailability due to metabolic processes *in vivo* and pH-dependent instability showcases the importance of studying modified polyphenols[36,37]. Polyphenol glycosides, while retaining health-promoting effects of corresponding aglycons, also offer improved metabolic stability and solubility[38,39].

Despite of their potential, chemistry of polyphenols poises a production challenge, as even the simplest compounds have multiple *O*- and *C*- glycosylation sites. Therefore, chemical polyphenol glycosylation requires multiple steps of group protection and deprotection in order to achieve regiospecificity. A notable example is synthesis of nothofagin (*C*-glucoside of phloretin), which currently can be achieved in an 8-step reaction[40]. Alternatively, nothofagin can be synthesized in a 100 g scale by employing a *C*-glucosyltransferase cascade[41].

Glycosyltransferases

Glycosyltransferases (GTs) are the enzymes that facilitate transfer of sugars from a sugar donor to a sugar acceptor. Leloir GTs, as opposed to non-Leloir GTs, use activated nucleotide sugars as sugar donors, and are the main focus of natural product glycosylation discussions. Glycosyltransferase reactions are nucleophilic substitutions at the glycosyl anomeric carbon, and are classified as either retaining or inverting, based on the change of glycosidic bond configuration[42].

Structure-wise, glycosyltransferases are typically split into a few families by fold, of which the most common are GT-A, GT-B, and GT-C. These folds showcase different structures, active sites and even mechanisms. In the GT-A superfamily, typical enzyme topology consists of closely related donor binding and acceptor binding domains, which form a binding pocket capable to accommodate both molecules[43]. Also, GT-A glycosyltransferases often bind a divalent metal ion (Mg^{2+} or Mn^{2+}) that coordinates the phosphate moiety of the sugar donor. A lot of bacterial and majority of mammalian GTs located in the Golgi apparatus and the endoplasmic reticulum, belong to the GT-A superfamily. Glycosyltransferases responsible for the production of A and B antigens that determine human blood groups are some of the most widely known GT-A fold adapting GTs[44]. In the GT-B superfamily, enzymes are composed of two Rossmann-like domains with several ß-sheets linked with α-helices, separated by a deep cleft and do not require metal ions for catalysis[45]. While being diverse, GT-B superfamily includes most of the bacterial secondary metabolite glycosylating enzymes, as well as many insect and plant GTs[46,47]. GT-C superfamily has only been identified more recently and consists of mostly hydrophobic integral membrane proteins that use lipid phosphate-linked sugar donors[48].

Glycosyltransferases are further classified to families by sequence similarity. Even though structurally they are rather homogenous, currently there are 115 known GT families (as of September 14th, 2022, according to CAZy database[49]), with just over a million sequences classified.

In terms of selectivity, GTs employ a varying degree of promiscuity. Individual enzymes can range from a single donor-acceptor pair to multiple combination and structural variations in both[50]. Some GTs also can carry out multiple glycosylations on a single molecule, either on different sites, or by elongating the glycan chain[29,51].

Glycosyltransferase family 1

With more than 34000 sequences classified, 57 experimental structures and 336 experimentally characterized (as of September 14th, 2022, according to CAZy), glycosyltransferase family 1 (GT1) is the most studied GT family so far. These enzymes most notably carry out ß-glycosylation of natural products by employing the inverting reaction type and using α-sugar nucleotides as donors[52]. UDP-sugars are the most commonly used donors, thus GT1s are also called UGTs (UDP-dependent glycosyltransferases)[53]. UGTs catalyze the formation of *O*-, *N*-, *S*-, and *C*- glycosides, and different reaction types can even be observed within individual enzymes[54]. Structurally, UGTs feature the GT-B fold, specifically the C-terminal domain, which contains conserved sugar donor binding site often characterized by PSPG motif[55], and the N-terminal domain, which contains the acceptor binding site and in general is much less conserved, corresponding to a higher variability in acceptor molecules[56].

UGT reaction mechanisms

The active site of most UGTs contains the catalytic base formed by a His-Asp dyad. During the *O*-glycosylation, this dyad deprotonates the glycosyl acceptor, which in turn facilitates the nucleophilic attack on the anomeric carbon of the glycoside donor (Fig. 2). *N*- and *S*- glycosylation mechanisms are slightly different in a way that the nucleophilic attack can happen without acceptor deprotonation, thus reducing the importance of His-Asp dyad. However, acceptor positioning relative to the donor is crucial in every reaction mechanism[54]. For *C*-glycosylation, several related mechanisms have been proposed, though it is yet to be firmly established[42,57,58].

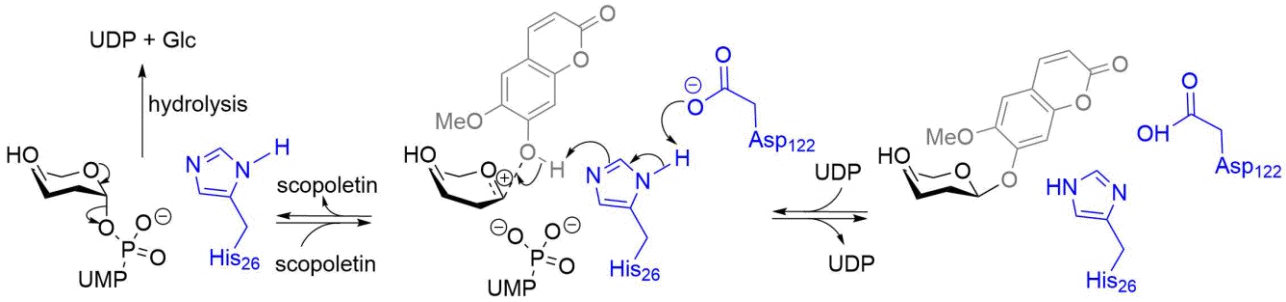


Figure 2. Scopoletin glucosylation by UGTs. Catalytic residues are represented in blue and numbered in reference to *Pt*UGT1[54]. Figure from Teze et al., 2022[59].

UGT regiospecificity

Promiscuity of UGTs is often thought to be correlated with the size of the acceptor binding pocket. It has been proposed that constrained pockets result in a strict substrate specificity, and large pockets can accommodate tens of diverse acceptors[60,61]. When it comes to regioselectivity, the molecular determinants are much less clear. Individual enzymes can produce a single glycosidic product given some substrates, but then end up producing a mixture with other, closely related compounds. In the screening performed by Feng and colleagues, a fungal UGT from *Mucor hiemalis* (*Mh*GT1) was able to produce from 1 to up to 6 different products with different polyphenols[61]. In contrast, plant GT from *Trollius chinensis* (*Tc*CGT1), albeit promiscuous, in majority of cases only produced one or two glycosides. It is evident that promiscuous UGTs often recognize structurally similar acceptor sites, but at the same time it is not enough to ensure enzymatic activity, and the structure of an entire acceptor molecule is important as well[55].

In the end, it is still widely postulated that regioselectivity of UGTs is determined by the positioning of the acceptor in the active site, thus following the lock-and-key model, which is supported by experimental evidence, usually by introducing geometrically sensible residue mutations that affect regio- and stereoselectivity[60,62,63].

Challenges in regiospecificity predictions

Prediction of enzyme regioselectivity is a longstanding challenge in computational chemistry and biology. Similar to structural selectivity, a lot of factors can play a role in this process, and the differences between the activity preferences are often miniscule. Apart from the well-known geometrical shape criterion, specific catalytic activity might also be affected by chemical environment in the active site, differences in energy barriers required to achieve a transitional state, and electrostatics around the ligand. As it was shown by Teze and colleagues, *O*-glycosylation of 3,4-dichlorophenol by *Pt*UGT1 was unachievable by any mutants incapable of deprotonating the substrate, even at high pH values, where the substrate would be predominantly in a deprotonated form. This suggested that

the active site microenvironment was acidic enough to protonate 3,4-dichlorphenol, and/or that deprotonation by the enzyme was essential for the reaction to happen[54].

When talking specifically about UGTs, it is known that phylogenetic relationships between them rarely do explain their functional differences. As shown by Lim and colleagues, UGTs from *Arabidopsis thaliana* could selectively glycosylate esculetin (6,7-dihydroxycoumarin) with a preference to either the 6-OH or 7-OH hydroxyl. However, while regioselectivity changing events could be observed in phylogeny, events that subsequently changed the regioselectivity back also appeared, as well as events that resulted in a complete loss of activity towards esculetin[64].

However, several machine learning models have recently been published for enzyme selectivity predictions[65,66], which benefit from the increasing amounts of data available on possible substrates. Unfortunately, databases still lack enough information about the regioselectivity of enzymes, thus complicating the application of data-driven methods in this case.

Computational methods in enzyme research

Continuous advances in sequencing technology, molecular biology, structural biology, and development of high-throughput experimental methods have enabled a large-scale expansion of data available in protein research. To be able to exploit it, sophisticated computational methods have also been developed. On the other hand, classical *in silico* methods are also improving constantly as a result of better fundamental understanding of natural processes, engineering advances, and, very importantly, growth of accessible computational power.

For enzyme discovery, sequence and structure search, alignment and phylogenetic methods are indispensable. As a starting point, a known protein sequence would often be used to find similar enzymes in publicly available databases by using one of the BLAST variations. Alternatively, if the structure of enzyme is known, similar folds can be detected by using DALI[67], or, with recent structure prediction breakthroughs, tools like FoldSeek can be used to search against millions of modelled structures[68].

Another set of methods is particularly important for enzyme engineering and their mechanistic studies. Rational enzyme engineering is already possible with multiple sequence alignment of related sequences but having a 3D model of the protein of interest enables a much more straightforward path for improvement. Molecular docking related methods allow the placement of substrates and other ligands into the binding pockets of proteins. Since proteins are dynamic structures, this is considered during molecular dynamics (MD) simulations, and with different restrains and additional forces, protein and ligand conformations can be sampled within a reasonable timeframe. For enzymatic reaction modelling, a quantum mechanics/molecular mechanics (QM/MM) hybrid method is still considered

as a "gold standard" after a few decades of usage[19,69]. The fundamental idea behind the QM/MM is to split the enzyme into the catalytic region (QM region) and the surroundings region (MM region). While the MM region is treated in a similar way to simple MD simulations, the molecules in the QM region (substrates, catalytic residues, structured water molecules, cofactors) are allowed to form and break covalent bonds, which leads to accurate reaction modelling.

With the explosive development of machine learning (ML), and the large amounts of data available from high-throughput experiments, the knowledge-based engineering strategies are getting continuously enhanced by data-driven ML methods. For example, multiple protocols for *in silico* directed evolution have been released in the last few years, based on sequence-activity data[70,71]. Moreover, models have been developed for enzyme thermostability predictions[72], predictions of Michaelis constant K_M[73], catalytic constant k_{cat} [74], and other property predictions[75]. As a separate field, *de novo* enzyme design has also largely benefited from the current developments, with several examples published where authors designed novel and functional enzymes[76,77].

Some of the methods relevant for this thesis are discussed further below.

Protein structure prediction

Historically, the prediction of protein 3D structures has most commonly been split in two parallel paths: physics-based approaches and evolutionary based approaches. The former essentially integrates our knowledge about molecular interactions and uses it to simulate the folding of proteins[78]. However, the computational costs required to achieve practically usable simulation durations are still too large, and protein physics models are yet to be sufficiently accurate, thus these approaches still face a long path of development. Alternatively, evolutionary history based methods rely on already known protein sequences, structures, and their historical relationship in a form of pairwise correlations from bioinformatics analyses[79,80]. Up until recently, some of the most commonly used tools for protein structure prediction either heavily relied on close structural homologs (MODELLER[81], SWISS-MODEL[82]) or on fold recognition, i.e., detection of structural templates when there are no homologous structures (I-TASSER[83]).

A breakthrough happened during CASP14 competition in 2020, when AlphaFold2 (AF2) beat all competition with a GDT_TS (roughly the fraction of correctly predicted part of protein) median score of 92.4, a significant improvement from CASP13, when the predecessor version of AlphaFold scored 58.9[84]. An advanced neural network architecture relies on evolutionary contact inference from multiple sequence alignments, and, to some extent, structural homologies. While the protein folding problem is still very much open and falls to the first structure prediction path, AlphaFold2 does very well with modelling many monomeric and, sometimes, multimeric proteins, which has been tested by

several independent research groups[85]. In addition, over 200 million predicted structures were deposited to the AlphaFold Protein Structure Database[86], which makes up approximately the entire UniProt database.

However, the release of RoseTTAFold (a method inspired by AF2, that reached similar accuracy)[87] and AF2 only marked the beginning of the next-generation protein structure prediction methods. In roughly a year, a number of novel folding approaches were released, both building on current knowledge, enhancing current methods, and introducing different ideas, such as the use of protein language models.

Molecular docking

Molecular docking is a simulation method, which studies interactions between molecules (ligands-receptors; protein-protein interactions)[88]. Building on the base of lock-and-key and induced-fit molecular recognition models, the first docking methods appeared as early as 1982[89]. Up until now, every docking protocol consists of two main parts: a) a good molecule positioning algorithm; and b) a robust ranking or scoring system. The complexity of molecule placement is proportional to the complexity of system definition. Usually, the ligand is defined as flexible, meaning that all chemically viable conformations must be considered when docking. While the active site residues can also be defined as flexible, more often they are treated as a rigid body due to otherwise rapidly increasing computational costs[28]. An ideal positioning algorithm therefore would sample the entire conformational space and generate a myriad of possible binding poses, which would then be ranked by the scoring function. In a realistic scenario, such algorithms must find a balance between maintaining a sufficient efficiency and not missing potentially valuable binding poses. Scoring binding poses is a more abstract challenge, as many scoring functions fail to accurately predict the binding affinity, which may or may not correlate with experimentally measured affinities[90]. Additionally, scoring functions are also often used to guide the positioning algorithms. For example, scores of temporary poses guide a Monte Carlo sampling method in the AutoDock Vina software[91].

Many different strategies have emerged to improve on some of the known shortcomings. Ensemble docking generally employs an additional initial step of generating multiple receptor conformations, which are then all used for traditional docking, thus efficiently mimicking flexible docking[92]. With the increasing amounts of structural data including proteins, proteins with ligands, and their annotations with experimentally determined binding energies have enabled ML-based approaches. Typically, ML scoring functions outperform classical scoring functions as they can exploit much larger datasets instead of using several expert-selected structural features[93]. This helps in both improving the sampling procedure, and eventually ranking the poses.

Historically, molecular docking has mostly been used in pharmaceutical industry for virtual structure-based drug screening[92,94]. However, it is also often used as either standalone method or in a combination with MD simulations as an enhanced docking technique in enzyme structural research[60,95], or as a starting point for subsequent MD studies[96].

Conventional molecular dynamics

Molecular dynamics is an advanced technique that allows simulation of biomolecules on an atomistic level[97]. MD simulations rely on the relationship between a specific configuration of atoms and its energy to propagate dynamics[98]. At the highest level, the movement of atoms in time is calculated by integrating Newton's equations of motions, as depicted in the following equation[99].

$$\frac{d^2 r_i(t)}{dt^2} = \frac{F_i(t)}{m_i}$$

with $F_i(t)$ describing the force exerted on atom i at time t; $r_i(t)$ being position of atom i in space at time t; m_i being the mass of atom i.

Time in simulations is split into timesteps δt, typically on a scale of femtoseconds (10 s^{-15}), to be able to capture development of processes on a 10 s^{-14} scale (e.g., rotations and C-C bond vibrations). Numerous algorithms have been developed to integrate equations of motion, including Verlet[100], velocity Verlet[101], and leapfrog[102]. The development of the system relies on the empirical force fields, energy functions, associated parameters that are used to compute energies and forces affecting the movement of atoms, also algorithms that control temperature and pressure.

Conventional MD simulations are often used to study protein structural dynamics[103], protein binding with ligands, substrates, other proteins[94,104], or in combination with other techniques (i.e., molecular docking, QM/MM)[92,95,105]. With the advances in computing hardware (i.e., GPUs) and corresponding software, it is now possible to perform longer and cheaper simulations[106]. However, they are still limited to hundreds of nanoseconds to tens of microseconds, while many biologically relevant processes happen on a millisecond to minute scale[107]. Numerous enhanced MD techniques have been developed that allow a reasonably fast and extensive sampling of systems conformational space.

Enhanced molecular dynamics

Methods of enhanced MD (also called enhanced sampling techniques) add an additional force/potential to the simulated system, to facilitate its escape from the local energy minima. These methods are mainly used to study big conformational changes (folding), and binding/unbinding processes[108].

They can be split into techniques that introduce potentials with or without collective variables (CVs). CVs are reduced descriptors of complicated molecular processes, such as a molecule conformational change or ligand binding. In the simplest systems, it can be a distance between two centers of mass (i.e., protein-protein, ligand-active site), and in more complicated ones they can grow to collections of dihedral angles, 3D contacts, and RMSDs of specific domains[109]. Different enhanced sampling techniques are summarized in Figure 3.

Replica Exchange Molecular Dynamics (REMD) is based on carrying out several parallel simulations from an identical starting point, but in different temperature, and then exchanging the replicas between neighboring temperature windows[110]. Accelerated MD introduces a bias potential to the system when it drops below a certain potential energy threshold, meaning that the protein is not allowed to fully stabilize[111].

Steered MD is the staple of CV utilizing methods. It introduces an external force potential applied along the CV[112]. Most often steered MD is used to pull ligands out of the protein active site to uncover the ligand-target unbinding mechanism, although it can be successfully used in other cases, such as pulling cellononaose (common substrate) across the interfacial active site of different cellulases[113] or exploring the conformational rearrangements of Rieske protein domains[114]. Random Acceleration MD (RAMD) uses a slightly different strategy for steering – instead of applying potential strictly along the single CV, it modifies the steering direction every time the ligand gets obstructed[115]. Umbrella Sampling is a powerful technique that utilizes steered MD to obtain multiple conformations of the system at defined intervals along the CV. After that, the simulations from obtained snapshots are reinitialized with a biased harmonic potential, thus eventually obtaining sufficient statistics along the entire reaction coordinate[116]. Metadynamics is another powerful method used to explore free energy landscapes[117]. It introduces a bias potential to the Hamiltonian of the system in the form of Gaussian-shaped function of specified CVs. However, in this case the system is not actively steered, but is rather prevented to go back to already visited conformations.

Enhanced MD techniques, while not entirely physically representative, enable simulated systems to evolve at a much faster pace, which would otherwise take unreasonable amounts of time in conventional MD simulations. Due to this, it is a very popular choice for cutting computational costs and focusing on specific processes instead of using brute-force until the event of interest (e.g., conformational change, ligand binding/unbinding) happens.

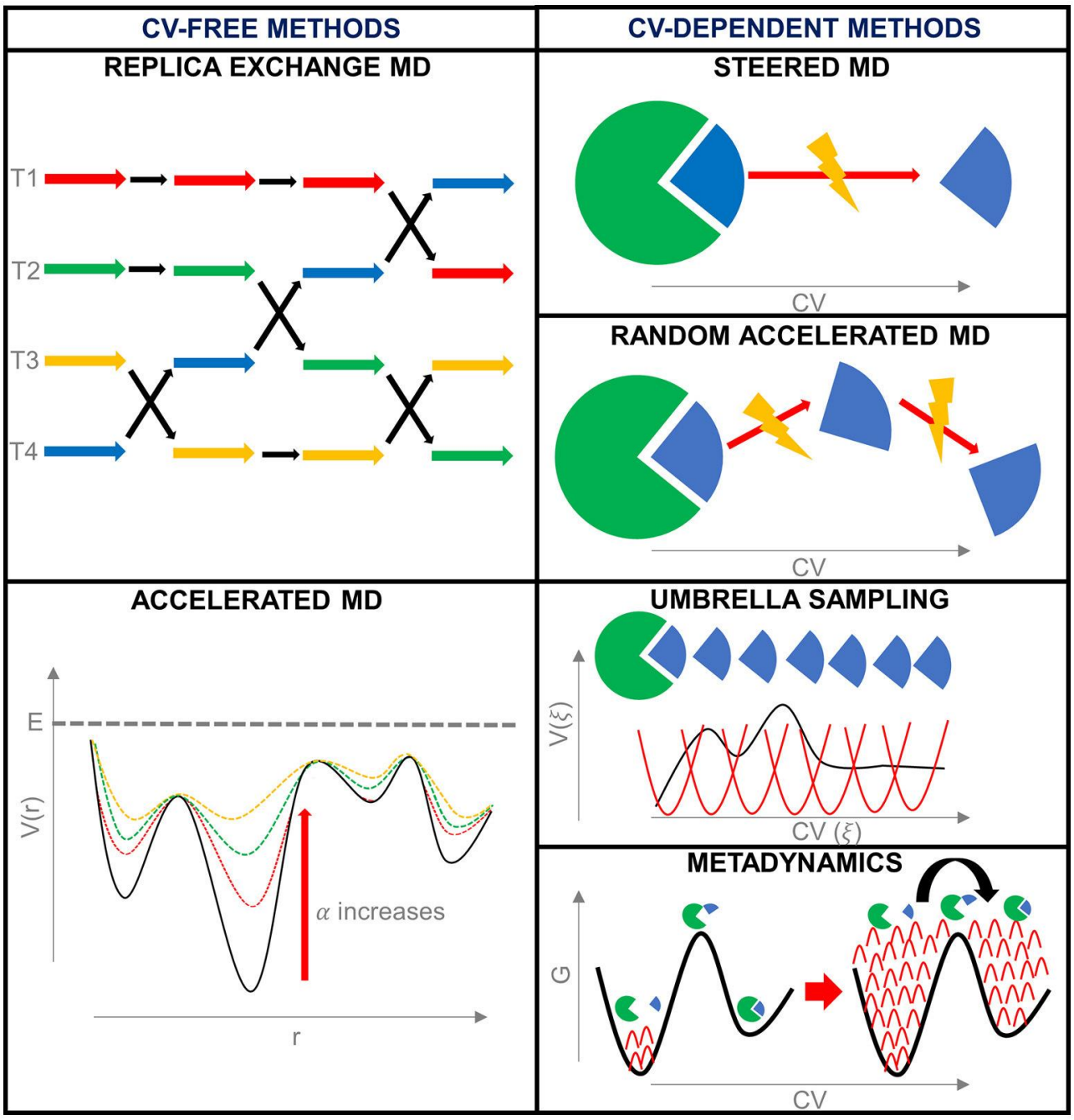


Figure 3. Summary of most popular enhanced sampling methods. Figure from Salmaso and Moro, 2018.

MOTIVATION

Natural enzymes promote catalysis by precise positioning of amino-acids and substrates to complement and stabilize the transition state ($TS^{‡}$). This can be reduced to two fundamental processes: formation of the Michaelis complex, and stabilization of $TS^{‡}$. In UGT research, enzymes are often considered to follow the lock-and-key model for specificity, suggesting that the formation of Michaelis complex is the governing factor for regiospecificity (see *UGT regiospecificity*). Indeed, most UGT rational engineering involves selection of geometrically sensible mutations, and explaining enzymatic reactivity and selectivity through the structural prism and binding of substrate[118–124]. On the other hand, $TS^{‡}$ stabilization is crucial for UGT reactivity, and structurally seemingly reactive UGT:substrate complexes could be observed *in silico* without experimental activity[54].

For this project, we 1) developed a methodology to quickly and computationally cheaply produce models of ternary Michaelis complexes; 2) assessed whether Michaelis complex formation based on molecular mechanics is enough to explain UGT regiospecificity.

RESULTS

Experimental setup

General UGT properties

All modeled and simulated UGTs obtain a similar tertiary fold, specifically the GT-B fold (See *Glycosyltransferases*). They consist of two tightly packed Rossmann-like domains, with a cleft in between that hosts substrate binding sites (Fig. 4, a). N-terminal domains consist of 7 stranded parallel ß-sheets surrounded by up to 9 α-helices. C-terminal domains consist of 5-6 stranded parallel ß-sheets surrounded up to 8 α-helices (Fig. 4, b). Both domains are also surrounded by less conserved loop regions of varying length.

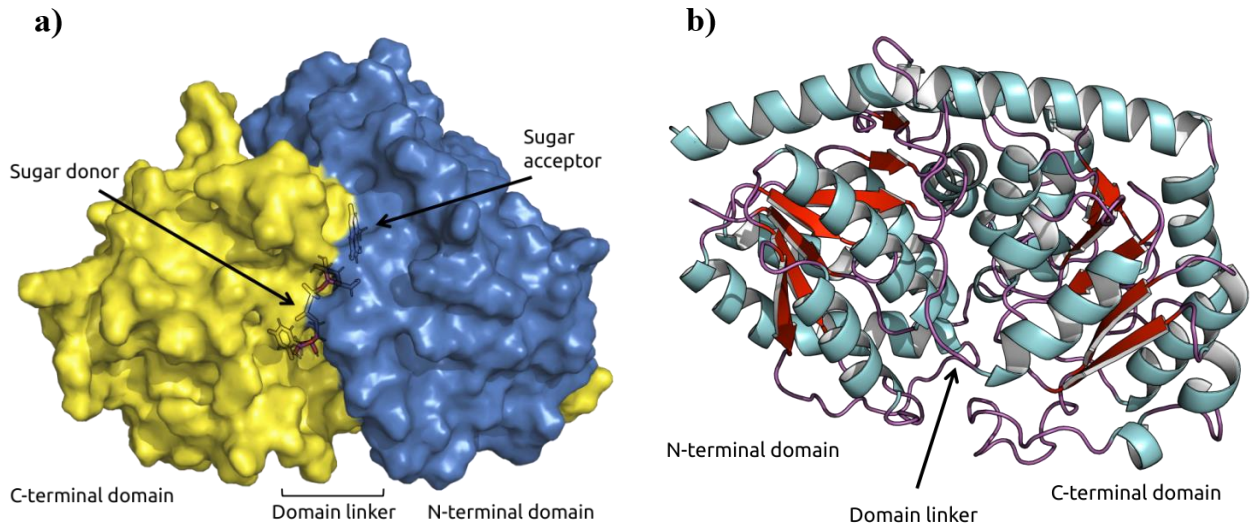


Figure 4. Structure of family 1 plant glycosyltransferases; a) surface view of UGT19 model, with sugar donor and acceptor (57DHMC) present in the binding site; b) cartoon view of UGT19 model with secondary structure components highlighted.

According to the reaction mechanism[54,125] reasonable thresholds for a catalytic acceptor conformation for *O*-glycosylation could be set to: the distance between the proton donor and the catalytic histidine ($N_{\varepsilon His}$-OH_{acc}) is below 3.5 Å, the angle between mentioned hydrogen, donor oxygen, and nitrogen is below 30°, the nucleophilic attack distance ($C1_{glc}$-OH_{acc}) is below 5 Å (Fig. 5, a), and the angle formed by $O1_{glc}$-$C1_{glc}$ bond and reactive oxygen of acceptor is above 130° (Fig. 5, b).

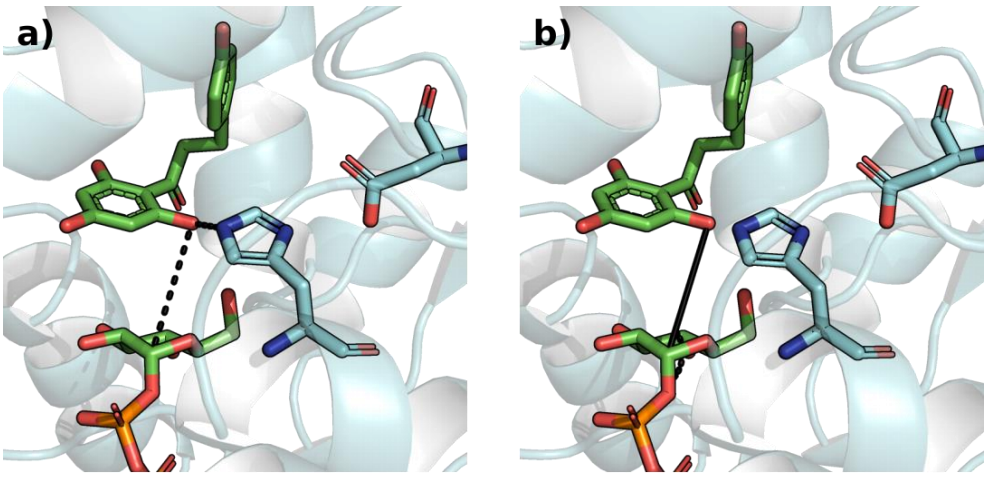


Figure 5. Parameters for *O*-glycosylation reactivity assertion: a) nucleophilic attack distance $C1_{glc}$-$OH2'_{phl}$ and hydrogen bond distance $N_{\varepsilon His}$-$OH2'_{phlo}$ are shown in dotted lines; b) angle between $O1_{glc}$-$C1_{glc}$ bond and $OH2'_{phl}$. Angle for hydrogen bond not shown.

For *C*-glycosylation of phloretin, same criteria for hydrogen bonding were used, considering any hydroxyl group close to catalytic His as potential donor. Nucleophilic attack distance was calculated from potential *C*-glycosylation sites to anomeric glucose carbon, with the same 5 Å threshold (Fig. 6, a). Also, angle between the vector perpendicular to the A aromatic ring of phloretin and the $C1_{glc}$-$O1_{glc}$ must be above 120° (Fig. 6, b).

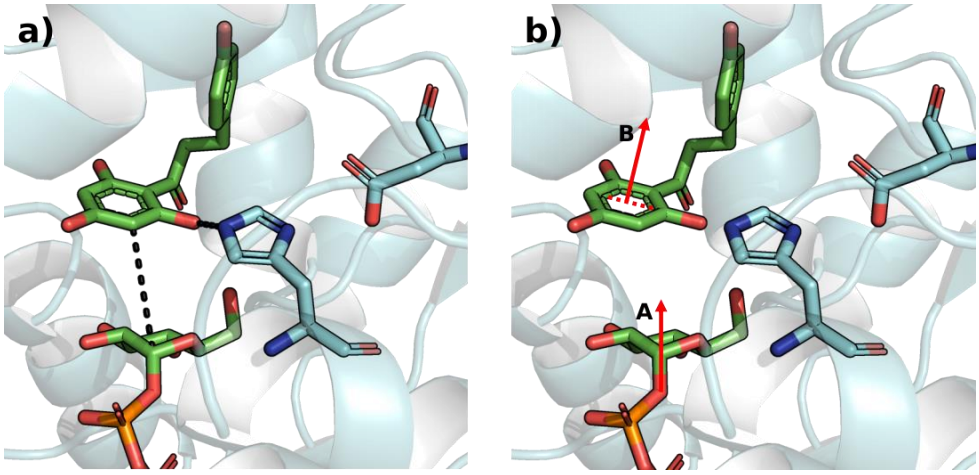


Figure 6. Parameters for *C*-glycosylation reactivity assertion: a) nucleophilic attack distance $C1_{glc}$-$C3'_{phl}$ and hydrogen bond distance $N_{\varepsilon His}$-$OH2'_{phlo}$ are shown in dotted lines; b) angle between vectors A and B. Angle for hydrogen bond not shown.

UGT datasets

Two different datasets were used in this study, both experimentally and computationally. A first set concerns itself with 13 enzymes assessed for the glycosylation of phloretin, and a second focuses on 34 plant GT1s, assessed for dihydroxy coumarins glycosylation.

For simulations with phloretin, 6 *O*-GTs, 1 hybrid *O*-/*C*-GT and 1 *C*-GT were selected from the in-house collection, with screening and/or kinetic data from several different experiments. 5 other *C*-GTs were taken from literature, provided they were able to *C*-glycosylate phloretin. Since data on their experimental yields could not be directly compared, only the categorical classification was used.

For simulations with coumarin derivatives, data was taken from a large-scale screening of 40 GT1s against 47 polyphenols (not published). Out of 40 enzymes, 5 originated from either bacteria, yeast, or animals, which results in them being structurally different from plant UGTs, thus were filtered out. One enzyme was missing the catalytic histidine residue in the active site and was also removed from the dataset. As a result, simulations were carried out with 34 plant UGTs. Notably, all in-house UGTs used in simulations with phloretin, were also simulated with coumarins.

Simulation pipeline considerations

In order to automate the simulation preparation and execution as much as possible, some compromises had to be made along the way. Due to the structural similarities across used UGTs, their sugar donor binding sites also align rather well. Therefore, all modeled structures were superimposed on a crystal structure of *Pt*UGT1 (6SU6.pdb), and UDP-glucose molecule from 6SU6 was copied to the models. This allowed the usage of the same reference coordinates, which also enabled automation of docking and usage of the same parametrized UDP-glucose molecule in all simulations.

Acceptor binding sites, contrary to donor binding sites, largely vary in size and layout. To accommodate for these differences, a large docking grid was selected, which might not be optimal for every UGT. Unfortunately, this introduced a step of manual docking pose selection to filter out “great” poses (according to the scoring algorithm), that might appear even outside of the active site.

While the rest of simulation preparation was mostly automated, some manual bottlenecks remained. GROMACS performs an assessment of charged residue protonation states based on the set pH and also considers chemical environment for histidines. For glycosylation reaction to be feasible, the catalytic His residues in all UGTs must be N_{δ} protonated (to simulate hydrogen bonding between catalytic His-Asp dyad). The algorithm assigned the states correctly is most of the cases, however, 8/39 system required manual assignment.

Lastly, whenever setting up simulations with a new acceptor, the exact atom names must be provided to the pipeline for applying restraints, and eventually analyzing the trajectories.

Simulations with phloretin

Phloretin is a dihydrochalcone, that is relatively easy glycosylated by multiple UGTs. Despite that, it remains difficult to predict which enzyme would produce which glycoside, as it exhibits multiple sites for both *O*- and *C*-glycosylation (Fig. 7).

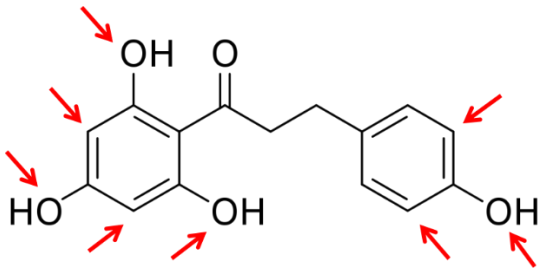


Figure 7. Phloretin molecule with potential glycosylation sites marked.

For this part of the project, we carried out 4 sets of unrestrained MD simulations on 13 UGTs known to produce nothofagin (phloretin 3-*C*-glucoside), phlorizin (phloretin 2-*O*-glucoside) or both. Multiple simulations were done both to have replicas, and to assess the impact of different starting conditions (velocities, docking poses). In order to assess the preference of glycosylation mode (*C*- vs *O*-), fractions of reactive conformations observed for both reaction types were calculated from the total number of snapshots from simulations.

From the first set of 100 ns simulations, only two out of six "pure" *C*-GTs appeared in a productive conformation for *C*-glycosylation, although both also appeared productive for *O*-glycosylation. Out of six "pure" *O*-GTs, three appeared productive for *O*-glycosylation, and two of them also for *C*-glycosylation. Both *C*- and *O*- glycoside producing UGT9 appeared productive for both reactions (Table 1, a). Seven complexes did not appear reactive even once for either reaction, thus we proceeded with a set of replicate simulations (Table 1, b), and a set of simulations with new binding poses for every complex (Table 1, c).

In the replicate set, only 3 complexes did not once appear reactive. However, two *C*-GTs now showed preference only for *O*-glycosylation, two appeared productive for both, and one – only for *C*-glycosylation. Regarding *O*-GTs, two showed only *O*-glycosylation, and two had mixed results.

In the set with alternative starting poses, 6 complexes were inactive. Two *C*-GTs and three *O*-GTs exhibited a reaction mixture, and only one *O*-GT showed complete for *O*-glycosylation.

The final set of "free" MD used a third selection of phloretin binding poses. In this case, 9 complexes did not appear reactive. One *C*-GT showed preference for *C*-glycosylation, and two *O*-GTs ended with a mixture (Table 1, d).

In all simulations, hybrid UGT9 resulted in a mixture of both *O*- and *C*-glycosylation of different proportions, always preferring *O*-glycosylation, which corresponded to the experimentally observed results.

Overall, 10 out of 13 UGTs (barring *Gg*CGT, UGT1, and UGT2), appear to yield reactive poses.

Table 1. Measured fractions of reactive poses for either *O*- or *C*- glycosylation across simulations for the 10/13 UGTs displaying reactive poses; a) 100 ns unrestrained; b) 50 ns unrestrained replica; c) 50 ns unrestrained, different acceptor docking poses; d) 20 ns unrestrained, different docking poses.

		a		**b**		**c**		**d**	
Name	Experi-mental	O_{react} (%)	C_{react} (%)	O_{react} (%)	C_{react} (%)	O_{react} (%)	C_{react} (%)	O_{react} (%)	C_{react} (%)
***Fc*CGT**	*C*-	0.05	13.28	42.02	0.64	0.00	0.00	0.00	0.00
***Mi*CGT**	*C*-	6.38	0.09	0.40	0.00	0.00	0.00	0.00	0.00
***Pp*CGT1**	*C*-	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.18
***Zm*CGT2**	*C*-	0.00	0.00	8.66	6.74	0.15	1.34	0.00	0.00
UGT5	*C*-	0.00	0.00	0.00	0.02	0.23	0.15	0.00	0.00
UGT10	*O*-	36.85	3.08	3.36	0.00	0.94	0.00	0.12	0.12
UGT17	*O*-	0.00	0.00	18.60	2.53	0.02	0.02	0.00	0.00
UGT42	*O*-	1.89	0.00	3.96	0.00	0.02	0.23	0.41	0.24
UGT53	*O*-	41.74	8.14	1.00	0.70	48.43	11.85	0.00	0.00
UGT9	*O-/C*-	7.49	0.65	3.45	0.13	1.17	0.36	96.53	52.53

Preference for glycosylation mode was calculated by subtracting the fraction of productive *C*-glycosylation modes from productive *O*-glycosylation modes. Notably, 3/4 active *O*-GTs only showed preference for *O*-glycosylation, and UGT42 preferred *O*-glycosylation in 3 out of 4 simulations. Out of 5 active *C*-GTs, however, none showed clear preference for *C*-glycosylation: *Fc*CGT, *Pp*CGT1, *Zm*CGT2, and UGT5 all scored higher for *C*-glycosylation once, same with *O*-glycosylation. *Mi*CGT preferred *O*-glycosylation in simulation sets **a** and **b**. Summarized results are shown in Figure 8.

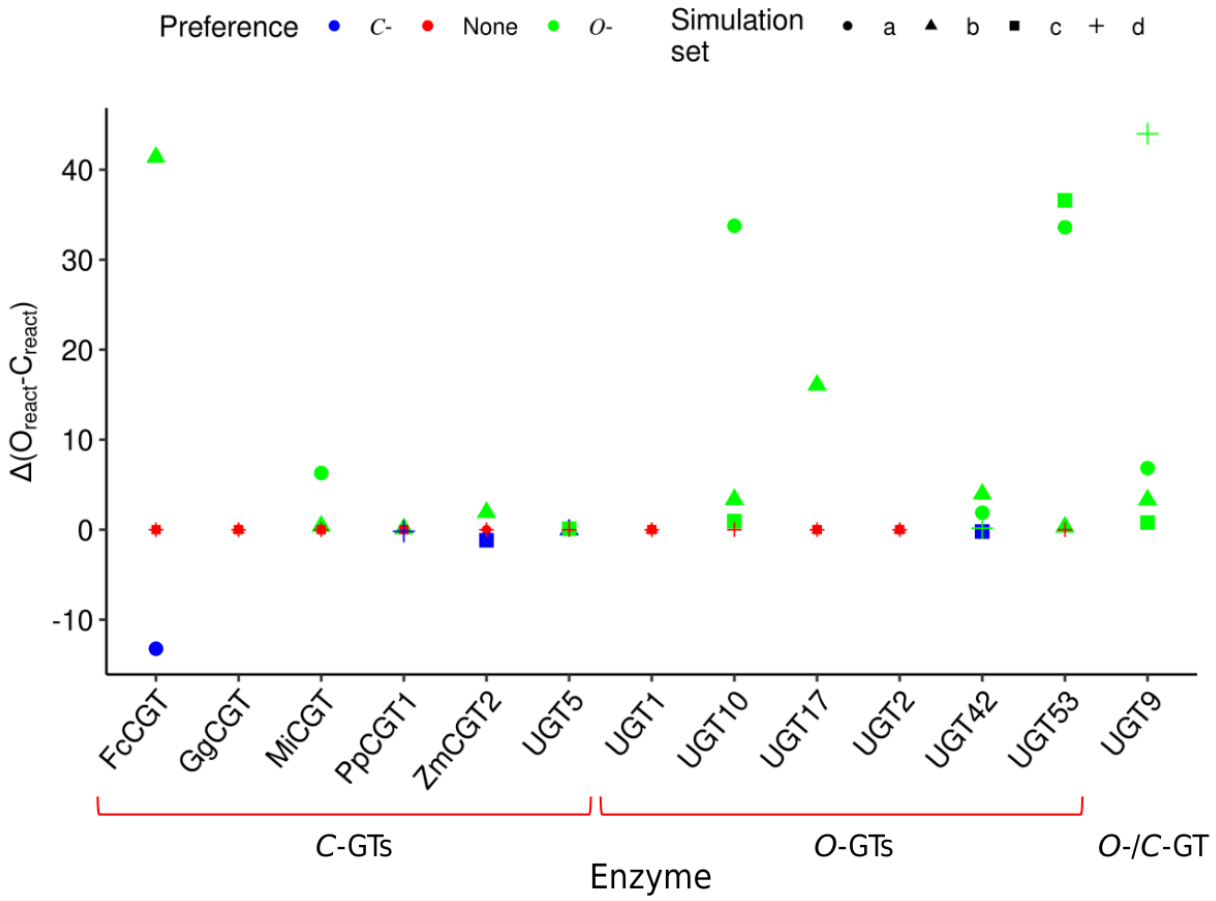


Figure 8. Summary of simulated UGT preferences in phloretin glycosylation. Positive numbers indicate excess of productive poses for *O*-glycosylation, and negative ones – for *C*-glycosylation. Simulation set identifiers match those of Table 1.

Inconsistencies across simulations showcased some drawbacks of using conventional MD. Indeed, simulation outcomes appeared to strongly depend on initial conditions. For example, between two replica simulations (**a** and **b),** *Fc*CGT showed completely opposite results, which rose from having identical initial coordinates but different sets of initial atomic velocities. Different initial docking poses greatly affected most systems in terms of reducing the total number of productive conformations. This, however, stems from the subjective choice of the "best" starting position. Since acceptor binding pockets are diverse across UGTs, there is no "one size fits all" solution when docking larger molecules.

Another drawback of using unrestrained MD is the simulation lengths. Even after 220 ns of accumulated simulation time, 3 complexes did not appear productive once. Additionally, the proportion of productive poses is often small.

Unfortunately, from the current knowledge it appears that both *C*- and *O*-glycosylating UGTs not only use the same machinery, but also bind the sugar acceptor in a similar fashion. High similarity therefore introduces more inconsistency and bias from reactivity criteria. Finally, it is known that

many UGTs also possess reverse glycosylation activities, specifically being able to form UDP-Glc from β-*O*-glycosides. As a result, some *C*-GTs might prefer *O*-glycosylation for specific substrates, but the reaction end-product is only the *C*-glycoside due to the reversibility of *O*-glycosylation. In contrast, *C*-glycosylation is irreversible. When characterizing *C*-GTs, authors often only present end-point reaction products, thus neglecting possible intermediate *O*-glycosylation activity. As an example, in-house experiments with *Mi*CGT and phloretin had indeed shown, that initially it produces phorizin, which is eventually replaced by nothofagin (data not shown).

Due to observed limitations, it was decided to continue the project by introducing restrained MD, continuing to work only on in-house UGTs, and focusing on simpler compounds.

Conformational changes of UGT42

During the **a** set of simulations, a shift was noticed in the conformation of UGT42 (Fig. 9). At roughly 52 ns, the interdomain cleft began opening, exposing both substrates to solute, and then closed again at 62 ns. During this window, conformation of phloretin changed from facilitating *O*-glycosylation into more suitable for *C*-glycosylation, except for the missing interaction with catalytic His. This illustrated two points: 1) flexibility of UGT interdomain linkage allows large scale movements, which temporarily alter binding sites of both sugar donor and acceptor; 2) substrate entry and exit from the binding site is likely to occur independently from the geometry of any “binding tunnels” observable in crystal structures and predicted models.

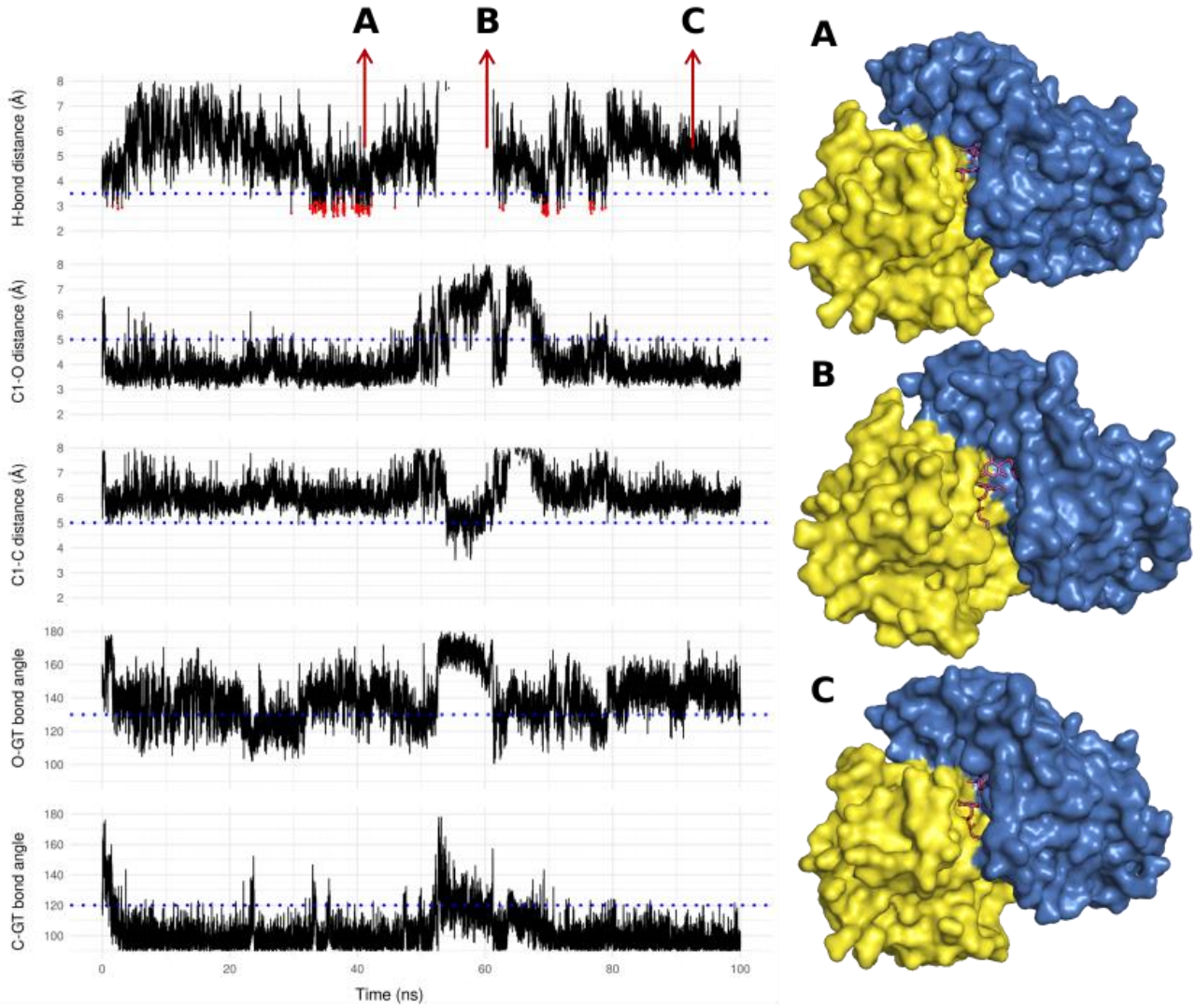


Figure 9. Development of UGT42 complex during the 100 ns unrestrained simulation, in respect to reactivity criteria. Red points in the H-bond distance plot indicate formation of a hydrogen bond between $N_{\varepsilon His}$-OH_{acc} Surface view of the system is shown: a) before opening (42 ns); b) during opening (60 ns); c) after closing (93 ns).

Simulations with coumarin derivatives

5,7-dihydroxy-4-methylcoumarin (57DHMC) and 5,7-dihydroxy-4-phenylcoumarin (57DHPC) are phenolic compounds, both with two *O*-glycosylation sites (Fig. 10). While *C*-glycosylation of these molecules is possible, it is much less common than *O*-glycosylation[126]. As a result, we have selected 34 *O*-glycoside producing plant UGTs from a screening experiment (See Materials and Methods), with varying conversion rates and regioselectivities for both 57DHMC and 57DHPC. The choice of acceptors is based on their small size and well-defined, limited options of glycosylation. We hypothesized that docking of smaller acceptors should not influence the simulation results heavily, as they would be able to move around the active site more freely.

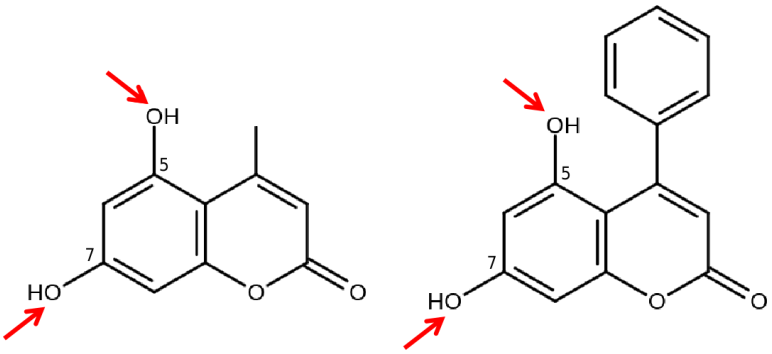


Figure 10. 5,7-dihydroxy-4-methylcoumarin and 5,7-dihydroyxy-4-phenylcoumarin molecules with potential *O*-glycosylation sites marked.

For this part of the project, a short-restrained MD step was introduced to every simulation, during which either one of the glycosylated hydroxyls was pulled close to the catalytic histidine. It was followed by a conventional MD part, during which the reactivity was assessed.

Initially, 136 simulations were carried out (34 enzymes, 2 acceptors, 2 simulations for steering each of hydroxyl groups), each limited to 2 ns, including a 0.5 ns step with distance restraints. Strikingly, 32/34 complexes with 57DHMC appeared in a catalytic conformation at least once, despite 8 of them being completely inactive on the acceptor *in vitro* (Table 2). Likewise, 32/34 complexes with 57DHPC appeared reactive, include 10 ones inactive *in vitro*.

Table 2. Relative experimental yields of 57DHMC and 57DHPC glycosylation screening, together with the fraction of catalytically suitable poses obtained during the unrestrained production MD.

	57DHMC				57DHPC			
Name	O5-glc yield (%)	O7-glc yield (%)	$O5_{react}$ (%)	$O7_{react}$ (%)	O5-glc yield (%)	O7-glc yield (%)	$O5_{react}$ (%)	$O7_{react}$ (%)
UGT1	11.6	2.9	94.0	4.3	0.0	0.0	88.7	47.3
UGT2	0.0	1.5	0.0	61.0	0.0	0.0	0.0	20.3
UGT5	31.1	7.4	0.0	34.7	79.2	0.0	79.3	0.0
UGT9	85.9	0.0	53.0	98.7	27.8	0.0	8.3	0.0
UGT10	41.1	48.0	5.3	8.0	40.5	0.0	0.0	42.3
UGT11	0.0	77.1	1.0	0.0	23.5	0.0	56.0	0.0
UGT16	47.7	26.1	76.7	54.3	3.5	0.0	1.3	22.3
UGT17	88.5	0.0	36.0	5.3	96.0	0.0	92.3	96.3
UGT19	91.6	3.3	0.0	10.3	82.1	0.0	6.7	26.7
UGT36	4.8	10.8	8.0	36.3	1.7	1.5	55.3	71.7
UGT38	94.9	0.0	86.0	88.7	1.2	58.5	0.0	4.3
UGT41	87.2	10.7	31.7	0.0	89.9	3.4	0.3	65.0
UGT42	27.7	66.1	49.7	32.3	56.0	27.5	18.0	57.0

UGT43	32.8	51.7	0.7	0.0	0.0	0.0	4.7	0.0
UGT46	57.1	9.4	11.0	0.0	89.8	0.0	0.0	78.0
UGT47	21.3	75.9	75.0	19.0	100.0	0.0	0.3	57.3
UGT50	94.2	0.0	32.0	97.0	19.1	0.0	0.0	0.0
UGT53	41.9	49.0	14.0	75.0	28.1	0.0	11.3	21.3
UGT117	0.0	0.0	8.0	40.0	0.0	0.0	80.3	58.0
UGT118	0.0	0.0	80.0	49.3	0.0	0.0	41.7	8.0
UGT121	63.7	18.6	87.0	0.0	11.1	20.4	0.0	98.0
UGT123	18.4	0.0	3.0	92.0	33.8	0.0	64.7	1.0
UGT124	1.9	0.0	82.7	7.7	0.0	0.0	0.7	0.0
UGT132	0.0	71.2	74.0	57.3	0.0	26.2	43.0	91.0
UGT140	0.0	0.0	0.0	61.7	0.3	0.0	0.3	0.7
UGT141	0.0	0.0	31.7	95.3	0.0	0.0	0.0	90.3
UGT143	29.5	2.2	47.0	0.3	48.1	0.0	0.0	86.3
UGT144	0.0	0.0	89.7	87.3	6.3	0.2	29.7	66.0
UGT149	0.0	0.0	66.7	0.0	0.0	0.0	7.3	69.3
UGT151	8.9	46.6	62.3	0.0	0.0	3.6	0.0	0.0
UGT152	11.0	84.4	90.3	52.7	17.0	1.0	0.0	13.3
UGT153	0.0	0.0	15.0	34.7	0.0	0.0	0.0	50.3
UGT154	0.0	0.0	0.0	0.0	0.0	0.0	89.7	0.3
UGT156	0.0	47.8	0.0	0.0	3.3	17.5	61.7	45.3

To assess whether there is correlation between experimental product yields and simulated fractions of productive poses, data was plotted (Fig. 11), and Kendall's correlation coefficient was calculated for all experiment-simulation pairs. Interestingly, with Kendall's $\tau = 0.09; -0.25; -0.01; 0.19$, for glycosylation of 57DHMC-O5, 57DHMC-O7, 57DHPC-O5 and 57DHPC-O7, respectively, we observed no correlation between experimental and simulated results.

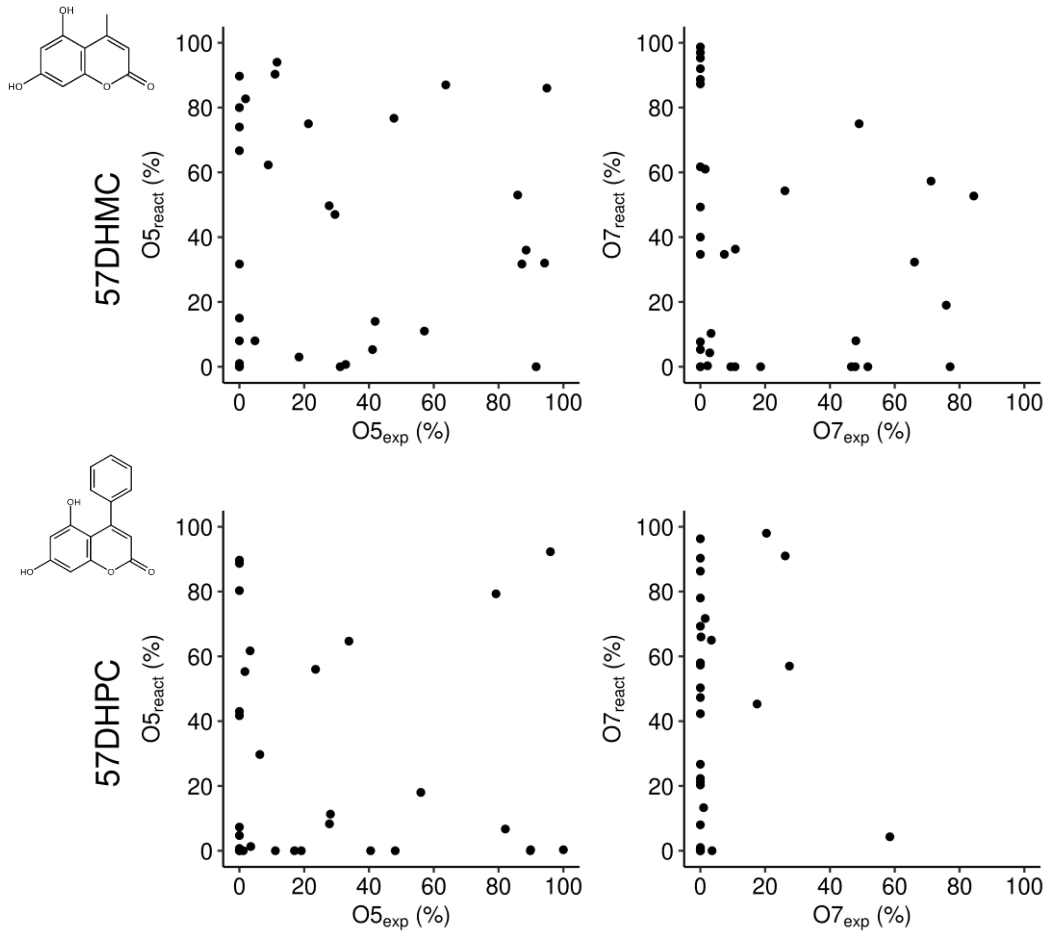


Figure 11. Fractions of reactive poses during simulations plotted against experimental yields of corresponding glycosides.

To assess the accuracy of predicted regioselectivity from simulations, two measures were calculated for each acceptor: 1) difference between experimental yield of O5-glucoside and O7-glucoside, and 2) difference between fractions of reactive poses for O5 glycosylation and O7 glycosylation. The relationship between these values can be seen in Figure 12. Kendall's correlation coefficients showed that in neither of the cases experimental and simulated preferences are correlated ($\tau = -0.08$; -0.09, for 57DHMC and 57DHPC, respectively). When excluding inactive enzymes and using a simple threshold (both differences below or above zero), predictions for 57DHMC regioselectivity were correct 12/26 times. In case of 57DHPC, predictions were correct only 7/24 times.

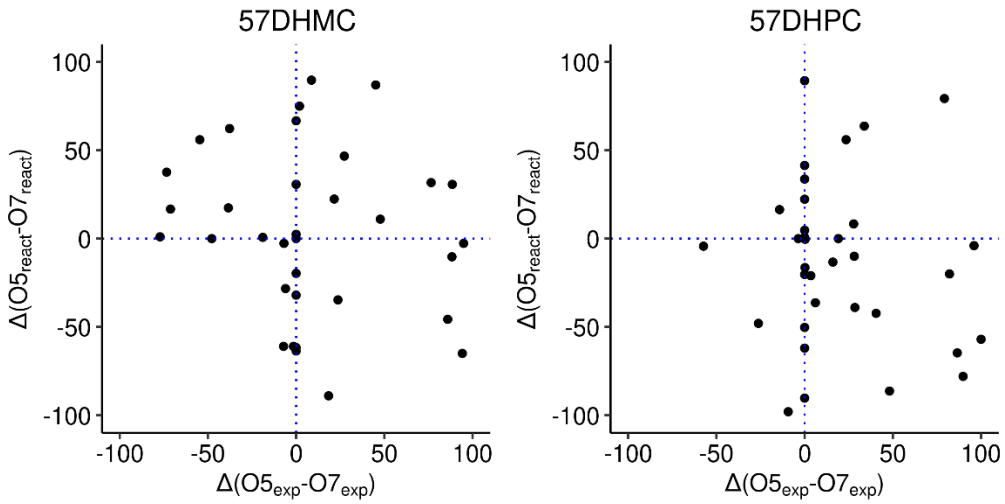


Figure 12. Simulated UGT regioselectivity preferences plotted against experimental preferences. Positive values on x axis indicate excess of O5-glucoside. Positive values on y axis indicate excess of reactive poses for O5-glycosylation. Dotted lines indicate boundaries between preferences.

Additionally, we decided to test other restraints for the steered MD step, as well as replicate some of the simulations from the first set. A representative subset of 14 UGTs was selected, mostly focusing on regioselective enzymes, and excluding most of the inactive ones. Three sets of 52 simulations were carried out with restraints on: a) hydrogen bond distance; b) nucleophilic attack distance; c) both. Results can be viewed in Table 3 for 57DHMC, and Table 4 for 57DHPC.

Table 3. Relative experimental yields of 57DHMC glycosylation screening, and fractions of reactive poses from unrestrained MD part of simulations with a) hydrogen bond distance; b) nucleophilic attack distance; c) both initial restraints. Systems for which no active Michaelis complexes were observed are bolded.

Name	O5-glc yield (%)	O7-glc yield (%)	**a**		**b**		**c**	
			$O5_{react}$ (%)	$O7_{react}$ (%)	$O5_{react}$ (%)	$O7_{react}$ (%)	$O5_{react}$ (%)	$O7_{react}$ (%)
UGT2	0.0	1.5	0.0	67.7	82.7	90.0	0.0	94.7
UGT5	**31.1**	7.4	**0.0**	55.3	**0.0**	0.0	**0.0**	32.3
UGT9	85.9	0.0	71.3	30.7	86.3	0.0	78.3	1.0
UGT11	0.0	77.1	1.3	0.0	15.0	0.0	3.0	15.0
UGT17	88.5	0.0	94.7	95.7	0.0	0.0	48.7	0.7
UGT19	91.6	3.3	0.0	13.7	0.0	0.3	0.3	9.3
UGT38	94.9	0.0	26.3	70.3	0.0	77.0	7.0	77.0
UGT42	27.7	66.1	20.7	0.7	6.7	66.3	76.3	0.0
UGT43	32.8	51.7	0.3	14.7	90.0	0.0	54.0	4.7
UGT50	94.2	0.0	8.0	99.3	18.7	98.7	40.7	99.3
UGT132	**0.0**	71.2	**0.0**	23.0	**0.0**	0.0	**0.0**	92.3
UGT143	29.5	2.2	55.7	0.7	7.0	81.0	9.3	4.7
UGT152	11.0	84.4	92.0	63.7	71.3	69.3	86.3	51.7
UGT156	0.0	47.8	41.3	11.0	0.0	8.7	22.0	5.7

Table 4. Relative experimental yields of 57DHPC glycosylation screening, and fractions of reactive poses from unrestrained MD part of simulations with a) hydrogen bond distance; b) nucleophilic attack distance; c) both initial restraints. Systems for which no active Michaelis complexes were observed are bolded.

Name	O5-glc yield (%)	O7-glc yield (%)	**a**		**b**		**c**	
			$O5_{react}$ (%)	$O7_{react}$ (%)	$O5_{react}$ (%)	$O7_{react}$ (%)	$O5_{react}$ (%)	$O7_{react}$ (%)
UGT2	0.0	0.0	0.0	4.3	5.0	0.0	0.0	8.0
UGT5	79.2	0.0	81.7	98.3	98.7	93.0	1.0	93.7
UGT9	27.8	0.0	49.7	0.3	88.3	0.0	37.3	0.0
UGT11	23.5	0.0	2.0	0.0	2.3	0.0	8.0	1.3
UGT17	96.0	0.0	0.0	68.0	0.0	74.3	39.0	83.7
UGT19	**82.1**	0.0	**0.0**	3.3	**0.0**	2.0	**0.0**	0.0
UGT38	1.2	58.5	7.3	58.7	0.0	0.0	0.0	92.3
UGT42	56.0	27.5	78.3	1.0	93.3	64.7	93.0	61.7
UGT43	0.0	0.0	0.7	0.3	0.0	0.0	3.0	43.0
UGT50	19.1	0.0	0.0	0.0	51.7	0.0	0.3	7.0
UGT132	0.0	26.2	65.7	3.7	66.7	0.0	51.3	42.7
UGT143	48.1	0.0	25.7	0.0	0.0	0.0	13.0	49.0
UGT152	17.0	1.0	0.0	78.0	2.7	76.7	54.7	96.0
UGT156	3.3	17.5	0.3	15.7	0.0	0.0	84.3	90.0

Across the new simulation sets, 26/28 possible enzyme-glycosylation combinations appeared reactive at some point for 57DHMC. Interestingly, 27/28 combinations appeared reactive with 57DHPC (28/28 with initial simulation set included), even though 8/14 enzymes are regiospecific and yield only one product. This could be explained by differences in activation energy. While Michaelis complex can form for both glycosylations, significantly higher energy barrier (i.e., >2 kcal/mol) for one of the reactions determines the regiospecificity for the alternative.

Despite the abundance of productive poses, correlation between experimental results of 57DHMC glycosylation and fractions of reactive poses was not observed (Kendall's $\tau = -0.21, -0.36, -0.08$ for sets **a**, **b**, and **c**, respectively). Simulations of **a** and **b** sets predicted 4/14 preferences correctly, while **c** got 5. For 57DHPC, correlation was also not observed (Kendall's $\tau = 0.04, -0.16, -0.02$ for sets **a**, **b**, and **c**, respectively). Simulations of **a** set resulted in 6/12 correct predictions, while **b** and **c** sets both scored 5/12 (Fig. 13). Notably, productive poses were obtained for UGT2 and UGT42, which are completely inactive towards 57DHPC experimentally.

In general, none of tested restraints improved the results significantly.

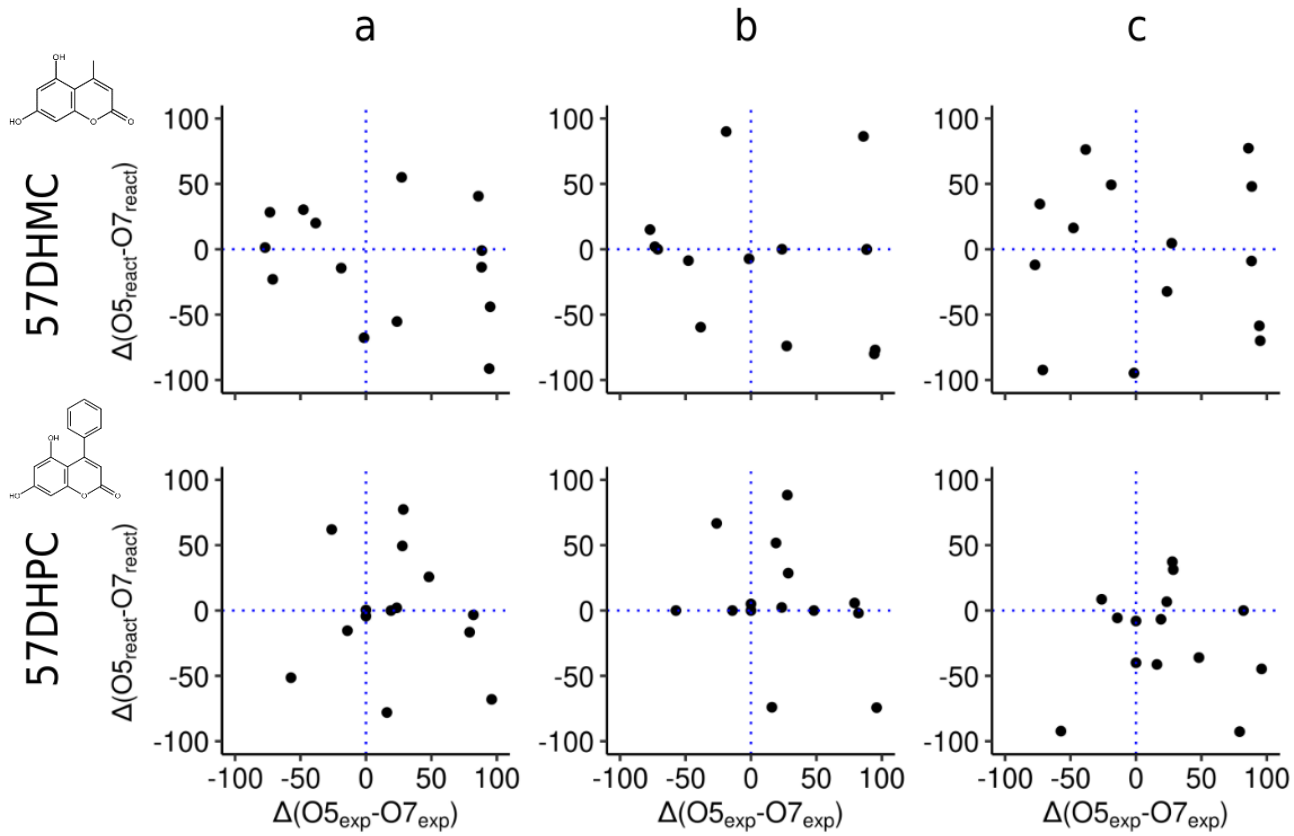


Figure 13. Simulated UGT regioselectivity preferences plotted against experimental preferences. Positive values on x axis indicate excess of O5-glucoside. Positive values on y axis indicate excess of reactive poses for O5-glycosylation. Dotted lines indicate boundaries between preferences. Simulation sets match those described in Table 3-4.

Introduction of restraints during the initial steps of simulation allowed to observe and analyze reactive poses for >90% of the studied systems, while reducing the simulation duration to as low as 2 ns, dramatically decreasing the computational cost. Moreover, it became clear that even unrestrained molecular mechanics allow to visualize potential reactivity for systems which are experimentally inactive. Together with the absence of correlation between regioselective preferences between simulations and experiments, it makes a strong case for a system that is not governed by a simple lock-and-key model.

Case study of UGT9

During the simulations with both coumarin derivatives, it was noticed that *Zm*C0HFA0 had one of the most consistent preferences for glycosylation. In most cases, it preferred O5 of both acceptors, which also agrees with experimental data. Upon closer inspection, E91 was determined to play a key role in acceptor stabilization. While this residue is not unique to UGT9 (Fig. 14, A), the acceptor

binding site positions 57DHMC and 57DHPC in such way that E91 can form an additional hydrogen bond with OH7' hydroxyl, positioning O5 in a productive pose (Fig. 14, B).

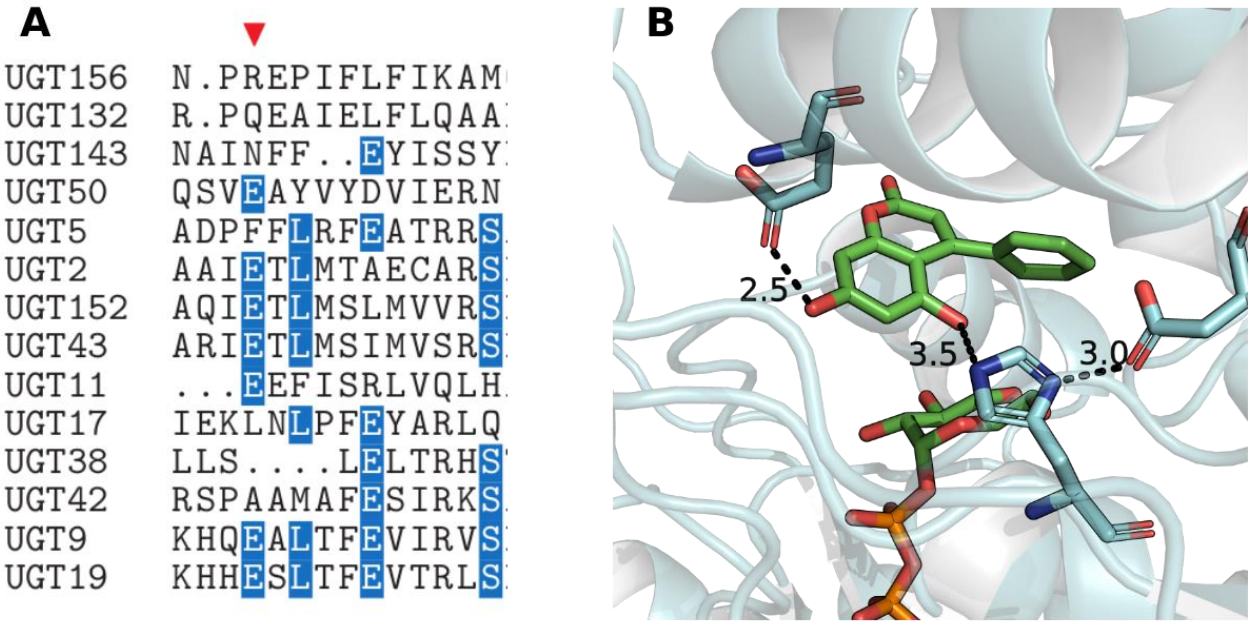


Figure 14. A) Multiple sequence alignment of 14 UGT subset. E91 is marked with a red arrow (residue number from UGT9). Light blue highlight indicates 50% conservation; B) active site of UGT9 with 57DHPC bound, catalytic His-Asp dyad and E91 shown. Distances are measured in Å.

Even in simulations where restrains are put on O7, most of the time the acceptor returns to the position illustrated in Fig 14B. An example trajectory analysis from the simulation set **a** with 57DHPC is presented in Figure 15. When hydrogen bond distance restraint is placed on O5, conformation is kept stable during the entire simulation length. When the restraint is on O7, molecule is kept in a reactive conformation while the force is applied but shifts away immediately after release.

Strict preference of acceptor binding conformation in UGT9 illustrates how some UGTs might follow the lock-and-key principle for regioselectivity. However, according to the studied dataset, it appears to be an exception instead of a rule.

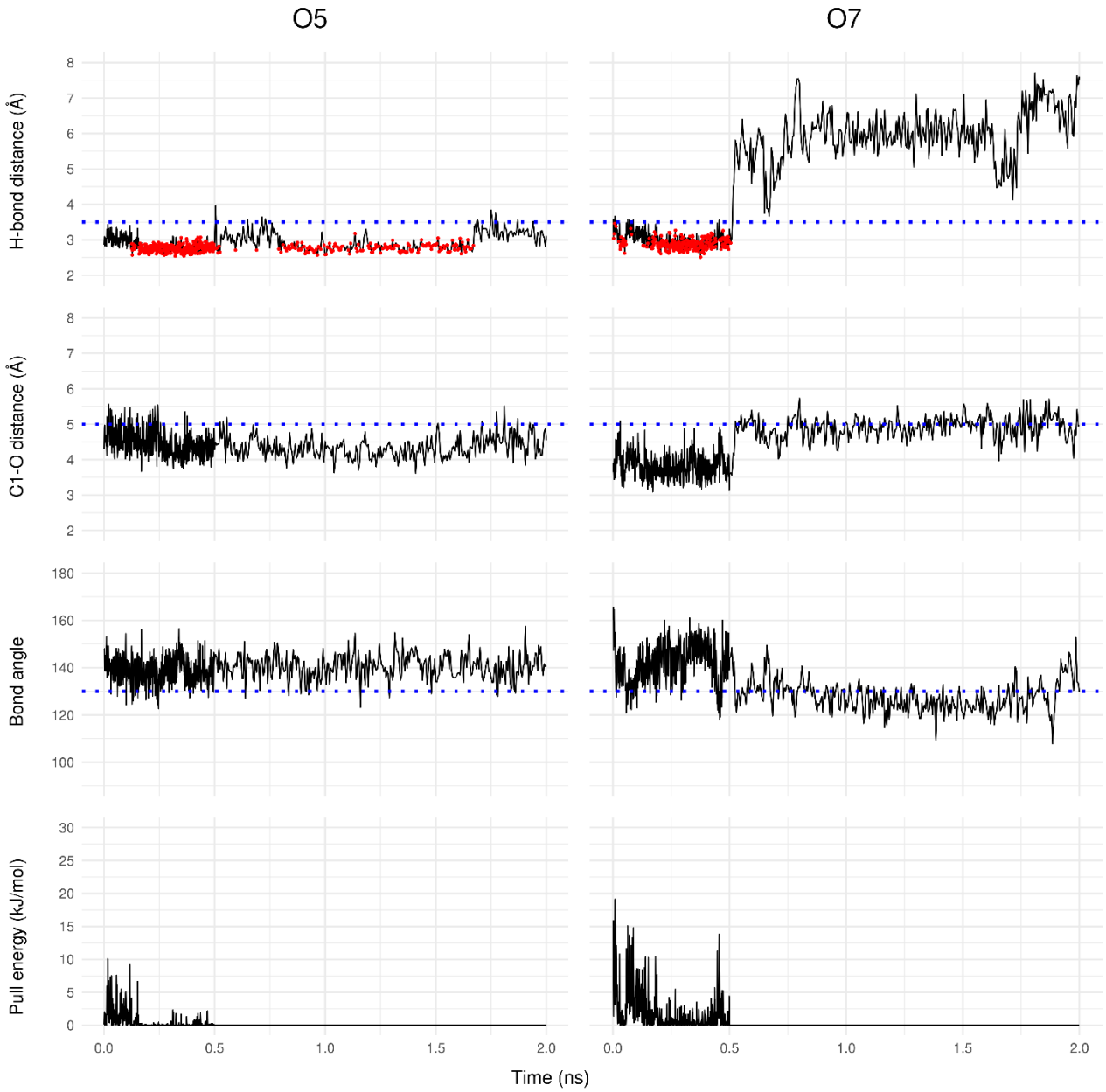


Figure 15. Trajectory analysis of simulations with UGT9 and 57DHPC, when restraint is placed on the hydrogen bond distance. Red points in the top plots mark the formation of hydrogen bond (suitable angle). Threshold values for reactivity criteria are marked with blue dashed lines.

DISCUSSION

Polyphenol glycosides are widely industrially used compounds. However, finding glycosyltransferases that would selectively yield products of interest is still challenging and relies on buying, purifying, and screening entire enzyme libraries. A computational method that would allow accurate screening of these libraries for regioselective, process-specific GTs would save time and resources otherwise spent in the lab.

With recent advances in protein structure predictions, it is now possible to easily obtain high quality glycosyltransferase models, thus unlocking a vast resource that up until now was named as one of the main bottlenecks in GT studies. Combined with ever-improving hardware capabilities and development of computational methods, breakthroughs in understanding how glycosylation works are bound to happen.

In this thesis project we set out to assess whether molecular mechanics can be directly used for predicting regiospecificity of family 1 glycosyltransferases. Specifically, we combined AlphaFold2 models of GT1s with MD simulations to assess their regiospecificity towards phloretin, 5,7-dihydroxy-4-methylcoumarin and 5,7-dihydroxy-4-phenylcoumarin. The main assumption was that if GT1s follow the lock-and-key principle, it would mean that enzyme:acceptor pairs form specific and stable complexes, favoring one glycosylation site over another.

During simulations with phloretin, we tested the possibility to distinguish phloretin *C*-glycosyltransferases from *O*-glycosyltransferases based on substrate binding poses. Results had shown that all 4 *O*-GTs that appeared in a productive conformation at least once, would all prefer *O*-glycosylation. However, 5 C-GTs that appeared reactive at least once, all showed either ambiguous results, or preference for *O*-glycosylation. Three enzymes did not appear reactive at all after 220 ns of accumulated simulation time, which prompted us to start using restrained MD. Additionally, strong inconsistencies were observed both in replicate simulations, and simulations starting from different acceptor docking poses.

Second part of the project focused on UGT regioselectivity towards two coumarin derivatives, 57DHMC and 57DHPC in particular. Here we tested the possibility to predict which one of two glycosylation sites would be preferred by an in-house UGT collection, previously screened against mentioned compounds. Now running much shorter simulations with a restrained MD step to restrain sugar acceptors movement, almost every enzyme was observed to form a reactive complex at least once. However, no correlation was observed between experimentally identified UGT regioselectivities and ones predicted with MD simulations. After running 3 more sets of simulations with a subset of GTs, no improvements were observed, and similar inconsistencies were seen as before. As a case study, UGT9 was investigated further, since its predictions corresponded well with experimental data.

It was found that overall binding site environment allowed both 57DHMC and 57DHPC form an additional hydrogen bond with E91, which stabilized the structure.

Enzymatic selectivity is often more complicated than a simple geometrical problem. Since MD simulations do not allow charge rearrangements or bond breaks/formations, they are intrinsically limited in predicting reactions. Apart from that, miniscule differences between the preferences of specific molecule binding can be too small to pick out from the noise of molecular mechanics. The ability of many UGTs to produce mixtures also suggests that activation energy differences between reactions is small: even a 90/10 ratio indicate a difference in activation energies of only 1.5 kcal.mol^{-1}, for reactions which have typically activation energies of the order of 18 kcal·mol^{-1}[54]. As shown by Serapian and van der Kamp in their work with an actinorhodin ketoreductase, even within a single enzyme and its point mutants, stereoselecticity is determined by different factors which they uncovered by applying MD, binding free-energy calculations and QM/MM simulations[127]. This showcased that precise selectivity control could not be reduced to one method or variable.

We showed that ternary complexes representing Michaelis complexes could be obtained in a computationally cheap fashion for the large majority of assayed systems. Those are particularly helpful to analyze the molecular determinants of a reaction that has been evidenced experimentally. However, given that complexes of experimentally unreactive pairs can be as easily obtained and no stability difference is observed, they are not to be used as an *in silico* predictor of reactivity. Although engineering based on the Michaelis complex might yield good results, stabilization of TS‡ appears to play a fundamental role in regioselectivity control.

MATERIALS AND METHODS

Data

For simulations with coumarin derivatives, 34 plant UGTs, that produce *O*-glycosides with acceptors of interest, were selected (Table 5). For simulations with phloretin, 8 UGTs were selected from the same dataset: 6 producing only *O*-glycoside, 1 producing *C*-glycoside (nothofagin), and 1 hybrid, that produces a proportionate mixture of both *O*-/*C*-glycosides. Also, 5 other nothofagin producing UGTs were selected from literature (Table 6). Structures of all the acceptors were taken from pubchem database[128].

Table 5. Overview of UGTs simulated with coumarin derivatives as acecptors.

Lab name	Name	Uniprot/GenBank ID	Organism
UGT1	*Zm*B6SRY5	B6SRY5	*Zea Mays*
UGT2	*Zm*B4F9H1	B4F9H1	*Zea Mays*
UGT5	*Zm*UGT708A6	A0A096SRM5	*Zea Mays*
UGT9	*Zm*C0HFA0	C0HFA0	*Zea Mays*
UGT10	*Zm*B4FG90	B4FG90	*Zea Mays*
UGT11	*Zm*FK974413	Sr_71E1	*Zea Mays*
UGT16	*Sl*D7S016	D7S016	*Solanum lycopersicum*
UGT17	*Rh*Q4R1I9	Q4R1I9	*Rosa hybrid cultivar*
UGT19	*Os*UGT88C1	Q8LJ11	*Oryza sativa*
UGT36	*Lc*UGT72B10	B6EWZ3	*Lycium barbarum*
UGT38	*Gm*UGT88E3	A6BM07	*Glycine max*
UGT41	*Fi*UGT88A10	D2KY82	*Forsythia intermedia*
UGT42	*Fe*UGT88J1	A0A0A1H7N8	*Fagopyrum esculentum*
UGT43	*Fe*UGT72B19	A0A0A1H7P3	*Fagopyrum esculentum*
UGT46	*Dc*UGT71F5	A7M6I2	*Dianthus caryophyllus*
UGT47	*Ct*UGT71E5	A0A1P8C3B1	*Carthamus tinctorius*
UGT50	*At*UGT88A1	O82383	*Arabidopsis thaliana*
UGT53	*At*UGT71C1	O82381	*Arabidopsis thaliana*
UGT117	*At*UGT74F2	NP_181910.1	*Arabidopsis thaliana*
UGT118	*At*UGT74F1	NP_973682.1	*Arabidopsis thaliana*
UGT121	*Av*UGT5	ARM65438.1	*Aloe vera*
UGT123	*Ac*UGT3	ARM65439.1	*Aloe arborescens*
UGT124	*Sl*Q5CAZ5	CAI62049.1	*Solanum lycopersicum*
UGT132	*At*UGT78D2	OAO89857.1	*Arabidopsis thaliana*
UGT140	*Vu*UGT73C	QCD86231.1	*Vigna unguiculata*

UGT141	*Lb*UGT75L5	BAG80544.1	*Lycium barbarum*
UGT143	*Na*UGT3_3	AQQ16706.1	*Nicotiana attenuate*
UGT144	*Lu*UGT85K6	AFJ52996.1	*Linum usitatissimum*
UGT149	*Gj*UGT14	BAM28983.1	*Gardenia jasminoides*
UGT151	*Nt*UGT	-	*Nicotiana tabacum*
UGT152	*Pt*UGT1	-	*Polygonum tinctorium*
UGT153	*At*UGT72E2	-	*Arabidopsis thaliana*
UGT154	*At*UGT72E3	-	*Arabidopsis thaliana*
UGT156	*Mt*UGT78G1	-	*Medicago truncatula*

Table 6. Overview of UGTs simulated with phloretin as acceptor.

Lab name	Name	UniProt/GenBank ID	Organism	Origin	Activity on phloretin
UGT53	*At*UGT71C1	O82381	*Arabidopsis thaliana*	In-house	*O*-
-	*Zm*CGT2	B6SWX3	*Zea mays*	Literature[129]	*C*-
-	*Pp*CGT1	MK616593	*Pueraria lobate*	Literature[129]	*C*-
UGT5	*Zm*UGT708A6	A0A096SRM5	*Zea mays*	In-house	*C*-
-	*Gg*CGT	MH998596	*Glycyrrhiza glabra*	Literature[130]	*C*-
-	*Mi*CGT	A0A0M4KE44	*Mangifera indica*	Literature[118]	*C*-
-	*Fc*CGT	A0A224AM54	*Fortunella crassifolia*	Literature[131]	*C*-
UGT2	*Zm*B4F9H1	B4F9H1	*Zea mays*	In-house	*O*-
UGT1	*Zm*B6SRY5	B6SRY5	*Zea mays*	In-house	*O*-
UGT42	*Fe*UGT88J1	A0A0A1H7N8	*Fagopyrum esculentum*	In-house	*O*-
UGT17	*Rh*Q4R1I9	Q4R1I9	*Rosa hybrid cultivar*	In-house	*O*-
UGT10	*Zm*B4FG90	B4FG90	*Zea mays*	In-house	*O*-
UGT9	*Zm*C0HFA0	C0HFA0	*Zea mays*	In-house	*O-/C*-

Preparation of ternary enzyme complexes

Even though some of the selected enzymes have their crystal structures solved and deposited in the PDB, it was decided to model all of the structures with AlphaFold2 (AF2) for the following reasons: 1) AF2 produces complete structural models, while crystal structures from PDB often have missing parts that have to be reconstructed prior to molecular dynamics simulations; 2) we have previously confirmed that AF2 can predict structures of GT1 family glycosyltransferases in nearly exact precision to experimental methods.

Protein structural models were generated by using AlphaFold v2.0 setup on Computerome, using all available structural homologs, and the database search preset was set to "reduced_dbs"[84]. After

predictions, built-in model relaxation was performed. Only the highest ranking (in pLDDT score) models were used downstream. Low scoring (pLDDT < 50) unstructured terminal regions were shortened (usually 1–10 amino acids) to reduce simulated system size in MD. Binary complexes of protein and sugar donor were obtained by structurally aligning protein model structures on the crystal structure of *Pt*UGT1 from *Polygonum tinctorium*, which has a bound UDP-glucose molecule in its active site (6SU6.pdb)[54]. The flexible acceptor molecules were added by docking into the acceptor binding site of the rigid binary complexes, using GNINA v1.0.1 software[132], a fork of SMINA[133], itself a fork of AutoDock Vina[91]. Docking grid was defined by a manually placed phloretin molecule in the 6SU6 binding site, using the *autobox_extend* option that expands the grid to a cube of longest distance between any two atoms in the placed phloretin molecule, and adding 4 Å to the final grid on all sides. Since 57DHMC and 57DHPC molecules are smaller than phloretin, only 3 Å were added to the final grid. Default settings of GNINA were otherwise used, meaning that positioning of acceptors was performed purely by AutoDock Vina, and only the final rescoring of poses was performed by a GNINA convolutional neural network. Best poses were selected by low binding energy and chemical intuition.

PyMOL (v2.4.0)[134] was used for superimposition and visualization of resulting structures.

Conventional molecular dynamics

Notably, completely conventional MD was only performed with phloretin as an acceptor. Simulations were performed with the GROMACS (v2021.3) software[135]. Proteins were parametrized with Amber ff14SB forcefield[136], acceptors with gaff2 forcefield[137], GLYCAM06[138] was used for the glucose moiety. The protonation state of all residues was assigned according to pH 7, and protonation of His residues was additionally assessed according to their chemical environment. A manual check was performed to adjust the protonation of catalytic His residues, to make sure that the proton is located at N_δ Substrates were prepared with antechamber module and converted to GROMACS format using the acpype Python package[139,140]. AM1-BCC partial atomic charges were used for sugar acceptors. The complex system was solvated in TIP3P water molecules[141] in a cubic box with minimum 10 Å edge distance. Random water molecules were replaced by Na^+ and Cl^- ions to neutralize the system.

Long-range electrostatics were treated with the particle-mesh Ewald method with a cutoff distance of 12 Å[142]**.** Van der Waals interactions were treated in a Verlet scheme with a cutoff distance of 12 Å and a switching function for the forces starting at 10 Å[143]. Hydrogen bonds were restrained using the LINCS algorithm[144]. Protein with substrates and water with ions were coupled to individual heat baths with a Bussi–Donadio–Parrinello thermostat[145]. Pressure coupling was done in either Berendsen[146] (simulations with phloretin) or Parrinello-Rahman[117] (simulations with 57DHMC and

57DHPC) barostat. Energy minimization was performed with steepest-descent algorithm for 50000 steps. NVT equilibration was performed for 100 ps with a reference temperature of 300 K, with restraints placed on protein and substrates. Afterwards, NPT equilibration with identical restrains was performed for 100 ps with a reference pressure of 1 bar. For the production run, the same barostat was kept on, and the restrains from substrate and protein were released. Production runs were carried out with varying lengths (20–100 ns), taking snapshots every 10 ps.

Trajectories were analyzed with built-in GROMACS command-line tools and visualized with VMD and PyMOL. Plots were generated by using ggplot and pubr packages in R.

Restrained molecular dynamics

Restrained MD simulations used the same system preparation and energy minimization protocol as conventional MD. All other system parameters are identical to conventional MD unless stated otherwise.

At the start of the production run, flat-bottomed distance restraints of 5000 $kJ/mol^{-1}nm^{-1}$ were placed on one or both of nucleophilic attack (4 Å) and/or deprotonation/hydrogen bond (2.8 Å) distances to simulate the process of substrate binding, therefore reducing the dependency on initial simulation conditions. Flat-bottomed restraint adds a simple harmonic potential to the system, except after the restrained distance drops below the reference value, at which point no force is applied. After 0.5 ns, restraints were removed, and simulations continued until 2 ns, taking snapshots every 5 ps. For every enzyme:acceptor:restraint combination, two parallel simulations were executed from the identical starting point after the NPT equilibration – one placing restraints on one possible glycosylation site, and one on another.

Data availability

All scripts required to prepare and run the simulations are provided in the GitHub repository: github.com/vaitkusd/GT1_MD_scripts. It also includes all used acceptor structures and UGT models, as well as plotted trajectories of simulations analyzed in this thesis.

ACKNOWLEDGEMENTS

My deepest gratitude goes to David Teze, who never got tired of sharing ideas and explaining fundamentals of chemistry, enzymology, and cheese. Special thanks go to Folmer, Natalia, Ditte, and the entire Enzyme Engineering & Structural Biology group at DTU Biosustain, who accepted me as a worthy member of the group and involved me in multiple research projects. Also, I wanted to thank Lluis Raich for expert MD insights, and Amelie Stein for providing internal support. Finally, I am thankful to James – a real human Brit who is never hesitant to proofread his friends' theses.

REFERENCES

1. Woodley, J. M., Breuer, M. & Mink, D. A future perspective on the role of industrial biotechnology for chemicals production. *Chem. Eng. Res. Des.* **91**, 2029–2036 (2013).
2. Leuchtenberger, W., Huthmacher, K. & Drauz, K. Biotechnological production of amino acids and derivatives: current status and prospects. *Appl. Microbiol. Biotechnol.* **69**, 1–8 (2005).
3. Chen, Y. & Nielsen, J. Biobased organic acids production by metabolically engineered microorganisms. *Curr. Opin. Biotechnol.* **37**, 165–172 (2016).
4. Becker, J., Lange, A., Fabarius, J. & Wittmann, C. Top value platform chemicals: bio-based production of organic acids. *Curr. Opin. Biotechnol.* **36**, 168–175 (2015).
5. Sheldon, R. A. & Woodley, J. M. Role of biocatalysis in sustainable chemistry. *Chem. Rev.* **118**, 801–838 (2018).
6. Kell, D. B., Swainston, N., Pir, P. & Oliver, S. G. Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis. *Trends Biotechnol.* **33**, 237–246 (2015).
7. Grant, C. *et al.* Identification and use of an alkane transporter plug-in for applications in biocatalysis and whole-cell biosensing of alkanes. *Sci. Rep.* **4**, 5844 (2014).
8. Goldsmith, M. & Tawfik, D. S. Enzyme engineering: reaching the maximal catalytic efficiency peak. *Curr. Opin. Struct. Biol.* **47**, 140–150 (2017).
9. Hughes, G. & Lewis, J. C. Introduction: Biocatalysis in Industry. *Chem. Rev.* **118**, 1–3 (2018).
10. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425 (1976).
11. Singh, R. K., Tiwari, M. K., Singh, R. & Lee, J.-K. From protein engineering to immobilization: promising strategies for the upgrade of industrial enzymes. *Int. J. Mol. Sci.* **14**, 1232–1277 (2013).
12. Kan, S. B. J., Lewis, R. D., Chen, K. & Arnold, F. H. Directed evolution of cytochrome c for carbon–silicon bond formation: Bringing silicon to life. *Science* **354**, 1048–1051 (2016).
13. Kan, S. B. J., Huang, X., Gumulya, Y., Chen, K. & Arnold, F. H. Genetically programmed chiral organoborane synthesis. *Nature* **552**, 132–136 (2017).
14. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
15. Reetz, M. T. What are the limitations of enzymes in synthetic organic chemistry? *Chem. Rec. N. Y. N* **16**, 2449–2459 (2016).
16. Romero, P. A. & Arnold, F. H. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* **10**, 866–876 (2009).
17. Zeymer, C. & Hilvert, D. Directed evolution of protein catalysts. *Annu. Rev. Biochem.* **87**, 131–157 (2018).
18. Choi, J.-M., Han, S.-S. & Kim, H.-S. Industrial applications of enzyme biocatalysis: Current status and future aspects. *Biotechnol. Adv.* **33**, 1443–1454 (2015).
19. Wu, L., Qin, L., Nie, Y., Xu, Y. & Zhao, Y.-L. Computer-aided understanding and engineering of enzymatic selectivity. *Biotechnol. Adv.* **54**, 107793 (2022).
20. Ebert, M. C. & Pelletier, J. N. Computational tools for enzyme improvement: why everyone can – and should – use them. *Curr. Opin. Chem. Biol.* **37**, 89–96 (2017).
21. Jiang, W. & Fang, B. Synthesizing chiral drug intermediates by biocatalysis. *Appl. Biochem. Biotechnol.* **192**, 146–179 (2020).
22. Calcaterra, A. & D'Acquarica, I. The market of chiral drugs: Chiral switches versus de novo enantiomerically pure compounds. *J. Pharm. Biomed. Anal.* **147**, 323–340 (2018).

23. Fischer, E. Einfluss der configuration auf die wirkung der enzyme. *Berichte Dtsch. Chem. Ges.* **27**, 2985–2993 (1894).
24. Ema, T. Mechanism of enantioselectivity of lipases and other synthetically useful hydrolases. *Curr. Org. Chem.* **8**, 1009–1025 (2004).
25. Koshland, D. E. Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl. Acad. Sci.* **44**, 98–104 (1958).
26. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
27. Prokop, Z. *et al.* Engineering of protein tunnels: keyhole-lock-key model for catalysis by the enzymes with active sites. in 421–464 (2012).
28. Tripathi, A. & Bankaitis, V. A. Molecular docking: from lock and key to combination lock. *J. Mol. Med. Clin. Appl.* **2**, 10.16966/2575-0305.106 (2017).
29. Thuan, N. H. & Sohng, J. K. Recent biotechnological progress in enzymatic synthesis of glycosides. *J. Ind. Microbiol. Biotechnol.* **40**, 1329–1356 (2013).
30. Quideau, S., Deffieux, D., Douat-Casassus, C. & Pouységu, L. Plant polyphenols: chemical properties, biological activities, and synthesis. *Angew. Chem. Int. Ed.* **50**, 586–621 (2011).
31. Tufarelli, V., Casalino, E., D'Alessandro, A. G. & Laudadio, V. Dietary phenolic compounds: biochemistry, metabolism and significance in animal and human health. *Curr. Drug Metab.* **18**, (2018).
32. Dandriyal, J., Singla, R., Kumar, M. & Jaitak, V. Recent developments of C-4 substituted coumarin derivatives as anticancer agents. *Eur. J. Med. Chem.* **119**, 141–168 (2016).
33. Peng, X.-M., Damu, G. L. V. & Zhou, C.-H. Current developments of coumarin compounds in medicinal chemistry. *Curr. Pharm. Des.* **19**, 3884–3930.
34. Behzad, S. *et al.* Health effects of phloretin: from chemistry to medicine. *Phytochem. Rev.* **16**, 527–533 (2017).
35. Yang, D., Wang, T., Long, M. & Li, P. Quercetin: Its Main Pharmacological Activity and Potential Application in Clinical Medicine. https://www.hindawi.com/journals/omcl/2020/8825387/.
36. Di Lorenzo, C., Colombo, F., Biella, S., Stockley, C. & Restani, P. Polyphenols and human health: the role of bioavailability. *Nutrients* **13**, 273 (2021).
37. Eran Nagar, E., Okun, Z. & Shpigelman, A. Digestive fate of polyphenols: updated view of the influence of chemical structure and the presence of cell wall material. *Curr. Opin. Food Sci.* **31**, 38–46 (2020).
38. Rawat, P. *et al.* Synthesis and antihyperglycemic activity of phenolic *C*-glycosides. *Bioorg. Med. Chem. Lett.* **21**, 228–233 (2011).
39. Teng, H. & Chen, L. Polyphenols and bioavailability: an update. *Crit. Rev. Food Sci. Nutr.* **59**, 2040–2051 (2019).
40. Yepremyan, A., Salehani, B. & Minehan, T. G. Concise total syntheses of aspalathin and nothofagin. *Org. Lett.* **12**, 1580–1583 (2010).
41. Bungaruang, L., Gutmann, A. & Nidetzky, B. Leloir glycosyltransferases and natural product glycosylation: biocatalytic synthesis of the C-glucoside nothofagin, a major antioxidant of redbush herbal tea. *Adv. Synth. Catal.* **355**, 2757–2763 (2013).
42. Nidetzky, B., Gutmann, A. & Zhong, C. Leloir glycosyltransferases as biocatalysts for chemical production. *ACS Catal.* **8**, 6283–6300 (2018).
43. Thibodeaux, C. J., Melançon, C. E. & Liu, H. Unusual sugar biosynthesis and natural product glycodiversification. *Nature* **446**, 1008–1016 (2007).

44. Yamamoto, F., Clausen, H., White, T., Marken, J. & Hakomori, S. Molecular genetic basis of the histo-blood group ABO system. *Nature* **345**, 229–233 (1990).
45. Hanukoglu, I. Proteopedia: Rossmann fold: a beta-alpha-beta fold at dinucleotide binding sites. *Biochem. Mol. Biol. Educ.* **43**, 206–209 (2015).
46. Hu, Y. & Walker, S. Remarkable structural similarities between diverse glycosyltransferases. *Chem. Biol.* **9**, 1287–1296 (2002).
47. Lim, E.-K. Plant glycosyltransferases: their potential as novel biocatalysts. *Chem. – Eur. J.* **11**, 5486–5494 (2005).
48. Gloster, T. M. Advances in understanding glycosyltransferases from a structural perspective. *Curr. Opin. Struct. Biol.* **28**, 131–141 (2014).
49. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* **37**, D233-238 (2009).
50. Caputi, L., Lim, E.-K. & Bowles, D. J. Discovery of new biocatalysts for the glycosylation of terpenoid scaffolds. *Chem. – Eur. J.* **14**, 6656–6662 (2008).
51. DeAngelis, P. L., Liu, J. & Linhardt, R. J. Chemoenzymatic synthesis of glycosaminoglycans: Re-creating, re-modeling and re-designing nature's longest or most complex carbohydrate chains. *Glycobiology* **23**, 764–777 (2013).
52. Louveau, T. & Osbourn, A. The sweet side of plant-specialized metabolism. *Cold Spring Harb. Perspect. Biol.* **11**, a034744 (2019).
53. Ross, J., Li, Y., Lim, E.-K. & Bowles, D. J. Higher plant glycosyltransferases. *Genome Biol.* **2**, reviews3004.1-reviews3004.6 (2001).
54. Teze, D. *et al. O*-/*N*-/*S*-specificity in glycosyltransferase catalysis: from mechanistic understanding to engineering. *ACS Catal.* **11**, 1810–1815 (2021).
55. Osmani, S. A., Bak, S. & Møller, B. L. Substrate specificity of plant UDP-dependent glycosyltransferases predicted from crystal structures and homology modeling. *Phytochemistry* **70**, 325–347 (2009).
56. Liang, D.-M. *et al.* Glycosyltransferases: mechanisms and applications in natural product development. *Chem. Soc. Rev.* **44**, 8350–8374 (2015).
57. Hirade, Y. *et al.* Identification and functional analysis of 2-hydroxyflavanone *C* -glucosyltransferase in soybean (*Glycine max*). *FEBS Lett.* **589**, 1778–1786 (2015).
58. Putkaradze, N., Teze, D., Fredslund, F. & Welner, D. H. Natural product *C*-glycosyltransferases – a scarcely characterised enzymatic activity with biotechnological potential. *Nat. Prod. Rep.* **38**, 432–443 (2021).
59. Teze, D., Bidart, G. N. & Welner, D. H. Family 1 glycosyltransferases (GT1, UGTs) are subject to dilution-induced inactivation and low chemo stability toward their own acceptor substrates. *Front. Mol. Biosci.* **9**, 909659 (2022).
60. He, J.-B. *et al.* Molecular and structural characterization of a promiscuous *C*-glycosyltransferase from trollius chinensis. *Angew. Chem.* **131**, 11637–11644 (2019).
61. Feng, J. *et al.* Regio- and stereospecific *O*-glycosylation of phenolic compounds catalyzed by a fungal glycosyltransferase from mucor hiemalis. *Adv. Synth. Catal.* **359**, 995–1006 (2017).
62. George Thompson, A. M., Iancu, C. V., Neet, K. E., Dean, J. V. & Choe, J. Differences in salicylic acid glucose conjugations by UGT74F1 and UGT74F2 from Arabidopsis thaliana. *Sci. Rep.* **7**, 46629 (2017).
63. Hiromoto, T. *et al.* Structural basis for acceptor-substrate recognition of UDP-glucose: anthocyanidin 3-O-glucosyltransferase from Clitoria ternatea. *Protein Sci.* **24**, 395–407 (2015).

64. Lim, E.-K. Evolution of substrate recognition across a multigene family of glycosyltransferases in Arabidopsis. *Glycobiology* **13**, 139–145 (2003).
65. Yang, M. *et al.* Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat. Chem. Biol.* **14**, 1109–1117 (2018).
66. Kroll, A., Ranjan, S., Engqvist, M. K. M. & Lercher, M. J. *The substrate scopes of enzymes: a general prediction model based on machine and deep learning*. http://biorxiv.org/lookup/doi/10.1101/2022.05.24.493213 (2022) doi:10.1101/2022.05.24.493213.
67. Holm, L. Using Dali for Protein Structure Comparison. *Methods Mol. Biol. Clifton NJ* **2112**, 29–42 (2020).
68. Kempen, M. van *et al.* Foldseek: fast and accurate protein structure search. 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
69. Warshel, A. & Levitt, M. Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227–249 (1976).
70. Cadet, F. *et al.* A machine learning approach for reliable prediction of amino acid interactions and its application in the directed evolution of enantioselective enzymes. *Sci. Rep.* **8**, 16757 (2018).
71. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* **16**, 687–694 (2019).
72. Li, G. *et al.* Learning deep representations of enzyme thermal adaptation. 2022.03.14.484272 (2022) doi:10.1101/2022.03.14.484272.
73. Kroll, A., Engqvist, M. K. M., Heckmann, D. & Lercher, M. J. Deep learning allows genome-scale prediction of Michaelis constants from structural features. *PLOS Biol.* **19**, e3001402 (2021).
74. Li, F. *et al.* Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction. *Nat. Catal.* **5**, 662–672 (2022).
75. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine learning in enzyme engineering. *ACS Catal.* **10**, 1210–1223 (2020).
76. Repecka, D. *et al.* Expanding functional protein sequence spaces using generative adversarial networks. *Nat. Mach. Intell.* **3**, 324–333 (2021).
77. Wang, J. *et al.* Scaffolding protein functional sites using deep learning. *Science* **377**, 387–394 (2022).
78. Brini, E., Simmerling, C. & Dill, K. Protein storytelling through physics. *Science* **370**, eaaz3041 (2020).
79. Shindyalov, I. N., Kolchanov, N. A. & Sander, C. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng.* **7**, 349–358 (1994).
80. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* **6**, e28766 (2011).
81. Fiser, A. & Šali, A. Modeller: generation and refinement of homology-based protein structure models. in *Methods in Enzymology* vol. 374 461–491 (Academic Press, 2003).
82. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42**, W252–W258 (2014).
83. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
84. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

85. Akdel, M. *et al. A structural biology community assessment of AlphaFold 2 applications*. http://biorxiv.org/lookup/doi/10.1101/2021.09.26.461876 (2021) doi:10.1101/2021.09.26.461876.
86. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
87. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
88. Tao, X. *et al.* Recent developments in molecular docking technology applied in food science: a review. *Int. J. Food Sci. Technol.* **55**, 33–45 (2020).
89. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288 (1982).
90. Spyrakis, F., Cozzini, P. & Kellogg, G. E. Docking and Scoring in Drug Discovery. in *Burger's Medicinal Chemistry and Drug Discovery* 601–684 (John Wiley & Sons, Ltd, 2010). doi:10.1002/0471266949.bmc140.
91. Trott, O. & Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* NA-NA (2009) doi:10.1002/jcc.21334.
92. Amaro, R. E. *et al.* Ensemble docking in drug discovery. *Biophys. J.* **114**, 2271–2278 (2018).
93. Ain, Q. U., Aleksandrova, A., Roessler, F. D. & Ballester, P. J. Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **5**, 405–424 (2015).
94. Hamza, A., Zhao, X., Tong, M., Tai, H.-H. & Zhan, C.-G. Novel human mPGES-1 inhibitors identified through structure-based virtual screening. *Bioorg. Med. Chem.* **19**, 6077–6086 (2011).
95. Kua, J., Zhang, Y. & McCammon, J. A. Studying enzyme binding specificity in acetylcholinesterase using a combined molecular dynamics and multiple docking approach. *J. Am. Chem. Soc.* **124**, 8260–8267 (2002).
96. Huang, J. *et al.* Exploring the catalytic function and active sites of a novel *C*-glycosyltransferase from Anemarrhena asphodeloides. *Synth. Syst. Biotechnol.* **7**, 621–630 (2022).
97. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
98. Lemkul, J. A. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J. Comput. Mol. Sci.* **1**, 5068–5068 (2019).
99. Adcock, S. A. & McCammon, J. A. Molecular Dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106**, 1589–1615 (2006).
100. Dama, J. F., Jin, J. & Voth, G. A. The theory of ultra-coarse-graining. 3. coarse-grained sites with rapid local equilibrium of internal states. *J. Chem. Theory Comput.* **13**, 1010–1022 (2017).
101. Latorraca, N. R., Venkatakrishnan, A. J. & Dror, R. O. GPCR dynamics: structures in motion. *Chem. Rev.* **117**, 139–155 (2017).
102. Melvin, R. L. *et al.* Uncovering large-scale conformational change in molecular dynamics without prior knowledge. *J. Chem. Theory Comput.* **12**, 6130–6146 (2016).
103. Lindorff-Larsen, K., Maragakis, P., Piana, S. & Shaw, D. E. Picosecond to millisecond structural dynamics in human ubiquitin. *J. Phys. Chem. B* **120**, 8313–8320 (2016).
104. Li, J. *et al.* Near-perfect control of the regioselective glucosylation enabled by rational design of glycosyltransferases. *Green Synth. Catal.* **2**, 45–53 (2021).

105. Ricci-Lopez, J., Aguila, S. A., Gilson, M. K. & Brizuela, C. A. Improving structure-based virtual screening with ensemble docking and machine learning. *J. Chem. Inf. Model.* **61**, 5362–5376 (2021).
106. Stone, J. E. *et al.* Evaluation of emerging energy-efficient heterogeneous computing platforms for biomolecular and cellular simulation workloads. in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)* 89–100 (IEEE, 2016). doi:10.1109/IPDPSW.2016.130.
107. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
108. Salmaso, V. & Moro, S. Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: an overview. *Front. Pharmacol.* **9**, (2018).
109. Chen, M. Collective variable-based enhanced sampling and machine learning. *Eur. Phys. J. B* **94**, 211 (2021).
110. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
111. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **120**, 11919–11929 (2004).
112. Isralewitz, B., Gao, M. & Schulten, K. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **11**, 224–230 (2001).
113. Kari, J. *et al.* Physical constraints and functional plasticity of cellulases. *Nat. Commun.* **12**, 3847 (2021).
114. Izrailev, S., Crofts, A. R., Berry, E. A. & Schulten, K. Steered Molecular Dynamics Simulation of the Rieske Subunit Motion in the Cytochrome bc1 Complex. *Biophys. J.* **77**, 1753–1768 (1999).
115. Lüdemann, S. K., Lounnas, V. & Wade, R. C. How do substrates enter and products exit the buried active site of cytochrome P450cam. *J. Mol. Biol.* **303**, 797–811 (2000).
116. Torrie, G. M. & Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **23**, 187–199 (1977).
117. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci.* **99**, 12562–12566 (2002).
118. Chen, D. *et al.* Probing the catalytic promiscuity of a regio- and stereospecific *C*-glycosyltransferase from *Mangifera indica*. *Angew. Chem. Int. Ed.* **54**, 12678–12682 (2015).
119. Härle, J. *et al.* Rational design of an aryl-*C*-glycoside catalyst from a natural product *O*-glycosyltransferase. *Chem. Biol.* **18**, 520–530 (2011).
120. Tam, H. K. *et al.* Structural characterization of *O*- and *C*-glycosylating variants of the landomycin glycosyltransferase LanGT2. *Angew. Chem. Int. Ed Engl.* **54**, 2811–2815 (2015).
121. Gutmann, A. & Nidetzky, B. Switching between *O*- and *C*-glycosyltransferase through exchange of active-site motifs. *Angew. Chem. Int. Ed.* **51**, 12879–12883 (2012).
122. Brazier-Hicks, M. *et al.* Characterization and engineering of the bifunctional N- and O-glucosyltransferase involved in xenobiotic metabolism in plants. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20238–20243 (2007).
123. Foshag, D., Campbell, C. & Pawelek, P. D. The *C*-glycosyltransferase IroB from pathogenic *Escherichia coli*: Identification of residues required for efficient catalysis. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1844**, 1619–1630 (2014).

124. Huang, W., He, Y., Jiang, R., Deng, Z. & Long, F. Functional and structural dissection of a plant steroid 3-*O*-glycosyltransferase facilitated the engineering enhancement of sugar donor promiscuity. *ACS Catal.* **12**, 2927–2937 (2022).
125. Lairson, L. L., Henrissat, B., Davies, G. J. & Withers, S. G. Glycosyltransferases: structures, functions, and mechanisms. *Annu. Rev. Biochem.* **77**, 521–555 (2008).
126. Chen, D. *et al.* Biocatalytic *C*-glucosylation of coumarins using an engineered *C*-glycosyltransferase. *Org. Lett.* **20**, 1634–1637 (2018).
127. Serapian, S. A. & van der Kamp, M. W. Unpicking the cause of stereoselectivity in actinorhodin ketoreductase variants with atomistic simulations. *ACS Catal.* **9**, 2381–2394 (2019).
128. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
129. Sun, Y. *et al.* Pathway-specific enzymes from bamboo and crop leaves biosynthesize anti-nociceptive C-glycosylated flavones. *Commun. Biol.* **3**, 1–11 (2020).
130. Zhang, M. *et al.* Functional characterization and structural basis of an efficient Di-*C*-glycosyltransferase from *Glycyrrhiza glabra*. *J. Am. Chem. Soc.* **142**, 3506–3512 (2020).
131. Ito, T., Fujimoto, S., Suito, F., Shimosaka, M. & Taguchi, G. *C*-Glycosyltransferases catalyzing the formation of di-*C*-glucosyl flavonoids in citrus plants. *Plant J. Cell Mol. Biol.* **91**, 187–198 (2017).
132. McNutt, A. T. *et al.* GNINA 1.0: molecular docking with deep learning. *J. Cheminformatics* **13**, 43 (2021).
133. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
134. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
135. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
136. Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
137. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25**, 1157–1174 (2004).
138. Kirschner, K. N. *et al.* GLYCAM06: A generalizable biomolecular force field. Carbohydrates: GLYCAM06. *J. Comput. Chem.* **29**, 622–655 (2008).
139. Sousa da Silva, A. W. & Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. Notes* **5**, 367 (2012).
140. Bernardi, A., Faller, R., Reith, D. & Kirschner, K. N. ACPYPE update for nonuniform 1–4 scale factors: Conversion of the GLYCAM06 force field from AMBER to GROMACS. *SoftwareX* **10**, 100241 (2019).
141. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
142. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
143. Verlet, L. Computer 'Experiments' on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **159**, 98–103 (1967).
144. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).

145. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
146. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).