DEVELOPING NEW INFORMATICS APPROACHES FOR INVESTIGATING SEQUENCE-STRUCTURE-FUNCTION AND EVOLUTIONARY RELATIONSHIPS IN GLYCOSYLTRANSFERASES

by

RAHIL TAUJALE

(Under the Direction of Natarajan Kannan and Arthur S. Edison)

ABSTRACT

Glycosyltransferases (GTs) play fundamental roles in nearly all cellular processes through the biosynthesis of complex carbohydrates and glycosylation of diverse protein and small molecule substrates. Although prevalent across the tree of life, the evolutionary basis for the complex and diverse modes of GT catalytic functions remain enigmatic. This is mainly due to the extensive structural and functional diversification of GTs that presents a major challenge in mapping the relationships connecting sequence, structure, fold and function.

In this dissertation, I develop and apply a combination of established and novel tools for large scale sequence based comparisons of glycosyltransferases across the tree of life. Using well curated structure-based sequence alignment profiles, I first align over half a million GT sequences adopting the GT-A fold to identify the conserved GT-A core and define the minimal active site and hydrophobic components required for GT-A function. Based on this conserved core, I build a phylogenetic framework connecting diverse GT-A families and propose a new evolutionary constraint based classification of GT-A sequences into evolutionarily related groups. Next, I use advances in deep learning to develop a GT fold classification and prediction model that extends the analysis from GT-A to other known and novel folds. I build this highly interpretable model to identify the core conserved features of all three major GT folds and predict GT families that are likely to adopt novel folds. Finally, I compile all the diverse datasets generated during these

studies into an interactive data analytics platform that can be used to infer novel hypotheses about GT-A fold enzymes. The results, data and tools developed during these projects for my dissertation have been helpful in multiple collaborations on different projects. These projects have also resulted in several scientific publications. In addition, I've been involved in developing bioinformatics tools for the analysis of metabolomics data that have also led to scientific publications. These projects have been summarized in the Appendix section.

INDEX WORDS: Glycosyltransferases, evolution, artificial intelligence, sequence alignment, data visualization

DEVELOPING NEW INFORMATICS APPROACHES FOR INVESTIGATING SEQUENCE-STRUCTURE-FUNCTION AND EVOLUTIONARY RELATIONSHIPS IN GLYCOSYLTRANSFERASES

by

RAHIL TAUJALE

B.Tech., Kathmandu University, Nepal, 2011

M.S., Northern Illinois University, 2015

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2021

DEVELOPING NEW INFORMATICS APPROACHES FOR INVESTIGATING SEQUENCE-STRUCTURE-FUNCTION AND EVOLUTIONARY RELATIONSHIPS IN GLYCOSYLTRANSFERASES

by

RAHIL TAUJALE

Major Professors:	Natarajan Kannan
	Arthur S. Edison
Committee:	Kelley W. Moremen
	Christopher M. West
	Jonathan Arnold

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2021

For my family. Thanks for always being there for me.

—♥—

In memory of my maternal grandpa, Tirtha Bahadur Shrestha.

That first computer you bought for us was all the motivation I ever needed.

&

In memory of my paternal grandpa, Baikuntha Taujale.

Your words of encouragement made this work possible.

ACKNOWLEDGEMENTS

I am grateful to my wonderful advisors, Dr. Natarajan Kannan and Dr. Art Edison, for their continued guidance and support during my Ph.D. study, without which this dissertation would not have been possible. I would also like to express my deepest appreciation to my committee members Dr. Kelley Moremen, Dr. Chris West and Dr. Jonathon Arnold, who have provided invaluable advice and guidance throughout my research. I would also like to thank all past and present members of the Kannan lab and the Edison lab who I've had the pleasure of working with. I also thank the Glycoscience Training Program and my fellow trainees for their support and NIH for the funding.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

1.1 Motivation

Complex carbohydrates (glycans) are one of the most abundant biopolymers that play essential roles in biological processes ranging from membrane organization, cell adhesion and interaction, pathogenicity, immunity, glycoprotein folding and structural stability [1]. In more than three billion years of evolution, every living cell has relied on glycans for many aspects of their crucial functions [1,2], much like other macromolecules of life (nucleic acids, proteins and lipids). However, in comparison to nucleic acids or amino acids, oligomerization of monosaccharides can result in a much greater number of structural variations. As an example, nucleic acids and proteins can have 6 different variations in a trimer compared to the estimates of 27,648 variations in a trisaccharide [3], highlighting the sheer diversity of possible carbohydrate molecules that could be encoded within cells. While biosynthesis of DNA, RNA and proteins are safely encoded in the genetic templates that are passed from parents to offspring, glycans undergo a non-template derived biosynthesis, despite this incredible variation. Instead, the fidelity of synthesizing these highly complex molecules rests on a large family of enzymes called glycosyltransferases (GTs), that act as the enzymatic templates for catalyzing the precise formation of glycosidic bonds that connect sugar monomeric units.

Because of this vital role, GTs have been implicated in many aspects of human life. Defects in GT functions have been associated with multiple diseases such as congenital disorders of glycosylation, neurodegenerative diseases and cancer [4–12]. GTs are also important in food and agriculture industries, animal husbandry, biofuel and vaccine production [13–21]. Understanding the natural roles of GTs in these sectors and using this knowledge to engineer and design synthetic GTs that are optimized for desired tasks has been heavily pursued for generating desired effects on downstream carbohydrate molecules [22–25]. However, these attempts have been hindered by the lack of understanding of the sequence-structure-function relationships connecting the diverse families of GT enzymes. To facilitate biosynthesis of diverse classes of glycans, GTs themselves have undergone extensive evolutionary changes that has led to large families of seemingly unrelated GTs performing very similar functions, making it extremely difficult to systematically relate shared mechanisms for function and regulation.

Here, we seek to address this gap in knowledge by applying rigorous statistical models along with evolutionary studies and deep learning to gain insights into shared conserved regions within sequences governing GT function and regulation. By applying these methods in a large collection of over half a million GT sequences, we provide a comprehensive overview of the conserved catalytic and hydrophobic core specific to the GT-A fold and develop a framework for fold prediction of novel GTs. We compile our findings into a user-friendly tool and make it available for general use, which will serve as a valuable resource for the glycoscience community.

1.2 Background

1.2.1 Complex carbohydrates

Carbohydrates are one of the four major classes of biological macromolecules found in living cells. In their basic form, they serve as a primary energy source to sustain life [26]. But instead of just simple sugars, they mostly exist as complex conjugates or glycans such as a short chain of 2-10 monosaccharides (oligosaccharides), a long linear chain (polysaccharide) or highly branched complex molecules attached to proteins (glycoproteins) and lipids (glycolipids) [27–29]. The presence of different sugar monomeric units and their ability to form different anomeric linkages at multiple different positions with or without branching allows glycans to be very diverse and heterogenous, as well as flexible and structurally less constrained [30–32]. These properties allow glycans to evolve rapidly making them well suited for a variety of functions within and on the surface of cells from signaling, recognition and adhesion to cell differentiation, growth and immunity [33–35]. The different blood group types that we have is a result of the changes in a single sugar at the tip of a glycan displayed on the surface of red blood cells [36]. Plant, fungi and bacteria have a thick coating made up of various polysaccharides providing a protective covering in the form of cell walls [37].

Because glycans are so important in so many diverse fields, it is very important to understand how they are built and associated with other physiological processes of the cell. Study of carbohydrates is far behind advances in nucleic acids and proteins research, mainly due to the lack of tools and technology to navigate their complex structures [38]. But this has been changing lately and in the past two decades, the field of glycoscience has made great strides towards understanding the diverse roles of glycans as well as identifying huge numbers of new glycan types across taxa

[39–46]. Most recently, in order to overcome the COVID-19 global pandemic, a lot of focus has been in understanding the coronavirus transmembrane spike (S) glycoprotein in SARS-CoV2, which mediates entry into the host cell through interactions with the human ACE2 receptor [47,48]. This just highlights the importance of glycans and its impact on daily life.

Still, there is a lot we do not know about the world of sugars, especially regarding their precision and biosynthesis which has been a major bottleneck in utilizing their potential. By improving our understanding of their biosynthetic pathways, we can facilitate design of glycans with desired properties for use in the treatment of diseases and in different industrial processes.

1.2.2 Glycosyltransferases

GTs are a large class of enzymes that catalyze the transfer of a sugar molecule from an activated donor (usually nucleoside di/mono-phosphate sugars or lipid-phosphate linked sugars) to a wide variety of acceptors by the formation of a glycosidic bond [49]. Due to their important role in the biosynthesis of glycans, they are essential for the survival of all living cells and are present across all taxonomic groups in varying numbers. In humans, more than 214 GTs have been identified and more than 180 of them have already been characterized [50,51], accounting for nearly 1% of the human genome. The total number of GTs increases drastically in *Arabidopsis thaliana* (thale-cress plant) with more than 500 coding sequences (nearly 2% of its genome), most likely owing to the fact that plant cells require synthesizing a much larger variety of polysaccharides and complex carbohydrates to form the rigid plant cell wall that covers them [52,53]. In general, 0.5–2% of the total number of genes in animal, plant, and fungi genomes are found to be GTs [54,55]. Protist genomes were found to have fewer relative GTs, accounting for only 0.26% of their genomes on average [56].

Across these organisms, GTs catalyze the biosynthesis of many different types of glycoconjugates, including glycolipids, glycoproteins, and polysaccharides [28]. The ones involved in polysaccharide biosynthesis are highly processive and achieve high polymerization efficiencies by not releasing the polymer product and continuously adding one sugar at a time [57]. Most other GTs involved in synthesizing glycoconjugates catalyze transfer of only a single sugar followed by the enzyme-product dissociation. GTs that initiate glycoconjugate biosynthesis need to be able to use a number of specific acceptor substrates from oligosaccharides, monosaccharides and polypeptides to lipids, small organic molecules, and DNA [58]. After the initiation, other GTs are then involved in core extension, elongation, branching and capping of the glycan chain to form a functional final glycan product [59]. Most GTs require a divalent cation as a cofactor (typically Mg++ or Mn++), although there are a number of examples where the activity is independent of a metal ion cofactor [60,61]. Many of these enzymes reside in the ER and the Golgi lumen and are Type II transmembrane proteins with a single transmembrane helix connected to the GT domain in the luminal side by a linker region [62,63]. A pH range of 5.0 to 7.0 is usually found to be the most suitable for these enzymes, which is also the pH values of the ER-Golgi-plasmalemma pathway [64].

Sequence based classification of GTs

A number of databases have been designed to identify and classify GTs based on their sequence similarities [65–69]. The one that stands out the most is the CAZy database [65], that maintains an updated catalogue of all carbohydrate-active enzymes including GTs and has served as an extremely helpful resource by identifying and classifying them into sequence similarity based families. Currently, CAZy lists over 500,000 sequences into 114 numbered families (03/01/2021) and 47 of them include a human GT sequence (Table 1.1). As new GTs are identified, they are constantly

updated in the database and if they do not share a significant homology with any of the existing families, a new family is created with this new sequence as its first member. Families GT2 and GT4 are two of the largest GT families with more than 150,000 sequences each whereas other families such as GT78 and GT93 have fewer than 20 members. Many families are specific to certain taxonomic groups such as family GT70 with GumKs (β-glucuronosyltransferases) [70] and Lgts in GT88 (α-glucosyltransferase) [71] that are only found in bacteria, GT67 (Scg) [72] found in certain protists, GT91 (β-1,2-mannosyltransferases, Wry and Bmt) [73] found only in Fungi or GT96 (peptidyl serine α-galactosyltransferases) [74] found only in plants.

This large number of families, combined with the prevalence of many taxonomic group-specific families and the highly uneven distribution of the number of sequences across families is indicative of the diversity in their primary sequences, which has made it challenging to study their relationships and evolution. Family specific insertions and changes in the core residues across millions of years of evolution have led to extremely divergent sequences making their systematic analysis a non-trivial task. Informatics studies have been limited to select closely related families or enzymes that act on similar glycoconjugates or are part of the same pathways [75,76]. An absence of evolutionary framework connecting GT families has left a knowledge gap, leaving experimental studies to rely on characterizing individual members across model species. But even experimental studies are quite tricky due to problems in expression, purification and crystallization of these enzymes [77].

Table 1.1: List of CAZy GT families with numbers of sequences across taxonomic groups.

GT-Fam	3D Fold Status	Structure	Archaea	Bacteria	Eukaryota	Characterized
GT1	GT-B	0	225	14279	7508	396
GT2	GT-A	13	4165	183436	9235	281
GT3	GT-B	2	45	214	707	11
GT4	GT-B	29	3921	144721	2837	190
GT5	GT-B	8	137	9545	6656	83
GT6	GT-A	4	0	63	208	14
GT7	GT-A	4	1	23	309	33
GT8	GT-A	11	10	11006	1816	53
GT9	GT-B	4	33	25547	0	15
GT10	GT-B	1	1	481	584	45
GT11	GT-u	0	4	963	140	39
GT12	GT-A	0	0	34	37	4
GT13	GT-A	2	0	1	148	20
GT14	GT-A	2	0	943	426	26
GT15	GT-A	2	0	1	1401	6
GT16	GT-A	1	0	0	86	9
GT17	GT-A	0	0	190	147	3
GT18	GT-B	1	0	0	27	6
GT19	GT-B	1	0	9785	13	6
GT20	GT-B	14	80	6876	1287	53
GT21	GT-A	0	66	1032	125	12
GT22	GT-C	0	0	81	1080	9
GT23	GT-B	2	0	666	97	10
GT24	GT-A	2	0	0	283	9
GT25	GT-A	0	0	5285	149	18
GT26	GT-u	1	2	9205	2	6
GT27	GT-A	8	0	42	328	54
GT28	GT-B	3	28	20080	84	11
GT29	GT-u	6	1	9	697	73
GT30	GT-B	2	0	9741	21	20
GT31	GT-A	1	0	8	1594	48
GT32	GT-A	0	1	2993	932	21
GT33	GT-B	0	0	0	267	3
GT34	GT-A	1	0	1	556	11
GT35	GT-B	10	237	14112	506	55
GT37	GT-B	1	0	0	126	4
GT38	GT-B	1	0	140	0	6
GT39	GT-C	2	1	1772	1114	23
GT40	GT-A	0	0	0	46	2
GT41	GT-B	5	0	3221	444	17
GT42	GT-u	2	0	355	0	21
GT43	GT-A	3	0	0	284	12
GT44	GT-u	7	0	904	0	10
GT45	GT-A	1	1	149	0	5
GT47	GT-B	0	0	6	949	13
GT48	GT-u	0	0	0	1662	19
GT49	GT-A	0	0	0	139	3
GT50	GT-C	0	0	0	295	3
GT51	GT-u	8	2	48007	12	15
GT52	GT-B	1	0	757	0	7
GT53	GT-u	1	0	1603	0	6
GT54	GT-A	0	0	2	184	4
GT55	GT-A	2	54	45	8	6
GT56	GT-B	0	0	2990	0	1
GT57	GT-C	0	0	0	486	5
GT58	GT-C	0	0	0	459	9
GT59	GT-C	0	0	0	233	3
GT60	GT-A	0	0	247	43	3

GT61	GT-B	0	0	0	3173	11
GT62	GT-A	1	0	27	584	1
GT63	GT-B	1	0	0	0	1
GT64	GT-A	1	0	0	194	10
GT65	GT-B	3	0	0	79	4
GT66	GT-C	11	622	345	511	17
GT67	GT-A	0	0	0	127	7
GT68	GT-B	2	0	0	68	1
GT69	GT-u	0	0	0	2110	1
GT70	GT-B	1	0	288	0	2
GT71	GT-u	0	0	15	1057	6
GT72	GT-B	1	0	0	0	1
GT73	GT-u	0	0	1785	0	2
GT74	GT-u	0	0	14	3	1
GT75	GT-A	0	58	10	141	5
GT76	GT-u	0	0	47	324	2
GT77	GT-A	0	0	3	291	10
GT78	GT-A	1	0	4	3	2
GT79	GT-B	0	0	0	11	1
GT80	GT-B	6	0	151	0	12
GT81	GT-A	3	140	1612	0	8
GT82	GT-A	0	0	342	0	2
GT83	GT-C	1	5	10183	1	8
GT84	GT-A	0	2	1294	0	1
GT85	GT-C	0	0	669	0	2
GT87	GT-C	0	2	2517	0	2
GT88	GT-A	2	0	189	0	4
GT89	GT-u	0	0	801	0	2
GT90	GT-B	2	0	0	696	4
GT91	GT-u	0	0	0	148	9
GT92	GT-u	0	0	0	170	5
GT93	GT-B	0	0	6	0	1
GT94	GT-B	0	4	361	0	1
GT95	GT-u	0	0	0	78	3
GT96	GT-u	0	0	0	132	5
GT97	GT-u	0	0	104	0	2
GT98	GT-u	0	0	0	115	1
GT99	GT-u	1	0	89	0	2
GT100	GT-u	0	0	298	0	1
GT101	GT-u	2	0	371	1	2
GT102	GT-u	0	0	147	0	2
GT103	GT-u	0	0	62	0	2
GT104	GT-B	3	0	1949	0	4
GT105	GT-u	0	0	4	357	4
GT106	GT-u	0	0	0	459	4
GT107	GT-B	3	0	3601	0	4
GT108	GT-u	4	0	322	64	6
GT109	GT-u	0	0	0	59	1
GT110	GT-u	0	0	0	265	1
GT111	GT-A	1	7	1691	0	1
GT112	GT-B	1	0	447	0	7
GT113	GT-B	2	2	1062	0	3
GT114	GT-u	0	0	20	1	1

GT structural folds

Despite high degrees of primary sequence diversity, GTs display an uncanny conservation when it comes to their 3D structural folds. Apart from a few notable exceptions, many of the recent crystal structures for representative GTs have revealed that a majority of the GT families adopt one of 3 known fold types: the GT-A, GT-B or the GT-C folds [61,78,79].

The GT-A fold is a single globular domain that consists of distinct donor and acceptor binding regions. The donor binding region is a typical Rossmann fold domain with β/α/β domains of different sizes forming a continuous central β sheets sandwiched by the α-helices [61,79]. Most GT-A enzymes also conserve a canonical Asp-X-Asp (DXD)motif that co-ordinates the divalent cation and the ribose moiety of the donor molecule [80–82]. However, this motif is certainly not invariant, with many examples of GT-A enzymes that do not conserve this motif, most of which are also metal-independent [83,84].

The GT-B fold has two distinct lobes, each with a Rossmann like domain, connected by a linker with a cleft in between where the substrate binding sites usually resides [78]. They are usually metal independent with the C-terminal domain usually contributing to donor substrate binding and the N-terminal domain containing features associated with acceptor binding [85].

The GT-C fold mostly consists of GTs with large hydrophobic integral membrane proteins with 8-13 transmembrane helices. The active site residues are generally located in long loop regions between these helices [86,87] . Nearly all of the families predicted to adopt this fold utilize a lipid linked phosphate sugar as a donor substrate instead of the more general nucleotide sugars.

Some GT families are found to adopt unique folds other than the above 3. For example, the peptidoglycan synthesizing GTs of the GT51 family have a lysozyme-type fold in the GT domain with five motifs conserved in both bifunctional and monofunctional GT51 enzymes [88]. Families

like the GT29 and GT42 sialyltransferases have been described to adopt variations of the GT-A fold with different order of the central β-sheets [89,90]. Some others like families GT99 and GT101 have been shown to have even more drastic differences in core structure and are described to adopt unique folds [91,92]. However, only about 62 out of the 114 families have a representative crystal structures, many of them only in their apo form without co-factors or ligands. So, there is an expectation that other new fold types might exist for GTs that have not been identified yet. This paucity of structures has further hampered our understanding of GT function and regulation. There is also an opportunity to utilize this remarkable structural conservation to develop novel computational methods for a comparative framework in order to understand the functional basis shared by GTs.

GT Mechanism

Depending on the anomeric configuration of the product, GTs can either be retaining that retain the α- configuration of the sugar or inverting if the glycosidic bond is inverted to a β- configuration [62]. This change in configuration can result in glycans with entirely different properties. For example, starch which is a polysaccharide of repeating glucose units connected by an α-1,4-glycosidic linkage, is degradable by enzymes in our body and serves as a source of energy. Whereas, cellulose in which glucose units are connected by a β-1,4-glycosidic linkage, cannot be degraded by our body and used as the main components of the plant cell wall. Based on this property of the end product formed by a GT, the mechanism for the transfer of sugar differs too. The mechanism for the inverting GTs is well understood and follows a direct displacement S_N2-like mechanism. Inverting GTs have a conserved active-site residue (usually an Asp or a Glu) that serves as a catalytic base to deprotonate the nucleophile of the acceptor, which then launches a nucleophilic attack on the anomeric carbon of the donor sugar allowing for a direct displacement of the leaving

phosphate group resulting in the inversion of anomeric configuration in the product [93]. Majority of these enzymes use the divalent cation to stabilize negative charge that develops during the departure of leaving group while some exceptional families like GT14 and GT42 use positively charged side chains or hydroxyls from conserved amino acids instead [84].

For the retaining GTs, a couple of mechanisms have been suggested, but there is still some debate on which mechanism they are most likely to follow [94]. Early observations extending from the mechanism followed by retaining glycosidases suggested that retaining GTs follow a double displacement mechanism that proceeds by forming a covalent glycosyl-enzyme intermediate [61]. This mechanism requires the presence of a suitably positioned residue to act as a nucleophile to form the intermediate. However, such a conserved residue has not been observed in many retaining GTs, nor a suitable catalytic base that could deprotonate the acceptor nucleophile. Additionally, it has proven difficult to experimentally capture an intermediate state, possibly due to its short-lived nature. The only trapped intermediate in fact involved a residue far from the active site [95]. Combined, these observations make the double displacement mechanism a contentious one. Another mechanism, which at this point seems more convincing, suggests a same side S_Ni-type mechanism that involves a direct attack by the acceptor involving a single nucleophilic displacement step [96]. In this scenario, the β-phosphate oxygen of the donor could act as the catalytic base and deprotonate the acceptor nucleophile priming it for an attack on the anomeric carbon of the donor from the same side as the leaving phosphate group, thus allowing for the retention of configuration [93]. A number of observations made in complex retaining GT structures [97–100] as well as QM/MM studies [101,102] have favored this mechanism making it the more likely route that retaining GTs follow.

1.2.3 An overview of evolutionary studies and approaches for GTs

As stated earlier, the vast diversity in the primary sequences has posed a challenge for generating meaningful sequence alignments to perform any comparative analyses across GT families. The CAZy database provides a comprehensive classification into the 114 families but lacks information on how these families are related to one another [65]. The fact that these enzymes share a limited number of highly conserved structural folds suggests that at least within a given fold, GT families could have a common ancestor that they diverged from. But the constant evolutionary pressures that require organisms to continuously synthesize novel glycans for existential competitive edge might have played a large role for the massive insertions and family specific attributes that GTs have developed over the course of their evolution [34]. This has led to sequence similarities of less than 10% across families resulting in a loss of evolutionary signals that most phylogenetic methods rely on to reconstruct evolutionary trajectories. Thus, most phylogenetic studies have instead focused on limited numbers of closely related families or sequences that work on the same pathways [75,76]. Some studies have attempted utilizing the structural conservation for the placement of families into broader related groups. A study conducted in 2003 used PSI-BLAST [103] and sequence alignments to delineate three distinct possibly monophyletic groups for GTs corresponding to the three major fold types (GT-A, B and C) in general [104]. However, a comprehensive phylogenetic tree was not built due to a lack of evolutionary signals and to make things even more complex, the number of sequences and new families have grown a lot and new fold types have been discovered albeit in select exceptional families.

Along with this increase in identified GT sequences, modern approaches to generate large quantities of purified protein and polished expression systems have allowed for a burst of new structures to become available as well, with a smaller number of them also able to provide valuable

insights into the mechanism of action by virtue of co-crystallization with co-factors and ligands [77,105,106]. Availability of these structures have also opened up new avenues to conduct a structure-centered evolutionary analysis of GTs and a number of studies have adopted this approach. For example, one study investigated the associations found between GT oligomers and structural similarity to make key observations on the evolution of specific structural sites for GT oligomeric states [107]. Other studies have also looked at the roles of GT oligomerization and complex formation towards evolution [108,109]. A number of studies have also focused on describing the overall plasticity of GT structures observed in large insertion regions around the active site, many of which have been linked to coordinated changes with donor and acceptor binding [110–113].

Stark differences between mammalian, plant or bacterial counterparts of similar GTs performing similar functions have also been noted which suggests a taxonomic group specific adaptation within these enzymes [114]. A good example of this are the α-1,3-Gal/GalNAc transferases of the GT6 family. Mammalian GT6 enzymes such as the Blood ABO transferases and GGTA1 are all dependent on a metal ion as co-factor for activity while all bacterial GT6 enzymes seem to have lost the canonical DXD motif and have a metal-independent activity [83]. Broader studies have systematically analyzed the effects of such adaptations in the evolution of GTs through phylogenetic profiling, which involves describing the presence or absence of GTs centered around specific glycan types across organisms from different taxonomic groups [56,115]. Such studies have made evident that presence of GTs within certain taxonomic niches are directly responsible for the biosynthesis of novel glycan types present only in those organismal groups, further highlighting the effects of the evolutionary pressures for diversified glycan synthesis and their role in GT evolution.

The field of glycoscience has been growing strongly in the past two decades and so has the volume of research conducted on GTs to understand their biosynthetic pathways. Collectively, these studies have provided a great detail of understanding of the mechanism, function and regulation of GTs. However, a comprehensive framework connecting the vast number of primary sequences with their structure and function is still missing. We still need to uncover the extent of shared features within a given GT fold that allow them to adopt similar structures despite sequence variability. The challenges associated with GT experimental studies make it difficult to identify and characterize novel GTs. In the absence of an evolutionary model, extending information from well-studied GTs to understudied GTs has not been feasible leading to knowledge gaps in describing their roles. Additionally, the diversity in glycans and GTs across organisms adds further complexity in describing common modes of function and regulation.

New approaches that are able to take advantage of the large number of primary sequences in conjunction with their observed conservation in structural folds need to be developed in order to make progress towards understanding the biological, functional and regulational roles played by GTs in complex cellular systems. This in turn will provide the groundwork for engineering synthetic GTs with desired roles to generate glycans that are beneficial for medical, commercial, and industrial needs.

1.2.4 Artificial intelligence strategies applied towards protein studies

With the advent of big data, like every other field, biology has also benefitted from generating tons of computational data from expression studies, whole genome sequencing, proteomics and metabolomics. New technologies and tools have made it ever so easy, accessible and cheap to generate much of these data and research labs around the world have been producing a lot of it. GT

sequences identified by homology detection have increased from less than 10,000 to more than 500,000 in the past 2 decades [65,78]. The number of GT structures have also seen a steady increase, although not as much as the primary sequences. Keeping up with this increase in data, computational power, accessibility and machine learning approaches have also risen quite a bit allowing for the adaptation of state of the art artificial intelligence techniques towards analyzing and solving biological problems. The last decade has seen a number of successful applications of machine learning approaches that utilize biological information encoded within primary protein sequences to make inferences about their functional properties and structure [116,117].

Specifically, deep learning techniques have been applied to make predictions about the protein secondary structure, disorder, solvent accessibility and post-translational modifications [118–123]. More recently, deep learning methods have been successful in accurately predicting the 3D structure of the protein given its primary sequence [124–127]. These methods, mostly multi-layer neural networks, rely on the evolutionary information encoded in the sequences to make these accurate predictions. The query sequence is used to collect related sequences and evolutionary information is represented as position specific scoring matrices that are used to make contact map predictions that assist in determining the 3D structural fold. While these methods have been a major breakthrough in protein structural studies, their reliance on using evolutionary information makes them challenging to be applied to proteins like GTs, where the sequences are very diverse and difficult to align. Deep learning methods that do not rely on sequence alignments and utilize intrinsic properties of the protein itself would be highly desirable and practical for approaching families like GTs.

1.3 Key challenges and unresolved questions

Given the wide array of biological roles mediated by glycans, understanding their biosynthetic pathways and using them to synthesize desired glycans has been pursued by researchers across different fields of science. As enzymatic templates responsible for carbohydrate biosynthesis, GTs are central to production of glycans across organisms [28,128]. Thus, understanding the biological underpinnings governing GT function and regulation is the first crucial step towards gaining control over the glycan repertoire and using them for significant contributions in the field of medicine, energy and materials science. However, the remarkable diversity of GT primary sequences combined with the limitations of experimental approaches have hindered our study of this important class of enzymes.

Traditional informatics approaches for comparative study of GTs have been ineffective since they cannot capture evolutionary signals in the primary sequences. And while sequence classification into multiple unrelated groups has helped studies within certain families, extending that knowledge to other understudied families has been extremely difficult [65]. Thus, understanding the evolutionary trajectory of GTs is critical for such comparative evolutionary studies, and also to help uncover conserved modes of function and regulation. Apart from the usually present DXD motif in the GT-A fold enzymes, not much is known about the extent of conservation or the relative locations of other catalytically important residues [62]. The relative position of the catalytic base, which is critical for the inverting reaction mechanism but not as essential for the retaining mechanism, seem to be well conserved in the solved crystal structures but there are many families that do not have representative structures where the identity and location of the catalytic base for those families remain unknown [129,130].

The different conserved folds and two mechanisms that pose different evolutionary constraints on GT sequences combined with the high sequence diversity and the presence of multiple large family specific loop insertions has posed a big challenge towards a comprehensive analysis of the evolution of GTs. Using strategies that are capable of aligning large number of divergent sequences such as profile-based methods in conjunction with structure-based sequence alignment could provide a way forward to generate accurate alignments and develop an evolutionary framework within specific folds [131–133]. Deep learning methods that do not rely on alignments could also provide a new avenue to build a framework for comparative analysis across GT families. Applying such methods systematically to GT families and compiling them into an interpretable and accessible format can provide the much needed groundwork to build testable hypotheses and prioritize experiments on GTs through computational comparative analyses.

1.4 Major research questions addressed

In order to overcome the major bottlenecks in GT research stated above, the following chapters aim to implement rigorous alignment based as well as novel alignment free methods to provide a comparative basis across all GTs. First, we provide a comprehensive evolutionary analysis of the GT-A fold families with key insights into the shared conserved GT-A core and delineate the specific regions responsible for donor and acceptor binding. Using statistically conserved patter positions within this conserved core, we present a functionally meaningful sub-classification of GT-A fold families and point out the evolutionary constraints that drive the functional adaptations in these subfamilies [132,134]. We further use this framework to train a machine learning model that can predict donor sugar specificities for novel GT-A enzymes based on the primary sequence properties. We extend this knowledge to other GT folds using a novel alignment free deep learning approach

that can identify and classify GT structural folds and predict families that are likely to adopt novel GT folds. These findings are also packaged into a user-friendly accessible interface and made available for general use to make interpretations and generate testable hypotheses.

1.4.1 Evolutionary relationships among GT-A fold enzymes

Rationale

The GT-A fold represents one of the major GT folds and includes some of the most important GTs ranging from the blood ABO transferases that determine the blood group specificity and the polypeptide GalNAc transferases that initiate the mucin-type O-glycosylation to the multitude of cell wall biosynthetic machineries responsible for cellulose and hemicellulose biosynthesis [135–138]. Even though this is one of the more well-studied GT folds, there is still very little information about the evolution and shared features of this fold type.

Research goals

In this study, I sought to utilize the structural conservation within GT-A fold sequences as a basis to generate accurate alignments of over half a million GT-A sequences. By statistical analyses within this large alignment, I determine the features of the GT-A conserved core that includes 6 alpha helices, 8 beta sheets and 3 large hypervariable regions with family specific features. I also identify the most conserved amino acid positions within this core that constitutes the residues of the active site and hydrophobic residues that make up the contiguous hydrophobic core for the GT-A enzymes. Identifying these regions and residue positions comprise the first comprehensive definition of the GT-A core providing a comparative basis for analyzing GT-A sequences across multiple families. I further utilize this novel definition to generate a phylogenetic tree highlighting

the evolutionary relationships across GT-A families. This is the first in-depth phylogenetic inference conducted on GT-A families that provides key novel insights into GT-A evolution such as the multiple evolutionary origins of the retaining and inverting mechanisms and the compensatory modes of substrate binding employed by different GT-A families. These key observations provide a rationale for the high sequence diversity and their roles in shaping glycan diversity across organisms. Finally, provided the conserved GT-A core, I train a machine learning model that learns the conserved features within this core to predict donor substrate specificity with high accuracy for known and novel GT-A fold enzymes. This tool provides a novel dependable computational tool to probe into the function and roles of GT-A sequences across organisms.

1.4.2 Fold type characterization and prediction across all known and novel GTs

Rationale

While our analysis of GT-A fold enzymes provides a comprehensive overview of GT-A evolution, it is extremely time-consuming and dependent on availability of crystal structures to extend to other folds. Generation of alignment profiles is labor intensive and requires a lot of manual curation and evaluation, which becomes much more complicated with longer domain regions and minimal sequence conservation. The fact that even structural alignment of representative GT-B crystal structures does not yield a reliable alignment highlights the need for alternative scalable strategies to analyze these GT fold types. More importantly, there are a large number of GT families with an unknown fold type . Variability in primary sequences and the limitations associated with experimental studies have made it challenging to learn more about the structural, functional and evolutionary basis for these families. Thus, it is crucial to develop methods that can informatically

mine the troves of sequence data available for these GT families and place them in a structural context bettering our understanding of the GT fold landscape and its implications on the expansion and evolution of GT families.

Research goals

In this study, I aim to build a deep learning framework that can learn from simple alignment-free inputs of primary sequences and predict fold types for GTs. I first train the model on all the major known GT fold types to learn their distinguishing features and classify them with high accuracy. It relies on the predicted secondary structures for training, with the rationale that the secondary structures are much more conserved across GT families than their primary sequences. In fact, using only this simple input, the model successfully identifies the core conserved regions for the GT-A, B and C fold types and further clusters them into groups of GT families with similar conserved structural features. I then use the model to predict which GT families are the most likely to adopt novel fold types providing informed targets for structural studies to discover novel GT fold types. Unlike many other deep learning tools, this architecture is designed to be highly interpretable and provides quantitative output in every step that is used to make meaningful biological inferences. Thus, this approach presents a novel alignment free method for comparative analyses of GT fold types that is easily scalable and can be applied to any other large protein family.

1.4.3 The GTXplorer portal for accessing and interpreting GT-A evolutionary information

Rationale

Our analyses of the GT-A fold families resulted in large amounts of computational data ranging from well-curated alignment profiles and phylogenetic models to quantitative evolutionarily conserved pattern positions, domain information and other sequence annotations. These datasets can be very helpful for interpreting existing results and generating new hypotheses for the glycoscience community, if presented in an organized and accessible format. The importance of such tools have been evident for other protein families [139,140] and would be valuable to make progress in understanding GT sequence-structure-function relationships.

Research goals

For this section, I aim to compile all the generated datasets for GT-A families into a user-friendly accessible resource that allows researchers to navigate the GT-A evolutionary data in a meaningful way. I organize the datasets into GTXplorer, an interactive tool that can be accessed online or built locally which provides users with two intuitive modes of navigating the data. Tree view allows users to click on the nodes of an interactive GT-A phylogenetic tree to view and download additional information about GT-A clades and families while the alignment view allows users to stack alignment information across clades, families and sub-families to perform comparative sequence analyses. The alignment view also provides annotations for all conserved regions within the GT-A domain alongside corresponding numbering for GT-A sequences from model organisms. With this tool, we intend to maximize the interpretability and usability of the valuable

computational data collected on GT-As so that they can be used in conjunction with the expertise of interested researchers around the world to boost our understanding the of the GT-A fold enzymes.

Bibliography

[1] Varki, A., Gagneux, P., in:, Varki A, Cummings RD, Esko JD, Stanley P, et al. (Eds.), *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2015.

[2] Marth, J.D., A unified vision of the building blocks of life. *Nature Cell Biology* 2008, 10, 1015–1015.

[3] Laine, R.A., Invited Commentary: A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 1012 structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* 1994, 4, 759–767.

[4] Haeuptle, M.A., Hennet, T., Congenital disorders of glycosylation: an update on defects affecting the biosynthesis of dolichol-linked oligosaccharides. *Human Mutation* 2009, 30, 1628–1641.

[5] Jaeken, J., Congenital disorders of glycosylation (CDG): it's (nearly) all in it! *Journal of Inherited Metabolic Disease* 2011, 34, 853–858.

[6] Freeze, H.H., Understanding Human Glycosylation Disorders: Biochemistry Leads the Charge. *J Biol Chem* 2013, 288, 6936–6945.

[7] Ashkani, J., Naidoo, K.J., Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. *Scientific Reports* 2016, 6, 26451.

[8] Fernández, L.P., Sánchez-Martínez, R., Vargas, T., Herranz, J., et al., The role of glycosyltransferase enzyme GCNT3 in colon and ovarian cancer prognosis and chemoresistance. *Scientific Reports* 2018, 8, 8485.

[9] Venkitachalam, S., Guda, K., Altered glycosyltransferases in colorectal cancer. *Expert Rev Gastroenterol Hepatol* 2017, 11, 5–7.

[10] Nagae, M., Kizuka, Y., Mihara, E., Kitago, Y., et al., Structure and mechanism of cancer-associated N -acetylglucosaminyltransferase-V. *Nat Commun* 2018, 9, 1–12.

[11] Gupta, R., Leon, F., Thompson, C.M., Nimmakayala, R., et al., Global analysis of human glycosyltransferases reveals novel targets for pancreatic cancer pathogenesis. *British Journal of Cancer* 2020, 122, 1661–1672.

[12] Moll, T., Shaw, P.J., Cooper-Knock, J., Disrupted glycosylation of lipids and proteins is a cause of neurodegeneration. *Brain* 2020, 143, 1332–1340.

[13] Liang, C., Zhang, Y., Jia, Y., Wenzhao Wang, et al., Engineering a Carbohydrate-processing Transglycosidase into Glycosyltransferase for Natural Product Glycodiversification. *Scientific Reports* 2016, 6, 21051.

[14] Yauk, Y.-K., Ged, C., Wang, M.Y., Matich, A.J., et al., Manipulation of flavour and aroma compound sequestration and release using a glycosyltransferase with specificity for terpene alcohols. *Plant J* 2014, 80, 317–330.

[15] Liu, X., Lin, C., Ma, X., Tan, Y., et al., Functional Characterization of a Flavonoid Glycosyltransferase in Sweet Orange (Citrus sinensis). *Front Plant Sci* 2018, 9.

[16] Mestrom, L., Przypis, M., Kowalczykiewicz, D., Pollender, A., et al., Leloir Glycosyltransferases in Applied Biocatalysis: A Multidisciplinary Approach. *International Journal of Molecular Sciences* 2019, 20, 5263.

[17] Russell, J., Kim, S.-K., Duma, J., Nothaft, H., et al., Deletion of a single glycosyltransferase in Caldicellulosiruptor bescii eliminates protein glycosylation and growth on crystalline cellulose. *Biotechnology for Biofuels* 2018, 11, 259.

[18] Welner, D.H., Shin, D., Tomaleri, G.P., DeGiovanni, A.M., et al., Plant cell wall glycosyltransferases: High-throughput recombinant expression screening and general requirements for these challenging enzymes. *PLOS ONE* 2017, 12, e0177591.

[19] Lao, J., Oikawa, A., Bromley, J.R., McInerney, P., et al., The plant glycosyltransferase clone collection for functional genomics. *The Plant Journal* 2014, 79, 517–529.

[20] Ebert, B., Birdseye, D., Liwanag, A.J.M., Laursen, T., et al., The Three Members of the Arabidopsis Glycosyltransferase Family 92 are Functional β-1,4-Galactan Synthases. *Plant and Cell Physiology* 2018, 59, 2624–2636.

[21] He, Y., Wu, L., Liu, X., Jiang, P., et al., TaUGT6, a Novel UDP-Glycosyltransferase Gene Enhances the Resistance to FHB and DON Accumulation in Wheat. *Front Plant Sci* 2020, 11, 574775.

[22] Hancock, S.M., Vaughan, M.D., Withers, S.G., Engineering of glycosidases and glycosyltransferases. *Current Opinion in Chemical Biology* 2006, 10, 509–519.

[23] McArthur, J.B., Chen, X., Glycosyltransferase engineering for carbohydrate synthesis. *Biochem Soc Trans* 2016, 44, 129–142.

[24] Schumann, B., Malaker, S.A., Wisnovsky, S.P., Debets, M.F., et al., Bump-and-Hole Engineering Identifies Specific Substrates of Glycosyltransferases in Living Cells. *Mol Cell* 2020, 78, 824-834.e15.

[25] Williams, G.J., Zhang, C., Thorson, J.S., Expanding the promiscuity of a natural-product glycosyltransferase by directed evolution. *Nature Chemical Biology* 2007, 3, 657–662.

[26] Stick, R.V., Williams, S., *Carbohydrates: The Essential Molecules of Life*, Elsevier, 2010.

[27] Prestegard, J.H., Liu, J., Widmalm, G., in:, Varki A, Cummings RD, Esko JD, Stanley P, et al. (Eds.), *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2015.

[28] Hölemann, A., Seeberger, P.H., Carbohydrate diversity: synthesis of glycoconjugates and complex carbohydrates. *Current Opinion in Biotechnology* 2004, 15, 615–622.

[29] Imperiali, B., Bacterial carbohydrate diversity — a Brave New World. *Current Opinion in Chemical Biology* 2019, 53, 1–8.

[30] Seeberger, P.H., in:, Varki A, Cummings RD, Esko JD, Stanley P, et al. (Eds.), *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2015.

[31] Herget, S., Toukach, P.V., Ranzinger, R., Hull, W.E., et al., Statistical analysis of the Bacterial Carbohydrate Structure Data Base (BCSDB): Characteristics and diversity of bacterial carbohydrates in comparison with mammalian glycans. *BMC Structural Biology* 2008, 8, 35.

[32] Werz, D.B., Ranzinger, R., Herget, S., Adibekian, A., et al., Exploring the Structural Diversity of Mammalian Carbohydrates ("Glycospace") by Statistical Databank Analysis. *ACS Chem. Biol.* 2007, 2, 685–691.

[33] Gagneux, P., Aebi, M., Varki, A., in:, Varki A, Cummings RD, Esko JD, Stanley P, et al. (Eds.), *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2015.

[34] Springer, S.A., Gagneux, P., Glycan Evolution in Response to Collaboration, Conflict, and Constraint. *Journal of Biological Chemistry* 2013, 288, 6904–6911.

[35] Varki, A., Nothing in Glycobiology Makes Sense, except in the Light of Evolution. *Cell* 2006, 126, 841–845.

[36] Patenaude, S.I., Seto, N.O.L., Borisova, S.N., Szpacenko, A., et al., The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nature Structural Biology* 2002, 9, 685–690.

[37] Amos, R.A., Mohnen, D., Critical Review of Plant Cell Wall Matrix Polysaccharide Glycosyltransferase Activities Verified by Heterologous Protein Expression. *Front. Plant Sci.* 2019, 10.

[38] Merry, A.H., Merry, C.L.R., Glycoscience finally comes of age. *EMBO Rep* 2005, 6, 900–903.

[39] Haynes, P.A., Phosphoglycosylation: A new structural class of glycosylation? *Glycobiology* 1998, 8, 1–5.

[40] Smith, B.A.H., Bertozzi, C.R., The clinical impact of glycobiology: targeting selectins, Siglecs and mammalian glycans. *Nature Reviews Drug Discovery* 2021, 20, 217–243.

[41] Woods, R.J., Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.* 2018, 118, 8005–8024.

[42] Varki, A., Cummings, R., Esko, J., Freeze, H., et al., *Structures Common to Different Types of Glycans*, Cold Spring Harbor Laboratory Press, 1999.

[43] Tra, V.N., Dube, D.H., Glycans in pathogenic bacteria – potential for targeted covalent therapeutics and imaging agents. *Chem Commun (Camb)* 2014, 50, 4659–4673.

[44] Comstock, L.E., Kasper, D.L., Bacterial Glycans: Key Mediators of Diverse Host Immune Responses. *Cell* 2006, 126, 847–850.

[45] Nothaft, H., Szymanski, C.M., New discoveries in bacterial N-glycosylation to expand the synthetic biology toolbox. *Current Opinion in Chemical Biology* 2019, 53, 16–24.

[46] Wells, L., Vosseller, K., Hart, G.W., Glycosylation of Nucleocytoplasmic Proteins: Signal Transduction and O-GlcNAc. *Science* 2001, 291, 2376–2378.

[47] Watanabe, Y., Allen, J.D., Wrapp, D., McLellan, J.S., Crispin, M., Site-specific glycan analysis of the SARS-CoV-2 spike. *Science* 2020, 369, 330–333.

[48] Vuksanaj, K., Understanding Glycans in COVID-19 Drug Design. *GEN - Genetic Engineering and Biotechnology News* 2020.

[49] Taniguchi, N., Honke, K., Fukuda, M., Narimatsu, H., et al., *Handbook of glycosyltransferases and related genes, second edition*, 2014.

[50] Hansen, L., Lind-Thomsen, A., Joshi, H.J., Pedersen, N.B., et al., A glycogene mutation map for discovery of diseases of glycosylation. *Glycobiology* 2015, 25, 211–224.

[51] Narimatsu, H., Human glycogene cloning: focus on beta 3-glycosyltransferase and beta 4-glycosyltransferase families. *Curr Opin Struct Biol* 2006, 16, 567–575.

[52] Caffall, K.H., Mohnen, D., The structure, function, and biosynthesis of plant cell wall pectic polysaccharides. *Carbohydrate Research* 2009, 344, 1879–1900.

[53] Hansen, S.F., Bettler, E., Wimmerová, M., Imberty, A., et al., Combination of Several Bioinformatics Approaches for the Identification of New Putative Glycosyltransferases in Arabidopsis. *J. Proteome Res.* 2009, 8, 743–753.

[54] Yu, J., Hu, F., Dossa, K., Wang, Z., Ke, T., Genome-wide analysis of UDP-glycosyltransferase super family in Brassica rapa and Brassica oleracea reveals its evolutionary history and functional characterization. *BMC Genomics* 2017, 18, 474.

[55] Henrissat, B., Surolia, A., Stanley, P., in:, Varki A, Cummings RD, Esko JD, Stanley P, et al. (Eds.), *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2015.

[56] Hashimoto, K., Tokimatsu, T., Kawano, S., Yoshizawa, A.C., et al., Comprehensive analysis of glycosyltransferases in eukaryotic genomes for structural and functional characterization of glycans. *Carbohydrate Research* 2009, 344, 881–887.

[57] Bi, Y., Hubbard, C., Purushotham, P., Zimmer, J., Insights into the structure and function of membrane-integrated processive glycosyltransferases. *Current Opinion in Structural Biology* 2015, 34, 78–86.

[58] Rini, J., Esko, J., Varki, A., in:, Varki A, Cummings RD, Esko JD, Freeze HH, et al. (Eds.), *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2009.

[59] Schjoldager, K.T., Narimatsu, Y., Joshi, H.J., Clausen, H., Global view of human protein glycosylation pathways and functions. *Nature Reviews Molecular Cell Biology* 2020, 21, 729–749.

[60] Zhang, Y., Wang, P.G., Brew, K., Specificity and Mechanism of Metal Ion Activation in UDP-galactose:β-Galactoside-α-1,3-galactosyltransferase *. *Journal of Biological Chemistry* 2001, 276, 11567–11574.

[61] Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* 2008, 77, 521–555.

[62] Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J., Imberty, A., Structures and mechanisms of glycosyltransferases. *Glycobiology* 2006, 16, 29R-37R.

[63] Young, W.W., Organization of Golgi Glycosyltransferases in Membranes: Complexity via Complexes. *J Membrane Biol* 2004, 198, 1–13.

[64] Varki, A., Cummings, R., Esko, J., Freeze, H., et al., *Glycosyltransferases*, Cold Spring Harbor Laboratory Press, 1999.

[65] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucl. Acids Res.* 2014, 42, D490–D495.

[66] Ekstrom, A., Taujale, R., McGinn, N., Yin, Y., PlantCAZyme: a database for plant carbohydrate-active enzymes. *Database (Oxford)* 2014, 2014.

[67] Zhang, H., Yohe, T., Huang, L., Entwistle, S., et al., dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* 2018, 46, W95–W101.

[68] Egorova, K.S., Toukach, P.V., CSDB_GT: a new curated database on glycosyltransferases. *Glycobiology* 2017, 27, 285–290.

[69] Yamada, I., Shiota, M., Shinmachi, D., Ono, T., et al., The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. *Nature Methods* 2020, 17, 649–650.

[70] Barreras, M., Salinas, S.R., Abdian, P.L., Kampel, M.A., Ielpi, L., Structure and Mechanism of GumK, a Membrane-associated Glucuronosyltransferase. *J Biol Chem* 2008, 283, 25027–25035.

[71] Hurtado-Guerrero, R., Zusman, T., Pathak, S., Ibrahim, A.F.M., et al., Molecular mechanism of elongation factor 1A inhibition by a Legionella pneumophila glycosyltransferase. *Biochem J* 2010, 426, 281–292.

[72] Dobson, D.E., Scholtes, L.D., Valdez, K.E., Sullivan, D.R., et al., Functional Identification of Galactosyltransferases (SCGs) Required for Species-specific Modifications of the

Lipophosphoglycan Adhesin Controlling Leishmania major-Sand Fly Interactions *. *Journal of Biological Chemistry* 2003, 278, 15523–15531.

[73] Mille, C., Bobrowicz, P., Trinel, P.-A., Li, H., et al., Identification of a new family of genes involved in beta-1,2-mannosylation of glycans in Pichia pastoris and Candida albicans. *J Biol Chem* 2008, 283, 9724–9736.

[74] Saito, F., Suyama, A., Oka, T., Yoko-O, T., et al., Identification of Novel Peptidyl Serine α-Galactosyltransferase Gene Family in Plants. *J Biol Chem* 2014, 289, 20405–20420.

[75] Taujale, R., Yin, Y., Glycosyltransferase Family 43 Is Also Found in Early Eukaryotes and Has Three Subfamilies in Charophycean Green Algae. *PLOS ONE* 2015, 10, e0128409.

[76] Lombard, J., The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. *Biol Direct* 2016, 11.

[77] Moremen, K.W., Ramiah, A., Stuart, M., Steel, J., et al., Expression system for structural and functional studies of human glycosylation enzymes. *Nat Chem Biol* 2018, 14, 156–162.

[78] Gloster, T.M., Advances in understanding glycosyltransferases from a structural perspective. *Current Opinion in Structural Biology* 2014, 28, 131–141.

[79] Breton, C., Fournel-Gigleux, S., Palcic, M.M., Recent structures, evolution and mechanisms of glycosyltransferases. *Current Opinion in Structural Biology* 2012, 22, 540–549.

[80] Bothe, M., Dutow, P., Pich, A., Genth, H., Klos, A., DXD Motif-Dependent and -Independent Effects of the Chlamydia trachomatis Cytotoxin CT166. *Toxins (Basel)* 2015, 7, 621–637.

[81] Götting, C., Müller, S., Schöttler, M., Schön, S., et al., Analysis of the DXD Motifs in Human Xylosyltransferase I Required for Enzyme Activity *. *Journal of Biological Chemistry* 2004, 279, 42566–42573.

[82] Li, J., Rancour, D.M., Allende, M.L., Worth, C.A., et al., The DXD motif is required for GM2 synthase activity but is not critical for nucleotide binding. *Glycobiology* 2001, 11, 217–229.

[83] Pham, T.T.K., Stinson, B., Thiyagarajan, N., Lizotte-Waniewski, M., et al., Structures of Complexes of a Metal-independent Glycosyltransferase GT6 from Bacteroides ovatus with UDP-N-Acetylgalactosamine (UDP-GalNAc) and Its Hydrolysis Products. *J Biol Chem* 2014, 289, 8041–8050.

[84] Pak, J.E., Arnoux, P., Zhou, S., Sivarajah, P., et al., X-ray Crystal Structure of Leukocyte Type Core 2 β1,6-N-Acetylglucosaminyltransferase Evidence for a Convergence of Metal Ion-independent glycosyltransfearse mechanism. *J. Biol. Chem.* 2006, 281, 26693–26701.

[85] Chang, A., Singh, S., Phillips, G.N., Thorson, J.S., Glycosyltransferase structural biology and its role in the design of catalysts for glycosylation. *Curr Opin Biotechnol* 2011, 22, 800–808.

[86] Maeda, Y., Watanabe, R., Harris, C.L., Hong, Y., et al., PIG-M transfers the first mannose to glycosylphosphatidylinositol on the lumenal side of the ER. *EMBO J* 2001, 20, 250–261.

[87] Strahl-Bolsinger, S., Immervoll, T., Deutzmann, R., Tanner, W., PMT1, the gene for a key enzyme of protein O-glycosylation in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* 1993, 90, 8164–8168.

[88] Lovering, A.L., Castro, L.H. de, Lim, D., Strynadka, N.C.J., Structural Insight into the Transglycosylation Step of Bacterial Cell-Wall Biosynthesis. *Science* 2007, 315, 1402–1405.

[89] Chiu, C.P.C., Watts, A.G., Lairson, L.L., Gilbert, M., et al., Structural analysis of the sialyltransferase CstII from Campylobacter jejuni in complex with a substrate analog. *Nature Structural & Molecular Biology* 2004, 11, 163–170.

[90] Meng, L., Forouhar, F., Thieker, D., Gao, Z., et al., Enzymatic Basis for N-Glycan Sialylation. *J Biol Chem* 2013, 288, 34680–34698.

[91] Ovchinnikova, O.G., Mallette, E., Koizumi, A., Lowary, T.L., et al., Bacterial β-Kdo glycosyltransferases represent a new glycosyltransferase family (GT99). *Proc Natl Acad Sci U S A* 2016, 113, E3120–E3129.

[92] Zhang, H., Zhu, F., Yang, T., Ding, L., et al., The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. *Nature Communications* 2014, 5, 4339.

[93] Moremen, K.W., Haltiwanger, R.S., Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat Chem Biol* 2019, 15, 853–864.

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

[95] Lairson, L.L., Chiu, C.P.C., Ly, H.D., He, S., et al., Intermediate trapping on a mutant retaining alpha-galactosyltransferase identifies an unexpected aspartate residue. *J Biol Chem* 2004, 279, 28339–28344.

[96] Persson, K., Ly, H.D., Dieckelmann, M., Wakarchuk, W.W., et al., Crystal structure of the retaining galactosyltransferase LgtC from Neisseria meningitidis in complex with donor and acceptor sugar analogs. *Nature Structural Biology* 2001, 8, 166–175.

[97] Yu, H., Takeuchi, M., LeBarron, J., Kantharia, J., et al., Notch-modifying xylosyltransferase structures support an SNi-like retaining mechanism. *Nat Chem Biol* 2015, 11, 847–854.

[98] Lira-Navarrete, E., Iglesias-Fernández, J., Zandberg, W.F., Compañón, I., et al., Substrate-guided front-face reaction revealed by combined structural snapshots and metadynamics for the polypeptide N-acetylgalactosaminyltransferase 2. *Angew Chem Int Ed Engl* 2014, 53, 8206–8210.

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

[100] Albesa-Jové, D., Mendoza, F., Rodrigo-Unzueta, A., Gomollón-Bel, F., et al., A Native Ternary Complex Trapped in a Crystal Reveals the Catalytic Mechanism of a Retaining Glycosyltransferase. *Angew Chem Int Ed Engl* 2015, 54, 9898–9902.

[101] Gómez, H., Polyak, I., Thiel, W., Lluch, J.M., Masgrau, L., Retaining Glycosyltransferase Mechanism Studied by QM/MM Methods: Lipopolysaccharyl-α-1,4-galactosyltransferase C Transfers α-Galactose via an Oxocarbenium Ion-like Transition State. *J. Am. Chem. Soc.* 2012, 134, 4743–4752.

[102] Lee, S.S., Hong, S.Y., Errey, J.C., Izumi, A., et al., Mechanistic evidence for a front-side, SNi-type reaction in a retaining glycosyltransferase. *Nat Chem Biol* 2011, 7, 631–638.

[103] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 1997, 25, 3389–3402.

[104] Liu, J., Mushegian, A., Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci* 2003, 12, 1418–1431.

[105] Kattke, M.D., Gosschalk, J.E., Martinez, O.E., Kumar, G., et al., Structure and mechanism of TagA, a novel membrane-associated glycosyltransferase that produces wall teichoic acids in pathogenic bacteria. *PLoS Pathog* 2019, 15.

[106] Liu, M., Wang, D., Li, Y., Li, X., et al., Crystal Structures of the C-Glycosyltransferase UGT708C1 from Buckwheat Provide Insights into the Mechanism of C-Glycosylation. *The Plant Cell* 2020, 32, 2917–2931.

[107] Hashimoto, K., Madej, T., Bryant, S.H., Panchenko, A.R., Functional states of homooligomers: insights from the evolution of glycosyltransferases. *J Mol Biol* 2010, 399, 196–206.

[108] Hassinen, A., Kellokumpu, S., Organizational Interplay of Golgi N-Glycosyltransferases Involves Organelle Microenvironment-Dependent Transitions between Enzyme Homo- and Heteromers. *J Biol Chem* 2014, 289, 26937–26948.

[109] Hassinen, A., Pujol, F.M., Kokkonen, N., Pieters, C., et al., Functional Organization of Golgi N- and O-Glycosylation Pathways Involves pH-dependent Complex Formation That Is Impaired in Cancer Cells. *J Biol Chem* 2011, 286, 38329–38340.

[110] Albesa-Jové, D., Guerin, M.E., The conformational plasticity of glycosyltransferases. *Current Opinion in Structural Biology* 2016, 40, 23–32.

[111] Jamaluddin, H., Tumbale, P., Withers, S.G., Acharya, K.R., Brew, K., Conformational Changes Induced by Binding UDP-2F-galactose to α-1,3 Galactosyltransferase-Implications for Catalysis. *Journal of Molecular Biology* 2007, 369, 1270–1281.

[112] Tsutsui, Y., Ramakrishnan, B., Qasba, P.K., Crystal Structures of β-1,4-Galactosyltransferase 7 Enzyme Reveal Conformational Changes and Substrate Binding. *J Biol Chem* 2013, 288, 31963–31970.

[113] Qasba, P.K., Ramakrishnan, B., Boeggeman, E., Substrate-induced conformational changes in glycosyltransferases. *Trends in Biochemical Sciences* 2005, 30, 53–62.

[114] Brockhausen, I., Crossroads between Bacterial and Mammalian Glycosyltransferases. *Front Immunol* 2014, 5.

[115] Tomono, T., Kojima, H., Fukuchi, S., Tohsato, Y., Ito, M., Investigation of glycan evolution based on a comprehensive analysis of glycosyltransferases using phylogenetic profiling. *Biophys Physicobiol* 2015, 12, 57–68.

[116] Shi, Q., Chen, W., Huang, S., Wang, Y., Xue, Z., Deep learning for mining protein data. *Briefings in Bioinformatics* 2021, 22, 194–218.

[117] Singh, A., Deep learning 3D structures. *Nature Methods* 2020, 17, 249–249.

[118] Green, J.R., Korenberg, M.J., Aboul-Magd, M.O., PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. *BMC Bioinformatics* 2009, 10, 222.

[119] Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K., et al., NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics* 2019, 87, 520–527.

[120] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., Learning Deep Features for Discriminative Localization. *arXiv:1512.04150 [cs]* 2015.

[121] Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., et al., DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 2018, 34, 2605–2613.

[122] Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* 2017, 33, 2842–2849.

[123] Cao, R., Freitas, C., Chan, L., Sun, M., et al., ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* 2017, 22, 1732.

[124] Gao, M., Zhou, H., Skolnick, J., DESTINI: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports* 2019, 9, 3514.

[125] Jumper, J., Evans, R., Pritzel, A., Green, T., et al., High Accuracy Protein Structure Prediction Using Deep Learning 2020.

[126] Yang, J., Anishchenko, I., Park, H., Peng, Z., et al., Improved protein structure prediction using predicted interresidue orientations. *PNAS* 2020, 117, 1496–1503.

[127] Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., et al., Improved protein structure prediction using potentials from deep learning. *Nature* 2020, 577, 706–710.

[128] Weijers, C.A.G.M., Franssen, M.C.R., Visser, G.M., Glycosyltransferase-catalyzed synthesis of bioactive oligosaccharides. *Biotechnology Advances* 2008, 26, 436–456.

[129] Soya, N., Fang, Y., Palcic, M.M., Klassen, J.S., Trapping and characterization of covalent intermediates of mutant retaining glycosyltransferases. *Glycobiology* 2011, 21, 547–552.

[130] Schuman, B., Evans, S.V., Fyles, T.M., Geometric Attributes of Retaining Glycosyltransferase Enzymes Favor an Orthogonal Mechanism. *PLOS ONE* 2013, 8, e71077.

[131] Neuwald, A.F., Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* 2009, 25, 1869–1875.

[132] Neuwald, A.F., A Bayesian Sampler for Optimization of Protein Domain Hierarchies. *J Comput Biol* 2014, 21, 269–286.

[133] Neuwald, A.F., Lanczycki, C.J., Hodges, T.K., Marchler-Bauer, A., Obtaining extremely large and accurate protein multiple sequence alignments from curated hierarchical alignments. *Database* 2020, 2020.

[134] Neuwald, A.F., Bayesian classification of residues associated with protein functional divergence: Arf and Arf-like GTPases. *Biol Direct* 2010, 5, 66.

[135] Turcot-Dubois, A.-L., Le Moullac-Vaidye, B., Despiau, S., Roubinet, F., et al., Long-term evolution of the CAZY glycosyltransferase 6 (ABO) gene family from fishes to mammals—a birth-and-death evolution model. *Glycobiology* 2007, 17, 516–528.

[136] Almeida, R., Amado, M., David, L., Levery, S.B., et al., A Family of Human β4-Galactosyltransferases: CLONING AND EXPRESSION OF TWO NOVEL UDP-GALACTOSE:β-N-ACETYLGLUCOSAMINE β1,4-GALACTOSYLTRANSFERASES, β4Gal-T2 AND β4Gal-T3 *. *Journal of Biological Chemistry* 1997, 272, 31979–31991.

[137] Revoredo, L., Wang, S., Bennett, E.P., Clausen, H., et al., Mucin-type O-glycosylation is controlled by short- and long-range glycopeptide substrate recognition that varies among members of the polypeptide GalNAc transferase family. *Glycobiology* 2016, 26, 360–376.

[138] Richmond, T., Higher plant cellulose synthases. *Genome Biol* 2000, 1, reviews3001.1-reviews3001.6.

[139] Metz, K.S., Deoudes, E.M., Berginski, M.E., Jimenez-Ruiz, I., et al., Coral: Clear and Customizable Visualization of Human Kinome Data. *Cell Syst* 2018, 7, 347-350.e1.

[140] McSkimming, D.I., Dastgheib, S., Baffi, T.R., Byrne, D.P., et al., KinView: a visual comparative sequence analysis tool for integrated kinome research. *Mol Biosyst* 2016, 12, 3651–3665.

CHAPTER 2

DEEP EVOLUTIONARY ANALYSIS REVEALS THE DESIGN PRINCIPLES OF FOLD A GLYCOSYLTRANSFERASES

Rahil Taujale, Aarya Venkat, Liang-Chin Huang, Zhongliang Zhou, Wayland Yeung, Khaled M Rasheed, Sheng Li, Arthur S Edison, Kelley W Moremen and Natarajan Kannan. 2020. *eLife* 9:e54532
Reprinted here with permission from the publisher.

Abstract

Glycosyltransferases (GTs) are prevalent across the tree of life and regulate nearly all aspects of cellular functions. The evolutionary basis for their complex and diverse modes of catalytic functions remain enigmatic. Here, based on deep mining of over half million GT-A fold sequences, we define a minimal core component shared among functionally diverse enzymes. We find that variations in the common core and emergence of hypervariable loops extending from the core contributed to GT-A diversity. We provide a phylogenetic framework relating diverse GT-A fold families for the first time and show that inverting and retaining mechanisms emerged multiple times independently during evolution. Using evolutionary information encoded in primary sequences, we trained a machine learning classifier to predict donor specificity with nearly 90% accuracy and deployed it for the annotation of understudied GTs. Our studies provide an evolutionary framework for investigating complex relationships connecting GT-A fold sequence, structure, function and regulation.

Author Contributions:

Conceptualization: RT, ASE, KWM, NK. Data curation and collection: RT, AV, ZZ. Formal analysis: RT, AV, LCH, ZZ. Methodology: RT, AV, ZZ, KMR, SL, KWM, NK. Validation: RT, LCH, ZZ, KMR, SL, NK. Visualization and figure generation: RT, AV, WY. Software: RT, LCH, ZZ, NK. Supervision: SL, ASE, KWM, NK. Writing – Original Draft: RT, AV, NK. Writing – review and editing: RT, AV, LCH, ZZ, WY, KMR, SL, ASE, KWM, NK.

2.1 Introduction

Complex carbohydrates make up a large bulk of the biomass of any living cell and play essential roles in biological processes ranging from cellular interactions, pathogenesis, immunity, quality control of protein folding and structural stability [1]. Biosynthesis of complex carbohydrates in most organisms is carried out by a large and diverse family of Glycosyltransferases (GTs) that transfer sugars from activated donors such as nucleotide diphosphate and monophosphate sugars or lipid linked sugars to a wide range of acceptors that include saccharides, lipids, nucleic acids and metabolites. Nearly 1% of protein coding genes in the human genome, and more than 2% of the *Arabidopsis* genome, are estimated to be GTs. GTs have undergone extensive variation in primary sequence and three-dimensional structure to catalyze the formation of glycosidic bonds between diverse donor and acceptor substrates. However, an incomplete understanding of the relationships connecting sequence, structure, function and regulation presents a major bottleneck in understanding pathogenicity, metabolic and neurodegenerative diseases associated with abnormal GT functions [2,3].

Structurally, GTs adopt one of three folds (GT-A, -B or -C) with the GT-A Rossmann like fold being the most common (Figure 2.1). The GT-A fold is characterized by alternating β-sheets and α-helices (α/β/α sandwich) found in most nucleotide binding proteins [4]. The majority of GT-A fold enzymes are metal dependent and conserve a DxD motif in the active site that helps coordinate the metal ion and the nucleotide sugar. Currently, 114 GT families have been catalogued in the Carbohydrates Active Enzymes (CAZy) database (accessed in March 2021) [5]. These families can be broadly classified into two categories based on their mechanism of action and the anomeric configuration of the glycosidic product relative to the sugar donor, namely, inverting or retaining (Figure 2.1). Inverting GTs generally employ an S_N2 single displacement reaction mechanism that

results in inversion of anomeric configuration for the product. In contrast, retaining GTs are believed to employ a dissociative S_Ni-type mechanism, where the anomeric configuration of the product is retained [6,7]. While the sequence basis for inverting and retaining mechanisms is not well understood, most inverting GT-As have a conserved Asp or Glu within a xED motif that serves as the catalytic base to deprotonate the incoming nucleophile of the acceptor, and initiate nucleophilic attack with direct displacement of the phosphate leaving group [7,8].

Retaining GT-As bind the sugar donor similarly to the inverting enzymes, but shift the position of the acceptor nucleophile to attack the anomeric carbon from an obtuse angle using a phosphate oxygen of the sugar donor as the catalytic base and employ a dissociative mechanism that retains the anomeric linkage for the resulting glycosidic product [6]. Such mechanistic diversity of GTs is further illustrated by recent crystal structures of GTs bound to acceptor and donor substrates which show that different acceptors are accommodated in the active site through variable loop regions emanating from the catalytic core [6,9–11]. However, whether these observations hold for the entire super-family is not known because of the lack of structural information for the vast number of GTs. The wealth of sequence data available on GTs provides an opportunity to infer underlying mechanisms through deep mining of large sequence datasets. In this regard, the CAZY database serves as a valuable resource [5] for generating new functional hypotheses by classifying GT enzymes into individual families based on overall sequence similarity. However, a broader understanding of how these enzymes evolved to recognize diverse donor and acceptor substrates requires a global comparison of diverse GT-A fold enzymes. Such comparisons are currently a challenge due to limited sequence similarity between families and the lack of a phylogenetic framework to detect evolutionary events associated with GT functional specialization. Previous efforts to investigate GT evolution have largely focused on individual families or pathways [12,13]

and have not explicitly addressed the challenge of mapping the evolution of functional diversity across families.

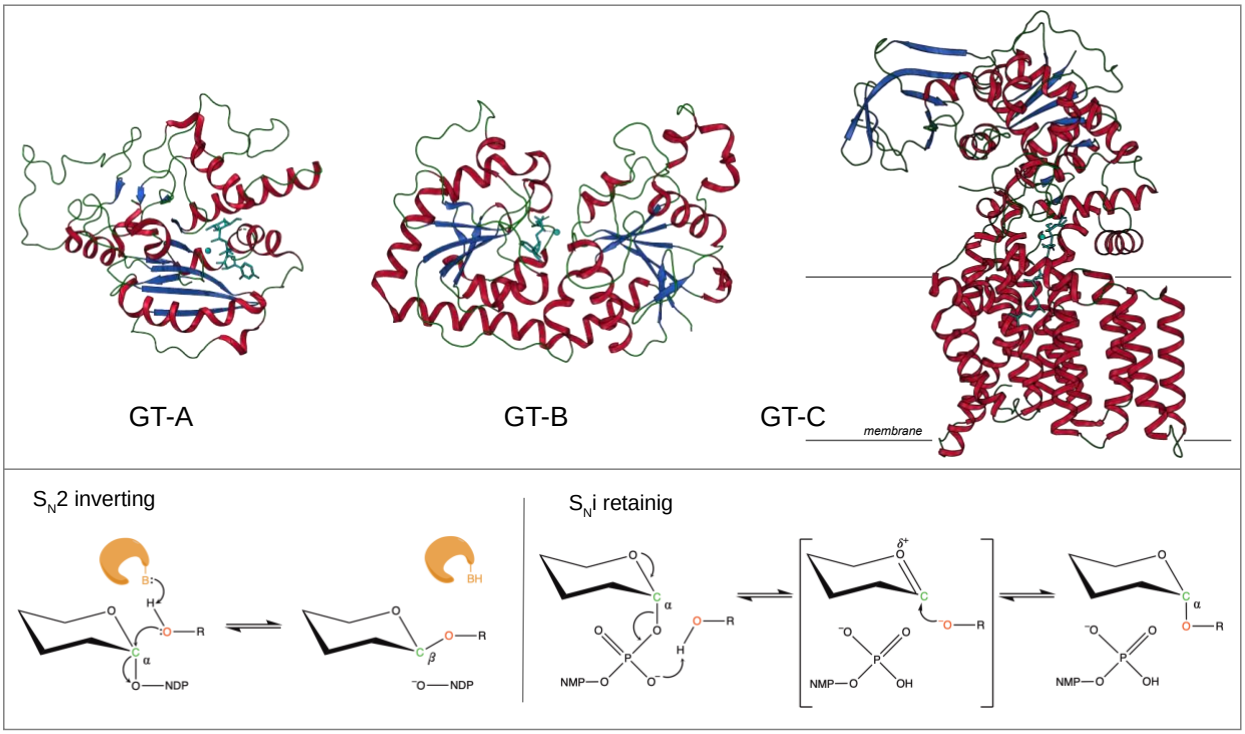


Figure 2.1: Glycosyltransferase folds and mechanisms.
Top: The three representative structural folds of glycosyltransferases. The GT-A fold is characterized by a single globular domain that contains a α/β/α Rossmann nucleotide binding domain (shown 2rj7;GT6). The GT-B fold enzymes are usually metal independent and contain two α/β/α domains separated by a flexible linker region with the substrate binding cleft in between (shown 1jg7;GT63). The GT-C fold enzymes are hydrophobic integral membrane proteins, generally use lipid phosphate linked sugar donors and have multiple transmembrane helices (shown 6gxc; GT66). Bottom: The mechanism of sugar transfer employed by glycosyltransferases. Inverting GTs follow a direct displacement SN-2-like mechanism that results in an inverted anomeric configuration. The mechanism for retaining GTs is still under debate although recently a same side SNi-type reaction has been proposed where the donor phosphate oxygen acts as a catalytic base and deprotonates the acceptor hydroxyl facilitating a same side attack, that results in the retention of anomeric configuration. The enzyme and catalytic base B are shown in orange. A generic hexose with α-linkage to a nucleoside diphosphate is used. Other mechanisms possibly employed by GTs is discussed in detail in [6].

Here through deep mining of over half a million GT-A fold related sequences from diverse organisms, and application of specialized computational tools developed for the study of large gene families [14,15], we define a common core shared among diverse GT-A fold enzymes. Using the common core features, we generate a phylogenetic framework for relating functionally diverse enzymes and show that inverting and retaining mechanisms emerged independently multiple times during evolution. We identify convergent modes of substrate recognition in evolutionarily divergent families and pinpoint sequence and structural features associated with functional specialization. Finally, based on the evolutionary and structural features gleaned from a broad analysis of diverse GT-A fold enzymes, we develop a machine learning (ML) framework for predicting donor specificity with nearly 90% accuracy. We predict donor specificity for uncharacterized GT-A enzymes in diverse model organisms and provide testable hypotheses for investigating the relationships connecting GT-A fold structure, function and evolution.

2.2 Results

2.2.1 An ancient common core shared among diverse GT-A fold enzymes

To define common features shared among diverse GT-A fold enzymes, we generated a multiple sequence alignment of over 600,000 GT-A fold related sequences in the non-redundant (NR) sequence database [16] using curated multiple-aligned profiles of diverse GTs. The alignment profiles were curated using available crystal structures (Methods) [17]. The resulting alignment revealed a GT-A common core consisting of 231 aligned positions. The common core is defined by eight β sheets and six α helices, including three β sheets and α helices from the N-terminal Rossmann fold (Figure 2.2A,B).

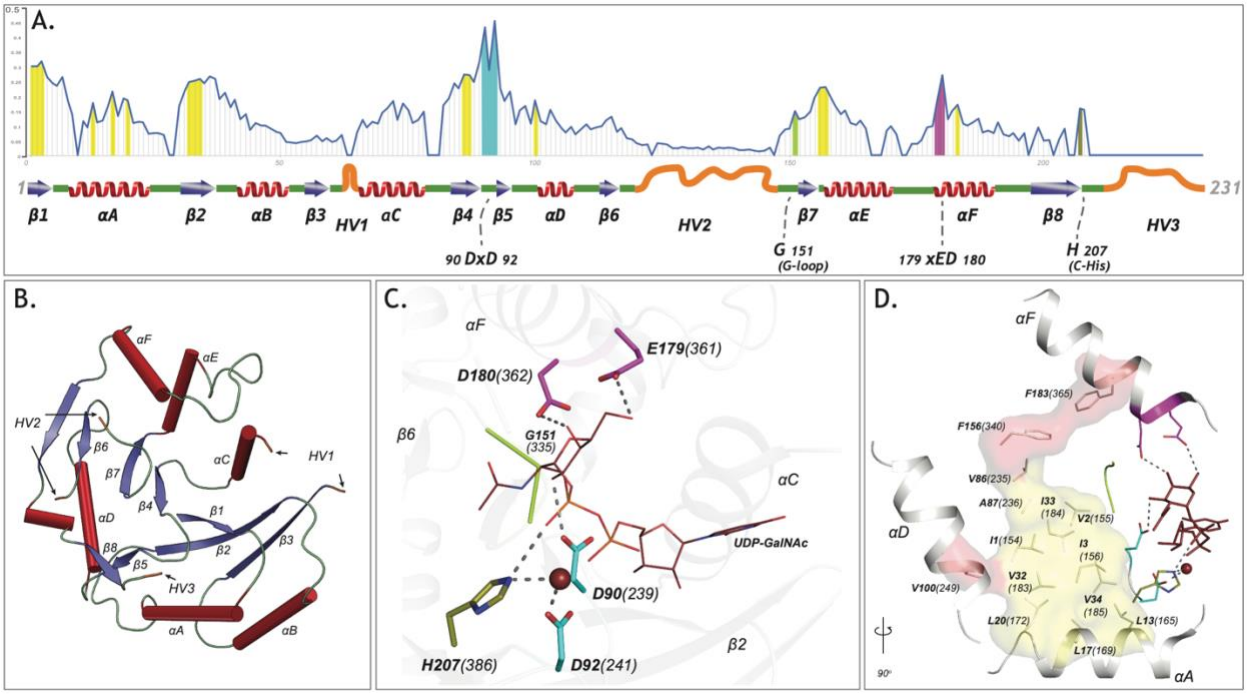


Figure 2.2: The GT-A common core and its elements.
A) Plot showing the schematics of the GT-A common core with 231 aligned positions. Conserved secondary structures (red α-helices, blue β-sheets, green loops) and hypervariable regions (HVs)(orange) are shown. Conservation score for each aligned position is plotted in the line graph above the schematics. Evolutionarily constrained regions in the core: the hydrophobic positions (yellow) and the active site residues (DxD: Cyan, xED: Magenta, G-loop: green, C-His: olive) are highlighted above the positions. B) The conserved secondary structures and the location of HVs are shown in the N-terminal GT2 domain of the multidomain chondroitin polymerase structure from E. coli (PDB: 2z87) that is used as a prototype as it displays closest similarity to the common core consensus. C) Active site residues of the prototypic GT-A structure. Metal ion and donor substrate are shown as a brown sphere and sticks, respectively. D) Architecture of the hydrophobic core (Yellow: core conserved in all Rossmann fold containing enzymes, Red: core elements present only in the GT-A fold). Residues are labeled based on their aligned positions. Numbers within parentheses indicate their position in the prototypic (PDB: 2z87) structure.

Quantification of the evolutionary constraints imposed on the common core reveal twenty residues shared among diverse GT-A fold families. These include the DxD and the xED motif residues involved in catalytic functions, and other residues not typically associated with catalysis (Figure 2.2A) such as the conserved glycine at aligned position 151 (G335 in 2z87) in the flexible

G-loop and a histidine residue (H386 in 2z87) in the C-terminal tail at aligned position 207, henceforth referred to as the C-His. Residues from the G-loop in some families, such as the blood ABOs (GT6) and glucosyl-3-phosphoglycerate synthases (GpgS; GT81), contribute to donor binding [18,19]. The C-His, likewise, coordinates with the metal ion and contributes to catalysis in a subset of GTs, such as polypeptide N-acetylgalactosaminyl transferases (ppGalNAcTs; GT27) and lipopolysaccharyl-α-1,4-galactosyltransferase C (LgtC; GT8) [20,21]. The conservation of these residues across diverse GT-A fold enzymes suggest that they likely perform similar functional roles in other families as well.

The remaining core conserved residues include fourteen hydrophobic residues that are dispersed in sequence, but spatially cluster to connect the catalytic site and the Rossmann fold. Eleven out of the fourteen residues (highlighted in yellow in Figure 2.2D) are shared by other Rossmann fold proteins suggesting a role for these residues in maintaining the overall fold. Three hydrophobic residues (V249, F340, F365; shown in red surface in Figure 2.2D), however, are unique to GT-A fold enzymes, and structurally bridge the αF helix (containing the xED motif), the αD helix and the Rossmann fold domain. Although the functional significance of this hydrophobic coupling is not evident from crystal structures, in some families (GT15 and GT55) the hydrophobic coupling between αF and the Rossmann fold domain is replaced by charged interactions. The structural and functional significance of these family specific variations are discussed below.

Our broad evolutionary analysis also reveals three hypervariable regions (HVs) extending from the common core. These include an extended loop segment connecting β3 strand and αC helix (HV1), a segment longer than 28 amino acids connecting β6 and β7 strand (HV2) and a C-terminal tail extending from the β8 strand (HV3) in the common core. These HVs, while conserved within families, display significant conformational and sequence variability across families (Figure 2.2A)

and encode family-specific motifs that contribute to acceptor specificity in individual families, as discussed below.

2.2.2 A phylogenetic framework relating diverse GT-A fold families

Having delineated the common core, we next sought to generate a phylogenetic tree relating diverse GT-A fold families using the core alignment. Because of the inherent challenges in the generation and visualization of large trees [22], we used a representative set of GT-A fold sequences for phylogenetic analysis by first clustering the ~600,000 sequences into functional categories using a Bayesian Partitioning with Pattern Selection (BPPS) method [23]. The BPPS method partitions sequences in a multiple sequence alignment into hierarchical sub-groups based on correlated residue patterns characteristic of each sub-group (Methods). This revealed 99 sub-groups with distinctive patterns. Representative sequences across diverse phyla from these sub-groups (993 sequences) were then used to generate a phylogenetic tree (Figure 2.3). Based on the phylogenetic placement of these sequences, we broadly define fifty-three major sub-groups, thirty-one of which correspond to CAZy-defined families. The remaining sub-groups correspond to sub-families within larger CAZy families. In particular, we sub-classified the largest GT family in the CAZy database, GT2, into ten phylogenetically distinct sub-families. Likewise, GT8 and GT31 were classified into seven and five sub-families, respectively. These sub-families are not explicitly captured in CAZy and are annotated based on overall sequence similarity to functionally characterized members. For example, “GT2-LpsRelated” corresponds to a sub-family within GT2 most closely related to the bacterial β-1-4-glucosyltransferases (lgtF) involved in Lipopolysaccharide biosynthesis (Figure 2.3). Such a hierarchical classification captures the evolutionary relationships between GT-A fold families/sub-families while keeping the nomenclature consistent with CAZy.

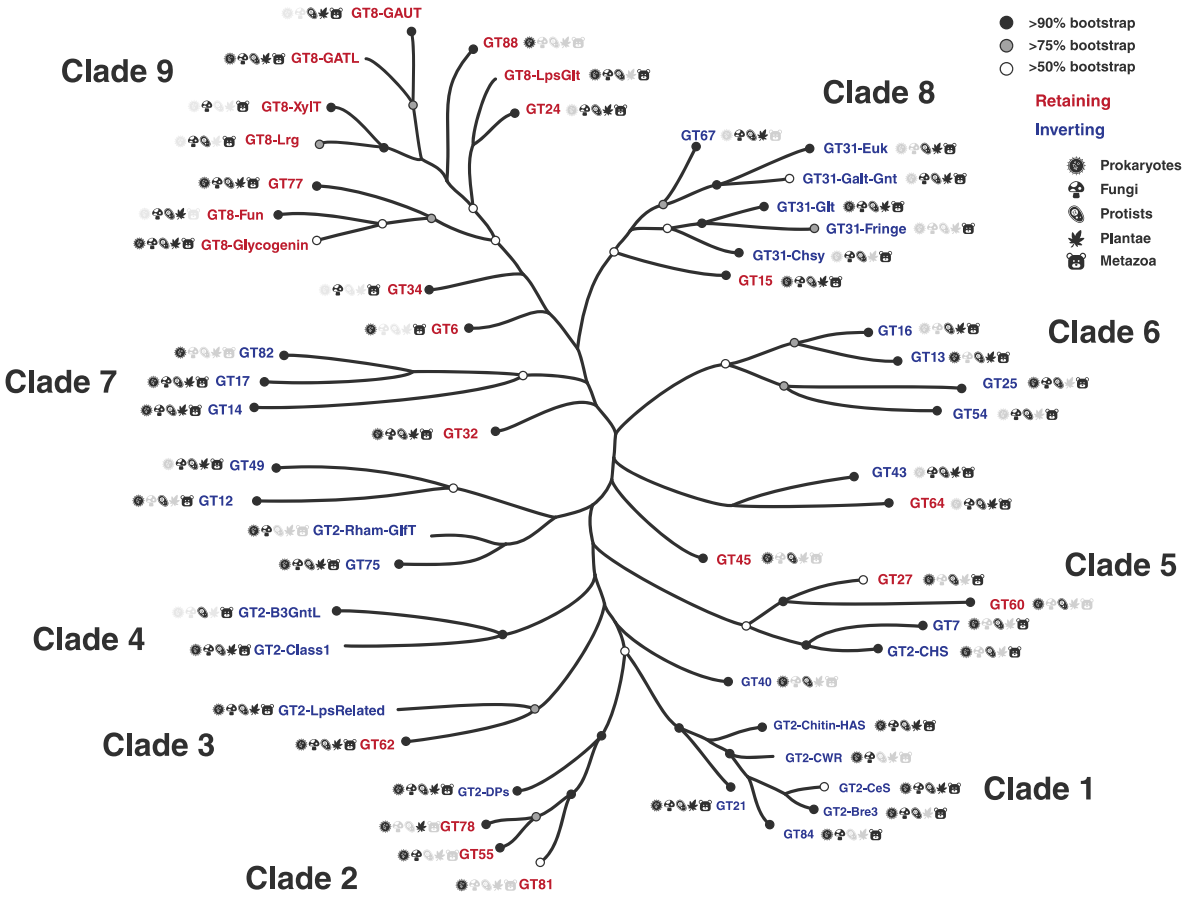


Figure 2.3: Phylogenetic tree highlighting the 53 major GT-A fold subfamilies.
Tips in this tree represent GT-A sub-families condensed from the original tree for illustration. Support values are indicated using different circles. Circles at the tips indicate bootstrap support for the GT-A family clade represented by that tip. Tips missing the circles represent GT-A families that do not form a single monophyletic clade. Nodes missing circles have a bootstrap support less than 50% and are unresolved. Icon labels indicate the taxonomic diversity of that sub clade. Colors indicate the mechanism for the families (blue: Inverting, red: Retaining). This condensed tree was generated by collapsing clades to the deepest node that includes sequences from the same family. For GT-A families that did not form a monophyletic clade, the clade that included the most sequences from that family was chosen. Branch lengths may approximate the original distances, but are not drawn to scale.

GT-A fold families and sub-families can be further grouped into clades based on shared sequence features and placement in the phylogenetic tree (Figure 2.3). For example, clade 1 groups four GT2 sub-families (GT2-CeS, GT2-CWR, GT2-Chitin-HAS and GT2-Bre3) with GT84 and GT21 with high confidence, as determined by bootstrap values (see Figure 2.3 legends). Members of these six families are all involved in either polysaccharide or glycosphingolipid biosynthesis. Additionally, the pattern-based classification identified a conserved [QR]XXRW motif in the C-terminal HV3 (Figure 2.4) which is unique to members of this clade. The [QR]XXRW motif residues coordinate with the donor and acceptor in a bacterial cellulose synthase (from GT2-CeS family) [24] and mutation of these residues in bacterial cyclic β-1,2-glucan synthetase (Cgs, GT84) abrogates activity [25], suggesting a critical role of this motif in functional specialization of clade 1 GT-As.

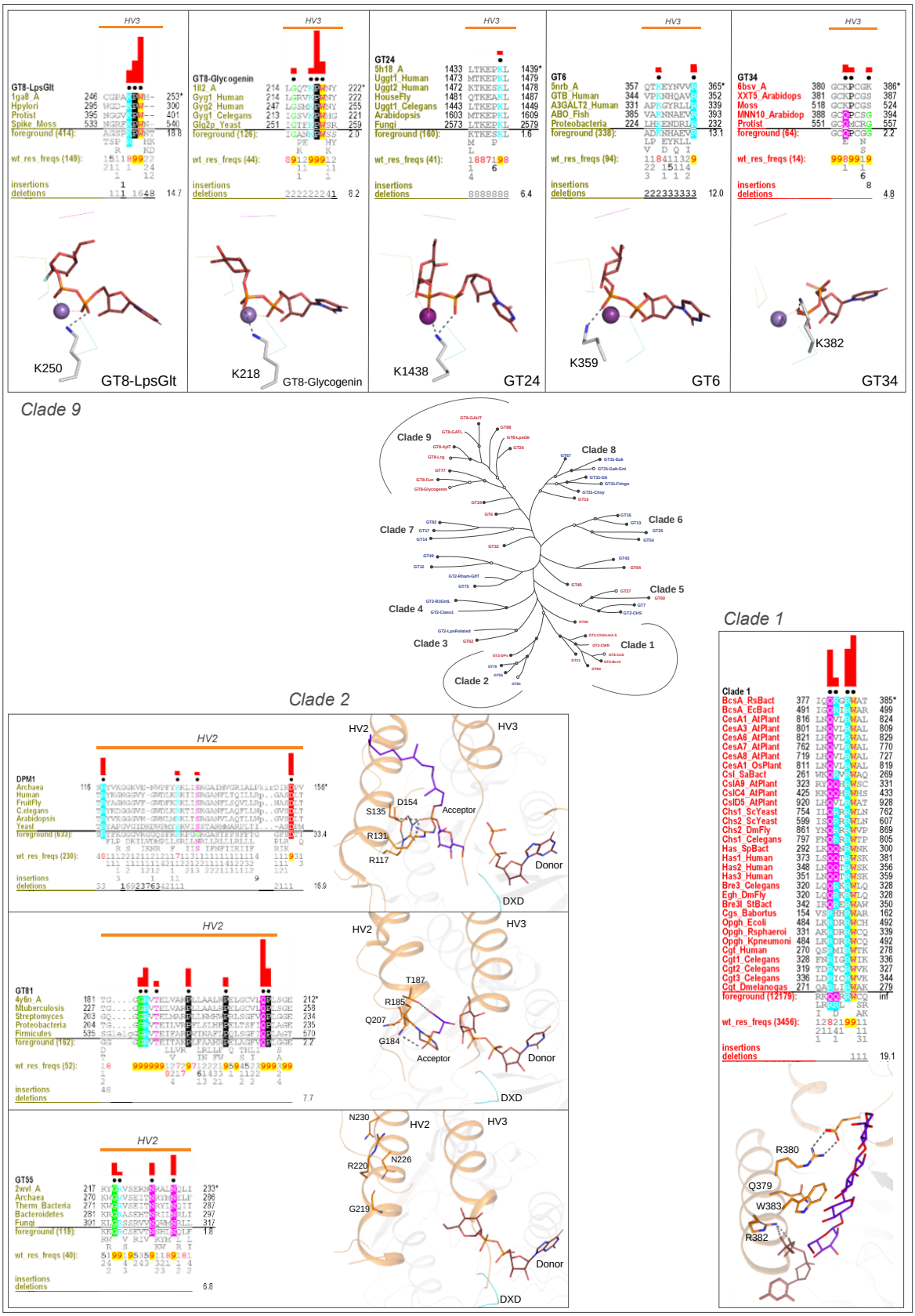
The GT8 sub-families form sub-clades within the larger clade 9. For example, GT8 sequences involved in the biosynthesis of pectin components group together in the GT8-GAUT and GT8-GATL families (Figure 2.3). The human LARGE1 and LARGE2 glycosyltransferases are multi-domain enzymes with two tandem GT-A domains. Their N-terminal GT-A domains fall into the GT8-Lrg subfamily that groups closely with GT8-xylosyltransferase (GT8-XylT) subfamily enzymes and places all the GT8 xylosyltransferases into a single well supported sub clade. The lipopolysaccharide α-glucosyltransferases (GT8-LpsGlt) group with the glucosyltransferases of the GT24 family, suggesting a common ancestor associated with glucose donor specificity. On the other hand, the GT8-Glycogenin sub family, which also includes members that transfer a glucose, is placed in a separate sub-clade, possibly indicating an early divergence for its unique ability to add glucose units to itself [26]. Clade 9 members also share common sequence features associated with substrate binding that includes a lysine residue within the commonly shared KPW motif in HV3

that coordinates with the phosphate group of the donor (e.g. bacterial LgtC GT8-LpsGlt) and other structures of clade 9 members)(Figure 2.4).

We noticed that three out of four MGAT GT-A families responsible for the branching of N-glycans (GT13 MGAT1, GT16 MGAT2 and GT54 MGAT4) fall in the same clade (clade 6), as expected (Figure 2.3). In contrast, the fourth family, GT17 MGAT3, which adds a bisecting GlcNAc to a core β-mannose with a β-1,4 linkage, is placed in a separate clade with GT14 and GT82 (clade 7), while a fifth MGAT member creating β-1,6-GlcNAc linkages (GT18 MGAT5) is a GT-B fold enzyme [27].

We further note that fifteen out of fifty-three GT-A families are found in both prokaryotes and eukaryotes. These fifteen families fall on different clades throughout the tree. GT-A families present only in prokaryotes, like GT81, GT82 and GT88, are also spread out in different clades (Figure 2.3). Similarly, other GT-A families that are present within restricted subsets of taxonomic groups (like GT40 and GT60 present only in prokaryotes and protists) are also scattered throughout the tree. These observations suggest that the divergence of most GT-A families predates the separation of prokaryotes and eukaryotes.

Figure 2.4: Clade specific conserved features in the HVs.
The conserved mode of donor binding in clade 9, conserved mode of acceptor binding in clade 2 and the conserved QXXRW motif in clade 1 are illustrated. HVs are shown in orange. Metal ions are shown as spheres. Red bars above the alignment indicate the extent of significance of conservation of residue in the column (Higher is more significantly conserved). Below every position in the alignment, numbers indicate the extent of conservation of residues at the position.


HV3
GT8-LpsGlt
K250
GT8-LpsGlt
GT8-Glycogenin
K218
GT8-Glycogenin
GT24
K1438
GT24
GT6
K359
GT6
GT34
K382
GT34
Clade 9
Clade 8
Clade 7
Clade 6
Clade 5
Clade 4
Clade 3
Clade 2
Clade 1
Clade 2
HV2
HV3
D154
S135
Acceptor
R131
R117
Donor
DXD
T187
R185
Q207
G184
Acceptor
Donor
DXD
N230
R220
N226
G219
Donor
DXD
Clade 1
R380
Q379
W383
R382

2.2.3 Multiple evolutionary lineages for inverting and retaining mechanisms

To obtain insights into the evolution of catalytic mechanism, we annotated the phylogenetic tree based on known mechanisms of action (inverting or retaining). Inverting GTs are colored in blue in the phylogenetic tree, while retaining GTs are colored in red (Figure 2.3). The dispersion of inverting and retaining families in multiple clades suggests that these catalytic mechanisms emerged independently multiple times during GT-A fold evolution. We find that natural perturbations in the catalytic base residue, an important distinction between the inverting and retaining mechanisms, correlates well with these multiple emergences across the tree.

The residue that acts as a catalytic base for inverting GTs (aspartate within the xED motif, xED-Asp) is variable across the retaining families consistent with its lack of role in the retaining S_Ni mechanism [6]. In the inverting families, the xED-Asp is nearly always conserved and appropriately positioned to function as a catalytic base (Figure 2.5A), though some exceptions have been noted [6,28]. Out of the five clades grouping inverting and retaining families, inverting families in three of these clades do not conserve the xED-Asp (GT2-DPs, GT2-LpsRelated and GT43). The heterogeneous nature of this residue in these families suggests that change of the catalytic base residue could be a key event in the transition between inverting and retaining mechanisms. Unlike families that conserve the xED-Asp, these families achieve inversion of stereochemistry through alternative modes that may relieve the constraints necessary to conserve the xED-Asp. For example, in GT43, the Asp base is replaced by a glutamate residue, which shifts the reaction center by one carbon bond [6]. Further, the dolichol phosphate transferases (DPMs and DPGs) in the GT2-DP family, which lack the xED-Asp entirely, transfer sugars to a negatively charged acceptor substrate (a phosphate group) and thus do not need a catalytic base to initiate nucleophilic attack [28]. Other GT-A inverting families lacking the xED-Asp (GT12, GT14, GT17,

GT49 and GT82) are grouped into separate monophyletic clades segregating them from inverting families with the conserved xED-Asp (Figure 2.3). Out of these, only GT14 has representative crystal structures where a glutamate serves as the catalytic base [29]. For other inverting families with a non-conserved xED-Asp, residues from other structural regions may serve as a catalytic base. On the other hand, retaining families like GT64 conserve the xED-Asp, yet do not use it as a catalytic base. Thus, there may be multiple ways in which inverting and retaining mechanisms diverge, with one path being mutation of the xED-Asp catalytic base.

One strongly supported clade that includes both inverting and retaining families is clade 2 that groups inverting GT-A family members that transfer sugars to phosphate acceptors (GT2-DPs) with three retaining GT-A families that also have phosphate-linked acceptors (GT55, GT78 and GT81). This placement is further supported by the observation that these families share structurally equivalent conserved residues in the HV2 region that coordinate the phosphate group of the acceptor. In the GT2-DP subfamily, R117, R131 and S135 (Figure 2.6A) in HV2 coordinate with the acceptor phosphate groups. The conservation of these residues in GT55 and GT81 suggests that they likely perform similar interactions in these latter subclades. Indeed, in the crystal structure of M. tuberculosis GpgS (GT81), HV2 adopts a conformation similar to GT2-DPs and the shared residues G184, R185 and T187 (equivalent to R117, R131 and S135) form similar interactions with the phosphate group of the acceptor (Figure 2.4).

Clade 5 places the inverting GT7 and GT2-CHS with the retaining GT27 and GT60 families (Figure 2.3). This supports the evolution of these families from a close common ancestor through gene duplication and divergence, which has been suggested through structural similarities between GT7 and GT27 [30]. After this initial divergence in mechanism within clade 5, the subclades group the β-1,4-GalNAc transferase domains of bacterial and protist chondroitin polymerases (involved

in the elongation of glycosaminoglycan chondroitin)(GT2-CHS) with the GT7 family. The GT7 family includes the higher organism counterparts of the β-1,4-GalNAc transferase domains of chondroitin synthases, along with β-1,4-Gal transferases. The close placement of GT60 and GT27 families in this clade is also directly supported by previous literature indicating that these families share a conserved mode of polypeptide Ser/Thr O-glycosylation [31]. Clade 5 thus consolidates previous independent findings and suggests a shared ancestor, potentially extending the common ancestry of GT2-CHS and GT7 to include GT27 and GT60, with an ancestral divergence in mechanism.

2.2.4 Variations in the core and hypervariable regions contribute to unique modes of substrate specificity

Analysis of the patterns of conservation and variation in the common core indicates that each residue position within the core has been mutated in some context during the course of evolution, highlighting the tolerance of the GT-A fold to extensive sequence variation. While some of these variations are confined to specific clades or families, such as replacement of DxD motif with DxH motif in GT27 and GT60, other variations are found independently across distal clades (Figure 2.5A). For example, GT14 and prokaryotic members of GT6 that fall on different clades, have independently lost the DxD motif and no longer require a metal ion for activity [29,32].

The C-His is also lost independently in multiple clades (Figure 2.5A). In order to investigate how the loss of metal binding C-His is compensated, we analyzed the C-His-metal ion interactions across all available crystal structures. Structural alignment of GT-A families lacking the C-His such as GT13, GT6 and GT64 families revealed a water molecule coordinating the metal ion in a manner similar to the C-His sidechain (Figure 2.5B).

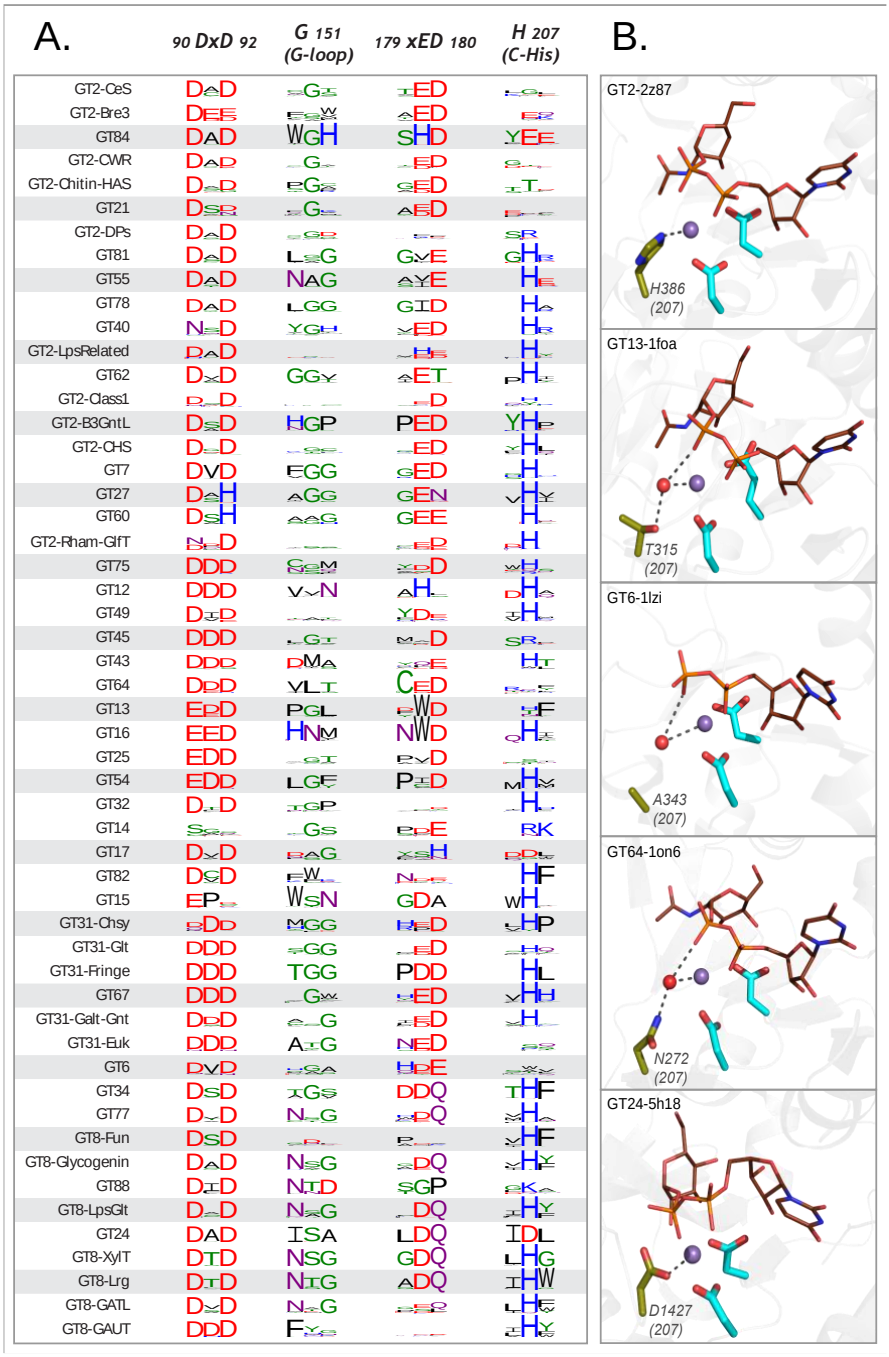


Figure 2.5: Variations in the GT-A conserved core.

A) Weblogo depicting the conservation of active site residues in the common core are shown for each of the GT-A families. Residues are colored based on their physiochemical properties. B) Variations in the C-His is compensated either using a water molecule (red sphere) or other charged residues (olive sticks) to conserve its interactions. The metal ion is shown as a purple sphere. The donor substrate is shown as brown lines. Interactions between the residues, metal ion and the donor are shown using dotted lines.

In other families, such as GT24, we found that the C-His is substituted by an aspartate (D1427), which coordinates with the metal ion similar to C-His (Figure 2.5B, bottom panel). Likewise, the conserved hydrophobic coupling between αF helix and the Rossmann domain is replaced by charged interactions (R388 and E274, respectively) in some retaining GTs such as GT15 and GT55. These substitutions point to the ability of GT-As to accommodate changes, even in conserved positions at the core, through compensatory mechanisms.

The HV regions show significant variability across GT-A families and extend from the common core to perform various roles from substrate binding to large conformational changes that position the donor and acceptor substrates for the enzymatic reaction [33–35]. Mutations within these HV regions, for example, at aligned position 126 in the HV2 region (Y177A,G in 4lw6, GT7), have also been shown to induce a shift in acceptor specificity [34].

Despite significant sequence variability, we find that these HV regions in fact conserve family specific residues that contribute to acceptor specificity. For example, a distinctive arginine (R117) and aspartate (D154) along with R131 and serine S135 within the HV2 of DPM1 (GT2-DP sub-family) contribute to specificity towards a dolichol phosphate acceptor by creating a charged binding pocket for the phosphate group (Figure 2.6A). Likewise, family-specific residues (R198, H221 and E224 in 5vcm) within the HV1 of MGAT2 (GT16) form a unique scaffold for recognizing the terminal GlcNAc of the N-glycan acceptor (Figure 2.6B). Similarly, the C-terminal GT64 domain of the multidomain EXTLs contain specific residues in HV2 (R181 and Y193) and HV3 (H289 and R293) that form a unique binding pocket for the tetrasaccharide linker acceptor used to synthesize glycosaminoglycans (Figure 2.6C). Together these examples illustrate the ability of HVs to evolve family specific motifs to recognize different acceptors.

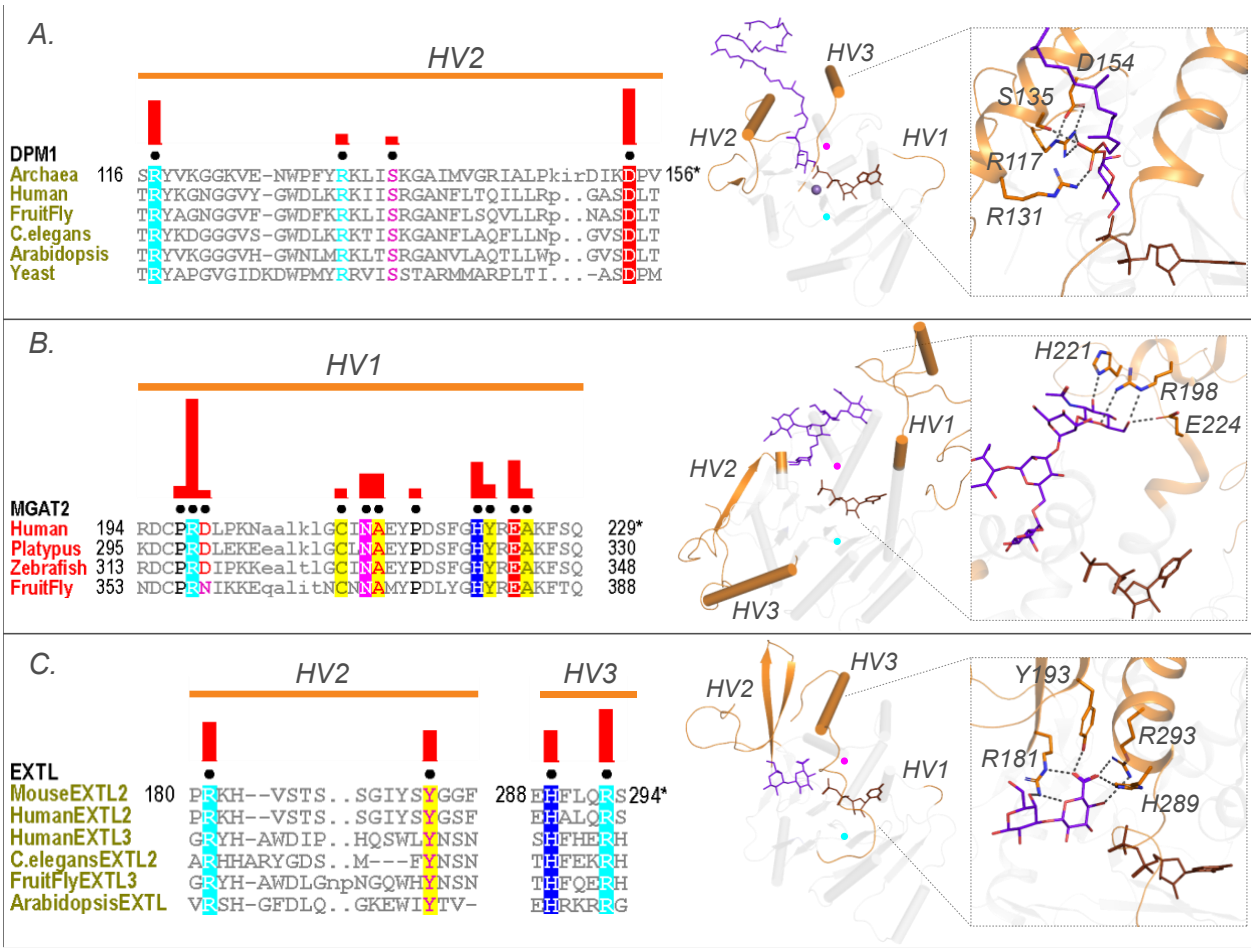


Figure 2.6: Family specific conserved features in the HV regions correlate with acceptor recognition and specificity.

Conserved residues in A) HV2 of the DPM1 sequences in the GT2-DP subfamily coordinate the phosphate group of the acceptor. B) HV1 of GT16 MGAT1 provide acceptor specificity. C) HV2 and HV3 of EXTL GT64 family (C-terminal GT domain of the multidomain sequences) coordinate the acceptor. Left: Alignments highlighting the constrained residues are shown for each family. The family specific conserved residues are shown using black dots above the alignment. Red bars above these dots indicate the significance of conservation (Higher bar corresponds to more significantly conserved position). Right: Representative pdb structures are shown for each family (GT2-DP:5mm1, GT16:5vcs, GT64:1on8); Donor substrates are colored brown. Acceptors are colored purple. HVs are highlighted in orange. The position of the conserved DxD and xED motif for each structure is shown as cyan and magenta circles respectively.

2.2.5 Machine learning to predict the donor specificity of GT-A sequences

As discussed above, the conserved catalytic residues dictate the mechanism of sugar transfer and metal binding while the extended HVs use family specific motifs to dictate acceptor specificity. We also find some clade specific features (such as the conserved Lys in clade 9, and QXXRW in clade 1) and G-loop residues involved in donor binding, however, the overall framework that dictates donor sugar specificity in GTs is largely unknown. Sequence homology alone is insufficient to predict donor specificity because evolutionarily divergent families can bind to common substrates, and sometimes even two closely related sequences bind to different donors [18]. For a subset of GT-B fold families, machine learning (ML) methods have been successfully applied towards predicting substrate specificities [36].

Our global analysis provides a comparative basis to expand such methods and contrast sequences that bind different donors across all GT-A families. To test whether evolutionary features gleaned from this global analysis can be used to better predict donor substrate specificity, we employed a ML framework that learns from the specificity-determining residues of functionally characterized enzymes to predict specificity of understudied sequences. In brief, using an alignment of a well curated set of 713 GT-A sequences with known donor sugars, we derived five amino acid properties (hydrophobicity, polarity, charge, side chain volume and accessible surface area) from each aligned position within the common core. These properties were then used as features to train multiple machine learning models. Among the seven methods used, the gradient boosted regression tree (GDBT) model achieved the best prediction performance (accuracy ~90%) based on a 10-fold cross validation (CV) using 239 contributing features (Figure 2.7A,B). This model adds an ensemble boosting to tree based learners used for predicting GT1 substrate specificities [36]. To further validate the model, we tested its performance on a validation set of 64 sequences that were

not used to train the ML model but have known sugar specificities. The GDBT classifier correctly predicted donor substrates for 92% of these sequences, 89% of which were predicted with high confidence.

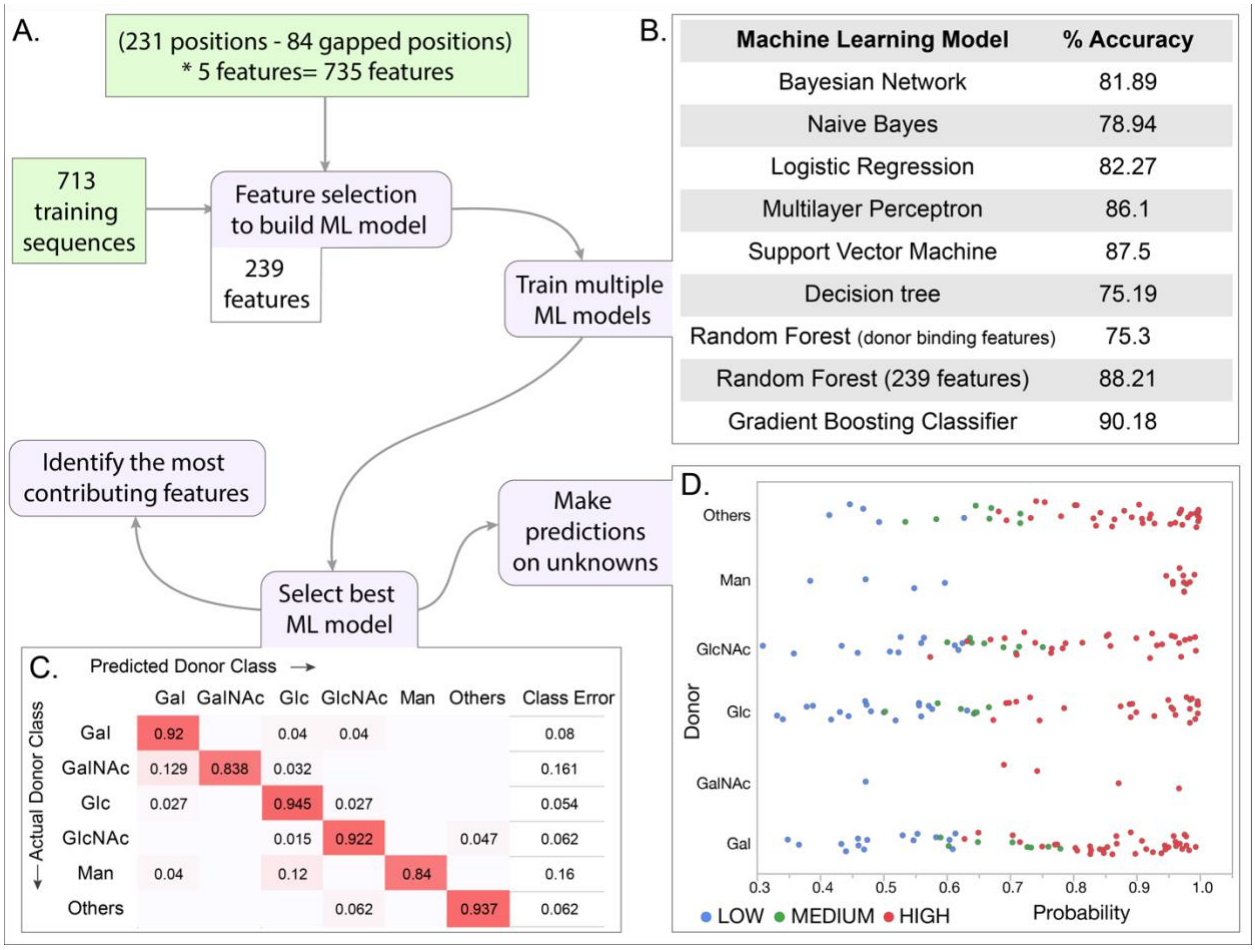


Figure 2.7: Machine learning approach for predicting donor class.
A) Brief pipeline of the machine learning analysis. Training set input into the pipeline are shown in green squares. Steps of the machine learning (ML) analysis in purple boxes are associated with different panels of the figure. B) Percent accuracy based on 10-fold cross validation (CV) for each of the trained ML models. C) Confusion matrix from the best model (GDBT using 239 features). D) Scatter plot showing the probability scores assigned for each predicted sequence by the predicted donor type. Colors indicate the confidence level of the prediction based on probability of assignment to a given donor class as well as confidence intervals of the predicted class i.e., difference in probability values between the 1st prediction class and the 2nd prediction class.

The GDBT model was then used to predict donor sugars for GT-A domains with unknown specificities from 5 organisms: *H. sapiens*, *C. elegans*, *D. melanogaster*, *A. thaliana* and *S.cerevisiae*. Each prediction is associated with a confidence level derived from the probability for each of the 6 donor classes (Methods). Nearly 77% of the predictions have high and moderate confidence levels and present good candidates for further investigation (Figure 2.7D). The remaining 23% of the predictions are low confidence. This likely reflects their promiscuity for donor preferences, as seen across many GT-As [19,37], or non-catalytic GT-As like C1GALT1C1 (Cosmc) [38].

Our predictions assign putative donors for 10 uncharacterized human GT-A domains. B3GNT9 is predicted to employ UDP-GlcNAc with high confidence like other GT31 β-3-N-acetylglucosaminyltransferases (B3GNTs) in humans [39]. The two procollagen galactosyltransferases in humans (COLGALT1 and COLGALT2) are multidomain proteins with two tandem GT-A domains. While their respective C-terminal domains catalyze β-Gal addition to hydroxylysine side-chains in collagen [40], our predictions assign a putative GlcNAc and Glc transferase role for their N-terminal GT domains, respectively. More interestingly, GLT8D1, a GT8 glycosyltransferase with an unknown function implicated in neurodegenerative diseases [41], is predicted to have a glucosyltransferase specificity. In other organisms, the GT2 sequences in *A. thaliana* (mostly involved in plant cell wall biosynthesis) are predicted to bind glucose and mannose substrates, the primary components of the plant cell wall. We also identify a novel N-acetylglucosyltransferase function for a GT25 enzyme in *C. elegans*. These predictions can guide characterization of new GT sequences with unknown functions.

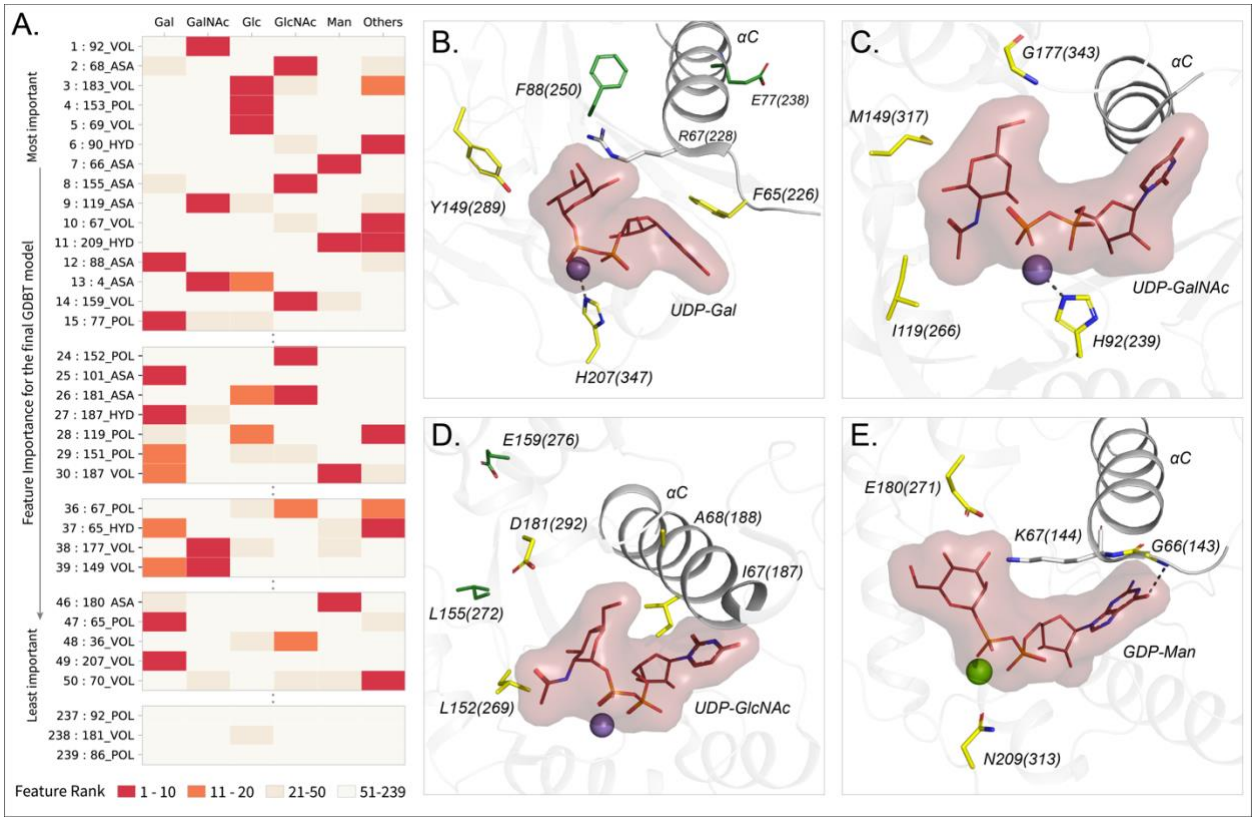


Figure 2.8: Top Contributing features from the GDBT model associated with sugar donor specificity.

A) Heatmap showing the contributions of representative features. Features are ordered based on their importance for the final GDBT model along the vertical axis. The heatmap colors indicate how important each feature is for a given sugar donor type with red indicating ranks 1-10 (highly important). B-E) Contributing features important for individual donor types are mapped onto representative structures. The amino acids at the feature positions are shown in yellow sticks and labelled. Feature positions distal from the donor binding site are shown in green sticks. Labels include the amino acid code, aligned residue position and the amino acid position in the crystal structure within parentheses. Donor substrate with the sugar is shown in lines with surface bounds. Divalent metal ions are shown as spheres. The αC helix is shown. B) Gal features mapped to a bovine β-1,4 Gal transferase (PDB ID: 1o0r). C) GalNAc features mapped to a human UDP-GalNAc: polypeptide alpha-N-acetylgalactosaminyltransferase (PDB ID: 2d7i). D)GlcNAc features mapped to a rabbit N-acetylglucosaminyltransferase I (PDB ID: 1foa). E) Man features mapped to a bacterial Mannosyl-3-Phosphoglycerate Synthase (PDB ID: 2wvl).

We next sought to identify features that contribute most to substrate (donor) prediction. To do this, we rank ordered the 239 features based on their contribution to predicting a donor subtype

using a six way classification (six donors). This revealed that the most contributing features of the GDBT model also contribute significantly to at least one specific donor type prediction, thereby enabling new inferences to be drawn between residue properties and donor sugar specificity (Figure 2.8A). As expected, some of the most contributing features include residues directly involved in substrate binding and catalytic functions such as the Asp within the DxD motif, residues in the G-loop, the catalytic base and the C-His [18,19,28]. Additionally, multiple residues from the alpha-C helix (aligned position 65-72; Y217-N224 in 2z87) immediately following HV1 are also identified as key specificity determining residues.

The C-helix is positioned close to the donor sugar binding pocket and many residues from this region have been shown to play roles in donor binding [42–44]. For example, Ramakrishnan *et. al.* showed that mutation of a single residue at position 67 in bovine β-1,4-galactosyltransferase T1 (R228K in 1o0r, GT7) resulted in relocation of the catalytic base and a change of donor specificity from Gal to Glc (Figure 2.8B)[45,46]. Our analysis identifies volume, polarity and accessible surface area of the residue at position 67 as an important contributor to donor specificity (Figure 2.8). In addition, our analysis identifies residue volume at position 149 as an important determinant of Gal specificity. Consistent with this observation, mutation of Y289 (position 149 in the consensus sequence) by a leucine broadens the specificity from Gal to GalNAc by creating additional space for accommodating the N-acetyl moiety [46,47].

While some of the highly ranked features are directly involved in donor binding, many others (such as aligned position 77, 88, 155 and 159, green sticks in Figure 2.8B,D) are distal from the donor binding site and are not directly involved in donor binding. An example of allostery has been observed in the human GT6 blood ABO α-1,3-galactosyltransfearse where mutation of a proline at position 117 (P234S in 5c4c) results in an alternative conformation of a methionine at position 150

(M266 in 5c4c) allowing for the accommodation of GalNAc instead of Gal [46,48]. Further, a Random forest model trained using features from only the donor binding residues performs with an accuracy of only 75%, indicating the importance of features other than those directly involved in donor binding. Thus, despite only a few residues being directly involved in donor interactions, additional contributions to donor specificity come from residues more distal from the active site. Contributions from these peripheral secondary shell features surrounding the donor binding site (Figure 2.8B-E) highlight the potential role of higher order (allosteric) interactions in determining donor substrate specificity.

2.3 Discussion

Prior studies on the evolution of GTs have generally focused either on distinct GT subfamilies or biosynthetic pathways with additional structural classifications of GTs into one of three distinct protein fold superfamilies [6,12,13]. In our present work we focused on the analysis of the largest of the GT superfamilies, those that comprise a GT-A protein fold characterized by an extended Rossmann domain with associated conserved helical segments. These enzymes generally employ the Rossmann domain for nucleotide sugar donor interactions and extended loop regions for acceptor glycan interactions [6]. Using an unbiased profile search strategy, we assembled a total of over 600,000 GT-A fold related sequences from all domains of life for deep evolutionary analysis. To support this profile-based assembly, we leveraged structural alignments on GT-A fold enzymes in PDB and secondary structure predictions when no crystal structures were available. The resulting alignment allowed the definition of a common structural core shared among the diverse GT-A fold enzymes and defined positions where hypervariable loop insertions were elaborated to provide additional functional diversification (Figure 2.2). In cases where data was available for enzyme-

acceptor complexes these latter loop insertions generally contribute to unique, family specific acceptor interactions. Thus, a structural framework is presented for GT-A fold enzyme evolution.

Since the common core is present across all kingdoms of life, it presumably represents the minimal ancestral structural unit for GT-A fold catalytic function by defining donor substrate interactions and minimal elements for acceptor recognition and catalysis. In fact, we find several archaeal and bacterial sequences that closely resemble this common core consensus sequence. Based on our studies, we propose a progressive diversification of glycosyltransferase function through evolution of donor specificity by accumulation of mutations in the common core region and divergence in acceptor recognition through expansion of the hypervariable loop regions. Consistent with this view, we find conserved family-specific motifs within the hypervariable regions that confer unique acceptor specificities in various families. These expansions likely contributed to the evolution of new GT functions and catalyzed new glycan diversification observed in all domains of life.

A surprising finding from our studies is the dispersion of inverting and retaining catalytic mechanisms among families in the GT-A fold evolutionary tree (Figure 2.3). Recent models indicate that distinctions between inverting and retaining catalytic mechanisms arise from differences in the angle of nucleophilic attack by the acceptor toward the anomeric center of the donor sugar [6]. Inverting mechanisms require an in-line attack and direct displacement by the nucleophile relative to the departing nucleotide diphosphate of the sugar donor and a conserved placement of the xED-Asp carboxyl group as catalytic base at the beginning of the αF helix. In contrast, retaining enzymes generally alter the angle of nucleophilic attack by the acceptor, use a donor phosphate oxygen as catalytic base, and employ a dissociative mechanism for sugar transfer [6]. The fundamental differences in these catalytic strategies would suggest an early divergence of

enzymes employing these respective mechanisms. However, the GT-A fold phylogenetic tree strongly suggests that inverting and retaining mechanisms evolved independently at multiple points in the evolution of GT-A families (Figure 2.3).

Since the main difference in these mechanisms is the change in position of the nucleophilic hydroxyl and catalytic base, substitutions at the catalytic base may have served as a catalyzing event in switching between mechanisms. The xED-Asp carboxyl group is highly conserved in the inverting enzymes and is appropriately placed for acceptor deprotonation. Variants of this motif either lack the residue entirely, as seen in many retaining enzymes, or use compensatory modes to accommodate changes at this position, as seen for the inverting enzymes in GT43, GT2-DPs, and GT2-LPSRelated. In fact, in each of the latter cases the respective inverting GT family is clustered with closely related GT families employing a retaining catalytic mechanism. Thus, inverting enzyme variants that accommodate changes to the xED motif group may represent examples of transitional phases in evolution between inverting and retaining catalytic mechanisms. Other inverting enzymes harboring variants in the xED motif segregate into separate clades and could represent outlier families that have developed alternative ways to compensate for the loss of xED-Asp. This ability to evolve distinct catalytic strategies, in some cases through presumed convergent evolution, could allow each family to evolve independent capabilities for donor and acceptor interactions as well as for anomeric linkage of sugar transfer, while retaining other essential aspects of protein structural integrity through the use of a conserved and stable Rossmann fold core.

In an effort to define the sequence constraints for the respective catalytic mechanisms we also employed a machine learning framework for prediction of the mechanism for unknown sequences and were able to assign the donor sugar nucleotide for a test set of enzymes with high accuracy. Our model expands on the approaches used in previous machine learning efforts focused on the

GT1 family of GT-B fold glycosyltransferases [36]. The phylogenetic and comparative framework presented here enables expansion of such models across all GT-A fold families with improved prediction accuracies. As additional functional data on GTs become available, the proposed machine learning framework can be extended to predict acceptor specificity and catalytic mechanisms, as described for the GT1 family. Surprisingly, the contributing features for accurate donor prediction include residues involved in donor binding as well as positions that are distal to the active site that likely contribute through secondary shell effects or allosteric interactions. Due to their indirect involvement, such positions are generally difficult to pinpoint using structural studies alone emphasizing the need for complementary machine learning based approaches in investigating GT functional specialization.

Numerous additional insights into GT function were also revealed through inspection of the aligned sequences and the phylogenetic tree. For example, the clustering of mammalian N-glycan GlcNAc branching enzymes (MGAT1 (GT13), MGAT2 (GT16), and MGAT4 (GT54)) in the same clade suggests a common origin for these enzymes, while placement of MGAT3 (GT17) in a separate clade could point to its unique role in adding a bisecting GlcNAc to the N-glycan core thereby regulating N-glycan extension [49]. In contrast, MGAT5 (GT18) involved in N-glycan β1,6-GlcNAc branching is a GT-B fold enzyme with a clearly distinct evolutionary origin. While most clades are well resolved, bootstrap support values for nodes at the base of the tree are low and need to be interpreted with caution. This low resolution results from high divergence between families and possibly other events like horizontal gene transfer and convergent evolution. However, trees generated using alternative strategies support the overall topology and clades are congruent with clusters obtained using an orthogonal Bayesian classification scheme, which adds confidence to the phylogeny.

For some GT-A fold enzymes variations in the catalytic site can also be accommodated by other compensatory changes. An example is the use of the C-His motif for coordination of the divalent cation in most GT-A fold enzymes in contrast with enzyme variants that employ water molecules to compensate for the loss of this residue (Figure 2.5B). Similarly, some inverting GTs dispense with the use of the divalent cation and the DxD motif and substitute interactions with the sugar donor through use of basic side chains (e.g. GT14). A further extreme is the duplication, divergence and pseudogenization within the GT31 family. Human C1GALT1C1 (GT31, COSMC) shares a high sequence similarity to another GT31 member, C1GALT1 (T-synthase), yet COSMC has lost both the DxD and the xED motifs and has no catalytic activity. Instead, COSMC acts as an important scaffold and chaperone for the proper assembly and catalytic function of T-synthase [38]. The ability of GT-As to harbor such structural variations that allow them to develop new functions make them well-suited to evolve rapidly and facilitate the synthesis of a diverse repertoire of glycans across all living organisms.

Our unbiased, top-down sequence-based analysis suggests new and unanticipated evolutionary relationships among the GT-A fold enzymes. Prior suggestions of such relationships have been inferred by the clustering of GT sequences into families in the CAZy database. However, the CAZy database of GT sequences does not provide access to the broader sequence relationships among the GT-A fold enzymes or how a general model of a core conserved GT-A fold scaffold can serve as a progenitor catalytic platform for binding sugar donors and facilitating glycan extension.

The sequence assembly, phylogenetic tree, and placement within the framework of known GT-A fold structures in the present studies provide key insights into conserved elements of the hydrophobic core, linkage to the DxD motif for cation and sugar donor interactions, and the conserved αF helix harboring the xED catalytic base. Additional hypervariable extensions at

defined positions from this conserved core were then progressively recruited to confer unique modes of acceptor interactions to develop new specificities and evolve new functions. Thus, the core of the protein scaffold can be maintained to facilitate protein stability while rapid evolution of the hypervariable loops can develop new glycan synthetic functionalities through presentation of novel acceptors to the catalytic site. Variation in the location of the acceptor hydroxyl nucleophile relative to the donor sugar anomeric center presents the opportunity for distinctions in catalytic mechanism and anomeric outcome for sugar transfer.

The result is a rapidly evolving set of GT enzymatic templates as the biosynthetic machinery for diverse glycan extension on cell surface and secreted glycoproteins and glycolipids. In such contexts the resulting glycoconjugates confer potential functional selective advantages at the cell surface, but also act as ligands and pathogen entry points for negative evolutionary pressure. These positive and negative selective pressures which force organisms to constantly adapt to an ever-changing environment is known as the Red Queen Hypothesis. These red queen effects on glycan synthesis have led to the remarkable diversity in GT enzymes and their resulting glycan structural products. We anticipate that the sequence and structural principles that drive GT-A fold evolution will also likely extend to GT-B and GT-C fold enzymes and represent a common theme for the elaboration of diverse glycan structures in all domains of life.

2.4 Methods

2.4.1 Generation of GT-A profiles and alignment

Building the GT-A profiles

Multiple alignments for 34 CAZy GT-A families, as determined based on literature [4,5,50,51], were collected from the Conserved Domain Database (CDD) [52] or were manually built using MAFFT v7.3 [53] from sequences curated at the CAZy database. Multiple separate alignments were generated for large families such as GT2 and GT8 to capture the diversity within these families. These alignments made up the seed profiles for the GT-A families. These seed profiles were then multiply aligned using the mapgaps scheme [17] guided by a structure based sequence alignments of all available pdb structures using Expresso [54] and MAFFT to generate the GT-A profiles. Alignments for families with no representative crystal structures were guided using secondary structure predictions performed using PCI-SS [55]. Finally, the alignment of secondary structures and conserved motifs were manually examined and corrected, where necessary. Very divergent GT-A families, such as GT29 and GT42 sialyltransferases, lack nearly all canonical GT-A motifs and do not align well with other GT-A families. Thus, they are noted as atypical GT-A fold families and not included in this analysis.

Sequence alignment and defining the GT-A common core

The GT-A profiles were then used for a sequence similarity search using mapgaps to identify and align ~600,000 GT-A domain sequences from the NCBI non redundant database. This alignment was filtered for fragmentary sequences and false hits. This filtered alignment was then used to

define the boundaries of the GT-A common core that extends from the first beta sheet of the Rossmann fold to a C-terminal helix with family specific motifs. This conserved alignment spanned 231 aligned positions. Sequences with multiple GT-A domains (like the GT8 and GT49 LARGE domains) or other accessory domains (like the GT27 and lectin domains) were separated into individual catalytic GT-A domains and treated separately throughout the analyses.

2.4.2 Structural alignment of Rossmann fold proteins

A select representative set of structures were collected from all Rossmann-fold containing protein domains using the SCOP database [56]. mTM-align [57] was used to align these structures with a subset of GT-A structures.

2.4.3 Bayesian Statistical analyses

A representative subset of 24,650 GT-A sequences were generated from the ~600,000 putative GT-A sequences by using a family-wise sequence similarity filtering (only keep <70% similar sequences; <50% for GT2 and GT8 families). This sequence set was then used to apply the Optimal multiple-category Bayesian Partitioning with Pattern Selection (omcBPPS) scheme [23]. omcBPPS identifies patterns of column-wise amino acid conservation and variation in the multiple sequence alignment. The resulting family specific positions were then used as statistical measures to classify the GT-As into 99 unique sets that correspond to the 53 families described in this study. omcBPPS also identified aligned positions that are conserved across all GT-A fold families. This revealed the 20 conserved positions within the core component, that were also verified by calculating conservation scores using the Jensen-Shannon divergence score as described and implemented by [58] (used in Figure 2.2A).

2.4.4 Phylogenetic analysis

Selection of sequences for phylogenetic analysis

A smaller subset of 993 sequences were used for phylogenetic reconstruction of the GT-A families. This set includes all the identified GT-A sequences from five model organisms: *H. sapiens* (human), *C. elegans* (worm), *D. melanogaster* (fly), *A. thaliana* (dicot plant) and *S. cerevisiae* (yeast) along with select sequences representing the diverse taxonomic group in each family. These representative sequences were selected by finding the union of top hits for every taxonomic group present within each of the 99 sets and the seed alignments for the 34 CAZy GT-A families. This selection criteria maximized the phylogenetic and taxonomic diversity while keeping the number of sequences to a minimum.

Details of the phylogenetic inference

The alignment for these 993 sequences was trimmed to remove the insert positions and keep only the 231 aligned positions described above. This trimmed alignment was used to build a phylogenetic consensus tree using IQTree v1.6.1 [59] with the following options: -nt AUTO -st AA -m MFP+MERGE -alrt 1000 -bb 1000 -wbt -nm 1000 -bnni. This implements ModelFinder [60] to select the best fit model based on Bayesian Information Criterion (BIC). Clade support for this tree was evaluated using bootstrapping which reports support values based on the number of times the same clade was observed on 1000 trees built using resampled alignment. Clades with bootstrap support values over 90% are well supported while values over 75% are moderately supported. Clades with bootstrap values less than 50% are considered unresolved in our analysis.

Orthogonal support for the phylogenetic tree

Further support for the phylogenetic tree was collected by comparing its topology to trees generated using orthogonal methods like Hidden Markov Model (HMM) distances and structural similarities, that have been used in previous studies [61,62]. The HMM-distance based phylogenetic tree was built using pHMM-Tree [61]. Briefly, hmm profiles were built for each of the 53 sub-families identified in our analyses. Pairwise distances between these profiles were calculated and the resulting distance matrix was used to build a neighbor joining tree. All trees were visualized using the interactive Tree of Life (iTOL) online tool [63]. For the structural similarity based clustering, pairwise root mean square distances (RMSD) were calculated for 50 unique representative GT-A structures using the cealign algorithm in PyMol v2.0.6 [64] to build a distance matrix. Only the defined GT-A catalytic domain spanning the 231 aligned positions along with insertions were used for the RMSD calculations. This RMSD matrix was then used for clustering using the "ward" method in python which resulted in a structural distance based hierarchical clustering of the pdb structures. The hierarchical topology obtained from the HMM distance-based method and the RMSD distance based clustering were then compared to the tree topology in Figure 2.3.

2.4.5 Defining the GT-A families and sub-families

The GT-A sequences were first classified into pattern-based groups using omcBPPS. Based on the placement of representative sequences from these groups in the phylogenetic tree, they were merged into GT-A families and sub-families. Sequences from some families did not form any distinct pattern-based groups due to either a low number of sequences for a statistically significant grouping (GT78) or a lack of distinguishing patterns within the aligned positions (GT25, GT88). Representative sequences for these families were collected from the seed alignments for these

families as described above. We also identified the N-terminal GT2 domain of the multidomain chondroitin polymerase structure from E. coli (Pdb Id: 2z87) as the prototypic GT-A structure to use as a comparative basis for structural analyses. This sequence was selected based on the lowest E-value and highest similarity score of a BLAST search of all pdb structures against the GT-A consensus sequences. Weblogos for the conserved active site residues were derived for each GT-A subfamily using Weblogo 3.6.0 [65].

2.4.6 Machine learning analysis

Gathering the training and validation dataset

In order to train an ML model for GT-A donor substrate prediction, we first curated a training dataset by mining the "characterized" tab of the CAZy GT database and the UniProt database [66] to find 713 GT-A domain sequences with known donor sugars. The donor sugar information for these sequences were extracted from their assigned protein names. Based on the availability of training sequences, 6 major donor type classes were defined: Glc, GlcNAc, Gal, GalNAc, Man, and "Others" with each class having more than 70 sequences in the training dataset. The "Others" category merged the least represented donor types with less than 50 training sequences each (Ara, Fuc, GalF, GlcA, ManNAc, Rham, and Xyl). An alignment of the 713 sequences was generated and then used to derive 5 amino acid properties (charge, polarity, hydrophobicity, average accessible surface area, and side chain volume) [67] for each aligned position. These properties were used as features for machine learning. We first removed highly gapped positions (>15% gaps) and implemented correlation-based feature selection (CFS) [68] with 5-fold CV by using WEKA version 3.8.3 [69] under default settings to select 239 informative features for building multiple multiclass classification models. In addition, we also curated 64 GT-A sequences with known donor

sugars for 5 model organisms (*H. sapiens*, *C. elegans*, *D. melanogaster*, *A. thaliana* and *S.cerevisiae*). These sequences were not used to train the ML model but set aside to be used as validation dataset to test the performance of the model.

Machine learning model training

We first trained random forest models by using an R package "randomForest" [70] with limited number of trees (ntree = 300) and limited maximum number of terminal nodes (maxnodes = 100) to avoid unrestricted tree expansion and potential overfitting. Two separate models were trained where the first one was trained with the larger set of 239 features and the second model was provided only 25 features coming from the donor binding residues. We used the GradientBoostingClassifier function of the sklearn package in python [71] to train the gradient boost regression tree (GDBT) model on the 239 features. This model was trained with the following parameters: learning_rate=0.1, n_estimators=1600, min_samples_split=25, min_samples_leaf=7, max_depth=4, max_features=18, subsample=0.75 and random_state=10. These parameters were chosen based on a grid search to fine tune the trade-off between the complexity of the model and the metrics on the testing data, thus ensuring meaningful predictions and avoiding overfitting.

The importance of each feature used in the GDBT model was measured based on the relative rank of the features in the decision nodes of a tree. To compare the performance of these models, we also trained Support Vector Machine (SVM), multilayer perceptron, Bayesian network, logistic regression, naive Bayes classifier, and decision tree models by using WEKA with 10-fold CV under default settings. 10-fold CV evaluates the ML models by iteratively training on 90% of the data selected at random and testing the prediction on the unseen 10% of the data. This is repeated 10 times and the results on the testing dataset are summarized into an accuracy measure. The GDBT

model trained with 239 features had the highest accuracy and overall performance and thus was selected as the model of choice for predicting donor sugar substrates for GT-A enzymes.

Evaluating the confidence of predictions

Confidence scores were assigned for each prediction based on the probability for each of the 6 donor classes. The class with the highest probability represents the predicted donor sugar. As such, larger differences in probability between the first and second predicted class result in more reliable predictions. To interpret this difference in score easily, we derive a three-category confidence level. If the probability for the first class is more than four times the probability of the second predicted class, then it is considered a high confidence prediction. If the difference is less than four times but more than double the probability by random chance (2*1/6 for a 6 class classification), it is considered a moderate confidence prediction. If it is neither, then it is a low confidence prediction.

Determining feature contributions for each donor sugar specificity

Feature importance for the GDBT model was first assessed using the relative rank of the features in the decision tree. Then, 6 separate GDBT models were trained as binary classifiers (Gal Vs everything else, Glc Vs everything else and so on for the 6 donor types). For each of these 6 classifiers, the features were rank ordered in the same way by assessing their rank in the decision tree nodes. This provided the contributions of each of the 239 features toward a specific donor specificity which was then compared to its rank in the full GDBT model.

Bibliography

[1] Varki, A., Gagneux, P., in:, Varki A, Cummings RD, Esko JD, Stanley P, et al. (Eds.), *Essent. Glycobiol.*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY) 2015.

[2] Ryan, P., Xu, M., Davey, A.K., Danon, J.J., et al., O-GlcNAc Modification Protects against Protein Misfolding and Aggregation in Neurodegenerative Disease. *ACS Chem. Neurosci.* 2019, 10, 2209–2221.

[3] Day, C.J., Semchenko, E.A., Korolik, V., Glycoconjugates Play a Key Role in Campylobacter jejuni Infection: Interactions between Host and Pathogen. *Front. Cell. Infect. Microbiol.* 2012, 2.

[4] Breton, C., Fournel-Gigleux, S., Palcic, M.M., Recent structures, evolution and mechanisms of glycosyltransferases. *Curr. Opin. Struct. Biol.* 2012, 22, 540–549.

[5] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 2014, 42, D490–D495.

[6] Moremen, K.W., Haltiwanger, R.S., Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat. Chem. Biol.* 2019, 15, 853–864.

[7] Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* 2008, 77, 521–555.

[8] Gloster, T.M., Advances in understanding glycosyltransferases from a structural perspective. *Curr. Opin. Struct. Biol.* 2014, 28, 131–141.

[9] Ramakrishnan, B., Qasba, P.K., Crystal structure of the catalytic domain of Drosophila beta1,4-Galactosyltransferase-7. *J. Biol. Chem.* 2010, 285, 15619–15626.

[10] Kadirvelraj, R., Yang, J.-Y., Sanders, J.H., Liu, L., et al., Human N-acetylglucosaminyltransferase II substrate recognition uses a modular architecture that includes a convergent exosite. *Proc. Natl. Acad. Sci. U. S. A.* 2018, 115, 4637–4642.

[11] Gordon, R.D., Sivarajah, P., Satkunarajah, M., Ma, D., et al., X-ray crystal structures of rabbit N-acetylglucosaminyltransferase I (GnT I) in complex with donor substrate analogues. *J. Mol. Biol.* 2006, 360, 67–79.

[12] Taujale, R., Yin, Y., Glycosyltransferase Family 43 Is Also Found in Early Eukaryotes and Has Three Subfamilies in Charophycean Green Algae. *PLOS ONE* 2015, 10, e0128409.

[13] Lombard, J., The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. *Biol. Direct* 2016, 11.

[14] Kannan, N., Taylor, S.S., Zhai, Y., Venter, J.C., Manning, G., Structural and Functional Diversity of the Microbial Kinome. *PLOS Biol.* 2007, 5, e17.

[15] Kwon, A., Scott, S., Taujale, R., Yeung, W., et al., Tracing the origin and evolution of pseudokinases across the tree of life. *Sci. Signal.* 2019, 12, eaav3810.

[16] Pruitt, K.D., Tatusova, T., Maglott, D.R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007, 35, D61-65.

[17] Neuwald, A.F., Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* 2009, 25, 1869–1875.

[18] Patenaude, S.I., Seto, N.O.L., Borisova, S.N., Szpacenko, A., et al., The structural basis for specificity in human ABO(H) blood group biosynthesis. *Nat. Struct. Biol.* 2002, 9, 685–690.

[19] Empadinhas, N., Pereira, P.J.B., Albuquerque, L., Costa, J., et al., Functional and structural characterization of a novel mannosyl-3-phosphoglycerate synthase from Rubrobacter xylanophilus reveals its dual substrate specificity. *Mol. Microbiol.* 2011, 79, 76–93.

[20] Fritz, T.A., Hurley, J.H., Trinh, L.-B., Shiloach, J., Tabak, L.A., The beginnings of mucin biosynthesis: The crystal structure of UDP-GalNAc:polypeptide α-N-acetylgalactosaminyltransferase-T1. *Proc. Natl. Acad. Sci. U. S. A.* 2004, 101, 15307–15312.

[21] Persson, K., Ly, H.D., Dieckelmann, M., Wakarchuk, W.W., et al., Crystal structure of the retaining galactosyltransferase LgtC from Neisseria meningitidis in complex with donor and acceptor sugar analogs. *Nat. Struct. Biol.* 2001, 8, 166–175.

[22] Sanderson, M.J., Driskell, A.C., The challenge of constructing large phylogenetic trees. *Trends Plant Sci.* 2003, 8, 374–379.

[23] Neuwald, A.F., A Bayesian Sampler for Optimization of Protein Domain Hierarchies. *J. Comput. Biol.* 2014, 21, 269–286.

[24] Morgan, J.L.W., Strumillo, J., Zimmer, J., Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* 2013, 493, 181–186.

[25] Ciocchini, A.E., Roset, M.S., Briones, G., de Iannino, N.I., Ugalde, R.A., Identification of active site residues of the inverting glycosyltransferase Cgs required for the synthesis of cyclic β-1,2-glucan, a Brucella abortus virulence factor. *Glycobiology* 2006, 16, 679–691.

[26] Alonso, M.D., Lomako, J., Lomako, W.M., Whelan, W.J., A new look at the biogenesis of glycogen. *FASEB J.* 1995, 9, 1126–1137.

[27] Nagae, M., Kizuka, Y., Mihara, E., Kitago, Y., et al., Structure and mechanism of cancer-associated N -acetylglucosaminyltransferase-V. *Nat. Commun.* 2018, 9, 1–12.

[28] Gandini, R., Reichenbach, T., Tan, T.-C., Divne, C., Structural basis for dolichylphosphate mannose biosynthesis. *Nat. Commun.* 2017, 8, 1–12.

[29] Briggs, D.C., Hohenester, E., Structural Basis for the Initiation of Glycosaminoglycan Biosynthesis by Human Xylosyltransferase 1. *Struct. England1993* 2018, 26, 801-809.e3.

[30] Ramakrishnan, B., Qasba, P.K., Structure-based Evolutionary Relationship of Glycosyltransferases: A Case Study of Vertebrate β1, 4-Galactosyltransferase, Invertebrate β1,4-N-acetylgalactosaminyltransferase and α-Polypeptidyl-N-acetylgalactosaminyltransferase. *Curr. Opin. Struct. Biol.* 2010, 20, 536–542.

[31] Heise, N., Singh, D., van der Wel, H., Sassi, S.O., et al., Molecular analysis of a UDP-GlcNAc:polypeptide α-N-acetylglucosaminyltransferase implicated in the initiation of mucin-type O-glycosylation in Trypanosoma cruzi. *Glycobiology* 2009, 19, 918–933.

[32] Pham, T.T.K., Stinson, B., Thiyagarajan, N., Lizotte-Waniewski, M., et al., Structures of Complexes of a Metal-independent Glycosyltransferase GT6 from Bacteroides ovatus with UDP-N-Acetylgalactosamine (UDP-GalNAc) and Its Hydrolysis Products. *J. Biol. Chem.* 2014, 289, 8041–8050.

[33] Jamaluddin, H., Tumbale, P., Withers, S.G., Acharya, K.R., Brew, K., Conformational Changes Induced by Binding UDP-2F-galactose to α-1,3 Galactosyltransferase- Implications for Catalysis. *J. Mol. Biol.* 2007, 369, 1270–1281.

[34] Tsutsui, Y., Ramakrishnan, B., Qasba, P.K., Crystal Structures of β-1,4-Galactosyltransferase 7 Enzyme Reveal Conformational Changes and Substrate Binding. *J. Biol. Chem.* 2013, 288, 31963–31970.

[35] Albesa-Jové, D., Romero-García, J., Sancho-Vaello, E., Contreras, F.-X., et al., Structural Snapshots and Loop Dynamics along the Catalytic Cycle of Glycosyltransferase GpgS. *Structure* 2017, 25, 1034-1044.e3.

[36] Yang, M., Fehl, C., Lees, K.V., Lim, E.-K., et al., Functional and informatics analysis enables glycosyltransferase activity prediction. *Nat. Chem. Biol.* 2018, 14, 1109–1117.

[37] Blixt, O., van Die, I., Norberg, T., van den Eijnden, D.H., High-level expression of the Neisseria meningitidis lgtA gene in Escherichia coli and characterization of the encoded N-acetylglucosaminyltransferase as a useful catalyst in the synthesis of GlcNAcβ1→3Gal and GalNAcβ1→3Gal linkages. *Glycobiology* 1999, 9, 1061–1071.

[38] Aryal, R.P., Ju, T., Cummings, R.D., Tight Complex Formation between Cosmc Chaperone and Its Specific Client Non-native T-synthase Leads to Enzyme Activity and Client-driven Dissociation. *J. Biol. Chem.* 2012, 287, 15317–15329.

[39] Togayachi, A., Kozono, Y., Kuno, A., Ohkura, T., et al., in:, Fukuda M (Ed.), *Methods Enzymol.*, vol. 479, Academic Press, 2010, pp. 185–204.

[40] Schegg, B., Hulsmeier, A.J., Rutschmann, C., Maag, C., Hennet, T., Core glycosylation of collagen is initiated by two beta(1-O)galactosyltransferases. *Mol. Cell. Biol.* 2009, 29, 943–952.

[41] Cooper-Knock, J., Moll, T., Ramesh, T., Castelli, L., et al., Mutations in the Glycosyltransferase Domain of GLT8D1 Are Associated with Familial Amyotrophic Lateral Sclerosis. *Cell Rep.* 2019, 26, 2298-2306.e5.

[42] Gagnon, S.M.L., Legg, M.S.G., Polakowski, R., Letts, J.A., et al., Conserved residues Arg188 and Asp302 are critical for active site organization and catalysis in human ABO(H) blood group A and B glycosyltransferases. *Glycobiology* 2018, 28, 624–636.

[43] Schuman, B., Persson, M., Landry, R.C., Polakowski, R., et al., Cysteine-to-serine mutants dramatically reorder the active site of human ABO(H) blood group B glycosyltransferase without affecting activity: structural insights into cooperative substrate binding. *J. Mol. Biol.* 2010, 402, 399–411.

[44] McArthur, J.B., Chen, X., Glycosyltransferase engineering for carbohydrate synthesis. *Biochem. Soc. Trans.* 2016, 44, 129–142.

[45] Ramakrishnan, B., Boeggeman, E., Qasba, P.K., Mutation of arginine 228 to lysine enhances the glucosyltransferase activity of bovine beta-1,4-galactosyltransferase I. *Biochemistry* 2005, 44, 3202–3210.

[46] Hancock, S.M., Vaughan, M.D., Withers, S.G., Engineering of glycosidases and glycosyltransferases. *Curr. Opin. Chem. Biol.* 2006, 10, 509–519.

[47] Ramakrishnan, B., Qasba, P.K., Structure-based design of beta 1,4-galactosyltransferase I (beta 4Gal-T1) with equally efficient N-acetylgalactosaminyltransferase activity: point mutation broadens beta 4Gal-T1 donor specificity. *J. Biol. Chem.* 2002, 277, 20833–20839.

[48] Marcus, S.L., Polakowski, R., Seto, N.O.L., Leinala, E., et al., A single point mutation reverses the donor specificity of human blood group B-synthesizing galactosyltransferase. *J. Biol. Chem.* 2003, 278, 12403–12405.

[49] Ikeda, Y., Ihara, H., Tsukamoto, H., Gu, J., Taniguchi, N., in:, Taniguchi N, Honke K, Fukuda M, Narimatsu H, et al. (Eds.), *Handb. Glycosyltransferases Relat. Genes*, Springer Japan, Tokyo 2014, pp. 209–222.

[50] Liu, J., Mushegian, A., Three monophyletic superfamilies account for the majority of the known glycosyltransferases. *Protein Sci. Publ. Protein Soc.* 2003, 12, 1418–1431.

[51] Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J., Imberty, A., Structures and mechanisms of glycosyltransferases. *Glycobiology* 2006, 16, 29R-37R.

[52] Marchler-Bauer, A., Bo, Y., Han, L., He, J., et al., CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 2017, 45, D200–D203.

[53] Katoh, K., Standley, D.M., MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* 2013, 30, 772–780.

[54] Armougom, F., Moretti, S., Poirot, O., Audic, S., et al., Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* 2006, 34, W604–W608.

[55] Green, J.R., Korenberg, M.J., Aboul-Magd, M.O., PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. *BMC Bioinformatics* 2009, 10, 222.

[56] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., Murzin, A.G., SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 2014, 42, D310–D314.

[57] Dong, R., Peng, Z., Zhang, Y., Yang, J., mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinforma. Oxf. Engl.* 2018, 34, 1719–1725.

[58] Capra, J.A., Singh, M., Predicting functionally important residues from sequence conservation. *Bioinformatics* 2007, 23, 1875–1882.

[59] Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* 2015, 32, 268–274.

[60] Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Haeseler, A. von, Jermiin, L.S., ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 2017, 14, 587–589.

[61] Huo, L., Zhang, H., Huo, X., Yang, Y., et al., pHMM-tree: phylogeny of profile hidden Markov models. *Bioinformatics* 2017, 33, 1093–1095.

[62] Hashimoto, K., Madej, T., Bryant, S.H., Panchenko, A.R., Functional states of homooligomers: insights from the evolution of glycosyltransferases. *J. Mol. Biol.* 2010, 399, 196–206.

[63] Letunic, I., Bork, P., Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 2019, 47, W256–W259.

[64] The PyMOL Molecular Graphics System, Version 2 Schrödinger,LLC. n.d.

[65] Crooks, G.E., WebLogo: A Sequence Logo Generator. *Genome Res.* 2004, 14, 1188–1190.

[66] UniProt Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019, 47, D506–D515.

[67] Kawashima, S., Ogata, H., Kanehisa, M., AAindex: Amino Acid Index Database. *Nucleic Acids Res.* 1999, 27, 368–369.

[68] Hall, M.A., *Correlation-based Feature Selection for Machine Learning*, 1999.

[69] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA 2016.

[70] Liaw, A., Wiener, M., Classification and Regression by randomForest 2002, 2, 5.

[71] Pedregosa, A Fabian, Varoquaux, G., Gramfort, A., Michel, V., et al., Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011, 12, 2825–2830.

CHAPTER 3

MAPPING THE GLYCOSYLTRANSFERASE FOLD LANDSCAPE USING DEEP LEARNING

Rahil Taujale, Zhongliang Zhou, Wayland Yeung, Kelley W Moremen, Sheng Li and Natarajan Kannan. *Submitted to Nature Communications*.

Abstract

Glycosyltransferases (GTs) play fundamental roles in nearly all cellular processes through the biosynthesis of complex carbohydrates and glycosylation of diverse protein and small molecule substrates. The extensive structural and functional diversification of GTs presents a major challenge in mapping the relationships connecting sequence, structure, fold and function. Here, we present a convolutional neural network with attention (CNN-attention) based deep learning model that leverages simple secondary structure representations generated from primary sequences to provide GT fold prediction with high accuracy. The model learned distinguishing features free of primary sequence alignment constraints and identified common cores within the protein folds, while classifying them into distinct clusters that group evolutionarily divergent families based on shared secondary structural features. We further extend our model to classify GT families of unknown folds and expand the GT fold landscape by identifying families with novel folds.

Author Contributions:

Conceptualization: RT, ZZ, SL, NK. Data curation and collection: RT, ZZ. Formal analysis: RT, ZZ. Methodology: RT, ZZ, SL, NK. Validation: RT, ZZ. Visualization and figure generation: RT, ZZ, WY. Software: RT, ZZ, WY. Supervision: SL, KWM, NK. Writing – Original Draft: RT, ZZ, NK. Writing – review and editing: RT, ZZ, WY, SL, KWM, NK.

3.1 Introduction

Glycosyltransferases (GTs) are a large family of enzymes tasked with the biosynthesis of complex carbohydrates that make up the bulk of biomass in cells [1]. Prevalent across the tree of life, these enzymes catalyze the transfer of a sugar molecule from an activated donor (mostly nucleotide sugars or dolichol-(pyro)phosphate linked sugars) to a wide variety of acceptors ranging from proteins and fatty acids to other carbohydrate molecules. The CAZy database [2] classifies over half a million GT sequences across organisms into 110 families based on overall sequence similarity. While sequences within families share detectable sequence similarity, sequences across families share little or no similarity [3]. The extensive diversification of GT sequences presents a major bottleneck in investigating the relationships connecting sequence, structure, fold and function.

As with other large protein families, GTs also exhibit a much higher conservation in 3D structural fold compared to primary sequences [4–6]. Across all 110 families, only 3 major folds have been identified (GT-A, -B and -C folds) with some families adopting other unique folds [1,7,8]. Recently, we proposed a phylogenetic framework relating diverse GT-A fold enzymes leveraging the common structural features identified through structure guided curation of large multiple sequence alignments [3]. While such multiple sequence alignment-based approaches have provided new insights into GT-A fold structure and evolution, such approaches are not scalable to other GT-folds for which there is limited structural data or limited structural homology.

The recent explosion of deep learning methods, in particular multi-layer neural networks, offer new opportunities for sequence classification and fold prediction through feature extraction and pattern recognition in large complex datasets [9,10]. The most recent successful application of these methods has been in the area of protein structure prediction in which the deep learning model

extracts residue co-variation from multiple sequence alignments to predict residue contacts in 3D structures [11–13]. Of note is Alphafold2 [14], an attention-based model that significantly outperformed previous best results in the biennial CASP assessment [15]. Other related efforts have focused on making residue level predictions such as disorder, solvent accessibility and post-translational modifications [16–19] using evolutionary information encoded in multiple sequence alignments. In these applications, the accuracy of predictions rely heavily on the quality of input multiple sequence alignments and these models cannot be directly extended for the study of divergent protein families such as GTs for which generating accurate multiple sequence alignments is a challenge for reasons mentioned above. Furthermore, the black-box nature of existing deep learning models prevents a direct biological interpretation of sequence or evolutionary features contributing to structure or fold prediction.

Here, we report a new convolutional neural network [20] with attention (CNN-attention) based model for GT fold type prediction solely based on secondary structure (ss) annotations as input. These coarse-grained input features are based on the premise that protein secondary and tertiary structures are far more conserved than primary sequences. Our model makes no use of amino acid physicochemical properties nor does it rely on generating evolutionary or alignment-based information and yet, achieves an average accuracy of 96% on fold prediction, and 77% on family classification. By using specially designed attention [21] modules, the trained model can generate highly interpretable activation maps that help locate conserved segments within sequences that point to the common cores within folds. We further leverage recent advances in open set recognition [22] and use a specially modified reconstruction error loss term to determine similarities between GTs so as to expand our model beyond known GT folds. The major advantages of our model are three-fold: 1) We propose an alignment free method to explore protein folds by leveraging ss

prediction as input data. 2) We focus on the interpretability of the model to mine features learned by the model and make meaningful biological inferences. 3) We extend our trained model to make predictions on GT families with unknown folds and report the ones most likely to adopt novel fold types to guide further research on discovery of novel glycoenzymes. The approach is applicable to other broad, heterogeneous protein families where challenges in primary sequence alignment approaches have hindered analysis of fold classification and evolutionary relationships.

3.2 Results

3.2.1 A deep learning framework to identify, classify and predict glycosyltransferase folds

We first sought to develop a deep learning model that could distinguish the features of glycosyltransferase (GT) structural folds from the large amount of readily available sequence information. To this end, we collected over half a million GT sequences from the CAZy database and filtered them based on sequence similarity, length and other criteria (see Methods) to generate a representative set of 44,620 GT sequences spanning all folds and families for training. Previous large-scale analysis of GTs [3,23] have revealed that the overall organization of the ss are far more conserved within folds than primary sequences. Therefore, we identified ss patterns using NetSurfP2.0 [19] and used them as the only input to train a 6-layer convolutional neural network (CNN) model for multitask fold and family classification (Figure 3.1). After refinement by the addition of attention modules and data augmentation strategies (Methods), the final optimized model achieves fold prediction with 96% accuracy and family classification with 77% accuracy,

based on 10-fold cross validation. Results for this final model highlighting the effects of data augmentation and the addition of multitasking and attention modules are provided in Table 3.1.

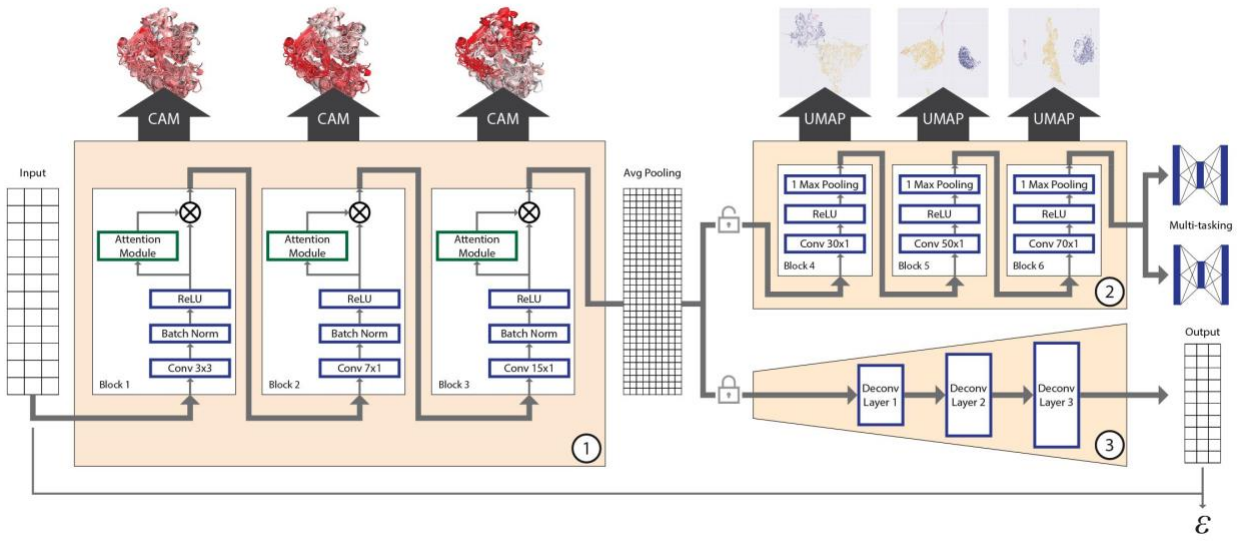


Figure 3.1: Overall schematics of the deep learning model used. A 3-state secondary structure prediction matrix for each sequence is used as input to Block 1. Block 1 includes the first 3 sequential one dimensional convolutional layers with attention for feature maps refinement. Feature maps from Block 1 are first passed through a global average pooling for dimension reduction and then fed into Block 2 with 3 additional convolutional layers, and finally used to make predictions for both fold and family. Blocks 1 and 2 constitute the deep CNN model for classification. Using GradCAM, features from block 1 are mapped back into sequences and structures for interpretation. Features from Block 2 are passed to UMAP for dimensionality reduction and visualization. Weights and features from Block 1 are frozen and used in an encoder that is passed to Block 3 which is the decoder with multiple deconvolution steps that completes an autoencoder model. Reconstruction error (ε) from this model is used to make predictions of fold type on GT families with unknown folds (GT-u).

The first three layers of our CNN model (Block 1, Figure 3.1) learn different levels of patterns in conserved ss features that can be projected as Class-specific Activation Maps using Grad-CAM [24,25] (CAM) on linear sequences and 3D structures. This enables us to highlight the

distinguishing features of a given GT fold recognized by the model. The last three layers (Block 2, Figure 3.1) further optimize the associated feature weights before feeding them into a fully connected multitask classifier to generate a classification with high accuracy. We extract these optimized feature embeddings and analyze them using dimensionality reduction by Uniform Manifold Approximation and Projection (UMAP) [26] to visualize the classification. In contrast to the more prevalent "black box" deep learning models, this architecture results in a highly interpretable model [27] with quantitative outputs to evaluate each step with high scrutiny and draw meaningful insights into ss patterns associated with GT function and fold.

Table 3.1: A complete comparison of different modules in the CNN-attention model. The model was trained with the same hyperparameter settings: learning rate at 5e-5 and weight decay rate at 1e-5 using Adam optimizer. Datasets were separated as augmented (Aug) or non-augmented (Non-Aug) to make comparison of the effect of the data augmentation method. The effects of multitask learning and attention modules were also tested.

Model	Dataset	Target	Precision	Recall	Accuracy	F1-score
CNN+ Multitask	Non-Aug	Fold	0.85	0.87	0.87	0.86
		Family	0.23	0.29	0.29	0.23
CNN+ Multitask+ Attention	Non-Aug	Fold	0.90	0.90	0.90	0.89
		Family	0.29	0.35	0.35	0.29
CNN+ Multitask	Aug	Fold	0.92	0.92	0.92	0.92
		Family	0.59	0.57	0.57	0.55
CNN+ Attention	Aug	Fold	0.96	0.96	0.96	0.96
		Family	0.72	0.69	0.69	0.69
CNN+ Multitask+ Attention	Aug	Fold	**0.96**	**0.96**	**0.96**	**0.96**
		Family	**0.78**	**0.77**	**0.77**	**0.77**

These two blocks also allow us to classify GT families of unknown structure into known folds, or assign them to novel folds. To classify GT families of unknown structure or fold, we integrate an autoencoder framework to our existing model in which the optimized weights from block 1 are frozen and used as a general feature extractor for the encoder. Block 3 (Figure 3.1) is then designed as a decoder with mirror structure of the CNN model that performs deconvolution operations. Applying the concepts of open set recognition framework that aim to extend knowledge from observed samples (closed set) to unseen samples (open set), we generate reconstruction errors (RE) using a modified mean square error, which measures how close a sequence with an unknown fold is to one of the known GT folds used in training (Methods) [28]. This measure is then used to identify GT families that are most likely to adopt novel folds. We discuss the results from the three blocks of our model in the following sections.

3.2.2 A landscape of all GT folds reveals distinct clusters within major fold types

We visualized feature maps generated from the three layers of Block 2 with the UMAP algorithm [26] (Figure 3.2A). As expected, we find separations between all the major GT folds, highlighting the model's ability to distinguish them. Sequences from the same GT family cluster together throughout, indicating the conservation of ss and the overall fold within individual families. Moreover, we find distinct substructures for the GT-A, -B and -C fold types. To further analyze these substructures, we first ran separate UMAP analyses on each of the three fold types and clustered the resulting projections using the Gaussian Mixture Model (GMM) algorithm [29] to identify clusters within the major GT fold types. This resulted in two GT-A clusters and three GT-B and GT-C clusters.

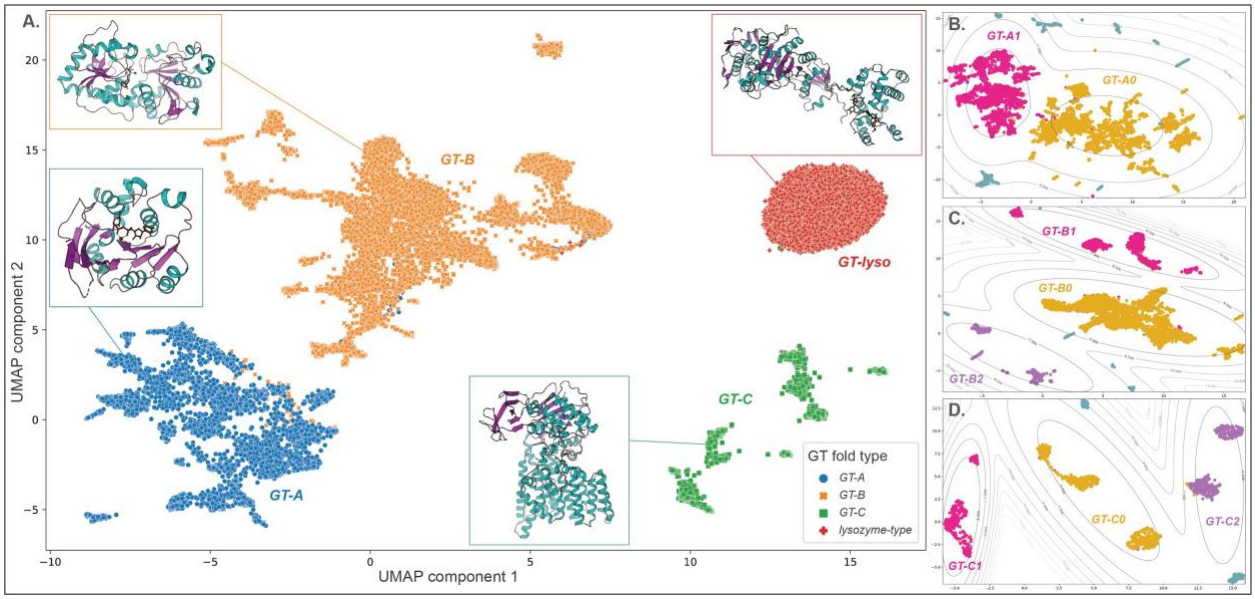


Figure 3.2: UMAP projection shows separation of the major GT fold types. Dots represent 2D UMAP projection of features for individual sequences. (A) Dots are colored based on their fold type and show a clear separation. Representative structures for each fold type are also shown. UMAP was applied separately on each major fold type and the projections for sequences belonging to the GT-A (panel B), GT-B (panel C) and GT-C (panel D) folds are shown. Clustering was done on these projections based on a Gaussian Mixture Model (GMM). Gray lines represent the contour for GMM scores around each cluster. Sequences that belong to a cluster are colored in yellow, magenta or purple and are labelled with the cluster name. Sequences that do not belong to any cluster are colored in teal.

The two distinct GT-A clusters accounted for most of the families with 17 out of 34 families grouping into a larger GT-A0 cluster. Ten families grouped into the GT-A1 cluster while the remaining 7 families did not group and scattered away from the two central clusters (Figure 3.2B). Sixteen out of 32 GT-B families used in training fall within the central GT-B0 cluster, while other families are spread out into smaller sub clusters (5 families in GT-B1, 6 in GT-B2 and 5 families ungrouped) (Figure 3.2C). Likewise, GT-C sequences are also scattered across three major clusters (Figure 3.2D) with only 2 out of 10 families (Alg10 glucosyltransferases of GT59 and the bacterial GT85 family) not grouped into any of the 3 clusters. We discuss the structural basis for the

separation of these GT-A, -B and -C clusters in the following sections. In contrast, the lysozyme type GT fold sequences (GT-lyso) all cluster into a single compact cluster indicating the structural similarity within this fold type and its stark difference from all the other fold types. A list of families belonging to each of these identified clusters is provided in Table 3.2. The 2D UMAP projection also shows several outlier sequences that do not fit within individual clusters. These sequences were either fragments that lack an entire GT domain, or display ss patterns significantly different from related family members.

Table 3.2: List of GT families and their corresponding fold and cluster. A Gaussian Mixture Model (GMM) was used to cluster families based on their 2D UMAP projections generated separately for each fold type. Families with a GMM score above -7.5 for GT-A, -7 for GT-B and -6.5 for GT-C were placed in a cluster.

Fold	Cluster	Families	GMM Score
GT-A	**GT-A0**	GT16	-5.464
		GT2	-5.524
		GT60	-5.642
		GT14	-5.688
		GT45	-5.694
		GT25	-5.785
		GT78	-5.786
		GT49	-5.945
		GT21	-6.060
		GT27	-6.150
		GT84	-6.307
		GT13	-6.346
		GT24	-6.381
		GT81	-6.477
		GT8	-6.500
		GT32	-6.599
		GT12	-7.134
	GT-A1	GT31	-5.309
		GT15	-5.494
		GT17	-5.566
		GT7	-5.657
		GT77	-5.716
		GT43	-5.777
		GT34	-5.940
		GT67	-5.945
		GT62	-6.155
		GT6	-6.906

	GT-A-Ungrouped	GT88	-7.635
		GT64	-8.277
		GT54	-8.510
		GT82	-9.230
		GT55	-9.308
		GT40	-10.256
		GT75	-11.761
GT-B	**GT-B0**	GT9	-4.539
		GT90	-4.614
		GT72	-4.620
		GT93	-4.639
		GT1	-4.824
		GT4	-4.837
		GT63	-4.837
		GT79	-4.847
		GT38	-4.904
		GT10	-5.109
		GT20	-5.238
		GT37	-5.267
		GT28	-5.505
		GT47	-5.912
		GT5	-5.965
		GT61	-6.830
	GT-B1	GT41	-5.310
		GT3	-5.656
		GT35	-5.760
		GT23	-6.101
		GT19	-6.834
	GT-B2	GT65	-6.139
		GT104	-6.484
		GT30	-6.673
		GT56	-6.829
		GT80	-6.840
		GT70	-6.889
	GT-B-Ungrouped	GT33	-7.144
		GT18	-7.231
		GT94	-7.454
		GT52	-8.463
		GT68	-8.781
GT-C	**GT-C0**	GT39	-5.025
		GT66	-5.590
		GT57	-5.851
	GT-C1	GT87	-4.413
		GT58	-5.296
		GT50	-6.058
	GT-C2	GT22	-5.371
		GT83	-5.807
	GT-C-Ungrouped	GT85	-6.935
		GT59	-8.834
GT-lyso	**GT-lyso**	GT51	-

3.2.3 CAM maps for GT-A clusters highlight differences in shared structural features

In order to understand the structural features of the major GT folds and their respective clusters, we mapped the CAM values obtained from each of the first three layers of the CNN model back to their respective sequences. In our previous work [3], we identified a common core shared by all GT-A fold enzymes. We first mapped the CAM values back to this GT-A common core alignment (Figure 3.3A,B). We find that the regions with highest conservation in the GT-A core (such as the DXD motif, G-loop and the first two beta sheets of the characteristic Rossmann-fold) correspond to the regions with the highest CAM values, indicating that the model is using these conserved regions to distinguish the GT-A fold from other GT fold types. It is important to emphasize that while our previous analysis required a laborious curation of the profiles and alignment to identify these regions, our current CNN-attention model was able to recognize and utilize these regions without any prior information or sequence alignment but only based on the predicted patterns of conserved ss across sequences.

CAM maps generated from layer 2 were the most informative and matched well with the core features of the GT-A fold. Layer 1 CAM values correspond to minute regions scattered throughout the domain and likely indicate local features learned by the model while CAM values from layer 3 extends over longer contiguous regions (Figure 3.3C), possibly capturing higher-order long-range correlations. We use these feature maps form all the 3 layers to identify conserved regions shared across different families within given clusters and folds.

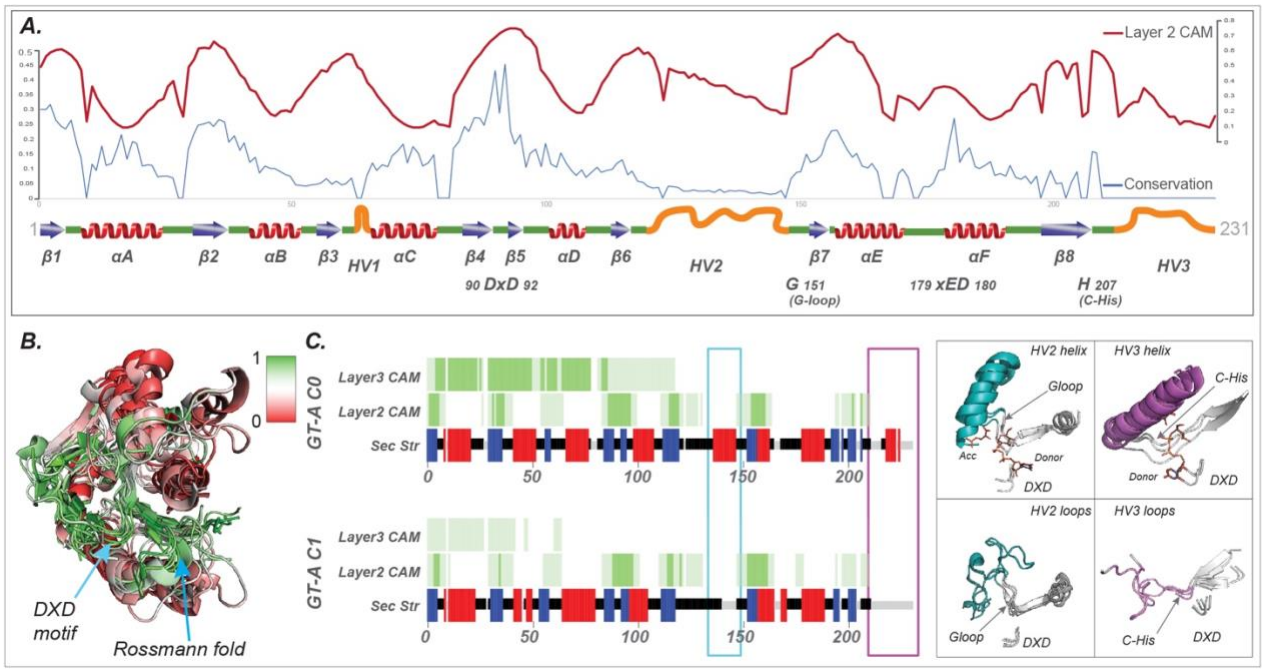


Figure 3.3: CAM highlights the GT-A fold core. A) The activation values from layer 2 are plotted in red line on top of the conserved secondary structures (blue arrows: Beta sheets; red: helices; green: loops, orange: Hypervariable regions) and the conservation scores in blue line. The most conserved regions generally have higher activation values. B) CAM values are mapped on to a structural alignment of the GT-A conserved core. The conserved regions are shown to have a high CAM value indicated by high intensity of green. C) Left: Consensus ss for the aligned positions in the 2 GT-A fold clusters are shown (blue: beta sheets; red: helices; green: loops). Average CAM values from layer 2 and layer 3 of the CNN-attention model are shown for each aligned position (higher intensity of green corresponds to a higher CAM value). Cyan and magenta boxes highlight the ss differences between the 2 clusters near the HV2 and HV3 region respectively. Right: The regions with differences in ss are shown in representative structures from each cluster (GT-A0: GT81 family structures 3ckq, 3o3p, 4y6n; GT-A1: GT6 family structures 5c4b, 5nrb and GT7 family structures 2ae7, 4lw6) and highlighted in cyan and magenta. The conserved DXD motif, G-loop and C-His are indicated for reference. Donor and acceptor substrates for GT-A0 are shown as sticks.

UMAP projection and clustering indicate the presence of 2 GT-A clusters (Figure 3.2B). GT-A cluster 0 (GT-A0) primarily constitutes large and phylogenetically distinct GT-A families such as GT2 and GT8 along with their closely related counterparts like GT84 (β-1,2-glucan synthases), GT21 (ceramide β-glucosyltransferases) and GT24 (glycoprotein α-glucosyltransferases) (Table 3.2). This cluster includes more than half of all the GT-A sequences used in training and represents a consensus ss that most closely matches the conserved core of the GT-A fold. The GT-A1 cluster includes GT31 and closely related families like GT15 and GT67. It also includes phylogenetically and functionally diverse families like GT7, GT77 and GT6. Meanwhile, families such as GT88 (bacterial Lgt1 sequences known to include large multi-helix insertions [30]), GT75 (that includes the self-glucosylating β-glucosyltransferases and UDP-L-arabinopyranose mutases), GT54 (MGAT4), and a few others are isolated away from the 2 main clusters, indicating some distinction in their ss patterns from other GT-A families.

In contrast to the GT-A1 consensus, GT-A0 families are distinguished by helical segments: the first one in the hypervariable region 2 (HV2) preceding the G-loop and the second one in the C-terminal HV3 region following the C-His position (Figure 3.3C). Both of these helices have been previously shown to include family specific residues directly involved in donor and acceptor binding [3], suggesting a conservation of these structural features of substrate binding within GT-A0 families mediated by the distinguishing helical segments. The ability of our model to cluster the evolutionarily divergent GT-A0 families based on the conservation of these helices highlights the value of our CNN-attention model in identifying convergent substrate binding mechanisms that are difficult to infer using traditional phylogenetic approaches.

3.2.4 The multiple levels of conserved core in GT-B and GT-C clusters

Our analysis identified a large central GT-B cluster (GT-B0) that includes some of the largest GT families such as GT4 with diverse functions and donor substrates, the UDP glucose/glucuronosyltransferases of GT1, GT5 sequences involved in glycogen and starch biosynthesis and lipopolysaccharide GlcNAc transferases of the GT9 family. Other families that cluster together include the fucosyltransferases from GT10 and GT37, trehalose phosphate synthases from GT20, the xylosyl-/glucosyl-transferases from GT90 and others. Clearly, families with a variety of functions including the largest and one of the most ancient families (GT4, which is also present in Archaea) are grouped together into a single cluster suggesting shared structural similarities within the GT-B fold. We additionally identify two other GT-B clusters, GT-B1 and GT-B2, both of which are slightly more sparse than GT-B0 and include fewer families.

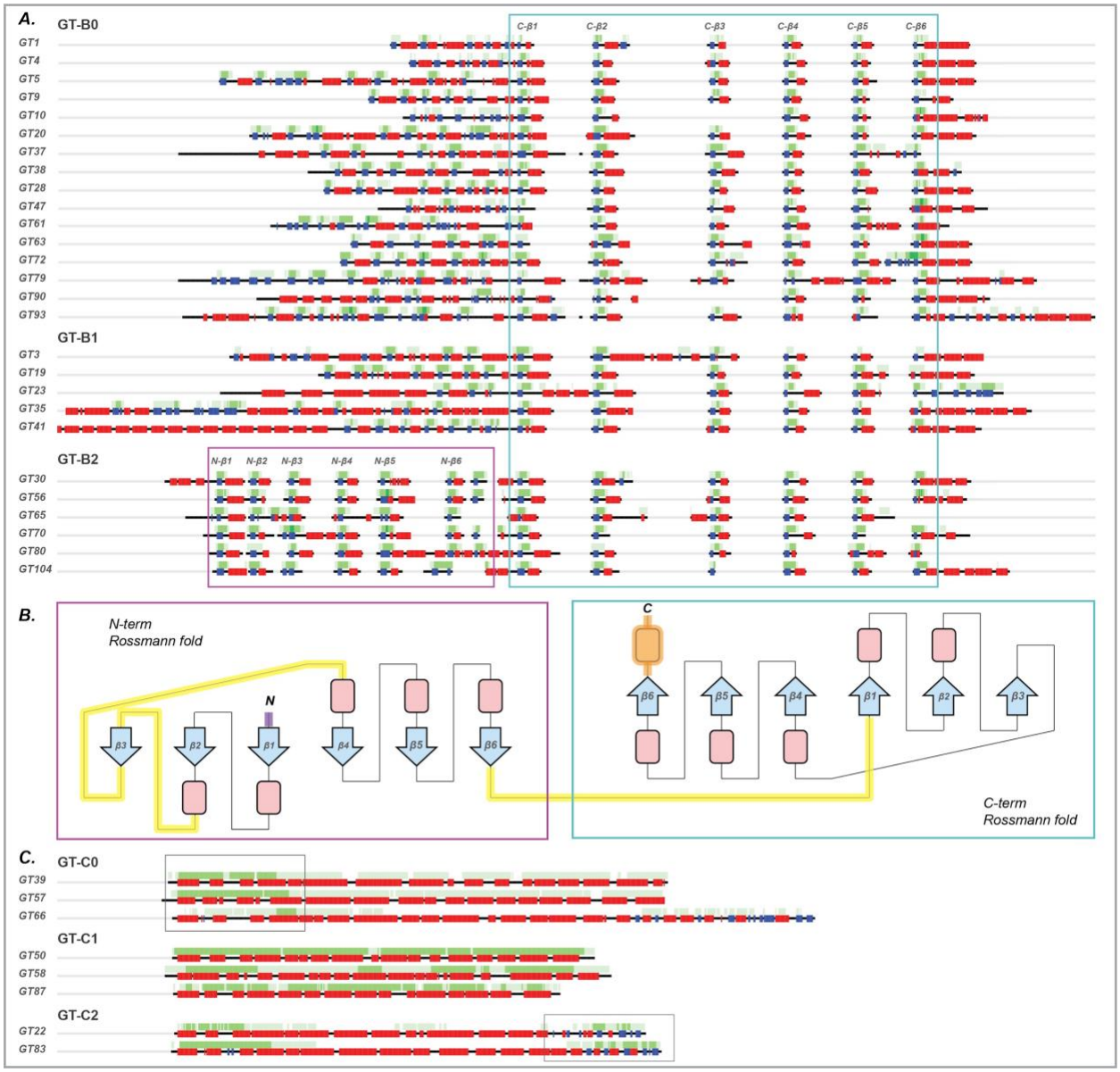
While it has been especially challenging to generate a GT-B fold-wide sequence alignment due to the lack of sequence conservation, in order to understand the patterns obtained from our CNN model, we generated family level alignments for each of the GT-B families. We then calculated a consensus ss and average layer two CAM map (Figure 3.4A) for each family. All of these families reflect the typical two β/α/β Rossmann fold domains characteristic of the GT-B fold. The most consistent pattern picked up by the CNN-attention model is the C-terminal Rossmann fold. Features associated with its 6 beta sheets are always significant in distinguishing GT-B families as indicated by the CAM value maps (cyan box in Figure 3.4A) and the conservation of this C-terminal region extends beyond GT-B0 to GT-B1, GT-B2 and other ungrouped GT-B families as well. Further, mapping the CAM values to representative structures revealed that the C-terminal Rossmann fold orientation and structure is well conserved across GT-B families with occasional family specific

insertions in the loop regions (Figure 3.4B). Thus, our study supports the C-terminal Rossmann domain as the common structural feature of GT-B fold families.

Upstream of the C-terminal Rossmann fold, CAM values are also higher in the ss of the N-terminal Rossmann fold region, likely indicating its importance in distinguishing the GT-B fold with 2 Rossmann folds versus the GT-A fold that has only a single Rossmann fold domain. However, these CAM value patterns are not consistent across families. Most families have a different number and order of beta sheets suggesting variability in the N-terminal domain, likely reflecting its function of binding different types of acceptor substrates, as shown in previous studies [1,31]. This variability is especially prominent in GT-B1 where families accommodate additional secondary structures in the N-terminal (for example, tetratricopeptide repeats in GT41 and coiled coils in GT23) (Figure 3.4A). Conversely, all families within the GT-B2 cluster are found to conserve a minimum of 6 beta sheets and 5 alpha helices in the N-terminal Rossmann fold, as indicated by the CAM values (magenta box in Figure 3.4A,B), highlighting the extension of the GT-B2 core to include both the N- and the C-terminal Rossmann fold domains. The functional implications of this extended core conservation in GT-B2 families is yet to be determined.

Figure 3.4: CAM maps for the different GT-B and GT-C fold clusters highlight their respective conserved cores. (A) Consensus ss (blue: beta sheets; red: helices; green: loops) and average CAM values (higher intensity of green corresponds to a higher CAM value) from layer 2 are shown for all families belonging to the 3 GT-B fold clusters. These average values were generated from sequence alignments within each family. High CAM values within the cyan box point to the C-terminal Rossmann fold conserved across all GT-B fold members and the magenta box points to the N-terminal Rossmann fold conserved in GT-B2. (B) A topological representation of the conserved features of GT-B. The conserved C-terminal Rossmann-like fold region is shown in the cyan box. The N-terminal Rossmann fold, which is most conserved in members of GT-B2 cluster is shown in magenta box. Conserved beta sheets are shown as blue arrows with labels and alpha helices are shown as red boxes. Loop regions that have the most variability across families are

indicated by yellow lines. Purple N-terminal loop and orange C-terminal helix indicate the presence of variable ss preceding the N-terminal and following the C-terminal Rossmann fold, respectively. (C) Consensus ss and average CAM values from layer 3 for GT-C families from clusters GT-C0, GT-C1 and GT-C2. Boxes indicate regions with higher average layer 3 CAM values for GT-C0 and GT-C2 in the N-terminal and the C-terminal regions, respectively. For GT-C1, layer 3 CAM is high throughout the full length of the sequences.



GT-Cs present an entirely different fold composed of 8 to 13 hydrophobic integral transmembrane helices with the active site and catalytic residues in long loop regions that makes them stand out from other GT fold enzymes [23]. The layer 3 CAM values of our CNN model responsible for capturing long range features recognized this trend and presented a consistent pattern for distinguishing the GT-C fold families (Figure 3.4C). In contrast to GT-A and GT-B, no other trends in CAM values from layer 1 and 2 exist for the GT-Cs suggesting that the layer 3 features were the most important and sufficient in distinguishing sequences adopting a GT-C fold.

For the first time, we define 3 major clusters within the GT-C fold families. The GT-C0 cluster families have higher layer 3 average CAM values towards the N-terminal helices, which most likely is enough to separate them from GT-C1 and GT-C2. In contrast, GT-C1 includes families that are generally shorter in sequence length with little to no contiguous loop segments. The layer 3 average CAM values for these families stay high throughout the length of the sequences. Moreover, all members of the 3 families in GT-C1 are mannosyltransferases (PigM family GT50, Alg3 family GT58 and bacterial pimE of GT87), with PigM and Alg3 also known to share detectable sequence similarity [32]. Finally, GT-C2 members are distinct from other GT-C clusters in the C-terminal region where they share a distinct region with an α/β/α arrangement. This region has been identified as a periplasmic domain in a bacterial aminoarabinose transferase ArnT of the GT83 family [33], which could interact with the donor substrate. Outside of the GT-C2 cluster, only GT66 family members (oligosaccharyltransferases) in GT-C0 have a similar extended C-terminal domain (Figure 3.4C).

3.2.5 Identifying families with novel GT folds using the convolutional autoencoder model

While our CNN model could successfully distinguish the known GT fold types, there are 30 CAZy GT families (GT-u) that could not be assigned to a known fold in the standard CNN-attention workflow. We wanted to extend our model to analyze and predict the fold types for these unknown families. To this end, we extended our existing CNN model to build an autoencoder that allows calculation of a reconstruction error (RE) for any given sequence (Methods). Sequences similar to the ones used in training (i.e., one of the known folds) would have a low RE whereas novel fold sequences would have a large RE. Figure 3.5A shows the distribution of RE for sequences with known (GT-A, -B, -C and -lyso in gray) and unknown folds (in red). There is a clear separation with the unknown fold sequences having a higher RE.

To statistically evaluate which GT-u families have a significantly higher RE than the known folds, we first fitted an extreme value distribution to our training data (RE from sequences with known folds) to calculate 95% and 99% confidence intervals (CI). We then compare a median RE value (mRE) for each GT-u family against these CI to make fold predictions. However, we note that the peak for unknown RE distribution falls within the 95% CI (below 0.107, Figure 3.5A) suggesting that a majority of GT-u sequences adopt one of the known folds. For families that are predicted to adopt a known fold, we also want to identify their closest known fold type. To achieve this, we further built 9 autoencoder models for each of the 2 GT-A, 3 GT-B, 3 GT-C and 1 GT-lyso clusters and calculated RE. Due to the low number of sequences in these cluster specific models, instead of fitting an extreme value distribution, we used a fold assignment score (FAS, one for each sub cluster totaling 9 FAS scorers for each GT-u family), to evaluate the best match for each of the GT-u families (Methods). Finally, based on the mRE and FAS scores for each GT-u family, we

predict their fold status with varying degrees of confidence using the criteria described in Methods (Figure 3.5, Figure 3.6, Table 3.3). In short, GT-u families with mRE higher than a threshold of 0.127 are considered to adopt novel folds. Families with mRE below 0.127 and at least one positive FAS score are assigned to the known fold with the highest FAS score. Families with mRE below 0.127 with all negative FAS scores are designated as variant fold types.

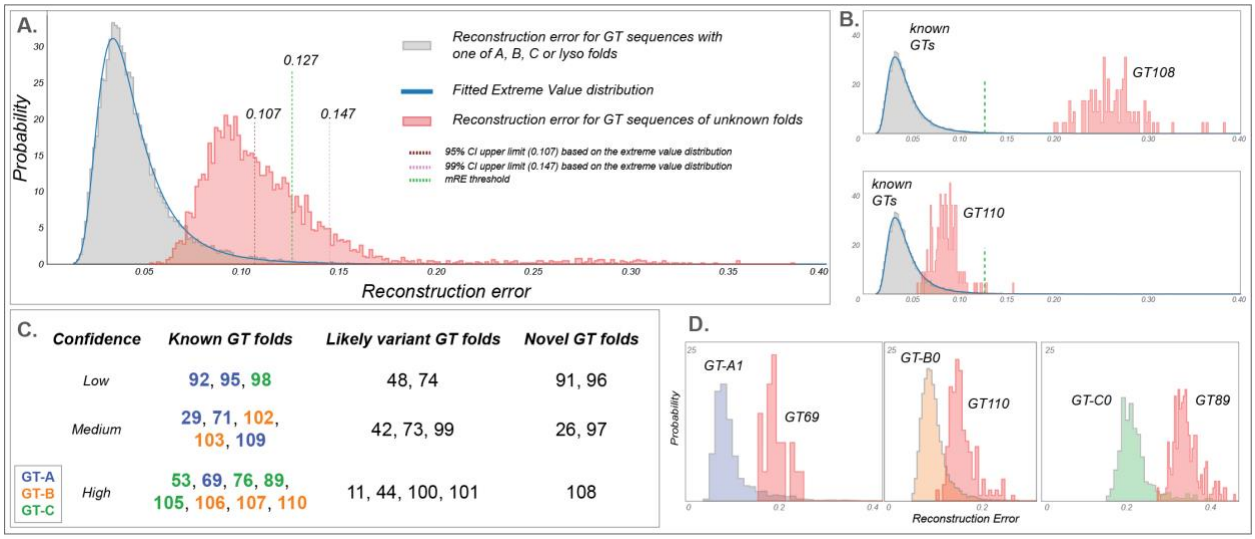


Figure 3.5: Fold prediction in GT-u families. Reconstruction error (RE) for the known GT fold families are shown in gray and GT-u in red. A) An extreme value distribution is fitted into RE for known fold to calculate a 95% and a 99% CI (upper limits in brown and pink dotted lines, respectively). The midpoint threshold RE at 0.127 (Methods) is marked with green dotted lines. B) As examples, RE for unknown fold families GT108 (upper panel) and GT110 (lower panel) are shown where RE for GT108 is very high and thus predicted to have a novel fold with high confidence. In contrast, the RE for GT110 is low and close to known fold families, thus predicted to have a known GT fold. C) Chart showing the fold prediction results for 29 GT families with unknown folds. Family names are placed based on their likelihood of adopting a novel fold and the confidence in that evaluation and are colored based on their assigned fold types. D) RE for three GT-u families predicted to have a known GT fold is plotted alongside RE for their predicted fold cluster with highest fold assignment score (FAS). Left: GT69 versus GT-A1; Middle: GT110 with GT-B0; Right: GT89 with GT-C0. mRE and FAS scores for all the GT-u families are provided in Table 3.3.

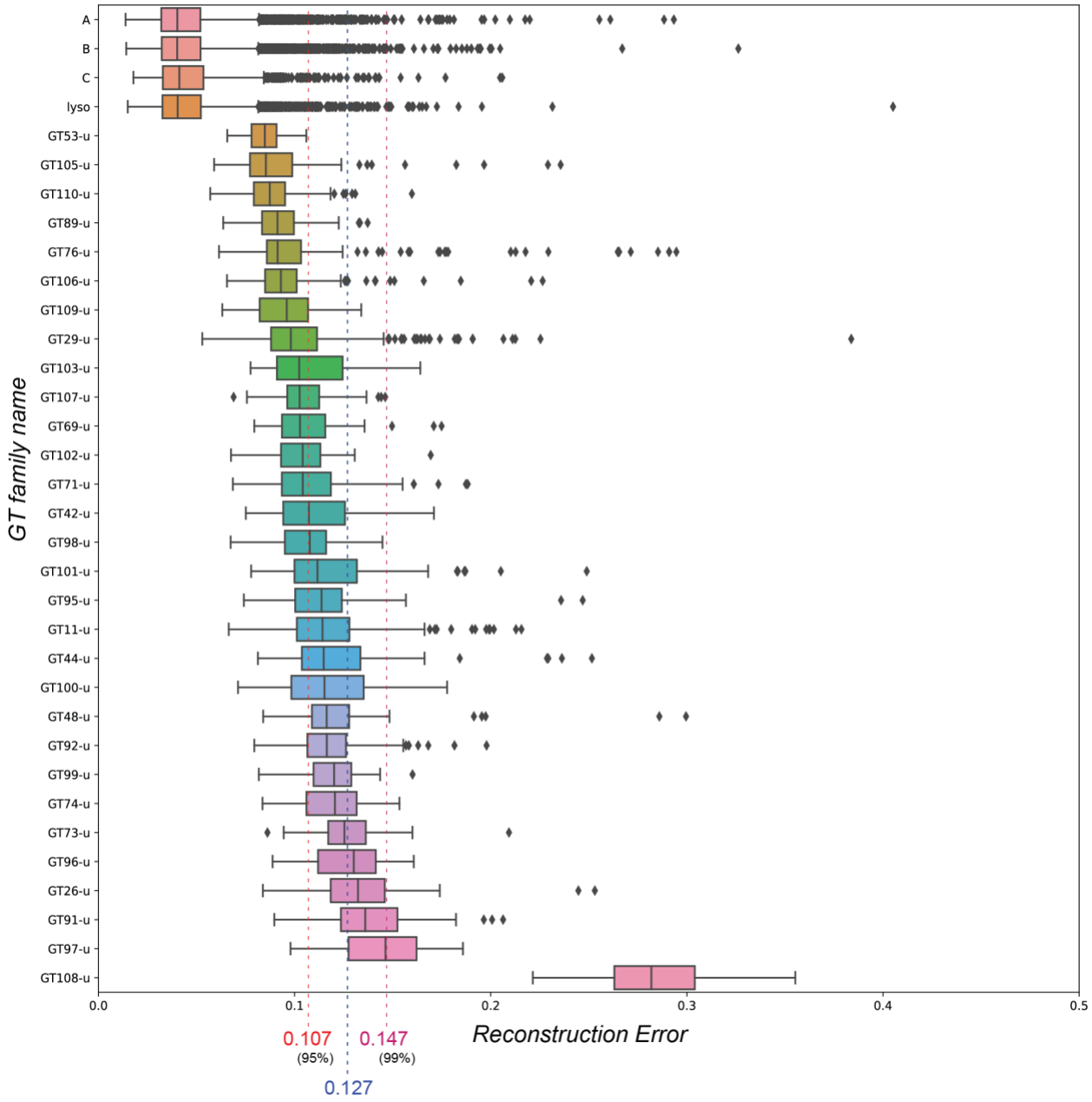


Figure 3.6: Boxplot showing the RE for each of the 4 known GT folds and all the GT-u families. Red line at 0.104 indicates the 95%CI upper bound based on the extreme value distribution of the training (known GT folds) sequences. Magenta line at 0.147 indicates the 99% CI upper bound. Blue line at 0.127 indicates the mid-point. GT-u families with median RE above this value are predicted to have novel GT folds with increasing confidence.

Table 3.3: Fold prediction results for the GT-u families. RE against all known GTs are shown. RE < 0.107 (Upper 95% CI) suggesting families most likely to adopt a known GT fold are highlighted in yellow and RE > 0.127 (Closer to or more than upper 99% CI) for families most likely to adopt a novel fold are highlighted in red. GT-u families are most likely to adopt the GT fold with the highest positive FAS. Families predicted to adopt a variant or a novel fold have negative FAS scores for all the clusters. The highest FAS scores for families predicted to adopt known folds are colored (>0.1 in green, 0-0.1 in orange). The predicted fold and confidence are indicated in the last two columns. Confidence is evaluated based on both the RE and FAS scores.

Family	Reconstruction Error RE (against All known GTs)	Fold Assignment Scores (FAS)									Max FAS score	Prediction Confidence	Predicted fold
		GT-A0	GT-A1	GT-B0	GT-B1	GT-B2	GT-C0	GT-C1	GT-C2	GT-lyso			
GT53-u	0.084	-0.991	-1.779	-4.399	-0.395	-0.566	0.218	0.413	0.013	-1.072	0.413	High	GT-C
GT110-u	0.087	0.059	0.257	0.261	-0.277	0.024	-2.764	-2.265	-0.653	-0.403	0.261	High	GT-B
GT89-u	0.091	-0.225	-0.465	-1.296	-0.001	-0.166	0.231	0.147	0.095	-0.522	0.231	High	GT-C
GT69-u	0.102	-0.040	0.177	-0.417	-0.398	-0.173	-2.902	-2.285	-0.695	-0.441	0.177	High	GT-A
GT76-u	0.091	-2.433	-2.774	-6.257	-0.603	-0.956	-0.374	0.151	-0.292	-1.327	0.151	High	GT-C
GT106-u	0.093	-0.123	0.058	0.130	-0.105	-0.052	-1.819	-1.470	-0.410	-0.415	0.130	High	GT-B
GT107-u	0.102	-0.227	-0.005	0.123	-0.062	0.057	-1.903	-1.344	-0.343	-0.363	0.123	High	GT-B
GT105-u	0.085	-2.243	-2.571	-5.841	-0.638	-0.821	-0.307	0.115	-0.294	-1.574	0.115	High	GT-C
GT102-u	0.104	-0.371	-0.196	0.094	-0.200	-0.027	-3.137	-2.288	-0.716	-0.872	0.094	Medium	GT-B
GT109-u	0.096	0.002	0.090	-0.043	-0.054	-0.076	-1.967	-1.645	-0.503	-0.524	0.090	Medium	GT-A
GT71-u	0.104	-0.090	0.086	-0.345	-0.119	-0.072	-1.534	-1.318	-0.308	-0.299	0.086	Medium	GT-A
GT29-u	0.098	-0.228	0.074	-0.115	-0.457	-0.200	-3.300	-2.393	-0.789	-0.580	0.074	Medium	GT-A
GT103-u	0.102	-0.132	-0.001	-0.133	-0.050	0.012	-1.682	-1.393	-0.300	-0.447	0.012	Medium	GT-B
GT98-u	0.107	-0.468	-0.719	-2.021	-0.130	-0.228	0.186	0.120	0.105	-0.801	0.186	Low	GT-C
GT92-u	0.116	-0.143	0.098	-1.105	-0.732	-0.252	-3.609	-2.659	-0.805	-0.512	0.098	Low	GT-A
GT95-u	0.113	-0.217	0.053	-0.569	-0.530	-0.308	-3.461	-2.664	-0.771	-0.584	0.053	Low	GT-A
GT74-u	0.120	-0.199	-0.042	-0.913	-0.609	-0.324	-3.785	-3.156	-0.872	-0.787	-0.042	Low	Variant
GT48-u	0.116	-0.285	-0.432	-1.054	-0.079	-0.219	-0.815	-0.805	-0.155	-0.618	-0.079	Low	Variant
GT73-u	0.125	-0.434	-0.148	-0.597	-0.632	-0.400	-4.530	-3.350	-1.028	-1.000	-0.148	Medium	Variant
GT42-u	0.107	-0.306	-0.170	-0.288	-0.349	-0.276	-3.295	-2.563	-0.818	-0.779	-0.170	Medium	Variant
GT99-u	0.120	-0.530	-0.341	-0.197	-0.339	-0.215	-3.569	-2.596	-0.891	-1.026	-0.197	Medium	Variant
GT101-u	0.111	-0.443	-0.234	-0.242	-0.313	-0.207	-3.407	-2.572	-0.761	-0.957	-0.207	High	Variant
GT44-u	0.114	-0.571	-0.730	-0.897	-0.250	-0.460	-2.647	-1.975	-0.684	-1.155	-0.250	High	Variant
GT100-u	0.115	-0.557	-0.312	-0.339	-0.558	-0.273	-4.503	-3.298	-1.001	-1.129	-0.273	High	Variant
GT11-u	0.114	-0.516	-0.378	-0.434	-0.628	-0.385	-4.813	-3.494	-1.135	-1.272	-0.378	High	Variant
GT96-u	0.130	-0.271	-0.004	-0.750	-0.275	-0.138	-2.133	-1.601	-0.451	-0.435	-0.004	Low	Novel
GT91-u	0.136	-0.294	-0.066	-1.731	-0.655	-0.253	-2.991	-1.993	-0.587	-0.300	-0.066	Low	Novel
GT26-u	0.132	-0.620	-0.719	-0.300	-0.625	-0.410	-4.231	-3.330	-1.111	-1.450	-0.300	Medium	Novel
GT97-u	0.146	-0.986	-0.698	-0.369	-0.790	-0.426	-6.031	-4.188	-1.513	-1.598	-0.369	Medium	Novel
GT108-u	0.281	-1.609	-1.238	-5.935	-3.291	-1.526	-9.169	-7.093	-2.221	-2.372	-1.238	High	Novel

Five families have very high mRE (larger than 0.127), and are predicted to adopt novel GT folds (Figure 3.5C, Figure 3.6, Table 3.3). The dual-activity mannosyltransferase/ phosphorylases of family GT108 have the highest mRE (0.281) and have indeed been shown to adopt a unique five-bladed β-propeller fold that is completely different from the four GT folds [34]. Another family predicted to have a novel fold, GT26, has a single representative crystal structure for a membrane associated GT TagA, from a bacteria *T. italicus*, which also adopts a novel fold [35]. Here, we predict three additional families, the fungal β-1,2-mannosyltransferases Bmt/Wry (GT91), plant peptidyl serine α-galactosyltransferases Sgt (GT96) and bacterial α-2,6-sialyltransferases (GT97), that likely adopt novel GT folds as well.

Using the mRE and the FAS scores, we assign 6 GT-u families as having a GT-A type fold (Table 3.3). Out of these, the GT29 mammalian sialyltransferases have been shown to adopt a modified GT-A fold with different orientations of the beta sheets in the Rossmann fold while conserving the overall Rossmann fold scaffold and specific sialyl motifs [36]. The human glycolipid glycosylphosphatidylinositol β-1,4-N-acetylgalactosaminyltransferase PGAP4 (GT109) has also been predicted to adopt a GT-A fold with transmembrane domain insertions [37]. In line with this study, GT109 family is predicted to have a GT-A fold with medium confidence. Our analysis further adds the α-1,3-mannosyltransferases (GT69) to the GT-A fold families with high confidence. Additionally, we predict that the α-mannosyltransferases Mnn (GT71), the plant GalS galactan synthases and other members of the GT92 family and members of the GT95 family (hydroxyproline β-L-arabinofuranosyltransferase HPATs) also adopt folds that are similar to the GT-A type fold.

We also identify 5 GT-u families that most likely adopt the GT-B fold (Figure 3.5C, Table 3.3). This includes the bacterial α-1,3-L-rhamnosyltransferase (GT102), the bacterial O-antigen-polysaccharide β-1,4-N-acetylglucosaminyltransferases (GT103), the GT106 family of plant

rhamnogalacturonan I 4-α-rhamnosyltransferases, the GT107 family of KDO transferases and the β-1,4-xylosyltransferases (Rxylt1/TMEM5) of the GT110 family. Similarly, 5 families have the highest positive FAS score against GT-C clusters and are predicted to adopt the GT-C fold (Table 3.3). In agreement with our predictions, cryo-EM based structures of representative bacterial Embs of the GT53 family have revealed a GT-C fold [38] and recent structural predictions on the human TMTCs of family GT105 have suggested that they adopt a GT-C fold [39]. In addition to these 2 families, we predict that PigVs (GT76), bacterial arabinofuranosyltransferases AftBs (GT89), and dpy-19 mannosyltransferases (GT98) also adopt a GT-C fold. In addition, all of these 5 families utilize lipid-linked sugar donor substrates similar to other known GT-C fold enzymes [23,38,40].

The remaining 9 families have a negative FAS score for all the GT-A, -B, -C and -lyso clusters and thus are not assigned a specific fold type. However, since they have an mRE below 0.127, these families are predicted to adopt a variant of the existing fold types rather than a novel fold type. Among them, families like the bimodular dual β-glucosyltransferases of GT101 and the multimodular bacterial β-KDO transferases of the GT99 family have representative crystal structures [41,42] revealing that they adopt unique folds consisting of the Rossmann-fold scaffold with the latter forming a variant of the GT-B fold type. The bacterial toxin glucosyltransferases of the GT44 family have also been shown to adopt a slightly modified structure highly similar to a GT-A fold [43,44]. Bacterial Csts from the GT42 family also have been shown to adopt a variant of the GT-A fold type that is highly similar to the GT29 sialyltransferases with both families conserving the sialyl motifs [45]. Yet, while GT29 scores higher against the GT-A1 cluster, GT42 does not and is correctly classified as a variant fold type suggesting key differences in other regions of the GT-A core. Here, we add variant fold predictions for the GT11 (fucosyltransferases), GT48 (glucan synthases), GT73 (bacterial KDO transferases), GT74 (includes few α-1,2-L-

fucosyltransferases) and GT100 (bacterial sialyltransferases) families. Additional details of these predictions are provided in the methods and the results summarized in Table 3.3.

3.3 Discussion

It has been well established that the structural folds of GTs, much like in many other large protein families, are far more conserved than primary sequence [5,6]. The functional diversification of GTs through extensive sequence variation and insertion of variable loops and disordered regions presents a major challenge for broad sequence or structural classification using alignment-based approaches. This inability to create a larger framework of GT structural classification has impeded understanding of the evolutionary relationships among the GTs during the expansion of glycan diversity in all domains of life [46,47].

Although GTs primarily adopt 3 major fold types, each have their own distinct features. GT-A and GT-B enzymes employ single or paired Rossmann folds, respectively, for donor and acceptor binding during catalysis. Less is known about GT-C fold enzymes that employ distinct features composed of multiple transmembrane helical domains. Identifying and distinguishing the GT-A and GT-B folds in the absence of solved structures is quite challenging and non-trivial, more so when the starting fold type is not known as is the case for multiple GT-u families. To overcome these challenges and produce reliable fold predictions, we use a CNN-attention based deep learning model that implements a completely alignment free approach relying simply on secondary structure patterns to classify all GT families into either the known fold types or predict novel fold types. As far as we know, this is the first attempt at utilizing this simple coarse-grained, dependable form of input for analyzing such a large group of enzymes using deep learning. We successfully built a model that classified known folds and families with a 96% and 77% accuracy, respectively. In

addition, we focus the design of our model on interpretability, where each layer generates outputs in the form of CAM maps (Block 1), features for UMAP visualization (Block 2) and reconstruction errors (Block 3) for biological interpretation and understanding of the model.

By mapping the features learned by the model using UMAP, we identified clusters of families within the major fold type that were found to share novel and distinct regions of similarity, as revealed by their CAM maps. Each of the two clusters within the GT-A fold include phylogenetically diverse families [3] yet each shares a unique set of secondary structural features within the hypervariable regions that distinguish the clusters and likely contribute to substrate recognition. Because such features shared by evolutionarily divergent families are difficult to detect through traditional phylogenetic approaches, the CNN model provides a valuable tool for inferring such shared structural mechanisms.

In the GT-B fold families, where previous attempts of sequence and structural alignments have proven difficult, we identify a central GT-B0 cluster which points to a limited conserved core in the C-terminal Rossmann fold, with insertions in the loop regions. We show that this conservation extends across the large and diverse GT-B0 cluster to other GT-B clusters as well (Figure 3.4A). In the smaller GT-B2 cluster, CAM maps point to additional structural similarities in the N-terminal Rossmann domain within this cluster. By virtue of these shared features, we present an alignment of predicted secondary structures across GT-B fold families providing a comparative basis for cross-cluster analyses (Figure 3.4). Similarly, we identify a subset of GT-C fold families (GT-C1) consisting entirely of mannosyltransferases where the CAM features extend throughout the entire length of the sequences (Figure 3.4C).

More importantly, we deploy an autoencoder model using the features from the CNN-attention model to make reliable predictions for GT-u families that are most likely to adopt a novel GT fold.

The 16 GT-u families found to adopt known folds (Figure 3.5C) provide a comparative basis for understanding their functions and associations. On the other hand, five GT-u families are predicted to adopt novel folds. Three out of the five families (GT91, GT96 and GT97) do not have a representative crystal structure. Coincidentally, each of these 3 families are found in select taxonomic groups (fungi, plants and bacteria, respectively) and have different functions. Moreover, out of the 12 families that are predicted to adopt variant folds, only 4 (GT42, 44, 99 and 101) have representative crystal structures, all of which point to novel structural adaptations and variations [41–45]. Our novel predictions for other families that lack representative structures provide informed targets for focused structural studies that could reveal divergent GT folds with novel mechanisms and modes of regulation to expand the GT fold space.

We use a combination of metrics (RE, FAS, number of sequences) to assign confidence levels for our predictions providing researchers with meaningful metrics of reliability for guiding future efforts. These predictions are based on the family level and utilize ss predictions on a large number of sequences from each family, thus providing robust results. However, interpretations for families such as GT78 (A fold), GT18 (B fold), GT103 or GT97 (novel fold) with very few unique sequences should be done with caution.

Finally, our approach employs a simple training dataset that is straightforward to prepare and is surprisingly adaptable for understanding fold diversity in any large protein family. Contrary to most “black box” deep learning models, the output of this workflow is a highly interpretable deep learning model that generates accurate fold predictions with quantitative outputs that provide meaningful biological insights without the need for primary sequence or structural alignment. Thus, the approach adds a novel and powerful new tool to the repertoire for computational and evolutionary analyses of large protein families.

3.4 Methods

3.4.1 Data collection and preprocessing

Sequence retrieval and ss prediction

We retrieved GenBank [48] IDs for sequences belonging to the 110 GT families from the CAZy database (accessed 04/05/2020). Sequences for these IDs were then collected from the NCBI GenBank database. These sequences were first filtered to sequence similarities of 60-95% depending on the number of sequences for each family to balance the number of sequences across families and to avoid overfitting. We predicted the secondary structures of our filtered dataset of 44,620 sequences using NetSurfP2.0 [19]. NetSurfP predicts both 3-state and 8-state secondary structures based on DSSP definitions [49]. Here, we only use the 3-state predictions as input features since these are reported with higher accuracy. Additionally, we make our predictions on the family level, that accounts for persistent ss predictions in multiple closely related sequences from the same family and makes our method robust to small inconsistencies in ss prediction for individual sequences.

Sequence length filtering

To allow batch training for neural networks, these sequences were padded to a consistent length of 798. We set this threshold by modeling the distribution of GT-A, B, C and lyso sequence lengths to a Gaussian distribution and setting our maximum length cutoff at $\mu+3\sigma$. However, for a subset of sequences that extend beyond 798 amino acids, we eliminated sequences flanking the GT-domain through domain mapping via Batch CD-search [50]. Sequences with multiple GT domains

were labeled separately and treated as different sequences. Sequences lacking an annotated GT domain or with an annotated GT domain longer than 798 amino acids were removed. Our final padded dataset contained 12,306 GT-A, 20,397 GT-B, 1518 GT-C, 5482 GT-lyso, and 4258 GT-unknowns where each sequence is represented by a 798X3 matrix of ss predictions and padding.

Data augmentation for balancing datasets across families

Skewed datasets can hinder the convergence of neural networks and negatively impact generalization. To mitigate this issue, we balanced our training dataset using data augmentation. Our data augmentation procedure randomly changes 5% of secondary structure positions to coil/ loop, excluding the padding region. This procedure can sometimes produce no changes, such as if only coil/ loop positions are randomly chosen. In these cases, the procedure is repeated until at least one change is made. For our balanced training set, each of the GT-A, GT-B and GT-C fold families were upsampled to 2000 sequences, whereas the single GT51 family of GT-lyso fold was downsampled to 5000 sequences. For two families with a very large number of sequences, GT2 and GT4, we selected 2000 divergent representatives from each family. This balanced training set was generated once and reused for parameter optimization unless otherwise indicated.

3.4.2 CNN model for fold and family classification

The model architecture involves a novel attention aided deep CNN model with six blocks. The first three blocks (Block 1, Figure 3.1) sequentially use a one dimensional convolutional layer, followed by a pooling layer and a batch normalization layer. This feeds into an attention [21] layer that performs a refinement of the generated feature maps. The convolution kernel sizes were set to 3, 7 and 15 with kernel numbers set to 256, 512 and 512 respectively for the first three blocks. Since

pooling operations lead to a loss of spatial information for the feature maps, such an operation is not applied on any of the first three layers, thus enabling mapping of the attention maps back to the sequence for interpretation.

The feature maps from these three layers are down-sampled using a global average pooling layer and then passed through three additional blocks before making the final prediction. In contrast to the first three layers that carry spatial information, these three layers (Block 2, Figure 3.1) use global max pooling operations that compute a single maximum value for each of the input channels, thus providing a single linearly independent representation for each sequence, regardless of sequence length. These representations can be transformed into high-dimensional vectors which can then be used by downstream clustering algorithms.

For the multitasking of fold and family classification, two separate fully connected layers were added with dropout. The model was trained on a single NVIDIA RTX 2080Ti graphic card for 6 hours. Dropout rate was set to 0.5 during training. Adam optimizers with a learning rate of 1e-4 and weight decay with a rate of 1e-5 were deployed during training.

3.4.3 Autoencoder framework for identification of novel fold GT-u families

We adopted a recent advancement in machine learning field named open set recognition [22] to extend the trained classifier's ability to distinguish an unseen pattern of ss from the seen dataset of known GT folds. In application, this framework is targeted to real-world scenarios where new classes (unknown classes), unseen during training, appear in the testing phase and requires the classifier to not only accurately classify seen classes but also effectively deal with unseen classes in testing [22]. This translates well to our problem of distinguishing GT families that most likely adopt a previously unseen fold which is considerably different from the GT-A, B, C, and lyso folds

that the model is trained on, while efficiently recognizing families that could adopt one of these known folds. We propose a CNN based autoencoder framework to accomplish this task, which is capable of reconstructing the known GT folds that it has learned on but unable to do so if a given sequence is quite different, resulting in a high reconstruction error.

The autoencoder (Block 3, Figure 3.1) comprises two parts: an encoder and a decoder. The encoder reused Block 1 of the CNN model trained on GT-A, B, C and lyso as a general feature extractor. Then, a mirror structure of the CNN model that includes multiple deconvolution operations and instance normalization is connected to the encoder to generate reconstruction of the inputs. Similarity between the seen and unseen classes is measured by calculating a reconstruction error (RE) of the input samples (Figure 3.5A). A modified loss function was proposed to calculate RE in order to omit the effects of padding regions in the reconstruction of sequences as follows:

$$Masked\ MSE\ = \frac{1}{n-2p} \sum_{1+p}^{n-p} (Y - \hat{Y})^2 \quad (1)$$

where p is the padding length at both ends of the sequence, n is the sequence length, $Y$is secondary structure input, $\hat{Y}$is the predicted secondary structure output.

In addition to this main autoencoder model, 9 additional autoencoder models were built with the same architecture but trained separately on the 9 clusters of GT folds: 2 GT-A, 3 GT-B, 3 GT-C and 1 GT-lyso clusters. RE against each of these clusters were used to derive a fold assignment score (FAS) that was used as a measure to indicate which known fold a given GT family would adopt, if it was predicted to adopt a known fold. The FAS score was calculated using the following equation:

$$FAS_{ab} = \frac{OOC_b - RE_a}{OOF_b - RE_b} - thres \quad (2)$$

where FAS_{ab} is the fold assignment score for GT-u family a against cluster b, RE_a is the median reconstruction error for sequences in family a, RE_b is the median RE for sequences in cluster b, OOC_b is the average RE for sequences with the same fold as sequences in cluster b but are not grouped in cluster b (called out of cluster (OOC)), OOF_b is the average RE for sequences from a different fold than sequences in cluster b (called out of fold (OOF)) and $thres$ is a threshold score for fold prediction. For clusters with a highly skewed OOC and OOF distribution ($OOC_{A0}, OOC_{A1}, OOC_{B0}, OOC_{B1}, OOF_{B0}, OOF_{C0}, OOF_{C1}$), median RE values were used instead of averages. Since GT-lyso had only one cluster, 20% of sequences were left out of training, unseen by the model and used to calculate OOC_{lyso}. The threshold $thres$ was set to 0.08 based on the RE distributions to account for the differences in $RE_{clusters}$ across different clusters.

3.4.4 Model interpretation

Structural mapping of layer-wise activation maps

To fully understand how the CNN-Attention model classifies GT fold types, we analyze feature maps generated from Blocks 1 and 2 using two different methods:

1) For all three layers of Block 1, we rely on making weakly supervised class specific localization through a label guided method named class-specific activation mapping using Grad-CAM [24] (CAM) that uses gradient descent to generate feature maps that target specific families. The attention layers inserted in Block 1 further enhance these activation values. These Grad-CAM results were used to generate activation maps that conserve spatial information and can be mapped

back into the sequence to identify the most contributing ss and sequence regions for fold classification and thus represent the core conserved features.

2) For the three layers of Block 2, we generate saliency maps that highlight activations by extracting the feature map values. These maps do not conserve spatial information but are used to generate representation vectors that are then subjected to dimensionality reduction using UMAP [26] to generate manifolds for visualization and clustering of the known GT fold types. To identify the major clusters within GT fold types, we clustered the 2D UMAP projections using the GMM algorithm [29]. UMAP was performed with multiple sets of parameters to find families that most consistently grouped together. When implementing the GMM algorithm, an appropriate cutoff for the GMM score was selected independently for each fold type in order to generate clusters robust to changes in parameters of UMAP.

Evaluation of the reconstruction error to identify novel fold type families

Since the RE for most training sequences would be very low, the RE distribution for the training data from the main autoencoder was first fitted to an extreme value distribution using the scipy [51] package. This was then used to evaluate a 95% CI and a 99% CI. Median RE (mRE) calculated for each GT-u family was then compared to these two CI limits to statistically evaluate their likelihood of adopting a novel fold. The value 0.127 marks the midpoint for the interval between the upper limits of the 95% and the 99% CI (0.107 and 0.147, respectively) and was used as a threshold for predicting GT-u families that adopt a novel fold (higher mRE than 0.127) or a variant of the known folds (mRE lower than 0.127). Further evaluation of the prediction was done using the FAS scores as follows:

1) For families with mRE higher than 0.127, FAS scores were always negative with increase in confidence with lowering FAS scores (If the highest FAS scores were between 0 and -0.1 low confidence; -0.1 to -0.5 medium confidence and below -0.5 high confidence).

2) For families with mRE lower than 0.107 (95% CI), FAS scores were always positive, and the GT-u family was assigned to the fold with highest FAS score, more confidence with higher FAS scores (FAS score higher than 0.1 high confidence, between 0-0.1 medium confidence)(For example, GT53-u has the highest FAS score of 0.413 for the GT-C1 cluster so it is assigned a GT-C fold with high confidence, Table 3.3).

3) For families with mRE between 0.107 and 0.127, if FAS scores were positive, assign them to the fold with highest FAS scores with low confidence. If the FAS scores were all negative, assign them as a variant of known folds.

Bibliography

[1] Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* 2008, 77, 521–555.

[2] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucl. Acids Res.* 2014, 42, D490–D495.

[3] Taujale, R., Venkat, A., Huang, L.-C., Zhou, Z., et al., Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases. *eLife* 2020, 9, e54532.

[4] Chothia, C., Lesk, A.M., The relation between the divergence of sequence and structure in proteins. *EMBO J* 1986, 5, 823–826.

[5] Sousounis, K., Haney, C.E., Cao, J., Sunchu, B., Tsonis, P.A., Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Hum Genomics* 2012, 6, 10.

[6] Bajaj, M., Blundell, T., Evolution and the tertiary structure of proteins. *Annu Rev Biophys Bioeng* 1984, 13, 453–492.

[7] Breton, C., Fournel-Gigleux, S., Palcic, M.M., Recent structures, evolution and mechanisms of glycosyltransferases. *Current Opinion in Structural Biology* 2012, 22, 540–549.

[8] Moremen, K.W., Haltiwanger, R.S., Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat Chem Biol* 2019, 15, 853–864.

[9] Shi, Q., Chen, W., Huang, S., Wang, Y., Xue, Z., Deep learning for mining protein data. *Briefings in Bioinformatics* 2021, 22, 194–218.

[10] Singh, A., Deep learning 3D structures. *Nature Methods* 2020, 17, 249–249.

[11] Gao, M., Zhou, H., Skolnick, J., DESTINI: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports* 2019, 9, 3514.

[12] Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., et al., Improved protein structure prediction using potentials from deep learning. *Nature* 2020, 577, 706–710.

[13] Yang, J., Anishchenko, I., Park, H., Peng, Z., et al., Improved protein structure prediction using predicted interresidue orientations. *PNAS* 2020, 117, 1496–1503.

[14] Jumper, J., Evans, R., Pritzel, A., Green, T., et al., High Accuracy Protein Structure Prediction Using Deep Learning 2020.

[15] Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., Moult, J., Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics* 2019, 87, 1011–1020.

[16] Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y., Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics* 2017, 33, 2842–2849.

[17] Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., et al., DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 2018, 34, 2605–2613.

[18] Cao, R., Freitas, C., Chan, L., Sun, M., et al., ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules* 2017, 22, 1732.

[19] Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K., et al., NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics* 2019, 87, 520–527.

[20] Kim, Y., in:, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Doha, Qatar 2014, pp. 1746–1751.

[21] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., CBAM: Convolutional Block Attention Module. *arXiv:1807.06521 [cs]* 2018.

[22] Geng, C., Huang, S., Chen, S., Recent Advances in Open Set Recognition: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* 2020, 1–1.

[23] Albuquerque-Wendt, A., Hütte, H.J., Buettner, F.F.R., Routier, F.H., Bakker, H., Membrane Topological Model of Glycosyltransferases of the GT-C Superfamily. *Int J Mol Sci* 2019, 20.

[24] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., et al., Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis* 2020, 128, 336–359.

[25] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., Learning Deep Features for Discriminative Localization. *arXiv:1512.04150 [cs]* 2015.

[26] McInnes, L., Healy, J., Melville, J., UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]* 2020.

[27] Zhang, Q., Zhu, S.-C., Visual Interpretability for Deep Learning: a Survey. *arXiv:1802.00614 [cs]* 2018.

[28] Oza, P., Patel, V.M., in:, IEEE Computer Society, 2019, pp. 2302–2311.

[29] Reynolds, D., in:, Li SZ, Jain A (Eds.), *Encyclopedia of Biometrics*, Springer US, Boston, MA 2009, pp. 659–663.

[30] Hurtado-Guerrero, R., Zusman, T., Pathak, S., Ibrahim, A.F.M., et al., Molecular mechanism of elongation factor 1A inhibition by a Legionella pneumophila glycosyltransferase. *Biochem J* 2010, 426, 281–292.

[31] Chang, A., Singh, S., Phillips, G.N., Thorson, J.S., Glycosyltransferase structural biology and its role in the design of catalysts for glycosylation. *Curr Opin Biotechnol* 2011, 22, 800–808.

[32] Oriol, R., Martinez-Duncker, I., Chantret, I., Mollicone, R., Codogno, P., Common Origin and Evolution of Glycosyltransferases Using Dol-P-monosaccharides as Donor Substrate. *Molecular Biology and Evolution* 2002, 19, 1451–1463.

[33] Petrou, V.I., Herrera, C.M., Schultz, K.M., Clarke, O.B., et al., Structures of aminoarabinose transferase ArnT suggest a molecular basis for lipid A glycosylation. *Science* 2016, 351, 608–612.

[34] Sernee, M.F., Ralton, J.E., Nero, T.L., Sobala, L.F., et al., A Family of Dual-Activity Glycosyltransferase-Phosphorylases Mediates Mannogen Turnover and Virulence in Leishmania Parasites. *Cell Host Microbe* 2019, 26, 385-399.e9.

[35] Kattke, M.D., Gosschalk, J.E., Martinez, O.E., Kumar, G., et al., Structure and mechanism of TagA, a novel membrane-associated glycosyltransferase that produces wall teichoic acids in pathogenic bacteria. *PLoS Pathog* 2019, 15.

[36] Meng, L., Forouhar, F., Thieker, D., Gao, Z., et al., Enzymatic Basis for N-Glycan Sialylation. *J Biol Chem* 2013, 288, 34680–34698.

[37] Hirata, T., Mishra, S.K., Nakamura, S., Saito, K., et al., Identification of a Golgi GPI-N-acetylgalactosamine transferase with tandem transmembrane regions in the catalytic domain. *Nat Commun* 2018, 9.

[38] Tan, Y.Z., Rodrigues, J., Keener, J.E., Zheng, R.B., et al., Cryo-EM structure of arabinosyltransferase EmbB from Mycobacterium smegmatis. *Nat Commun* 2020, 11, 3396.

[39] Eisenhaber, B., Sinha, S., Jadalanki, C.K., Shitov, V.A., et al., Conserved sequence motifs in human TMTC1, TMTC2, TMTC3, and TMTC4, new O-mannosyltransferases from the GT-C/PMT clan, are rationalized as ligand binding sites. *Biol Direct* 2021, 16.

[40] Larsen, I.S.B., Narimatsu, Y., Joshi, H.J., Siukstaite, L., et al., Discovery of an O-mannosylation pathway selectively serving cadherins and protocadherins. *Proc Natl Acad Sci U S A* 2017, 114, 11163–11168.

[41] Ovchinnikova, O.G., Mallette, E., Koizumi, A., Lowary, T.L., et al., Bacterial β-Kdo glycosyltransferases represent a new glycosyltransferase family (GT99). *Proc Natl Acad Sci U S A* 2016, 113, E3120-3129.

[42] Zhang, H., Zhu, F., Yang, T., Ding, L., et al., The highly conserved domain of unknown function 1792 has a distinct glycosyltransferase fold. *Nature Communications* 2014, 5, 4339.

[43] Pruitt, R.N., Chumbler, N.M., Rutherford, S.A., Farrow, M.A., et al., Structural Determinants of Clostridium difficile Toxin A Glucosyltransferase Activity. *J Biol Chem* 2012, 287, 8013–8020.

[44] Chen, P., Lam, K., Liu, Z., Mindlin, F.A., et al., Structure of the full-length Clostridium difficile toxin B. *Nature Structural & Molecular Biology* 2019, 26, 712–719.

[45] Chiu, C.P.C., Watts, A.G., Lairson, L.L., Gilbert, M., et al., Structural analysis of the sialyltransferase CstII from Campylobacter jejuni in complex with a substrate analog. *Nature Structural & Molecular Biology* 2004, 11, 163–170.

[46] Schmid, J., Heider, D., Wendel, N.J., Sperl, N., Sieber, V., Bacterial Glycosyltransferases: Challenges and Opportunities of a Highly Diverse Enzyme Class Toward Tailoring Natural Products. *Front. Microbiol.* 2016, 7.

[47] Breton, C., Šnajdrová, L., Jeanneau, C., Koča, J., Imberty, A., Structures and mechanisms of glycosyltransferases. *Glycobiology* 2006, 16, 29R-37R.

[48] Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., GenBank. *Nucleic Acids Res* 2016, 44, D67–D72.

[49] Kabsch, W., Sander, C., Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, 22, 2577–2637.

[50] Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., et al., CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 2011, 39, D225–D229.

[51] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., et al., SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* 2020, 17, 261–272.

CHAPTER 4

GTXPLORER: A PORTAL TO NAVIGATE AND VISUALIZE THE EVOLUTIONARY INFORMATION ENCODED IN FOLD A GLYCOSYLTRANSFERASES

Rahil Taujale, Saber Soleymani, Amitabh Priyadarshi, Aarya Venkat, Wayland Yeung, Krys Kochut and Natarajan Kannan. *To be submitted to Glycobiology*.

Abstract

Glycosyltransferases (GTs) play a central role in sustaining all forms of life through the biosynthesis of complex carbohydrates. Despite significant strides made in recent years to establish computational resources, databases and tools to understand the nature and role of carbohydrates and related glycoenzymes, a data analytics framework that connects the sequence-structure-function relationships to the evolution of GTs is currently lacking, which hinders characterization of understudied GTs and the synthetic design of GTs for medical and biotechnology applications. Here, we present GTXplorer as an integrated platform that presents evolutionary information of GTs adopting a GT-A fold in an intuitive format enabling *in silico* investigation through comparative sequence analysis to derive informed hypotheses about their function and regulation. The tree view mode provides an overview of the evolutionary relationships of GT-A families and allows users to select phylogenetically relevant families for comparisons. The selected families can then be compared in the alignment view at the residue level using annotated weblogo stacks of the GT-A core specific to the selected clade, family or subfamily. All data are easily accessible and can be downloaded for further analysis. GTXplorer can be accessed at https://uga-gta.netlify.app/ or from github at https://github.com/esbgkannan/GTXplorer to deploy locally. By packaging multiple data streams into an accessible, user friendly format, GTXplorer presents the first evolutionary data analytics platform to promote glycosyltransferase research.

Author Contributions:

Conceptualization: RT, NK. Data curation and collection: RT. Formal analysis: RT, AV. Methodology: RT, SS, NK. Validation: RT, SS, AP, AV. Visualization and figure generation: RT,

SS, AP, AV, WY. Software: RT, SS, AP, WY. Supervision: KK, NK. Writing – Original Draft: RT. Writing – review and editing: RT, SS, AP, AV, WY, KK, NK.

4.1 Introduction

Glycosyltransferases (GTs) catalyze the biosynthesis of complex carbohydrates, one of the most abundant biopolymers present in any cell. Defects in GTs have been implicated in several diseases such as congenital disorders of glycosylation, neurodegeneration and cancer [1–5]. In addition, there is a significant interest in the optimization, synthesis and engineering of GTs for applications in biofuels, food products, agricultural and livestock industries and more recently, for production of a new generation of highly efficient glycoconjugate vaccines [6–8]. Glycosyltransferases are currently classified into 114 families by the CAZy database, most of which adopt one of 3 major folds: GT-A, B or C [9]. The diversity in their primary sequences with large insertion loops in the catalytic domain and the presence of very few conserved catalytic residues have presented a major challenge in comparative analyses across this large family of enzymes.

Recently, we relied on the structural conservation within the GT-A fold families to present a comprehensive evolutionary framework connecting the functional diversity across all GT-A fold enzymes [10]. This study has generated a wealth of curated information such as curated alignments of over half a million GT-A sequences, phylogenetic profiles and annotations for statistically conserved amino acid positions that can provide key insights into the relationships between sequence, structure, function and regulation of GT-As through comparative sequence analyses and serve as a conceptual starting point for generating testable hypotheses. However, the general usability and interpretability of such complex data structures relies on the availability of a clear user-friendly tool that presents this contextual information in a coherent visualization scheme.

The past decade has seen much progress towards bioinformatics resources for glycoscience research. For example, GlyGen [11] provides an integrative portal to retrieve glycoconjugate related

data. Multiple glycoenzymes related databases provide curated interlinked information about their primary sequences [9,12,13]. Resources focusing on the structural aspects of glycoscience, specifically the conversion of carbohydrate sequences to 3D models [14,15], cataloguing the 3D structural information of carbohydrates and related enzymes [16] and visualization of carbohydrate structures [17,18] have already proven to be immensely valuable in driving glycoscience research. However, in order to make sense of the layered information encoded in the expansive range of different glycans, we first need to understand the evolution of biosynthetic machinery that led to this diversification. Yet, among these budding compendiums of resources, a data analytics framework for mining the relationships connecting sequence, structure, function and regulation of GTs is currently lacking.

Here, we present GTXplorer, an interactive tool for the visualization of GT-A evolutionary information which provides an integrated platform for comparative analysis of more than half a million sequences that adopt a GT-A type fold and are spread across 9 clades, 53 families and 99 subfamilies. By presenting complex data sources linking the evolutionary relationships between these diverse families of enzymes in an intuitive browser, GTXplorer enables researchers to extract and use the evolutionary data encoded in thousands of GT-A sequences from diverse organisms to interpret and predict the functional impact of natural variants in primary sequences. Through an interactive phylogenetic tree navigator and an annotated contrast alignment viewer, GTXplorer offers modular access to perform family, subfamily or amino acid residue level comparisons. Built using the REACT and the jQuery libraries, users can easily navigate as well as download relevant information for generating publication-quality images and to support other in-house analyses. GTXplorer connects data from disparate sources and format into an interpretable form that is readily accessible as a web portal or a locally built application for use in performing comparative

evolutionary analyses of GT-A fold enzymes and generating novel testable hypotheses driving experimental research in glycoscience.

4.2 Results

4.2.1 The GTXplorer portal

We present GTXplorer as a user-friendly interactive portal to navigate through the large volumes of data related to the GT-A fold families and subfamilies. It includes a diverse portfolio of information related to the classification and distribution of GT-A sequences along with a repository of sequence, alignment and phylogenetic data (Figure 4.1).

Data is organized into three layers of GT-A classification (clades, families and subfamilies) as outlined in [10]. The topmost level are the nine major GT-A clades that include families, some of which are further subdivided into subfamilies based on conserved pattern positions. Two modes of navigation options are provided: 1) A tree view mode which provides an interactive phylogenetic tree that users can click on to access detailed information about specific GT-A clades and families and 2) An alignment mode where users can access a hierarchical layout of GT-A clades, families and subfamilies that can be selected to generate comparative alignments using weblogos [19]. Both views offer insights into different aspects of GT evolution. The phylogenetic tree depicts the evolutionary relationships across GT-A families. Clicking on the clades or families in the tree view brings up a card that includes a short description of the known roles of sequences within that clade/family along with additional information about the mechanism, domain organization, sub-classification and quantified presence across major taxonomic groups.

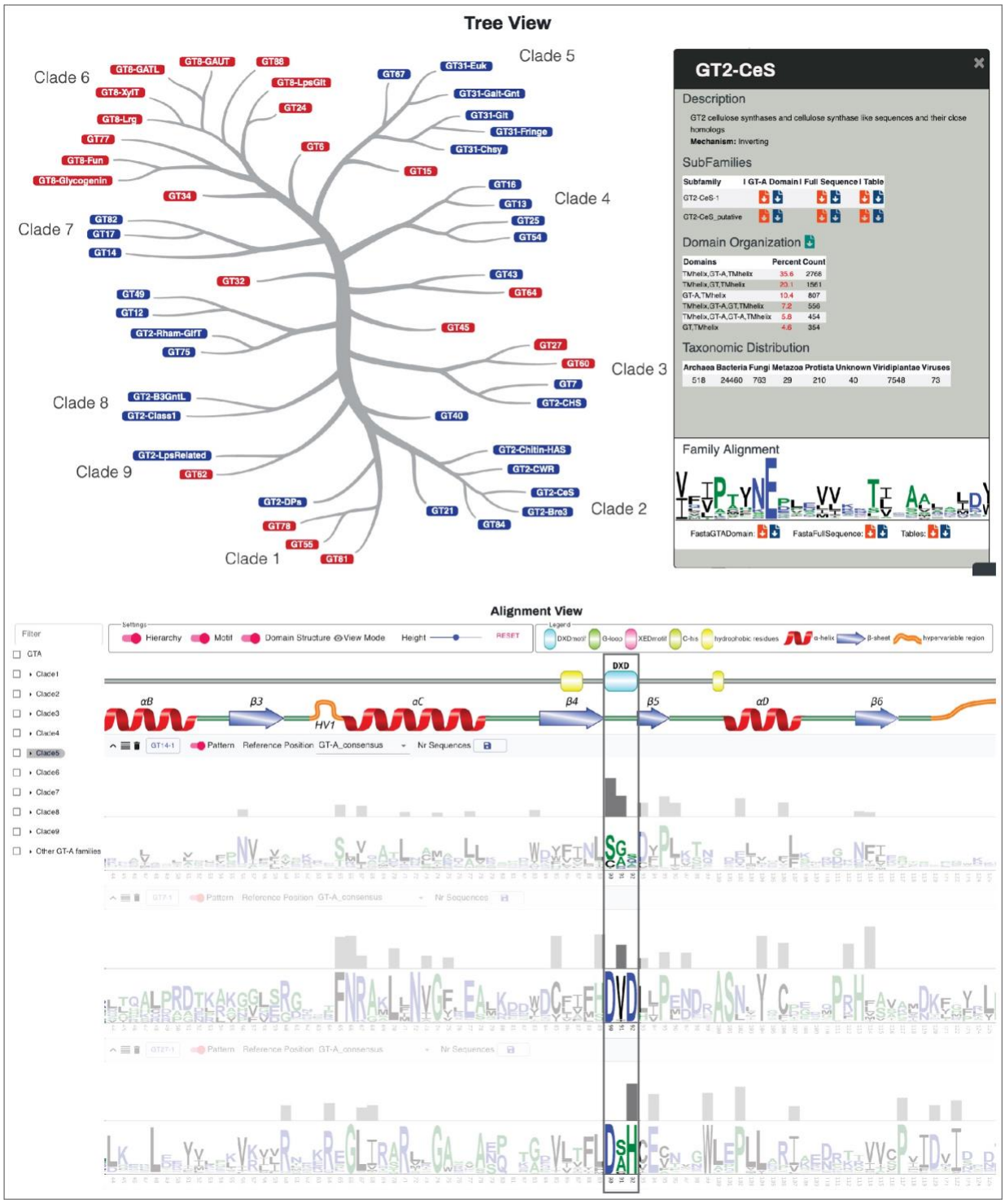


Figure 4.1: The GTXplorer web interface. Top: Tree view panel shows the interactive tree on the left. Upon clicking a node (in this case, family GT2-CeS), a card appears on the right with all the information about that node. Bottom: A snapshot of the alignment view highlighting the variation in the DXD motif between the metal independent GT14-1 subfamily and the metal dependent GT7-1 and GT27-1 subfamilies.

On the other hand, the alignment view offers a way to compare large sequence alignments across clades/families/subfamilies to identify conserved positions at different levels of the hierarchy. The alignment is annotated with conserved structural features, motifs and other regions of functional importance and any sequence from the hierarchical level can be used to number the aligned position, thus allowing easy mapping of these features to any sequence of interest. Also, we annotate the conserved pattern positions identified in our previous Bayesian pattern based analysis for all the subfamilies which provide informed targets for mutational analyses in functional and regulation studies [10] (gray bars in bottom panel of Figure 4.1).

These intuitive view modes with multiple customization options can also help generate snapshots with the most relevant information to be used in publications. The datasets are generated from two large sequence repositories (NCBInr [20] and UniProt proteomes [21]) and download links are conveniently placed throughout the navigation pages to download any specific piece of information desired from any of the 2 repositories. Users can download raw full length sequences in fasta format for any specific clade/family/subfamily. In addition, users can also download fasta files with an alignment of the GT-A domains or tabular formatted files with additional information.

4.2.2 Example case studies with GTXplorer to infer catalytic mechanisms and donor specificity

We illustrate the use of GTXplorer in hypothesis generation using 2 case studies outlining the observations made in [10]. For the first case, we observe the natural variation in the catalytic base position (xED-Asp) in the context of multiple evolutionary origins of retaining and inverting mechanisms (Figure 4.2A). GT-A families with an inverting mechanism use a catalytic base to

deprotonate the acceptor nucleophile, facilitating the initiation of a nucleophilic attack with direct displacement of the phosphate leaving group [22,23].

In contrast, for the retaining GT-As, in light of crystal structures with acceptor complexes that lack a suitably positioned conserved residue that could act as a catalytic base supported by mutational studies have led to the proposal of an alternative SNi mechanism where the phosphate leaving group acts as the catalytic base instead [24]. Thus, the two mechanisms exert different evolutionary pressures on the catalytic base position which would have to be conserved in the inverting enzymes, since it is essential for catalytic activity. This can be visualized using GTXplorer where a user could: Step 1- find and select families from inverting or retaining clades (for example GT2-CWR, GT21 and GT16 inverting from clades 1 and 6 and GT8-GAUT and GT88 retaining from clade 9) in the phylogenetic tree, Step 2 - turn on the radio button for motif annotation and Step 3 - scroll to find the xED motif position in the alignment viewer. In the alignment, the conservation of the xED-Asp is clearly highlighted for the selected inverting families.

However, in our previous study, we observed that there were a number of families that did not adhere to this rule where retaining families conserved the xED-Asp residue and inverting families like GT2-DPs, GT43, GT14 and others found alternative modes of inversion relieving the constraints to conserve the xED-Asp [10]. These exceptional families generally occurred in clades where inverting and retaining families were grouped together or closer to one another. This observation can be readily accessed in GTXplorer by simply selecting additional families from other clades (such as GT43, GT82, GT64 and GT45) and visualizing their alignment at the xED motif in the alignment viewer (Figure 4.2A). By making such comparative sequence analyses readily accessible, GTXplorer enables users to find trends and derive hypotheses based on the evolutionary conservation of specific residue positions.

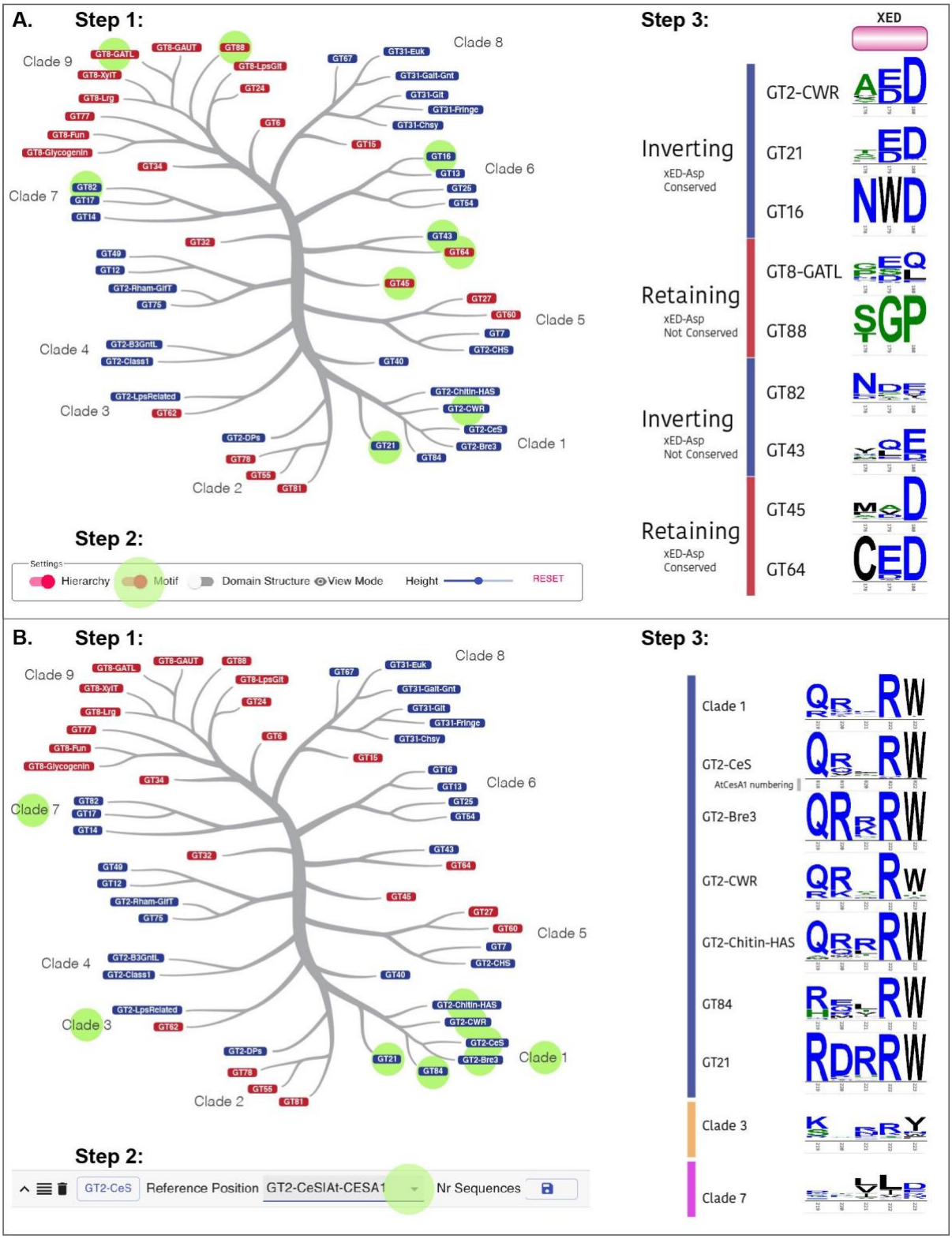


Figure 4.2: Example cases for the use of GTXplorer. (A) Steps to visualize changes in the catalytic base position (xED-Asp) based on the mechanism of different families. (B) Steps to identify Clade 1 specific [QR]XXRW motif involved in donor and acceptor binding.

For the second case, we use GTXplorer to identify a clade-specific motif essential for substrate binding. Clade 1 groups processive GT-A families generally involved in the biosynthesis of large polysaccharides or glycosphingolipids. A [QR]XXRW motif has been identified in its members where it has been shown to be important for binding both the donor and the acceptor [25]. We illustrated the extent of conservation of this motif within the Clade 1 members.

GTXplorer allows users to conveniently locate such conserved motifs starting from their sequence of interest (Figure 4.2B). For example, a user can easily start from a sequence of interest, like the *Arabidopsis thaliana* cellulose synthase 1 (AtCesA1) which is one of the subunits required for cellulose biosynthesis in *A. thaliana* [26]. This sequence is part of the GT2-CeS family which can be identified from the table available for download through GTXplorer. By displaying the weblogo alignment for GT2-CeS by selecting that family in the phylogenetic tree, users can locate the conserved QXXRW motif and even number the alignment positions using the AtCesA1 sequence as reference (Figure 4.2B Step 2).

To make comparisons of this motif with other members of the clade, other families of Clade 1 can be selected, all of which show up in the alignment viewer and display the conserved motif. To contrast this with other clades, users could select for example, Clade 3 and 7 where these motif positions are not conserved at all, highlighting the conservation of [QR]XXRW motif only within clade 1 members. Thus, GTXplorer can be a very useful tool in finding such evolutionarily constrained positions that point to conserved modes of function such as substrate binding.

4.3 Conclusion

The wide array of well curated information and its accessibility provided by GTXplorer makes it a user-friendly, interactive community resource that can strengthen both computational and

experimental research into the GT-A fold families, as highlighted by the specific examples provided. Coherent presentation of evolutionary information encoded in sequences allows experts and novices to easily visualize and perform comparative analyses of sequence conservation across evolutionary groups of GT-As. This analysis can then either be used to make interpretations of their results at hand or derive testable hypotheses to guide future experiments in glycoscience.

4.4 Methods

The interactive tree used in the tree view is generated using an annotated SVG file. Upon clicking on the nodes of the tree, users interact through a jquery which pulls the relevant card along with the contained information.

The alignment view is built on a component-based framework using the React application that provides flexibility, interoperability and reusability required by our data analytics platform to handle multiple sources of data. Within the React application, 3 main components are defined: 1) a hierarchical tree viewer for the selection of clades, families and subfamilies, 2) the main comparative alignment viewer and 3) the settings section. The hierarchical tree is rendered by parsing a JSON file in the backend that includes associated information about the clades, families and subfamilies. The comparative alignment viewer displays a summarized view of the alignment rendered as weblogo images in separate components. As such, these individual sub components are movable and adjustable independently and have options to download or display additional information. The settings section provides functions to customize the alignment view. This provides a flexible implementation that can easily be adjusted using auxiliary files to accommodate new datasets or adapt the tool for visualizing an entirely different protein family as well.

The web version of GTXplorer is hosted under Netlify. Users can also download GTXplorer from github and deploy a local version of the application using the provided instructions ensuring continued access to both the data and the interactive platform.

The GT-A phylogenetic tree, sequences, alignments and taxonomic information were obtained from [10]. The phylogenetic tree was redrawn in Adobe Illustrator for better aesthetics. The alignment weblogos were generated using the Weblogo tool [19]. Domain bounds and domain organization information was collected by querying the GT sequences using Batch CD-Search [27].

Bibliography

[1] Clarke, E., Green, R.C., Green, J.S., Mahoney, K., et al., Inherited deleterious variants in GALNT12 are associated with CRC susceptibility. *Hum Mutat* 2012, 33, 1056–1058.

[2] Kinoshita, M., Mitsui, Y., Kakoi, N., Yamada, K., et al., Common glycoproteins expressing polylactosamine-type glycans on matched patient primary and metastatic melanoma cells show different glycan profiles. *J Proteome Res* 2014, 13, 1021–1033.

[3] Joshi, H.J., Hansen, L., Narimatsu, Y., Freeze, H.H., et al., Glycosyltransferase genes that cause monogenic congenital disorders of glycosylation are distinct from glycosyltransferase genes associated with complex diseases. *Glycobiology* 2018, 28, 284–294.

[4] Gupta, R., Leon, F., Thompson, C.M., Nimmakayala, R., et al., Global analysis of human glycosyltransferases reveals novel targets for pancreatic cancer pathogenesis. *British Journal of Cancer* 2020, 122, 1661–1672.

[5] Moll, T., Shaw, P.J., Cooper-Knock, J., Disrupted glycosylation of lipids and proteins is a cause of neurodegeneration. *Brain* 2020, 143, 1332–1340.

[6] Himmel, M.E., Ding, S.-Y., Johnson, D.K., Adney, W.S., et al., Biomass recalcitrance: engineering plants and enzymes for biofuels production. *Science* 2007, 315, 804–807.

[7] McArthur, J.B., Chen, X., Glycosyltransferase engineering for carbohydrate synthesis. *Biochem Soc Trans* 2016, 44, 129–142.

[8] Micoli, F., Bino, L.D., Alfini, R., Carboni, F., et al., Glycoconjugate vaccines: current approaches towards faster vaccine design. *Expert Review of Vaccines* 2019, 18, 881–895.

[9] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 2014, 42, D490-495.

[10] Taujale, R., Venkat, A., Huang, L.-C., Zhou, Z., et al., Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases. *eLife* 2020, 9, e54532.

[11] York, W.S., Mazumder, R., Ranzinger, R., Edwards, N., et al., GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* 2020, 30, 72–73.

[12] Egorova, K.S., Smirnova, N.S., Toukach, P.V., CSDB_GT, a curated glycosyltransferase database with close-to-full coverage on three most studied nonanimal species. *Glycobiology* 2020.

[13] Yamada, I., Shiota, M., Shinmachi, D., Ono, T., et al., The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. *Nature Methods* 2020, 17, 649–650.

[14] Kirschner, K.N., Yongye, A.B., Tschampel, S.M., González-Outeiriño, J., et al., GLYCAM06: a generalizable biomolecular force field. Carbohydrates. *J Comput Chem* 2008, 29, 622–655.

[15] Engelsen, S.B., Hansen, P.I., Pérez, S., POLYS 2.0: An open source software package for building three-dimensional structures of polysaccharides. *Biopolymers* 2014, 101, 733–743.

[16] Pérez, S., Sarkar, A., Rivet, A., Breton, C., Imberty, A., Glyco3D: a portal for structural glycosciences. *Methods Mol Biol* 2015, 1273, 241–258.

[17] Sarkar, A., Pérez, S., PolySac3DB: an annotated data base of 3 dimensional structures of polysaccharides. *BMC Bioinformatics* 2012, 13, 302.

[18] Pérez, S., Tubiana, T., Imberty, A., Baaden, M., Three-dimensional representations of complex carbohydrates and polysaccharides—SweetUnityMol: A video game-based computer graphic software. *Glycobiology* 2015, 25, 483–491.

[19] Crooks, G.E., WebLogo: A Sequence Logo Generator. *Genome Research* 2004, 14, 1188–1190.

[20] Pruitt, K.D., Tatusova, T., Maglott, D.R., NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 2007, 35, D61-65.

[21] UniProt Consortium, UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019, 47, D506–D515.

[22] Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* 2008, 77, 521–555.

[23] Gloster, T.M., Advances in understanding glycosyltransferases from a structural perspective. *Current Opinion in Structural Biology* 2014, 28, 131–141.

[24] Moremen, K.W., Haltiwanger, R.S., Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat Chem Biol* 2019, 15, 853–864.

[25] Morgan, J.L.W., Strumillo, J., Zimmer, J., Crystallographic snapshot of cellulose synthesis and membrane translocation. *Nature* 2013, 493, 181–186.

[26] Taylor, N.G., Laurie, S., Turner, S.R., Multiple Cellulose Synthase Catalytic Subunits Are Required for Cellulose Synthesis in Arabidopsis. *The Plant Cell* 2000, 12, 2529–2539.

[27] Marchler-Bauer, A., Bo, Y., Han, L., He, J., et al., CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* 2017, 45, D200–D203.

CHAPTER 5

DISCUSSION AND CONCLUDING REMARKS

5.1 Achievement of goals

By applying novel computational approaches towards a holistic study of the glycosyltransferases, I have made original contributions towards understanding the evolutionary relationships across divergent families within this large class of enzymes and in the process, addressed each of the questions stated in the first chapter. The studies presented here specifically provide, for the first time, a phylogenetic model encompassing all known GT-A fold families providing valuable insights into the multiple evolutionary origins of their mechanism of transfer and specific observations that outline multiple adaptations towards substrate specificity. These findings demonstrate how the need to synthesize novel glycan types for survival has driven rapid evolution of its biosynthetic machinery shaping the highly adaptable and diverse GT repertoire. Using recent advances in machine learning, I provide accurate models that are able to predict donor substrates for the GT-A fold enzymes and a larger deep learning based fold prediction model that provides valuable insights into the GT structural fold diversity. Both methods serve as valuable tools to generate hypotheses and guide future experimental efforts. Finally, by packaging the huge amount of data generated during these studies into a user-friendly platform, I develop a data analytics tool

that provides clear and customizable ways to view complex datasets in easily interpretable integrative formats.

5.1.1 Evolution of the GT-A fold enzymes

Leveraging on the structural conservation within the GT-A fold, I applied a structure guided profile based sequence alignment strategy to align over half a million GT-A fold sequences which enabled a comprehensive evolutionary analysis of these sequences. For the first time, this study describes features of the common core conserved across all GT-A sequences and highlights the conserved catalytic residues as well as hydrophobic residues that make up a contiguous hydrophobic core keeping the GT-A fold together. While previous evolutionary studies focused on only enzymes within certain GT families, or GTs working on the same pathways [1,2], this study provides the first comprehensive phylogenetic model that describes the evolutionary trajectory of all known GT-A fold enzymes and classifies them into evolutionarily related groups. By identifying constraints shared by these groups, this study also provides valuable insights into shared features that can help understand function and regulation of GT-A enzymes. These observations can be very helpful in the design of synthetic GTs to identify regions that are crucial for minimal GT-A fold activity from regions that can be grafted to obtain the desired effect in terms of higher efficiency or substrate specificity.

5.1.2 A framework for GT fold prediction and classification

Although the primary sequences are very diverse, the 3D structure of GTs is very well conserved. Only 3 major folds have been described with some families known to adopt variant or different folds [3,4]. However, there are many GT families for which a fold type is yet to be assigned and it

is also likely that they may adopt novel folds that haven't been discovered [5]. This study employs a deep learning method to first understand the most distinguishing features of each fold type and them use those features to classify GT families of unknown fold to either one of the known classes or predict how likely they are to adopt a novel fold. This framework presents a flexible interpretable approach to learning fold-specific features which enabled me to highlight the core features for all the 3 major GT fold types. Identifying these core structural features provides a platform for comparative analyses within and across GT folds. Additionally, the predictions for the novel fold families provide targets for crystallographic studies to uncover novel GT folds and understand new modes of function. The dep learning framework adopted for this problem presents a highly interpretable extendable model that can easily be extended to the study of other large protein families as well. More importantly, this approach presents a completely alignment-free method of comparative analysis and fold prediction, which can be very useful in the study of families like GTs where insertions and evolution has made primary sequences very diverse and difficult to align.

5.1.3 Interactive tool for visualizing GT-A evolutionary relationships

Our comprehensive analysis of the GT-A fold families generated a wealth of curated information in different data formats ranging from a phylogenetic model describing the evolutionary relationships between the GT-A families, large sequence alignments and associated annotations to conserved pattern position annotations. To use these datasets for drawing meaningful observations, they have to be first organized and presented in a coherent way facilitating easy access and browsing to draw connections between the datasets. I developed GTXplorer as a possible solution to this problem which offers a data analytics platform for all users to easily access and perform comparative sequence analysis to draw meaningful hypotheses and guide future experiments.

GTXplorer presents the first evolutionary data framework focused on glycosyltransferases and plugs into the myriad of other carbohydrate and glycoenzyme related databases [6–9] by providing an evolutionary perspective to interpret observations regarding differences in glycoforms across organisms or changes associated with function and regulation. It can also serve as a valuable resource for the design of synthetic GTs that are optimized for desired function by providing annotated insights into the important aspects of GT-A enzymes for minimal structural stability or specific regions important for substrate binding activity.

5.1.4 Insights into the primary sequence diversity of the GT-A enzymes

Nearly 650,000 GT-A domain sequences from nr and more than 160,000 from UniProt proteomes are organized and annotated in the GTXplorer tool based on our analysis of the GT-A families in the second chapter, providing an unprecedented opportunity for a computational analysis of the distribution of GT-A sequences across different families and taxonomic groups. In this section, I provide some key insights into the distribution and nature of sequences within the novel classification of GT-A enzymes that can be helpful in future characterization of these families. Considering sequences from the well curated UniProt proteomes database, families like GT2-Class1 and GT2-DPs are found to house more than twenty thousand sequences spanning across all taxonomic groups (Figure 5.1A). These are the most widespread and well conserved GT-As likely indicating that they are closest to the GT-A ancestral sequences. On the other hand, families like GT78 and GT88 represent very small families with less than 20 sequences, and are found only in select taxonomic groups (mostly bacteria), indicating their divergence and organism-specific roles (Figure 5.1A). In general, most GT2 families were found to have high numbers of sequences.

Figure 5.1: Insights into the sequence diversity of GT-A families. A) Bar plot showing the number of sequences found in each GT-A family. B) Percent of GT-A sequence found from major taxonomic groups. C) Percent of sequences in each GT-A family predicted to have a transmembrane helix. D) Scatter plot showing average GT-A domain length vs average full length of GT-A sequences within each family.

78% of all GT-A sequences included in GTXplorer come from bacterial species, while 10% are accounted for by metazoans (Figure 5.1B). We also find few sequences spread across families from viruses. While this is in part a result of the number of organisms with full proteomes in the

UniProt database, this also reflects the diversity of GTs in different taxonomic hierarchies. From the prediction of transmembrane helices, we find that all members of Clade 1 tend to conserve transmembrane helices while the retaining members of Clade 2 (GT55, GT78 and GT81) have a lesser propensity to conserve transmembrane helices than its inverting members (GT2-DPs) (Figure 5.1C). We also find that in general, most families have a GT-A domain of length 220-280. GT24 that includes the human UGGTs involved in the glycoprotein folding quality control have the longest sequences (average length >1400 amino acids) yet have a GT-A domain of average length close to 240 (Figure 5.1D). On the other hand, GT88 family of bacterial α-glucosyltransferases (Lgts) were found to have the longest GT-A domains, most likely owing to the large multi-helix insertion in the HV2 region. Similarly, families GT8-GAUT, GT16, GT49, GT55 and GT75 also have large insertions in the HV regions that provide specificity for donor and acceptor binding and thus were found to have longer GT-A domains. On the other hand, GT31-Glt sequences were found to have short GT-A domains as well as full length sequences.

5.2 Future directions

The observations and findings presented here provide a rich dataset and valuable tools to understand the relationships connecting sequence, structure function and evolution of GTs. I highlight some avenues for future investigation in light of the presented novel studies.

5.2.1 Comparison of disease mutation and natural variation in GTs

GTs have important roles to play in nearly all aspects of cellular function through glycosylation of vital proteins and lipid molecules and the biosynthesis of polysaccharides and glycans that are essential for cell growth and development, membrane functions, cellular interactions and protein

folding. As such, defects in GTs often lead to breakdown in a cascade of mechanisms throughout the cell resulting in severe phenotypic consequences. GTs have been directly implicated in several diseases, most notably a collection of glycosylation disorders called the congenital disorders of glycosylation (CDG), with more than 100 CDGs already identified caused by deficiencies in some of the more than 200 glycogenes involved in different glycosylation pathways [10–12]. GTs play important roles in bacterial and viral infections [13] and defects in GTs have more recently also been linked to neurodegenerative diseases including familial amyotrophic lateral sclerosis (ALS), Alzheimer's disease, Huntington's disease and Parkinson's disease [14]. GTs have also been found to play important roles in cancer progression [15–17]. Specifically, changes in GT expression profiles have been associated with cancer subtypes [18] and have been suggested as markers for early tumorigenesis [19]. Clearly, there is a lot to learn about the functional roles of GTs and how they can be targeted for preventive measures against these diseases.

There have been scientific efforts at characterizing and mining disease variations in GTs [20] that have aimed at cataloging mutations across glycogenes to understand their association with disease phenotypes. However, such studies have been limited in scope to individual genes or closely related families due to lack of a comparative framework across GT families. Here, we provide well curated sequence alignments and a phylogenetic framework connecting the relationships across GT-A fold families. This opens up the possibility to, for the first time, understand these disease variations in the context of the GT-A common core elements and learn from the mutational effects of one family to predict its implications in a related family. Moreover, by virtue of the large curated alignments, the disease variations can directly be compared against the natural variations occurring at subsites to understand directly the functional consequences of a specific point mutation and how that might affect protein structure, function and stability.

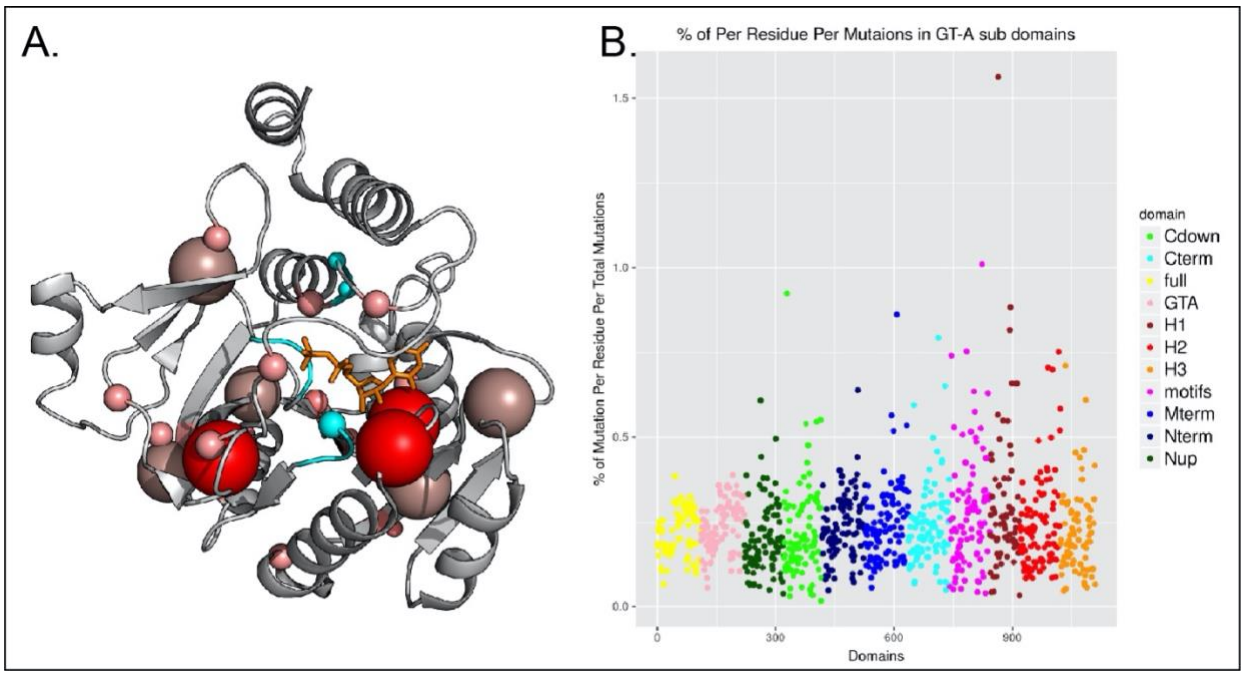


Figure 5.2: Mapping the disease mutations to the GT-A common core. (A) Disease mutations are mapped to a representative GT-A core structure (PDB ID:2d7i). Number of mutations are reflected by the size and color of the spheres with bright red and large spheres indicating the highest number of mutations mapping to that position. The active site residue positions (DXD, xED and G-loop) are highlighted in cyan. The UDP-donor is shown as orange sticks. (B) Categorical scatter plot indicating the percent of mutations per residue per total mutations within any given region in the GT-A core of all the human GT-A sequences. The different regions are Nup: Upstream from the GT-A domain; Nterm: N-terminal GT-A core upto the DXD motif; Mterm: Middle region of the GT-A domain from the DXD motif to the xED motif; motifs: The catalytic motif (DXD, xED, G-loop and C-His) positions; H1,H2,H3: hypervariable regions 1, 2 and 3 respecitively; Cterm: C-terminal GT-A core regions from the xED motif to the HV3 region; Cdown: Downstream from the GT-A domain;full: Full length sequence; GTA: The entire GT-A domain.

I have made some progress in this front where preliminary results have suggested a widespread distribution of disease mutations throughout the GT-A core domain. I collected all the disease mutation information available on the human GT-A enzymes from The Cancer Genome Atlas (TCGA) and the Catalogue Of Somatic Mutations In Cancer (COSMIC) databases, two of the most comprehensive databases collecting disease variation data from whole genome and targeted

sequencing efforts on patients. These mutations were then mapped into the protein sequence positions for each GT in the context of the GT-A core alignment allowing for comparing the mutational types and positions across all GT sequences. Figure 5.2A shows the prevalence of disease mutations in the GT-A core across all human GT-A sequences. We find that the regions that are in structural proximity in the active site are more prone to disease mutations compared to other peripheral sites in the GT domain. It is likely that these mutations are destabilizing the active site of the enzyme thus losing its catalytic activity resulting in the disease phenotypes. However, the specific effects of these mutations on the functions is yet to be studied.

We also notice that along the primary sequence of the GT-A domain, some regions are more prone to disease mutations than others. Specifically, the 3 HV regions were found to harbor slightly more mutations (Figure 5.2B). These regions are family specific and have conserved residues that are directly involved in substrate interactions. It is thus important to note that higher frequency of disease mutations in these regions could have a direct effect on the enzyme's ability to recognize and bind its specific substrates thus leading to the disease phenotype. Thus, our preliminary analyses suggest disease mutations are occurring in the functional regions of human GT-As and could have a direct impact on their activity.

Another aspect that is important to note about the GTs is their ability to accommodate incredible diversity even within the active site residues. Compared to other large protein families like kinases, GTs do not have a single invariant amino acid position that is conserved throughout all families. For the GT-A fold enzymes, the DXD motif is typically considered a hallmark but even that motif has a significant variation (Figure 5.3). For our preliminary analyses, we mark GT-A enzymes lacking a DXD motif as non-canonical. As shown in Figure 5.3, we find that some GT-A families like GT14, GT15 and GT2-Chitin-HAS are much richer in non-canonical sequences

suggesting that these families are either metal independent and do not require to conserve the DXD motif anymore, or have developed alternative modes of binding the metal ion. With our analysis, the disease mutations can be analyzed in the context of natural variations to find associations in how variations affect different modes of function within the GT-A fold families.

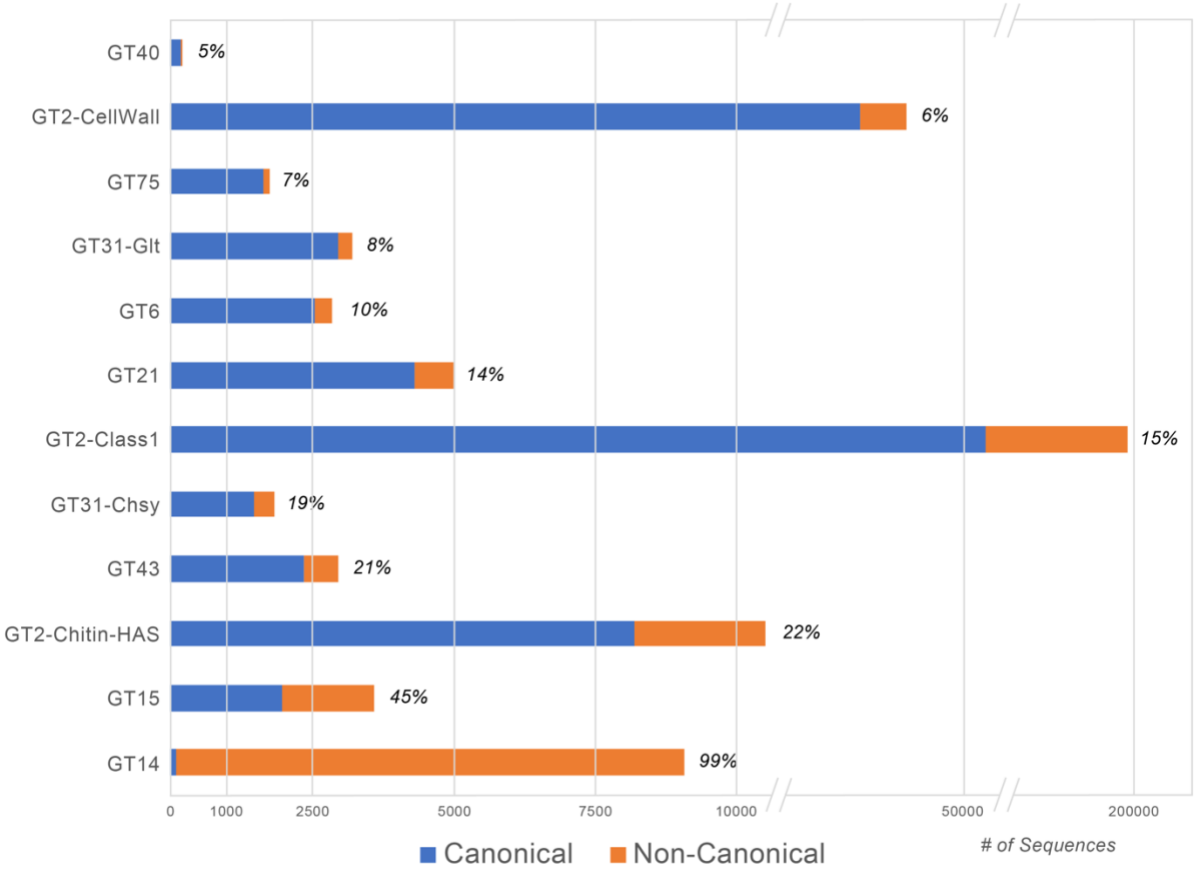


Figure 5.3: Distribution showing the number of canonical (with the DXD motif) and non-canonical (lacking the DXD motif) sequences in GT-A fold families. Only the 12 families with the highest number of non-canonical sequences are shown. Numbers at the tip of the bars indicate the percentage of sequences in that family that are non-canonical.

5.2.2 Directed evolution towards a minimal GT-A fold enzyme template

GTs have been a focus of attention as useful synthetic tools in designing natural oligosaccharides and glycoconjugates for a long time [21,22]. As natural catalysts for the biosynthesis of these molecules, they offer well optimized systems that can be fine-tuned to generate glycodiversity. Many GTs have been known to be promiscuous towards their substrates thus opening an opportunity to find one that could catalyze a required sugar transfer reaction [23]. However, it is not always possible to find such promiscuous GTs with required efficiencies. In many cases, the desired carbohydrates may not exist in nature or require unnatural modifications and the biological processes synthesizing a desired glycan may not be scalable or efficient for wider use. Even when a naturally occurring GT is found, the complexities of expression and purification could make it unsuitable [24,25]. As such, alternative strategies through GT engineering has been pursued for more than a decade for optimizing the large scale production of glycans of interest for medical and industrial purposes [24,26–29]. With the increase in GT structures, especially ones with donor and substrate analogues, structure guided directed evolution methods have been increasingly successful in optimizing GT function [24,30]. High throughput screening strategies for the performance of such methods and selection of mutants with best traits have also been developed [31]. However, a proper understanding of the key elements for GT function and activity is essential for exploring all possibilities with GT engineering.

Our evolutionary study of the GT-A fold families enabled us to identify their core conserved features based on a rigorous statistical analysis of over half a million sequences providing residue level annotations on specific positions that comprise the hydrophobic core important for stability and folding, the active site residues required for minimal catalytic activity and the positional bounds for family specific insertions involved in substrate recognition and regulation through various

mechanisms (Figure 5.4). This provides rich information for the design of an ancestral GT-A sequence that only has the minimal elements required for activity and proper folding that could serve as a template upon which additional loop regions and insertions can be grafted for providing the desired functional specificities.

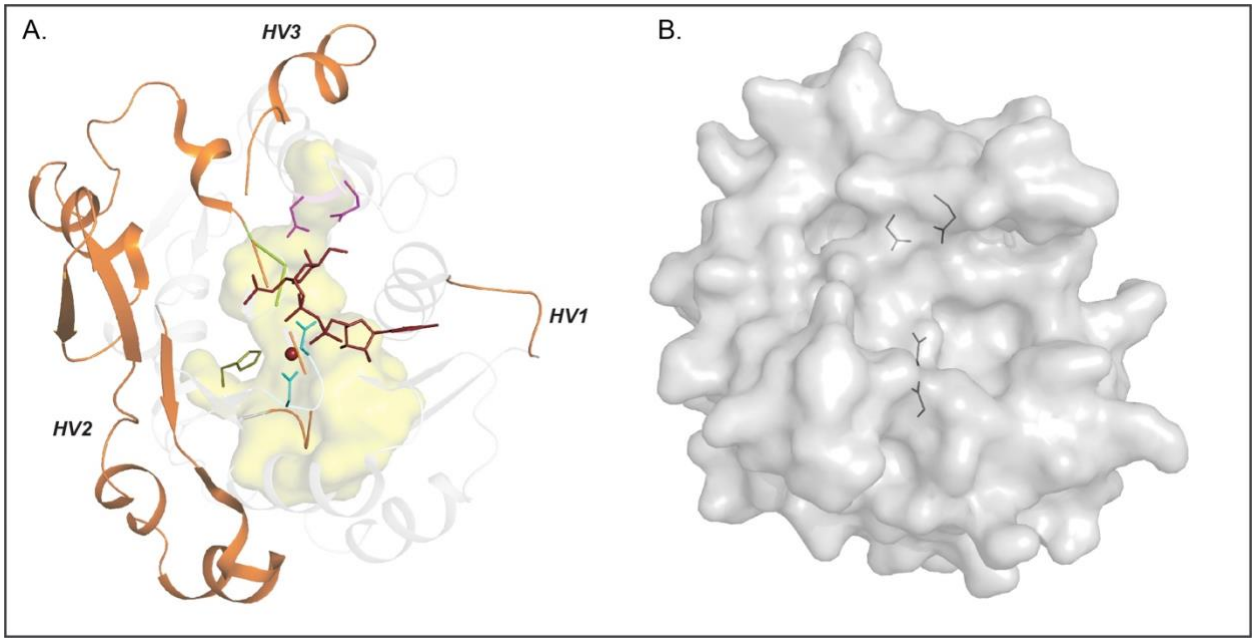


Figure 5.4: The minimal GT-A core unit. (A) Figure highlights the important parts of the GT-A core. The core catalytic motifs in the active site are shown in colored sticks (DXD: cyan, xED: magenta, G-loop: green, C-His: olive). The hydrophobic core is shown as yellow surface. Donor substrate is shown as brown sticks and the metal ion as brown sphere. Hypervariable regions that are added to this minimal core unit are shown as orange cartoon. (B) Surface representation of only the minimal core unit of GT-A after removing the hypervariable regions. Location of the catalytic motifs (DXD and xED) are indicated using gray sticks.

For example, our analysis identified 3 hypervariable regions (HV1,2 and 3, marked in Figure 5.4A) that point to specific positions where family specific insertions with conserved pattern positions involved in substrate binding can be found across diverse GT-As. Many GT-A families present natural variations with only a short loop replacing the elaborate HV regions suggesting that evolution has found ways of accommodating these insertions independently of the core GT-A fold. In fact, in our study, we found several naturally occurring bacterial and archaeal GT-A sequences that have minimal insertions in these regions and conserve only the core components. These sequences likely represent the closest surviving relatives to the ancestral GT-A sequences that all current families evolved from. These sequences serve as excellent templates for the design of a minimal GT-A unit.

We also identified fourteen hydrophobic residue positions that were significantly conserved across all GT-A enzymes that form a contiguous hydrophobic core originating in the Rossmann domain and extending to connect the xED helix and other helices that make up the conserved active site of GT-A enzymes (yellow surface in Figure 5.4A). Identifying these residue positions that connect all the critical components required to make up the GT-A core provides additional rationale for designing a synthetic enzyme that conserves appropriate hydrophobic residues at these positions ensuring the presence of hydrophobic effects driving proper folding. Additionally, identification of the catalytic motif positions that make up the active site provides us with a repertoire of natural variation in over half a million sequences across billions of years of evolution that have been accommodated in these positions. This provides an unprecedented opportunity to learn co-evolution of residues at these positions to provide an informed list of catalytic residue combinations for desired functional effect within the active site.

Building upon these observations and findings, the studies presented here facilitate a rich source of information required for the design of a minimal functional GT-A unit that can then be used for engineering strategies to act as a modular template for generating GTs with desired functional roles (Figure 5.4B). Members in our lab have already started investigating these prospects by computationally optimizing a closely matching GT-A consensus sequence for folding and stability. This synthetic protein sequence is currently being expressed in multiple expression systems and experimentally analyzed for proper folding and stability.

Bibliography

[1] Taujale, R., Yin, Y., Glycosyltransferase family 43 is also found in early eukaryotes and has three subfamilies in charophycean green algae. *PloS One* 2015, 10.

[2] Lombard, J., The multiple evolutionary origins of the eukaryotic N-glycosylation pathway. *Biol. Direct* 2016, 11.

[3] Lairson, L.L., Henrissat, B., Davies, G.J., Withers, S.G., Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* 2008, 77, 521–555.

[4] Moremen, K.W., Haltiwanger, R.S., Emerging structural insights into glycosyltransferase-mediated synthesis of glycans. *Nat. Chem. Biol.* 2019, 15, 853–864.

[5] Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P.M., Henrissat, B., The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 2014, 42, D490–D495.

[6] Egorova, K.S., Smirnova, N.S., Toukach, P.V., CSDB_GT, a curated glycosyltransferase database with close-to-full coverage on three most studied nonanimal species. *Glycobiology* 2020.

[7] Pérez, S., Sarkar, A., Rivet, A., Breton, C., Imberty, A., Glyco3D: a portal for structural glycosciences. *Methods Mol. Biol. Clifton NJ* 2015, 1273, 241–258.

[8] Yamada, I., Shiota, M., Shinmachi, D., Ono, T., et al., The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences. *Nat. Methods* 2020, 17, 649–650.

[9] York, W.S., Mazumder, R., Ranzinger, R., Edwards, N., et al., GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* 2020, 30, 72–73.

[10] Freeze, H.H., Understanding Human Glycosylation Disorders: Biochemistry Leads the Charge. *J. Biol. Chem.* 2013, 288, 6936–6945.

[11] Haeuptle, M.A., Hennet, T., Congenital disorders of glycosylation: an update on defects affecting the biosynthesis of dolichol-linked oligosaccharides. *Hum. Mutat.* 2009, 30, 1628–1641.

[12] Jaeken, J., Congenital disorders of glycosylation (CDG): it's (nearly) all in it! *J. Inherit. Metab. Dis.* 2011, 34, 853–858.

[13] Bilverstone, T.W., Garland, M., Cave, R.J., Kelly, M.L., et al., The glucosyltransferase activity of C. difficile Toxin B is required for disease pathogenesis. *PLOS Pathog.* 2020, 16, e1008852.

[14] Moll, T., Shaw, P.J., Cooper-Knock, J., Disrupted glycosylation of lipids and proteins is a cause of neurodegeneration. *Brain* 2020, 143, 1332–1340.

[15] Venkitachalam, S., Guda, K., Altered glycosyltransferases in colorectal cancer. *Expert Rev. Gastroenterol. Hepatol.* 2017, 11, 5–7.

[16] Venkitachalam, S., Revoredo, L., Varadan, V., Fecteau, R.E., et al., Biochemical and functional characterization of glycosylation-associated mutational landscapes in colon cancer. *Sci. Rep.* 2016, 6, 23642.

[17] Fernández, L.P., Sánchez-Martínez, R., Vargas, T., Herranz, J., et al., The role of glycosyltransferase enzyme GCNT3 in colon and ovarian cancer prognosis and chemoresistance. *Sci. Rep.* 2018, 8, 8485.

[18] Ashkani, J., Naidoo, K.J., Glycosyltransferase Gene Expression Profiles Classify Cancer Types and Propose Prognostic Subtypes. *Sci. Rep.* 2016, 6, 26451.

[19] Andergassen, U., Liesche, F., Kölbl, A.C., Ilmer, M., et al., Glycosyltransferases as Markers for Early Tumorigenesis. *BioMed Res. Int.* 2015, 2015, e792672.

[20] Hansen, L., Lind-Thomsen, A., Joshi, H.J., Pedersen, N.B., et al., A glycogene mutation map for discovery of diseases of glycosylation. *Glycobiology* 2015, 25, 211–224.

[21] Chang, A., Singh, S., Phillips, G.N., Thorson, J.S., Glycosyltransferase structural biology and its role in the design of catalysts for glycosylation. *Curr. Opin. Biotechnol.* 2011, 22, 800–808.

[22] Palcic, M.M., Glycosyltransferases as biocatalysts. *Curr. Opin. Chem. Biol.* 2011, 15, 226–233.

[23] Hancock, S.M., Vaughan, M.D., Withers, S.G., Engineering of glycosidases and glycosyltransferases. *Curr. Opin. Chem. Biol.* 2006, 10, 509–519.

[24] McArthur, J.B., Chen, X., Glycosyltransferase engineering for carbohydrate synthesis. *Biochem. Soc. Trans.* 2016, 44, 129–142.

[25] Moremen, K.W., Ramiah, A., Stuart, M., Steel, J., et al., Expression system for structural and functional studies of human glycosylation enzymes. *Nat. Chem. Biol.* 2018, 14, 156–162.

[26] Schumann, B., Malaker, S.A., Wisnovsky, S.P., Debets, M.F., et al., Bump-and-Hole Engineering Identifies Specific Substrates of Glycosyltransferases in Living Cells. *Mol. Cell* 2020, 78, 824-834.e15.

[27] Williams, G.J., Zhang, C., Thorson, J.S., Expanding the promiscuity of a natural-product glycosyltransferase by directed evolution. *Nat. Chem. Biol.* 2007, 3, 657–662.

[28] Cho, K.W., Kim, T.-S., Le, T.T., Nguyen, H.T., et al., Altering UDP-glucose Donor Substrate Specificity of Bacillus licheniformis Glycosyltransferase towards TDP-glucose 2019, 29, 268–273.

[29] Williams, G.J., Goff, R.D., Zhang, C., Thorson, J.S., Optimizing glycosyltransferase specificity via 'hot spot' saturation mutagenesis presents a new catalyst for novobiocin glycorandomization. *Chem. Biol.* 2008, 15, 393.

[30] Akere, A., Chen, S.H., Liu, X., Chen, Y., et al., Structure-based enzyme engineering improves donor-substrate recognition of Arabidopsis thaliana glycosyltransferases. *Biochem. J.* 2020, 477, 2791–2805.

[31] Aharoni, A., Thieme, K., Chiu, C.P.C., Buchini, S., et al., High-throughput screening methodology for the directed evolution of glycosyltransferases. *Nat. Methods* 2006, 3, 609–614.

APPENDIX A

EXTENDED RESULTS

In addition to the results presented in the main dissertation, I contributed to multiple collaborative projects during my Ph.D. study that resulted in multiple co-authored journal publications and manuscript submissions. The abstract for these bodies of work are listed below.

Characterization of a cytoplasmic glucosyltransferase that extends the core trisaccharide of the Toxoplasma Skp1 E3 ubiquitin ligase subunit

Kazi Rahman, Msano Mandalasi, Peng Zhao, M. Osman Sheikh, Rahil Taujale, Hyun W.Kim, Hanke van der Wel, Khushi Matta, Natarajan Kannan, John N. Glushka, Lance Wells, Christopher M. West

Skp1 is a subunit of the SCF (Skp1/Cullin 1/F-box protein) class of E3 ubiquitin ligases that are important for eukaryotic protein degradation. Unlike its animal counterparts, Skp1 from *Toxoplasma gondii* is hydroxylated by an O_2-dependent prolyl-4-hydroxylase (PhyA), and the resulting hydroxyproline can subsequently be modified by a five-sugar chain. A similar modification is found in the social amoeba *Dictyostelium*, where it regulates SCF assembly and O_2-dependent development. Homologous glycosyltransferases assemble a similar core trisaccharide in both organisms, and a bifunctional α-galactosyltransferase from CAZy family GT77 mediates the addition of the final two sugars in *Dictyostelium*, generating Galα1, 3Galα1,3Fucα1,2Galβ1,3GlcNAcα1-. Here, we found that *Toxoplasma* utilizes a cytoplasmic glycosyltransferase from an ancient clade of CAZy family GT32 to catalyze transfer of the fourth sugar. Catalytically active Glt1 was required for the addition of the terminal disaccharide in cells,

---

Kazi Rahman, Msano Mandalasi, Peng Zhao, M. Osman Sheikh, Rahil Taujale, Hyun W.Kim, Hanke van der Wel, Khushi Matta, Natarajan Kannan, John N. Glushka, Lance Wells, Christopher M. West. 2017. Journal of Biological Chemistry 292(45):18644-18659

and cytosolic extracts catalyzed transfer of [^{3}H]glucose from UDP-[^{3}H]glucose to the trisaccharide form of Skp1 in a *glt1*-dependent fashion. Recombinant Glt1 catalyzed the same reaction, confirming that it directly mediates Skp1 glucosylation, and NMR demonstrated formation of a Glcα1,3Fuc linkage. Recombinant Glt1 strongly preferred the full core trisaccharide attached to Skp1 and labeled only Skp1 in *glt1*Δ extracts, suggesting specificity for Skp1. *glt1*-knock-out parasites exhibited a growth defect not rescued by catalytically inactive Glt1, indicating that the glycan acts in concert with the first enzyme in the pathway, PhyA, in cells. A genomic bioinformatics survey suggested that Glt1 belongs to the ancestral Skp1 glycosylation pathway in protists and evolved separately from related Golgi-resident GT32 glycosyltransferases.

Contributions:

I conducted the evolutionary analysis of the GT32 family of Skp1 related glycosyltransferases across diverse organisms and defined the evolutionary relationships across these sequences.

Dereplication of plant phenolics using a mass-spectrometry database independent method

Ricardo M. Borges, Rahil Taujale, Juliana Santana de Souza, Thaís de Andrade Bezerra, Eder Lana e Silva, Ronny Herzog, Francesca V. Ponce, Jean-Luc Wolfender, Arthur S. Edison

INTRODUCTION: Dereplication, an approach to sidestep the efforts involved in the isolation of known compounds, is generally accepted as being the first stage of novel discoveries in natural product research. It is based on metabolite profiling analysis of complex natural extracts.

OBJECTIVE: To present the application of LipidXplorer for automatic targeted dereplication of phenolics in plant crude extracts based on direct infusion high-resolution tandem mass spectrometry data.

MATERIAL AND METHODS: LipidXplorer uses a user-defined molecular fragmentation query language (MFQL) to search for specific characteristic fragmentation patterns in large data sets and highlight the corresponding metabolites. To this end, MFQL files were written to dereplicate common phenolics occurring in plant extracts. Complementary MFQL files were used for validation purposes.

Ricardo M. Borges, Rahil Taujale, Juliana Santana de Souza, Thaís de Andrade Bezerra, Eder Lana e Silva, Ronny Herzog, Francesca V. Ponce, Jean-Luc Wolfender, Arthur S. Edison. 2018. *Phytochemical Analysis*. 29(6):601-612

RESULTS: New MFQL files with molecular formula restrictions for common classes of phenolic natural products were generated for the metabolite profiling of different representative crude plant extracts. This method was evaluated against an open-source software for mass-spectrometry data processing (MZMine®) and against manual annotation based on published data.

CONCLUSION: The targeted LipidXplorer method implemented using common phenolic fragmentation patterns, was found to be able to annotate more phenolics than MZMine® that is based on automated queries on the available databases. Additionally, screening for ascarosides, natural products with unrelated structures to plant phenolics collected from the nematode Caenorhabditis elegans, demonstrated the specificity of this method by cross-testing both groups of chemicals in both plants and nematodes.

Contributions:

I was responsible for writing the scripts for the MFQL file conversion scripts and performed the comparisons of the *C. elegans* results to generate the comparative heatmap figure.

Extending compound identification for molecular network using the LipidXplorer database independent method: A proof of concept using glycoalkaloids from Solanum pseudoquina A. St.-Hil.

Vitor Soares, Rahil Taujale, Rafael Garrett, Antonio Jorge R. da Silva, Ricardo M. Borges

INTRODUCTION: Molecular networks are now established as the method of choice for tandem mass spectrometry dereplication and similarity-based structure elucidation. Node identification can be used to start the propagation of the structure elucidation of unknown compounds progressively.
OBJECTIVE: To demonstrate the capabilities of using the LipidXplorer data results along with molecular networking to identify nodes and aid sequential structure elucidation of unknown compounds.
MATERIAL AND METHODS: Molecular fragmentation query language (MFQL) files were written to identify glycoalkaloids based on known structures described for Solanum species. A dataset generated from liquid chromatography-high resolution mass spectrometry (LC-HRMS) analysis of Solanum pseudoquina sample were submitted to dereplication on both LipidXplorer software and Global Natural Products Social Molecular Network (GNPS) online system. The

Vitor Soares, Rahil Taujale, Rafael Garrett, Antonio Jorge R. da Silva, Ricardo M. Borges. 2018. *Phytochemical Analysis*. 30(2):132-138

resulting attribute table from GNPS calculations was merged with the LipidXplorer results and this merged file was used for network visualisation in Cytoscape. Nodes in the molecular network were labelled using the LipidXplorer identifiers, thus assisting the structure elucidation of unidentified compounds.

RESULTS: The combination of the LipidXplorer glycoalkaloids list and GNPS analysis was used in Cytoscape to label nodes in the molecular network. The analysis of the network using these labelled starting points triggered the structure elucidation of closely related nodes leading to the identification of 30 compounds using the LipidXplorer output and four purified and structure elucidated compounds, including a new glycoalkaloids identified as 3-O-(β-d-xylopyranosyl)-(20R,25S)-22,26-epimino-16-acetyl-cholesta-5,22(N)-diene.

CONCLUSION: A significant compound identification completely based on molecular formula and fragmentation queries was achieved. This new and effective approach could help researches to expand the identification rate of compounds in dereplication studies using molecular networks.

Contributions:

I was involved in designing the informatics aspect of this project, specifically writing the scripts to parse the MFQL files and automate generating files for LipdXplorer and also combine this information with relevant GNPS analysis data.

Tracing the origin and evolution of pseudokinases across the tree of life

Annie Kwon, Steven Scott, Rahil Taujale, Wayland Yeung, Krys J. Kochut, Patrick A. Eyers, Natarajan Kannan

Protein phosphorylation by eukaryotic protein kinases (ePKs) is a fundamental mechanism of cell signaling in all organisms. In model vertebrates, ~10% of ePKs are classified as pseudokinases, which have amino acid changes within the catalytic machinery of the kinase domain that distinguish them from their canonical kinase counterparts. However, pseudokinases still regulate various signaling pathways, usually doing so in the absence of their own catalytic output. To investigate the prevalence, evolutionary relationships, and biological diversity of these pseudoenzymes, we performed a comprehensive analysis of putative pseudokinase sequences in available eukaryotic, bacterial, and archaeal proteomes. We found that pseudokinases are present across all domains of life, and we classified nearly 30,000 eukaryotic, 1500 bacterial, and 20 archaeal pseudokinase sequences into 86 pseudokinase families, including ~30 families that were previously unknown. We uncovered a rich variety of pseudokinases with notable expansions not only in animals but also in plants, fungi, and bacteria, where pseudokinases have previously received cursory attention.

Annie Kwon, Steven Scott, Rahil Taujale, Wayland Yeung, Krys J. Kochut, Patrick A. Eyers, Natarajan Kannan. 2019. *Science Signaling*. 12(578)

These expansions are accompanied by domain shuffling, which suggests roles for pseudokinases in plant innate immunity, plant-fungal interactions, and bacterial signaling. Mechanistically, the ancestral kinase fold has diverged in many distinct ways through the enrichment of unique sequence motifs to generate new families of pseudokinases in which the kinase domain is repurposed for noncanonical nucleotide binding or to stabilize unique, inactive kinase conformations. We further provide a collection of annotated pseudokinase sequences in the Protein Kinase Ontology (ProKinO) as a new mineable resource for the signaling community.

Contributions:

For this project, I was involved in the design of the overall study and determine best methods to identify and classify pseudokinases. I helped design the pipeline and conducted the analyses to determine the relationships between different pseudokinases through profile comparisons and phylogenetic methods. I was specifically involved in describing the relationships and domain organization of the *R. irregularis* specific groups Rig1, Rig2 and Rig3.

A *Toxoplasma* Prolyl Hydroxylase Mediates Oxygen Stress Responses by Regulating Translation Elongation

Celia Florimond, Charlotte Cordonnier, Rahil Taujale, Hanke van der Wel, Natarajan Kannan, Christopher M. West, Ira J. Blader

As the protozoan parasite Toxoplasma gondii disseminates through its host, it responds to environmental changes by altering its gene expression, metabolism, and other processes. Oxygen is one variable environmental factor, and properly adapting to changes in oxygen levels is critical to prevent the accumulation of reactive oxygen species and other cytotoxic factors. Thus, oxygen-sensing proteins are important, and among these, 2-oxoglutarate-dependent prolyl hydroxylases are highly conserved throughout evolution. *Toxoplasma* expresses two such enzymes, TgPHYa, which regulates the SCF-ubiquitin ligase complex, and TgPHYb. To characterize TgPHYb, we created a *Toxoplasma* strain that conditionally expresses TgPHYb and report that TgPHYb is required for optimal parasite growth under normal growth conditions. However, exposing TgPHYb-depleted parasites to extracellular stress leads to severe decreases in parasite invasion, which is likely due to decreased abundance of parasite adhesins. Adhesin protein abundance is reduced in TgPHYb-

---

Celia Florimond, Charlotte Cordonnier, Rahil Taujale, Hanke van der Wel, Natarajan Kannan, Christopher M. West, Ira J. Blader. 2019. *mBio*. 10(2)

depleted parasites as a result of inactivation of the protein synthesis elongation factor eEF2 that is accompanied by decreased rates of translational elongation. In contrast to most other oxygen-sensing proteins that mediate cellular responses to low O2, TgPHYb is specifically required for parasite growth and protein synthesis at high, but not low,O2tensions as well as resistance to reactive oxygen species. *In vivo*, reduced TgPHYb expression leads to lower parasite burdens in oxygen-rich tissues. Taken together, these data identify TgPHYb as a sensor of high O2 levels, in contrast to TgPHYa, which supports the parasite at low O2.

Contributions:

My role in this project was to conduct an evolutionary analysis of PHYa, PHYb and other closely related sequences to define evolutionarily related clades that could help understand the evolution of specific function of the TgPHYa and TgPHYb proteins.

Conservation of Atypical Allostery in *C.elegans* UDP-Glucose Dehydrogenase

Nathaniel R. Beattie, Nicholas D. Keul, Tiffany N. Hicks Sirmans, Weston E. McDonald, Trevor M. Talmadge, Rahil Taujale, Natarajan Kannan, Zachary A. Wood

Human UDP-glucose dehydrogenase (hUGDH) oxidizes uridine diphosphate (UDP)-glucose to UDP-glucuronic acid, an essential substrate in the phase II metabolism of drugs. The activity of hUGDH is controlled by an atypical allosteric mechanism in which the feedback inhibitor UDP-xylose competes with the substrate for the active site and triggers a buried allosteric switch to produce an inactive complex (EΩ). Previous comparisons with a nonallosteric UGDH identified six large-to-small substitutions that produce packing defects in the protein core and provide the conformational flexibility necessary for the allosteric transition. Here, we test the hypothesis that these large-to-small substitutions form a motif that can be used to identify allosteric UGDHs. *Caenorhabditis elegans* UGDH (cUGDH) conserves this motif with the exception of an Ala-to-Pro substitution in position 109. The crystal structures of unliganded and UDP-xylose bound cUGDH show that the A109P substitution is accommodated by an Asn-to-Ser substitution at position 290. Steady-state analysis and sedimentation velocity studies show that the allosteric transition is

---

Nathaniel R. Beattie, Nicholas D. Keul, Tiffany N. Hicks Sirmans, Weston E. McDonald, Trevor M. Talmadge, Rahil Taujale, Natarajan Kannan, Zachary A. Wood. 2019. *ACS Omega*. 4:16318-16329

conserved in cUGDH. The enzyme also exhibits hysteresis in progress curves and negative cooperativity with respect to NAD+ binding. Both of these phenomena are conserved in the human enzyme, which is strong evidence that these represent fundamental features of atypical allostery in UGDH. A phylogenetic analysis of UGDH shows that the atypical allostery motif is ancient and identifies a potential transition point in the evolution of the UGDH family.

Contributions:

In this project, I was involved in conducting the phylogenetic analysis of the UGDH sequences across diverse organisms and define their evolutionary relationships.

Exploring the understudied human kinome for research and therapeutic opportunities

Nienke Moret, Changchang Liu, Benjamin M. Gyori, John A. Bachman, Albert Steppi, Rahil Taujale, Liang-Chin Huang, Clemens Hug, Matt Berginski, Shawn Gomez, Natarajan Kannan, and Peter K. Sorger

The functions of protein kinases have been heavily studied and inhibitors for many human kinases have been developed into FDA-approved therapeutics. A substantial fraction of the human kinome is nonetheless understudied. In this paper, members of the NIH Understudied Kinome Consortium mine public data on "dark" kinases to estimate the likelihood that they are functional. We start with a re-analysis of the human kinome and describe the criteria for creation of an inclusive set of 710 kinase domains and a curated set of 557 protein kinase like (PKL) domains. Nearly all PKLs are expressed in one or more CCLE cell lines and a substantial number are also essential in the Cancer Dependency Map. Dark kinases are frequently differentially expressed or mutated in The Cancer Genome Atlas and other disease databases and investigational and approved kinase inhibitors appear to inhibit them as off-target activities. Thus, it seems likely that the dark human kinome contains multiple biologically important genes, a subset of which may be viable drug targets.

Nienke Moret, Changchang Liu, Benjamin M. Gyori, John A. Bachman, Albert Steppi, Rahil Taujale, Liang-Chin Huang, Clemens Hug, Matt Berginski, Shawn Gomez, Natarajan Kannan, and Peter K. Sorger. *To be submitted.*

Contributions:

For this project, I started with the consortium provided list of 753 human proteins tagged with the term “kinase” in the UniProt database and informatically identify which of these sequences in fact were likely to be true human protein kinases. I was involved in classifying these sequences into eukaryotic protein kinases, eukaryotic-like protein kinases, atypical protein kinases and unrelated to protein kinases.

KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases

Liang-Chin Huang, Rahil Taujale, Nathan Gravel, Aarya Venkat, Wayland Yeung, Dominic P Byrne, Patrick A Eyers, and Natarajan Kannan

Protein kinases are among the largest druggable family of signaling proteins, involved in various human diseases, including cancers and neurodegenerative disorders. Despite their clinical relevance, nearly 30% of the 545 human protein kinases remain highly understudied. Comparative genomics is a powerful approach for predicting and investigating the functions of understudied kinases. However, an incomplete knowledge of kinase orthologs across fully sequenced kinomes severely limits the application of comparative approaches for illuminating understudied kinases. Here, we propose KinOrtho, a query-and graph-based orthology inference method that combines full-length and domain-based approaches to map one-to-one kinase orthologs across 17 thousand species. Using multiple metrics, we show that KinOrtho performed better than existing methods in identifying kinase orthologs across evolutionarily divergent species and eliminated potential false positives by flagging sequences without a proper kinase domain for further evaluation. We demonstrate the advantage of using domain-based

---

Liang-Chin Huang, Rahil Taujale, Nathan Gravel, Aarya Venkat, Wayland Yeung, Dominic P Byrne, Patrick A Eyers, and Natarajan Kannan. *Submitted to BMC Bioinformatics.*

approaches for identifying domain fusion events, highlighting a case between an understudied serine/threonine kinase TAOK1 and a metabolic kinase PIK3C2A with high co-expression in human cells. We also identify evolutionary fission events involving the understudied OBSCN kinase domains, further highlighting the value of domain-based orthology inference approaches. Using KinOrtho-defined orthologs, Gene Ontology annotations, and machine learning, we propose putative biological functions of several understudied kinases, including the role of TP53RK in cell cycle checkpoint(s), the involvement of TSSK3 and TSSK6 in acrosomal vesicle localization, and potential functions for the ULK4 pseudokinase in neuronal development. The well-curated kinome ortholog set can serve as a valuable resource for illuminating understudied kinases, and the KinOrtho framework can be extended to any gene-family of interest.

Contributions:

In this project, I contributed by designing the study and formulating the KinOrtho pipeline. I was involved in conducting the evolutionary analyses including building phylogenetic trees and phylogenetic profiling. I helped perform parts of the benchmark tests by downloading and running other orthology inference tools, generating finalized figures and statistics.

APPENDIX B

SUPPLEMENTARY INFORMATION

In addition to the figures, tables and data presented here in the main dissertation sections, there are some additional large tables and supplementary figures that present relevant results and observations further supporting the studies presented here. For the published eLife paper, this supplementary tables and figures can be accessed through the eLife website here: https://elifesciences.org/articles/54532/figures#content

For the GT deep learning project, the related datasets and workflows can be accessed here: https://github.com/esbgkannan/GT-CNN