Principles of Modularity in Enzyme Evolution

by

Aarya Venkat

(Under the Direction of Natarajan Kannan)

Abstract

Modularity - the concept that complex proteins can be broken down into simpler, interdependent components - manifests in enzymatic evolution and underpins their functional diversification. Glycosyltransferases and kinases, enzymes with crucial roles in cellular processes, serve as exemplars to decipher modular evolution. Detailed comparative analysis of these enzymes provides novel insights into their inherent functional plasticity, illuminating versatile allosteric mechanisms to regulate catalysis. This research further extends to enzyme engineering, demonstrating how understanding enzymatic modularity can facilitate the design of new enzymes with desired functions. Ultimately, this dissertation presents a comprehensive framework for the modular understanding of enzymatic evolution and its implications for bioengineering, which can aid in the development of novel therapeutic and biotechnological applications.

Index words: Bioinformatics, Evolution, Structure-Function, Catalytic Mechanism, Allostery, Enzyme Engineering

Principles of Modularity in Enzyme Evolution

by

Aarya Venkat

M.S., University of California, San Diego, 2017

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

2023

Principles of Modularity in Enzyme Evolution

by

Aarya Venkat

Major Professor:	Natarajan Kannan
Committee:	Kelley W. Moremen
	Robert Haltiwanger
	Robert J. Woods

Electronic Version Approved:

Ron Walcott

Dean of the Graduate School

The University of Georgia

December 2023

Dedication

To my undergraduate mentees, Claire, Cat, and Grace. I am incredibly proud of all of you and have been so lucky to be your mentor. You are all excellent scientists and I am so excited to see where you go and what you do in the future!

To my dog Freya, who has patiently listened to me rehearse my talks over and over, even providing constructive (sometimes hurtful) feedback.

Acknowledgments

I would like to thank my advisor Dr. Natarajan Kannan for his advice and support during this PhD, as well as members of my committee: Dr. Kelley Moremen, Dr. Robert Haltiwanger, and Dr. Robert Woods. I additionally thank the members and alumni of my lab for their helpful advice and contributions: Dr. Rahil Taujale, Dr. Samiksha Katiyar, Dr. Annie Kwon, Dr. Liang-Chin Huang, Dr. Wayland Yeung, Safal Shrestha, Zhongliang Zhou, Nathan Gravel, Mariah Salcedo, Brady O'Boyle, and my undergraduate mentees, Claire Bunn, Catalina Ney, and Grace Watterson.

I additionally would like to thank the Institute of Bioinformatics and the Biochemistry and Molecular Biology department, both students and faculty, for taking me in and helping me adjust to graduate student life so many years ago. In particular, I would like to thank Dr. Zachary Wood, Dr. Breeanna Urbanowicz, Dr. Erin Dolan, Dr. Lance Wells, and Dr. Christopher West for their constant support, encouragement, and advice.

Finally, I thank my parents and my sister for their endless support during this time.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background and Motivation

1.1.1 A preamble to systems biology

In the ever-evolving narrative of biology, every sub-discipline boasts its own distinct language, enabling precise communication and understanding of intricate processes. Systems biology employs the dynamic process of reading, writing, and erasing information to the functions of proteins (Meyer and Jaffrey 2017; Gillette and Hill 2015; Dedola et al. 2020). **Readers** interpret 'writing'. They recognize and bind specifically to certain post-translational modifications, such as specific carbohydrate assemblies, ensuring that the message carried by the PTM is correctly understood and responded to within the biological system. **Writers** like kinases and glycosyltransferases, 'write' or modify biomolecules by addition of a phosphate or sugar (Figure 1.1). These modifications contribute to a wide-variety of functions, from cell-signalling to triggers for the unfolded protein response to altered recognition or evasion of the immune system (A. Chakrabarti, A. W. Chen, and Varner 2011; Varki, Richard D Cummings, et al. 2022). **Erasers** remove or modify written information. Together, this process of reading, writing, and erasing, regulate cellular processes.

My focus in this dissertation involves two enzymatic writers, kinases and GTs, and how they have evolved modular elements to finely-tune this writing process to dictate biological processes.

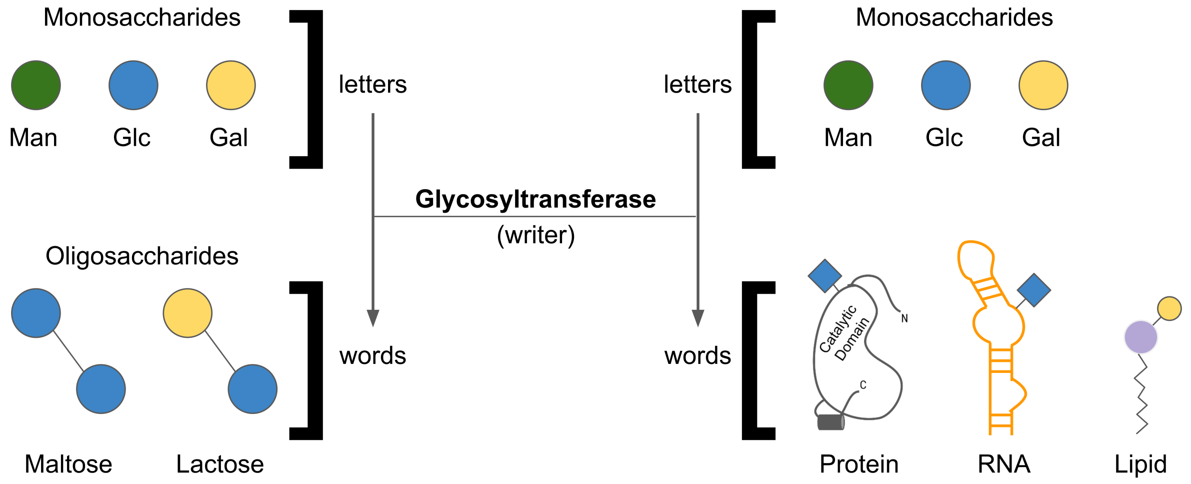


Figure 1.1: Cartoon example of the writing function performed by a glycosyltransferase. It can transfer monosaccharides to build bigger sugars or transfer them onto other biomolecules regulating cellular processes.

1.1.2 Modular evolution

Protein evolution is a multi-dimensional process shaped by various selection factors including entropy, environment, and genetics (Pál, Papp, and Lercher 2006). Together, these factors not only affect the physical characteristics of individual organisms but also contribute to broader evolutionary trends across species and lineages.

This complexity is further nuanced by the concept of modular evolution. Modular evolution suggests that proteins, and biological systems as a whole, may evolve in 'modules' or distinct units that can function independently but also interact with other modules. These modules can be as small as individual protein domains or as large as entire metabolic pathways (Moore et al. 2008). Modularity allows for more flexible adaptations, as one module can change or evolve without necessarily disrupting the function of others. Modularity in evolution thus adds another layer to the multi-dimensional landscape in which proteins evolve, enabling more targeted changes and potentially accelerating the pace of evolutionary innovation.

The modular nature of protein evolution has a profound impact on the diversity of proteins we observe. Because modules can evolve independently, it becomes possible for multiple variants of a module to coexist and combine in different ways, leading to a rich tapestry of protein functions. This is somewhat analogous to building blocks: by rearranging a set number of blocks, you can create a vast array of structures. In biological terms, this improves adaptation of an organism, allowing it to more readily handle red queen effects or changing environmental conditions (Chr and Smith 1984).

But how can we trace the origins and diversification of these modules? One powerful approach is through the study of evolutionary conservation. Modules that serve critical functions are likely to be conserved, or remain relatively unchanged, across different species (Hirsh and Fraser 2001). For example, the protein hydrophobic core is a slow-evolving conserved module around which a protein folds; it is less likely to be variant because folding and packing defects fundamentally affect overall protein stability. By comparing the sequences of conserved modules across various organisms, we can gain insights into their evolutionary history and perhaps even pinpoint the emergence of new functional modules (Echave and Wilke 2017).

Therefore, modular evolution and evolutionary conservation are interlinked concepts that together offer a comprehensive view of protein evolution. They help us understand not just how proteins change, but also how these changes are orchestrated to produce the incredible biological diversity we see today.

1.1.3 Protein hydrophobic cores

Building upon the framework of modular evolution, it is noteworthy to consider its implications at the structural level within proteins. As mentioned earlier, one of the essential structural modules within proteins is the hydrophobic core. Traditionally considered as static entities, these cores are integral in maintaining protein stability and guiding the folding process. Formed by the aggregation of hydrophobic amino acids, these cores exist in the interior of the protein, shielded from the aqueous environment, thereby minimizing the system's free energy (Kronberg 2016).

While the hydrophobic core has been extensively studied for its role in providing stability to protein structure, the traditional view of these cores as rigid, unchanging entities has been challenged (Sarina Bromberg and Ken A Dill 1994; S. Taylor et al. 2004; Lazar and Handel 1998). Particularly in protein kinases, recent studies have revealed an astonishing degree of flexibility within the hydrophobic core (S. Taylor et al. 2004; Susan S. Taylor and Alexandr P. Kornev 2011). This flexibility suggests a more dynamic role for the hydrophobic core, potentially allowing for adaptive changes in protein function or interactions.

This newfound flexibility in kinases raises intriguing questions for the broader landscape of enzymatic functions. Could similar levels of flexibility in the hydrophobic cores of other enzymes be a generalized phenomenon? Such a possibility would further emphasize the modularity and adaptability of protein structures in response to evolutionary pressures. Alterations in the hydrophobic core are not without consequences. Disruptions in this structural module, whether due to mutations or external perturbations, can significantly affect protein stability and function (Sarina Bromberg and Ken A Dill 1994; Lazar and Handel 1998). For example, mutations that destabilize the hydrophobic core can result in protein misfolding, a pathological condition implicated in a variety of diseases, including neurodegenerative disorders like Alzheimer's disease (Knowles, Vendruscolo, and Dobson 2014; Selkoe 2004).

The hydrophobic core can be conceptualized as a structural module that plays a critical role in both the stability and adaptability of proteins. This perspective not only enriches the existing understanding of protein architecture but also opens new avenues for exploring the evolution and diversity of enzymatic functions. My research explores the impact and significance of this core module on protein structure-function and allosteric regulation in two distinct enzyme superfamilies: Eukaryotic Protein Kinases (EPKs) and Fold A Glycosyltransferases (GT-As).

1.1.4 Kinase biology

EPKs are a critical family for cell signaling due to their role in phosphorylation and are thus one of the largest druggable families in the proteome (Manning et al. 2002). All kinases have a heavily conserved chassis, formed by an N-lobe and C-lobe which function to bind ATP and phosphorylate substrates (Gógl et al. 2019). Spanning these lobes is are two spines of a hydrophobic core that dynamically assemble to regulate kinase activity. These spines are called the Catalytic spine (C-spine) and Regulatory spine (R-spine) (S. Taylor et al. 2004). The C-spine is not a contiguous set of residues; it requires ATP to bind, improving core packing of the disjointed C-spine (S. Taylor et al. 2004; Susan S. Taylor and Alexandr P. Kornev 2011). This may be why kinases show increased thermostability upon ATP binding.

Kinases phosphorylate. Phosphorylation acts as a molecular switch, modulating protein functions, and consequently affecting the operational and regulatory processes within cells. By altering the structure and activity of proteins, kinases control cellular responses to environmental cues, facilitate inter- and intra-cellular communication, and maintain cellular homeostasis (Cowan and Storey 2003).

Substrate recognition by kinases is not merely sequence-based but is also influenced by the spatial and temporal distribution of both the kinase and substrate (Johnson et al. 2023; Zhou et al. 2023), ensuring highly regulated and context-dependent cellular functions. This complex interplay between kinases and substrates underscores the versatility and adaptability of cellular signaling networks, allowing cells to respond efficiently to varied physiological demands and environmental conditions. Kinases occupy a central role in cellular biology, orchestrating a myriad of cellular processes through the modification of a diverse range of substrates (Manning et al. 2002). The convergence of specificity and diversity within kinases ensures the accurate propagation of cellular signals, allowing organisms to adapt and respond to their environment effectively. The intricate regulatory mechanisms and the conserved catalytic core among the diverse kinase family highlight the evolutionary significance and the complexity of cellular signaling networks (Manning et al. 2002; S. Taylor et al. 2004).

The diversity within the kinase family is reflected not only in their substrate preferences but also in their structural configurations, regulatory mechanisms, cellular localizations, and physiological roles. Despite this diversity, a commonality among kinases is their conserved catalytic core (S. Taylor et al. 2004; Gógl et al. 2019), responsible for the binding of ATP and the transfer of the phosphate group, providing a universal mechanism for phosphorylation across different kinase classes. Because they are so well conserved, they are often differentiated by N-terminal and C-terminal segments that flank the kinase chassis. Kinases interact with a multitude of scaffold proteins, adaptor proteins, and other signaling molecules, forming multi-protein complexes that facilitate signal transduction and ensure the spatial and temporal specificity of signaling events. Previously, we published a paper on holozoan tyrosine kinases revealing how differences in the N-terminus of various kinases contribute to alterations in function and compartmentalize signaling components, allowing for the synchronized regulation of cellular processes (Yeung, Kwon, et al. 2021). It is critical to understand how a diversity of N and C-terminal modules that flank the kinase yield the functional diversity we see today. The intricate regulatory mechanisms governing kinase activity underscore their importance in maintaining cellular equilibrium.

Finally in this dissertation, we compare kinases to other enzyme families such as glycosyltransferases uncovering the convergent and divergent evolutionary paths that have shaped the family's enzymatic landscape. We reveal the shared principles between kinases and GTs, as well as unique modular adaptations that underlie the functional diversity and specificity of different enzyme classes, offering a holistic perspective on the orchestration of cellular processes and the evolutionary design of biological systems.

1.1.5 Glycosyltransferase biology

The various folds of GTs

Glycosyltransferases (GTs) are a diverse group of enzymes that play central roles in the glycosylation processes across myriad organisms. Their diversity extends not just to their substrate preferences and reaction types, but also to structural makeup. One of the intriguing aspects of GTs is their distinct structural conformations or "folds". These folds provide insights into the evolutionary pathways, functional specificities, and mechanistic operations of these enzymes. Here, we delve into the different established folds of GTs, namely Fold A, B, C, and lysozymal-type (Varki, Richard D Cummings, et al. 2022; Moremen and Haltiwanger 2019; Venkat, Tehrani, et al. 2022; Taujale, Venkat, et al. 2020; Taujale, Zhou, et al. 2021).

1. **Fold A GTs (GT-A)**:

General Structure: Fold A GTs are primarily characterized by a Rossmann-fold, a common motif in proteins that bind nucleotides. This motif consists of alternating β-strands and α-helices in a specific $\beta/\alpha/\beta$-fold pattern.

Function & Mechanism: Enzymes of this fold operate via either an inverting (S$_N$2), which use an xED-catalytic base to deprotonate the acceptor, or retaining (S$_N$i) mechanism, where the acceptor is deprotonated by the NDP-sugar β-phosphate (Varki, Richard D Cummings, et al. 2022). This action often requires the presence of a divalent metal cation, typically Manganese, bound through the DxD motif, however, some variant GT-As may no longer need to bind a metal cation, independently losing the need for the DxD. Instead, they coordinate the NDP-sugar through family-specific variations of amino acids that may mimic the divalent cation (e.g. DGK-lysine in GT116 (Amos et al. 2022) or mutation of the C-His to basic amino acids in GT14 (Taujale, Venkat, et al. 2020)).

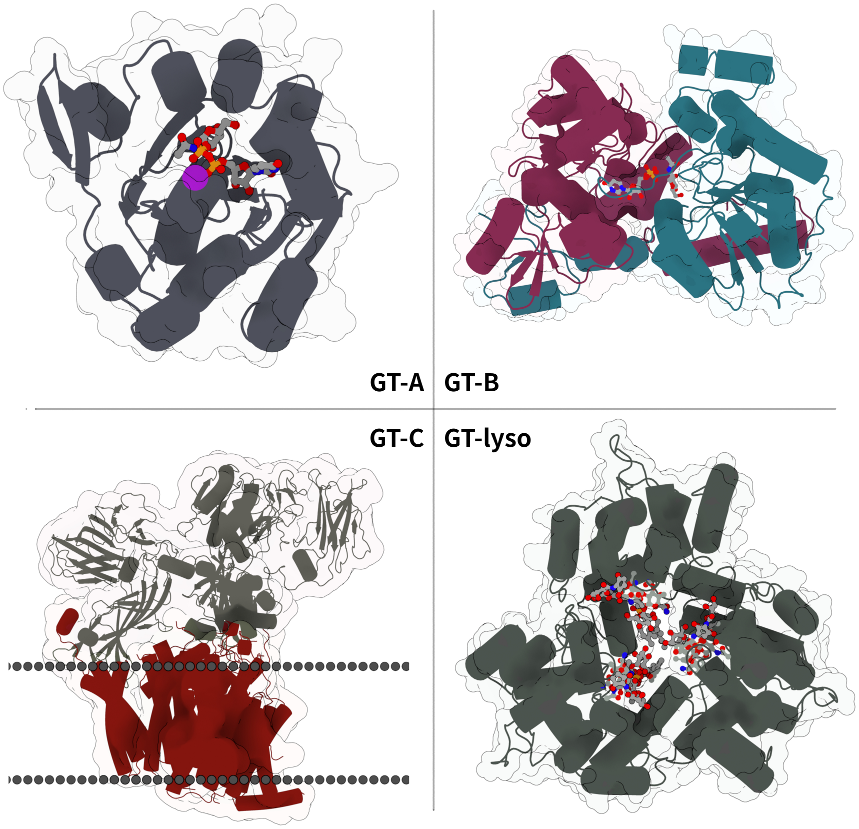


Figure 1.2: Cartoon representation of different GT folds.

2. **Fold B GTs (GT-B)**:

General Structure: Similar to GT-As, GT-Bs have dual Rossmann-fold domains. The C-terminal domain binds the nucleotide sugar donor, where the acceptor lies between the cleft of both domains.

Function & Mechanism: Like GT-As, enzymes of this fold have both inverting and retaining mechanisms. In rare cases, retaining GT-B enzymes may also use a double-displacement type mechanism instead of the front-facing (S_Ni) mechanism (Venkat, Tehrani, et al. 2022). However, GT-Bs forego the use of a DxD motif and do not use a divalent cation for catalytic activity (Varki, Richard D Cummings, et al. 2022).

3. **Fold C GTs (GT-C)**:

General Structure: This fold is less structurally represented as compared to Folds A and B. They structurally diverge significantly from these enzymes as well, with an additional massive membrane bound domain.

Function & Mechanism: Unlike GT-A and GT-B enzymes, these enzymes predominantly use lipid-linked donors (Varki, Richard D Cummings, et al. 2022).

4. **Lysozymal-type GTs (GT-lyso)**:

General Structure: These GTs share structural similarities with lysozymes, enzymes known for breaking down bacterial cell walls within the lysosome. The core structure usually incorporates a prominent β-sheet configuration.

Function & Mechanism: Akin to GT-C enzymes, these use lipid-linked donors. They have structural kinship with lysozymes and these GTs do not appear structurally cluster with any other GT families (Taujale, Zhou, et al. 2021).

The structural folds of glycosyltransferases offer a fascinating lens into their evolutionary history, functional diversity, and catalytic mechanisms. The variety in their configurations—ranging from the well-studied Rossmann-folds in Fold A GTs to newly characterized lysozyme-type GTs underscores the adaptability and versatility of these crucial enzymes. As research into GTs continues to expand, understanding these structural nuances will be paramount in decoding their intricate roles in biology and potential applications in biotechnology.

GT-As across the tree of life

Glycosyltransferases (GTs), particularly those belonging to the GT-A family, exhibit a unique and conserved structural framework. We previously delineated that most GT-As generally share 231 residues that make up this common scaffold (Taujale, Venkat, et al. 2020). Structurally, this scaffold is made up of a Rossmann-like fold characterized by a central β-sheet flanked by α-helices, creating an $\alpha/\beta/\alpha$ sandwich architecture, interspersed with highly divergent hypervariable regions, which altogether coordinate the binding specificity of a diversity of donor and acceptor substrates, as well as fine-tuning binding kinetics and kinetic efficiency for the transfer of sugars.

From bacteria to eukaryotes, GT-As exhibit a vast array of functional diversities, participating in the synthesis of a myriad of glycoconjugates essential for cellular structure, recognition, and signaling. Their presence across the tree of life denotes the evolutionary conservation and functional versatility of GT-As in mediating glycosylation reactions crucial for the survival and adaptation of organisms (Taujale, Venkat, et al. 2020). Exploring GT-As across different species reveals the evolutionary trajectories and adaptive innovations of these enzymes, offering insights into the co-evolutionary dynamics of glycosylation pathways and cellular networks. Analyzing the distribution and diversification of GT-As can unravel the evolutionary pressures and ecological contexts that have shaped the functional landscape of glycosyltransferases. Understanding how GTs relate to one another and the influence of the hydrophobic core on these relationships is essential for elucidating the evolutionary patterns and functional dynamics of these enzymes.

As mentioned previously, the hydrophobic core has been long considered a static component. However, its potential dynamic nature may be instrumental in understanding the evolutionary and functional adaptations of GTs. This flexibility might play a critical role in shaping enzyme specificity and determining functional pathways, thereby influencing the evolutionary trajectories of glycosyltransferases. An in-depth study of the evolutionary patterns of GTs and the role of their hydrophobic cores can enhance our understanding of molecular evolution. This exploration can provide a comprehensive perspective on the factors influencing the structural and functional variations among enzymes, highlighting the evolutionary mechanisms underpinning the diverse functionalities of these proteins.

1.2 Research gaps and major questions addressed

The research in this dissertation will dive into the modular evolution of glycosyltransferases and kinases through three pivotal studies. These studies aim to bridge the existing gaps in our understanding of the evolutionary and functional intricacies of these enzymes.

1.2.1 GT-A evolution and modularity of the hydrophobic core

The overarching evolution of GT-As and the structure and conservation of the GT-A hydrophobic core was first elucidated by our published work in 2020 (Taujale, Venkat, et al. 2020), where we constructed systematic sequence profiles of glycosyltransferases using bayesian sequence methods (Andrew F. Neuwald 2009; Andrew F Neuwald 2014). These methods extracted evolutionary patterns from GT-As through classification of GT-A enzyme families into distinct sets, separated by constraints unique to each set. Constraints shared across GT-As were used to identify residue features like the hydrophobic core and key motifs like the DxD motif, G-loop, xED motif, and C-His, as well as three hypervariable regions which share almost no sequence conservation between families.

The existing literature lacks comprehensive studies unraveling the intricate relationships between the evolutionary trajectories of GT-As and the dynamic nature of their hydrophobic cores. There is an unmet need for research addressing the possible functional adaptations and structural diversities emerging from the modifications in the hydrophobic core, potentially revealing novel insights into the evolution of enzymatic specificity and diversity.

My research explored the foundational structural elements of proteins, specifically hydrophobic cores, traditionally related to protein folding and stability, and their impacts on protein evolution and function within fold A glycosyltransferases (GT-As). The focus was on a large superfamily of enzymes—GT-As, renowned for catalyzing the formation of glycosidic linkages across a spectrum of donor and acceptor substrates through distinct catalytic mechanisms (inverting versus retaining).

Through the application of hidden Markov models and comprehensive protein structural alignments, this study unearthed resemblances in the phosphate-binding cassette (PBC) of GT-As and those found in unrelated nucleotide-binding proteins like UDP-sugar pyrophosphorylases. The exploration substantiates that GT-As have undergone divergent evolution from other nucleotide-binding proteins due to structural expansions of the PBC and its unique hydrophobic linkage to the F-helix, which encompasses the catalytic base (xED-Asp).

The research illustrated that while hydrophobic tethering is a conserved trait across various GT-A fold enzymes, anomalies exist, exemplified by families like B3GNT2. This study conducted meticulous experimental mutational analysis and molecular dynamics simulations to assess the structural and functional repercussions of variations in core packing and tethering interactions, discovering that specific core mutations, like T336I in B3GNT2, augment catalytic efficiency by influencing the conformational positioning of the catalytic base.

This exploration has culminated in a published paper, presenting a groundbreaking model of evolution where the GT-A core experienced progressive evolution, elaborating upon an ancient PBC, seen in various nucleotide-binding proteins. This modifiable core has been a pivotal structural platform, allowing the evolution of novel catalytic and substrate-binding functions within contemporary GT-A fold enzymes. The findings illuminate the intricate evolutionary paths and functional adaptabilities within GT-As, contributing to the enriched understanding of protein structures and functions, and offering insights that have the potential to impact future studies in protein evolution.

1.2.2 Kinase Allostery by Flanking Segments: DCLKs Introduction

Protein kinases are extraordinarily conserved enzymes involved in the dynamic orchestration of cellular signaling. This conservation begs the question of how different kinases, in spite of their highly conserved chassis, regulate activity. Recent research unveils how kinases vary flanking segments at the N-terminus resulting in evolutionary and functional diversification (Yeung, Kwon, et al. 2021). But there are still gaps in the understanding of the molecular and regulatory roles of C-terminal flanking segments in allostery and the isoform-specific differences therein. Dual specificity protein kinase (DCLK) family members, characterized by their diverse isoforms and regulatory modules, offer a promising avenue to explore these gaps. Detailed studies focusing on the isoform-specific modules from alternative splicing and their impact on kinase function in DCLKs are currently scarce but critical for understanding the functional intricacies and regulatory diversity of kinases.

DCLKs have intrigued the scientific community, due to their unique microtubule-associated properties and, notably, their distinct C-terminal tail segments, which appear to be involved in autoregulation of function. In our study, we hone in on this C-tail variability, driven by alternative splicing events, to unravel its implications on the holistic functional mechanism of DCLK kinases.

With the varying tail lengths across DCLK isoforms, we were curious about the molecular basis of how these changes affect kinase autoregulation. To tackle this conundrum, we employed a series of approaches: statistical sequence analysis, molecular dynamics simulations, and meticulous in vitro mutational analysis. These methodologies, combined, shed light on the evolutionary intricacies within the DCLK family. We identified, within the DCLK1 sub-family, distinct splice variants that utilize alternative codons, enhancing the inhibitory capacity of the DCLK1 C-tail.

Moreover, our investigations have delineated specific co-conserved motifs, which not only demarcate DCLKs from the broader Calcium Calmodulin Kinase (CAMK) ensemble but also emphasize the assembly of pivotal motifs anchoring the C-terminal tail for auto-regulatory purposes.

When the C-terminal tail undergoes structural modifications, the ramifications span protein stability, nucleotide/inhibitor-binding affinity, and enzymatic activity, suggesting a complex, isoform-driven regulatory matrix.

This work provides a scaffold for dissecting kinome regulation, with a spotlight on the regulatory mechanisms of intrinsically disordered regions. These insights, I believe, will be instrumental in guiding the future design of DCLK1 modulators, with therapeutic potential that may very well reshape our current landscape of kinase-targeted interventions.

1.2.3 AI for Glycosyltransferases

Accurate prediction of glycosyltransferase specificity remains a complex challenge, due to the inherent complexity of expression and purification of GTs, as well as a lack of structural data availability. As previously described, GTs are exhibit great functional diversity. Given their pivotal role in various metabolic and signaling pathways, there exists a critical need to predict and classify GT function to aid in hypothesis generation and testing for glycobiologists.

With the advent of deep learning and protein language models, how can advanced computational tools be developed to classify glycosyltransferase function in a robust and accessible manner? We introduce "Glydentify," an advanced tool designed for classification of glycosyltransferase function. Distinct from traditional sequence classification or conventional machine learning methods, Glydentify leverages state-of-the-art protein language models, specifically ESM2. This enables the tool to extract high-dimensional sequence embeddings, providing a rich dataset for accurate classification. The results indicate that Glydentify can classify GT families with a confidence level of 92% and also predict potential donor binding with an 89% confidence, using input fasta sequences.

Furthermore, the utility of Glydentify extends beyond its computational capabilities. Integrating Gradio ensures the tool provides an intuitive interface, eliminating the need for extensive programming knowledge. This design choice ensures broader accessibility to researchers, facilitating the application of cutting-edge machine learning in diverse research settings.

Available as an open-source tool on GitHub and also accessible via web browser (https://huggingface.co/spaces/arikat/Glydentify), Glydentify aims to drive advancements in GT research. By combining sophisticated computational methodologies to serve practical research needs, it provides a foundational platform for deeper insights into glycosyltransferase functional diversities and evolutionary patterns.

References

Bromberg, Sarina and Ken A Dill (1994). "Side-chain entropy and packing in proteins". *protein Science* 3.7, pp. 997–1009.

Chakrabarti, Anirikh, Aaron W Chen, and Jeffrey D Varner (2011). "A review of the mammalian unfolded protein response". *Biotechnology and bioengineering* 108.12, pp. 2777–2793.

Chr, Nils and J Maynard Smith (1984). "Coevolution in ecosystems: Red Queen evolution or stasis?" *Evolution*, pp. 870–880.

Cowan, Kyra J and Kenneth B Storey (2003). "Mitogen-activated protein kinases: new signaling pathways functioning in cellular responses to environmental stress". *Journal of Experimental Biology* 206.7, pp. 1107–1115.

Dedola, Simone et al. (2020). "Revisiting the language of glycoscience: readers, writers and erasers in carbohydrate biochemistry". *ChemBioChem* 21.3, pp. 423–427.

Echave, Julian and Claus O Wilke (2017). "Biophysical models of protein evolution: understanding the patterns of evolutionary sequence divergence". *Annual review of biophysics* 46, pp. 85–103.

Gillette, Thomas G and Joseph A Hill (2015). "Readers, writers, and erasers: chromatin as the whiteboard of heart disease". *Circulation research* 116.7, pp. 1245–1253.

Gógl, Gergő et al. (Apr. 2019). "Disordered Protein Kinase Regions in Regulation of Kinase Domain Cores". eng. *Trends in Biochemical Sciences* 44.4, pp. 300–311. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2018.12.002.

Hirsh, Aaron E and Hunter B Fraser (2001). "Protein dispensability and rate of evolution". *Nature* 411.6841, pp. 1046–1049.

Johnson, Jared L et al. (2023). “An atlas of substrate specificities for the human serine/threonine kinome”. *Nature* 613.7945, pp. 759–766.

Knowles, Tuomas PJ, Michele Vendruscolo, and Christopher M Dobson (2014). “The amyloid state and its association with protein misfolding diseases”. *Nature reviews Molecular cell biology* 15.6, pp. 384–396.

Kronberg, Bengt (2016). “The hydrophobic effect”. *Current Opinion in Colloid & Interface Science* 22, pp. 14–22.

Lazar, Greg A and Tracy M Handel (1998). “Hydrophobic core packing and protein design”. *Current Opinion in Chemical Biology* 2.6, pp. 675–679.

Manning, G. et al. (Dec. 2002). “The protein kinase complement of the human genome”. eng. *Science (New York, N.Y.)* 298.5600, pp. 1912–1934. ISSN: 1095-9203. DOI: 10.1126/science.1075762.

Meyer, Kate D and Samie R Jaffrey (2017). “Rethinking m6A readers, writers, and erasers”. *Annual review of cell and developmental biology* 33, pp. 319–342.

Moore, Andrew D et al. (2008). “Arrangements in the modular evolution of proteins”. *Trends in biochemical sciences* 33.9, pp. 444–451.

Moremen, Kelley W and Robert S Haltiwanger (2019). “Emerging structural insights into glycosyltransferase-mediated synthesis of glycans”. *Nature chemical biology* 15.9, pp. 853–864.

Neuwald, Andrew F (2014). “A Bayesian sampler for optimization of protein domain hierarchies”. *Journal of Computational Biology* 21.3, pp. 269–286.

— (Aug. 2009). “Rapid detection, classification and accurate alignment of up to a million or more related protein sequences”. eng. *Bioinformatics (Oxford, England)* 25.15, pp. 1869–1875. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp342.

Pál, Csaba, Balázs Papp, and Martin J Lercher (2006). “An integrated view of protein evolution”. *Nature reviews genetics* 7.5, pp. 337–348.

Selkoe, Dennis J (2004). "Cell biology of protein misfolding: the examples of Alzheimer's and Parkinson's diseases". *Nature cell biology* 6.11, pp. 1054–1061.

Taujale, Rahil, Saber Soleymani, et al. (2021). "GTXplorer: A portal to navigate and visualize the evolutionary information encoded in fold A glycosyltransferases". *Glycobiology* 31.11, pp. 1472–1477.

Taujale, Rahil, Aarya Venkat, et al. (2020). "Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases". *Elife* 9, e54532.

Taujale, Rahil, Zhongliang Zhou, et al. (2021). "Mapping the glycosyltransferase fold landscape using interpretable deep learning". *Nature Communications* 12.1, p. 5656.

Taylor, SS et al. (2004). "PKA: a portrait of protein kinase dynamics". *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics* 1697.1-2, pp. 259–269.

Varki, Ajit, Richard D Cummings, et al. (2022). "Essentials of Glycobiology [internet]".

Venkat, Aarya, Daniel Tehrani, et al. (2022). "Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases". *Journal of Biological Chemistry* 298.8.

Yeung, Wayland, Annie Kwon, et al. (Dec. 2021). "Evolution of Functional Diversity in the Holozoan Tyrosine Kinome". eng. *Molecular Biology and Evolution* 38.12, pp. 5625–5639. ISSN: 1537-1719. DOI: 10.1093/molbev/msab272.

Zhou, Zhongliang et al. (2023). "Phosformer: an explainable transformer model for protein kinase-specific phosphorylation predictions". *Bioinformatics* 39.2, btad046.

Chapter 2

Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases

Venkat, Aarya, et al. "Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases." Journal of Biological Chemistry 298.8 (2022).
Reprinted here with permission of the publisher.

2.1 Abstract

Hydrophobic cores are fundamental structural properties of proteins typically associated with protein folding and stability; however, how the hydrophobic core shapes protein evolution and function is poorly understood. Here, we investigated the role of conserved hydrophobic cores in fold-A glycosyltransferases (GT-As), a large superfamily of enzymes that catalyze formation of glycosidic linkages between diverse donor and acceptor substrates through distinct catalytic mechanisms (inverting versus retaining). Using hidden Markov models and protein structural alignments, we identify similarities in the phosphate-binding cassette (PBC) of GT-As and unrelated nucleotide-binding proteins, such as UDP-sugar pyrophosphorylases. We demonstrate that GT-As have diverged from other nucleotide-binding proteins through structural elaboration of the PBC and its unique hydrophobic tethering to the F-helix, which harbors the catalytic base (xED-Asp). While the hydrophobic tethering is conserved across diverse GT-A fold enzymes, some families, such as B3GNT2, display variations in tethering interactions and core packing. We evaluated the structural and functional impact of these core variations through experimental mutational analysis and molecular dynamics simulations and find that some of the core mutations (T336I in B3GNT2) increase catalytic efficiency by modulating the conformational occupancy of the catalytic base between "D-in" and acceptor-accessible "D-out" conformation. Taken together, our studies support a model of evolution in which the GT-A core evolved progressively through elaboration upon an ancient PBC found in diverse nucleotide-binding proteins, and malleability of this core provided the structural framework for evolving new catalytic and substrate-binding functions in extant GT-A fold enzymes.

Keywords

glycobiology – glycosyltransferase – protein evolution – structure–function – bioinformatics

2.2 Introduction

Glycosyltransferases (GTs) are a diverse family of enzymes that catalyze the formation of glycosidic linkages between sugars and other macromolecules (Varki, Richard D. Cummings, et al. 2015). These enzymes are found across the tree of life and are involved in a number of critical cellular functions through post-translational modifications, including protein folding, signaling, and stability (Varki, Richard D. Cummings, et al. 2015).

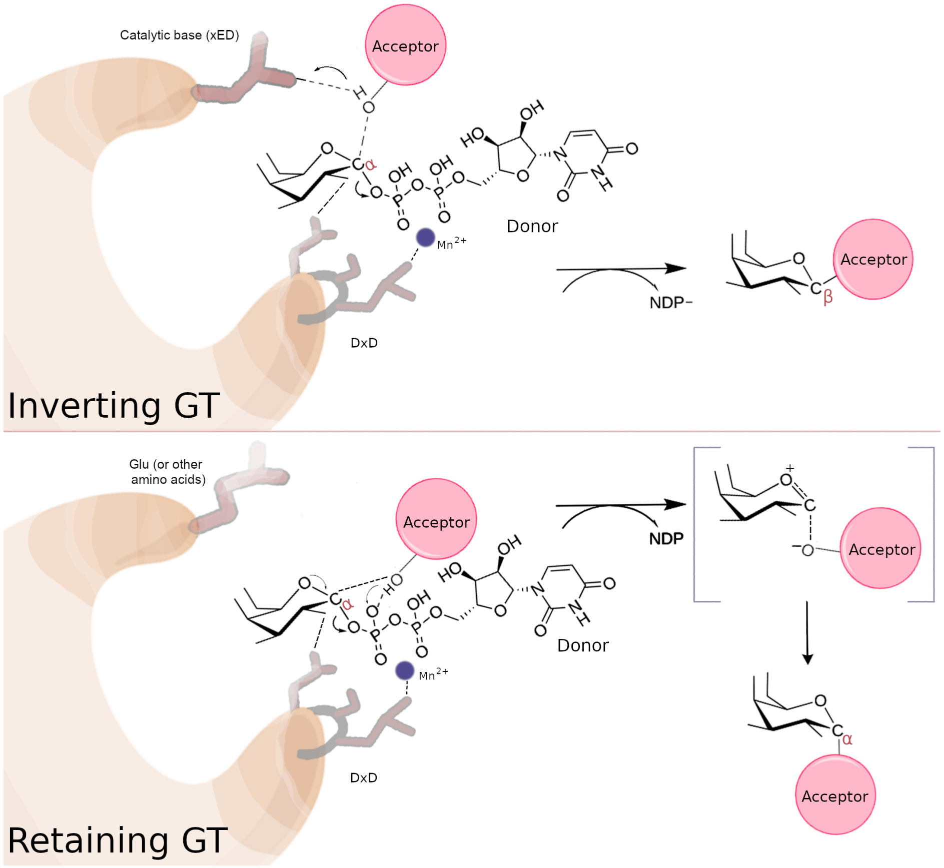


Figure 2.1: Cartoon Mechanism

Misregulation, or aberrant glycosylation, is implicated in a wide range of diseases, including Alzheimer's, Parkinson's, muscular dystrophies, and human cancers (Agrawal et al. 2017; Chugh et al. 2015; Grewal et al. 2001; Kitazume, Saido, and Hashimoto 2004; Moll, Shaw, and Cooper-Knock 2020; Yoshida et al. 2001). Based on the catalytic mechanism, GTs are broadly classified as "inverting" or "retaining" based on the stereochemistry of the glycosidic bond they generate (Figure 2.1). Inverting GTs generally employ a direct S_N2 displacement mechanism with a protein-associated catalytic base that deprotonates the acceptor nucleophile hydroxyl leading to attack on the anomeric center and displacement of the nucleotide diphosphate–leaving group. By contrast, retaining GTs do not use an enzyme side chain as catalytic base but instead are generally considered to employ a same-side S_Ni-type mechanism where the acceptor hydroxyl nucleophile is deprotonated by the donor β-phosphate oxygen and attacks the anomeric carbon atom of the donor sugar from the same side as the leaving nucleotide (Moremen and Haltiwanger 2019). While there are rare examples of unusual GTs that presumably employ a double-displacement mechanism (Kimber et al. 2020; Ovchinnikova et al. 2016), in general, the differences in catalytic machinery between inverting and retaining GTs are the location and use of a catalytic base in acceptor deprotonation and the location of the acceptor nucleophile hydroxyl relative to the nucleotide sugar donor (Moremen and Haltiwanger 2019).

Independent of the catalytic mechanism, GTs can be classified into one of four major folds (A, B, C, and lyso) (Varki, Richard D. Cummings, et al. 2015; Moremen and Haltiwanger 2019; Taujale, Zhou, et al. 2021) or variants of known folds (Taujale, Zhou, et al. 2021) based on primary sequence similarity and 3D topology. A vast majority of GTs fall within the GT-A fold, which is characterized by the Rossmann fold–like $\alpha/\beta/\alpha$ sandwich topology adopted by a diverse class of nucleotide-binding proteins unrelated to GTs (Varki, Richard D. Cummings, et al. 2015; Breton et al. 2006), but the structural basis for how GTs evolutionarily diverged from other Rossmann fold proteins is not known. We recently reported a deep evolutionary classification of GT-A fold sequences into 53 (sub)families that broadly fall into nine different clades and identified the core structural features shared among diverse GT-A fold enzymes (Taujale, Venkat, et al. 2020; Kadirvelraj et al. 2021). These core features include two motifs (DxD and xED) involved in catalytic functions as well as an extended network of hydrophobic residues connecting the catalytic and nucleotide-binding sites. While a majority of these conserved hydrophobic residues are present in other Rossmann fold enzymes, a subset of GT-A families such as GT6 and GT8, the GT-A–specific residues are frequently mutated in cancer subtypes (Table S1). However, the structural and functional roles of these natural and disease variations in the core are largely unknown.

Nearly all folded proteins are characterized by hydrophobic residues in the core that contribute to protein folding and stability (Baldwin and Matthews 1994; Maxwell and Davidson 1998; Szilágyi and Závodszky 2000) While most protein cores are optimally packed, in many regulatory and signaling proteins, the core packing is nonoptimal resembling a "nuts-and-bolts" in a jar model (S. Bromberg and K. A. Dill 1994), in which some core residues are rigid, whereas others are flexible. The overall fitness of a hydrophobic core is determined by energetic favorability of packing interactions (J. Chen and Stites 2001), and packing efficiency has been correlated with protein dynamics and allosteric functions (Bhardwaj and Gerstein 2009; Ben-David et al. 2019). The nonoptimal packing of the core provides a selective advantage in some proteins, such as protein kinases, which are dynamically assembled during regulation of catalysis. Protein kinases contain an extended hydrophobic network connecting the ATP and substrate-binding lobes, termed the "spines," which are dynamically assembled during kinase activation (Alexandr P. Kornev and Susan S. Taylor 2010) and the suboptimal packing of the spine residues enable dynamic regulation of catalytic activity (J. Chen and Stites 2001; Susan S. Taylor and Alexandr P. Kornev 2011; J. Kim et al. 2017). Indeed, malleable cores have been implicated in allosteric regulation or inhibition in other enzyme families as well (Hardy et al. 2004; Horn and Shoichet 2004; Mei et al. 2018), but the role of conserved core in GT-A evolution and function has not been systematically investigated.

Here using a combination of structural bioinformatics and experimental studies, we investigate the role of conserved hydrophobic core in GT-A structure, function, and evolution. Based on the identification of an ancient phosphate-binding cassette (PBC; (Longo et al. 2020), Fig. 1) shared by GT-As and other nucleotide-binding proteins, we dissect the hydrophobic core of GT-A enzymes into three categories: residues shared among PBC-containing enzymes, residues shared by

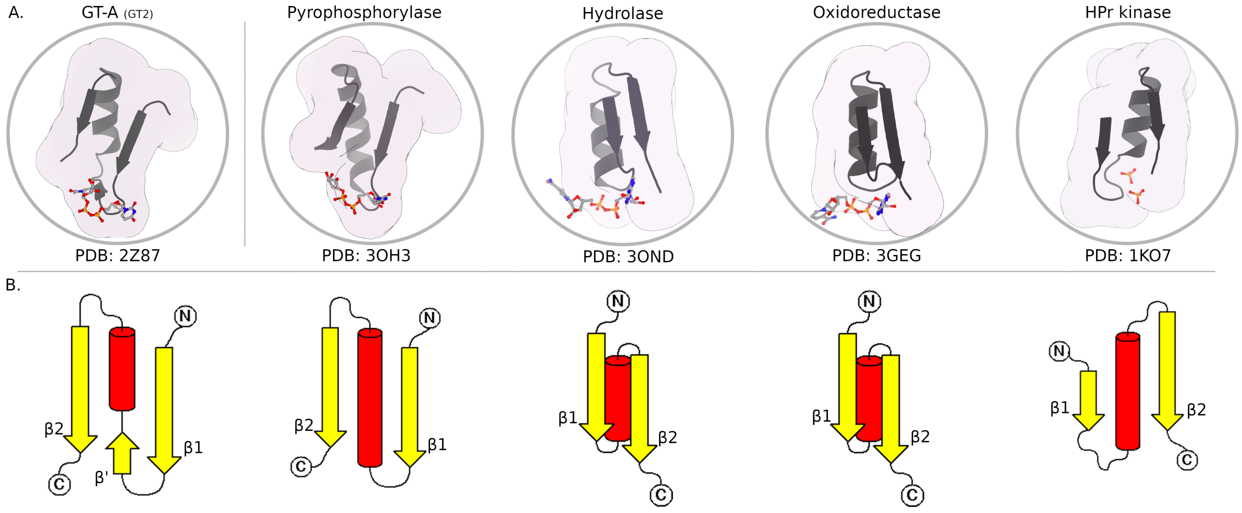


Figure 2.2: Structural comparison of the PBC in selected enzyme superfamilies. A) cartoon representations of different enzyme superfamilies with a GT-A structure at the left, demonstrating superfamily specific variations to a shared ancestral β-α-β phosphate-binding region. B) comparison of a subset of GT-A, Rossmann fold, and P-loop NTPase PBC topologies as cartoons to show how GT-As structurally differ from most other Rossmann fold enzymes. Many topologies exist to bind the phospho-nucleotide ligand.

Rossmann fold proteins, and residues unique to the GT-A core. We perform an in-depth structural analysis of the GT core–specific residues (residues 156 and 183) connecting the PBC and the αF-helix and find a strong correlation between hydrophobic packing and catalytic mechanism (inverting versus retaining). We propose that a dynamic GT-A core provides a selective advantage by enabling new modes of donor- and acceptor-binding functions. Our studies support a model in which the GT-A core evolved progressively through elaboration of an ancient PBC found in diverse nucleotide phosphate–binding proteins. Implications of our findings in the synthetic design of GTs and characterization of oncogenic mutations mapping to the core are discussed.

2.3 Results

2.3.1 Delineation of the PBC and modular evolution of the GT-A hydrophobic core

Recently, an ancestral PBC shared among P-loop NTPases and Rossmann fold enzymes was reported (Longo et al. 2020). This includes several major enzyme superfamilies, such as pyrophosphorylases, oxidoreductases, epimerases, and hydrolases. Now, based on further structural comparisons (see the Experimental procedures section), we extend the presence of this ancestral PBC to GT-As (Fig. 1). We used hidden Markov models (HMMs) from previously published PBC themes (Longo et al. 2020), which produced significant hits to the PBC of GT-As. Different enzyme families have variable structural topologies of the PBC (28). By performing an all-versus-all structural comparison of a representative set of these different PBCs, we identify clusters of PBCs that further support structural and functional similarities between the GT-A PBC and NDP-sugar pyrophosphorylases (Fig. S4). GT-A PBCs closely resemble that of Rossmann pyrophosphorylases in terms of overall topology. Notably, both pyrophosphorylases and GT-As consistently use metal ions to bind the dinucleotide phosphate. Specifically, UDP-sugar pyrophosphorylases bind a UTP donor and sugar-1-phosphate acceptor and catalyze the formation of a UDP-sugar substrate, which is used as a donor substrate for both GT-A and GT-B fold enzymes (Varki, Richard D. Cummings, et al. 2015). Structural alignment of the PBCs (using Protein Data Bank [PDB] IDs: 3OH3 and 2Z87) reveals similar PBC topologies for cofactor and nucleotide binding in these two enzymes (Figs. S2–S5). Matching homology from the HMM analysis and the structural alignment suggest a shared ancestry between these two protein families, although the possibility of convergent evolution of a common phosphate-binding mode cannot be ruled out.

GT-A PBCs differ from most other Rossmann fold enzymes and P-loop NTPases by flipping the topological orientation and replacing the glycine-rich loop (located between the β1 sheet and

$\alpha 1$ helix) with an additional pseudo beta bridge (β'), shifting the binding site for both the ligand and divalent cation (Figs. 1B and S2). Likewise, elaboration of the loop connecting $\beta 1$ and $\alpha 1$ helix in GT-A through insertion of the metal coordinating DxD motif further contributes to structural and functional divergence of GT-A PBC from other Rossmann enzymes (Figs. 1B and S2).

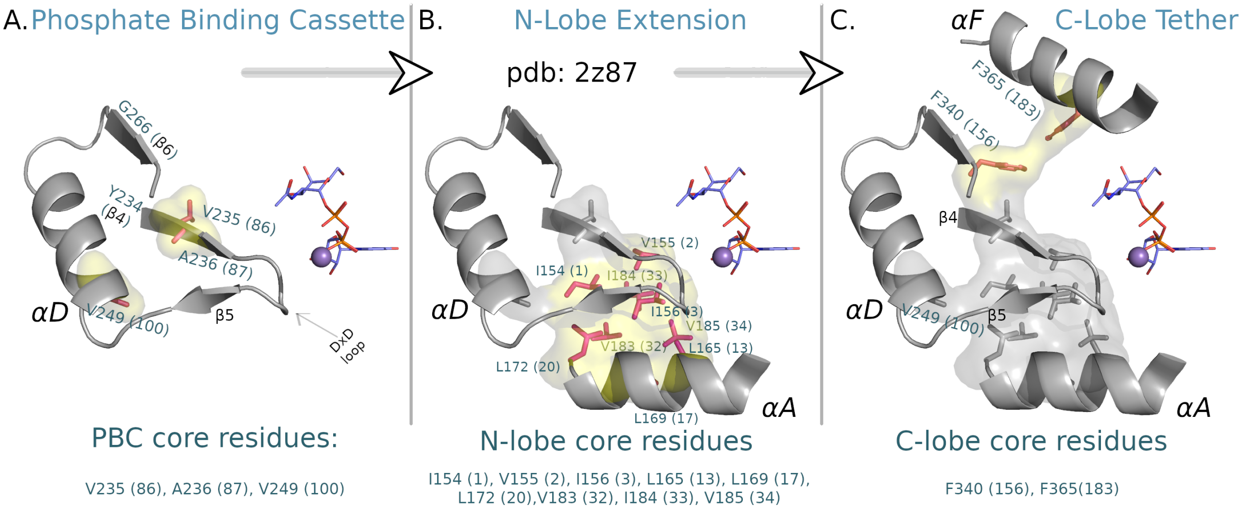


Figure 2.3: The GT-A hydrophobic core is separable into three modules over evolutionary time. A, structural depiction of the ancestral phosphate-binding cassette (PBC) in GT2 (Protein Data Bank ID: 2Z87), which contains three of the hydrophobic residues of the GT-A core (surface representation). B and C, extension of the hydrophobic core from the PBC, showing the insertion of an N-lobe core, common to all Rossmann fold enzymes, and a GT-A specific C-lobe tether which connects the αF-helix to the PBC.

In GT-As, the PBC corresponds to $\beta 4$, αD, and $\beta 6$ (residues Y234 to G266 in PBC; Fig. 2A) containing the classic metal-binding DxD motif and a miniature hydrophobic core (Fig. 2A). Delineation of the PBC allows us to further dissect the anatomy of the GT-A core into three hierarchical categories based on the depth of conservation of hydrophobic residues. We denote these residues based on the GT2 structure (PDB ID: 2Z87) and the consensus alignment numbering published in a previous study (alignment position indicated parenthetically). Residues present in the PBC include V235 (86), A236 (87), and V249 (100) (Figs. 2A and S6). Residues shared by Rossmann fold enzymes include I154 (1), V155 (2), I156 (3), L165 (13), L169 (17), L172 (20), V183 (32), I184 (33), V185 (34), V235 (86), and A236 (87) (Figs. 2B and S6); and residues unique to GT-A fold enzymes include V249 (100), F340 (156), and F365 (183) (Figs. 2C and S6). Hydrophobic residues shared by Rossmann fold enzymes tether the PBC to the N-lobe (αA-helix), whereas residues unique to GT-A fold enzymes tether the PBC to the αF-helix in the C-lobe. In particular, the GT-A–specific hydrophobic residue in the F-helix (F365; position 183 in Fig. 2C) mediate a van der Waals interaction with hydrophobic residues in the PBC (F340 position in Fig. 2C) and a backbone hydrogen bond with the catalytic xED-Asp. Because the C-lobe tethering of the PBC is unique to GT-As and represent the most recent addition in GT-A core evolution, we focus on the C-lobe tethering interaction (F340 and F365) in the following sections.

2.3.2 GT-A–specific extension of the ancestral core is malleable and contributes to conformational flexibility, acceptor recognition, and catalysis

We performed a detailed analysis of the structural interactions mediated by tether residues (at positions 156 [F340] and 183 [F365]) in representative crystal structures to investigate their role in GT-A fold structure. Analysis of the contact distances between these residues indicates significant variability in side-chain contact distances (ranging from 4 to 14 Å) across diverse GT-A enzymes. Further analysis of these distances in inverting and retaining enzymes revealed strong correlation between contact distance and catalytic mechanism (p = 1.61E-13, using a two-tailed t test) (Fig. 3A, Supp File 156-183dist).

In inverting GT-As, the hydrophobic contact distance between 156 and 183 is in the range of 4 to 7 Å, whereas in the majority of retaining GT-As, the median distance between these residues increases significantly, with a normalized maxima around 10 Å. Retaining GT-As form a bimodal distribution, where several retaining GT-As have a contact distance between 4 and 7 Å. We observe these retaining GT-As to appear in clades containing previously phylogenetically

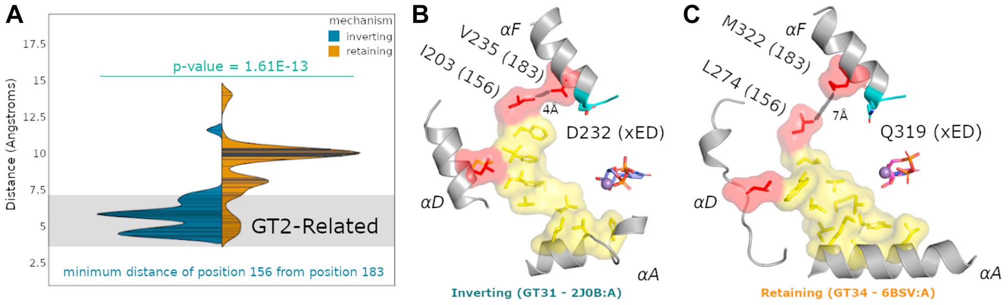


Figure 2.4: Structural conservation and variability in the C-lobe tether. A, Violin plot of representative GT-A Protein Data Bank structures, separated by mechanism, measuring the minimum distance from hydrophobic core positions 156 and 183, with a line of fit for histogram density showing significant separation between retaining and inverting GT-As (p = 1.61E-13). The gray bar indicates the range for a typical hydrophobic contact. Retaining GT-As show a higher variation than inverting GT-As for this region, with most retaining GT-As having a minimum distance between 9 and 10 Å, greater than a hydrophobic contact. Inversely, most inverting GT-As appear to maintain a contact distance of ≈3 to 6 Å, within contact distance. B and C, structural differences between retaining and inverting GTs, using two representative GT-A structures reveal a separation in most retaining GTs that appears to extend the size of the hydrophobic core. Core residues in yellow are conserved across all Rossmann fold enzymes, whereas red residues are GT-A specific. Where most inverting cores (blue) can directly make contacts in the tether, many retaining GTs have a gap between these conserved residues from packing defects.

classified subfamilies (Breton et al. 2006) of the large GT2 CAZy family, thus we term these as "GT2 related" (Figs. 3A and S7). GT2s are more primordial (Breton et al. 2006), and as such, we note that retaining enzymes related to GT2 have largely maintained a spacing consistent with the more constrained inverting enzymes. More distant retaining GT-As appear to have a less tightly packed C-lobe tether (Fig. 3, B and C).

While the catalytic base (xED-Asp) is conserved in inverting GTs, in retaining enzymes, the xED-Asp is often replaced by a glutamine or a glutamate, which shifts the site of catalysis by >2 Å (**moremen_emerging_2019**), preventing it from being used as a catalytic base. Instead of the xED motif, retaining GTs use the β-phosphate oxygen of the UDP-sugar donor as a catalytic base and perform a dissociative SNi-type reaction mechanism (**moremen_emerging_2019**). To determine whether the loss of constraint on the xED-Asp in retaining enzymes correlates with packing in the C-lobe tether, we analyzed the nature of residues surrounding the tether in primary sequences and 3D structures (Fig. 4). Comparisons of inverting and retaining GTs indicate differences in both xED-Asp position as well as residues involved in C-lobe tether (Fig. 4A). We further compare core packing interactions between representative GT-A crystal structures, and note that the retaining GT-As have a less tightly packed tether because of a substitution of a flexible methionine (M322) by a valine (V235), which alters core packing (Figs. 4, B and C and S3). In a subset of GTs, such as GT15, the hydrophobic tether is replaced by a salt bridge interaction (Fig. S9). Likewise, in B3GNT2 (GT31), a conserved water molecule is involved in the tethering interaction (Fig. S9E). The structural and functional implications of these family specific variations are discussed later.

2.3.3 B3GNT2-specific variations in the C-lobe tether contribute to catalytic activity, stability, and dynamics

We next investigated the structural and functional implications of B3GNT2-specific variation in the C-lobe tether. In B3GNT2 crystal structures, the threonine (T336) side chain forms van der Waals interactions with hydrophobic residues (F156) in the phosphate-binding module to maintain the C-lobe tether. Also, the small size of the threonine side chain creates internal cavities that are occupied by a water molecule, which coordinate with the hydroxyl group of T336 side chain as well as the xED-Asp. To investigate the structural and functional implications of these B3GNT2-specific variations, we performed a computational and experimental screen of different variants at position 183 (T336). A computational screen using Rosetta predicted a subset of stabilizing and destabilizing mutations (Fig. 5A).

With these predicted sets of stabilizing and destabilizing mutations, we then experimentally expressed a subset of single and double mutants (F309W, T336I, Y311I/T336V, Y311F/T336I, Y311F/T336Y, and Y311F/T336V) through recombinant expression in human embryonic kidney 293 cells (Kadirvelraj et al. 2021). All the generated mutants expressed at detectable levels and

A.

Inverting

154	%	155	%	156	%	178	%	179	%	180	%	181	%	182	%	183	%
G	20.5%	A	16.7%	L	15.8%	G	10.8%	E	23.4%	D	39.0%	D	25.7%	D	20.3%	L	20.1%
F	11.5%	L	14.3%	F	15.1%	A	10.4%	D	10.2%	E	18.5%	E	9.6%	E	15.0%	F	11.8%
S	7.3%	V	8.7%	V	12.5%	V	8.0%	F	7.3%	G	4.8%	V	8.2%	V	10.2%	D	7.3%
Y	7.3%	G	8.5%	I	10.2%	L	6.8%	P	6.1%	L	4.3%	M	7.5%	W	6.1%	I	7.1%
T	6.1%	I	7.7%	Y	8.2%	S	6.4%	S	5.7%	S	4.3%	L	5.7%	L	6.0%	E	6.9%
V	5.1%	Y	7.7%	M	6.3%	Y	6.3%	W	5.7%	R	3.6%	S	4.6%	F	5.5%	M	6.7%
A	4.9%	F	7.3%	A	6.0%	P	5.9%	G	5.0%	H	3.5%	A	4.5%	I	4.6%	V	6.2%
L	4.7%	M	6.3%	W	4.9%	F	5.4%	T	4.5%	Y	3.1%	T	4.3%	Y	4.3%	T	5.1%
C	4.5%	C	5.6%	S	4.7%	D	5.2%	A	4.2%	P	2.8%	I	4.1%	R	3.7%	R	4.7%
W	4.4%	S	5.4%	G	4.6%	T	5.0%	L	4.0%	I	2.6%	G	3.6%	T	3.7%	Y	4.4%

Retaining

154	%	155	%	156	%	178	%	179	%	180	%	181	%	182	%	183	%
L	14.0%	F	18.0%	V	15.3%	G	20.3%	D	18.5%	D	18.2%	D	19.5%	D	12.2%	L	15.0%
F	13.3%	L	16.3%	F	13.2%	L	7.9%	G	15.4%	E	17.5%	Q	12.5%	L	11.1%	I	10.5%
G	12.6%	V	13.8%	G	10.8%	D	7.2%	E	12.1%	Q	13.1%	N	8.8%	A	10.5%	E	8.8%
V	7.0%	G	9.9%	L	10.1%	A	6.9%	A	4.7%	A	7.4%	E	8.4%	E	9.1%	D	7.8%
Y	7.0%	I	8.1%	I	8.0%	I	5.5%	F	4.7%	G	5.7%	S	8.4%	Q	8.8%	V	7.1%
A	6.7%	T	6.7%	A	7.0%	V	5.2%	K	4.7%	T	5.1%	L	5.4%	G	7.8%	P	6.8%
I	5.6%	A	4.2%	Y	6.6%	N	4.8%	N	4.7%	N	4.4%	I	5.1%	V	7.4%	F	6.1%
M	5.3%	M	3.2%	M	6.3%	S	4.8%	I	4.4%	V	4.0%	P	5.1%	I	6.4%	G	5.8%
N	4.6%	Y	2.8%	T	5.6%	W	4.8%	L	4.0%	S	3.4%	G	4.7%	H	5.7%	A	5.4%
S	4.2%	Q	2.5%	S	3.8%	H	4.5%	V	3.4%	L	2.7%	A	3.7%	P	5.7%	M	4.8%

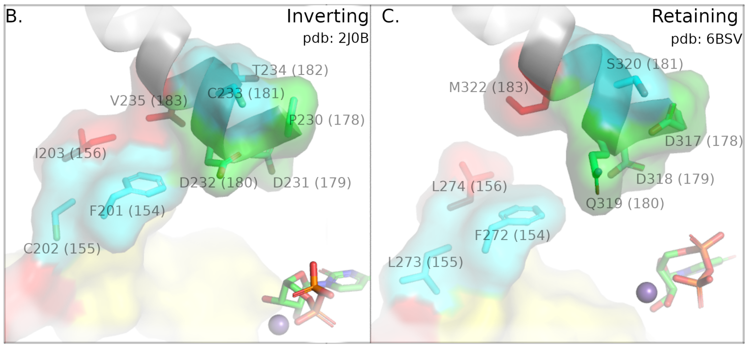
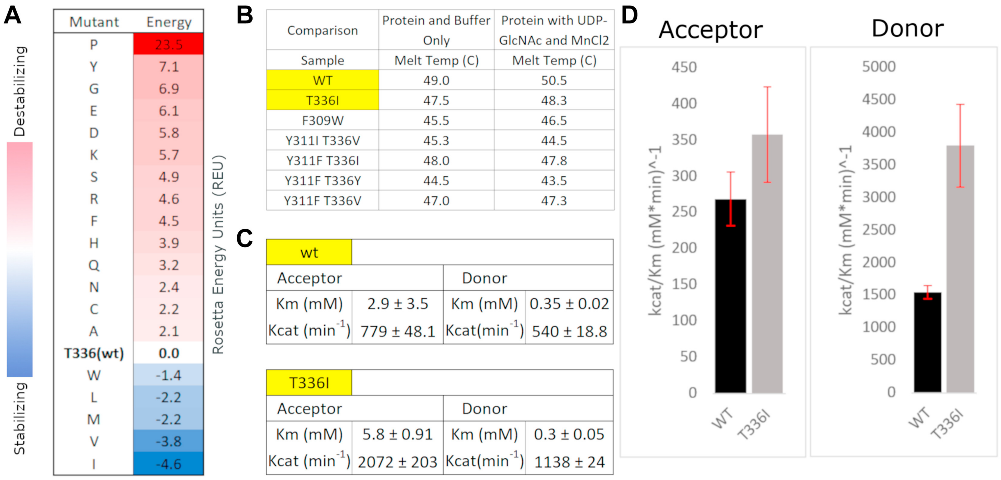


Figure 2.5: Amino acid preferences in the C-lobe tether of inverting and retaining enzymes. A, array of the top ten residue frequencies from a sequence alignment of inverting and retaining GTs, showing higher conservation and constraints in the C-terminal tether (156, 183) and xED-Asp (180) in inverting GT-As. A full table of these residue frequencies is shown in Table S4. B and C, a comparison of representative inverting and retaining GT-A core packing in the same orientation, showing that the retaining pocket is less packed, as compared with inverting GT-As. The xED is highlighted in green, the C-lobe tether residues are highlighted in red, and in blue are residues in the logo adjacent to the C-lobe tether.

did not impair folding or secretion (Table S3 and Fig. S10). We next examined the thermostability and catalytic activity of these mutants using thermal shift assays and Promega UDP-Glo assays, respectively. The mutants altered thermal stability to varying degrees. While T336I, Y311F/T336I, and Y311F/T336V were partially destabilizing (≈2 °C relative to wt), F309W, Y311I/T336V, and Y311F/T336V were more destabilizing (>4 °C relative to wt) (Fig. 5B).



Mutant	Energy
P	23.5
Y	7.1
G	6.9
E	6.1
D	5.8
K	5.7
S	4.9
R	4.6
F	4.5
H	3.9
Q	3.2
N	2.4
C	2.2
A	2.1
T336(wt)	**0.0**
W	-1.4
L	-2.2
M	-2.2
V	-3.8
I	-4.6

Comparison	Protein and Buffer Only	Protein with UDP-GlcNAc and MnCl2
Sample	Melt Temp (C)	Melt Temp (C)
WT	49.0	50.5
T336I	47.5	48.3
F309W	45.5	46.5
Y311I T336V	45.3	44.5
Y311F T336I	48.0	47.8
Y311F T336Y	44.5	43.5
Y311F T336V	47.0	47.3

wt			
Acceptor		Donor	
Km (mM)	2.9 ± 3.5	Km (mM)	0.35 ± 0.02
Kcat (min^{-1})	779 ± 48.1	Kcat(min^{-1})	540 ± 18.8

T336I			
Acceptor		Donor	
Km (mM)	5.8 ± 0.91	Km (mM)	0.3 ± 0.05
Kcat (min^{-1})	2072 ± 203	Kcat(min^{-1})	1138 ± 24

Figure 2.6: Computational and experimental screen of B3GNT2-specific variations in the C-lobe tether. A, computational mutational screen of the T336 mutants to identify potential stabilizing mutations. B, thermostability data of T336I mutant and wt B3GNT2, with all other mutants. C, table of kinetic parameters for acceptor and donor saturation in wt and T336I. D, kinetic efficiency (Kcat/Km) of B3GNT2 wt versus T336I upon acceptor and donor saturation, demonstrating a 1.3-fold and 2.5-fold increase, respectively, for the T336I relative to wt.

Analysis of the kinetic efficiency (kcat/Km) of the mutants revealed varying impact on substrate affinity (Km) and turnover (kcat). In particular, catalytic activity of T336I increases by approximately twofold relative to wt, under acceptor and donor saturation (Fig. 5, C and D and Table S2). The Km of T336I increased twofold under acceptor saturation and decreased by 0.15-fold under donor saturation. The catalytic efficiency of T336I increased by 1.3-fold and 2.5-fold under acceptor and donor saturations, respectively (Fig. 5D and Table S2). On the other hand, the F309W mutant displayed catalytic efficiency comparable to wt upon acceptor saturation, and a 1.93-fold increase in efficiency upon donor saturation, despite reduced thermostability. The other mutants, generally, displayed decreased catalytic efficiency relative to wt (Table S2).

To investigate the structural basis for the increased activity observed for the T336I mutant, we performed microsecond time-scale molecular dynamics (MD) simulations of wt and mutant B3GNT2 (Fig. 6), focusing on the conformational changes associated with the xED-Asp. In the crystal structure, the xED-Asp (D333) exists in two distinct conformations: D-in and D-out.

In the D-in conformation, the xED-Asp is pointing toward the hydrophobic core and forms a water-mediated hydrogen bonding network with T336. In the D-out conformation, the xED-Asp points out toward the acceptor-binding site and forms a hydrogen bond with a hydroxyl group in the acceptor-bound complex where it acts as catalytic base (Fig. 6A). In the MD simulations of wt B3GNT2, both these conformations are equally sampled in the apo and acceptor-bound complexes (Fig. 6B). However, in the T336I mutant, the xED-Asp is predominantly observed in the D-out conformation. The D-in conformation is not sampled as frequently in the mutant, since the Ile substitution occludes the water-binding site in wt B3GNT2. The shift in the conformational occupancy of the xED-Asp in the acceptor-bound "out" conformation may explain the partial increase in catalytic activity observed for the T336I mutant because the xED-Asp is readily able to deprotonate the acceptor. We further note that in the crystal structure of the closest relative, Manic fringe (PDB ID: 2J0A (29)), which contains a valine in place of the threonine, the xED-Asp adopts the D-out conformation in the crystal structure. Indeed, MD simulation with a valine mutant also demonstrates a preference for the D-out conformation (Fig. S11). Finally, we note that protonation of the xED-Asp also alters conformational dynamics (Figs. S12 and S13) primarily through changes in the chi-2 dihedral, as noted in other systems (P. Chakrabarti 1994; Shan et al. 2009). Based on these MD simulations, we hypothesize that changes in pKa may influence B3GNT2 catalytic activity. Together, our simulations provide additional support for our hypothesis that GT-A fold catalytic activities and mechanisms can be fine-tuned through mutations in the GT-A–specific C-lobe tether.

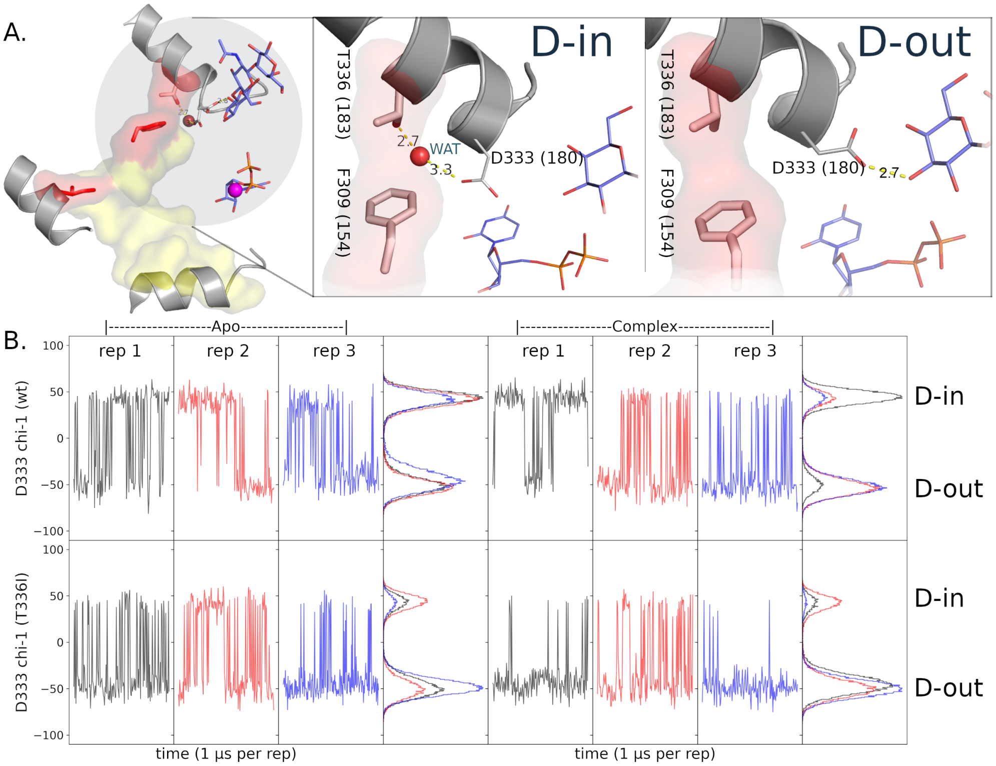


Figure 2.7: Molecular dynamics simulations of wt and mutant B3GNT2. A, snapshots from an MD simulation of the wt complex, showing two unique conformations of the xED-Asp. The D-in and D-out conformations are termed as such depending on their orientation inward, interacting with the threonine aided through a hydrogen bond interaction with a water molecule, or outward toward the acceptor–donor complex. B, 12 MD simulations (three replicates, 1 μs each) demonstrating the conformational shift of mutant T336I to the D-out conformation. Replicates show the dynamic switching between the D-in and D-out conformations over the course of the simulation, with the histograms showing the total ratio of D-in:D-out for each replicate.

2.4 Discussion

2.4.1 A proposed modular evolution of GT-As

In our previous study comparing GT-A fold enzymes from diverse organisms, we identified a conserved hydrophobic core under strong selective pressure, as reflected by the low evolutionary rates of these residues among the 231 aligned positions in the GT-A catalytic domain (Figs. 7A and S14). Here, we further dissect the anatomy of the core based on a broader analysis of diverse nucleotide-binding Rossman fold enzymes. Our studies reveal three distinct GT-A core modules added over evolutionary time (Fig. 7B) that are further embellished by family specific hypervariable regions. The first module is contained within an ancestral PBC, common to many nucleotide phosphate–binding enzymes. Ancestral phosphate-binding enzymes embellished upon this core to maintain its phosphate-binding function while resulting in the functionally diverse superfamilies that exist today. This core serves a similar function in GT-As by conserving motifs (specifically, the DXD motif) that are directly involved in binding the phosphate moiety of the donor substrate. GT-As, along with many other enzyme families, build upon this PBC to form the Rossmann fold, which binds a diverse array of cofactors including nucleotide sugars (Shin and Kihara 2019). We note different topological orientations of the PBC in enzyme families, even within the P-loop NTPases (Longo et al. 2020). However, the similarities between pyrophosphorylases and GT-As, in terms of shared PBC topologies, nucleotide, and divalent cation binding, suggests either convergent evolution, or a common ancestor connecting these enzyme families.

Extant GT-A fold enzymes extended the phosphate-binding module through addition of a unique C-terminal extension of the hydrophobic core, facilitated by the residues 156 and 183 (F340 and F365 in GT2), which tethers the F-helix and xED catalytic base to the PBC. The tether aids in positioning the catalytic base residue for inverting GTs critical for their SN2 displacement mechanism (**moremen_emerging_2019**). Among retaining GTs, the tether to the F-helix and positioning of the xED motif is maintained, but since catalytic base function for most retaining enzymes is accomplished by the β-phosphate oxygen of the sugar nucleotide donor (**moremen_emerging_2019**), selective pressure for maintaining the position of the catalytic base relative to the sugar donor is no longer needed. As a result, residues flanking the xED in retaining GT-As may be more malleable and likely to mutate, allowing these GT-As to sample new acceptor interactions and other functions, resulting in increased tethering variation.

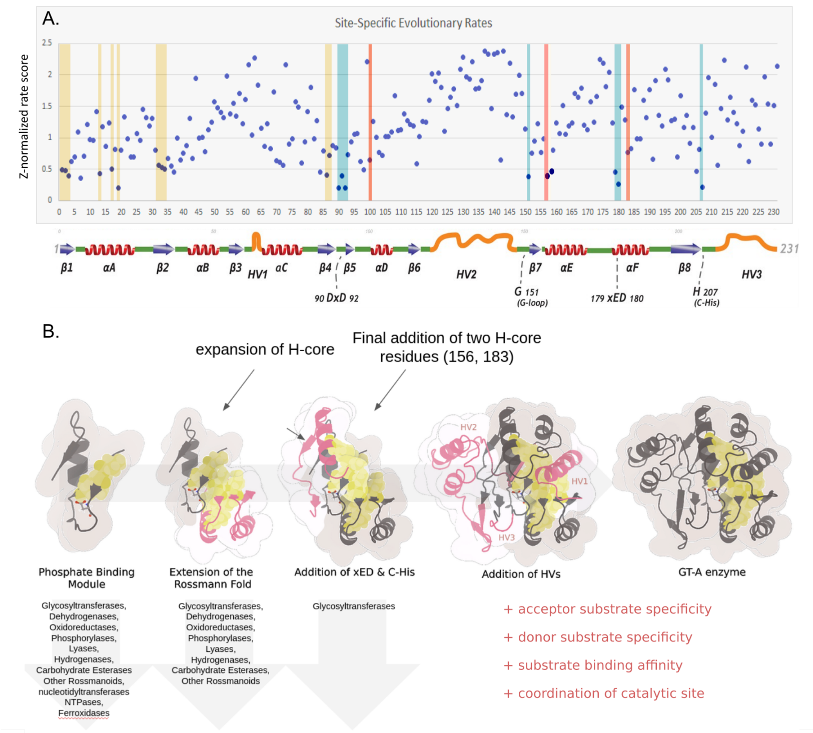


Figure 2.8: Modular evolution of GT-As. A, site-specific rate conservation of each residue of the 231 aligned positions. Dots in yellow bars reflect hydrophobic residues common to all Rossmann fold enzymes. Dots in blue bars reflect functional motifs, including DxD, G-loop, xED, and the C-His. Dots in red bars are GT-A–specific residues of the hydrophobic core. B, model of the evolutionary progression of fold A glycosyltransferases. Beginning from the elementary phosphate-binding cassette, GT-As gained a Rossmann fold that extended the hydrophobic core. Following this, various GT-As make use of the xED motif as a catalytic base, the presence of this motif correlates with mechanistic variations. Finally, family specific hypervariable regions are introduced to further regulate GT-A function. New additions in pink.

We previously proposed that inverting and retaining mechanisms evolved multiple independent times during GT-A enzyme evolution by generating a phylogenetic tree of diverse GT-A fold enzymes (Taujale, Venkat, et al. 2020). Here, we show that variations in the C-lobe tether may have contributed to this multiple independent evolution by altering core packing and xED-base positioning for either an associative mechanism or a dissociative mechanism. Consistent with this view, retaining GTs, mostly the ones that are further away from inverting families in the phylogenetic tree (GT2 unrelated, Fig. S7), tend to elongate the C-lobe tether with distances around 9 to 10 Å, often even accommodating extra residues between these positions (Figs. 3C and S8). In contrast, inverting GTs and GT2-related retaining GT-As have a tightly packed tether with inter-residue distances of around 3 to 4 and 5 to 7 Å, respectively. Multiple GTs show variability in this tether, even going so far as to change the packing interactions from van der Waals to salt bridges (Fig. S9).

We note that the retaining GTs, GT55 (mannosyl-3-phosphoglycerate synthase) and GT15 (glycolipid 2-α-mannosyltransferase) that are divergent (located in different branches of the tree), have a salt-bridge tether in common, suggesting that this variation may not just be structural but may have a functional role. Notably, both GTs are mannosyltransferases that catalyze transfer to unique acceptors; GT55 to a phosphate-linked glycerate acceptor and GT15 to a glycolipid (Gonçalves et al. 2010; Possner, Claesson, and Guy 2015). These two mannosyltransferases, accommodating different acceptor substrates, may suggest a convergent evolution of this tether and one of multiple solutions that influences accommodation of a vast diversity of acceptor–donor complexes. Thus, variability and malleability of the C-lobe tether provides the structural framework for multiple independent paths for evolutionary interconversion of retaining and inverting mechanisms on a common fold.

The regulatory functions of a flexible hydrophobic core have been well articulated in large protein superfamilies such as kinases (Susan S Taylor, Meharena, and Alexandr P Kornev 2019). Here, through computationally aided mutational analyses and MD simulations of the C-lobe tether in B3GNT2, we demonstrate that this GT-A–specific extension contributes to the functional stability of the enzyme. Introduction of the more canonical hydrophobic packing in the C-lobe tether favored the D-out conformation of the xED-Asp. This D-out conformation was also observed in the native crystal structures of a related GT31 enzyme, Manic fringe (Jinek et al. 2006; Moloney et al. 2000), which has a valine in place of B3GNT2's threonine. By changing the conformational occupancy of the catalytic base, wt B3GNT2 may illustrate an evolutionary mechanism to fine-tune catalytic activity. Accumulation of such mutations provides the basis for large-scale transitions in enzyme function during evolution (S. Bromberg and K. A. Dill 1994; J. Chen and Stites 2001; Tyzack et al. 2017).

An analysis of cancer variants cataloged in The Cancer Genome Atlas and COSMIC (the Catalogue Of Somatic Mutations In Cancer) reveals nearly 420 nonsynonymous mutations mapping to the GT-A hydrophobic core, 47 of which map to the C-lobe tether (Table S1 and Fig. S15). Most of these mutations are predominantly located in the GT8 subfamilies, such as GT8-LARGE, and change the size or biochemical properties of the hydrophobic residues. Investigating how these oncogenic mutations impact GT structure and regulation will further illuminate the functions of the understudied GT-A core in disease states. The ability to switch substrate preferences and control enzyme kinetics through malleable cores could mark the fine margins to ensure proper glycosyl transfer. As such, understanding the intricate mechanisms that guide the activity of these diverse enzyme families allows us to engineer new regulatory functions, and we believe that the identification of the critical rheostat functions played by the hydrophobic core could pave the way for rational design and engineering of GTs with new functional properties.

2.5 Methods

Hydrophobic core distance plots

To get minimum distances for each aligned hydrophobic residue in each PDB, we first split each chain from 470 GT crystal structures taken from the CAZy database into 972 PDBs. We then wrote a script using the Biopython module (Cock et al. 2009) to measure the minimum distances of each aligned hydrophobic position amongst each other. We only used structures with a resolution under 2.5 Å. We generated csv files of these positions and minimum atomic distance values, generating plots of each residue distance, as well as all-versus-all median distances for each hydrophobic core position (Fig. S16). With this table, we were able to categorize these GTs by (sub)family and mechanism and generate plots of the extended core. To avoid bias by PDBs that are overrepresented in the available GT-A structures, we performed a CD-HIT query on all available PDB sequences at 90% sequence similarity to generate a diverse and representative set of PDBs for structural informatics studies.

Rosetta modeling

Structural minimization and loop modification were performed, in preparation for MD simulations, using Rosetta's kinematic loop generation protocol (Susan S. Taylor and Alexandr P. Kornev 2011). Structures underwent 10,000 cycles of minimization to prevent atomic clashes in silico.

Oncogenic variant analyses

Full-length GT-A sequences were mined from The Cancer Genome Atlas (Tomczak, Czerwińska, and Wiznerowicz 2015) and COSMIC databases. These sequences were mapped to previously published GT-A profiles (Taujale, Venkat, et al. 2020). Mutations falling at hydrophobic core positions were collected, and duplicate counts were pruned based on patient and sample IDs to get a final count.

Mutational analyses

For the B3GNT2 structure, we computed mutations for every amino acid for the equivalent positions at 154 and 183 (F309 and T336 in B3GNT2 [PDB ID: 6WMN]). These mutations were performed using the cartesian DDG protocol (Frenz et al. 2020; Park et al. 2016), with three replicates. Rosetta energies were averaged to produce the table of energy values in Table S2. From this table, we picked, based on Rosetta energy scores, sets of stabilizing and destabilizing mutations. A critical caveat to note is that the Rosetta energy score only gives a relative indication of whether a structure is stabilizing or destabilizing. This method does not consider backbone rearrangement upon a mutation that changes packing; thus, the score does not always reflect in vitro data. Nevertheless, these scores provide an adequate basis for selecting mutations.

Mutant expression and purification

The B3GNT2 wt construct was generated as previously described (Kadirvelraj et al. 2021). Site-directed mutagenesis was performed using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs) to generate the six mutant B3GNT2 samples. Recombinant B3GNT2 and mutants were generated by transfection of 100 ml cultures of FreeStyle 293-F cells (Thermo Fisher Scientific) as previously described (Kadirvelraj et al. 2021). Six days after transfection, the samples were harvested using centrifugation, and enzyme in the culture supernatant was purified by Ni2+–nitrilotriacetic acid chromatography. Final samples were buffer exchanged into 25 mM Hepes and 300 mM NaCl, pH 7.5, concentrated by ultrafiltration, and protein concentration was determined using GFP-fluorescence and UV absorbance using a Nanodrop spectrophotometer. The samples were buffer exchanged into 25 mM Hepes and 300 mM NaCl and verified for purity and length using SDS-PAGE gels.

Sequence analysis

Sequence logos were generated using WebLogo 3.0 and GTXplorer (Crooks et al. 2004; Taujale, Soleymani, et al. 2021), using sequence alignments generated in our previous article (Taujale, Venkat, et al. 2020). We performed the structure-based sequence alignment using PROMALS3D and visualized the sequence alignment using ESPript3 (Gouet, Robert, and Courcelle 2003; Pei, B.-H. Kim, and Grishin 2008). The secondary structure representation in the alignment was generated using data from the DSSP output (Kabsch and Sander 1983) on the GT2 crystal structure (PDB ID: 2Z87). Calculation of deletions was performed by counting the percentage of gaps in a position across the sequence alignment (Fig. S17).

HMM analysis

Utilizing HMMs produced from Ref. (Kolodny et al. 2021), we ran searches across available GT-A sequences using HMMsearch (Mistry et al. 2013). These searches detected significant similarities in the PBC of P-loop NTPases and a subset of Rossmann fold enzymes, including GT-As. We then took a broad number of the PBCs from the published HMMs along with a set of representative PBCs from GT-As and pyrophosphorylases and performed an all-versus-all structural comparison using the TMalign algorithm (Zhang and Skolnick 2005). These RMSDs were then used in a network graph in Cytoscape (National Resource for Network Biology) (Shannon et al. 2003), where nodes represent each PDB and edges represent the RMSD similarity between each node. We used an edge-weighted spring embedded layout to organize the nodes into clusters of closely related proteins. We used a cutoff filter of 2.5 Å to remove the noise of distant connections. This resulted in clusters of closely related proteins, placing UDP-sugar pyrophosphorylases and GT-As next to each other.

Dihedral analyses

Python code was written for analyzing dihedral angles of residues in PBDs and MD frames (Figs. 5, S11–S13 and S18). This code can be found in the GitHub link in the Data availability section.

Kinetics

Promega UDP-Glo GT assays were used to analyze the B3GNT2 kinetic parameters as previously described (Kadirvelraj et al. 2021). Reactions were performed in a buffer containing 100 mM Hepes, pH 7, 2 mM MnCl2, and 1 mg/ml bovine serum albulin in 10 μl reactions using varied concentrations of lacto-N-neotetraose (0.3125–5 mM) as acceptor and UDP-GlcNAc (0.0625–1 mM) as donor to determine the KM and kcat values for wt and mutant B3GNT2 (Table S2 and Fig. S19). Enzyme input varied from 0.156 ng for wt B3GNT2 to 10 ng for severely destabilizing mutations, and each sample was run in biological duplicates.

Molecular Dynamics

Multiple MD simulations were run on the B3GNT2 crystal structures (PDB IDs: 6WMN and 6WMO). We first performed loop modeling using the Kinematic Loop Modeling Protocol in Rosetta to address any missing regions in the structure and then minimized the structure to avoid steric clashes (Stein and Kortemme 2013). Long time-scale unbiased MD simulations were performed on B3GNT2 at the microsecond level, with two replicates (each 1 μs long). All MD simulations used the Amber99SB-ILDN force field, commonly used for long time-scale protein simulations, along with the GLYCAM06 force field for glycan parameterization (Case et al. 2005; Kirschner et al. 2008; Lindorff-Larsen et al. 2010). Long-range electrostatics were calculated via particle mesh Ewald algorithms. All simulations used the TIP3P water model (Price and Brooks 2004). Energy minimization was run for a maximum of 10,000 cycles, performed using the steepest-descent algorithm, followed by the conjugate-gradient algorithm. The system was heated from 0 K to a temperature of 300 K. MD analyses were facilitated in python using the MDAnalysis module (Michaud-Agrawal et al. 2011). After two equilibration steps that lasted 50 ps, microsecond-long simulations were run at a 2 fs timestep.

Single-molecule charge calculations

We derived the protocol for parameterization of the UDP-donor substrate for the GTs from the GLYCAM force-field article (Kirschner et al. 2008). Ab initio QM was performed using Gaussian16 to optimize the donor ligand at the HF/6-31G* level. We then calculated the charge of the compound using antechamber. The electric charge of the aglycon was previously calculated to be -0.194 au. These parameters were then used to generate ligand input files for use with MD simulations.

Molecular modeling

The structures were visualized and analyzed in Schrodinger PyMOL 2.0. Structural alignments were performed in PyMOL 2.0 using the cealign algorithm (Shindyalov and Bourne 1998). Cartoon models of these structures were created using The Protein Imager (Tomasello, Armenia, and Molla 2020) to aesthetically portray these structures, after alignment in PyMOL 2.0.

Site-specific relative evolutionary rate conservation

To produce a normalized conservation value for each aligned position, we used a previously generated alignment, published in our previous article (Taujale, Venkat, et al. 2020), as input into the program Rate4Site (Pupko et al. 2002). This software employs an empirical Bayesian method to calculate a neighbor-joining tree with maximum likelihood distances to output a relative conservation score at each site.

Thermal Shift

ThermoFluor assays were performed in 96-well PCR plates in duplicates with each well containing 45 μl of GFP-tagged protein in the desired buffer at a concentration of 2 μM. The buffer consisted of 25 mM Hepes, 300 mM NaCl, pH 7.5, with 5 μl of 100× SYPRO Orange (Thermo Fisher Scientific). After a 15 min preincubation at room temperature, a melt curve program was run on a Bio-Rad CFX96 machine using a 50 μl total sample volume, from 25 to 95 °C, with a ramp speed of 1 °C/min. The B3GNT2 melt curve was observed in the 40 to 70 °C temperature range based on an increase in SYPRO Orange fluorescence, whereas the GFP fusion tag exhibited an additional melt curve at $\approx$88 °C.

AlphaFold2 models

AlphaFold2 produced several previously unknown GT-A structures (Jumper et al. 2021). For subfamilies not found in the AlphaFold2 database, we ran AlphaFold2 on a supercomputer cluster to produce models. After mapping these sequences to known profiles, as described in our previous article (Taujale, Venkat, et al. 2020), we wrote a python script to map alignment positions to these structural models and then visualized the hydrophobic core positions in PyMOL 2.0.

References

Agrawal, Praveen et al. (June 2017). "A Systems Biology Approach Identifies FUT8 as a Driver of Melanoma Metastasis". eng. *Cancer Cell* 31.6, 804–819.e7. ISSN: 1878-3686. DOI: 10.1016/j.ccell.2017.05.007.

Baldwin, E. P. and B. W. Matthews (Aug. 1994). "Core-packing constraints, hydrophobicity and protein design". eng. *Current Opinion in Biotechnology* 5.4, pp. 396–402. ISSN: 0958-1669. DOI: 10.1016/0958-1669(94)90048-5.

Ben-David, Moshe et al. (Jan. 2019). "Allosteric Modulation of Binding Specificity by Alternative Packing of Protein Cores". eng. *Journal of Molecular Biology* 431.2, pp. 336–350. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2018.11.018.

Bhardwaj, Nitin and Mark Gerstein (June 2009). "Relating protein conformational changes to packing efficiency and disorder". eng. *Protein Science: A Publication of the Protein Society* 18.6, pp. 1230–1240. ISSN: 1469-896X. DOI: 10.1002/pro.132.

Breton, Christelle et al. (2006). "Structures and mechanisms of glycosyltransferases". *Glycobiology* 16.2, 29R–37R.

Bromberg, S. and K. A. Dill (July 1994). "Side-chain entropy and packing in proteins". eng. *Protein Science: A Publication of the Protein Society* 3.7, pp. 997–1009. ISSN: 0961-8368. DOI: 10.1002/pro.5560030702.

Case, David A. et al. (Dec. 2005). "The Amber biomolecular simulation programs". eng. *Journal of Computational Chemistry* 26.16, pp. 1668–1688. ISSN: 0192-8651. DOI: 10.1002/jcc.20290.

Chakrabarti, P (1994). "Conformational analysis of carboxylate and carboxamide side-chains bound to cations". *Journal of molecular biology* 239.2, pp. 306–314.

Chen, J. and W. E. Stites (Dec. 2001). "Packing is a key selection factor in the evolution of protein hydrophobic cores". eng. *Biochemistry* 40.50, pp. 15280–15289. ISSN: 0006-2960. DOI: 10.1021/bi011776v.

Chugh, Seema et al. (Dec. 2015). "Pathobiological implications of mucin glycans in cancer: Sweet poison and novel targets". eng. *Biochimica Et Biophysica Acta* 1856.2, pp. 211–225. ISSN: 0006-3002. DOI: 10.1016/j.bbcan.2015.08.003.

Cock, Peter J. A. et al. (June 2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics". eng. *Bioinformatics (Oxford, England)* 25.11, pp. 1422–1423. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp163.

Crooks, Gavin E et al. (2004). "WebLogo: a sequence logo generator". *Genome research* 14.6, pp. 1188–1190.

Frenz, Brandon et al. (2020). "Prediction of Protein Mutational Free Energy: Benchmark and Sampling Improvements Increase Classification Accuracy". eng. *Frontiers in Bioengineering and Biotechnology* 8, p. 558247. ISSN: 2296-4185. DOI: 10.3389/fbioe.2020.558247.

Gonçalves, Susana et al. (2010). "Structural analysis of Thermus thermophilus HB27 mannosyl-3-phosphoglycerate synthase provides evidence for a second catalytic metal ion and new insight into the retaining mechanism of glycosyltransferases". *Journal of Biological Chemistry* 285.23, pp. 17857–17868.

Gouet, Patrice, Xavier Robert, and Emmanuel Courcelle (2003). "ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins". *Nucleic acids research* 31.13, pp. 3320–3323.

Grewal, P. K. et al. (June 2001). "Mutant glycosyltransferase and altered glycosylation of alpha-dystroglycan in the myodystrophy mouse". eng. *Nature Genetics* 28.2, pp. 151–154. ISSN: 1061-4036. DOI: 10.1038/88865.

Hardy, Jeanne A. et al. (Aug. 2004). "Discovery of an allosteric site in the caspases". eng. *Proceedings of the National Academy of Sciences of the United States of America* 101.34, pp. 12461–12466. ISSN: 0027-8424. DOI: 10.1073/pnas.0404781101.

Horn, James R. and Brian K. Shoichet (Mar. 2004). "Allosteric inhibition through core disruption". eng. *Journal of Molecular Biology* 336.5, pp. 1283–1291. ISSN: 0022-2836. DOI: 10.1016/j.jmb.2003.12.068.

Jinek, Martin et al. (2006). "Structural insights into the Notch-modifying glycosyltransferase Fringe". *Nature structural & molecular biology* 13.10, pp. 945–946.

Jumper, John et al. (Aug. 2021). "Highly accurate protein structure prediction with AlphaFold". eng. *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

Kabsch, Wolfgang and Christian Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers: Original Research on Biomolecules* 22.12, pp. 2577–2637.

Kadirvelraj, Renuka et al. (June 2021). "Comparison of human poly-N-acetyl-lactosamine synthase structure with GT-A fold glycosyltransferases supports a modular assembly of catalytic subsites". eng. *The Journal of Biological Chemistry* 296, p. 100110. ISSN: 1083-351X. DOI: 10.1074/jbc.RA120.015305.

Kim, Jonggul et al. (2017). "A dynamic hydrophobic core orchestrates allostery in protein kinases". *Science advances* 3.4, e1600663.

Kimber, Matthew S et al. (2020). "The Structurally Unusual Retaining β-Kdo Glycosyltransferase WbbB Uses a Double-Displacement Mechanism with an Intermediate Adduct Rearrangement Step". *The FASEB Journal* 34.S1, pp. 1–1.

Kirschner, Karl N. et al. (Mar. 2008). "GLYCAM06: a generalizable biomolecular force field. Carbohydrates". eng. *Journal of Computational Chemistry* 29.4, pp. 622–655. ISSN: 1096-987X. DOI: 10.1002/jcc.20820.

Kitazume, Shinobu, Takaomi C. Saido, and Yasuhiro Hashimoto (2004). "Alzheimer's beta-secretase cleaves a glycosyltransferase as a physiological substrate". eng. *Glycoconjugate Journal* 20.1, pp. 59–62. ISSN: 0282-0080. DOI: 10.1023/B:GLYC.0000016743.25495.45.

Kolodny, Rachel et al. (2021). "Bridging themes: short protein segments found in different architectures". *Molecular biology and evolution* 38.6, pp. 2191–2208.

Kornev, Alexandr P. and Susan S. Taylor (Mar. 2010). "Defining the conserved internal architecture of a protein kinase". eng. *Biochimica Et Biophysica Acta* 1804.3, pp. 440–444. ISSN: 0006-3002. DOI: 10.1016/j.bbapap.2009.10.017.

Lindorff-Larsen, Kresten et al. (June 2010). "Improved side-chain torsion potentials for the Amber ff99SB protein force field". eng. *Proteins* 78.8, pp. 1950–1958. ISSN: 1097-0134. DOI: 10.1002/prot.22711.

Longo, Liam M. et al. (Dec. 2020). "On the emergence of P-Loop NTPase and Rossmann enzymes from a Beta-Alpha-Beta ancestral fragment". eng. *eLife* 9, e64415. ISSN: 2050-084X. DOI: 10.7554/eLife.64415.

Maxwell, K. L. and A. R. Davidson (Nov. 1998). "Mutagenesis of a buried polar interaction in an SH3 domain: sequence conservation provides the best prediction of stability effects". eng. *Biochemistry* 37.46, pp. 16172–16182. ISSN: 0006-2960. DOI: 10.1021/bi981788p.

Mei, Longcan et al. (Mar. 2018). "Site-Mutation of Hydrophobic Core Residues Synchronically Poise Super Interleukin 2 for Signaling: Identifying Distant Structural Effects through Affordable Computations". eng. *International Journal of Molecular Sciences* 19.3, E916. ISSN: 1422-0067. DOI: 10.3390/ijms19030916.

Mistry, Jaina et al. (2013). "Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions". *Nucleic acids research* 41.12, e121–e121.

Moll, Tobias, Pamela J. Shaw, and Johnathan Cooper-Knock (May 2020). "Disrupted glycosylation of lipids and proteins is a cause of neurodegeneration". eng. *Brain: A Journal of Neurology* 143.5, pp. 1332–1340. ISSN: 1460-2156. DOI: 10.1093/brain/awz358.

Moremen, Kelley W and Robert S Haltiwanger (2019). "Emerging structural insights into glycosyltransferase-mediated synthesis of glycans". *Nature chemical biology* 15.9, pp. 853–864.

Ovchinnikova, Olga G et al. (2016). "Bacterial β-Kdo glycosyltransferases represent a new glycosyltransferase family (GT99)". *Proceedings of the National Academy of Sciences* 113.22, E3120–E3129.

Park, Hahnbeom et al. (Dec. 2016). "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules". eng. *Journal of Chemical Theory and Computation* 12.12, pp. 6201–6212. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.6b00819.

Pei, Jimin, Bong-Hyun Kim, and Nick V Grishin (2008). "PROMALS3D: a tool for multiple protein sequence and structure alignments". *Nucleic acids research* 36.7, pp. 2295–2300.

Possner, Dominik DD, Magnus Claesson, and Jodie E Guy (2015). "Structure of the glycosyltransferase Ktr4p from Saccharomyces cerevisiae". *PLoS One* 10.8, e0136239.

Price, Daniel J. and Charles L. Brooks (Nov. 2004). "A modified TIP3P water potential for simulation with Ewald summation". eng. *The Journal of Chemical Physics* 121.20, pp. 10096–10103. ISSN: 0021-9606. DOI: 10.1063/1.1808117.

Pupko, Tal et al. (2002). "Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues". *Bioinformatics* 18.suppl_1, S71–S77.

Shan, Yibing et al. (2009). "A conserved protonation-dependent switch controls drug binding in the Abl kinase". *Proceedings of the National Academy of Sciences* 106.1, pp. 139–144.

Shannon, Paul et al. (2003). “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. *Genome research* 13.11, pp. 2498–2504.

Shin, Woong-Hee and Daisuke Kihara (2019). “55 Years of the Rossmann Fold”. eng. *Methods in Molecular Biology (Clifton, N.J.)* 1958, pp. 1–13. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-9161-7_1.

Shindyalov, Ilya N and Philip E Bourne (1998). “Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.” *Protein engineering* 11.9, pp. 739–747.

Stein, Amelie and Tanja Kortemme (2013). “Improvements to robotics-inspired conformational sampling in rosetta”. eng. *PloS One* 8.5, e63090. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0063090.

Szilágyi, A. and P. Závodszky (May 2000). “Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey”. eng. *Structure (London, England: 1993)* 8.5, pp. 493–504. ISSN: 0969-2126. DOI: 10.1016/s0969-2126(00)00133-7.

Taujale, Rahil, Saber Soleymani, et al. (2021). “GTXplorer: A portal to navigate and visualize the evolutionary information encoded in fold A glycosyltransferases”. *Glycobiology* 31.11, pp. 1472–1477.

Taujale, Rahil, Aarya Venkat, et al. (2020). “Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases”. *Elife* 9, e54532.

Taujale, Rahil, Zhongliang Zhou, et al. (2021). “Mapping the glycosyltransferase fold landscape using interpretable deep learning”. *Nature Communications* 12.1, p. 5656.

Taylor, Susan S, Hiruy S Meharena, and Alexandr P Kornev (2019). “Evolution of a dynamic molecular switch”. *IUBMB life* 71.6, pp. 672–684.

Tomasello, Gianluca, Ilaria Armenia, and Gianluca Molla (2020). “The Protein Imager: a full-featured online molecular viewer interface with server-side HQ-rendering capabilities”. *Bioinformatics* 36.9, pp. 2909–2911.

Tomczak, Katarzyna, Patrycja Czerwińska, and Maciej Wiznerowicz (2015). "Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge". *Contemporary Oncology/Współczesna Onkologia* 2015.1, pp. 68–77.

Tyzack, Jonathan D et al. (2017). "Understanding enzyme function evolution from a computational perspective". *Current opinion in structural biology* 47, pp. 131–139.

Varki, Ajit, Richard D. Cummings, et al., eds. (2015). *Essentials of Glycobiology*. eng. 3rd. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press. URL: http://www.ncbi.nlm.nih.gov/books/NBK310274/.

Yoshida, Aruto et al. (2001). "Muscular dystrophy and neuronal migration disorder caused by mutations in a glycosyltransferase, POMGnT1". *Developmental cell* 1.5, pp. 717–724.

Zhang, Yang and Jeffrey Skolnick (2005). "TM-align: a protein structure alignment algorithm based on the TM-score". *Nucleic acids research* 33.7, pp. 2302–2309.

Chapter 3

Mechanistic and evolutionary insights into isoform-specific 'supercharging' in DCLK family kinases

Venkat, Aarya, et al. "Mechanistic and evolutionary insights into isoform-specific 'supercharging' in DCLK family kinases." eLife (2023).
Reprinted here with permission of the publisher.

3.1 Abstract

Catalytic signaling outputs of protein kinases are dynamically regulated by an array of structural mechanisms, including allosteric interactions mediated by intrinsically disordered segments flanking the conserved catalytic domain. The Doublecortin Like Kinases (DCLKs) are a family of microtubule-associated proteins characterized by a flexible C-terminal autoregulatory 'tail' segment that varies in length across the various human DCLK isoforms. However, the mechanism whereby these isoform-specific variations contribute to unique modes of autoregulation is not well understood. Here, we employ a combination of statistical sequence analysis, molecular dynamics simulations and in vitro mutational analysis to define hallmarks of DCLK family evolutionary divergence, including analysis of splice variants within the DCLK1 sub-family, which arise through alternative codon usage and serve to 'supercharge' the inhibitory potential of the DCLK1 C-tail. We identify co-conserved motifs that readily distinguish DCLKs from all other Calcium Calmodulin Kinases (CAMKs), and a 'Swiss-army' assembly of distinct motifs that tether the C-terminal tail to conserved ATP and substrate-binding regions of the catalytic domain to generate a scaffold for auto-regulation through C-tail dynamics. Consistently, deletions and mutations that alter C-terminal tail length or interfere with co-conserved interactions within the catalytic domain alter intrinsic protein stability, nucleotide/inhibitor-binding, and catalytic activity, suggesting isoform-specific regulation of activity through alternative splicing. Our studies provide a detailed framework for investigating kinome–wide regulation of catalytic output through cis-regulatory events mediated by intrinsically disordered segments, opening new avenues for the design of mechanistically-divergent DCLK1 modulators, stabilizers or degraders.

Keywords

kinase – allostery – protein evolution – structure–function – bioinformatics

3.2 Introduction

Protein kinases are one of the largest druggable protein families comprising 1.7% of the human genome and play essential roles in regulating diverse eukaryotic cell signaling pathways (Manning et al. 2002). The Doublecortin-like kinases (DCLKs) are understudied members of the calcium/calmodulin-dependent kinase (CAMK) clade of serine-threonine kinases (Agulto et al. 2021; Bayer and Howard Schulman 2019; Gógl et al. 2019). There are three distinct paralogs of DCLK (1, 2, and 3), the last of which is annotated as a "dark" kinase due to the lack of information pertaining to its function (Berginski et al. 2021). Human DCLK1 (also known as DCAMKL1) was initially identified in 1999 (Sossey-Alaoui and Srivastava 1999), followed by the cloning of human DCLK2 and 3 paralogs (Ohmae et al. 2006). Full-length DCLK proteins contain N-terminal Doublecortin-like (DCX) domains, microtubule-binding elements that play a role in microtubule dynamics, neurogenesis, and neuronal migration (Couillard-Despres et al. 2005; Horesh et al. 1999). DCLKs have garnered much interest as disease biomarkers, since they are upregulated in a variety of cancer pathologies (Cheng et al. 2022; Gao et al. 2016; Westphalen, Quante, and T. C. Wang 2017), as well as neurodegenerative disorders such as Huntington's Disease (Galvan, Francelle, Gaillard, Longprez, Carrillo-de Sauvage, Liot, Cambon, Stimmer, Luccantoni, Flament, et al. 2018b). However, the mechanisms by which DCLK activity is auto-regulated, and how and why they have diverged from other protein kinases is not well understood.

Like all protein kinases, the catalytic domain of DCLKs adopts a bi-lobal fold (Gógl et al. 2019), with an N-terminal ATP binding lobe and C-terminal substrate binding region. Canonical elements within the two lobes include the DFG motif, a Lys-Glu salt bridge that is associated with the active conformation, Gly-rich loop, and ATP-binding pocket, which are all critical elements for catalysis. Many protein kinases, including CAMKs, Tyrosine Kinases (TKs) and AGCs, elaborate on these core elements with unique N-terminal and C-terminal extensions that flank these catalytic lobes (Kannan et al. 2007; Yeon et al. 2016; T. Nguyen et al. 2015; Yeung, Kwon, et al. 2021), allowing them to function as allosteric regulators of catalytic activity (Gógl et al. 2019). Indeed, CAMKs are archetypal examples of kinases that can exist in an active-like structural conformation, yet still remain catalytically inactive (Gógl et al. 2019). This is in large part due to the presence of unique C-terminal tails that are capable of blocking ATP or substrate binding in well-studied kinases such as CAMK1 and CAMKII. In canonical CAMKs, autoinhibition may be released upon Ca2+/Calmodulin (CaM) interaction with the CAMK C-tail, which makes the substrate-binding pocket and enzyme active-site accessible (Rellos et al. 2010). In CAMKII, the N and C-terminal segments flanking the kinase domain are variable in length across different isoforms and the level of kinase autoinhibition or autoactivation has been reported to be dependent on the linker length (Bhattacharyya et al. 2020). The CAMKII C-tail can be organized into an autoregulatory domain and an intrinsically disordered association domain. The autoregulatory domain also serves as a pseudosubstrate, which physically blocks the substrate binding pocket until it is competed away by CaM (Hudmon and Howard Schulman 2002). Notably, this autoregulatory pseudosubstrate can be phosphorylated (Rellos et al. 2010), and phosphorylation of the C-tail makes CAMKII insensitive to CaM binding. Across the CAMK group, several other kinases share autoinhibitory activity via interactions between Ca2+/CaM binding domains and the C-terminal tail (Huse and Kuriyan 2002; Wayman et al. 2008), and a major feature of these kinases is variation in the tail length across the distinct genetic isoforms.

Name (this study)	Isoform number	Uniprot ID	Alternate Names
DCLK1.1	1	O15075-2	DCAMKL1 alpha
DCLK1.2	2	O15075-1	DCAMKL1 beta
ΔDCLK1.1	3	O15075-3	-
ΔDCLK1.2	4	O15075-4	-

The human genome encodes four distinct DCLK1 isoforms, termed DCLK1.1-1.4 in UniProt (Table 1, Figure 3.1, (Omori et al. 1998)), which display differential activity– and tissue-specific expression profiles. Human DCLK1.1 (also known as DCLK1 alpha) is expressed in a variety of tissues, but is enriched in cells derived from the fetal and adult brain, whereas DCLK1.2 (also known as DCLK1 beta) is expressed exclusively in the embryonic brain (Matsumoto, Pilz, and Ledbetter 1999). DCLK1.3 and 1.4, which lack tandem microtubule-binding DCX domains (Figure 3.1) but are otherwise identical to DCLK1.1 and DCLK1.2, respectively, are also highly expressed in the brain. To aid with clarity, the names of the human DCLK1 genes and their isoforms used in this paper are summarized in Table 1. Recent structural and cellular analyses have begun to clarify the mechanisms by which the DCLK1.2 isoform is regulated by the C-tail (Agulto et al. 2021; Patel et al. 2021; Cheng et al. 2022). Mechanistically, autophosphorylation of Thr 688, which is present only in the C-tail of DCLK1.2 (and DCLK1.4), blunts kinase activity and subsequently inhibits phosphorylation of the N-terminal DCX domain and thus drives DCLK microtubule association in cells (Agulto et al. 2021). Consistently, deletion of the C-tail or mutation of Thr 688 restores DCLK1.2 kinase activity, subsequently leading to DCX domain phosphorylation and the abolition of microtubule binding. The length and sequence of the C-tail varies across the DCLK1 isoforms; however, how these variations contribute to isoform-specific functions and how they emerged during the course of evolution is not known.

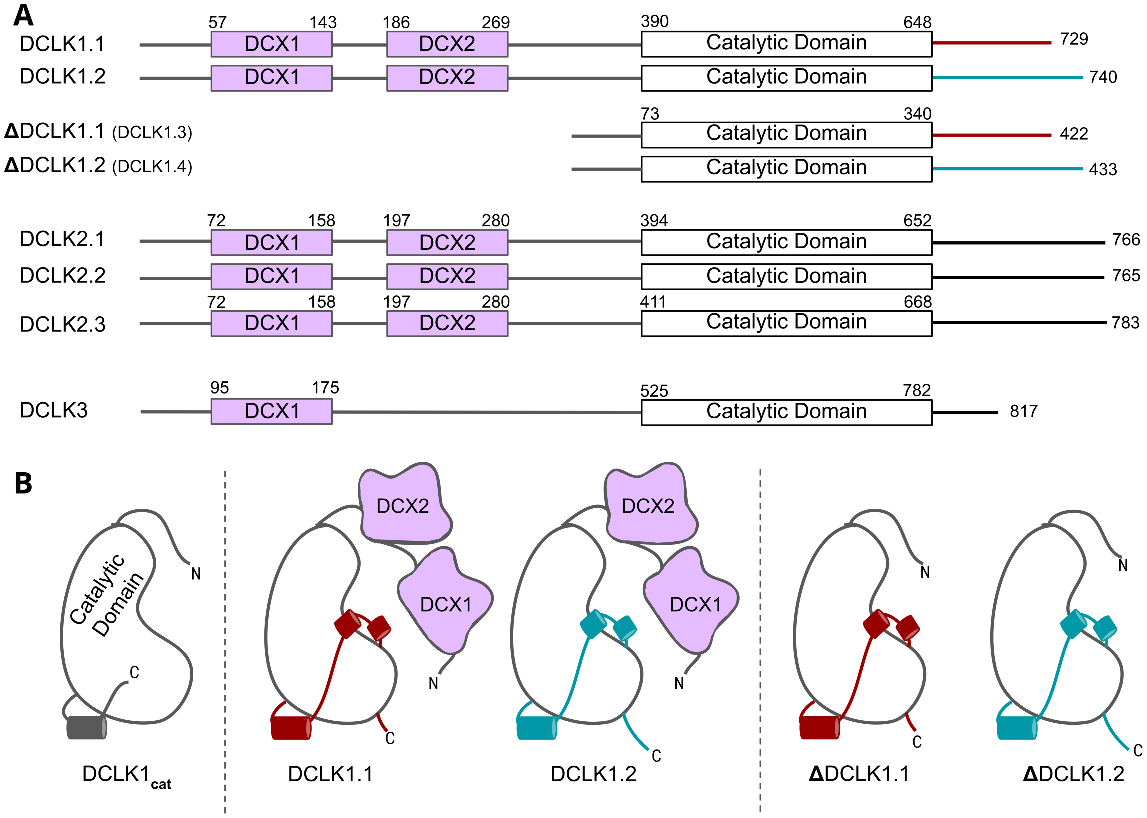


Figure 3.1: A) Schematic representation of domain organization for the known isoforms of the three human DCLK paralogs. Domain boundaries are annotated according to the representative amino acid sequences derived from UniProt. B) DCLK1 isoforms visualized as cartoons, showing key structural differences between the four human DCLK1 isoforms and a DCLK1 catalytic domain with artificially short linker regions (DCLK1cat).

In this paper, we employ an evolutionary systems approach that combines statistical sequence analysis with experimental studies to generate new models of DCLK evolutionary divergence and functional specialization. We identify the C-terminal tail as the hallmark of DCLK functional specialization across the kingdoms of life and propose a refined model in which this regulated tail functions as a highly adaptable ‘Swiss-Army knife’ that can ‘supercharge’ multiple aspects of DCLK signaling output. Notably, a conserved segment of the C-tail functions as an isoform-specific autoinhibitory motif, which mimics ATP functions through direct tail docking to the nucleotide-binding pocket, where it forms an ordered set of interactions that aligns the catalytic (C) spine of the kinase in the absence of ATP binding. Furthermore, molecular modelling demonstrates that a phosphorylated threonine in the C-tail of DCLK1.2, which is absent in DCLK1.1, is positionally-poised to competitively mimic the gamma phosphate of ATP, perhaps in a regulated manner. Other segments of the tail function as a pseudosubstrate by occluding the substrate-binding pocket and tethering to key functional regions of the catalytic domain. Thermostability analysis of purified DCLK1 proteins, combined with molecular dynamics simulations, confirms major differences in thermal and dynamic profiles of the DCLK1 isoforms, while catalytic activity assays reveal how specific variations in the G-loop and C-tail can rescue DCLK1.2 from the autoinhibited conformation. Together, these studies demonstrate that isoform-specific variations in the C-terminal tail co-evolved with residues in the DCLK kinase domain, contributing to regulatory diversification and functional specialization.

3.3 Results

3.3.1 Origin and evolutionary divergence of DCLK family members

The human DCLKs repertoire is composed of three genes, termed DCLK1, 2 and 3 (Figure 3.1A, Table 1). The experimental model employed in this study, DCLK1, is composed of multiple spliced variants in human cells. Those full-length proteins that contain N-terminal DCX domains are usually referred to as DCLK1.1 or DCLK1.2 and the variants that lack the DCX domains are termed here (for simplicity) ΔDCLK 1.1 and ΔDCLK1.2 (also referred to as DCLK1.3 and DCLK1.4). The core catalytic domain with minimal flanking regions (DCLK1cat, Figure 3.1B) is identical in all DCLK1 proteins, whereas the length of the tail, or the presence of the DCX domains, generates considerable diversity from the single human DCLK1 gene (Figure 3.1B, Figure 3.1-figure supplement 1). To infer evolutionary relationships of DCLK paralogs, and especially the evolution of the C-terminal tail regions that lie adjacent to the kinase domain (Figure 3.1), we performed phylogenetic analysis of 36 DCLK sequences with an outgroup of closely related CAMK sequences (Figure 3.2A, Figure 3.2-source data 1). These DCLK sequences are from a representative group of holozoans, which consist of multicellular eukaryotes (metazoans) and closely related unicellular eukaryotes (pre-metazoans). The analysis generated four distinct clades: pre-metazoan DCLK, metazoan DCLK3, vertebrate DCLK2 and vertebrate DCLK1. Interestingly, DCLK genes demonstrated significant expansion and diversification within metazoan taxa. The pre-metazoan DCLK sequences were the most ancestral and showed no DCLK diversity, suggesting the DCLK expansion and diversification correlated with the evolution of multicellular organisms. Within the metazoan expansion of DCLK, DCLK3 is the most ancestral and can be broken down into two sub-clades: protostome DCLK3 and deuterostome DCLK3. Within invertebrates, only two DCLK paralogs were present, one that was identified as a DCLK3 ortholog and another that was not clearly defined as either DCLK1 or DCLK2. This suggests that the diversification into DCLK1 and DCLK2 paralogs from an ancestral DCLK1/2-like paralog occurred after the divergence of invertebrates and vertebrates, which is further supported by the monophyletic DCLK1 and DCLK2 clades in vertebrates (bootstrap value: 99).

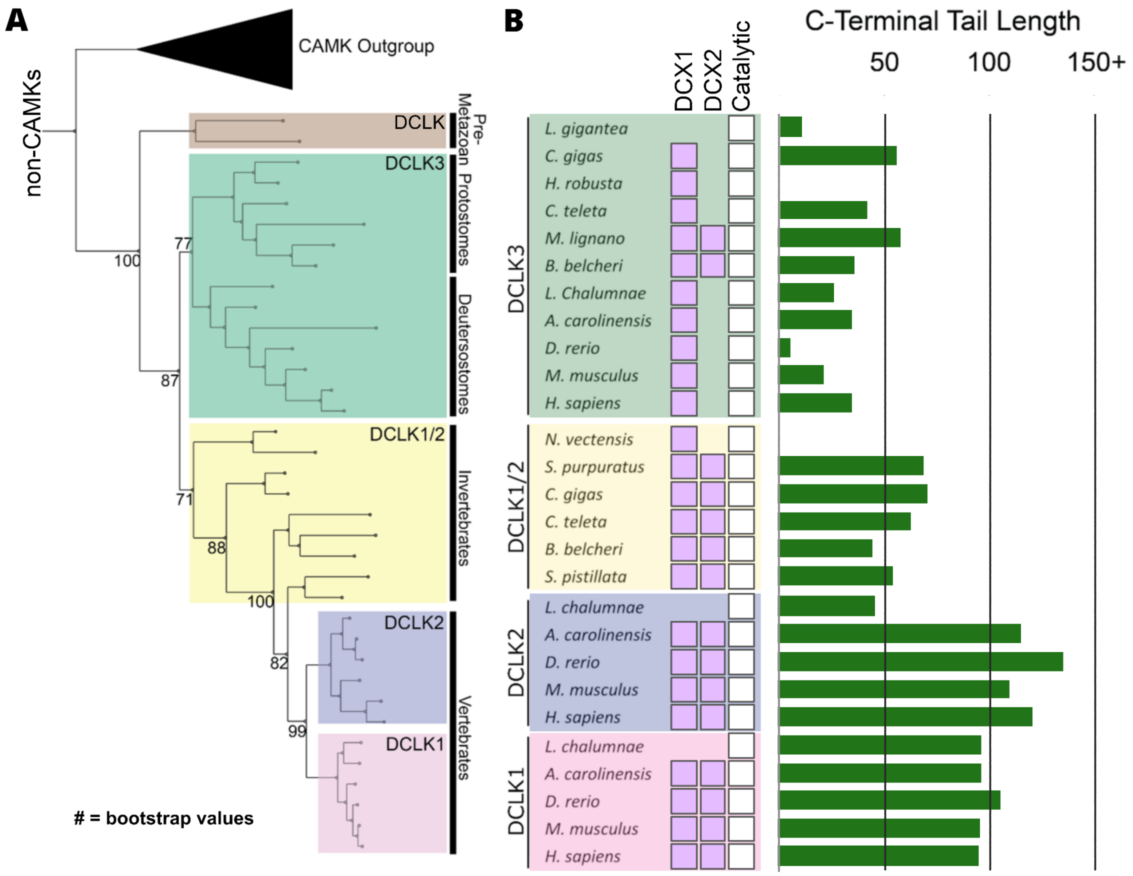


Figure 3.2: Evolution of the DCLK family. A) Phylogenetic tree showing the divergence and grouping of DCLK sub-families in different taxonomic groups. Bootstrap values are provided for each clade. B) Shows domain annotations for sequences included in the phylogenetic tree. The length of C-terminal tail segment for these sequences is shown as a histogram (green). The original tree generated using IQTREE is provided in Figure 2-source data 1.

Interestingly, the expansion of DCLK in metazoans and the diversification of DCLK1 and DCLK2 within vertebrates correlates well with the length and sequence similarity of the C-terminal tail, which also varies between the different DCLK1 splice variants (Figure 3.1A, Figure 3.1-figure supplement 1 and 2). Within both protostome and deuterostome DCLK3, the length of the C-terminal tail is ≈50 residues or less. This is in marked contrast to the tail lengths of vertebrate DCLK1 and DCLK2, which are ≈100 residues long. In addition to the C-terminal tail, an analysis of the domain organization of these DCLKs reveals that DCLK3 predominantly contains only a single N-terminal Doublecortin domain (DCX), whereas invertebrate DCLKs, and vertebrate DCLK1 and DCLK2 predominantly contain two DCX domains at the N-terminus of the long isoforms (Figure 3.2B). In addition, we identified a putative active site-binding motif, VSVI, and a phosphorylatable threonine conserved within vertebrate DCLK1 and DCLK2, which is absent in all other DCLK sequences, including invertebrate DCLK1/2. This raises the possibility that the DCLK1/2 tail extensions are employed for vertebrate-specific regulatory functions.

Next, we compared the type of DCLK1 protein sequence encoded by a range of chordate mammalian genomes. The domain organization of each DCLK1 isoform was compared based on annotated sequences from UniProt, demonstrating the presence of at least one DCLK1 protein that lacks the DCX domains in every species examined, with a mixture of ΔDCLK1.1 and ΔDCLK1.2 splice variants. Interestingly, it was only in the human DCLK1 gene that definitive evidence for ΔDCLK1.1 and ΔDCLK1.2 variants erewere found (Figure 3.3A). To establish a model for DCLK1 biophysical analysis, we constructed a recombinant hybrid human DCLK1 catalytic domain with a short C-tail sequence that is equivalent to DCLK1.1 amino acids 351-689, containing the catalytic domain with a short C-tail region. As shown in Figure 3.3B, incubation of size-exclusion chromatography (SEC) purified GST-tagged DCLK1 with 3C protease generated the mature untagged DCLK1 protein for biophysical analysis. Analytical SEC revealed that purified DCLK1.1 and DCLK1.2 isoforms are monomeric in solution (Figure 3.3-figure supplements 1-3). We evaluated catalytic activity for DCLK1.1351-689 using a validated peptide phosphorylation assay (Figure 3.3C), which revealed efficient phosphorylation of a DCLK1 substrate peptide. The KM[ATP] for peptide phosphorylation was close to 20 μM in the presence of Mg2+ ions (Figure 3.3C, left panel), similar to values measured for other Ser/Thr kinases that are autophosphorylated and active after expression from bacteria (Dominic P. Byrne, Shrestha, et al. 2020). DCLK-dependent peptide phosphorylation was completely blocked (Figure 3.3C, right panel) by prior incubation of the reaction mixture (containing 1mM ATP) with the chemical inhibitor

DCLK1-IN-1, as expected (29). In addition to enzyme activity, we monitored thermal denaturation of purified, folded, DCLK1351-689 protein in the presence of ATP, either alone or as a Mg:ATP complex, which is required for catalysis. As shown in Figure 3.3D, DCLK1 was stabilized by 2.1°C upon incubation with an excess of Mg:ATP, and this protective effect was completely blocked by mutation of Asp 533 (of the conserved DFG motif) to Ala, consistent with canonical ATP interaction in the nucleotide-binding site. Finally, we assessed the thermal effects of a panel of DCLK1 inhibitors on the model DCLK1.1351-689 protein. Prior incubation with DCLK1-IN-1, LRRK2-IN-1, the benzopyrimidodiazipinones XMD8-92 and XMD8-85, which have been reported to potently (though not specifically) inhibit DCLK1 activity (Patel et al. 2021), led to marked protection from thermal unfolding (Figure 3.3-figure supplement 4). Consistently, the negative control compound DCLK1-Neg (Ferguson et al. 2020) was ineffective in stabilizing DCLK1.

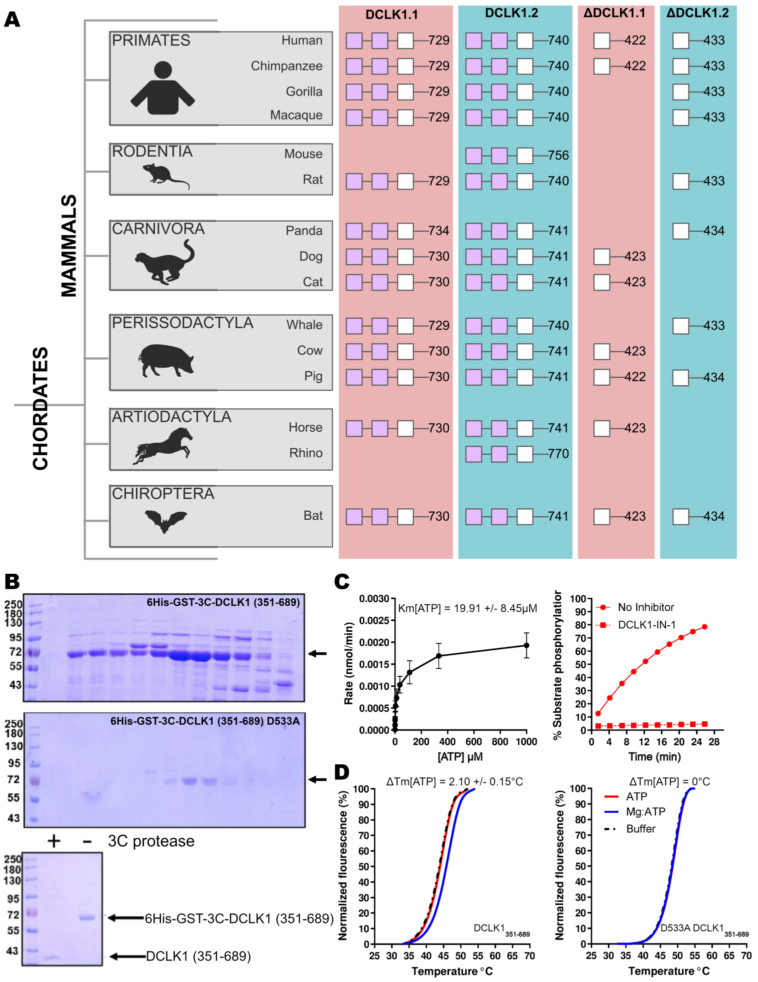
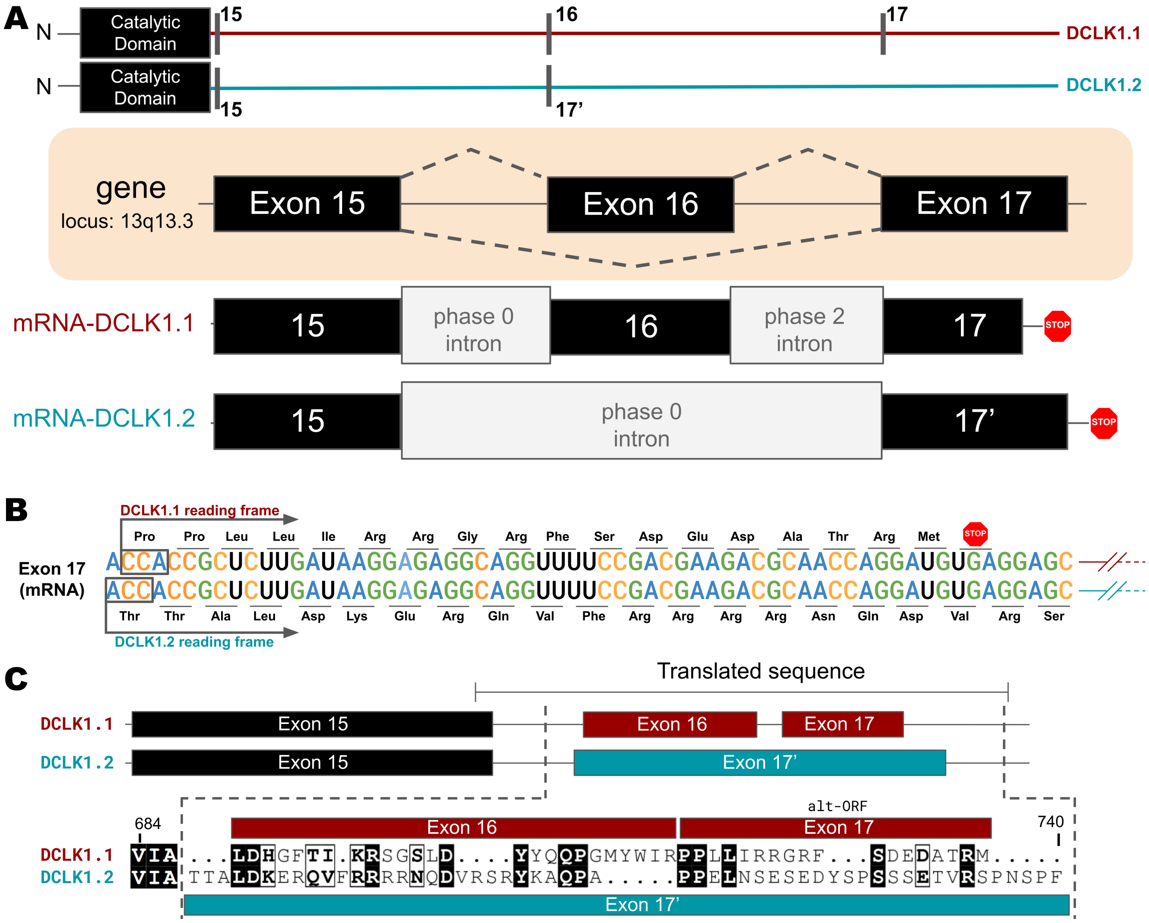


Figure 3.3: A) Cartoon cladogram of mammalian species showing the domain organization of each DCLK1 isoform from representative annotated sequences from UniProt. UniProt IDs for each sequence are provided in Figure 3-Source File 1. B) SDS-PAGE of 6His-GST-3C-DCLK1.1 (351-689, Top) or a D533A mutant in which the DFG Asp is mutated to Ala (Middle). Proteins were separated by size exclusion chromatography, and high-purity fractions were pooled. The affinity tag was removed prior to analysis by incubation with 3C protease, leading to a demonstrable shift in mobility (bottom) C) Evaluation of catalytic activity towards DCLK1 peptide. DCLK1.1 351-689 possesses a Km [ATP] 20 μM in vitro (left) and real-time substrate phosphorylation was inhibited by prior incubation with the small molecule DCLK1-IN-1, right). D) Thermal shift assay demonstrating a 2.1°C increase in the stability of DCLK1 351-689 in the presence of Mg:ATP (left), which was absent in the D533A protein (right).

3.3.2 Key differences between isoforms in the C-tail of DCLK1 arise from alternative-splicing and different open-reading frames

Higher-order vertebrates have multiple isoforms of DCLK1 and DCLK2, where sequence variations occur in either or both the N and C terminal regions attached to the kinase domain. Human DCLK1, for example, has four unique isoforms. Isoforms 1 and 2 differ in C-terminal tail length due to variations in exon splicing (Figure 3.3.4A). Further examination of the intron and exon boundaries indicates that human DCLK1.1 contains an additional exon (exon 16) that is not spliced in DCLK1.2. Exon 16 is spliced with exon 17 with a phase 2 intron, which introduces a shift in the reading frame and an earlier translated stop codon (UGA) in exon 17 (Figure 3.3.4B). In DCLK1.2, exon 15 is spliced with exon 17, with a non-disruptive phase 0 intron, resulting in the full translation of exon 17. These changes introduce multiple indels (insertions and deletions) and result in the insertion of a phosphorylatable threonine (T688) in DCLK1.2 that is absent in the DCLK1.1 variant (Figure 3.3.4B-C), suggesting a possible exon duplication for adaptive regulation of DCLK1 function by phosphorylation. DCLK1.2 is the best-characterized isoform in terms of structure and function (Cheng et al. 2022), and to compare it with DCLK1.1, we generated a series of C-terminal tail deletion mutants to evaluate how variations in the C-terminal tail contribute to isoform specific DCLK1 functions.

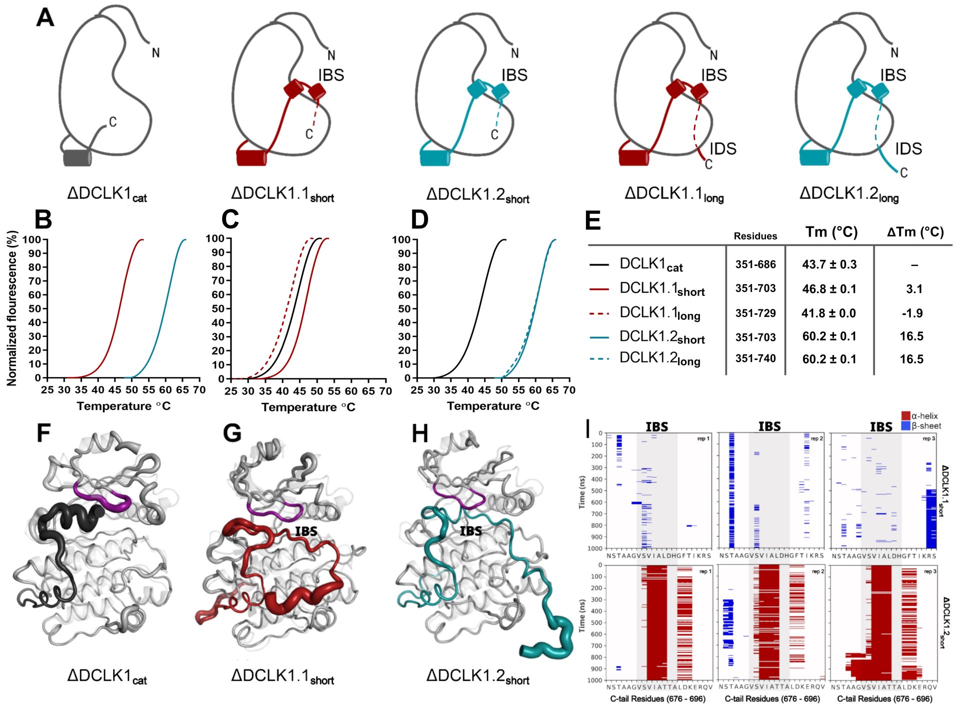


A
N
Catalytic Domain
15
16
17
DCLK1.1
N
Catalytic Domain
15
17'
DCLK1.2
gene
locus: 13q13.3
Exon 15
Exon 16
Exon 17
mRNA-DCLK1.1
15
phase 0 intron
16
phase 2 intron
17
STOP
mRNA-DCLK1.2
15
phase 0 intron
17'
STOP
B
DCLK1.1 reading frame
Pro Pro Leu Leu Ile Arg Arg Gly Arg Phe Ser Asp Glu Asp Ala Thr Arg Met STOP
Exon 17 (mRNA)
ACCACCGCUCUUGAUAAGGAGAGGCAGGUUUUCCGACGAAGACGCAACCAGGAUGUGAGGAGC
ACCACCGCUCUUGAUAAGGAGAGGCAGGUUUUCCGACGAAGACGCAACCAGGAUGUGAGGAGC
Thr Thr Ala Leu Asp Lys Glu Arg Gln Val Phe Arg Arg Arg Arg Asn Gln Asp Val Arg Ser
DCLK1.2 reading frame
C
Translated sequence
DCLK1.1
Exon 15
Exon 16
Exon 17
DCLK1.2
Exon 15
Exon 17'
alt-ORF
684
740
Exon 16
Exon 17
DCLK1.1 VIA...LDHGFTI.KRSGSLD....YYQQPGMYWIRPPLLIRRGRF...SDEDATRM.....
DCLK1.2 VIATTALDKERQVFRRRRNQDVRSRYKAQPA.....PPELNSESEDYSPSSSETVRSPNSPF
Exon 17'

Figure 3.4: A) Gene and intron-exon organization of DCLK1 human isoforms in the C-terminal tail. The DCLK1 gene is present on locus 13q13.3, and isoforms 1 and 3, contain an additional exon (exon 16), in the C-terminal tail that is absent in DCLK1.2. B) A phase 2 intron results in the alternative transcript of exon 17 in isoform 1, translating a different open-reading frame and early stop codon, resulting in the shorter sequence. C) Cartoon organization of the C-tail exons (exon 15, 16, and 17) of the DCLK1

3.3.3 Isoform-specific variations encode changes in molecular dynamics, thermostability and catalytic activity in DCLK1

To study isoform specific differences in the C-tail, we employed experimental techniques to compare protein stability and catalytic activity between purified DCLK1 proteins alongside molecular dynamics simulations for DCLK1.1 and DCLK1.2 with different tail lengths. Isoforms 1.1 and 1.2 share identical sequences across the kinase domain and within the first 38 residues of the C-tail, and we used this information to design a new recombinant protein, termed DCLK1cat (residues 351-686). C-terminal to this totally conserved region, both isoforms possess extended tail segments, which includes the putative inhibitory binding segment (IBS; residues 682-688) and an additional intrinsically-disordered segment (IDS; residues 703-end). To study the role of the C-tail in modulating kinase stability and activity, we purified the DCLK1cat, and C-tail containing (long and short) variants of each isoform, each of which lack the N-terminal DCX domains (Figure 3.3.5A). SDS-PAGE demonstrated that protein preparations were essentially homogenous after affinity and gel filtration chromatography (Figure 3.3.5-figure supplement 1). The short forms of the recombinant proteins (DCLK1.1351-703 and DCLK1.2351-703) possess a partially truncated C-tail and were designed to match the amino acid sequence previously used to solve the structure of DCLK1.2 protein (Cheng et al. 2022). Notably, these proteins exclude the IDS. The long forms of the DCLK1 proteins include the full-length C-tail for each isoform (DCLK1.1351-729 and DCLK1.2351-740) and incorporate IDS domains. We first performed comparative thermal shift analyses to quantify variance in thermal stability between the different purified proteins. When contrasting DCLK1.1short and DCLK1.2short which do not differ in size or tail length but encode unique sets of amino acids in their partially truncated C-tail as a result of alternative splicing (Figure 3.3.4), we observed that DCLK1.2 was some 14°C more stable than DCLK1.1 (Figure 3.3.5B). When compared with DCLK1cat, both DCLK1.1 short and long exhibited only subtle changes in thermal stability (Figure 3.3.5C & E), whereas both DCLK1.2 proteins (DCLK1.2short and DCLK1.2long) were significantly stabilized (relative ΔTm >16°C, Figure 3.3.5D & E).



	Residues	Tm (°C)	ΔTm (°C)
DCLK1$_{cat}$	351-686	43.7 ± 0.3	–
DCLK1.1$_{short}$	351-703	46.8 ± 0.1	3.1
DCLK1.1$_{long}$	351-729	41.8 ± 0.0	-1.9
DCLK1.2$_{short}$	351-703	60.2 ± 0.1	16.5
DCLK1.2$_{long}$	351-740	60.2 ± 0.1	16.5

Figure 3.5: Structural and dynamic variations between DCLK1 isoforms. A) Cartoons of DCLK1 construct used in our assays, portraying the locations of the Inhibitory Binding Segment (IBS) and the Intrinsically Disordered Segment (IDS). B-E) DSF thermal denaturation profiling of the purified DCLK1 core catalytic domain, or tail-matched DCLK1.1 and DCLK1.2 proteins. Unfolding curves and changes in Tm values (ΔTm) for each protein relative to WT DCLK1cat are indicated. F-H) B-factor structural representations of DCLK1short proteins shown in A). The width of the region indicates the extent of flexibility based on averaged RMSF data from three one microsecond MD replicates. I) DSSP analysis of three replicates of one microsecond MD simulations showing the residues surrounding the IBS in the C-tail of DCLK1.1short and DCLK1.2short. Blue indicates the presence of a Beta-sheet or Beta-bridge secondary structures and red indicates the presence of alpha-helical structures.

We next performed MD simulations to study the dynamics within the distinct DCLK1 C-tails that might explain the observed difference in protein stability. The crystal structure of DCLK1.2 (PDB: 6KYQ) was employed for the DCLK1.2short model and AlphaFold2 was used to model the other proteins (DCLK1cat and DCLK1.1short). Comparison of the root mean square fluctuations (RMSF) of the two isoforms in three different replicates of molecular dynamics simulations indicates strikingly different thermal fluctuations in the C-terminal tails and catalytic domains (Figure 3.3.5-source data 1). In particular, the IBS segment (between 682-688) is stably docked in the ATP binding pocket in DCLK1.2 and an alpha helical conformation is maintained during the microsecond time scale across different replicates (Figure 3.3.5I, bottom). In contrast, the IBS is more unstable in DCLK1.1, as indicated by high thermal fluctuations and a lack of secondary structure propensity (Figure 3.3.5I, top). A caveat to bear in mind is that DCLK1.1 is an AlphaFold2 model, which will also account for increased RMSF. Analysis of sequence variations and structural interactions provides additional insights into the differential dynamics of the two isoforms. The helical conformation of the IDS in DCK1.2 is maintained during the simulation due, in part, to a capping interaction with Thr 687, which is absent in DCLK1.1 due to the alternative splicing event detailed above. Likewise, another key residue in DCLK1.2, Lys 692, anchors the tail to the catalytic domain through directional salt bridges with the conserved aspartates (Asp 511 and Asp 533) in the HRD and DFG motifs (Figure 3.3.5, figure-supplement 2A). These interactions are not observed in DCLK1.1 simulations because Lys 692 is substituted to a histidine (His 689), which is unable to form a corresponding interaction with the catalytic domain (Figure 3.3.5, figure-supplement 2B). We also evaluated the effects of T688A (non-phosphorylated) or T688E (phosphomimetic) mutations through DCLK1 MD simulations and found that the two mutations slightly destabilize the tail relative to WT. Three replicates of the two mutants show increased RMSF of the tail region relative to WT DCLK1.1 (Figure 3.3.5, figure-supplement 3). Either mutation was not sufficiently destabilizing on its own to unlatch the C-tail, and we hypothesize that other residues in addition to T688 are also likely to be important for contributing to conformational regulation of the kinase domain by the C-terminal tail. The variable docking of the C-tail within the kinase domains of the two DCLK1 isoforms, and the extent to which this contributes to more transient or stable autoinhibited states are explored in more detail in the next section.

3.3.4 Residues contributing to the co-evolution and unique tethering of the C-terminal tail to the DCLK catalytic domain

To identify specific residues that contribute to the unique modes of DCLK regulation by the C-terminal tail, we performed statistical analysis of the evolutionary constraints acting on DCLK and related CAMK family sequences. We aligned the catalytic domain of DCLK and related CAMK sequences from diverse organisms and employed the Bayesian Partitioning with Pattern Selection (BPPS) method (30) to identify residues that most distinguish DCLK sequences (foreground alignment in Figure 3.3.6B) from CAMK sequences (background alignment). Beyond the catalytic domain, DCLKs share sequence and structural similarities in the first helix of the tail (αR1 in CAMK1) (31,32), with other CAMKs but share no detectable sequence similarity beyond this helical segment. DCLKs also share a CAMK-specific insert segment located between F and G helices in the catalytic domain, although the nature of residues conserved within the insert is unique to individual CAMK families (Figure 3.3.6-figure supplement 1). BPPS analysis revealed DCLK-specific constraints in different regions of the kinase domain, most notably, the ATP binding G-loop, N terminus of the C-helix, the activation loop, and C-terminus of the F-helix (Figure 3.3.6-figure supplement 2).

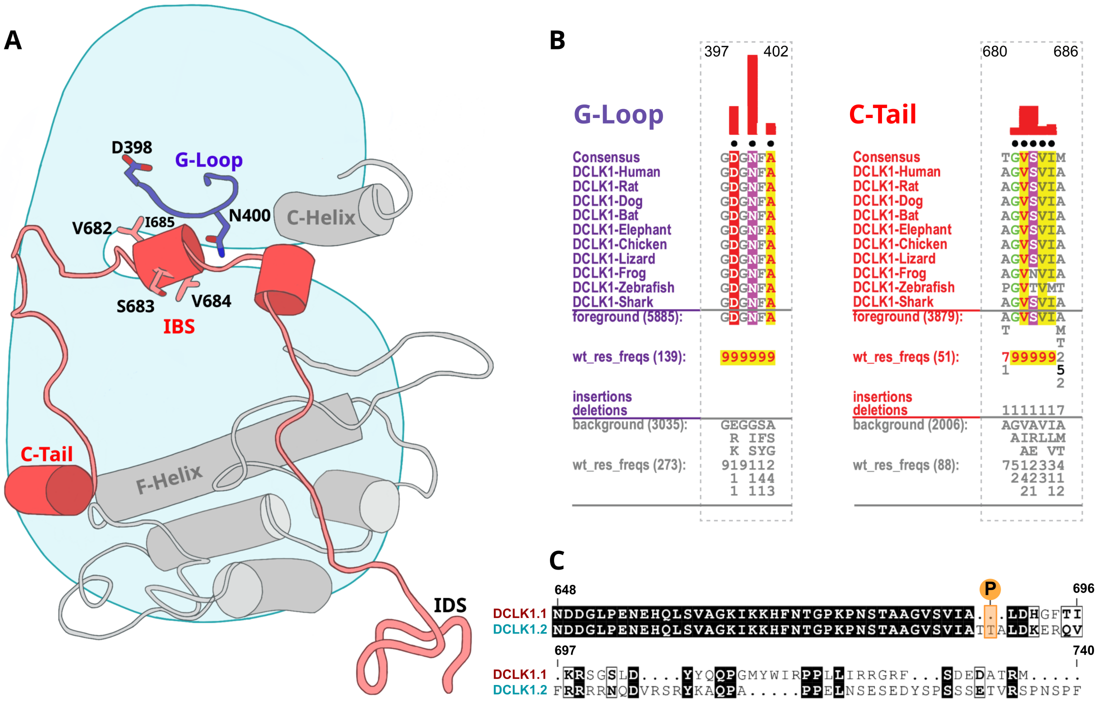


Figure 3.6: Identification of DCLK specific constraints. A) Cartoon of DCLK1.2 and the intrinsically disordered segment (IDS) with evolutionary constraints mapped to the kinase domain and C-tail. B) Sequence constraints that distinguish DCLK1/2/3 sequences from closely related CAMK sequences are shown in a contrast hierarchical alignment (CHA). The CHA shows DCLK1/2/3 sequences from diverse organisms as the display alignment. The foreground consists of DCLK sequences while the background alignment contains related CAMK sequences. The foreground and background alignments are shown as residue frequencies below the display alignment in integer tenths (1–9). The histogram (red) indicates the extent to which distinguishing residues in the foreground diverge from the corresponding position in the background alignment. Black dots indicate the alignment positions used by the BPPS (Neuwald, 2014) procedure when classifying DCLK sequences from related CAMK sequences. Alignment number is based on the human DCLK1.2 sequence (UniProt ID: O15075-2). C) Sequence alignment of human DCLK1 isoforms.

Some of the most significant DCLK specific residue constraints map to the ATP binding G-loop (GDGNFA motif) (Figure 3.3.6A-B). In particular, Asp 398, Asn 400 and Ala 402 are unique to DCLKs as the corresponding residues are strikingly different in other CAMKs. Asp 398 is typically a charged residue (K/R) in other CAMKs while Asn 400 and Ala 402 are typically hydrophobic and polar residues, respectively (see residue frequencies in background alignment; Figure 3.3.6B). Notably, both Asn 400 and Asp 398 make direct interactions with residues in the C-tail either in the crystal structure or molecular dynamics simulations (see below). Likewise, DCLK conserved residues in the C-helix and activation loop tether the C-terminal tail to functional regions of the kinase core, suggesting co-option of the DCLK catalytic domain to uniquely interact with the flanking cis regulatory tail.

3.3.5 An autoinhibitory ATP-mimic completes the C-spine and mimics the gamma phosphate of ATP

The most stable segment of the C-tail based on the B-factor and RMSF fluctuations in MD simulations is a unique region (682-688 in DCLK1.2) that docks into the ATP binding pocket through both hydrophobic and hydrogen-bonding interactions. Remarkably, this peptide segment mimics the physiological ATP ligand, and stabilizes the catalytic domain through 'completion' of the hydrophobic catalytic spine (Figure 3.3.7A, (33)). The residues that mimic adenosine and complete the C-spine of DCLK1 are Val 682, Val 684, and Ile 685 (PDB: 6KYQ), which are part of the α-helix that docks to the ATP-binding pocket (Figure 3.3.7B). Interestingly, based on our BPPS analyses, these C-tail residues are uniquely vertebrate DCLK1-specific pattern constraints. At the tail end of this α-helix are two Thr residues, Thr 687 and Thr 688. As previously noted, these Thr residues mark the beginning of exon 17, and are one of the key variations between human DCLK1 isoforms, found only in DCLK1.2 variants.

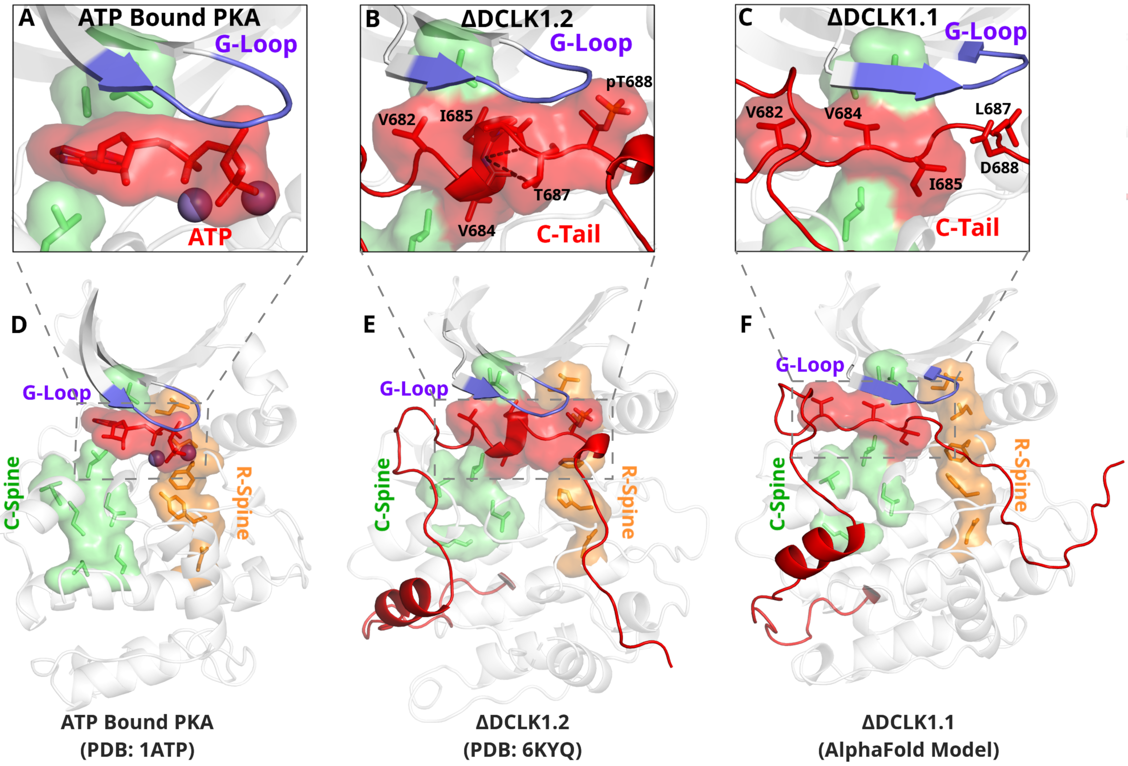
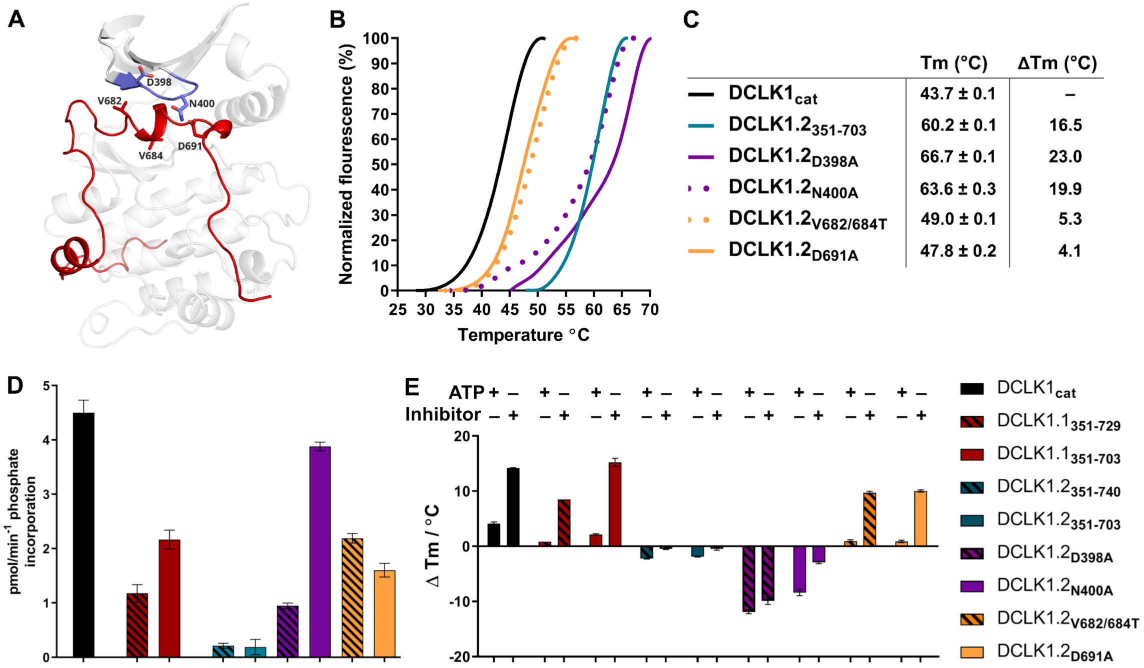


Figure 3.7: The DCLK1 C-tail 'completes' the regulatory C-spine (green). A) PKA crystal structure (pdb: 1ATP) with bound ATP in red and Mg2+ in purple. The C-spine is completed by the adenine ring of ATP. The gamma phosphate of ATP hydrogen bonds with the second glycine of the G-loop. B) DCLK1.2 crystal structure (pdb: 6KYQ) showing how the C-tail (red) docks underneath the pocket and mimics the ATP structure. The C-spine is completed by V682 and V684 in the C-tail and helical segments defined using DSSP are shown. T687 is also depicted making multiple hydrogen bonds with the backbone of V684 and I685 (dashed lines). C) DCLK1.1 AlphaFold2 model showing an unstructured loop in the C-tail docking into the ATP binding pocket, where V684 and I685 are predicted to complete the C-spine. The average per-residue confidence of the C-tail is 49%. D-F) Zoomed out versions of A-C, demonstrating how the DCLK1 C-tail docks into the ATP binding cleft, akin to ATP in PKA.

Structural analysis and MD simulations reveal that Thr 687 in DCLK1.2 caps the stable α-helix that extends the C-spine (Figure 3.3.7B, Figure 3.3.7-figure supplement 1B). In comparison, the same region in DCLK1.1, which lacks Thr 687, is predicted to be unstructured. Upon phosphorylation, Thr 688 in DCLK1.2 can mimic the gamma phosphate of ATP by maintaining a stable hydrogen bonding distance with the backbone of the second glycine of the G-loop (G399) (Figure 3.3.7B, Figure 3.3.7-figure supplement 1C). We additionally observe that the sidechain of Asn 400, a DCLK-specific G-loop constraint, further stabilizes the phosphate group in pThr 688 through hydrogen bonding. As previously described, Thr 688 is unique to DCLK1.2. The lack of this functional site in DCK1.1 is correlated with increased RMSF and instability of the ATP-mimic segment in isoform 1 MDs (Figure 3.3.5C-D, Figure 3.3.7-figure supplement 1B). Comparatively, MD analysis of DCLK1.2 and a phosphothreonine-containing DCLK1.2 demonstrates reduced C-tail fluctuations, suggesting the potential regulatory involvement of Thr 688 phosphorylation for further modulation of the autoinhibited conformation (Figure 3.3.7-figure supplement 1C), consistent with previous findings (Agulto et al. 2021).

3.3.6 Mutational analysis support isoform-specific allosteric control of catalytic activity by the C-terminal tail

To evaluate how sequence differences between DCLK1.1 and DCLK1.2 affected both thermal stability and catalytic potential, we generated targeted mutations at contact residues within the Gly-rich loop and C-tail of DCK1.1 and DCLK1.2 (at the indicted residues depicted in Figure 3.3.8A) which we predicted would disrupt or destabilize C-tail docking within the domain. All proteins were purified to near homogeneity by IMAC and size exclusion chromatography (Figure 3.3.5-figure supplement 1), and the thermal stability of a panel of DCLK1.2 mutant and WT proteins were compared side-by-side with the DCLK1cat (Figure 3.3.8B). The ΔTm values obtained (Figure 3.3.8C) demonstrate that mutation of Asp 398 or Asn 400 in the Gly-rich loop are by themselves insufficient to destabilize DCLK1.2. In marked contrast, dual mutation of the hydrophobic pair of Val 682 and Val 684 residues to Thr, or mutation of the acidic tail residue Asp 691, resulted in a pronounced reduction in DCLK1.2 thermal stability. Moreover, the recorded Tm values for these latter two mutations quite closely resembled the Tm of DCLK1cat (which lacks the C-tail entirely), which is consistent with the uncoupling of the C-tail and a commensurate decrease in thermal stability associated with loss of this interaction.



	Tm (°C)	ΔTm (°C)
DCLK1$_{cat}$	43.7 ± 0.1	–
DCLK1.2$_{351\text{-}703}$	60.2 ± 0.1	16.5
DCLK1.2$_{D398A}$	66.7 ± 0.1	23.0
DCLK1.2$_{N400A}$	63.6 ± 0.3	19.9
DCLK1.2$_{V682/684T}$	49.0 ± 0.1	5.3
DCLK1.2$_{D691A}$	47.8 ± 0.2	4.1

Figure 3.8: A) Structural depiction of DCLK1.2 (PDB: 6KYQ) showing the location of modified DCLK1 amino acids on the G-loop (purple) or C-tail (red). B-C) Differential Scanning Fluorimetry assays depicting thermal denaturation profiles of each protein along with the calculated Tm value. D) Kinase assays. DCLK1-dependent phosphate incorporation (pmol/min-1) into the DCLK1 peptide substrate was calculated for DCLK1cat, long and short DCLK1.1 and the indicated DCLK1.2 variants. E) Thermal stability analysis in the presence of ATP or DCLK1-IN-1 for DCLK1 proteins. For DCLK1.2, all proteins were generated in the DCLK1.2 short background.

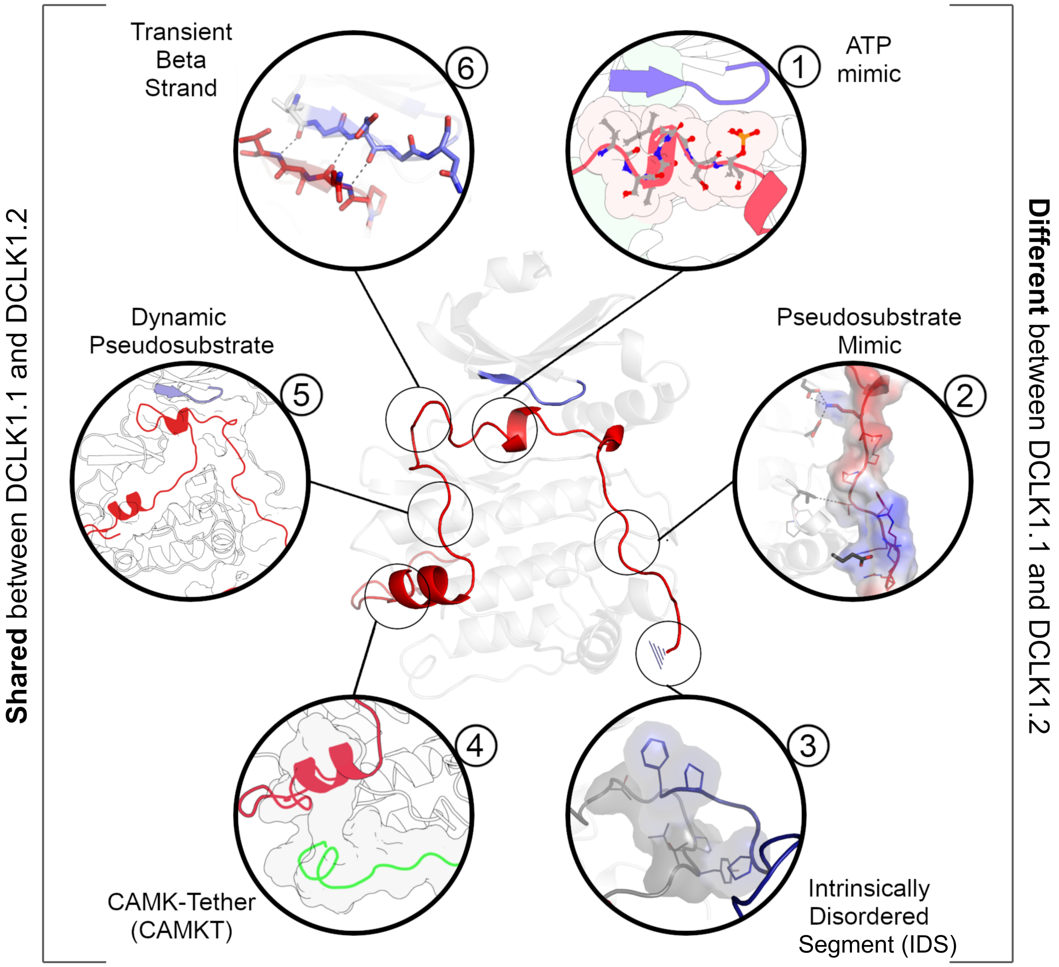
We next determined the catalytic activity of our recombinant DCLK1.1 and DCLK1.2 proteins side-by-side (Figure 3.3.8D, Figure 3.3.8-figure supplement 1). Although partially diminished in relation to DCLK1cat, both DCLK1.1short (351–703) and DCLK1.1long (351–729) variants possess robust catalytic activity. This suggested ineffective ATP-competitive auto-inhibition mediated by the C-tail segment of DCLK1.1 and is consistent with their closely matched Tm values to DCLK1cat (Figure 3.3.5C). Interestingly, both C-tail containing variants of DCLK1.1 (and particularly DCLK1.1351-729) exhibited lower affinity for ATP (inferred from KM[ATP] for peptide phosphorylation), which is consistent with partial-occlusion of the ATP binding pocket (Figure 3.3.8-figure supplement 1). In marked contrast, the detectable kinase activity for short (351–703) or long (351–740) DCLK1.2 proteins was significantly blunted compared to DCLK1cat, exhibiting just ≈5% of the activity of the catalytic domain alone, and consistently, the calculated KM[ATP] was ≈4 fold higher compared to the catalytic domain lacking the C-tail. We also utilized autophosphorylation as a proxy for overall kinase activity. Quantitative tandem mass spectrometry (MS/MS) analysis of site-specific autophosphorylation within DCLK1.1short and DCLK1.2short demonstrate a marked reduction in the site-specific abundance of phosphate in DCLK1.2 when compared to DCLK1.1 at two separate sites that could be directly and accurately quantified by MS (S438 and S660, DCLK1.1 relative abundance set to 1, Figure 3.3.8-figure supplement 2). LC-MS/MS also indicated that several autophosphorylation sites identified in isoform 1 were absent in DCLK1.2 (Ser 683 and Thr 692, the latter of which is an amino acid that is unique to the C-tail of DCLK1.1, Figure 3.3.8-figure supplement 2). Interestingly, amino acid substitutions in the G loop or the C-tail of DCLK1.2 designed to subvert C-tail and ATP site interactions also had major effects on DCLK1.2 phosphorylation and catalytic activity. DCLK1.2 D398A was activated some 5-fold when compared to the WT form, whereas DCLK1.2 N400A was almost as active as the DCLK1cat.

Consistently, DCLK1.2 V682T/V684T and D691A proteins were also much more active than the WT form of DCLK1.2. Kinetic analysis also confirmed higher Vmax (but broadly similar KM[ATP]) values for DCLK1.2 D398A and V682T/V684T relative to the WT protein (Figure 3.3.8-figure supplement 1). Moreover, comprehensive LC-MS/MS phosphosite mapping revealed a marked increase in the total number of phosphorylated amino acids in all of the mutant DCLK1.2 proteins, consistent with the enhanced catalytic activity of these proteins when compared to WT DCLK1.2 (Figure 3.3.8-figure supplement 2). Together, these observations confirm that targeted mutations are sufficient to relieve ATP-competitive C-tail autoinhibition by physical tail uncoupling; this model is also strongly supported by marked changes in the biophysical stability of the mutant proteins, particularly for V682/684T and D691A (Figure 3.3.8B).

To expand on this finding, we investigated the interaction of Mg:ATP or DCLK1-IN-1 to our panel of DCLK1.1 and 1.2 proteins (Figure 3.3.8E), using changes in thermal stability as a reporter of ligand binding. DCLK1cat (351–686), DCLK1.1short (351–703) and DCLK1.1long (351–729) proteins all behaved similarly in the presence of either Mg:ATP or DCLK1-IN-1, inducing marked stabilization. In contrast, DCLK1.2short (351–703) or DCLK1.2long (351–740) proteins registered negligible thermal shifts in the presence of the same concentration of either ligand, which is in-line with the C-tail tightly occupying the ATP-binding site and obstructing their binding. Remarkably, D398A and N400A DCLK1.2, whose high basal Tm values (compared to DCLK1cat) are consistent with stabilization by docking of the C-tail within the kinase domain, were markedly destabilized in the presence of either ligand. This suggests that incorporation of these G-loop mutations in isolation is insufficient to dislodge the C-tail, but rather that the stability of the interaction is compromised to the extent that either ATP or DCLK-IN-1 can competitively dislodge the bound tail from the ATP active site, resulting in a net destabilization caused by lack of tail engagement. This observation is corroborated by the results of our kinase assays, where both D398A and N400A mutants were more active than WT DCLK1.2, confirming appropriate ATP binding (a pre-requisite for catalysis). Finally, DCLK1.2 V682T/V684T and D691A, which exhibit lower basal Tm values than the WT protein (indicating a loss of tail interaction), were both stabilized in a ligand-dependent manner to a similar degree to that observed for DCLK1cat and DCLK1.1 proteins (Figure 3.3.8E). Collectively, these observations clearly demonstrate that the C-tail section of DCLK1.2 can both stabilize the canonical DCLK kinase domain and inhibit kinase activity (by impeding the binding and structural coordination of Mg:ATP) much more effectively than that of DCLK1.1. Our targeted mutational analysis of key contact residues in DCLK1.2 also clearly shows that this is a consequence of specific amino acid interactions that are absent in DCLK1.1 due to alternative-splicing and subsequent sequence variation.

3.3.7 Classification of DCLK regulatory segments

We synthesized our experimental findings by classifying the DCLK C-tail into six functional segments, based on interactions with different regions of the catalytic domain and their conservation between the C-tail splice variants in our analysis (Figure 3.3.9): First is an ATP-mimetic peptide segment (residues 682-688 in DCLK1.2) that readily mimics physiological ATP binding by completing the C-spine in the nucleotide-binding site. The inhibitory peptide also contains a phosphorylatable Thr residue, which sits adjacent to the highly characteristic Gly-rich loop (GDGNFA, residues 396-402). Second, we define a pseudosubstrate mimic (PSM, residues 692-701), which interacts with the acidic HRD and DFG Asp side-chains and docks in the substrate pocket occluding substrate access. Third, at the C-terminus of the tail, lies an intrinsically disordered segment (IDS, residues 702-749, Figure 3.3.9-figure supplement 1), which packs dynamically against DCLK conserved residues in the kinase activation loop. Fourth, at the beginning of the C-tail lies a CAMK-tether (residues 654-664), a set of residues that pack against a CAMK-specific insert in the C-lobe. In many CAMK crystal structures, this insert makes multiple contacts with the F-helix and C-tail (Figure 3.3.6-figure supplement 1). Fifth, this is followed by a highly dynamic pseudosubstrate region (residues 672-678) that occludes the substrate pocket and will thus interfere with substrate phosphorylation. Sixth, a transient beta-strand is formed in DCLK1.2 through amino acid specific sequences that help modulate and potentially strengthen binding of the C-tail in this isoform (Figure 3.3.7-figure supplement 1A-B).



Transient Beta Strand
6
ATP mimic
1
Dynamic Pseudosubstrate
5
Pseudosubstrate Mimic
2
CAMK-Tether (CAMKT)
4
Intrinsically Disordered Segment (IDS)
3
Shared between DCLK1.1 and DCLK1.2
Different between DCLK1.1 and DCLK1.2

Figure 3.9: A DCLK1 C-tail can act as a multi-functional Swiss Army Knife, using six distinct segments for a variety of regulatory functions including mimicking ATP binding/association, stabilizing the G-loop, occluding the substrate binding pocket, and packing against the kinase activation loop.

Collectively, these segments and their associated interaction sites demonstrate that co-evolution of the unique C-tail with the catalytic domain is the central hallmark of DCLK functional divergence, and that changes in these segments possess the ability to 'supercharge' catalytic output of the kinase. In particular, the variable C-terminal segments of the tail might contribute to isoform-specific functional specialization. The combinatorial diversity of events that modulate C-tail function may allow DCLKs to nimbly coordinate various tasks including ATP-binding, substrate-based phosphorylation, regulation of DCX domain phosphorylation and structural disposition, kinase autoinhibition and allosteric regulation. Isoform-specific variability provides additional nuance to regulatory and catalytic signaling events and may even contribute to differences in cellular localization (e.g. cytoplasm or nucleoplasm) and tissue-specific activity, enabling contextual DCLK regulation through these modular sequence segments.

3.4 Methods

3.4.1 Ortholog Identification

To identify orthologs, we used the software KinOrtho (L.-C. Huang et al. 2021) to query one-to-one orthologous relationships for DCLK1/2/3 across the proteome. After collation of the various orthologs, we parsed the sequence data for taxonomic information and classified each sequence by family. We further separated human DCLK1 into each unique isoform and aligned them.

3.4.2 Phylogenetic Analysis

We identified diverse DCLK orthologs from the UniProt database (UniProt Consortium 2021) using an profile-based approach (Andrew F. Neuwald 2009). From this dataset, we manually curated a taxonomically diverse set of DCLK orthologs composed of 36 sequences spanning 16 model organisms. These sequences were used to generate a maximum-likelihood phylogenetic tree using IQTREE version 1.6.12 (L.-T. Nguyen et al. 2015). Branch support values were generated using ultrafast bootstrap with 1000 resamples (Hoang et al. 2018). The consensus tree was selected as the final tree. The optimal substitution model for our final topology was determined to be LG (Le and Gascuel 2008) with invariable sites and discrete gamma model (Gu, Fu, and W. H. Li 1995) based on the Bayesian Information Criterion as determined by ModelFinder (Kalyaanamoorthy et al. 2017). We rooted our final tree against an outgroup of 17 closely related human CAMK kinases using ETE Toolkit version 3.1.2 (Huerta-Cepas, Serra, and Bork 2016).

3.4.3 Sequence and Structure Analysis

MAFFT (Katoh et al. 2002) generated multiple sequence alignments were fed into the Bayesian Partitioning with Pattern Selection (BPPS) tool to determine evolutionarily conserved and functionally significant residues (Andrew F Neuwald 2014). Constraints mapped onto AlphaFold-predicted structures were visualized in PyMOL to analyze biochemical interactions.

3.4.4 Rosetta Loop Modeling

Loop modeling was performed on the crystal structure (6KYQ) using the Kinematic Loop Modeling protocol (Mandell, Coutsias, and Kortemme 2009) to model missing residues. Following this, the structure underwent five cycles of rotamer-repacking and minimization using the Rosetta Fast-relax protocol (Tyka et al. 2011).

3.4.5 DSSP Analysis

To analyze changes in secondary structure over our MDs, we employed the DSSP command in GROMACS (**kabsch_dictionary_1983**). This produces an output that contains an array of secondary structure values against each residue. The MDAnalysis python module (Michaud-Agrawal et al. 2011) was used to plot these values.

3.4.6 Molecular Dynamics

PDB constructs were generated by retrieving structural models RCSB and the AlphaFold2 database. Post-translational modifications were performed in PyMOL using the PyTMs plugin. All structures were solvated using the TIP3P water model (Jorgensen et al. 1983). Energy minimization was run for a maximum of 10,000 steps, performed using the steepest-descent algorithm, followed by the conjugate-gradient algorithm. The system was heated from 0K to a temperature of 300K. After two equilibration steps that each lasted 20 picoseconds, 1 microsecond long simulations were run at a two femtosecond timestep. Long-range electrostatics were calculated via particle mesh Ewald (PME) algorithms using the GROMACS MD engine (Pronk et al. 2013). We utilized the CHARMM36 force field (J. Huang and MacKerell 2013). The resulting output was visualized using VMD 1.9.3 (Humphrey, Dalke, and Schulten 1996). All molecular dynamics analysis was conducted using scripts coded in Python using the MDAnalysis module (Michaud-Agrawal et al. 2011).

3.4.7 Computational Mutational Analysis

Cartesian ddG in Rosetta (Park et al. 2016) was utilized to predict potential stabilizing and destabilizing mutations in the enzyme structure. We performed three replicates per mutation and averaged the Rosetta energies. All mutant energies were then subtracted by the wt Rosetta energy to generate a panel of ddG values relative to wt. Combined with our sequence analyses, we mutated kinase and DCLK-specific constraints to identify destabilizing interactions in the c-tail.

3.4.8 Exon-Intron Boundary Mapping

The precise gene structure of DCLK1 isoforms were mapped onto the human genome with each isoform used as a query protein sequence in order to generate exon-intron borders. This was achieved using Scipio (version 1.4.1) (Keller et al. 2008) with default settings. Exons were numbered based on Ensembl annotations (Cunningham et al. 2022). The translation of each annotated gene sequence to protein sequence was provided with the output file (Figure 3.3.4-source file 1-4).

3.4.9 DCLK1 cloning and recombinant protein expression

6His-DCLK1 catalytic domain (351–686), DCLK1.1 351-703 (short C-tail) or 351-729 (full C-tail), DCLK1.2 351-703 (short C-tail) or 351-740 (full C-tail), and DCLK1.2 351-703 containing D398A, N400A, V682T/V684T or D691A substitutions were synthesized by Twist Biosciences in pET28a. 6His-GST-(3C) DCLK1.1 351-689 was amplified by PCR and cloned into pOPINJ. Kinase dead, D533A 6His-GST-(3C) DCLK1.1 351-689 was generated by PCR-based site directed mutagenesis (Figure 3.3-source data 3). All plasmids were sequenced prior to their use in protein expression studies. All proteins, including 6His-GST-(3C) DCLK1 351-689, with a 3C-protease cleavable affinity tag, were expressed in BL21(DE3)pLysS E. coli (Novagen) and purified by affinity and size exclusion chromatography. The short N-terminal 6-His affinity tag present on all other DCLK1 proteins described in this paper was left in situ on recombinant proteins, since it does not appear to interfere with DSF, biochemical interactions or catalysis. For analytical SEC chromatography, 1 mg of each DCLK1 protein was assayed on a Superdex 200 Increase 10/300 GL (Cytiva), and the eluted fractions were also analysed by SDS-PAGE and Coomassie blue staining to confirm composition. The molecular weight standards were loaded in a mixture of 200 ug of Bovine Serum Albumin (BSA), Carbonic Anhydrase (CA), and Alcohol Dehydrogenase (AD) each.

3.4.10 Mass Spectrometry

Purified DCLK1 proteins (5 μg) were diluted ($\approx$40-fold) in 100 mM ammonium bicarbonate pH 8.0 and reduced (DTT) and alkylated (iodoacetamide, as previously described (Ferries et al. 2017), and digested with a 25:1 (w/w) trypsin gold (Promega) at 37 °C for 18 hours with gentle agitation. Digests were then subjected to strong cation exchange chromatography using in-house packed stage tip clean-up (Leonard A. Daly et al. 2021). Dried tryptic peptides were solubilized in 20 μl of 3% (v/v) acetonitrile and 0.1% (v/v) TFA in water, sonicated for 10 min, and centrifuged at 13,000 x g for 10 min at 4°C and supernatant collected. LC-MS/MS separation was performed using an Ultimate 3000 nano system (Dionex), over a 60-min gradient (Ferries et al. 2017). Briefly, samples were loaded at a rate of 12 μL/min onto a trapping column (PepMap100, C18, 300 μm $\times$ 5 mm) in loading buffer (3% (v/v) acetonitrile, 0.1% (v/v) TFA). Samples were then resolved on an analytical column (Easy-Spray C18 75 μm $\times$ 500 mm, 2 μm bead diameter column) using a gradient of 97% A (0.1% (v/v) formic acid): 3% B (80% (v/v) acetonitrile, 0.1% (v/v) formic acid) to 80% B over 30 min at a flow rate of 300 nL/min. All data acquisition was performed using a Fusion Lumos Tribrid mass spectrometer (Thermo Scientific). Samples were injected twice with either higher-energy C-trap dissociation (HCD) fragmentation (set at 32% normalized collision energy [NCE]) or Electron transfer dissociation (ETD) with supplemental 30% NCE HCD (EThcD) for 2+ to 4+ charge states using a top 3s top speed mode. MS1 spectra were acquired at a 120K resolution (at 200 m/z), over a range of 300 to 2000 m/z, normalised AGC target = 50%, maximum injection time = 50 ms. MS2 spectra were acquired at a 30K resolution (at 200 m/z), AGC target = standard, maximum injection time = dynamic. A dynamic exclusion window of 20 s was applied at a 10 ppm mass tolerance. Data was analysed by Proteome Discoverer 2.4 in conjunction with the MASCOT search engine using a custom database of the UniProt Escherichia coli reviewed database (Updated January 2023) with the DCLK1 mutant variant amino acid sequences manually added, and using the search parameters: fixed modification = carbamidomethylation (C), variable modifications = oxidation (M) and phospho (S/T/Y), MS1 mass tolerance = 10 ppm, MS2 mass tolerance = 0.01 Da, and the ptmRS node on; set to a score $>$ 99.0. For HCD data, instrument type = electrospray ionization–Fourier-transform ion cyclotron resonance (ESI-FTICR), for EThcD data, instrument type = EThcD. For label free relative quantification of phosphopeptide abundances of the different DCLK1 variants, the minora feature detector was active and set to calculate the area under the curve for peptide m/z ions. Abundance of phosphopeptide ions were normalised against the total protein abundance (determine by the HI3 method (Silva et al. 2006), as in the minora feature detector node) to account for potential protein load variability during analysis.

3.4.11 DCLK1 DSF

Thermal Shift Assays (TSA), were performed using Differential Scanning Fluorimetry (DSF) in a StepOnePlus Real-Time PCR machine (Life Technologies) in combination with Sypro-Orange dye (Invitrogen) and a thermal ramping protocol (0.3°C per minute between 25 and 94°C). Recombinant DCLK1 proteins were assayed at a final concentration of 5 μM in 50 mM Tris–HCl (pH 7.4) and 100 mM NaCl in the presence or absence of the indicated concentrations of ligand (ATP or Mg:ATP) or DCLK1 inhibitor compounds, with final DMSO concentrations never higher than 4% (v/v). Thermal melting data were processed using the Boltzmann equation to generate sigmoidal denaturation curves, and average Tm/ΔTm values were calculated as described using GraphPad Prism software, as previously described, from 3 technical repeats (Dominic P. Byrne, Clarke, et al. 2020).

3.4.12 DCLK1 kinase assays

DCLK1 peptide-based enzyme assays (Dominic P. Byrne, Vonderach, et al. 2016; Omar et al. 2023) were carried out using the LabChip EZ Reader platform, which monitors and quantifies real-time phosphorylation-induced changes in the mobility of the fluorescently-labelled DCLK1 peptide substrate 5-FAM-KKALRRQETVDAL-CONH2. To assess DCLK1 catalytic domains, or DCLK1.1 or DCLK1.2 variants, 100ng of purified protein were incubated with a high (1 mM) concentration of ATP (to mimic cellular levels of nucleotide) and 2 μM of the fluorescent substrate in 25 mM HEPES (pH 7.4), 5 mM MgCl2, and 0.001% (v/v) Brij 35. DCLK1-IN-1 and DCLK1-NEG (kind gifts from Dr Fleur Ferguson, UCSF) enzyme inhibition was quantified under identical assay conditions in the presence of 10 μM of each compound. Assays are either reported as rates (pmoles/min phosphate incorporation) during linear phosphate incorporation (e.g total substrate phosphorylation limited to <20-30% to prevent ATP depletion and to ensure assay linearity), or presented as time-dependent percentage substrate phosphorylation (kinetic mode). Rates of substrate phosphorylation (pmol phosphate incorporation per min) were determined using a fixed amount of kinase and linear regression analysis with GraphPad Prism software; Vmax and KM[ATP] values were calculated at 2 μM substrate peptide concentration, as previously described (McSkimming et al. 2016). Rates are normalized to enzyme concentration and all enzyme rate and kinetic data are presented as mean and SD of 4 technical replicates.

3.5 Discussion

The kinase domain is a conserved switch for phosphorylation-based catalytic regulation. Yet the complexity of cell signaling pathways demands other nuanced forms of regulation beyond binary "on" or "off" switch-based mechanisms. For many Ser/Thr kinases, including AGC and CAMK families, these distinct regulatory functions come from segments which flank the kinase domain, N– and C-terminal regions, which serve to modulate activity through allosteric activation, inhibition, or rheostatic behaviors that change based on environmental conditions (Gógl et al. 2019). In this study, we expand on our knowledge of allosteric diversity in the human kinome by revealing how alternative splicing of the DCLK1 C-tail contributes to isoform-specific behaviors, coupling regulation of catalytic output, phosphorylation, protein dynamics and stability, substrate binding, and protein-protein interactions. Our "Swiss Army Knife" model for DCLK1 expands our view of allosteric regulation as not just a dynamic process facilitated by proteins, but one where adaptive genetic mechanisms, like differential splicing, dexterously tune isoform-specific functions for specific cellular signaling roles; in the case of DCLK1.1, this allows 'supercharging' of catalysis between splice variants due to key amino acid differences in the C-tail that are lacking in the DCLK1.2 isoform.

Multiple members of the human kinome have independently evolved C-tail regions that dock to the N or C-lobe of the kinase domain in cis. In the case of the AGC kinases, the C-terminal tail is a very well-studied in-cis modulatory element that serves to explain a variety of regulatory properties in this kinase sub-family (Kannan et al. 2007; Romano et al. 2009; Baffi and Newton

2022; Susan S. Taylor and Alexandr P. Kornev 2011). Classical deletion studies with members of the CAMK family, have also revealed a cis-acting inhibitory element lying C-terminal to the catalytic domain of both CAMK1 (Yokokura et al. 1995) and CAMK2 (Yang and H. Schulman 1999). More recent examples of C-tail functional diversity in CAMK family members are presented by the human pseudokinases TRIB1 and TRIB2, which employ C-tail sequences to either latch (and structurally restrict) the atypical kinase domain or to bind competitively to the Ubiquitin E3 ligase COP1 (Patrick A. Eyers 2015; Patrick A. Eyers, Keeshan, and Kannan 2017). Functional disengagement of the TRIB1 or TRIB2 tail through deletion, mutagenesis or small molecule binding has marked effects on pseudokinase conformation, intrinsic stability and cellular transformation (Foulkes et al. 2018; Harris et al. 2022; Keeshan et al. 2010; James M. Murphy et al. 2015).

3.5.1 A novel pseudosubstrate region encoded by DCLK1

In addition to the marked differences between DCLK1 splice variants relevant to nucleotide binding, small molecule interactions and catalysis, our work also reveals two unique pseudosubstrate segments present before and after the IBS. Before the IBS segment, we observe the formation of an anti-parallel transient beta sheet with the beta1 strand in the catalytic domain (Figure 3.3.5I, Figure 3.3.7-figure supplement 1A-B). During the formation of this transient structure, the C-tail dynamically occludes part of the substrate binding pocket. A beta-sheet is observed in all three MD replicates of DCLK1.1, but only in a single replicate of DCLK1.2 dynamic analysis. At the other end of the IBS is another pseudosubstrate segment whose structure and dynamics change as a result of alternative exon splicing. In DCLK1.2, the pseudosubstrate segment is stable, with an average RMSF of 1.8 Angstroms, facilitated by key interactions from Lys 692, which coordinates acidic residues in the HRD and DFG motifs (Figure 3.3.5-figure supplement 2A). In DCLK1.1, Lys 692 is replaced by a His, which weakly coordinates with the HRD and DFG motifs, resulting increased dynamics of the segment and an RMSF of 3.1Å (Figure 3.3.5, Figure 3.3.5-figure supplement 2B). Together, residue variation between isoforms contribute to differences in stability and alteration of dynamics of the tail. HPCAL1 was recently proposed as a possible substrate that activates DCLK1 in a calcium-dependent manner, but it is unclear how it may bind DCLK1 (Patel et al. 2021). Because only exon 15 of the C-tail is conserved between the isoforms, it is possible the location of binding occurs in this dynamic pseudosubstrate segment prior to the IBS, where increased flexibility and occlusion of the substrate pocket is reflective of the absence of HPCAL1, or a similar calmodulin-like substrate.

3.5.2 Discovery of the DCLK1 ATP-Mimic region; a splice-variant specific regulatory module

Our structural analysis of DCLK1 reveals a remarkable structural mimic of ATP located in the C-tail, which differs markedly between DCLK1.1 and DLCK1.2 splice variants. We note for the first time that a set of three residues in the ATP-mimic, Val 682, Val 684, and Ile 685, are conserved across all isoforms of DCLK1 and DCLK2 (Figure 3.1C) and serve to extend the kinase C-spine. Mutation of these residues in DCLK1.2 uncouples tail binding and activates the kinase. Proximal to these residues are two Thr residues (Thr 687 and Thr 688), which are present in DCLK1.2, but absent in DCLK1.1. Based on published phosphoproteomics data, both Thr residues can be phosphorylated (2) and are thus likely to be regulatory in DCLK1.2. Although we could not detect phosphorylation of either of these predicted regulatory sites in the WT form of DCLK1.2, we consistently observed pThr 688 in activated mutant DCLK1.2 variants (Figure 3.3.8-figure supplement 2). By analyzing DCLK1 dynamics using MD simulations, we observed multiple key interactions between the G-loop and the C-tail in DCLK1, such as dipole interactions with the second glycine in the G-loop by the phosphothreonine. In addition, Thr 687 contributes to increased stabilization of DCLK1.2 tail by forming a C-cap with the helical ATP-mimic segment. We aligned the intensively-studied protein kinase (PKA, PDB: 1ATP), a DCLK1.2 structure (PDB: 6KYQ), and frames of our MD trajectory, which demonstrate remarkable overlap of the ATP gamma phosphate and C-tail phosphothreonine, which seemingly acts as a mimic for the ATP co-factor. As phosphorylation is reported to lead to DCLK1 inhibition, this suggests a complex mechanism of regulation, in which the DCLK-specific constraints in the G-loop, the intrinsic flexibility of the C-tail, and threonine phosphorylation, by cis or trans-mediated modification, systematically prevent hyperphosphorylation of the doublecortin domains and cellular effects. Somewhat paradoxically, we could only identify phosphorylated Thr 688 in activated DCLK1.2 mutants, but not in the autoinhibited (WT) versioin. form. This suggests that the selected mutations exhibit a regulatory hierarchal dominance over inhibitory Thr 688 phosphorylation and are sufficient to liberate DCLK1.2 from its auto-inhibited, C-tail bound state. This also implies that phosphorylation of Thr 688 may only be minimally-required for autoinhibition, especially given its association with the hyperactive variants obtained by mutagenesis (Figure 3.3.8-figure supplement 1-2).

Finally, we have evaluated the terminal residues in the DCLK1 C-tail, which are predicted to be intrinsically disordered. Side-by-side analysis of DCLK1.1 and DCLK1.2 in which this region is added from a common core terminating at residue 703, shows that increasing the length of the tail in both DCLK1.1 and DCLK1.2 has little additional effect on the inhibitory or stabilizing effects driven by the highly ordered tail regions that precede it. Kinase domains are regulated by IDRs in a multitude of ways, but the CAMK family is specifically enriched for adapted C-terminal extensions that, as we show here, can block the ATP and substrate binding, and enzymatically inactivate the kinase domain by occlusion of the activation loop through a flexible helical IDS on their C-tail (4). We note that the DCLK1.2 kinase domain crystallizes in an 'active' closed conformation, despite binding of the C-tail in an autoinhibitory manner (Cheng et al. 2022). Repeated packing motions of the IDS against the activation loop in all replicates of DCLK1.1 MD simulations, suggest that tail may occlude the activation loop, similar to other CAMKs, possibly pointing to a mode of cis autoregulation. Conversely, AlphaFold2 predicts the placement of the DCX domains as adjacent to the IDS in both DCLK1.1 and DCLK1.2 (Figure S1). It is possible, like other CAMKs, the IDS facilitates protein binding, whether to the DCX domains, calcium-modulated proteins, or other kinases.

3.5.3 Evolutionary divergence and functional specialization of DCLKs

For the DCLK family as a whole, we discovered phylogenetic divergence between DCLK1 and 2 as a relatively recent event, (Figure 3.1A) in which metazoan DCLK3 is the more ancestral DCLK gene from which DCLK1 and 2 emerged after duplication. Because of shared evolutionary constraints and the recent divergence between DCLK1 and 2, we surmise the functional specialization of the DCLK1 tail is shared between these paralogs. Moreover, we quantify key differences between human DCLK1.1 and 1.2 activity that are impacted by amino acid changes that contribute to the function of the C-tail. The differences between DCLK1 isoforms 1 and 2 are generated by variations in exon splicing, which change both the C-tail protein sequence, and introduce or exclude potential phosphorylation sites. Expression of the highly autoinhibitable DCLK1.2 isoform is believed to be predominant in the brain during embryogenesis, although DCLK1.1 is also thought to be present in the adult brain (Burgess and Reiner 2002). It is therefore possible that an altered ratio in DCLK1.1/1.2 expression, accompanied by changes in the requirement for DCLK1 auto-regulation, are relevant for early neurogenesis. The overall sequence similarity at the protein level, despite the loss of a pair of putative phosphorylation sites, suggests a possible exon duplication in the C-tail, whereby polymorphisms have allowed for adaptive regulation during development and proliferation. Indeed, we also speculate that the induced expression of DCLK1.2 splice variants in multiple cancer subtypes (Qu et al. 2019) is likely to be indicative of a survival and drug-resistance role that could be targetable with new types of small molecule. Although nanomolar DCLK1 ATP-site inhibitors such as DCLK1-IN-1 have been developed that can bind tightly to the DCLK1 ATP site (Figure 3.3 and Figure 3.3-figure supplement 1), the 'problematic' existence of human DCLK1.1 and DCLK1.2 splice variants with distinct auto-inhibitory properties may present a challenge to compound engagement in the cell, where relief of auto-inhibition through C-tail undocking in DCLK1.2 is likely to require a high concentration of compound in order to compete and disengage interactions at the ATP site. Indeed, although potent chemical DCLK1 inhibitors such as DCLK1-IN-1 are known to influence DCLK1 autophosphorylation and cell motility, they have relatively modest effects in cells in terms of cytotoxicity (Ferguson et al. 2020; Ding et al. 2021). Therefore, we propose that the dual inhibitory effects of the C-tail and the transmission of this information to adjacent DCX domains, which control adaptive cellular phenotypes such as EMT in cancer cells (Major et al. 1985), may make allosteric classes of DCLK1 inhibitor a preferred therapeutic option, especially if they can be tailored specifically towards DCLK1.1 or DCLK1.2, whose autoregulation is different in terms of the varied molecular details we have uncovered here.

References

Agulto, Regina L et al. (July 2021). "Autoregulatory control of microtubule binding in doublecortin-like kinase 1". *eLife* 10. Ed. by Andrew P Carter and Piali Sengupta. Publisher: eLife Sciences Publications, Ltd, e60126. ISSN: 2050-084X. DOI: 10.7554/eLife.60126. URL: https://doi.org/10.7554/eLife.60126.

Baffi, Timothy R. and Alexandra C. Newton (Apr. 2022). "mTOR Regulation of AGC Kinases: New Twist to an Old Tail". en. *Molecular Pharmacology* 101.4. Publisher: American Society for Pharmacology and Experimental Therapeutics Section: Axelrod Symposium Protein Kinases in Tune - Special Section, pp. 213–218. ISSN: 0026-895X, 1521-0111. DOI: 10.1124/molpharm.121.000310. URL: https://molpharm.aspetjournals.org/content/101/4/213.

Bayer, K. Ulrich and Howard Schulman (Aug. 2019). "CaM Kinase: Still Inspiring at 40". eng. *Neuron* 103.3, pp. 380–394. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2019.05.033.

Berginski, Matthew E. et al. (Jan. 2021). "The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases". eng. *Nucleic Acids Research* 49.D1, pp. D529–D535. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa853.

Bhattacharyya, Moitrayee et al. (Mar. 2020). "Flexible linkers in CaMKII control the balance between activating and inhibitory autophosphorylation". eng. *eLife* 9, e53670. ISSN: 2050-084X. DOI: 10.7554/eLife.53670.

Burgess, Harold A. and Orly Reiner (May 2002). "Alternative splice variants of doublecortin-like kinase are differentially expressed and have different kinase activities". eng. *The Journal of*

Biological Chemistry 277.20, pp. 17696–17705. ISSN: 0021-9258. DOI: 10.1074/jbc.M111981200.

Byrne, Dominic P., Christopher J. Clarke, et al. (July 2020). "Use of the Polo-like kinase 4 (PLK4) inhibitor centrinone to investigate intracellular signalling networks using SILAC-based phosphoproteomics". eng. *The Biochemical Journal* 477.13, pp. 2451–2475. ISSN: 1470-8728. DOI: 10.1042/BCJ20200309.

Byrne, Dominic P., Safal Shrestha, et al. (July 2020). "Aurora A regulation by reversible cysteine oxidation reveals evolutionarily conserved redox control of Ser/Thr protein kinase activity". eng. *Science Signaling* 13.639, eaax2713. ISSN: 1937-9145. DOI: 10.1126/scisignal.aax2713.

Byrne, Dominic P., Matthias Vonderach, et al. (Oct. 2016). "cAMP-dependent protein kinase (PKA) complexes probed by complementary differential scanning fluorimetry and ion mobility-mass spectrometry". eng. *The Biochemical Journal* 473.19, pp. 3159–3175. ISSN: 1470-8728. DOI: 10.1042/BCJ20160648.

Cheng, Linna et al. (Jan. 2022). "DCLK1 autoinhibition and activation in tumorigenesis". eng. *Innovation (Cambridge (Mass.))* 3.1, p. 100191. ISSN: 2666-6758. DOI: 10.1016/j.xinn.2021.100191.

Couillard-Despres, Sebastien et al. (2005). "Doublecortin expression levels in adult brain reflect neurogenesis". en. *European Journal of Neuroscience* 21.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/1
9568.2004.03813.x, pp. 1–14. ISSN: 1460-9568. DOI: 10.1111/j.1460-9568.2004.03813.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-9568.2004.03813.x.

Cunningham, Fiona et al. (Jan. 2022). "Ensembl 2022". eng. *Nucleic Acids Research* 50.D1, pp. D988–D995. ISSN: 1362-4962. DOI: 10.1093/nar/gkab1049.

Daly, Leonard A. et al. (July 2021). "Oxygen-dependent changes in binding partners and post-translational modifications regulate the abundance and activity of HIF-1/2". eng. *Science Signaling* 14.692, eabf6685. ISSN: 1937-9145. DOI: 10.1126/scisignal.abf6685.

Ding, Ling et al. (Nov. 2021). "Inhibition of DCLK1 with DCLK1-IN-1 Suppresses Renal Cell Carcinoma Invasion and Stemness and Promotes Cytotoxic T-Cell-Mediated Anti-Tumor Immunity". eng. *Cancers* 13.22, p. 5729. ISSN: 2072-6694. DOI: 10.3390/cancers13225729.

Eyers, Patrick A. (Nov. 2015). "TRIBBLES: A Twist in the Pseudokinase Tail". eng. *Structure (London, England: 1993)* 23.11, pp. 1974–1976. ISSN: 1878-4186. DOI: 10.1016/j.str.2015.10.003.

Eyers, Patrick A., Karen Keeshan, and Natarajan Kannan (Apr. 2017). "Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease". eng. *Trends in Cell Biology* 27.4, pp. 284–298. ISSN: 1879-3088. DOI: 10.1016/j.tcb.2016.11.002.

Ferguson, Fleur M. et al. (June 2020). "Discovery of a selective inhibitor of doublecortin like kinase 1". eng. *Nature Chemical Biology* 16.6, pp. 635–643. ISSN: 1552-4469. DOI: 10.1038/s41589-020-0506-0.

Ferries, Samantha et al. (Sept. 2017). "Evaluation of Parameters for Confident Phosphorylation Site Localization Using an Orbitrap Fusion Tribrid Mass Spectrometer". eng. *Journal of Proteome Research* 16.9, pp. 3448–3459. ISSN: 1535-3907. DOI: 10.1021/acs.jproteome.7b00337.

Foulkes, Daniel M. et al. (Sept. 2018). "Covalent inhibitors of EGFR family protein kinases induce degradation of human Tribbles 2 (TRIB2) pseudokinase in cancer cells". eng. *Science Signaling* 11.549, eaat7951. ISSN: 1937-9145. DOI: 10.1126/scisignal.aat7951.

Galvan, Laurie, Laetitia Francelle, Marie-Claude Gaillard, Lucie de Longprez, Maria-Angeles Carrillo-de Sauvage, Géraldine Liot, Karine Cambon, Lev Stimmer, Sophie Luccantoni, Julien Flament, et al. (May 2018b). "The striatal kinase DCLK3 produces neuroprotection against mutant huntingtin". eng. *Brain: A Journal of Neurology* 141.5, pp. 1434–1454. ISSN: 1460-2156. DOI: 10.1093/brain/awy057.

Gao, Tianbo et al. (Oct. 2016). "DCLK1 is up-regulated and associated with metastasis and prognosis in colorectal cancer". en. *Journal of Cancer Research and Clinical Oncology* 142.10, pp. 2131–2140. ISSN: 1432-1335. DOI: 10.1007/s00432-016-2218-0. URL: https://doi.org/10.1007/s00432-016-2218-0.

Gógl, Gergő et al. (Apr. 2019). "Disordered Protein Kinase Regions in Regulation of Kinase Domain Cores". eng. *Trends in Biochemical Sciences* 44.4, pp. 300–311. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2018.12.002.

Gu, X., Y. X. Fu, and W. H. Li (July 1995). "Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites". eng. *Molecular Biology and Evolution* 12.4, pp. 546–557. ISSN: 0737-4038. DOI: 10.1093/oxfordjournals.molbev.a040235.

Harris, John A. et al. (2022). "Analysis of human Tribbles 2 (TRIB2) pseudokinase". eng. *Methods in Enzymology* 667, pp. 79–99. ISSN: 1557-7988. DOI: 10.1016/bs.mie.2022.03.025.

Hoang, Diep Thi et al. (Feb. 2018). "UFBoot2: Improving the Ultrafast Bootstrap Approximation". eng. *Molecular Biology and Evolution* 35.2, pp. 518–522. ISSN: 1537-1719. DOI: 10.1093/molbev/msx281.

Horesh, David et al. (Sept. 1999). "Doublecortin, a Stabilizer of Microtubules". *Human Molecular Genetics* 8.9, pp. 1599–1610. ISSN: 0964-6906. DOI: 10.1093/hmg/8.9.1599. URL: https://doi.org/10.1093/hmg/8.9.1599.

Huang, Jing and Alexander D. MacKerell (Sept. 2013). "CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data". eng. *Journal of Computational Chemistry* 34.25, pp. 2135–2145. ISSN: 1096-987X. DOI: 10.1002/jcc.23354.

Huang, Liang-Chin et al. (2021). "KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases". *BMC bioinformatics* 22, pp. 1–25.

Hudmon, Andy and Howard Schulman (June 2002). "Structure-function of the multifunctional Ca2+/calmodulin-dependent protein kinase II". eng. *The Biochemical Journal* 364.Pt 3, pp. 593–611. ISSN: 0264-6021. DOI: 10.1042/BJ20020228.

Huerta-Cepas, Jaime, François Serra, and Peer Bork (June 2016). "ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data". eng. *Molecular Biology and Evolution* 33.6, pp. 1635–1638. ISSN: 1537-1719. DOI: 10.1093/molbev/msw046.

Humphrey, W., A. Dalke, and K. Schulten (Feb. 1996). "VMD: visual molecular dynamics". eng. *Journal of Molecular Graphics* 14.1, pp. 33–38, 27–28. ISSN: 0263-7855. DOI: 10.1016/0263-7855(96)00018-5.

Huse, Morgan and John Kuriyan (May 2002). "The conformational plasticity of protein kinases". eng. *Cell* 109.3, pp. 275–282. ISSN: 0092-8674. DOI: 10.1016/s0092-8674(02)00741-9.

Jorgensen, William L. et al. (July 1983). "Comparison of simple potential functions for simulating liquid water". *The Journal of Chemical Physics* 79.2. Publisher: American Institute of Physics, pp. 926–935. ISSN: 0021-9606. DOI: 10.1063/1.445869. URL: https://aip.scitation.org/doi/10.1063/1.445869.

Kabsch, Wolfgang and Christian Sander (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". *Biopolymers: Original Research on Biomolecules* 22.12, pp. 2577–2637.

Kalyaanamoorthy, Subha et al. (June 2017). "ModelFinder: fast model selection for accurate phylogenetic estimates". eng. *Nature Methods* 14.6, pp. 587–589. ISSN: 1548-7105. DOI: 10.1038/nmeth.4285.

Kannan, Natarajan et al. (Jan. 2007). "The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module". eng. *Proceedings of the National Academy* *of Sciences of the United States of America* 104.4, pp. 1272–1277. ISSN: 0027-8424. DOI: 10.1073/pnas.0610251104.

Katoh, Kazutaka et al. (July 2002). "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform". eng. *Nucleic Acids Research* 30.14, pp. 3059–3066. ISSN: 1362-4962. DOI: 10.1093/nar/gkf436.

Keeshan, Karen et al. (Dec. 2010). "Transformation by Tribbles homolog 2 (Trib2) requires both the Trib2 kinase domain and COP1 binding". eng. *Blood* 116.23, pp. 4948–4957. ISSN: 1528-0020. DOI: 10.1182/blood-2009-10-247361.

Keller, Oliver et al. (June 2008). "Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species". eng. *BMC bioinformatics* 9, p. 278. ISSN: 1471-2105. DOI: 10.1186/1471-2105-9-278.

Le, Si Quang and Olivier Gascuel (July 2008). "An improved general amino acid replacement matrix". eng. *Molecular Biology and Evolution* 25.7, pp. 1307–1320. ISSN: 1537-1719. DOI: 10.1093/molbev/msn067.

Major, J. et al. (Mar. 1985). "Increased SCE inducibility by low doses of methylcholanthrene in lymphocytes obtained from patients with Down's disease". eng. *Mutation Research* 149.1, pp. 51–55. ISSN: 0027-5107. DOI: 10.1016/0027-5107(85)90008-9.

Mandell, Daniel J., Evangelos A. Coutsias, and Tanja Kortemme (Aug. 2009). "Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling". en. *Nature Methods* 6.8. Number: 8 Publisher: Nature Publishing Group, pp. 551–552. ISSN: 1548-7105. DOI: 10.1038/nmeth0809-551. URL: https://www.nature.com/articles/nmeth0809-551.

Manning, G. et al. (Dec. 2002). "The protein kinase complement of the human genome". eng. *Science (New York, N.Y.)* 298.5600, pp. 1912–1934. ISSN: 1095-9203. DOI: 10.1126/science.1075762.

Matsumoto, N., D. T. Pilz, and D. H. Ledbetter (Mar. 1999). "Genomic structure, chromosomal mapping, and expression pattern of human DCAMKL1 (KIAA0369), a homologue of DCX (XLIS)". eng. *Genomics* 56.2, pp. 179–183. ISSN: 0888-7543. DOI: 10.1006/geno.1998.5673.

McSkimming, Daniel Ian et al. (Nov. 2016). "KinView: a visual comparative sequence analysis tool for integrated kinome research". eng. *Molecular bioSystems* 12.12, pp. 3651–3665. ISSN: 1742-2051. DOI: 10.1039/c6mb00466k.

Michaud-Agrawal, Naveen et al. (July 2011). "MDAnalysis: a toolkit for the analysis of molecular dynamics simulations". eng. *Journal of Computational Chemistry* 32.10, pp. 2319–2327. ISSN: 1096-987X. DOI: 10.1002/jcc.21787.

Murphy, James M. et al. (Nov. 2015). "Molecular Mechanism of CCAAT-Enhancer Binding Protein Recruitment by the TRIB1 Pseudokinase". eng. *Structure (London, England: 1993)* 23.11, pp. 2111–2121. ISSN: 1878-4186. DOI: 10.1016/j.str.2015.08.017.

Neuwald, Andrew F (2014). "A Bayesian sampler for optimization of protein domain hierarchies". *Journal of Computational Biology* 21.3, pp. 269–286.

— (Aug. 2009). "Rapid detection, classification and accurate alignment of up to a million or more related protein sequences". eng. *Bioinformatics (Oxford, England)* 25.15, pp. 1869–1875. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp342.

Nguyen, Lam-Tung et al. (Jan. 2015). "IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies". eng. *Molecular Biology and Evolution* 32.1, pp. 268–274. ISSN: 1537-1719. DOI: 10.1093/molbev/msu300.

Nguyen, Tuan et al. (2015). "Co-conserved MAPK features couple D-domain docking groove to distal allosteric sites via the C-terminal flanking tail". eng. *PloS One* 10.3, e0119636. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0119636.

Ohmae, Shogo et al. (July 2006). "Molecular identification and characterization of a family of kinases with homology to Ca2+/calmodulin-dependent protein kinases I/IV". eng. *The*

Journal of Biological Chemistry 281.29, pp. 20427–20439. ISSN: 0021-9258. DOI: 10.1074/jbc.M513212200.

Omar, Mitchell H. et al. (June 2023). "Classification of Cushing's syndrome PKAc mutants based upon their ability to bind PKI". eng. *The Biochemical Journal* 480.12, pp. 875–890. ISSN: 1470-8728. DOI: 10.1042/BCJ20230183.

Omori, Y. et al. (1998). "Expression and chromosomal localization of KIAA0369, a putative kinase structurally related to Doublecortin". eng. *Journal of Human Genetics* 43.3, pp. 169–177. ISSN: 1434-5161. DOI: 10.1007/s100380050063.

Patel, Onisha et al. (Sept. 2021). "Structural basis for small molecule targeting of Doublecortin Like Kinase 1 with DCLK1-IN-1". eng. *Communications Biology* 4.1, p. 1105. ISSN: 2399-3642. DOI: 10.1038/s42003-021-02631-y.

Pronk, Sander et al. (Apr. 2013). "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit". eng. *Bioinformatics (Oxford, England)* 29.7, pp. 845–854. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btt055.

Qu, Dongfeng et al. (2019). "Overexpression of DCLK1-AL Increases Tumor Cell Invasion, Drug Resistance, and KRAS Activation and Can Be Targeted to Inhibit Tumorigenesis in Pancreatic Cancer". eng. *Journal of Oncology* 2019, p. 6402925. ISSN: 1687-8450. DOI: 10.1155/2019/6402925.

Rellos, Peter et al. (July 2010). "Structure of the CaMKIIdelta/calmodulin complex reveals the molecular mechanism of CaMKII kinase activation". eng. *PLoS biology* 8.7, e1000426. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1000426.

Romano, Robert A. et al. (2009). "A chimeric mechanism for polyvalent trans-phosphorylation of PKA by PDK1". en. *Protein Science* 18.7, pp. 1486–1497. ISSN: 1469-896X. DOI: 10.1002/pro.146.

Silva, Jeffrey C. et al. (Jan. 2006). "Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition". eng. *Molecular & cellular proteomics: MCP* 5.1, pp. 144–156. ISSN: 1535-9476. DOI: 10.1074/mcp.M500230-MCP200.

Sossey-Alaoui, K. and A. K. Srivastava (Feb. 1999). "DCAMKL1, a brain-specific transmembrane protein on 13q12.3 that is similar to doublecortin (DCX)". eng. *Genomics* 56.1, pp. 121–126. ISSN: 0888-7543. DOI: 10.1006/geno.1998.5718.

Taylor, Susan S. and Alexandr P. Kornev (Feb. 2011). "Protein kinases: evolution of dynamic regulatory proteins". en. *Trends in Biochemical Sciences* 36.2, pp. 65–77. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2010.09.006. URL: https://www.sciencedirect.com/science/article/pii/S0968000410001830.

Tyka, Michael D. et al. (Jan. 2011). "Alternate states of proteins revealed by detailed energy landscape mapping". eng. *Journal of Molecular Biology* 405.2, pp. 607–618. ISSN: 1089-8638. DOI: 10.1016/j.jmb.2010.11.008.

UniProt Consortium (Jan. 2021). "UniProt: the universal protein knowledgebase in 2021". eng. *Nucleic Acids Research* 49.D1, pp. D480–D489. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa1100.

Wayman, Gary A. et al. (Sept. 2008). "Calmodulin-kinases: modulators of neuronal development and plasticity". eng. *Neuron* 59.6, pp. 914–931. ISSN: 1097-4199. DOI: 10.1016/j.neuron.2008.08.021.

Westphalen, C. Benedikt, Michael Quante, and Timothy C. Wang (July 2017). "Functional implication of Dclk1 and Dclk1-expressing cells in cancer". *Small GTPases* 8.3. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/21541248.2016.1208792, pp. 164–171. ISSN: 2154-1248. DOI: 10.1080/21541248.2016.1208792. URL: https://doi.org/10.1080/21541248.2016.1208792.

Yang, E. and H. Schulman (Sept. 1999). "Structural examination of autoregulation of multifunctional calcium/calmodulin-dependent protein kinase II". eng. *The Journal of Biological Chemistry* 274.37, pp. 26199–26208. ISSN: 0021-9258. DOI: 10.1074/jbc.274.37.26199.

Yeon, Ju Hun et al. (Feb. 2016). "Systems-wide Identification of cis-Regulatory Elements in Proteins". eng. *Cell Systems* 2.2, pp. 89–100. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.02.004.

Yeung, Wayland, Annie Kwon, et al. (Dec. 2021). "Evolution of Functional Diversity in the Holozoan Tyrosine Kinome". eng. *Molecular Biology and Evolution* 38.12, pp. 5625–5639. ISSN: 1537-1719. DOI: 10.1093/molbev/msab272.

Yokokura, H. et al. (Oct. 1995). "The regulatory region of calcium/calmodulin-dependent protein kinase I contains closely associated autoinhibitory and calmodulin-binding domains". eng. *The Journal of Biological Chemistry* 270.40, pp. 23851–23859. ISSN: 0021-9258. DOI: 10.1074/jbc.270.40.23851.

Chapter 4

Glydentify, a deep learning tool for classifying glycosyltransferase function

Venkat, Aarya et al. Glydentify, a deep learning tool for classifying glycosyltransferase function. To be submitted to *Bioinformatics*.

4.1 Abstract

Protein language models have emerged as a powerful tool for predicting protein function by capturing the underlying grammar and syntax of protein sequences. Here, we introduce Glydentify, an open-source and user-friendly application that uses protein language models for the classification of glycosyltransferases (GTs). Utilizing the state-of-the-art ESM2 protein language model, Glydentify extracts high-dimensional sequence embeddings to accurately classify GTs into fold A families with 92% accuracy. The tool also predicts GT-A donor binding preferences with an accuracy of 88%. Notably, Glydentify identifies key residues that contribute to a prediction, thereby adding an explainable component to the application. With an intuitive interface powered by Gradio, Glydentify requires no programming experience from the user, democratizing access to cutting-edge deep learning technologies for GT research. The application is freely available on GitHub and can be accessed directly through any web browser (https://huggingface.co/spaces/arikat/Glydentify).

Keywords:

Glycobiology, Bioinformatics, Deep Learning, Sequence Classification, Protein Language Model

4.2 Introduction

Identification and characterization of protein sequences is a critical task in biological research, from understanding fundamental biochemical processes to advancing enzymatic synthesis. Yet there are few tools available for biomedical scientists to study and make hypotheses about Glycosyltransferases (GTs) (Taujale, Soleymani, et al. 2021; York et al. 2020), an underappreciated enzyme superfamily that catalyzes glycosidic linkages. These enzymes are highly divergent and difficulties with expression and purification hinder progress with classification.

Currently, GTs are classified by sequence similarity using traditional alignment approaches. The largest repository for GT family classification is the Carbohydrate Active Enzyme (CAZy) database (Cantarel et al. 2009), the authority on the classification of new GT families, among other enzymes. However, because GTs are highly divergent, the effectiveness of traditional methods is diminished. Through extensive evolutionary, phylogenetic, and deep learning approaches, we previously uncovered a number of new subfamilies and even novel GT folds (Taujale, Soleymani, et al. 2021; Taujale, Venkat, et al. 2020; Taujale, Zhou, et al. 2021; Venkat, Tehrani, et al. 2022). With the emergence of deep learning protein language models, like ESM2 (Evolutionary Scale Modeling) (Z. Lin et al. 2023), alignment-free approaches (Yeung, Zhou, Mathew, et al. 2023) are now possible using sequence embeddings, which encapsulate high-dimensional data about protein evolution, structure, and chemistry. These embeddings can then be leveraged to train classification models for function prediction.

Here, we present GlyIdentify (Figure 4.1), a novel, open-source application released with two modules, one which uses a classifier trained for predicting fold A GT (GT-A) family membership, and another which uses a classifier to predict potential donor substrates for GT-A sequences. Users can input fasta sequences and seamlessly receive a prediction. GlyIdentify is open-source and freely available on GitHub. In addition, it can be accessed online directly through the online HuggingFace webtool.

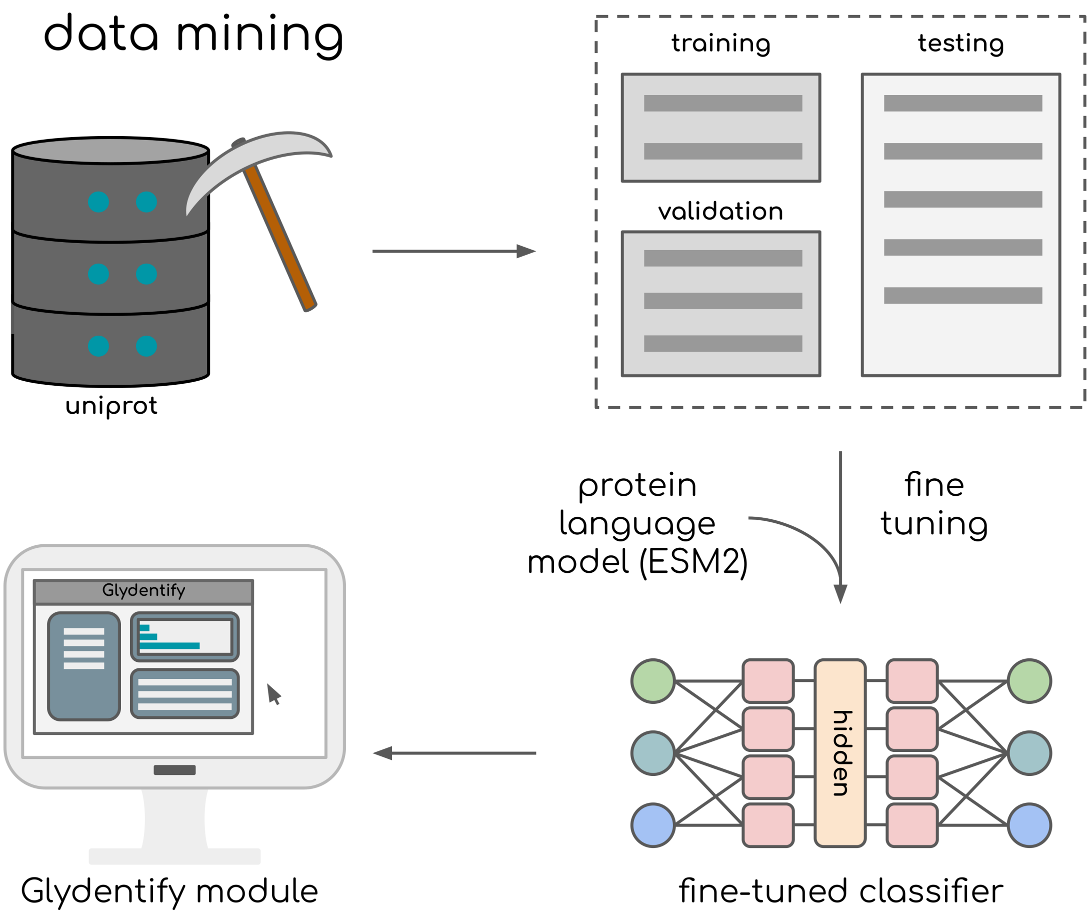


Figure 4.1: A cartoon workflow for developing a Glydentify module.

4.3 Methods

Family data collection and preprocessing

Glycosyltransferase sequences were obtained by searching the Uniprot database using sequence-based profiles, as previously described (Taujale, Venkat, et al. 2020). About 200,000 sequences were obtained; CD-Hit (ref) was used to purge sequences above 90% similarity. 10,000 sequences equally distributed across the 72 known GT-A families (Taujale, Venkat, et al. 2020) were pulled from the remaining sequences (183,212). These sequences had an amino acid length between

200-600 amino acids, to prevent capture of fragmentary sequences or sequences with multiple domains. Each sequence was labeled with the family it belonged to, using the CAZy numbering scheme. We took a subset of our 10,000 sequence training set for validation testing using an 80-20 split. The remainder of the 173,212 sequences was used for testing.

Donor data collection and preprocessing

GT sequences were obtained by scraping the Uniprot database for sequences with information about catalytic activity. Five datasets were created for GTs that bound mannose, glucose, galactose, GalNAc, and GlcNAc, totaling about 50,000 sequences. Each dataset was purged at 80% similarity using CD-Hit. Finally, we split the dataset at 70-15-15, for training, validation, and testing, using sklearn, with stratification enabled to ensure equal distribution of labels. Like before, the amino acid length of trained sequences was between 200-600 amino acids to prevent capture of sequences with multiple domains or fragmentary GT sequences. Each sequence was labeled with the donor substrate it would bind.

Embedding Generation and Training

The ESM2 protein language model was used to generate sequence embeddings. The sequences were tokenized and padded or truncated to a maximum length of 512 before passing them to the model. ESM2 has been shown to effectively capture protein sequence properties (Z. Lin et al. 2023). Employing the softmax function we converted the model's output logits into probabilities, which serve as a statistical basis for prediction. These softwmax-transformed logits are used to construct a bar graph using MatPlotLib to allow users to appreciate the confidence level of the top predictions. The models were trained using an nVidia RTX 5000 GPU, but only require a single cpu for running the program.

Validation and Testing

The validation set was evaluated during the training and the loss/accuracy curves were plotted to help tune hyperparameters and evaluate if the model was overfitting the training set. Using a validation set helped evaluate how generalizable our results would be without running the data through the large testing set.

Explainability

To explain which residues contribute to a particular prediction, we programmed a post-hoc method which masks N residues during the prediction, and re-runs the prediction for (Total Residues/N) iterations. By masking a set of residues, the model takes account for alternative predictions which manifest based on hidden residues.

App Development and Packaging

The trained model was integrated into a Gradio-based web application, providing a user-friendly interface to input glycosyltransferase sequences and receive predictions. The application is freely available on the HuggingFace platform (https://huggingface.co/spaces/arikat/Glydentify). We also uploaded this application into a downloadable GitHub package for easy deployment and use.

4.4 Results

Traditional sequence-based approaches, like Conserved Domain Database (CDD) search, rely heavily on pre-defined, curated domain databases. While they are effective for identifying known domains in sequences, these methods often struggle with novel sequences or sequences with low similarity to known domains. Additionally, they may fail to capture functionally important but less-conserved elements of the sequence, leading to incomplete or less accurate annotations.

Glydentify, on the other hand, is fine-tuned from the ESM2 embeddings, allowing it to draw from evolutionary and structural patterns within Glycosyltranserase (GT) families. This empowers it to classify GTs based on subtle patterns that may not be easily identified by traditional sequence approaches or raw ESM2 embeddings alone.

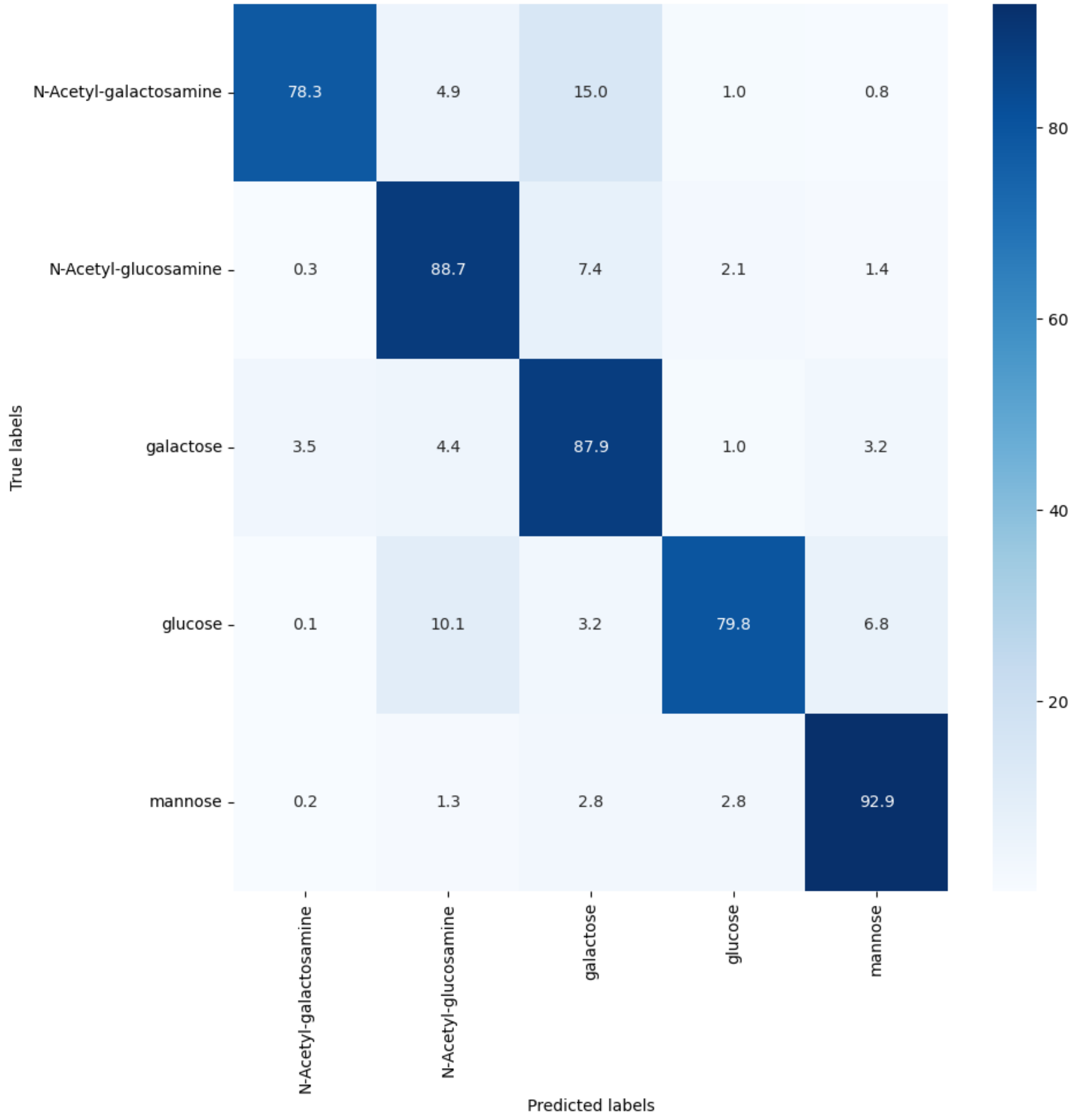


Figure 4.2: A confusion matrix showing model accuracy for each donor label in our testing dataset.

The classification performed by ESM-2 in Glydentify is based on a transformer architecture, which embeds protein sequences into high-dimensional spaces, capturing complex patterns and dependencies in the sequence data. The model then uses a form of logistic regression at the output layer, specifically a softmax layer, to classify the sequences into different categories, and it is trained with a cross-entropy loss to optimize its predictions.

Our two modules fine-tuned from the T33-150M ESM2 model offer accuracies of 92% for family classification and 87% for donor classification (Figure S1). We can further break these accuracies down through the use of a confusion matrix (Figure S2, Figure 4.2), which indicates model performance based on comparison of true and predicted labels. Our family classification is highly accurate with most GT families. However, some accuracy is lost when classifying subfamilies. Predominantly, we see the model confuses closely related GT2-subfamilies, understanding they are within the broader GT2 superfamily, but having trouble distinguishing the distinct features that differentiate closely related subfamilies. This may be due to lower representation of certain GT2 subfamilies. Alternatively, our choice to truncate the sequence to the GT-domain to save on computational cost, may also contribute to model inaccuracy.

Similarly, our donor classification indicates that the model largely performs donor prediction well, but has difficulty with classification of highly similar donors, like N-acetyl galactosamine (GalNAc) and Galactose (Gal). These discrepancies may arise from situations like the ABO glycosyltransferases which transfer the aforementioned sugars, resulting in an A or B blood phenotype (Patenaude et al. 2002). There are only four distinct residues which differ between ABO GTs (Patenaude et al. 2002), which distinguish whether they transfer GalNAc or Gal. There are likely many more such cases, and with larger and more fine-grained explainable deep learning models, the particular residues involved in changing donor specificity may be identified.

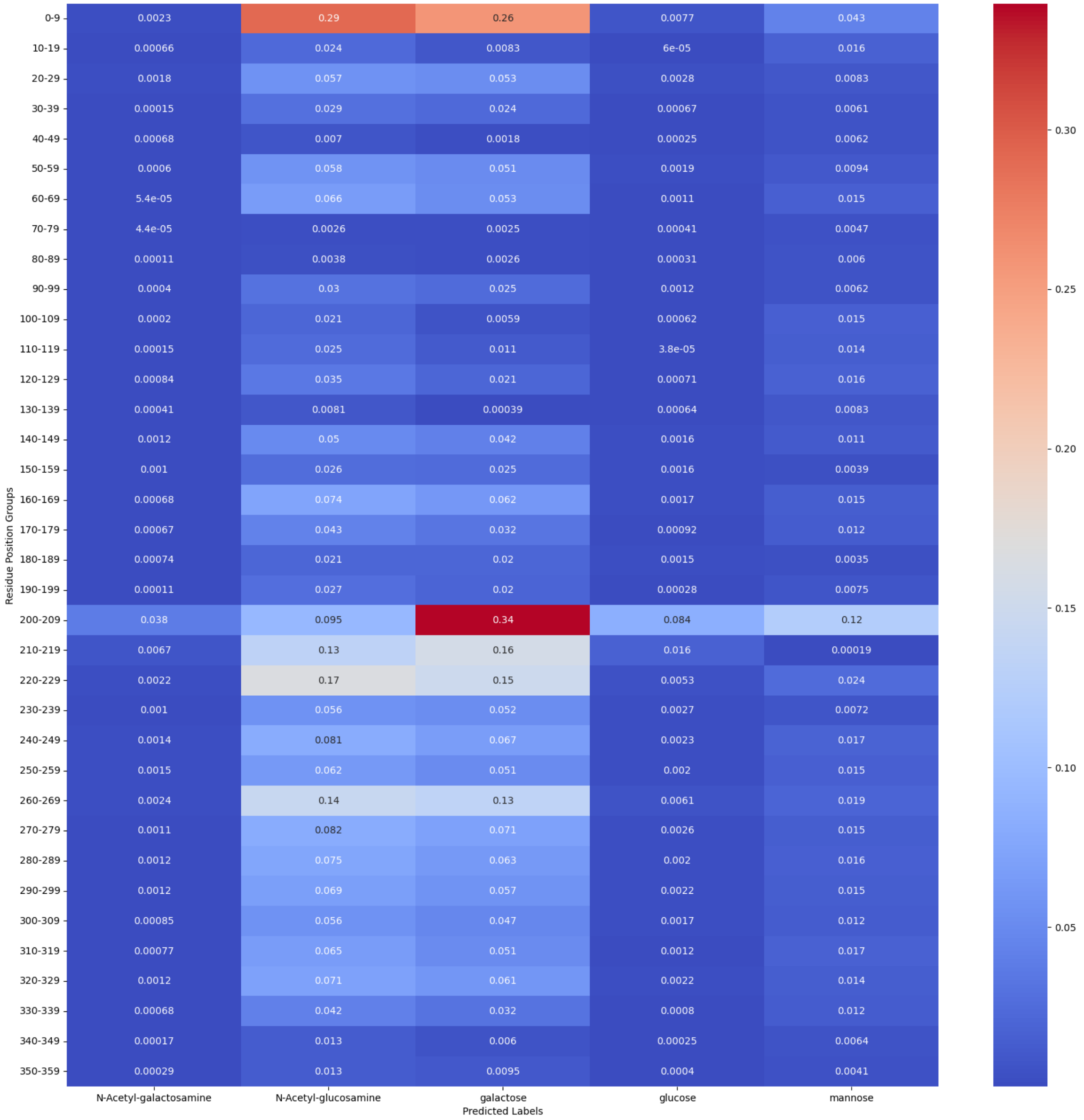


Figure 4.3: Explainable heatmap where the y-axis represents sequence range in groups of ten residues, and the X-axis represents categorical prediction (which sugar the enzyme sequence is predicted to bind). Collectively, the heatmap outlines the residues which contribute the most to a given prediction, as well as highlights potential residues in other predictions that may be helpful for engineering bifunctional or promiscuous GT-A enzymes.

To this avail, we added an explainable component to the donor prediction model in Glydentify (see Methods). We masked every N residues and re-ran the prediction to generate a heatmap of possible predictions (Total Residues/N), when N residues are masked. This heatmap then highlights possible residue groups that contributed to a prediction (Figure 4.3). We provide the option to change granularity of the heatmap, by allowing the user to select the number of residues masked (N). Lower masking results in increased compute time, but higher resolution regarding which residues may contribute to a prediction. In our example, our model correctly predicts Galactose for B-1,3-Galactosyltranferase (B blood-type GT-A enzyme).

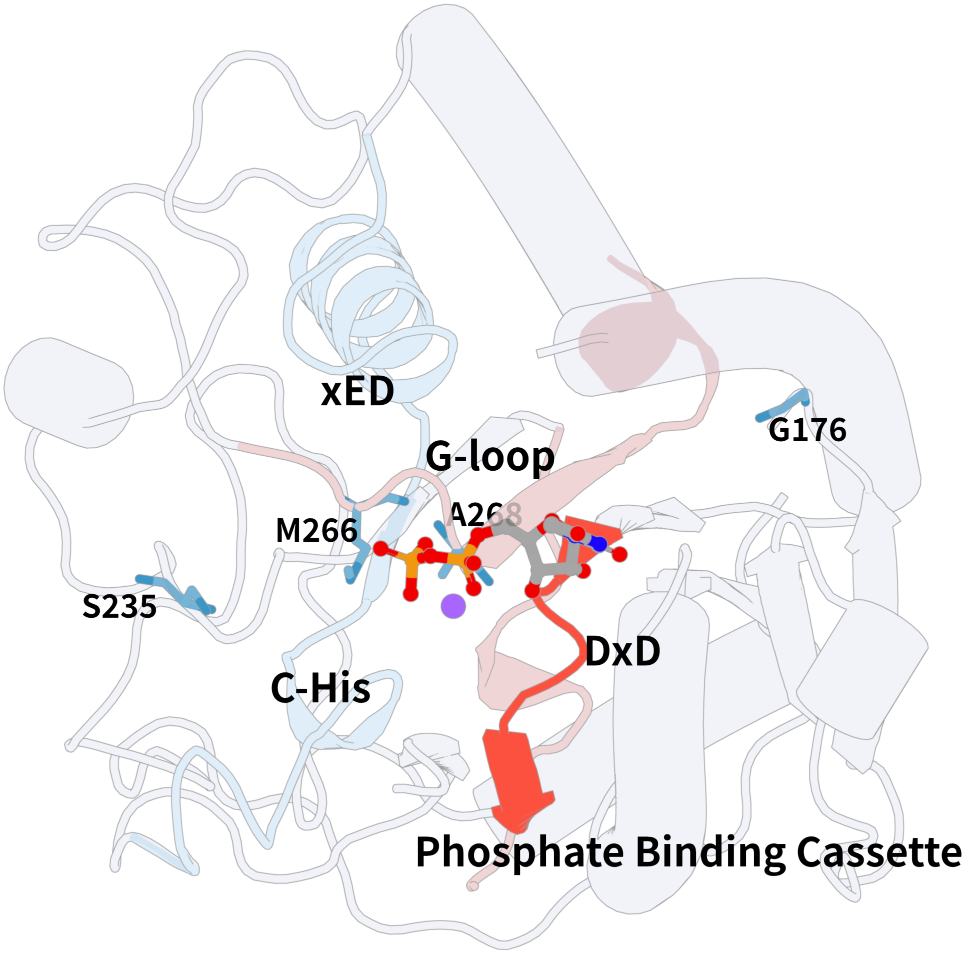


Figure 4.4: A heatmap mapped to a pdb structure of B-1,3-Galactosyltransferase, a B blood-type ABO GT (pdb: 2RIX).

We mapped the explainable heatmap onto an ABO GT structure (Figure 4.4) and it indicates that some of the residues of interest for the galactose prediction are the conserved phosphate binding cassette in GTs (Venkat, Tehrani, et al. 2022), which is necessary for the binding the donor, as well as the glycine rich loop (G-loop) (Taujale, Venkat, et al. 2020), which contain two of four residues which differ between A and B blood-type GT-As (Patenaude et al. 2002). The model also asserts slight contributions are made by regions containing the catalytic xED motif and C-His, which helps bind the metal cation. With an explainable component, biologists may be more inclined to comprehend a particular prediction with biochemical or sequence expertise.

4.5 Conclusion

GIydentify leverages evolutionary and structural patterns concealed within sequences to effectively classify fold A GTs by family and by donor function. It provides a robust solution to the problem of sequence diversity in protein families. GTs, for instance, are incredibly diverse in terms of sequence, structure, and function (Taujale, Soleymani, et al. 2021; Taujale, Venkat, et al. 2020; Taujale, Zhou, et al. 2021; Venkat, Tehrani, et al. 2022; Moremen and Haltiwanger 2019; Varki, Richard D Cummings, et al. 2022). Thus, traditional sequence methods may struggle to accurately classify such a diverse group, but GIydentify's alignment-free embedding-based approach is well-suited to handle this variability. GIydentify also does not rely on pre-defined, curated databases, making it better equipped to handle novel sequences or sequence regions that may not be well-represented in existing databases.

Despite its proficiency, it's important to note that protein language models do not fully incorporate three-dimensional structure information. As the field of bioinformatics continues to advance, with developments like AlphaFold2 improving 3D structure prediction and emerging sequence-structure methods such as contrastive-learning approaches, the prospects for holistic protein learning models become increasingly promising. Such advancements are anticipated to predict critical attributes including protein structure, binding location and substrate specificity, and mechanisms of allostery. Ultimately, as more comprehensive and accurate models are developed, it will be critical to improve large-scale accessibility and usage for non-specialists to verify and enhance our understanding of protein function and contribute to efforts in bioengineering and substrate synthesis.

Finally, the Glydentify application is modular, allowing for easy addition of new classification models. The open-source code can be forked on github by interested scientists, seeking to train their own classifiers on glyco-enzymes of interest. Well-performing models can easily be added to the Glydentify interface, enhancing its utility and further establishing it as a platform that democratizes access to AI for glycobiology.

References

Cantarel, Brandi L et al. (2009). "The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics". *Nucleic acids research* 37.suppl_1, pp. D233–D238.

Lin, Zeming et al. (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". *Science* 379.6637, pp. 1123–1130.

Moremen, Kelley W and Robert S Haltiwanger (2019). "Emerging structural insights into glycosyltransferase-mediated synthesis of glycans". *Nature chemical biology* 15.9, pp. 853–864.

Patenaude, Sonia I et al. (2002). "The structural basis for specificity in human ABO (H) blood group biosynthesis". *Nature structural biology* 9.9, pp. 685–690.

Taujale, Rahil, Saber Soleymani, et al. (2021). "GTXplorer: A portal to navigate and visualize the evolutionary information encoded in fold A glycosyltransferases". *Glycobiology* 31.11, pp. 1472–1477.

Taujale, Rahil, Aarya Venkat, et al. (2020). "Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases". *Elife* 9, e54532.

Taujale, Rahil, Zhongliang Zhou, et al. (2021). "Mapping the glycosyltransferase fold landscape using interpretable deep learning". *Nature Communications* 12.1, p. 5656.

Varki, Ajit, Richard D Cummings, et al. (2022). "Essentials of Glycobiology [internet]".

Venkat, Aarya, Daniel Tehrani, et al. (2022). "Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases". *Journal of Biological Chemistry* 298.8.

Yeung, Wayland, Zhongliang Zhou, Liju Mathew, et al. (2023). "Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies". *Briefings in Bioinformatics* 24.1, bbac619.

York, William S et al. (2020). "GlyGen: computational and informatics resources for glycoscience". *Glycobiology* 30.2, pp. 72–73.

Chapter 5

Concluding remarks and future directions

5.1 Achievement of Goals

Evolutionary biology is the applied philosophy through which we attempt to comprehend the machinations of natural selection. This approach allows us to capture information from millions of years of natural engineering producing the diverse and highly regulated enzymes that exist today. Just as bricks can be assembled in various configurations to create structures with differing functions, enzymes too evolve through modular principles, allowing for diversification while retaining core functionalities. This dissertation explored the underlying principles of modularity in enzyme evolution, illuminating how nature, like an expert architect, repurposes, combines, and innovates to craft the intricate web of biological pathways that sustain life. Through our findings, we gain a deeper appreciation for the evolutionary dynamics that have shaped the enzymatic repertoire of living organisms and provide insights that may guide future endeavors in synthetic biology and bioengineering.

5.2 Modular evolution of GT-As

Here we dived into the evolution of fold A glycoslytransferases (Venkat, Tehrani, et al. 2022), mining the sequential and structural patterns hidden within glycosyltransferase sequences to construct an evolutionary trajectory of GT-As. Specifically, we established that GT-As evolve in three distinct modules originating from the hydrophobic core, with further hypervariable modules embellishing the canonical GT-A scaffold, allowing for the extensive diversity of functions performed by modern GT-As today. We also uncover that the latest hydrophobic core module tethers the phosphate binding cassette, which contains the DxD motif, to the F-helix, which contains the catalytic base, the xED motif, differs between inverting and retaining glycosyltransferases. We further demonstrate through mutations in the tether that we can allosterically regulate GT-A function, elucidating how nuanced effects from variation of the tether help control GT-A function.

5.2.1 Future directions

This work is the foundation for multiple research directions. We focused on GT-A function and allosteric regulation of these enzymes, but no bioinformatics analyses have been performed on non-functional or pseudoenzyme variants of this superfamily. Pseudoenzymes are a burgeoning topic and have been identified in multiple families. One may ask what the purpose of a non-catalytic enzyme is, for what is the purpose of a brick, if not for building. Yet, nature has found many ways to repurpose these biological bricks to serve other functions, as chaperones, scaffolds, for localization and more (Ribeiro et al. 2019). While many pseudoenzymes have been identified for other enzymes, only a single pseudo GT-A has been identified (COSMC) (Yingchun Wang et al. 2010), but with existing sequence profiles of GT-As, it seems feasible (although no doubt difficult) to uncover non-canonical and potential pseudoenzymes through systematic analysis of GT-A sequence. However, this is not trivial because extensive diversity in glycosyltransferases often do not result in loss of function, but simply variance of it. I investigated non-canonical and potential pseudoenzymes using COSMC as a case-study (see Appendix A). However, I believe that the emergence of deep learning and specifically protein language models like ESM2, will improve our ability to discern hidden patterns of functionality, or lackthereof, in the GT-A sequence space.

While we have focused predominantly on fold A glycosyltransferases, our work also reveals that GT-B fold enzymes appear to have convergently evolved glycosyltransferase function, based on differences in the topology of their phosphate binding cassette (PBC), as compared to GT-As. GT-B fold enzymes appear to share a PBC similar to traditional Rossmann fold enzymes, whereas the PBC of GT-As matches enzymes like pyrophosphorylases. Pyrophosphorylases also bind metal cations akin to GT-As, whereas GT-B enzymes do not bind metal ions at all for function, aking to traditional Rossmann fold enzymes. This additionally may explain why little sequence similarity exists between GT-A and GT-B fold enzymes, despite a shared function and Rossmann-like fold. I hypothesize that while these enzymes may share an ancient ancestor, they may have independently converged on sugar-transferring function.

Thus, a natural direction is to focus on GT-B enzymes independently to understand how they evolved and unique patterns that contributed to functional diversication in these enzymes. Compared to other enzymes, glycosyltransferases have less structural representation, with a large number of GT-A and GT-B families still left uncharacterized. But with the advent of AlphaFold2, alignment of these enzymes and identification of the limits of their domains is now feasible. I first generated Hidden Markov Model profiles of every known GT-B family and then pulled all sequences matching these GT-B profiles from the Uniprot database. The resultant 4̃00,000 GT-B sequences were then filtered and aligned (see Appendix B), resulting in the first global alignment of GT-B sequences. The resultant sequence alignment contained over 20,000 columns, and was therefore cleaned of inserts (see Appendix C), to create an easy-to-visualize sequence alignment and sequence logo. With the availability of this alignment, we can now create comparative phylogenetic trees of GT-B enzymes, through both a sequence-based nearest-neighbor tree, as well as an RMSD-based distance tree by structural alignment of accurate AlphaFold structures (see Appendix D). These trees would allow for evolutionary and functional insights into the GT-B family and would represent the first effort to globally align this enzyme family. GT-B enzymes are an vastly important family of enzymes, and the availability of easier compute resources and deep learning tools enables study of the superfamily as a whole.

5.3 Evolution of a multi-functional tail in Doublecortin-like Kinases

In my second project, we elucidated the mechanisms of a multi-functional C-tail in Doublecortin-like kinases, which utilize this tail as a swiss army knife of regulatory functions in microtubule dynamics and neuronal development. We show how engineering of specific mutations in the autoinhibited DCLK1.2 isoform rescues DCLK1 activity, pointing to an evolutionary adaptation to autoregulate function depending on environmental influences (Venkat, Watterson, et al. 2023).

5.3.1 Future directions

Our focus on DCLK1 can really be expanded to kinase allostery as a whole, given that their conserved domains are like a molecular chassis, which kinases have embellished upon over evolutionary time through N and C terminal modular additions. DCLKs belong to the CAMK family of kinases and though some information has been uncovered about how these enzymes uniquely regulate kinase function through variable and intrinsically disordered C-tails. We can further expand into looking at DCLK2 and DCLK3 as models for understanding orthologous function of the DCLK family as a whole. DCLK3, especially, is one of least studied kinases, with a single paper identifying it as a neuroprotectant in Huntington's disease (Galvan, Francelle, Gaillard, Longprez, Carrillo-de Sauvage, Liot, Cambon, Stimmer, Luccantoni, Flament, et al. 2018a). Investigation of its structure and how it is post-translationally modified, as well as mechanistic insights into its function and potential substrates will greatly improve our ability to understand its role and potential interaction with the Huntingtin protein.

Additionally, and importantly, we can use what we have learned from DCLK1 as a model for how kinases may use alternative splicing mechanisms to directly modulate enzyme function through minor changes in intron/exon placement. Little exploration aside from our paper has delved into the genetic basis of kinase regulation. We show in our own paper that availability of isoform data is often quite poor due to misannotation or lack of annotation in protein sequence databases. It is therefore necessary to establish automated protocols to study the global effect of intron/exon variations that mediate kinase activity.

5.4 Development of a deep learning predictive classifier of GT-A function

Through use of novel protein language models, we created a classifier that predicts GT-A family and donor substrate binding. Sugar donors are expensive, thus this model offers a lucrative method to cut-down on testing sugar-donors for unknown sequences by computational prediction and hypothesis generation.

5.4.1 Future directions

Naturally, we have started with five well-studied and well-represented donors. However, there are nine mammalian GT-A donors available for testing. Because the rest of these donors are not as well represented in Uniprot, they are difficult to incorporate without loss of accuracy. With appropriate dataset augmentations, loss functions, and regularization, I believe these donors could also be included within the model. Eventually, all donors could be included and the classifier could extend to GT-B families as well with sufficient data availability (see appendix B). As sequence embeddings are higher-dimensional numerical representations of sequences (vector representations), we do not require an alignment (Yeung, Zhou, S. Li, et al. 2023). Instead we can directly compare sequence embeddings to derive function-predictions. This opportunity is ripe to pursue as GT-Bs are enormously difficult to compare through traditional alignment methods because of little sequence conservation. As protein language models consider a holistic set of features of a protein, capturing the grammar and syntax of amino acids from chemistry to evolution, their applications will be invaluable in future bioinformatics studies.

References

Galvan, Laurie, Laetitia Francelle, Marie-Claude Gaillard, Lucie de Longprez, Maria-Angeles Carrillo-de Sauvage, Géraldine Liot, Karine Cambon, Lev Stimmer, Sophie Luccantoni, Julien Flament, et al. (2018a). "The striatal kinase DCLK3 produces neuroprotection against mutant huntingtin". *Brain* 141.5, pp. 1434–1454.

Ribeiro, António JM et al. (2019). "Emerging concepts in pseudoenzyme classification, evolution, and signaling". *Science Signaling* 12.594, eaat9797.

Taujale, Rahil, Aarya Venkat, et al. (2020). "Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases". *Elife* 9, e54532.

Varki, Ajit, Richard D Cummings, et al. (2022). "Essentials of Glycobiology [internet]".

Venkat, Aarya, Daniel Tehrani, et al. (2022). "Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases". *Journal of Biological Chemistry* 298.8.

Venkat, Aarya, Grace Watterson, et al. (2023). "Mechanistic and evolutionary insights into isoform-specific 'supercharging'in DCLK family kinases". *bioRxiv*.

Yeung, Wayland, Zhongliang Zhou, Sheng Li, et al. (2023). "Alignment-free estimation of sequence conservation for identifying functional sites using protein sequence embeddings". *Briefings in Bioinformatics* 24.1, bbac599.

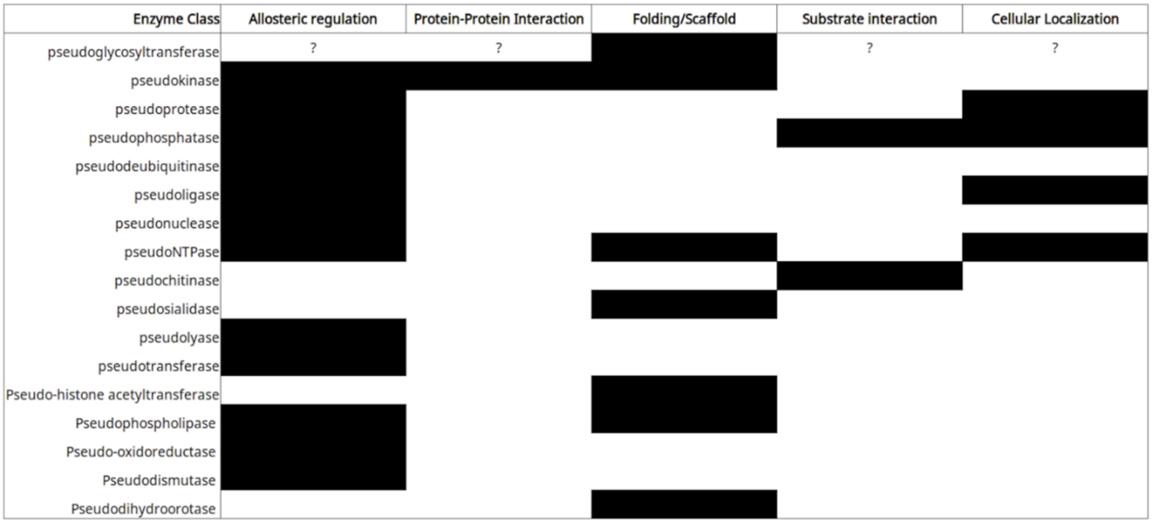
Appendix A

Evolutionary Origin and Emergence of Pseudoglycosyltransferases

A.1 Abstract

A.2 Introduction

Across evolutionary time, enzymes have demonstrated an abundance of ways to catalyze a particular reaction. For glycosylation, this means varying donor and acceptor substrates, varying the location of a glycosidic linkage, as well as the stereochemistry of the glycosidic bond (alpha or beta). Glycosyltransferases (GTs), which catalyze glycosidic linkages, also come in four distinct folds (A, B, C, and lyso), which differ by structure and subcellular location (Varki, Richard D Cummings, et al. 2022; Taujale, Zhou, et al. 2021). We previously mined over 600,000 GT-A fold sequences from all domains of life, showing extraordinary sequence diversity across the tree of life, facilitating numerous critical cellular functions, including protein folding, signaling, and stability (Taujale, Venkat, et al. 2020; Taujale, Soleymani, et al. 2021; Venkat, Tehrani, et al. 2022). In the face of this incredible sequence diversity, we find numerous GT-As that contain amino acid variations at key catalytic and functional motifs, possibly indicating catalytic impairment. Notably, catalytic inactivation of enzymes has been noted before, prevailingly in the kinase field (Kwon et al. 2019; Shrestha et al. 2020). These so-termed "pseudokinases" and other pseudoenzymes still function in central roles, working as dynamic scaffolds, allosteric modulators, chaperones, and used for cellular localization (Figure 1). However, due to difficulties with expression and crystallization with glycosyltransferases, the exploration of possible catalytically inactive GTs has been limited, but the recent availability of the sequence data and predicted structures via AlphaFold2 reignite the exploration into classifying non-canonical and pseudoGTs.



Enzyme Class	Allosteric regulation	Protein-Protein Interaction	Folding/Scaffold	Substrate interaction	Cellular Localization
pseudoglycosyltransferase	?	?	■	?	?
pseudokinase	■	■	■		
pseudoprotease	■				■
pseudophosphatase	■			■	■
pseudodeubiquitinase	■				
pseudoligase	■				■
pseudonuclease	■				
pseudoNTPase	■		■		■
pseudochitinase				■	
pseudosialidase			■		
pseudolyase	■				
pseudotransferase	■				
Pseudo-histone acetyltransferase			■		
Pseudophospholipase	■		■		
Pseudo-oxidoreductase	■				
Pseudodismutase	■				
Pseudodihydroorotase			■		

Figure A.1: Table of pseudoenzymes and experimentally determined pseudoenzyme functions

Pseudoenzymes are a burgeoning topic of study (Ribeiro et al. 2019) with close sequence homology to extant enzymes but lacking the equivalent catalytic function. While pseudoenzymes are commonly defined by the loss of canonical motifs or differences in key sequence variations (James M Murphy, Farhan, and Patrick A Eyers 2017), classifying pseudoGTs by this metric is not a trivial matter. GTs vary active site residues for a variety of reasons, from modulating binding of a divalent cation to swapping a mechanism from an Sn2 to an Sni mechanism. This high level of sequence variation, even in key catalytic motifs, makes it difficult to apply the pseudo label to GTs, as variation in these motifs often defines variation of chemistry and catalytic mechanism, rather than absence of function. Yet catalytically dead glycosyltransferases, such as COSMC, have been identified, losing their primary catalytic function and instead adopting a new function, such as a molecular chaperone (Ju and Richard D Cummings 2002; Yingchun Wang et al. 2010). Here we seek to delineate the features of canonical, non-canonical, and pseudoGTs to establish a method of classification for these diverse enzymes.

The closest evolutionary relative to COSMC is the catalytic T-Synthase enzyme, yet COSMC has no catalytic activity itself. It is instead predicted to act as a scaffold for T-synthase folding. T-Synthase activity is dependent on COSMC (Ju and Richard D Cummings 2002; Yingchun

Wang et al. 2010). Utilizing state-of-the-art sequence and structure methods, we explore the structure-function and evolution of the dynamic interplay between COSMC and T-Synthase as a case-study to understand how pseudogenization may occur in this diverse family.

A.3 Results

Our investigation into the distribution and characteristic features of non-canonical GTs revealed their widespread occurrence across the tree of life, a trend similar to what is observed with pseudokinases (Figure 2). However, unlike pseudokinases, the variation of catalytic motifs within these GTs does not signify a loss of function, but rather a diversification of it.

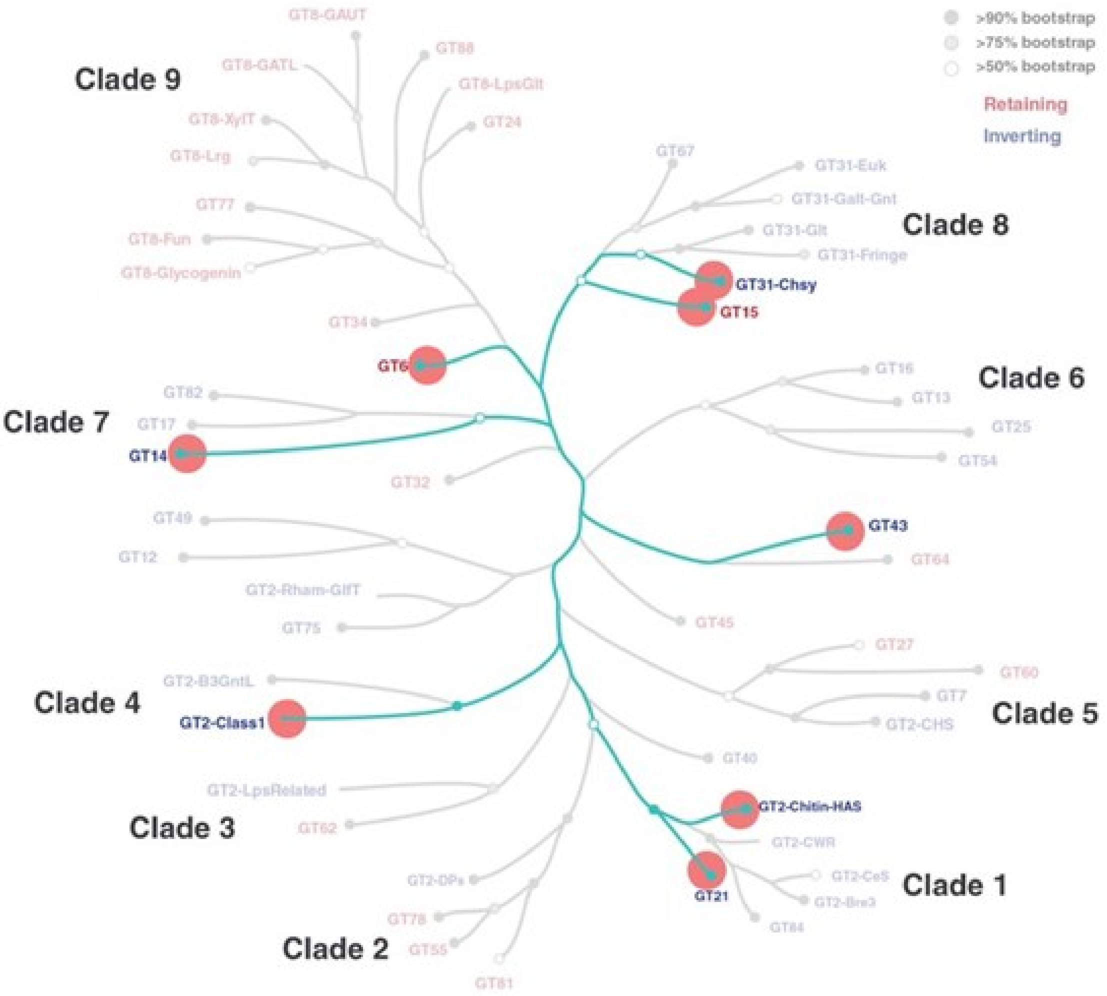
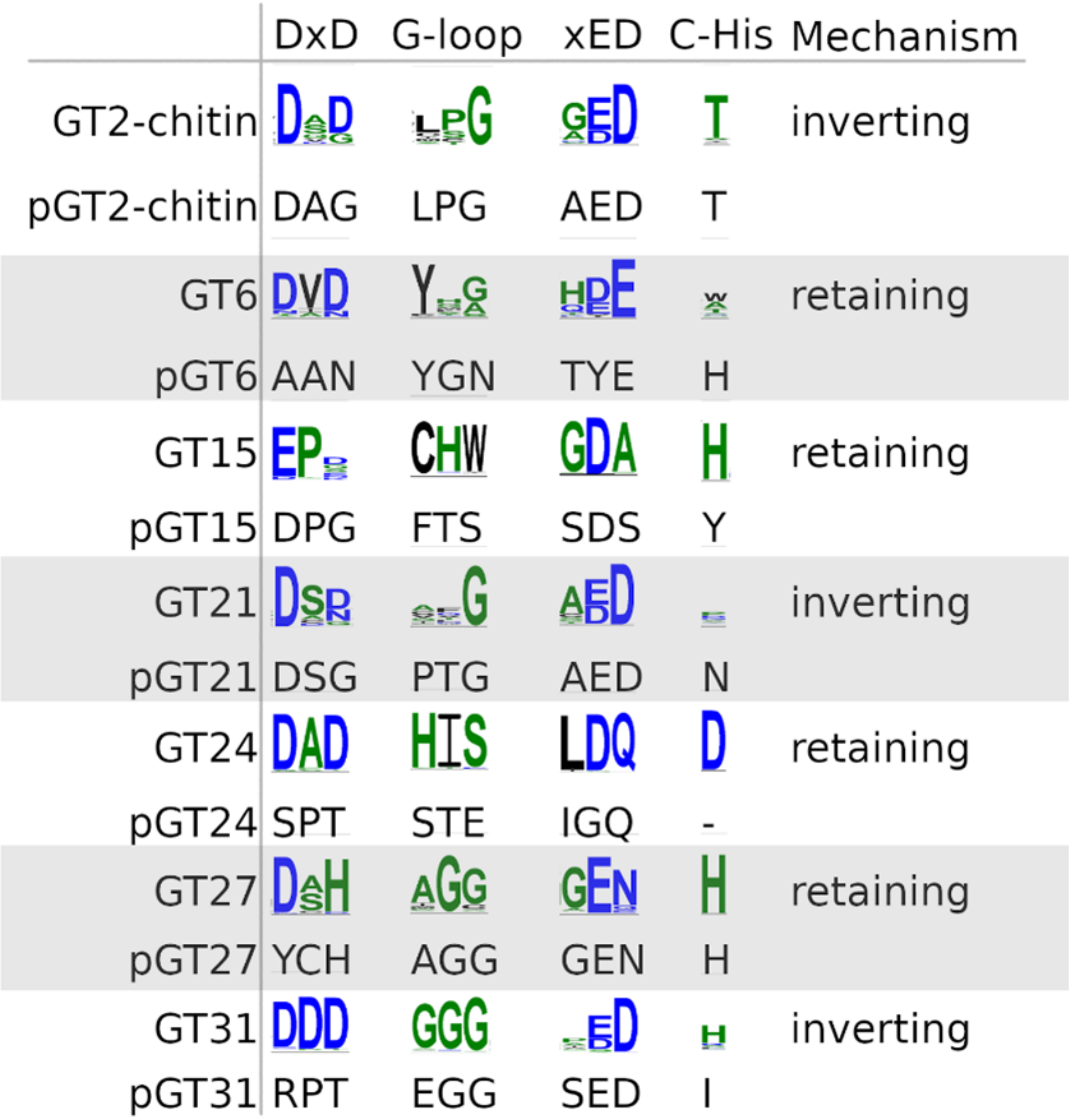


Figure A.2: GT-A phylogenetic tree highlighting non-canonical GT-As, GT-As lacking the DxD motif, across the tree of life

One striking example in GT-As is the catalytic base Aspartate (the xED motif) whose function varies in necessity: it is crucial for inverting GTs, but not for retaining ones (Venkat, Tehrani, et al. 2022). This variation is linked to the functional diversification of the GT-As (Taujale, Venkat, et al. 2020). Similarly, we observed variation in the G-loop, which is modulated to accommodate the specific needs of binding different acceptor substrates.



	DxD	G-loop	xED	C-His	Mechanism
GT2-chitin					inverting
pGT2-chitin	DAG	LPG	AED	T	
GT6					retaining
pGT6	AAN	YGN	TYE	H	
GT15					retaining
pGT15	DPG	FTS	SDS	Y	
GT21					inverting
pGT21	DSG	PTG	AED	N	
GT24					retaining
pGT24	SPT	STE	IGQ	-	
GT27					retaining
pGT27	YCH	AGG	GEN	H	
GT31					inverting
pGT31	RPT	EGG	SED	I	

Figure A.3: Short list of non-canonical enzymes using the CAZy-numbering scheme and how they vary motifs critical to GT-A function

Moreover, we noticed a significant dependency of GT-A function on metal binding, largely mediated by the DxD motif. However, non-canonical GTs such as GT14 and RRGAT1 deviate from this norm, presenting variations in the DxD motif. They exhibit seemingly metal-independent functions, utilizing other pocket residues for donor substrate binding. Upon closer examination of the canonical motifs, we identified that a large proportion of the non-canonical GTs we studied exhibited variations in the DxD motif (Figure 3). These findings suggest that non-canonical GTs could use different molecular strategies for function, hinting at a broader functional diversity within the GT family than previously understood.

Much like pseudokinases, non-canonical GTs are found across the tree of life (Figure 2). However, unlike pseudokinases, variation of catalytic motifs in GTs is associated with variation in function, rather than absence of it. GT-As, for example, vary in the use of the catalytic base Aspartate (xED motif), which is critical for function in inverting GTs but not in retaining GTs (Taujale, Venkat, et al. 2020; Moremen and Haltiwanger 2019). The G-loop is varied based on the needs of binding a specific acceptor substrate (Taujale, Venkat, et al. 2020). GT-A function is also largely metal-dependent, due to the metal binding DxD motif, yet GT14 and RRGAT1 are non-canonical GTs that vary the DxD motif and seemingly have metal-independent function, where other residues in the pocket are used to bind the donor substrate (Amos et al. 2022). Looking strictly at the canonical motifs, many of the non-canonical GTs we have identified tend to vary the DxD motif (Figure 3).

Because of the great variability of GTs, variation in key canonical motifs may not indicate pseudogenization. But combining sequence analyses, literature searches, and the availability of accurate structure-prediction methods, such as AlphaFold2, we can link how key variations in sequence affect structure and consequently function of an enzyme. We observe several mutations in the DxD motifs of non-canonical GTs (Figs 3 and 4). Intriguingly, both GT14 and RRGAT1 have lost the DxD motif entirely at the family level, suggesting an evolutionary shift towards alternate mechanisms for donor substrate binding. Considering that the majority of GTs rely on metal binding for their catalytic activity, the loss of residues crucial for binding metal cations bears significant implications for the enzyme's affinity for the donor substrate.

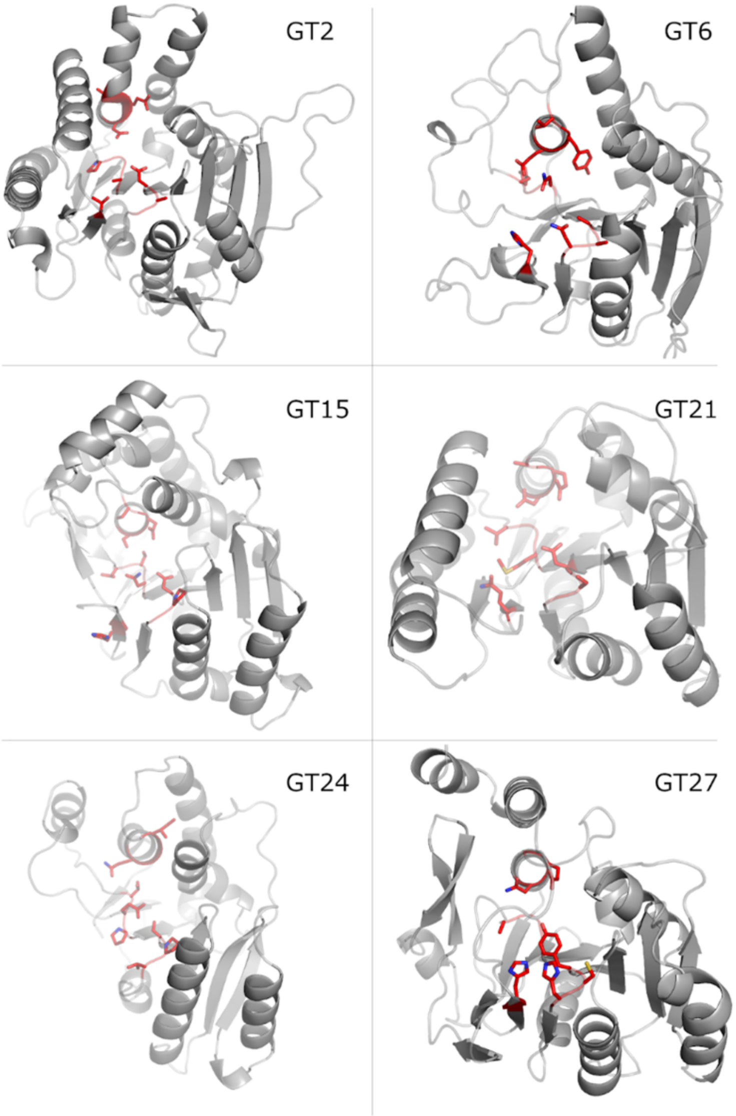


Figure A.4: AlphaFold structures of non-canonical GT-A enzymes, with DxD (bottom) and xED (top) motifs highlighted in red sticks

This finding elevates the importance of those individual sequences within a family that loses the DxD motif while the motif is otherwise conserved. The likely evolutionary timescale suggests it's improbable that an alternate substrate-binding method has had sufficient time to evolve. Therefore, these outlier sequences may be particularly valuable subjects for further study, potentially providing insights into the early stages of functional adaptation and diversification in GTs.

GT14 and GT116 (RRGAT1) have collectively lost the DxD motif at a family level (Briggs and Hohenester 2018; Amos et al. 2022), evolving alternate methods of binding the donor substrates. Because most GTs are metal-dependent for catalytic activity, losing the key residues that facilitate binding the metal cation is significant because it directly affects enzyme affinity for the donor substrate. Therefore, individual sequences that lose the DxD in a family that conserves this motif may be worth inspecting because it is unlikely enough time has passed to evolve an alternate method of binding the donor substrate.

A.3.1 Exploration of Non-Canonical GT-As taken from 185,000 AlphaFold2 structures

Given the limited number of crystallized GT-A sequences available, we relied on the structure-prediction tool AlphaFold2 for structural analysis. While these models are predictions and come with associated uncertainties, the majority of predicted AlphaFold2 structures matched their respective crystallographic counterparts within a margin of 1.1 angstroms (Figure S1). This consistency gives us confidence in the utility of the tool for our analysis.

For our study, we applied AlphaFold2's pLDDT metric to evaluate the confidence level of predicted structures. By analyzing the architecture of these anomalous structures, we could make initial inferences regarding potential pseudogenization within GT-As, offering new insights into the functional evolution of these diverse enzymes.

As most GT-A sequences are not crystallized, AlphaFold2 (Jumper et al. 2021) has become an excellent source for structural analysis. While it is necessary to include the caveat that these are predicted models, we note that a majority of predicted AlphaFold structures match their crystallographic matches within 1.1 angstroms [Fig S1]. Moving on this assumption and filtering structures using AlphaFold2's pLDDT metric for evaluating structure-confidence (see appendix D), we plan to pull structures that deviate from the canonical DxD motif out of the the 185,000 available structures and analyze the architecture to make possible inferences about modifications to the active site of GT-As.

A.3.2 COSMC as a case study for pseudoGTs.

Building on our findings on the potential pseudogenization within GT-As, we proceeded with a case study to explore features that might define a pseudoGT. We chose to focus on the GT-A COSMC, which is known to lack catalytic function (Ju and Richard D Cummings 2002; Yingchun Wang et al. 2010) but plays an essential role as a chaperone for an evolutionary relative, T-synthase.

We found that COSMC likely evolved through a duplication event from an ancestral T-synthase (Figure 5). Unlike their ancestral counterparts, eukaryotic T-synthase enzymes require COSMC as a chaperone for proper folding. Notably, key differences exist in the DxD motif and the C-His sequence between T-synthase and COSMC. The former alters the DxD motif to RPT in COSMC, while the latter is completely absent in COSMC (Figure 5B-C).

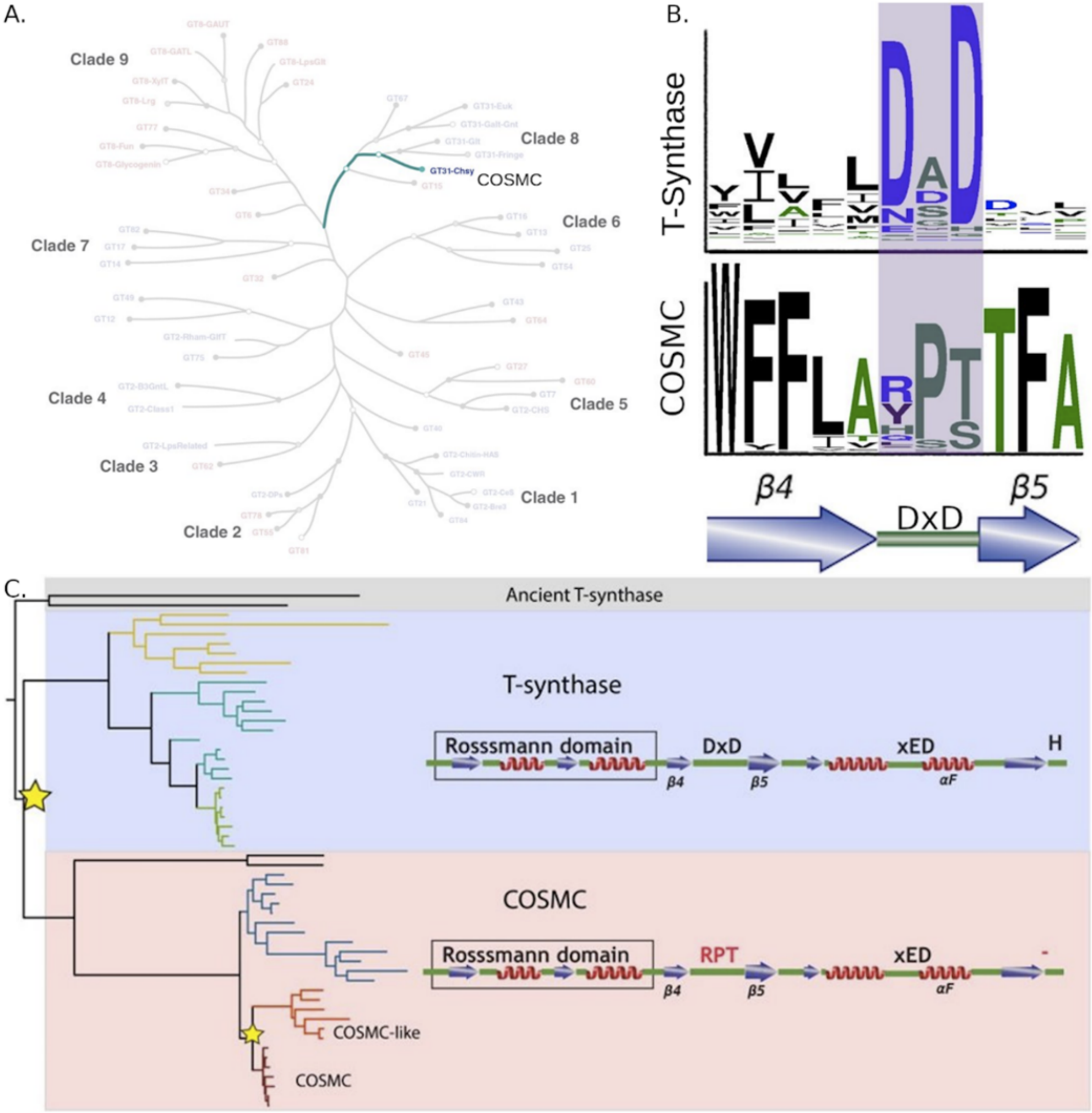


Figure A.5: Evolution of COSMC and T-Synthase via duplication event. A) Phylogenetic tree highlighting the location of COSMC on the GT-A phylogenetic tree. B) Sequence logo depicting the loss of the DxD constraint between closely related sequences COSMC and T-Synthases. C) Phylogenetic tree of T-Synthase and COSMC evolution, where stars depict a duplication event. Secondary structure and key motifs of each human sequence is shown.

The DxD motif and C-His sequence are critical for metal cation binding in GTs. We hypothesize that the absence of these motifs in COSMC results in its inability to bind the Mn2+ ion that

T-Synthase binds. However, COSMC is still capable of binding Zn2+ and Fe2+ with its C-terminal tail, a process important for oligomerization. This C-terminal tail forms part of the hypervariable region 3 in the COSMC sequence, a region involved in a range of functions including substrate affinity and binding.

To further probe the relationship between COSMC and T-Synthase, we used AlphaFold2-Multimer to generate an oligomeric structure comprised of two molecules each of T-Synthase and COSMC. Our model predicts a strand-exchange interaction between the C-terminal tails of COSMC and T-Synthase, as well as an interaction between the hypervariable region 1 (HV1) of COSMC and the N-terminal recognition domain of T-Synthase.

Further investigations using Bayesian analyses revealed key constraints in both COSMC and T-Synthase. Combined with interactions in the hypervariable regions, these constraints suggest a network of interactions that extends from the catalytic pocket to HV regions 1 and 3, which seem to mediate the binding of T-Synthase to COSMC and vice-versa (Figure 6).

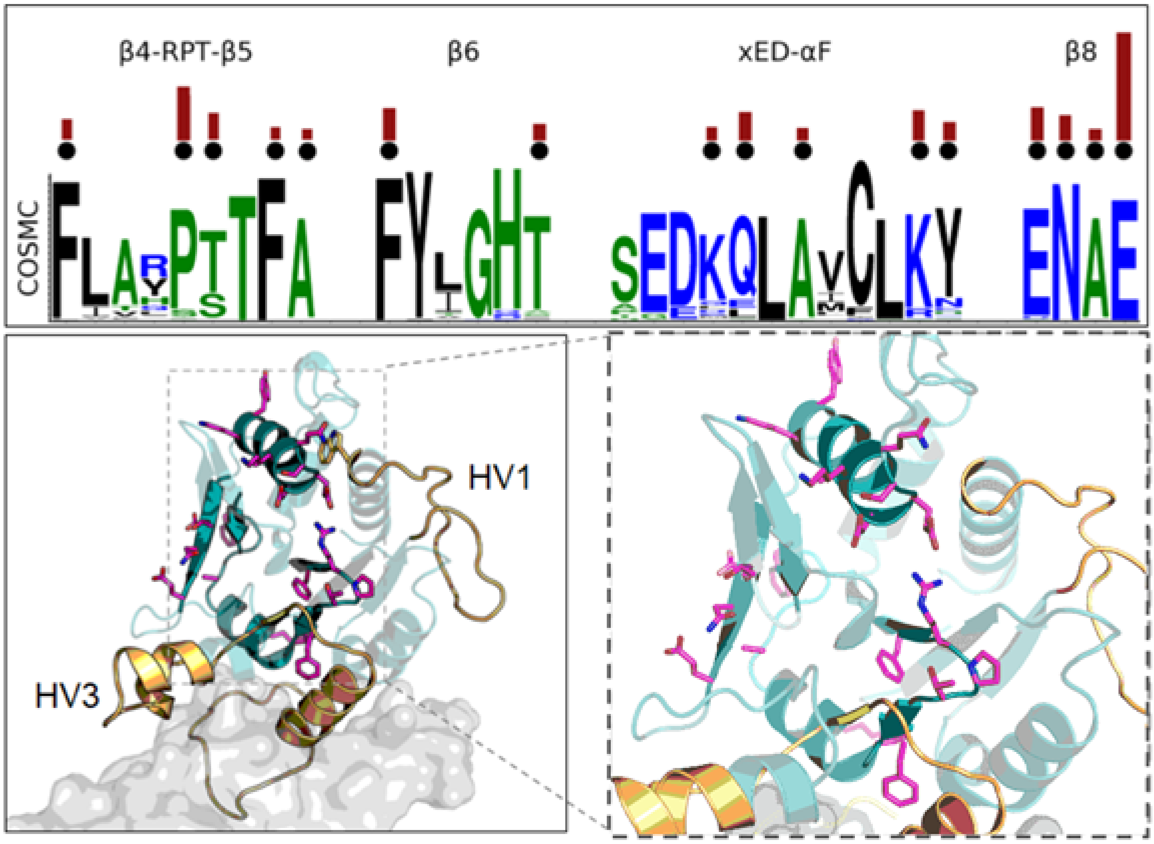


Figure A.6: Sequence logo of COSMC-specific patterns, identified via Bayesian analysis. Red histograms indicate level of significance of a given pattern residue. These patterns were then mapped onto an AlphaFold structure, showing a network of connected residues across the active site, connecting the active site to HV regions.

Intriguingly, COSMC's HV1 appears to be lengthened compared to most HV1s in GT-As, and is predicted to interface around the catalytic pocket of T-Synthase. Also, the N-terminal recognition domain of T-Synthase, which is experimentally demonstrated to be specific for COSMC binding, is predicted to interface with COSMC's HV1.

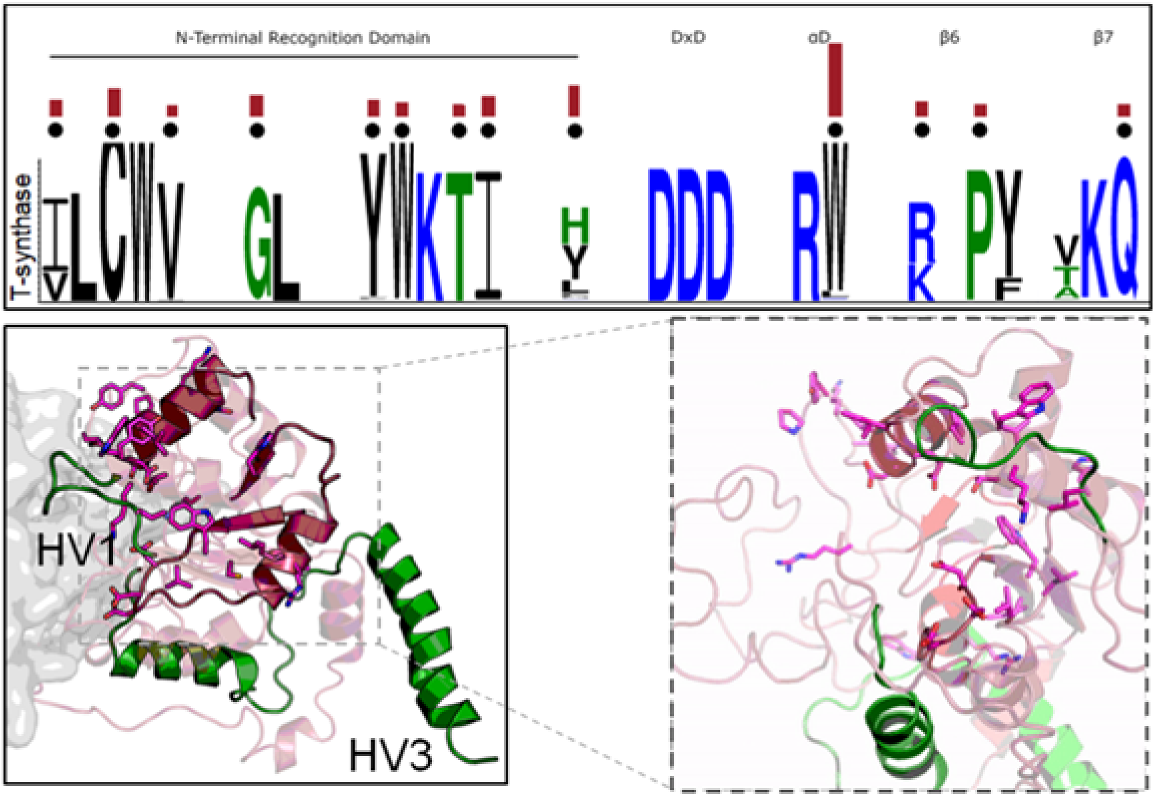


Figure A.7: Sequence logo of T-synthase-specific patterns, identified via Bayesian analysis. Red histograms indicate level of significance of a given pattern residue. These patterns were then mapped onto an AlphaFold structure, showing a network of connected residues across the active site, connecting the active site to HV regions.

To further investigate the function of these constraints identified by bayesian sequence analyses, we performed molecular dynamics simulations at the microsecond timescale for both the monomer and the predicted tetrameric structure.

To investigate features that may define a pseudoGT, we employed our analyses on the GT-A COSMC as a case study. COSMC has previously been stated to lack catalytic function (Ju and Richard D Cummings 2002). Additionally, it has been implicated as a necessary component for the folding of an evolutionary relative, T-synthase. We show COSMC evolved in a duplication event from an ancient T-synthase (Fig 5). Where ancestral T-synthase enzymes are capable of folding without a chaperone, eukaryotic T-synthase requires COSMC as a chaperone (Hanes, Moremen, and Richard D Cummings 2017). The key differences in catalytic motifs between T-synthase and COSMC are in the DxD motif, where COSMC mutates the DxD to RPT, and the C-His, a deletion in the COSMC sequence (Fig 5B-C).

COSMC as being unable to bind the Mn divalent cation that T-Synthase binds. Interestingly, COSMC is capable of binding Zn2+ and Fe2+ with the C-terminal tail, which is pertinent for oligomerization (Hanes, Moremen, and Richard D Cummings 2017). This C-terminal tail is part of hypervariable region 3 in the COSMC sequence. Hypervariable regions are considered to be used for substrate affinity, binding, and a host of other functions (Taujale, Venkat, et al. 2020; Venkat, Tehrani, et al. 2022). To evaluate the interactions between COSMC and T-Synthase, we used AlphaFold2-Multimer to generate an oligomeric structure composed of two T-synthase molecules and two COSMC molecules. Interestingly, our AlphaFold model predicts a strand-exchange interaction between the C-terminal tails of COSMC and T-Synthase, as well as an interaction between HV1 of COSMC and the N-terminal recognition domain of T-Synthase (70-93% accuracy measured by pLDDT).

Using Bayesian analyses, we further identify key constraints in COSMC and T-Synthase which, combined with interactions in the hypervariable (HV) regions, suggest a network of interactions extending from the catalytic pocket to HV regions 1 and 3, which seem facilitate the binding of T-Synthase to COSMC and vice-versa (Fig 6). Specifically, we note that COSMC seems to extend HV1, where most HV1s in GT-As are about four residues (Taujale, Venkat, et al. 2020). The length of this HV1 is 12 residues, predicted by AlphaFold2 to interface around the catalytic pocket of T-Synthase. Further, residues classified as the N-terminal recognition domain of T-Synthase, a region experimentally demonstrated to be specific for binding COSMC (Hanes, Moremen, and Richard D Cummings 2017), is predicted to interface with COSMC's HV1.

We investigated the function of BPPS-identified constraints using molecular dynamics simulations. Simulations were performed at the microsecond timescale for the monomer and the predicted tetrameric structure.

A.4 Methods

Phylogenetic and sequence analysis.

We used a previously published phylogenetic tree (elife 2020). Sequences containing variations of the DxD motif were identified by scanning through previously generated sequence profiles. These sequences were used to highlight branches on the phylogenetic tree to place where pseudoGTs appeared among GT-A families. Sequence logos were generated using WebLogo 3 and GTXplorer (Crooks et al. 2004; Taujale, Soleymani, et al. 2021). Sequence constraints were generated using the Bayesian Partitioning with Pattern Selection (BPPS) software (Andrew F Neuwald 2014), highlighting specific amino acid constraints for T-synthase and COSMC, along with histograms that detailed the significance of a given pattern.

Structure prediction, analysis, and visualization.

Predicted monomers and oligomers were generated for COSMC and T-synthase using AlphaFold2 (Jumper et al. 2021) with the multimer option. Full length sequences were provided as input and structures predicted were first filtered using the pLDDT confidence metric before analysis. A python script was written to map AlphaFold2 pdb numbering to existing sequence profiles, allowing for us to match sequence constraints onto the pdb structure. These structures were visualized using Schrodinger PyMol 2.6.

Comparison of AlphaFold and crystal structures.

GT-A Crystal structures were pulled from RCSB. Corresponding Uniprot IDs were pulled from the same RCSB page and compared to the AlphaFold structure produced for the given Uniprot sequence. The structures were then aligned using the ceAlign algorithm in Schrodinger PyMol 2.0, producing an RMSD value. An additional structural alignment was performed using the tmAlign algorithm as a validation step, producing a TM score. A box-and-whisker plot was used to represent the corresponding RMSD and TM values.

Appendix B

Deep Evolutionary Analysis of Fold-B Glycosyltransferases

B.1 Abstract

Glycosyltransferases (GTs) catalyze the transfer of sugar moieties and are pivotal in the biosynthesis of carbohydrates, glycoproteins, and glycolipids. Among GTs, fold B glycosyltransferases (GT-Bs) are particularly intriguing due to their unique two-domain architecture and a vast substrate range. GT-Bs play a critical role in numerous physiological processes, with implications ranging from bacterial cell-wall synthesis to the modulation of host-pathogen interactions. Despite their importance, a comprehensive understanding of GT-Bs has been hampered by challenges associated with aligning sequences across diverse GT-B families. Traditional alignment methodologies often fall short in capturing the nuanced variations and conserved motifs due to the GT-Bs' inherent sequence diversity. To address this gap, this study introduced a pioneering methodology employing deep learning, specifically utilizing the tool "learnMSA." Our approach not only efficiently aligns GT-B sequences across various families, but also reveals previously obscured conserved motifs and functional nuances. By leveraging a Bayesian approach for sequence grouping and integrating high-confidence structural data, we present a comprehensive landscape of GT-Bs. Our study paves the way for deeper insights into the functional, evolutionary, and therapeutic potentials of GT-Bs, underscoring the transformative power of integrating machine learning into bioinformatics challenges.

B.2 Background

Glycosyltransferases (GTs) are a diverse group of enzymes responsible for catalyzing the transfer of sugar moieties from donor molecules to acceptor molecules, playing an indispensable role in the biosynthesis of carbohydrates, glycoproteins, and glycolipids (Varki, Richard D Cummings, et al. 2022). Among the diverse classes of GTs, fold B glycosyltransferases, often referred to as GT-Bs, stand out due to their unique structural features and functionalities. GT-Bs are characterized by a distinctive two-domain architecture, with both domains contributing to the active site formation (Varki, Richard D Cummings, et al. 2022; Moremen and Haltiwanger 2019). Despite sharing almost no significant sequence similarity with GT-As, another primary group of GTs, the two-domain structure of GT-Bs is reminiscent of the Rossmann fold. This evolutionarily conserved domain is typically associated with binding nucleotide cofactors (Venkat, Tehrani, et al. 2022), underscoring the significance of GT-Bs in the realm of enzymology and biochemistry. Previously we classified GT-As with a landmark paper that defined the landscape of GT-A fold enzymes (Taujale, Venkat, et al. 2020), now we seek to do the same with GT-Bs.

However, the diversity of GT-Bs goes beyond their structural uniqueness. Their vast substrate range, from simple sugars to complex polysaccharides, underpins numerous physiological processes. GT-Bs are implicated in a myriad of biological phenomena, including cell-wall synthesis in bacteria, glycosylation of proteins in eukaryotes, and the modulation of host-pathogen interactions (Moremen and Haltiwanger 2019). Their dysfunction can lead to a plethora of pathological conditions, emphasizing their importance in health and disease.

Despite their significance, a comprehensive understanding of GT-Bs has remained elusive due to challenges associated with sequence alignment across GT-B families. Traditional sequence alignment methodologies often fail to capture the nuanced variations and conserve motifs among the vast and varied members of the GT-B clan. These difficulties arise from the inherent sequence diversity of GT-Bs, punctuated by sporadic conservation patches, making it a daunting task to delineate meaningful patterns.

With the advent of deep learning, new horizons in bioinformatics and computational biology have emerged. Deep learning's ability to identify patterns from vast, seemingly unrelated data offers a promising avenue to address the longstanding challenge of GT-B alignment. By embracing a paradigm shift from traditional alignment methodologies to deep learning-based strategies, we venture into a novel approach to decode the mysteries of GT-Bs.

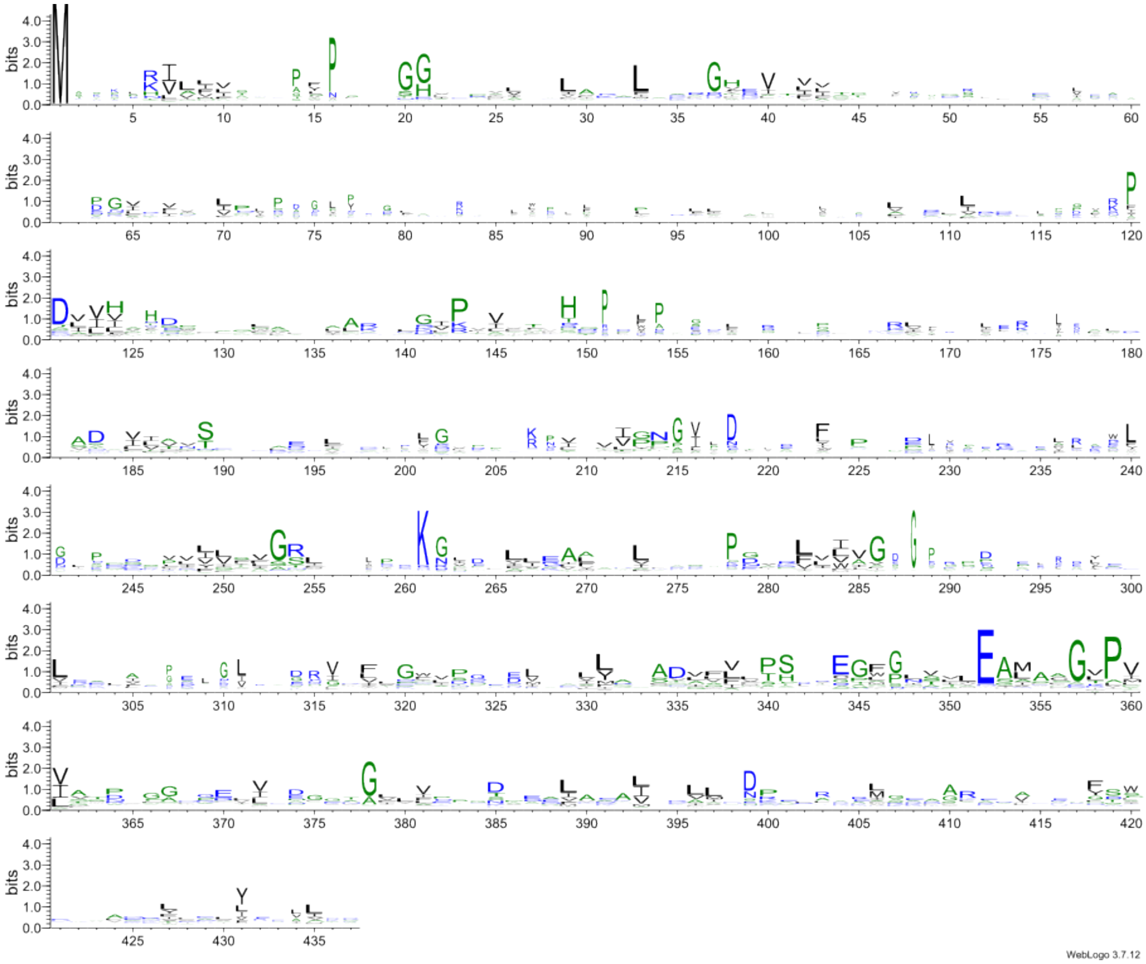


Figure B.1: Sequence logo of an alignment of GT-B sequences. N = 160,665

In this endeavor, we introduce a pioneering methodology that not only efficiently aligns GT-B sequences from various families but also unveils previously obscured conserved motifs and functional nuances. By offering a comprehensive, high-resolution view of the GT-B landscape, this study sets the stage for future investigations into the functional, evolutionary, and therapeutic potentials of this intriguing class of enzymes. By combining deep learning methods for sequence alignment, learnMSA (Becker and Stanke 2022), with traditional sequence methods, such as Hidden Markov Model profile alignments (Eddy 1996), we can generate precise profiles for enzyme domains across the GT-B family; I generated profiles from GT-B families classified in CAZy and used them to mine the Uniprot database for sequences related to my GT-B profiles. These sequences were aligned with learnMSA and GapClean (see Appendix C) was subsequently utilized to generate a clean alignment of sequences, with threshold-limited gaps for ease of visualization and interpretability. We demonstrate the power of this approach by presenting a sequence logo of over 160,000 aligned GT-B sequences, spanning the known GT-B families found in CAZy (Figure B.1). Our approach not only efficiently aligns GT-B sequences across various families, but also reveals previously obscured conserved motifs and functional nuances.

I performed a preliminary phylogenetic analysis using a trimmed alignment from the initial 160,000 sequences. This alignment was filtered at 60% using CD-Hit. After filtering, I constructed a radial rooted phylogenetic tree using the Jones-Taylor-Thornton maximum-likelihood model with FastTree (Figure B.2). An initial observation of the tree indicates that the GT1 family is widely spread throughout the tree of life. The tree is rooted between GT1 and GT4 families, suggesting a potential evolutionary GT-B ancestor from these families. It is critical to note that while this initial large tree was used to lay out relationships between GT-B families, many sequences lack sufficient annotation to make functional conclusions. There are two approaches that will be critical to interpreting functional and evolutionary information from these families. The first is careful curation of well-annotated sequences from Uniprot; a focus on model organisms to produce a phylogenetic tree as performed in (Venkat, Watterson, et al. 2023), will be necessary for confident interpretation.

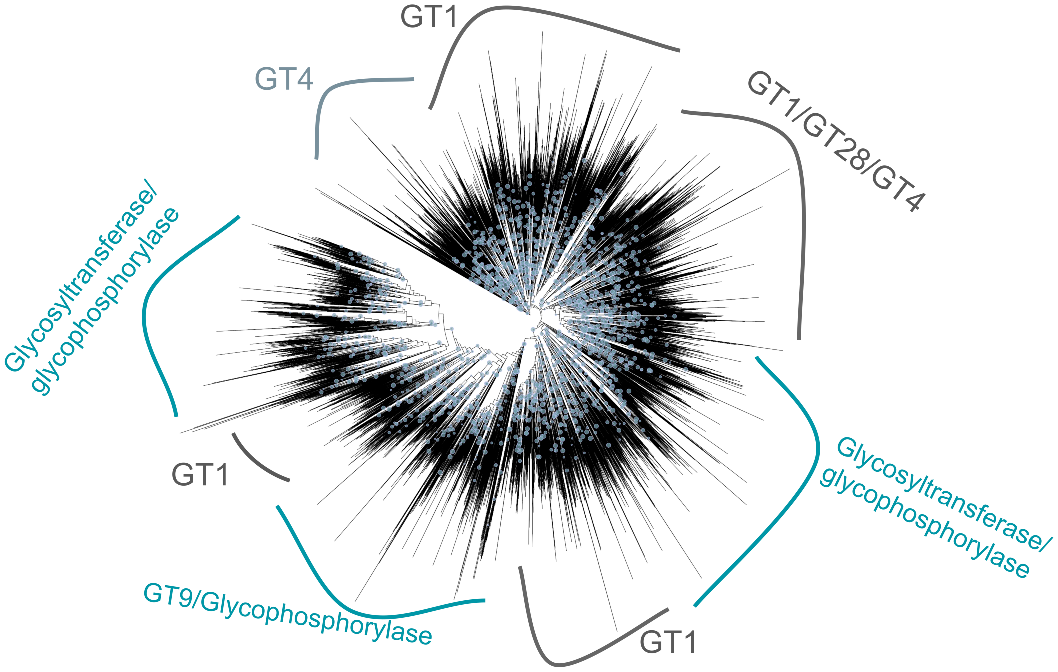


Figure B.2: Phylogenetic tree of GT-B families filtered at 60% from an alignment of 160,000 sequences. N = 53,028. Blue circles indicate bootstrap values above 98%.

Additionally, by leveraging bayesian approaches for sequence clustering and delineation of family-specific constraints (Taujale, Venkat, et al. 2020; Kwon et al. 2019), we can elucidate how evolutionary patterns led to functional divergence across GT-B enzymes. By integrating high-confidence structural data as well as highly confident predicted structures from AlphaFold2, we can present a comprehensive landscape of GT-Bs. This study can pave the way for deeper insights into the structure-function and evolution of GT-Bs.

Appendix C

GapClean: A tool for cleaning up sequence alignments.

C.1 Background

At the heart of bioinformatics lies the foundational task of sequence alignment, a method used to identify regions of similarity between biological sequences. This similarity can arise from functional, structural, or evolutionary relationships between the sequences. The primary goal of sequence alignment is to identify the optimal way to line up two sequences so that the highest number of matching characters (nucleotides or amino acids) can be achieved, taking into account possible gaps that might be introduced due to deletions, insertions, or evolutionary divergences.

There are generally two types of sequence alignment methods: global alignment and local alignment. Global alignment attempts to align every residue in every sequence, often useful when the sequences in question are of roughly equal size and are suspected to share a common ancestry. Local alignment, on the other hand, identifies regions of similarity within long sequences that are often widely divergent overall.

One of the pioneering global alignment algorithms is the Needleman-Wunsch algorithm, proposed by Saul B. Needleman and Christian D. Wunsch in 1970. This algorithm employs a dynamic programming approach to ensure that the optimal alignment is found. It systematically builds a matrix that keeps track of alignment scores at each position, ensuring that the final path chosen through this matrix represents the best possible alignment of the two sequences. The beauty of the Needleman-Wunsch algorithm lies in its ability to guarantee an optimal solution, but this comes at the cost of increased computational complexity, especially for long sequences.

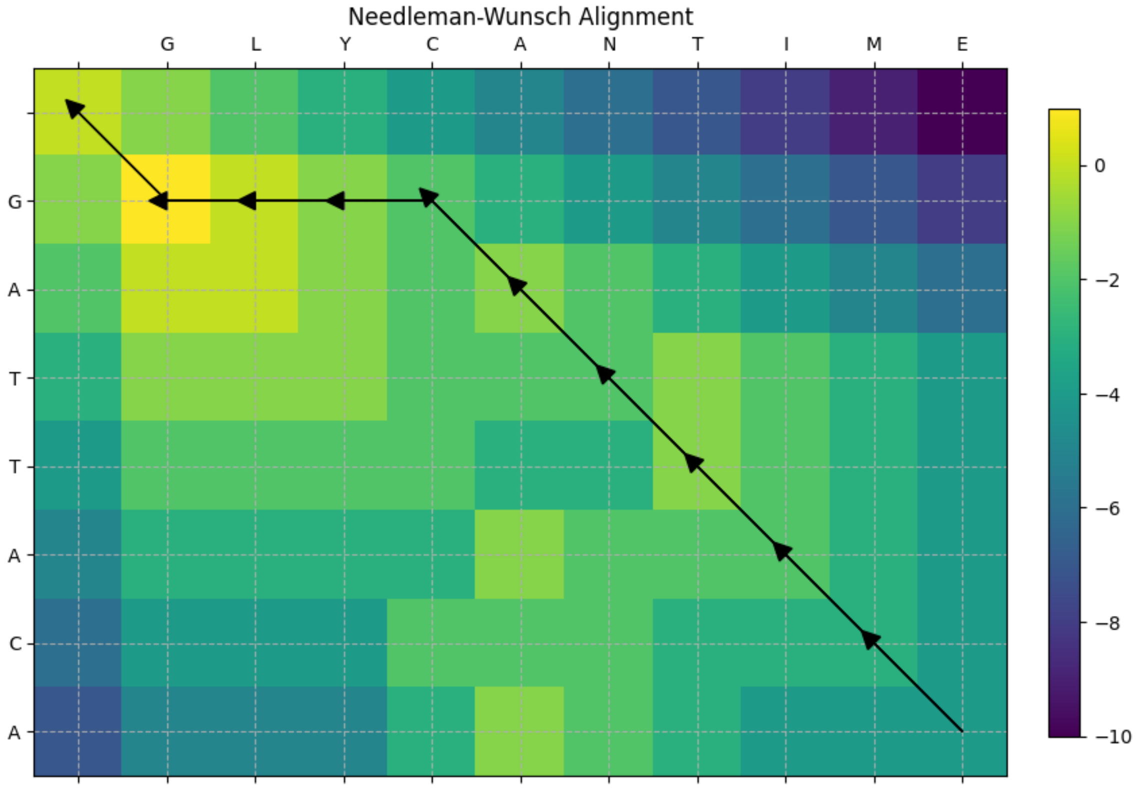


Figure C.1: Example of a Needleman-Wunsch Alignment, with a black arrow representing the optimal traceback, based on gap, match, and mismatch scores.

The Needleman-Wunsch algorithm and its derivatives have set the stage for many subsequent innovations in sequence alignment. Over the decades, a plethora of software tools have been developed to tackle the challenge of aligning sequences, where modern tools are now capable of aligning millions of sequences.

Naturally, the alignment of a million sequences also poses a critical problem. It is difficult to visualize and comprehend a large sequence alignment. How do we handle millions of inserts and gaps that vary between every sequence?

I developed a tool, GapClean, for this exact task. GapClean takes a gappy multiple sequence alignment and removes columns at a threshold value. For example, at a threshold of 70%, any position in a sequence alignment containing more than 70% gaps across the entire alignment, will be removed. This refocuses the alignment on the larger conserved segments of the protein domain. It would aid in identifying key motifs and for building an overall consensus sequence over a massive sequence alignment.

The application turns a sequence alignment into an array of i x j characters, where i represents a row, and j represents a column. Because a sequence alignment ensures that all sequences are the same length, accounting for inserts and deletions, we can simply count the presence or absence (indicated by a dash, "-") of an amino acid across the column. If we iterate over i, over all columns, we can tabulate the number of deletions in a column, or alignment position, and delete the column if it does not meet an input threshold. Through this method, we can eliminate family or subfamily-specific inserts in a sequence alignment, and output an alignment that highlights shared features across the sequence space. It is also critical to note that arrays are generally easier to compute with, rather than strings ("text"), thus the computational time required for processing sequences, even as large as one million sequences in an alignment, would dramatically decrease.

As a case study, I used this tool to truncate the GT-B alignment of 1̃60,000 sequences to capture the GT-B domain (see appendix B), despite extensive diversity across the GT-B sequences. This trimmed a 13GB sequence alignment with 2̃0,000 aligned positions to a 300MB file with 4̃00 aligned positions. GapClean can therefore be an incredibly useful tool for handling large sequence alignments. I plan to create an easy user-interface for biologists of all backgrounds to be able to use for their own needs.

In summary, GapClean addresses a simple, but unmet need in the field of bioinformatics, specifically in handling the complexities of multiple sequence alignments. By efficiently removing columns exceeding a specified gap threshold, GapClean streamlines the alignment, enabling a clearer focus on conserved protein domains and key motifs. This is particularly beneficial for large-scale sequence analyses, where traditional methods may falter under the sheer volume of data. The transformation of sequence alignments into an array further optimizes computational efficiency, making it feasible to process alignments with up to a million sequences within a manageable timeframe. While further benchmarking is necessary to quantify processing times accurately, preliminary observations suggest a promising linear relationship between the number of sequences and processing duration. As bioinformatics continues to evolve, tools like GapClean will play a pivotal role in interpretation of complex protein data.

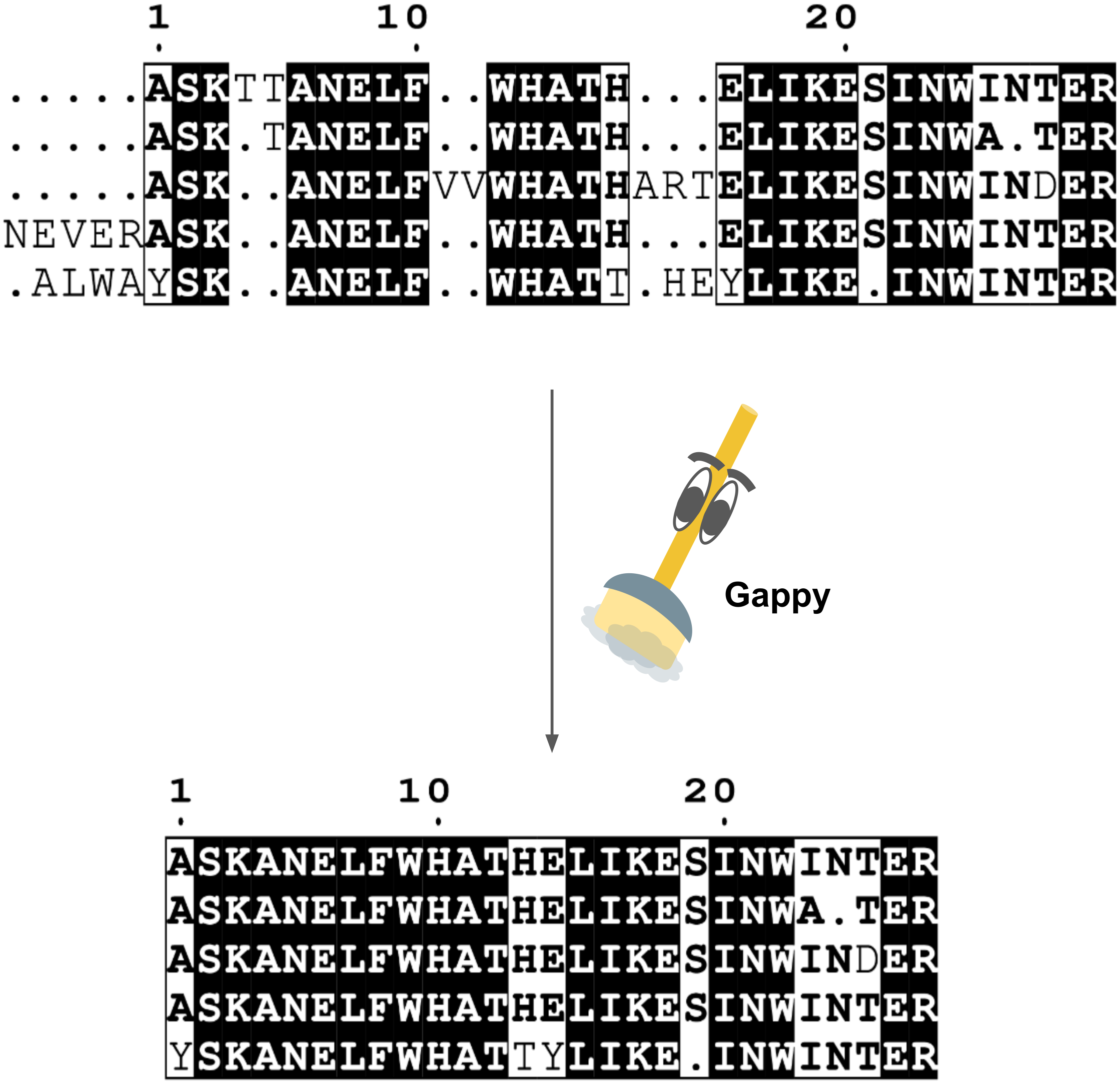

1
10
20
.....ASKTTANELF..WHATH...ELIKESINWINTER
.....ASK.TANELF..WHATH...ELIKESINWA.TER
.....ASK..ANELFVVWHATHARTELIKESINWINDER
NEVERASK..ANELF..WHATH...ELIKESINWINTER
.ALWAYSK..ANELF..WHATT.HEYLIKE.INWINTER
Gappy
1
10
20
ASKANELFWHATHELIKESINWINTER
ASKANELFWHATHELIKESINWA.TER
ASKANELFWHATHELIKESINWINDER
ASKANELFWHATHELIKESINWINTER
YSKANELFWHATTYLIKE.INWINTER

Figure C.2: Example of GapClean-ed alignment, at a 40% threshold, removing extraneous gaps and sequence-specific inserts.

C.2 GapClean code

```bash
#!/bin/bash
# Function to display help
display_help() {
    echo
    echo
    echo "  ====================================================  "
    echo
    echo "                      GapClean (v0.5)                   "
    echo
    echo "                  Written by Aarya Venkat               "
    echo
    echo "  ====================================================  "
    echo
    echo "Description: GapClean takes a gappy multiple sequence alignment"
    echo "and removes columns with gaps at a specified threshold value to"
    echo "produce a \"cleaner\" and easier to visualize sequence alignment."
    echo
    echo "Usage: $0 [options]"
    echo
    echo "   -i   Input file       Required."
    echo "   -o   Output file      Required."
    echo "   -t   Threshold value  Optional. Default is 99."
    echo "   -h   Display this help message."
    echo
    echo
    echo "  Example: gapclean -i input.fa -o output.fa -t 95  "
    echo
    exit 1
}

# Initialize variables
INPUT=""
OUTPUT=""
THRESHOLD=99  # Default value

# Parse command line arguments
while getopts "i:o:t:h" opt; do
  case $opt in
    i) INPUT="$OPTARG"
    ;;
    o) OUTPUT="$OPTARG"
    ;;
    t) THRESHOLD="$OPTARG"
    ;;
    h) display_help
    ;;
    \?) echo "Invalid option -$OPTARG" >&2
    ;;
  esac
done

# Check if required arguments are provided
if [ -z "$INPUT" ] || [ -z "$OUTPUT" ]; then
    echo
    echo "Invalid options. -i (input) and -o (output) are required."
    echo
    display_help
    echo
    exit 1
fi
```

```bash

# Dynamic naming based on the input name
NEWLINE_TMP="${INPUT}.newline.txt"
HEADERS_TMP="${INPUT}.headers.txt"
SEQUENCES_TMP="${INPUT}.sequences.txt"
PROCESSED_SEQUENCES_TMP="${INPUT}.sequences.processed.txt"
echo

perl ${0%gapclean}bucket/remove_newline.pl $INPUT > $NEWLINE_TMP

echo "Initial cleanup of newlines from fasta sequence"

echo

echo "splitting $INPUT into sequences and headers"

grep -v ">" $NEWLINE_TMP > $SEQUENCES_TMP
grep ">" $NEWLINE_TMP > $HEADERS_TMP

echo

# Call the python script with the parsed arguments
python ${0%gapclean}bucket/gapclean.py -i $SEQUENCES_TMP -o
    $PROCESSED_SEQUENCES_TMP -t "$THRESHOLD"

# Combine headers and processed sequences into the output file
paste -d’\n’ $HEADERS_TMP $PROCESSED_SEQUENCES_TMP > $OUTPUT

# Remove temporary files
rm $HEADERS_TMP $SEQUENCES_TMP $PROCESSED_SEQUENCES_TMP $NEWLINE_TMP
```

Listing C.1: bash code for GapClean

Appendix D

alphaFilter: a tool for filtering AlphaFold2 models.

D.1 Background

AlphaFold2 was an incredibly impactful tool in the biochemistry world, releasing millions upon millions of protein structures from the Uniprot database. The ease of using the tool with straightforward interpretability led to its widespread use among computational and experimental scientists alike. While the value of AlphaFold2 is easily recognizable, it has normalized an excessive availability of poor quality and biased models, which many non-experts in molecular modeling may find difficult to parse and deal with. To combat this, I generated alphaFilter, a simple tool to filter out poor quality and fragmentary AlphaFold2 structures.

Fragmentary structures oversaturate the AlphaFold2 database, they are often high accuracy structures but only represent part of the overall proteins. They often only add noise to a given dataset and are typically excised during data curation. Similarly, poor accuracy models, identified by median pLDDT, a per-residue confidence metric AlphaFold2 places in the B-factor column of a structure, can lead to faulty interpretations of a protein structure-function.

To aid scientists needing to operate on multitudes of AlphaFold2 structures, I created a simple pythonic tool to filter structures by median pLDDT (accuracy) and number of residues (to remove fragmentary structures). Users can individually define the thresholds they need for each, but hopefully the availability of this tool can aid scientists in their analyses of AF2 structures. It is a simple python script that counts the total residues and takes the median pLDDT value from the B-factor column and compares these counts against an input threshold. It outputs only pdbs that meet this threshold.

To use alphaFilter, one must simply provide the path to the directory containing the AlphaFold2 models, the desired plddt confidence threshold, and the minimum number of residues a model should have.

Usage: python3 alphafilter.py -d DIRECTORY -t THRESHOLD -r RES_MINIM

Example: python3 alphafilter.py -d /home/aarya/Desktop/alphafold_pdbs/ -t 80 -r 120

The example above takes all alphafold pdbs located in /home/aarya/Desktop/alphafold_pdbs/, identifies only pdbs with a median pLDDT above 80% and contains over 120 residues in the structure, and finally outputs a filtered list of pdbs for you to then perform further analyses on. Scientists can make their alphafold2 datasets more robust to interpretibility through using this tool. The tool can be found at https://github.com/arikat/alphaFilter.

D.2 alphaFilter code

```python
import os
import statistics
import argparse

def extract_plddt(pdb_file):
    plddt_values = []
    amino_acid_count = 0
    last_residue_number = None

    with open(pdb_file, ’r’) as f:
        for line in f:
            if line.startswith("ATOM"):
                # B-factor column is columns 61-66 in PDB format.
                plddt = float(line[60:66].strip())
                plddt_values.append(plddt)

                # Residue number is columns 23-26 in PDB format.
                current_residue_number = int(line[22:26].strip())
                if current_residue_number != last_residue_number:
                    amino_acid_count += 1
                    last_residue_number = current_residue_number

    return plddt_values, amino_acid_count

def print_plddt(directory, threshold=90, min_amino_acids=250):

    # Print PDB files in a directory with a median pLDDT above a given threshold and a minimum amino acid count.

    for file in os.listdir(directory):
        if file.endswith(".pdb"):
            plddt_values, amino_acid_count = extract_plddt(os.path.join(directory, file))
            if statistics.mean(plddt_values) > threshold and amino_acid_count >= min_amino_acids:
                print(file)


def main():
    parser = argparse.ArgumentParser(description=’Process a directory of AlphaFold2 model and output a list based on a mean confidence threshold.’)
    parser.add_argument(’-d’, ’--directory’, help=’Path to file directory’, required=True)
    parser.add_argument(’-t’, ’--threshold’, help=’Threshold for plddt confidence (default is 90)’, type=int, default=90, required=True)
    parser.add_argument(’-r’, ’--res_minim’, help=’Filter minimum number of residues in AlphaFold2 structure’, type=int, default=90, required=True)


    args = parser.parse_args()

    print_plddt(args.directory, args.threshold, args.res_minim)

if __name__ == "__main__":
    main()
```

Listing D.1: Python code for alphaFilter

Appendix E

Prevalence and Homology of the Pneumococcal Serine-Rich Repeat Protein at the Global Scale

Aceil, J.*, Venkat, A.* et al. "Prevalence and Homology of the Pneumococcal Serine-Rich Repeat Protein at the Global Scale" Microbiology Spectrum (2023).
Reprinted here with permission of the publisher.

E.1 Abstract

Pneumococcal pneumonia remains a WHO high-priority disease despite multivalent conjugate vaccines administered in clinical practice worldwide. A protein-based, serotype-independent vaccine has long-promised comprehensive coverage of most clinical isolates of the pneumococcus. Along with numerous pneumococcal surface protein immunogens, the pneumococcal serine-rich repeat protein (PsrP) has been investigated as a potential vaccine target due to its surface exposure and functions toward bacterial virulence and lung infection. Three critical criteria for its vaccine potential - the clinical prevalence, serotype distribution, and sequence homology of PsrP - have yet to be well characterized. Here, we used genomes of 13,454 clinically isolated pneumococci from the Global Pneumococcal Sequencing project to investigate PsrP presence among isolates, distribution among serotypes, and interrogate its homology as a protein across species. These isolates represent all age groups, countries worldwide, and types of pneumococcal infection. We found PsrP present in at least 50% of all isolates across all determined serotypes and nontypeable (NT) clinical isolates. Using a combination of peptide matching and HMM profiles built on full-length and individual PsrP domains, we identified novel variants that expand PsrP diversity and prevalence. We also observed sequence variability in its basic region (BR) between isolates and serotypes. PsrP has a strong vaccine potential due to its breadth of coverage, especially in nonvaccine serotypes (NVTs) when exploiting its regions of conservation in vaccine design. **IMPORTANCE**: An updated outlook on PsrP prevalence and serotype distribution sheds new light on the comprehensiveness of a PsrP-based protein vaccine. The protein is present in all vaccine serotypes and highly present in the next wave of potentially disease-causing serotypes not included in the current multivalent conjugate vaccines. Furthermore, PsrP is strongly correlated with clinical isolates harboring pneumococcal disease as opposed to pneumococcal carriage. PsrP is also highly present in strains and serotypes from Africa, where the need for a protein-based vaccine is the greatest, giving new reasoning to pursue PsrP as a protein vaccine.

Keywords: PsrP, Streptococcus pneumoniae, bioinformatics, epidemiology, protein vaccine

Appendix F

Published PhD manuscripts

1. Bendzunas, G., Byrne, D.P., Shrestha, S., Daly, L.A., Oswald, S.O., Katiyar, S., **Venkat, A.**, Yeung, W., Eyers, C.E., Eyers, P.A. and Kannan, N., 2023. Redox Regulation of Brain Selective Kinases BRSK1/2: Implications for Dynamic Control of the Eukaryotic AMPK family through Cys-based mechanisms. bioRxiv, pp.2023-10.

2. **Venkat, A.***, Watterson, G.*, Byrne, D.P.*, O'Boyle, B., Shrestha, S., Gravel, N., Fairweather, E.E., Daly, L.A., Bunn, C., Yeung, W. and Aggarwal, I., 2023. Mechanistic and evolutionary insights into isoform-specific 'supercharging'in DCLK family kinases. bioRxiv.

3. Aceil, J.*, **Venkat, A.***, Pan, E., Kannan, N. and Avci, F.Y., 2023. Prevalence and Homology of the Pneumococcal Serine-Rich Repeat Protein at the Global Scale. Microbiology Spectrum, pp.e03252-22.

4. Yeung, W., Zhou, Z., Mathew, L., Gravel, N., Taujale, R., O'Boyle, B., Salcedo, M., **Venkat, A.**, Lanzilotta, W., Li, S. and Kannan, N., 2023. Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. Briefings in Bioinformatics, 24(1), p.bbac619.

5. Amos, R.A., Atmodjo, M.A., Huang, C., Gao, Z., **Venkat, A.**, Taujale, R., Kannan, N., Moremen, K.W. and Mohnen, D., 2022. Polymerization of the backbone of the pectic polysaccharide rhamnogalacturonan I. Nature plants, 8(11), pp.1289-1303.

6. **Venkat, A.**, Tehrani, D., Taujale, R., Yeung, W., Gravel, N., Moremen, K.W. and Kannan, N., 2022. Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases. Journal of Biological Chemistry, 298(8).

7. Yeung, W., Kwon, A., Taujale, R., Bunn, C., **Venkat, A.** and Kannan, N., 2021. Evolution of functional diversity in the holozoan tyrosine kinome. Molecular Biology and Evolution, 38(12), pp.5625-5639.

8. Huang, L.C., Taujale, R., Gravel, N., **Venkat, A.**, Yeung, W., Byrne, D.P., Eyers, P.A. and Kannan, N., 2021. KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases. BMC bioinformatics, 22, pp.1-25.

9. Taujale, R., Soleymani, S., Priyadarshi, A., **Venkat, A.**, Yeung, W., Kochut, K.J. and Kannan, N., 2021. GTXplorer: A portal to navigate and visualize the evolutionary information encoded in fold A glycosyltransferases. Glycobiology, 31(11), pp.1472-1477.

10. Gosztyla, M.L., Kwong, L., Murray, N.A., Williams, C.E., Behnke, N., Curry, P., Corbett, K.D., DSouza, K.N., de Pablo, J.G., Gicobi, J., Javidnia, M., Lotay, N., Prescott, S.M., Quinn, J.P., Rivera, S.V.G., Smith, M.A., Yang, K.T.Y., **Venkat, A.**, Yamoah, M.A., 2021. Responses to 10 common criticisms of anti-racism action in STEMM. PLoS Computational Biology, 17(7), p.e1009141.

11. Zhang, A., **Venkat, A.**, Taujale, R., Mull, J.L., Ito, A., Kannan, N. and Haltiwanger, R.S., 2021. Peters plus syndrome mutations affect the function and stability of human 1, 3-glucosyltransferase. Journal of Biological Chemistry, 297(1).

12. Huang, L.C., Yeung, W., Wang, Y., Cheng, H., **Venkat, A.**, Li, S., Ma, P., Rasheed, K. and Kannan, N., 2020. Quantitative Structure–Mutation–Activity Relationship Tests (QSMART) model for protein kinase inhibitor response prediction. BMC bioinformatics, 21, pp.1-22.

13. Taujale, R., **Venkat, A.**, Huang, L.C., Zhou, Z., Yeung, W., Rasheed, K.M., Li, S., Edison, A.S., Moremen, K.W. and Kannan, N., 2020. Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases. Elife, 9, p.e54532.

Appendix Bibliography

Amos, Robert A et al. (2022). "Polymerization of the backbone of the pectic polysaccharide rhamnogalacturonan I". *Nature plants* 8.11, pp. 1289–1303.

Becker, Felix and Mario Stanke (2022). "learnMSA: learning and aligning large protein families". *GigaScience* 11, giac104.

Briggs, David C and Erhard Hohenester (2018). "Structural basis for the initiation of glycosaminoglycan biosynthesis by human xylosyltransferase 1". *Structure* 26.6, pp. 801–809.

Crooks, Gavin E et al. (2004). "WebLogo: a sequence logo generator". *Genome research* 14.6, pp. 1188–1190.

Eddy, Sean R (1996). "Hidden markov models". *Current opinion in structural biology* 6.3, pp. 361–365.

Hanes, Melinda S, Kelley W Moremen, and Richard D Cummings (2017). "Biochemical characterization of functional domains of the chaperone Cosmc". *PLoS One* 12.6, e0180242.

Ju, Tongzhong and Richard D Cummings (2002). "A unique molecular chaperone Cosmc required for activity of the mammalian core 1 β3-galactosyltransferase". *Proceedings of the national academy of sciences* 99.26, pp. 16613–16618.

Jumper, John et al. (Aug. 2021). "Highly accurate protein structure prediction with AlphaFold". eng. *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

Kwon, Annie et al. (2019). "Tracing the origin and evolution of pseudokinases across the tree of life". *Science Signaling* 12.578, eaav3810.

Murphy, James M, Hesso Farhan, and Patrick A Eyers (2017). "Bio-Zombie: the rise of pseudoenzymes in biology". *Biochemical Society Transactions* 45.2, pp. 537–544.

Neuwald, Andrew F (2014). "A Bayesian sampler for optimization of protein domain hierarchies". *Journal of Computational Biology* 21.3, pp. 269–286.

Ribeiro, António JM et al. (2019). "Emerging concepts in pseudoenzyme classification, evolution, and signaling". *Science Signaling* 12.594, eaat9797.

Shrestha, Safal et al. (2020). "Cataloguing the dead: breathing new life into pseudokinase research". *The FEBS Journal* 287.19, pp. 4150–4169.

Taujale, Rahil, Saber Soleymani, et al. (2021). "GTXplorer: A portal to navigate and visualize the evolutionary information encoded in fold A glycosyltransferases". *Glycobiology* 31.11, pp. 1472–1477.

Taujale, Rahil, Aarya Venkat, et al. (2020). "Deep evolutionary analysis reveals the design principles of fold A glycosyltransferases". *Elife* 9, e54532.

Taujale, Rahil, Zhongliang Zhou, et al. (2021). "Mapping the glycosyltransferase fold landscape using interpretable deep learning". *Nature Communications* 12.1, p. 5656.

Varki, Ajit, Richard D Cummings, et al. (2022). "Essentials of Glycobiology [internet]".

Venkat, Aarya, Daniel Tehrani, et al. (2022). "Modularity of the hydrophobic core and evolution of functional diversity in fold A glycosyltransferases". *Journal of Biological Chemistry* 298.8.

Wang, Yingchun et al. (2010). "Cosmc is an essential chaperone for correct protein O-glycosylation". *Proceedings of the National Academy of Sciences* 107.20, pp. 9228–9233.