THE PROTEIN SUGAR INTERACTOME

by

Samuel William Canner

A dissertation submitted to Johns Hopkins University in conformity with the requirements for

the degree of Doctor of Philosophy

Baltimore, Maryland

August 2025

[This page is intentionally left blank]

Abstract

Carbohydrates are essential biomolecules involved in myriad cellular processes, regulating protein folding, providing cellular structure, and mediating cell-cell communication. Despite their widespread importance across the cellular landscape, carbohydrates remain one of the least characterized biomolecules due to their chemical diversity, structural flexibility, and lack of a templated biosynthetic pathway. These intrinsic complexities result in non-covalent protein–carbohydrate that are inherently weak and transient, posing significant challenges to crystalizing and resolving experimental structures. Accordingly, computational approaches have advantages to predict and evaluate how novel proteins interact with carbohydrate ligands. However, prior to my dissertation research, no computational or experimental tools were able to systematically identify the protein-sugar interactome.

In this dissertation, I present several advancements in computational glycobiology for predicting the protein-sugar interactome. Firstly, working alongside Dr. Sudhanshu Shanker, I developed a deep learning method CArbohydrate Protein Site IdentiFier (CAPSIF). CAPSIF was created with two variants: (1) a 3D-UNet voxel-based neural network model (CAPSIF:V) and (2) an equivariant graph neural network model (CAPSIF:G). While both models outperform previous surrogate methods used for carbohydrate binding site prediction, CAPSIF:V performs better than CAPSIF:G, achieving test Dice scores of 0.597 and 0.543 and test set Matthews correlation coefficients (MCCs) of 0.599 and 0.538, respectively. We further tested CAPSIF:V on AlphaFold2-predicted protein structures. CAPSIF:V performed equivalently on both experimentally determined structures and AlphaFold2 predicted structures. Finally, we demonstrated how CAPSIF models can be used in conjunction with local glycan-docking protocols, such as GlycanDock, to predict bound protein-carbohydrate structures.

Expanding on this work, I addressed the grand challenge of identifying the human and *E. coli* protein-sugar interactomes. Given the impracticality of experimental screening of the entire proteome against extensive libraries of glycans, computational screening of proteins for carbohydrate-binding provides an attractive and ultimately testable alternative. Current estimates label 1.5 to 5% of proteins as carbohydrate-binding proteins; however, 50-70% of proteins are known to be glycosylated, suggesting a potential wealth of proteins that bind to carbohydrates. I therefore developed a neural network architecture, named **P**rotein **i**nteraction of **Ca**rbohydrates **P**redictor (PiCAP), to predict whether a protein non-covalently binds to a carbohydrate. I trained PiCAP on a novel dataset of known carbohydrate binders and selected proteins that I identified as likely *not* to bind carbohydrates, including transcription factors, cytoskeletal components, and small-molecule-binding proteins. PiCAP achieves a 90% balanced accuracy on protein-level predictions of carbohydrate binding/non-binding. Using the same dataset, I developed a model named **Ca**rbohydrate **P**rotein **S**ite **I**denti**f**ier 2 (CAPSIF2) to predict protein residues that interact non-covalently with carbohydrates. CAPSIF2 achieves a Dice coefficient of 0.57 on residue-level predictions on our independent test dataset, outcompeting all previous models for this task. To demonstrate the biological applicability of PiCAP and CAPSIF2, I investigated cell surface proteins of human neural cells and further predicted the likelihood of three proteomes, notably *E. coli, M. musculus,* and *H. sapiens*, to bind to carbohydrates. PiCAP predicts that approximately 35-40% of proteins in these proteomes bind carbohydrates. In the human proteome, PiCAP predicts that 75% of extracellular and cell surface proteins are putative carbohydrate binders. The PiCAP predicted binders are highly enriched for functions and processes such as growth factor receptor binding, inflammatory responses, and cell-cell adhesion.

Throughout my dissertation, I have developed a set of models to predict the protein-sugar interactome, with the critical next step being the structural docking of non-covalent protein-carbohydrate complexes on a proteome-wide scale. Current all-atom structure prediction models like AlphaFold3 (AF3), Boltz-1, Chai-1, DiffDock, and RosettaFold-All Atom (RFAA) were validated on protein-small molecule complexes; however, no benchmark or evaluation exists specifically for noncovalent protein-carbohydrate docking. To address this, I developed a high-quality dataset of experimental structures – Benchmark of CArbohydrate Protein Interactions (BCAPIN). Using BCAPIN and a novel evaluation metric, DockQC, I assessed the performance of all-atom structure prediction models on non-covalent protein-carbohydrate docking. I found all methods achieved comparable results, with an 85% success rate for structures of at least acceptable quality. However, I found that the predictive power of all models declined with increasing carbohydrate polymer length. With the capabilities and limitations assessed, I evaluated AF3's ability to predict binding for a set of putative human carbohydrate binding and carbohydrate non-binding proteins. While current models show promise, further development is needed to enable high-confidence, high-throughput prediction of the complete protein-sugar interactome.

In summary, my work advances the field of glycobiology by enabling comprehensive characterization of the protein-sugar interactome on ‘omic scales.

Thesis Committee

Jeffrey J. Gray (Primary Advisor, Reader)
Professor
Department of Chemical and Biomolecular Engineering
Johns Hopkins Whiting School of Engineering

Albert Lau (Chair, Reader)
Associate Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins School of Medicine

Stephen Fried (Member)
Associate Professor
Department of Chemistry
Johns Hopkins Krieger School of Arts and Sciences

Doug Barrick (Member)
Professor
Department of Biophysics
Johns Hopkins Krieger School of Arts and Sciences

Margaret Johnson (Member)
Associate Professor
Department of Biophysics
Johns Hopkins Krieger School of Arts and Sciences

Ronald L. Schnaar (Observational Member)
Professor
Department of Pharmacology and Molecular Sciences
Johns Hopkins School of Medicine

Natasha Zachara (Member)
Professor
Department of Pharmacology and Molecular Sciences
Johns Hopkins School of Medicine

Dedicated to:

Rutuj Gavankar,

Stephen R. Wassall,

and all my pacers on the trails, formal and informal, along the way.

Acknowledgements

A PhD is not simply a journey of a student sitting alone in a basement surrounded by state-of-the-art equipment built in the 1980s; rather, it is a journey involving a person in a lab led by a principal investigator, in a department, monitored by a university, residing inside a city, within an uncertain nation and world – with every layer influencing the journey. Here is how my journey was shaped by all these factors.

I joined the lab of **Dr. Jeffrey J. Gray** in the summer of 2021. His mentorship has truly allowed me to understand the importance of clarity and simplicity in scientific communication. His positivity was always welcome and appreciated during our one-on-one meetings.

I have had the honor of working alongside several scientists in the Gray lab during my time, of which I will forever have a strong appreciation of **Dr. Sai Pooja Mahajan**, **Dr. Ameya Harmalkar**, **Michael Chungyoun**, and **Fatima Hitawala**. Pooja was always incredibly supportive – inside and outside of the lab. I will always remember my weekly runs at R House and in Druid Hill with Ameya. Mikey was always a fantastic and studious individual whom I always appreciated talking with daily. Fatima and I worked together on BioComp 2024, teaching high schoolers over the summer. Other incredibly influential scientists from the Graylab that inspired me include **Dr. Sudhanshu Shanker**, whose path I am directly following, **Dr. Rituparna Samanta**, my fellow membrane expert, and **Dr. Morgan Nance,** who helped me navigate my department better.

Prior to joining the Gray lab, I joined the Johns Hopkins University Department of Biophysics in fall of 2020, the year that didn't exist. Although it was a difficult start to a PhD, I was grateful to spend that year with my cohort-mate **Edgar Manriquez-Sandavol**, especially for

Thanksgiving 2020. Throughout my time at the Biophysics department, I've gotten to know many scientists, such as **Amy Fernandez**, my fellow ultra-running colleague, **Andrea Ori**, with whom I watched The Bachelor during 2020, as well as **Dr. Adip Jhaveri**, **Soumya Behera, Sushil Pangeni, Mankun Sang**, and **Meera Joshi**. However, the most important student in the Biophysics PhD program to me was **Dr. Daniel Evans** (Danny), who was always a friend throughout my PhD: always incredibly supportive and whom I cannot thank enough.

The department, however, is not just students, but also a massive organization of post-doctoral researchers, faculty, and staff. Weekly, I've bothered the post-docs in the Johnson lab – **Dr. Yiben Fu**, my former mentor during my rotation in the Johnson lab, and **Dr. Samuel Foley**, whom I had the pleasure of originally meeting in 2020 before the world ended and with whom I continued to bond with over our mutual love of membranes and politics. I greatly appreciate the work and advice of many faculty in my department, such as the notably the director of the program, **Dr. Karen Fleming**, who was always expedient, kind, and understanding in every interaction I've had, **Dr. Stephen Fried**, who offered a skeptical but well-attuned eye to my research, **Dr. Doug Barrick**, a fellow runner and exercise enjoyer, **Dr. Albert Lau**, my committee chair who has been incredibly responsive and diligent in all matters, **Dr. Margaret Johnson**, who is incredibly knowledgeable and welcoming, **Dr. Natasha Zachara** who graciously joined my committee, and **Dr. Ronald L. Schnaar**, who has been a pleasure to collaborate with and share a ride with during the Society of Glycobiology.

Finally, the department is held together by the amazing staff who connect it all together: **Liz Wilson,** one of the kindest humans I've ever met, **Jessica Appel,** an incredibly knowledgeable person for all financial information, **Nancy Foltz**; and the academic program administrator, both **Nicole Goode**, a fellow runner, and **Alexandre Labat**, who have both been wonderful.

A true strength of Johns Hopkins University, however, is the diversity of research going on across multiple departments and dual-commitments of many of their employees – so everyone can communicate across disciplines. Although my primary appointment was with biophysics, I had a secondary appointment working in the Department of Chemical and Biomolecular Engineering under Dr. Gray. Through these appointments, I have had the pleasure of spending time with **Brett O'Brien**, who I always appreciated running into on campus, **Anastasia Georgiou**, an expert in the craft of ice cream making, **Yue (Moon) Ying**, who I'd consistently bother in the Johnson Lab, and many more. In addition, I appreciate the outside advice of **Dr. Robert Lessick** and **Dr. Peter Armitage** on how Hopkins generally works and how to navigate different situations.

In addition to Johns Hopkins, I've had the pleasure of presenting and meeting many scientists across a multitude of universities at various conferences, meetings, and Zoom calls. At these conferences, I appreciate the comradery and discussions with fellow researchers, notably **James Jeffries, Dr. Benjamin Kellman,** and **Dr. Jacob Hoffman**. In a similar vein, I appreciate my continued friendship with a fellow Johns Hopkins biophysics program interviewee who instead matriculated into the UPenn biophysics program, **Nick Palmer**.

The academic environment was conducive to my scientific developments; however, the outside community is what allowed me to truly thrive – life is just as important to a PhD as work. I moved to Baltimore in August of 2020 – an uncertain and isolated time, which made making any sort of friendship nearly impossible. So, in order to be more outside, my cousin **Sarah Franiak-Jaeger** convinced me to download Nike Run Club, allowing me to finally blossom into a true bona fide runner. Because he knew I had become obsessed with running, Danny invited me to a run club – **Will Walker's** first child: **A Tribe Called Run**. Through Tribe, I was able to join an amazing community and attended Bachelor(ette) watch parties with **Dr. Nicole Arulanantham**, **Alex**

Collado, Sam Collado Esq., Rebecca Ogus, Dr. Zach Nasipak, Dr. Ifunanya Nwogbaga, and **Angelo Tigol**. Nicole was one of my strongest confidants throughout my time in Baltimore, and I will always appreciate our conversations. Sam always provides a non-runner (e.g. sane) perspective on all things, Rebecca is a fantastic priest with a calm and collected demeanor, and Zach has always been supportive in all efforts. It has been amazing to meet all these individuals and see them change over time into who they are today from who we all were coming out of the pandemic – especially Angelo, who is fully embracing who he truly is.

Although Sarah's insistence on the Nike Run Club app may have been the initial motivating factor to run, I have always been inspired to run by my uncle, **Joe Franiak**. Joe finished in second place two years in a row at the world's hardest ultramarathon: Badwater 135. In summer 2021 he inspired me to transition into an ultramarathoner and sign up for my first 50 miler: The Stone Mill. At this event, I am grateful for **Mac McComas** and **Doug Jones** for sticking alongside me for the first 35 miles, until I could no longer run and had to hobble to the finish.

A natural part of a runner's training is speedwork, which I had only done alone at that point; however, I must thank **Jessica Brennan** for convincing me to join her at Track Tuesdays with **Dr. Thomas Athey**. Although my stomach and sleep schedule may have hated it, my soul always appreciated our early 6AM morning sessions and conversations. Those mornings were my first introduction to the **Faster Bastards**, led by **Thomas Neuberger**, which, much like Tribe, is a fantastic and supportive community that I am grateful to be a part of.

Both Tribe and Bastards allowed me to learn more about the city and its people. Through Tribe I met the sociable and incredibly talented **Melody Thomas**, who I am so grateful for repeatedly inviting me to photograph events run by the **Waterfront Partnership** – and for helping me learn so much more about Baltimore through this.

Although I lived in Mount Vernon, I began to learn much more about Remington through Mac and our mutual close friends: the **Bar Raccoons**: Zach, **Elliot Madre, Forrest Wrenn, Nora Frankel, Kate Allen, Chris Cobbs,** and **Ryan Detter.** Elliot has an infectious positivity that seemingly can never be destroyed. Forrest is a fantastic human who took me in on Thanksgiving and has become an amazing father to Cora. Nora always kindly hosted our Hellraiser watch parties at the insistence of Kate. Chris would always be there as support if I needed it with a thirst to try new craft beers in search of the best. Ryan, a fellow Midwesterner, was a fantastic running buddy and great person to share a martini with.

One of the most pivotal moments in my life was running my first hundred miler (Rocky Raccoon) alongside Mac, Doug, and **Dan Frank** in February 2023 in Texas. I am so happy I was lapped by Dan at mile 60 (he was at 80) and got to hear him complain – it was the perfect encouragement. After my panic attack at mile 69 and intense knee inflammation, I was only walking and ready to drop while texts from **Best Aunt** (Loretta Franiak II) and **Lindsay Kohan** were incredibly encouraging. Although I appreciated those supportive texts– I will always remember sitting in the van, holding back tears as Mac and Dan gave me the courage to finish. After everything, I cannot overstate how impactful their support was, and how much their words meant.

Naturally, the only next step after a hundred miler is to do an Ironman while exclusively eating Jimmy Dean sandwiches without any knowledge of how to swim. I cannot thank **Fausto Bonilla** enough, despite not even knowing me, he came out and did the swim and the bike with me – bonding us into close friends. I cannot wait for our future adventures together. And through Fausto, I joined the amazing group **Too Hot for Classic** (later **Cute Eyes**), a close-knit family with

Fausto, **Adrian Alday**, **Meg Rorison, Rigo Sanchez, Paul Turner,** and **Jon Ober.** Meg is a kind videographer who, despite the hardship she's gone through, still picks herself up. Rigo, a man I met at a fateful Tribe speedwork, has been amazing to be around, to spend New Years with, and to watch as a fantastic father. Paul, although now MIA, always brought a positive energy unmatched elsewhere. Jon Ober, someone I ran alongside at one of my first Faster Bastards' classics, is among the kindest and most welcoming people I've met. I cannot wait for the next cosmo/pumpkintini night.

In addition, I am ecstatic to be a part of the running **Monthly Mayhem** group of Mac, Forrest, Cobbs, Angelo, **Dr. Macaroni Peacock** (Martin Michalcek)**, Zulu Toboggan** (Zachary Thomas)**, Dr. Matthew Newmeyer, Andy Smith,** and **Dr. Sai Bharath**. Macaroni was always a fantastic person for philosophical and esoteric conversations. I cannot thank Zulu enough for taking me in on Thanksgiving in 2021 and inspiring me during my first Tribe run. Matt has always brought a calm presence to any conversation. Andy is an amazing graphic designer and father. Sai has always been a very conscious individual who knows how to give everything and go all in on whatever he's doing.

The community of runners I've met through Tribe, Bastards, and **Believe in the Run** has been astounding and wonderful. **Dave Carpenter** was always there to help me as an astute brewer, egg nog distiller, and kind shoulder to lean on. My cousin **Gavin Tabb** is one of the best mezcal experts in the DMV area. **Eric Erdman**, who took me in as an orphan during Christmas. **Eric Schulman**, a fantastic father and husband and steadfast pillar in the community. **Karl Mulligan**, a fellow Hopkins PhD student who made me a Canton Kayak Club aficionado. **Kelly Kitzmiller**, the most shout-at-able runner in all of Baltimore. **Heather Chou**, the most infectiously positive person who is also my sister in sciatica. **Josh Sanchez**, who keeps his positive and youthful energy with him and imparts it on everyone else. **William Pinkney**, who although a man of illusion, has been a pleasure to get to know. **Ryan Haines** always stands for what is right. **Morgan Jones**, who was the best volleyball player in my Volo league. **Dr. Samuel Curtis**, the best dog dad of all time. **Dr. Melinda Martinez**, an expert in conservation. **Paul Gochar**, a fantastic advisor in all matters. **Laura Van Oudenaren**, a fantastic confidant. **Kira Wisniewski** contains a positive presence in every interaction she is a part of. **Trent Lackey**, the master of blueberries and assisting people. **Chris Hauger**, a fellow Midwesterner endlessly provides support to all athletes of all backgrounds. **Yehudah Silver**, an incredible ice skater and devoted med student. **Dr. Chuck Dave Brezz** (real doctor), an incredible person to talk to and knowledgeable in all domains of life. **Jessica Honeycutt,** runs 70+ miles all while smiling the entire time. **Dr. Marcia Croft**, an even toned kind human, with an immeasurable ability to play the bassoon.

In running, I've had the chance to run alongside many. Very few experiences however rival that of my third time running The Stone Mill alongside **Ken Ivanetch** and **Johnny Lyons**. Whenever I get into a bad mindset, I love to think about how you both pulled me by my collar from mile 5 to 40, until I was able to return the favor on that miserably fateful day. Further, I am grateful to those I conscripted into service during Umstead 100 (my second hundred miler): Elliot for pacing the last 37 miles as I talked about low back pain, alongside **Alex Canner** for crewing the entire time (and driving home in our rental car).

The months following Umstead, however, were miserable. I herniated a disc in training, somewhere, somehow. "Miserable" however, is too light a word, as I could not run and if I tried, I'd be bedridden for days. I also got hit by a car while biking to the community pool, among other issues. I am especially grateful to the following people for assisting me throughout my recovery: **Dr. Chris Heydrick** (DPT), **Jeremy Ardanuy**, **Katherine Cunningham**, **David McFeeters**, **Dr.**

Steve Mitchell, and **Matt Fedderly.** Heydrick was a welcome sight every week at PT. Jeremy was my coach during Umstead, but became my bosom buddy in injury, I truly value our many walks together. Katherine was always there when I needed her. Steve has always been a fantastic pillar to lean on. And finally, Fedderly was always willing to help, and showed me the number one thing that helped in my recovery: the work of **Dr. Stuart McGill** (who I must acknowledge and thank for his many years of work in the field). For the bike issue, I would like to thank **Jed Weeks**, the foremost bike expert in Baltimore, and **Juan Carlos Puga**, for his legal expertise.

The end-goal of my recovery was to run another hundred miler, so I would like to thank Heydrick for his consistent help, and Johnny for joining me to run the entire length of the NCR as training. Further, I would like to thank Johnny for pacing me for the first 65 miles of the C&O 100, and Katherine for the miserable final 20 miles of walking.

Several more running-related people deserve thanks: **Eddie O'Keefe, Anne Rosenthal, Juan Francisco Lucas** and **Amanda Phillips de Lucas.** Eddie, much like the other brewers Elliot and Dave, is infectiously positive and constantly giving back to the community. Anne is a positive force in the Baltimore area, and I am grateful for her helping connect me with it. Juan and Amanda were the reason I was able to move apartments, and it has been a pleasure receiving their REI catalog.

In addition, I would like to thank the climbing community I've joined: **Sarah Little, Vince Filardi,** and **Jay Eastman**. Sarah is a great friend to grab ice cream with. Vince is an even-keel, kind man with a helping hand. Jay is not just a great man to climb alongside but also run alongside.

But there is a world outside of Baltimore that shaped everything around my five-year journey. First, my wonderful undergraduate PI, **Dr. Stephen Wassall**, for whom I performed

Molecular Dynamics simulations of lipid membranes for, was how I first learned of and pursued biophysics. Steve was (and still is) a fantastic investigator, one I deeply admire. He fostered the beginning of my academic journey, and, though I am still working at the time of my dissertation defense to publish my undergraduate research, he remains one of my closest confidants.

Moving to Baltimore was scary; I knew nobody – and it was the year that didn't exist. Well, not nobody. I knew one person – my cousin **Dr. Chris Kapp** (but like a real doctor – MD, not this PhD nonsense), who I leaned on throughout my time and am still waiting on to finish watching Season 2 of The Boys with. Although I may have been mostly alone in 2020, I did have some friends on the East Coast in NYC that I could spend time with ranging from a friend from high school who's let me crash on his couch and I always visit, **Jack Zhang,** to those I met at the Met – **Terrence Schroeder, Francisco Lupini, Karina Schroeder, Michael Koh,** and **Josephina Tell** – who helped me when I needed it most after the passing of **Rutuj Gavankar**.

And surrounding those individuals are my family who supported me. My mother **Lisa Franiak-Canner**, father **Mark Canner,** and brothers **Dr. Mark Edward Canner** (real doctor), and Alex, as well as **Busia** (Loretta Franiak I) and Best Aunt, who were just as critical in raising me as my parents. I will never forget the days playing Pooh sticks and getting Dairy queen with Busia in my youth, and how Best Aunt served as an unwavering pillar for my family during my teenage years and recently for Busia. Other family members who have supported me along the way (not yet mentioned) include **Barb Kapp** and **Greg Kapp,** who came out to visit me in Baltimore (unrelated to Chris' graduation), **Dr. Randall Franiak** (real doctor), and **Beanack** (Edwin Franiak). In addition, I'd like to thank my friends from undergrad who have supported me along this journey as well: **Damanveer Singh, Madelyn Nolting**. I would also like to express my sincerest appreciation for **Jerry** (from UConn) and **Dr. Guillermo Rambo,** who, although a little unconventional, are the strongest people I've met – physically and emotionally – able to withstand any storm.

To those willing to (unwittingly) spend my birthday with me, offer me a space at their table at Thanksgiving, have a late night at 29th street tavern with cosmopolitans and/or pumpkintinis, or just go for a weekly stroll around Hopkins - thank you.

Contents

List of Figures

Chapter 1

Introduction

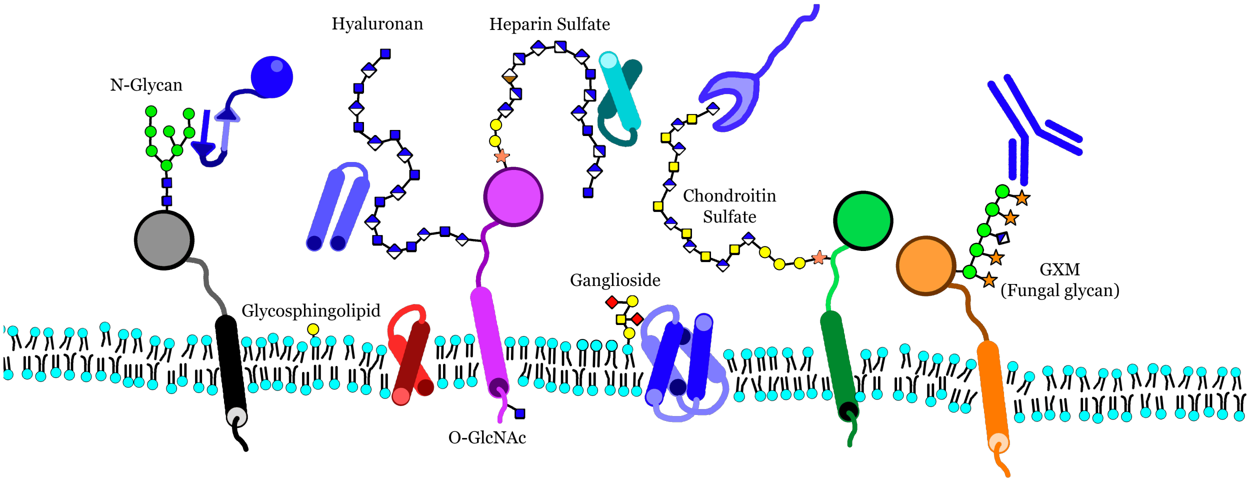


Figure 1.1: Cartoon of protein-carbohydrate interactions and glycoproteins in the cell. (not to scale)

Carbohydrates are ubiquitous across life

The unique structure and diversity of carbohydrates

All life is composed of an immense number of biomolecules, inorganic compounds, and elements; however, biology is often framed through the lens of the The Central Dogma. The Central Dogma states that deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA) which is then translated into proteins, which carry out cellular functions.[1] While this simplification of biology is incredibly useful; this model naturally overlooks the interplay among all biopolymers inside the cell. In particular, it fails to capture the importance of other key biopolymers in the cell, notably lipids and carbohydrates. Lipids are the essential components of cell membranes, defining what is or is not a part of an organelle, cell, or organism.[2–4] Carbohydrates however serve unique purposes of energy metabolism and in the functional modulation of all other biopolymers.[5]

Carbohydrates, also known as sugars, are hydrated carbon-based polymers with the basic chemical formula $C_i(H_2O)_j$ where *i* and *j* are positive integers. The foundational building blocks of carbohydrates are monosaccharides. Carbohydrates can exist in either a linear or cyclic (ring) form, with the cyclic form being the most common in biological environments. These cyclic forms can be five-membered rings (furanose) or a six-membered ring (pyranose). Additionally, these rings can adopt different conformations: where furanoses typically exist in envelope or twist conformations, and pyranoses in a $^4C_1$ conformation or, less often, a $^1C_4$ conformation. The conformation of these carbohydrates is further specified by (1) their stereoisomer, L or D - with D being the primary eukaryote conformation, and (2) the anomeric carbon existing in an α or β conformation.[6]

Monosaccharides are distinguished based on the (1) epimerization of the hydroxyls and (2) functional group modification of hydroxyls. Epimerization refers to the changes in relative orientation of hydroxyl groups to the carbohydrate rings, which can be either equatorial or axial. Common chemical modifications include acetylation, methylation, and deoxygenation which impart unique properties to the saccharide.[6]

Figure 1.2 shows the most common mammalian carbohydrate monosaccharides. Glucose (Glc), galactose (Gal), and mannose (Man) are all epimers of one another; they have the same stoichiometry, but differ in hydroxyl orientations. Glc has all hydroxyls equatorial, where Gal has C_4-OH is axial and Man has C_2-OH is axial (Figure 1.2A). One of the most common modifications is the addition of an amine group, as seen in GlcNAc and GalNAc, where the C_2 position is modified to contain an N-acetyl group (NHAc) (Figure 1.2B). One of the most studied monosaccharides is sialic acid (Sia), also known as neuraminic acid (Neu), with special interest in the Neu5Ac variant. Neu5Ac, the only version produced by humans, boasts nine carbons, a negative charge, an acetyl group, a three-carbon chain decorated in hydroxyls, and a carboxyl group.[6]

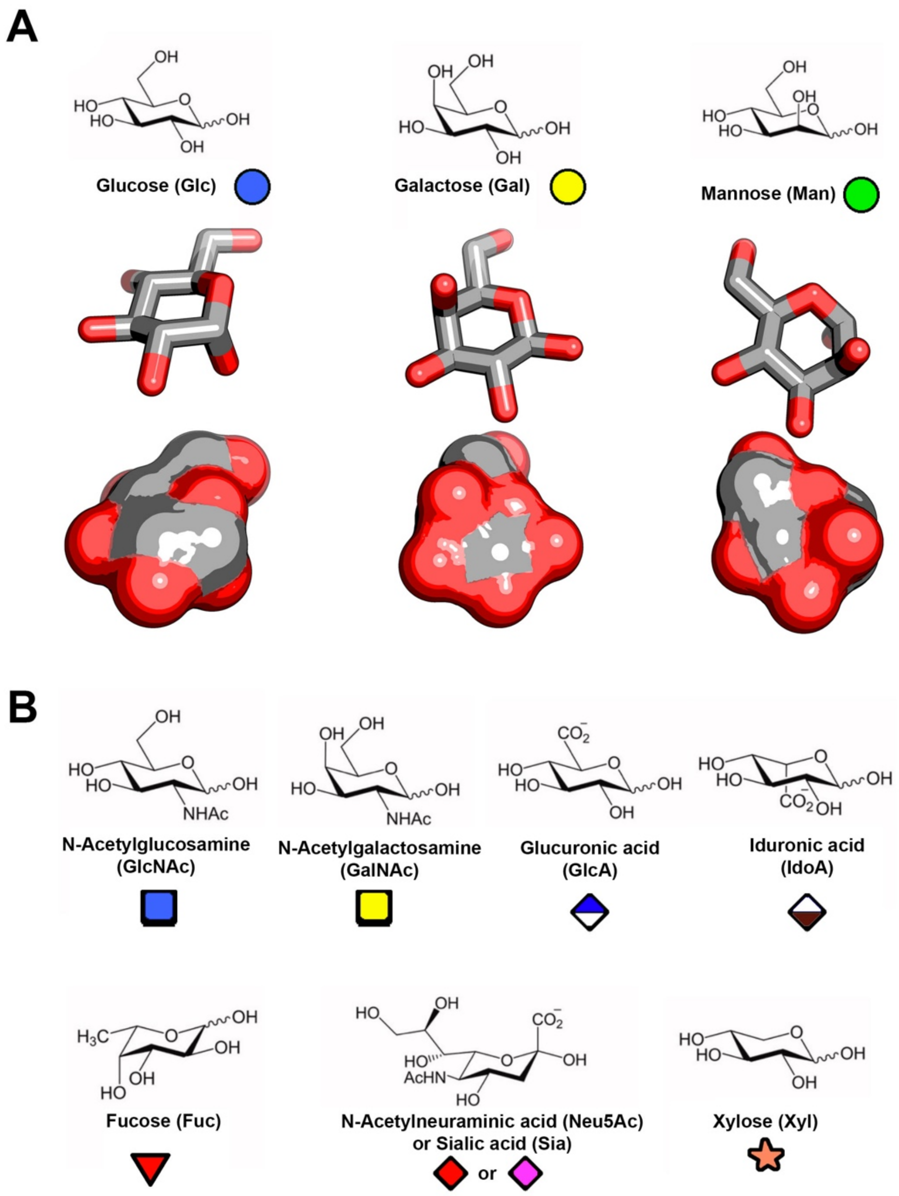


Figure 1.2: Chemical diagram and cartoon representations of common mammalian monosaccharides. (A) Lewis structure, 3D sticks, and 3D surface representation of common D-pyranoses glucose, galactose, and mannose. (B) Lewis structures and cartoons of other common pyranoses.[6]

The diversity of monosaccharides arises from their chemical orientations, modifications, and conformations, but even greater diversity is achieved when they are linked together to oligosaccharides (less than 12 monosaccharides) or polysaccharides (greater than 12 monosaccharides). When a monosaccharide is covalently linked to another molecule, the resulting saccharide is called a glycan. Although the number of commonly observed unique, unmodified monosaccharides (10) is significantly less than unmodified amino acids (20), carbohydrate structures are distinguished by their myriad possible linkages. In theory, any monosaccharide can be covalently bonded to any other via condensation reactions between hydroxyl groups. For example, it is theoretically possible to connect four Glc monosaccharides together in 1,792 distinct structures. In practice, a very small subset of these structures is observed due to organisms lacking the enzymes required to create those structures. As a result, carbohydrate chains are more often categorized into broad categories of N-linked, O-linked structures, and glycolipids.[6]

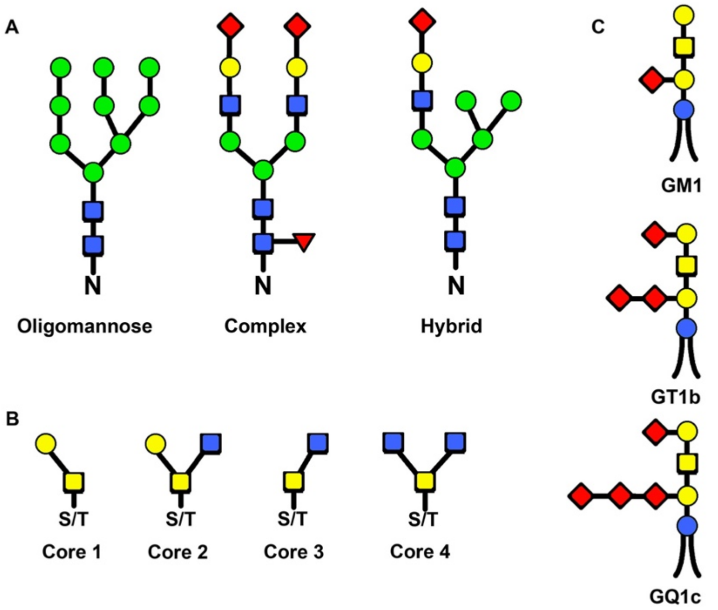


Figure 1.3: Common mammalian glycosylation patterns. (A) N-linked glycosylation patterns. (B) O-linked GalNAc glycosylation cores. (C) GM1 ganglioside.

N-linked glycans, or N-glycans, are glycans that are covalently linked by an N-glycosidic bond to an asparagine (Asn) residue of a protein or peptide. The consensus N-glycosylation sequon is always an NX(S/T) motif, where X is any amino acid that is not proline, and S and T are serine and threonine, respectively.[7] In eukaryotes, the first monosaccharide to be attached to the Asn residue is always a GlcNAc. N-glycans are synthesized in the endoplasmic reticulum (ER) on dolichol phosphate (Dol-P), after which the glycan is transferred "en bloc" to an acceptor protein by an oligosaccharyltransferase (OST).[6] The glycosylated protein later transported through the Golgi apparatus, where enzymes further modify the initial glycan bloc, growing and shrinking the glycan tree into unique substructures.[8,9] These structures are grouped into three common categories (1) oligomannose, (2) complex, and (3) hybrid. Figure 1.3A shows these structures: oligomannose structures terminate in Man residues, complex structures terminate in Sia residues, and hybrid structures terminate have at least one branch ending in Sia and at least one in Man.[6]

O-linked glycans are covalently linked to Ser or Thr residues. Figure 1.3B shows the four well-described O-linked O-GalNAc core structures, which may be extended to linear or branched chains, similar to that of N-linked glycans, terminating in the ABO and Lewis blood group epitopes. Many mammalian O-linked glycans are initiated by the transfer of O-GalNac to Ser/Thr by GALNT, after which the chain is extended one monosaccharide at a time. One well established purpose of O-linked glycans is occupying the extracellular environment, with glycosaminoglycans (GAGs) such as heparan sulfate, hyaluronan, dermatan sulfate, and chondroitin sulfate. Heparan sulfate (HS) is a heterogenous linear polymer with a high degree of polymerization (DP). HS contains repeating -4GlcA1β-4GlcNAcα1- units with domains that are either highly sulfated or unmodified.[6] Other O-linked glycans in mammals include O-mannosylation, which can account for 33% of tissue O-glycosylation, O-fucosylation, and O-glucosylation, typically observed in epidermal growth factor (EGF) and thrombospondin repeat (TSR) domains.[6]

Glycolipids are lipids covalently modified to glycans. The most abundant glycolipids in mammals are glycosphingolipids (GSLs), where the lipid is a sphingolipid, such as sphingomyelin. Glycosylation of these begins with the addition of a Glc or Gal monosaccharide to the ceramine (Cer) backbone, yielding a GlcCer or GalCer. These glycan chains can then be elongated and classified into groups, with the most well-known being gangliosides (ganglio-series lipids). Although official nomenclature requires a neutral ganglio-series core; all sialyated GSLs are colloquially referred to as gangliosides. Several example GSLs, including ganglioside GM2, are shown in Figure 1.3C.[6]

Other categories of glycans found throughout cells include NDP-monosaccharides (sugar precursors), O-GlcNAc, and glycation. Sugar precursors are monosaccharides activated by covalent attachment to nucleotide diphosphate (NDP), allowing for the addition to growing glycan chains. O-GlcNAc is a dynamic modification added by O-GlcNAc transferase (OGT) and removed by O-GlcNAcase (OGA) in the cytoplasm, mitochondria, and nucleus of eukaryote. Glycation refers to the non-enzymatic linkage of a glycan and a receptor molecule, such as proteins and DNA.[6,10,11]

Carbohydrates in the cell

Glycans and carbohydrates serve many purposes within biological systems, with the most studied categories being (1) metabolism, (2) structural contributions, and (3) roles as information carriers. Carbohydrates are the preferred source of energy for cells, especially glucose, the ubiquitous equatorial pyranose. Cells possess enzymes to convert carbohydrates and other biopolymers into glucose to fuel energy production through the citric acid cycle.[5]

Carbohydrates play various structural roles to allow cellular propagation and proliferation. The cell wall of plant cells is composed of primarily glycans such as cellulose, with a repeating unit of [4Glcβ1]$_n$ unit, and pectins (GalA polymers). The cell wall provides structural support, allowing the cell to withstand mechanical stress and osmostic pressure differences. The bacterial cell wall is primarily composed of peptidoglycan, a polymer of GlcNAc and MurNAc crosslinked by peptides. During neural cells differentiation, neural cell adhesion molecules (NCAMs) are glycosylated to have long, linear polysialic acid chains, with a degree of polymerization (DP) greater than 100. The high DP and negative charge of Sia saccharides repel neighboring cells, facilitating proper neuron migration and spacing. Additionally, on smaller scales, carbohydrates influence the protein structural dynamics and folding.[6]

The field of glycobiology is primarily interested in the role of carbohydrates as information carriers. This function is primarily mediated through cell-cell interactions, where a protein-carbohydrate handshake is the first step in many physiological processes.[12] Proteins that specifically bind carbohydrates for this purpose are known as glycan binding proteins (GBPs), with lectins and antibodies being of special interest. Lectins are a family of proteins with the specific purpose of carbohydrate binding.[6]

GBPs typically recognize material in the extracellular space for cell-cell interactions. Intrinsic GBPs recognize self-glycans and mediate cell-cell interactions. The mammalian sialoadhesin protein (Siglec-1, CD169) binds Sia, preferentially α2-3Sia, on neighboring cells and is implicated in macrophages for antigen presentation.[13,14] Extrinsic GBPs originate from exogenous organisms and viruses, recognizing non-self-glycans in parasitic or symbiotic relationships. For example, the SARS-COV2 viral spike protein interacts with heparan sulfate, suggesting a mechanism for targeting the human ACE2 receptor.[15]

Common GBPs include lectins, carbohydrate active enzymes (e.g. glycosyltransferases and glycosylsidases), and antibodies. Formal definitions of lectins and GBPs however exclude other enzymes, carriers, or native sugar sensors.[6] For example, the human tetraspanins and integrins interact with *cis* gangliosides, but they exist outside the canonical lectin/GBP nomenclature.[16,17] Figure 1.1 provides a cartoon depiction of glycosylation and protein-carbohydrate interactions at a cellular level.

The molecular mechanisms and experimental identification of protein-carbohydrate interactions

Protein-carbohydrate interactions are typically weak, with dissociation constants (K_d) in the mM to μM range. Protein-carbohydrate interactions however are usually multivalent. GBPs can possess multiple binding sites for a carbohydrate epitope, and carbohydrates themselves often present repeating binding units clustered on extracellular surfaces. Therefore, these interactions are commonly measured *in vitro* by the more biologically relevant avidity (combined binding strength) rather than affinity (single-site strength).[6]

The molecular mechanism of protein-carbohydrate binding often involves a fold or motif containing β-sheets. This binding mechanism uses hydrogen bonds with saccharide hydroxyls, indirect (water mediated) interactions, and/or π orbital interactions to bind a carbohydrate. In Figure 1.4, I show a direct hydrogen bond (1.4A), indirect hydrogen bond (1.4B), and CH-π bond (1.4C). In the polysialic acid binding antibody scFv735, the protein is stabilized by six (6) direct hydrogen bonds and eleven (11) indirect water mediated interactions. [18,19] Due to the number of indirect interactions, protein-carbohydrate interactions have proved challenging to computationally model.[18,19] In a structural analysis of the protein data bank (PDB), Hudson et al. (2015) found that the carbohydrate-binding pockets of proteins have a higher preference for aromatic residues, notably Trp and Tyr, for CH- π bonding.[20]

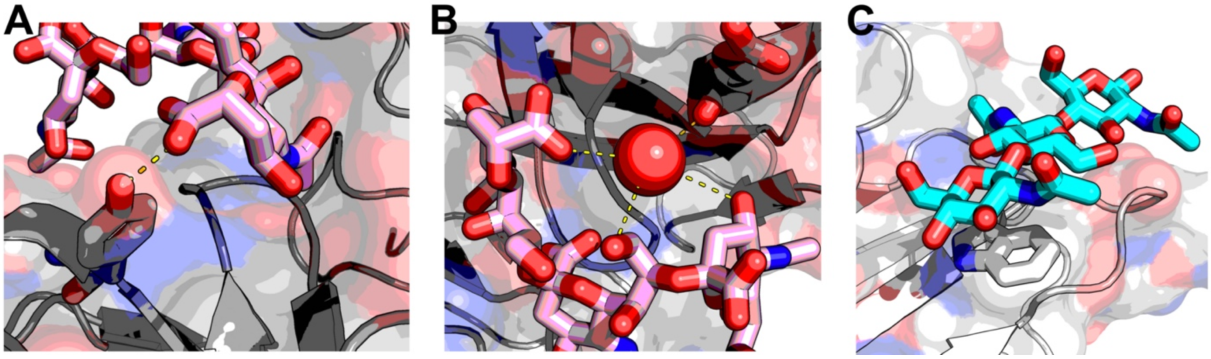


Figure 1.4: Mechanisms of protein-carbohydrate interactions. (A) Hydrogen bond of a Tyrosine-Sia (pink) interaction (PDB: 3WBD). (B) indirect (water mediated) interaction of Aspargine-Sia (pink) (PDB: 3WBD) (C) CH-π interaction of Tryptophan-GlcNAc (blue) (PDB: 8AD2).

While structural information is critical for understanding protein-carbohydrate interactions, traditional methods like crystallography are labor-intensive and not high throughput. Currently, the state-of-the-art methods for identifying protein-carbohydrate interactions without structural data include glycan arrays and diazirine linkers. Glycan arrays are solid supports with immobilized saccharides, enabling the non-covalent binding of a protein of interest.[21] Diazirine linkers are photoaffinity probes that can be attached to most glycans (provided the correct chemistry), delivered into a cell *in vivo*, and then irradiated to crosslink with the nearby (potentially binding) protein.[22] Recently, Zhang et al. performed the first diazirine linker experiments on gangliosides, identifying the first ever ganglioside interactome of 873 putative proteins.[17,23] Both glycan arrays and diazirine linkers allow high-throughput screening of protein-carbohydrate interactions; however, these methods are qualitative and thus fail to provide quantitative binding values and the precise carbohydrate-binding region of these proteins.

Although scientists have discovered many proteins that bind to carbohydrates through specific motifs, experimentally identifying these proteins, or the residues that bind the carbohydrates, remains difficult. As a result, identification of the entire protein-sugar interactome (the complete set of carbohydrate-binding proteins in a species) has not yet been possible.

In this dissertation, I computationally explore carbohydrate-binding proteins without restricting by protein family or function. My goal is to identify the protein-sugar interactome: to find all proteins that interact with carbohydrates across metabolic, structural, and molecular recognition functions. Throughout this dissertation, I use deep learning methods, which are explored in the following section.

Computational methods

Deep Learning Overview

Deep learning (DL) is a subset of machine learning (ML) leveraging a data-driven approach to classify input data with mathematical models of neurons (nodes). DL achieved remarkable performance in all areas of science, including image recognition and language processing, and is now emerging as a powerful tool in biology. In my dissertation research on glycobiology, I leveraged novel DL techniques. Here I provide a brief overview of DL fundamentals to familiarize readers.

Fundamentals and Dense Neural Network Framework

The most common and simplest neural network is the fully connected (FC) dense neural network (DNN). A simple multilayer DNN takes input features (*X*) and performs successive matrix multiplications to generate a series of hidden (h) representations and ultimately produce a predicted output ($\hat{Y}$) that estimates the true value Y. A single dense layer takes the form:

$$h_{i+1} = \sigma_i(W_i h_i + b_i) ,$$

where h_i is the **h**idden embedding at layer i, W_i is the **w**eights of layer i, and b_i is the bias of layer i, and σ is the activation function.[24–26] The input *X* is h_o, the embedding at layer 0, and $\hat{Y}$ is h_f, the final embedding of the model. The embeddings h_i are typically *n x 1* vectors, with W_i matrices shaping the output of each layer. $W_i h_i + b_i$ is a linear equation; therefore, activation functions σ are used to introduce non-linearity to the model. Several example activation functions, such as rectified linear unit (ReLU) and sigmoid are shown in Figure 1.5D. In Figure 1.5A, I show a schematic of a DNN.

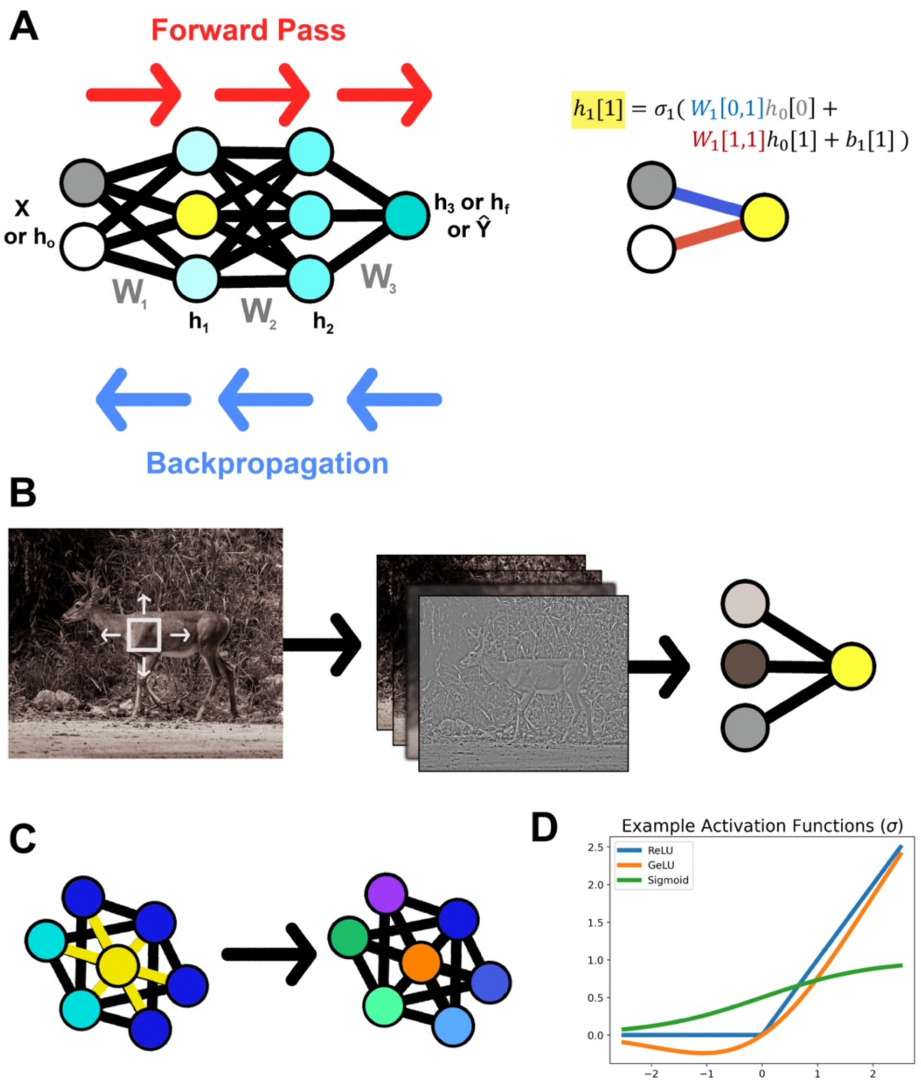


Figure 1.5: Annotated neural network architectures. (A) Simple three-layer dense neural network (DNN). (B) Simple convolutional neural network on an image. (C) A cartoon representation of an equivariant graph convolution. (D) Example activation functions.

Neural networks are not manually tuned equations: they are highly parametrized algorithms. Therefore, finding an optimal solution for W_i and b_i (the weights and biases) requires non-trivial methods. For this process, neural networks use back propagation to determine the parameters of the weights. The goal of a neural network is to map input features X to an output $\hat{Y}$

within some margin of error of the true value Y. We measure the difference between $\hat{Y}$ and Y using a loss function ($\mathcal{L}$). A simple loss function is the mean squared error (MSE), which is the loss function of linear regression, shown below.

$$\mathcal{L} = \left\| Y - \hat{Y} \right\|^2$$

The difference between the predicted and true values are measured by the loss is then used to update the weights of the neural network through backpropagation. Backpropagation leverages calculus, most notably the chain rule, to update the weights to reduce the error.[27] One algorithm for iterative backpropagation is stochastic gradient descent (SGD), shown below:

$$W = W - \eta \, \nabla\mathcal{L}(w)$$

where η is the learning rate. With these equations, we can construct a simple DNN to predict or classify input data of fixed size.[27]

Although the framework described above is simple, in practice, many variations have been developed to improve performance. The earliest activation functions were rectified linear units (ReLU) and sigmoid; however, more recent activation functions such as leaky ReLU, Gaussian error linear units (GeLU), and SoftMax are now common.[28] To improve generalization, batches (e.g. predicting on multiple inputs at once) are often be used for training alongside batch normalization, layer normalization, and dropout.[29,30] Finally, common variations of weight updating include the Adam optimizer,[31] weight decay, and stochastic weight averaging (SWA). [32]

Convolutions capture patterns

DNNs are fantastic tools for one-dimensional data, showing strong predictive power on many non-trivial tasks. Despite their power on 1D data, DNNs are not optimal for higher dimensional data, such as images, because object position can vary tremendously throughout the inputs. To better capture patterns in such data, convolutional neural networks (CNNs) are used. CNNs employ the common mathematical operation of a convolution:

$$(f * g)(t) = \int f(x)g(x - t)dx$$

where $g(t)$ is the function (or input) and f is the convolving function (or filter). However, because digital images are discrete (composed of pixels, or voxels in 3D images); a discrete convolution operation is used:

$$h_{k,l+1} = \sum_{m} \sum_{n} f_{k,l}[m, n] * h_l[i - m, j - n]$$

where $f_{k,l}$ is the k'th convolutional filter at layer l, h_l is the embedding at layer l which is a concatenation of all $h_{k,l}$ values - the result of the filter k on h_l.[24–26] This convolution process is illustrated in Figure 1.5B.

A typical CNN stacks several convolutional layers, before "flattening" the result (converting from the 2D or 3D matrix to a 1D vector) for forward propagation by dense layers.[24–26] Common convolutional layer variations include padding, dilation, stride, and pooling.[33]

3-Dimensional data requires equivariant information

Protein structures are typically represented in the protein data bank (PDB) format, which lists the fixed-point Cartesian coordinates of each atom in Angstroms (Å). Cartesian coordinates are versatile, allowing the calculation of protein features that are invariant or equivariant to rotation and translation, such as dihedral angles, bond angles, bond lengths, and residue-residue contacts. These invariant and equivariant properties reflect the intrinsic features of the protein and are independent of any position or orientation in Cartesian space. Therefore, I leverage equivariant algebra to describe protein structures, mapping values from the input coordinate domain to an equivariant codomain defined by the appropriate symmetry group. Formally, a function $f(x)$ is equivariant if, for a symmetry operation G:

$$G(f(x)) = f(G(x))$$

For proteins, G is the 3-dimensional roto-translation group $SE(3) = SO(3) \ltimes \mathbb{R}^3$.[34] To model 3D proteins in their native 3D space, we require frameworks that *only* use equivariant and invariant features to these symmetry operations. Predictions on raw Cartesian space would depend entirely on the arbitrary placement of the protein in space, not its true intrinsic properties.

In this dissertation, I leverage a $SE(3)$–equivariant neural network framework called equivariant graph neural network (EGNN). EGNN employs equivariant graph convolutional layers (EGCLs) to recognize patterns of graphs. The foundational equations used by EGCLs are:

$$m_{ij} = \sigma_e(h_i^l, h_j^l, ||x_i - x_j||^2, a_{ij})$$

$$m_i = \sum_{i \neq j} m_{ij}$$

$$h_i^{l+1} = \sigma_h(h_i^l, m_i)$$

Where h_i^l is the embedding of node *i* at layer *l*, m_{ij} is the message from node *j* to node *i*, x_i is the coordinates of node *i*, a_{ij} is the edge attributes of nodes i and *j*, σ_e is the message activation function, and σ_h is the node activation function.[35] Messages are calculated for all neighboring nodes, typically determined by a distance cutoff or by *k*-nearest neighbors. Edge attributes for proteins often include distance (represented by a radial basis function (RBF)), orientation, and direction between neighboring nodes.[36] Using this approach, I develop models that propagate and process protein structural information in the natural 3D graph space.

3D equivariant graph neural networks are an area of active study. EGNN is one of the simplest frameworks for 3D graph predictions; most alternative methods use spherical harmonics to propagate information.[37] I chose EGNN for its straightforward mathematics and improved performance relative to the spherical harmonics-based methods.

Biophysical Deep Learning Models

Although DL theory has existed since the 1950s, the advent of GPU acceleration has finally enabled the practical applications of DL algorithms. Since 2020, DL applications in biophysics have grown exponentially. Two recent general biophysical models of interest are AlphaFold2 and ESM (evolutionary scale modeling).[38,39]

Google DeepMind employees recently received the 2024 Nobel Prize in Chemistry for the development of AlphaFold2 (shared with Dr. David Baker for his pioneering work in *de novo* protein design).[40] AlphaFold2 (AF2) is a DL model that inputs a protein sequence and predicts the complete 3D protein structure.[38] AF2 uses a dual-track approach, integrating multiple sequence alignment (MSA) information with 2D representations to predict amino acid positions in a canonical frame.[38]

AF2 was evaluated in the 14th Critical Assessment of Protein Structure Prediction (CASP14) challenge, where it outperformed every competing method and predicted the best model for 89 of 97 targets.[41] AF2's strong performance is due in large part to training on the entire PDB, a testament to the open scientific sharing of innumerable independent researchers across the world for the past 50+ years.

Recently, AF2 was updated to AlphaFold 3 (AF3), which uses a generative diffusion network for protein coordinate prediction.[42] AF3 improves on AF2 as it does require inputs to be canonical amino acids, enabling it to model post-translational modifications (PTMs), DNA, and ligands of arbitrary input.[42] AF3 achieves a 76% success rate on the small molecule docking

PoseBusters benchmark, but, prior to this dissertation research, had not been evaluated for carbohydrate docking prediction tasks.[42]

Following the release of ChatGPT, computational biologists adapted transformer architectures to protein sequence data. One notable transformer-based deep learning model is evolutionary scale modeling (ESM).[39] ESM is a large language model (LLM) trained on masked protein sequences (where certain amino acids were hidden), with the goal of predicting the identify of those masked residues.[39]

The strength of ESM is in its representation of proteins. ESM2 has a 34-layer architecture, where its final layer provides a 1280-dimensional embedding that can be extracted and used as input for other deep learning models.[39] This embedding contains evolutionary information about the protein, similar to an MSA, improving downstream performance.[39,43]

Advances in Computational Glycobiology

Due to the scarcity of experimental data, computational glycobiology has also been constrained, but is currently poised for significant growth and advancement. Here I provide a non-extensive list of the current algorithms and methods for computational glycobiology, spanning from glycosite prediction, glycosylation prediction, binding prediction, and protein-carbohydrate docking.

Although the only N-linked glycosylation sequon is the well-known NX(S/T) motif, glycosylation events are not homogenously distributed across all such motifs in a protein. Different regions are preferentially glycosylated by various enzymes. LMNglyPred is a neural network using an LLM to predict N-linked glycosylation sites.[44] Similarly Stack-OglyPred-PLM predicts O-linked glycosylation sites.[45]

Given a glycosylation site, determining which glycan is preferentially expressed at the position is imperative. The (proprietary) InSaNNE neural network was trained on the GlyConnect database[46] to predict the specific glycan given the surrounding sequence (*n*-5 to *n*+5) of the NX(S/T) motif.

InSaNNE uses an LLM, named SweetNet,[47] which is trained on biologically observed glycans. SweetNet represents each glycan as a graph, with saccharides as nodes and their covalent connections as edges. LectinOracle concatenates SweetNet glycan embeddings and ESM-1b protein embeddings to predict which glycans a provided lectin can bind.[48]

Currently, most carbohydrate research is fueled by protein sequences; however, my predecessor in the lab, Dr. Morgan Nance, concentrated on structural modeling of protein-carbohydrate interactions. Nance developed the Rosetta-based method of GlycanDock, a local refinement technique for protein-carbohydrate docking.[49] GlycanDock was the first algorithm designed specifically for docking carbohydrate-protein complexes within the Rosetta suite. Previous tools, such as AutoDock,[50] required stand-alone protocols requiring manual interventions for pipelines such as protein design.

While the aforementioned methods are critical to better understand structural glycobiology, **no high-throughput approach exists to identify the protein-sugar interactome**. Next-generation sequencing has made available more than 25,000 reference genomes, with high confidence *de novo* structures for over 80 of those species. Although numerous protein-protein interactome maps have been generated, no equivalent map exists for the protein-sugar interactome. **Here, in this dissertation, I have developed a new method to uncover the protein-sugar interactome for any species.**

Dissertation Overview

Prior to my doctoral work, no publicly available methods existed for predicting non-covalent binding of proteins and carbohydrates – either for determining *whether* a protein binds carbohydrates or for identifying specific binding residues. The Gray lab focuses on *de novo* therapeutic development, with a long-term goal of engineering proteins that bind glycoproteins, such as viral receptors, with high specificity. However, with only a limited understanding of protein-glycan interactions, *de novo* protein design of is constrained by our basic scientific understanding of the protein-sugar interactome. **Therefore, the objectives of my research are to elucidate how proteins bind to carbohydrates, to discover all proteins capable of carbohydrate-binding, and to identify current limitations in *de novo* protein-carbohydrate docking predictions.**

Chapter 1 summarizes the biological roles of carbohydrates and the computational techniques employed in my doctoral studies. Chapter 2 details a method I developed in collaboration with Dr. Sudhanshu Shanker: **CA**rbohydrate **P**rotein **S**ite **I**denti**F**ier (CAPSIF). Chapter 3 presents two deep learning algorithms I developed: CAPSIF2, an updated version of CAPSIF, and **P**rotein **i**nteraction of **CA**rbohydrate **P**redictor (PiCAP), which predicts whether a protein binds carbohydrates. Additionally, Chapter 3 analyzes the *E. coli*, *M. musculus*, and *H. sapiens* proteomes for carbohydrate binding. Chapter 4 evaluates the performance of current *de novo* all-atom structure prediction on protein-carbohydrate docking. Chapter 5 summarizes my contributions to the field of protein-carbohydrate modeling and highlights potential directions for future research in the field of computational glycobiology.

Chapter 2

CAPSIF: Structure-based neural network protein-carbohydrate predictions at a residue level

Adapted from: Canner, S. W.[‡], Shanker, S.[‡] & Gray, J. J. Structure-based neural network protein–carbohydrate interaction predictions at the residue level. *Frontiers in Bioinformatics*. 3:1186531, (2023).

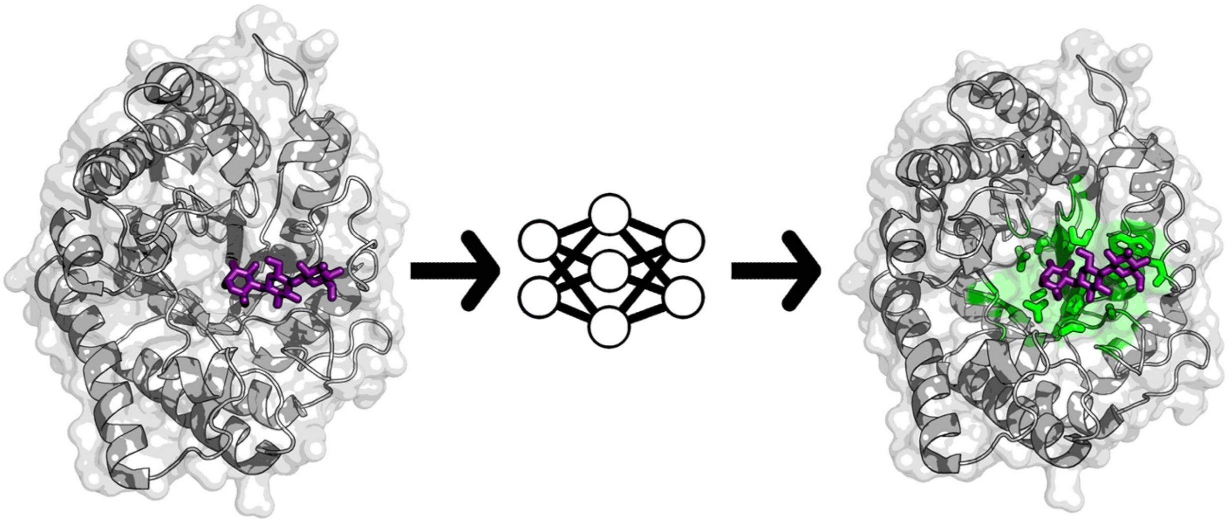


Figure 2.1: CArbohydrate Protein Site IdentiFier (CAPSIF) analyzes protein structures to identify carbohydrate binding pockets.

Attribution of credit: SWC (Writing, Methods, Analysis, Figures), SS (Conceptualization, Methods, Analysis), JJG (Conceptualization, Writing, Analysis).

[‡]indicates equal contribution of the work. As this work was performed in equal contribution with Dr. Shanker, I will use the term "we" and "our" in this chapter to properly attribute credit to Dr. Shanker's efforts.

Overview

Carbohydrates dynamically and transiently interact with proteins for cell-cell recognition, cellular differentiation, immune response, and many other cellular processes. Despite the molecular importance of these interactions, there are currently few reliable computational tools to predict potential carbohydrate binding sites on any given protein. Here, we present two deep learning models named CArbohydrate-Protein interaction Site IdentiFier (CAPSIF) that predict non-covalent carbohydrate binding sites on proteins: (1) a 3D-UNet voxel-based neural network model (CAPSIF:V) and (2) an equivariant graph neural network model (CAPSIF:G). While both models outperform previous surrogate methods used for carbohydrate binding site prediction, CAPSIF:V performs better than CAPSIF:G, achieving test Dice scores of 0.597 and 0.543 and test set Matthews correlation coefficients (MCCs) of 0.599 and 0.538, respectively. We further tested CAPSIF:V on AlphaFold2-predicted protein structures. CAPSIF:V performed equivalently on both experimentally determined structures and AlphaFold2 predicted structures. Finally, we demonstrate how CAPSIF models can be used in conjunction with local glycan-docking protocols, such as GlycanDock, to predict bound protein-carbohydrate structures.

Introduction

The carbohydrate-protein handshake is the first step of many pathological and physiological processes.[12] Pathogens attach to host cells after their lectins successfully bind to surface carbohydrates (or glycans)[6,51–53]. The innate and adaptive immune systems utilize carbohydrate signatures present on cellular and subcellular surfaces to recognize and destroy foreign components[54,55]. Glycosaminoglycans (GAGs) bind to membrane proteins of adjacent cells for cell-cell adhesion and to regulate intracellular processes[56–58]. Despite the biological importance of these carbohydrate-protein interactions, there are few carbohydrate-specific tools leveraging the vast Protein DataBank (PDB) and recent advances in machine learning (ML) to elucidate the binding of carbohydrates at a residue level.

Knowledge of carbohydrate-protein interactions has been leveraged to develop therapeutic candidates to neutralize infections and inspire proper health function.[59] One bottleneck in designing carbohydrate-mimetic drugs is obtaining residue-level interaction knowledge through methods such as structural data and/or mutational scanning profiles [60–62]. Further, in some studies, computational tools have been used to predict docked structures, refine bound carbohydrates, or extract dynamic information[62–64].

Recent developments in deep learning (DL) have substantially enhanced the theoretical modeling of proteins and protein-protein interactions. For example, neural networks can design stable proteins with unique folds using graph representations.[36] 3D structures can be predicted with programs such as IgFold [65] and Alphafold2 (AF2).[40] Predicted 3D atomic coordinates can be probed to determine ligand or protein binding capabilities using neural networks such as Kalasanty or dMaSIF.[66,67]

Recent computational studies have demonstrated new ways to explore protein-carbohydrate interactions. Our lab has also contributed to the advancement of this field by adding the following, (1) a shotgun scanning glycomutagenesis protocol to predict the stability and activity of protein glycovariants,[68] and (2) the GlycanDock algorithm to refine protein-glycoligand bound structures.[49]

Recently there have been developments in small molecule binding site predictors. Small molecule binding site predictors typically fall into four categories: template, geometry, energy, or machine learning based.[69] Template based strategies, such as 3DLigandSite,[70] search datasets for sequence and/or structurally related ligand binding proteins to assess prospective binding sites. Geometry based methods, like FPocket,[71] search the surface of proteins for pockets and cavities. Energy based methods, such as FTMap,[72] use probe molecules to scan the surface of a protein to determine the energetic favorability of binding. Recently, machine learning techniques, such as Kalasanty,[66] have emerged and outperformed previous classical site prediction algorithms, commonly with convolutions on a 3D voxel grid containing atomistic information.[73,74]

Although there are many general small molecule binding site predictors,[66,72,75] few tailored algorithms exist for prediction of protein-carbohydrate sites. In 2000, Taroni *et al.* performed an analysis of carbohydrate binding spots using the solvation potential, residue propensity, hydrophobicity, planarity, protrusion, and relative accessible surface area to construct a function to predict carbohydrate binding sites.[76] In 2007, Malik and Ahmad created a neural network to predict the carbohydrate binding sites using their constructed Procarb40 dataset, a collection of 40 proteins, with leave one out validation.[77] In 2009, Kulharia built InCa-SiteFinder to predict carbohydrate and inositol binding sites by leveraging a grid to construct an energy-based method for predicting binding sites.[78] Tsai *et al.* constructed carbohydrate binding probability density maps using an encoding of 30 protein atom types as an input to a machine learning algorithm.[79] Later, Zhou, Yang and colleagues developed two methods to predict carbohydrate binding sites, (1) a template-based approach named SPOT-Struc[80] and (2) a support vector machine (SVM) named SPRINT-CBH that leverages sequence-based features.[81] Tsia[79] and SPOT-Struc[80] have achieved Matthews correlation coefficients (MMCs) of 0.45 on test sets of 108 and 14 proteins, respectively. The increased size of the protein databank and the improvements in deep learning methods now presents an opportunity to train and test more broadly.

Larger protein-carbohydrate structural databases now include UniLectin3D[82] and ProCaff.[83] UniLectin3D focuses on lectins bound to carbohydrates, containing 2406 structures; however, it contains many redundant structures and is currently limited to 592 unique sequences. ProCaff lists 552 carbohydrate-binding protein structures and their binding affinities under various conditions; however, many structures are only available in the unbound form.

Many drug targets, from pathogen-lectins to aberrant selectins, are carbohydrate binding proteins. [59,84] Understanding the physiological response and determining a glycomimetic drug to neutralize the infection requires residue-level knowledge. [84] Currently, DL algorithms LectinOracle[48] and GlyNet[85] predict lectin-carbohydrate binding on a protein level; however, pharmaceutical development requires residue-level information.

In this work, we develop two DL methods for residue-level carbohydrate-binding site prediction for non-covalently bound carbohydrates. The two methods have different architectures, one using voxel convolutions and one using graph convolutions. We also present a dataset of 808 non-covalently bound nonhomologous protein chain-carbohydrate structures and use it to train and test both models. We compare the performance of the models with each other and with FTMap[72] and Kalasanty.[66] Then, we evaluate the performance of the models on AlphaFold2[40] predicted versus experimentally determined structures. Finally, we present a proof-of-concept pipeline to predict bound protein-carbohydrate structures.

Results

Dataset for carbohydrate-protein structures

To construct a method to predict carbohydrate-protein interactions, we needed a large and reliable dataset to use for training and testing. The dataset should contain as many non-homologous structures as possible to avoid biasing to specific folds. By filtering the PDB[86] we constructed a dataset of 808 high accuracy (< 3 Å resolution), nonhomologous (30% sequence identity), and physiologically relevant experimental structures (by manually removing buffers), spanning 16 carbohydrate monomer species. When multiple copies were present in the same PDB file, we used only a single protein chain and all adjacent carbohydrate chains. In these structures, 5.2% of the protein residues contact carbohydrates. The final dataset consists of 808 structures, which we split into 521 training structures, 125 validation structures, and 162 test structures. These structures only contain single chain protein interactions with non-covalently bound carbohydrates.

CAPSIF uses deep neural networks to predict carbohydrate interaction sites

We constructed convolutional neural networks (CNNs) named CArbohydrate-Protein Site IdentiFier (CAPSIF) to predict carbohydrate binding residues from a protein structure. CNNs were initially developed for images, exploiting the spatial relationship of nearby pixels for prediction tasks. They have been applied to predict protein structure[87–89] and small molecule binding pockets of proteins.[66] To predict carbohydrate binding residues using structural information, we created two CAPSIF CNN architectures, CAPSIF:Voxel (CAPSIF:V) and CAPSIF:Graph (CAPSIF:G).

Since a protein can change its side chain conformations upon binding a small molecule or carbohydrate (from *apo* to *holo*), we sought a protein representation that is robust to these and other binding induced changes. We chose a residue-level representation, using only the Cβ positions of all residues (or Cα in glycine), since the Cβ position is frequently equivalent in both the *apo* and *holo* states.[90] Both CAPSIF architectures use the following features: unbound solvent accessible surface area (SASA) of each residue, a backbone orientation (architecture specific), and encodings of amino acid properties, including hydrophobicity index (0 to 1),[91] "aromatophilicity" index (0 to 1),[92] hydrogen bond donor capability (0,1), and hydrogen bond acceptor capability (0,1) (Methods/Table 2.3).

The first CAPSIF architecture, CAPSIF:V, is a 3D voxelized approach to learn carbohydrate binding pockets. CAPSIF:V uses a UNet architecture, which comprises a grid with a series of convolutions compressing and then decompressing the data to its original size with residual connections to previous layers of the same size. For each grid, we used an 8 Å^3 voxel size where CAPSIF:V encodes each residue's β carbon (Cβ) into a corresponding voxel. CAPSIF:V predicts a label *P*(carbohydrate-binding residue) for each voxel on the initial grid (Figure 2.2A; Methods/Figure 2.7**)**.

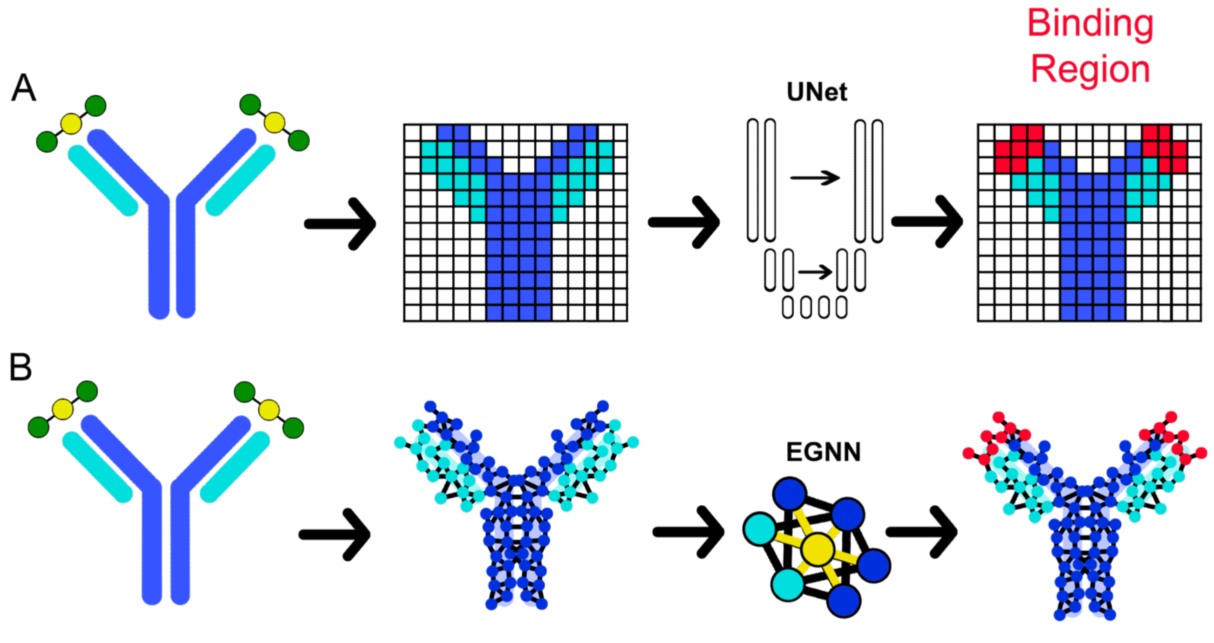


Figure 2.2: Two deep learning models that predict where proteins bind carbohydrates. (A) The first model (CAPSIF:V) maps the β carbon (Cβ) coordinates into voxels, utilizes a convolutional UNet architecture, and predicts the binding residues. (B) The second model (CAPSIF:G) converts the Cβ coordinates into network nodes with edges for residue-residue neighbors, performs convolutions on nodes with respect to neighbors with an equivariant graph neural network (EGNN) architecture, and predicts which residues bind sugars.

The second architecture, CAPSIF Graph (CAPSIF:G), is an equivariant graph neural network (EGNN),[93] with each Cβ represented as a node on the graph and edges connected between all neighbor residues within 12 Å (Figure 2.2B). EGNNs use graph-based convolutions with message passing between connected nodes based on node features and the edge features (distances).[93] We explored many variations of these neural network architectures; the Supporting Information includes data supporting our architecture and data representation choices.

The carbohydrate-binding residues comprise 5.2% of the dataset. To ameliorate the effect of data imbalance, we followed Stepniewska-Dziubinska *et al.* and chose the complement of the Dice similarity coefficient (d) as our loss function ($L = 1 - d$).[66] The Dice coefficient is normalized by both the correctly and incorrectly predicted residues:

$$d = \frac{2*TP}{(TP+FP)+(TP+FN)}, \text{(Eq 2.1)}$$

where *TP* = true positives, *FP* = false positives, and *FN* = false negatives. Since *d* does not depend on true negative labels, this loss function is insensitive to imbalanced datasets where the positive label is observed much less than the negative label.[66]

CAPSIF predicts carbohydrate-binding residues with encouraging accuracy

CAPSIF:V and CAPSIF:G are novel architectures for predicting carbohydrate binding residues; however, they use 512 structures to train with a substantial data imbalance. We therefore investigated the performance of CAPSIF on a held-out test set to determine whether the architectures accurately predict carbohydrate-binding regions despite the small amount of training data. Four representative CAPSIF:V predictions are shown in Figure 2.3, highlighting *TP* residue predictions, (green), *FP* residues (blue), and *FN* residues (red). CAPSIF:V captures the binding pocket visually for an endoglucanase (2.3A), xylanase (2.3B), and β-glucanase (2.3C), but it performs poorly on the HINT protein that binds ribose (2.3D), a five membered ring carbohydrate that is commonly associated with nucleotides.

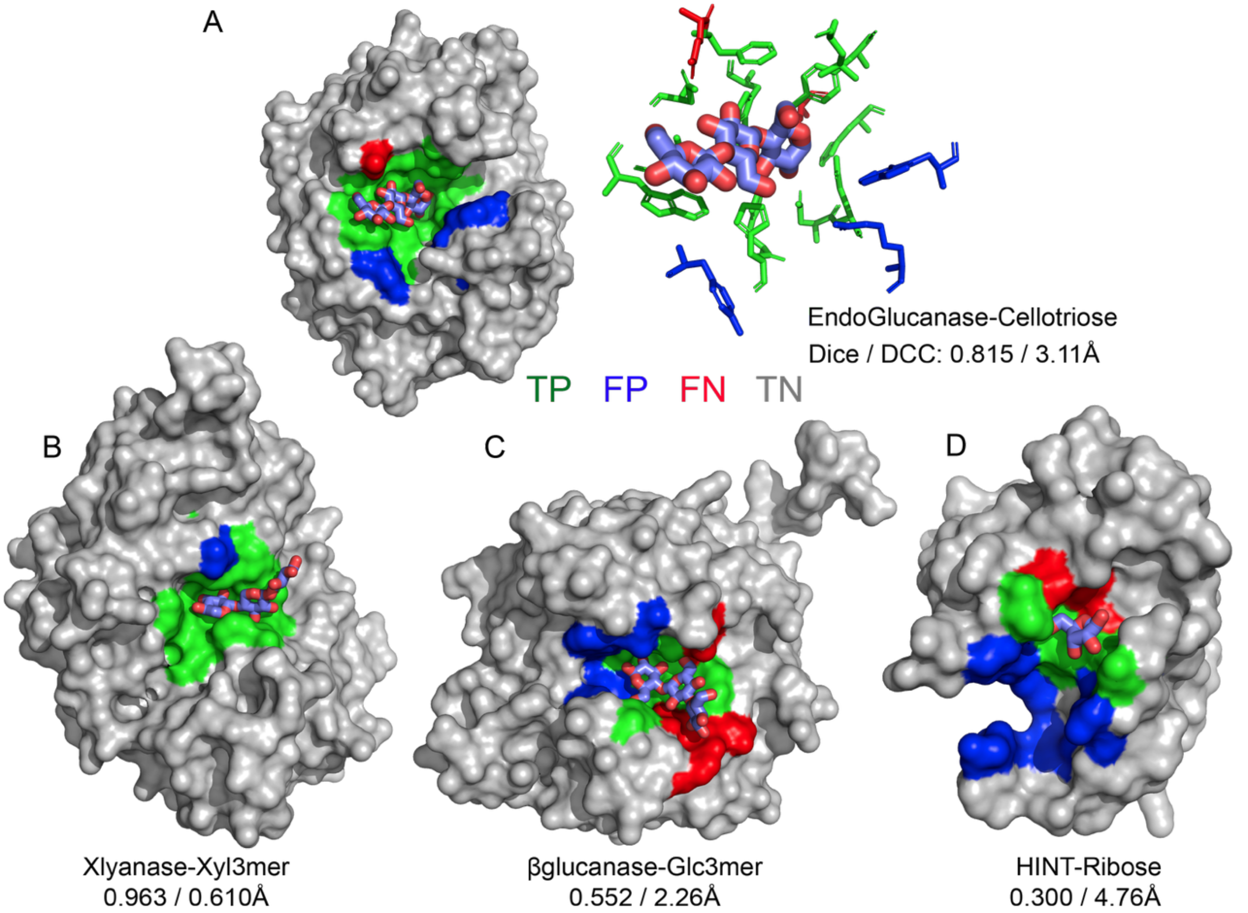


Figure 2.3: Prediction of carbohydrate binding sites on a protein surface using CAPSIF:Voxel. (A) Two representations of binding residues for cellotriose bound to endoglucanase (6GL0), surface (left) and sticks (right); Predicted surface representation of (B) xylanase bound to a xylose 3-mer (3W26), (C) β-glucanase bound to a glucose 3-mer (5A95), and (D) HINT protein bound to a ribose monomer (4RHN) predictions. True positive residue predictions are colored green, false positives are blue, false negatives are red, true negatives are gray, and the bound carbohydrate is cyan; Dice is defined by eq (1) and DCC is distance from center to center of the predicted binding regions.

For comparison, we evaluated how small molecule binding site predictors FTMap[72] and Kalasanty[66] perform for carbohydrate-binding tasks. We assessed these methods using the following metrics: the Dice coefficient (*Eq 2.1*), distance from the center of the crystal to the center of the predicted binding location (DCC) of each independent binding site, positive predictive value (PPV), sensitivity, and Matthews correlation coefficient (MCC). Similar to the Dice coefficient, the MCC is suited for unbalanced datasets; it has been reported in previous carbohydrate binding site studies.[79–81] MCC is:

$$MCC = \frac{(TP*TN-FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \text{ (Eq 2.2)}$$

where *TN* = true negatives. MCC ranges from -1 (worst) to +1 (best). The Dice coefficient measures the overlap of correctly predicted interacting residues to all predicted interacting residues. We define a success as a Dice score greater than 0.6 or, following Stepniewska-Dziubinska *et al.*, a DCC under 4 Å.[66]

On the CAPSIF test set, FTMap achieved an average Dice coefficient of 0.351 and average DCC of 10.5 Å, and Kalasanty achieved an average Dice of 0.108 and average DCC of 14.6 Å (Table 2.1). Further, FTMap predicted 16.8% of test structures with greater than 0.6 Dice and 16.8% of test structures with less than 4 Å DCC, while Kalasanty predicted 0% of test structures with greater than 0.6 Dice and 21.4% of test structures with less than 4 Å DCC (Table 2.1, Figure 2.4A,B).

Table 2.1: Average metric for each method on test set. Dice similarity coefficient is defined by eq (2.1), PPV is positive predictive value = TP / (TP + FP), Sensitivity = TP / (TP + FN), DCC is distance from center to center of predicted versus experimentally determined residues and only calculated for proteins that yield predictions (coverage), MCC is Matthews correlation coefficient and defined by eq (2.2). Bold face indicates best performance for each metric.

Model	**Dice**	**PPV**	**Sensitivity**	**DCC (Å)**	**MCC**	**Coverage (%)**
FTMap	0.351	0.284	0.505	10.56	0.222	**100.0**
Kalasanty	0.108	0.080	0.207	14.62	-0.624	90.0
CAPSIF:V	**0.597**	**0.598**	**0.647**	**4.48**	**0.599**	94.4
CAPSIF:G	0.543	0.541	0.590	5.85	0.538	83.2

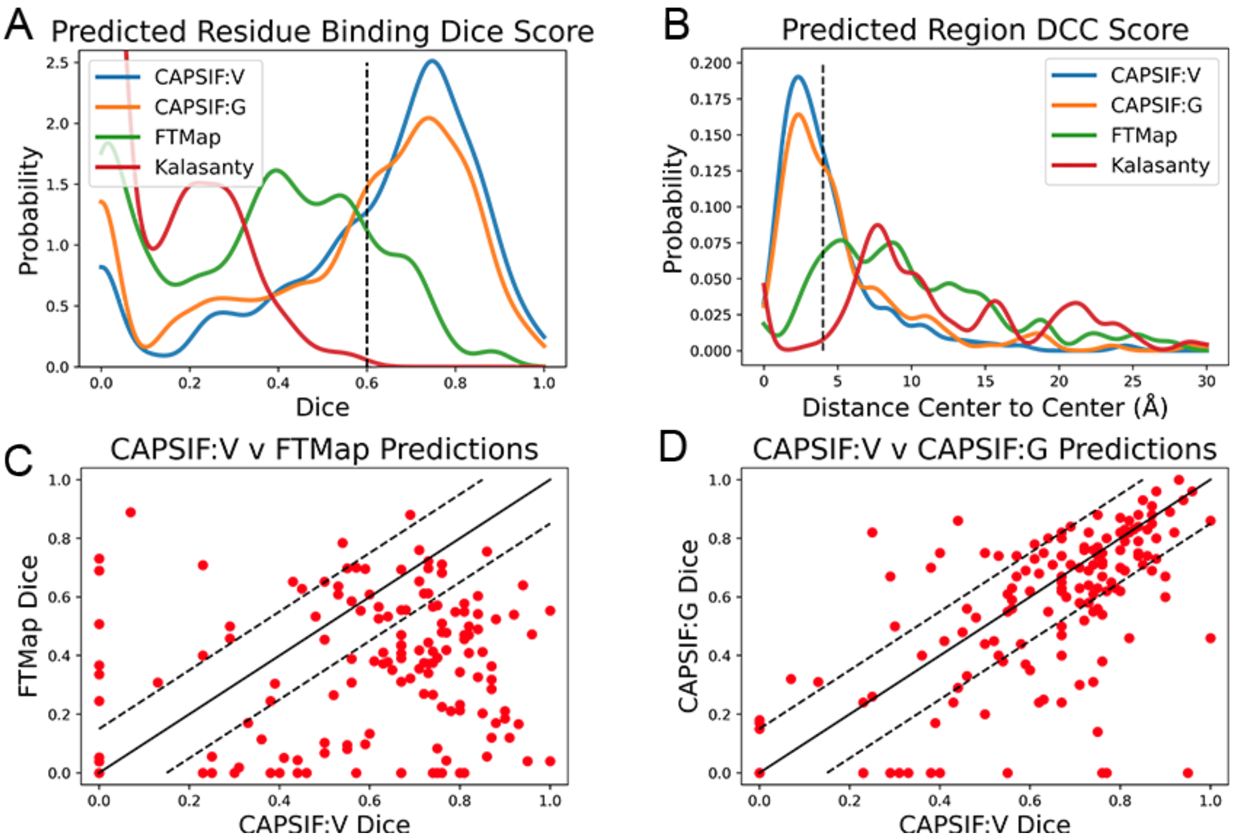


Figure 2.4: Distributions of CAPSIF:V and CAPSIF:G assessment metrics compared to FTMap[72] and Kalasanty.[66] (A) Distribution of Dice similarity coefficient for all methods smoothed with a Gaussian kernel density estimate (KDE, bandwidth h = 0.04); (B) Distance from center to center (DCC) of predicted to experimental carbohydrate binding residues (smoothed with a Gaussian KDE, h = 0.75 Å); (C) Per-target comparison of CAPSIF:V to FTMap and (D) CAPSIF:G Dice coefficients.

We then investigated whether our CAPSIF models, which are specifically tuned for carbohydrate binding, predict the carbohydrate binding regions more accurately than Kalasanty and FTMap. On the held-out CAPSIF test set, CAPSIF:V achieves an average 0.596 Dice coefficient and 4.48 Å DCC metric, and CAPSIF:G achieves an average 0.543 Dice coefficient and 5.85 Å DCC metric (Table 2.1). Further CAPSIF:V successfully predicts 62.7% of test structures with greater than 0.6 Dice and 56.5% of test structures with less than 4 Å DCC, and CAPSIF:G successfully predicts 55.2% of test structures with less than 0.6 Dice and 46.0% of test structures with less than 4.0 Å DCC. Both CAPSIF models have a most probable prediction at 0.77 Dice and 2.5 Å DCC (Table 2.1, Figure 2.4A,B).

Since CAPSIF is ML based and FTMap is energy based, FTMap may predict more accurately on different cases compared to CAPSIF. We compared the CAPSIF:V and FTMap Dice scores for each structure (Figure 2.4C). FTMap achieves a significantly higher Dice coeffiecents (difference greater than 0.15 Dice) than CAPSIF:V in 10.9% of cases, and CAPSIF:V predicts the binding region with a significantly greater Dice coefficient than FTMap in 67.9% of cases. We also compared the computer time. On The FTMap server, FTMap requires an hour or more to predict the binding region for a single structure, whereas both CAPSIF:V and CAPSIF:G predict binding sites within seconds on a single CPU. Thus, on average, CAPSIF:V and CAPSIF:G outperform current small molecule binding site predictors for carbohydrate binding.

Finally, we compared the CAPSIF:V architecture to the CAPSIF:G architecture. CAPSIF:V has an average Dice coefficient of 0.596 and CAPSIF:G has an average Dice coefficient of 0.543 across the test dataset (Table 2.1). When comparing the Dice on the test set, CAPSIF:V predicts 27.3% of structures with greater than 0.15 Dice than CAPSIF:G, while CAPSIF:G predicts 11.2% of structures with greater than 0.15 Dice than CAPSIF:V (Figure 2.4D). Thus, CAPSIF:V outperforms CAPSIF:G on carbohydrate binding site prediction.

Carbohydrates are unique biomolecules that bind to different lectins with high specificity. Both CAPSIF architectures treat all carbohydrates agnostically, meaning that all sugar residue types are considered equivalent for predictions. Nonetheless, we compared prediction results across different sugar residue types. (Appendix). CAPSIF:V performs best on glucose (Glc), galactosamine (GalN), arabinose (Ara), xylose (Xyl), ribose (Rib), and galacturonic acid (GalNAc). It predicts regions that bind neuraminic acid (Neu/Sia), fucose (Fuc), and Glucuronic acid (GlcNAc) with less than an average 0.5 Dice coefficient. The weaker performance could stem from the chemical differences or differences in the size of the training data. Neu and Fuc are substantially chemically distinct carbohydrates, as Neu is a 9-carbon structure and Fuc adopts an (*L*) conformation; both are sparse in our dataset. Further, CAPSIF:V performs best on transport proteins, membrane proteins, and hydrolases; however, it performs weakly on viral proteins and lyases.

CAPSIF:Voxel in most cases performs similarly on AlphaFold2 structures

Both CAPSIF models were trained and tested on bound crystal structures; however, experimental protein structure determination can be expensive, even in the absence of a carbohydrate. We therefore investigated whether CAPSIF:V could usefully predict carbohydrate binding structures from computationally modeled structures. We reconstructed the test protein structure dataset with the Colab implementation[94] of AlphaFold2 (AF2)[40] and predicted the carbohydrate binding residues of the predicted structures and evaluated the same performance metrics (Table 2.2). CAPSIF:V predicts the carbohydrate binding regions with similar Dice coefficients of 0.597 and 0.586 for protein databank versus AF2 predicted structures, respectively. Figure 2.5A shows that the Dice distribution is similar between PDB and AF2 structures. CAPSIF:V predicts the center of the carbohydrate binding region more accurately on AF2 structures with a DCC of 3.8 Å, compared to 4.5 Å on crystal structures.

Table 2.2: Metrics for CAPSIF:Voxel inputting PDB or AF2 structures. Dice, PPV, Sensitivity, DCC, MCC, and defined in Table 1.

Structures	Dice	PPV	Sensitivity	DCC (Å)	MCC	Coverage (%)
PDB	0.597	0.598	0.647	4.48	0.599	94.4
AF2	0.586	0.508	0.744	3.76	0.598	85.0

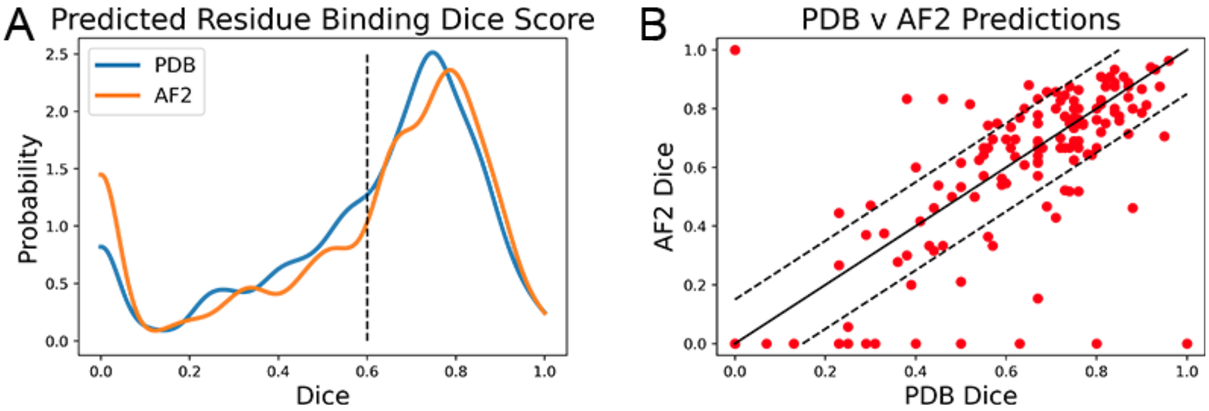


Figure 2.5: Dice coefficient assessment of CAPSIF:Voxel on PDB and AlphaFold 2 (AF2) structures. (A) Kernel density estimate (h = 0.04) showing the distribution of Dice coefficient for both methods; (B) Comparison of each test structure between CAPSIF:V on PDB and AF2 structures.

Although CAPSIF:V has a lower average DCC on AF2 structures compared to experimental structures, CAPSIF:V fails to predict any sites at all on 15% of AF2 structures, whereas it fails in only 5% of PDB structures, suggesting that the signal about the sugar binding is removed for some of the small backbone errors produced by AF2.

The multiple outliers where CAPSIF:V fails to predict the region of carbohydrate binding in only AF2 predicted structures are sorted in Figure 2.4B. CAPSIF:V predicts a Dice coefficient of at least 0.15 units higher for PDB structures in 14.9% of structures and predicts AF2 structures with a 0.15 Dice coefficient or higher for 8.7% of test structures. AF2 generated structures can be inaccurate; however, in most of the test cases, AF2 captures the structures with angstrom level accuracy and the carbohydrate binding residues with high pLDDT confidence; unfortunately, the pLDDT confidence measure does not correlate with the CAPSIF success rate (Figure 2.15).

CAPSIF assists *ab initio* prediction of bound protein-carbohydrate structures

CAPSIF:V predicts the carbohydrate binding site on the majority of proteins with high accuracy, suggesting that it might be used in a pipeline to predict bound protein-carbohydrate structures. As a proof-of-concept, we developed a prospective pipeline and tested it on five proteins from the GlycanDock[49] test dataset that were not included the CAPSIF dataset.

We constructed the following rudimentary pipeline. We predicted the binding site from each unbound protein's experimentally determined structure with CAPSIF:V and constructed the known carbohydrate with Rosetta. The carbohydrate's center of mass (CoM) was then placed in the CoM of the predicted binding region and manually rotated to align with the binding region shape. Next, we used the Rosetta FastRelax[95] protocol to remove steric clashes. Then we used Rosetta's standard GlycanDock[49] to predict the bound structures. To find the highest rated bound structure, we filtered 9,500 decoys by their computed interaction energy.

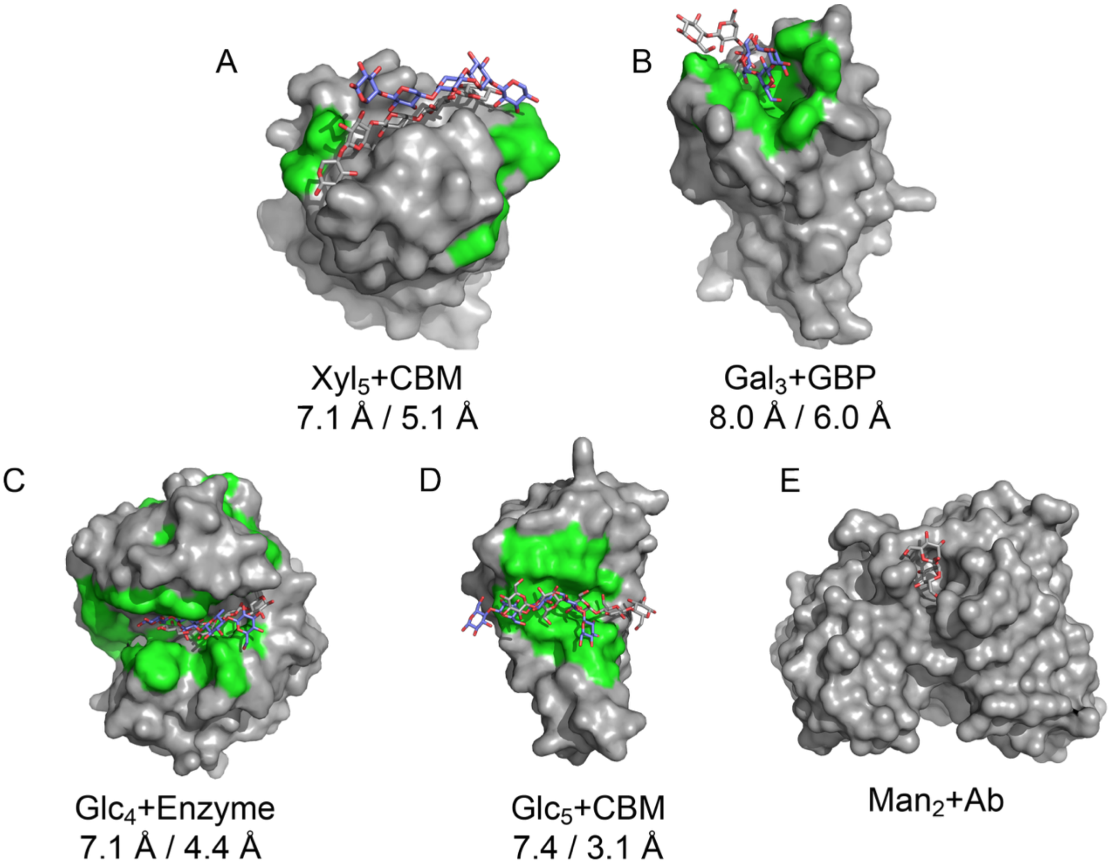


Figure 2.6: Results of CAPSIF:V-GlycanDock pipeline. CAPSIF-predicted residues are shown in green. Wild type unbound structures are shown in surface representation in gray with the experimentally determined carbohydrate in gray sticks and predicted bound carbohydrate in purple sticks. RMSD of entire ligand and RMSD of register-adjusted ligand are shown below. (A) a carbohydrate binding module (CBM), 1GMM (unbound PDB)/1UXX (bound PDB), (B) a glycan binding protein (GBP), 1L7L/2VXJ, (C) an enzyme, 1OLR/1UU6, (D) a CBM, 2ZEW/2ZEX, and (E) an antibody (Ab), 6N32/6N35.

We tested the pipeline on five experimentally solved unbound proteins: *P. aeruginosa* lectin 1, a glycan binding protein (GBP, 1L7L), two carbohydrate binding modules (CBMs, 1GMM and 2ZEW), a glycoside hydrolase enzyme (1OLR), and an anti-HIV-1 antibody (Ab) (6N32). Figure 2.6 shows structures and the root mean squared deviation (RMSD) of each predicted carbohydrate structure from the experimental structure. CAPSIF:V predicted carbohydrate binding residues near the correct site on four of the five proteins, but it failed to predict any binding residues on the antibody (6N32). For three of the proteins, CAPSIF:V predicts the region with high accuracy, but on 1GMM, CAPSIF:V predicts regions flanking the binding site, but still provides a similar CoM to the actual binding region. For the for carbohydrates with identified sites, the standard GlycanDock protocol was able to refine the carbohydrate structure to an RMSD of less than 8 Å for the entire ligand and less than 6 Å for register-adjusted values, where the termini were removed before calculating RMSD. The 3-mer Gal GBP (1L7L) has the worst RMSD (6 Å register adjusted, Figure 2.6B), likely because the *holo* conformation (2VXJ) undergoes a conformational change at the carbohydrate-binding site. Although this Ab case example failed, CAPSIF successfully predicted the carbohydrate binding regions of 9 of the 11 Abs tested from the Glycan Dock test set, which has no overlap with the CAPSIF training set. These predictions demonstrate the potential of CAPSIF to help inform experimental hypotheses or for high throughput predictions of bound protein-carbohydrate structures.

Discussion

We demonstrated that both CAPSIF models predict residues of proteins that bind carbohydrates with much higher accuracy than prior approaches. CAPSIF:V uses a voxelized approach and predicts 62.7% of crystal structures with a distance from the center of the predicted region to the center of the experimentally determined region (DCC) within 4 Å. CAPSIF:G performs strongly on the dataset, predicting 55.2% of crystal structures with a DCC less than 4 Å, with CAPSIF:V performing similarly or outperforming CAPSIF:G in 88.8% of cases. CAPSIF:V is robust to most errors in protein structure of the magnitude in AF2 structures (ångström-level): the algorithm predicts similar carbohydrate-binding residue regions independent of whether the input structure is experimentally determined or predicted by AF2. This algorithm is a substantial improvement over surrogate ligand site predictors Kalasanty and FTMap.

Further, CAPSIF outperforms previous methods specifically tuned for carbohydrate binding. CAPSIF:V achieves a 0.599 MCC and CAPSIF:G achieved a 0.538 MCC on the test dataset. Tsia *et al*'s method using probability density maps achieved a 0.45 MCC on their independent test dataset of 108 proteins,[79] SPOT-Struc achieved a ~0.45 MCC on their test dataset of 14 proteins,[80] and SPRINT-CBH achieves a MCC of 0.27 MCC on their test set of 158 proteins.[81] While these datasets differ from ours, ours is a similarly constructed non-homologous dataset of 162 structures, and CAPSIF has markedly stronger MCC. Although CAPSIF:V performs best, we advocate for usage of CAPSIF:V and CAPSIF:G in tandem to predict carbohydrate-binding residues since there are numerous cases where one CAPSIF model outperforms the other.

Although CAPSIF accurately captures the protein-carbohydrate binding interface, there are limitations. CAPSIF is carbohydrate-agnostic, so it only predicts that a protein residue will bind one of 16 carbohydrate monomers. That is, CAPSIF predicts the location of carbohydrate binding but not which carbohydrate preferentially binds there. Further, CAPSIF was only trained and tested on known non-covalent carbohydrate binding proteins, therefore CAPSIF may not be informative on non-carbohydrate binding proteins or proteins that bind glycoconjugates such as ribose in nucleic acids, ATP/ADP, or GTP/GDP (Figure 2.17). CAPSIF was trained on a small set of sixteen sugar residue types, and it will be most useful for non-modified sugar residues. Another limitation is that CAPSIF fails to predict any binding on about three times as many AF2 predicted structures as crystal structures. Unfortunately, CAPSIF prediction accuracy on AF2 structures is not correlated with pLDDT confidence metrics so it is not possible to know when it will fail. Further, CAPSIF was tested on AF2 predicted structures for proteins that already exist and may already exist in the AF2 training set. CAPSIF additionally is unable to predict whether a protein is a carbohydrate binding protein (Figure 2.18).

The scope of CAPSIF makes it well suited for a computational pipeline. We suggest the use of DeepFRI,[96] a deep learning model that predicts protein function, to first determine if the protein is a carbohydrate binding protein. If the protein is a carbohydrate binding protein, then LectinOracle [48] and GlyNet [85] can be used to predict which carbohydrates bind the protein. CAPSIF can then predict binding locations, either from an experimental structure or AF2 generated structures, and then GlycanDock[49] can predict a docked protein-carbohydrate structure.

We tested part of this pipeline by predicting the binding region using CAPSIF:V and docking the known carbohydrate binder to the region with GlycanDock.[49] CAPSIF:V predicted binding sites on four of the five proteins. The antibody case, which failed, binds a carbohydrate at the complementary determining region (CDR) loops, split over two chains, but CAPSIF was trained only on single chain data. When register adjusted, each structure yielded a ligand RMSD less than 6 Å. We anticipate that a more well-tuned pipeline could yield higher accuracy structures *ab initio* from sequence only.

To our knowledge, voxelized and graph-based site prediction has not been presented simultaneously before. Existing studies have used graphs to either predict binding affinity[97] or a docked structure (in coordination with diffusion techniques),[43] but they have not been used to determine small molecule binding regions. We tested two architectures utilizing either voxel or graph representations. We showed that CAPSIF:V outperforms CAPSIF:G, both of which use convolutions to predict the carbohydrate binding ability of residues with the same residue representation. We can speculate about the reason by considering the differences between the approaches. CAPSIF:V discretizes the protein structure over a 3D grid, which can obscure the Cβ position by a few Ångströms, whereas CAPSIF:G uses the coordinates without any loss of spatial information. CAPSIF:V encodes the initial ~1.4M feature input to a lower dimensionality of a 512-feature vector to encode the entire structure, whereas CAPSIF:G lifts the data from an N_{res} x 30 to a higher dimensionality of N_{res} x 64. CAPSIF:V has ~102M parameters and CAPSIF:G has ~236K parameters, reflecting how graph-based methods capture the spatially equivariant information in fewer parameters. One characteristic of using the voxel representation is that the grid contains voxels with the protein and the voxels outside the protein, including binding pocket cavities, whereas the graph representation only contains the protein. The voxel network reasoning over the surface pocket volume may be the key factor for improved carbohydrate-binding residue prediction.

Building on this initial screen, future studies could focus on improving the CAPSIF data representation for improved accuracy and extending these models to predict which carbohydrate monomer a residue most preferentially binds as well as whether the protein is a carbohydrate-binding protein. In the future, the dataset could include oligomeric structures that bind carbohydrates at the oligomeric interface. Further, one could improve model performance by leveraging homologous structures with data splits across families. Although lectins are well known for carbohydrate binding, some protein families, such as G protein coupled receptors (GPCRs) and antibodies, do not exclusively bind carbohydrates.[98,99] Additionally, with our carbohydrate binding site data set, one could test small molecule binding site predictor neural networks like Kalasanty[66] or PeSTo [100] by fine-tuning them for sugars. High throughput methods like these could enable proteomic scale sorting of carbohydrate binding capabilities.

Methods

Dataset

No dataset of nonhomologous bound protein-carbohydrate structures existed that leveraged the total size of the current PDB, so we constructed one. Simply selecting all RCSB [86] structures with carbohydrates gives all docked protein-carbohydrate structures but also inherently returns all glycosylated proteins, glycosylated peptides, as well as all protein structures that use carbohydrates as crystallization agents. We desired to determine all true physiological protein-carbohydrate interactions, so therefore we manually removed nonspecific crystallization buffers or glycoproteins. Next, we removed all proteins with resolution over 3 Å. Then we removed all homologous protein structures over 30% sequence identity to remove all sequentially redundant proteins, only accounting for chain homology and not domain homology. Some structures containing sugars with modified monosaccharides and cyclic carbohydrates were unreadable in the PyRosetta[101] software and therefore additionally removed.

The final dataset consists of 808 structures, with a split of 521 training structures, 125 validation structures, and 162 test structures. Each structure has one or more of the following carbohydrate monomers: glucose (Glc), glucosamine (GlcNAc), glucuronic acid (GlcA), fucose (Fuc), mannose (Man), mannosamine (ManNAc), galactose (Gal), galactosamine (GalNAc), galacturonic acid (GalA), neuraminic acid (Neu)/sialic acid (Sia), arabinose (Ara), xylose (Xyl), ribose, rhamnose (Rha), abequose (Abe), and fructose (Fru). We split the training, validation, and test sets pseudo-randomly to ensure equal representation of all carbohydrate species in each split. The numbers of each monomer per structure and Dice coefficient for each carbohydrate monomer type and each protein family in the test set from CAPSIF:V are included at our github link (Data

Availability). For all following work, we defined a carbohydrate-interacting residue as residues with any heavy atom that is within 4.2 Å of a carbohydrate heavy atom.

CAPSIF:V Data Processing

Convolutional neural networks are not rotation invariant, and so data augmentation by rotations improves their performance.[102] Therefore, we augmented the input data for CAPSIF:V during training to overcome the rotational variance. Each time a structure was used in training, it was rotated in Cartesian space by a random angle in {-180°,180°} around an axis defined by a randomly-chosen residue's location and the protein center-of-mass. With the random rotation for each epoch, the network learned approximately 1,000 different orientations of each structure in the data set. If the protein was too large for the grid size, the protein was split into separate grids and run separately (about 22% of the training points).

Neural Network Architectures

Features

Due to the small dataset size of 808 structures, we chose residue-level representations instead of atomistic. We assigned all residue information to the Cβ atom of each residue because the position of the Cβ is similar in *apo* and *holo* states.[90] The features are listed in **Table 2.3**. The SASA, hydrophobicity, H bond donor/acceptor indices were calculated using pyRosetta,[101] and aromatophilicty was indexed by Hirano and Kameda.[92]

CAPSIF:Voxel

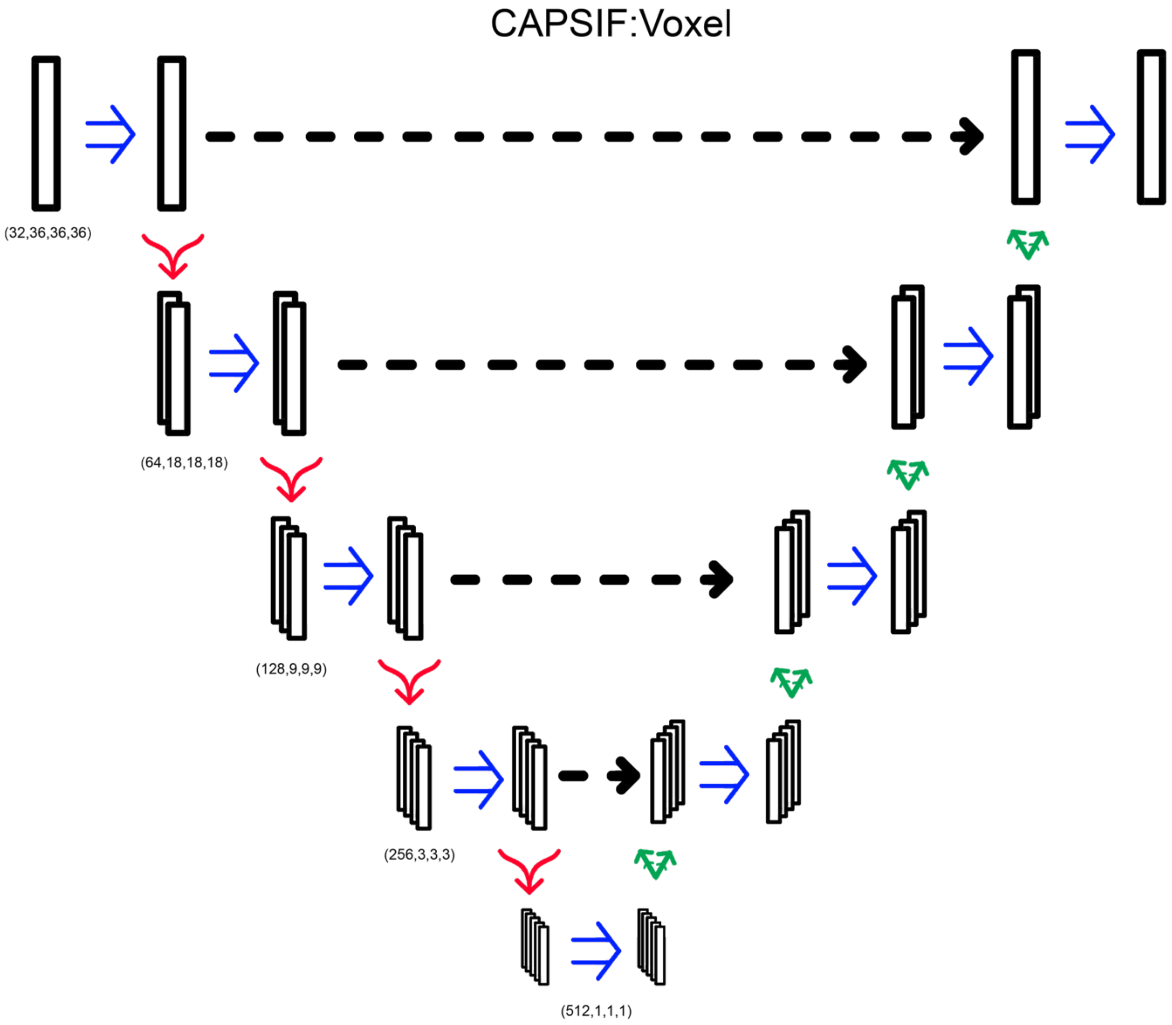


Figure 2.7: CAPSIF:V architecture. Blue arrows indicate a double convolution, red arrows indicate an encoding layer, and green arrows indicate a decoding layer.

CAPSIF:V utilizes a UNet architecture, encoding and decoding the input structure to predict carbohydrate binding residues with residual connections. CAPSIF:V inputs a grid of 36 x 36 x 36 voxels with each voxel representing 2 Å x 2 Å x 2 Å. We input a tensor of size (28,36,36,36), with the 28 features from Table 2.3, where orientation is the normalized components of the Cα to Cβ bond vector. All voxels without a Cβ within are input as zero-vectors.

Table 2.3: List of features and the associated encoding size used for both CAPSIF models.

Feature Type	Encoding Size
Amino acid (one-hot)	20
SASA	1
Hydrophobicity	1
Aromatophilicity	1
H Bond Donor/Acceptor	2
Orientation (Voxel only)	3
Torsion (Graph only)	4

CAPSIF:V contains an embedding layer and 9 convolutional blocks where 4 blocks encode the structure, 1 block forms the bottleneck, and 4 blocks decode the structural information. The embedding layer lifts the 28-channel input into a 32-dimension space. Each block has a double convolution, performing the following methods twice: 3D convolution, with the same number of input channels as number of output channels, (5x5x5) kernel with a stride of 1 and padding of 2, a batch normalization layer, and rectified linear units (ReLU) activation function. In addition, each encoding block also has a MaxPooling layer to double the size of the channels (32,64,128,256,512) while reducing 3D cubic voxel number (36,18,9,3,1). Each decoding block first concatenates the results of the encoding layer of the same size and then performs a double convolution and a 3D-transposed convolution operator, reducing the number of channels (256,128,64,32) while increasing the 3D cubic voxel number (3,9,18,36). After the 9 blocks, there is a single convolutional layer condensing the input channels (32) into a single output channel, which is then followed by a sigmoid activation function to output the probability that the voxel contains a residue that binds a sugar (Figure 2.7). CAPSIF:V contains 102,676,001 parameters.

CAPSIF:V was trained for 1,000 epochs with a learning rate of 10^{-4} and batch size of 20 grids using the Adam optimizer[31] with the loss function $L = 1 - d$, where d is defined by $Eq\ 2.1$.

In optimizing CAPSIF:V, we explored several model variations. We tested various combinations of 3x3x3, 5x5x5, and 7x7x7 convolutional filters. We used four convolutions per layer instead of the double convolution in the primary model. Further, we used larger voxel grid sizes (72x72x72 instead of 36x36x36) with another decoding/encoding layer in the UNet architecture. We also attempted different configurations of skip connections, such as UNet++.[103] These models required slower learning rates and showed slower convergence with no improvement in prediction quality than the presented model. The best model from validation accuracy is detailed above.

CAPSIF:EGNN

CAPSIF:G is an equivariant graph neural network[35] that performs convolutions on each node (chosen as each Cα for glycine and Cβ for all others). Graph edges are connected between neighbors (defined as all other nodes` within 12 Å) and the edge attribute is the distance between node Cβ atoms. In addition to the features used in CAPSIF:V, we include a torsional component in the node features as the sine and cosine of the φ and ψ angles of each residue (Table 2.3).

CAPSIF:G first lifts the 29-feature input node into a 64-dimension space. The 64-feature vector, alongside the edge features (distances) is then input to eight consecutive equivariant graph convolutional layers (EGCLs).[35] Each EGCL contains an edge multilayer perceptron (MLP), a node MLP, a coordinate MLP, and attention MLP. The edge MLP consists of two blocks of a linear layer and a rectified linear units (ReLU) activation function. The node MLP consists of a linear layer, a ReLU activation layer, and linear layer. The coordinate MLP contains a linear layer, a

ReLU activation layer, and a linear layer. The attention MLP contains a linear layer and a sigmoid activation function. All layers input and output a 64-feature vector. Finally, CAPSIF returns the embedding to a 29-feature vector per node, adds the initial input features to the final vector, performs batch normalization, and then uses a sigmoid activation function to output a probability of carbohydrate binding of all residues. CAPSIF:G contains 236,009 parameters.

This model was trained for 1,000 epochs with a learning rate of 10^{-4} and batch size of one protein using the Adam optimizer[31] with the loss function $L = 1 - d$, where d is defined by (*Eq 2.1*).

In optimizing CAPSIF:G, we explored changing the number of graph convolutional layers and the latent space dimensionality. We tested the number of layers (L = 4,6,8,16) and used the different dimensionalities of the latent space (d = 16,32,64). The best performing model is detailed above.

Data Availability

The datasets and the code for each model are available for non-commercial use at `https://github.com/Graylab/CAPSIF.`

Appendix

Dataset

The dataset is composed of the following monomers: glucose (Glc), glucosamine (GlcNAc), glucuronic acid (GlcA), fucose (Fuc), mannose (Man), mannosamine (ManNAc), galactose (Gal), galactosamine (GalNAc), galacturonic acid (GalA), neuraminic acid/sialic acid (Neu/Sia), arabinose (Ara), xylose (Xyl), ribose, rhamnose (Rha), abequose (Abe), and fructose (Fru). There are either no training structures or very few for Fru, ManNAc, Abe, Rha, ribose, GalA, and GlcA. Although some have a high average test Dice similarity coefficient, CAPSIF may not accurately predict protein residues that bind those carbohydrate species well. Finally, CAPSIF:Voxel does not perform well on predicting residues that bind Neu and Fuc, likely due to their 9-carbon structure and (*L*) conformation, respectively, as well as GlcNAc.

Determination of Data Representation

For voxel locations, we compared three representation choices, (1) α carbon (Cα), (2) β carbon (Cβ), or (3) Cα and Cβ positions for the location of voxels. We trained and tested each of these models as described in the Methods. We compared the Dice coefficient, sensitivity and positive predictive value to determine which representation performs best (Figure 2.8, Table 2.4). The Cβ-only representation has an average test Dice coefficient of 0.551, with the Cα representation having a test Dice coefficient of 0.545, where when both the Cα and Cβ are included together in the representation, the architecture has an average test Dice coefficient of only 0.528.

Finally, we further included orientation information of the residues themselves by concatenating the unit vector of the Cα to Cβ bond to the Cβ only representation. This representation had an average test metric of 0.597 (Cβ: Cα → Cβ vec) (Figure 2.7, Table 2.4). This method performed the best of all three representations, having the largest coverage and highest average test metrics. For these reasons, we chose Cβ: Cα → Cβ as our representation of coordinates and orientation for CAPSIF:V.

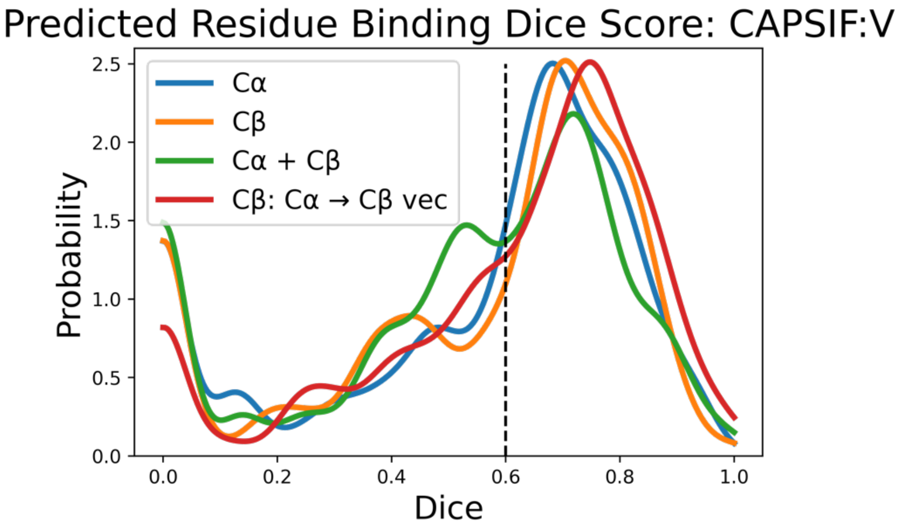


Figure 2.8: Test Dice coefficient assessment for different representations with CAPSIF:V architectures: Blue shows a Cβ representation including a normalized vector for alpha carbon (Cα) to Cβ, orange shows only a Cβ representation, green shows Cα representation, and red shows Cα and Cβ representation with all voxels.

Table 2.4: Performance for each CAPSIF:V model. Dice coefficient is defined by (Eq 1); PPV and Sensitivity are same as Table 2.1.

Voxel **Representation**	**Dice**	**PPV**	**Sensitivity**
Cβ	0.551	0.563	0.583
Cα	0.545	0.535	0.620
Cα + Cβ	0.528	0.555	0.554
Cβ: Cα → Cβ	**0.597**	**0.598**	**0.647**

Next, we investigated CAPSIF:G node representations, with the architecture described in Methods. We constructed the following variants: Cβ nodes with φ and ψ angles, Cβ and N, Cα, and C backbone nodes (and one-hot encoding for atom type, without φ and ψ angles). The Cβ only node representation performed the best with a Dice coefficient of 0.543. Further, Cβ takes a fraction of the time for predictions compared to the backbone due to graph construction time, therefore we chose the CAPSIF:G to be the Cβ model (Figure 2.9, Table 2.5).

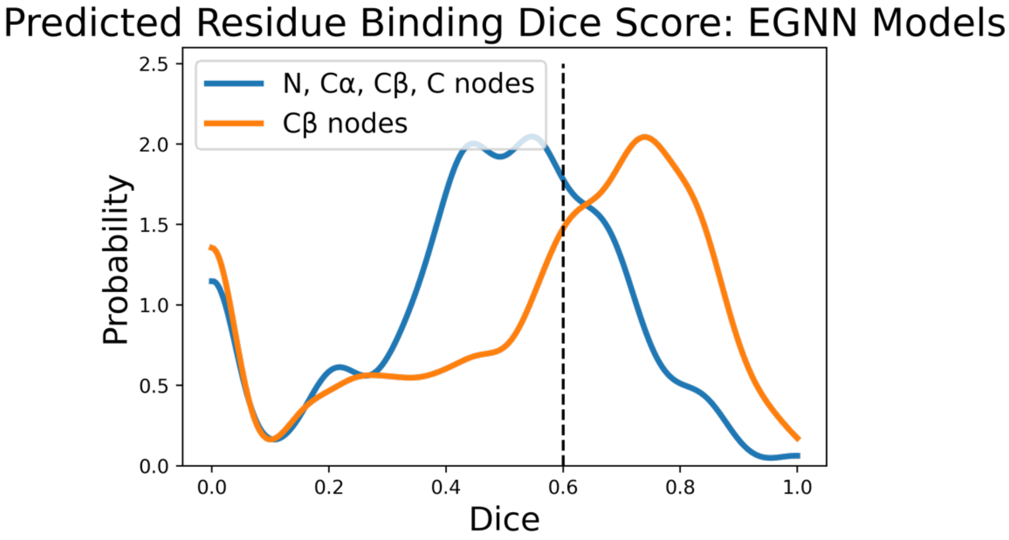


Figure 2.9: Test Dice coefficient assessment for different representations with EGNN architectures. Blue shows all backbone atoms node representation, orange shows a Cβ node representation.

Table 2.5: **Performance for EGNN model node representation**. Dice coefficient is defined by (Eq 1); PPV and Sensitivity are same as Table 1.

EGNN **Representation**	**Dice**	**PPV**	**Sensitivity**
Cβ	**0.543**	**0.541**	**0.590**
Backbone	0.458	0.396	0.647

Random Assignment of Carbohydrate Binding Regions

As a control, we compared CAPSIF to a random baseline. For example, for 200 amino acids with a 5.0% positivity rate, we randomly select 10 residues as a true label (sugar binding) and computed the Dice similarity coefficient (*Eq 1.1*). Using 1,000 trials for an endoglucanase (6GL0), which has 331 total residues with 14 that experimentally bind carbohydrates, we observe a theoretical maximum Dice coefficient at approximately 0.08 when all residues are predicted as carbohydrate binders. At a rate of 5%, we observe a mean Dice coefficient of 0.046, where CAPSIF:V predicts that protein with a Dice coefficient of 0.963 (Figure 2.10A).

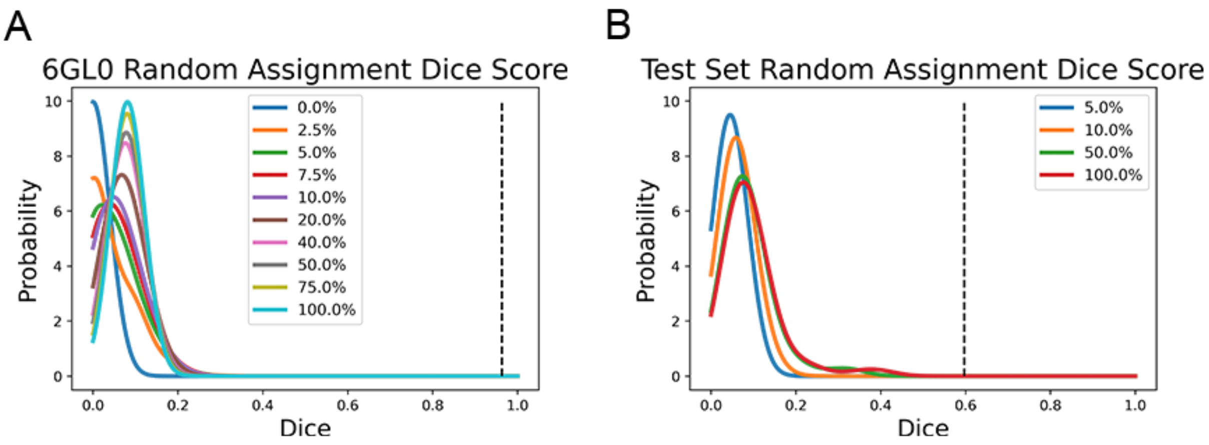


Figure 2.10: Dice coefficient assessment with random assignment smoothed with a kernel density estimate with bandwidth h = .04. **(A)** Dice evaluation of random assignment of an endoglucanase (6GL0). **(B)** Dice evaluation over entire test set.

The dataset has, on average, 5.16% of protein residues bind carbohydrates. With random assignment over the entire dataset, random assignment at 5.16% yields an average 0.046 Dice score, where CAPSIF:V outperforms random assignment by over 12-fold at an average 0.593 Dice (Figure 2.10B).

Determination of CAPSIF probability threshold

To determine the best probability cutoff value for the final activation function, we altered the threshold on the test dataset (Figure 2.11). CAPSIF:V differs minimally for all thresholds while CAPSIF:G negatively correlates with increasing threshold and drops more sharply after a cutoff of 0.6. For both architectures we chose a threshold of 0.5.

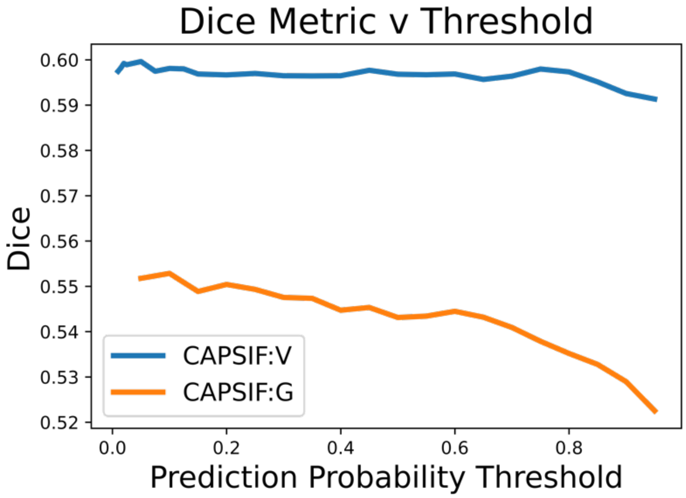


Figure 2.11: Test Dice coefficient assessment for CAPSIF architectures for various thresholds for the final sigmoid activation function. Blue represents CAPSIF:V, orange represents CAPSIF:G.

Comparison of Dice and DCC metrics

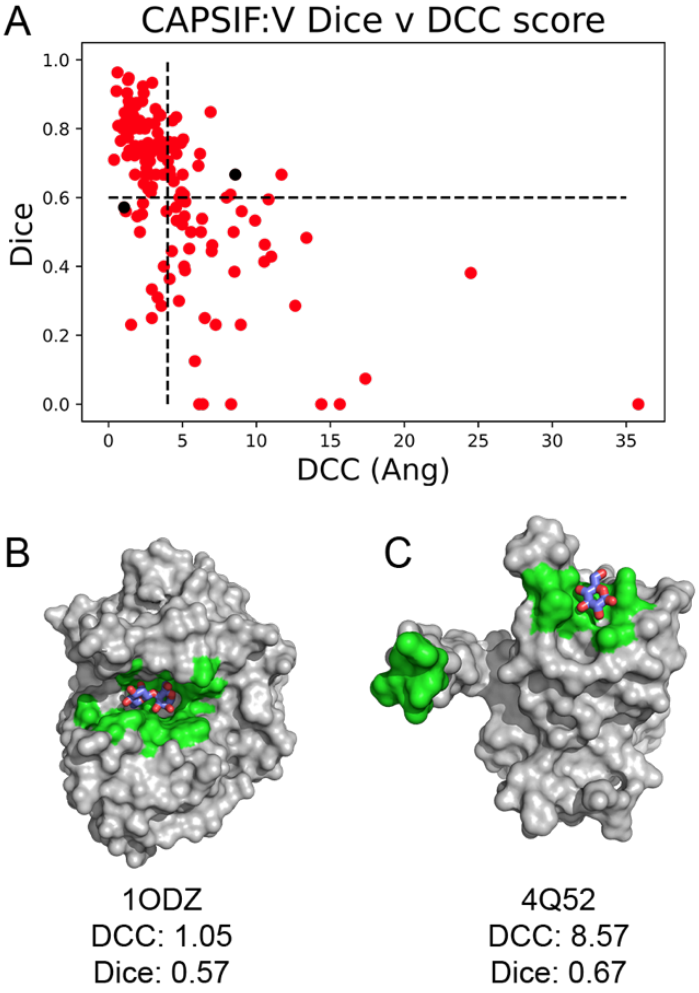


Figure 2.12: Comparison of Dice score and DCC. **(A)** Per-target comparison of Dice and DCC for CAPSIF:V predictions on the test set. CAPSIF:V predictions (green) on **(B)** endo-1,4-β-mannosidase 1ODZ and **(C)** *C. pinesis* DSM 2588 (4Q52) (gray).

Figures comparing CAPSIF:Voxel and CAPSIF:Graph predictions

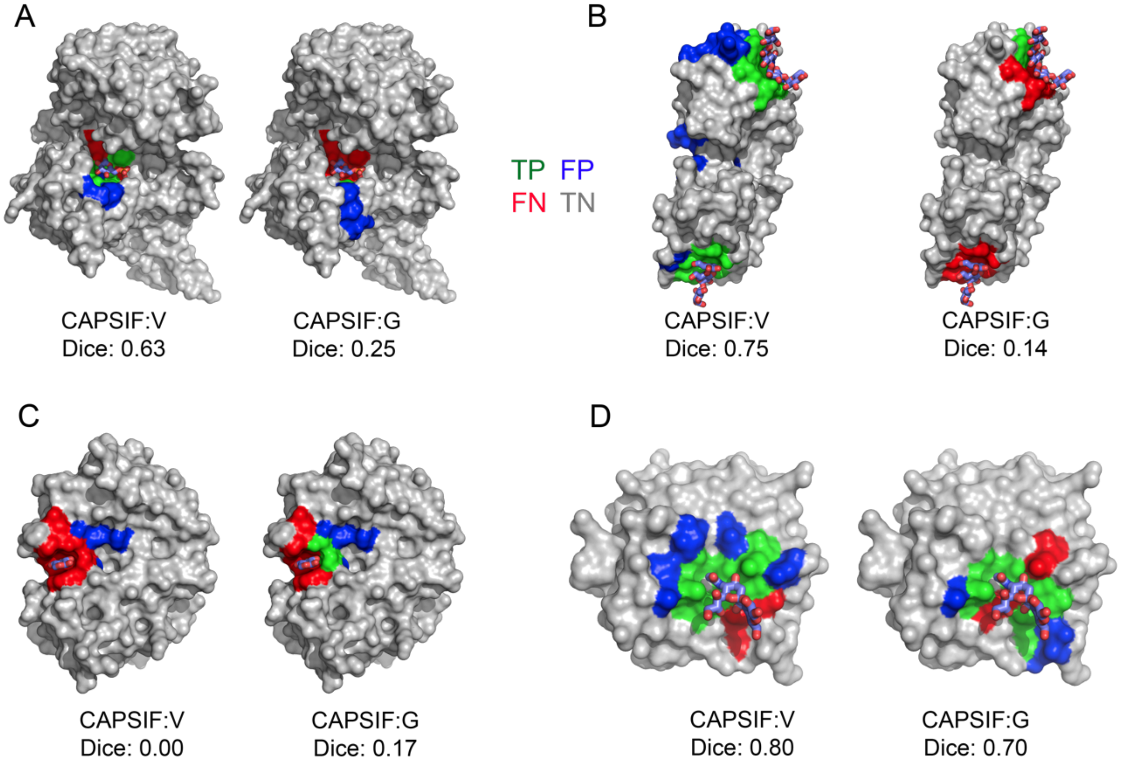


Figure 2.13: Prediction of carbohydrate binding sites on a protein surface using CAPSIF:V and CAPSIF:G. **(A)** Glc 6-phosphate dehydrogenase (PDB:5UKW), **(B)** streptococcal virulence factor (PDB:2J44), **(C)** MCR-1 catalytic domain (PDB:5ZJV), and **(D)** CBM40 (PDB:6ER3). Residue labels - green: true positive, blue: false positive, red: false negative, gray: true negative, cyan: bound carbohydrate; Dice coefficient is defined by eq (2.1) and DCC is distance from center to center of the predicted binding regions.

Comparison of RCSB and AF2 predicted structures

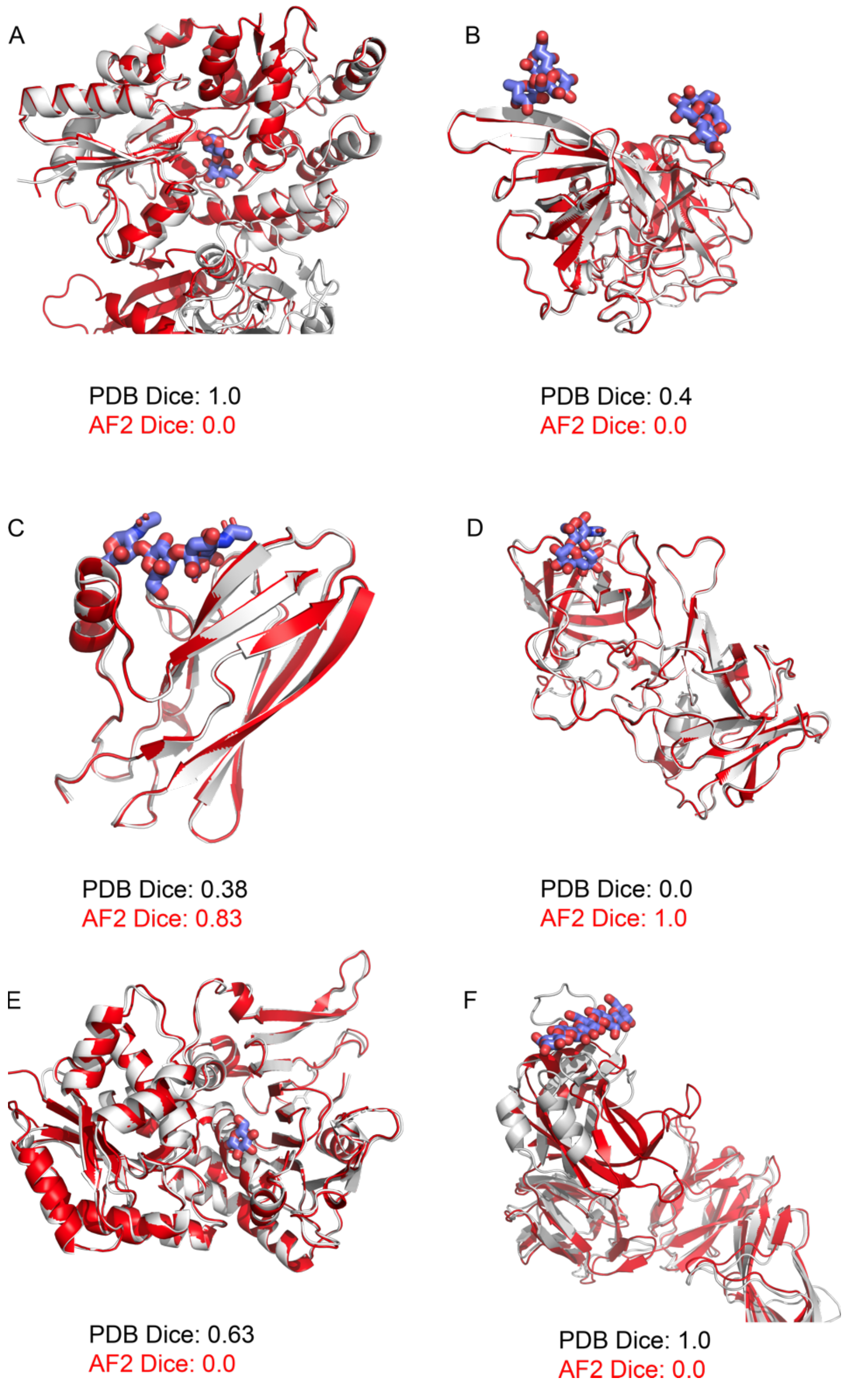


Figure 2.14: AF2 structure prediction (red) of carbohydrate (purple) binding proteins compared to experimentally solved structures (white); **(A)** SUFU (PDB:4BL8) **(B)** E. coli aminopeptidase N (PDB:4XO5), **(C)** GspB siglec domain (PDB:5IUC), **(D)** GII.13 novovirus capsid P domain (PDB:5ZVC), **(E)** Glc 6-phosphate dehydrogenase (PDB:5UKW), and **(F)** surface GBP B (PDB:6E57). Dice coefficient is defined by eq (2.1).

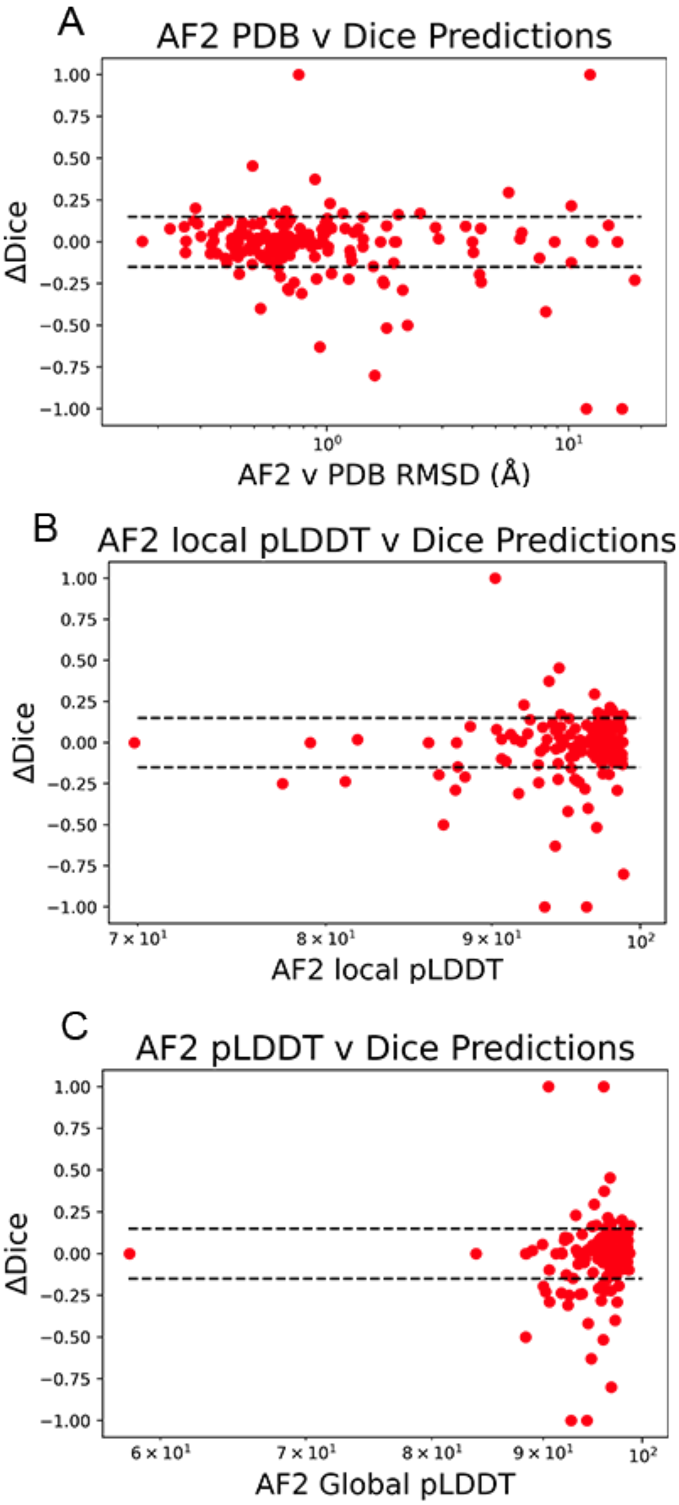


Figure 2.15: CAPSIF:V accuracy is not correlated with AF2 accuracy or confidence. CAPSIF:V predictions on AF2 structure prediction metrics of carbohydrate binding proteins compared to RCSB structures. Change in Dice metric (ΔDice = AF2 Dice – RCSB Dice) compared to **(A)** the total Cα RMSD (log scale), **(B)** Local average pLDDT score of the carbohydrate binding region, and **(C)** total average pLDDT score of the entire structure.

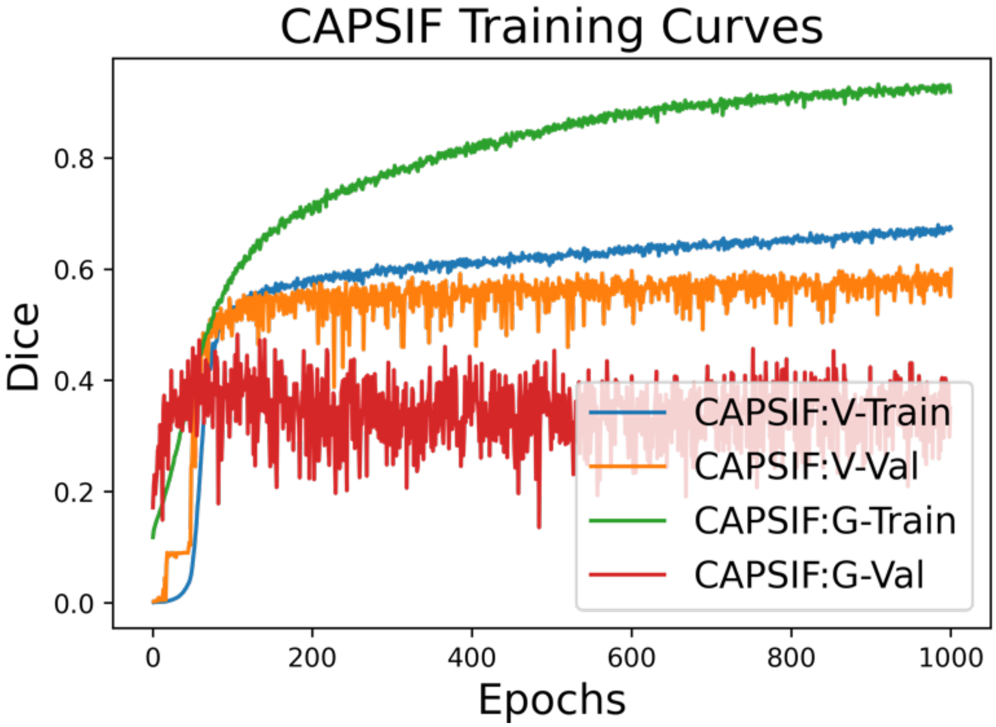


Figure 2.16: Training and validation curves of both CAPSIF models.

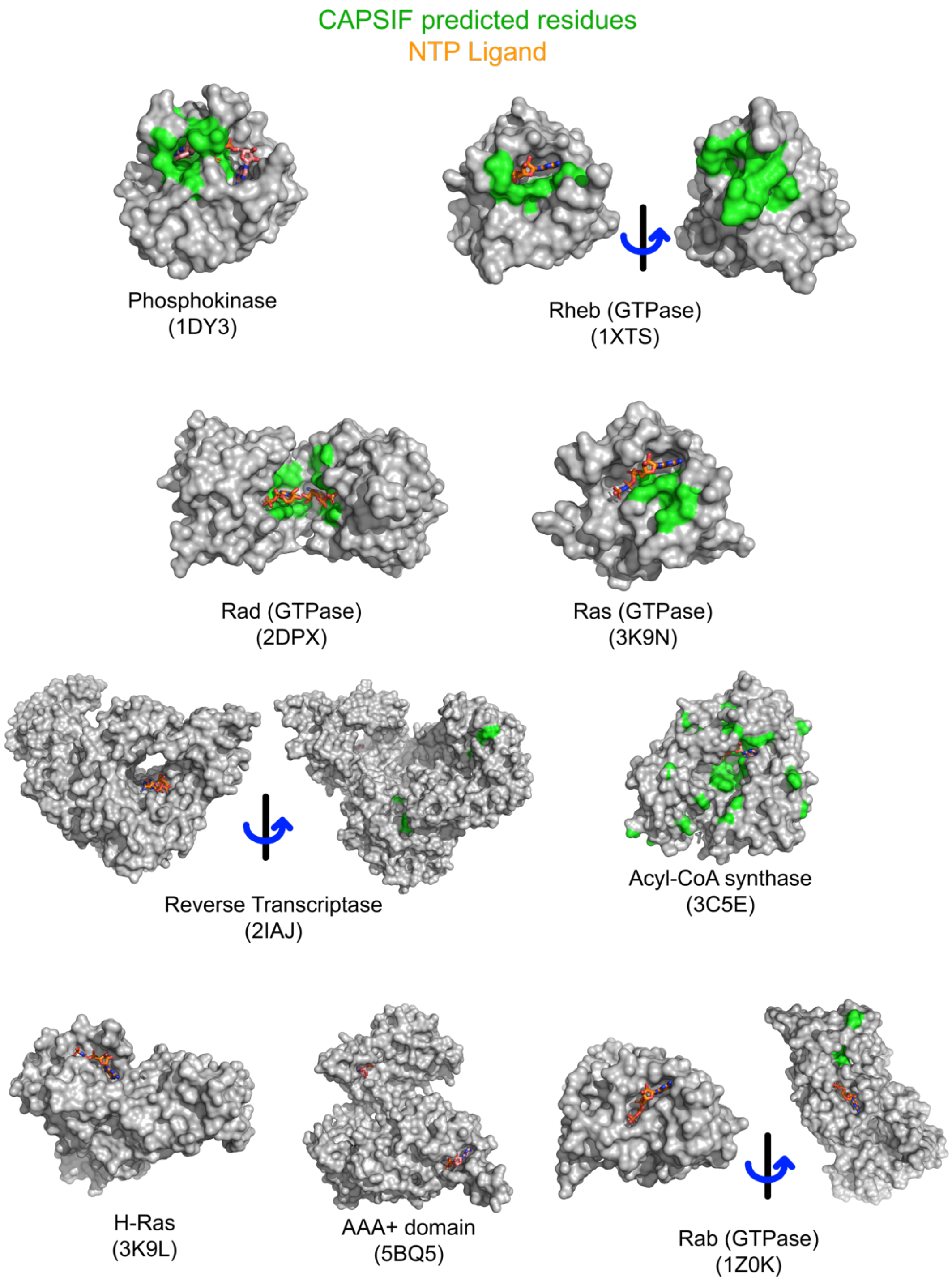


Figure 2.17: Prediction of CAPSIF:V and CAPSIF:G on ATP and GTP-binding proteins. Both CAPSIF models predict similar regions on the ATP/GTP binding proteins, but only qualitatively capture the binding region of the phosphokinase, Acyl-CoA synthase, Rad, and Ras.

CAPSIF cannot distinguish carbohydrate binding proteins from non-binding proteins

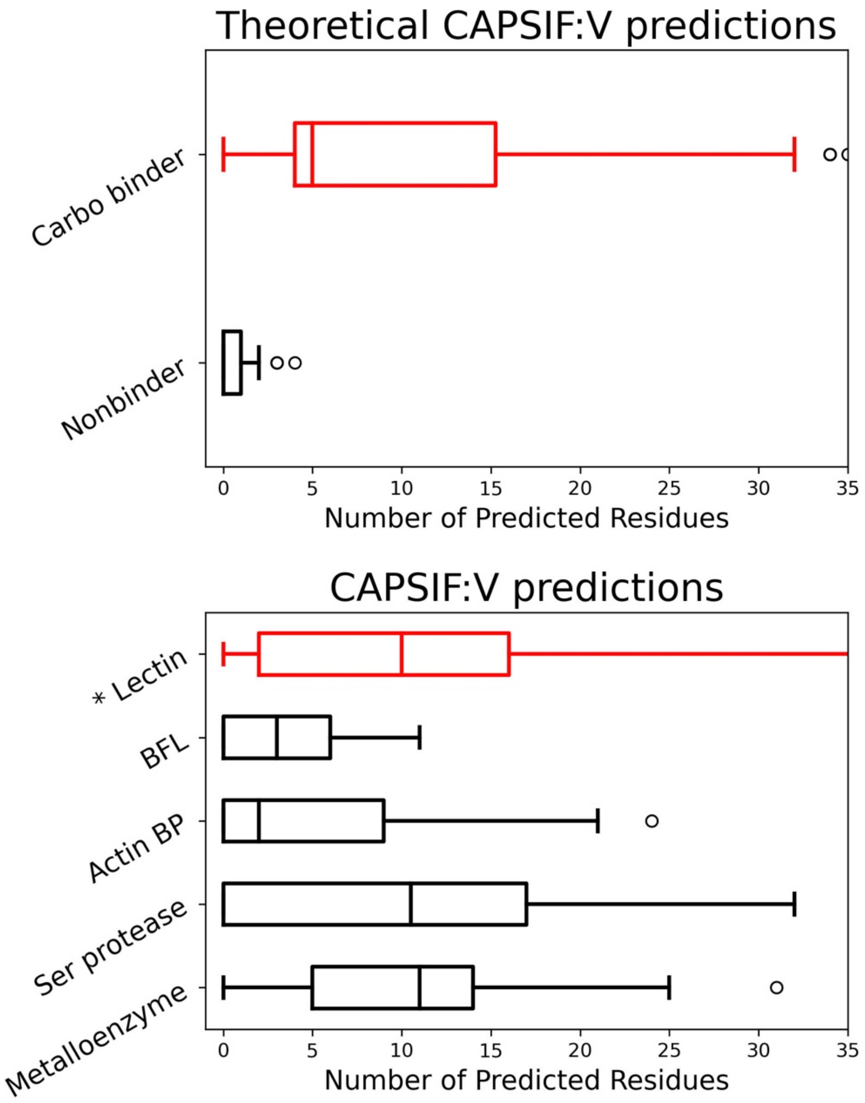


Figure 2.18: Theoretical CAPSIF:V predictions compared to experimental predictions. (Top) Ideal number of residues predicted by CAPSIF:V for carbohydrate binding proteins (red) compared to non-carbohydrate binding proteins (red). (Bottom) CAPSIF:V predictions on protein structures from the families of lectins (red), BFLs, actin binding proteins (BP), serine (Ser) proteases, and metalloenzymes.

Chapter 3

PiCAP: Predictions from Deep Learning Propose Substantial Protein-Carbohydrate Interplay

Adapted from: Canner, S. W., Schnaar, R. L. & Gray, J. J. Predictions from Deep Learning Propose Substantial Protein-Carbohydrate Interplay. *bioRxiv*, (2025).

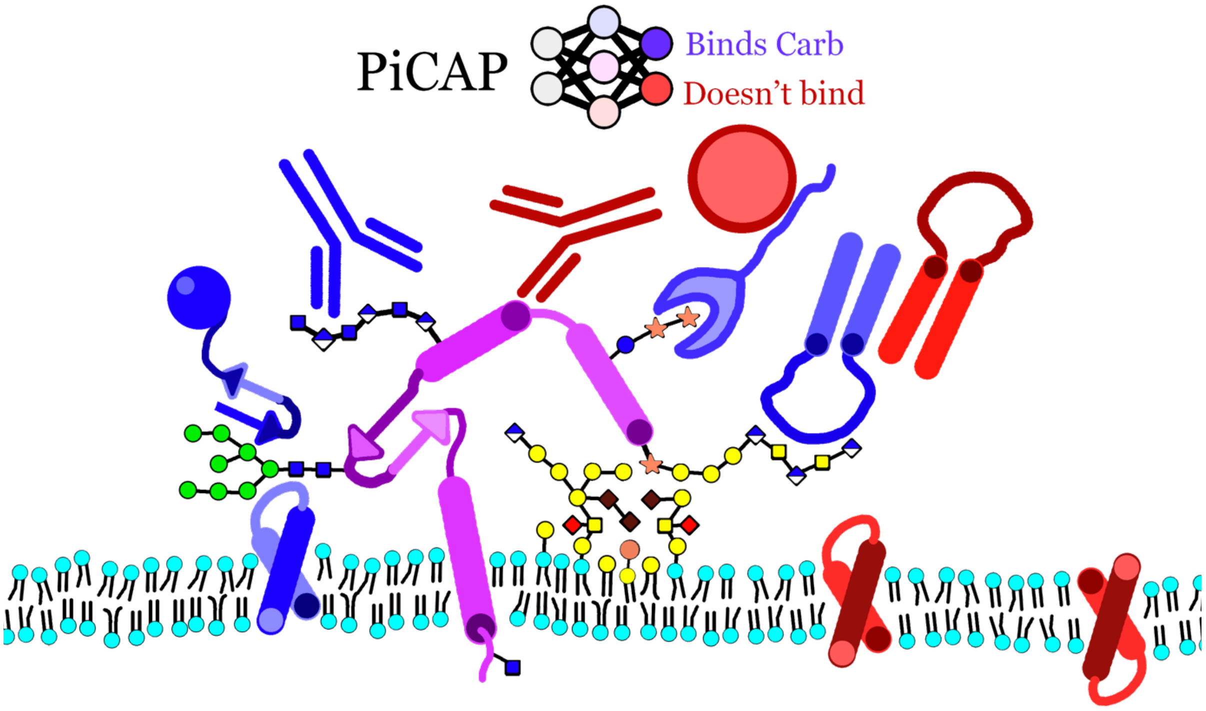


Figure 3.1: Protein interaction of CArbohydrates Predictor (PiCAP) identifies the likelihood a protein non-covalently interacts with carbohydrates.

Attribution of credit: SWC (Conceptualization, Writing, Methods, Analysis, Figures), RLS (Writing, Analysis), JJG (Conceptualization, Writing, Analysis)

Overview

Although I previously developed a method for identifying residues implicated in protein-carbohydrate interactions, researchers are unable to determine whether a protein binds to carbohydrates. I therefore sought to address the grand challenge to identify protein-sugar interactome in an organism. Direct experiments would require extensive libraries of glycans to definitively distinguish binding from non-binding proteins. Computational screening of proteins for carbohydrate-binding provides an attractive and ultimately testable alternative. Current estimates label 1.5 to 5% of proteins as carbohydrate-binding proteins; however, 50-70% of proteins are known to be glycosylated, suggesting a potential wealth of proteins that bind to carbohydrates. I therefore developed a neural network architecture, named **P**rotein **i**nteraction of **Ca**rbohydrates **P**redictor (PiCAP), to predict whether a protein non-covalently binds to a carbohydrate. I trained PiCAP on a novel dataset of known carbohydrate binders and selected proteins that I identified as likely *not* to bind carbohydrates, including transcription factors, cytoskeletal components, and small-molecule-binding proteins. PiCAP achieves a 90% balanced accuracy on protein-level predictions of carbohydrate binding/non-binding. Using the same dataset, I developed a model named **Ca**rbohydrate **P**rotein **S**ite **I**denti**f**ier 2 (CAPSIF2) to predict protein residues that interact non-covalently with carbohydrates. CAPSIF2 achieves a Dice coefficient of 0.57 on residue-level predictions on our independent test dataset, outcompeting all previous models for this task. To demonstrate the biological applicability of PiCAP and CAPSIF2, I investigated cell surface proteins of human neural cells and further predicted the likelihood of three proteomes, notably *E. coli, M. musculus,* and *H. sapiens*, to bind to carbohydrates. In the human proteome, PiCAP predicts that 75% of extracellular and cell surface proteins are putative carbohydrate binders. The PiCAP predicted binders are highly enriched for functions and processes such as growth factor receptor binding, inflammatory responses, and cell-cell adhesion.

Significance Statement

The totality of the protein-sugar interactome remains elusive, in part due to the inability to test a proteome versus a glycome in a high throughput manner. Here I show the first high-throughput methodology to predict protein-carbohydrate interactions at proteomic scales by using structural and sequence information. To provide a better grasp of the role of carbohydrates in cellular functions, I created a computational method to predict the carbohydrate binding profiles of the human, mouse, and *E. coli* proteomes.

Introduction

In mammalian biology, carbohydrates are studied as two distinct families of molecules that are the focus of two disciplines. As metabolic precursors, from food or stored reserves, polysaccharides and monosaccharides (primarily glucose) are transported into and stored in the cytoplasm where they are subject to catabolic transformations to produce energy.[5] In contrast, distinct covalent groupings of varied monosaccharide building blocks covalently bound to proteins (glycoproteins and proteoglycans) and lipids (glycolipids) are relatively stable and are abundant at the cell surface and in the extracellular milieu.[6] A notable exception is O-GlcNAcylation, the reversible covalent attachment of the single sugar N-acetylglucosamine (GlcNAc) to serines and threonines of many cytoplasmic and nuclear proteins.[104]

Like their structures, the functions of carbohydrates are diverse. Among other functions they play essential roles in metabolism, they contribute to protein, cell and tissue structures, and they engage in molecular recognition upstream of cell-cell adhesion and cell regulation.[6] Most of these functions involve engagement of glycans by proteins.[105] These protein-carbohydrate interactions have predominantly been studied using chemical and biochemical methods, despite recent advances in the computational field. [106]

With the advent of the third generation of machine learning and large datasets, many novel algorithms have been created to better understand biophysical phenomena.[2,107] Deep learning methods have recently overtaken most traditional algorithms for all biomolecular methods on all biopolymers, including prediction of protein structure, protein-small molecule interactions, and *de novo* protein design.[38,42,43,108,109] Two of the largest computational steps in biophysics made recently are the releases of AlphaFold 2 (AF2)[38] and ESM[39]. AF2 revolutionized the protein structure landscape by creating a public, easily accessible, and accurate method for protein structure prediction. AF2 additionally predicted the protein structures of 48 organisms that are publicly accessible.[108] ESM (named for evolutionary scale modeling) revolutionized protein sequence representations through its transformer architecture, with ability to richly encode the language of protein sequences.[39]

Leveraging recent computational advances, scientists are beginning to explore the breadth of protein-carbohydrate interactions. I expect some of these protein-carbohydrate interactions to be involved in carbohydrate metabolism, some in intermolecular recognition and regulation of protein functions (e.g. O-GlcNAc), and others in cell adhesion and cell regulation. The goal of this work is to use computational advances to predict the protein-sugar interactome: all proteins amenable to carbohydrate binding, in its broadest interpretation. Conventionally, researchers have focused on the carbohydrate binding protein family of lectins, which excludes enzymes, carriers, or native sugar sensors. Here I computationally explore carbohydrate-binding proteins without excluding them based on function; I expect to capture proteins across metabolic, structural and molecular recognition functions. Although this approach is agnostic to carbohydrate species; as discovery progresses, our work may be expanded to provide further sub-characterization to identify the functional definitions of carbohydrate binding.

In the previous chapter, I developed a dataset and two models, named CArbohydrate Protein Site IdentiFier (CAPSIF):Graph and CAPSIF:Voxel, to predict the protein residues involved in noncovalent carbohydrate-protein interactions.[106] CAPSIF:V and CAPSIF:G are trained and tested on the same dataset and use the same residue level encodings, but CAPSIF:V encodes proteins onto a 3D voxelized grid with a UNet architecture whereas CAPSIF:G uses an equivariant graph neural network (EGNN) message passing framework; CAPSIF:V slightly outperformed CAPSIF:G by all measured metrics.

Since both CAPSIF models were released, two similar models have been created. Bibekar et al. released Protein Structure Transformer (PeSTo)-Carbs, which uses a geometric transformer architecture to predict residues involved in protein-carbohydrate interactions.[110,111] PeSTo-Carbs employs a query-key-value attention mechanism with message passing across atoms that are then pooled for residue-wise predictions.[110] He et al. released DeepGlycanSite, which leverages a geometric message-passing architecture to predict a glycan binding site in both the case of a known ligand and an unknown ligand.[111] PeSTo-Carbs modestly outperforms both CAPSIF models on all reported metrics, whereas DeepGlycanSite focuses on binding to nucleotide structures as compared to carbohydrate-only polymers.

Most carbohydrate-protein interaction algorithms rely on multiple datasets to extract experimental coordinates for prediction.[106,110] Currently the standard protein-carbohydrate dataset is UniLectin; however, UniLectin focuses only on proteins in the lectin family and thereby does not include other carbohydrate binding proteins.[112] Recently, DIONYSUS was released detailing an immense set of experimental carbohydrate binding proteins with non-covalently bound carbohydrate as well as glycosylated proteins.[113]

Since experimentally solved structures can be difficult to obtain, especially in the presence of a carbohydrate ligand, some datasets of sequences exist that identify carbohydrate binding proteins. The Carbohydrate Active enZYmes (CAZY) dataset identifies sequences of catalytically active proteins that act on glycosidic bonds.[8] LectomeXplore is a dataset that identifies known lectins, their associated structures (if known) and potential lectins as identified by sequence similarity via a hidden Markov model (HMM).[16] Rather than limiting their work to known lectins, Zhang et al. developed high throughput experiments with a ganglioside probe that identified 873 putative human proteins that likely interact with gangliosides.[17] These works are limited by their scope, requiring either specific protein families or specific carbohydrate species to interact.

Here, I present novel frameworks to both predict whether a protein can bind to carbohydrates and where on that protein the carbohydrate binds, entitled **P**rotein **i**nteraction of **CA**rbohydrate **P**redictor (PiCAP) and **CA**rbohydrate **P**rotein **S**ite **I**denti**F**ier **2** (CAPSIF2). Both models leverage a large dataset with two training stages, first using all small molecule binding interfaces and then fine tuning with carbohydrate-specific data. I assess the ability of these models in their tasks. I then validate PiCAP against the work of Zhang et al.[17] and identify potential outliers in their dataset. Finally, I use these models to make the first prediction of carbohydrate binding proteins, and residues of these proteins, of three proteomes. While these first proteome-wide predictions are likely noisy, they define the broad scope of the problem and invite refinement by future experimental and computational methods.

Results

NoCAP: a novel non-binder dataset

Many datasets exist for protein-carbohydrate interactions, with the most notable being DIONYSUS (Table 3.1). However, there is no dataset of proteins that do *not* bind to carbohydrates; therefore, I developed a novel dataset consisting of proteins known to bind carbohydrates and proteins that likely do not bind carbohydrates based on biophysical intuition. Although the non-binder dataset is likely mildly contaminated with some currently unknown carbohydrate binding proteins, I believe this dataset to be generally representative of proteins that do not bind carbohydrates. I denote this novel combined dataset as Nonbinder and binder of CArbohydrate Protein interactions (NoCAP) (Table 3.1). In addition, I created a subset of NoCAP, named DIONYSUS-Residue (DR) as all binding proteins in NoCAP with a bound ligand, retaining the DIONYSUS name as most protein structures were retrieved from the DIONYSUS dataset (Table 3.1).

Table 3.1: **Experimental structural datasets**. Columns 2 and 3 indicate dataset inclusion in the NoCAP or DR datasets.

Dataset	NoCAP	DR	Description	n proteins
CAPSIF[106]	✓	✓	Bound protein-glycan complexes	802
TS 90	✓	✓	Test set for Pesto-Carbs; a subset of the CAPSIF test set	90
DIONYSUS[113]	✓	✓	Bound protein-glycan complexes	5,461
UniLectin[112]	✓		Lectin structures and sequences	2,881
ProGen[114]	✓		*De novo* designed lysozymes	69
Designed-NB[115]	✓	✓	List of crystal and complementary designed non-binders	2,800
SAbDab[116]	✓	✓	Crystalized antibodies to their antigen (filtered)	2,925
PDB-Bind[117]	✓	✓	Small molecule binders (filtered to remove carbs)	17,191
PDIDB[118]	✓		DNA-binding proteins (putative non-carb binders)	922
Manual selection	✓		Biophysically putative non-binding proteins (fatty acyl synthases, cytoskeletal components, flippases, ion transporters, ribosomal proteins)	606

To provide a more comprehensive view of the physiologic characteristics of protein-carbohydrate interactions, I curated the datasets to contain complete structures (e.g. not separate chains) and incorporated both ligand-bound *holo* and unbound *apo* forms. While DIONYSUS already aggregated several sources including Unilectin3D and SabDAb, I additionally use Unilectin3D to obtain unbound *apo* structures of lectins and leverage SabDAb to access antibody-protein and antibody-nucleic acid complexes. In total, NoCAP contains 30,429 structures, with 9,509 carbohydrate binding proteins and 21,339 putative nonbinders. The DR set, which is that of bound protein-carbohydrate complexes, contains 6,263 structures in total.

CAPSIF2 outcompetes all previous models identifying carbohydrate-binding residues

I constructed an equivariant graph neural network (EGNN) named Carbohydrate Protein Site IdentiFier 2 (CAPSIF2) leveraging the same general architecture of my previous work CAPSIF:Graph (CAPSIF:G). Although CAPSIF:G underperformed CAPSIF:V, I chose the EGNN architecture because it is scalable to proteins of any size, while CAPSIF:V is limited by the size of the underlying convolutional voxels. Although the dataset of this work (6,724 protein structures) is substantially larger than my previous work (~800 protein structures), there is still an intrinsic data imbalance in that most protein residues (~95%) do not bind carbohydrates. To address this, I once again leveraged the Dice loss (Table 3.2) to emphasize the residues that bind carbohydrates (see methods).

Table 3.2: **Average metrics for each deep learning architecture on test sets.** Dice coefficient is described as 2TP / (2TP + FP + FN), where TP, FP, and FN are the counts of the true positives, false positives, and false negatives, respectively. MCC is the Matthews correlation coefficient. Boldface indicates the best performance for each metric.

Model	DR Dice	DR MCC	TS 90 Dice	TS 90 MCC
CAPSIF2	**0.573**	**0.574**	0.616	0.607
Pesto-Carbs	0.493	0.492	**0.638**	**0.624**
CAPSIF:V	0.226	0.202	0.608	0.622

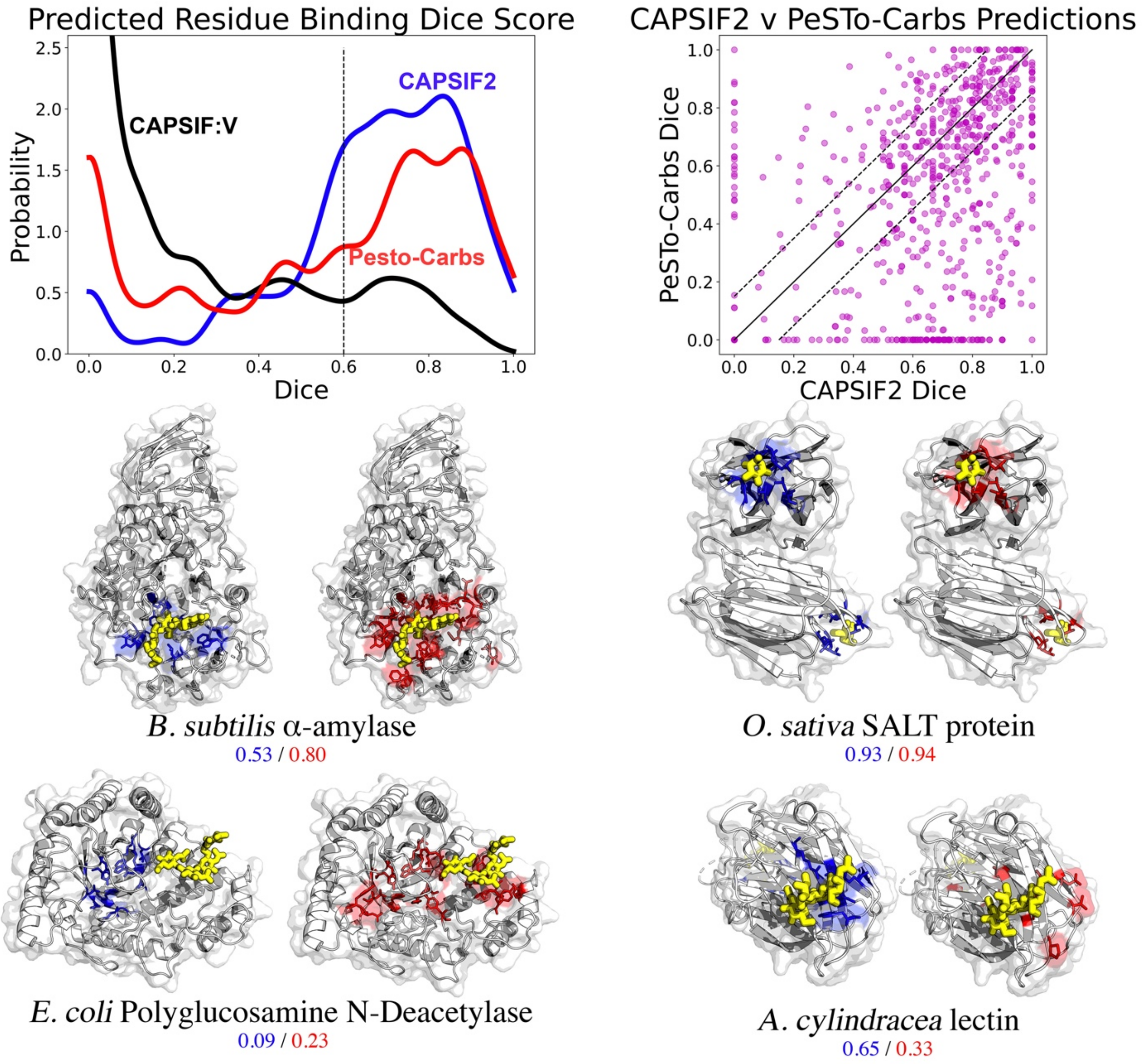


Figure 3.2: Comparison of CAPSIF2 and PeSTo-Carbs on residue-wise prediction tasks. (A) Distribution of Dice coefficient across prediction targets (proteins) for CAPSIF2 (blue), PeSTo-Carbs (red), and CAPSIF:V (black) on the DR test set. Densities smoothed with a Gaussian kernel density estimate (KDE, bandwidth h = 0.04) . (B) Per-target comparison of CAPSIF2 to PeSTo-Carbs. (C) Side-by-side comparison of carbohydrate (yellow) bound proteins (gray) predictions by CAPSIF2 (blue, left) and PeSTo-Carbs (orange, right) on *B. Subtilis* α-amylase (1BAG), *O. sativa* SALT protein (5GVY), *E. coli* poly-β-1,6-N-acetyl-D-glucosamine N-deacetylase C-terminal domain (4P7R), and galactose binding lectin (5XFD). Per-target Dice coefficients shown below.

In Figure 3.2 and Table 3.2, I compare my results to PesTo-Carbs[110] and my previous model CAPSIF:V[106]. On the TS-90 test set, CAPSIF2 achieves 0.616 Dice and 0.607 MCC metrics and PesTo-Carbs outcompetes my model on this test set with a 0.638 Dice coefficient and 0.624 MCC (Table 2). Contrarily on the DR test set, CAPSIF2 achieves 0.573 Dice coefficient and 0.574 MCC, while PesTo-Carbs only achieves 0.493 Dice and 0.492 MCC metrics. On a per target basis, CAPSIF2 performs greater than 0.15 Dice better than PesTo-Carbs on 40% of targets and PesTo-Carbs performs greater than 0.15 Dice than CAPSIF2 on 15% of targets (Figure 1B).

I further show the results of specific targets in Figure 3.2C. In most of these cases, PesTo-Carbs and CAPSIF2 can successfully find the binding region, with varying accuracy; however, they both appear to fail on some targets, such as N-acetyl-D-glucosamine N-deacetylase. This target notably has an observable pocket in the center of the structure, which CAPSIF2 and PesTo-Carbs incorrectly identifies as the binding region, wherein the experimentally solved oligosaccharide is proximal to the pocket.

PiCAP accurately predicts carbohydrate binding and non-binding on experimental structures

Leveraging the same foundational network structure as CAPSIF2, I constructed the equivariant graph neural network (EGNN) named Protein interaction of Carbohydrate Predictor (PiCAP) with five additional layers to yield a single value prediction of whether a protein does or does not bind a carbohydrate. PiCAP assesses the spatial relationship of residues over an increasing context window, pooling the sequence into a fixed size 2D image, and providing a singular classification prediction based on that 2D representation. To my knowledge, PiCAP is the first DL model to assess protein-noncovalent binding of carbohydrates at a protein level.

Table 3.3: Metrics for PiCAP on the NoCAP test set and associated subsets with the number of proteins in parentheses. BACC is balanced accuracy. TPR is True Positive Rate TPR = TP / (TP + FP). TNR is True Negative Rate TNR = TN / (TN + FN).

Test Set	**Accuracy**
NoCAP BACC (4,411)	0.896
NoCAP TPR (2,374)	0.963
NoCAP TNR (2,037)	0.828
TNR Ribosome (7)	1.0
TNR Holdout (92)	0.902
TPR ProGen Lysozymes (69)	0.841
TNR Designed Nonbinders (186)	0.608
Antibody BACC (50)	0.562

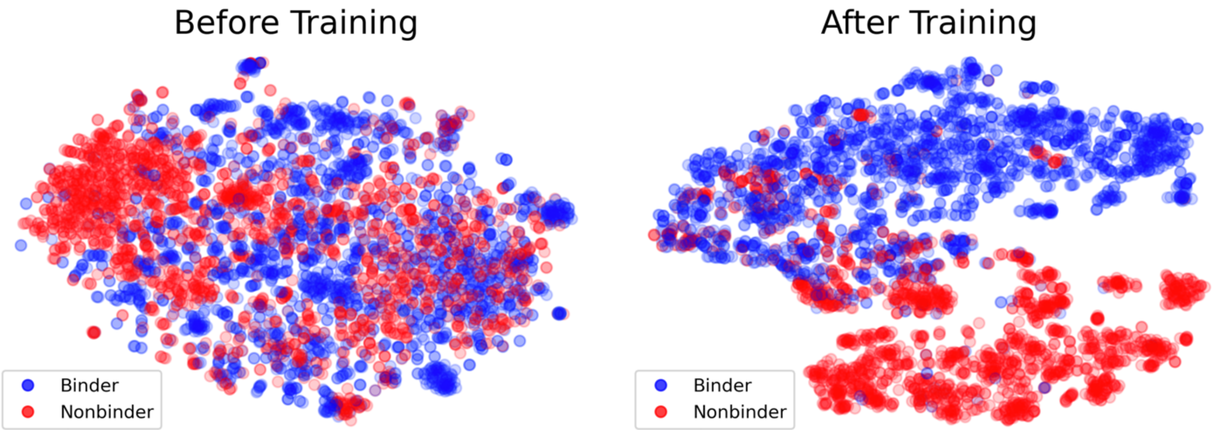


Figure 3.3: T-distributed stochastic neighbor embedding (T-SNE) diagrams of the PiCAP final layer embeddings of the NoCAP test set. (A) The randomly initialized model's final layer output. (B) The final trained model's final layer output.

I tested PiCAP on a holdout set based on sequence similarity, finding that PiCAP achieves an 89.6% balanced accuracy (BACC), with a 96.3% true positive rate (TPR) and 82.8% true negative rate (TNR) (Table 3.3). The ability to separate out carbohydrate binding (blue) and non-binding (red) proteins is further demonstrated in 2D t-distributed stochastic neighbor embedding (T-SNE) plots (Figure 3.3)[119]. Despite my best efforts, I do expect that the nonbinder dataset is likely contaminated with some carbohydrate binding proteins, therefore I must further discriminate PiCAP's ability to predict on specific test set subsets.

When inspecting subsets of NoCAP (Table 3.3), I find PiCAP correctly predicts all the protein chains of the ribosome assembly as non-binders. I further have a holdout set of multiple proteins from various protein families, consisting of fatty acyl synthases, actin, myosin, and flippases, where PiCAP achieves an encouraging 90.2% accuracy on this negative subset. I observed that PiCAP performed well on designed lysozymes from the ProGen language model[114] with an 84.1% accuracy. This high accuracy may be a result from the high redundancy of the

ProGen lysozymes, which span only five families. Contrarily, PiCAP achieves poor accuracy on computationally designed non-binder proteins, these being poor designs regarded as non-binding to the carbohydrate on the designed pocket, with an accuracy of only 60.8%. As a final test, I asked how my model performed on antibodies – specifically to identify antibodies that bind proteins or the glycans of glycoproteins. Of the 50 tested antibody structures, PiCAP achieved a 79% TNR and 33% TPR for a BACC of 56%. The antibodies and designed nonbinders are proteins hypervariably mutated at the binding site for specificity, which has the poorest performance of PiCAP, whereas PiCAP performs encouragingly on more evolutionarily and biologically defined proteins.

PiCAP agrees with LectomeXplore and experimental evidence

The NoCAP dataset for training and testing PiCAP comprises *experimentally* solved structures; therefore, I decided to investigate how the model performs on two datasets of *computationally predicted* structures. The first dataset is LectomeXplore published by Bonnardel et al., which identifies likely lectins across 37,794 organisms using a hidden Markov model (HMM) based on sequence and structural similarity.[16] I also investigated the ganglioside interactome as published by Zhang et al., where they developed a high throughput assay to identify putative human proteins that interact with gangliosides.[17] Both of these datasets have only sequence/UniProt gene IDs, therefore, for input into my algorithm, I used the predicted structures of the AF2 model proteomes,[108] only retaining confident segments of the structure (pLDDT larger than 70).

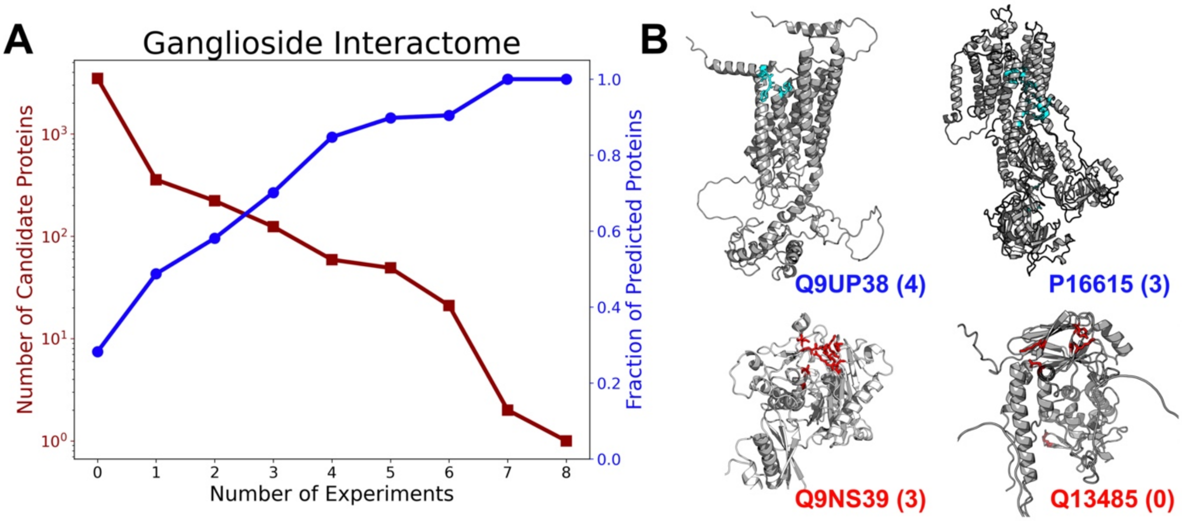


Figure 3.4: PiCAP validation against computational datasets. (A) Plot of Zhang et al. identified proteins across nine experiments alongside the fraction of proteins in each bin predicted as a carbohydrate binder by PiCAP. (B) PiCAP and CAPSIF2 predictions of selected ganglioside interactome proteins. The top row (blue) indicates proteins predicted as carbohydrate binders by PiCAP and bottom row (red) as proteins predicted as non-binders by PiCAP with the number of experiments the protein was identified by in parentheses. Highlighted residues in cyan (top column) and red (bottom column) are the predicted binding regions by CAPSIF2.

LectomeXplore

The most closely related work to PiCAP is LectomeXplore, which identified putative lectins through sequence and structure homology. Unlike LectomeXplore, PiCAP does not limit proteins to be only of the lectin superfamily. I compared the likelihood of all predicted LectomeXplore lectins (greater than 0.25 confidence) present in the AF2 reference proteomes of two model species, *M. Musculus* and *H. Sapiens*, finding the agreement between the available AF2 structures of LectomeXplore and PiCAP to be 100% (225 of 225) for *M. Musculus* and 99.6% (229 of 230) for *H. Sapiens*. Further, PiCAP has a 100% (109 of 109) agreement between the available

AF2 structures on all confirmed human lectins from HumanLectome.[120] These results suggest a strong true positive rate (TPR) of PiCAP on the simplest class of sugar binding proteins.

Ganglioside Interactome

Zhang et al. developed a high throughput method to identify proteins that interact with gangliosides. They created ganglioside probes with photoaffinity tags that covalently linked the probe to nearby proteins, and then they used mass spectroscopy and statistical methods to identify those proteins. They used six different probes in two different cell lines (A431 and SH-SY5Y), and for a total of nine experiments; I filtered the putative proteins by experiment. I selected the top 250 proteins above background from each experiment and removed CRAPome proteins.[121] This identified 873 unique proteins across all nine experiments.[17] As a high throughput method, and the first and largest of its kind, the error rates of their method have yet to be explored and cross-validated across other experimental methods. I therefore will use PiCAP to investigate the putative proteins of the ganglioside interactome work.

Of the 873 identified candidate ganglioside binding proteins, I was able to identify 848 proteins in the AF2 reference human proteome. PiCAP predicts 506 (60%) of these proteins as carbohydrate binders. Further, PiCAP also predicts that 988 of 3,500 putative non-ganglioside binding proteins (28%) as likely carbohydrate binders. Although these numbers at first suggest a substantial disagreement between my works, I observe a strong positive increase in the fraction of proteins predicted as carbohydrate binders compared to the number of experiments that identified a binding protein (Figure 3.4A).

To explore the agreement and disagreement between the experiments, I selected four representative proteins: Frizzled-1 (Entry Name: FZD1; UniProt: Q9ULW2), ATPase sarcoplasmic/endoplasmic reticulum Ca2+ transporting 2 (AT2A2; P16615), Double-stranded RNA-specific editase B2 (RED2; Q9NS39), and mothers against decapentaplegic homolog 4 (SMAD4; Q13485). FZD1 is involved in the Wnt signalling pathway and was identified by four of Zhang et al.'s experiments; PiCAP predicts FZD1 as a carbohydrate binder, and FZD1 was a subject of close scrutiny in the ganglioside interactome work[17]. ATP2A2 is an intracellular calcium/ATP pump and was identified by three experiments and predicted as a carbohydrate binding protein by PiCAP. ATP2A2 has a specific role in ATP-mediated transport of calcium ions and likely little specific affinity for carbohydrates, let alone gangliosides[122]. RED2 is an enzyme that converts adenosine to inosine in pre-mRNA and was identified by three experiments[123]. PiCAP disagrees with the experimental results and predicts RED2 as a non-binder, which could indicate a potential error in the experimental evidence. Finally, SMAD4 is a transcription factor[124], which was identified to not interact with gangliosides in all experiments, where PiCAP agrees and predicts the protein as a carbohydrate non-binder.

PiCAP and CAPSIF2 can predict putative proteome scale interactomes

With PiCAP validated to an acceptable level, I sought to understand the protein-carbohydrate interactome with greater breadth than studied before. I chose three model organisms from the AF2 proteome datasets[108], *E. coli*, *M. musculus,* and *H. sapiens*. Of the 4,363 proteins in the AF2 *E.coli strain K12* proteome (UP000000625), PiCAP yielded predictions on 4,339 accessible proteins and predicted 1,677 (39%) proteins as carbohydrate binders. Of the 21,615 proteins in the AF2 *M. musculus* proteome (UP000000589), PiCAP yielded predictions on 21,304 proteins and predicts 8,177 (38%) proteins as carbohydrate binders. Of the 20,650 proteins in the AF2 *H. sapiens* proteome (UP000005640), PiCAP yielded predictions on 20,067 proteins and predicts 7,029 (35%) proteins as carbohydrate binders (Figure 3.5A). I further provide the results of three additional model species: *Drosophila Melanogaster*, *C. elegans*, and *S. cerevisiae* in the supplemental information, without detailed analysis.

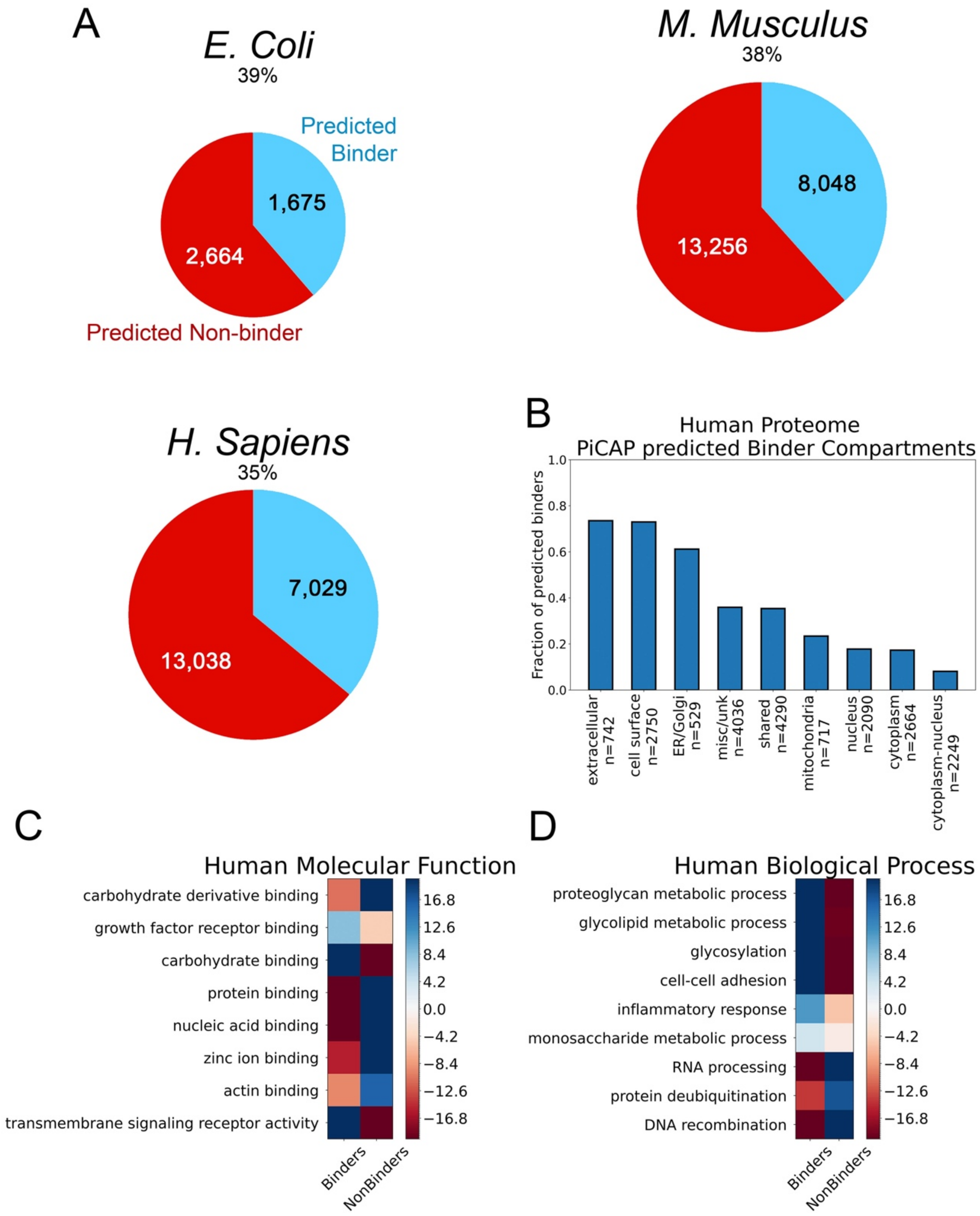


Figure 3.5: PiCAP predictions of proteomes. (A) Comparison of the fraction of proteins predicted as carbohydrate binders by PiCAP across three proteomes. (B) Cellular components of human proteome predicted carbohydrate binding and non-binding proteins (see also Table 3.6). Human proteome statistical tests showing the -log$_{10}$ of the false discovery rate (FDR) and overrepresentation (blue) and underrepresentation (red) for select (C) molecular functions and (D) biological processes. FDR measures the expected proportion of false positives among the list of putative carbohydrate binders and non-binders.

Human Proteome

The primary proteome I analyzed was the AF2 human proteome (UP000005640), which contains 20,650 unique proteins with substantial resolution. PiCAP predicted 7,029, or 34%, of proteins to bind to carbohydrates. For comparison, the total number of lectins identified by LectomeXplore is 230, or 1.1% of the UniProt reference proteome[16], and the number known by CaZy is 349, or 1.7%[8]. In contrast, the number of proteins experimentally identified as likely to bind gangliosides, a unique glycan family, is 873, or 4.2%[17]. To reconcile the differences between my work and the work of many others, I analyzed the subcellular localization, molecular functions, and biological processes of predicted binding and non-binding proteins.

I investigated the subcellular compartments wherein PiCAP predicted carbohydrate binding proteins and non-binders reside based on Gene Ontology (GO) terms (Figure 3.5, Table 3.6). The compartments with the highest fraction of sugar-binding proteins are extracellular (75%), cell surface (75%) and ER/Golgi (50%), which aligns with these being subcellular compartments involved in intercellular communication. The regions mostly devoid of carbohydrates and glycans are the nucleus and cytoplasm; PiCAP predicts 85-97% of these proteins as non-carbohydrate-binding proteins. To further investigate the binding profiles of PiCAP, I queried co-factor binding. A significant portion of carbohydrate-binding proteins depend on a co-factor such as calcium in C-type lectins[6], whereas zinc is more dominantly oriented as a DNA/RNA binding co-factor. [125] PiCAP predicts 58% of calcium-binding proteins and 22% of zinc-binding proteins as carbohydrate binding proteins, indicating that PiCAP does not conflate co-factor binding for carbohydrate binding (Figure 3.7). Additionally, PiCAP predicts 94% of human GO associated carbohydrate binding proteins as binders and 91.5% of GO associated DNA/RNA binding proteins as non-binders, indicating an overall ~93% accuracy, which agrees with the NoCAP dataset evaluation (Figures 3.7, 3.11).

To discern higher specificity from the AF2 human proteome predicted by PiCAP, I selected representative GO terms with PANTHER.[126,127] Using the false discovery rate (FDR) for human proteome related molecular functions, I find PiCAP-predicted binding proteins are significantly overrepresented to act in growth factor receptor binding, transmembrane signaling, and unsurprisingly, carbohydrate binding (Figure 3.5C). Additionally, PiCAP-predicted binders are highly underrepresented for carbohydrate derivative binding but also protein binding, nucleic acid binding, zinc ion binding, and actin binding. Next, I analyzed the biological processes of human PiCAP predicted carbohydrate binding proteins, finding them overrepresented in proteoglycan and glycolipid metabolic processes, cell-cell adhesion, inflammation response, monosaccharide metabolic process, and unsurprisingly glycosylation (Figure 3.5D). Comparatively, I found that carbohydrate binding proteins are underrepresented in RNA processing, protein deubiquitinization and DNA recombination cellular processes. Analysis for *E. coli* strain K12 and *M. Musculus* AF2 proteomes is provided in the supporting information (Figures 3.8-11), showing similar predictions.

Discussion

I have demonstrated (1) an updated protein carbohydrate site identifier CAPSIF2 that outcompetes all current models on a generalized dataset and (2) a novel model named PiCAP that predicts *whether* a protein binds to carbohydrates or not. I validate the models against other models and datasets and applied to proteome scale analysis to garner more information about the protein-carbohydrate interactome.

CAPSIF2 boasts modest improvements in prediction accuracy on the original CAPSIF/TS90 dataset compared to CAPSIF:G and CAPSIF:V, but it underperforms PesTo-Carbs. CAPSIF2 however excels the most at a larger dataset containing ~1k structures with substantially larger sequence variability, outcompeting all tested models. CAPSIF2 leverages a graph neural network operating on residues, using the same foundational approach as CAPSIF:G, while CAPSIF:V used a 3D voxelized CNN approach. PesTo-Carbs also leverages a graph neural network approach; however, it operates at an atom-wise level and only pools to the residue level late in the architecture. These graph architectures however have a similar level of parameters, where CAPSIF:G has 236K parameters, CAPSIF2 has 1.6M parameters, and PesTo-Carbs has 1.1M parameters; while CAPSIF:V has substantially more with 102M parameters.

I believe that the differences in performance are primarily not attributable to the architectures themselves, but rather the datasets. All models perform in a Matthews correlation coefficient (MCC) range from 0.55 to 0.63; thus, I attribute the largest differences to the stochastic training of these models and the slight variations in architectures. Structural protein-carbohydrate datasets are limited currently by the size of the PDB, as these interactions must be strong and stable to observe with experimental methods, where in physiology these interactions are often guided by avidity over affinity and/or enzymatic activity on the carbohydrates themselves. I believe larger datasets is only one part to improving these models, but improving the datasets with manual interrogation of all structures and with the identification of continuous biophysical pockets is necessary to improve the models' performance.

To improve carbohydrate-protein structural datasets and improve the general biological understanding of the carbohydrate-protein interactome, I created PiCAP. PiCAP is the first model of its kind as it predicts the protein-sugar interactome - carbohydrate binding of proteins independent of family/function – whether it be a cell surface protein for adhesion and communication or for metabolic enzymatics. The dataset I used to train PiCAP primarily separates known carbohydrate binders and proteins that are unlikely to bind to carbohydrates physiologically inside the cell – ranging from small molecule binders to cytoskeletal components. Although this approach is imperfect, it is the first attempt of this kind and, while limited by the underlying skewed PDB species distribution toward soluble proteins from human and simple prokarya, it still leverages a generalizable biophysical intuition of the cellular systems. Further augmentation of the NoCAP dataset could improve on the breadth of the training data by using sequence databases in conjunction with structure predictions such as AlphaFold3[42], or Boltz-1[128]. Additional positive sugar binders can come from CAZY[8,9] which contains 5M+ enzymes. Additional negatives can be identified across many species using specific GO terms with known presence in the cytoplasm, nucleus, or nuclear membrane.

Ultimately PiCAP achieves 89.6% accuracy on the experimental NoCAP dataset with 1.8M parameters. PiCAP predicts most subsets of the test set with equivalent accuracy (designed lysozymes, cytoskeletal proteins, flippases, and fatty acyl binding proteins); however, it proves notably worse on designed-non binders and antibodies. The designed non-binders were created using Rosetta, where the binding pocket itself was designed but the remainder of the protein remained untouched[115]. These designs were labeled as non-binders by positive Rosetta binding energy scores – and never experimentally expressed nor tested. In a similar vein, to bind carbohydrates, antibodies use their hypervariable regions which are local regions that undergo somatic hypermutation. My input to the protein is ESM2 embeddings, which uses full sequence context to extract a large 1280-dimensional embedding of each residue. As the ESM2 model is only trained and tested on biological proteins, the signal specificity of the binding pocket sequence of the designed non-binders may be masked by its more evolutionarily conserved residue, leading PiCAP to predict these non-biological proteins as carbohydrate binders. PiCAP studies protein sequence and structural information together, indicating PiCAP as a strong candidate for proteome wide studies of protein-carbohydrate interactions.

Since PiCAP performed well on NoCAP data, I sought to validate the model against other methods that predicted carbohydrate binders: the ganglioside interactome and LectomeXplore. While I saw only 60% of proteins in the ganglioside interactome as positive, after closely evaluating a subset of the data, I reconciled the difference with the error of the high throughput experimental method. Although PiCAP appears to disagree with a good fraction of the ganglioside dataset, it has a strong linear relationship with the high throughput experiments. The more experiments that identified a protein, the higher likelihood that PiCAP predicted the protein as a carbohydrate binder. I further observe strong agreement between LectomeXplore and PiCAP, with an average of 95% agreement across three model species.

With experimental and computational validation, I then leveraged PiCAP against the AF2 proteome datasets. PiCAP predicts 35~40% of all proteins in three biological model species to be carbohydrate binding proteins – the highest prediction to date. As carbohydrates are ubiquitous across all species and are the foundational building block of energy storage and integral to most all extracellular communication, it is unsurprising for such a high fraction of proteins bind to carbohydrates. PiCAP results can be further validated by proteomic evaluation by experiments such as the pull-downs from Zhang et al.[17] or liquid glycan arrays[129,130]. The computational predictions can help elucidate more functionality of proteins and provide a larger context to their roles inside the cell and the suggestion of more protein moonlighting than previously understood.

Despite their biophysical importance in most all cellular functions, carbohydrates remain elusive with few studies determining the exact extent of protein-carbohydrate interactions. My work expands to all proteins/carbohydrates in an agnostic manner that abstains from any limits on protein family or carbohydrate species. I released the results of CAPSIF2 and PiCAP of six model system proteomes for all proteins for open-source scientific use. Additional steps can now be taken for the ultimate goals to design proteins to carbohydrate and glycoprotein targets for therapeutic purposes. Firstly, I encourage the expansion of this work or LectinOracle [48] or GlyNet[85] to predict carbohydrate species to all carbohydrate binding proteins. One simple step would be to predict whether proteins bind to just a specific species of carbohydrate – such as chitins or sialic acids. Another step would a high throughput computational docking of those carbohydrate species to the identified proteins, using CAPSIF2 or PesTo-Carbs[110] or DeepGlycanSite[111] to identify an initial hypothesis to feed GlycanDock[49], or directly *de novo* with programs like DiffDock[43], RosettaFold-All Atom (RF-AA)[131], AlphaFold3[42], or Boltz-1[128] (although there are currently no validation studies testing whether these methods provide high accuracy on carbohydrate-specific docking). In addition, all these methods leverage deep learning techniques. Deep learning methods require multitudes of data, and although I was able to demonstrate impressive results on low accuracy/messy data, I believe a clean dataset is integral and necessary for the future of this field. A better annotated set of proteins that do not bind carbohydrates would be helpful, as well as all structural proteins to have all ligands together, where currently there is a high redundancy in protein structures with slightly different ligands or crystallization techniques, which reduce the accuracy of the test metrics in comparing CAPSIF2 and PesTo-Carbs. I also believe tandem experiments, such as those done by Zhang et al.[17] or selective exo-enzymatic labeling (SEEL) glyco-engineering high throughput methods [22] to validate these models could further demonstrate a larger wealth of carbohydrate binding proteins, alongside their specificity, allowing for further annotation of the genome on a large scale.

In total, I present a novel framework to predict the protein-sugar interactome across any species. Taking a sugar agnostic approach, categorizing glycans and metabolic glucose together, PiCAP accurately predicts evolutionarily conserved proteins as carbohydrate binding proteins with approximately 90% accuracy (Table 3.3, Figures 3.7, 3.11). PiCAP's predictions align with established biophysical principles, indicating that carbohydrate binding is largely absent from the cytoplasm and nucleus and approximately 75% of all cell surface and extracellular proteins bind carbohydrates (Figure 3.5). This suggests that a majority of membrane, surface, and extracellular proteins may predominantly interact with glycans for localization and binding, rather than entirely relying on protein-protein specific interactions. These findings highlight the potential of PiCAP to not only accelerate glycoproteomic research but also refine the understanding of protein function in the broader context of cellular communication and molecular recognition.

Methods

Dataset

Carbohydrate-binding proteins were selected by combining multiple datasets. I selected carbohydrate binding antibodies from SAbDab,[116] all experimentally solved proteins from UniLectin[112] (with and without bound carbohydrates), the CAPSIF dataset,[106] and most notably, the DIONYSUS dataset,[113] which was filtered for only saccharide containing complexes. Further, I included the computationally designed and experimentally viable lysozymes from ProGen,[114] with structures predicted by the Colab distribution of AlphaFold2.[94]

There are several datasets of protein-carbohydrate interactions; however, there is no dataset of proteins that do not bind to carbohydrates, so I constructed one (Table 3.1). In the creation of such dataset, an intrinsic difficulty is that it is not possible to prove that a protein does not bind to a carbohydrate of any kind; therefore, I selected proteins that biophysically have low likelihood to bind to carbohydrates due to their function or location inside the cell. The experimentally solved proteins selected were primarily chosen as small molecule binding proteins, DNA binding proteins, nuclear pore complex proteins, serine proteases, cytoskeletal proteins, aminotransferases, flippases, fatty acid binding proteins, selected antibodies (antibodies), and ribosomal proteins. In addition to these proteins, Luo et al. computationally constructed a dataset of carbohydrate non-binder proteins with the Rosetta software[115].

Small molecule data constitutes the largest portion of the non-binders (~18k pdbs), as I used the PDB-Bind 2020 dataset.[117] Some proteins in the PDB-Bind dataset contain carbohydrates as the ligand, in which case I identified those ligands using PyRosetta[101] and removed them from the non-binder dataset and added them to the binder dataset. Antibodies were selected using the SAbDab dataset by finding all proteins that were bound to proteins or nucleic acids and further filtering to structures not containing any carbohydrates in the structure nor an NX(S/T) motif in the antigen.[116] Ribosomal proteins were selected from the bacterial ribosome structure.[132] The remainder of protein structures were selected by inspection from the RCSB PDB.[133]

After combining the datasets and adjusting for duplicate PDBs across different datasets, the final NoCAP dataset contains 30,849 total unique protein structures. Of these structures, 9,608 bind to carbohydrates, with 6,724 having an experimentally bound carbohydrate. Of the 21,412 non carbohydrate binders, 17,191 have an experimentally resolved small molecule bound to it, leaving 4,221 as nonbinders. To encourage generalizability to minor errors in structure predictions, I also reconstructed the 12,021 shortest sequence proteins of the 30,849 with the Colab implementation of AlphaFold 2[94], where I only kept the 11,042 of predicted structures with a pLDDT greater than 80.

Preprocessing

With the dataset, I desired to leverage both sequence and structural information to predict carbohydrate-binding capabilities of proteins. Family information of sequence similarity can strongly indicate carbohydrate binding capabilities, while structural motifs can be present across protein families for carbohydrate binding, and I desire my method to identify both. I extracted the sequence and the Cβ positions of all protein residues (Cα for glycine) using PyRosetta.[101] Next, I used ESM2[39] to provide a high-dimensional sequence embedding for each protein residue of each protein chain. I labeled protein residues that were within 4.2 Å of a non-covalently bound carbohydrate (or small molecule) as a binding residue.

Most previous work has used single protein chains for protein-carbohydrate predictions[106,110]; however, many proteins only exist in the context of multiple chains. For this reason, I preprocessed all protein structures with all chains in the PDB file, except the initial CAPSIF dataset and antibodies. To limit the redundancy of the training set, I used MMseqs to cluster protein sequences by 60% sequence identity into distinct clusters for training/testing.[86] I then split the clusters into an 80/5/15 train/validation/test, maintaining the same proteins from CAPSIF remain in the same dataset distribution. This left 24,957 structures in training, 1,479 structures for validation, and 4,413 structures for testing.

Secondary validation set

I have a primary dataset of carbohydrate-protein binding; however, I need to demonstrate the ability of PiCAP to predict outside of the crystally solved structures. To do this, I gathered all UniProt[134] accession codes from Zhang et al.[17] and LectomeXplore[16] and matched them to the AF2 publicly accessible organism proteomes. This captured 848 of 878 (97%) of putative ganglioside binding proteins and 3,400 of 4,335 (78%) of non-ganglioside binding proteins. LectomeXplore uses sequence and structural protein information, alongside infectious pathogens that affect these species, and lists all reference sequences and structures (UniProt, ensembl, NCBI, RCSB, etc.) with severe redundancy. Therefore, for a direct quantitative comparison, I therefore used only those that existed singly as UniProt values inside the reference proteomes of AF2. I used the confidence metric of 0.25 for identification of lectins, which yielded 230 human proteins and 225 mouse proteins.

For the proteome analysis, I used the AF2 publicly accessible organism proteomes.[108] AF2 generates structures with an internal confidence metric called pLDDT, where low confidence regions will have pLDDTs under 70. I therefore performed analysis and studies on AF2 protein regions with high confidence, or residues with greater than 70 pLDDT, independent of structural continuity. I applied the analysis to the following model organisms: *E. Coli, M. Musculus,* and *H. Sapiens*. I further provide the results of the full-length sequence, independent of pLDDT in the Appendix alongside the results of three other model organisms: *C. elegans*, *D. melanogaster*, and *S. cerevesia*.

Architectures

I fed the residue coordinates and sequence embeddings into the CAPSIF2 and PiCAP architectures (Figure 1A) into the main model block, which uses a message passing equivariant graph neural network (EGNN) of equivariant graph convolutional layers (EGCL).[35] Each layer sums the outputs of a multilayer perceptron (MLP) that inputs the features of the central node and the features of all of its neighboring nodes and the edge attributes of the neighbors. Following Ingraham et al.[36], the edge attributes are a radial basis function (RBF) of the distance, the orientation, and direction of the neighboring residues.

CAPSIF2, a carbohydrate binding residue predictor, has 12 residual ECGLs with an embedding dimension of 128 (Figure 3.6). The neighborhood context window is fixed at the 16 nearest neighbors. After the graph convolutions, each residue is passed to a two-layer dense decoder, finally outputting the carbohydrate-binding likelihood of each residue. CAPSIF2 contains 1,600,387 parameters.

PiCAP, a predictor of whether a protein binds to carbohydrates, has 12 total residual EGCLs with an embedding dimension of 128 and leverages an increasing neighborhood context window for information propagation, as inspired by PeSTo[100] and PeSTo-Carbs[110]. The first three layers use the 10 nearest neighbors, layers 4 to 6 use the 20 nearest neighbors, layers 7 to 9 use 40 neighbors, and layers 10 to 12 use 60 neighbors. (Figure 3.6). The model specific block is a pooling block that uses an adaptive pool to truncate or slightly expand the size of the protein to a fixed length (150), where the model then uses two convolutional layers and three dense layers to predict the likelihood of a protein to bind to carbohydrates. PiCAP contains 1,798,895 parameters.

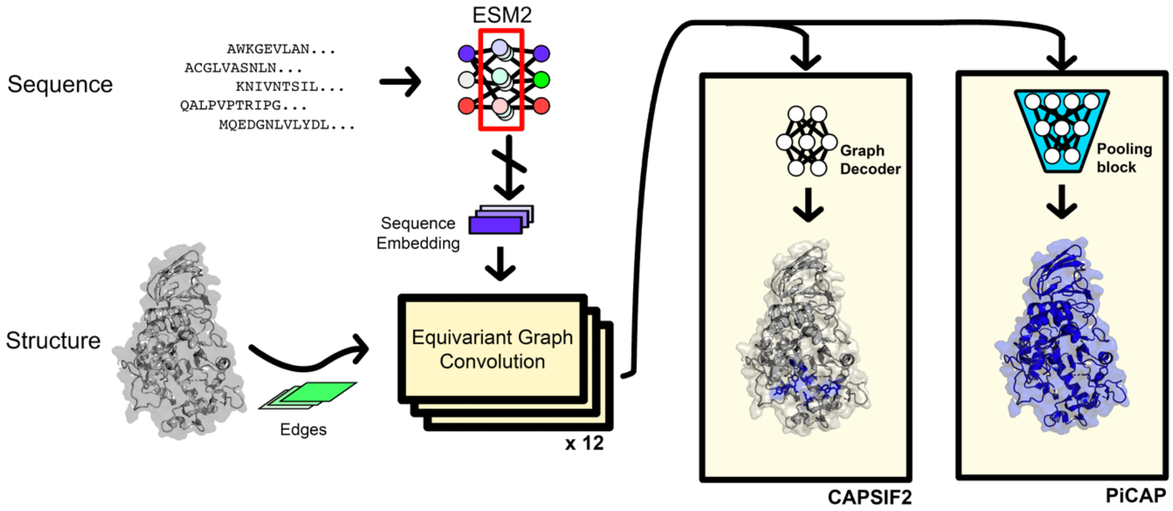


Figure 3.6: Architectures of CArbohydrate Protein Site IdentiFier 2 (CAPSIF2) and Protein interaction of CArbohydrate Predictor (PiCAP).

Training

I trained both models using two cycles: small molecule binding residue prediction and the model specific task (protein or residue level predictions). The first training cycle used the CAPSIF2 base architecture with randomized initial weights $\sim N(0{,}0.02)$ for the residue level prediction. The model was trained for a maximum of 1,000 epochs, with training prematurely stopped once the validation loss did not decrease after 35 epochs. This training cycle had a learning rate of 2 x 10^{-6} and a weight decay of 10^{-7} with the Adam optimizer with the loss function $L = 1 - d$, where d is the Dice-Sorenson coefficient (also known as the F1 score) and a batch size of 1. To improve model generalization, each epoch sampled a single protein from every training cluster available from the small molecule dataset. The smallest 12,000 protein sequences were modeled structurally with the colab distribution of AF2,[38,94] and if the selected protein was available via AF2, I selected the crystal structure 40% of the time and the AF2 structure 60% of the time.

For the second training cycle, CAPSIF2 used the same architecture as the first training cycle and required no randomization. CAPSIF2 was trained only on proteins with experimentally determined carbohydrate binding sites with learning rate of 2 x 10^{-5} and weight decay of 10^{-6} with the Adam optimizer and the same loss function of $L = 1 - d$. Similar to the first training cycle, I randomly selected an available AF2 structure 60% of the time.

For the second training cycle, PiCAP used the weights where available from the first training iteration of CAPSIF2 and randomized weights for the model specific block $\sim N(0{,}0.02)$. PiCAP was trained on the entire training set for binary classification with a learning rate of 2 x 10^{-5} and weight decay of 10^{-6} with the Adam optimizer and binary cross entropy (BCE) loss function. Similar to the first training cycle, I randomly used an available AF2 structure 60% of the time.

Data Availability

Data, code, and datasets are available at Github, where CAPSIF2 and PiCAP can be run at: https://github.com/Graylab/picap. With the assistance of Matt Mulqueen, I further provide a webserver on ROSIE where CAPSIF2 and PiCAP can additionally be run: https://r2.graylab.jhu.edu/apps/index.

Appendix

Dataset Description

I provide all proteome data as an Excel document (xlsx) on the Github (Data Availability). This Excel document contains all prediction information of PiCAP and CAPSIF2 on the AlphaFold 2 proteomic data[108]. Since PiCAP can produce false negative binders, the separate predictions of CAPSIF2 in all cases may assist in hypotheses of known carbohydrate and small molecule binding proteins. In addition, I show predictions on all proteins in the dataset, even in cases with low pLDDT, although the analysis in the primary text only analyzes predictions of proteins with greater than an average of 70 pLDDT. In all sheets, I provide the following columns:

- UniProt Entry
- Common Gene Name (Entry_Name)
- Protein_name
- Gene Ontology terms
- PiCAP prediction on only residues with greater than 70 pLDDT
- CAPSIF2 predicted binding residues on residues with greater than 70 pLDDT

The PiCAP output is a probability value in the range from 0 to 1. In the main text I use the cutoff value of 0.23 to indicate that any protein with predicted probability greater than 0.23 is predicted as a carbohydrate binding protein. The PiCAP prediction may be used as a confidence metric, where the higher probabilities suggest more confidence PiCAP has that the model is a carbohydrate non-binder or binder, respectively.

Model Hyperparameterization

To optimize performance on a neural network, I assessed multiple hyperparameters in both models to achieve their performance. I focused primarily on the following hyperparameters: embedding dimension, *k*-nearest neighbors (knn), and number of layers. For simplicity, I treat each network as a series of four (4) blocks, composed of a certain number of layers where I vary knn per block.

CAPSIF2 parameterization

In my previous work on CAPSIF:G, I used one-hot encodings of amino acid type and biophysical properties with simple edge embeddings. To contain more information, in this work, I altered the node features to ESM2 embeddings and edges. With these input features, I then focused on the size and depths of the network.[106] A full account of all tested hyperparameters is listed below in Table 3.4. I selected CAPSIF2 as the model that performed the best on the DR test set, which was composed of 12 layers with a static number of k nearest neighbors of 16.

Table 3.4: **Performance of various CAPSIF2 models on the Dionysus Residue (DR) test set.** Dice and Matthews correlation coefficient (MCC) are as defined in the main text. Boldface indicates the best performance in each metric. Selected CAPSIF2 model is highlighted in yellow.

Layers per block	**KNN per block**	**DR Dice**	**DR MCC**	**TS90 Dice**
3	6,6,6,6	0.541	0.542	0.533
3	8,8,8,8	0.477	0.484	0.431
3	8,12,16,20	0.391	0.407	0.366
3	6,10,14,18	0.528	0.528	0.572
3	10,20,40,60	0.289	0.312	0.364
3	16,16,16,16	**0.573**	**0.574**	0.616
3	20,20,20,20	0.498	0.496	0.575
4	8,12,16,20	0.566	0.567	**0.639**
4	6,10,14,18	0.408	0.419	0.319
4	8,8,8,8	0.491	0.486	0.548
CAPSIF:V	N/A	0.226	0.202	0.608

PiCAP parameterization

I followed the same methodology as CAPSIF2 to identify the strongest performing PiCAP model parameters. The hyperparameter search is provided below in Table 3.5. The decision on which model performed strongest was less straightforward than CAPSIF2, as all multiple models performed strongly across the NoCAP test set. I selected a model that performed well across most metrics placing just below the top of every other category to encourage generalizability, as some of the top performing models were prone to overfitting and unstable predictions. The chosen PiCAP model consisted of 12 layers, with the knn gradually increasing from 10 to 60 neighbors across the layers.

Table 3.5: **Performance of various PiCAP models on the NoCAP test set.** BACC is balanced accuracy. TPR is True Positive Rate TPR = TP / (TP + FP). TNR is True Negative Rate TNR = TN / (TN + FN).

Layers per block	**KNN per block**	**cutoff**	**NoCAP BACC**	**NoCAP TPR**	**NoCAP TNR**	**Nonbinders TNR**	**Ribosome TNR**	**Holdout TNR**
3	6,6,6,6	0.94	0.85	0.87	0.83	0.624	**1.0**	0.857
3	**8,8,8,8**	**0.33**	0.892	0.964	0.82	0.667	0.857	**0.929**
3	20,20,20,20	0.21	0.856	0.88	0.833	0.683	**1.0**	0.429
3	10,20,40,60	0.23	0.896	0.963	0.828	0.608	**1.0**	0.902
3	6,10,14,18	0.99	0.779	0.927	0.631	0.656	0.857	0.571
4	6,6,6,6	0.77	0.885	**0.976**	0.794	0.731	**1.0**	0.857
4	8,8,8,8	0.19	**0.897**	0.951	**0.842**	0.704	**1.0**	0.857
4	6,10,14,18	0.32	0.861	0.974	0.745	0.134	0.857	0.643
4	8,12,16,20	0.84	0.877	0.954	0.801	**0.785**	**1.0**	0.643

Proteomic Data

In the excel document in the Github (see Data Availability), I provide a list of all proteins from six organisms. Here I list the overall metrics of the three organisms in the excel document that were not discussed in the main text. PiCAP predicts that in *C. elegans* (nematode worm) 9,278 of the 19,227 proteins (48%) bind carbohydrates. PiCAP predicts that in *D. melanogaster* (fruit fly) 5,248 of the 13,351 proteins (39%) bind carbohydrates. PiCAP predicts that in S. *cerevisiae* (yeast) 1,749 of the 5,849 proteins (29%) bind carbohydrates.

Cellular component analysis

To analyze the *E. coli* strain K12*, M. musculus,* and *H. sapiens* AF2 reference proteomes, I employed the use of Gene Ontology (GO) terms and PANTHER.[126,127] For Figure 3.5B, I analyzed the GO terms representative of cellular compartments, I performed a limited search limited to Table 3.6, where any protein observed in multiple of the compartments (excluding just nucleus and cytoplasm) were placed in the "shared" compartment. PANTHER provided the statistical overrepresentation tests and false discovery rates (FDRs) of all cellular compartments, molecular functions, and cellular processes (when the FDR was less than 0.05).

Table 3.6: Simplified cellular compartment GO Terms

Compartment	GO Terms
Cell Surface	cell surface [GO:0009986] plasma membrane [GO:0005886] extracellular space [GO:0005615] extracellular matrix [GO:0031012]
Cytoplasm	cytosol [GO:0005829] cytoplasm [GO:0005737]
Nucleus	nucleus [GO:0005634]
Mitochondrion	mitochondrion [GO:0005739]
ER/Golgi	endoplasmic reticulum [GO:0005783] Golgi apparatus [GO:0005794]

Supplemental Human Proteome Analysis

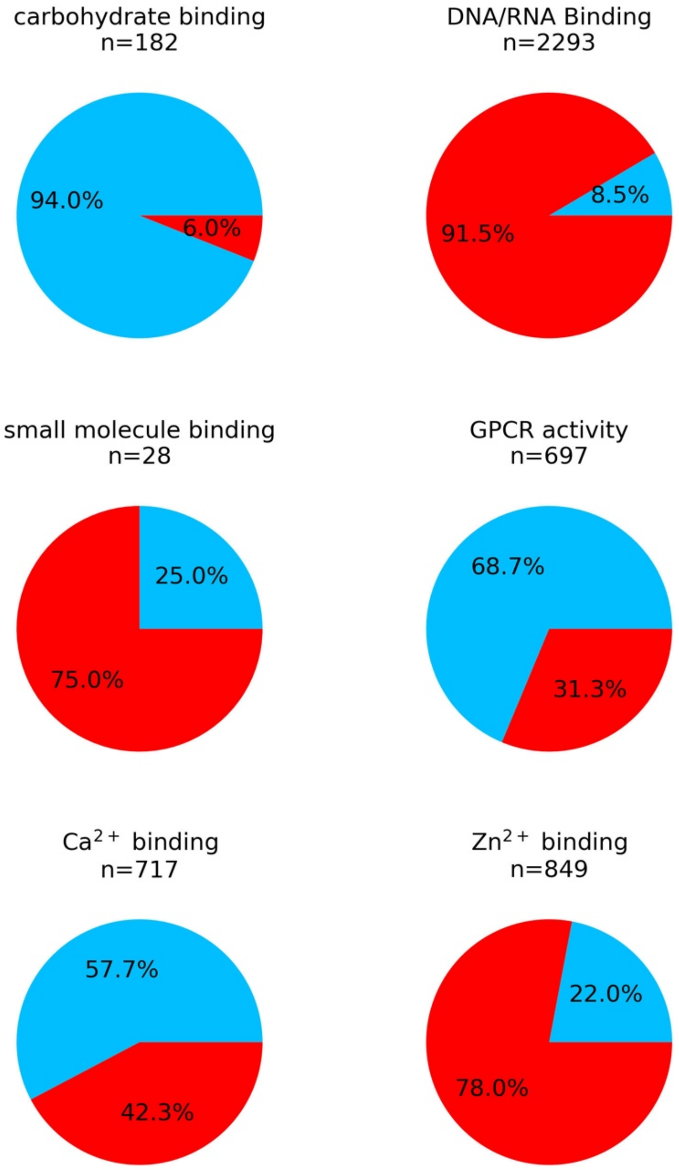


Figure 3.7: Human carbohydrate binding protein functionality. Percentage of proteins with known binding functions predicted as carbohydrate binding (blue) and non-carbohydrate binding (red) proteins for sets of proteins with Gene ontology terms for carbohydrate binding [GO:0030246], DNA and RNA binding [GO:0003677, 0003723], small molecule binding [GO:0036094], GPCR Activity (G protein-coupled receptor activity [GO:0004930]), Calcium binding (calcium ion binding [GO:0005509]), and (Bottom Right) Zinc binding (zinc ion binding [GO:0008270]).

To assess the overall accuracy of PiCAP, I identified the percentage of proteins with GO terms associated with carbohydrate binding, DNA/RNA binding, small molecule binding, G protein coupled receptor (GPCR) activity, calcium ion binding, and zinc ion binding (Figure 3.7). Of the 182 proteins with a GO term for carbohydrate binding in the human proteome, PiCAP predicts 94% of these proteins as carbohydrate binding proteins, indicating a 6% false negative rate.

Here, I defined nucleic acids as non-carbohydrates, and PiCAP identifies 91.5% of DNA/RNA binding proteins as carbohydrate non-binders. PiCAP predicts 8.5% of DNA/RNA binding proteins as carbohydrate binding; where several nucleic acid binding proteins are known to bind carbohydrates, such as DNA polymerase I (such as in PDB 1NK4), so those predicted binders cannot be completely ignored as false positives. There are limited proteins with small molecule binding GO terms associated (28), and PiCAP predicts only 25% of these proteins to bind carbohydrates, which could include small molecules with hydrated ring structures, mimicking carbohydrate epitopes. Additionally, zinc is an ion commonly associated with nucleic acid binding with zinc finger motifs; however, zinc is established in other pathways like neuron excitability. PiCAP predicts 22% of known zinc ion binding proteins as carbohydrate binding proteins, and 78% as carbohydrate nonbinders.

I further assessed calcium ion binding, where calcium is ubiquitous across many cellular processes from muscular contraction to being a secondary ligand necessary for C-type lectin binding. In NoCAP, there are 2263 structures containing calcium, where 946 (41%) of those proteins bind carbohydrates. In the training set specifically, there are 1619 proteins that have calcium ions present and only 614 (38%) of those bind carbohydrates. On the human proteome, PiCAP predicts 58% of calcium binding proteins as carbohydrate binders and 42% as carbohydrate non-binders, indicating that PiCAP effectively distinguishes C-type lectins and calcium cofactor-carbohydrate binding proteins from other calcium binding proteins.

Finally, I investigated GPCRs, where proteins with GPCR activity are integral across many unrelated intercellular communication pathways, where PiCAP predicts 69% of proteins with GPCR activity as carbohydrate binding proteins, suggesting a wealth of agonists and antagonists being carbohydrates (or carbohydrate-like molecules with hydrated rings) may be critical in GPCR binding.

E. Coli Proteome analysis

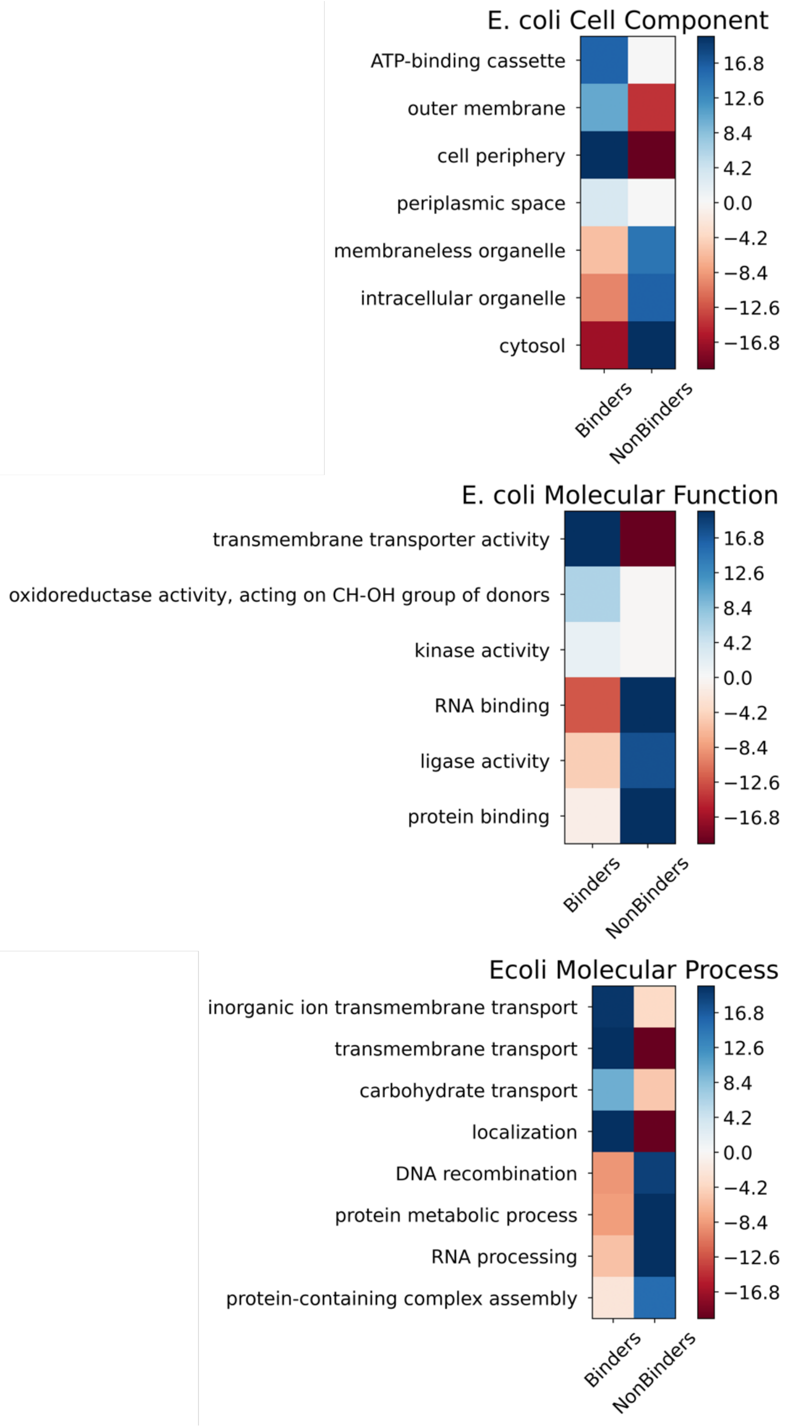


Figure 3.8: Statistical analysis of *E. coli* strain K12 PiCAP predicted carbohydrate binding and non-binding proteins. The false discovery rate (FDR) alongside the overrepresentation (blue) and underrepresentation (red) are shown for select cellular compartments, molecular functions, and cellular processes.

Mouse Proteome Analysis

Cellular Component

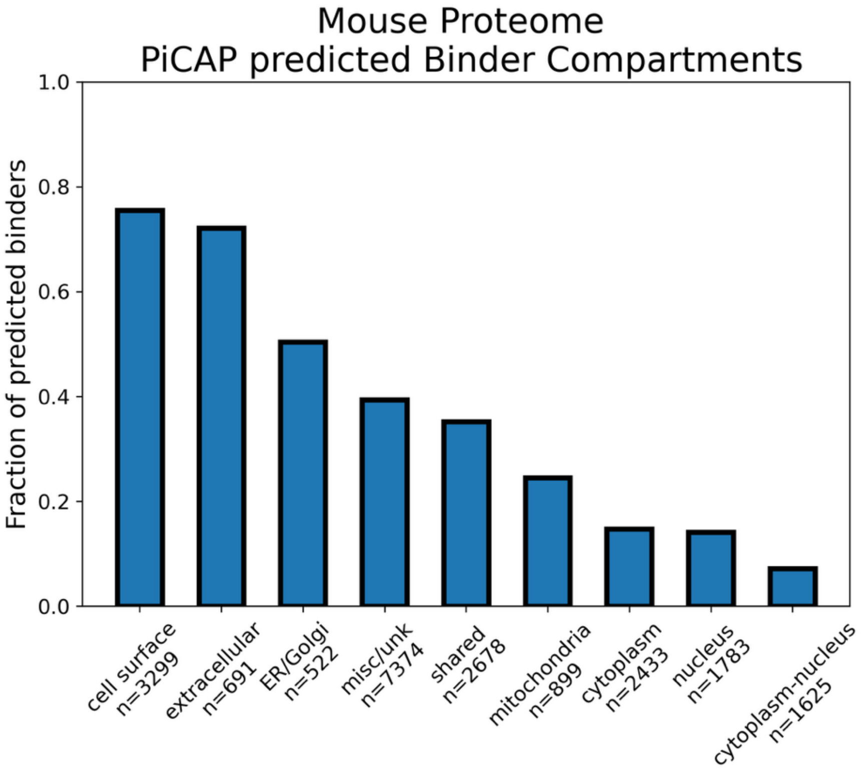
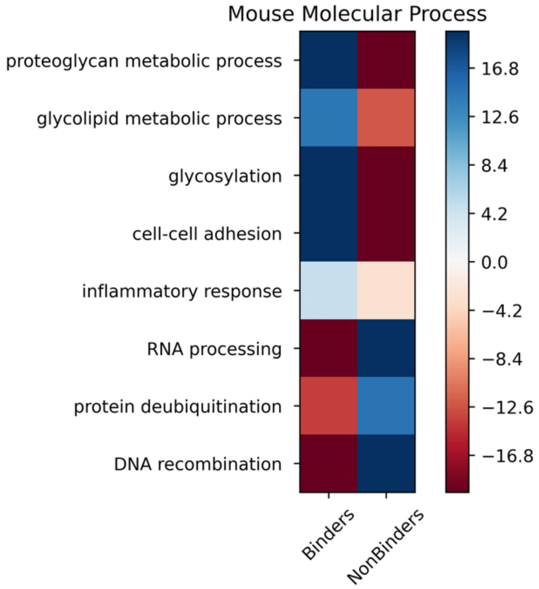


Figure 3.9: Cellular components of *M. musculus* proteome predicted carbohydrate binding and non-binding proteins according to Table 3.6.

Overrepresentation of protein processes and functions



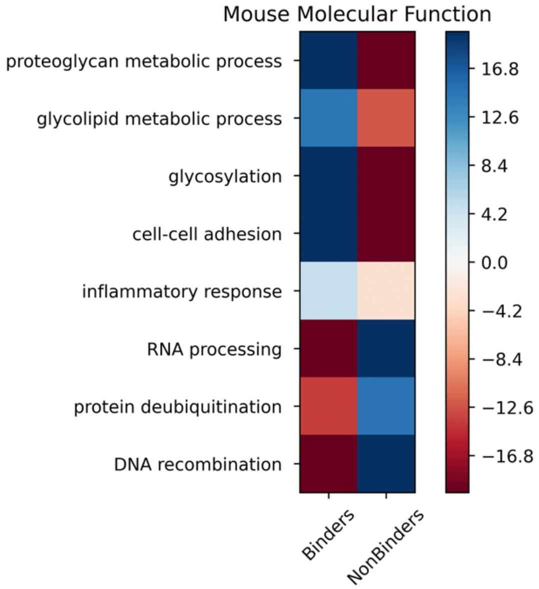


Figure 3.10: Statistical analysis of *M. musculus* PiCAP predicted carbohydrate binding and non-binding proteins. The FDR and overrepresentation (blue) and underrepresentation (red) are shown for select cellular functions and molecular processes.

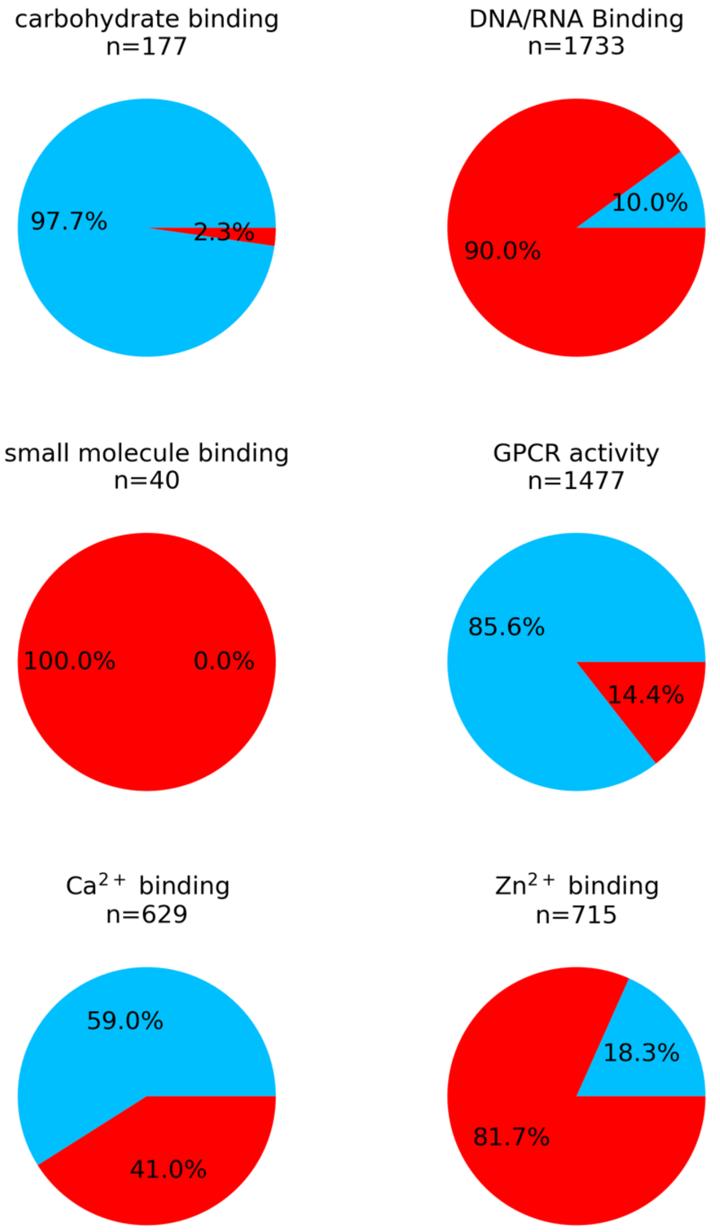


Figure 3.11: ***M. musculus*** **carbohydrate binding protein functionality.** Percentage of proteins with known binding functions predicted as carbohydrate binding (blue) and non-carbohydrate binding (red) proteins for Gene ontology terms for carbohydrate binding [GO:0030246], DNA and RNA binding [GO:0003677, 0003723], small molecule binding [GO:0036094], GPCR Activity (G protein-coupled receptor activity [GO:0004930]), Calcium binding (calcium ion binding [GO:0005509]), and (Bottom Right) Zinc binding (zinc ion binding [GO:0008270]).

Chapter 4

BCAPIN: Evaluation of De Novo Deep Learning Models on the Protein-Sugar Interactome

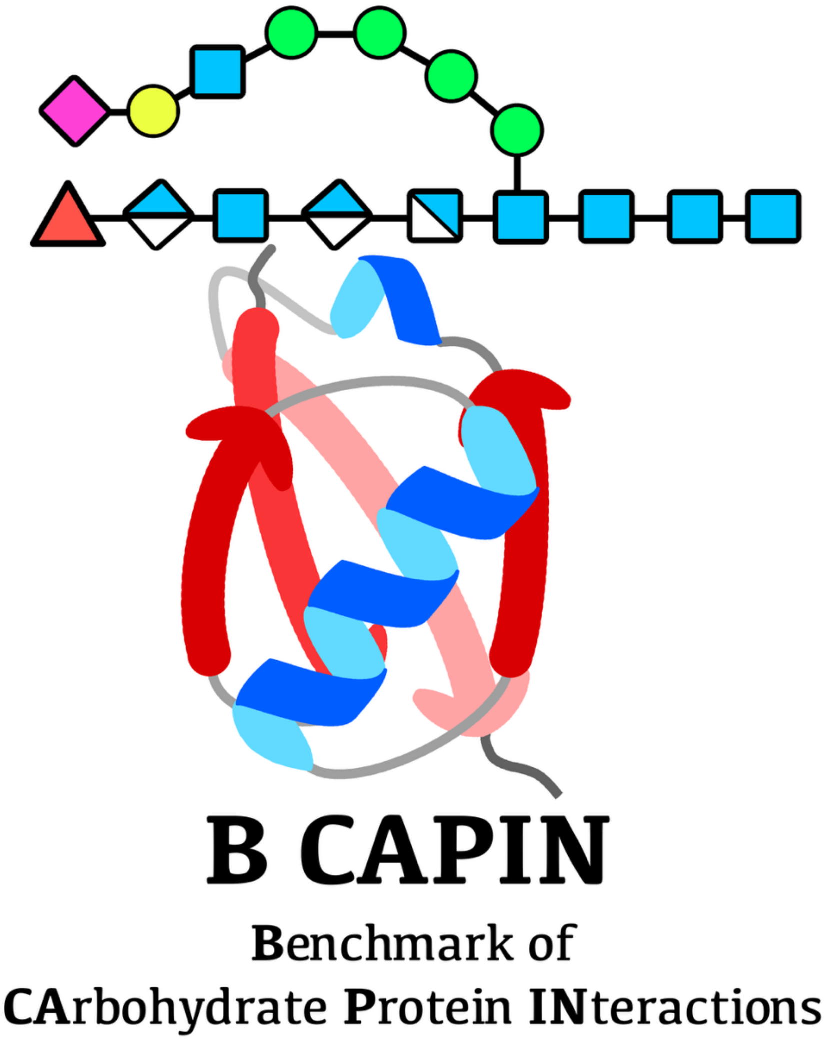


Figure 4.1: Benchmarking Carbohydrate Predictions with BCAPIN

Attribution of credit: this work was performed primarily alongside Dr. Lei Lu, with support from Sho S. Takeshita, and advised by Dr.'s William F. DeGrado and Jeffrey J. Gray.

Overview

Advances in deep learning have produced a range of models for predicting the protein-sugar interactome; however, structural docking of noncovalent protein-carbohydrate complexes remains largely unexplored. Although all-atom structure prediction models like AlphaFold3 (AF3), Boltz-1, Chai-1, DiffDock, and RosettaFold-All Atom (RFAA) were validated on protein-small molecule complexes, no benchmark or evaluation exists specifically for noncovalent protein-carbohydrate docking. To address this, I developed a high-quality dataset of experimental structures – Benchmark of CArbohydrate Protein Interactions (BCAPIN). Using BCAPIN and a novel evaluation metric, DockQC, I assessed the performance of all-atom structure prediction models on non-covalent protein-carbohydrate docking. I found all methods achieved comparable results, with an 85% success rate for structures of at least acceptable quality. However, I found that the predictive power of all models declined with increasing carbohydrate polymer length. With the capabilities and limitations assessed, I evaluated AF3's ability to predict binding for a set of putative human carbohydrate binding and carbohydrate non-binding proteins. While current models show promise, further development is needed to enable high-confidence, high-throughput prediction of the complete protein-sugar interactome.

Introduction

Many new computational prediction tools have recently been developed to decode the protein-sugar interactome. Bonnardel et al. created LectomeXplore, which annotates all known proteomes with a hidden Markov model (HMM) for lectins (glycan-binding proteins).[16] If the protein is identified as a lectin, one could use Lundstrøm et al.'s model LectinOracle to predict which carbohydrate the lectin binds.[48] However, not all carbohydrate binding proteins are lectins, for example native sugar sensors and antibodies.[6] Leveraging this gap, some of us (Canner, Gray) developed PiCAP to predict *whether* a protein binds to carbohydrates, irrespective of protein family, and released predicted annotations on six different species, predicting a putative list of all the proteins present in the protein-sugar interactome.[23] To further elucidate these protein-carbohydrate interactions, Canner, Shanker et al. and Bibekar et al. created CAPSIF[106] and PeSTo-Carbs,[110] respectively, to predict which residues a protein uses to bind to carbohydrates. The combination of these breakthrough models can be used to predict whether any given protein binds to a carbohydrate (as a lectin or non-lectin), what carbohydrate it binds to (if the protein is a lectin), and what residues are implicated in the protein for carbohydrate binding. Now, with all-atom biomolecular prediction software like AlphaFold3 (AF3),[42] protein-carbohydrate complex structures can be readily predicted. AF3 and other deep learning models thereby make possible the development of a complete putative structural dataset of the entire protein-sugar interactome: all protein-carbohydrate interactions across a species. First however, researchers must evaluate the performance of AF3 and other all-atom biomolecular structure prediction models on protein-carbohydrate complexes.

The development of AlphaFold3[42] built upon a string of advances in protein structure prediction, such as the Nobel Prize research of David Baker, Demis Hassabis, and John Jumper: Rosetta and AlphaFold2.[40] In the past few years, the leap in the most recent generation of *de novo* prediction methods was the ability to model any molecule with all-atom structure prediction. The Jaakkola lab developed DiffDock to predict small molecule docking on a provided protein structure.[43,135] The Baker lab developed RosettaFold-AllAtom (RFAA), becoming the first end-to-end all-atom biomolecule structure prediction.[131] Google DeepMind released their first end-to-end all atom prediction model AlphaFold3 (AF3). Building on previous work from DiffDock, the

Jaakkola lab developed Boltz-1.[128] With partnerships from OpenAI and other industry representatives, the Chai Discovery team released their (proprietary) model Chai-1.[42,136]

Given this growing suite of models (albeit non-exhaustive), identification of their performance on specific tasks is critical, with one of the most used metrics being the success rate when benchmarked against a dataset called Posebusters.[137] Posebusters contains non-covalent protein-small molecule complexes. Posebusters provides well-defined specificity of the small molecule and binding protein pocket, with a model's success measured by its ability to predict small molecule complexes under 2 Å RMSD from the solved structure. In total, DiffDock and RFAA both achieve 42% success on PoseBusters,[131] while AF3 and Chai-1 achieve 76% and 77% success on PoseBusters,[42,136] respectively. No success rate was reported on PoseBusters for Boltz-1.[128]

While PoseBusters emphasizes strong specific protein-ligand binding, protein-carbohydrate interactions present unique challenges. Unlike protein-small molecule interactions, protein-carbohydrate interactions are commonly less specific, with proteins containing multiple binding sites for long linear heterogenous polymers containing various epitopes, and therefore sugars require extra attention that is not provided in the dataset.[20,105,138] Moreover, proteins stabilize carbohydrates through a combination of direct contacts (hydrogen bonding, electrostatics), indirect (water mediated) interactions, and by CH-π bonds via aromatic residues.[18,20] Finally, the binding affinity of protein-carbohydrate complexes are commonly weak (μM – mM), but rather driven by high avidity (nM) of multiple binding sites on the protein or multiple repeats of the glycan epitope.[6]

Due to the distinct binding mechanisms involved in noncovalent protein-carbohydrate interactions, solved experimental structures of bound non-covalent carbohydrates to proteins are limited. From all solved structures in the Protein Data Bank,[86] DIONYSUS identifies protein structures with non-covalent specific interactions with carbohydrates to be 2.5% (5,461).[113] With the advent of high-throughput diazirine photoaffinity linker experimental data of protein-carbohydrate interactions, [22] researchers are attaining more knowledge of protein-carbohydrate interactions on a protein level.[17] I therefore propose that all-atom deep learning (DL) structure prediction pipeline may enhance general understanding of the protein-sugar interactome.

Here, I benchmark DL structural models: AF3, Boltz-1, Chai-1, RFAA, and DiffDock on the task of predicting docked *de novo* protein-carbohydrate structures. To benchmark the models, I constructed a novel dataset of proteins unseen during each model's training. I identify the strengths and shortcomings of these models and evaluate test cases where all models perform poorly. With strengths and limitations identified, I then use AF3 as a proof-of-concept tool for predicting the structural *de novo* human protein-sugar interactome. This work sets the stage for future integration of deep learning tools in structural glycobiology to fully characterize the protein-sugar interactome across all species.

Results

BCAPIN and DockQC: Novel datasets and analysis

To assess the capabilities of AlphaFold3 (AF3), Boltz-1, Chai-1, DiffDock, and RosettaFold All-Atom (RFAA) at predicting protein-carbohydrate complexes, it is essential to have an independent test set of high-quality experimentally resolved protein-carbohydrate structures and a suitable evaluation metric. For the dataset, I leveraged DIONYSUS,[113] which aggregates all experimentally determined protein-glycan structures from the PDB. I first excluded all protein-nucleic acid complexes and clustered the remaining protein sequences at 50% identity.

I removed clusters with structures solved before the latest model's training cutoff dataset (September 2021). Importantly, due to experimental limitations, not all experimental structures are of equal quality. To ensure structural reliability, I applied a filter using the real space correlation coefficient (RSCC)[139], which measures the agreement between the calculated and experimental density. Structures with an RSCC greater than or equal to 0.9 were retained (Figure 4.10). The resulting Benchmark of CArbohydrate Protein INteractions (BCAPIN) test set consists of 20 structures: 9 structures that bind sugar monomers, 3 structures that bind dimers, 5 structures that bind polymers, and 3 structures that bind at least a nucleotide (NTP) and a saccharide (Table 4.1).

Table 4.1: **Benchmark of CArbohydrate Protein INteractions (BCAPIN) test set.** The table lists the PDB 4-letter ID, protein name, UniProt ID, glycan input string for GlyLES, and any secondary ligands if present.

PDB	Protein Name	UniProt	GlyLES Input String	Secondary Ligand
7blg	Carbohydrate-binding protein family 32	Q8A3D9	Gal	
7en5	HTH-type transcriptional regulator MurR	P77245	MurNAc	
7exj	Probable galactinol-sucrose galactosyltransferase 6	Q8RX87	Fruf(b2-1)[Gal(a1-6)]Glc	
7exo	Putative L-type lectin	Q58791	Man	
7f9g	Thrombocorticin	C0HM62	Fucp	
7jnf	F5/8 type C domain protein	A0A0H2YN38	GalNAc	
7jwf	Glycoside hydrolase Family 110	N/A	Gal(a1-3)Galb	
7mzs	Fimbrial adhesin	A0A2X2BLR9	Gal	
7rft	SAS protein 20	N/A	Glc(a1-4)Glc(a1-4)Glc(a1-4)Glc(a1-4)Glca	
7rpy	Cohesin containing protein	N/A	Glc(a1-4)Glc(a1-4)Glca	
7vi7	β-N-acetylhexosaminidase	B2UPP0	GlcNAc	
7w11	3-O-Glycosyltransferase	A0A385Z7H9	Glc	UDP
7w18	Alginate lyase	D2KX85	ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManAb	
7zon	Glycoside hydrolase family 18	A0A979GQH9	Glcb	
8axs	Exo-α-Sialidase	N/A	Neu5Ac	
8bf3	Feruloyl esterase wtsFae1B	A0A5S8WFA0	Xyl(b1-4)Xylb	
8d0r	Fucosyltransferase	Q9Y231	Fucp(a1-2)Gal(b1-4)GlcNAcb	GDP
8dzd	MS3494 : putative secreted protein	A0QY10	Fru(b2-1)Glca	
8ic1	Endo-α-D-arabinanase mutant	N/A	Araf(a1-5)Araf(a1-5)Araf(a1-5)Araf(a1-5)Araf(a1-5)Araf(a1-5)Araf	
8inp	7-O-uridine diphosphate glycosyltransferase	N/A	Glc	UDP

To evaluate the performance of predicted protein-carbohydrate complexes, I developed a single continuous scoring metric named DockQC. DockQC is inspired by the DockQ metric from the CASP-CAPRI challenge, averaging the fraction of native contacts (F_{nat}), interface root mean squared deviation (IRMS), and ligand RMS (LRMS) to designate a predicted structure's quality. While DockQ is widely used for protein-protein docking, the native code is unusable on the test cases, and, when reimplemented, it tends to overestimate the quality for protein-carbohydrate complexes, often assigning medium-to-high scores even when the predicted ligand position is incorrect (Figure 4.9, Table 4.2).

DockQC addresses these issues by averaging three terms: F_{nat}, ring-ring RMSD (rRMS), and LRMS. F_{nat}measures the fraction of native residue-residue contacts, *rRMS* is a novel metric that measures the RMSD between the center of mass (COM) of each carbohydrate ring in the aligned predicted and experimental structures, and *LRMS* measures the RMSD of all aligned ligand heavy atoms.

With the BCAPIN test set and evaluation metrics established, I investigated the performance of five methods, AlphaFold3, (AF3), Boltz-1, Chai-1, RosettaFold All-Atom (RFAA), and DiffDock, at predicting protein-carbohydrate structure. I first evaluated the behavior of DockQC on the set. Thresholds were chosen after inspecting many predictions and tuning metric weights, some examples are described next.

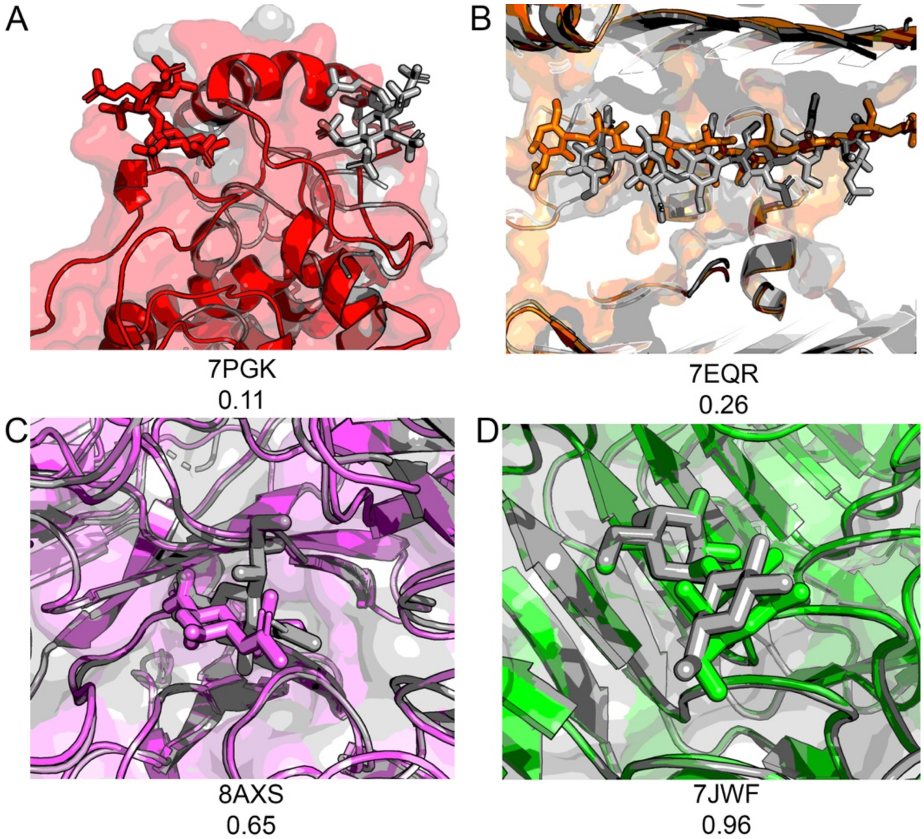


Figure 4.2: Protein-carbohydrate docked structures across DL methods. (A) Incorrect prediction of Chai-1 (red) on 7PGK (DockQC = 0.11). (B) Acceptable quality prediction of RFAA (orange) on 7EQR (DockQC = 0.26). (C) Medium quality prediction of Boltz-1 (violet) on 8AXS (DockQC=0.65). (D) High quality prediction of AF3 (green) on 7JWF (DockQC=0.96)

On hedgehog interacting protein (7PGK), which binds a disaccharide heparin analog, Chai-1 failed to predict the protein structure accurately, leading to an incorrect carbohydrate placement with a low DockQC score of 0.11 (Figure 4.2A). For chitoporin (7EQR), a β-barrel protein that binds an oligosaccharide with a degree of polymerization (DP) of six, RFAA captured the binding pocket of the carbohydrate, but lacked broader structural accuracy, yielding an acceptable prediction a DockQC of 0.26 (Figure 4.2B). With sialidase-sialic acid complex (8AXS), Boltz-1 achieved a medium quality prediction, correctly modeling the binding pocket and ring position (but not its orientation), with a DockQC of 0.65 (Figure 4.2C). In contrast, on glycoside hydrolase family 110 protein binding a Gal dimer (7JWF), AF3 nearly recapitulated the experimental structure delivering a high-quality structure with a 0.96 DockQC (Figure 4.2D). In total, my DockQC quality thresholds chosen to be incorrect (DockQC < 0.25), acceptable (0.25 <= DockQC < 0.50), medium (0.50 <= DockQC < 0.80), and high (DockQC >= 0.80) (Figure 4.9, Table 4.2).

DL Methods achieve medium or high accuracy on over 80% of cases

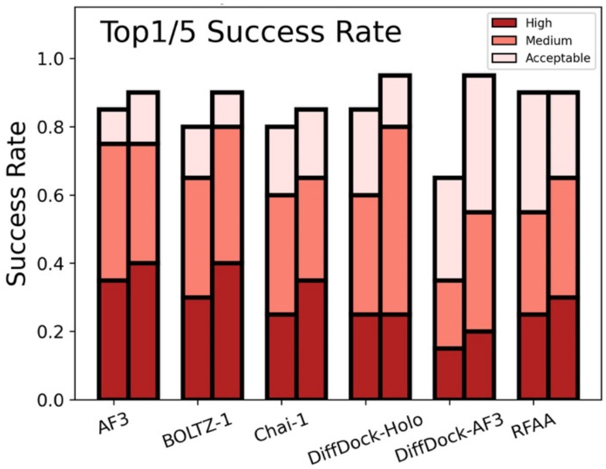


Figure 4.3: DL model success rates on BCAPIN Test Set. Each labeled method has the top-1 model on the left and top-5 model on the right

After tuning the DockQC metric, I evaluated overall model performances on all BCAPIN targets (Figure 4.3). Across methods, I found comparable results for all end-to-end models, at least 80% of their highest confidence predictions (top-1) scored with at least acceptable quality. Expanding scoring to include the most accurate of each model's top 5 confidence predictions (top-5) led to only marginal improvements. AF3 was the best-performing model: its top-1 predictions yielded 10% acceptable, 40% medium, and 35% high quality structures; top-5 predictions improved slightly to 15% acceptable, 35% medium, and 40% high quality structures.

Given the strong performance of end-to-end models on BCAPIN, I next examined how starting structure influences DiffDock's predictive power. DiffDock-*holo* (initialized with the experimentally solved *holo* protein structure) performed equivalently to the end-to-end models, achieving at least acceptable quality on 85% of all top-1 predictions. In contrast, Diffdock-*AF3* (initialized from AF3-predicted *apo* protein) achieved only 60% acceptable or better quality in top-1 predictions. However, when extending to the top-5 predictions, Diffdock-AF3 improved substantially, yielding 85% acceptable quality structures. Thus, DiffDock is sensitive to the initial input structure.

Methods fail to capture all interactions

Although all models perform strongly on BCAPIN, I sought to identify cases where all models still struggle. Notably, all models fail to predict on two complexes: 8DZD and 7ZON (Figure 4.4).

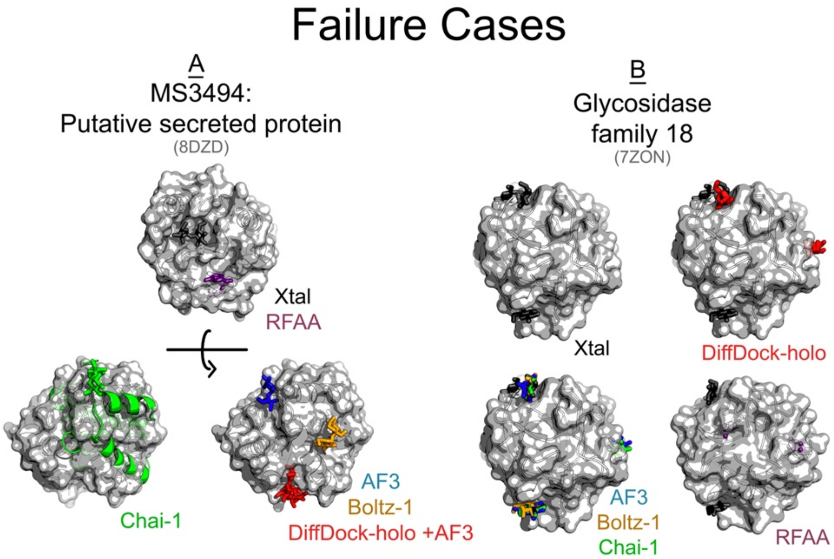


Figure 4.4: Failure of DL prediction algorithms on select proteins from the BCAPIN test set. Experimentally solved structures of (A) secreted protein (8DZD) and (B) glycosidase family 18 (7ZON, right) in gray, alongside AF3 predictions (blue), Boltz-1 (orange), Chai-1 (green), Diffdock (red), and RFAA (magenta).

8DZD is a *Mycobacterium smegmatis* secreted protein composed entirely of α-helices bound to a fructose-glucose disaccharide. While most models (except Chai-1) accurately predict the protein backbone, none correctly dock the ligand. RFAA places the ligand inside the protein. 7ZON is a glycosidase primarily composed of β-sheets bound to three independent glucose monosaccharides. Although most models correctly predict two of the binding sites, the models consistently misplace the third monosaccharide on the opposite side of the protein surface.

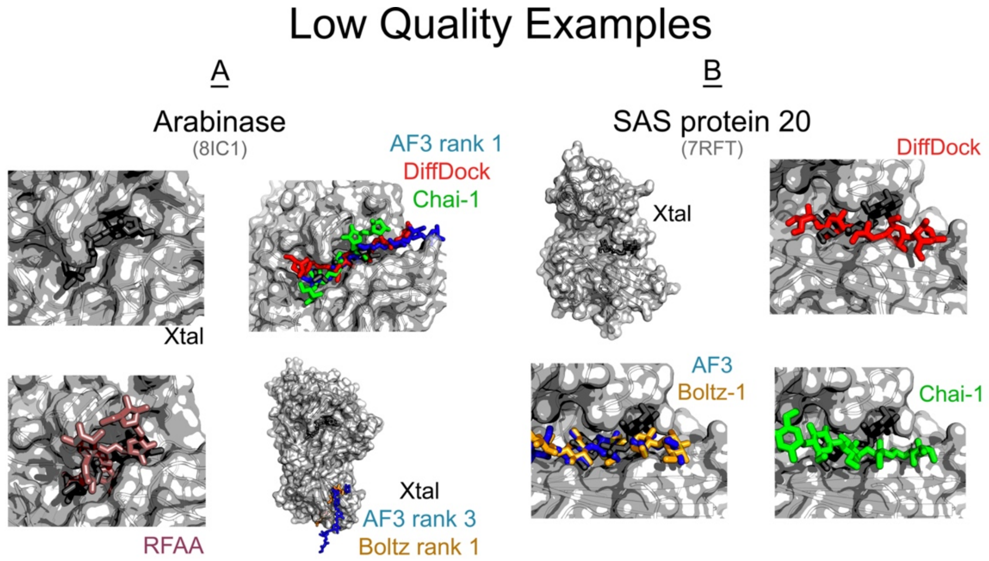


Figure 4.5: Low Quality DL predictions select proteins from the BCAPIN test set. We show the experimentally solved structures of (A) arabinose (8IC1) and (B) SAS protein 20 (7RFT), in gray, alongside AF3 predictions (blue), Boltz-1 (orange), Chai-1 (green), Diffdock (red), and RFAA (magenta).

I further scrutinized all predictions to identify additional cases of sub-optimal performance. I found that all models produced only acceptable to medium quality on 8IC1 and 7RFT (Figure 4.5).

8IC1 is an arabinose that binds a homogenous arabinofuranose oligosaccharide of DP 4 along a β-sheet. Several models, such as AF3 and Boltz-1, incorrectly predict binding at an alternative β-sheet, while others (DiffDock and RFAA) incorrectly predict the saccharide conformation (Figure 4.5A). 7RFT is a SAS protein 20 that binds a glucose oligosaccharide of DP 3 at a β-sheet. Although all methods identify the binding pocket of 7RFT correctly, none accurately reproduce the specific experimental conformation, particularly the orientation of the terminal Glc, which experimentally makes minimal contact with the β-strand (Figure 4.5B). These data suggest that current models may have difficulty on α-helical binding pockets of saccharides, simultaneous binding of multiple ligands, and docking longer saccharides.

Prediction Power decreases with carbohydrate length

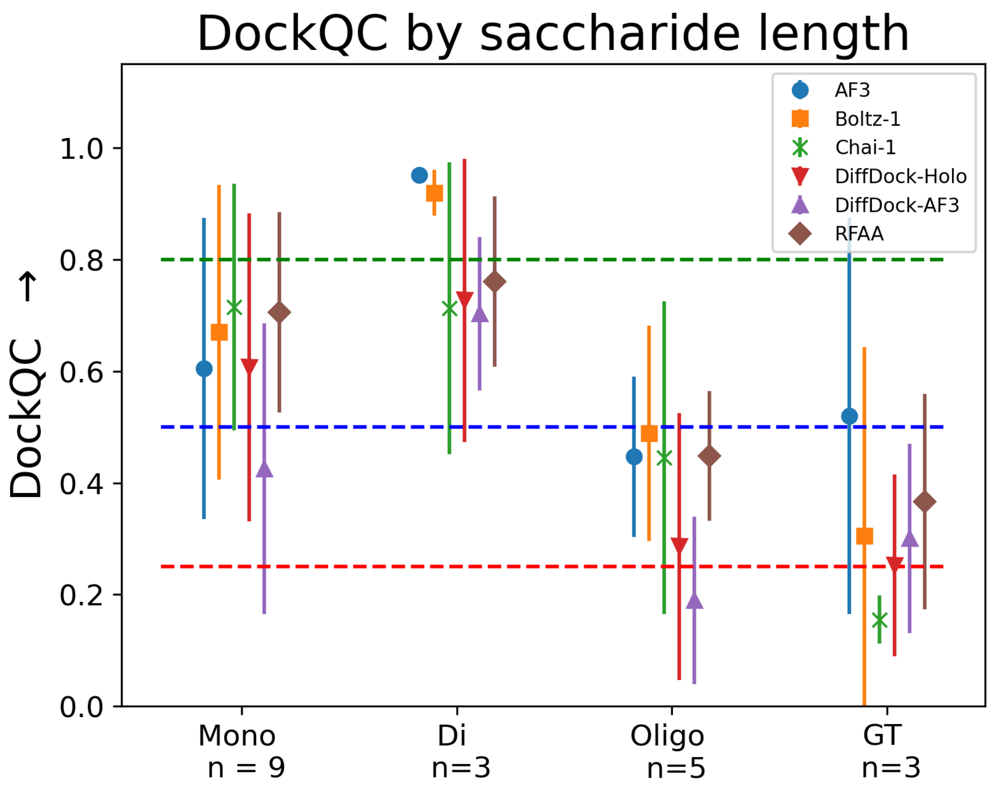


Figure 4.6: Comparison of average and standard deviation DockQC of predicted structures versus saccharide length. I group saccharide length into a degree of polymerization (DP) of 1 (mono), 2 (di), and 3+ (oligo), and further group all glycosyltransferases (GTs) together that require multiple inputs (e.g. a saccharide and NTP) and with the number of proteins in each group listed. Dashed lines indicate the DockQC cutoffs between acceptable (red), medium (blue), and high (green) quality structures. Top-1 prediction on BCAPIN with AF3 (blue circle), Boltz-1 (orange square), Chai-1 (Green X), Diffdock-*holo* (red triangle), Diffdock-*AF3* (purple triangle), and RFAA (brown diamond).

I next hypothesized that model performance may correlate with saccharide complexity. To explore the role of DP on performance, I plotted the top-1 DockQC score against saccharide length (Figure 4.6). In total, all models showed similar trends across saccharide length categories: medium quality for monosaccharides, medium to high quality for disaccharides, acceptable quality for oligosaccharides, and acceptable quality for glycosyltransferases (GTs). Thus, I observed a decline in performance as complexity increased from simple mono and disaccharides to DP of three or greater and coordination of small ligands, in the case of GTs.

Prediction confidence is a mediocre metric

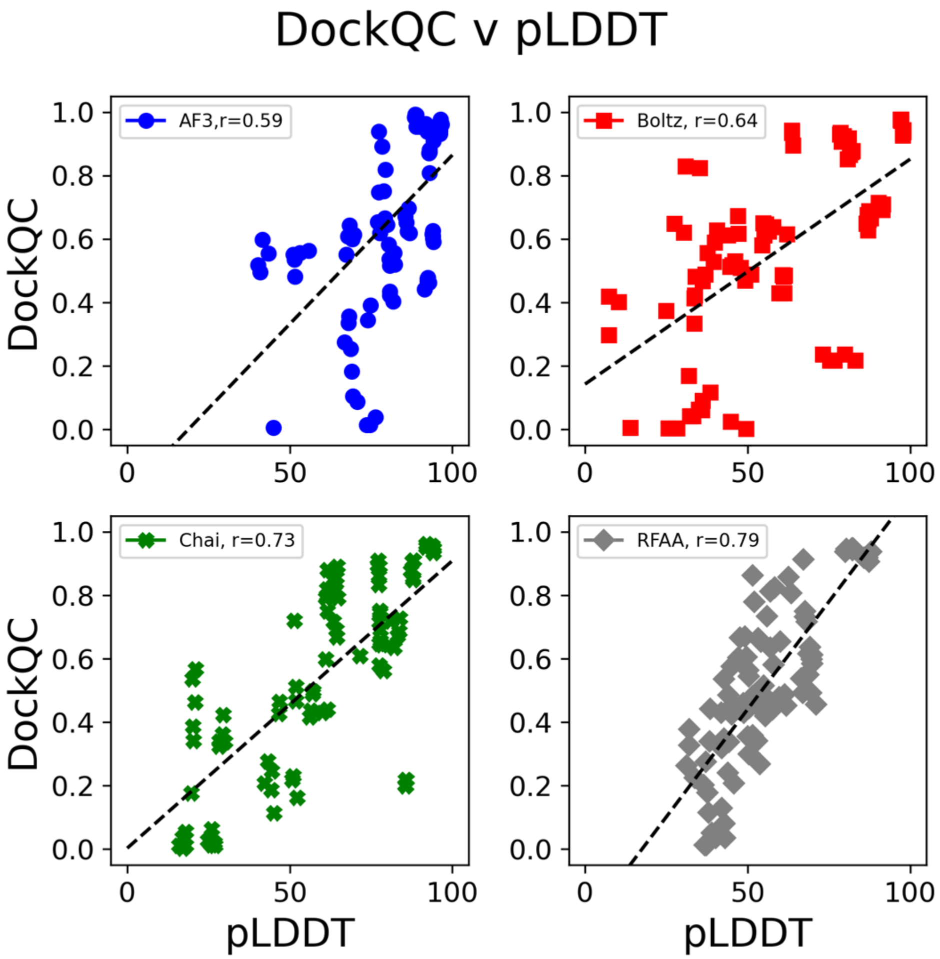


Figure 4.7: Comparison of confidence metrics and DockQC accuracy on the BCAPIN test set. Lines of best fit are provided for each plot. (A) Comparison of DockQC and ligand pLDDT for AF3 (blue circle), Boltz-1 (red square), Chai-1 (green X) and RFAA (gray diamond). (B) Comparison of DiffDock confidence for both DiffDock-*holo* (red circle) and DiffDock-*AF3* (blue square). (C) Comparison of DockQC and ipTM for AF3, Boltz-1, and Chai-1. (D) Comparison of DockQC versus the pAE for AF3, Boltz-1 (called pDE), and RFAA.

Although all current models perform strongly on BCAPIN, performance varies across predictions. I therefore assessed whether models can reliably self-assess the accuracy of their own predictions using internal confidence metrics, such as predicted local distance difference test (pLDDT), interface predicted template modeling score (ipTM), and predicted absolute error (PAE). For average ligand pLDDT, AF3 and Boltz-1 show moderate correlations with DockQC, whereas Chai-1 and RFAA produce strong correlations (Figure 4.7). Since pLDDT reflects only the ligand confidence, I also evaluated ipTM, which incorporates the protein-ligand interface. Among models reporting ipTM (AF3, Boltz-1, Chai-1), all show moderate correlations, with Chai-1 performing best (Figure 4.11). For PAE, Boltz-1 showed a weak negative of -0.26, AF3 a moderate correlation, with RFAA a strong correlation of -0.7 with DockQC (Figure 4.12).

Contrary to the end-to-end models, DiffDock provides only one confidence metric. While both DiffDock-holo and DiffDock-AF3 use the same scoring, DiffDock-*AF3*'s provides a significantly weaker correlation than DiffDock-*holo*, reinforcing DiffDock's sensitivity to the starting structure (Figure 4.13).

Overall, all end-to-end models show moderate correlations between their internal confidence metrics to the DockQC, with RFAA demonstrating the strongest predictive reliability. Contrarily, DiffDock's confidence metric is more susceptible to small perturbations in the input structure, limiting its reliability.

Proteome scale predictions require refinement

The BCAPIN dataset is limited to small (less than 600 residues) single- or two-domain structurally resolved proteins with strong binding affinities. Despite being implicated in important physiological interactions, binding characteristics of large multidomain or multichain structures with carbohydrates are less well characterized due to their relative low binding affinity (but high avidity). To elucidate the protein-sugar interactome, researchers currently employ photoaffinity tag experiments[17] or use computational tools like LectinOracle[48] or PiCAP.[23] However, these tools do not provide structural protein-carbohydrate complex predictions. Therefore, I aimed to assess if any end-to-end all atom structure prediction models could provide a high-throughput *de novo* approach for predicting docked protein-carbohydrate complexes with high confidence. To evaluate a *de novo* protein-carbohydrate docking pipeline, I selected nine proteins from the human proteome and used AF3 with its ipTM confidence metric to predict their structures in complex with either a GM1 ganglioside or a hybrid N-glycan (Figure 4.8). I used GM1 ganglioside ligands for proteins experimentally identified to interact with GM1 gangliosides in Zhang et al. and the hybrid N-glycan ligand for all others, as it is a common covalent modification on membrane and secreted proteins.

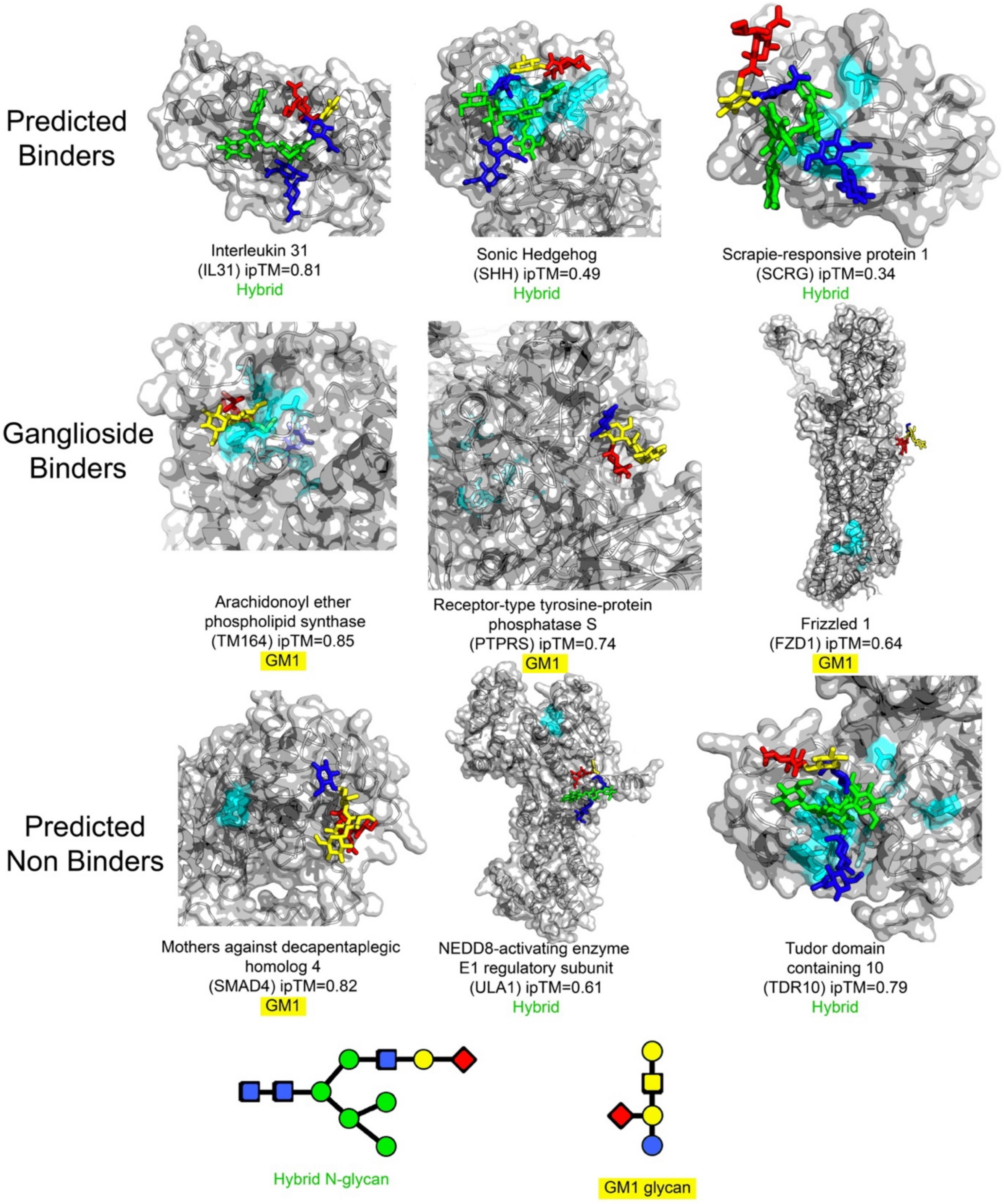

Predicted Binders
Interleukin 31 (IL31) ipTM=0.81
Hybrid
Sonic Hedgehog (SHH) ipTM=0.49
Hybrid
Scrapie-responsive protein 1 (SCRG) ipTM=0.34
Hybrid
Ganglioside Binders
Arachidonoyl ether phospholipid synthase (TM164) ipTM=0.85
GM1
Receptor-type tyrosine-protein phosphatase S (PTPRS) ipTM=0.74
GM1
Frizzled 1 (FZD1) ipTM=0.64
GM1
Predicted Non Binders
Mothers against decapentaplegic homolog 4 (SMAD4) ipTM=0.82
GM1
NEDD8-activating enzyme E1 regulatory subunit (ULA1) ipTM=0.61
Hybrid
Tudor domain containing 10 (TDR10) ipTM=0.79
Hybrid
Hybrid N-glycan
GM1 glycan

Figure 4.8: AlphaFold3 predictions on selected human protein-glycan interactions. PiCAP provides the protein-level prediction, and CAPSIF2 residue predictions (cyan). The bound glycan is either a complex N-glycan (green) or a GM1 ganglioside (yellow), with the initial GlcNAc of the N-glycan highlighted in blue and all sialic acids highlighted in magenta.

PiCAP predicted interleukin 31 (IL31), sonic hedgehog (SHH), and scrapie-responsive protein 1 (SCRG) as putative carbohydrate binding proteins. Here, I used AF3 to dock these proteins with a hybrid N-glycan, a common branched saccharide where one branch terminates in an oligomannose chain and the other in a sialic acid. CAPSIF2 predicted no carbohydrate-binding residues on IL31; however, AF3 predicted the glycan to bind at an unstructured region of the protein with a high interaction confidence (ipTM = 0.81). Conversely, AF3 docked the N-glycan at the CAPSIF2 predicted residues of SHH and SCRG with a lower confidence (ipTM = 0.49).

Experimentally, arachindonyl ether phospholipid synthase (TM164), receptor-type tyrosine-protein phosphatase S (PTPRS), and Frizzled 1 (FZD1) were identified in multiple experiments as ganglioside binding proteins.[17] These proteins were also predicted by PiCAP to bind a carbohydrate. I therefore modeled these proteins in complex with the GM1 ganglioside glycan (Figure 4.8). AF3 predicted TM164 to bind GM1 in the CAPSIF2 predicted pocket with high confidence (ipTM = 0.85). However, AF3 however predicts PTPRS and FZD1 to bind the ganglioside glycan at sites outside of the CAPSIF2 predicted pockets. Notably, CAPSIF2 predicts on intracellular binding pocket for FZD1, whereas both AF3, experimental data, and CAPSIF:V suggest binding occurs in the extracellular region.[17]

While PiCAP predicts approximately 7,000 human proteins to bind carbohydrates, it also predicts ~13,000 human proteins as non-binders. To assess whether AF3 could also discriminate between physiologically relevant and irrelevant interactions, I selected three proteins: mothers against decapentaplegic homolog 4 (SMAD4), NEDD-8 activating enzyme E1 regulatory subunit (ULA1), and Tudor domain containing 10 (TDR10). Since SMAD4 was previously investigated by Zhang et al. and identified as a putative *non*-binder of GM1, I modeled the protein with GM1. AF3 however predicts the SMAD4-GM1 complex with a high confidence (ipTM = 0.82).

Similarly, AF3 predicted moderate to high confidence interactions for an N-glycan in complex with ULA1 (ipTM = 0.61) and TDR10 (ipTM = 0.79). These findings suggest that ipTM values alone may not be sufficient to distinguish between physiologic and non-physiologic interactions in a high-throughput manner.

Discussion

I present an evaluation of multiple end-to-end all-atom prediction frameworks for carbohydrate-protein docking and interrogate their capabilities at unveiling the structural secrets of the protein-sugar interactome. Overall, all methods perform incredibly well at this task – all end-to-end models capture 80% of their highest confidence models at least acceptable quality (Figure 4.3). These models improve upon previous energy-based protein-carbohydrate docking methods like GlycanDock[49] and HADDOCK[140], which are useful for refinement but not full *de novo* docking. Although the models I tested improve upon previous methods and models, they still have limitations, including reduced performance with increased complexity. Specifically, the models perform worse on multi-ligand targets (GTs) and saccharides with DP greater or equal to three. Also, the models lack robust confidence metrics for protein-carbohydrate complexes.

The BCAPIN dataset is the first study of protein-carbohydrate noncovalent docking, including all protein-carbohydrate complexes in the PDB. However, BCAPIN primarily comprises small, globular, single-domain proteins bound to linear glycan chains, which is not representative of the diverse protein-carbohydrate interactions found in physiological contexts. Thus, as more experimental data becomes available, alongside further developments in these prediction techniques, the framework presented here can be iterated to better elucidate the protein-sugar interactome.

The largest limitation in continually iterating and benchmarking this structure prediction software is the availability of high-quality experimental structures. Although the DIONYSUS dataset is impressive in its scope, containing 5,461 protein-carbohydrate complexes, only 1,842 unique protein structures remain after 95% sequence similarity[86,113]. Further, when assessing the individual unique binding pockets of these DIONYSUS proteins, there are only 258 unique clusters of binding pockets.[141] With this limited set of ~1,800 unique structures and ~250 unique binding mechanisms, data science and machine learning approaches are restricted. Therefore, discovery of novel carbohydrate binding proteins and their structural interactions is critical.

To better improve computational approaches, I believe that one of the most promising sources of future data future lies in liquid glycan arrays and photoaffinity labeling experiments (e.g. those using diazirine linkers).[17,129,130] These *in vivo* high throughput techniques enable identification of protein-carbohydrate interactions on a proteome-wide scale; however, they currently lack immediate structural resolution. Computational modeling stands poised to fill this gap by providing structural hypotheses at atomic level detail, thereby accelerating the validation and functional understanding of these experimentally identified interactions. To push the scope of the BCAPIN test set, I selected two branched polysaccharides with distinct properties to explore AF3's capabilities. Although this study does not demonstrate that AF3 is yet ready to support full scale high-throughput experiments comparable to photoaffinity labeling, it shows that AF3 can generate useful, testable hypotheses on a case-by-case basis that may expedite wet lab investigations.

To aid wet lab experiments, my lab has computationally studied protein-carbohydrate structural interactions. I developed GlycanDock[49], CAPSIF[106], and PiCAP[23] as ways to elucidate these interactions. PiCAP in particular, represents a significant advancement, as it was the first model to predict *whether* a protein binds to carbohydrate, irrespective of protein family on a proteome scale. However, these current models rely on the fundamental work of thousands of scientists solving crystal structures of protein-carbohydrate complexes. While high-throughput technologies are likely to uncover many more non-covalent protein carbohydrate interactions *in vivo*, reliably obtaining the bound structure or identifying the full glycan repertoire for each protein remains a computational bottleneck.

I envision a full suite of models and methods will fill the gap to identify the full protein sugar interactome of a species. I advocate for a model that would improve upon LectinOracle[48], integrating the glycan embeddings from methods like SweetNet[47] or Gifflar[142] using sequence and structural information insights from structure prediction models, current photoaffinity experiments, and CAZY[8] can predict the glycan binding repertoire of all proteins. With this addition, one can use PiCAP to predict whether a protein binds carbohydrates, use CAPSIF2 or PeSTo-Carbs to predict how the protein binds the carbohydrate structurally, and finally, use the proposed model to predict which carbohydrates are recognized, all at high-throughput scales. This integrated approach will be essential to fully map the protein-sugar interactome, advancing general understanding of glycan-mediated biology, enabling translational applications in therapeutics and diagnostics.

Methods

Dataset

To evaluate how all-atom prediction software extrapolates to glycans, I used DIONYSUS (access date: October 8, 2024), to construct the dataset. I first selected all protein-carbohydrate complexes after the September 2021 training cutoff date used by all models. Of the 5,461 identified structures by DIONYSUS, 614 proteins were deposited in the PDB after the training date cutoff. I then clustered all 5,461 protein sequences using MMSEQS[86] into 50% sequence identity clusters and removed any post-cutoff proteins with sequence homology with any protein published before the training date cutoff, leaving 105 structures. I then selected a single structure from each cluster, selecting the complex with the highest degree of polymerization (DP), leaving 35 protein structures. Of these 35 protein structures, 11 experimentally bind monomers, 6 experimentally bind dimers, 13 bind polymers (3+ saccharides), and 5 bind a saccharide and nucleotide triphosphate (NTP).

For each structure, I analyzed the ligand structure quality measures, notably real space R factor (RSR) and real space correlation coefficient (RSCC) (Figure 4.10).[139] When these metrics weren't available, (7TOH, 7YWF, 8CSF) I provide their root-mean-squared deviation Z-scores (RMSZ). I define the set of high-quality structures with an RSCC greater than 0.9,[139] which contains 20 structures: 9 that bind monomers, 3 that bind dimers, 5 that bind polymers, and 3 that bind at least an NTP and a saccharide. I named the dataset the Benchmark of CArbohydrate Protein INteractions (BCAPIN).

Prediction methodology

To provide an equivalent and biologically relevant input ligand for all structures, I generated the SMILES strings of the original PDB ligand using GlyLES[143] (part of the Glycowork[144] Python package). In the case of homogenous polymers, I extended the length of the original carbohydrate by a DP of 2 to provide additional biological context. AF3, Boltz-1, Chai-1 and Diffdock input a SMILES string,[42,43,128,136], but RFAA requires an SDF file input (ligand coordinates) to perform the calculations, which I used RDKit to calculate the initial ligand coordinates. In the case of heparin binding proteins (8EDI and 7PGK), I used the SMILES retrieved from the PubChem compound instead.[145] For the five glycosyltransferases (GTs) targets, I input both the carbohydrate(s) and NTP to the software for multi-body docking.

To replicate the process of a simple *de novo* pipeline, I ran all methods without modifications or customizations. AF3, Boltz-1, and RFAA were run with a local distribution with five random seeds using the SMILES strings (or RDKit generated SDF from the SMILES for RFAA). Chai-1 was run using the Chai-1 servers, which uses five random seeds for predictions. All confidence metrics were extracted from provided mmCIF and json files. For predicted absolute error, I rather used Boltz-1's interface predicted distance error (ipde).

Diffdock is not an end-to-end method, therefore I ran DiffDock in two different contexts, (1) with the solved experimental structure, which I call DiffDock-*holo*, and (2) with a predicted AF3 protein structure, which I call Diffdock-*AF3*. The AF3 structure for the input into DiffDock-*AF3* was chosen as the best ranking AF3 *apo* model running from 5 random seeds. I ran both DiffDock methods using the HuggingFace server with the SMILES strings, resulting in 10 total models. On GTs with multiple ligands, I concatenate the structures of the same rank together for a singular prediction.

Metrics

Carbohydrates differ substantially from conventional small molecules, as they range from small monosaccharides to branched polymers. I therefore selected the following metrics to analyze protein-carbohydrate complex predictions: full ligand F_{nat} ($F_{\text{nat,full}}$), residue F_{nat} ($F_{\text{nat,res}}$), ligand RMSD (LRMS), and ring-ring RMSD (rRMS).

F_{nat} is the fraction of native contacts, defined as all residue-residue contacts (any heavy atom to any heavy atom) within 5 Å:

$$F_{\text{nat}} = \frac{TP}{TP+FN},$$

where TP (True Positives) is the overlap between predicted contacts and experimentally known contacts and FN (False negatives) are all experimental contacts not observed in the predicted structure. I use this formal definition of residue-residue contacts which I call $F_{\text{nat,res}}$. In addition, as these are small molecule-like ligands, I additionally define $F_{\text{nat,full}}$, which instead of carbohydrate residue-protein residue contacts, instead is the full ligand F_{nat}, or any carbohydrate heavy atom-protein residue contacts (effectively treating the full ligand as a singular residue).

In addition to F_{nat}, I leverage the root means squared deviation (RMSD) metric:

$$RMSD(x, y) = \sqrt{\frac{1}{n}\sum_{i}^{n}\|x_i - y_i\|^2},$$

where x_i are the coordinates of select heavy atoms of the predicted structure and y_i are the same heavy atom coordinates of the experimentally determined structure after optimal superposition of the protein's binding pocket (all residues within 10 Å of the ligand). I chose two different RMSDs to indicate the fine-grained nature of carbohydrate polymers: ligand RMSD and ring RMSD. Ligand RMSD (*LRMS*) measures the distance between the predicted and experimental structures of the ligand's heavy atoms. For LRMS, I use the RDKit implementation that compares the maximal similar substructures.[146,147] Ring RMSD (*rRMS*) simplifies the problem to only measuring the distance between the center of mass (COM) of each carbohydrate ring. I use a greedy implementation of *rRMS*, where each saccharide species is equivariant to any other saccharide species along the polymer chain.

I combine the four separate measurements to "DockQC," which represents the overall quality of the predicted protein-carbohydrate structure on a scale from 0 to 1. This metric is inspired by the foundational DockQ metric for measuring protein-protein docking.[148,149] DockQ measures on a scale from [0,1] by combining the fraction of natural contacts (F_{nat}), LRMS, and interface RMS (iRMS).[148,149]

$$DockQ = \frac{1}{3}(F_{\text{nat}} + iRMS_{\text{scaled},d_1} + LRMS_{\text{scaled},d_2})$$

where d_1 = 1.5 Å and d_2 = 8.5 Å and

$$RMS_{\text{scaled},d_i} = \frac{1}{1 + \left(\frac{RMS}{d_i}\right)^2}$$

Currently, DockQ does not allow the ligands to differ in size between the crystal and predicted structure. Additionally for small molecules, DockQ only reports the LRMS value.[149] When I reimplemented the DockQ metric with these values accounted for, I found it unrepresentative of the predictions (Table 4.2, Figure 4.9). I therefore constructed the DockQC based on the metrics as follows:

$$DockQC = \frac{1}{3}\left(\frac{1}{2}(F_{\text{nat,res}} + F_{\text{nat,full}}) + rRMS_{\text{scaled},d_1} + LRMS_{\text{scaled},d_2}\right)$$

where d_1 = 2.0 Å, d_2 = 4.0 Å. I tuned the scaling factors of d_1 and d_2 to fit the DockQC into the four different categories: incorrect (DockQC < 0.25), acceptable (0.25 <= DockQC < 0.50), medium (0.50 <= DockQC < 0.80), and high (DockQC >= 0.80) (Figure 4.9, Table 4.2).

Human proteome predictions

I selected nine proteins from the human proteome to evaluate *de novo* docking on proteomic scales, where PiCAP predicts six of these proteins as carbohydrate binding proteins and three as non-binding proteins. I used the following purported glycans for docking based on the function of each protein: GM1, Gal(β1-3)GalNAc(β1-4)[Neu5Ac(α1-3)]Gal(β1-4)Glcβ, for ganglioside binding proteins and a hybrid N-glycan for the remaining proteins, Neu5Ac(α1-6)Gal(β1-4)GlcNAc(β1-2)Man(α1-3)[Man(α1-6)[Man(α1-3)]Man(α1-6)]Man(β1-4)GlcNAc(β1-4)GlcNAcβ.

Data Availability

The BCAPIN dataset and all model inputs, code, and analysis data are available at Github: `github.com/graylab/dockqc`.

Appendix

DockQC parameterization

Table 4.2: DockQC with different d_i values compared to DockQ[148] on 8 targets. Human labels for "High" quality, "Medium" quality, "Acceptable" (Acc) quality, and "Low" quality are provided.

	PDB	DockQ	DockQC d_1=3.0, d_2=1.5	DockQC d_1=5.0, d_2=1.5	**DockQC** **d_1=4.0, d_2=2.0**	DockQC d_1=3.0, d_2=2.0
High	7BLG	0.94	0.93	0.94	**0.94**	0.93
	8AD2	0.94	0.80	0.83	**0.85**	0.83
Med	7EN5	0.86	0.47	0.55	**0.56**	0.52
	7W18	0.68	0.51	0.52	**0.56**	0.56
Acc	7F9G	0.73	0.32	0.40	**0.39**	0.35
	7EQR	0.61	0.35	0.36	**0.40**	0.40
Low	8DZD	0.37	0.01	0.02	**0.01**	0.01
	8EDI	0.61	0.21	0.23	**0.22**	0.21

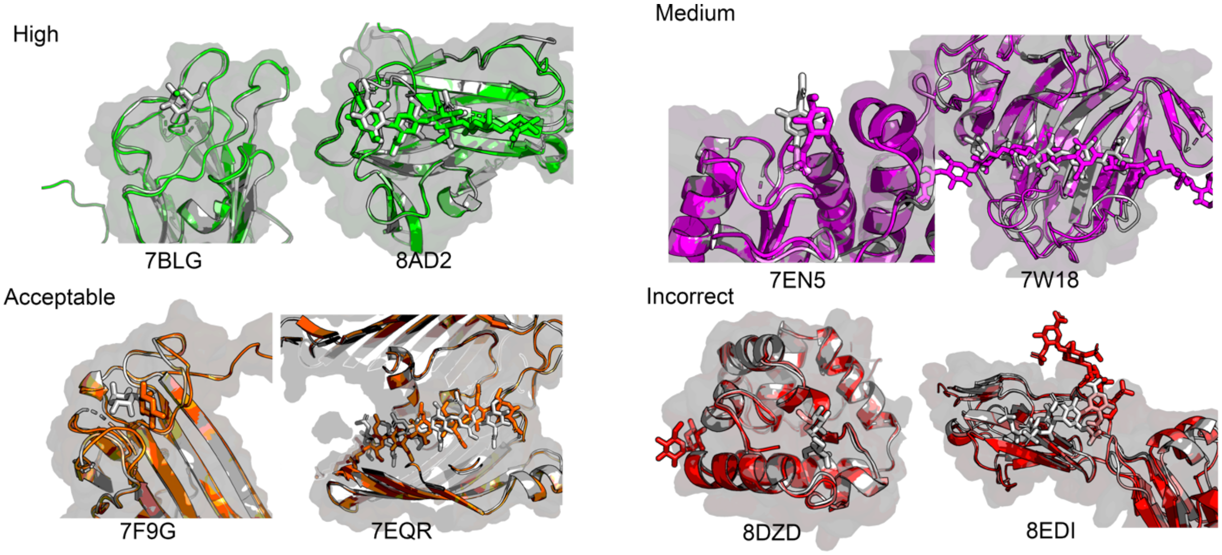


Figure 4.9: Predicted protein structures deemed to be of high, medium, acceptable, and incorrect quality.

BCAPIN RSC and RSRR values

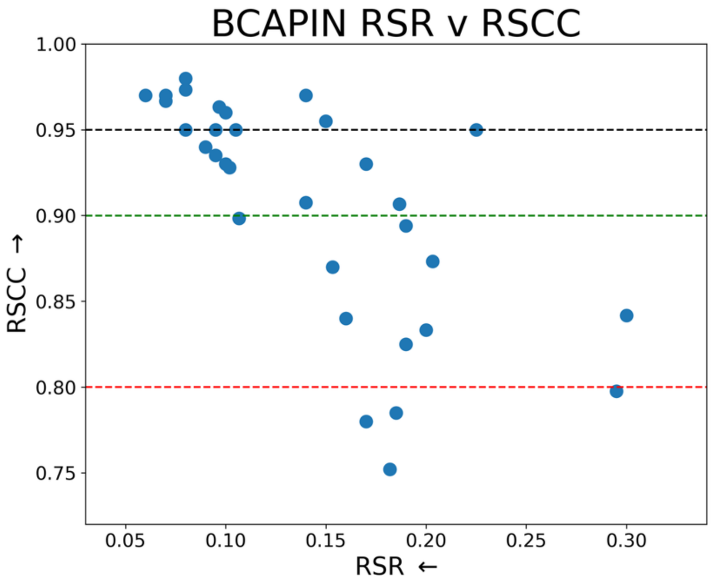


Figure 4.10: Analysis of the BCAPIN real space R factor (RSR) and real space correlation coefficient (RSCC). Lines at RSCC values of 0.95, 0.9 (the cutoff for the BCAPIN set), and 0.8 are shown.[139]

BCAPIN confidence metrics

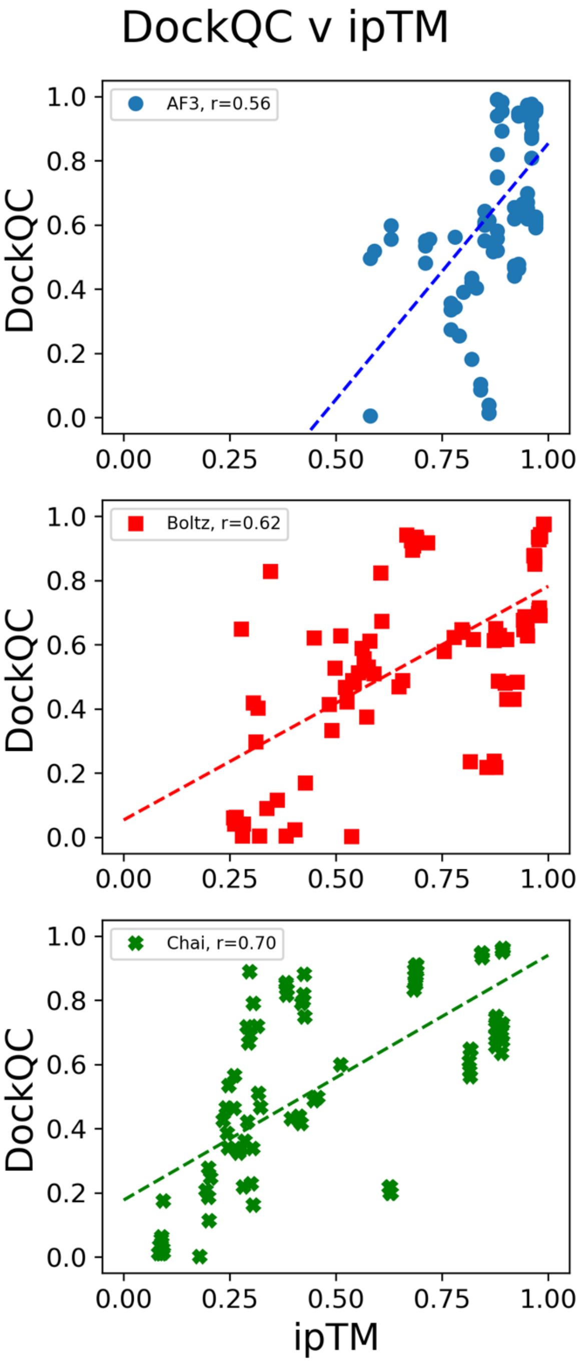


Figure 4.11: Comparison of ipTM and DockQC accuracy on BCAPIN. AF3 (blue circle), Boltz-1 (red square), and Chai-1 (green X) were the only models which reported such metric.

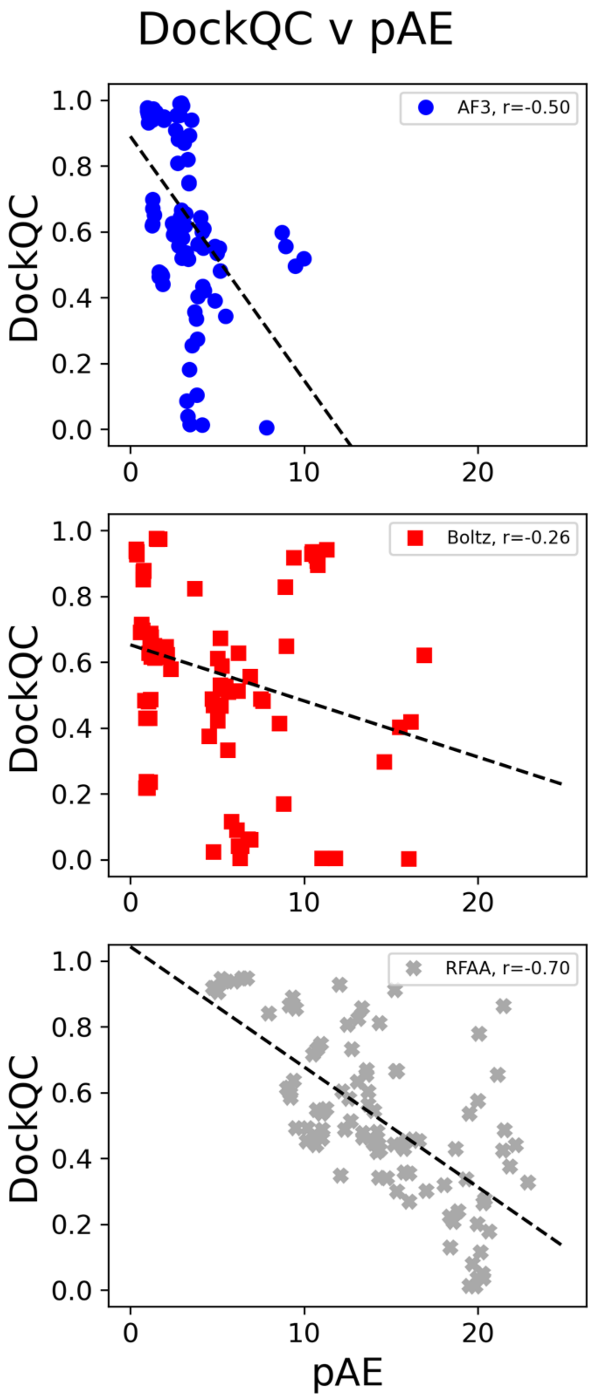


Figure 4.12: Comparison of pAE and DockQC accuracy on BCAPIN. AF3 (blue circle), Boltz-1 (red square), and RFAA (gray Diamond) were the only models which reported such metric.

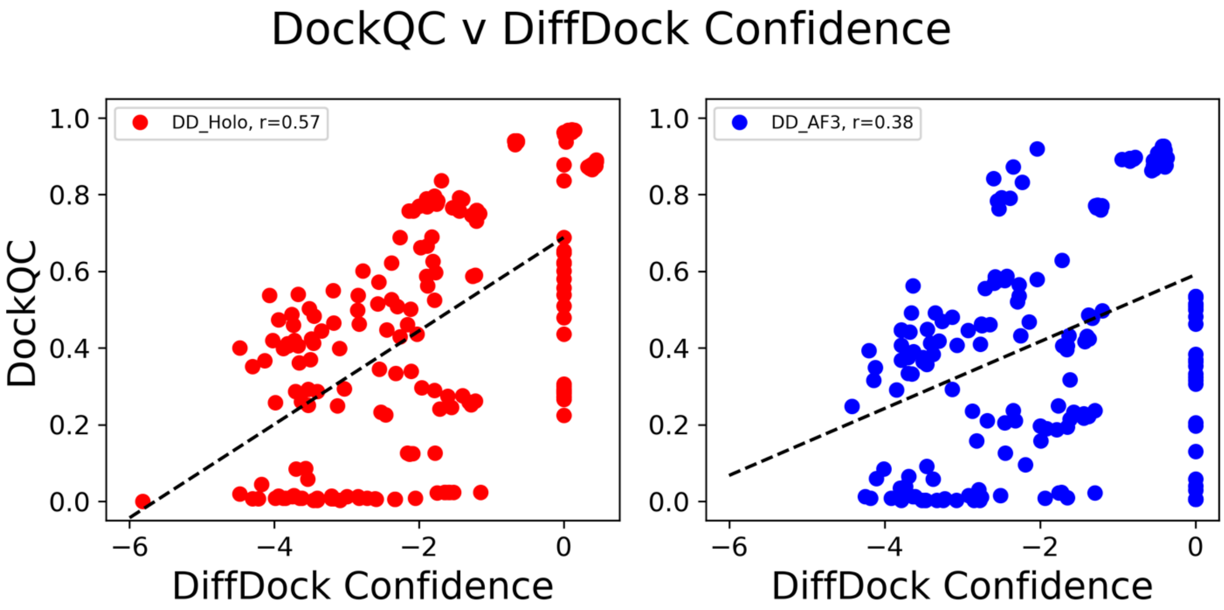


Figure 4.13: Comparison of pAE and DockQC accuracy on BCAPIN. AF3 (blue circle), Boltz-1 (red square), and RFAA (gray Diamond) were the only models which reported such metric.

BCAPIN: Full Set analysis

To identify any discrepancies between high quality and low quality protein-carbohydrate complexes, I analyzed the total 35 structures (20 high quality, 15 low quality) below. I found similar results to analyzing only the high quality structure predictions, all models perform strongly, capturing at least 80% of all top-5 complexes with at least acceptable quality DockQC.

Table 4.3: Benchmark of CArbohydrate Protein INteractions (BCAPIN) low quality proteins (RSCC < 0.9). The table lists the PDB 4-letter ID, protein name, UniProt ID, glycan input string for GlyLES, and any secondary ligands if present.

PDB	Protein Name	UniProt	GlyLES Input String	Secondary Ligand
7eqr	Chitoporin	L0RVU0	GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc	
7lvy	UDP-glycosyltransferase 203A2	T1K1R5	Glc	UDP
7p8g	Glucosyl-3-phosphoglycerate synthase	K5B7Z4	Glc	
7pgk	Hedgehog-interacting protein	Q96QV1	Heparin_analog: PDB SMILES used	
7pug	GH115	N/A	Xyl(b1-4)Xyl(b1-4)Xyl(b1-4)Xyl(b1-4)Xyl(b1-4)Xyl(b1-4)Xylb	
7toh	SGNH hydrolase	A0A5M4AV20	GlcA4Me(a1-2)Xylb	
7tvp	Ciral AMG chitosanase	Unk	GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc	
7vu1	Chitoporin	P75733	GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc	
7vwb	17 kDa phloem lectin	Q8LK69	Gal(b1-4)GlcNAca	
7xtn	α-1,3-mannosyl-glycoprotein 4-β-N-acetylglucosaminyltransferase A-like isoform X1	A0A6J2K041	GlcNAc	
7ywf	Dirigent protein	Q306J3	Gal(a1-3)Galb	
8ad2	Nictaba	Q94EW1	GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc(b1-4)GlcNAc	
8csf	WbbB D232C-Kdo adduct	Q6U8B0	Rha(a1-3)GlcNAcb	Ligand 2: GDP Ligand 3: Kdo
8edi	Netrin receptor unc-5	Q26261	Heparin_analog: PDB SMILES used	
8ped	Alginate lyase	A0A7I9C8Z1	ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManA(b1-4)ManAb	

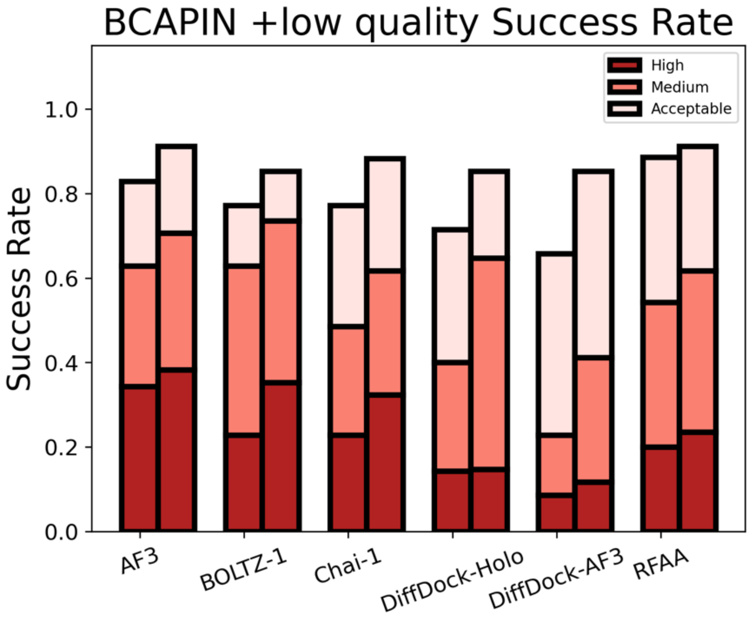


Figure 4.14: Analysis of the BCAPIN test set, including the low quality structures (RSCC < 0.9) (n=35). Top-1 (left) and top-5 (right) predictions of each method.

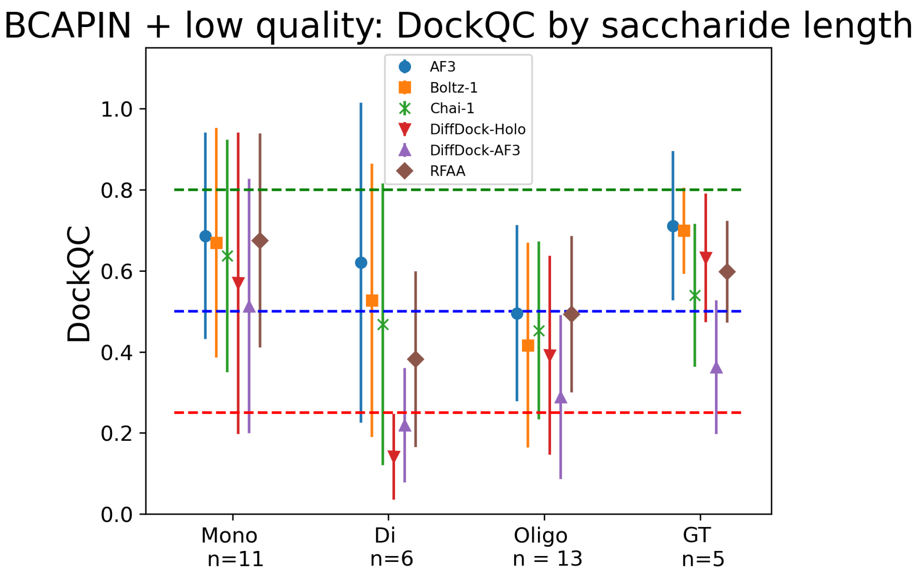


Figure 4.15: Comparison of average and standard deviation DockQC of predicted structures versus saccharide length for BCAPIN set with the low quality structures (RSCC < 0.9). I group saccharide length into a degree of polymerization (DP) of 1 (mono), 2 (di), and 3+ (oligo), and further group all glycosyltransferases (GTs) together that require multiple inputs (e.g. a saccharide and NTP) and with the number of proteins in each group listed. I also show the DockQC cutoffs between acceptable (red), medium (blue), and high (green) quality structures. I show the top-1 predictions for AF3 (blue circle), Boltz-1 (orange square), Chai-1 (Green X), Diffdock-*holo* (red triangle), Diffdock-*AF3* (purple triangle), and RFAA (brown diamond).

Chapter 5

Conclusion

As the primary means of intercellular communication, protein-carbohydrate interactions are critical to multicellular organism survival and proliferation; however, their weak, dynamic, and avidity-driven nature make them challenging to purify and experimentally study. Given these experimental difficulties, it is critical to also investigate carbohydrates and generate hypotheses with computational algorithms. However, prior to my dissertation work, computational methods for analyzing protein-carbohydrate interactions were notably limited.

My Contributions

I began my dissertation research at an inflection point in the field of biophysics. Rapid advancements in DL algorithms, 3D molecular representations, and computational hardware ushered in a data-driven wave of novel approaches for the analysis of biological data at unprecedented scales.[35,38] Leveraging these developments, I aimed to synthesize innovative structural glycobiology methods to uncover new biological phenomena. This dissertation details my efforts in developing and benchmarking deep learning tools to characterize the protein-sugar interactome at both the protein and residue levels.

Computational modeling of protein-carbohydrate complexes is a difficult task. The available structural data is sparse, often low-resolution, and sometimes lacking the full biological context of the bound glycan. To address these limitations, I systematically utilized all accessible data, while prioritizing experimentally solved, high-resolution structures for more critical data analysis.

CAPSIF is the first open and accessible structural deep learning algorithm designed to predict protein-carbohydrate interactions at a residue level. Dr. Sudhanshu Shanker and I trained two different models, a 3D voxel-based model (CAPSIF:V) and a 3D graph-based model (CAPSIF:G). We analyzed the capabilities of both models and demonstrated the power of these models by developing a *de novo* pipeline with AlphaFold2, CAPSIF, GlycanDock,[49] and Rosetta tools for protein-carbohydrate docking. Overall, CAPSIF achieved a 0.59 Matthews correlation coefficient (MCC); indicating strong performance with opportunities for future improvement.

To advance CAPSIF, I expanded the dataset and enhanced the input protein representation using ESM2, while also investigating the distinction between carbohydrate binding and non-binding proteins. I curated the Nonbinder and binder of CArbohydrate Protein interactions (NoCAP) dataset, which comprises experimentally determined carbohydrate binding proteins (with and without their associated ligand) as well as proteins presumed to not bind carbohydrates (such as small molecule binders, DNA binding proteins, cytoskeletal components).

Utilizing NoCAP, I trained two new models – CAPSIF2 and Protein interaction of CArbohydrates Predictor (PiCAP). CAPSIF2 surpassed both CAPSIF:V and CAPSIF:G across all performance metrics, achieving a 0.62 MCC on the original CAPSIF dataset and 0.57 MCC on the expanded dataset. PiCAP, however, predicts whether a protein has carbohydrate binding capabilities. PiCAP distinguishes carbohydrate-binding proteins with a 90% accuracy on experimentally solved structures and strongly correlates with results from high-throughput experiments, such as those profiling the ganglioside interactome. I then investigated how proteomes interact with carbohydrates, finding that PiCAP predicts 35-40% of proteins in the human, mouse, and *E. coli* proteomes bind to carbohydrates. I further analyzed the subcellular components, molecular functions, and biological processes of these predicted binding proteins, finding that 75% of human and mouse extracellular and cell surface proteins are predicted to bind carbohydrates.

While identification of carbohydrate-binding sites aids experimental design, predicting the full structure of protein-carbohydrate complexes is even more lucrative for hypothesis generation and testing. To this end, I benchmarked all atom structure prediction deep learning models, including AlphaFold 3, Boltz-1, Chai-1, DiffDock, and RosettaFold-AllAtom on their ability to predict non-covalently bound protein-carbohydrate complexes. To test these models, I first compiled a dataset of protein-carbohydrate complexes excluded from their training sets and evaluated the experimental fits of these structures. I also developed DockQC, a novel scalar metric (0 to 1) to quantitatively assess prediction accuracy relative to the native structure. In general, the tested models attained comparable performance in protein-carbohydrate docking with strong predictive capabilities on monosaccharides and disaccharides and reduced performance on oligosaccharides and multi-ligand targets. I then used AlphaFold3 to predict the protein-carbohydrate binding of several PiCAP predicted carbohydrate binding and non-binding human proteins, finding that further advances are necessary to enable high-confidence high-throughput predictions on the protein-sugar interactome.

Future Directions

The field of biophysics has changed dramatically since I began my dissertation. Like most other fields, biophysics became increasingly data-driven, leveraging the immense corpus of protein structures and sequences amassed over past decades. The most influential models in this field focused on solving general biological problems like protein structure prediction; however, there was limited focus on smaller subfields, such as glycobiology, where the data are much more sparse. I believe continued innovation in dataset curation, transfer learning of PiCAP/CAPSIF, creation of virtual glycan arrays, and new approaches to protein design will pave way to deeper insights in glycobiology.

Dataset Cleaning

In Chapter 4, I observed that the DIONYSUS dataset, despite listing 5,460 carbohydrate-binding proteins, includes significant redundancy: at 95% sequence homology DIONYSUS yields only 1,842 unique protein structures.[113] Many of these are slight sequence or ligand variants, and about 40% have low resolution glycan data (as measured by RSCC). Additionally, experimental limitations often prevent the capture of complete binding interfaces (Figure 5.1). In Chapter 4, I cleaned 105 structures to 35 representative structures; however, the remaining 5,355 structures are also in need of refinement and cleaning to serve as an appropriate training set and benchmark for future work.

Two example cases are provided in Figure 5.1 where data cleaning and in-filling are clear. Figure 5.1A shows two structures of a glucanase solved under similar conditions, where there is a near perfect overlap of the bound ligand between the structures, indicating the structures should be merged. Figure 5.1B shows a single structure of an alginate lyase with two bound mannose

(Man) trimer ligands; however, there is a clear pattern in the binding, indicating that computational infilling could provide a more complete bound ligand. These structures could be computationally infilled with classical techniques like Rosetta, or *de novo* predicted by AF3 or Boltz-2 and retained in the dataset when they have a high quality DockQC with the solved structure.

Further, one dataset ripe for *de novo* structure prediction is Carbohydrate Active enZYmes (CAZY) which lists all sequences of known carbohydrate binding enzymes, commonly alongside their bound ligands. Using the RSCC of crystal structures, DockQC of computationally merged/infilled/extrapolated predictions, and pLDDT of predicted CAZY proteins, researchers can construct a well classified dataset with quality assessments for proper training and testing of future models.

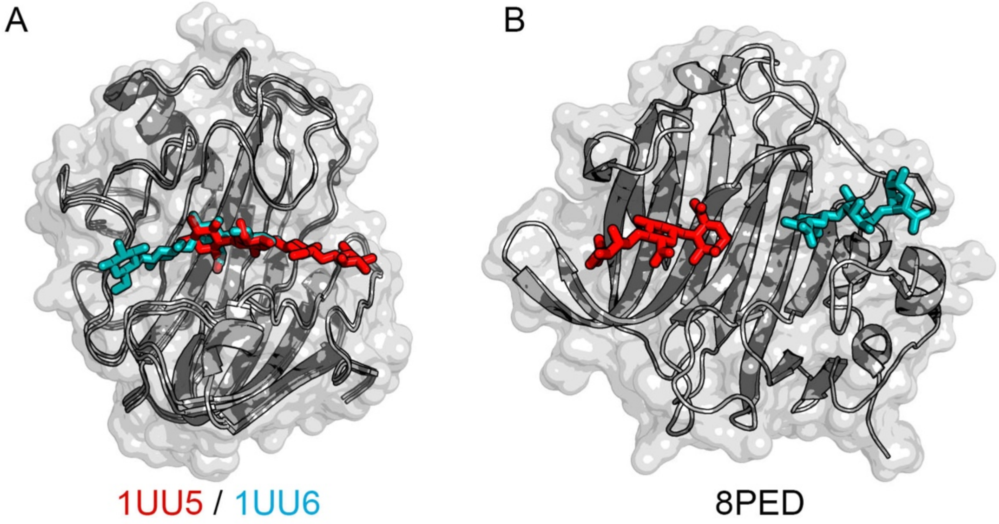


Figure 5.1: Examples of solved protein-carbohydrate structures requiring manual refinement. (A) 1UU5 (red) /1UU6 (teal) have the same bound carbohydrate but in different overlapping locations. (B) 8PED binds two of the same trisaccharide suggesting a continuous pocket that binds an oligosaccharide DP 8 that is not experimentally solved.

Sialic Acid PiCAP and CAPSIF

Gangliosides and sialic acids are incredibly important to most mammalian systems, being critical to neural development and proper immune signaling. In chapter 3, I validated PiCAP against the ganglioside interactome work of Zhang et al. I next suggest the development of PiCAP and CAPSIF2 models specific to sialic acids. These models perform encouragingly on generalized sugars, and with a single targeted training step of transfer learning, these models can be fine-tuned to predict on protein-sialic acid interactions. Sialic acid is one of the most unique saccharides, containing nine carbons, an acetyl group, and carboxyl group (Figure 1.2), and therefore likely the easiest saccharide signal for a model to specifically understand on an individual basis.

To create a sialic acid-specific PiCAP/CAPSIF2, we would leverage the DR/NoCAP dataset (Table 3.1) for negative samples and select all sialic acid containing structures in the PDB (n=616) to provide positive samples. A single transfer learning step with the current training/testing scheme would provide the foundation, with a similar analysis. Naturally, problems may emerge since 616 proteins is a very limited number, especially before filtering for sequence redundancy; however, with the litany of glycan arrays and the ganglioside interactome work, external validation can be attained.

Virtual Glycan Arrays

Glycan arrays are widely used to qualitatively probe protein-carbohydrate interactions by presenting a protein to diverse glycans.[21] Current computational methods for glycan arrays are LectinOracle and GlyNet. LectinOracle uses a SweetNet glycan embedding and an ESM lectin embedding to predict if a lectin binds to the given glycan.[48] GlyNet rather predicts that for a given glycan, which of 352 select lectins binds the glycan.[85] Both methods however are constrained by the limited set of known lectins.

PiCAP predicts 35% of the human proteome could bind carbohydrates, where only 2.5% of those predicted binders are known lectins, underscoring the need for algorithms capable of predicting specific protein-carbohydrate binding beyond the current scope of lectins. For this, I propose the development of a virtual glycan array. A virtual glycan array would be a neural network model trained to identify the glycan binding epitopes of a provided input protein. A novel algorithm to perform virtual glycan arrays would leverage input data from lectins, native sugar sensors, and the incoming influx of data from high throughput photoaffinity tag pulldown assays. Such an algorithm could concatenate GIFFLAR[142] glycan embeddings, ESM3[150] protein sequence embeddings, and AlphaFold3 [42] *apo* structures to predict the binding profiles of proteins. This proposed method could then be applied to PiCAP predictions of the human protein-sugar interactome to unveil the specific interactions for direct experimental validation.

Design

My dissertation work has been focused on elucidating the protein-sugar interactome. A natural next step with these interactions identified is protein design targeting glycoproteins, such as viral spike proteins. Although limited experimental data hamper current deep learning approaches, iterative development and application of computational and experimental strategies will continually improve all aspects of protein therapeutic design. I envision protein design can occur in three different domains to further glycobiology: glycosylation, binding site design, and protein design.

Glycosylation

Glycosylation, the enzymatic attachment process of carbohydrates to biomolecules, is central to protein function. Reliably predicting N-linked and O-linked glycosylation sites on ‘omic scales, as well as *which* glycans are present at those sites, would facilitate experimental procedures and protein design. Although tools are beginning to emerge in this field, there is substantial room for improvement, as with all introductory techniques.[68] For example, the INSANNE model predicts glycosylation of only human proteins in human systems;[7] however, does not account for what expression system, nor that typical human therapeutic expression happens in Chinese hamster ovary (CHO) cells.[151] Such predictive capabilities would enable *de novo* design of specific glycosylation on glycoproteins, thus advancing researchers’ ability to perform experiments and uncover more protein-glycan interactions.

CAPSIF Site and Design

Currently, CAPSIF models provide binary classification of carbohydrate-binding sites without confidence metrics or specificity to carbohydrate type. I believe that a next-generation CAPSIF model should aim to predict not just the presence, but the identity of carbohydrate ligands. More specifically, the model should leverage an improved loss function (binary cross entropy (BCE) instead of Dice) and more granular binding site categories (e.g. Hex, HexNAc, Sia). Achieving this resolution would enable “hallucination” design strategies[152] to engineer proteins with high-confidence experimentally testable binding sites using virtual and *in vitro* glycan arrays.

De Novo Protein Design

While protein structure prediction has been revolutionized by deep learning, the same strides are being made in the world of *de novo* protein design. A fine-tuned algorithm for *de novo* design of proteins for carbohydrate ligands would be incredibly lucrative: as it could assist in probing cell/tissue glycosylation patterns, cellular targeting of therapeutics, development of antibodies to emerging viruses, and components for tissue engineering.

Current techniques for *de novo* protein design began with a slightly round-about methodology. To design a protein to bind a target protein, a researcher would use a diffusion model, such as RFDiffusion[153] to generate a protein backbone, which would then be fed into ProteinMPNN[154] to predict a protein sequence without the given binding context, and then validating the structure with AF2. Although the external use of AF2 provided some validation; the compounding biases and errors of no context-ProteinMPNN sequence predictions from RFDiffusion yielded low hit rates. Now, as these methods have been maturing, more context has been incorporated by new models like RFDiffusion All-atom[131] and LigandMPNN[155].

Creating a singular end-to-end DL algorithm for designing proteins to bind small molecules could streamline the process and reduce cumulative model error. And with this developed model, one could then transfer learn and fine-tune the model to protein-carbohydrate interactions on a high-quality curated dataset, as envisioned in the Dataset Cleaning section. With these models trained and validated against *de novo* protein-carbohydrate design, researchers can then expand to bind novel glycoprotein motifs and epitopes with therapeutic relevance to more quickly respond to future pathogenic outbreaks, such as the one that my PhD began with.

Chapter 6

References

1. Koonin, E. V. Why the Central Dogma: on the nature of the great biological exclusion principle. *Biol Direct* **10**, 52 (2015).
2. Canner, S. W., Feller, S. E. & Wassall, S. R. Molecular Organization of a Raft-like Domain in a Polyunsaturated Phospholipid Bilayer: A Supervised Machine Learning Analysis of Molecular Dynamics Simulations. *J Phys Chem B* **125**, 13158–13167 (2021).
3. Leng, X. *et al.* All n-3 PUFA are not the same: MD simulations reveal differences in membrane organization for EPA, DHA and DPA. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1860**, 1125–1134 (2018).
4. Wassall, S. R. *et al.* Docosahexaenoic acid regulates the formation of lipid rafts: A unified view from experiment and simulation. *Biochimica et Biophysica Acta (BBA) - Biomembranes* **1860**, 1985–1993 (2018).
5. Chandel, N. S. Carbohydrate Metabolism. *Cold Spring Harb Perspect Biol* **13**, (2021).
6. Varki, A. *et al. Essentials of Glycobiology*. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2022).
7. Kellman, B. P. *et al.* Decoding glycosylation potential from protein structure across human glycoproteins with a multi-view recurrent neural network. *bioRxiv* (2024) doi:10.1101/2024.05.15.594334.

8. Henrissat, B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J* **280 (Pt 2)**, 309–16 (1991).

9. Henrissat, B. & Davies, G. Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* **7**, 637–644 (1997).

10. Martinez, M. *et al.* Quantitative Proteomics Reveals that the OGT Interactome Is Remodeled in Response to Oxidative Stress. *Mol Cell Proteomics* **20**, (2021).

11. Narayanan, B. *et al.* Differential Detection of O-GlcNAcylated proteins in the heart using antibodies. *Anal Biochem* **678**, 115262 (2023).

12. De Schutter, K. & Van Damme, E. Protein-Carbohydrate Interactions as Part of Plant Defense and Animal Immunity. *Molecules* **20**, 9029–9053 (2015).

13. Kim, H.-J. *et al.* Blood monocyte-derived CD169+ macrophages contribute to antitumor immunity against glioblastoma. *Nat Commun* **13**, 6211 (2022).

14. Hartnell, A. *et al.* Characterization of human sialoadhesin, a sialic acid binding receptor expressed by resident and inflammatory macrophage populations. *Blood* **97**, 288–296 (2001).

15. Zhang, Q. *et al.* Heparan sulfate assists SARS-CoV-2 in cell entry and can be targeted by approved drugs in vitro. *Cell Discov* **6**, 80 (2020).

16. Bonnardel, F., Mariethoz, J., Pérez, S., Imberty, A. & Lisacek, F. LectomeXplore, an update of UniLectin for the discovery of

carbohydrate-binding proteins based on a new lectin classification. *Nucleic Acids Res* **49**, D1548–D1554 (2021).

17. Zhang, G.-L. *et al.* The Human Ganglioside Interactome in Live Cells Revealed Using Clickable Photoaffinity Ganglioside Probes. *J Am Chem Soc* **146**, 17801–17816 (2024).

18. Nagae, M. *et al.* Crystal Structure of Anti-polysialic Acid Antibody Single Chain Fv Fragment Complexed with Octasialic Acid. *Journal of Biological Chemistry* **288**, 33784–33796 (2013).

19. Smith, S. T., Shub, L. & Meiler, J. PlaceWaters: Real-time, explicit interface water sampling during Rosetta ligand docking. *PLoS One* **17**, (2022).

20. Hudson, K. L. *et al.* Carbohydrate–Aromatic Interactions in Proteins. *J Am Chem Soc* **137**, 15152–15160 (2015).

21. Oyelaran, O. & Gildersleeve, J. C. Glycan arrays: recent advances and future challenges. *Curr Opin Chem Biol* **13**, 406–413 (2009).

22. Babulic, J. L., De León González, F. V. & Capicciotti, C. J. Recent advances in photoaffinity labeling strategies to capture Glycan–Protein interactions. *Curr Opin Chem Biol* **80**, 102456 (2024).

23. Canner, S. W., Schnaar, R. L. & Gray, J. J. Predictions from Deep Learning Propose Substantial Protein-Carbohydrate Interplay. Preprint at https://doi.org/10.1101/2025.03.07.641884 (2025).

24. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

25. Alzubaidi, L. *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* **8**, 53 (2021).

26. Sarker, I. H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput Sci* **2**, 420 (2021).

27. Sclocchi, A. & Wyart, M. On the different regimes of stochastic gradient descent. *Proceedings of the National Academy of Sciences* **121**, (2024).

28. Dubey, S. R., Singh, S. K. & Chaudhuri, B. B. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* **503**, 92–108 (2022).

29. Labach, A., Salehinejad, H. & Valaee, S. Survey of Dropout Methods for Deep Neural Networks. *ArXiv* (2019).

30. Ba, J. L., Kiros, J. R. & Hinton, G. E. Layer Normalization. *ArXiv* (2016).

31. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (2015).

32. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A. G. Averaging Weights Leads to Wider Optima and Better Generalization. *ArXiv* (2018).

33. Indolia, S., Goswami, A. K., Mishra, S. P. & Asopa, P. Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach. *Procedia Comput Sci* **132**, 679–688 (2018).

34. Fei, J. & Deng, Z. Rotation invariance and equivariance in 3D deep learning: a survey. *Artif Intell Rev* **57**, 168 (2024).

35. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) Equivariant Graph Neural Networks. *Proceedings of the 38 th International Conference on Machine Learning* **139**, (2021).

36. Ingraham, J., Garg, V. K., Barzilay, R. & Jaakkola, T. Generative Models for Graph-Based Protein Design. *Adv Neural Inf Process Syst* (2019).

37. Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun* **13**, 2453 (2022).

38. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

39. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science (1979)* **379**, 1123–1130 (2023).

40. The Royal Swedish Academy of Sciences. *The Nobel Prize in Chemistry 2024*. (2024).

41. Jumper, J. *et al.* Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function, and Bioinformatics* **89**, 1711–1721 (2021).

42. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

43. Corso, G., Stärk, H., Jing, B., Barzilay, R. & Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *ICLR* (2023).

44. Pakhrin, S. C. *et al.* LMNglyPred: prediction of human N-linked glycosylation sites using embeddings from a pre-trained protein language model. *Glycobiology* **33**, 411–422 (2023).

45. Pakhrin, S. C. *et al.* Prediction of human O-linked glycosylation sites using stacked generalization and embeddings from pre-trained protein language model. *Bioinformatics* **40**, (2024).

46. Alocci, D. *et al.* GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. *J Proteome Res* **18**, 664–677 (2019).

47. Burkholz, R., Quackenbush, J. & Bojar, D. Using graph convolutional neural networks to learn a representation for glycans. *Cell Rep* **35**, 109251 (2021).

48. Lundstrøm, J., Korhonen, E., Lisacek, F. & Bojar, D. LectinOracle: A Generalizable Deep Learning Model for Lectin–Glycan Binding Prediction. *Advanced Science* **9**, (2022).

49. Nance, M. L., Labonte, J. W., Adolf-Bryfogle, J. & Gray, J. J. Development and Evaluation of GlycanDock: A Protein–Glycoligand Docking Refinement Algorithm in Rosetta. *J Phys Chem B* **125**, 6807–6820 (2021).

50. Forli, S. *et al.* Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nat Protoc* **11**, 905–919 (2016).

51. Kato, K. & Ishiwa, A. The Role of Carbohydrates in Infection Strategies of Enteric Pathogens. *Trop Med Health* **43**, 41–52 (2015).

52. Karlsson, K.-A. Pathogen-Host Protein-Carbohydrate Interactions as the Basis of Important Infections. in 431–443 (2001).

53. Dyason, J. C. & von Itzstein, M. Viral surface glycoproteins in carbohydrate recognition. in *Microbial Glycobiology* 269–283 (Elsevier, 2010).

54. Haji-Ghassemi, O., Blackler, R. J., Martin Young, N. & Evans, S. v. Antibody recognition of carbohydrate epitopes. *Glycobiology* **25**, 920–952 (2015).

55. Kappler, K. & Hennet, T. Emergence and significance of carbohydrate-specific antibodies. *Genes and Immunity* vol. 21 224–239 Preprint at https://doi.org/10.1038/s41435-020-0105-9 (2020).

56. Funderburgh, J. L. Keratan sulfate: structure, biosynthesis, and function. *Glycobiology* **10**, 951–958 (2000).

57. Yip, G. W., Smollich, M. & Götte, M. Therapeutic value of glycosaminoglycans in cancer. *Mol Cancer Ther* **5**, 2139–2148 (2006).

58. Angata, K. *et al.* Polysialic Acid-Directed Migration and Differentiation of Neural Precursors Are Essential for Mouse Brain Development. *Mol Cell Biol* **27**, 6659–6668 (2007).

59. Lu, W. & Pieters, R. J. Carbohydrate–protein interactions and multivalency: implications for the inhibition of influenza A virus infections. *Expert Opin Drug Discov* **14**, 387–395 (2019).

60. Seabright, G. E., Doores, K. J., Burton, D. R. & Crispin, M. Protein and Glycan Mimicry in HIV Vaccine Design. *Journal of Molecular Biology* vol. 431 2223–2247 Preprint at https://doi.org/10.1016/j.jmb.2019.04.016 (2019).

61. Kieber-Emmons, T., Saha, S., Pashov, A., Monzavi-Karbassi, B. & Murali, R. Carbohydrate-mimetic peptides for pan anti-tumor responses. *Frontiers in Immunology* vol. 5 Preprint at https://doi.org/10.3389/fimmu.2014.00308 (2014).

62. Del, M. *et al. Protein-Carbohydrate Interactions Studied by NMR: From Molecular Recognition to Drug Design*. *Current Protein and Peptide Science* vol. 13 (2012).

63. Hao, D. *et al.* Mechanism of Glycans Modulating Cholesteryl Ester Transfer Protein: Unveiled by Molecular Dynamics Simulation. *J Chem Inf Model* acs.jcim.1c00233 (2021) doi:10.1021/acs.jcim.1c00233.

64. Crawford, C. J. *et al.* A glycan FRET assay for detection and characterization of catalytic antibodies to the Cryptococcus neoformans capsule. (2020) doi:10.1073/pnas.2016198118/-/DCSupplemental.

65. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat Commun* **14**, 2389 (2023).

66. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Improving detection of protein-ligand binding sites with 3D segmentation. *Sci Rep* **10**, (2020).

67. Sverrisson, F., Feydy, J., Correiácorreiá, B. E. & Bronstein, M. M. Fast end-to-end learning on protein surfaces. doi:10.1101/2020.12.28.424589.

68. Li, M. *et al.* Shotgun scanning glycomutagenesis: A simple and efficient strategy for constructing and characterizing neoglycoproteins. *Proceedings of the National Academy of Sciences* **118**, (2021).

69. Xie, Z. R. & Hwang, M. J. Methods for predicting protein–ligand binding sites. *Methods in Molecular Biology* **1215**, 383–398 (2015).

70. McGreig, J. E. *et al.* 3DLigandSite: structure-based prediction of protein–ligand binding sites. *Nucleic Acids Res* **50**, W13–W20 (2022).

71. le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).

72. Kozakov, D. *et al.* The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc* **10**, 733–755 (2015).

73. Mylonas, S. K., Axenopoulos, A. & Daras, P. DeepSurf: a surface-based deep learning approach for the prediction of ligand binding sites on proteins. *Bioinformatics* **37**, 1681–1690 (2021).

74. Kandel, J., Tayara, H. & Chong, K. T. PUResNet: prediction of protein-ligand binding sites using deep residual neural network. *J Cheminform* **13**, (2021).

75. Evans, D. J. *et al.* Finding Druggable Sites in Proteins Using TACTICS. *J Chem Inf Model* **61**, 2897–2910 (2021).

76. Taroni, C., Jones, S. & Thornton, J. M. Analysis and prediction of carbohydrate binding sites. *Protein Engineering, Design and Selection* **13**, 89–98 (2000).

77. Malik, A. & Ahmad, S. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Struct Biol* **7**, (2007).

78. Kulharia, M., Bridgett, S. J., Goody, R. S. & Jackson, R. M. InCa-SiteFinder: A method for structure-based prediction of inositol and carbohydrate binding sites on proteins. *J Mol Graph Model* **28**, 297–303 (2009).

79. Tsai, K.-C. *et al.* Prediction of Carbohydrate Binding Sites on Protein Surfaces with 3-Dimensional Probability Density Distributions of Interacting Atoms. *PLoS One* **7**, e40846 (2012).

80. Zhao, H., Yang, Y., von Itzstein, M. & Zhou, Y. Carbohydrate-binding protein identification by coupling structural similarity searching with binding affinity prediction. *J Comput Chem* **35**, 2177–2183 (2014).

81. Taherzadeh, G., Zhou, Y., Liew, A. W.-C. & Yang, Y. Sequence-Based Prediction of Protein–Carbohydrate Binding Sites Using Support Vector Machines. *J Chem Inf Model* **56**, 2115–2122 (2016).

82. Bonnardel, F. *et al.* UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acids Res* **47**, D1236–D1244 (2019).

83. Siva Shanmugam, N. R., Jino Blessy, J., Veluraja, K. & Michael Gromiha, M. ProCaff: protein–carbohydrate complex binding affinity database. *Bioinformatics* **36**, 3615–3617 (2020).

84. Ernst, B. & Magnani, J. L. From carbohydrate leads to glycomimetic drugs. *Nat Rev Drug Discov* **8**, 661–677 (2009).

85. Carpenter, E. J., Seth, S., Yue, N., Greiner, R. & Derda, R. GlyNet: a multi-task neural network for predicting protein–glycan interactions. *Chem Sci* **13**, 6669–6686 (2022).

86. Hauser, M., Steinegger, M. & Söding, J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics* **32**, 1323–1330 (2016).

87. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences* **117**, 1496–1503 (2020).

88. Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **3**, (2022).

89. Du, Z. *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc* **16**, 5634–5651 (2021).

90. Clark, J. J., Benson, M. L., Smith, R. D. & Carlson, H. A. Inherent versus induced protein flexibility: Comparisons within and between apo and holo structures. *PLoS Comput Biol* **15**, (2019).

91. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132 (1982).

92. Hirano, A. & Kameda, T. *Aromaphilicity Index* of Amino Acids: Molecular Dynamics Simulations of the Protein Binding Affinity for Carbon Nanomaterials. *ACS Appl Nano Mater* **4**, 2486–2495 (2021).

93. Satorras, V. G., Hoogeboom, E. & Welling, M. E(n) Equivariant Graph Neural Networks. *Proceedings of the 38th International Conference on Machine Learning (PMLR)* **139**, 9323–9332 (2021).

94. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679–682 (2022).

95. Tyka, M. D. *et al.* Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *J Mol Biol* **405**, 607–618 (2011).

96. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**, (2021).

97. Jones, D. *et al.* Improved Protein–Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *J Chem Inf Model* **61**, 1583–1592 (2021).

98. Yang, D. *et al.* G protein-coupled receptors: structure- and function-based drug discovery. *Signal Transduct Target Ther* **6**, 7 (2021).

99. Dingjan, T. *et al.* Structural biology of antibody recognition of carbohydrate epitopes and potential uses for targeted cancer immunotherapies. *Mol Immunol* **67**, 75–88 (2015).

100. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun* **14**, 2175 (2023).

101. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–91 (2010).

102. Villar, S., Hogg, D. W., Storey-Fisher, K., Yao, W. & Blum-Smith, B. Scalars are universal: Equivariant machine learning, structured like classical physics. in *Advances in Neural Information Processing Systems* (eds. Ranzato, M., Beygelzimer,

A., Dauphin, Y., Liang, P. S. & Vaughan, J. W.) vol. 34 28848–28863 (Curran Associates, Inc., 2021).

103. Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. in *Lecture Notes in Computer Science* 3–11 (Springer, Charm, 2018). doi:10.1007/978-3-030-00889-5_1.

104. Lazarus, M. B., Nam, Y., Jiang, J., Sliz, P. & Walker, S. Structure of human O-GlcNAc transferase and its complex with a peptide substrate. *Nature* **469**, 564–567 (2011).

105. Zhang, S., Chen, K. Y. & Zou, X. Carbohydrate-protein interactions: advances and challenges. *Commun Inf Syst* **21**, 147–163 (2021).

106. Canner, S. W., Shanker, S. & Gray, J. J. Structure-based neural network protein–carbohydrate interaction predictions at the residue level. *Frontiers in Bioinformatics* **3**, (2023).

107. Jänes, J. & Beltrao, P. Deep learning for protein structure prediction and design—progress and applications. *Mol Syst Biol* **20**, 162–169 (2024).

108. Varadi, M. *et al.* AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* **52**, D368–D375 (2024).

109. Zhang, F. *et al.* MetalNet2: an enhanced server for predicting metal-binding sites in proteomes. *Natl Sci Rev* **11**, (2024).

110. Bibekar, P., Krapp, L. & Peraro, M. D. PeSTo-Carbs: Geometric Deep Learning for Prediction of Protein–Carbohydrate Binding Interfaces. *J Chem Theory Comput* **20**, 2985–2991 (2024).

111. He, X. *et al.* Highly accurate carbohydrate-binding site prediction with DeepGlycanSite. *Nat Commun* **15**, 5163 (2024).

112. Imberty, A., Bonnardel, F. & Lisacek, F. UniLectin, A One-Stop-Shop to Explore and Study Carbohydrate-Binding Proteins. *Curr Protoc* **1**, (2021).

113. Gheeraert, A. *et al.* DIONYSUS: a database of protein–carbohydrate interfaces. *Nucleic Acids Res* (2024) doi:10.1093/nar/gkae890.

114. Madani, A. *et al.* Large language models generate functional protein sequences across diverse families. *Nat Biotechnol* **41**, 1099–1106 (2023).

115. Luo, Y. & Parmeggiani, F. CLIMBS: assessing Carbohydrate-Protein interactions through a graph neural network classifier using synthetic negative data. *bioRxiv* Preprint at https://doi.org/10.1101/2025.02.27.640667 (2025).

116. Dunbar, J. *et al.* SAbDab: the structural antibody database. *Nucleic Acids Res* **42**, D1140–D1146 (2014).

117. Liu, Z. *et al.* Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions. *Acc Chem Res* **50**, 302–309 (2017).

118. Norambuena, T. & Melo, F. The Protein-DNA Interface database. *BMC Bioinformatics* **11**, 262 (2010).

119. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

120. Schnider, B. *et al.* HumanLectome, an update of UniLectin for the annotation and prediction of human lectins. *Nucleic Acids Res* **52**, D1683–D1693 (2024).

121. Mellacheruvu, D. *et al.* The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nat Methods* **10**, 730–736 (2013).

122. Mikkelsen, S. A., Vangheluwe, P. & Andersen, J. P. A Darier disease mutation relieves kinetic constraints imposed by the tail of sarco(endo)plasmic reticulum Ca2+-ATPase 2b. *Journal of Biological Chemistry* **293**, 3880–3889 (2018).

123. Cenci, C. *et al.* Down-regulation of RNA Editing in Pediatric Astrocytomas. *Journal of Biological Chemistry* **283**, 7251–7260 (2008).

124. Tan, X. *et al.* Loss of Smad4 promotes aggressive lung cancer metastasis by de-repression of PAK3 via miRNA regulation. *Nat Commun* **12**, 4853 (2021).

125. Cassandri, M. *et al.* Zinc-finger proteins in health and disease. *Cell Death Discov* **3**, 17071 (2017).

126. Mi, H. *et al.* Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nat Protoc* **14**, 703–721 (2019).

127. Thomas, P. D. *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* **31**, 8–22 (2022).

128. Wohlwend, J. *et al.* Boltz-1 Democratizing Biomolecular Interaction Modeling. Preprint at https://doi.org/10.1101/2024.11.19.624167 (2024).

129. Lin, C.-L., Carpenter, E. J., Li, T., Ahmed, T. & Derda, R. Liquid Glycan Array. in *Phage Engineering and Analysis* 143–159 (2024). doi:10.1007/978-1-0716-3798-2_10.

130. Lima, G. M. *et al.* The liquid lectin array detects compositional glycocalyx differences using multivalent DNA-encoded lectins on phage. *Cell Chem Biol* (2024) doi:10.1016/j.chembiol.2024.09.010.

131. Krishna, R. *et al.* Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science (1979)* **384**, (2024).

132. Watson, Z. L. *et al.* Structure of the bacterial ribosome at 2 Å resolution. *Elife* **9**, (2020).

133. Berman, H. M. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).

134. Bateman, A. *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **51**, D523–D531 (2023).

135. Corso, G. *et al.* Deep Confident Steps to New Pockets: Strategies for Docking Generalization. (2024).

136. Boitreaud, J. *et al.* Chai-1: Decoding the molecular interactions of life. Preprint at https://doi.org/10.1101/2024.10.10.615955 (2024).

137. Buttenschoen, M., Morris, G. M. & Deane, C. M. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci* **15**, 3130–3139 (2024).

138. Fontana, C. & Widmalm, G. Primary Structure of Glycans by NMR Spectroscopy. *Chem Rev* **123**, 1040–1102 (2023).

139. Smart, O. S. *et al.* Validation of ligands in macromolecular structures determined by X-ray crystallography. *Acta Crystallogr D Struct Biol* **74**, 228–236 (2018).

140. Ranaudo, A., Giulini, M., Pelissou Ayuso, A. & Bonvin, A. M. J. J. Modeling Protein–Glycan Interactions with HADDOCK. *J Chem Inf Model* **64**, 7816–7825 (2024).

141. Gheeraert, A., Guyon, F., Pérez, S. & Galochkina, T. Unraveling the diversity of protein-carbohydrate interfaces: Insights from a multi-scale study. *Carbohydr Res* **550**, 109377 (2025).

142. Joeres, R. & Bojar, D. Higher-Order Message Passing for Glycan Representation Learning. (2024).

143. Thomès, L., Burkholz, R. & Bojar, D. Glycowork: A Python package for glycan data science and machine learning. *Glycobiology* **31**, 1240–1244 (2021).

144. Joeres, R., Bojar, D. & Kalinina, O. V. GlyLES: Grammar-based Parsing of Glycans from IUPAC-condensed to SMILES. *J Cheminform* **15**, 37 (2023).

145. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Res* **53**, D1516–D1525 (2025).

146. Meli, R. & Biggin, P. C. spyrmsd: symmetry-corrected RMSD calculations in Python. *J Cheminform* **12**, 49 (2020).

147. Landrum, G. RDKit: Open-source cheminformatics. *RDKit* (2006).

148. Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **11**, e0161879 (2016).

149. Mirabello, C. & Wallner, B. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics* **40**, (2024).

150. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science (1979)* **387**, 850–858 (2025).

151. Xu, X. *et al.* The genomic sequence of the Chinese hamster ovary (CHO)-K1 cell line. *Nat Biotechnol* **29**, 735–741 (2011).

152. Mahajan, S. P., Ruffolo, J. A., Frick, R. & Gray, J. J. Hallucinating structure-conditioned antibody libraries for target-specific binders. *Front Immunol* **13**, (2022).

153. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

154. Dauparas, J. *et al.* Robust deep learning–based protein sequence design using ProteinMPNN. *Science (1979)* **378**, 49–56 (2022).

155. Kuhlman, B. *et al.* Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science (1979)* **302**, 1364–1368 (2003).

SAMUEL WILLIAM CANNER

scanner1@jhu.edu

Johns Hopkins University (JHU) PhD candidate developing and applying deep learning methods to study protein-carbohydrate interactions on cellular scales.

WORK EXPERIENCE

JHU BIOPHYSICS RESEARCH: JEFFREY J. GRAY LAB **Baltimore, MD**

PhD Candidate *August 2021 – Current*

- Develop deep learning architectures to predict protein-carbohydrate binding interfaces
- Model protein-carbohydrate and protein-protein interactions using Rosetta software and deep learning
- Utilize a suite of software to pre-process and organize 30,000+ crystal protein structures
- Analyze six organism proteomes in terms of protein-carbohydrate interactions to assess biological functionality
- Dock carbohydrates to proteins using the Rosetta software to gain physiological insight
- Benchmark and evaluate protein-small molecule complex predictions across multitude of DL models
- Predict critical residues implicated in enzymatic activity of antibody-carbohydrate complexes
- Collaborate with experimental colleagues to design experimental protocols
- Work alongside the lab to predict bound protein-protein complexes in the CASP15-CAPRI competition

AUGMENT BIOLOGICS **Remote**

Founding AI Scientist *February 2025 – Current*

- Migrate academic code into a server-based platform (Modal, AWS S3)
- Implement unit tests and API calls, architectures, and benchmarks for CI/CD
- Integrate LLM approaches to predict protein glycosylation patterns
- Perform hyper parameterization and compare benchmark capabilities
- Perform fundamental R&D of the DL algorithm to better capture the scientific goals of glycosylation prediction

IUPUI BIOPHYSICS RESEARCH: STEVEN R. WASSALL LAB **Indianapolis, IN**

Researcher *August 2016 – May 2020*

- Create and run lipid membrane Molecular Dynamics (MD) simulations on high performance computing (HPC) clusters
- Use all atom (AA) and coarse grained (CG) approaches to determine molecular interactions
- Perform umbrella sampling simulations to determine the binding of molecules to lipid membranes
- Parameterize new computational models of molecules
- Run fluorescence spectroscopy experiments to validate computational results
- Mentor students in both computational and experimental attributes of laboratory work
- Secured an allocation on the ANTON2 supercomputer through a competitive PSC grant
- Presented at numerous IUPUI campus research poster sessions

REGENERON: PROTEIN BIOCHEMISTRY INTERNSHIP **Tarrytown, NY**

Researcher *May 2019 – August 2019*

- Perform modulated differential scanning calorimetry (mDSC) on antibodies
- Analyze mDSC to identify thermostability of drug candidates
- Run MD simulations to develop an *in situ* method to screen antibody candidates

BIOCOMP TEACHER **Baltimore, MD**

High School Bootcamp Teacher *July 2023*

- Collaborate alongside a colleague to create and run a high school program for underrepresented students in STEM
- Create lecture materials to cover the fundamentals molecular biology
- Create python notebooks to assist students in learning the fundamentals of coding
- Organize two weeks of coursework and material with a final project of *ab initio* folding of unique proteins
- Outreach to all JHU departments and industry affiliates for guest lectures and discussions on different research paths
- Apply and receive funding from multiple internal JHU and Rosetta Community sources
- Create outreach materials to present and send to different organizations in the Baltimore City area
- Photograph the event to create materials for future years

TEACHING ASSISTANT, TUTOR **Indianapolis, IN and Baltimore, MD**

Tutor and Teaching Assistant *August 2018 – May 2020; July - October 2023*

- Led recitations and individual support for 250.649: Introduction to Computing in Biology, helping students strengthen computational and biological analysis
- Aid students in introductory physics classes, requiring strong understanding of physics material and good communication to explain difficult concepts
- Assist students in PHYS 29900: Introduction to Computational Physics, requiring both solid foundation in computational tools and physics understanding

FREELANCE PHOTOGRAPHER **Indianapolis, IN and Baltimore, MD**

Photographer *August 2018 – Current*

- Event photography for events ranging from weddings, parties, dances, concerts, and community gatherings
- Consistently photograph for the Baltimore City Waterfront Partnership's events, notably Baltimore by Baltimore
- Photograph for Baltimore City Downtown Sailing Center (DSC)'s Regatta

COMPUTATIONAL LANGUAGE AND TOOLS PROFICIENCY

C, C++, Python (PyTorch, TensorFlow), Tcl/Tk, Java, MATLAB, R, LaTeX, Node.js, React.js, Linux/Unix, Modal

PUBLICATIONS

- **Canner, S.W.,** Kelley, P., Phillips, A.Q., Feller, S.E., Wassall, S.R. 2025. α-Tocopherol and a Polyunsaturated Phospholipid Prefer Each Other's Company in Mixed Membranes with Raft-forming Sphingomyelin and Cholesterol: MD Simulations. *J. Phys Chem B.*
  - https://doi.org/10.1021/acs.jpcb.5c05553
- **Canner, S.W.,** Lu, L., Takeshita, S.S., Gray, J.J. 2015. Evaluation of De Novo Deep Learning Models on the Protein-Sugar Interactome. *bioRxiv.*
  - https://doi.org/10.1101/2025.09.02.673778
- Harmalkar, A., Chu, L.S., **Canner, S.W.**, Samanta, R., Frick, R., Davila-Hernandez, F.A., Sarma, S., Hitawala, F. Gray, J.J. 2025. Docking With Rosetta and Deep Learning Approaches in CAPRI Rounds 47-55. *Proteins*.
  - https://doi.org/10.1002/prot.70016
- **Canner, S.W.**, Schnaar, R.L., Gray, J.J. 2025. Predictions from Deep Learning Propose Substantial Protein-Carbohydrate Interplay. *bioRxiv*.
  - https://doi.org/10.1101/2025.03.07.641884

- Zhang G.L., Porter M.J., Awol A.K., Orsburn B.C., **Canner, S.W.,** Gray J.J., O'Meally R.N., Cole R.N., Schnaar R.L. 2024. The Human Ganglioside Interactome in Lie Cells Revealed Using Clickable Photoaffinity Ganglioside Probes. *JACS*, 149(26):17801-17816.
    - https://doi.org/10.1021/jacs.4c03196
- Martins M., dos Santos A.M., da Costa C.H.S., **Canner S.W.**, Chungyoun M., Gray J.J., Skaf M.S., Ostermeier M., Goldbeck R. 2024. Thermostability Enhancement of GH 62 a-l-Arabinofuranosidase by Directed Evolution and Rational Design, *ACS Journal of Agricultural and Food Chemistry*
    - https://doi.org/10.1021/acs.jafc.3c08019
- Lensink, M., Brysbaert, G. Raouraoua, N… **Canner, S.W.,** … S. Wodak J. 2023. Impact of AlphaFold on Structure Prediction of Protein Complexes: The CASP15-CAPRI Experiment. *Proteins,* 91(12):1658-1683.
    - https://doi.org/10.22541/au.168888815.53957253/v1
- **Canner, S.W.,** Shanker S., and Gray J.J. 2023. Structure-based neural network protein-carbohydrate interaction predictions at the residue level. *Front. Bioinform.* 3.
    - https://doi.org/10.3389/fbinf.2023.1186531
- **Canner, S.W.,** Feller S.E., and Wassall S.R. 2021. Molecular Organization of a Raft-like Domain in a Polyunsaturated Phosopholipid Bilayer: A Supervised Machine Learning Analysis of Molecular Dynamics Simulations. *J Phys Chem B.* 125(48):13158-13167.
    - https://doi.org/10.1021/acs.jpcb.1c06511
- Wassall S.R., Leng X., **Canner S.W.**, Pennington E.R., Kinnun J.J., Cavazos A.T., Dadoo S., Johnson D., Heberle F. A., J. Katsaras and S.R. Shaikh. 2018. Docosahexaenoic acid regulates the formation of lipid rafts: A unified view from experiment and simulation. *Biochim. Biophys. Acta*, 1860(10):1985-1993.
    - https://doi.org/10.1016/j.bbamem.2018.04.016
- Leng, X., Kinnun J.J., Cavazos A.T.**, Canner S.W.**, Shaikh S.R., Feller S.E., and Wassall S.R. 2018. All n-3 PUFA are not the same: MD simulations reveal differences in membrane organization for EPA, DHA and DPA. *Biochim. Biophys. Acta*. 1860(5):1125-1134.
    - https://doi.org/10.1016/j.bbamem.2018.01.002

PRESENTATIONS

- The Sugar Science, Th1nk Tank, May 8
    - "The Proteome as a Lectome: predictions from deep learning propose substantial protein-carbohydrate interplay"
- Society For Glycobiology 2024, Poster and Oral Presentation
    - "The Proteome as a Lectome: predictions from deep learning propose substantial protein-carbohydrate interplay"
- Jr. Mathematical Institute for Data Science Seminar, Oral Presentation

    - "Leveraging imperfect datasets to elucidate the cellular functionality of protein-glycan interactions on proteomic scales"
- Institute for Biophysical Research (IBR) at Johns Hopkins 2024, Oral Presentation
    - "The Proteome as a Lectome: predictions from deep learning propose substantial protein-carbohydrate interplay"
- EuroCarb21 2023, Oral Presentation
    - Structure based neural network predictions of protein carbohydrate interactions
- NIH-FDA Glycosience Research Day 2024, Poster
- Summer RosettaCon 2022, Poster
- Winter RosettaCon 2022, Poster
- Biophysical Society, Poster: 2024, 2020, 2019, 2018

EDUCATION

JOHNS HOPKINS UNIVERSITY **Baltimore, MD**

Program in Molecular Biophysics *August 2020 – October 2025*

Lab of Jeffrey J. Gray

INDIANA UNIVERSITY-PURDUE UNIVERSITY AT INDIANAPOLIS **Indianapolis, IN**

Physics BS, Computer Science BS, Mathematics and Creative Writing MinorAugust 2016 – May 2020

- GPA: 3.985
- Credit Hours: 151

AWARDS AND COMMUNITY INVOLVEMENT

- NSF GRFP Honorable Mention 2021
- IUPUI Chancellor's Award for Outstanding Research (2020): Top Undergraduate Researcher Award
- IUPUI Top 100 Undergraduate Students (2020)
- IUPUI Honors College Chancellor's and Dean of Science Scholar
- 2018 IUPUI Undergraduate Physics Student
- Former board member of Naptown Stomp
- Taught two introductory coding classes at the Johnson County White River Library
- IUPUI School of Science Ambassador
- Former Alpha Lambda Delta/Phi Eta Sigma Community Service Committee Member
- Crochet blankets/prayer shawls for the sick and elderly

Fin.