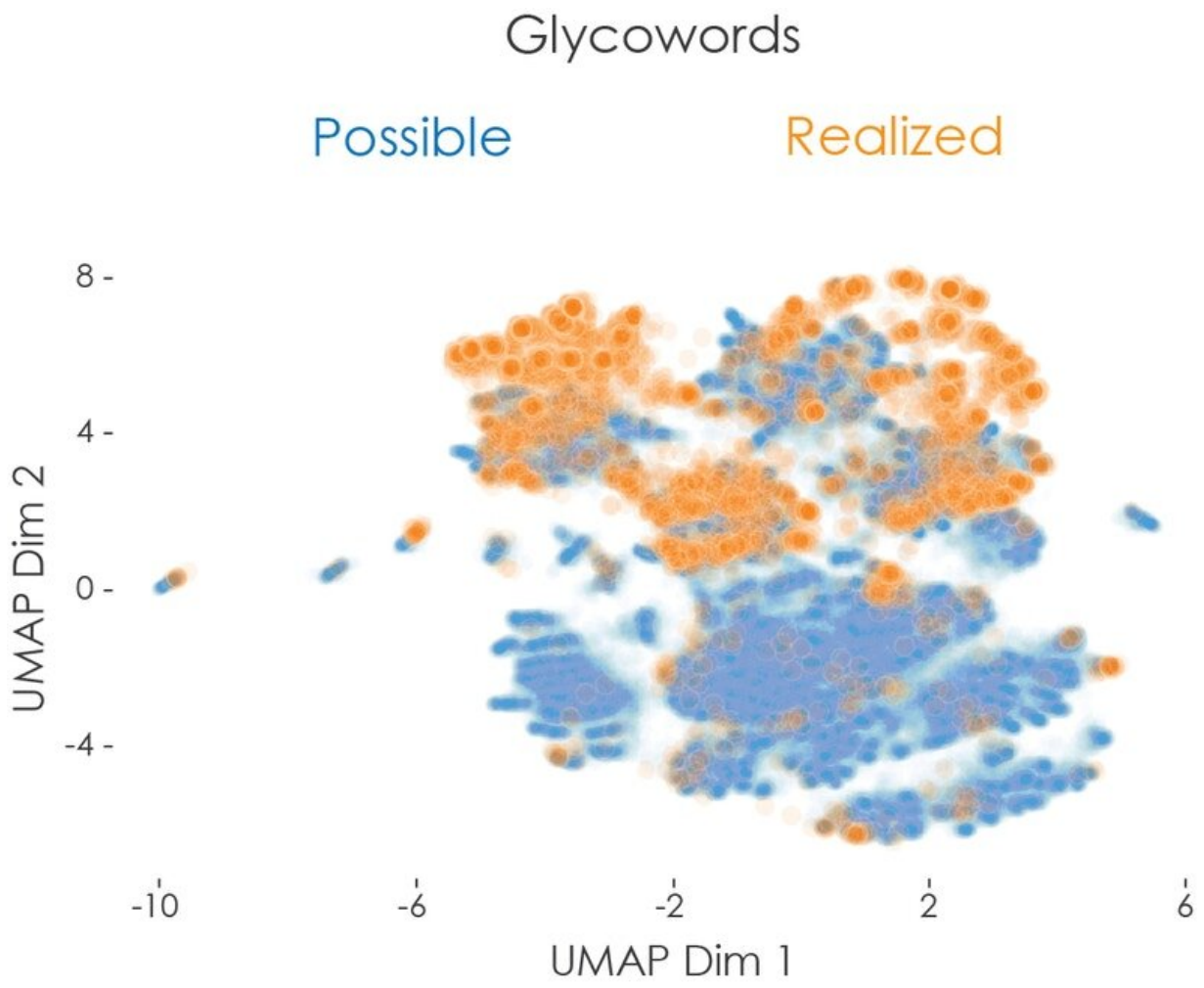
Deep learning and bioinformatics tools enable in-depth study of glycan molecules for understanding infections

October 28 2020, by Lindsay Brownell



Based on the existing glycoletters present in SugarBase, the researchers generated a graph of all the possible combinations that could produce glycowords (blue). When they analyzed the glycans sequences documented in SugarBase,

they found only a subset (orange) of the possible glycowords, offering insight into glycans' sequence evolution. Credit: Wyss Institute at Harvard University

We're told from a young age not to eat too much sugar, but in reality, our bodies are full of the stuff. The surface of every living cell, and even viruses, is covered in a mess of glycans: long, branching chains of simple sugars linked together by covalent bonds. These cell-surface sugars are crucial for regulating cell-cell contact, including the attachment of bacteria to healthy host cells. Glycans are also found on all other biological polymers, including proteins and RNA, and their presence impacts the polymers' stability and function.

Despite their ubiquity and importance, glycans remain poorly understood because of their complexity. Rather than just the four nucleotide "letters" that make up DNA and RNA molecules, glycans have an "alphabet" of hundreds of different monosaccharides that can be strung together into sequences with a seemingly infinite array of lengths and branches. In addition, an individual glycan sequence can be changed due to the interplay of multiple enzymes and conditions both within and outside a cell, without the need for genetic mutations.

Now, a team of scientists from the Wyss Institute for Biologically Inspired Engineering at Harvard University and the Massachusetts Institute of Technology (MIT) has cracked the glycan code by developing new machine learning and bioinformatics methods that enable researchers to systematically study glycans and identify sequences that play a role in the interactions of microbes and their host cells, as well as other still-unknown functions. The tools are presented in a new paper published today in *Cell Host & Microbe,* and are available online as a free Wyss WebApp that researchers can use to perform their own analyses of thousands of glycans.

"The language-based models that we have created can be used to predict whether and how a given glycan will be detected by the human immune system, thus helping us determine whether a strain of bacteria that harbors that glycan on its surface is likely to be pathogenic," said first author Daniel Bojar, Ph.D., a Postdoctoral Fellow at the Wyss Institute and MIT. "These resources also enable the study of glycan sequences involved in molecular mimicry and immune evasion, expanding our understanding of host-microbe interactions."

Glycan grammar rules

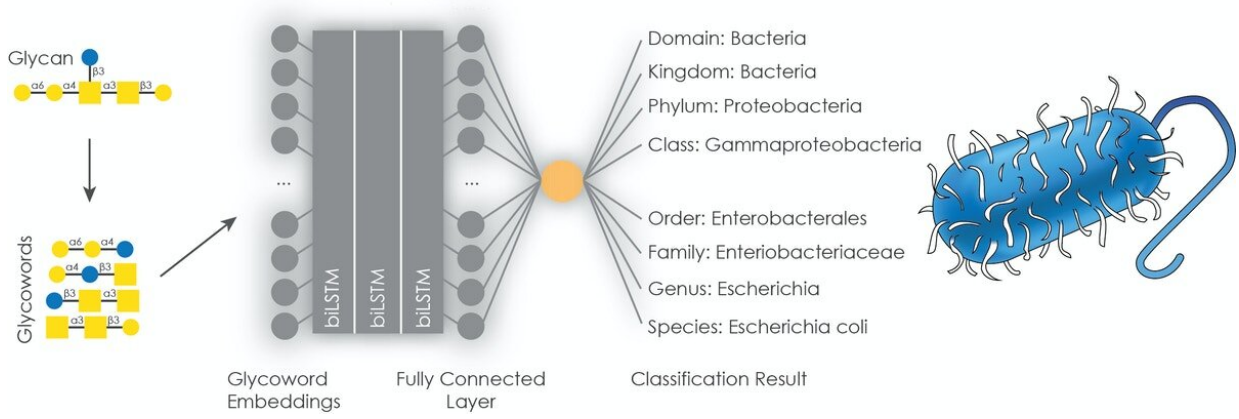
Because glycans are the outermost layer of all living cell types, they are necessarily involved in the process of infection, both in the interaction of a prokaryotic bacterium binding to a eukaryotic host cell, and the interactions between the cells of the immune system. This has created an evolutionary arms race, in which bacterial glycans evolve to mimic those found on their hosts' cells to evade immune detection, and hosts' glycans are modified so that pathogens can no longer use them to gain access. To trace this history of glycan sequence development and identify meaningful trends and patterns, the research team turned to machine learning algorithms—specifically, natural language processing (NLP), which has previously demonstrated success in analyzing other biopolymers, like RNA and proteins.

"Languages are actually quite similar to molecular sequences: the order of the elements matters, elements that are not next to each other can still influence each other, and their structures evolve over time," said co-author Rani Powers, Ph.D., a Senior Staff Scientist at the Wyss Institute.

First, the team needed to assemble a large database of glycan sequences on which an NLP-based algorithm could be trained. They combed through existing datasets both online and in the academic literature to create a database of 19,299 unique glycan sequences, which they dubbed

SugarBase. Within SugarBase they identified 1,027 unique glycan molecules or bonds they termed "glycoletters" making up the glycan alphabet, which could theoretically be combined into "glycowords" that the team defined as three glycoletters and two bonds.

To develop an NLP-based model that could analyze sequences of glycoletters and pick out distinct glycowords, the team chose to use a bidirectional recurrent neural network (RNN). RNNs, which also underlie the "autocomplete" feature of text messaging and email software, predict the next word in a sequence given the preceding words, enabling them to learn complex, order-dependent interactions. They trained their glycoletter-based language model, dubbed SweetTalk, on sequences from SugarBase, and used it to predict the next most probable glycoletter in a glycan sequence based on the preceding glycoletters, in the context of glycowords.



The research team constructed a language model-based classifier, SweetOrigins, to predict the taxonomic origin of a given glycan sequence. They replicated this structure for each level of classification, from individual species all the way up to domains, creating eight SweetOrigins models that were able to classify the taxonomic group of a glycan with high accuracy. Credit: Wyss Institute at Harvard University

SweetTalk revealed that from the close to 1.2 trillion theoretically possible glycowords, only 19,866 distinct glycowords (~0.0000016%) were present in the database of existing glycans. The observed glycowords also tended to be clustered together in groups with highly similar sequences, partly indicating the taxonomic groups in which the glycowords are found, rather than distributed evenly among all possible sequence combinations. These outcomes likely reflect the high "cost" to an organism of evolving dedicated enzymes to construct specific glycan substructures—in that scenario, it is more evolutionarily efficient to tweak existing glycowords rather than generate completely new ones.

Given the important role glycans play in human immunity, the researchers fine-tuned SweetTalk using a smaller, curated list of glycans that are known from the literature to cause an immune response. When predicting the immunogenicity of glycan sequences from SugarBase, the SweetTalk model achieved an accuracy of ~92%, compared to an accuracy of ~51% for a model trained on scrambled glycan sequences. For example, glycans that are rich in a simple sugar called rhamnose, which is found in bacteria but not in mammals, were unambiguously labeled as immunogenic by SweetTalk. The model's excellent performance indicated that language-based models could be used to study characteristics of glycans on a large scale and with many potential applications, such as the exploration of glycan-immune system interactions.

Pour some sugar on me

Based on the success of their first glycan-focused deep learning model, the team had a hunch that deep learning could also illuminate the "family tree" of glycan sequences. To achieve this, they constructed a language model-based classifier called SweetOrigins. They first pre-trained

SweetOrigins with a SweetTalk model, then used the language-like properties of glycans to fine-tune the new model on a different task: predicting the taxonomic group of glycans by learning species-specific features of glycans that indicate their evolutionary history. They replicated this structure for each level of classification, from individual species all the way up to domains (e.g., Bacteria, Eukarya), creating eight SweetOrigins models that were able to classify the taxonomic group of a glycan with high accuracy. For example, the model accurately predicted glycans from the kingdoms Animalia (91.1%) and Bacteria (97.2%), allowing a glycan of unknown origin to be quickly classified as either animal-associated, microbe-associated, or found on both cell types.

The researchers then used SweetOrigins to investigate host-pathogen interactions, reasoning that differences in the glycans associated with various strains of E. coli bacteria could be used to predict how infectious the strains are. They trained a deep learning-based classifier with the same language model architecture as SweetOrigins on E. coli-specific glycan sequences, and were able to predict E. coli strain pathogenicity with an accuracy of ~89%. It also placed the majority of glycans that are associated with E. coli strains of unknown pathogenicity at various places along the spectrum of infectiousness, helping to identify strains that are likely to be pathogenic to humans.

"Interestingly, the glycans that our model predicts are most associated with infection bear a striking resemblance to glycans found on the cells that form the mucosal barriers in animals' bodies, which keep pathogens out," said Diogo Camacho, Ph.D., a co-corresponding author of the paper and Senior Bioinformatics Scientist at the Wyss Institute. "This suggests that the glycans on pathogenic bacteria have evolved to mimic those found on the hosts' cells, facilitating their entry and evasion of the immune system."

To more deeply probe how glycans function in host-microbe

interactions, the team developed a glycan sequence alignment method, which compares individual glycan sequences to determine regions that are conserved between glycans and, therefore, likely serve a similar function. They chose a specific polysaccharide sequence from the pathogen Staphylococcus aureus that is known to increase bacterial virulence and hypothesized that this glycan helped the bacterium escape immune detection. When they compared that polysaccharide to similar glycan sequences in the dataset, they found the best alignment result with the enterobacterial common antigen (ECA), a glycan found on the Enterobacteriaceae family of symbiotic and pathogenic bacteria.

The team also found ECA-like sequences associated with bacteria in the Staphylococcus, Acinetobacter, and Haemophilus genera, which are not part of the Enterobacteriaceae family that typically carries the ECA. This insight suggests that, in addition to mimicking the glycans found on their hosts, bacterial glycans can also evolve to mimic those found on other bacteria such as those in our microbiome, and that pathogenicity can arise via glycans on microbes that are not traditionally thought to be dangerous.

"The resources we developed here—SugarBase, SweetTalk, and SweetOrigins—enable the rapid discovery, understanding, and utilization of glycan sequences, and can predict the pathogenic potential of bacterial strains based on their glycans," said co-corresponding author Jim Collins, Ph.D., a Wyss Core Faculty member who is also the Termeer Professor of Medical Engineering & Science at MIT. "As glycobiology progresses, these tools can be readily expanded and updated, eventually allowing for the precise classification of glycans and facilitating the glycan-based study of host-microbe interactions at unprecedented resolution, potentially leading to new antimicrobial therapeutics."

"This achievement is yet another example of the power of applying

computational approaches to biological problems that have so far defied resolution because of their complexity. I am also very impressed with this team for making their tools openly available to researchers around the world, which promises to accelerate the pace of our collective understanding of glycans and their impact on human health," said Wyss Institute Founding Director Don Ingber, M.D., Ph.D. Ingber is also the Judah Folkman Professor of Vascular Biology at Harvard Medical School and the Vascular Biology Program at Boston Children's Hospital, as well as Professor of Bioengineering at Harvard's John A. Paulson School of Engineering and Applied Sciences.

More information: Daniel Bojar et al, Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions, *Cell Host & Microbe* (2020). DOI: 10.1016/j.chom.2020.10.004

Provided by Harvard University