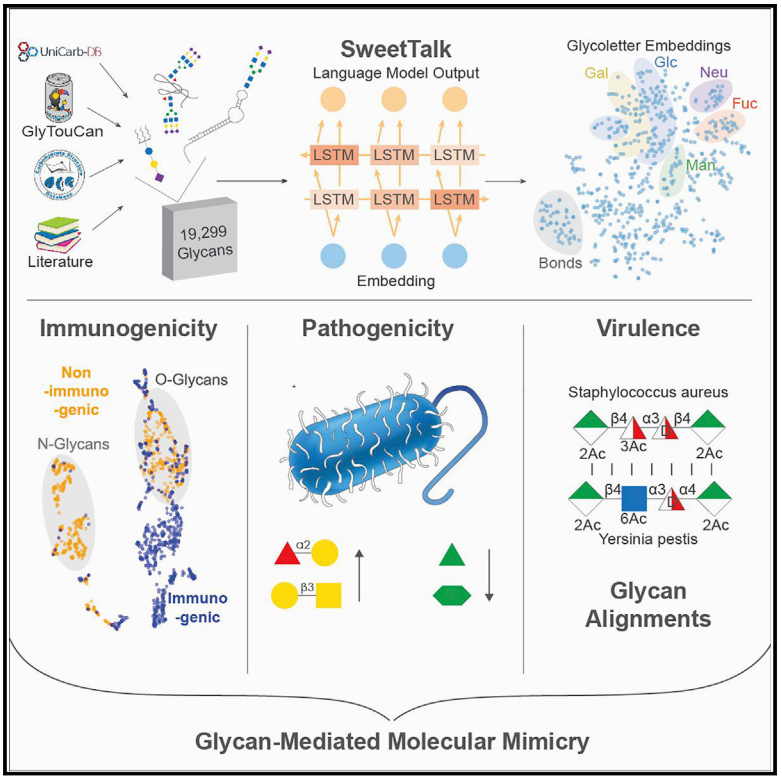
Resource

Cell Host & Microbe

Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions

Graphical Abstract



Authors

Daniel Bojar, Rani K. Powers, Diogo M. Camacho, James J. Collins

Correspondence

diogo.camacho@wyss.harvard.edu (D.M.C.), jimjc@mit.edu (J.J.C.)

In Brief

Bojar et al. present a workflow that combines machine learning and bioinformatics techniques to analyze the prominent role of glycans in host-microbe interactions. The herein developed glycan-focused language models and alignments allow for the prediction and analysis of glycan immunogenicity, association with pathogenicity, and taxonomic classification.

Highlights

- Glycan-focused language models can be used for sequence-to-function models
- Information in glycans predicts immunogenicity, pathogenicity, and taxonomic origin
- Glycan alignments shed light into bacterial virulence

Bojar et al., 2021, Cell Host & Microbe *29*, 132–144
January 13, 2021
https://doi.org/10.1016/j.chom.2020.10.004

CellPress

CellPress
OPEN ACCESS

Cell Host & Microbe

Resource

Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions

Daniel Bojar,[1,2] Rani K. Powers,[1,2] Diogo M. Camacho,[1,4,*] and James J. Collins[1,2,3,4,5,*]
[1]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA
[2]Department of Biological Engineering and Institute for Medical Engineering & Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4]These authors contributed equally
[5]Lead Contact
*Correspondence: diogo.camacho@wyss.harvard.edu (D.M.C.), jimjc@mit.edu (J.J.C.)
https://doi.org/10.1016/j.chom.2020.10.004

SUMMARY

Glycans, the most diverse biopolymer, are shaped by evolutionary pressures stemming from host-microbe interactions. Here, we present machine learning and bioinformatics methods to leverage the evolutionary information present in glycans to gain insights into how pathogens and commensals interact with hosts. By using techniques from natural language processing, we develop deep-learning models for glycans that are trained on a curated dataset of 19,299 unique glycans and can be used to study and predict glycan functions. We show that these models can be utilized to predict glycan immunogenicity and the pathogenicity of bacterial strains, as well as investigate glycan-mediated immune evasion via molecular mimicry. We also develop glycan-alignment methods and use these to analyze virulence-determining glycan motifs in the capsular polysaccharides of bacterial pathogens. These resources enable one to identify and study glycan motifs involved in immunogenicity, pathogenicity, molecular mimicry, and immune evasion, expanding our understanding of host-microbe interactions.

INTRODUCTION

In contrast to RNA and proteins, whose sequences can be elucidated from their associated DNA sequence, glycans are the only biopolymer outside the rules of the central dogma of molecular biology. Although glycans are synthesized by DNA-encoded enzymes (Lairson et al., 2008), an individual glycan sequence is dependent on the interplay between multiple enzymes and cellular conditions. Additionally, the expansive glycan alphabet of hundreds of different monosaccharides allows for a large number of potential oligosaccharides, built with different monosaccharides, lengths, connectivity, and branching. Glycans are present as modifications on all other biopolymers (Varki, 2017), exerting varying effects on biomolecules, including stabilization and modulation of their functionality (Dekkers et al., 2017; Solá and Griebenow, 2009). Apart from influencing the function of individual proteins, glycans are also crucial for cell-cell contact in the case of glycan-glycan interactions during the attachment of pathogenic bacteria to host cells (Day et al., 2015), and they mediate essential developmental processes such as nervous system development (Haltiwanger and Lowe, 2004). Recently, Lauc et al. hypothesized that the plethora of available glycoforms and their plasticity facilitated the evolution of complex multicellular lifeforms (Lauc et al., 2014), reasoning that is supported by the essential roles of glycans in developmental processes and cell-cell communication and emphasizes the evolutionary information in glycans.

Because glycans make up the outermost layer of both eukaryotic and prokaryotic cells, cross-kingdom interactions will necessarily involve these molecules (Day et al., 2015). The prominent role of glycans in host-pathogen interactions (Varki, 2017) has resulted in evolutionary pressures and opportunities on both sides of the interaction—natural selection can modify host glycan receptors used by pathogens without losing their functionalities, whereas pathogens and commensals need to alter their glycans to evade the host immune system. These interactions provide a window into understanding glycan-mediated host-microbe relationships. Glycans display great phenotypic variability: sequences can be changed depending on environmental conditions, such as the level of extracellular metabolites (Park et al., 2017), without the need for genetic mutations, potentially facilitating rapid responses to changes in host-microbe relationships.

Given the aforementioned glycan-mediated host-microbe interactions, glycans could provide insights into pathogenicity and commensalism determinants, as, for instance, molecular mimicry of host glycans by both pathogens and commensals facilitates their immune evasion (Carlin et al., 2009; Varki and Gagneux, 2015). Additional therapeutic potential is enabled by the widespread usage of glycans by viruses for cell adhesion and

CellPress
OPEN ACCESS

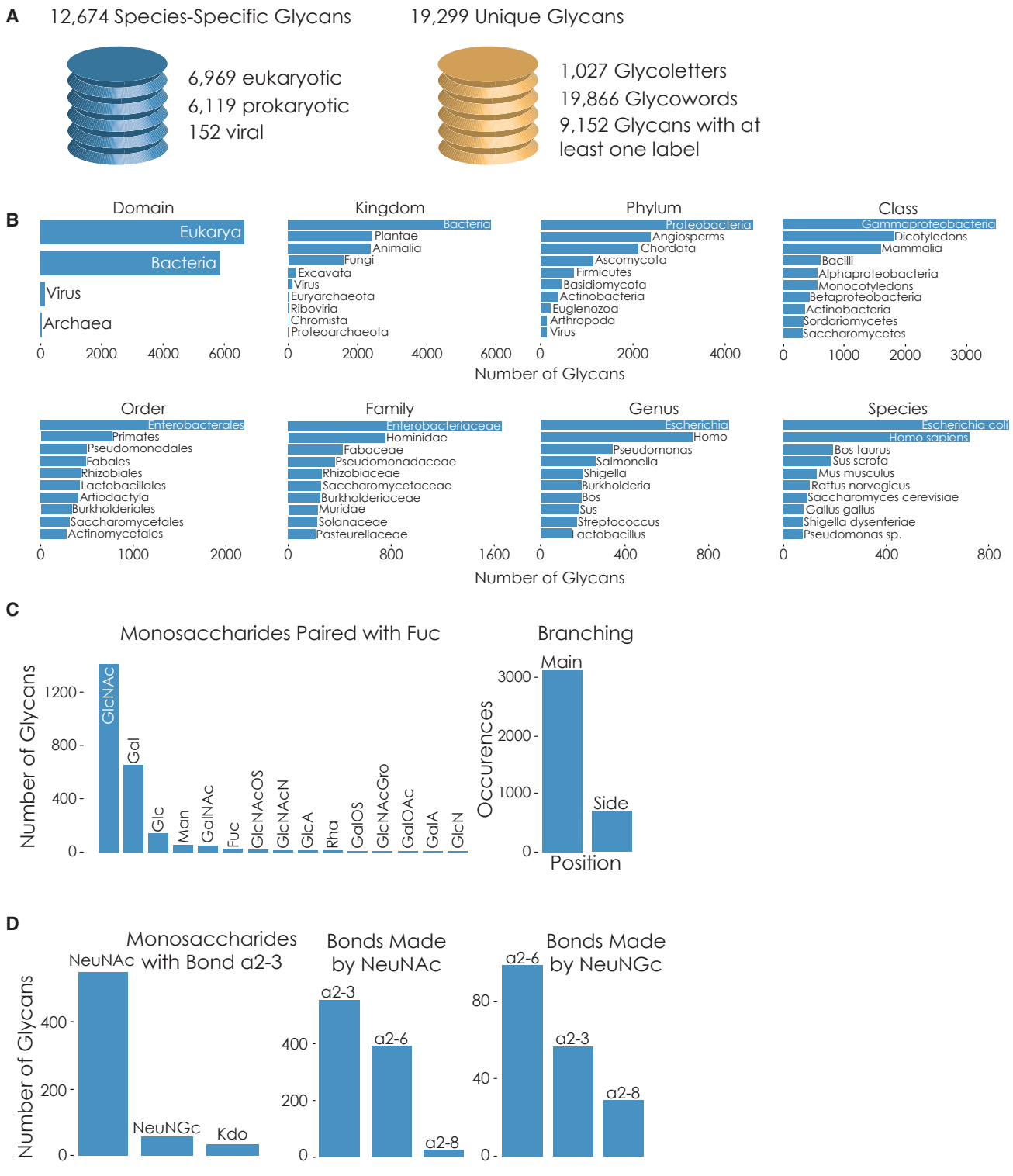


Figure 1. Using a Curated Glycan Dataset as a Resource for Glycobiology and Analyzing Host-Microbe Interactions

(A) Building curated datasets of species-specific and unique glycan sequences. Glycans stemming from proteins, lipids, small molecules, or cellular surfaces were gathered from UniCarbKB, CSDB, GlyTouCan, and the academic literature. We deposited these datasets in our database SugarBase, containing additional associated metadata, such as linkage and immunogenicity information.

(legend continued on next page)

entry (Thompson et al., 2019) and pathogenic bacteria (Poole et al., 2018).

In addition to previous work developing computational approaches to glycan analysis (McDonald et al., 2016; Spahn et al., 2016), identifying relevant glycan motifs and their roles in host-microbe interactions at scale would benefit from pattern-learning algorithms, such as machine learning, that can uncover statistical dependencies in biological sequences (Camacho et al., 2018). Research on other biopolymers has shown that language models, originally developed for the analysis of human languages, perform best in this task (Alley et al., 2019; Almagro Armenteros et al., 2020; Strodthoff et al., 2020), because they can leverage evolutionarily conserved regularities and language-like properties in such sequences. Language models, with their memory-like features, are well suited for leveraging patterns and implicit structure in biopolymers such as those underlying nucleic acids (Valeri et al., 2020) and proteins (Alley et al., 2019), because information in these sequences is order dependent, and non-neighboring residues can have meaningful interactions. Applying a natural language-processing approach to biological sequences also enables learning a representation of a molecule that can be used to analyze sequence motifs and predict functional properties. These types of models are therefore a suitable starting point for the analysis of glycan sequences.

Here, we present a resource toolkit comprising machine learning and bioinformatics methods as well as a large glycan database to leverage the evolutionary information present in glycans for predictive purposes in the context of host-microbe interactions, e.g., by understanding pathogenicity-associated glycan motifs. This toolkit can be used as a complete workflow for investigating host-microbe interactions, from a glycan dataset to glycan motifs identified by machine learning and further investigated by glycan alignments, or as separate modules. Underlying all of this is our language model for glycans, SweetTalk, trained on a dataset of 19,299 unique glycan sequences. With this, we demonstrate that similarities between glycans can be visualized and used to predict glycan properties such as human immunogenicity. Another part of our platform is SweetOrigins, a language-model-based classifier predicting the taxonomic origin of glycans that we use to obtain evolution-informed representations of glycans. To achieve this in the context of glycan-mediated host-microbe interactions, we manually curated a comprehensive dataset comprising 12,674 glycans with species annotations. These datasets were combined into a database, SugarBase, that is amenable to programmatic access and integration into deep-learning pipelines, thus providing resources for analyses involving host-microbe interactions.

In this work, we demonstrate the potential and generalizability of using SugarBase, SweetTalk, SweetOrigins, and a glycan-alignment methodology for studying glycan-mediated host-microbe interactions. We show that a language-model-based classifier trained on glycan sequences can accurately predict glycan immunogenicity and the pathogenicity of *E. coli* strains, revealing predictive glycan motifs. We also leverage the evolutionary information gained by SweetOrigins to analyze glycan motifs that could be used for molecular-mimicry-mediated immune evasion by commensals and pathogens. Applying our glycan-alignment methodology to the example of the capsular polysaccharides of *Staphylococcus aureus* and *Acinetobacter baumannii*, we uncover a potential connection to the enterobacterial common antigen and hypothesize a mechanism for the increased virulence mediated by these glycan motifs. Taken together, these resources offer a powerful and generalizable platform for studying and understanding the role of glycans in host-microbe interactions.

RESULTS

Curating Glycan Datasets for Glycobiology and Glycan-Mediated Host-Microbe Interactions

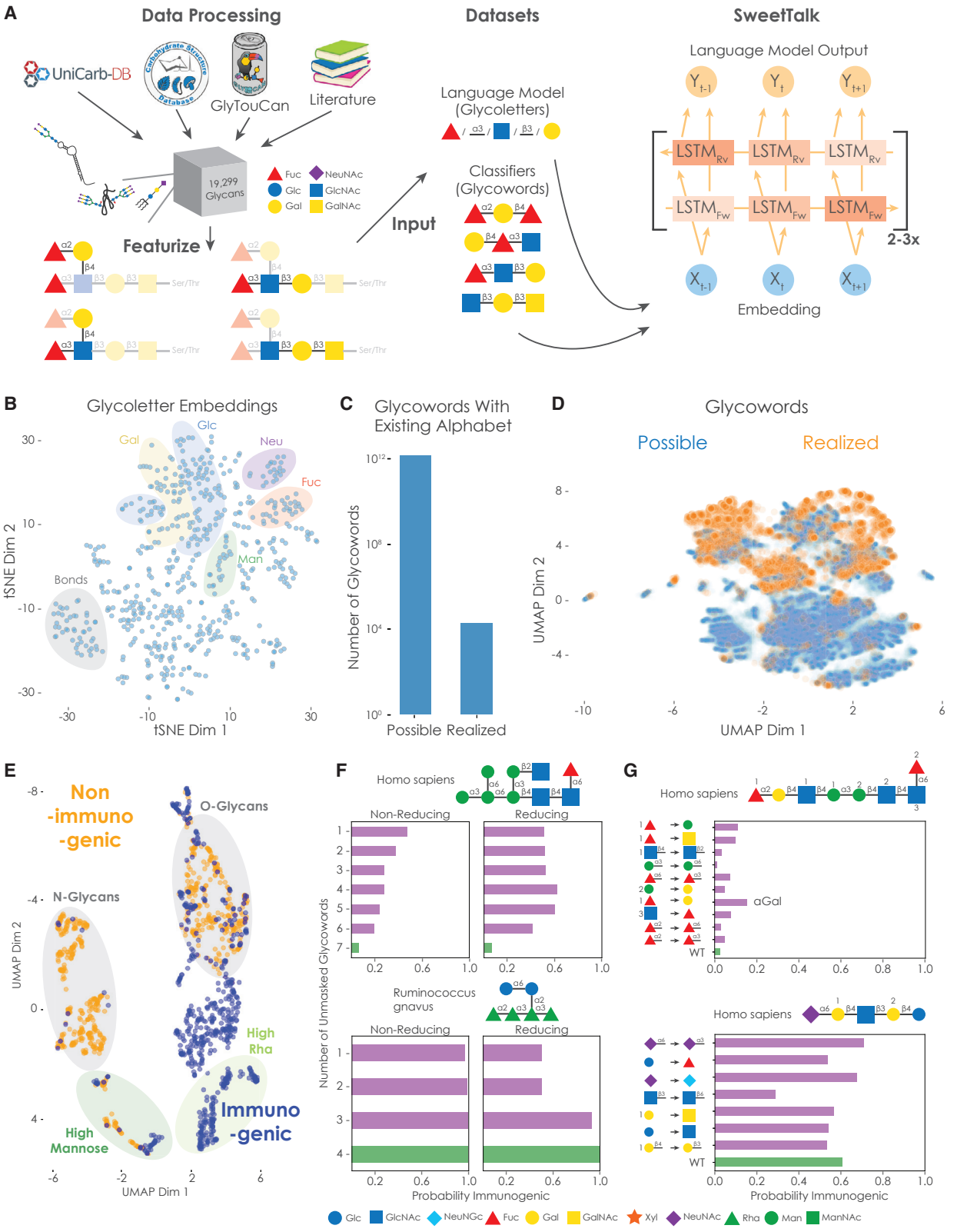
To investigate the role of glycans in host-microbe interactions, we constructed a dataset of species-specific glycan sequences that could be used to train machine-learning models. For this, we gathered and curated a dataset with glycans from GlyTouCan (Tiemeyer et al., 2017), UniCarbKB (Campbell et al., 2014), the Carbohydrate Structure Database (CSDB) (Toukach and Egorova, 2016), and targeted literature searches (see STAR Methods). To facilitate training deep-learning models on glycan sequences, we only included glycans with fully elucidated sequences, including the determination of linkages between monosaccharides. Our dataset contained 12,674 highly diverse glycans with a deposited species association (Figure 1A; Table S1) and included glycans from 1,726 species (corresponding to 39 taxonomic phyla; Figure 1B). Specifically, our dataset contained 6,969 eukaryotic, 6,119 prokaryotic, and 152 viral glycans. Because we included all species for which we could find glycans, this dataset constituted a comprehensive snapshot of currently known species-specific glycans, with glycans from numerous bacteria, facilitating the study of glycan-mediated host-microbe interactions.

We further reasoned that the inclusion of glycan sequences without a deposited species label would strengthen the language models we describe below. This approach is supported by the success of transfer learning in the field of machine learning (Howard and Ruder, 2018), in which models are initially trained on large datasets without labels and then finetuned on smaller datasets with labels. This makes more data available to learn general patterns, such as sequence motifs, that can be leveraged to predict glycan properties. Accordingly, we curated a separate dataset in which we used the databases mentioned above to gather 19,299 unique glycan sequences, irrespective of whether species information was available (Figure 1A; STAR Methods; Table S2). To gain a comprehensive view of glycobiology, we included all glycan categories, encompassing

(B) Glycan species distribution in the species-specific glycan dataset. For all glycans with species information, up to the 10 most abundant classes for each taxonomic level are shown with their number of glycans.

(C and D) Analyzing the local structural context of glycoletters. We identified the most frequent monosaccharides following fucose in glycans (C), highlighting its local structural context together with its likely position in the glycan structure (main versus side branch). Additionally, we compared the binding behavior of several sialic acids (D).

CellPress
OPEN ACCESS



(legend on next page)

protein-, lipid-, and small molecule-associated glycans, as well as capsular and extracellular polysaccharides.

In our dataset, we observed 1,027 unique monosaccharides or bonds that were present in glycan sequences and comprised the smallest units of an alphabet for a glycan language. Analogous to natural language processing, we termed these entities "glycoletters" and constructed "glycowords" by considering trisaccharides (i.e., three monosaccharides and two connecting bonds, or five glycoletters), yielding 19,866 unique glycowords in our dataset. With this, we sought to incorporate local structural information into our models and enable the discovery of relevant motifs, which usually contain subsequences larger than a single monosaccharide. Even larger substructures would preclude the analysis of shorter glycans and lead to an exponential increase in the size of the resulting vocabulary. We would also like to note that although we chose trisaccharides as building blocks, glycan substructures of any length can be used to build a vocabulary for our models without considerable changes.

To make these data and analysis resources readily accessible and facilitate further advances in glycobiology, we created SugarBase, a comprehensive glycan database with metadata and analytical tools based on this work (Figure S1A; Table S2; https://webapps.wyss.harvard.edu/sugarbase). SugarBase offers accessible glycan data, explorable glycan representations learned by our language models, and many of the methods developed here as tools, such as the local structural context of any glycoletter (Figure S1B) and glycan alignments, described below.

Reasoning that our glycan datasets constitute broad resources for glycobiology and host-microbe interactions, we set out to investigate host glycan substructures that could be emulated by microbes for molecular mimicry. Analyzing the environment of the monosaccharide fucose as an example, we observed N-acetylglucosamine (GlcNAc) and galactose (Gal) as typical connected monosaccharides (Figure 1C), which is consistent with the fucosyltransferase substrate specificities annotated in glycosyltransferase family 10 (Lombard et al., 2014). Thus, microbial glycans containing fucose could potentially include either GlcNAc or Gal in direct proximity to maximize similarity with host glycans. This insight aids in formulating hypotheses and identifying glycan motifs relevant for molecular mimicry, as we describe below. We also differentiated binding orientation preferences for different sialic acids, a crucial monosaccharide type in host-pathogen interactions (Figure 1D; Haines-menges et al., 2015), revealing a preference for the characteristic human monosaccharide NeuNAc to be (α2-3)-linked, relative to other sialic acids such as NeuNGc. These types of analyses can directly lead to hypotheses of glycan motifs that can be investigated by using the methods presented in this work.

Using Natural Language Processing to Learn the Grammar of Glycans

Next, we used our curated dataset of 19,299 glycan sequences (Table S2) to develop a deep-learning-based language model, SweetTalk. For this, we chose a bidirectional recurrent neural network (RNN; Figure 2A; Sherstinsky, 2020), because this type of model has delivered state-of-the-art results for other biopolymers, such as protein sequences (Alley et al., 2019; Almagro Armenteros et al., 2020; Strodthoff et al., 2020). Originally developed for human languages, RNNs exhibit memory-like elements by predicting the next word given the preceding words (Sherstinsky, 2020); this enables RNNs to learn complex, order-dependent interactions in proteins by viewing amino acids as letters and predicting the next amino acid given the preceding sequence (Alley et al., 2019). Two of the main usages for a trained language model are as follows: (1) extracting a learned representation for each word and (2) finetuning the model for predicting structural or functional properties of a sequence. For the former, a representation or embedding that characterizes a word in terms of context, usage, and meaning is constructed in the parameters of the trained model for each word in the vocabulary. This learned representation can be used to quantify the similarity of two glycan sequences or analyze language properties, which we demonstrate with the analysis of molecular mimicry in host-microbe interactions. The latter—finetuning a general language model on a predictive task such as predicting pathogenicity—is also known as transfer learning (Howard and Ruder, 2018; Tan et al., 2018), and in our case it involves general glycan features that are learned by the language model to predict functional properties.

Figure 2. Learning the Language of Glycans Revealed Regularities in Substructures and Can Be Used to Predict Glycan Immunogenicity

(A) Building a language model for glycobiology. We used glycowords, overlapping units consisting of three monosaccharides and two bonds, for our glycoletter-based bidirectional RNN, SweetTalk, that was trained by predicting the next glycoletter given previous glycoletters. Glycans are drawn in accordance with the symbol nomenclature for glycans (SNFG).

(B) Learned representation of glycoletters by SweetTalk. We visualized the embedding for every glycoletter by t-distributed stochastic neighbor embedding (t-SNE). Areas enriched for modified monosaccharides of one type are colored.

(C) Comparing the abundance of possible and observed glycowords. Possible glycowords were calculated from the pool of observed glycoletters and their exhaustive combination (36 bonds and 991 monosaccharides).

(D) Comparing the distribution of possible and observed glycowords. We generated 250,000 glycowords by randomly sampling from the observed pool of monosaccharides and bonds and formed their embedding by averaging their constituent glycoletter embeddings. A uniform manifold approximation and projection (UMAP) of these generated glycowords (blue) and all observed glycowords (orange) is shown.

(E) Glycan embeddings learned by the immunogenicity classifier. Embeddings for glycans from our immunogenicity dataset are shown via UMAP and colored according to whether they were immunogenic (blue) or non-immunogenic (orange).

(F) Glycoword masking to probe the immunogenicity classifier. Glycowords were progressively exchanged with padding ("masking") from both termini ("Non-Reducing"/"Reducing") and used as input for the trained immunogenicity classifier. Inferred immunogenicity probability indicates how crucial each region of a glycan is for prediction, with the bar representing the full-length glycan at the bottom.

(G) Glycan *in silico* alterations to probe immunogenicity classifier. For 4,000 iterations, single monosaccharides or bonds were replaced with a random monosaccharide or bond. If the resulting glycowords were observed, we used them as input for the trained immunogenicity classifier. Inferred immunogenicity probability is plotted together with the altered glycan sequences, with the wildtype glycan found at the bottom. In case of ambiguity, a number indicates which monosaccharide was modified. The addition of an "S" implies a sulfurylated monosaccharide, whereas "Me" implies a methylated monosaccharide.

Glycans are the only nonlinear biopolymer, with up to multiple branches per sequence. To enable a language model despite this branching, we extracted partially overlapping ''glycowords'' from the non-reducing end to the reducing end of glycans in the bracket notation (Figure 2A), comprising three monosaccharides and two bonds. These glycowords represented snapshots of structural contexts that characterize a glycan sequence. By using monosaccharides and bonds as ''glycoletters,'' we then trained a glycoletter-based language model, SweetTalk, predicting the next most probable glycoletter given the preceding glycoletters in the context of these glycowords (Table S3). This operation, instead of directly training on full sequences, avoids learning specious relationships between glycoletters that are close in the bracket notation but far apart in the actual glycan structure due to branching. We then demonstrated the necessity of accounting for the order-dependent information in glycans by training SweetTalk on scrambled glycan sequences, randomizing the order but keeping the composition of a sequence—this resulted in severely degraded model performance, emphasizing the language-like elements inherent in glycan sequences (Table S3). Analyzing the learned embeddings of glycoletters after training SweetTalk revealed similar positions in embedding space for monosaccharides and their modified counterparts (e.g., sulfurylated galactose, GalOS, and sulfurylated N-acetylgalactosamine, GalNAcOS; Figure 2B), implying similarity in their language characteristics and context. This finding is reminiscent of observations made on the popular word2vec embeddings that also learn a representation of words in a human language by considering their neighboring words/context, in which semantically similar words form clusters (Mikolov et al., 2013).

We then constructed glycoword embeddings by averaging the embeddings of their constituent glycoletters. Our first observation was that from the close to 1.2 trillion possible glycowords (given our observed glycoletters), only 19,866 distinct glycowords (~0.0000016%) were observed here (Figure 2C). Moreover, these 19,866 glycowords were not evenly distributed in the learned embedding space, as existing glycowords formed clusters compared to *in silico*-generated, possible glycowords (Figure 2D). The observation that the glycoword space (and, thus, glycan space) is sparsely populated is potentially a consequence of having to evolve dedicated enzymes for constructing specific glycan substructures from a species-specific set of monosaccharide building blocks, making most combinations inaccessible.

Predicting Glycan Immunogenicity with a Glycan-Based Language Model

Given the important role glycans play in human immunity (Kappler and Hennet, 2020; Reusch and Tejada, 2015), we curated known immunogenic glycans from the literature (Table S2) to finetune a SweetTalk-based classifier with glycan sequences as input to predict their immunogenicity to humans. On an independent validation dataset, our model achieved an accuracy of ~92% (F1 score or balanced F score: 0.915), in comparison with an accuracy of ~51% for a model trained on scrambled glycan sequences (Figures 2E–2G; Table S4). Alternative machine-learning models that did not treat glycan sequences as a language, such as random forest classifiers, only achieved accuracies ranging from ~80%–88% for this task (Table S4), emphasizing the importance of order and patterns for elucidating glycan properties.

Rhamnose-rich glycans, a common monosaccharide in bacteria but not in mammals, were unambiguously assigned to an immunogenic cluster by our RNN-based model and presented the most striking motif for glycan immunogenicity (Figure 2E). The cluster containing high-mannose glycans provided additional ambiguity, because it included both immature human glycans and immunogenic fungal glycans, potentially suggesting the immunogenicity of unintentionally exposed immature human glycans. Indeed, the presence of immature high-mannose glycans on viral surfaces has been noted to influence immunogenicity, with many broadly neutralizing antibodies targeting the high-mannose glycans on HIV glycoproteins (Lavine et al., 2012). We also found that human mucosal O-glycans, characterized by their interactions with bacteria, were interspersed with bacterial immunogenic glycans in the embedding space, in contrast to N-linked glycans. This adds to the notion of an immunological compromise of recognizing these bacterial glycans at the expense of targeting human O-glycans with shared motifs, such as the ABH blood group antigens (Kappler and Hennet, 2020). These analyses indicate that embeddings from glycan-focused language 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.

Using Deep Learning to Provide Evolution-Informed Glycan Representations

We next hypothesized that the evolutionary pressures on glycans stemming from host-pathogen interactions could be extracted by a deep-learning model. For this, we constructed a language-model-based classifier, SweetOrigins, to predict the taxonomic origin of a glycan (Figure 3A). In distinguishing taxonomic classes, SweetOrigins could learn species-specific features of glycans that are indicative of their evolutionary history. Based on a bidirectional RNN, we first pre-trained SweetOrigins with a SweetTalk model as described above. We then used the language-like properties learned in this process to finetune the model on a different task—predicting the taxonomic group of glycans. By doing this for every taxonomic level, from the species level up to the domain level, we obtained eight SweetOrigins models with the same basic model architecture except for different final layers. These final layers could learn how to combine the extracted information from glycans for predicting their taxonomic group, and they differed in terms of their number of output nodes, as the number of classes varied for each taxonomic level. This strategy was successful in extracting evolutionary information from glycans, as SweetOrigins models classified the taxonomic group of a glycan with high accuracy (Table 1).

In contrast to other biological sequences such as DNA or proteins, the number of available sequences for glycans is still limited, which is compounded by their high diversity. This is especially visible in prediction tasks in which only few glycans per class are available, such as for the species-level SweetOrigins model, resulting in lower model performance for rare classes and less useful glycan representations for downstream analyses. As knowledge of host-microbe interactions at the species

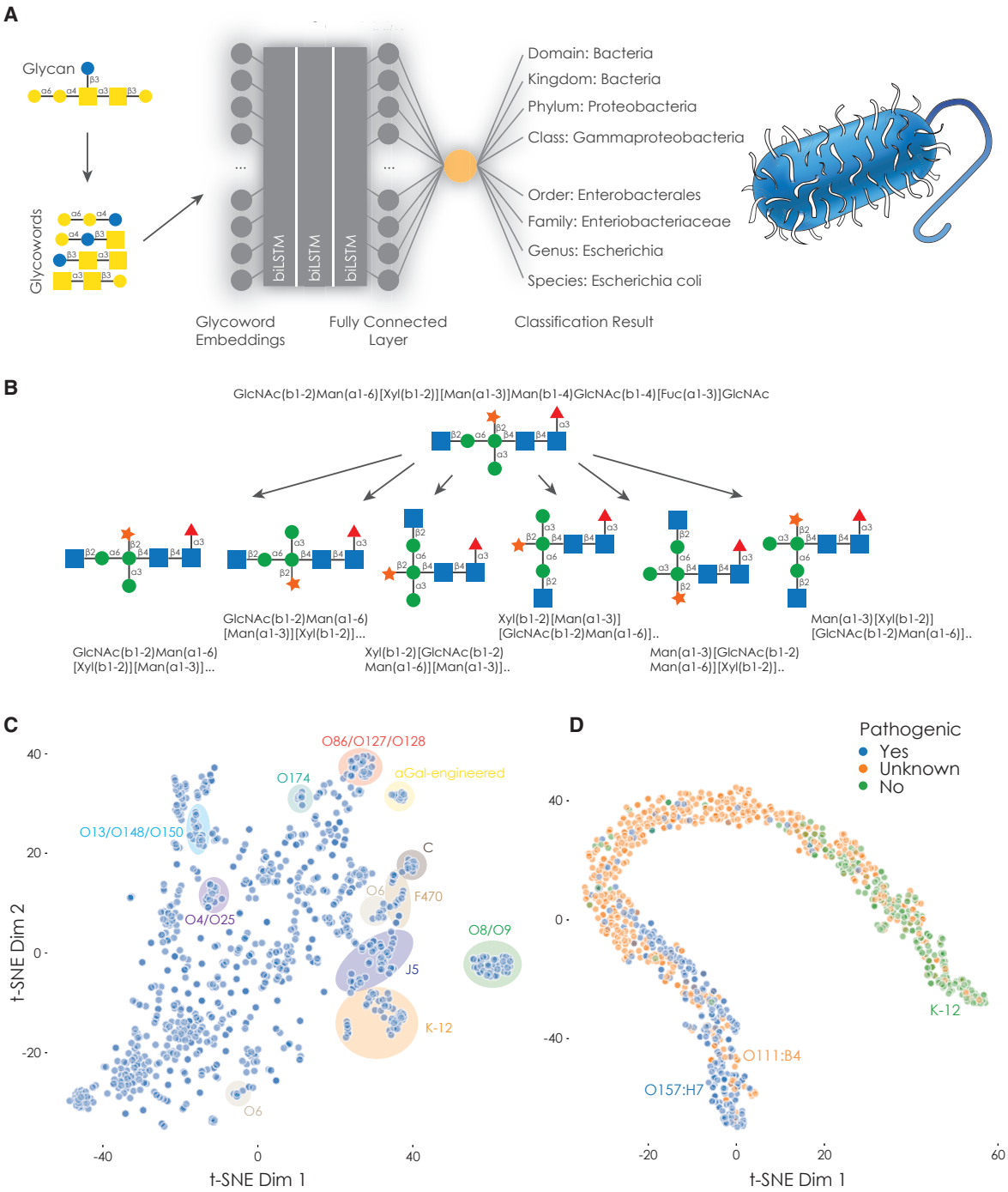


Figure 3. Deep-Learning-Based Classifiers Use Glycans to Predict Taxonomic Origin and Pathogenicity

(A) Exemplary schematic of SweetOrigins to predict taxonomic origin from glycans. Lists of glycowords are used as input for a SweetOrigins model to predict the taxonomic class ranging from the domain level down to the species level.

(B) Glycan data augmentation strategy. Different bracket notations describing the same glycan can be generated by alternating double branches as well as replacing side branches with main branches to increase model robustness.

(C) Glycans of *E. coli* in embedding space distinguish strains. The embedding for all 1,010 *E. coli*-derived glycans with strain information from the trained species-level SweetOrigins model is plotted via t-SNE and colored for areas enriched for annotated *E. coli* strains.

(D) *E. coli* glycans predict pathogenicity. For all *E. coli*-derived glycans, representations learned by a model predicting pathogenicity are plotted via t-SNE and colored as to whether they stem from pathogenic, non-pathogenic, or unlabeled *E. coli*. Example strains for all cases are annotated.

Table 1. Metrics of Trained SweetOrigins Models

Taxonomic Level	Classes	Baseline Accuracy		Cross-Entropy Loss		Accuracy		MCC	
		Random	Max	Base	Aug	Base	Aug	Base	Aug
Domain	4 (4)	0.2500	0.99	0.2841	**0.1906**	0.9128	**0.9313**	0.8134	**0.8693**
Kingdom	9 (11)	0.1111	0.98	0.3844	**0.3249**	0.8733	**0.8953**	0.8001	**0.8390**
Phylum	33 (39)	0.0303	0.98	0.8685	**0.7543**	0.7779	**0.8008**	0.7018	**0.7341**
Class	71 (101)	0.0141	0.96	1.3283	**1.1729**	0.6803	**0.7149**	0.6218	**0.6638**
Order	145 (207)	0.0069	0.92	2.2498	**2.1132**	0.4937	**0.5333**	0.4602	**0.5066**
Family	258 (411)	0.0039	0.90	2.9834	**2.7068**	0.4134	**0.4660**	0.3873	**0.4428**
Genus	405 (919)	0.0025	0.86	3.6588	**3.4081**	0.3658	**0.3849**	0.3505	**0.3682**
Species	581 (1,726)	0.0017	0.86	4.3704	**3.9550**	0.3052	**0.3651**	0.2870	**0.3496**

Taxonomic groups with fewer than five unique glycans were not used for model training or validation. Number of classes indicates the number of included taxonomic groups, whereas the full number of taxonomic groups in our dataset is given in parentheses. Models were trained with the standard set of glycans (Base) or after data augmentation (Aug). As an accuracy baseline, a random prediction of classes was used for each model. Max indicates the maximum theoretically possible accuracy given shared glycan sequences across taxonomic groups. Cross-entropy loss, accuracy, and Matthew's correlation coefficient (MCC) of the trained model on a separate validation set are given for each taxonomic level. For each metric and taxonomic level, the superior value is bolded.

level could offer insights, we developed methods that enable training glycan-focused machine-learning models on small datasets. This goal motivated our transfer-learning approach of pretraining a language model on all glycan sequences and then fine-tuning the model on a smaller dataset, because this approach in natural language processing has in some cases reduced the necessary dataset size by a factor of 100 (Howard and Ruder, 2018). In other domains of deep learning, such as image classification, data augmentation routinely results in improved model quality and robustness by providing the model with slightly modified versions of the data (Perez and Wang, 2017), such as rotating images or changing their brightness. We reasoned that the same could be achieved for biomolecules such as glycans; we thus designed a data-augmentation method, specifically for glycans, by conceptualizing glycans as graphs and forming a set of isomorphic graphs comprising slightly different lists of glycowords that we used as inputs for SweetOrigins (Figure 3B; STAR Methods). Capitalizing on the ambiguity of the bracket notation (Tanaka et al., 2014), we generated bracket notations that differed in their ordering of branches but still described the same glycan. This led to model performance improvements at every classification level, with absolute accuracy increases of up to 6%, by effectively increasing the amount of available data. As we envisioned, classifications with less data per class, such as the species level, benefited most from data augmentation (Table 1), paving the way for using glycan-based deep-learning models with smaller datasets.

In general, our predictions were robust, and we could, for example, accurately predict glycans from the kingdoms Animalia (91.1%) and Bacteria (97.2%), as well as glycans from the phyla Chordata (91.9%) and Firmicutes (90.4%) in our validation dataset (Figures S2A–S2C). This demonstrates that SweetOrigins can learn glycan representations from both hosts and microbes, enabling the analyses presented below. Any misclassifications occurred among closely related groups, such as viral glycans misclassified as those of their hosts (Figures S2A–S2C). Glycan embeddings from our trained SweetOrigins model illustrated clusters reminiscent of taxonomic groups (Figure S2D). We next used our trained SweetOrigins models to infer the taxonomic origin of the 10,333 glycans without a species label in our dataset (Table S2). For several randomly selected glycans, we performed literature searches to validate the predictions made by SweetOrigins (Figure S2E; Table S5), indicating that our trained SweetOrigins models had accurately learned species- or group-specific glycan motifs.

We next used SweetOrigins models to investigate host-pathogen interactions, specifically in the context of the well-studied bacterium *E. coli*. Although SweetOrigins classifiers were only trained up to the species level, we hypothesized that subspecies-level information could be extracted from the rich glycan representation learned by the species-level SweetOrigins model. To test this, we gathered 1,010 glycan sequences from *E. coli* with strain-level annotation from CSDB and used these as inputs to our trained model, yielding learned representations that we used to differentiate serotypes. We could readily identify clusters enriched for several strains in the representations, such as the serotypes O8/O9, characterized by a special polymannose O-antigen (Greenfield et al., 2012), and the K-12 strain popular in molecular biology research (Figure 3C), demonstrating the diversity and characteristic features of glycans for different *E. coli* strains.

We next reasoned, given the prominent role of glycans in host-microbe interactions, that these glycan differences could be used to predict *E. coli* pathogenicity, because *E. coli* strains can range from being non-colonizing to commensal or pathogenic (Lim et al., 2010). Accordingly, we trained a deep-learning-based classifier with the same language-model architecture as SweetOrigins on glycan sequences to elucidate whether information in glycans can predict pathogenicity. With a threshold of 0.5 in the predicted probability of pathogenicity, we found that we were able to predict *E. coli* strain pathogenicity with an accuracy of ~89% on a separate validation dataset (Figure 3D; F1 score: ~0.906). This positioned *E. coli* strains along a continuum of predicted pathogenicity and supported the role of glycans in mediating pathogenicity. Interestingly, *E. coli* strains such as O111:B4, which were labeled as "unknown" in the

dataset and therefore not available during model training, were predicted to be among the pathogenic strains and confirmed to cause gastric disease (Viljanen et al., 1990). Our trained model placed the majority of *E. coli* glycans from unknown pathogenicity strains between pathogenic and non-pathogenic strains, adding to the notion of a continuum of pathogenicity (Casadevall, 2017).

Because glycans appear to be predictive of pathogenicity, we reasoned that certain glycan motifs in *E. coli* strains on the pathogenic end of the spectrum might provide further insight into pathogenesis. To address this notion, we identified glycan motifs that are enriched in regions populated by predominantly pathogenic *E. coli* strains in the representation learned by our model (Figure 3D). Motifs in these pathogenicity-associated glycans exhibited a striking resemblance to host mucosal glycans, with an enrichment for α1-2-linked fucose and the core 1 O-glycan structure (also known as T antigen; Gal(β1-3)GalNAc) prevalent in mucins (Figures S3A and S3B). Consistent with our local structural context analysis (Figure 1B), the majority of α1-2-linked fucose residues in pathogenic *E. coli* strains were linked to galactose (Figure S3C), forming part of the human blood group H antigen. Indeed, when analyzing the glycan motifs most predictive of *E. coli* strain pathogenicity, both Gal(β1-3)GalNAc and Fuc(α1-2)Gal disaccharides were among the top 20 motifs (Figure S3D). On the other hand, the presence of typical bacterial glycan components, such as rhamnose or L-Glycero-D-Manno-Heptose (LDManHep), was associated with lower predicted pathogenicity (Figure S3D).

Using Glycan Alignments to Study Virulence Determinants in Bacterial Pathogens

To better understand the function of glycans in host-microbe interactions, we developed a sequence-alignment method. For DNA and protein sequences, alignments use sequence changes due to mutations and insertions to enable, for example, the identification of conserved motifs in protein families (Doğan and Karaçalı, 2013). To facilitate analogous analyses for glycans and capitalize on the evolutionary influence of host-pathogen interactions on glycans, we developed methods for gapped, pairwise alignments of glycan sequences based on the Needleman-Wunsch alignment algorithm (Needleman and Wunsch, 1970). For this, we constructed a substitution matrix (which we termed GLYSUM; Table S6), analogous to the BLOSUM matrices used in protein alignments, that utilizes the likelihood of substituting two monosaccharides to calculate alignment scores. To assess whether our glycan alignments performed as envisioned, we analyzed viral glycans that are predominantly derived from their host organisms and thus should align to host glycans. As expected, the optimal alignment for the viral glycans was indeed from their host organisms (Figures 4A and 4B), supporting the validity of our glycan-alignment method.

We reasoned that functionally relevant glycan motifs for host-pathogen interactions are likely conserved to some extent and could be analyzed with glycan alignments. As an example, we used our glycan-alignment method to align the serotype 5 capsular polysaccharide of the clinically relevant pathogen *S. aureus*, which is known to increase bacterial virulence (Tzianabos et al., 2001), against our dataset. Because the capsular polysaccharides of *S. aureus* mediate its evasion of the immune system (Weidenmaier and Lee, 2015), we hypothesized that comparing these to similar sequences might offer insights to understand their pathogenicity. Notably, the best alignment results were achieved with the enterobacterial common antigen, ECA (Figure 4C), conserved in the Enterobacteriaceae family, which has been shown to be important for virulence (Gilbreath et al., 2012) and outer membrane permeability (Mitchell et al., 2018). These findings are supported by experiments demonstrating that ECA deficiency in *E. coli* can be rescued by the expression of enzymes from serotype 5 *S. aureus* (Kiser and Lee, 1998). Such a phenotype complementation could suggest that this ECA-like glycan motif fulfills a similar role in *S. aureus* as the canonical ECA in *E. coli*.

To further probe the connection of ECA-like glycans and increased virulence, we aligned the canonical ECA motif against our dataset to compile a list of ECA-like sequences and their alignment distances; we used these distances to construct a dendrogram detailing the relationships between ECA-like glycan sequences (Figure 4D). Although most of the *S. aureus*-derived ECA-like sequences formed a separate cluster, the type 5 capsular polysaccharide was located in a different cluster with the canonical ECA sequences. Of note, we observed an ECA-like motif in the capsular polysaccharide of *A. baumannii* (Figure 4D, bold), one of the most problematic hospital-acquired pathogens, in the same cluster dominated by canonical ECA sequences. The capsular polysaccharide of *A. baumannii* has been implicated with antibiotic resistance and virulence (Geisinger and Isberg, 2015), providing an intriguing potential link to the functions of the canonical ECA. For other pathogens, such as *Haemophilus ducreyi*, the expression of a gene cluster synthesizing a putative ECA-like glycan has also been linked to increased virulence (Banks et al., 2008), further suggesting a connection of this motif with virulence. Notably, the genera Staphylococcus, Acinetobacter, and Haemophilus are not part of the Enterobacteriaceae family that is typically associated with the ECA, highlighting the importance of our glycan alignments for screening thousands of glycans to aid in understanding motifs important for pathogenicity, such as the ECA-like glycans from *S. aureus* and *A. baumannii*.

DISCUSSION

Here, we presented a set of resources—a collection of deep-learning and bioinformatics methods, together with large, curated datasets of glycan sequences—that can be used to gain insights into many facets of glycan-mediated host-microbe interactions. The aggregation of many glycan sequences in our datasets leads to robust machine-learning models that are largely unaffected by data-entry errors, thereby adjusting for database errors. By training a language model to understand the hidden grammar of glycan sequences, we demonstrated that the information in glycans can be used to predict a range of glycan properties, such as immunogenicity or pathogenicity. We also showed that sequences can be compared and clustered by learning a representation for each glycan via our trained models. For applications involving glycoproteins, the distribution of variant glycans on a protein (Wu et al., 2018) could be accounted for by averaging their representations, potentially even weighted by their relative abundance. By developing both

CellPress
OPEN ACCESS

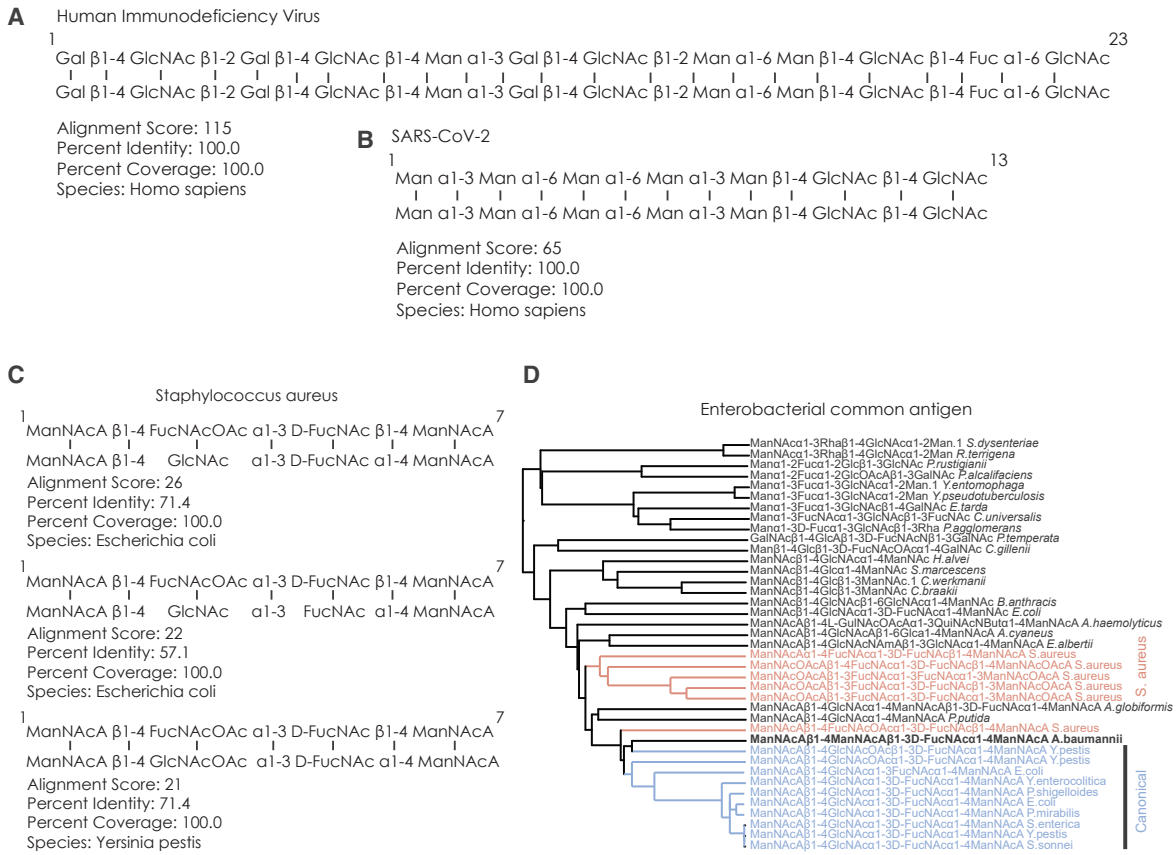


Figure 4. Glycan Alignments Identify Pathogenicity-Associated Glycan Motifs

(A and B) Viral glycans aligned to host glycans. We aligned viral glycans to all glycans and depicted the highest scoring alignment.

(C) Glycan alignments using serotype 5 capsular polysaccharide of *S. aureus*. The repeating unit of the glycan was aligned against our database, and the best three alignments are shown.

(D) ECA and ECA-like glycans. We aligned the canonical ECA sequence against our entire dataset, curated ECA-like sequences from the best 50 alignments, and constructed a dendrogram from alignment distances.

transfer-learning and data-augmentation methods for glycan-focused machine learning, we also addressed the pressing issue of the limited availability of glycan sequences due to experimental difficulties, enabling machine learning for many applications in glycobiology.

Our deep-learning strategies enabled us to introduce language models for glycans, while our curated datasets offer a state-of-the-art coverage for glycan sequences across a multitude of organisms. In contrast to word2vec-type models (Mikolov et al., 2013), our language-model-based approach captured sequential information beyond mere co-occurrences in glycan sequences and thus achieved better predictive results than alternative machine-learning techniques. This also enabled us to analyze glycan motifs, such as those important for immunogenicity and pathogenicity, that are dependent on sequential information and their relative position in glycans. Additionally, starting from a glycoletter-based model allowed for the construction of embeddings for close to 1.2 trillion glycowords, making SweetTalk easily extendable to the full diversity of glycobiology. SweetTalk can also incorporate position-specific modifications, illustrating its flexibility and potential for the analysis of information-rich glycosaminoglycans to predict, for instance, viral binding such as required for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cell entry (Liu et al., 2020).

Our resources can be utilized as a complete workflow, from a glycan dataset to motifs obtained by machine learning and further analyzed by glycan alignment, or as separate modules. The accuracy exhibited by our SweetOrigins models demonstrated that glycans can be used to distinguish closely related taxonomic groups and provided the means to leverage the evolutionary information in glycans for predictive purposes. Our observation that *E. coli* glycans are predictive of pathogenicity adds to the role of glycans as mediators of host-microbe relationships (Poole et al., 2018). The continuum of pathogenicity of *E. coli* strains, suggested by our deep-learning model, further adds to the redefinition of the notion of pathogenicity from a binary concept to a gradual, environmentally controlled process (Casadevall, 2017), mediated and influenced by glycans.

Both glycan alignments and glycan classification can connect glycan functions with sequence patterns, which we have used to derive insight from glycan motifs by analyzing glycans that could potentially be used for molecular-mimicry-mediated immune evasion by pathogenic *E. coli* strains. We further hypothesized

that glycan-based molecular mimicry, in addition to mimicking host glycans, could also extend to approximating glycans from other bacteria for increased virulence, e.g., as in the case of the capsular polysaccharides of *S. aureus* and *A. baumannii*, in which we hypothesized that they potentially mimicked the ECA of other bacteria. Our glycan-alignment method readily facilitated a hypothesis of the ECA mimicry performed by glycans of these pathogens, with a potentially broader relevance of this phenomenon in other pathogens, such as *H. ducreyi*, that are predicted to engage in ECA mimicry as well. In general, the resources developed here enable rapid discovery, understanding, and utilization of functionally relevant glycan motifs from glycan datasets, especially in the context of host-pathogen interactions. Another important feature of trained machine-learning models is the prediction of properties for newly acquired samples, such as predicting the pathogenic potential of newly identified *E. coli* strains based on their glycans. As glycobiology progresses, SugarBase and our deep-learning models could be readily expanded and updated, enabling an even more comprehensive investigation of glycan-mediated host-microbe interactions. This will eventually allow for precise classification at the subspecies level using language-model-based approaches, facilitating the glycan-based study of host-microbe interactions at unprecedented resolution.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- METHOD DETAILS
  - Dataset
  - Data Processing
  - Analyzing Links in Glycan Sequences
  - Glycan *In Silico* Modification
  - Glycan Alignment
  - Model Training
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.chom.2020.10.004.

ACKNOWLEDGMENTS

The authors would like to thank Jacqueline Valeri and Mathieu Groussin for helpful discussions. This work was supported by the Predictive BioAnalytics Initiative at the Wyss Institute for Biologically Inspired Engineering.

AUTHOR CONTRIBUTIONS

D.B. conceived the method. D.B., D.M.C., and J.J.C. designed the experiments. D.B. performed the experiments and implemented the method. R.K.P. developed the SugarBase web tool. D.M.C. and J.J.C. supervised the work. D.B., R.K.P., D.M.C., and J.J.C. wrote and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 29, 2020
Revised: September 9, 2020
Accepted: October 8, 2020
Published: October 28, 2020

REFERENCES

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat. Methods *16*, 1315–1322.

Almagro Armenteros, J.J., Johansen, A.R., Winther, O., and Nielsen, H. (2020). Language modelling for biological sequences – curated datasets and baselines. bioRxiv. https://doi.org/10.1101/2020.03.09.983585.

Banks, K.E., Fortney, K.R., Baker, B., Billings, S.D., Katz, B.P., Munson, R.S., Jr., and Spinola, S.M. (2008). The enterobacterial common antigen-like gene cluster of *Haemophilus ducreyi* contributes to virulence in humans. J. Infect. Dis. *197*, 1531–1536.

Bardor, M., Faveeuw, C., Fitchette, A.-C., Gilbert, D., Galas, L., Trottein, F., Faye, L., and Lerouge, P. (2003). Immunoreactivity in mammals of two typical plant glyco-epitopes, core alpha(1,3)-fucose and core xylose. Glycobiology *13*, 427–434.

Bashir, S., Leviatan Ben Arye, S., Reuven, E.M., Yu, H., Costa, C., Galiñanes, M., Bottio, T., Chen, X., and Padler-Karavani, V. (2019). Presentation Mode of Glycans Affect Recognition of Human Serum anti-Neu5Gc IgG Antibodies. Bioconjug. Chem. *30*, 161–168.

Bovin, N., Obukhova, P., Shilova, N., Rapoport, E., Popova, I., Navakouski, M., Unverzagt, C., Vuskovic, M., and Huflejt, M. (2012). Repertoire of human natural anti-glycan immunoglobulins. Do we have auto-antibodies? Biochim. Biophys. Acta *1820*, 1373–1382.

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. Cell *173*, 1581–1592.

Campbell, M.P., Peterson, R., Mariethoz, J., Gasteiger, E., Akune, Y., Aoki-Kinoshita, K.F., Lisacek, F., and Packer, N.H. (2014). UniCarbKB: building a knowledge platform for glycoproteomics. Nucleic Acids Res. *42*, D215–D221.

Carlin, A.F., Uchiyama, S., Chang, Y.-C., Lewis, A.L., Nizet, V., and Varki, A. (2009). Molecular mimicry of host sialylated glycans allows a bacterial pathogen to engage neutrophil Siglec-9 and dampen the innate immune response. Blood *113*, 3333–3336.

Casadevall, A. (2017). The Pathogenic Potential of a Microbe. MSphere *2*, e00015–e00017.

Day, C.J., Tran, E.N., Semchenko, E.A., Tram, G., Hartley-Tassell, L.E., Ng, P.S.K., King, R.M., Ulanovsky, R., McAtamney, S., Apicella, M.A., et al. (2015). Glycan:glycan interactions: High affinity biomolecular interactions that can mediate binding of pathogenic bacteria to host cells. Proc. Natl. Acad. Sci. USA *112*, E7266–E7275.

Dekkers, G., Treffers, L., Plomp, R., Bentlage, A.E.H., de Boer, M., Koeleman, C.A.M., Lissenberg-Thunnissen, S.N., Visser, R., Brouwer, M., Mok, J.Y., et al. (2017). Decoding the Human Immunoglobulin G-Glycan Repertoire Reveals a Spectrum of Fc-Receptor- and Complement-Mediated-Effector Activities. Front. Immunol. *8*, 877.

Doğan, T., and Karaçalı, B. (2013). Automatic identification of highly conserved family regions and relationships in genome wide datasets including remote protein sequences. PLoS One *8*, e75458.

Dotan, N., Altstock, R.T., Schwarz, M., and Dukler, A. (2006). Anti-glycan antibodies as biomarkers for diagnosis and prognosis. Lupus *15*, 442–450.

Geisinger, E., and Isberg, R.R. (2015). Antibiotic modulation of capsular exopolysaccharide and virulence in Acinetobacter baumannii. PLoS Pathog. *11*, e1004691.

Gilbreath, J.J., Colvocoresses Dodds, J., Rick, P.D., Soloski, M.J., Merrell, D.S., and Metcalf, E.S. (2012). Enterobacterial common antigen mutants of

Salmonella enterica serovar Typhimurium establish a persistent infection and provide protection against subsequent lethal challenge. Infect. Immun. *80*, 441–450.

Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Presented at the Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp. 249–256.

Greenfield, L.K., Richards, M.R., Li, J., Wakarchuk, W.W., Lowary, T.L., and Whitfield, C. (2012). Biosynthesis of the polymannose lipopolysaccharide O-antigens from *Escherichia coli* serotypes O8 and O9a requires a unique combination of single- and multiple-active site mannosyltransferases. J. Biol. Chem. *287*, 35078–35091.

Haines-menges, B.L., Whitaker, W.B., Lubin, J.B., and Boyd, E.F. (2015). Host Sialic Acids: A Delicacy for the Pathogen with Discerning Taste. In Metabolism and Bacterial Pathogenesis, C. Conway, ed. (American Society of Microbiology), pp. 321–342.

Haltiwanger, R.S., and Lowe, J.B. (2004). Role of glycosylation in development. Annu. Rev. Biochem. *73*, 491–537.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. Neural Comput. *9*, 1735–1780.

Hong, Y., and Reeves, P.R. (2014). Diversity of o-antigen repeat unit structures can account for the substantial sequence variation of wzx translocases. J. Bacteriol. *196*, 1713–1722.

Howard, J., and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. arXiv.

Kappler, K., and Hennet, T. (2020). Emergence and significance of carbohydrate-specific antibodies. Genes Immun. *21*, 224–239.

Khasbiullina, N.R., Shilova, N.V., Navakouski, M.J., Nokel, A.Yu., Blixt, O., Kononov, L.O., Knirel, Y.A., and Bovin, N.V. (2019). The Repertoire of Human Antiglycan Antibodies and Its Dynamics in the First Year of Life. Biochemistry (Mosc.) *84*, 608–616.

Kiser, K.B., and Lee, J.C. (1998). Staphylococcus aureus cap5O and cap5P genes functionally complement mutations affecting enterobacterial common-antigen biosynthesis in Escherichia coli. J. Bacteriol. *180*, 403–406.

Knirel, Y.A. (2011). Structure of O-Antigens. In Bacterial Lipopolysaccharides, Y.A. Knirel and M.A. Valvano, eds. (Springer Vienna), pp. 41–115.

Lairson, L.L., Henrissat, B., Davies, G.J., and Withers, S.G. (2008). Glycosyltransferases: structures, functions, and mechanisms. Annu. Rev. Biochem. *77*, 521–555.

Lauc, G., Krištić, J., and Zoldoš, V. (2014). Glycans - the third revolution in evolution. Front. Genet. *5*, 145.

Lavine, C.L., Lao, S., Montefiori, D.C., Haynes, B.F., Sodroski, J.G., and Yang, X.; NIAID Center for HIV/AIDS Vaccine Immunology (CHAVI) (2012). High-mannose glycan-dependent epitopes are frequently targeted in broad neutralizing antibody responses during human immunodeficiency virus type 1 infection. J. Virol. *86*, 2153–2164.

Lim, J.Y., Yoon, J., and Hovde, C.J. (2010). A brief overview of Escherichia coli O157:H7 and its plasmid O157. J. Microbiol. Biotechnol. *20*, 5–14.

Liu, L., Chopra, P., Li, X., Wolfert, M.A., Tompkins, S.M., and Boons, G.-J. (2020). SARS-CoV-2 spike protein binds heparan sulfate in a length- and sequence-dependent manner. bioRxiv. 2020.05.10.087288. https://doi.org/10.1101/2020.05.10.087288.

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

Lundberg, S.M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems, *Volume 30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. *(Curran Associates, Inc), pp. 4765–4774*.

McDonald, A.G., Tipton, K.F., and Davey, G.P. (2016). A Knowledge-Based System for Display and Prediction of O-Glycosylation Network Behaviour in Response to Enzyme Knockouts. PLoS Comput. Biol. *12*, e1004844.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv.

Mitchell, A.M., Srikumar, T., and Silhavy, T.J. (2018). Cyclic Enterobacterial Common Antigen Maintains the Outer Membrane Permeability Barrier of Escherichia coli in a Manner Controlled by YhdP. mBio *9*, e01321-18.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Park, D., Xu, G., Barboza, M., Shah, I.M., Wong, M., Raybould, H., Mills, D.A., and Lebrilla, C.B. (2017). Enterocyte glycosylation is responsive to changes in extracellular conditions: implications for membrane functions. Glycobiology *27*, 847–860.

Paschinger, K., Fabini, G., Schuster, D., Rendić, D., and Wilson, I.B.H. (2005). Definition of immunogenic carbohydrate epitopes. Acta Biochim. Pol. *52*, 629–632.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Perez, L., and Wang, J. (2017). The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv.

Pochechueva, T., Jacob, F., Fedier, A., and Heinzelmann-Schwarz, V. (2012). Tumor-associated glycans and their role in gynecological cancers: accelerating translational research by novel high-throughput approaches. Metabolites *2*, 913–939.

Poole, J., Day, C.J., von Itzstein, M., Paton, J.C., and Jennings, M.P. (2018). Glycointeractions in bacterial pathogenesis. Nat. Rev. Microbiol. *16*, 440–452.

Reusch, D., and Tejada, M.L. (2015). Fc glycans of therapeutic antibodies as critical quality attributes. Glycobiology *25*, 1325–1334.

Samraj, A.N., Bertrand, K.A., Luben, R., Khedri, Z., Yu, H., Nguyen, D., Gregg, C.J., Diaz, S.L., Sawyer, S., Chen, X., et al. (2018). Polyclonal human antibodies against glycans bearing red meat-derived non-human sialic acid N-glycolylneuraminic acid are stable, reproducible, complex and vary between individuals: Total antibody levels are associated with colorectal cancer risk. PLoS One *13*, e0197464.

Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Phys. Nonlinear Phenom. *404*, 132306.

Silipo, A., and Molinaro, A. (2010). The Diversity of the Core Oligosaccharide in Lipopolysaccharides. In Endotoxins: Structure, Function and Recognition, X. Wang and P.J. Quinn, eds. (Springer Netherlands), pp. 69–99.

Solá, R.J., and Griebenow, K. (2009). Effects of glycosylation on the stability of protein pharmaceuticals. J. Pharm. Sci. *98*, 1223–1245.

Spahn, P.N., Hansen, A.H., Hansen, H.G., Arnsdorf, J., Kildegaard, H.F., and Lewis, N.E. (2016). A Markov chain model for N-linked protein glycosylation–towards a low-parameter tool for model-driven glycoengineering. Metab. Eng. *33*, 52–66.

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2020). UDSMProt: universal deep sequence models for protein classification. Bioinformatics *36*, 2401–2409.

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. (2018). A Survey on Deep Transfer Learning. arXiv.

Tanaka, K., Aoki-Kinoshita, K.F., Kotera, M., Sawaki, H., Tsuchiya, S., Fujita, N., Shikanai, T., Kato, M., Kawano, S., Yamada, I., and Narimatsu, H. (2014). WURCS: the Web3 unique representation of carbohydrate structures. J. Chem. Inf. Model. *54*, 1558–1566.

Thompson, A.J., de Vries, R.P., and Paulson, J.C. (2019). Virus recognition of glycan receptors. Curr. Opin. Virol. *34*, 117–129.

Tiemeyer, M., Aoki, K., Paulson, J., Cummings, R.D., York, W.S., Karlsson, N.G., Lisacek, F., Packer, N.H., Campbell, M.P., Aoki, N.P., et al. (2017). GlyTouCan: an accessible glycan structure repository. Glycobiology *27*, 915–919.

Toukach, P.V., and Egorova, K.S. (2016). Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts. Nucleic Acids Res. *44* (D1), D1229–D1236.

Tsuchiya, S., Yamada, I., and Aoki-Kinoshita, K.F. (2019). GlycanFormatConverter: a conversion tool for translating the complexities of glycans. Bioinformatics *35*, 2434–2440.

Tzianabos, A.O., Wang, J.Y., and Lee, J.C. (2001). Structural rationale for the modulation of abscess formation by Staphylococcus aureus capsular polysaccharides. Proc. Natl. Acad. Sci. USA *98*, 9365–9370.

Valeri, J.A., Collins, K.M., Ramesh, P., Alcantar, M.A., Lepe, B.A., Lu, T.K., and Camacho, D.M. (2020). Sequence-to-function deep learning frameworks for engineered riboregulators. Nat Commun *11*, 5058, https://doi.org/10.1038/s41467-020-18676-2.

Varki, A. (2017). Biological roles of glycans. Glycobiology *27*, 3–49.

Varki, A., and Gagneux, P. (2015). Biological Functions of Glycans. In Essentials of Glycobiology, A. Varki, R.D. Cummings, J.D. Esko, P. Stanley, G.W. Hart, M. Aebi, A.G. Darvill, T. Kinoshita, N.H. Packer, and J.H. Prestegard, et al., eds. (Cold Spring Harbor Laboratory Press).

Viljanen, M.K., Peltola, T., Junnila, S.Y., Olkkonen, L., Järvinen, H., Kuistila, M., and Huovinen, P. (1990). Outbreak of diarrhoea due to Escherichia coli O111:B4 in schoolchildren and adults: association of Vi antigen-like reactivity. Lancet *336*, 831–834.

Weidenmaier, C., and Lee, J.C. (2015). Structure and Function of Surface Polysaccharides of Staphylococcus aureus. In Staphylococcus Aureus, F. Bagnoli, R. Rappuoli, and G. Grandi, eds. (Springer International Publishing), pp. 57–93.

Wu, D., Struwe, W.B., Harvey, D.J., Ferguson, M.A.J., and Robinson, C.V. (2018). N-glycan microheterogeneity regulates interactions of plasma proteins. Proc. Natl. Acad. Sci. USA *115*, 8763–8768.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
PyTorch	Paszke et al., 2019	https://github.com/pytorch/pytorch
Scikit-learn	Pedregosa et al., 2011	https://github.com/scikit-learn/scikit-learn
Apex	N/A	https://github.com/NVIDIA/apex
Python-alignment	N/A	https://github.com/eseraygun/python-alignment
SHAP	Lundberg and Lee, 2017	https://github.com/slundberg/shap
SweetTalk	This paper	https://github.com/midas-wyss/sweettalk
SweetOrigins	This paper	https://github.com/midas-wyss/sweetorigins
SugarBase	This paper	https://webapps.wyss.harvard.edu/sugarbase

RESOURCE AVAILABILITY

Lead Contact

Communication should be directed to the lead contact, James J. Collins (jimjc@mit.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Data used for all analyses can be found in the supplementary tables. All code and trained models can be found at https://github.com/midas-wyss/sweettalk and https://github.com/midas-wyss/sweetorigins.

METHOD DETAILS

Dataset

To create a comprehensive glycan dataset annotated with species labels, we manually curated 12,674 glycan sequences from three sources: UniCarbKB (Campbell et al., 2014), the Carbohydrate Structure Database (CSDB) (Toukach and Egorova, 2016), and the peer-reviewed scientific literature. From UniCarbKB, we compiled all glycans with species information, a length of at least three monosaccharides to facilitate usage with machine learning models, and a working link to PubChem to retrieve their sequences. We further complemented and extended this list by gathering glycans deposited in the Carbohydrate Structure Database (CSDB) up to December 2019 with a length of at least three monosaccharides. For species with more than 15 strains available on CSDB, only glycans from the first 15 strains were recorded to prevent taxonomic bias. For the model organism *E. coli*, all available glycan sequences were recorded to facilitate a strain-based analysis. Labels for *E. coli* strain pathogenicity were assigned, if possible, via the peer-reviewed academic literature. Finally, we performed additional literature searches, predominantly adding viral and archaeal glycans, which are underrepresented in the other databases. We revised and completed the annotations for all species' taxonomic characterization (species, genus, family, order, class, phylum, kingdom, domain) based on the NCBI Taxonomy Browser. In total, the dataset contained sequences from 1,726 different species from a range of 39 taxonomic phyla. To the best of our knowledge, this database represents the most comprehensive and current resource of glycans and their species information to date (Table S1).

To enable transfer learning by first pre-training a language model, we also added glycan sequences that lacked species information, by extracting the Web3 Unique Representation of Carbohydrate Structures (WURCS) representation (Tanaka et al., 2014) of the set of all glycans with at least three monosaccharides deposited on GlyTouCan (Tiemeyer et al., 2017) that were also available on PubChem (n = 18,926) and the databases mentioned above; this resulted in an augmented database containing 19,299 unique glycan sequences (Table S2). For all glycans, we relied on the quality control of the respective database. All glycans in WURCS representation were reformatted into the IUPAC condensed representation, using the GlycanFormatConverter software (Tsuchiya et al., 2019). For the immunogenicity classifier, all GlycoEpitope (https://www.glycoepitope.jp) entries with a minimum length of at least three monosaccharides were extracted. This list was further complemented by targeted literature searches (Bardor et al., 2003; Bashir et al., 2019; Bovin et al., 2012; Dotan et al., 2006; Hong and Reeves, 2014; Khasbiullina et al., 2019; Knirel, 2011; Paschinger et al., 2005; Pochechueva et al., 2012; Samraj et al., 2018; Silipo and Molinaro, 2010) resulting in the final set of immunogenic glycans (n = 685, Table S2). We included protein-, lipid-, and small molecule-associated glycans as well as capsular and extracellular

polysaccharides in our dataset of 19,299 glycans. All these glycans were paired with an ID to allow for our relational database SugarBase, linking all available information (linkage type, species information, human immunogenicity, etc.) to a glycan sequence (Table S2). Additionally, we included representations learned by our language model for all observed glycoletters (monosaccharides or bonds) as well as glycowords (trisaccharides).

Data Processing

Glycan sequences were processed by removing dangling bonds (e.g., '(α1-'). Analogous to word stemming in natural language processing, unifying different inflections of the same word, we removed position-specific information of monosaccharide modifications to reduce vocabulary size. Then, we harmonized capitalization and, in the case of glycan repeat structures, appended the first monosaccharide to their end to capture more sequence context. Additional steps to exclude duplicated glycans included strict ordering of multiple branches with equal lengths by ascending connection to the main branch (e.g., branch ending in 'α1-2' before branch ending in 'β1-4'). For branches closest to the non-reducing end, the longest branch was defined as the main chain. Observed monosaccharide modifications necessitated a hierarchy of order (in case of multiple modifications on the same monosaccharide) to avoid duplicates or mislabeling: NAc > OAc > NGc > OGc > NS > OS > NP > OP > NAm > OAm > NBut > OBut > NProp > OProp > NMe > OMe > CMe > NFo > OFo > OPPEtn > OPEtn > OEtn > A > N > SH > OPCho > OPyr > OVac > OPam > OEtg > OFer > OSin > OAep > OCoum > ODco > OLau > OSte > OOle > OBz > OCin > OAch > OMal > OMar > OOrn > rest.

Data processing for model training included featurization of glycan sequences into glycoletters (e.g., 'Gal'), as well as glycowords (three monosaccharides connected by two bonds). The conversion of a glycan sequence into glycowords, from the non-reducing to the reducing end, resulted in a list of partially overlapping glycowords, with maximum overlap so that two subsequent glycowords only differed in one monosaccharide and one bond. The aim of these glycowords is to capture representative characteristics and local structural contexts of a given glycan. The dataset comprising all glycowords (n = 113,112) was then used to train a context-specific, glycoletter-based language model. For scrambled glycan sequences, the order of glycoletters in any given glycan was randomly shuffled to maintain composition but erase patterns. All abbreviations for glycan nomenclature in this work can be found in Table S7.

Analyzing Links in Glycan Sequences

To determine typical local structural contexts of monosaccharides and bonds, we quantified the frequency of a given monosaccharide co-occurring with any other monosaccharide in our extensive database of unique glycans. Additionally, we also compared the relative frequencies of a particular monosaccharide being observed in the glycan main branch versus a side branch in our database.

Glycan *In Silico* Modification

We performed *in silico* modification of glycans by replacing monosaccharides and/or bonds with other observed monosaccharides/bonds. We used exhaustive modification, replacing glycoletters with all possible glycoletters, while only retaining modified glycans comprising previously observed glycowords. This ensured physiological relevance, given the extreme sparsity of observed glycan sequences compared to the theoretical number of possibilities.

Glycan Alignment

Global sequence alignment of glycans was implemented according to the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) by adapting the Python Alignment library (https://github.com/eseraygun/python-alignment). For our GLYcan SUbstitution Matrix (GLYSUM; Table S6), the exhaustive list of *in silico* modifications resulting in glycans with observed glycowords was generated (n = 1,238,879). All thereby observed monosaccharide and/or bond substitutions were recorded in a symmetric matrix and converted into substitution frequencies by dividing them by the total number of retained modifications. The substitution score S_{ij} for each possible substitution was then calculated with the following formula:

$$S_{ij} = \lambda \log\left(\frac{p_{ij}}{q_i * q_j}\right)$$

The substitution frequency is hereby denoted as p_{ij}, while q_i and q_j describe the observed base frequencies of the respective glycoletters. Additionally, we used λ as a scaling factor (a value of four in this work) to arrive at suitable integer values by rounding all values up or down. Substitutions never observed during this procedure received a final value of −5, lower than any of the observed substitution scores, while the diagonal values of the substitution matrix were set at 5, higher than any of the observed substitution scores. The penalty for gaps for alignments in this work was set at −5, to match the minimal substitution score.

Model Training

All models were trained on an NVIDIA® Tesla® K80 GPU using PyTorch (Paszke et al., 2019). For all models, architecture and hyperparameters were optimized by minimizing the respective loss function. For the language models, we used mixed precision training utilizing the Apex library (https://github.com/nvidia/apex). For language models and classifiers, we randomly split the respective dataset into 80% for training and 20% for validation. A modified stratified shuffle split was used to randomly split glycans into training and validation sets for the species classifier so that, for every class, 80% of the glycans were present in the training set and 20% in the validation set. Further, only classes comprising at least five glycans were used for training and testing the SweetOrigins models. We

employed data augmentation by forming a generalizable subset of all possible isomorphic glycans if a glycan sequence had isomorphic glycans. Specifically, we swapped the order of double branches and exhaustively exchanged the main branch with the side branches closest to the non-reducing end in the bracket notation (Figure 3B). The resulting sequence in the bracket notation still described the same glycan in a slightly different way, increasing model robustness during training. Glycans were converted into lists of glycowords describing the glycans, brought to equal lengths using a padding token facilitating model training, and used in batches of 32 glycans for training and testing.

SweetTalk and the SweetOrigins models for each taxonomic level consisted of a three-layered, bidirectional recurrent neural network using long short-term memory (LSTM) units (Hochreiter and Schmidhuber, 1997) with 128 nodes per layer, including an embedding layer for the glycowords. The concatenated hidden representation learned by the bidirectional LSTMs was then projected to a fully connected layer at the end for the final prediction. The language model SweetTalk was trained by predicting the next glycoletters, given preceding glycoletters, in the context of glycowords, thereby learning the local structural context of glycoletters. The embedding layer for classifiers was derived by first training a glycoletter-based language model and then extracting the learned glycoletters embedding and calculating initial glycoword embeddings for SweetOrigins. The last, fully connected layer in all models was initialized by Xavier initialization (Glorot and Bengio, 2010) and the number of nodes was determined by the number of classes for each classifier. We used a cross-entropy loss function and the ADAM optimizer with a starting learning rate of 0.0001 (decaying it with a cosine function over 100 epochs during training) and a weight decay of 0.005. Additionally, we employed an early stopping criterion after 10 epochs without improvement in validation loss for regularization.

The model for predicting *E. coli* strain pathogenicity followed the same architecture except for using 150 nodes per layer, a binary cross-entropy loss function, and a learning rate of 0.0005. Machine learning models used for comparison comprised random forest classifiers and support vector machines for classification. For the implementation of these models, we used the scikit-learn implementation (Pedregosa et al., 2011). Feature importances were extracted using SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017). Hyperparameters for all methods were optimized by maximization of accuracy via 5-fold cross-validation.

QUANTIFICATION AND STATISTICAL ANALYSIS

This study did not use statistical analysis. All experimental details can be found in the STAR Methods section.