CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Glycoinformatics in the Artificial Intelligence Era

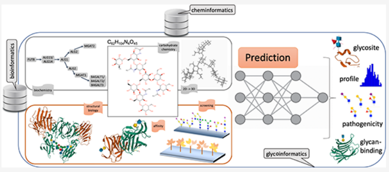
Daniel Bojar* and Frederique Lisacek*

Cite This: *Chem. Rev.* 2022, 122, 15971−15988

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Artificial intelligence (AI) methods have been and are now being increasingly integrated in prediction software implemented in bioinformatics and its glycoscience branch known as glycoinformatics. AI techniques have evolved in the past decades, and their applications in glycoscience are not yet widespread. This limited use is partly explained by the peculiarities of glyco-data that are notoriously hard to produce and analyze. Nonetheless, as time goes, the accumulation of glycomics, glycoproteomics, and glycan-binding data has reached a point where even the most recent deep learning methods can provide predictors with good performance. We discuss the historical development of the application of various AI methods in the broader field of glycoinformatics. A particular focus is placed on shining a light on challenges in glyco-data handling, contextualized by lessons learnt from related disciplines. Ending on the discussion of state-of-the-art deep learning approaches in glycoinformatics, we also envision the future of glycoinformatics, including development that need to occur in order to truly unleash the capabilities of glycoscience in the systems biology era.



CONTENTS

1. INTRODUCTION

Glycoinformatics, sometimes also called glycobioinformatics,[1] can be straightforwardly defined as the application of bioinformatics to glycoscience. Bioinformatics, according to Wikipedia, refers to the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data [https://en.wikipedia.org/wiki/Bioinformatics]. With the rise of systems

Special Issue: Glycosciences

Received: February 10, 2022
Published: August 12, 2022

biology and the expansion of -omics technologies, bioinformatics has become an integral part of research in life science.

The sheer size of experimental -omics data sets has grounded bioinformatics into data science. In recent years, emphasis has been put on the generation of findable, accessible, interoperable, and reusable (FAIR) biological data.[2] Findable is indispensable, because data search is a frequent task that should obviously be made easy to the largest community of life scientists. However, as simple as this task seems, it still primarily requires that data and related metadata (information supplementing data) be associated with a unique and persistent identifier and, secondarily, readability by both humans and computers. Accessible is highly practical because it involves retrieval using these identifiers with a standardized protocol such as hypertext transfer protocol (HTTP). Interoperable is a crucial constraint in attempts to merge or integrate data from different sources. To become interoperable, data need to be described with standard languages reflecting knowledge representations, commonly known as ontologies, otherwise also qualified as controlled vocabularies. For example, gene ontology[3] has revolutionized biomolecular data annotation and enabled rational cross-referencing between data resources. The first three FAIR principles precede the fourth, which ultimately is the goal of these efforts for data sustainability. Reusability can finally be achieved through well-described metadata, including data provenance and community standards. In the end, FAIRness rights to reuse data can be regulated by licensing but FAIR/O cancels any possible limitation and allows for free data reuse and open science. The surge of data generation, sharing, and usage in the recent SARS-CoV-2 pandemic is a good example of application of FAIR principles for everyone's benefit.

Large volumes of consistent data are the ideal input for developing models and methods to predict a biological outcome. Myriads of solutions to predicting molecular shapes/structures, locations, expressions, as well as interactions populate bioinformatics toolboxes. A significant proportion of them rely on artificial intelligence (AI), mostly learning methods. Nonetheless, to achieve robustness and accuracy, these tools require not only quality data but also fine-tuning over time. A striking example is the prediction of protein 3D structure from sequence. Most learning approaches predicting structure from sequence leverage evolutionary information (e.g., via multiple sequence alignments) and/or existing structural information from homologous proteins. Neural networks were actually first used in the late 1980s to predict protein secondary structure from sequence,[4] but the application of such AI-based prediction to 3D structure was delayed for over a decade. Early implementations relied on the prediction of amino acid contact maps,[5] and, again, it took over another decade to bring this approach to the next level with RaptorX-Contact.[6] Such progress was easy to follow through the critical assessment of protein structure prediction (CASP) competition designed to assess the quality of 3D structure prediction tools every second year since 1994 [https://predictioncenter.org/]. RaptorX-Contact used residual convolutional neural networks to predict contact maps from evolutionary coupling and sequence conservation with superior results on CASP11, the 2014 edition. This paved the way to AlphaFold2[7] that outshined CASP14, the 2020 edition, with further improvements and fined-tuned deep learning-based methods. AlphaFold2 predictions are increasingly accessible to users of major bioinformatics reference databases and portals (e.g., UniProt[8]), or to experienced bioinformaticians using community implementations (e.g., ColabFold[9]). In a nutshell and unsurprisingly, decades were needed to reach such excellence.

As a subset of bioinformatics, glycoinformatics faces similar challenges. Glyco-data, much like broad biological data, are spread across biology and chemistry, yet the complexity and the diversity of carbohydrate molecules, as well as their nontemplate driven biosynthesis, have created a wider gap between the two fields.

Carbohydrate chemistry research has been internationally coordinated for many decades through the International Union of Pure and Applied Chemistry (IUPAC), of which it became an associated organization in 1970 [https://ico.chemistry.unimelb.edu.au/]. This age-old grounding in international exchange prompted the need for collecting data which eventually happened in the form of CarbBank,[10] setting the premises of glycoinformatics at a time when bioinformatics was in its infancy. Unfortunately, the path to expansion was rough and long before the field was recognized, as reported in several reviews[11−14] and a dedicated chapter in a reference manual,[15] and despite the hurdles and various intermediary short-lived initiatives for collecting and storing glycan structural data, these have now found a safer place in the universal repository named GlyTouCan, first released in 2016 and hosted in Japan.[16−18]

In parallel, glycobiologists have concentrated their effort on multiple forms of functional studies to reveal that glycosylation is site-specific,[19] tissue-dependent,[20] and influenced by environment.[21] Glycomics and glycoproteomics have matured to provide increasingly comprehensive data sets[22] that have just begun to populate databases.[23] Furthermore, the development of array technology, starting with the Consortium for Functional Glycomics (CFG) initiative, channelled screening data into a single location [http://www.functionalglycomics.org/glycomics/publicdata/home.jsp].

Up to this point in this Introduction, it appears that carbohydrate chemists determine the structural pieces of the puzzle while glycobiologists attempt to place them into a biological context. Yet biochemists hold another key with the elucidation of carbohydrate metabolism and catabolism. Bridging information provided by these different views is challenging. The attachment of solved structures on their conjugate(s) is often unspecified, and the correlation between a set of structures and their biosynthetic pathways is not obvious because chronology may be hard to establish and glycosyltransferase availability is frequently unknown. Quantitative evidence can be sought in transcriptome analyses that can shed light on the expression of carbohydrate biosynthetic enzymes that is notoriously different in distinct tissues, cell types, or diseases. Additional structural constraints can be determined because protein glycosylation is mainly regulated at the level of both the enzymatic machinery and the glycoprotein structure.[24] Nonetheless, glycan-binding experiments are centered on ligands independent of their natural occurrence, making it difficult to reconcile all viewpoints. This situation is clearly presented in a recently published comprehensive overview[25] that will not be reproduced or paraphrased here. Rather, the present review extends this prior description of the glycoinformatics landscape with a focus on learning methods and their applications in glycobiology. To do so, we briefly survey the specificity of glyco-data in the life science data ecosystem as well as the long-standing presence of AI methods in bioinformatics and glycoinformatics. The two aspects, data and AI, are tightly interrelated as exposed throughout this review. Importantly, AI methods are data hungry, and, unavoidably, our coverage is

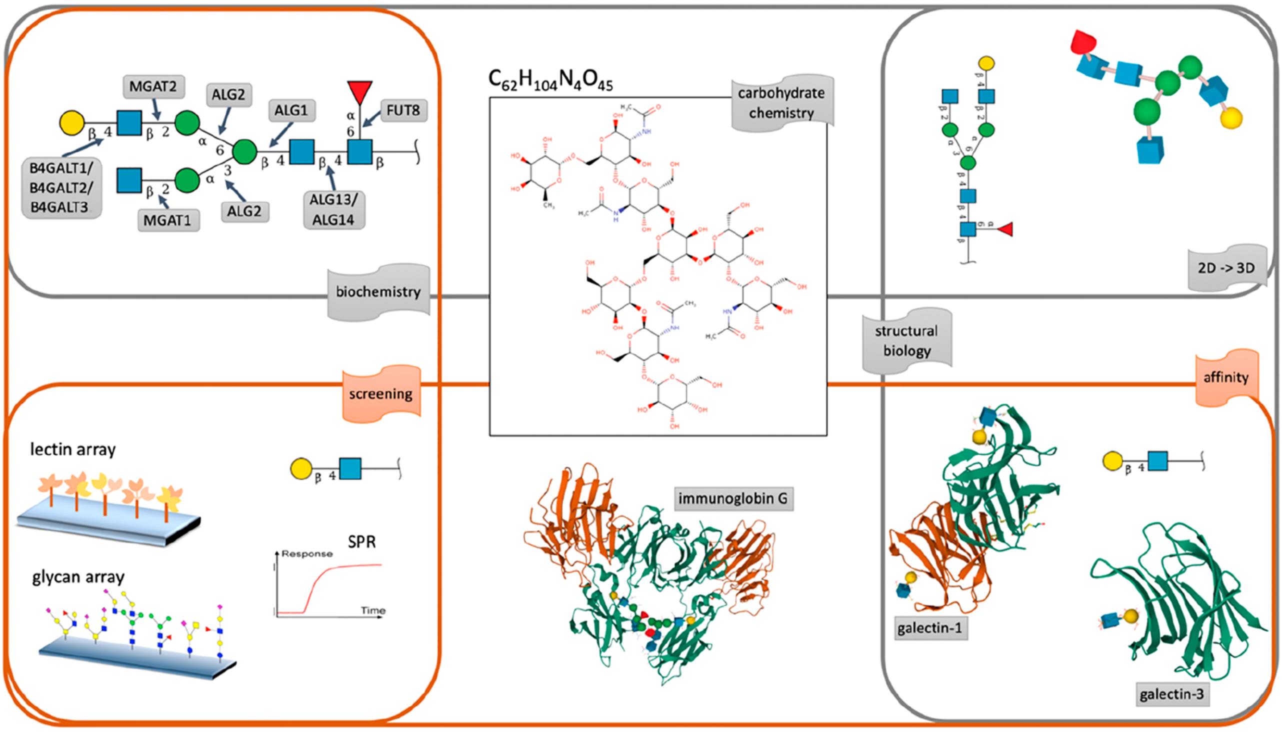


Figure 1. Summary of information collectable from the structure of the $C_{62}H_{104}N_4O_{45}$compound, reflecting the diversity of questions addressed in glycoscience. Whether information is chemical (upper center), biochemical (upper left), and structural (upper right), it requires functional complements dependent on affinity or screening methods. In this example, $C_{62}H_{104}N_4O_{45}$ is shown to be attached to an immunoglobulin γ (lower center) and its terminal *N*-acetyllactosamine moiety recognized by galectin 1 and 3 (lower right) and possibly screened with array technologies (lower left).

therefore biased toward the most abundantly generated glycodata, which tends to be related to glycans in association with glycoproteins (N-/O-linked) and, to a lesser extent, glycan-binding proteins. This panorama is followed by a focus on the spreading of deep learning approaches in glycoscience. Finally, we summarize our view of future development in AI-based applications in this context.

2. IDIOSYNCRASIES OF EXPERIMENTAL DATA IN GLYCOSCIENCE

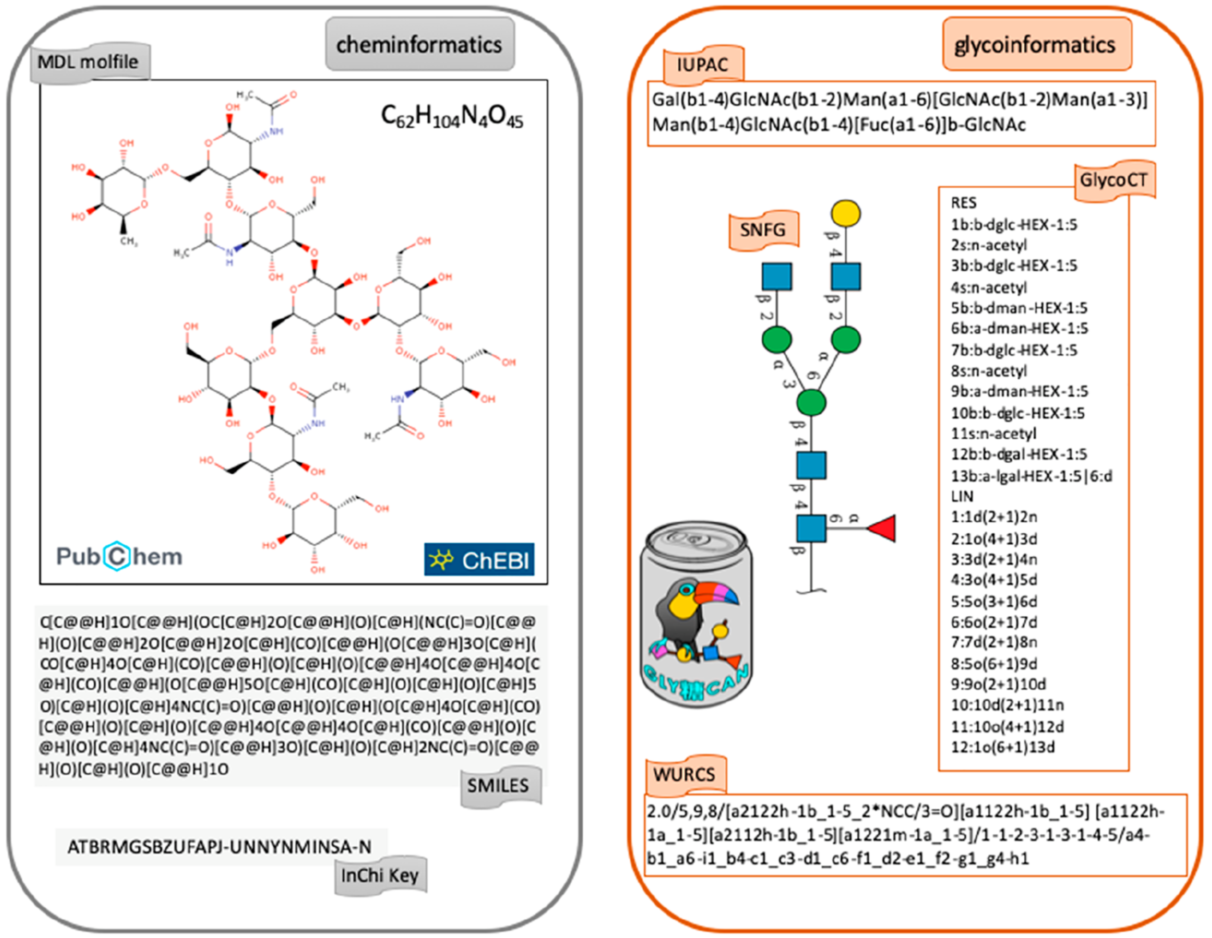
Any prediction or modeling tool requires data processing, and the more precise the definition of the possible solution space, the better the tool will perform. Recalling the fragmented situation presented in the introduction, glycoscience data have unique features that need to be considered.

2.1. Sparsity

At this point in time, the estimate of the "glycan space" dimension is controversial and reminiscent of the debated estimate of the human genome content prior to sequencing it. Speculation about the gene count ranged between 30 000 and 500 000, and actual data forced everyone into a more or less drastic downscaling. Our current knowledge of glycan biosynthesis makes it difficult to set boundaries. In theory, there could be billions of structures considering all known species, but, practically, GlyTouCan currently contains close to ~51 000 structures (version 3.1.0), many of which are redundant due to varying degrees of resolution. Considering one species at a time, *Homo sapiens* is probably the most studied and the figures are not any more precise. At present, the range is often suggested to be on the order of magnitude of 10^4 and it is not clear whether the array of experimental techniques used to solve structures guarantees an exhaustive coverage of glycan structures. In fact, the regular occurrence of paradigm-shifting discoveries of a new glycan type with unconventional strategies tends to suggest that "standard" workflows may miss unexpected structures. The latest examples have revealed bisecting Lewis X structures in the human brain[26] and even to-be-confirmed glycosylated RNA.[27] Additionally, the attention of researchers is predominantly focused on protein-associated N- and O-linked glycans, with less consideration for glycolipids, which is even worse for glycosaminoglycans, lipopolysaccharides, or polymeric glycans. Reasons for this can be seen in the lack of accessible, large-scale, and comprehensive methods to study these molecules as well as their intrinsically higher heterogeneity. In a nutshell, the extent of the glycome remains a very open question. In this situation, data can be qualified as sparse because of their uneven spread in the glycan space. Specifically, the sparsity stems from at least two main sources: (i) the fraction of glycans yet unknown to us in any given species and (ii) the fraction of glycans that is not being measured (or cannot be annotated) in glycomics experiments due to sample processing, low abundance, ionization difficulties, isomers, and many other potential issues. Now, if we consider the estimate of 10^4 total human glycans, the pool of currently known human glycans can be placed somewhere around 3000, while a typical glycomics experiment merely measures dozens to low hundreds of glycans. Both types of sparsity are not only substantial but also due to systematic biases, some of which are mentioned above, making a systems perspective more difficult.

2.2. Heterogeneity

As mentioned earlier, the full qualification of the structure and the function of a glycan usually requires a set of experiments spanning chemistry, biochemistry, affinity, and screening



```
C[C@@H]1O[C@@H](OC[C@H]2O[C@@H](O)[C@H](NC(C)=O)[C@@H](O)[C@@H]2O[C@@H]2O[C@H](CO)[C@@H](O[C@@H]3O[C@H](CO[C@H]4O[C@H](CO)[C@@H](O)[C@H](O)[C@@H]4O[C@@H]4O[C@H](CO)[C@@H](O[C@@H]5O[C@H](CO)[C@H](O)[C@H](O)[C@H]5O)[C@H](O)[C@H]4NC(C)=O)[C@@H](O)[C@H](O[C@H]4O[C@H](CO)[C@@H](O)[C@H](O)[C@@H]4O[C@@H]4O[C@H](CO)[C@@H](O)[C@H](O)[C@H]4NC(C)=O)[C@@H]3O)[C@H](O)[C@H]2NC(C)=O)[C@@H](O)[C@H](O)[C@@H]1O
```

```
ATBRMGSBZUFAPJ-UNNYNMINSA-N
```

```
Gal(b1-4)GlcNAc(b1-2)Man(a1-6)[GlcNAc(b1-2)Man(a1-3)]
Man(b1-4)GlcNAc(b1-4)[Fuc(a1-6)]b-GlcNAc
```

```
RES
1b:b-dglc-HEX-1:5
2s:n-acetyl
3b:b-dglc-HEX-1:5
4s:n-acetyl
5b:b-dman-HEX-1:5
6b:a-dman-HEX-1:5
7b:b-dglc-HEX-1:5
8s:n-acetyl
9b:a-dman-HEX-1:5
10b:b-dglc-HEX-1:5
11s:n-acetyl
12b:b-dgal-HEX-1:5
13b:a-lgal-HEX-1:5|6:d
LIN
1:1d(2+1)2n
2:1o(4+1)3d
3:3d(2+1)4n
4:3o(4+1)5d
5:5o(3+1)6d
6:6o(2+1)7d
7:7d(2+1)8n
8:5o(6+1)9d
9:9o(2+1)10d
10:10d(2+1)11n
11:10o(4+1)12d
12:1o(6+1)13d
```

```
2.0/5,9,8/[a2122h-1b_1-5_2*NCC/3=O][a1122h-1b_1-5] [a1122h-1a_1-5][a2112h-1b_1-5][a1221m-1a_1-5]/1-1-2-3-1-3-1-4-5/a4-b1_a6-i1_b4-c1_c3-d1_c6-f1_d2-e1_f2-g1_g4-h1
```

Figure 2. Contrast of encoding schemes for the same compound in cheminformatics and glycoinformatics. The left panel shows the depiction of $C_{62}H_{104}N_4O_{45}$ as provided in the PubChem and ChEBI databases of chemical compounds. These resources rely on the SMILES and InChi or InChi Key encodings that are used as popular input formats in many cheminformatics and bioinformatics tools. The right panel displays the most commonly used formats in glycoinformatics, namely IUPAC, GlycoCT, and WURCS. In the center, $C_{62}H_{104}N_4O_{45}$ is depicted in the symbol nomenclature for glycans (SNFG), now spreading both in glycoinformatics and in the literature.

technologies that are diverse and the results of which are difficult to corroborate. The situation is illustrated in Figure 1, where the various ways of collecting information on Gal(β1-4)GlcNAc(β1-2)Man(α1-6)[GlcNAc(β1−2)Man(α1-3)]Man(β1-4)GlcNAc(β1-4)[Fuc(α1-6)]β-GlcNAc are highlighted. In each case, the nature of the information and its extraction entails substantially different means, resulting in challenging matching and adjustment tasks in order to rationalize the presence of the glycan in the conditions where it was observed.

Many techniques used in glycoscience require a compromise between efficiency (breadth) and precision (depth), especially in the high-throughput era. Mass spectrometry (MS) will be preferred to nuclear magnetic resonance (NMR) if the experiment entails minimizing the sample load and maximizing the throughput, but it is likely to lower the level of structural details. This aspect as well as further details of glycan structural data acquisition are extensively covered in two recent reviews.[28,29] For our purposes, suffice to say that distinguishing mass isomers is challenged by the multiplicity of commonly occurring monosaccharides such as hexoses present in the form of equally measurable glucose, mannose, or galactose, with the same chemical composition and mass. Here, various techniques such as digestion by monosaccharide-specific exoglycosidases or further fragmentation with mass spectrometry (MS2 to MSn) may result in less ambiguous identification. Likewise, glycomics MS data usually provide quality structural data but no protein site-specific information, while glycoproteomics MS data are precise on mapping glyco-sites but only with low resolution glycan compositions. In the end, the various sources are difficult to merge into a single and clear view of a glycome.

2.3. Field-Specific Encoding

The complications involved in determining a full glycome have two immediate consequences. Generally, the required time and expertise prevent glycoscientists from venturing into any other related-omic field. In turn, life scientists with no training in glycoscience are often disinclined to undertake sizable extra work to investigate glycosylation. In the end, a partial disconnection with biology tends to characterize the production of glyco-data.

From a bioinformatics point of view, the divide also exists. In the past decade, the mapping of metabolic pathways from genomes has brought cheminformatics closer to bioinformatics. This entails sharing data formats to promote data exchange so that reactions can be precisely described,[30,31] with unambiguous substrates and products as well as definite enzymes initially translated from genomic sequences. All chemical compounds of the reference PubChem[32] and ChEBI[33] databases are described with SMILES[34] and InChi/InChi Key[35] encodings that are

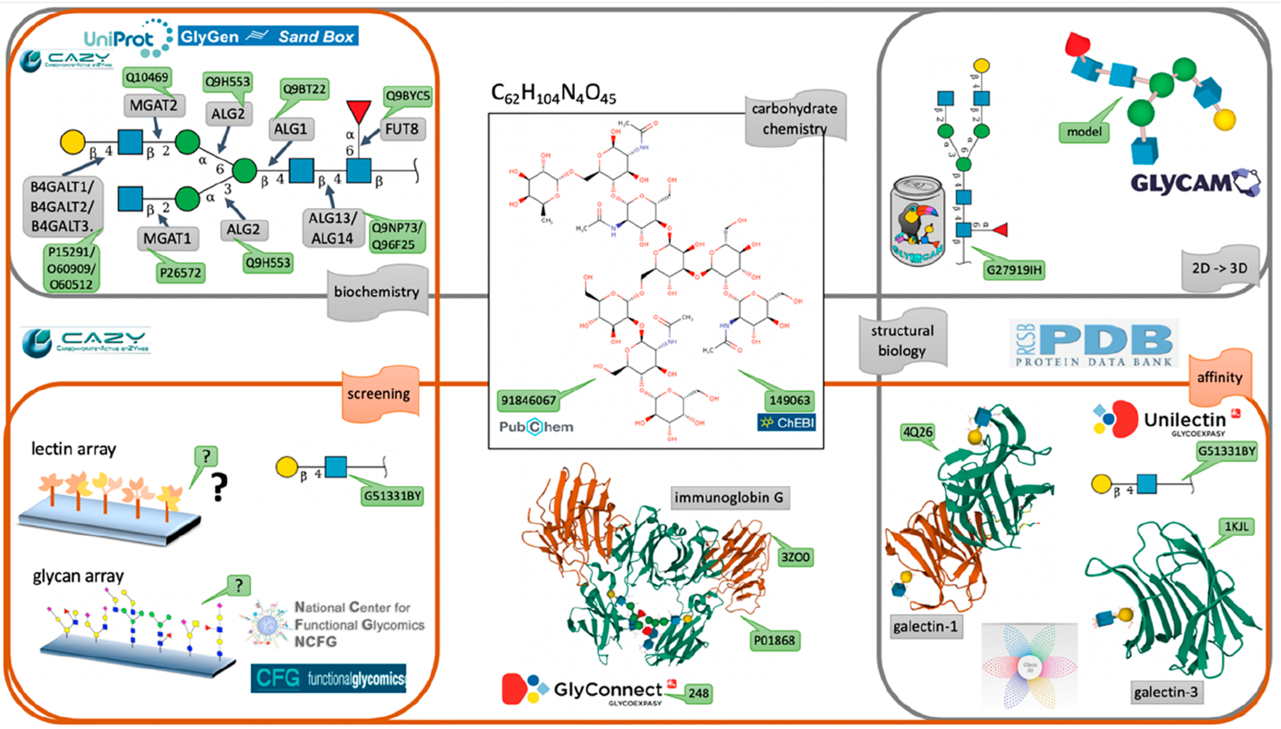


Figure 3. Diversity of bioinformatics resources with database identifiers. The exact same illustration of Figure 1 is kept and complemented with IDs (green tags) from the selection of relevant databases. Enzyme data can be found in both the CAZy and the UniProt databases. The GlyGen Sand Box provides the details of each step of biosynthesis. Structural details of glycoproteins and glycan-binding proteins are provided by the PDB channelled through the GlyConnect and UniLectin3D databases, respectively. Screening data are not precisely specified.

readable by the vast majority of cheminformatics tools. All corresponding depictions are generated from MDL molfiles. The specification of biochemical pathways relies on the knowledge stored in PubChem and ChEBI.

The convergence of cheminformatics with glycoinformatics is not as clear. All glycans of GlyTouCan are encoded in in IUPAC,[36] GlycoCT,[37] and WURCS.[38] Each structure in this database is represented in the symbol nomenclature for glycans (SNFG) that has been adopted as a standard in glycoscience.[39,40] Nonetheless, in recent years, closer interactions between GlyTouCan, PubChem, and ChEBI led to include the WURCS encoding and the SNFG notation in glycan entries of the latter two databases. Figure 2 illustrates the parallel options taken in cheminformatics and glycoinformatics. It should be noted that Figure 2 highlights the best case scenario of a fully defined structure. In reality, glyco-data often lack compositional or linkage information that is better handled by glycoinformatics-specific formats.[41]

The usage of multiple nomenclatures requires continuous harmonization. For example, when a new monosaccharide substituent (or modification, such as added phosphate or methyl) is discovered, it must be included in the encoding format (except for WURCS that was created partly to avoid this situation). This, in turn, impacts conversion software that must be maintained.[42] These efforts are costly but essential in keeping a connected community. Another illustration of confusion created by multiple and independent contributions is the case of drawing glycan structures. The uncoordinated development of web interfaces have resulted in a panoply of different tools,[43] which allow researchers to easily visualize glycan structures in their presentations and publications. This gave rise to a substantial variety of (i) specific colors for monosaccharides, (ii) depictions of linkage (undirected, directed, dashed) (iii) and linkage label (β1-4, β4, 4, 14β), and (iv) choice of reducing end depiction (nothing, protein backbone, OH). In the case of plant or fungal polysaccharides, there is also an abundance of trivial names (such as arabinan or glucan) as well as the ambiguity of where the repeating unit starts. Not only is such a variety detrimental to the implementation of a universal depiction such as SNFG, it also makes it confusing for newcomers to the field to understand meaningful associations in the visualization of glycan structures.

3. GLYCO-DATA REPRESENTATION

3.1. Lessons Learned from Bioinformatics

The precise recording and depiction of the heterogeneous information illustrated in Figure 1 is a definite glycoinformatics challenge. Figure 3 highlights the possibility of referring each and every entity: a glycan, its biosynthetic pathway, or the epitopes it contains, in an appropriate database, with a unique and stable identifier. This view is widely spread in bioinformatics and not completely realistic in glycoinformatics. To be effective, FAIR principles mentioned in the introduction apply to data and metadata (information about that data). For that reason, minimizing ambiguity is of the essence. The precision of description is guaranteed by associating each piece of data with a database identifier, shown as green tags in Figure 3 in a reproduction of Figure 1. Most of the cited databases also contain metadata.

Biochemical knowledge has been traditionally covered by the CAZy database, where cazymes (carbohydrate-active enzymes) are collected and classified.[44] CAZy revolves around amino acid sequence annotation and has grown in the past decades in close relation with NCBI genomes,[45] Swiss-Prot,[46] and UniProt,[8] allowing the unambiguous characterization of cazymes via

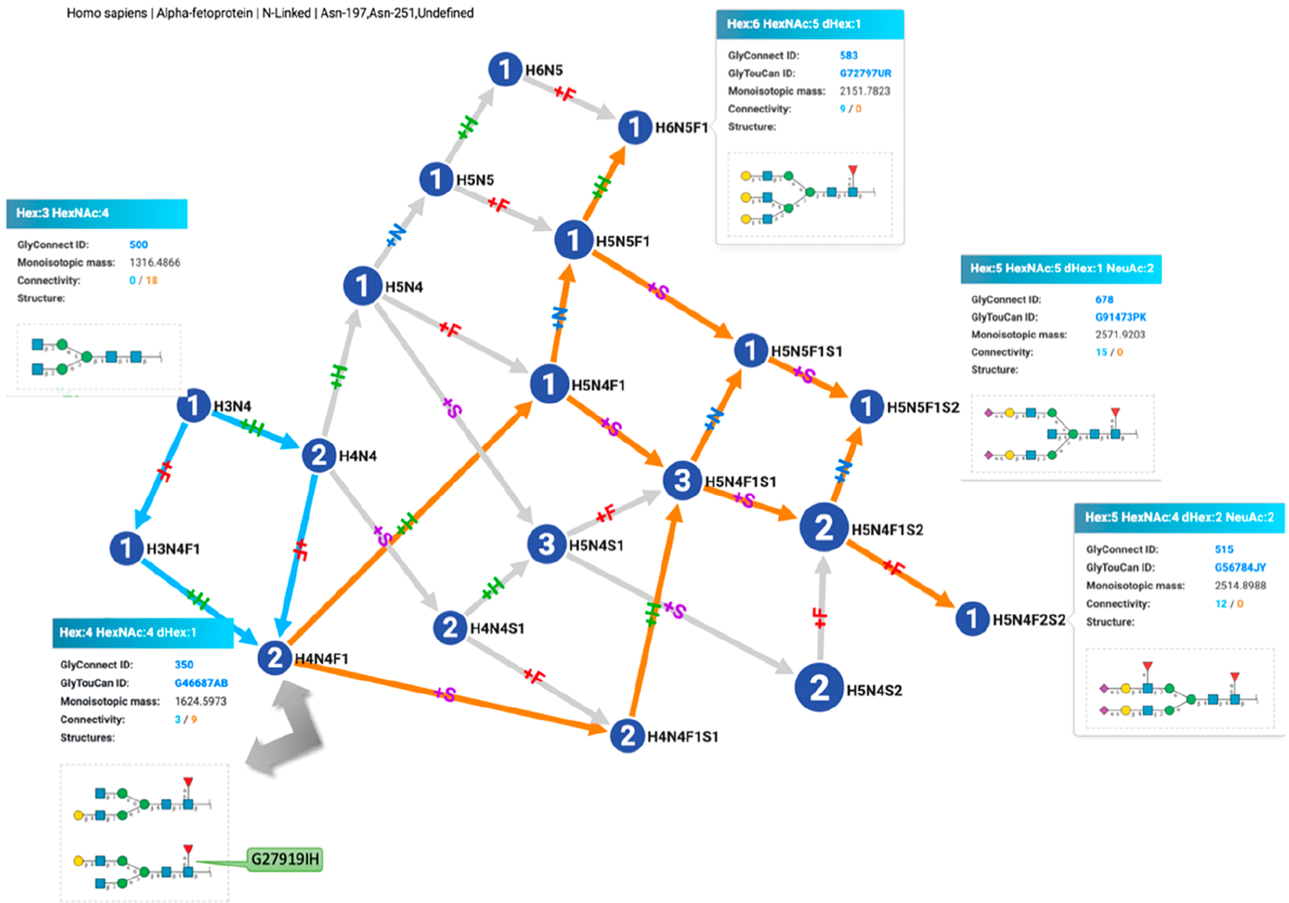


Figure 4. Relatedness of glycan structures of the human α-fetoprotein glycome (mapped with GlyConnect Compozitor). This representation emphasizes how listed glycans composing a protein glycome are tied together in terms of shared substructures. In this graph, cyan incoming paths connect glycan compositions and associated structures to GlyTouCanID G27919IH, as its substructures while orange outgoing paths connect G27919IH to glycan compositions and associated structures that include it.

sequence accession numbers. In 2021, the GlyGen project[47] released an interface to visualize the stepwise synthesis of GlyTouCan registered structures, called the SandBox [https://glygen.ccrc.uga.edu/sandbox/]. From a structural biology point of view, precision is brought by the knowledge of three-dimensional protein structures stored in the Protein Data Bank (PDB).[48] A plugin to the LiteMol structure visualization software conveniently represents carbohydrates attached or bound in the 3D-SNFG representation;[49] in this example, the known *N*-acetyllactosamine terminal motif of the example *N*-glycan structure, referred to as ID G27919IH in GlyTouCan. *N*-Acetyllactosamine is also recorded in GlyTouCan as ID G51331BY to provide a precise ligand definition, which can be used in turn by UniLectin3D[50] that covers knowledge of lectins, also known as carbohydrate-binding proteins. UniLectin3D uses PDB IDs to reference lectins that are human galectins in this example. GlyTouCan ID G27919IH also appears in the GlyConnect database[51] as attached to human immunoglobulin gamma (GlyConnect ID 278; UniProt P01868; PDB 3ZO0). Figure 3 also reveals the weakness of the screening information. Many array experiments are undertaken, but very few are collected. It was the purpose of the Consortium for Functional Glycomics[52] at the turn of the century, but this initiative has ended. The National Center for Functional Glycomics has taken over and is preparing the launch of a new repository [https://ncfg.hms.harvard.edu/microarrays]. Other initiatives, such as GlyMDB[53] or CarbArrayArt, have made provision for storing array data in a rational manner.[54]

Glycan data management and exchange is significantly helped by the Minimum Information Required About a Glycomics Experiment (MIRAGE) project initiated by the Beilstein Institute in 2011 [https://www.beilstein-institut.de/en/projects/mirage/].[55] It follows the Minimum Information Standard movement that has produced sets of guidelines and formats for reporting experimental data in the past two decades, especially those generated with high-throughput methods [https://en.wikipedia.org/wiki/Minimum_information_standard]. At this point in time, very few glycoinformatics resources collect raw data to make them accessible to the community. As discussed for mass spectrometry data,[56] channelling data through a pipeline is needed but to date still incomplete. GlycoPOST[57] is the first implementation of a working MS data repository. Cited glycan array-related projects are compliant with the corresponding guidelines.[58]

Each of the data sources mentioned above attempt to comply with existing controlled vocabularies and ontologies as pointed out as mandatory in the FAIR principles.

3.2. Lessons Learned from Proteomics

The dominance of mass spectrometry (MS) in proteomics sets a precedent for glycomics. In particular, the evolution of peptide MS data processing offers clues to handling glycan and glycopeptide MS data. In the early days of proteomics, the main objectives of MS data processing were the improvement of protein identification and the increase of its rate via automation,[59,60] giving rise to lists of identified proteins in association with a tissue or a cell line. Rapidly, the need for making sense of those lists spurred the implementation of tools, enabling comparative methods.[61] Finally, the concomitant development of interactomics led to map protein interaction networks to support the interpretation of coidentified proteins in a sample.[62] As glycomics lags behind proteomics, the progression is similar but not as advanced. At this point in time, lists of identified glycans are being published yet often still lacking a precise identifier despite the existence of a universal glycan data repository.[18] These lists are often provided as independent items and their possible relatedness limited to the determination of trends. Many publications report sialylation, fucosylation, or bisecting GlcNAc (among others) as prevalent features of a glycome content, which are in turn considered as a summary representation of a list. Nonetheless, the dependency of listed structures is reflected in glycan synthesis, which is a stepwise process easily visualized with graphs in which each connection represents the addition of a single monosaccharide. Figure 4 shows such a graph where GlyTouCanID G27919IH is now shown as a component of the human alpha-fetoprotein (UniProtID P02771). This representation was generated by GlyConnect Compozitor that processes glycan compositions,[63] as opposed to defined structures, to handle current glycoproteomics data stored along glycomics data in the GlyConnect database. In Figure 4, the graph represents the glycome of human α-fetoprotein (UniProt ID: P02771) as recorded in GlyConnect and curated from seven publications. The view is centered on the composition corresponding to G27919IH and shows highlighted paths in cyan to map G27919IH substructures and in orange to map all structures, of which G27919IH is a substructure. Each leaf of the graph in that example is shown to emphasize the possible diversity in a single protein glycome.

4. CLASSICAL MACHINE LEARNING IN BIOINFORMATICS

4.1. Decades of Trials and Errors

It took almost two decades to realize the power of applying dynamic programming[64] to amino acid sequence alignment[65] but only two years for early bioinformaticians (not designated as such at the time) to implement revitalised neural networks[66] in gene promoter[67] or protein secondary structure[4,68] prediction from sequence data. From then on, the most efficient sequence motif/pattern prediction methods have heavily relied on machine learning (ML) methods. This approach was soon popularized in bioinformatics through the dissemination of a reference manual[69] (2nd edition in 2001).

ML methods in the form of neural net(work)s (NN), support vector machine (SVM), and some versions of hidden Markov models (HMM) have been applied to a broad variety of prediction, classification, and discovery related problems. The point of this review is not to cover these topics in detail and repeat previous work (see numerous references in *Briefings in Bioinformatics*, Oxford Press) but to provide a few landmarks in order to set the scene for introducing the application of ML in glycoscience. Note that ML in cheminformatics was previously and extensively described in this journal for text mining.[70]

In a nutshell, ML techniques require a set of examples (training set) from which regular features are extracted to define the profile of elements of the training set. A scoring function is then defined and used to decide whether a new object matches the learned profile. This very short summary emphasizes the importance of describing the examples with appropriate descriptors that will provide the salient features to be extracted. Furthermore, the examples need to be carefully selected to be considered as representative of the aimed-for trend and, unless maintained and/or providing the option of retraining, the application of an ML method is not valid for long.

Numerous ML-based applications have come and gone with the expansion of -omics and systems biology. On the one hand, the exponential growth of data sets has regularly challenged bioinformatics tools that could not scale or keep up with updates. On the other hand, examples in genomics suggest that it can be exceedingly difficult to avoid bias and overconfidence when applying ML to biological data.[71] In contrast, the prediction of protein export cleavage sites illustrates the value of a well-defined problem that finds a suitable ML solution withstanding the test of time. SignalP[72] was first released in 1997 and, in its current sixth version, is still widely used for predicting the presence of an N-terminal signal peptide. The full coverage of the development and evolution of the tool up to version five was recently reviewed.[73] Suffice to say that SignalP 1.0 was implemented a neural network (NN), while SignalP 2.0 included a supplementary HMM prediction. This addition aimed at distinguishing signal peptides from anchors. SignalP 3.0 introduced a new D-score to strengthen the specificity of signal peptides compared to other sequences complementing the HMM prediction. SignalP 4.0 was shaped back as a pure NN-based method that turned out having blind spots that were corrected in SignalP 4.1. To keep up with learning method improvement, a shift to deep learning was initiated in 2018 with the release of SignalP 5.0,[74] designed to account for signal diversity across species. In fact, SignalP 5.0 and its recent successor SignalP 6.0[75] belong to section 5 of this review, and they are cited here simply to highlight the value of a well-formulated problem relying on well-defined data. In this case, an ML-based prediction tool was adapted over 25 years, to evolving technology through thoughtful upgrades. SignalP 6.0 now considers five types of signal peptide at each position and relies on a BERT-type language model (see section 5).

4.2. Useful Toolboxes

ML methods have been democratised with the release of libraries such as WEKA,[76] Shogun [https://www.shogun-toolbox.org/], and mlpack,[77] to name a few, that allow for fast implementation and integration into computational analyses . Next to providing predictions from biological data, ML methods also allow researchers to work with learned similarities (also called representations or embeddings), to visually and quantitatively compare samples. Ranging from protein sequences to tandem mass spectrometry spectra, complex data is usually very high-dimensional and therefore hard to visualize and understand in a granular manner. A benefit of applying machine learning to these types of data is that a numerical representation is learned, which can be projected into two-dimensional space (amenable to plotting it) by a variety of methods. Examples of this are distributed stochastic neighbor embedding (t-SNE) or uniform manifold approximation and projection (UMAP[78]), both of which aim at finding a two-dimensional formulation of the data that best preserves distances in the original dimensionality.[79] These methods are particularly popular in the single cell-based -omic fields with a tendency to create opposing communities.[80,81]

4.3. First-Generation AI for Glycomics (1990−2004)

4.3.1. Optimizing Mass Spectrometry Processing. Some parts of the mass spectrometry (MS) pipeline provide opportunities for automation and enhancement via machine learning. Although initially designed for proteomics, the fine-tuning of data analysis offers valid points for glycomics. Probabilistic models were first introduced to train the prediction of peptide fragment intensities in MS/MS.[82] Then, the generalized use of a false discovery rate (FDR) as a confidence assessment of peptide identification has led to the design of efficient scoring schemes supporting the discrimination between experimental and decoy mass spectra, as implemented in Percolator[83] or in Barista[84] software tools. Note that FDR calculations are still lacking in most glycoproteomics data analysis software, as revealed in a recent community challenge.[85]

De novo sequencing is another challenging area of MS-based identification both in proteomics and glycomics, although not yet conclusive in the latter case. A support vector machine (SVM) model was for instance proposed to optimize the score of cross-ring ions and other structural features in order to improve structure assignment from MS/MS spectra.[86]

4.3.2. Predicting Glycosylation Sites. Searching for glycosylation patterns has very early on provided challenging questions for ML methods. In *N*-glycosylation, glycans are bound to a nitrogen atom of an asparagine (Asn or N) residue. The attachment on this amino acid was shown to be characterized by a short consensus sequence: N-X-S/T where S is a serine (Ser), T is a threonine (Thr), and X is any amino acid except proline[87] (Pro or P). In *O*-glycosylation, glycans are linked to an oxygen atom of a serine (S) or a threonine (T) and, occasionally, to a hydroxyproline or a tyrosine. No consensus sequence has been observed around *O*-glycosylated sites except for an unusual abundance of hydroxylated amino acids. Sequence alignments emphasized the frequent occurrence of proline residues close to the glycosylation sites, especially just before and three positions after the glycosylated residue. In contrast, the presence of charged amino acids at these sites tends to prevent the placement of glycans.[88] In *C*-glycosylation, also called *C*-mannosylation, glycans get attached to the carbon atom of a tryptophan (Trp or W) residue and the W-X-X-W motif was identified as the acceptor consensus sequence for glycan binding.

Notable efforts in bioinformatics have boosted both the discovery and mapping of *O*- and *N*-glycosylation sites. Results of large-scale mapping of human *N*-sites has been collected in UniPep[89] (found on proteins isolated from plasma, cerebrospinal fluid, various tissues, and cell sources) and *N*-glycosite Atlas[90] (22 human tissues/body fluids). A genetic engineering approach using human cell lines has enabled proteome-wide discovery of GalNAc-type *O*-glycosylation sites.[91] Despite the value of these data sets and the development of new technologies, the experimental determination of glycosylation sites only roughly characterizes glycans. To compensate for the lack of data, in silico prediction has been, and continues to be, developed mainly based on ML methods. These were trained primarily to recognize the protein sequence context but also secondary structure features and accessibility of glycosylated residues.

Neural networks were implemented early on in a set of online tools made available from the late 1990s on a server of the Technical University of Denmark. The collection spanned the prediction of *N*-glycosites with NetNGlyc, *O*-glycosites with NetOGlyc, as well as in O-GlcNAc sites with YinOYang.[92] The training sets used in each case were very limited at the time and, expectedly, did not yield highly reliable results. Nonetheless, they settled as references in the field. Years later, the sensitivity of NetOGlyc considerably improved following the massive information input provided by Clausen and colleagues.[91]

4.4. Second Generation (2005−2015)

4.4.1. Random Forest and Support Vector Machines. Two approaches destined to increase learning quality of neural networks were introduced a few years later, namely support vector machines (SVMs)[93] and random forest (RF),[94] and both were used to refine the quality of glycosite prediction. All of the corresponding contributions demonstrated how they outperformed NetNGlyc and NetOGlyc following this enhancement. On the basis of a random forest implementation, GlycoMine[95] was shown to improve the prediction performance for the three major types of glycosylation in human. The method relies on employing intensive feature selection techniques and integrating several informative features (e.g., sequence-based, structural, and functional features) to predict the glycosylation sites in a protein of interest. However, the online version of GlycoMine was last updated in August 2017 and is currently accessible only via an archived web-link [https://web.archive.org/web/20170812131319/http://www.structbioinfor.org/Lab/GlycoMine/]. The GPP (glycosylation prediction program)[96] also relies on a random forest predictor exclusively based on protein sequence features. Other software solutions hinge on the implementation of SVMs, whether single in GlycoEP[97] on multiple protein features (sequence, structure, etc.) or combined in EnsembleGly[98] on strictly sequence features or part of a multistage approach as in N-GlyDE,[99] also relying on multiple structural protein features and calculated features such as gapped dipeptides. Table 1 recapitulates this range of published tools and helps comparison. In general, preexisting methods to calculate protein features such as secondary structure,[100] are selected and integrated in the prediction workflow.

Table 1. Summary Table of Cited Methods for Glycoprotein Site Prediction

tool name	N/O glycans	method	features
NetNGlyc	N	NN	sequence
NetOGlyc	O	NN	sequence, structure
GlycoMine	N/O	RF	sequence, structure, functional
GPP	N/O	RF	sequence
GlycoEP	N/O	SVM	sequence, structure, evolutionary
EnsembleGly	N/O	SVM	sequence
N-GlyDE	N	SVM	sequence (gapped dipeptides), structure
SPRINT-Gly	N/O	SVM/ NN	sequence, structure, physicochemical, evolutionary
DeepNGlyPred	N	NN	sequence (gapped dipeptides), structure, evolutionary

The prediction of *O*-GlcNAcylated sites in DB-OGap[101] was also considered an improvement over YinOYang with an amino acid sequence-based SVM model and more recently with a combination of techniques spanning SVM and random forest.[102]

In the end, none of these methods include training data accounting for the actual structures attached on *N*- or *O*-sites despite the earlier suggestion that distinct structural features of glycans correlate with protein structure[103] and the observed changes in protein sequence alignment surrounding sites depending on the properties of attached structures, e.g., core-fucosylated vs nonfucosylated.[104]

Naturally, improvement was also sought in *N*-glycopeptide identification from tandem MS data. SVMs[105] as well as random forest[106] were introduced to define more effective scoring schemes for intact glycopeptide-spectrum matches. Yet the limited reliability of intact glycopeptide identification from MS data, as recently demonstrated,[85] is not related to the use or absence of use of ML-based scoring functions.

4.4.2. Classifying Glycans. Probabilistic models and machine learning methods are commonly brought together as complementary approaches in bioinformatics and this observation transfers to glycoinformatics. In particular, the tree-shaped structures of glycans have inspired bioinformaticians to use hidden Markov models (HMMs) to classify glycan structures.[107] Other attempts produced SVM-based glycan classification,[108] but these paths have not been further explored.

5. NEXT-GENERATION MACHINE LEARNING

5.1. Background

Deep learning has recently revolutionized the analysis of large volumes of biological sequences. One key advantage of deep learning is the ability to work with "unstructured" data, such as sequences or images, without having to define or calculate features that are used for establishing correlations or for training a model.[109] Instead, the deep learning model (ideally) learns the relevant features in a data-driven manner that are then used for prediction. This makes large data sets, such as sequence repositories, directly accessible for being used for model training and has the potential to be less biased and more potent, as features are not restricted to human-defined data characteristics. A key distinction in deep learning models is their separation into "supervised" and "unsupervised" models. Supervised models are trained with known labels, which means that, given the features (e.g., sequence etc.) of a data point, the model aims to predict the label (e.g., protein function). Unsupervised models, on the other hand, only have access to features, not labels. In this case, the model merely learns similarities of data points based on their features. An example would be an unsupervised model trained on protein sequences that would allow clustering of these proteins based on their learned similarity.

A multitude of different model architectures, algorithms of how to learn mapping the input to the output, are available and used in deep learning applied to biology. The ones relevant in the context of glycobiology are feed-forward neural networks, recurrent neural networks, graph neural networks, and transformer-based models. In feed-forward neural networks, a number of layers is constructed, in which each layer learns a function for combining the output from the previous layer to achieve a meaningful prediction of the output at the final layer.[110] This architecture was one of the first historical deep learning methods and remains relatively simple yet popular in usage. Language models, such as recurrent neural networks (RNNs) or transformer-based methods, were originally developed to analyze human languages such as English, yet have been applied to the analysis of biological "languages", such as DNA or proteins, as well. Another prominent architecture that recently also surfaced in computational biology is the transformer. In RNNs, inputs are analyzed sequentially, e.g., using each word of a sentence as an input to the model and propagating the intermediate output of this procedure to the analysis of the next word.[111] As a consequence, RNNs exhibit a limited form of memory, important in analyzing sequential data. The key principle of transformer-based methods is to abandon the sequential approach and analyze language-based data simultaneously.[112] This is made possible by the application of "attention", an algorithmic trick to let the model learn which parts of the sentence/sequence are relevant in the context of prediction. Finally, graph neural networks have been developed to analyze nonlinear data formats, from social media networks to protein 3D structures.[113] The key operations here are convolutions and pooling to convert the irregularly shaped graph input into a fixed-size set of numbers that can be used to reach a prediction. Convolution operations describe nodes in a graph by the features of its neighbors, while pooling operations condense the information from this process, for instance, by collecting maximal or mean values from the convolutions across the graph.

The most prominent application of deep learning in biology in recent history can be seen with the emergence of protein structure prediction models such as AlphaFold2.[7] Providing a scalable means of predicting 3D structures of proteins from their sequence, AlphaFold2 also already impacted glycobiology, with the potential of integrating modeled carbohydrate structures with AlphaFold2-derived protein structures.[114] Another example of how deep learning, applied to biological sequences, has advanced glycobiology can be found in the prediction of glycosites using neural networks and protein sequences,[91] as discussed above. From a chemistry perspective, Ardejani et al. recently presented the combination of quantum mechanical calculations and machine learning to study protein−*N*-glycan interactions in detail.[115] These applications and others predominantly build on the prolific literature on protein-focused deep learning and large associated data sets.

5.2. Deep Learning and Glycobiology

Despite the proliferation of the literature on this topic, the deep learning-driven analysis of proteins has only begun relatively recently, with the development of algorithms such as UniRep[116] in 2019, that was designed to learn the fitness of point mutation variants. In the case of UniRep and most other deep learning-based language models, protein sequences are viewed as a biological language, with amino acids as characters of a (very long) protein "word". The model is then trained by predicting the next character (i.e., amino acid), given the preceding characters. For this prediction task, models such as UniRep not only consider the identity of amino acids but also can learn amino acid properties (akin to size, polarity, etc.) via a trainable embedding or, synonymously, representation. This representation vector, once properly trained, can be viewed as expressing amino acid similarity, as the distance of the representation of two physicochemically similar amino acids such as aspartate and glutamate, should be smaller than two very dissimilar amino acids such as glycine and arginine. Once trained on this character-by-character approach, the final model can then consider the full protein sequence to predict protein properties.

Machine and deep learning approaches for protein sequences can have immediate implications for glycobiology, such as when they are applied to glycosyltransferases. Using calculated protein sequence properties as input for a gradient-boosted regression tree model, monosaccharide donor specificity could be predicted for fold A glycosyltransferases.[117] Shortly after, the

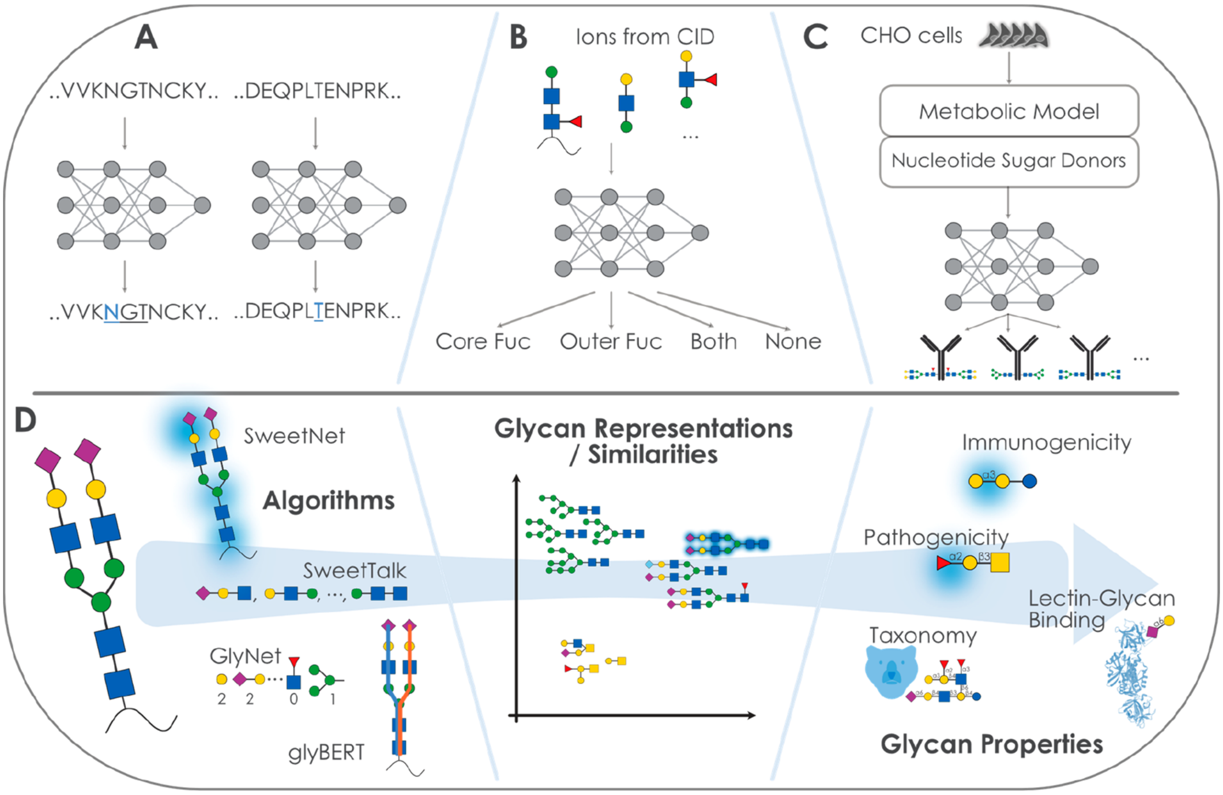


Figure 5. Applying next-generation machine learning to glycobiology. (A) Neural network-based models can be used to predict glycosylation sites from protein sequences. (B) The fucosylation state of glycopeptides can be predicted via neural networks or support vector machines from ion fragments. (C) On the basis of a metabolic model and a neural network, the distribution of glycosylation states of recombinant proteins could be predicted in CHO cells. (D) Using deep learning to predict glycan properties. Several algorithms with their conceptual approach to analyzing glycan sequences are shown. These algorithms learn a glycan similarity or representation that can be used for clustering. The same learned features can also be used to predict a variety of glycan properties, some of which are shown here.

same authors presented a deep learning approach, using a convolutional neural network with attention, to predict the fold of glycosyltransferases from their sequence.[118] This allowed them to identify glycosyltransferase families which are likely to exhibit novel folds that should be further investigated with regard to their structure.

Corresponding analyses for applying deep learning directly to glycans have emerged later than those for proteins, such as SweetTalk.[119] Many reasons for this delay could be mentioned: (i) a relative lack of available sequence data that could support large unsupervised model training, (ii) a relative lack of available labeled data that connects glycans with properties or outcomes for supervised model training, and (iii) the sequence diversity of glycans, comprising hundreds of building blocks and nonlinear sequences due to branching.

The first advances of deep learning applied to glycan sequences have largely mirrored developments in applying AI to proteins. This included the development of models treating glycans as a list of features (see GlyNet[120]), a biological language (see SweetTalk[119] and glyBERT[121]), or as molecular graphs (see SweetNet[122]). A general observation here is that, over time, models have changed to accommodate the branched, nonlinear nature of glycans, which has led to substantial improvement in the quality of predictions.[122]

Models treating glycans as a list of features stem from a long history in drug development, for which "fingerprinting" is a common practice.[123] Here, a chemical is described by a list of standardized features, for instance the absence/presence of certain chemical moieties or the connectivities of atoms. Analogously, glycans can be characterized by a vector cataloguing the absence/presence/number of sequence motifs, for instance on the monosaccharide or disaccharide level. Importantly, fingerprints of this type are not bijective, as it is not generally possible to uniquely reconstruct a glycan from its fingerprint, and one fingerprint may describe several distinct glycans exhibiting these motifs in different configurations. The fingerprint is then used as input for, typically, a feed-forward neural network, a model with multiple layers that learns a function to combine information from previous layers into subsequent layers. Here, each neuron of the first layer has access to information on one feature of the fingerprint, which then is further combined in later layers of the model before arriving at a prediction that is informed by information from the fingerprint.

Glycan language models are closest to the protein language models mentioned above. However, differences arise between the different language models. One distinctive aspect is what a token signifies in a model. One such model type would be the transformer, a model designed for handling sequential data. A transformer does not process a sequence from beginning to end but rather learns from parts of the sequence. In other words, a transformer identifies the meaningful bits. In the transformer-based glyBERT, a token corresponds to a monosaccharide, while the recurrent neural network-based SweetTalk considers larger units, trisaccharides, as tokens in the language of glycans. A recurrent neural network is also designed for sequential data but does process a sequence from beginning to end, keeping previous parts of the sequence in memory to arrive at a final prediction. Next, the process of model training also differs between the two model types. SweetTalk-type models function analogous to UniRep described before by predicting the next token given the previous tokens. Transformer-based models such as glyBERT, however, operate by the principle of attention:[124] given the whole sequence, the model learns which parts are salient (i.e., important for prediction) in order to focus on them with regard to prediction. Both types of models also have an embedding layer, as described in the context of UniRep, to learn similarities of monosaccharides or larger structures.

Finally, graph neural networks consider glycans to be akin to molecular graphs, with precedent models in drug discovery that treated chemicals as molecular graphs.[125] This is fitting in the context of glycans, as glycans are trees, branched sequences with no full cycles, and therefore a special case of graphs. From this viewpoint, a monosaccharide can be considered a node in this graph and, similar to the other discussed techniques, each node type can have its own features, namely a trainable embedding vector. Graph convolutional neural networks, such as the mentioned SweetNet, learn graph "neighbourhoods" via several convolution operations. A convolution in this context is a filter that is iteratively applied over the whole graph to only continue with relevant information in the rest of the model. Each convolution considers the features (i.e., embeddings) of neighboring nodes to describe a node. Each subsequent convolution then has a wider definition of what constitutes "neighbouring", describing a larger portion of the graph in the process. After doing this for every node and its neighborhood, these description features are then passed along to a neural network that uses this learned graph representation to arrive at a prediction.

As already mentioned, deep learning models for glycans have the additional advantage of learning glycan similarity, a concept which can of course also be expressed without machine learning in principle, for instance via counting motifs.[126] In various contexts, learned glycan similarities have been shown to cluster by glycan class and/or characteristic motifs.[122] Next to being used for downstream models that can use these similarities as new features for prediction, learned glycan similarities can be used to visualize clusters of related glycans, for instance via t-SNE or UMAP mentioned above, and allow for interpretation of learned glycan associations. In a recent study, this has been for instance used for an in-depth investigation of the role and properties of different fucose-containing motifs across taxonomic kingdoms.[127]

5.3. Glycosite Prediction

As mentioned in section 4.4.1, glycosite prediction remains a challenge without the inclusion of glycan data. Nonetheless, the two most recent additions to the available models for the prediction of *N*-glycosylation have implemented deep learning methods (see Figure 5A). SPRINT-Gly[128] trained an SVM and deep NNs on calculated amino acid, evolutionary, structural, and physicochemical features, and, in the same vein, DeepNGlyPred[129] is a deep NN, trained on sequence-based features (e.g., gapped dipeptides), predicted structural features, and evolutionary information. Table 1 recapitulates the glycosite predictive tools cited in sections 4.3.2 and 4.4, as well as in the present one.

5.4. From Mass Spectrometry to Glycosylation

To solve, or at least ameliorate, the relative lack of data in glycobiology, improvements in the data acquisition pipeline are necessary for advancing glycan-focused deep learning. Recent work has focused on evaluating the potential of machine learning and deep learning to achieve these improvements. For example, both support vector machines (SVM; machine learning) and neural networks (NN; deep learning) were assessed to predict core and outer fucosylation from glycopeptides, specifically from CID-based ions[130] (see Figure 3B). Encouragingly, both the SVM and the NN model matched manual interpretation, resulting in a near-perfect prediction when assuming that the manual interpretation reflects the biological reality. Approaches such as this could lead to a considerable improvement in speed, cost-efficiency, and, thereby, throughput. Other recent endeavors rely on deep learning, such as on the usage of bidirectional recurrent neural networks, to predict fragment ion intensities.[131] Deep learning was also used in Prosit,[132] with which both fragment ion intensity and retention time are predicted. Prosit combines a bidirectional recurrent neural network, applied to protein sequences, with an attention layer. Similar methods for the prediction of glycan fragmentation would be advantageous for advancing glycomics.

In data-independent acquisition, the prediction of peptide fragmentation in tandem mass spectrometry is a major focus for deep learning approaches, with great progress for predicting peptide fragmentation,[133] while glycopeptide fragmentation prediction still may require further advances until spectral libraries are no longer necessary.[134]

Generating glycosylation information without mass spectrometry would represent a paradigm shift in glycobiology. Thus, approaches that build on the knowledge base constructed so far and aim at this goal are not only a worthwhile focus but also dependent on advanced algorithms such as from deep learning. A recent neural network based on a metabolic model predicting nucleotide sugar donor concentrations as inputs, demonstrated that the proportions of a limited set of *N*-linked glycans could be predicted on four recombinant glycoproteins from three CHO cell lines[135] (see Figure 5C). Extension and generalization of such a model could hold promise to advance glycoengineering efforts in the production of biopharmaceuticals.[136]

5.5. Using Deep Learning to Predict Glycan Properties

While the above section discussed advances in obtaining glycan sequences, we now turn to the question of what to eventually do with obtained glycan data to gain further insight into biological contexts. Analogous to deep learning models predicting protein properties from sequences, such as GO terms or EC numbers,[137] recent developments in glycobiology have introduced a new generation of deep learning-based sequence-to-function models (see Figure 5D).

Current examples of predicting glycan properties from sequences include prediction of glycan class, taxonomy, immunogenic potential, and association with bacterial pathogenicity.[119,122] Trained models in these tasks not only can be used to infer the properties of newly discovered glycans but also can be used to retrieve motifs that are important to endow a glycan with a property, such as human-like glycan motifs used in molecular mimicry by pathogenic *Escherichia coli* strains. Other notable efforts in this area, using machine learning, are the investigation of the association of clinical characteristics with glycans from cancer patients[138] or the prediction of α-fetoprotein (AFP)-negative hepatocellular carcinoma using glycan fingerprints,[139] both of which could offer a promising target for deep learning in the future. The key limiting factor in all applications of this type of supervised learning in glycobiology is the lack of labeled data of glycan sequences with known information about their properties or functions that could be used to train a model. Often this information may exist in sufficient quantity across the literature yet is scattered and would require exhaustive manual curation that may be prohibitively expensive.

One domain which has made considerable progress in solving this bottleneck via organized databases and resources, as mentioned above, is the study of lectin−glycan interactions. Here, resources such as the CFG array database or the UniLectin3D database for lectin−glycan crystal structures offer data that are more suitable for machine learning and related endeavors. Therefore, several machine and deep learning approaches have recently joined the statistical motif-counting methods to analyze glycan-binding data. An example is a recent study[140] that used the structural data of UniLectin3D to train random forest-based models on computed physicochemical and geometric features of proteins to predict their binding to observed glycan fragments.

Other approaches have used glycan-binding data from glycan arrays, where lectins are probed for their binding to immobilized glycans on a glass slide.[141] One example for this would be GlyNet,[142] a neural network-based approach for using the motifs that occur in a glycan to predict its binding to the lectins that have historically been assayed on the CFG platform. Combining interpretable, rule-based machine learning with expert annotation has also recently resulted in the detailed elucidation of the binding motifs of a large set of commonly used lectins.[143] In other work, glyBERT has been introduced as a transformer-based model for glycans that can also be used to learn the binding specificity of a lectin, provided that binding data exist. All the approaches above require dedicated binding data of the lectin that is being studied. Another recent approach to solve this limitation and the problem of predicting the binding specificity of a lectin has been put forward with LectinOracle,[144] a deep learning model that combines a language model for the lectin sequence (the ESM-1b transformer model[145]) and a graph neural network for the glycan sequence (SweetNet) to predict lectin-glycan binding. By also considering the information of proteins, LectinOracle can generalize to new proteins as well as new glycans. For the relatively data-sparse field of glycobiology, such a strategy is crucial to at least provide predictions that can in turn generate new hypotheses for contexts with many uncharacterized lectins, such as the microbiome.

6. GRADUAL IMPACT ON GLYCOSCIENCE AND DEVELOPMENT PROSPECTS

At this stage of method development in glycoscience, glycoinformatics provides a real chance for unifying a view of the molecular interactions mediated by glycans. From measurement (e.g., mass spectrometry fragmentation prediction) to biological context (e.g., glycosite prediction) and glycan properties as well as functions, glycoinformatics is advancing every facet of glycoscience and has the potential to continue doing so in the future.

6.1. Expected Evolution of AI in Glycoinformatics

6.1.1. Evolution of Data. At this point in time, a simple explanation for glycomics lagging behind other -omics lies in the absence of high-throughput sequencing of glycans. Consequently, data accumulates substantially slower than in genomics or transcriptomics. This is especially true for glycan classes outside of N- and O-linked glycans. The contrast is even more striking as bioinformatics is now geared to process petabytes of nucleotide sequences and run smart searches to reveal hidden information. Such massive sequence data crunching has already led to the identification of ∼10^5 unknown viral species.[146] In that sense, the future of glycomics hinges on new technological development that may enable glycan high-throughput sequencing and also improve the analysis of other types of glycans such as glycosaminoglycans. Efforts in other fields, such as the recent advances toward protein sequencing, demonstrate that sequencing in principle can also be applied to non-nucleic acid biopolymers.[147]

Another key point regarding data collection is the current limited availability of quantitative data that would allow more accurate profiling. At present, immunoglobulin profiling is by far the most advanced in comparison to other glycoconjugates.[148,149] Other prospects can be expected to expand from improved techniques in glycan and tissue imaging.[150,151]

6.1.2. Improved Prediction. Like in many scientific fields, AI methods are increasingly implemented to improve classification and prediction. Machine learning applied in various aspects of glycoscience (e.g., glycosite prediction or monosaccharide donor prediction for glycosyltransferases) still predominantly rely on human-devised, calculated features as model input. This is presumably the reason why classical machine learning methods (SVM, RF, etc.) currently often still outperform deep learning approaches on these tasks. One of the most important advantages of deep learning is that it allows for access to information beyond the rationally chosen features of a sample. It is therefore to be expected that deep learning approaches using raw sequences in a proper format will yield improved performance in the future. Another promising direction is the combination of calculated features and raw inputs, such as sequences, that has been shown to improve performance for small molecule property prediction.[152]

Additionally, while existing models are largely inclusive of less well-studied glycan classes, such as plant and fungal polysaccharides, in terms of their model architecture, there might be approaches that could perform better on tasks involving these polymers, for instance by considering their repeat structure. However, existing data for most prediction tasks described in this manuscript are largely restricted to N- and O- linked glycans as well as glycolipids and, in a limited manner, glycosaminoglycans. Therefore, both available data and existing models will likely need to be improved to fully leverage the information in polymeric glycans.

On the other end, the purpose of predicting glycan-binding is to design specific ligands, for instance to inhibit glycan-binding proteins of pathogens, but more context-sensitive information will be required to qualify specificity. In particular, realistic binding prediction is likely to depend on additional characteristics, such as expression of the lectin and physiological conditions. Ultimately, models will need to account for all of these aspects, as well as the structural features of glycoconjugates and glycan-binding proteins. These will be helped if 3D models are more systemically built while taking glycans into consideration, as made possible with AlphaFold2 predictions.[114]

6.1.3. Improved Representation. The learned numerical representation by ML models can also be used to find the most similar known data point, given a new unknown data point. In the context of tandem mass spectrometry in proteomics, this has been used to quickly assign unidentified spectra to peptides.[153] A similar procedure in glycomics or glycoproteomics could advance these fields as well. Next to similarity, the learned representation gained by an unsupervised model can also be viewed as learned features of, say, a protein sequence, which can be used by another downstream model. An example for this can be found in the case of evolutionary scale modeling 1b (ESM-1b), a transformer-based language model trained on protein sequences.[145] The learned representations from ESM-1b are multipurpose and can be used to predict various properties such as protein structure, stability, or function, with downstream models that do not need as many parameters as one would need if the model would be trained on protein sequences from scratch. While these learned protein features can be very useful in the glycosciences (e.g., glycosite prediction, glycan-binding prediction, etc.), eventually an analogous model for glycans might be needed to improve prediction tasks. This is especially relevant as the number of available labels for glycans, that could be used for training a model, is typically much lower than is the case for proteins, necessitating the usage of such a pretrained model for satisfactory performance.

Current models all focus on glycan sequences, processed in various ways. However, future models might have to include chemical as well as 3D information to achieve optimal results.[154] In the case of proteins, joint representations of sequence and structure have shown improved performance for downstream models.[155] There are already numerous indications that glycan conformation influences function and including this information into predictive models is bound to place them closer to the biological reality. Key challenges here are (i) how best to incorporate this information into existing or future glycan-focused artificial intelligence models given the conformational flexibility of glycans and (ii) how to obtain a sufficient number of glycan conformations. As reviewed earlier, the answers lie in the constant improvement of experimental techniques (for example, intact glycoprotein resolution can benefit from cryoelectron microscopy) as well as that of molecular dynamics (MD) simulation algorithms.[156] The latter critically depend on the further increase of computer power and on the refinement of molecular aspects such as glycan placement on proteins.

6.2. Bridging Glycoinformatics and Bioinformatics

The importance of applying FAIR principles, promoted in bioinformatics and emphasized in the introduction as well as in section 3, has beneficial effects on glycoinformatics development that increasingly bridges with resources of other -omics and provides the means to expand systematism.

Single cell technologies have spread like wildfire in most -omics applications, providing each discipline with more specific and refined information on molecular activity and interactions. Glycomics has not yet benefited from such a step forward. It remains open to debate whether information is easier to obtain from genes and biosynthetic pathways that regulate glycosylation than from the direct analysis of glycan structures. At present, glycoengineering tends to be more advanced when handling genes,[157] but this does not rule out a yet-to-come strictly single cell glycomics approach. First steps in this direction have added partial/fragmentary glycan information to the analysis of single cells and/or a combination with their transcriptome.[158,159] Data integration will be facilitated by collecting same level information from different and complementing -omics. Nonetheless, attempts at combining views on regulation have already emerged, for example, by considering the interplay between microRNA and glycan expression[160] or between gene transcription and glycan expression,[161] and these are likely to expand in the very near future. In all cases, both bioinformatics and glycoinformatics resources are needed in this context. This bridging process is the core of an international cooperative approach named GlySpace,[23] designed to facilitate structural and functional glycomic data sharing and information exchange and committing to provide high quality, reliable, well-referenced, and accurate data to the benefit of users. In parallel, the release of software libraries tailored for glycoinformatics applications consolidates this initiative. Several examples spanning the management of mass spectrometry data in Python[162] or Java[163] or handling ML tools[164] are readily available for software developers.

6.3. Multiscale View

As mentioned in the introduction, the variety and disparity of sources of information that are needed to understand the details of glycan structure and function are still a hindrance to rapid progress in glycobiology. Ultimately, the goal of glycoinformatics is to restore a more thorough picture from pieces created artificially by technological constraints. For as long as this puzzle is, if not complete, at least advanced enough to allow for reliable predictions, it will concentrate efforts within glycoscience. However, the key contribution of glycans to biological processes, especially in cell−cell communication, cannot be ignored and, as mentioned above, glycomics should be combined with other omics. In fact, the ideal view of understanding living organisms is dynamic and starts from the atomic to the cellular, tissue, and organ levels. Multiscale modeling as the holy grail of systems biology must be fed with a refined knowledge of glycoscience.

AUTHOR INFORMATION

Corresponding Authors

Frederique Lisacek − *Proteome Informatics Group, Swiss Institute of Bioinformatics, CH-1227 Geneva, Switzerland; Computer Science Department & Section of Biology, University of Geneva, CH-1227 Geneva, Switzerland;* orcid.org/0000-0002-0948-4537; Email: frederique.lisacek@sib.swiss

Daniel Bojar − *Department of Chemistry and Molecular Biology and Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg 41390, Sweden*; Email: daniel.bojar@gu.se

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrev.2c00110

Notes

The authors declare no competing financial interest.

Biographies

Daniel Bojar is a tenure-track assistant professor in bioinformatics at the Department of Chemistry and Molecular Biology and the Wallenberg Centre for Molecular and Translational Medicine at the University in Gothenburg. After a Ph.D. in mammalian synthetic biology at ETH Zurich and a postdoctoral stay at the Wyss Institute for Biologically Inspired Engineering at Harvard University, his group now focuses on research at the intersection of glycobiology and machine learning. Dr. Bojar has received a Branco Weiss Fellowship—Society in Science, a Foresight Fellowship, and is on the 2022 Forbes 30 Under 30 Europe list.

Frederique Lisacek received a Ph.D. in Computer Science (Artificial Intelligence) from the University Pierre & Marie Curie, Paris, France. Then, she has held research positions in bioinformatics in France, Japan, and Australia, working on knowledge representation and sequence analysis in various fields of molecular biology. She has been involved in early proteomics projects within two companies Proteome Systems Ltd in Sydney, Australia, and Geneva Bioinformatics (GeneBio) SA. In 2006, she joined the Swiss Institute of Bioinformatics (SIB) in the Proteome Informatics Group that she has managed since 2008 and where she initiated several projects on the study of protein posttranslational modifications, with a strong focus on glycosylation. She holds a lecturer position at the University of Geneva.

ACKNOWLEDGMENTS

D.B. acknowledges funding from the Branco Weiss Fellowship − Society in Science, the Knut and Alice Wallenberg Foundation, and the University of Gothenburg, Sweden. F.L. acknowledges funding from the Swiss National Science Foundation (SNSF) grant 443 #31003A/179249, Glyco@Alps (ANR-15-IDEX-02) and the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI).

ABBREVIATIONS/ACRONYMS

AI = artificial intelligence
BERT = bidirectional encoder representations from transformers
CASP = critical assessment of protein structure prediction
CAZy = carbohydrate-active enzymes
CID = collision-induced dissociation
CFG = Consortium for Functional Glycomics
ChEBI = chemical entities of biological interest
DL = deep learning
EC = enzyme commission number
ESM-1b = evolutionary scale modeling 1b
FAIR = findable, accessible, interoperable, and reusable
FDR = false discovery rate
Fuc = fucose
Gal = galactose
GalNAc = *N*-acetylgalactosamine
Glc = glucose
GlcNAc = *N*-acetylglucosamine
GO = gene ontology
HMM = hidden Markov model
HTTP = hypertext transfer protocol
InChI = IUPAC international chemical identifier
IUPAC = International Union of Pure and Applied Chemistry
MIRAGE = minimum information required about a glycomics experiment
Man = mannose
MD = molecular dynamics
ML = machine learning
MS = mass spectrometry
NMR = nuclear magnetic resonance spectroscopy
NN = neural network
PDB = Protein Data Bank
RF = random forest
SMILES = simplified molecular input line entry system
SNFG = symbol nomenclature for glycans
SVM = support vector machine
t-SNE = t-distributed stochastic neighbor embedding
UMAP = uniform manifold approximation and projection
WURCS = Web3 unique representation of carbohydrate structures

REFERENCES

(1) Aoki-Kinoshita, K. F.; Lisacek, F.; Karlsson, N.; Kolarich, D.; Packer, N. H. GlycoBioinformatics. *Beilstein J. Org. Chem.* **2021**, *17*, 2726−2728.

(2) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018.

(3) Carbon, S.; Douglass, E.; Good, B. M.; Unni, D. R.; Harris, N. L.; Mungall, C. J.; Basu, S.; Chisholm, R. L.; Dodson, R. J.; Hartline, E.; et al. The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Res.* **2021**, *49*, D325−D334.

(4) Qian, N.; Sejnowski, T. J. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.* **1988**, *202* (4), 865−884.

(5) Fariselli, P.; Olmea, O.; Valencia, A.; Casadio, R. Prediction of Contact Maps with Neural Networks and Correlated Mutations. *Protein Eng. Des. Sel.* **2001**, *14* (11), 835−843.

(6) Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLOS Comput. Biol.* **2017**, *13* (1), No. e1005324.

(7) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.

(8) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivetova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; et al. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.

(9) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold - Making Protein Folding Accessible to All; preprint. *Nature Methods* **2021**, *19*, 679.

(10) Doubet, S.; Bock, K.; Smith, D.; Darvill, A.; Albersheim, P. The Complex Carbohydrate Structure Database. *Trends Biochem. Sci.* **1989**, *14* (12), 475−477.

(11) Pérez, S.; Mulloy, B. Prospects for Glycoinformatics. *Curr. Opin. Struct. Biol.* **2005**, *15* (5), 517−524.

(12) Lütteke, T. The Use of Glycoinformatics in Glycochemistry. *Beilstein J. Org. Chem.* **2012**, *8*, 915−929.

(13) Li, F.; Glinskii, O. V.; Glinsky, V. V. Glycobioinformatics: Current Strategies and Tools for Data Mining in MS-Based Glycoproteomics. *PROTEOMICS* **2013**, *13* (2), 341−354.

(14) Egorova, K. S.; Toukach, P. V. Glycoinformatics: Bridging Isolated Islands in the Sea of Data. *Angew. Chem., Int. Ed.* **2018**, *57* (46), 14986−14990.

(15) Campbell, M. P.; Aoki-Kinoshita, K. F.; Lisacek, F.; York, W. S.; Packer, N. H. Glycoinformatics. In *Essentials of Glycobiology*; Varki, A., Cummings, R. D., Esko, J. D., Stanley, P., Hart, G. W., Aebi, M., Darvill, A. G., Kinoshita, T., Packer, N. H., Prestegard, J. H., et al., Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2015.

(16) Aoki-Kinoshita, K.; Agravat, S.; Aoki, N. P.; Arpinar, S.; Cummings, R. D.; Fujita, A.; Fujita, N.; Hart, G. M.; Haslam, S. M.; Kawasaki, T.; et al. GlyTouCan 1.0 − The International Glycan Structure Repository. *Nucleic Acids Res.* **2016**, *44* (D1), D1237−D1242.

(17) 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. GlyTouCan: An Accessible Glycan Structure Repository. *Glycobiology* **2017**, *27* (10), 915−919.

(18) Fujita, A.; Aoki, N. P.; Shinmachi, D.; Matsubara, M.; Tsuchiya, S.; Shiota, M.; Ono, T.; Yamada, I.; Aoki-Kinoshita, K. F. The International Glycan Repository GlyTouCan Version 3.0. *Nucleic Acids Res.* **2021**, *8*, D1529−D1533.

(19) Sumer-Bayraktar, Z.; Nguyen-Khuong, T.; Jayo, R.; Chen, D. D. Y.; Ali, S.; Packer, N. H.; Thaysen-Andersen, M. Micro- and Macroheterogeneity of *N* -Glycosylation Yields Size and Charge Isoforms of Human Sex Hormone Binding Globulin Circulating in Serum. *PROTEOMICS* **2012**, *12* (22), 3315−3327.

(20) Medzihradszky, K. F.; Kaasik, K.; Chalkley, R. J. Tissue-Specific Glycosylation at the Glycopeptide Level. *Mol. Cell. Proteomics* **2015**, *14* (8), 2103−2110.

(21) Carvalho-cruz, P.; Alisson-Silva, F.; Todeschini, A. R.; Dias, W. B. Cellular Glycosylation Senses Metabolic Changes and Modulates Cell Plasticity during Epithelial to Mesenchymal Transition: Cellular Glycosylation and Changes During EMT. *Dev. Dyn.* **2018**, *247* (3), 481−491.

(22) Oliveira, T.; Thaysen-Andersen, M.; Packer, N. H.; Kolarich, D. The Hitchhiker's Guide to Glycoproteomics. *Biochem. Soc. Trans.* **2021**, *49* (4), 1643−1662.

(23) Lisacek, F.; Aoki-Kinoshita, K. F.; Vora, J. K.; Mazumder, R.; Tiemeyer, M. Glycoinformatics Resources Integrated Through the GlySpace Alliance. In *Comprehensive Glycoscience*; Elsevier, 2021; pp 507−521.

(24) Losfeld, M.-E.; Scibona, E.; Lin, C.-W.; Villiger, T. K.; Gauss, R.; Morbidelli, M.; Aebi, M. Influence of Protein/Glycan Interaction on Site-Specific Glycan Heterogeneity. *FASEB J.* **2017**, *31*, 4623−4635.

(25) Kellman, B. P.; Lewis, N. E. Big-Data Glycomics: Tools to Connect Glycan Biosynthesis to Extracellular Communication. *Trends Biochem. Sci.* **2021**, *46*, 284−300.

(26) Helm, J.; Grünwald-Gruber, C.; Thader, A.; Urteil, J.; Führer, J.; Stenitzer, D.; Maresch, D.; Neumann, L.; Pabst, M.; Altmann, F. Bisecting Lewis X in Hybrid-Type *N* -Glycans of Human Brain Revealed by Deep Structural Glycomics. *Anal. Chem.* **2021**, *93* (45), 15175−15182.

(27) Flynn, R. A.; Pedram, K.; Malaker, S. A.; Batista, P. J.; Smith, B. A. H.; Johnson, A. G.; George, B. M.; Majzoub, K.; Villalta, P. W.; Carette, J. E.; Bertozzi, C. R.; et al. Small RNAs Are Modified with N-Glycans and Displayed on the Surface of Living Cells. *Cell* **2021**, *184*, 3109−3124.

(28) Gray, C. J.; Migas, L. G.; Barran, P. E.; Pagel, K.; Seeberger, P. H.; Eyers, C. E.; Boons, G.-J.; Pohl, N. L. B.; Compagnon, I.; Widmalm, G.; et al. Advancing Solutions to the Carbohydrate Sequencing Challenge. *J. Am. Chem. Soc.* **2019**, *141* (37), 14463−14479.

(29) Grabarics, M.; Lettow, M.; Kirschbaum, C.; Greis, K.; Manz, C.; Pagel, K. Mass Spectrometry-Based Techniques to Elucidate the Sugar Code. *Chem. Rev.* **2022**, *122* (8), 7840−7908.

(30) Morgat, A.; Axelsen, K. B.; Lombardot, T.; Alcantara, R.; Aimo, L.; Zerara, M.; Niknejad, A.; Belda, E.; Hyka-Nouspikel, N.; Coudert, E.; et al. Updates in Rhea–a Manually Curated Resource of Biochemical Reactions. *Nucleic Acids Res.* **2015**, *43* (D1), D459−D464.

(31) Morgat, A.; Lombardot, T.; Coudert, E.; Axelsen, K.; Neto, T. B.; Gehant, S.; Bansal, P.; Bolleman, J.; Gasteiger, E.; de Castro, E.; et al. Enzyme Annotation in UniProtKB Using Rhea. *Bioinformatics* **2019**, btz817.

(32) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; et al. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2019**, *47* (D1), D1102−D1109.

(33) Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. *Nucleic Acids Res.* **2016**, *44* (D1), D1214−D1219.

(34) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97−101.

(35) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminformatics* **2015**, *7*, 23.

(36) Sharon, N. IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN). Nomenclature of Glycoproteins, Glycopeptides and Peptidoglycans: JCBN Recommendations 1985. *Glycoconj. J.* **1986**, *3* (2), 123−133.

(37) Herget, S.; Ranzinger, R.; Maass, K.; Lieth, C.-W. v. d. GlycoCT—a Unifying Sequence Format for Carbohydrates. *Carbohydr. Res.* **2008**, *343* (12), 2162−2171.

(38) Tanaka, K.; Aoki-Kinoshita, K. F.; Kotera, M.; Sawaki, H.; Tsuchiya, S.; Fujita, N.; Shikanai, T.; Kato, M.; Kawano, S.; Yamada, I.; et al. WURCS: The Web3 Unique Representation of Carbohydrate Structures. *J. Chem. Inf. Model.* **2014**, *54* (6), 1558−1566.

(39) Varki, A.; Cummings, R. D.; Aebi, M.; Packer, N. H.; Seeberger, P. H.; Esko, J. D.; Stanley, P.; Hart, G.; Darvill, A.; Kinoshita, T.; et al. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **2015**, *25* (12), 1323−1324.

(40) Neelamegham, S.; Aoki-Kinoshita, K.; Bolton, E.; Frank, M.; Lisacek, F.; Lütteke, T.; O'Boyle, N.; Packer, N. H.; Stanley, P.; Toukach, P.; et al. Updates to the Symbol Nomenclature for Glycans Guidelines. *Glycobiology* **2019**, *29* (9), 620−624.

(41) Daponte, V.; Hayes, C.; Mariethoz, J.; Lisacek, F. Dealing with the Ambiguity of Glycan Substructure Search. *Molecules* **2022**, *27*, 65.

(42) Tsuchiya, S.; Yamada, I.; Aoki-Kinoshita, K. F. GlycanFormatConverter: a conversion tool for translating the complexities of glycans. *Bioinformatics* **2019**, *35* (14), 2434−2440.

(43) Lal, K.; Bermeo, R.; Perez, S. Computational Tools for Drawing, Building and Displaying Carbohydrates: A Visual Guide. *Beilstein J. Org. Chem.* **2020**, *16*, 2448−2468.

(44) Drula, E.; Garron, M.-L.; Dogan, S.; Lombard, V.; Henrissat, B.; Terrapon, N. The Carbohydrate-Active Enzyme Database: Functions and Literature. *Nucleic Acids Res.* **2022**, *50* (D1), D571−D577.

(45) Sayers, E. W.; Bolton, E. E.; Brister, J. R.; Canese, K.; Chan, J.; Comeau, D. C.; Connor, R.; Funk, K.; Kelly, C.; Kim, S.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2022**, *50* (D1), D20−D26.

(46) Bairoch, A.; Apweiler, R. The SWISS-PROT Protein Sequence Data Bank and Its Supplement TrEMBL in 1999. *Nucleic Acids Res.* **1999**, *27* (1), 49−54.

(47) York, W. S.; Mazumder, R.; Ranzinger, R.; Edwards, N.; Kahsay, R.; Aoki-Kinoshita, K. F.; Campbell, M. P.; Cummings, R. D.; Feizi, T.; Martin, M.; et al. GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* **2020**, *30* (2), 72−73.

(48) Behzadi, P.; Gajdács, M. Worldwide Protein Data Bank (wwPDB): A Virtual Treasure for Research in Biotechnology. *Eur. J. Microbiol. Immunol.* **2022**, *11*, 77−86.

(49) Sehnal, D.; Grant, O. C. Rapidly Display Glycan Symbols in 3D Structures: 3D-SNFG in LiteMol. *J. Proteome Res.* **2019**, *18* (2), 770−774.

(50) Imberty, A.; Bonnardel, F.; Lisacek, F. UniLectin, A One-Stop-Shop to Explore and Study Carbohydrate-Binding Proteins. *Curr. Protoc.* **2021**, *1*, 305.

(51) Alocci, D.; Mariethoz, J.; Gastaldello, A.; Gasteiger, E.; Karlsson, N. G.; Kolarich, D.; Packer, N. H.; Lisacek, F. GlyConnect: Glycoproteomics Goes Visual, Interactive, and Analytical. *J. Proteome Res.* **2019**, *18* (2), 664−677.

(52) Raman, R.; Venkataraman, M.; Ramakrishnan, S.; Lang, W.; Raguram, S.; Sasisekharan, R. Advancing Glycomics: Implementation Strategies at the Consortium for Functional Glycomics. *Glycobiology* **2006**, *16* (5), 82R−90R.

(53) Cao, Y.; Park, S.-J.; Mehta, A. Y.; Cummings, R. D.; Im, W. GlyMDB: Glycan Microarray Database and Analysis Toolset. *Bioinformatics* **2020**, *36* (8), 2438−2442.

(54) Akune, Y.; Arpinar, S.; Silva, L. M; Palma, A. S; Tajadura-Ortega, V.; Aoki-Kinoshita, K. F; Ranzinger, R.; Liu, Y.; Feizi, T. CarbArrayART: a new software tool for carbohydrate microarray data storage, processing, presentation, and reporting. *Glycobiology* **2022**, *32* (7), 552−555.

(55) York, W. S.; Agravat, S.; Aoki-Kinoshita, K. F.; McBride, R.; Campbell, M. P.; Costello, C. E.; Dell, A.; Feizi, T.; Haslam, S. M.; Karlsson, N.; et al. MIRAGE: The Minimum Information Required for a Glycomics Experiment. *Glycobiology* **2014**, *24* (5), 402−406.

(56) Rojas-Macias, M. A.; Mariethoz, J.; Andersson, P.; Jin, C.; Venkatakrishnan, V.; Aoki, N. P.; Shinmachi, D.; Ashwood, C.; Madunic, K.; Zhang, T.; et al. Towards a Standardized Bioinformatics Infrastructure for N- and O-Glycomics. *Nat. Commun.* **2019**, *10*, 3275.

(57) Watanabe, Y.; Aoki-Kinoshita, K. F.; Ishihama, Y.; Okuda, S. GlycoPOST Realizes FAIR Principles for Glycomics Mass Spectrometry Data. *Nucleic Acids Res.* **2021**, *49*, D1523−D1528.

(58) Liu, Y.; McBride, R.; Stoll, M.; Palma, A. S.; Silva, L.; Agravat, S.; Aoki-Kinoshita, K. F.; Campbell, M. P.; Costello, C. E.; Dell, A.; et al. The Minimum Information Required for a Glycomics Experiment (MIRAGE) Project: Improving the Standards for Reporting Glycan Microarray-Based Data. *Glycobiology* **2017**, *27*, 280−284.

(59) Traini, M.; Gooley, A. A.; Ou, K.; Wilkins, M. R.; Tonella, L.; Sanchez, J.-C.; Hochstrasser, D. F.; Williams, K. L. Towards an Automated Approach for Protein Identification in Proteome Projects. *Electrophoresis* **1998**, *19* (11), 1941−1949.

(60) Levander, F.; Rögnvaldsson, T.; Samuelsson, J.; James, P. Automated Methods for Improved Protein Identification by Peptide Mass Fingerprinting. *Proteomics* **2004**, *4* (9), 2594−2601.

(61) Lisacek, F.; Cohen-Boulakia, S.; Appel, R. D. Proteome Informatics II: Bioinformatics for Comparative Proteomics. *Proteomics* **2006**, *6* (20), 5445−5466.

(62) Bensimon, A.; Heck, A. J. R.; Aebersold, R. Mass Spectrometry−Based Proteomics and Network Biology. *Annu. Rev. Biochem.* **2012**, *81* (1), 379−405.

(63) Robin, T.; Mariethoz, J.; Lisacek, F. Examining and Fine-Tuning the Selection of Glycan Compositions with GlyConnect Compozitor. *Mol. Cell. Proteomics* **2020**, *19* (10), 1602−1618.

(64) Bellman, R. On the Theory of Dynamic Programming. *Proc. Natl. Acad. Sci. U. S. A.* **1952**, *38* (8), 716−719.

(65) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48* (3), 443−453.

(66) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323* (6088), 533−536.

(67) Lukashin, A. V.; Anshelevich, V. V.; Amirikyan, B. R.; Gragerov, A. I.; Frank-Kamenetskii, M. D. Neural Network Models for Promoter Recognition. *J. Biomol. Struct. Dyn.* **1989**, *6* (6), 1123−1133.

(68) Holley, L. H.; Karplus, M. Protein Secondary Structure Prediction with a Neural Network. *Proc. Natl. Acad. Sci. U. S. A.* **1989**, *86* (1), 152−156.

(69) Baldi, P.; Brunak, S. *Bioinformatics: The Machine Learning Approach; Adaptive computation and machine learning*; MIT Press: Cambridge, MA, 1998.

(70) Krallinger, M.; Rabal, O.; Lourenço, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **2017**, *117* (12), 7673−7761.

(71) Whalen, S.; Schreiber, J.; Noble, W. S.; Pollard, K. S. Navigating the Pitfalls of Applying Machine Learning in Genomics. *Nat. Rev. Genet.* **2022**, *23*, 169−181.

(72) Nielsen, H.; Engelbrecht, J.; Brunak, S.; von Heijne, G. Identification of Prokaryotic and Eukaryotic Signal Peptides and Prediction of Their Cleavage Sites. *Protein Eng. Des. Sel.* **1997**, *10* (1), 1−6.

(73) Nielsen, H.; Tsirigos, K. D.; Brunak, S.; von Heijne, G. A Brief History of Protein Sorting Prediction. *Protein J.* **2019**, *38* (3), 200−216.

(74) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks. *Nat. Biotechnol.* **2019**, *37* (4), 420−423.

(75) Teufel, F.; Almagro Armenteros, J. J.; Johansen, A. R.; Gíslason, M. H.; Pihl, S. I.; Tsirigos, K. D.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models. *Nat. Biotechnol.* **2022**. DOI: 10.1038/s41587-021-01156-3

(76) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explor. Newsl.* **2009**, *11* (1), 10−18.

(77) Curtin, R. R.; Edel, M.; Lozhnikov, M.; Mentekidis, Y.; Ghaisas, S.; Zhang, S. Mlpack 3: A Fast, Flexible Machine Learning Library. *J. Open Source Softw.* **2018**, *3*, 726.

(78) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* **2020**, arXiv:1802.03426.

(79) van der Maaten, L.; Hinton, G. *J. Machine Learning Res.* **2008**, *9*, 2579−2605.

(80) Kobak, D.; Berens, P. The Art of Using T-SNE for Single-Cell Transcriptomics. *Nat. Commun.* **2019**, *10*, 5416.

(81) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat. Biotechnol.* **2019**, *37* (1), 38−44.

(82) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-Based Protein Identification by Machine Learning from a Library of Tandem Mass Spectra. *Nat. Biotechnol.* **2004**, *22* (2), 214−219.

(83) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-Supervised Learning for Peptide Identification from Shotgun Proteomics Datasets. *Nat. Methods* **2007**, *4* (11), 923−925.

(84) Spivak, M.; Weston, J.; Tomazela, D.; MacCoss, M. J.; Noble, W. S. Direct Maximization of Protein Identifications from Tandem Mass Spectra. *Mol. Cell. Proteomics* **2012**, *11*, M111.012161.

(85) Kawahara, R.; Alagesan, K.; Bern, M.; Cao, W.; Chalkley, R. J.; Cheng, K.; Choo, M. S.; Edwards, N.; Goldman, R.; Hoffmann, M. Community Evaluation of Glycoproteomics Informatics Solutions Reveals High-Performance Search Strategies of Glycopeptide Data. *bioRXiv* **2021**, 2021.03.14.435332.

(86) Kumozaki, S.; Sato, K.; Sakakibara, Y. A Machine Learning Based Approach to de Novo Sequencing of Glycans from Tandem Mass Spectrometry Spectrum. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12* (6), 1267−1274.

(87) Bause, E. Structural Requirements of *N* -Glycosylation of Proteins. Studies with Proline Peptides as Conformational Probes. *Biochem. J.* **1983**, *209* (2), 331−336.

(88) Nehrke, K.; Ten Hagen, K. G.; Hagen, F. K.; Tabak, L. A. Charge Distribution of Flanking Amino Acids Inhibits O-Glycosylation of Several Single-Site Acceptors *in Vivo*. *Glycobiology* **1997**, *7* (8), 1053−1060.

(89) Zhang, H.; Loriaux, P.; Eng, J.; Campbell, D.; Keller, A.; Moss, P.; Bonneau, R.; Zhang, N.; Zhou, Y.; Wollscheid, B.; et al. UniPep−a Database for Human N-Linked Glycosites: A Resource for Biomarker Discovery. *Genome Biol.* **2006**, *7* (8), R73.

(90) Sun, S.; Hu, Y.; Ao, M.; Shah, P.; Chen, J.; Yang, W.; Jia, X.; Tian, Y.; Thomas, S.; Zhang, H. N-GlycositeAtlas: A Database Resource for Mass Spectrometry-Based Human N-Linked Glycoprotein and Glycosylation Site Mapping. *Clin. Proteomics* **2019**, *16*, 35.

(91) Steentoft, C.; Vakhrushev, S. Y.; Joshi, H. J.; Kong, Y.; Vester-Christensen, M. B.; Schjoldager, K. T.-B. G.; Lavrsen, K.; Dabelsteen, S.; Pedersen, N. B.; Marcos-Silva, L.; et al. Precision Mapping of the Human O-GalNAc Glycoproteome through SimpleCell Technology. *EMBO J.* **2013**, *32* (10), 1478−1488.

(92) Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence. *Proteomics* **2004**, *4* (6), 1633−1649.

(93) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273−297.

(94) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.

(95) Li, F.; Li, C.; Wang, M.; Webb, G. I.; Zhang, Y.; Whisstock, J. C.; Song, J. GlycoMine: A Machine Learning-Based Approach for Predicting N-, C- and O-Linked Glycosylation in the Human Proteome. *Bioinformatics* **2015**, *31* (9), 1411−1419.

(96) Hamby, S. E.; Hirst, J. D. Prediction of Glycosylation Sites Using Random Forests. *BMC Bioinformatics* **2008**, *9*, 500.

(97) Chauhan, J. S.; Rao, A.; Raghava, G. P. S. In Silico Platform for Prediction of N-, O- and C-Glycosites in Eukaryotic Protein Sequences. *PLoS One* **2013**, *8* (6), No. e67008.

(98) Caragea, C.; Sinapov, J.; Silvescu, A.; Dobbs, D.; Honavar, V. Glycosylation Site Prediction Using Ensembles of Support Vector Machine Classifiers. *BMC Bioinformatics* **2007**, *8*, 438.

(99) Pitti, T.; Chen, C.-T.; Lin, H.-N.; Choong, W.-K.; Hsu, W.-L.; Sung, T.-Y. N-GlyDE: A Two-Stage N-Linked Glycosylation Site

Prediction Incorporating Gapped Dipeptides and Pattern-Based Encoding. *Sci. Rep.* **2019**, *9*, 15975.

(100) Smolarczyk, T.; Roterman-Konieczna, I.; Stapor, K. Protein Secondary Structure Prediction: A Review of Progress and Directions. *Curr. Bioinforma.* **2020**, *15* (2), 90−107.

(101) Wang, J.; Torii, M.; Liu, H.; Hart, G. W.; Hu, Z.-Z. DbOGAP - an Integrated Bioinformatics Resource for Protein O-GlcNAcylation. *BMC Bioinformatics* **2011**, *12*, 91.

(102) Jia, C.; Zuo, Y.; Zou, Q. O-GlcNAcPRED-II: An Integrated Classification Algorithm for Identifying O-GlcNAcylation Sites Based on Fuzzy Undersampling and a K-Means PCA Oversampling Technique. *Bioinformatics* **2018**, *34* (12), 2029−2036.

(103) Thaysen-Andersen, M.; Packer, N. H. Site-Specific Glycoproteomics Confirms That Protein Structure Dictates Formation of N-Glycan Type, Core Fucosylation and Branching. *Glycobiology* **2012**, *22* (11), 1440−1452.

(104) Gastaldello, A.; Alocci, D.; Baeriswyl, J.-L.; Mariethoz, J.; Lisacek, F. GlycoSiteAlign: Glycosite Alignment Based on Glycan Structure. *J. Proteome Res.* **2016**, *15* (10), 3916−3928.

(105) Wu, S.-W.; Liang, S.-Y.; Pu, T.-H.; Chang, F.-Y.; Khoo, K.-H. Sweet-Heart — An Integrated Suite of Enabling Computational Tools for Automated MS2/MS3 Sequencing and Identification of Glycopeptides. *J. Proteomics* **2013**, *84*, 1−16.

(106) Liang, S.-Y.; Wu, S.-W.; Pu, T.-H.; Chang, F.-Y.; Khoo, K.-H. An Adaptive Workflow Coupled with Random Forest Algorithm to Identify Intact N-Glycopeptides Detected from Mass Spectrometry. *Bioinformatics* **2014**, *30* (13), 1908−1916.

(107) Mamitsuka, H. Glycoinformatics: Data Mining-Based Approaches. *Chim. Int. J. Chem.* **2011**, *65* (1), 10−13.

(108) Yamanishi, Y.; Bach, F.; Vert, J.-P. Glycan Classification with Tree Kernels. *Bioinformatics* **2007**, *23* (10), 1211−1216.

(109) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40−55.

(110) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85−117.

(111) Sherstinsky, A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network. *Phys. D: Nonlinear Phenomena* **2018**, *404*, 132306.

(112) Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A Survey of Transformers. *arXiv* 2021, arXiv:2106.04554, DOI: 10.48550/arXiv.2106.04554.

(113) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Networks Learning Syst.* **2019**, *32*, 4−24.

(114) Bagdonas, H.; Fogarty, C. A.; Fadda, E.; Agirre, J. The Case for Post-Predictional Modifications in the AlphaFold Protein Structure Database. *Nat. Struct. Mol. Biol.* **2021**, *28*, 869.

(115) Ardejani, M. S.; Noodleman, L.; Powers, E. T.; Kelly, J. W. Stereoelectronic Effects in Stabilizing Protein−N-Glycan Interactions Revealed by Experiment and Machine Learning. *Nat. Chem.* **2021**, *13* (5), 480−487.

(116) Alley, E. C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G. M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16* (12), 1315−1322.

(117) Taujale, R.; Venkat, A.; Huang, L.-C.; Zhou, Z.; Yeung, W.; Rasheed, K. M.; Li, S.; Edison, A. S.; Moremen, K. W.; Kannan, N. Deep Evolutionary Analysis Reveals the Design Principles of Fold A Glycosyltransferases. *eLife* **2020**, *9*, No. e54532.

(118) Taujale, R.; Zhou, Z.; Yeung, W.; Moremen, K. W.; Li, S.; Kannan, N. Mapping the Glycosyltransferase Fold Landscape Using Interpretable Deep Learning. *Nat. Commun.* **2021**, *12*, 5656.

(119) Bojar, D.; Powers, R. K.; Camacho, D. M.; Collins, J. J. Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions. *Cell Host Microbe* **2021**, *29*, 132.

(120) Carpenter, E. J.; Seth, S.; Yue, N.; Greiner, R.; Derda, R. GlyNet: A Multi-Task Neural Network for Predicting Protein-Glycan Interactions. *Chem. Sci.* **2022**, *13*, 6669.

(121) Dai, B.; Mattox, D. E.; Bailey-Kellogg, C. Attention Please: Modeling Global and Local Context in Glycan Structure-Function Relationships. *bioRxiv* **2021**, PPR407970.

(122) Burkholz, R.; Quackenbush, J.; Bojar, D. Using Graph Convolutional Neural Networks to Learn a Representation for Glycans. *Cell Rep.* **2021**, *35* (11), 109251.

(123) Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent Applications of Deep Learning and Machine Intelligence on in Silico Drug Discovery: Methods, Tools and Databases. *Brief. Bioinform.* **2019**, *20* (5), 1878−1912.

(124) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.

(125) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A Compact Review of Molecular Property Prediction with Graph Neural Networks. *Drug Discovery Today Technol.* **2020**, *37*, 1−12.

(126) Rademacher, C.; Paulson, J. C. Glycan Fingerprints: Calculating Diversity in Glycan Libraries. *ACS Chem. Biol.* **2012**, *7* (5), 829−834.

(127) Thomès, L.; Bojar, D. The Role of Fucose-Containing Glycan Motifs Across Taxonomic Kingdoms. *Front. Mol. Biosci.* **2021**, *8*, 755577.

(128) Taherzadeh, G.; Dehzangi, A.; Golchin, M.; Zhou, Y.; Campbell, M. P. SPRINT-Gly: Predicting N- and O-Linked Glycosylation Sites of Human and Mouse Proteins by Using Sequence and Predicted Structural Properties. *Bioinformatics* **2019**, *35* (20), 4140−4146.

(129) Pakhrin, S. C.; Aoki-Kinoshita, K. F.; Caragea, D.; Kc, D. B. DeepNGlyPred: A Deep Neural Network-Based Approach for Human N-Linked Glycosylation Site Prediction. *Molecules* **2021**, *26* (23), 7314.

(130) Hwang, H.; Jeong, H. K.; Lee, H. K.; Park, G. W.; Lee, J. Y.; Lee, S. Y.; Kang, Y.-M.; An, H. J.; Kang, J. G.; Ko, J.-H.; et al. Machine Learning Classifies Core and Outer Fucosylation of N-Glycoproteins Using Mass Spectrometry. *Sci. Rep.* **2020**, *10*, 318.

(131) Tiwary, S.; Levy, R.; Gutenbrunner, P.; Salinas Soto, F.; Palaniappan, K. K.; Deming, L.; Berndl, M.; Brant, A.; Cimermancic, P.; Cox, J. High-Quality MS/MS Spectrum Prediction for Data-Dependent and Data-Independent Acquisition Data Analysis. *Nat. Methods* **2019**, *16* (6), 519−525.

(132) Gessulat, S.; Schmidt, T.; Zolg, D. P.; Samaras, P.; Schnatbaum, K.; Zerweck, J.; Knaute, T.; Rechenberger, J.; Delanghe, B.; Huhmer, A.; et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* **2019**, *16*, 509−518.

(133) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat. Methods* **2019**, *16* (1), 63−66.

(134) Ye, Z.; Vakhrushev, S. Y. The Role of Data-Independent Acquisition for Glycoproteomics. *Mol. Cell. Proteomics* **2021**, *20*, 100042.

(135) Kotidis, P.; Kontoravdi, C. Harnessing the Potential of Artificial Neural Networks for Predicting Protein Glycosylation. *Metab. Eng. Commun.* **2020**, *10*, No. e00131.

(136) Štor, J.; Ruckerbauer, D. E.; Széliová, D.; Zanghellini, J.; Borth, N. Towards Rational Glyco-Engineering in CHO: From Data to Predictive Models. *Curr. Opin. Biotechnol.* **2021**, *71*, 9−17.

(137) Sanderson, T.; Bileschi, M. L.; Belanger, D.; Colwell, L. J. ProteInfer: Deep Networks for Protein Functional Inference. *bioRXiv* **2021**, 2021.09.20.461077.

(138) Mészáros, B.; Járvás, G.; Kun, R.; Szabó, M.; Csánky, E.; Abonyi, J.; Guttman, A. Machine Learning Based Analysis of Human Serum N-Glycome Alterations to Follow up Lung Tumor Surgery. *Cancers* **2020**, *12* (12), 3700.

(139) Huang, C.; Fang, M.; Feng, H.; Liu, L.; Li, Y.; Xu, X.; Wang, H.; Wang, Y.; Tong, L.; Zhou, L.; et al. N-glycan Fingerprint Predicts Alpha-fetoprotein Negative Hepatocellular Carcinoma: A Large-scale Multicenter Study. *Int. J. Cancer* **2021**, *149* (3), 717−727.

(140) Mattox, D. E.; Bailey-Kellogg, C. Comprehensive Analysis of Lectin-Glycan Interactions Reveals Determinants of Lectin Specificity. *PLoS Comput. Biol.* **2021**, *17*, e1009470.

(141) Gao, C.; Wei, M.; McKitrick, T. R.; McQuillan, A. M.; Heimburg-Molinaro, J.; Cummings, R. D. Glycan Microarrays as Chemical Tools for Identifying Glycan Recognition by Immune Proteins. *Front. Chem.* **2019**, *7*, 833.

(142) Carpenter, E. J.; Seth, S.; Yue, N.; Greiner, R.; Derda, R. GlyNet: a multi-task neural network for predicting protein-glycan interactions. *Chem Sci.* **2022**, *13* (22), 6669−6686.

(143) Bojar, D.; Meche, L.; Meng, G.; Eng, W.; Smith, D. F.; Cummings, R. D.; Mahal, L. K. A Useful Guide to Lectin Binding: Machine-Learning Directed Annotation of 57 Unique Lectin Specificities. *ACS Chem. Biol.* **2022**, 689.

(144) Lundstrøm, J.; Korhonen, E.; Lisacek, F.; Bojar, D. LectinOracle: A Generalizable Deep Learning Model for Lectin−Glycan Binding Prediction. *Adv. Sci.* **2022**, *9* (1), 2103807.

(145) Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C. L.; Ma, J.; Fergus, R. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2016239118.

(146) Edgar, R. C.; Taylor, J.; Lin, V.; Altman, T.; Barbera, P.; Meleshko, D.; Lohr, D.; Novakovsky, G.; Buchfink, B.; Al-Shayeb, B.; et al. Petabase-Scale Sequence Alignment Catalyses Viral Discovery. *Nature* **2022**, *602* (7895), 142−147.

(147) Brinkerhoff, H.; Kang, A. S. W.; Liu, J.; Aksimentiev, A.; Dekker, C. Multiple Rereads of Single Proteins at Single−Amino Acid Resolution Using Nanopores. *Science* **2021**, *374* (6574), 1509−1513.

(148) Lippold, S.; de Ru, A. H.; Nouta, J.; van Veelen, P. A.; Palmblad, M.; Wuhrer, M.; de Haan, N. Semiautomated Glycoproteomics Data Analysis Workflow for Maximized Glycopeptide Identification and Reliable Quantification. *Beilstein J. Org. Chem.* **2020**, *16*, 3038−3051.

(149) Trbojević-Akmačić, I.; Lageveen-Kammeijer, G. S. M.; Heijs, B.; Petrović, T.; Deriš, H.; Wuhrer, M.; Lauc, G. High-Throughput Glycomic Methods. *Chem. Rev.* **2022**, DOI: 10.1021/acs.chemrev.1c01031.

(150) Agard, N. J.; Bertozzi, C. R. Chemical Approaches To Perturb, Profile, and Perceive Glycans. *Acc. Chem. Res.* **2009**, *42* (6), 788−797.

(151) Drake, R. R.; West, C. A.; Mehta, A. S.; Angel, P. M. MALDI Mass Spectrometry Imaging of N-Linked Glycans in Tissues. In *Glycobiophysics*; Yamaguchi, Y., Kato, K., Eds.; Advances in Experimental Medicine and Biology; Springer Singapore: Singapore, 2018; Vol. *1104*, pp 59−76.

(152) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; et al. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370−3388.

(153) Bittremieux, W.; May, D. H.; Bilmes, J.; Noble, W. S. A Learned Embedding for Efficient Joint Analysis of Millions of Mass Spectra. *Nature Methods* **2018**, *19*, 675−678.

(154) Perez, S.; Makshakova, O. Multifaceted Computational Modeling in Glycoscience. *Chem. Rev.* **2022**, DOI: 10.1021/acs.chemrev.2c00060.

(155) Mansoor, S.; Baek, M.; Madan, U.; Horvitz, E. Toward More General Embeddings for Protein Design: Harnessing Joint Representations of Sequence and Structure. *bioRXiv* **2021**, 2021.09.01.458592.

(156) Woods, R. J. Predicting the Structures of Glycans, Glycoproteins, and Their Complexes. *Chem. Rev.* **2018**, *118* (17), 8005−8024.

(157) Narimatsu, Y.; Büll, C.; Chen, Y.-H.; Wandall, H. H.; Yang, Z.; Clausen, H. Genetic Glycoengineering in Mammalian Cells. *J. Biol. Chem.* **2021**, *296*, 100448.

(158) Minoshima, F.; Ozaki, H.; Odaka, H.; Tateno, H. Integrated Analysis of Glycan and RNA in Single Cells. *iScience* **2021**, *24* (8), 102882.

(159) Kearney, C. J.; Vervoort, S. J.; Ramsbottom, K. M.; Todorovski, I.; Lelliott, E. J.; Zethoven, M.; Pijpers, L.; Martin, B. P.; Semple, T.; Martelotto, L.; et al. SUGAR-Seq Enables Simultaneous Detection of Glycans, Epitopes, and the Transcriptome in Single Cells. *Sci. Adv.* **2021**, *7*, eabe3610.

(160) Thu, C. T.; Mahal, L. K. Sweet Control: MicroRNA Regulation of the Glycome. *Biochemistry* **2020**, *59* (34), 3098−3110.

(161) Groth, T.; Gunawan, R.; Neelamegham, S. A Systems-Based Framework to Computationally Describe Putative Transcription Factors and Signaling Pathways Regulating Glycan Biosynthesis. *Beilstein J. Org. Chem.* **2021**, *17*, 1712−1724.

(162) Klein, J.; Zaia, J. Glypy: An Open Source Glycoinformatics Library. *J. Proteome Res.* **2019**, *18* (9), 3532−3537.

(163) Horlacher, O.; Nikitin, F.; Alocci, D.; Mariethoz, J.; Müller, M.; Lisacek, F. MzJava: An Open Source Library for Mass Spectrometry Data Processing. *J. Proteomics* **2015**, *129*, 63−70.

(164) Thomès, L.; Burkholz, R.; Bojar, D. Glycowork: A Python Package for Glycan Data Science and Machine Learning. *Glycobiology* **2021**, *31*, 1240−1244.