Tools to cut the sweet layer-cake that is glycoproteomics

Proteins and sugars interact in sweet, complex ways. More tools are emerging, but working in this field still takes endurance.

Vivien Marx

Pop a sample into an instrument and get a cell's glycome—its entire set of sugars—at single-cell resolution. And it reads out the number of sugars and where they are attached to the cell's proteins. For now, however, there's no simple analyzer for glycans, the complex sugars found within living cells and on the cell surface. They may be attached to other biomolecules such as proteins, nucleic acids or lipids. "We are making major progress but signal sensitivity is an important problem as we do not have a simple way to 'amplify' glycan sequence information yet," says MIT researcher Laura Kiessling. As in genomics, as sequencing has made clear, "I think we will see more patterns when we have more information." To 'see' patterns, researchers can use chemical labels[1] such as those from the Kiessling lab. To identify many peptides and the glycans attached to them, they can use mass spectrometry and any number of informatics tools.

"The glycoproteomics field is at an exciting time," says Nicholas Riley, a post-doctoral fellow in the Stanford University lab of Carolyn Bertozzi. With efforts that predate his own, the research community has kept pushing ahead. The field has grown and over the last two to three years it has reached "critical mass." Enrichment techniques are an active area of research, as are data acquisition approaches with mass spec–based methods, he says, "but what feels like the greatest arena of activity has been in the informatics space."

Glycoproteomic data are challenging, says Riley, so there's "a great need for informatics tools that can handle all of the caveats and considerations." Some tools can handle the tricky task of localizing where on the protein a glycan is attached. Before delving into how some labs describe their tool strategies, it's useful to first consider how studying glycoproteins leads to a collaborative mindset and why glycoproteins matter.

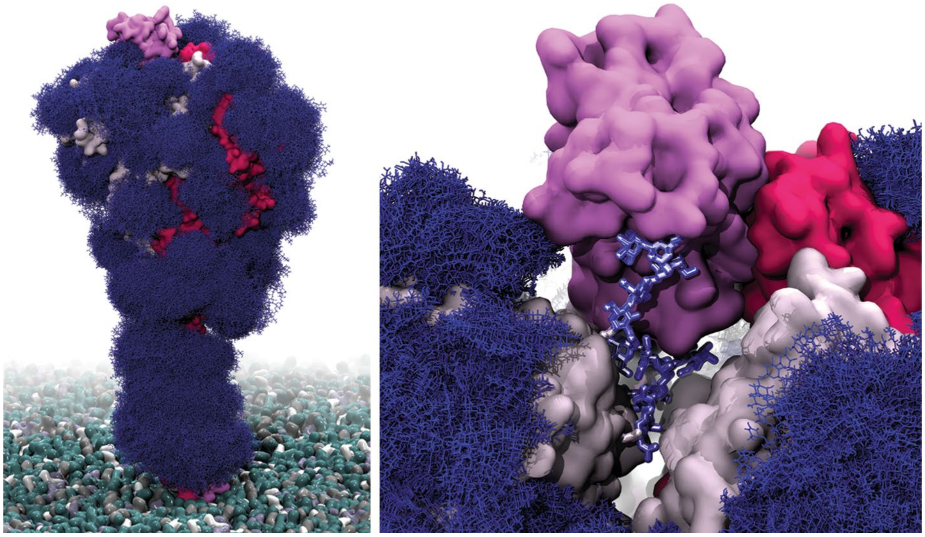
Proteins and carbohydrates connect, and not just on a dessert plate. Protein-bound sugar chains are found in and on cells of many organisms. In glycoproteomics, scientists seek to learn which glycans are where on these proteins. Credit: Kenishirotie/Alamy Stock Photo

Collaborative sugars

"I think sometimes the exact structure of a glycan really matters, and other times the glycan imparts more of an aggregate chemical behavior," says Lloyd Smith from the University of Wisconsin–Madison. About glycoproteins, much detail is still to be revealed, he says, and "there is much to learn." Studying glycosylation is quite collaborative, so teams with different expertise find one another, says Kay-Hooi Khoo, a researcher at the Academia Sinica in Taiwan. As a glycobiologist, depending on the scientific question, he might work with colleagues in cryo-electron microscopy (cryo-EM) or immunology. It's taken time for glycobiology and glycoproteomics to be acknowledged, and full recognition of their scientific importance will take a while. Khoo says that telling a colleague that a protein has certain glycans attached to it can still lead to a "Wow, so what?" response.

There's a methods issue. "In glycans, you don't have a CRISPR equivalent," he says. One cannot target and alter a particular glycan residue without touching any other. The hope is that it will eventually be possible to define the glycans on all of a protein's potential glycosylation sites and explore how this changes, for instance, during development or because of cell cycle phase or immune stimulation.

There's also a numbers issue. A protein with just ten glycosylated sites, also called glycosites, can have different types of glycans attached to each site, says Khoo. For now, labs do not have a comprehensive view of a sample's glycoproteome; they report what they can see. "The reality is we are only seeing the tip of an iceberg, we are not seeing the whole ice," he says. Just having this iceberg tip is leading the field quite far, so Khoo and others are glad to see the field's progress and its prospects.



Glycans hide the SARS-CoV-2 virus from the immune system. At two distinct positions (blue structures, right image) they also support the spike protein in the ‘up’ conformation needed for infection. Credit: L. Casalino, UCSD

Glycoproteins in action

Glycans on a cell surface can be hard to tell apart, especially on bacterial cells, says Kiessling. That’s because most current glycoproteomics methods are directed at mammalian glycans. It was long thought that bacteria do not glycosylate their proteins, but evidence has reversed this. Glycosylation in prokaryotes is “an accepted fact”[2], and glycoproteins in bacteria appear to play roles in infection and pathogenesis. Kiessling and her group develop and use ways to tag bacterial glycans to explore their functional roles by examining the effects of chemical perturbations. One type of probe is a selective label equipped with a bioorthogonal handle, such as azido-arabinofuranose. “This handle can be used to pull down the glycans for further analysis,” she says. Arabinofuranose derivatives selectively label glycans in mycobacteria. “We are taking advantage of endogenous pathways to install groups with chemical handles,” she says. Kiessling and her collaborator MIT biologist Barbara Imperiali also explore ‘pulling down’ relevant glycans using human lectins[3], a kind of glycoprotein, many of which bind microbial glycans. The team hopes probes can ease the way to isolating bacterial glycans and help to identify and characterize protein–carbohydrate interactions at the cell surface. What intrigues Kiessling and her group is that sugars on the surfaces of cells interact with proteins on the surfaces of other cells, both pathogens and friendlies, but these are weak interactions: anywhere from one thousand to one million times weaker than the forces at work in protein–protein interactions. With the probes, the scientists can explore these delicate interactions.

Glycoproteins also matter with SARS-CoV-2, the virus that causes COVID-19. To begin infection, the virus’s spike protein uses the angiotensin-converting enzyme 2 (ACE2) receptor on our cells as its unfortunate doorknob. Infected cells produce the virus and lend their glycosylation biosynthesis system to make the virus’s sugar coat, says Shisheng Sun, a glycoproteomics researcher at Northwest University in Xi’an city, China. Cells are coated with sugars, and so is the virus. Sugars on the spike protein aid and abet infection by shielding the protein from detection by the host’s immune system. The spike protein protrudes from behind this glycan curtain as needed.

But the glycans are also “playing a more active role than just shielding,” says Khoo, who also studies SARS-CoV-2 variants. He points to research[4] that shows how N-glycans at particular sites on the spike protein have dynamic control. N-glycans link to the amino acid asparagine in a protein. At two distinct positions, the N-glycans stabilize the spike protein in the ‘up’ position, the conformation it needs for infection. Removing these glycans destabilizes the viral receptor-binding domain and reduces binding to ACE-2. Khoo says this insight informs the development of treatments and vaccines and allows a better understanding of the virus.

The N and the O

Glycoproteomics also matters in cancer research. Cancer cells have altered glycoproteins, says Sun. For now, he says, most researchers have found only associations between glycoproteomic shifts and cancer, so “more studies are needed to confirm whether these changes are the reasons to cause cancer or just results.” Sun thinks that StrucGP[5], his software for analyzing tandem mass spec spectra, can help cancer biologists explore how altered glycans link to their cancer question of interest, in particular for N-glycans. The N-glycosylation site, says Sun, usually has a restricted consensus motif, NXS/T or asparagine-X-serine/threonine, in which X can be any amino acid except proline. The software can connect information about which N-glycans are modified on glycoproteins and find which of the protein’s glycosites the glycans inhabit. It can also help researchers link the glycans changed in cancer with information about modified glycoproteins and glycosites, and then connect that to what they know about related pathways, cellular locations, possible molecular functions and other relevant information. This network of information feeds into exploration of the specific roles that biological glycosylation and glycoproteins play in cancer.

In a liver cancer cell line treated with human growth factor, Sun and colleagues used StrucGP to identify over 2,000 intact glycopeptides, and they identified a certain type of N-glycan, core fucosylated glycans, as the most common[6]. They looked at site-specific glycosylation and found that cells can increase their folate uptake when there is increased core fucosylation on the protein folate receptor α (FOLR1), especially at one particular glycosite, asparagine 201. They followed up with molecular methods to confirm that the level of core fucosylation on FOLR1, especially at this glycosite, positively regulates the capacity of cells to take up folates. This enhanced folate uptake can promote epithelial-to-mesenchymal transition in these cells, which can render cancers invasive and capable of metastasis. To find the changed glycans at the structural and site-specific level, software users need a little proteomics background. But the lab works on making StrucGP more widely usable.

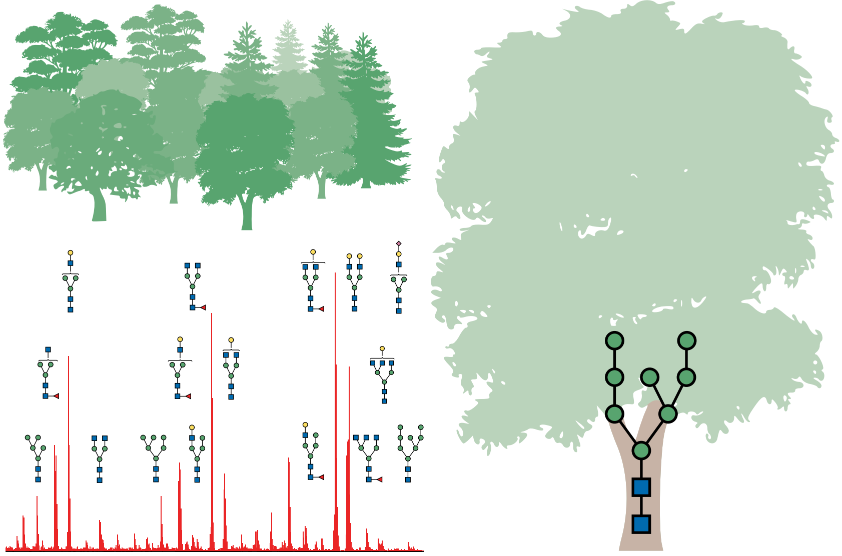
As Sun delves into cancer and other disease-related cell biological questions, using his and other tools, he seeks to characterize the sequences and structures of glycans and glycoproteins in cells. Tandem mass spec delivers much useful data, he says, “but they are just like a book with unknown words, we need to find a way to understand their meanings.”

Human cells use four major monosaccharides, and software reveals how many monosaccharides make up these glycans. With mass spec, Sun's and other labs reap intact glycopeptides, which are sugars attached to the peptides they modify. The results show a basket of b/y ions, which are fragmented from peptides. There are B/Y ions from the glycan portions. To identify the composition of glycans and its peptides in these baskets, he says, scientists can use open software tools such as GPQuest, pGlyco 2.0, MSFragger-Glyco and the commercial software Byonic.

StrucGP decodes the detailed N-glycan structures at each attachment site on a protein. Sun thinks of N-glycans like a tree with a trunk and branches. StrucGP reveals what the trunk and branches look like. Typically, when tandem mass spec yields spectra, labs use a database to identify glycans. StrucGP does not use a database, he says. Instead, it applies what is known about N-glycan biosynthesis and the modules that make up N-glycans. The software predicts branch structures and uses 'feature fragment ions' to identify the whole glycan structure. For example, he says, Y ions of each of the four types of 'trunks', each with two or three monosaccharides, have recognizable patterns in the spectra. The branch structures, reflected by the B ions, are recognizable, too. With the molecular weight of the whole glycan, which is the molecular weight of the intact glycopeptide minus that of the peptide, the tool can calculate all possible branch structures. The idea for this method occurred to Sun when he just had joined the university in 2017. The lab had no mass spec instrument then.

At the time, he told his students that, to fully understand proteomics and glycoproteomics, they had to learn to read "raw-data level" spectra manually. It made him think about developing high-throughput computational analysis methods to identify glycan structures at a glycosite-specific level and to do so de novo, in a database-independent manner. After all, he says, no comprehensive database of all glycan structures exists. And given the many structural isoforms for each glycan composition, it's challenging to build a 'theoretical' glycan structure database.

For their de novo analysis route, the researchers separate the analysis of glycans into 'trunk' and 'branches'. It's a modular strategy that shows that thousands of N-glycans are made up of four core structures with dozens of branch structures, each with at least one, and up to six, branches. The whole glycan structure can be identified by separately recognizing each module of a glycan from its feature fragment ions, says Sun.



Cells are covered with glycans, much like a forest, and glycans are structured like trees with branches, says Shisheng Sun. His lab's software StrucGP uses mass spec spectra to decode N-glycan structures at each attachment site on a protein. Credit: S. Sun, Northwest Univ.; T. Phillips, Springer Nature

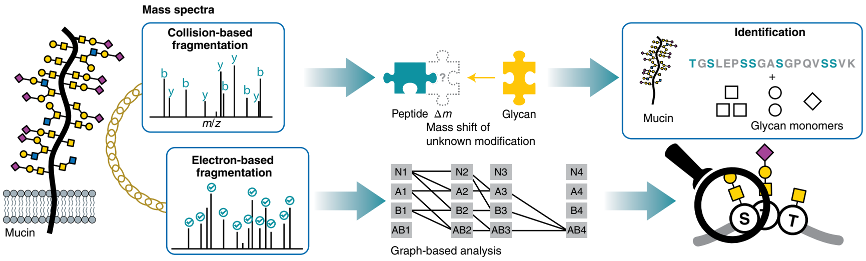
In glycoproteomics and glycobiology, "I think the lack of convenient and effective approaches for site-specific glycan interpretation and intervention are two major bottlenecks," says Sun. His method delivers site-specific glycan structural analysis, but he sees much room for improvement.

To investigate the functional role of glycosylation, one needs to knock down or overexpress a given glycan on a given glycoprotein and glycosite. But, says Sun, the current methods can only knock down or overexpress genes involved in glycosylation, known as glyco-genes, and that will affect all glycoproteins that are modified by the corresponding glycans. One can mutate the glycosite, but then all possible glycans attached at the glycosites will be removed. This is why current approaches cannot readily provide direct evidence to confirm the exact functions of a glycan on a given glycoprotein or glycosite.

Another bottleneck is the large-scale synthesis of standard glycopeptides. When his team addressed manuscript revisions, the team needed standard glycopeptides to assess the software's performance. The researchers explored options that finally led to a single standard glycopeptide. Synthesizing standard glycopeptides remains very challenging, says Sun. Another challenge is O-glycopeptide analysis, which is much harder than analysis of N-glycans.

The difficulty with O-glycopeptides, says Sun, is that whereas the N-glycosylation site usually has the NXS/T motif, O-glycans can be added onto any of a protein's serines or threonines, which makes them hard to find in the spectra. But on the plus side, one O-glycan modification, O-GlcNAc, always involves the same monosaccharide attached at the glycosylation site, and the modification process is controlled by only one synthetase and one hydrolase. This makes it easier to study the biological functions of O-GlcNAc than those of N-glycans, he says. All other types of O-glycans, especially the O-GalNAc type—also known as mucin-type glycans—are hard to analyze. O-GalNAc glycans have at least eight different core structures, they lack a common motif at the glycosylation site and they often occur in clusters. This means that many O-glycans can be attached to one peptide. "All these features make studying O-glycans more difficult than N-glycans," says Sun.

O-Pair Search[7] is a computational way to identify O-glycopeptides and localize where the O-glycans attach on peptides. The tool is integrated in the open platform MetaMorpheus. According to the developers, their tool finds more O-glycopeptides than others, it can work with different O-glycan databases and it works both with higher-energy collisional dissociation spectra and with those from electron-transfer dissociation, through a



Mucins are densely O-glycosylated proteins on the cell surface. O-Pair Search is software that can use paired mass spectra to identify O-glycopeptides from mucins and localize the O-glycans within the peptide sequence. Credit: N. Riley, Stanford Univ.

technique called higher-energy collisional dissociation (HCD) supplemental activation. It can, say the developers, reduce the O-glycopeptide search time by up to 2,000 times compared to the widely used commercial tool Byonic.

Lei Lu led the work on O-Pair Search during his PhD research in the Smith lab at the University of Wisconsin–Madison. Lu is now a postdoctoral fellow in the DeGrado lab at the University of California, San Francisco. The tool identifies O-glycopeptides as well as localizes O-glycosites, which, says Lu, is especially important for glycopeptides with multiple glycosylated sites. For example, mucin-type O-glycosylation, which plays roles in immune response, is especially hard to analyze. Dozens of mucin-type O-glycans have been identified in people. What makes O-linked glycans tough to detect by mass spec are aspects such as their lability and ionization inefficiency.

Echoing his colleague Riley, Lu says that the ability to detect these O-glycans has improved through new enrichment methods, new instrument methods that include fragmentation types such as electron transfer disassociation and new data acquisition strategies. Among the bioinformatic challenges are low search speed and inaccurate assignment of O-glycans to serine and threonine residues. Mucins have dense regions of serines and threonines, and each serine and threonine can be modified by different types of O-glycans. It appears, he says, that nature likes to put different numbers of O-glycans on mucin peptides.

To localize O-glycans on peptides with a database search method, scientists traditionally build an O-glycopeptide database by assigning all possible O-glycans to the peptide sequences in silico. To Lu, the challenge brings to mind the math thought experiment called the wheat and chessboard problem in which one moves across a chessboard, grains in hand. At each square, one doubles the number of grains compared to the previous square. A full chessboard will have quintillions of grains, which exceeds by around 2,000 times the world's grain output. “You can imagine when the number and type of O-glycans increase per mucin peptide, the possibility of O-glycosylation will be increased by multiple orders of magnitude and the brute-force method is unsupportable,” says Lu. This computational intractability led his team to develop a graph-based approach with dynamic programming, which takes a different approach to database searching. It's widely used in computational gene sequence comparison and top-down proteomics analysis. With O-Pair Search, the team designed and optimized this apprpach for O-glycopeptide analysis.

The methods developers take advantage of the ion-indexed open search strategy to identify peptide sequences, says Lu. To obtain localization, they don't need to assign O-glycans to each peptide and build up a huge glycopeptide database. Instead, the assignment of O-glycans to the peptide sequence happens simultaneously with the matching process. This approach shrinks the number of potential matches.

O-PairSearch has broad applicability in the diverse glycoproteomics field, says Riley. Previously some informatics tools were designed for specific workflows or particular spectral and fragmentation types. “That restricted which algorithms could be used for any given experiment,” he says. More recently, developers have better addressed researcher needs by building platforms that are more flexible in the data types they can handle. Most tools are designed to handle N-glycoproteomics data but struggle with O-glycoproteomics data, he says.

N-glycoproteomics is significantly easier than O-glycoproteomics, says Riley. Many tools handle N-glycopeptides well, he says, which motivated the team to design O-Pair Search specifically for O-glycopeptides. O-glycosylation affects dense regions of serines and threonines that all can be modified with a large number of glycans, says Riley. That causes “search space issues,” meaning that the combinatorial space involved in considering all possible modification sites is huge.

Difficulties with identification arise because many sequences have similar amino acid residue compositions and similar glycan mass distributions. One still has to assign the correct glycan mass and peptide sequence, and find which residues harbor which glycans, says Riley. Then there are several types of O-glycosylation. The team tried to keep all of this in mind as they worked on O-Pair Search.

Speaking more generally about O-PairSearch, Smith says, “I am really proud of our group making open-source software that others can access understand, use and build upon to improve.” In his view, commercial software suffers from being both costly and closed source, which “is a drag upon scientific progress.” But he also realizes that software needs support, “and the major benefit of commercial software is that it can be maintained, which is also important.“

ProteinMetrics sells the Byos platform, a proteomics software suite of workflows for mass spec. One ‘node’ is the widely used Byonic search engine for identifying and localizing modifications. Eric Carlson, president and CEO of ProteinMetrics, says that a deep understanding of protein glycosylation is critical to basic biology research and for developing biotherapeutics. Sensitive tools help to characterize the inherent heterogeneity of samples, he says. “Specialized tools focused only on glycosylation are insufficient,” he says.

The search engine can handle complex samples with multiple modifications, he says. Researchers have used this platform to find out about membrane-spanning proteins in gram-negative bacteria. These microbes have a sandwich of a cell envelope: there's an outer membrane layer, a peptidoglycan layer beneath that and an inner membrane layer. Carlson points to a study[8] that Marshall Bern, Protein Metrics co-founder and the company's vice president of research, co-authored.

In a number of gram-negative bacterial species, the team found that beta-barrel outer membrane proteins are covalently tethered to the peptidoglycan layer, and the attachments can vary depending on cell cycle phase. Their analysis included genetic analysis, molecular dynamics simulations, structural modeling, cryo-EM imaging, immunoblotting and liquid chromatography with tandem mass spec analysis, and they used the Byos platform to identify and compute the extracted ion chromatograms for the peptidoglycan-bound proteins. In Carlson's view, such work stands to help with developing the "next generation" of antibiotics. And, he says, it shows that "what might appear to be an area in need of a specialized tool, is anything but."

An FDR drive

As the many tools have emerged, major challenges and bottlenecks have remained, says Riley, including the speed of informatics tools, their robustness and their breadth of applicability. Speed matters because the search space of a glycoproteomics experiment is very large relative to the proteome, he says. Several approaches are helping to make speed less of a factor.

Robustness is another issue due to, for example, the large search space and the fact that different modifications can combine to equal the same mass. This means that false discovery rates (FDR) in glycoproteomic experiments remain high. "This is still a major issue in the field that has not been fully resolved," says Riley.

In O-Pair Search, says Riley, the team built in some quality metrics and worked to improve the robustness of the way they calculate FDR. "We calculate FDR separately for modified peptides and non-modified peptides," he says. Not doing so can lead to a higher number of incorrect hits for modified peptides. The team also set up a quality measure for localization levels in the spectra. "That said, every search algorithm and informatics tool that is identifying glycopeptides still has room to improve in its FDR scores," he says.

Overall, some things in glycoproteomics have become easier, says Riley, such as setting up mass spec methods to acquire glycopeptide data. "It is now easier than ever to design a mass spec method to acquire high-quality glycoproteomics data," he says. It's still hard, he says, to get glycopeptides to the point at which they can be analyzed, such as by digesting glycoproteins or properly enriching glycopeptides, and to analyze the data to get confident identifications and quantifications.

It matters to be able to delineate "high-confidence, localized glycopeptide identifications from other identification," says Riley. Non-localized glycopeptide identifications can be useful, but fully localized identifications are better. Having the ability to quickly understand which identifications are which in a dataset is helpful. "Our localization level assignments in O-Pair Search may not be 'the' answer," he says. "But I hope it is an answer that continues an important conversation for future informatic tool development."

Cells are covered in diverse sugars, as this artist's rendering shows. Sugars can be attached to proteins. Software and hardware advances help labs to localize these sugars. Credit: MedicalRF.com

Tough trade-offs

N-glycans are quite ubiquitous and have often been the focus of studies and tool development, whereas the many types of O-glycans call out for further research. "Currently the technology is geared towards solving the more ubiquitous ones," says Khoo. This has led to more tools and discussion about N-glycosylation. Of late, however, he has seen more of a focus on O-glycosylation.

Localization, he says, gets particularly challenging, for example, when identifying a peptide with more than one site carrying a glycan[9]. Beyond querying localization, signal is crucial with mass spec, and it's where ionization efficiency comes into play. Software cannot work without a good signal, he says. If a researcher needs five fragment ions to be confident about a "hit" but the software only has two ions to work with, that software cannot address identity assignment. Only an improved spectrum, he says, will yield more confident localization, more confident glycan identification and a push of the FDR down to zero or close to zero.

Annotation with an FDR can be used when assigning which peptide carries which glycans and at which site. But scientists face a trade-off, says Khoo. They have to choose between wanting more 'hits' in which they will have less confidence or wanting fewer hits that offer greater confidence. Biologists might be looking whether certain conditions lead to upregulation or downregulation of certain proteins. When spectra do not find "their protein," they are uninterested in the data, he says. "Whatever we can see in the data, we want to improve the fidelity of," he says. In the next three to five years, more new tools will emerge, and he thinks the community is "within reach" of solving this challenge.

In proteomics, FDR can be addressed relatively easily, says Khoo. A database lookup yields a 'true hit', and then labs can look at a 'reverse sequence', sometimes called a garbage sequence, in which they find no hit. In glycoproteomics, it's less clear how to control the FDR for glycan identification. This is not a software issue, it's an "inherent problem," says Khoo. Byonic and other tools take a "peptide-first" search approach. Other tools involve a "glycan-first" search approach, which can provide more accurate glycan assignment.

When a researcher makes a hard choice between numbers and accuracy, says Khoo, he or she might find that a Byonic search of 8,000 glycopeptide hits leads to 10-20% inaccurate assignments. That researcher might choose analysis with other software that delivers 8,000 hits of which 7,500 are accurate. A cancer biologist's "favorite glycoprotein" will be more likely to be found with a method that delivers a higher number of hits. But that will be accompanied by a higher FDR. There is dynamic range to contend with, too: some glycopeptides are highly abundant, others are not.

Generally speaking, Khoo tells his students that artificial intelligence will be doing much of the heavy lifting in data analysis of the future. But one area likely to be excluded from this progress, in his view, is glycobiology. "Even ten years after you graduate," he says, "every other problem will be solved, except glyco, so you have a bright future." ❐

Vivien Marx ✉
Nature Methods.
✉*e-mail: v.marx@us.nature.com*

Published online: 17 August 2021
https://doi.org/10.1038/s41592-021-01253-w

References

1. Bertozzi, C. & Kiessling, L. *Science* **29**, 2357–2364 (2001).
2. Schmidt, M. A., Riley, L. W. & Benz, I. *Trends Microbiol.* **11**, 554–561 (2003).
3. Wesener, D. A. *Nat. Struct. Mol. Biol.* **22**, 603–618 (2015).
4. Casalino, L. et al. *ACS Cent. Sci.* **6**, 1722–1734 (2020).
5. Shen, J. et al. *Nat. Methods* https://doi.org/10.1038/s41592-021-01209-0 (2021).
6. Jia, L. et al. *Theranostics* **11**, 6905–6921 (2021).
7. Lu, L. et al. *Nat. Methods* **17**, 1133–1138 (2020).
8. Sandoz, K. et al. *Nat. Microbiol.* **6**, 19–26 (2021).
9. Khoo, K. H. *Biochem. Soc. Trans.* **49**, 55–69 (2021).