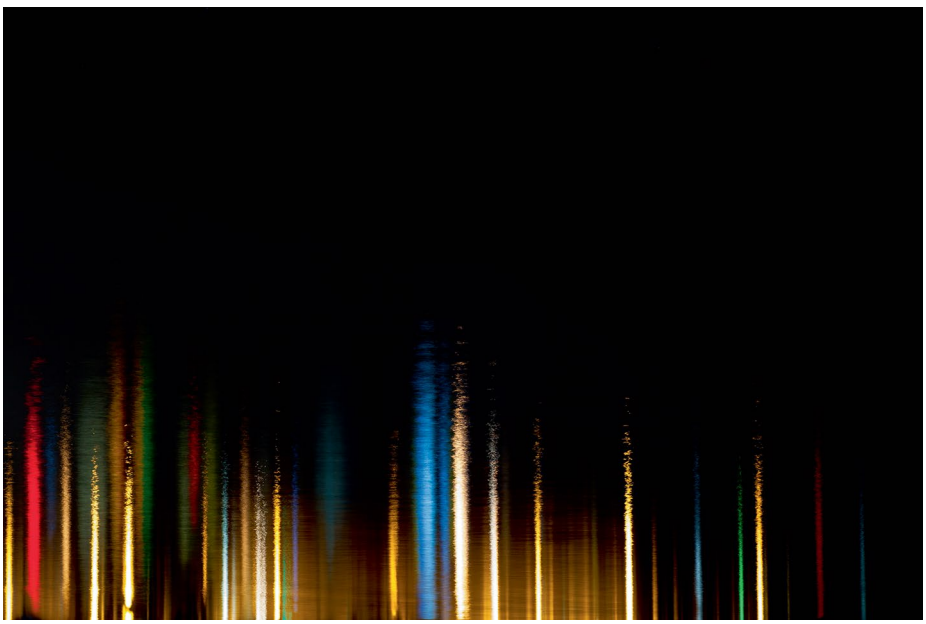
Diving deeper into the proteome

As new technology enables researchers to find and characterize less-common post-translational modifications that drive gene expression and cellular metabolism, the movement to catalog the entire human proteome gains momentum

Caroline Seydel

Sequencing the human genome provided biologists with a potential 'parts list' of the human proteome, but little functional understanding of all those parts. Most research focuses on an elite group of about 5,000 proteins, but the human genome directly encodes some 20,000 gene products—a number that balloons to millions of distinct 'proteoforms' when alternative splicing variants and post-translational modifications (PTMs) are taken into account. This cycle of the best-known proteins getting the most attention is self-reinforcing, because at grant proposal time, it's safer to propose a hypothesis with a lot of solid background data supporting it, rather than an exploratory project on an obscure protein with little connection to known pathways. But a movement is gaining steam among researchers that contends that it's time to break the cycle and reduce this imbalance in attention and funding.

"There is a vocal minority who say, 'Look, we need big science approaches,'" says Neil Kelleher, professor in the department of molecular biosciences at Northwestern University. "I fit into this category of folks who are calling for a moonshot—the Genome Project equivalent for the world of proteins." A Human Proteome Project was launched almost immediately upon completion of the Human Genome Project, but so far, characterizing the vast diversity of proteins has seemed well out of reach given the available technology. "From 20,300 human genes, you can get alternative splicing, different isoforms," Kelleher says. "Add in the genetic sources of variation like SNPs [single-nucleotide polymorphisms] and coding mutations, and then PTMs, it really made the proteome get complicated in an exponential way. But it's not as bad as people think. In terms of the protein explosion, biology really constricts the numbers of PTMs that it uses intentionally." Kelleher is part of the team that's launching the Human Proteoform Project, an effort to generate a reference set of every protein produced by the genome. The organizers hope that a coordinated effort like this will help harness the various sources of funding and resources to pull together and accelerate the development of faster, more efficient proteomics technology. Already, advances in mass spectrometry analysis have enabled more detailed study of how PTMs combine on proteins. In addition, more sensitive detection of rare and transient PTMs is revealing entirely new types of modifications and thereby opening up new avenues of functional research. Yet no matter how sophisticated these tools and methods become, they can only illuminate the proteins they're applied to. That's why some researchers want to start by surveying the field to find out which human proteins could use a little more attention.



Credit: Graham Jepson / Alamy Stock Photo

Understudied Protein Initiative

For years, genomics has been writing checks that proteomics can't yet cash. Tools such as genome-wide association studies identify scores of proteins potentially associated with a particular disease or condition, but many of these proteins haven't been well characterized. Too often, researchers assume that if a protein hasn't been studied much, it probably isn't interesting, says Juri Rappsilber, professor of proteomics at the University of Edinburgh and professor of bioanalytics at the Technical University in Berlin. "But how are we to know if we don't study them?"

Biologically interesting, and even medically relevant, proteins may remain understudied for a variety of reasons. From a practical standpoint, it's harder to study proteins that are small, low in abundance or otherwise difficult to work with in the lab. But also, Rappsilber points out, "the extent of previous data on a protein determines how many ideas we have about what more could be studied on a protein." It can be hard for a researcher to find an entry point to start studying a new protein when little is known about its structure, function or interactions with other proteins.

To help biologists get over that hurdle, Rappsilber is helping launch the Understudied Protein Initiative[1], a project

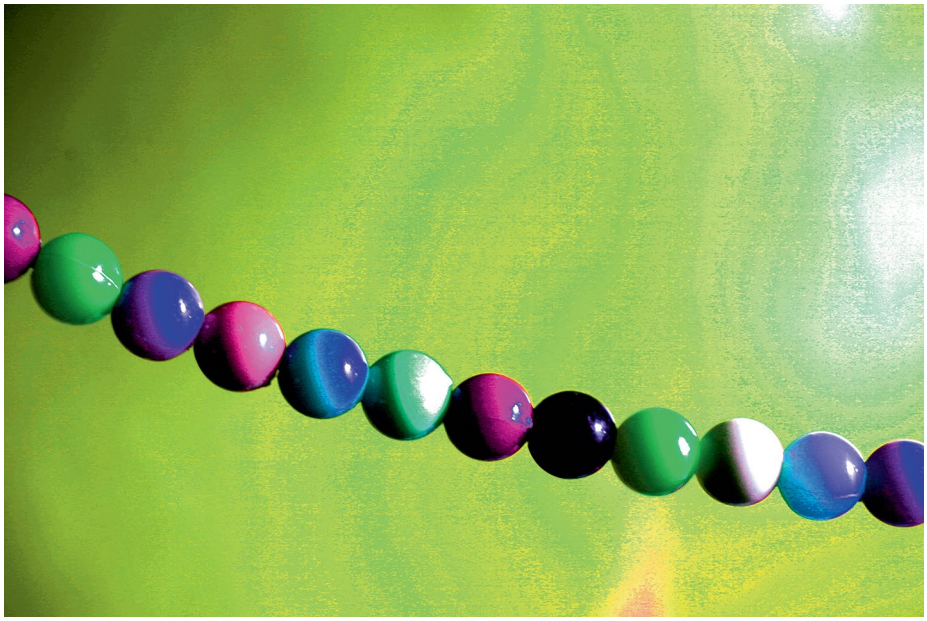
funded by the Wellcome Trust to identify the most promising uncharacterized proteins and the best methods to study them. "Our pitch is that proteomics is capable of seeing a lot, and if we only had enough initial information on the proteins, then molecular biologists and cell biologists would be able to pick the proteins they are specifically interested in and work them up in detail," Rappsilber says.

The project has published a survey to solicit opinions from researchers in the field about which proteins people consider "understudied." Just because a protein has some annotation available in UniProt, for example, doesn't mean that a biologist would find enough information to build a research project around. "What makes the difference from [biologists] saying 'ooh, high risk, let's not touch it,' to 'okay, risky, but we have a good idea how to approach this'?" Rappsilber says. "We want to ask many of them and integrate their decisions to get a human, intuitive-based definition of the problem."

Emma Lundberg, associate professor of bioengineering at Stanford University and professor of cell biology proteomics at KTH Royal Institute of Technology, in Sweden, serves on the organizing committee for the Understudied Protein Initiative workshop. "I think it's key to make people aware of the bias that we have in protein databases right now," she says. "If you do a gene-set enrichment analysis, your results are going to be biased towards the well-studied proteins where we have lots of functional data. We want to start the discussion on how we can make even better use of the large-scale technologies available to us, and avoid biasing our follow-up studies."

Lundberg is also the director of the Cell Atlas, part of the Human Protein Atlas program, whose goal is to document the distribution and localization of proteins within the cell. Data from the Cell Atlas revealed that more than 50% of proteins localize to multiple compartments in the cell. "This means they might have different functions in different compartments," she says. "We can't tell without further studies, functional studies." She points out that most of the proteins with multiple documented functions are drug targets. "They're often highly characterized enzymes," she says. "So of course, there's a bias there. It's probably not that drug targets are more likely to have dual functions. It's that they are so well studied and that's why we know of their dual functions." The Cell Atlas data, then, may inspire more functional studies of proteins that localize to multiple cellular compartments.

In addition to defining which proteins the biology community considers "well studied" or "understudied," the Understudied Proteins Project will collect information about which proteomics techniques yield the most fruit. "What makes me so excited about this initiative is that we start by defining the problem from the point of wanting to solve it," Rappsilber explains. "For funding agencies, it's very attractive, because they can take hard data to decide where they want to put their funding."



Credit: Andy Hallam / Alamy Stock Photo

The Human Proteoform Project

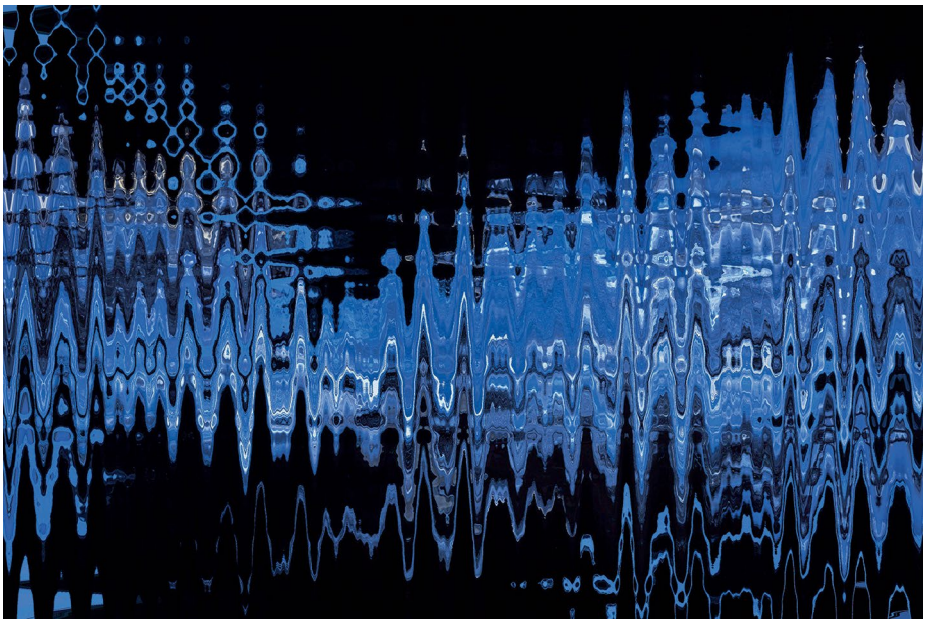
It's tempting to make comparisons to the Human Genome Project, but unlike DNA, proteins come in a dazzling array of chemical configurations. A single gene product can assume dozens or hundreds of different forms, thanks to alternative splicing, PTMs and mutations that alter protein folding. In 2013 the proteomics community coined the word "proteoforms" to refer to all the different unique protein molecules produced from a gene. Defining and characterizing every proteoform will require a coordinated effort involving many different techniques and tools. "The Genome Project was big," says Lundberg, "but this project will be massive." Despite this daunting scale, momentum is building behind the idea of cataloguing the entire collection of human proteoforms. Last year, the nonprofit Consortium for Top-Down Proteomics launched the Human Proteoform Project[2], bringing together researchers from academia, government and industry to collaborate and finance these efforts. The organizers acknowledge that proteomics technology isn't quite up to the job yet, but one of the project's goals is to spur the development of better proteomics tools and techniques, just as the Human Genome Project provided that boost to DNA sequencing.

"It's the arc of history," says Kelleher, who serves as the Consortium's president. "In 1993, it wasn't at all clear that you could sequence DNA for a dollar a base, but by 1996 it was," he points out. By taking that leap, the project itself drove the creation of the technology it needed. A similar moment exists now for proteomics, he says. "We need about a 100-fold increase in the rate of proteoform discovery," says Kelleher, because it's estimated that a single cell contains around 1 million proteoforms. "Right now, from a couple hundred thousand cells of the same type grown in a dish, I can give you about 10,000–20,000 proteoforms in a month," Kelleher says. "There's lots of reasons to believe that in this decade we will invent technology that will match the scale of our proteoform biology with the scale of the measurements needed to bring proteomics on par with genomics."

Over the past decade or so, mass spectrometry has emerged as the most reliable way to distinguish proteoforms, but mass spectrometry techniques typically require cutting up the protein with a protease, such as trypsin, to generate short peptides that can be analyzed. This 'shotgun' or 'bottom-up' proteomics is very robust, allowing high-throughput identification of thousands of proteins from a cell lysate. Because the proteins are digested before analysis, however, it's incredibly difficult to

distinguish individual proteoforms. Imagine taking apart a hundred cars and then counting up all the different parts. You could probably sort them by make and model, but you wouldn't be able to tell whether the bike rack and the heated seats came from two different Honda Civics or the same one.

Kelleher is a pioneer of a method called top-down proteomics, in which the proteins are analyzed intact rather than in pieces. In a project called the Blood Proteoform Atlas, he and colleagues used top-down methods to identify 29,620 different proteoforms from 1,690 genes across 21 different blood and bone marrow cell types[3].Top-down proteomics generates huge datasets, so the researchers had to develop new, cloud-based solutions to analyze the quantity of data being generated. The atlas represents a major technological achievement, but it also has medical relevance: by testing blood samples from liver transplant patients, the team created a proteoform signature associated with transplant rejection.



Credit: Belozerova Daria / Alamy Stock Photo

The 'Goldilocks' approach

Although top-down proteomics methods can now successfully identify, characterize and quantify thousands of intact proteoforms from a cell sample, they're still not quite as robust as bottom-up methods. Various technical challenges remain, including efficient solubilization of large proteins, improved affinity reagents to capture low-abundance proteins and better separation methods for complex mixtures of intact proteins. Some researchers are trying for the best of both worlds. Benjamin Garcia, professor and head of the biochemistry and molecular biophysics department at Washington University School of Medicine in St. Louis, has focused on optimizing novel approaches using 'middle-down' mass spectrometry. Unlike top-down, these approaches still involve cutting the protein into manageable pieces, but the peptides are large enough to reveal combinations of PTMs, an improvement over bottom-up methods[4]. "Middle-down proteomics is kind of like the 'Goldilocks' approach," he says. "You don't use trypsin because it cuts too frequently. You use proteases that cut at amino acids that don't occur as frequently, and you get larger pieces." Identifying combinations of PTMs on the same protein can convey important functional information. Garcia points to the SARS-CoV-2 spike protein as a key example. Glycosylation of the spike protein both helps the virus bind to cellular receptors and disguises it from the immune system. "If you really want to understand everything that's going on, you have to look at all the different proteoforms of that spike protein," he says. "That's going to be the next big push in the proteomics field that's really going to take off, is counting all these different combinatorial proteoforms."

Histones, the proteins that help secure long strands of DNA into tightly packed chromosomes, could be considered poster children for the complexity of PTMs. Using both top-down and middle-down approaches, Garcia has dramatically increased the number of documented histone proteoforms. "For histone H3, probably the most highly modified protein in the human body, we've been able to detect somewhere around 2,500 proteoforms of that same protein," says Garcia. "Theoretically, it should be millions, or a billion if you do the permutations, and there's only 2,500, so that means there are a lot of rules, hierarchy and patterns." PTMs that change the chemical structure of histones can have a significant effect on the cell by influencing gene expression, so it makes sense that they aren't just popped on indiscriminately. "We're getting enough datasets that we can start understanding which histone modifications occur together, which do not like to be found together and which are indifferent," Garcia says.

Now the challenge is to uncover how the different combinations behave in terms of gene expression. "What if you have a histone mark that silences genes at the same time as one that activates genes, which one wins out?" he says. Some histone modifications behave as master regulators, he says, superseding any other modifications found on the protein. Others seem to act like boundary elements, preventing the modification from spreading. "With all this data, it's really interesting to now try to understand the function," he says. "It's been a fun ride, and there's still a lot to do."

Digging deeper into post-translational modifications

"Overall, there's a lot of excitement about identifying and quantifying new post-translational modifications," says Garcia. "Every year, there's a handful that are discovered for the first time which are really interesting or even change what we know about biological pathways."

Over the last 15 years, Yingming Zhao, a professor at the University of Chicago, has uncovered an impressive number of histone lysine modifications, including propionylation, butyrylation, succinylation, malonylation, glutarylation and crotonylation, using bottom-up methods. Zhao and his colleagues developed software that identifies the mass shift caused by protein modifications, known or unknown. "From the mass shift, we can deduce the elemental composition, and from the elemental composition we can deduce the possible structures," Zhao explains. Next, they make a synthetic peptide with the proposed modification and compare it to the cell-derived peptide, to verify that they coelute on HPLC and have similar mass spectral fragmentation patterns. They confirm the specific isomer using an antibody designed against it, among other methods. Last year, using this strategy, Zhao's group discovered a new modification,

methacrylation, which is a structural isomer of crotonylation[5]. "We always make all the synthetic peptides of the structural isomers," Zhao said. When the cell-derived peptide didn't match the crotonylated synthetic peptide on the HPLC, they knew it had to be something else. "Methacrylation matched perfectly," Zhao says. Treating cells with isotopically labeled methacrylate confirmed that it could be added to lysine residues.

These modifications haven't been discovered sooner because they're rare, but a PTM doesn't have to be abundant or long-lasting to have far-reaching effects. If a modification switches on a signaling pathway, the cell may not need it to stick around very long. "These modifications are pretty transient, usually," Garcia says. "To capture them is pretty tough." Advances in enzyme inhibition have enabled researchers to capture reversible modifications that only exist briefly in the cell. Additionally, lysing the cells can disrupt the cell's careful balance of adding and removing modifications, he says. "You start getting a lot of deacetylation or dephosphorylation that normally wouldn't occur," Garcia says. "If you can inhibit that in the lysate or even a few minutes before lysing the cells, it makes a big difference."

Recently, Zhao's group reported a new metabolite-derived histone modification, lactylation, and presented evidence that it constitutes a mechanism by which metabolic changes can influence gene expression[6]. Inside a cell, the concentration of various nutrients and metabolites constantly fluctuates. Histone modifications allow the cell to turn genes on and off in response to these changing conditions. Lactate is a well-known metabolic byproduct, and recent studies had found that it can induce changes in gene expression in immune cells, though it wasn't clear how. Zhao looked at macrophages, immune cells that produce a burst of inflammation to kill bacteria and recruit other immune cells to the infection site. When this happens, the cells ramp up glycolysis, which produces lactate. Zhao's work showed that in macrophages, lactate-derived lactylation turned on genes associated with wound healing, suggesting that the excess lactate may help the cell turn around and repair the damage caused by the barrage of bactericidal inflammation.

Hui Ye at China Pharmaceutical University is taking a different approach to look for lactylation sites. Tandem mass spectrometry generates a cyclic immonium ion from a lactylated lysine, which reliably points to the presence of lactylation. Ye's group combed through dozens of publicly available proteome datasets searching for proteins bearing the distinctive lactylation signature. "We are using the signature ion to explore the 'dark matter' in the proteomic data that's already out there," Ye says. "It's a very cool approach that complements the conventional methods."

Credit: Martin Wierink / Alamy Stock Photo

They found a great deal of lactylation on enzymes involved in glycolysis, including fructose-bisphosphate aldolase A, or ALDOA. To home in on lactylation sites that were functionally important, the researchers used a specially evolved pyrrolysyl-tRNA synthetase–tRNA pair to add the unnatural amino acid Klac—a lysine with a lactyl group already attached—to the enzyme. "Usually, you do validation by introducing mutations," explains Ye. "This approach lets the protein incorporate the exact modified amino acid, and that's definitely better than the mutation approach."

They discovered that lactylation of a certain lysine reduces the enzyme's activity, suggesting a negative feedback loop[7]. As lactate starts to accumulate, it can slow down the glycolytic pathway by modifying ALDOA and other glycolytic enzymes. "We found a new route for how end-product inhibition can be conducted," says Ye. "Usually, it's a non-covalent interaction, but in this case we found that lactylation can do the job."

Ye points out that these kinds of studies benefit from the availability of a great deal of publicly shared proteomics data. "The proteomics community is really doing a great job," she says. "These open resources will greatly benefit biologists and inspire translational research."

Old PTMs get a second look

Discovery of a new type of protein modification opens all sorts of research doors, but some investigators are taking a fresh look at previously described PTMs using state-of-the-art methods. "In the last few years, we've seen an explosion, not so much in the number of new modifications found but of really being able to characterize them well," Garcia says. ADP-ribosylation, for instance, was first discovered in the 1960s, but it has been challenging to study in the lab. For one thing, the ADP-ribose group often breaks off when the protein is fragmented for mass spectrometric analysis. Also, the size of the added side chain can range from one to hundreds of ADP-ribose subunits, so researchers had to develop a method of reliably trimming the long polymers down to a single, recognizable unit to identify modification sites.

Now, researchers are investigating the modification's function in the cell. Anthony Leung and Marc Greenberg at Johns Hopkins devised a system using probes comprising different lengths of poly-ADP-ribose, or PAR, molecules to pull out potential binding partners. So far, Leung and colleagues have uncovered about 700 proteins that bind to PAR. Not surprisingly, given PAR's resemblance to RNA, some of these have been previously characterized as RNA-binding proteins.

One, called FUS, is important in certain neurodegenerative disorders, such as Alzheimer's disease and frontotemporal dementia. When it binds to PAR, FUS forms a biomolecular condensate, a sort of membraneless organelle or cloud inside the cell[8]. It turns out that "PAR is a very potent trigger of the formation of this biomolecular condensate," Leung says. Whereas FUS and RNA bind in a 1:1 ratio, Leung's lab, together

with Sua Myong's lab at Johns Hopkins, found that "one molecule of PAR can cause thousands of FUS molecules to condense," he says. "This was a very surprising finding." In addition, large amounts of PAR will cause the condensates to become aggregates, clumps of protein more solid than condensates. Protein aggregates, of course, are implicated in a number of neurodegenerative diseases, and data suggest that PAR abundance is increased in some of these diseases. "This may have some therapeutic implications," Leung says. And FUS is just one example; PAR can induce the formation of biomolecular condensates associated with cancer and viral infection as well.

Even for well-studied modifications like phosphorylation, much remains unknown about how modified and unmodified proteoforms differ in chemical reactivity and function. Activity-based protein profiling is a method that can single out reactive residues that may be functionally important in the cell. "We started activity profiling with the goal of bringing chemistry and chemistry methods into native biological systems," says Benjamin Cravatt, professor of chemistry at the Scripps Research Institute in La Jolla, California. "Activity profiling can provide a path to understanding dynamic changes in the functional state of proteins and biological systems, and also a robust way to discover ligands for those proteins."

The technique involves using small-molecule probes that target chemically reactive sites on proteins, such as the active sites of enzymes, so even if a protein doesn't have structural homology to a functional class, the activity probe can find it. For instance, Cravatt and his colleagues identified new members of one of the largest enzyme classes, the serine hydrolase family, using activity profiling.

Activity profiling is also useful for characterizing how PTMs affect protein function. Recently, Cravatt investigated how phosphorylation changed the reactivity of cysteine residues, which are often targets for modification and can reside at protein–protein interfaces[9]. The team isolated proteins from cells during mitosis, when phosphorylation is increased on many proteins. Then they compared cysteine reactivity before and after treating the proteome with a phosphatase to strip the phosphorylation. In this way, they found that phosphorylation can significantly alter cysteine reactivity elsewhere on the same proteoform. Although some of the significant phosphorylation-dependent changes they detected occurred on proteins already known to have important functions in cell division, others may signify novel functions in mitosis.

On the brink of a new era of proteomics

Cravatt points out that one of the challenges facing the field is how to enrich rare proteoforms for functional studies. "Most of the proteomics methods are bottom-up," he says. "If there's a rare proteoform that has a hyper-reactive cysteine, and the majority of that protein doesn't, we're not going to see that if we grind it all up and look at it in aggregate." As the technology to characterize and quantify rare proteoforms improves, the new discoveries continue to attract attention and money, and the momentum snowballs. "[Top-down and middle-down] approaches are starting to get a lot more people interested," Garcia says. "It's still not at the ease level of bottom-up experiments, but people are seeing how important it is to look at all these different proteoforms, and then they want to make investments to improve the entire pipeline."

Ultimately, of course, understanding proteoform biology goes hand in hand with studying protein function. "Everybody wants to detect and assign a function to PTMs," says Kelleher. "People are right in their desire to get at the function, get at the mechanisms of biology. But we've been ignoring the fundamental truth that our biology is proteoform biology. If we ignore proteoform measurement, then we're going to make our efforts less efficient." ❐

Caroline Seydel ✉
Los Angeles, CA, USA.
✉*e-mail: caroline.seydel@gmail.com*

Published online: 25 August 2022
https://doi.org/10.1038/s41592-022-01599-9

References
1. Kustatscher, G. et al. *Nat. Methods* **19**, 774–779 (2022).
2. Smith, L. M. et al. *Sci. Adv.* **7**, eabk0734 (2021).
3. Melani, R. D. et al. *Science* **375**, 411–418 (2022).
4. Coradin, M. et al. *Methods* **184**, 86–92 (2020).
5. Delaney, K. et al. *Cell Discov.* **7**, 122 (2021).
6. Zhang, D. et al. *Nature* **574**, 575–580 (2019).
7. Wan, N. et al. *Nat. Methods* **19**, 854–864 (2022).
8. Rhine, K. et al. *Mol. Cell* **82**, 969–985.e11 (2022).
9. Kemper, E. K., Zhang, Y., Dix, M. M. & Cravatt, B. F. *Nat. Methods* **19**, 341–352 (2022).