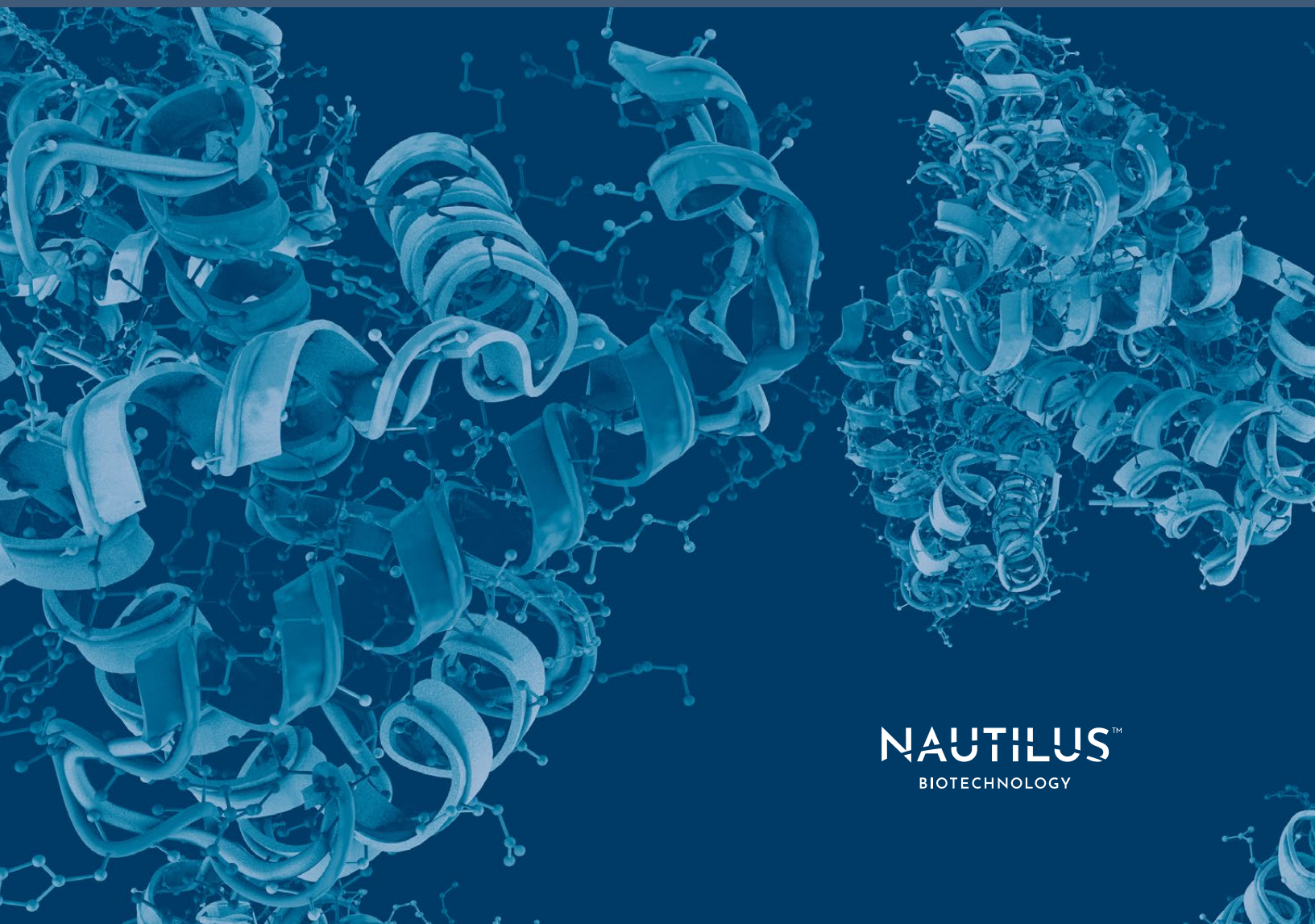
2024 Gilbert S. Omenn Computational Proteomics Award Lecture

Parag Mallick, Ph.D.



NAUTILUS™
BIOTECHNOLOGY


US HUPO
from genes to function
2024
Gilbert S. Omenn
Computational Proteomics Award
Presented to
Parag Mallick

May, 2024

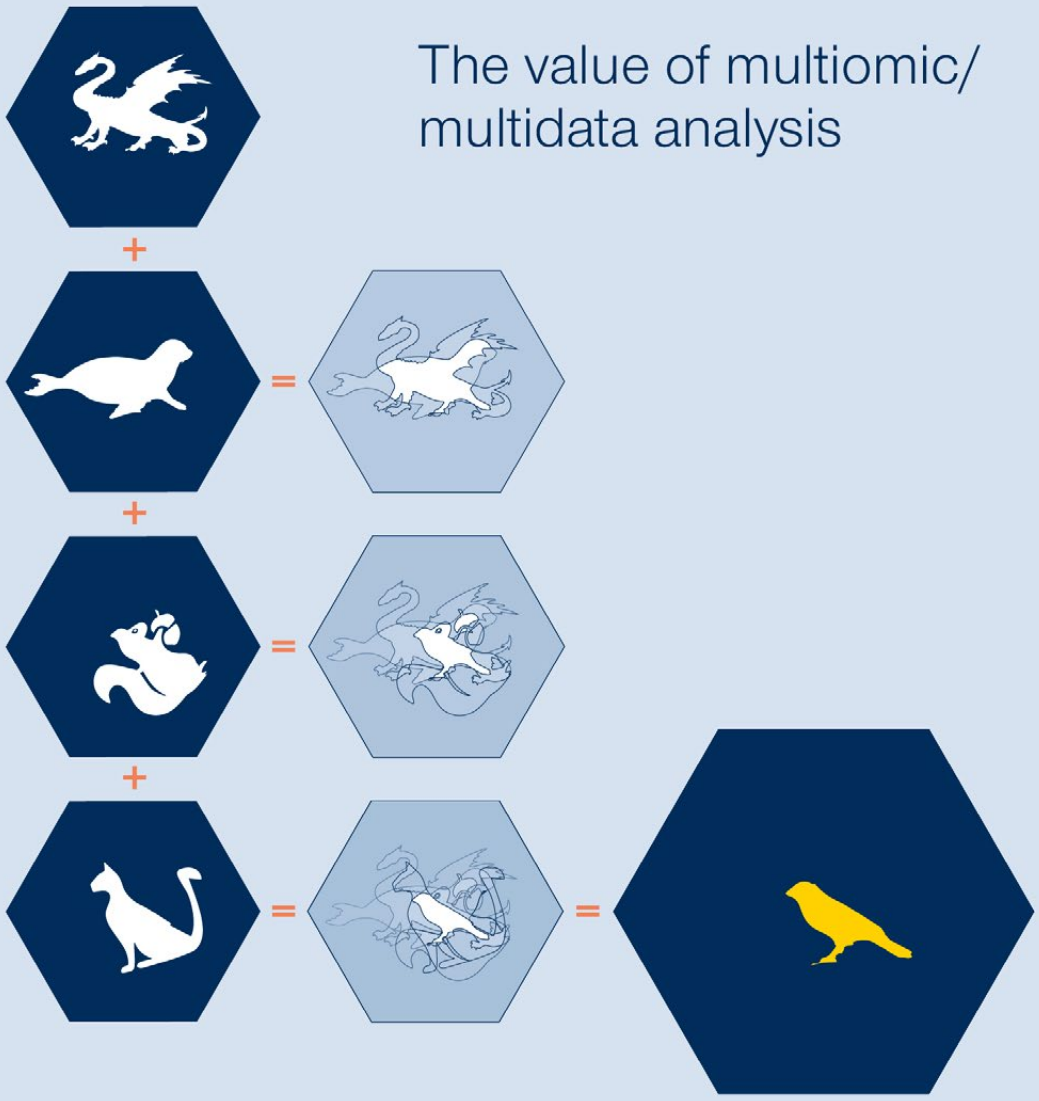
When I was first told that my longtime friend and colleague, Parag Mallick, had been named the 2024 recipient of the US HUPO Gilbert S. Omenn Computational Proteomics Award, I recall thinking that it would be difficult to identify anyone more deserving of the recognition. I knew that he would have thoughtful and insightful comments about the future of proteomic data when he presented the Omenn lecture. That certainly proved to be true, as you will see when reading this transcript.

Parag is an out-of-the-box thinker. Or rather, if you know Parag as well as I do, he is probably better described as a pull-it-out-of-the-hat thinker. (I have been "lucky" enough to be on the receiving end of Parag's penchant for magic.) So, it is no surprise that Parag's contributions to computational proteomics, including ProteoWizard, and far beyond, have foundationally and positively impacted researchers' ability to study proteomics; not just by improving the tools used to interpret data, but at the root levels of how we think about measuring proteins, defining data standards and improving core data analysis methods. His vision for the future – articulated in this lecture – will inspire all of us in the field to think about how to make data, analyses and models available to the broader community so that we can seize this opportunity, as he says, "to go much further and accelerate the pace at which important discoveries are made."

Having witnessed Parag's deep commitment and contributions to computational proteomics, I can think of no one more deserving of this recognition or someone to whom we should all pay more attention when envisioning how the data we create can be leveraged to create the greatest possible impact.

Joshua LaBaer, M.D., Ph.D.
Executive Director, Biodesign Institute at Arizona State University
Former President, US HUPO

The value of multiomic/ multidata analysis



Let us know what you think about the lecture and learn more about Nautilus

Reproducibility and robustness are essential in multiomic studies if we hope to leverage their results for the greatest possible impact. The proteomics field has come a long way in making its datasets available to the broader community. Yet, today I'd like to posit that it is not enough to simply share the data that comes from omics studies, and that we have an opportunity to go much further and accelerate the pace at which important discoveries are made.

A crucial focus in my lab is multiomics. Multiomic studies bring together many different types of data to ask and probe biological questions whose answers involve the integration of diverse regulatory processes. We believe this is valuable because the intersection of data is so much more powerful than any single kind of data alone. On their own, each type of data gives us a lens through which we can view biology. So, for instance, even when looking at the same biological system, we might look at the genome and see a dragon. We might look at the transcriptome and see a seal. Look at the proteome and see a squirrel. Or look at the metabolome and see a cat. On their own, these are great views of the data, but when we put them together, perhaps something that we didn't anticipate comes into view.

Did you all see the canary from the beginning? Be honest.

Multiomics can offer profound insights into biology, but the challenge is to ensure these insights are easy to access, while also reproducible, robust, and extensible.

This is particularly challenging for a number of reasons. First of all is just the shear volume of data. There already exist exabytes of biological data and this volume is growing exponentially. Additionally, data analysis approaches themselves can be quite complicated. Furthermore, expertise is often siloed. Researchers who are experts in proteomics data are likely distinct from those that are experts in spatial pathomic or lipidomic data.

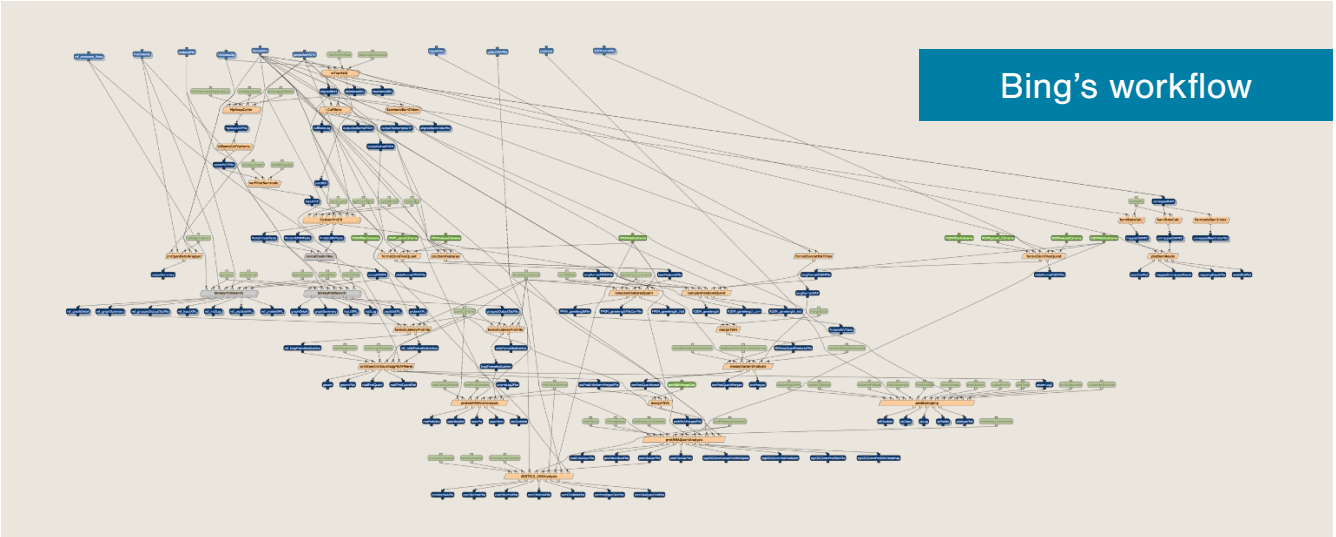
Attempting a multiomics reproduction

So in my lab, we wanted to start tackling this challenge of multiomics data analysis by examining the factors that are impediments to multiomics. We focused on not only lowering the barrier to multiomics, but also on investigating reproducibility and robustness of multiomic analysis. Our ultimate goals are to build a system to make multi-data analysis a little easier, to enable sharing of methods and, at the same time, improve reproducibility and robustness of multiomic studies.

Working with Yolanda Gil's lab, which specializes in intelligent and semantically aware workflows, we began looking at a seminal CPTAC study led by Bing Zhang. This paper involves some amazing proteogenomic work and the authors went to great lengths to make all their data available. Knowing this, we asked, "Can we do a figure for figure reproduction of this paper?"

The original work was a substantial undertaking. It took multiple bioinformaticians multiple years to do the initial analysis, but, we figured, "It's all done now, right? The data has been deposited. It should be easy for us to build something really quickly, and then click a button and reproduce this paper." The original researchers had great documentation and we had fantastic collaborators in Bing's lab who were amazing to work with. So our approach was to use a semantic workflow engine to build intelligent workflows that could reproduce the work. Along the way, we wanted to identify any friction points and quantify their impact on reproducibility and robustness.

Just for those who are not familiar, workflows are essentially highly organized and detailed assemblies of tools that you pass data through. They contain inputs for parameters, files, and applications that you'd like to run with various possible ways to run them, and the particular workflows we used are semantically aware and generalizable. So, for instance, you could do things like change out a module and say, "Okay, this could be done with many different tools. How might inputting a different tool change the results?" Or you could do something like swap RNA-seq data for micro-array data and see how that changes the results. In addition, workflows record every operation that was performed, the version of an application that was used, intermediates created along the way, and more.



We started by reading the paper and building the workflows that seemed implied by what was written in the methods section. We then wrote higher level workflows to take those methods and do comparisons between the results we produced and what was in the paper. When something didn't match up, we would go back and harass the authors and be like, "Hey, this doesn't quite match your paper. What's not there?" Then, we repeated this process over and over and over.

It is daunting to think about just how complex our multiomic workflows are. Essentially there's a piece for transcriptomics and proteomics and then lots of analyses. We were very excited when we finally put this all together and hoped we could click a button, run this massive proteogenomic analysis and reproduce the results from the paper. Yet, when we clicked that button, the result was totally different from what was reported in the paper.

In fact, we identified nearly twice as many proteins as they did. That's usually considered a good problem to have, but we were going for accuracy, not number of proteins so this was kind of terrifying.

We went back to the authors and asked questions. We discovered that some of the differences resulted from issues with the identifiers or the databases or parameters not being available and iterated. After a lot of tweaking, we eventually got to a place where we could reproduce the results pretty well. It wasn't a perfect match, but we got within say 20% and the exciting thing about having the workflow set up was that we could in fact click a button and generate all the figures from the paper.

Our workflow was also able to generate the key results figures. The key result of the paper, that patients fall into distinct proteogenomic cancer subtypes or clusters, was recapitulated. While we were at it, we generated all the supplemental figures as well. Again we were very happy to see that we could substantially reproduce the subtypes and get very similar supplemental figures.

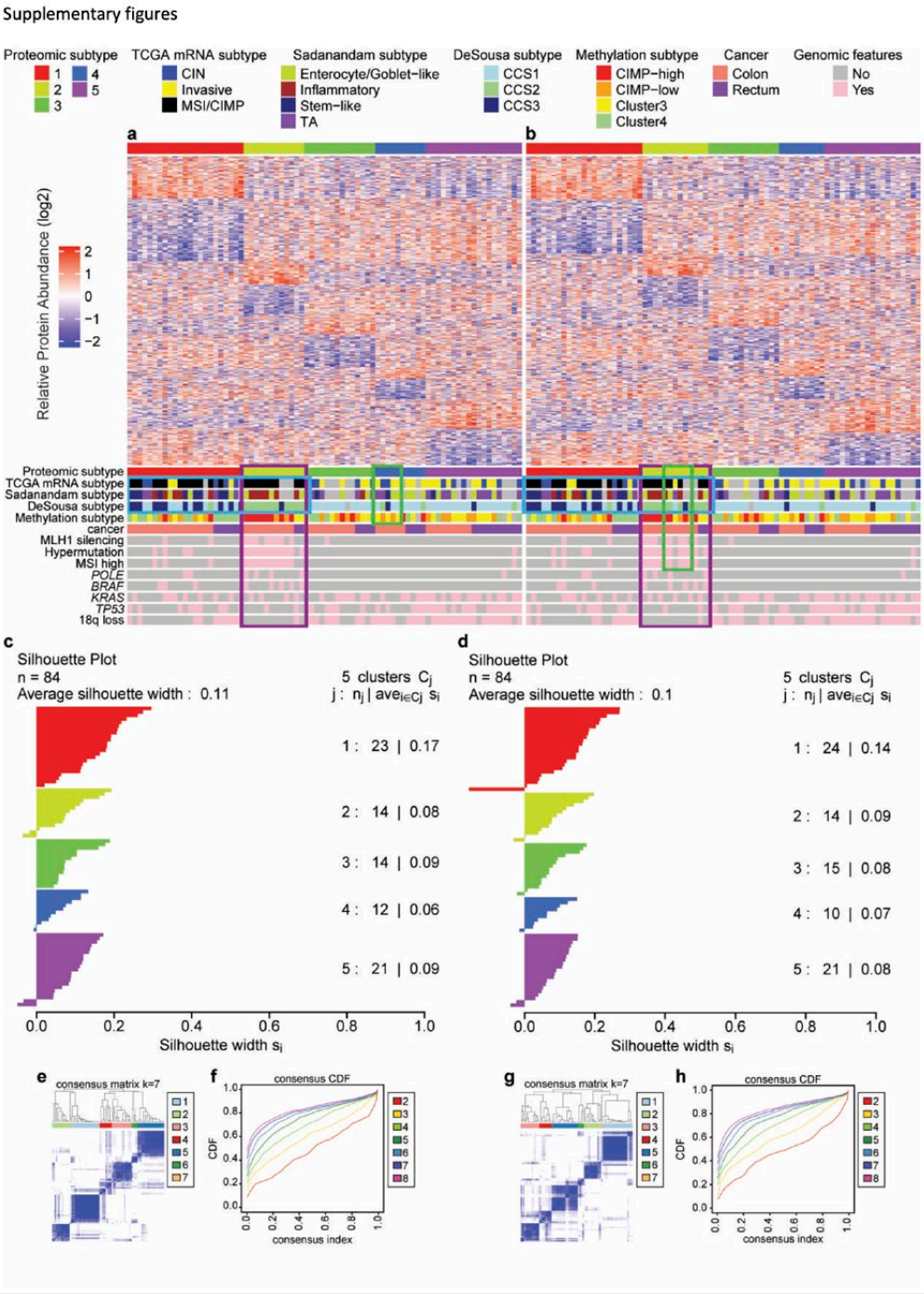
Yet, after all of this work and all these iterations, there was still 10% of the data that didn't match and we put a tremendous amount of effort into figuring out why. Well, it turns out there were fundamental differences in things like individual peptides being assigned differently. These differences had cascading effects throughout the analysis. For instance, in one case, a change of 1 amino acid in one peptide changed its assigned protein group. This then led to differences downstream that could potentially alter which patients got assigned to which cancer subtype.

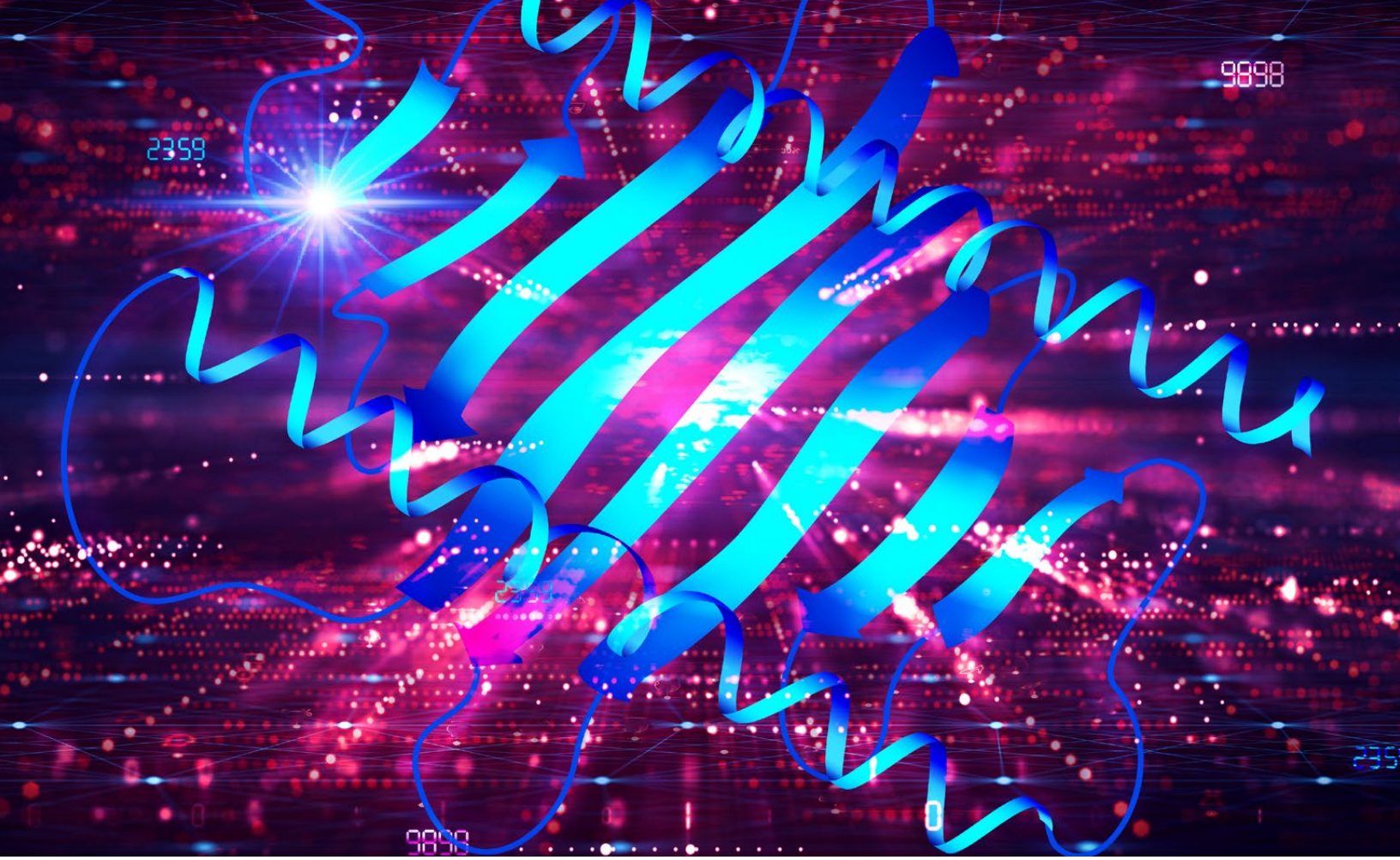
Nonetheless, I want to reiterate that one of the major findings of the study was that there were several different subtypes of colon cancer, and the vast majority of the samples realigned to the original subtypes. So even though there was a lot of chaos at the peptide level, once you rolled it up to the patient level, a lot of that variation wasn't incredibly consequential.

Let us know what you think about the lecture and learn more about Nautilus 〉

In the end, our work revealed that method sections really don't capture the fullness of the analyses. To reproduce a study accurately, we need not just the raw data files, but all the auxiliary files, the different tools that were used in the analysis, the sets of parameters selected, and versions for everything. In some cases, older versions of tools may stop being available, or may stop working in modern operating systems, so containers holding archival versions of all the pieces of an analysis need to be maintained. This is a new frontier in reproducibility that we really need to pay attention to as a field. For true reproducibility, we need to archive the entire path from question to hypothesis to data to analysis to conclusion.

Comparisons of the original and reproduced figures





The power of workflow engines

So when we think about not just reproducibility but also the robustness of a finding, the ideal findings are robust to these kinds of subtle variations. Putting this ideal top of mind allows us to think about new ways to do analyses such that we report which findings are the most robust and have the best strength of evidence. And then, you know, report the other ones too, but clarify what their strength of evidence is.

Thinking more deeply about robustness, one of the questions we can ask with a workflow engine is, “How sensitive are the results to the particular tools used?” When we change search engines for instance, we get different proteins, but does that extend all the way to the final results? One of the ways we can look more deeply at this question is to leverage our workflows to analyze new data with new methods. What was so great about this CPTAC study is they analyzed about 100 patient samples and then a few years later they collected 100 more. So, looking at the two groups and using our workflows, we could ask, “Do the results hold up?”

When we ran the workflows again, what was really interesting is, as we started with the initial cohort of samples and then added in the new samples, there were points where you saw the patient clusters jumping around. Some of these new patients jumped to other places. Yet, ultimately, once you added all the new patient samples, they actually did recapitulate the original 5 subtypes or clusters, which was really exciting.

Taking this together, I think we have an opportunity in the field to think about these workflow engines as a fundamental layer that we can add to our studies. These workflows can then be reused, shared across labs and leveraged to examine the robustness and reproducibility of even just one study. That is, we can use workflows to assess how sensitive a study’s results are to shifting parameters in the workflow. Without workflows, that kind of sensitivity analysis is hard to do, but with them, we may ultimately get to the point where we can continuously reanalyze data.

A robust, reproducible, and AI-enabled future

With that, I know I'm almost out of time but I'm going to end on a dream for the future. When we think about the future of science, we think about asking questions. Often, we start with a hypothesis, then we formulate a set of experiments and computational analyses to test that hypothesis. Later, we collect data out in the wild, test it to get a sense of it, and then repeat the process refining our hypothesis and analysis methods, collecting more data, testing again, and so on around this loop.

This iterative process is unfortunately very arduous, but I'd like you to imagine a world where it is not a heavy lift. In this world, it is instead easy to continuously accelerate that hypothesize, test, evaluate cycle, and we can quickly run the cycle 100s or even 1000s of times.

In that world, we can take advantage of emerging AI tools and computational workflows. In that world, scientists at the bench have a hypothesis, they chat with a little interactive discovery agent and say, "Hey, I've got a hypothesis and here's some stuff I think we should think about." Then, the interactive discovery agent goes, formulates lines of inquiry, sends some of them off to a cloud lab to have the experiments done, and everything is tracked through the workflow system.

Later the discovery agent comes back and says, "Hey, here's what I found." Then maybe every week it monitors the science surrounding the inquiry. It notifies the scientist when a new paper is published, determines if it can incorporate the data into the established workflows, grabs the data, does the appropriate stats, puts it into the system, and sends the scientist back the summarized results in an email detailing how the new data supports or refutes the scientist's previous findings.

In this world, science evolves into a continuous process of refining and evaluating hypotheses, not in discrete increments of graduate students, but instead, in a more continuous and ongoing manner.

In this world, we think about sharing information not just as sharing data, but also sharing hypotheses, models, and supporting material. With this kind of sharing, research can evolve into a system where we think at the level of the biology, not at the level of the individual data elements.

So, how do we get to this imagined future? Well, step zero is to create integrated standards across multiomic tools. Today we have a lot of support for standards in mass spec, and don't have standards integrated across all our tools. They're coming. We need to anticipate how we will integrate all these different data elements together and then start building towards repositories of containers, workflows and models.

And I'm going to stop there because I think that's already a lot to do, but I would like you all to think about a world where we can take advantage of these sophisticated computational tools and AI resources to vastly accelerate the pace of science and get excited for what's coming.

Let us know what you think about the lecture and learn more about Nautilus 〉

SELECT QUESTIONS FROM THE US HUPO COMMUNITY

Let us know what you think about the lecture and learn more about Nautilus

Question 1

In the work you presented, you had to scale back your analysis to better match the result from the original CPTAC paper. My lab has the same issue. We analyze data when we get it and then spend years interpreting the biology, but, by the time we're ready to publish, the original data analysis is way out of date. How do you incorporate this problem into your thought process?

ANSWER 1

In our current approach, a bioinformatician writes a script, analyzes the data, and then we go off and do all the interpretation and don't come back to the analysis until several years later. If, instead, the whole process was done in the context of workflow systems, the lift to reanalyze the data or incorporate new parameters would go way down.

So, the first thing we have to do is think about this as a continuous process as opposed to a serial process and create the infrastructure necessary for us to carry out that continuous process. This will make it possible for us to easily do things like pull in new FASTA files and update parameters. Then, we need to think about our reported results as what they are - snapshots in time. We need to be able to report the most recent version of our results while being okay with the fact that science evolves and findings evolve. This is good and normal and does not mean that we're wrong or lying. We're just capturing a moment.

Question 2

It's also important to think about the psychology of researchers. Students tend to do things the way they were done before because they assume that was the right way to do it, and people do not like to start from zero.

ANSWER 2

Yes, so I think there's a tension there that's really interesting. It's the tension between building on what's been done before versus innovating around it. My perspective on that is we should have the ability to grab a workflow off the shelf and repeat what was done before if that's what we want to do. Yet, we should also have the ability to intentionally say, "I want to poke around and futz with the old workflow. I want to change parameters and intentionally bring new knowledge to evolve the way that I analyze the data." Then we can run both the old and updated workflows and see how they change the results. These kinds of things happen sort of accidentally or implicitly now, but I'd like us to have a system where we can do them intentionally and explicitly. Additionally, by having a meta-layer on top of the workflow engine, it allows us to benchmark workflows against each other more easily to see if that arcane workflow from back in the day was actually better.

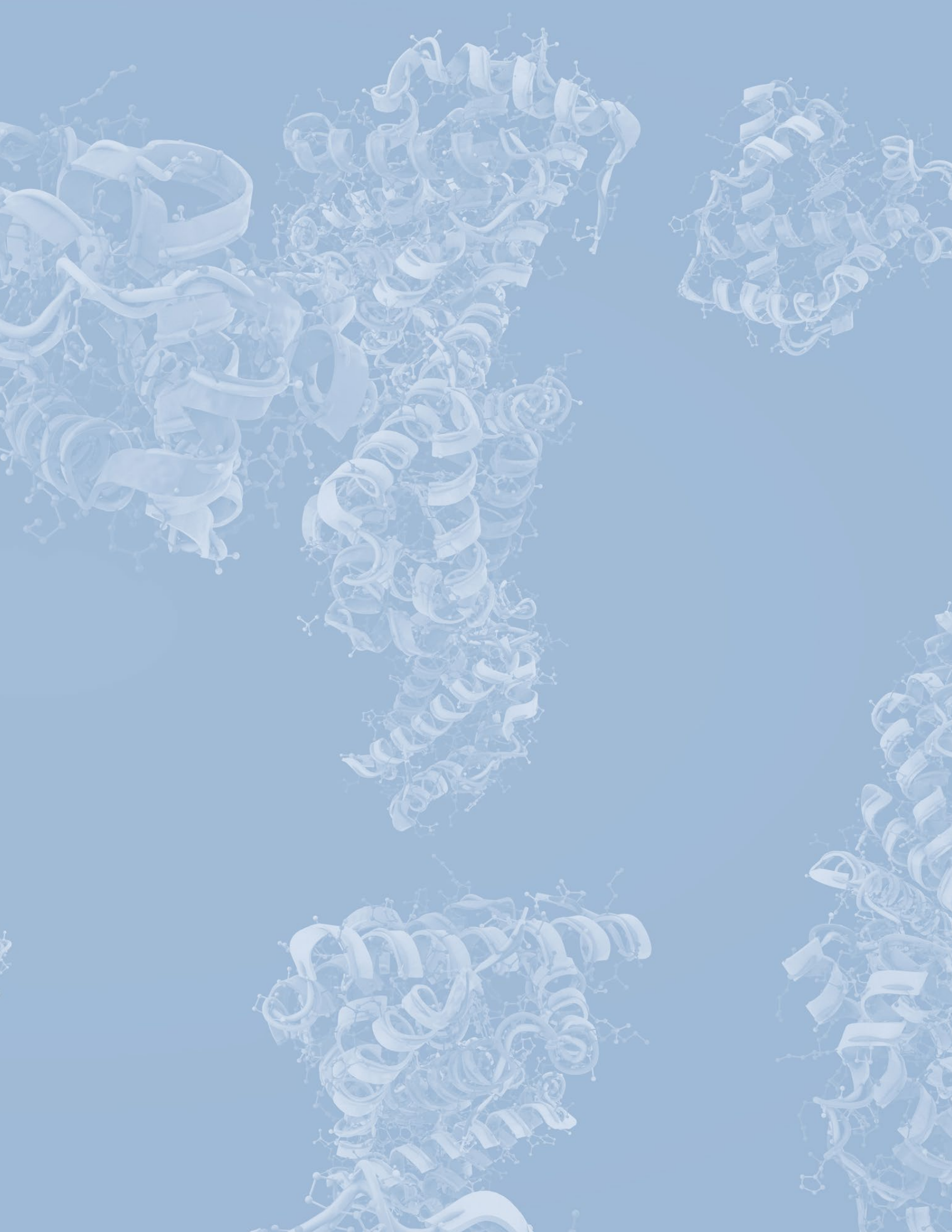
Question 3

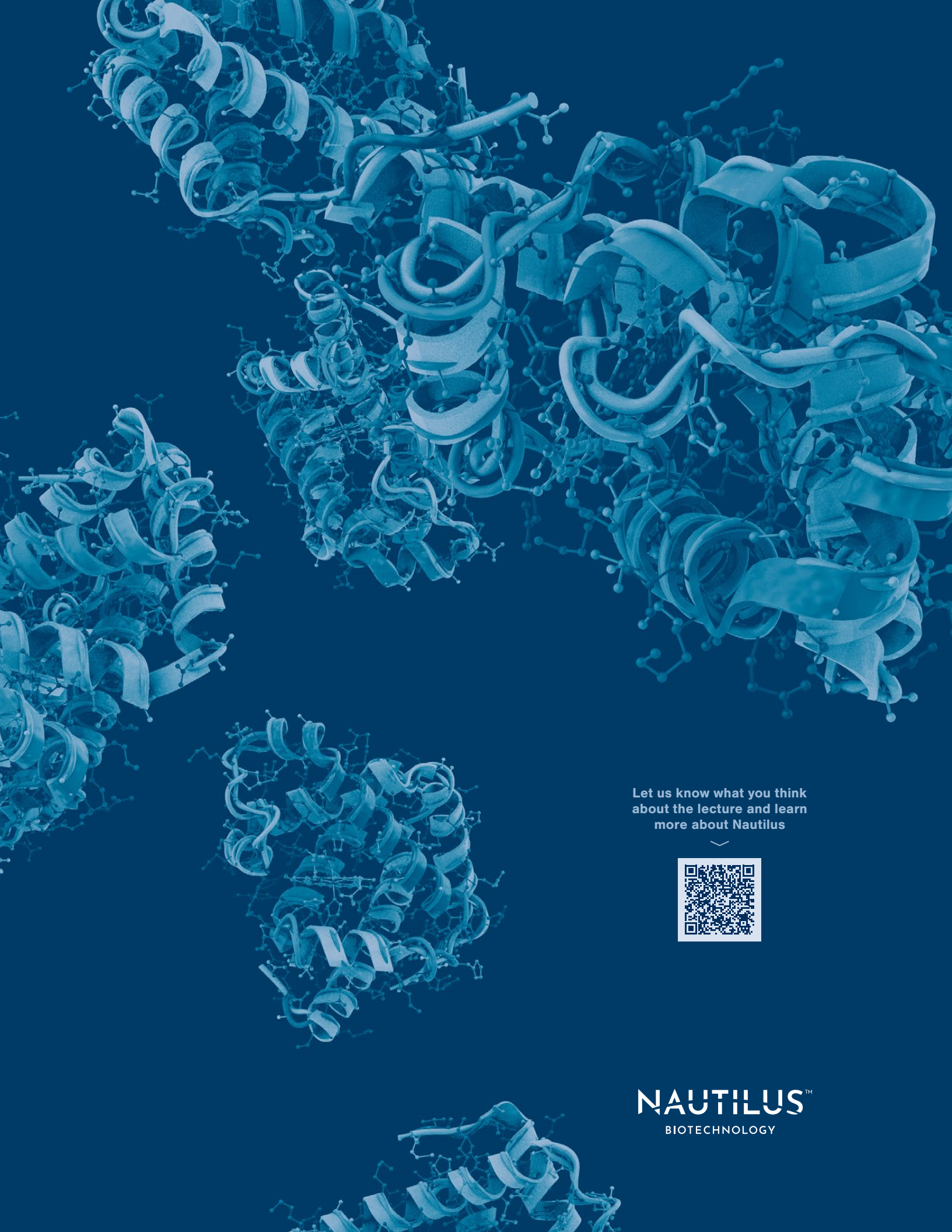
In the past you wrangled instrument vendors to develop data formats that were accessible to all, but now the people that you have to wrangle are researchers. So, I'm curious, what combination of evangelism and rewards will help us precipitate the adoption of the practices you outlined in your lecture?

ANSWER 3

What I noticed early on is that we get a ton of power from combining data from a lot of different places. Currently, we lose power because we throw data into an archive and then someone picks it up and runs with it without much involvement from the original team of researchers, but I don't think this has to be the case. With accessible workflow systems, I think we can have the original team of researchers be partners on re-analyses and thereby make their chain of influence larger. Additionally, we can actually recognize that a dataset from X was used in an analysis – and that could be something that goes on their CV!

So, really, the opportunity here is to recognize that our local, small experiments are better when they incorporate all the available data through these workflow systems. Then, thinking about our experiments in the context of everything else that is out there also helps us recognize that, once our experiments and data are placed in context, they can contribute to all the other work being done. Thus, making all our data and analyses available as part of workflow systems vastly increases their possible impact.





Let us know what you think about the lecture and learn more about Nautilus

NAUTILUS™
BIOTECHNOLOGY