©revvity/shutterstock

New modalities, new tools:

choosing a flexible digital platform for drug discovery

Produced by Europa Science

SCIENTIFIC COMPUTING WORLD

revvity signals

Contents

Introduction

With small molecule research being the dominant modality for decades, many legacy data platforms are struggling to cope with the demands of newer modalities, such as RNA therapeutics, gene therapies, ADCs (antibody drug conjugates) and protein degraders.

For drug discovery teams, these new modalities are an exciting prospect, but multiple modalities throw up a range of challenges around data management.

At a recent roundtable discussion, hosted by *Scientific Computing World*, leading drug discovery experts discussed the data impact of working with these multiple modalities, covering areas such as data standards, ontologies, cultural challenges and software needs.

Roundtable participants

Simon Andrews, Head of Bioinformatics, The Babraham Institute

Mattias Bood, Associate Principal Scientist, AstraZeneca

Irene Choi, Senior Director, Head of Drug Discovery, Verge Genomics

Nick Lynch, Director, Curlew Research

Evelien Micholt, Operational Lead Discovery Sciences, Galapagos

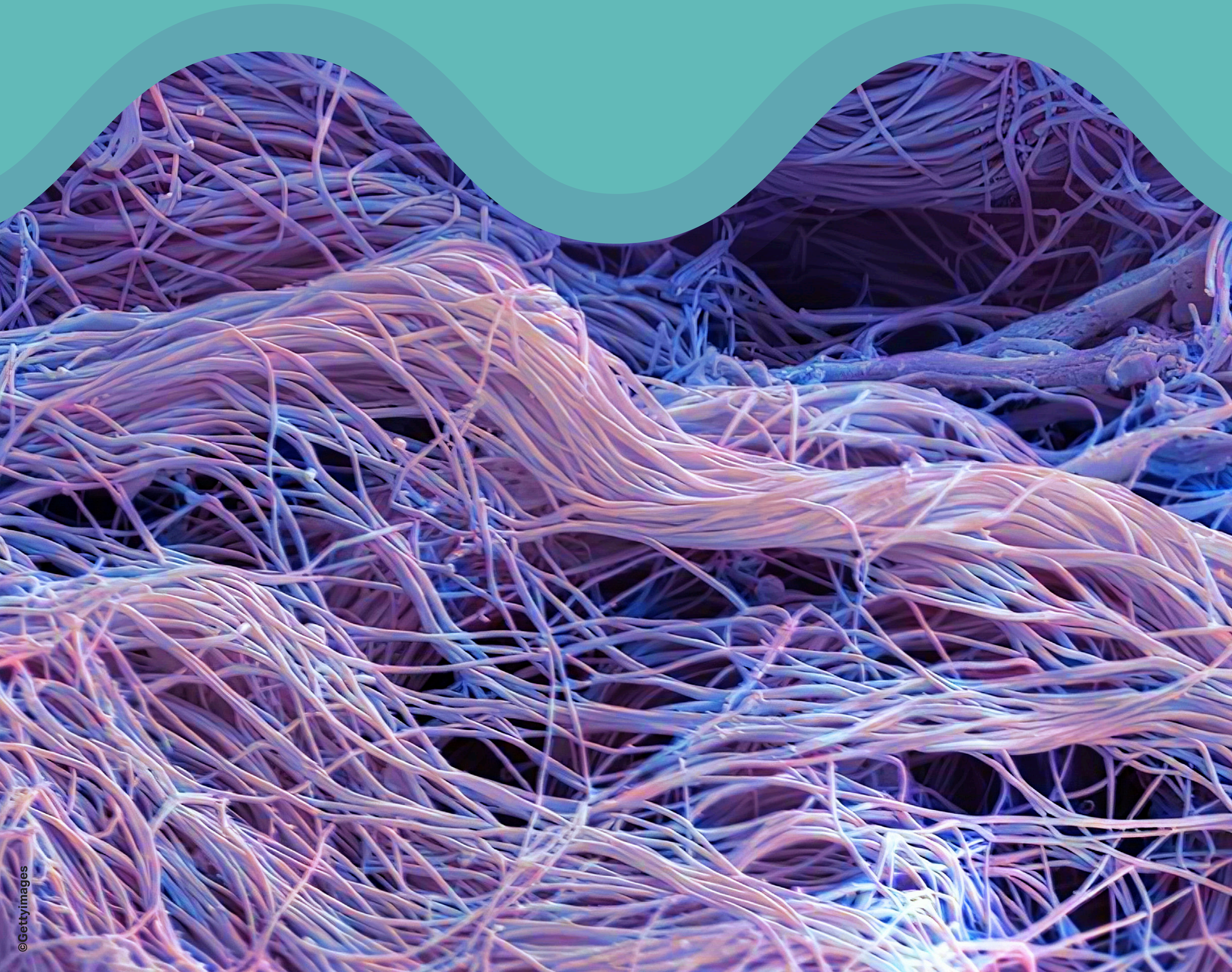
Roberto Olivares-Amaya, former VP Omics & AI, LifeMine Therapeutics

Sarah Sirin, Director and Head of Computational Chemistry, Remix Therapeutics

Nicolas Triballeau, Director, Drug Discovery, Revvity Signals

All panellists in this discussion participated independently and any comments should not be interpreted as an endorsement or recommendation for the products and services of the sponsor.

Managing data issues with multiple modalities



©Gettyimages

"We're still debating whether we want to go to the time and expense of doing a more formal integration"

Simon Andrews, Head of Bioinformatics, The Babraham Institute

The search for viable targets and drug candidates is expensive and hugely competitive, with pharmaceutical companies investing millions in talent and tools to deliver the next major medical breakthrough.

In recent years, that search has begun to cover new areas in the form of modalities such as RNA therapeutics, gene therapies, ADCs (antibody drug conjugates) and protein degraders. Recording, storing, retrieving and interpreting data from these multiple modalities is making the data challenge ever more complex.

Simon Andrews, Head of Bioinformatics at the Babraham Institute, based just outside Cambridge, UK, says the complexity is causing his team to re-evaluate its data approach from the ground up.

"When it comes to dealing with data from multiple modalities," he says, "we've been 'thinking about' this rather than 'doing' a lot of it. We manage data for a whole series of data-generating facilities across multiple platforms, such as sequencing proteomics, flow cytometry, imaging and more. While these data platforms have historically been quite siloed, we're now refreshing our data management approach and trying to decide how integrated we want them to be.

"People are certainly designing studies that cross those modalities. There is a big trade-off: one can build a formal infrastructure that formally links bits of data we hold in those different centres to the point that we can track that automatically, or at least lay the groundwork, so that if somebody wanted to do it in future, it would be easy – or at least feasible – to link those things together. On the other hand, it is a huge effort to build, establish and run those platforms on behalf of the people creating the data and the samples, as well as on the experimental side to annotate the data to a level where it is feasible to be shared.

"At the moment, we are getting people that are running those experiments, but they are very much using the individual facilities to generate the data. The integration is not happening until after it has passed out of the facilities, then it's being processed and filtered to deliver a completely bespoke downstream analysis. That works okay, but it doesn't scale so well. We're still debating whether we want to go to the time and expense of doing a more formal integration."

"Getting the right data strategy from the beginning is something that can give you great advantages"

Roberto Olivares-Amaya, former VP Omics & AI, LifeMine Therapeutics

Galapagos, a Belgian pharmaceutical research company, initially worked on small molecules, but has since adopted cell therapy. Evelien Micholt, Operational Lead Discovery Sciences, says the company thought carefully about how to treat data from these distinct modalities: "We made a distinction between the data of the modality creation versus the data of the actual experiments done with modalities on any form of disease model. The different modalities have their own dedicated master system for registration and inventory. The experiment data is stored in a joint system, enabling users to compare data of different modalities."

Dealing with data from two distinct modalities is also a challenge that Roberto Olivares-Amaya has faced in his previous role as VP Omics & AI at LifeMine Therapeutics. "Our omics annotation was very robust," he says, "as it took multiple versions to get it right. To get mass spectrometry data in as soon as possible, we faced the challenge of data ingestion and integration with the rest of the omics data. It provided an opportunity to ask how we could make it useful for scientists, both on the data side and the experimental side. We needed to bridge this analysis together from a software and data perspective in a way that wasn't bespoke, so that we could scale it and run it on a continuous basis.

"Getting the right data strategy from the beginning is something that can give you great advantages, but it's hard to think in advance about what other new modalities are coming in as you're trying to find new therapeutics."

At Verge Genomics, a biotechnology company based in San Francisco, USA, the team is trying to align across different types of modalities, as Irene Choi, Senior Director, Head of Drug Discovery, explains: "It's complex, because if you look at small molecules or antibodies or other modalities that really engage a biological mechanism, they often follow a similar pathway, but not necessarily in the same way. You try to find the common threads, but they don't necessarily replace one another. It's about teasing apart those differences, while also looking for the threads of commonality. You then need to apply what modality could be the most effective in terms of what you're trying to achieve.

"It is important to annotate the data sets that go into whatever you're using properly, to then drive your predictions.

"The additional component of time plays a crucial part in all of this. For example, neurodegenerative disease is an age-related disorder, so the phase of the disease when your data sets are captured affects what your data will show and can present a completely different picture. When you overlay this data on top of different modalities, you have a giant puzzle where you're looking for commonalities."

"To go back and compare old data with new data is not always so trivial either, especially for programs that might span back very far in time"

Mattias Bood, Associate Principal Scientist, AstraZeneca

From a big pharma perspective, such as that experienced by Mattias Bood, Associate Principal Scientist at AstraZeneca, budgetary restrictions might not be such an issue. "We perform very thorough modality assessments," says Bood, "looking at every single piece that we can think of to get to the biology we need for a given target. Of course, we can afford to be agnostic in that regard, given our resources, which is very exciting from a therapeutic point of view. It also facilitates both new and novel combinations of various drug modalities."

Budgets might not be an issue, but data challenges remain. "We have decent ways to compile the data from these multiple modalities and make it fairly accessible but, in most cases, it's bespoke," continues Bood. "It's siloed in different IT systems and, sometimes, different departments can, at times, be difficult to piece together. To go back and compare old data with new data is not always so trivial either, especially for programs that might span very far back in time."

Remix Therapeutics is a small biotech working in small molecule drug discovery – an established modality space. However, the company is interested in RNA processing, so it does generate different types of data.

Sarah Sirin, Director and Head of Computational Chemistry, says: "Sometimes we use novel modalities such as ASOs (antisense oligonucleotides) as tool reagents and those require different registration and data collection considerations, etc.

"We often generate two types of data: large volume data sets across hundreds of different targets using various sequencing technologies, which can be very noisy; and, separately, really high quality robust data as part of an LO (low observation) programme or hit-to-lead programme. We have been agile in creating different data repositories to store these types of data, but we are mindful with the integration of those data sets. A project chemist doesn't necessarily need to see all of the high (noisy) content.

"To make the data challenges a bit easier to overcome, we have an agile AWS (Amazon Web Services) environment outfitted with open source tools. We can create an SQL query and pull out bespoke data sets, depending on the questions that we are asking. We're not a huge company, so the AWS resources and open source toolkits make data communication easy from our perspective. In the early stages of company building, we don't always think about all the usability scenarios five years down the road, because five years is a long time in the biotech world, but that creates other problems, such as sustainability."

"One also needs to consider what will provide the greatest value to both the data producer and the data consumer"

Nick Lynch, Director, Curlew Research

As a consultant working with several clients over many years, Nick Lynch, Director at Curlew Research, is well positioned to advise on what teams need to consider when working with multiple modalities.

"We like to use the data life cycle as a way to make decisions around whether to use general data workflow tools versus specific tools," he says. "We think some of the challenges are around, perhaps, the registration and uniqueness, because you may need different tooling for a small molecule versus some of the other modality types. Then it's trying to work through where you need to compare data across modalities versus where you don't need to do so. We try to use the business questions that people want to ask to help shape what the data needs to look like and how it should be stored.

"One also needs to consider what will provide the greatest value to both the data producer and the data consumer, using ontologies, for example."

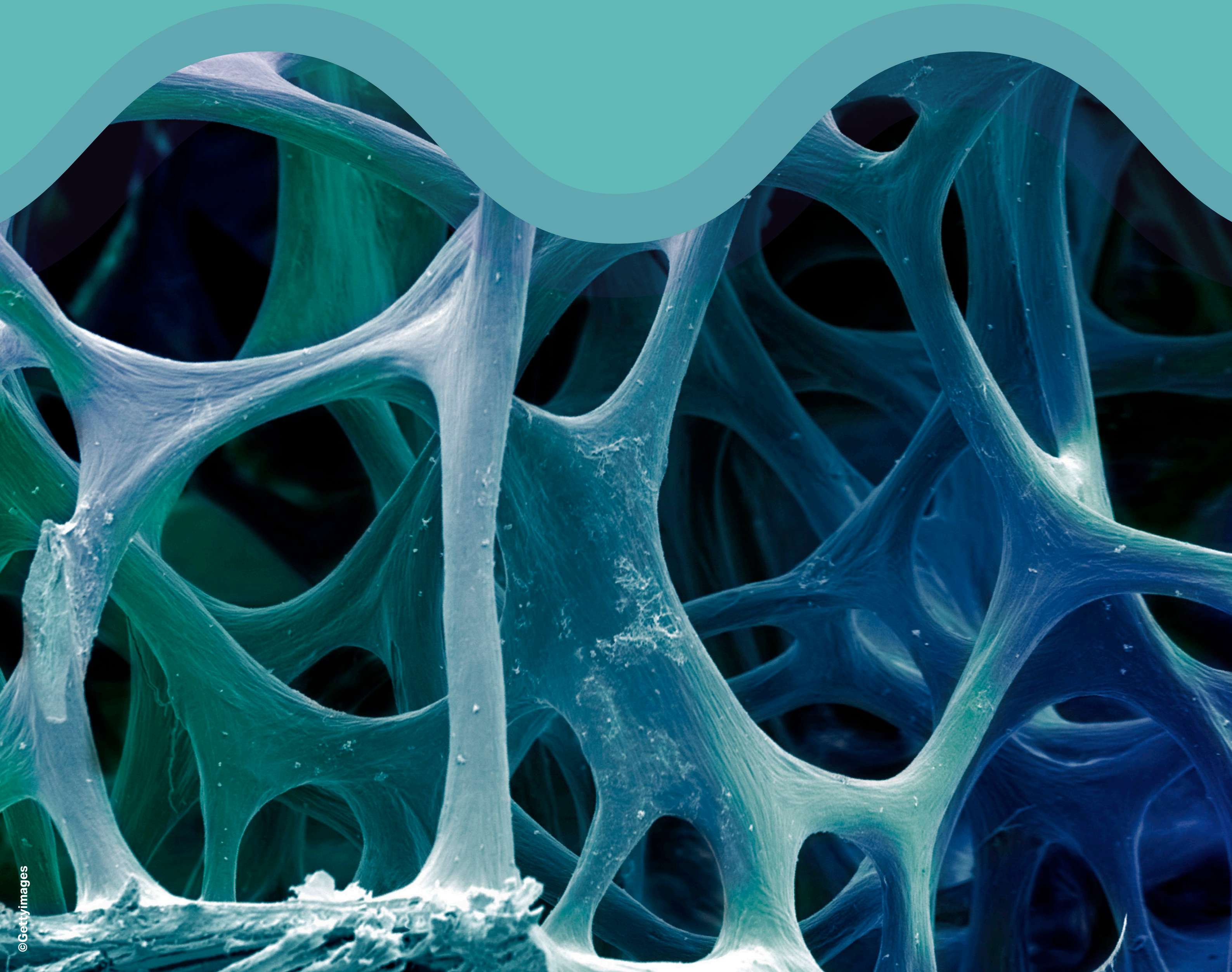
Nicolas Triballeau, Director, Drug Discovery at Revvity Signals, previously worked as a chemist at a European biotech. "We had very interesting discussions with colleagues about these different silos," he says. "They were operating in their different labs and we were trying to understand what we shared in common between, for instance, people who were working with small molecules and those working with antibodies, or with cell therapies, or with ASOs. One of the critical common aspects is when we do biological experiments on these modalities.

"We created an experiment data model, which we then turned into an ontology. The idea was to find a way of describing an experiment at a high level, such as the environment with its temperature, or describing all the different entities that come into play in an experiment.

"For instance, we called the modality the perturbagen, making it completely independent from whether it was a small molecule or an antibody. Then we had the reactive system, which could be as simple as a protein, a cell or a whole animal. Finally, we had a precise description of the measurements that we were doing during the course of these experiments – completely agnostic of the modality.

"This enabled us to create this high-level, common, generic tool, but, of course, we still needed specialised tools to capture the uniqueness of these entities. That solution would not be the same for small molecules as it would be for cell therapies."

Decision-making on data structures



©Gettyimages

"If scientists are taking a separate copy of centralised data to work on, you've just doubled your storage costs"

Simon Andrews, Head of Bioinformatics, The Babraham Institute

With the additional complexity that working with multiple modalities creates, decision-making on how data should be structured becomes ever more important.

At the Babraham Institute, teams attempt to pull together the various stakeholders that have an interest in that decision-making.

"They often have priorities that conflict with each other," says Andrews. "It's not generally the research groups generating the primary data – there's normally a central facility that runs big mass spectrometers or sequencers doing that. The facilities themselves care very much about the process of actually running the equipment and managing the facility and dealing with their modality. They're less concerned about what anyone else is doing. They just want their part to work efficiently.

"The challenge for us as an institution is making sure that we are only storing what we have to store, that we store it efficiently and that we can find it again. The costs for storing this data are enormous. If scientists are taking a separate copy of centralised data to work on, you've just doubled your storage costs. So, wherever possible, it's important to have a method of presenting the data that allows people to manipulate it without duplicating it – that has a huge financial incentive.

"There are also the technical limitations of the system you use; some pipelines can pass data straight into a database, while others require you to go through a proprietary, licensed Windows program before they are in a usable form. There is an awful lot of plumbing to make all of these different bits connect together well enough to present any kind of unified view."

Inevitably, teams want to interrogate and, in some cases, manipulate data, leading to issues with version control. This is something that Micholt's team at Galapagos has considered.

"We have certain protocols around our experiments and related data," she explains. "For example, up to a point, experiment data remains comparable, but – from a certain iteration or small change – that data set is no longer comparable. We try to capture that by having an assay catalogue, in which we actually define all the parameters of an experiment and expected values that are in or out of bounds, and also identify whether this experiment data is comparable to a previous version or not. We can then use any new version of the data to either feed into a standard model, so it can be averaged, for example, or it could be seen as a new experiment with a standalone data set."

"The collaborative effort should also have the user interface in mind, as users want to be able to access the data they need quickly"

Irene Choi, Senior Director, Head of Drug Discovery, Verge Genomics

For a smaller company, such as Verge Genomics, the 'start-up mindset' creates flexibility that, perhaps, larger, more established institutions may not have. "We can self-organise and imagine how the structure could be so that it can be the most efficient to all the end users," says Choi. "We really pride ourselves on having designed our own platform and developed our own data sets, procuring a bunch of patient tissue from various repositories. One of the beauties of creating your own data sets is having the liberty to structure that.

"We have a head of computational biology who does the prime lifting in terms of how data needs to be organised and structured, but it's also done in collaboration with all of the scientists in the room. On my drug discovery team, I have a bunch of bench scientists that generate data, but then also use the data, and then feed that back into our database to generate the next set of targets or validate the initial target predictions. There needs to be this iterative process in which all the parties in the room can have usability.

"For us, everyone's in the room when we are designing a bench study that would generate data that would go into our database; similarly, when we are obtaining data sets from external sources, the scientists are in the room to be able to give some insight into what role that data set could play and how relevant it might be to the other data types that we may have.

"This collaborative effort should also have the user interface in mind, as users want to be able to access the data they need quickly. You can't do that if your data set is not organised or well connected."

Sirin agrees that small companies have the agility to learn and adapt as they go. "We might start out working on some new methodology with a template that was pre-agreed on, based on the data set that you have at that time," she says. "What we've done at Remix is often go back and learn from the initial rounds of data sets even years after.

"We then go back to the data curation templates and change the way that we publish data or change the parameters or the metadata that get exposed to computational chemists or data scientists. Having the agility to go back and republish or revisit some of the data sets, or reanalyse the raw data using different parameters, is key to having homogeneous data sets that we can then use for knowledge extraction."

For Olivares-Amaya, that cross-discipline approach is vital. "It's important to bring both data scientists and experimentalists together to figure out how the data is going to be used and iterated to make decisions," he says. "You then let a research informatics team decide on the details of storage, availability and so on.

“It’s important to bring both data scientists and experimentalists together to figure out how the data is going to be used and iterated to make decisions”

Roberto Olivares-Amaya, former VP Omics & AI, LifeMine Therapeutics

“The other thing to think about is the value of the data itself: is it going to be just for experiments or is it going to be part of the overall value of the company? At my previous company, for example, we needed a more robust infrastructure as we had a lot of fungal strains that were essential to the value of the company – we had to fully sequence those strains. On the other hand, we had experiments that were part of the decision-making process week-in, week-out to find new small molecules from fungi. Those two parts required different speeds of decision-making and therefore needed different approaches.”

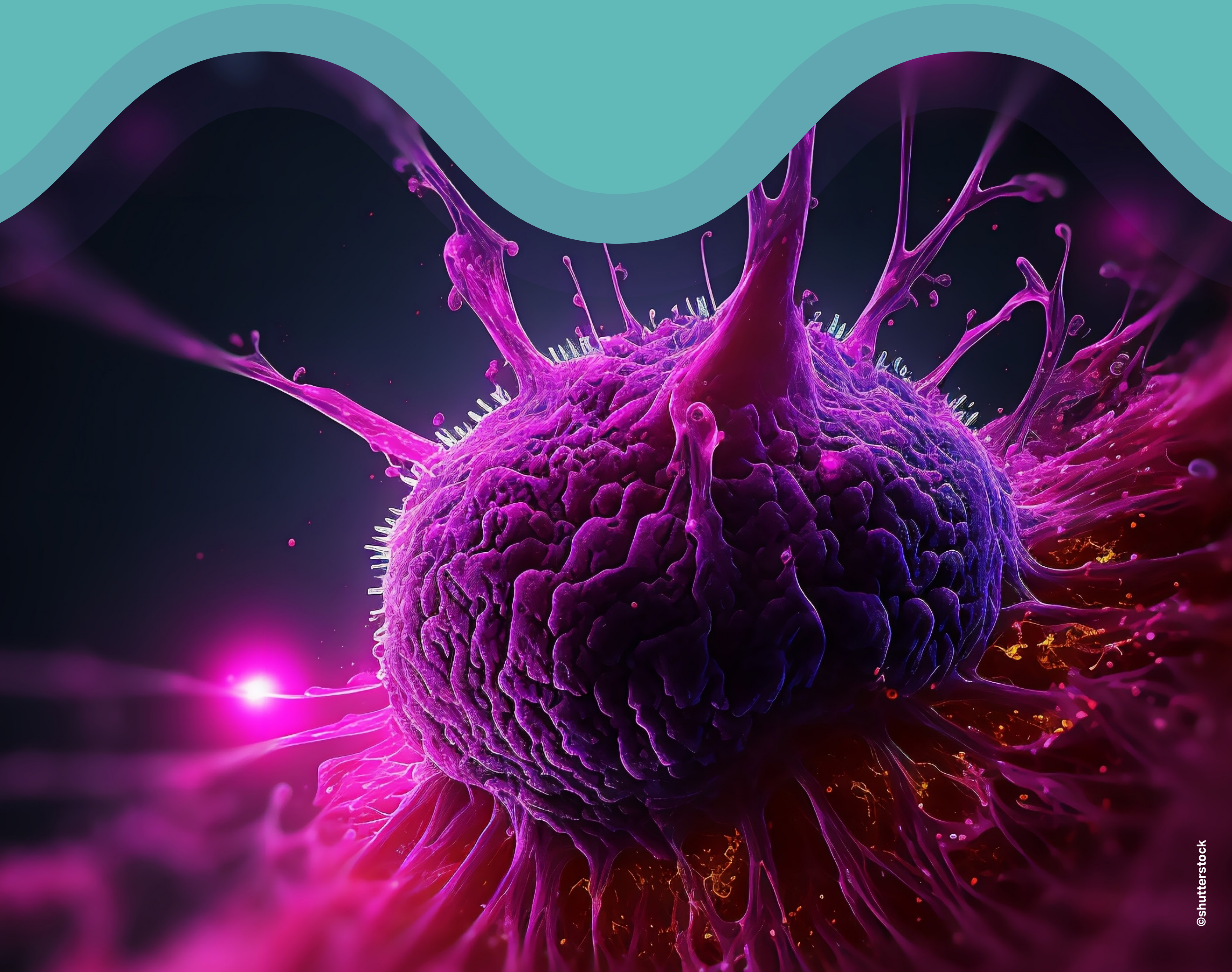
Revvity’s Triballeau adds that what’s discussed in those stakeholder meetings is important and can also help set some ground rules. “It might include asking what is our reference list of genes, our reference list of proteins, our reference list of cells and so on,” he says. “Deciding to go with a particular standard means you can implement an ontology management system that everyone has to use. For instance, when a scientist needs to report something about a certain species or a certain gene, it’s never a free text field that needs to be filled in – you have to select from a drop-down menu. That ensures harmonisation of data.”

At the other end of the scale, in big pharma, Bood says there are pros and cons. “We have a controlled vocabulary for everything,” he says, “ranging from which solvents you pick for our organic chemistry reaction to, basically, building our entire own Helm language. All this control helps facilitate training AI [artificial intelligence] and ML [machine learning] models.

“For early-stage companies in the oligonucleotide business looking to hit a target, it’s relatively straightforward to do sequence searching and alignments and generate early hit series, but a few years down the line, there might be a big risk that you will have a difficult time tracking exactly what modifications those particular early hits had, or how particular sequences related to assay data. At the very least, without stringent control and controlled vocabularies with clear experiment designs and recording of outcomes, training ML on your datasets will be incredibly challenging. For new modalities, we have done a lot of ground work to get it right with these pieces in the early days.

“With the advent of machine learning and neural network approaches – and in an environment with tens of thousands of compounds and associated sequencing – it’s not easy to map these things out.”

Encouraging collaboration among software platforms and silos



©shutterstock

“You need to do some decision manangement or tracking in a structured manner”

Evelien Micholt, Operational Lead Discovery Sciences, Galapagos

Where organisations are involved in multiple modalities, not only can this mean siloed teams, but also software platforms that are operating entirely independently of each other. Getting all of these to talk to each other is a significant challenge.

“Every discipline has its own language in which it communicates,” says Verge’s Choi, “so it’s about finding that common ground on which we can all understand each other.

“To try to address that, we use a project diary in which all of the parties that are involved can look at, contribute to and ask questions. It also captures the history of how a hypothesis or an objective has evolved over time. As data gets generated, your objectives can change or you can have a different purpose to which you want to apply that piece of data and analysis.

“We also ensure all members contribute; sometimes people can shy away from giving their two cents if they’re not the expert in the room, but still their opinions matter because they’re the end user.”

Olivares-Amaya agrees with this. “Creating an environment where everyone can speak up, no matter the hierarchy, is key to gaining a full understanding of data structure and usage,” he says.

Again, with small companies, there are aspects that make this easier, while others are more difficult.

“You do need to do some decision management or tracking in a structured manner,” says Galapagos’ Micholt. “We’ve achieved this up to a point, in that we now have documents, but they are not very queryable.

“As a smaller company, we may have an advantage in that all targets are mostly linked under a single project lead, who is aware of all types of modality attempts happening within that project. I can understand that would be a problem at scale, where you have several project leads on the same target with different modalities.”

“It’s certainly simpler at a smaller company,” says Remix’s Sirin. “We have an open book policy: everyone has access to the LIMS [laboratory information management system] and all of our data visualisation dashboards.

“Our clinical data gets siloed and firewalled for obvious reasons, but in terms of preclinical research, it’s open books.

“Nothing beats coming together as a team and having regular check-ins, so we’re not falling into various pitfalls with the data. We sit down with the data scientist or the person that generates the data and we try to extract that knowledge and incorporate that into the various dashboards with caveats and warnings if there need to be.

“Software platforms do help significantly with trying to prevent various data silos. With that being said, not everyone’s going to know how to use Spotfire or Vortex or whatever tool it may be. When I was at a big pharma, one of the things that it did well was putting a search engine that sits on top of all of the data. I find this a little challenging to enable that in a small company with a small IT organisation.”

> **"Experts responsible for data generation and analysis in the various tracks meet and share their respective data. This is key, as the data package for an antibody will be significantly different from a small molecule; neither IT systems nor individual teams speak the same language in these expert functions."**
>
> **Mattias Bood, Associate Principal Scientist, AstraZeneca**

As one of those working at a larger institution, Babraham's Andrews agrees that, with scale, it's not always possible to know what everyone else is doing in detail.

"We have individuals that are extremely knowledgeable about individual modalities," he says. "You may think you understand their data set, but they will have layers of knowledge and experience that provide a level of detail that is almost impossible to derive from looking at the data set in isolation. That detailed knowledge is very hard to carry through when you start combining data sets from different modalities.

"By necessity, you have to present a very simplified view of each of the modalities to be able to integrate them. In order to counter this, we run internal training courses so that our teams can learn about other techniques; for example, you can learn how proteomics works if you're mainly involved in sequencing. You may never use the information you learn directly, but at least you need to be aware of the limitations of the technique.

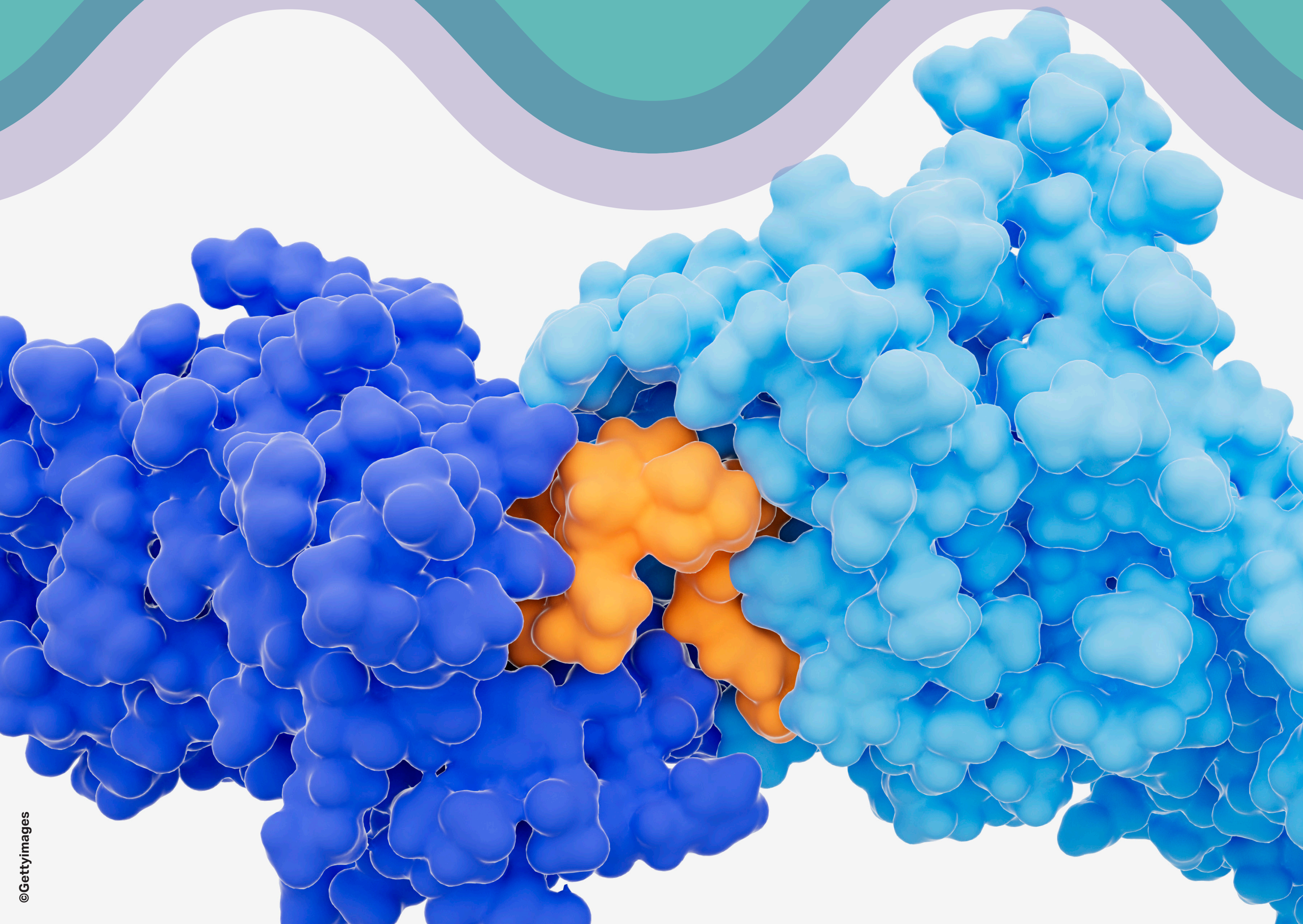
"We get the domain experts to help give the best possible guidance for processing the data. They might present some sort of QC (Quality Control) type view, so that you can weed out problems early on and then carry through the minimal set of metrics that we need to be able to interpret that data.

"We combine that knowledge with the expertise of groups such as mine [in bioinformatics], who are more generalist in nature, but need to understand enough about those specific domains to avoid mistakes when bringing data together."

AstraZeneca's Bood is very much in the large company camp when it comes to issues faced in data sharing. "We work with a plethora of drug targets," he says, "some with a single chosen modality and others with several modalities at once. Typically, we would have a project leader for each individual modality, but naturally they are collaborating across the project as a whole. Usually, experts responsible for data generation and analysis in the various tracks meet and share their respective data. This is key, as the data package for an antibody will be significantly different from a small molecule; neither IT systems nor individual teams speak the same language in these expert functions. We do try to harmonise things as much as possible and have significant frameworks set up for how data should be acquired, documented, reported and stored."

Curlew's Lynch says you don't always need to resolve all siloes. "We don't necessarily have to integrate all data for all people," he says. "As an adviser, we would want to engage with the different scientist personas. Some will only want to query data that's earlier in the work stream, while an ML data scientist will want to see all the data. It's about having different views to suit the different roles that they have, because different questions require different amounts of data."

Bespoke vs public data languages and ontologies



©Gettyimages

“I’m sure, 10 years down the road, we’ll look back with hindsight and have a different perspective, having learned some lessons along the way”

Sarah Sirin, Director and Head of Computational Chemistry, Remix Therapeutics

In smoothing communications among data stakeholders, there is a choice to be made between using established, recognised ontologies and developing a more specific one for use solely at your institution. It’s not a black-and-white choice though, as Curlew’s Lynch says: “There’s always going to be a blend of internal and publicly-available ontologies, but you probably want to try and use the public ones to start with. Those will have historical adoption, although, of course, some are better curated than others.

“There will be cases within any company where you’ll have your own particular targets or mutations and you’ll need to extend the public ones. That’s where a reasonable ontology management system can help. It should always be 95% public, with a small element of bespoke development. Communities like ours can help push these public resources to perhaps keep up with the science.”

The blended approach is exactly what Micholt’s team uses at Galapagos. “Specifically for proteins, we had quite a challenge to make something that was workable,” she says. “For naturally occurring proteins, we took the ontologies that were publicly available, but for synthetic proteins, mutations or isoforms, we created a different master system.”

The same is true for Sirin at Remix. “We’re targeting different pieces of RNA,” she says. “We often start out with the publicly available annotations, but we have to make up our own ontology along the way. For example, if we’re working with an in vitro system that might have some nuances, we may be moving away from the native RNA structure. As we make these modifications and mutations, we naturally come up with our own scheme and language to distinguish the modified RNA from the naturally occurring variants.

“We have a semi-flexible registration infrastructure that allows us to capture some of these nuances with some guardrails to serve as a guide, ensuring data integrity. I’m sure, 10 years down the road, we’ll look back with hindsight and have a different perspective, having learned some lessons along the way.”

Triballeau’s previous experience also used publicly available ontologies. “It was rarely the case that we could use them as they were,” he says. “For example, the standard living species reference database is the NCBI (National Center for Biotechnology Information), which has about 1 million species registered in its taxonomy. The work we did only required maybe a couple of hundred species, so it was really more manageable if we actually built the ontology ourselves.

“By the way, the NCBI taxonomy is not an ontology. It’s a taxonomy, so it’s not the same format. We wanted to have everything as an ontology, so that it’s independent. Whether you were using a species or a gene protein, it would be understood by the same ontology.

“Sometimes, it can take less time to build an ontology from the ground up than it does to tweak an existing one to fit your needs.”

"Bridging between ontologies can be challenging," adds Verge's Choi. "We start using public ontologies, but often find that they're either misaligned when you start bridging between ontologies or they're just too vague. That's not good enough when you're looking at identifying targets or disease mechanisms and pathways."

Sometimes, the positioning of one's organisation will dictate the choice of ontologies, as is the case at the Babraham Institute.

"We need to satisfy a need for publication of both the data and the actual papers that come from this," explains Andrews, "so we're largely guided by the annotation structure that's required from the public repositories. There are relatively few cases where we would invent our own systems on top of that, because we usually fall within the scope of the public repositories."

Public ontologies can lag behind the needs of users, though, as AstraZeneca's Bood explains. "In my field of oligonucleotide chemistry," he says, "we're often not really speaking the same language, whether that's in academia or industry. We can say that we are speaking Helm, or we have the Pistoia Alliance mandating in some cases, but the rules and the framework are just not moving fast enough. It's not developed enough. It doesn't enable the science we do or plan to do.

"We might start with something that would be publicly available and then develop something completely bespoke on top of that. We can't fully utilise or publish all the nuances and building blocks, because this is part of our IP, which naturally complicates data sharing on a wider level."

Ontology challenges aren't limited to the drug discovery part of the chain, says Olivares-Amaya: "The ontologies challenge that I've faced has been more to do with mapping together clinical trial data with electronic health records across different providers. There, you face human problems, such as what are the decisions that have been made in the past that put you in this position that maybe constrains your ontology."

Public data sets and reproducibility

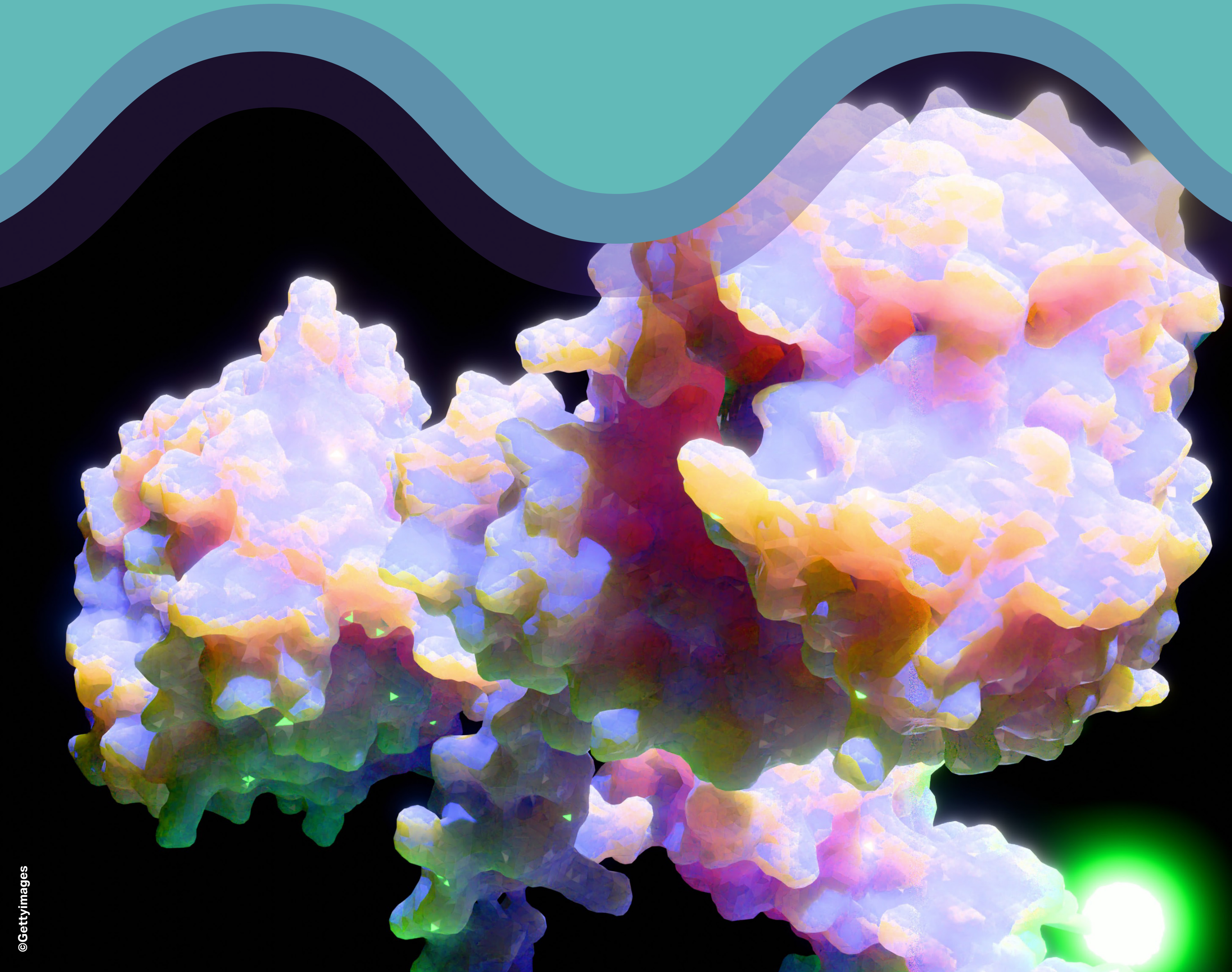
It is recommended to exercise a degree of caution when using public data sets, as they may not always be reliable.

"One of the challenges of ingesting public data sets is inconsistency," says Verge's Choi. "We're really reliant on generating our own internal data where we have robust numbers, case controls, enough depth in sequencing and so on. We do this largely to show how it does or doesn't mirror what we see in the public data sets as a way to cross reference."

For Babraham's Andrews, it's about pausing at that ingestion point to think about the wider data, rather than the headline conclusions. "We encourage all of our scientists – particularly the younger ones – to look beyond the headlines of a research paper and into the underlying data," he says. "In many ways, taking a paper at face-value is the same attitude AI has, in that it will ingest all of the information and inherently trust it.

"The nice thing now for most public data sets is you can go to the underlying raw data where some of the interpretation is taken off it – but that comes at a computational cost."

Cultural challenges in changing data platforms and structures



©Gettyimages

> **"Achieving widespread cultural adoption of new data policies is tough. We can use mandates or help with adoption. We like to focus on the collective need and the reason for the change to help the scientist understand and rationalise the temporary adoption pains. We try all the tricks. We throw sticks and carrots and all of that, but we haven't really found a great approach yet"**
>
> Sarah Sirin, Director and Head of Computational Chemistry, Remix Therapeutics

While different modalities and software platforms create issues with data sharing, there is also the cultural issue of persuading teams to adopt new data practices.

"Achieving widespread cultural adoption of new data policies is tough," says Remix's Sirin. "We can use mandates or help with adoption. We like to focus on the collective need and the reason for the change to help the scientist understand and rationalise the temporary adoption pains.

"We try all the tricks. We throw sticks and carrots and all of that, but we haven't really found a great approach yet."

"Resistance to change is part of human nature," adds Triballeau. "We have to live with natural inertia. Some of that resistance comes from a fear that the change will slow them down, because of a perceived additional administrative burden when they register a new molecule or a new modality or log on new results. They want this to be as efficient as possible, so they can get on with the science.

"The trick here is to explain that the new platform that they are going to use will semi-automate what they used to do manually beforehand. You need to demonstrate that upfront if you're going to get them to change from a tool that they've been using for 10 years or more.

"New platforms can help verify the data at the point of input, ensuring it has rich metadata and allowing reuse at a later date."

Olivares-Amaya advocates "starting small" with any changes. "One of the ways to drive adoption is to figure out how to break things into iterations to get feedback early," he says. "That can also mean getting a smaller team to work out a version that works, before rolling it out to the larger organisation – that can help break the inertia of the status quo.

"What also helps is to train all new starters at the company only on the new platform – that can really help adoption."

Verge's Choi once again champions the benefits of getting stakeholders together. "One of the things that I found most useful in transitioning to any new platform is finding common pain points that are shared, kind of, across the different groups," she says. "By doing this, everyone feels like they are co-creating the new platform, which in turn leads them to be more motivated to adopt it when it comes online."

“The most successful places where I’ve seen widespread adoption of a new platform have been where there was decent management buy-in”

Evelien Micholt, Operational Lead Discovery Sciences, Galapagos

“We try to make sure that there is some immediate early win for the people that are using any new platform,” says Babraham’s Andrews. “Yes, you can insist that they use it in the correct way, but people are busy. We have been really wary of putting up apparent roadblocks that dictate that certain fields need to be completed before users can submit their data. There’s a real temptation just to fill anything in there to make the question go away. This can lead to additional – and useless – data that just creates noise in your downstream analysis.

“If you can solve an immediate problem for someone with the new system – even if that’s not the main point of it – then that’s the thing that you sell it on: ‘If you do this, then, later on when you want to do that, you can now press one button and it happens, rather than you spending half a day doing it’ – that’s an incentive.

“We spend a lot of time working with the scientists to make the user experience that they have on the front end as slick as possible. Streamline the process as much as you can. So, even if it’s as simple as: ‘Well, I’m entering 20 samples. Okay, I’ve put one in. Let’s populate the next one with all the same details, because it’s probably going to be in the same experiment.’ Then, you only have to change the bits that vary.”

Support from the very top of an organisation can be a real accelerator for circumnavigating cultural barriers, according to Galapagos’s Micholt. “The most successful places where I’ve seen widespread adoption of a new platform have been where there was decent management buy-in. If you don’t have a champion who takes the time to explain the benefits, then people are going to take the easy way and carry on doing things the way they’ve always done them.”

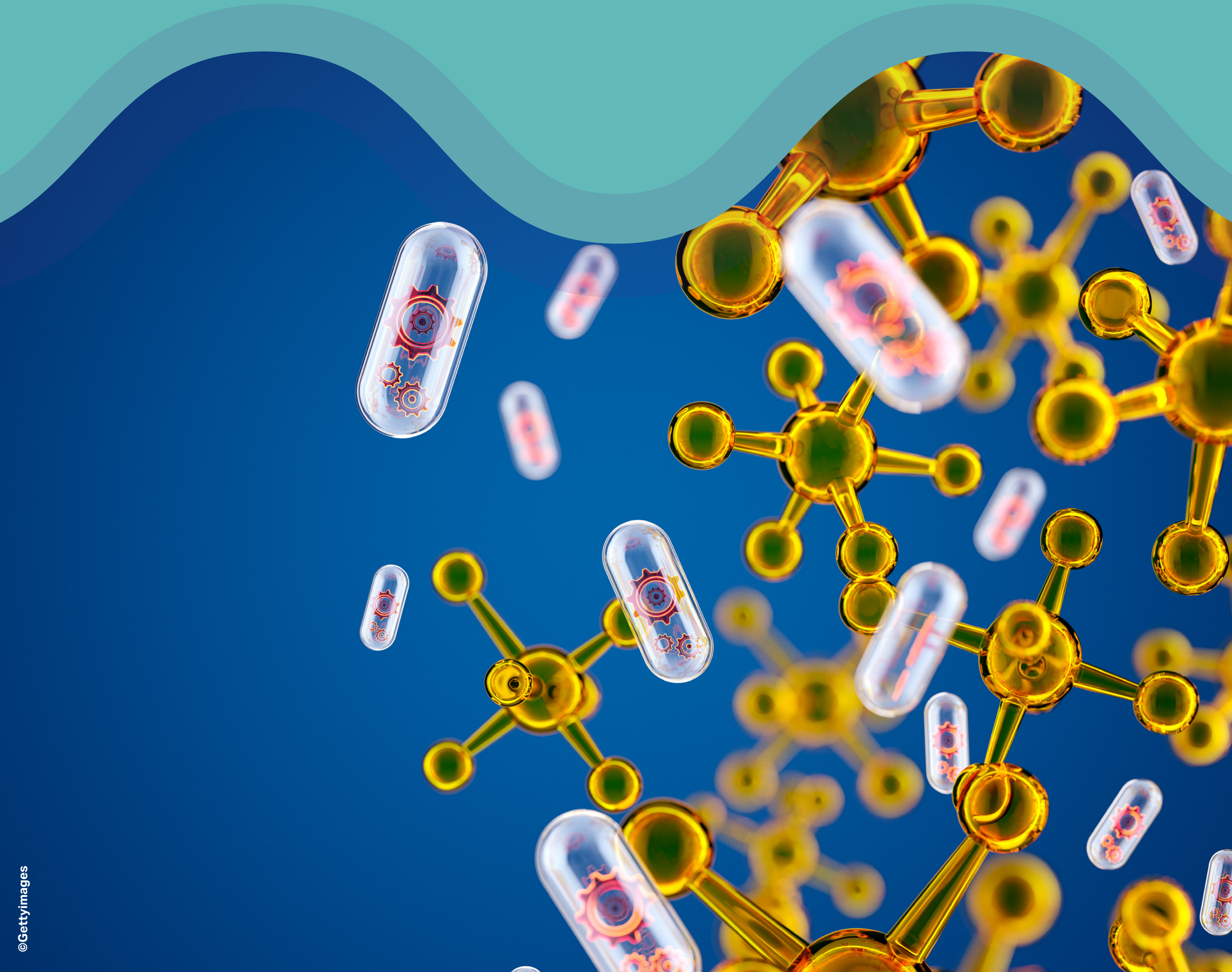
Curlew’s Lynch agrees: “Rather than the term ‘carrot and stick’, we encourage using ‘invest and reward’. That helps the top-level stakeholders to pay attention, as well as the ‘what’s in it for me?’ attitude of individuals at the user level. With new platforms, there’s always going to be ‘investment’, not just financially, but in people’s time too.”

AstraZeneca’s Bood says product champions and canvassing widespread opinion can be a winning combination.

“Compliance is important and necessary,” he says. “We mandate it as required by local legislation. However, whenever possible, we try to involve the end users in the process by selecting super-users or champions for any new software rollout; they get to provide significant input and suggest improvements to the platform, as well as going on to help teach other staff.

“We also try to crowdsource when possible. For instance, if launching a new platform with new ontologies, we’ll set up possibilities for end-users to add to the framework by adding their own phrases or building blocks and so on under controlled forms. This helps involve the users and is greatly appreciated.”

Artificial intelligence (AI) / Machine learning (ML) in drug discovery



©Gettyimages

The drive to adopt AI/ML appears to be ubiquitous in drug discovery, but it's not a switch one can flick to solve all problems; it requires careful planning and development, as Babraham's Andrews says: "There seems to be an opinion coming from above to us that AI/ML is going to solve a lot of the intractable problems that we've been sitting on for a long time. That opinion needs to meet more reality than it currently is.

"AI systems are just really trying to pick signals out of a lot of raw data. Unless you have an enormous collection of data, ideally one very specifically designed and collected in a consistent way with consistent processing and every other sort of extraneous factor removed, or a very large data set where every other potential extraneous factor is well annotated and can be incorporated, then the signals that it can pick up on are very often not the ones that you would want. You can spend an awfully long time thinking that you're going down a productive route, but actually end up spending more time than you would have done by a more conventional analysis disproving what the automated model has been doing.

"Having a degree of planning in the very early stages to ensure that you're actually feeding it the right kind of problems, and that you've collected the right data to do it to be able to address those with the right sort of technology, is super important."

When it has been implemented well, as is often the case in big pharma, AI/ML can obviously make a huge difference.

"We have an LLM (large language model) sitting on top of everything and it's like magic when it comes to learning about specific subjects," says AstraZeneca's Bood. "As an organic chemist by training, it is immensely useful when it comes to explaining how certain biology works for specific drug projects, enabling me to start to see connections that I never really had before.

"Moreover, we have plenty of custom AI/ML predictive models for a range of individual siloed tasks, from predicting organic chemistry outcomes to designing the best anti-sense oligonucleotide for a specific mRNA and much more."

AI/ML tools are rendered virtually useless unless the underlying data is of sufficient volume and quality. "The data helps to drive and accelerate the application of AI/ML in the development of different modalities," says Verge's Choi. "The more data you can put back in to let it learn, then the better the predictions will be for the next set of targets or the biological mechanisms in which you're interested. What we have developed at Verge is an iterative process where we use human data sets to create novel targets for treating a neurodegenerative disease. We then get a rank order for those targets, which we then test in human model systems. This generates data to put back into our algorithm to see how it did or didn't rescue a disease network or signature that we were interested in.

"We can then iterate on this multiple times to increase the probability of rank ordering targets that would be most effective in mitigating a disease signature.

"One of the other challenges to that is that AI/ML can be misled. If it's generated by limited data sets that are skewed or perhaps not appropriate for the given indication that you're pursuing, it can lead you down some really weird tangents. It becomes important that, in the application of how we use a AI/ML, there is some level of expert supervision that sits on top of it."

Olivares-Amaya agrees that the quality of data is key. "It is becoming more and more relevant to access

> **"There are tools that are able to process histology images swiftly and in a more reliable way than an expert would... on the microscope. That indeed impacts drug discovery and health in general"**
>
> **Nicolas Triballeau, Director, Drug Discovery, Revvity Signals**

metadata, particularly with the advent of AI and ChatGPT," he says. "In the multimodal space, you are likely to have text and data together – and search methods are changing as a result.

"AI is going to be very interesting when we figure out how to encode the central dogma into an LLM. There are some papers out there that are starting to research this, but putting in a DNA structure or a protein sequence and then going up and down the central dogma to predict outcomes could be transformative."

There are already some widely used AI tools out there. "We have AlphaFold nowadays to predict protein 3D structures," says Revvity's Triballeau, "and there are tools that are able to process histology images swiftly and in a more reliable way than an expert would… on the microscope. That indeed impacts drug discovery and health in general.

"In my previous company, we had some successes with predicting some ADME [absorption, distribution, metabolism, and excretion] endpoints, toxicities or plasma protein bindings, which was only possible because there was a lot of quality data that could be used to train these models. Day-to-day work is often on smaller data sets, meaning there isn't enough data to train the AI."

AI/ML is not a 'fit and forget' application, according to Curlew's Lynch. "You need to be careful not to develop pure blackbox AI/ML models," he explains. "It's important to have a holistic view of your models and be willing to challenge them through continuous review."

Furthermore, it is also important to remember that human intervention and supervision will always be necessary, as Remix's Sirin points out.

"There's always a need for human expert knowledge on top of any AI/ML model," she says. "Scientists that are relying on these models are going to be more successful in the long run than scientists that perhaps just avoid them completely."

How software platforms can improve

©revvity

“Software platforms should augment the role of the scientist; they’re not there to replace them”

Nick Lynch, Director, Curlew Research

Choosing the right software platform (or, in many cases, the right combination of platforms) to suit one’s drug discovery needs is a case of balancing the pros and cons of each one, depending on the outcomes required. For the most part, though, users still believe software platforms can do better.

“Getting platforms to talk to each other is the one thing that I ask all software platforms to really think about,” says Verge’s Choi. “We have two different ELN [electronic laboratory notebook] systems; we have our own internal web portal of data and just trying to get them to talk to each other is still a challenge.”

Revvity’s Triballeau agrees: “Defining a common language across tools – that’s the difficult thing,” he says. “There is now this concept of an ontology management system, which can push terminology and semantics to different tools. The problem is that not all tools can absorb these terminologies. One of the key challenges is for software companies to develop tools that have APIs that have capabilities to actually ingest control vocabularies and ontologies.”

It’s inevitable that systems and needs evolve, as Babraham’s Andrews says: “One of the real difficulties here is that we don’t necessarily have a platform that’s going to be our platform for the rest of eternity.

“So, having something where we can have a nice front end that the scientists might see, but also an equally efficient backend and open storage of what’s behind there so that, should we need to extend that or transition it to something new, it’s easy to do that.

“Increasingly, having something where we can publicise our data directly from the platform would be really useful. We’re being urged to make our data more sharable and publicly available and having to do that through a second step is difficult.”

AstraZeneca’s Bood urges vendors to change tack to help collaboration and integration. “It feels like both software and hardware vendors are going towards making proprietary solutions where they lock in their data,” he says. “Take mass spec, for instance; for some of the common big names, you’re trying to process their data on their own software suites, but it’s so difficult to export and analyse it in other tools.”

Returning to the topic of AI/ML, Remix’s Sirin pleads for better usability of the tools that are available. “ML and AI are still expert tools that are, at times, difficult to interact with,” she says. “Better dashboards, intuitive user interfaces, or explainable models are all things that can help scientists play around with the models, educate themselves and get comfortable with ML nuances.”

The final word goes to Curlew’s Lynch, who reminds us that software platforms alone are not the answer. “We have to get the balance right between human and machine interaction with these platforms,” he says. “Software platforms should augment the role of the scientist; they’re not there to replace them.”

Conclusions

- There is a trade-off between the time and expense incurred in restructuring data versus the benefits gained through being able to access data freely from multiple modalities.
- Consider separating experiment data, allowing that to exist (and be available for interrogation) independent of modality.
- Not all data is the same – large volume data sets can be useful but 'noisy', while the detail of experiment-specific data may need to be viewed in isolation. Your data structure should reflect these different needs.
- Mapping out your data life cycle can help assess the data structure required and the tools you may need to interrogate it.
- A centralised data resource that allows interrogation and manipulation without the need to duplicate data sets can save on storage costs.
- Bring together as many stakeholders as possible from all stages of the discovery process when first building (or rebuilding) a data structure to ensure everyone can access the data in the way they need. This will also help with adoption of the new system, as stakeholders will be invested in the result.
- Ensure that all levels of the organisation are involved in these discussions, not just team leaders, since bench-level scientists will often bring as much insight from the ground as anyone else.
- Being stricter with data structures and annotations will make the deployment of AI/ML tools much more effective.

- For larger companies, consider running training courses so that teams in other disciplines can get a better understanding of how each other works – and, therefore, how data demands may differ between teams.

- Use publicly available, recognised ontologies wherever possible, recognising that some organisation-specific customisation may be required.

- When introducing a new data policy, try to deliver quick wins that demonstrate the benefits of the new approach – these will usually be about saving time in the long-run.

- In order to be effective, AI/ML tools need both human management on top, and quality data underneath.

- Choose a software platform that uses common data formats and/or integration options through APIs.

For more information, visit: scientific-computing.com
and revvitysignals.com

This White Paper was produced by Scientific Computing World
in partnership with Revvity Signals