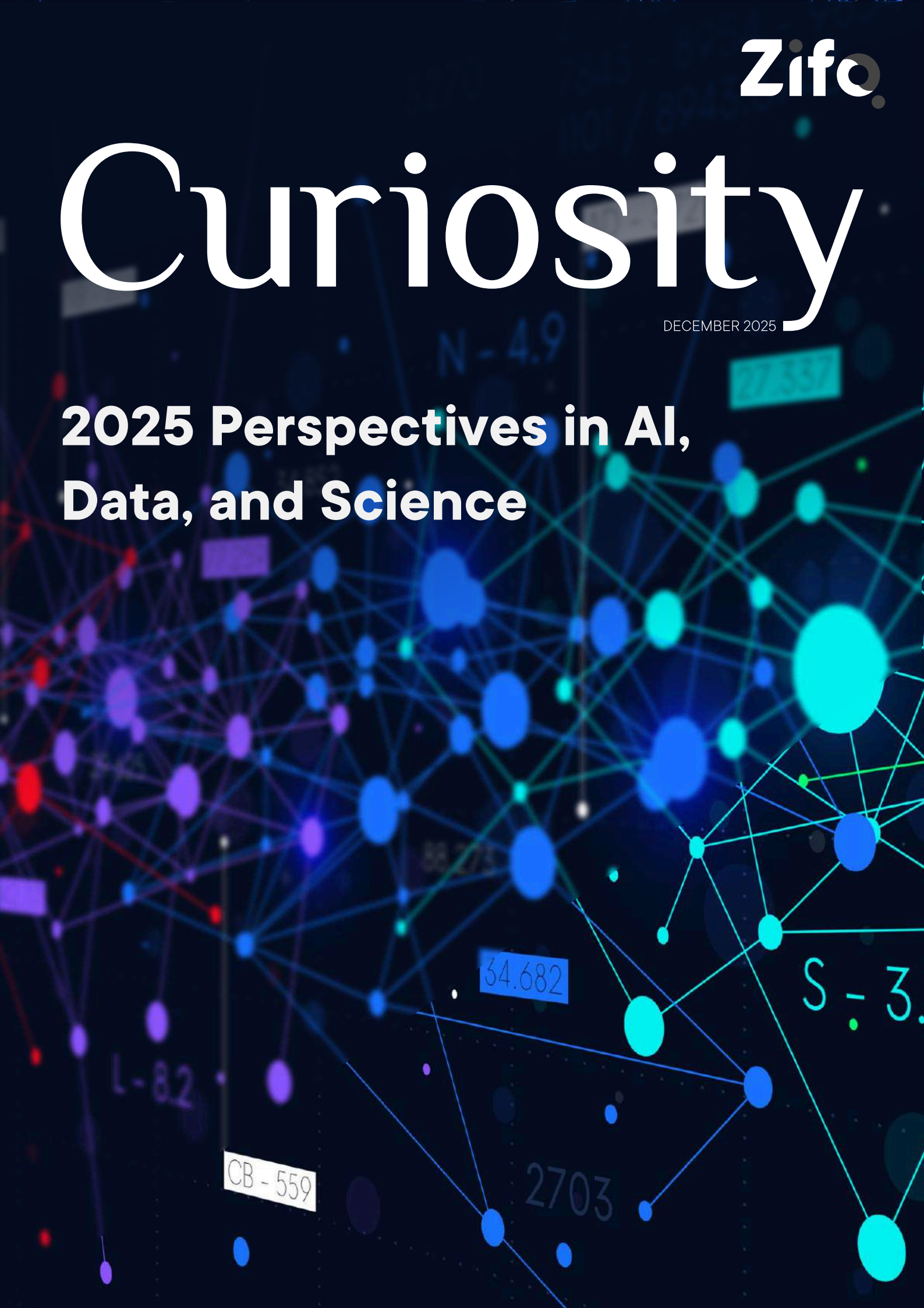
Zifo

Curiosity

DECEMBER 2025

2025 Perspectives in AI, Data, and Science



Preface

AI at Curious Crossroads

AI sits at a curious crossroads in biopharma and other science focussed industries. Google's claim that its Gemma model helped uncover a new cancer-therapy pathway has reignited debate about what truly counts as "AI," what qualifies as a scientific discovery, and how the role of the scientist is shifting as algorithms enter the lab.

Meanwhile, translational genomics is experiencing its own "Google Maps" moment: biobanks are producing unprecedented volumes of data, yet navigating from raw information to therapeutic insight still feels like using a paper map. Broader industry sentiment mirrors this tension.

While the AI hype is reaching feverish pitch, Gen AI has entered Gartner's "Trough of Disillusionment," and the analyst firm places "Tech Transfer" there as well, reflecting widespread frustration with undelivered promises.

Zifo's recent Data Readiness Survey echoes this reality: scientifically driven companies are investing across R&D, manufacturing, and clinical trials, but most of the work remains foundational – cleaning data, aligning standards, and establishing shared metadata and ontologies. AI initiatives are just beginning their long road to maturity.

In this landscape of heightened expectations and uneven progress, the question becomes where AI can deliver real, grounded value – sometimes in the flashy breakthroughs that grab headlines, but just as often in the quieter, behind-the-scenes work that keeps science moving forward.

Read our perspectives on AI, Data, and Science in this year-end special edition.

Team Zifo

Table of Contents

Executive Viewpoint

Expert View

Table of Contents

Zifo Data Survey

Interview: AI Not Yet a Miracle Worker, But a Powerful Scientific Prompt Responder

Tech giant Google in collaboration with Yale University released an AI model known as Cell2Sentence-Scale 27B (C2S-Scale), a new 27 billion parameter foundation model designed to understand the language of individual cells. Built on Google's Gemma family of open models, C2S-Scale represents a new frontier in single-cell analysis.

Given this scenario, and even as AI fever grips Biopharma and other science-focused industries, along with its fair share of disillusionment, we spoke with **Paul Denny-Gouldson, Zifo's Chief Scientific Officer,** to make sense of what Google's announcement means for AI in science and for scientists.

Q.1

Google announced this partnership with Yale, and their AI model has generated a cancer hypothesis. This, of course, is the initial stage of research, and the company says they achieved in vitro success, which then requires testing and confirmation in vivo. Firstly, what are your thoughts on this development? Secondly, what does it truly mean for an AI model to generate a scientific hypothesis, and why is that significant? I am asking this because I want to know: is this a 'Sputnik moment' for AI, or is that an exaggeration?

Firstly, I would like to congratulate the Google and Yale team for achieving this impressive breakthrough. The first thing we have to consider is what this thing that's been created actually is: it's a model of a single cell that enables it to answer questions a scientist asks. For instance, they might ask, "Can you find me this?" or "Can you tell me about this?"

The model's power lies in the vast volume of different, multimodal data types used to build it. This essentially creates a statistical model of how the single cell will behave. The data types include imaging, biochemical, and phenotypic data, all merged with open-source information on pathways, genes, and proteins.

This specific development is a single-cell model for cancer, making it disease-specific. The paper discusses the concept of developing similar models for other diseases and cell types. By linking different data types, the model offers predictions.

> ***However, we must be careful about calling it an "AI" in the sense of an independent creator; it is a model that is only as good as the data used to create it.***

The authors themselves discuss extending the model with more data types to increase the fidelity of the information and the model's ability to offer alternatives or predict things.

But it always comes down to the question you ask of the model. The model itself won't just create something out of nothing regarding potential treatments or outcomes for the disease in question. It responds to a prompt like, "I am investigating this; can you find me an alternative or something that works with this?"

That is why it's so important to really distinguish this from the de novo creation of a hypothesis by the model or the AI on its own.

Back to the question – is this a Sputnik moment – maybe, maybe not. I am not sure. My healthy scepticism has me thinking if this has been done before, but behind "closed doors" – so in that way it may not be the "first". There is no doubt, however, it will fuel the research into this space as it shows what's possible – and being open-source means others can build on it, which is the best way to accelerate development.

So, in this way, it is just like Sputnik – it will incentivise and challenge others to build on it and do better, like getting a man on the moon, or in this case, even Mars.

Q.2

Okay, this is fantastic. They built a model, applied a "specialised LLM" trained on a multimodal corpus of over 50 million cells and associated text on top of it, and the scientists used a highly structured, computational query to interact with the model. Hence, this isn't a classic LLM query like "Hey C2S, what drug combination works for cold tumors?"

Yes, that's correct. Now, let's be clear: this is a very, very large model. While I don't know if private organizations have produced bigger ones, the sheer number of parameters they used – in the billions – means it can accommodate a vast variety of different questions. This is because all cell behavior is interlinked, and the model's main function is to map all those links.

The power of having such a large model to describe the overall behavior of an individual cell – remember, this is a single-cell model – is its ability to find relationships that a human cannot see or even contemplate. We can only actively manipulate perhaps two or three parameters in our minds at once; whereas this model has billions.

That is the true power I see in this methodology: building a model using the highest quality and largest volume of data possible. This allows the model to be used effectively to test or generate hypotheses.

Q.3

Now that the model has come up with the hypothesis, the scientists will still have to test it as the next step, through experiments.

Yes, now that the model has come up with the hypothesis, the normal process goes on. Now we have the classic design, make, test, analyze (DMTA) cycle.

> ***What the Google-Yale model offers is the ability to generate ideas for testable hypotheses in the lab, all derived from a vast pool of quality data.***

This brings us back to the issue of good data: if the data used to create the model is not trustworthy or high-quality, the model will inherently be flawed. Regardless, the scientist, organization, or lab environment still needs to proceed with testing that hypothesis.

The fascinating aspect here lies in the cost and subsequent cycle. The investment to build and train this model is significant – likely multiple millions, if not tens of millions of dollars, factoring in compute power, personnel, and data curation. This endeavor is a real showcase of what is possible with major investment from partners like Google and Yale.

Now that the model exists, as we go through the design, make, test, analyze cycle in the lab – the "wet work" – and generate new data, that data can be used to augment and retrain the model.

If an experimental result confirms the model's prediction, the model's trust in that relationship is enhanced. If the experiment result "fails", the significance of that perceived relationship is lowered.

> ***This entire cycle must continue: you continually add new information based on the hypotheses the model generated.***

This process is called reinforcement or checking of the model. As you conduct more testing, the model becomes progressively better, because you are exploring, extending and validating the entire "cell process space" that exists within the model.

Q.4

Okay, so how was this hypothesis being validated? I am asking because we are so used to the consumerization of LLM models. People can ask ChatGPT or any other LLM any question, and it throws out an answer, but we don't know if it's true or if it's just hallucinating, and so on. From the paper, were you able to deduce any guardrails that Google and Yale put in place to ensure that the hypothesis generated by the model is actually validated, and not just hallucinating? How do they ensure that?

There are different ways to reduce the risk of hallucination. One is ensuring the data used is of good quality.

This alone is a labor-intensive problem, requiring knowledgeable people to validate the data. However, a number of automated and other tools can be used to check the data for consistency.

As for the guardrails, this is where the concept of a data foundation comes in. Beyond the raw data, the critical piece is the ontology used to govern the relationship between all the entities.

> ***You can start with a gene, a pathway, a hormone, or a metabolite and traverse the relationships between all those entities.***

The ontology is critical; they used and curated a number of external ontologies to ensure they had a good definition of all the relationships and terminology. This ontology can then be used to check the data.

It asks: “Does this new data fit the ontology, or does it suggest claims or relationships that are either new or contradict existing information?”
If a contradiction arises, scientists must go in and examine the data to see if it was tagged incorrectly or produced inconsistently.

> ***The large language model (LLM) itself also needs guardrails to prevent it from making incorrect assumptions or stating that two things are related when they are not. The ontology helps serve as the quality checker or QC element to support the LLM.***

The final piece, of course, is that you cannot trust anything without experimental testing and checking.

Q.5 ***What excites you about this AI model?***

The exciting part here is using this concept of a model to empower scientists to explore hypotheses. The next step would be to have an agent that can take that hypothesis and turn it into an experimental design – for example, “I can help you design an experiment to test the relationship between this gene-pathway-protein-disease”.

This could then be executed by another agent that controls an automated robotics laboratory – and a third agent could pull all that data and analysis together for the scientists, before pushing it back into the model after the scientist intervention.

This creates a virtuous cycle of support. As you add more data from successful or failed experimental studies, the model is continually refined through reinforcement. If something works, it enhances the trust in that relationship. If it doesn't, it lowers the significance of that potential relationship.

> ***The more testing conducted, the better the model becomes, because you are exploring and validating the entire experimental space.***

Playing this forward, the ultimate goal is to have a whole set of models that mimic a human, starting at the cell level, scaling up to an organ, linked organs, and eventually the entire system – a human – the "moon shot" perhaps.

Previous attempts at this were small-scale, and while useful, they are limited. The power of this new wave of large compute is that it enables us to build much better, larger-scale models.

Q.6 ***Okay, moving on, Paul. As a scientist yourself, how important do you think scientists are in the current day, now that these AI models are slowly emerging, and what makes their job so important?***

I think it's really interesting talking to fellow scientists.

As I do most weeks, many of them are a little distrustful of AI generally, mainly because of the hype and the outlandish claims some people have made about what it can do. I think it's all about being very thoughtful and employing a healthy scepticism – and that's exactly what I see with my science colleagues and believe myself.

There are some scientists who are "all in" and pursuing it aggressively, and others who are completely against it. But the biggest group is that middle band utilizing healthy scepticism and asking, "What's in it for me? How will this help me do my science?"

You have to imagine there are so many different types of scientists and problems that scientifically driven organizations are trying to solve.

> ***The example with Google and Yale – a single-cell cancer model – is perhaps only applicable to 0.1% of the scientific community.***

This is where the hype has taken over a little: the declaration that "we have done this, and it's amazing" isn't relevant if you talk to chemists, food scientists, or people who don't work in cancer or cells. These other scientists are asking, "I need something that will help me."

I think having healthy scepticism is very good. The job now for all organizations is to determine what kind of support will work for each specific scientist or scientific group.

Will it be a "model" like this, or a different concept of how machine learning or an LLM-based search interface can interrogate information?

Fundamentally, there is a whole host of different scientific problems to solve. Whether it is physics, chemistry, biology, or material science, all have their own unique issues.

The key is to put the scientist at the center and determine how we can help them do their job: help them ask questions they perhaps didn't think of and see patterns they can't see.

This is the ultimate power of LLM interrogation of big models: surfacing patterns, links, and relationships that are impossible for a human being to perceive.

Q.7

The hypothesis that Google generated for a single-cell cancer model – does this qualify as a discovery? And with AI entering the process, does that generation itself qualify as a discovery?

Personally, I don't believe it qualifies as discovery. Discovery is the finding of proof of something – being able to definitively state, "We now know this behaves this way and this is a new discovery."

What Google's single-cell model has done is create an environment for people to ask questions they perhaps couldn't have asked and to generate hypotheses, or "multipaths," that then need to be tested.

> ***Until that relationship is proven in the lab – both in vitro, which they have successfully done, and then in vivo, which they are currently working on – it does not qualify as a discovery for me.***

The model didn't just tell the scientists there's a relationship "out of the blue." You had to apply their scientific knowledge of disease biology and pathways to formulate the initial queries – as they describe in the paper – to build up the investigation.

Only then is the model able to suggest a potential link and the treatment regime for the specific problem. It's an interesting journey through the semantics of what constitutes discovery.

> ***I believe these models, at this stage, will support scientists in exploring areas and relationships they had never seen before – but the scientist still must prompt and interrogate the model.***

Q.8

All of the scientific AI we hear about at the moment is, so to speak, primarily focused on the discovery side – specifically drug discovery – which is arguably the most glamorous part. However, discovery alone is not all of science, right? There is a whole host of processes involved in bringing a drug to the market, and currently, all AI efforts are concentrated on discovery.

It's becoming a bit like a fairy tale where some people claim, "Oh, I just need to ask GPT to find the next prototype for paracetamol, or whatever, and it gives me an answer."

Do you think there is too much focus on the discovery side because it's obviously glamorous – the idea of just finding the next magical drug?

> ***The chatter suggests that drug discovery might move from wet labs to computation. I know these are very far-fetched, futuristic pronouncements, but we shouldn't dismiss them either, as we don't know how technology will evolve.***
>
> ***As a scientist for a long time, Paul, when you hear such pronouncements, what comes to your mind? What do you think is going to happen?***

You are absolutely right; the discovery topic is the one that grabs the headlines.

> ***However, behind the scenes, at all the scientifically driven organizations I talk to in Pharma-Biotech, FMCG, Speciality Chemicals, Food & Beverage, Agrotech etc., their understanding of AI is evolving very rapidly. They are looking at how AI can impact all parts of the lifecycle: not just discovery, but also research, development, manufacturing, and product trials.***

They are exploring concepts like AI augmentation, decision support, pattern recognition, and workflow support to enable things they couldn't do before. Crucially, the time to value is often faster the further up the chain we go toward development and manufacturing.

This is because there are many problems in those stages that are manual and stepwise: creating documentation, checking data quality, and so on. These tasks are perfectly suited to well-trained models and AI and Agentic concepts. Furthermore, the amount of money you need to spend to address these issues is much lower.

The problems are more discreet and defined – you can precisely put a box around them. For example, if the goal is to create an Investigational New Drug (IND) application, you can work backward, mapping all the required data sources, variables, and parameters. This process can then be monitored and augmented with AI.

In contrast, the discovery problems are massive. Look at the investment required to get a single-cell model to partially describe one disease type: multiple millions to tens of millions of dollars, plus the cost of aggregating data from many different places. These discovery problems are so hard, yet they dominate the headlines.

> ***We have seen many startups in this drug discovery space that require a huge amount of capital that haven't succeeded yet despite spending heavily on aggregating old data and generating new data.***

I feel we are definitely heading toward, or are already in, the trough of disillusionment in this space. The actual wins we are seeing are in more process-oriented, stepwise concepts for AI, and that success will build trust.

But who knows? We could eventually see multiple massive models – a "model of the heart," a "model of the lungs" – each described at this huge scale. Models exist now that describe these organs, but they aren't built on these massive datasets.

The interesting future question will be to compare the massive models with the smaller ones and determine how much better they are versus how much more money you have to spend to build them.

Q.9 ***And, personally, what do you think is the importance of a scientist in this age of AI?***

I think we are heading toward a state where these AI models will adopt the co-pilot concept – an assistant that is simply a part of our life as scientists. How fast we get there, I don't know, but we are already seeing it. Product companies are implementing it, and organizations are building their own support tools.

This raises many questions: How do I know that a product-embedded AI is compatible with my own corporate AI strategy and models, or any orchestration layer I am using? Can I leverage its functionality, but turn it off if I want to use my own AI models in the scientist workflow?

This is because you have to trust the AI. You must trust the model, how it's been developed, the data it was created on, and its output. Without that trust, it's like a game of Russian roulette; you don't know if you'll get the right answer. This brings us back to: How do we avoid or pick up hallucinations and ensure the AI is giving a quality, proper result or inference?

Then we enter the realm of governance – AI governance, model governance, and the concepts of AIOps and GenAIOps, similar to DevOps in many ways. I believe this will become highly connected to the quality groups within organizations, as they are responsible for the quality of the final product.

> ***If we are using tools that directly impact that quality, there will be a much bigger emphasis on governance.***

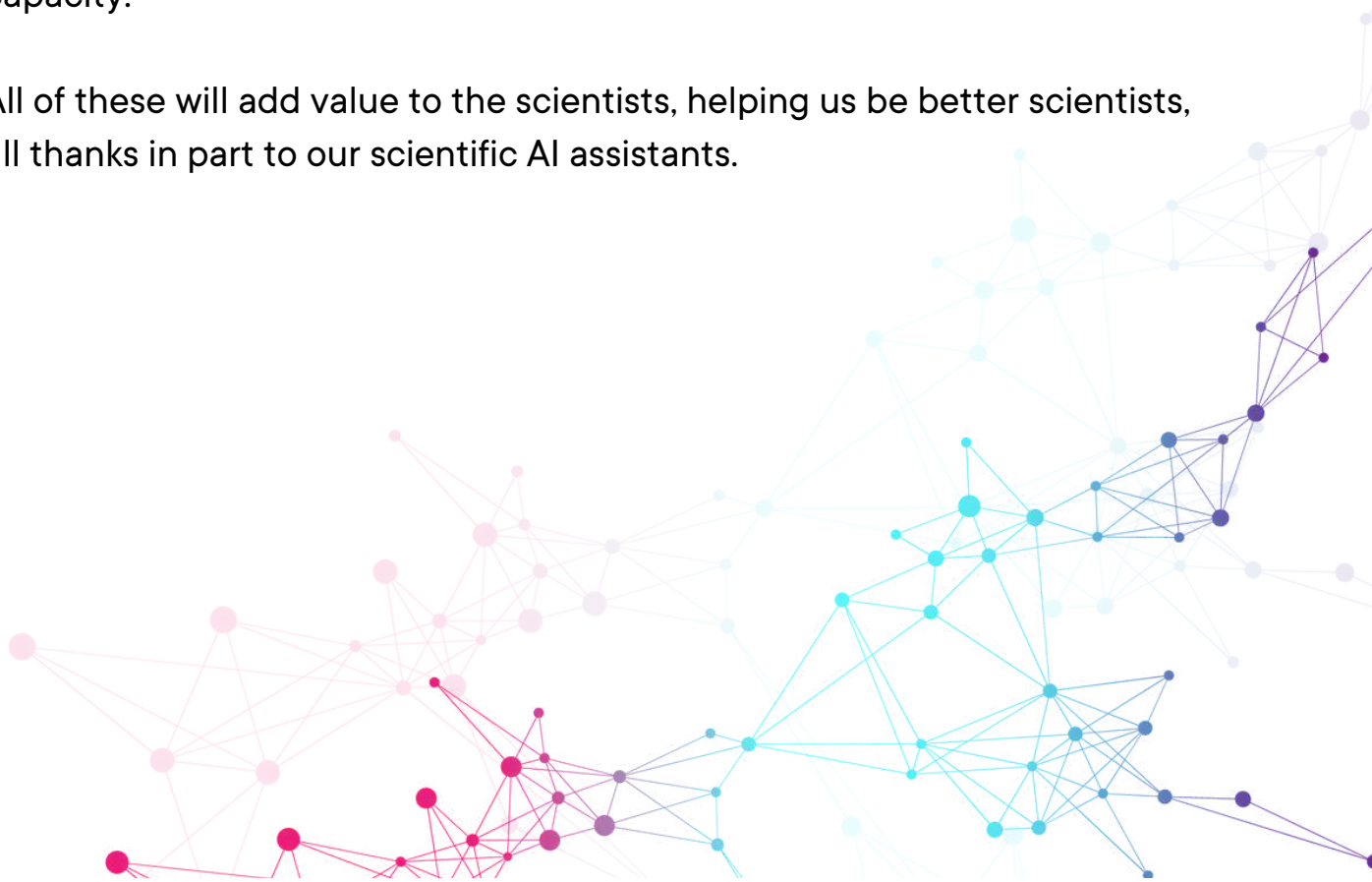
Organizations are not fully there yet; some are just beginning to build this concept. It all comes back to: How do I build trust?

The answer is: I need really good, trustworthy data. This is where the scientific data foundations become essential again – they must be in place to deliver reliable data for the models to consume and provide quality inference or hypothesis support.

> ***What does it take to be a scientist in the age of AI? I think it's all down to being a scientist – healthy scepticism with an equal amount of curiosity and willingness to learn and try new things.***

Applying the DMTA cycle to AI in the same way we typically use it to our chosen area of science. Embracing the technology will help us in finding new ways to work that let us explore new innovative approaches, speed up current things, increase precision and accuracy or enable more capacity.

All of these will add value to the scientists, helping us be better scientists, all thanks in part to our scientific AI assistants.



Interview: Tackling AI Anxiety and Technology Obsolescence to Drive Better Science

While Gen AI has taken the world by storm, as is evident by the consumerization of this technology, scepticism is creeping into the highest level of business regarding the ROI of AI investments.

A recent study released by the MIT Media Lab/Project NANDA found that 95% of investments in generative AI have produced zero returns.

For its part, in what could be termed as a “watershed moment” in the life of this fledgeling technology, Gartner has said that Gen AI has entered the “Trough of Disillusionment” era, the third step in its five-stage technology adoption framework, which is an indicator of widespread dissatisfaction with a technology.

To bolster its claim, Gartner report states that less than 30% of AI leaders have reported that their CEOs are happy with AI investment return.

Given that this breakthrough technology is driving individual productivity, albeit sporadically, but struggling to deliver cohesive P&L level ROI, scientific leaders find themselves at an interesting intersection.

They are compelled to balance the promise of AI with a realistic estimate of investment ROI – a dichotomy that is potentially anxiety-inducing and laced with uncertain business outcomes.

Given this scenario, we spoke with **Margaret Difilippo, Zifo's Head of Commercial in North America,** who brings nearly 35 years of experience in the Biopharma industry, about the lingering anxiety and uncertainty surrounding shiny new objects like AI – and how best to balance the differing expectations of the scientific and IT/informatics departments, the two most important stakeholders in driving science forward.

Q.1

Right up front, there's a lot happening in the market today. From an economic point of view, and considering the impact of tariffs, there's a great deal of business uncertainty. At the same time, many significant AI investments are also being made.

Given this charged atmosphere, what are you hearing from the market? What are top executives telling you when you meet with them? Are you picking up on any signals that business newspapers aren't publishing – any insights that would be interesting for our readers?

I think right now there's a lot of anxiety and uncertainty, especially in large pharma when it comes to how best to make use of AI. Also, alongside AI induced anxiety, we're also seeing many executives being replaced, specifically at the VP level, with new people coming in. This is causing a lot of turmoil.

When these transitions happen, budgets often get frozen for a bit because the new executives tend to bring in their own people. This creates uncertainty among the levels below them, who are worried about their jobs and the new direction the company will take.

Right now, I feel there's a lot of uncertainty. However, in the last few weeks, I've noticed things are starting to settle down as the new people get into place. It felt like things came to a halt earlier this summer, around May and June, but now they're starting to open up again with the new leadership in place at many companies.

Q.2

When it comes to technology choices, I mean, if you take any industry, technology is the backbone. We may work in a bank, a lab, or on an airplane, but technology is what drives almost all workflows.

So, regarding the biopharma sector, what is your take, especially given the AI buzz? Who usually holds the power to make technology decisions? Is it the scientific side i.e., the business side, or the IT/Informatics side? Who decides what technology the business should be using?

I think there are different cultural components across biopharma.

> ***In some companies, the power is with the business; with others, it's with the IT group; and with some, it's a mix.***

What I see is that successful organizations are those where IT and science work collaboratively. When you "throw things over the wall" – if IT has an idea that they need to update their technology and they just throw it over the wall – it will be met with huge resistance.

I think organizations that have senior-level people at the top providing a clear direction and vision on how technology will enable them to do their science better with greater efficiency – for instance, helping them identify targets faster or stratify patients faster – that's where you have the most success.

> ***So where is the power held? IT may have the power to implement something, but if they can't convince the scientists to make that change, it's just not going to happen.***

It's a bit surprising to me, in a sense: some scientists still like to use Excel. Their mentality is, "Why should I change? Give me a reason to change. Why change when I've been doing this for the last 30 years and it's been working for me?"

In research specifically, there needs to be a compelling "why" for them to change. It's up to the IT department, along with the cultural leaders on the science side, to make that case. The younger generation of new scientists are trained in data science and computational biology, so they are comfortable with AI.

But for those who have been in the lab and have been successful for 30 years, there needs to be a whole change management and adoption management process.

People will only make a change if they see value in it, and it's up to the entire organization to demonstrate that value.

Q.3

Okay, that's interesting. So, what are some of the most common, or perhaps surprising, forms of resistance you've been facing lately? This is especially in the context of large digital transformation programs or even with well-established technologies like ELN or LIMS, not just new ones like AI, which is still in its nascent phase. What are some of the surprising forms of resistance?

I think there's a lot of fear about what this means for people. Fear of how it will affect their day-to-day job, and whether they might be out of a job.

> ***I believe if you look downstream on the manufacturing side, that's where most of the resistance will be – because of all the compliance, GxP, and GMP regulations.***

When systems like LIMS are in place, they're expected to stay for a decade. It's amazing to me to find that at some of the big pharma companies, their systems are 12 years behind. They don't upgrade because the time and money required for an upgrade in a compliant, validated system is a huge effort.

So, they tend to kick the can down the road and don't keep their technology up to date. It's working and validated, they feel secure with it, and there's always the risk: if we update the system, are we going to introduce errors? Will we be able to answer the auditors' questions? Fear is probably the biggest factor.

In the clinical space, it's also about how it's going to impact the integrity of the data, which is a big concern.

I also think that, at least in the many decades I've been in this industry, there have been a lot of hypes. Back in the day, computational chemistry was a huge hype. Everyone thought it was going to solve everything and create all these new drugs, so people jumped on the bandwagon.

Then combinatorial chemistry came along, and everyone thought we'd be able to develop new drugs using silicon graphics systems and wouldn't need to do as much lab work. All of these hypes went through a cycle and then fizzled out.

Based on who I talk to, there's a belief that the current hype won't fizzle out completely but will come to a more standardized level where people will see the true value.

> ***So, some people are taking a wait-and-see approach. They don't want to be part of the hype. Once it settles down and they can see where the value is and where it can be successfully applied, then they will embrace it and get more involved.***

Q.4

That makes a lot of sense. Companies do change their platforms, even if they'd prefer to stick with them for a decade. While it might happen at a glacial pace, digital transformations are indeed taking place.

When a company decides to replace its LIMS, for example, it's a huge undertaking. They must overcome all the hurdles you mentioned – regulations, effort, cost, and investment.

If they've managed to jump through all those hoops, there must be a very compelling reason for them to do so. What typically pushes them over the edge?

At some point, older systems just won't be able to run on current platforms. Eventually, they will break.

Most companies are now cloud-based, though I still see manufacturing operations that are on-premise. The research side has moved to SaaS; I don't know any research organizations I deal with that haven't moved to the cloud. However, I still see resistance to moving to the cloud in manufacturing.

But with executives seeing the need to break down silos – because it's all about the data – we need to have things more centralized so we can use data across R&D, clinical, and manufacturing and make the best use of that data.

> ***These silos need to be broken down. So, people are starting to see, and it's coming from the top down, that we need to break them down for a competitive advantage, for speed to market, and so on.***

Q.5

You mentioned data, and since you have spent decades in this industry, I was wondering: since when have people started realizing the importance of it? When I read articles now, I get the sense that all research basically comes down to data.

It's so much so that the term "Biotech" has been flipped; now there are "TechBio" startups where tech geeks come in with quantum and physics ideas to find patterns in data, similar to what happened in Wall Street trading couple of decades ago. Historically, when did this trend start?

I think it's been a journey, but it's really picked up speed in the last seven years. That would be my take. Maybe 10 years ago, it was still a theory, but in the last seven years, it has become more actionable. And honestly, in the last three years, it's grown exponentially.

> ***People are starting to feel nervous, like "am I missing something?" I went to a forum recently where executives from a big-name pharma said, "We feel so far behind because we're not using agentic AI."***

There's so much hype and talk about using AI that executives are saying, "We need to roll this out. We need to make sure every scientist in the organization understands what it can do for them." So, in the last two or three years, I think there has been real pressure from the industry, with companies feeling like they are behind the curve.

Q.6

Big Pharma went through the patent cliff. Did that play a role? Are there any specific reasons you can think of? Something must have happened about seven years ago that pushed people to realize how important data is.

It's a natural evolution. I think it's also been influenced by the cross-pollination of executives from outside of biopharma who are now entering the industry and bringing a different perspective.

They look at what the industry is doing and say, "Wait, you guys are really far behind. We've been doing this for the last decade. Why aren't you?"

The pharmaceutical industry is far behind other sectors like banking or finance in terms of digital transformation. Why is that? I would go back to fear and regulatory issues that have really prevented people from embracing this transformation as quickly as other industries.

Q.7

But could it also be because finding a drug target is getting that much more difficult, and perhaps all the low-hanging fruit have already been taken? Is that something you hear from executives?

I think yes – it's gotten harder. I also think the new drugs coming out aren't just based on biology or chemistry alone. They're a combination of new modalities, such as cell and gene therapy or combination therapy. This means there's now a greater need to understand how data science can connect these different, siloed modalities.

> ***There's much more emphasis on how data can bring these things together, which is a shift from a time when a scientist might have focused solely on creating the next small molecule.***

Q.8

You interact with lot of professionals in the industry and exchange a lot of ideas. So, regarding the importance of data, how does the IT or Informatics side see it versus the scientists? Do they perceive it differently? Is there a gap?

Yes, I believe so. Scientists have typically focused narrowly on their own data and on owning it. In contrast, IT's perspective is often, "Let's do an enterprise-wide solution for everybody." But when you do something for everyone, each person feels like it doesn't do exactly what they need for their specific workflow.

> ***I've seen instances where IT just threw a platform over the wall, ripping out of scientists' hands what they were used to and giving them something that just doesn't work for them.***

What they should have done is brought people together so they could be part of the solution. People don't like being told what to use. But when they're involved, they might accept a solution that's not 100% perfect for them because they understand it's for the good of the organization – so that data can be used not just by them and their department, but by others as well. They need buy-in.

Instead of a two- or three-year transformation that people don't want to be a part of, organizations should look for quick wins to show value. They need to demonstrate that a new system won't replace a person's job. Instead, it will handle mundane tasks so they can focus on more creative, scientific work.

You need to sell them on the value it brings and the reason for the change.

Q.9

What's the mix like? Let's say out of 10 companies you've worked with, what's the ballpark breakdown? In how many did IT hold the power, and in how many did scientists hold the power?

I'd say it's about 50/50. The power dynamic can shift, especially when there are changes in upper-level management, as I mentioned earlier. At times, I've seen the business – that is, the scientists – have more power than IT, and at other times, I've seen IT have more power than the business. So, it really depends on the situation.

Q.10

Margaret, where do you see the industry heading in the next two to three years? What is your reading of the situation, given that you are picking up AI-induced anxiety signals from a section of the industry, and how do you see this playing out?

AI is not going away. I believe that organizations are now in the process of essentially trying to accelerate the deployment of AI and are doing more on education and enablement.

They are also understanding where there could be some quick wins and are getting people brought in and on the bandwagon. I think there is a bit of hype around AI, and that will settle down where it will become a part of everybody's day-to-day routine, and people will embrace it more.
I think that's where we are right now.

> *It's here to stay, but we need to make sure that people understand it is not replacing humans.*

It's making people more efficient and advancing science to ultimately improve human lives.

Interview: AI by Itself Isn't Going to Magically Deliver a Drug Today

In a recent survey by Zifo Technologies on Data Readiness for AI, it was found that only about one-third of scientists and informaticians feel confident in leveraging scientific data for AI initiatives, highlighting foundational data challenges across biopharma, which makes AI adoption that much more difficult in due course, if not addressed soon.

AI gained prominence and popularity after the introduction of OpenAI's ChatGPT: ordinary consumers got to experience the GenAI technology first hand, creating unprecedented buzz and excitement.

At the same time, one should remember that AI/ML has been around for a long time. In fact, the Finance industry is much ahead of all other industries when it comes to gainful deployment of latest technologies.

In the U.S. stock market, about 70% of trading volume is initiated through **algorithmic trading**, which is a forerunner for AI trading. According to Greenwich Coalition study, about a quarter of buy-side equity traders surveyed plan on incorporating internal AI technologies into their trade execution workflow in the coming year.

In this fast-evolving scenario, to gain a deep understanding of the current state of new technology adoption in the Biopharma sector, we spoke to **Sujeegar Jeevanandam, Zifo's Principal Consultant, Data and AI,** about its real-world impact, moving beyond the glamorous headlines of drug discovery.

Q.1

By its very nature, the biopharma industry avoids risk. It must, because of strict regulations, complex manufacturing, and data privacy. You don't see the kind of rapid tech adoption that Wall Street saw with algorithmic trading, which now uses such technologies to trade trillions.

So where does that leave biopharma with AI? Based on your interaction with various industry stakeholders, are they genuinely enthusiastic and making it a priority, or is their cautious nature creating a more lukewarm response? What's really driving their approach?

Firstly, I don't subscribe to the idea that biopharma is "risk averse." In my view, the industry thrives on risk. The sheer number of clinical trials that launch each year – and the high failure rate they accept – is proof of this. You never start a trial with a guarantee of success; it's inherently a betting game.

Q.2

Yeah, I take your point, maybe I should have qualified it as risk averse when it comes to embracing new technologies.

Correct. The industry tries to be cautious and control the journey they go through to take a drug to market. That's true whether it's with the talent they acquire, the partners they choose, the technologies they adopt, and, most importantly, the processes they follow. The idea is, "I am already taking such a high risk, so I'll go with a strategy that can mitigate some of those risks."

So, from a pure technology point of view, the industry can be seen as a laggard compared to other industries, like finance, as you mentioned. But when you think from a larger perspective, if you take a risk in one area, you try to mitigate it in another. Right?

> ***Now let me take up the second part of your question: the biggest driving factor, I feel, influencing the adoption of AI is FOMO – everybody feels they will miss out on this wave if they don't get involved now.***

Compared to how the industry adopted technologies like electronic lab notebooks (ELN), LIMS, cloud, and other technical advancements, I feel the pharma industry is taking a very driven approach to embrace AI. The ultimate question is going to be: how well will you adopt AI?

Q.3 ***Is it hype? I mean, if I ask you, is AI just hype, or are there genuine use cases for it?***

It's not hype. It has the potential to significantly reduce the cost of developing and bringing a new drug to market.

> *However, AI by itself isn't going to magically deliver a drug today. Everyone wants that – everyone wants to ask a question and have AI say, "Here's a drug; it's going to cure." That's not what I believe AI is capable of right now.*

But the opportunities for AI to create an impact in a localized manner throughout the drug's 10-year lifecycle in the industry are tremendous. If you can even create a 5% impact at various stages of the pipeline, it's going to translate into millions, if not billions, in savings. It will also reduce the cycle time to bring a new drug to market, and it has the potential to significantly reduce challenges and adverse events that can happen once a drug goes to market.

Q.4

Many companies already have IT systems in place, and AI essentially sits on top of them. You can use it to harvest data from various instruments and systems to reveal patterns or insights that scientists can study. So, AI's primary use is to extract these insights from existing data.

This brings us to the relationship between basic IT plumbing and AI. Think of it like algebra versus calculus in building a house. Algebra represents the fundamental building blocks – the bricks, cement, and tiles. You can't build a house without these basic materials. Calculus, on the other hand, is like the design phase: deciding where rooms go, how to arrange furniture, and so on. Many people try to jump straight to calculus without mastering algebra first, finding it difficult.

> ***That's because algebra provides the foundational building blocks applied in real-world situations by calculus.***
> ***Similarly, you can't do AI without basic IT plumbing. The question is, where do we draw the line? In what situations will basic IT plumbing solve the problem, and when do you need AI? For some situations, just like algebra is enough for certain problems, basic IT plumbing will suffice; you don't need to move to the "calculus" level of AI.***

I'm going to use a different example, if that's fine. I will use the analogy of a laptop and the program that we do in a laptop. To me, the LLM is the CPU. Billions of people use laptops. Most of us don't have to understand how a CPU works. They all, I hope, know that a CPU is critical for a laptop or a machine, but they have no knowledge about how a CPU functions. They still use the laptop very effectively, right?

When we talk about Gen AI in general, the work that all the tech leaders are performing today – OpenAI, Perplexity, and others – they are building a laptop for the industry to use, on top of the LLM, which is the CPU.

So, if I'm a team leader who wants to leverage AI, I don't need my team to understand the LLMs. I need my team to be capable of working with a laptop, right?

> ***So, what it means is LLM is yet another technology that is available for knowledge workers to perform their business function. So, in a way, yes, IT Systems and Data are the "Algebra" while AI application is the "Calculus".***

I used to start programming in BASIC and Mainframe, and when a new technology like Python and C# evolved, I embraced it. I learned it.

The tools changed, but the fundamental problems didn't change – I was still solving the same scientific problem for my customers. This means AI will eventually fold into this term that you called IT plumbing.

Q.5 ***Okay, in that case, scientists don't need to be an expert in AI or LLM, right?***

I don't think scientists have to go out of their way to become an expert in AI or LLMs or any technology to be able to use it.

If I know how to use a laptop – if I understand the purpose of the laptop, keyboards, mouse, and so on – I can do my job effectively without having to know the specifics. And that's what the industry and society are moving towards: making it easy for consumption. If you make AI easy for consumption, more people will consume it. This means a scientist doesn't have to worry about whether it's too tech-heavy for them to leverage. I don't think that's what customers, or any individual users, should worry about.

Q.6 ***When we talk about AI in biopharma, most people discuss discovery, specifically the drug discovery phase.***

Of course, that's the most often cited use case. However, I'm looking for some non-glamorous areas in biopharma where AI can be equally and potently deployed. Can you give two or three specific or general examples?

The most non-glamorous example I can think of is simply getting answers to questions.

We document, capture, and record every aspect of research. Sometimes, we have a question and know the answer lies in a specific document or experiment record. Today, you might spend hours trying to find that record. You get tired and decide to just redo the entire experiment to generate a new copy of the data for your reporting.

With AI, you could get those answers very quickly or at least be pointed to the exact location where the answer resides. I cannot overstate how much in savings we'll realize by being able to find a fact that we know already exists.

Beyond that, there are numerous opportunities to optimize operational activities, particularly with the help of Gen AI combined with digital twin capabilities.

Routine activities a scientist or analyst performs in the lab can be automated through AI. A good example is: I'm a scientist running an experiment. I observe something that I need to document. Today, I can't stop the experiment to do the documentation; I just hope I'll remember and diligently capture that observation at the end.

But with AI, I could ask ChatGPT or Copilot to make a note of that observation through a voice command.

The impact of being able to capture these observations, which could become incredibly valuable when developing new models and finding answers, is tremendous.

Q.7

Okay, so AI is not just at the large, glamorous end of discovery. It can be surgically deployed across the process, across the value chain, basically wherever there is a pain point in the entire R&D and manufacturing value chain. Are you saying that it could be surgically deployed to solve a particular pain point, and then one can be imaginative and come up with a use case to solve a problem?

> ***Absolutely. In fact, viewing the impact of AI only through the glamorous aspects will result in a lot of pain and disappointment.***

I do not believe that even if you put together all the knowledge from all the biotech and pharma companies in the world, we would have the information to accurately map human biology.

This is why every clinical trial is a risk; it's a lottery where you hope all the other pieces that you have not thought of would click into place.
And it is going to take quite a while, in my opinion, to be able to fully model human biology, to be able to say that "here's the question, here's the answer, and the answer is the cure."

I truly believe that is where the industry should be going. You have to leverage AI to completely model human biology. But we are not going to get that today. I don't even think we are going to get there in the next five years or so, maybe much longer after that.

Q.8

Beyond speeding up the trial (time as a benefit) or reducing its cost (cost as a benefit) – because everyone cites time and cost as the two distinct benefits for using AI – what other benefits could you think of from AI?

The third dimension, I'd say, is knowledge. Today, your knowledge space is limited by the talent you have.

This can result in focusing on a specific indication or pathway or sticking to traditional manufacturing processes to make a new drug, which falls under the CMC domain. The power of AI is that when used correctly, your knowledge space can grow manyfold. You could tap into potential areas that you are currently completely blind to.

For example, let's look at translational sciences, where we generate a huge volume of multi-modal data. Target identification and target validation are two core milestones at the beginning of a drug's life cycle.

Any pharma or biotech company has a specific indication or targets on which they have a huge knowledge base, and they continue to exploit that because of their in depth understanding.

Without AI, they might not even attempt to step out of this comfort zone to look at possibilities like the impact an approved drug could have on a totally novel target or an indication they aren't focusing on.
Why would they? They have limited resources and talent. They operate in a space where they believe they have the best chances of developing a new cure or drug. With AI, however, you can explore new targets and new indications with relatively minimal effort and time.

This expansion of knowledge, this opportunity to play on a much larger ground, comes at a cost. You need good quality data to be able to expand your knowledge. To me, compared to the cost and time benefits, knowledge benefits demand larger data capital. As we all know, that's a space where a lot of improvements driven by FAIR are happening and still need to happen.

Q.9

How about privacy concerns? Because it came up in the latest Zifo survey as well. We conducted a survey, and in it, privacy was mentioned as the top concern for why someone might not adopt AI, given that these are all third-party tools.

The concerns around privacy or security of data are no different from those brought upon when the industry was trying to adopt the cloud.

> ***I believe the industry has learned a lot about how we can protect IP and secure data, even if the data is stored or processed by a third party.***

Naturally, everyone has doubts about how various players are securing and isolating one customer's data from another, but I do not believe that is going to limit the adoption of AI by the pharma biotech industry. They are doing it already; they are using ChatGPT Enterprise version, for example, leveraging the cloud.

Many customers have their own implementation of these models on-premise or in their own private cloud, to be sure. However, we already see customers leveraging models hosted by third-party providers the same way they are leveraging cloud services hosted by AWS and Google Cloud.

Q.10

So, what does success look like? I know these are still very early days, but let's say a pharma company gets on the AI bandwagon – putting aside the fact that I know people will say, "Okay, we will discover that success" – let's take that glamorous part out of the equation.

From a non-glamorous perspective, how can they enumerate success in the sense that they have created value? What kind of KPIs can they show for it – because they're making investments that would require ROI. As you said, they will have to invest in data, data capture, and all of that is an investment.

What would a truly successful integration of AI look like, where the business value is shown, and that does not necessarily mean faster drug discovery? So, if you take that out of the equation, what would success look like?

As a life sciences consultant, I see AI becoming a norm in the research and development process, the way ELN and LIMS have become the norm. If anyone isn't using ELN and LIMS, they're an outlier; they're seen as a laggard.

When the industry treats AI adoption with the same lens, it means AI has become core to life sciences research .

> ***However, there are still miles to go because interestingly, even today, believe it or not, there are multiple pharma and biotech companies that still run on paper. They have hundreds, if not thousands, of scientists documenting everything on paper, not even using ELN.***

Q.11 ***You mentioned ELN, how about LIMS?***

I would say a lack of a proper LIMS means you're sending a lot of instructions through emails, while depending on handwritten notes and Excel sheets. This shows how slow the industry has been in the past when it comes to technology adoption.

ELNs first rolled out at the beginning of the millenium. Twenty-five years later, while some people still use paper, I observe significant adoption of AI in the industry and among the customers I have worked with.

So, I do not believe the industry is going to take 25 years to adopt AI. That's why I do not believe the risk-averse nature of the industry is impacting or hindering the inclusion of AI into research and development.

Q.12

So, you're saying, from what you have observed, the adoption of AI is faster, or rather robust than adoption of say, ELN or LIMS?

Yes, that's correct. The rate of AI adoption is faster than ELN or LIMS adoption, in my opinion.

Q.13

Everyone understands the importance of AI – nobody's disputing it. But if you take any large biopharma company, leaving aside the startups, if you take the large ones, that's where the complexity lies. So, who's responsible for this AI strategy? Because they'll be working on different targets, different projects, different teams, different systems, geographies, languages, and whatnot.

So, who's responsible for coming up with an overall AI strategy and then also deciding when, where, and how to use AI?
Do you see any such thing happening in the industry currently?

Or is it all being done individually by the teams themselves, with no overarching, centrally driven mandate from top management saying, "Okay, use AI here and here"? Or is it more decentralized?

> ***I'm going to give you a very boring answer: everyone is responsible for incorporating AI into their operations.***

However, from what I have seen with my customers and across the industry, there is a top-down approach. Organizations are setting up specific functional units to drive and incorporate AI across the enterprise.

These units are reporting directly to the Chief Scientific Officer (CSO) or even the CEO to ensure there is an enterprise-wide strategy for AI incorporation.

But that kind of a top-down approach takes a long time to reach the "leaf nodes," and the impact is also not very tangible. Some of the most innovative pharma and biotechs are also taking a bottom-up approach, enabling individual teams to be able to fast-track value realization through AI and not have to wait for the entire organization to embrace it.

This kind of "pocket approach" has its own challenges. We talk about privacy, security, the cost, and how you define success.

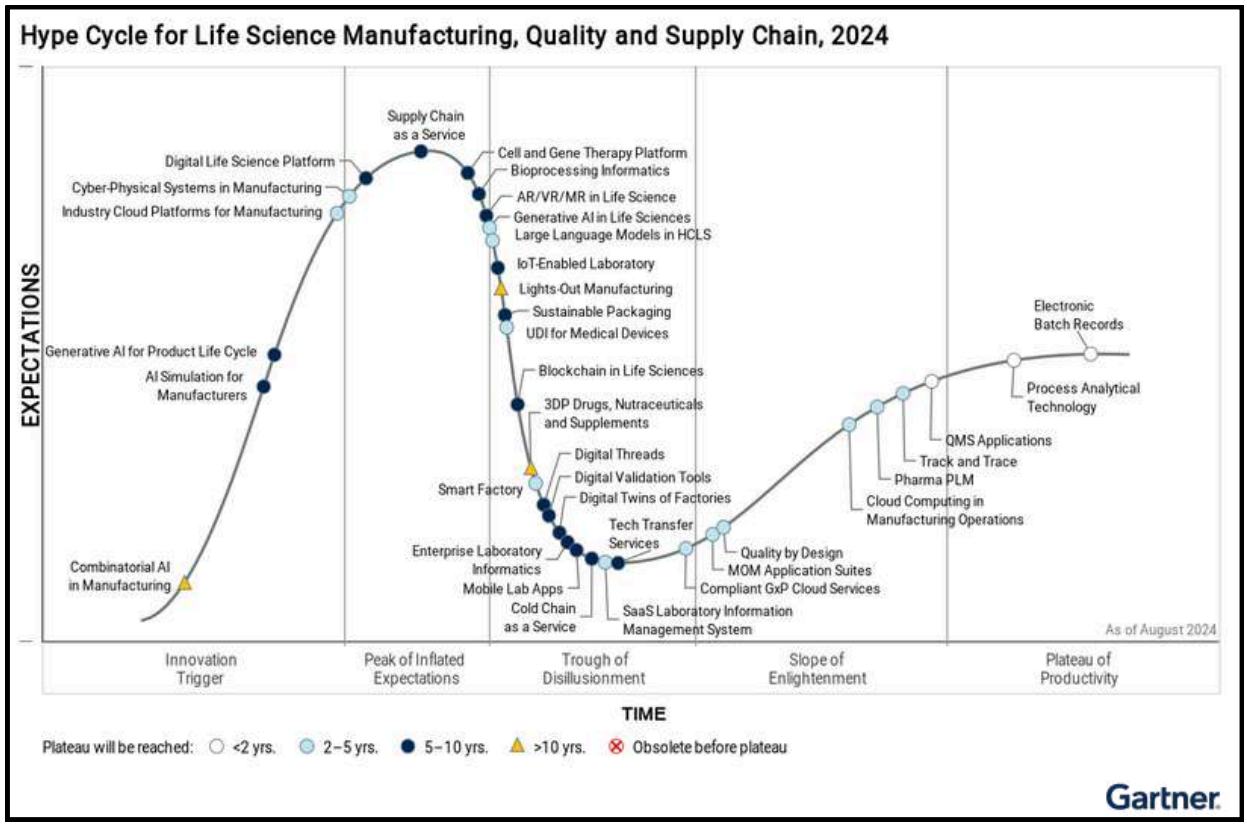
But when done the right way, these pocket exercises bring increased momentum to the organization's drive towards AI.

Ultimately, I think the boring answer is that everyone, from IT to business to scientists to the CEO, they all have to embrace AI, find ways to integrate it into their workflow. It means asking questions, seeking help, or simply starting to adopt it.

Beyond the Hype: Biopharma's Tech Transfer Process Needs a Fundamental Reset

In a Hype Cycle report released by Gartner, Tech Transfer has been placed in the "Trough of Disillusionment." This signifies that the promised benefits from solution providers are not being delivered, leading to disappointment among customers.

The suggested path out of this disillusionment is to bring in external help by collaborating with IT service providers to get the process back on track.



To gain a deep understanding of the current state of tech transfer – an important segment of the biopharma value chain – we spoke to **Adam Paton, Zifo Technology's tech transfer and life sciences informatics expert**, to learn what can be done to help companies struggling with the tech transfer process.

Q.1 ***1. What is tech transfer in simple terms – explain it with couple of practical examples.***

Tech transfer is the vital process of transitioning a successfully developed item, such as a new product or compound, from one area of an organization to another. Its primary application is often seen when moving a product to the manufacturing floor for commercial-scale production.

This process, as the name suggests, involves the systematic provision of technology and information between different points in a chain. A practical example is the journey of an analytical method: from its initial development (often called method development) to its validation, documentation, approval, and finally, its transfer into a production facility.

> ***The goal is to ensure that what has been carefully created and verified can be reliably scaled up for industrial use.***

In simpler terms, consider the passing down of a beloved family cake recipe. When a grandmother shares her recipe with her grandchild, she's not just giving them a list of ingredients. She's transferring comprehensive knowledge: the tools, the specific steps, and even what the finished cake should look like. This complete sharing of information is, at its heart, tech transfer in action.

Q.2 ***Considering the vast amounts of data generated during R&D, how does early planning for data transfer and integration (e.g., data format standardization, metadata management etc) minimize costly delays and errors during later tech transfer stages?***

Tech transfer, at its core, is about knowledge transfer, which is fundamentally data.

> ***If we consider the journey from a scientific hypothesis to a new medicine, it's a path paved with significant time, effort, skill, and investment, all geared towards generating vast amounts of data at every step.***

Adding to this complexity is the rise of new modalities like biologics, gene therapies, and cell therapies. These aren't simple chemical molecules; a single cell contains millions, even billions, of data points compared to the relatively few elements in a traditional medicine like aspirin. This means the sheer volume of data is exploding across research, development, and manufacturing.

Furthermore, advancements in lab equipment and automation are accelerating data production even more. Consequently, data isn't just crucial; it's becoming a major bottleneck.

> ***If data remains siloed and isn't considered from end-to-end, trying to piece it together retrospectively becomes incredibly difficult and unsustainable. This approach, while perhaps viable decades ago, is no longer feasible today.***

Therefore, we must prioritize strategies that foster incremental optimizations and improvements towards data portability from the very beginning. Even seemingly simple measures, like establishing consistent terminology across research, development, and manufacturing, can have a profound and positive impact on this process.

Q.3

From what I understand, this sounds like insurance. Everyone understands the importance of insurance, and no one would dispute the value of a good insurance contract. However, sometimes people still fail to secure one, right? So, why do companies miss out on these crucial, seemingly obvious steps, and subsequently find themselves unprepared?

That's an excellent question, and it points to what I see as an endemic problem in how life sciences sector has embraced the digital age.

The prevailing mindset, often epitomized by Electronic Data Capture (EDC), has historically focused on capturing information after an event has already occurred.

The issue with this approach is that it's reactive: you're addressing a problem after it's been created. While data is indeed being captured, its value is diminished because it's not being done "in-flight." This leads to a massive, often retrospective, effort at the end of a process.

> ***The attitude has largely been, "if it's not broken, don't fix it" – this method has worked for decades, so why change?***

However, a closer analysis of current operations reveals this approach is hugely inefficient. I believe much of this stems from historical practices and a lack of awareness. Fortunately, the emergence of newer technologies and the current disruption within the industry are now bringing this critical issue to the forefront, demanding that we address it.

Q.4 ***Okay, so if I understand correctly, you're saying the very process of data capture thus far happens after the data is created. Is that right?***

Yes, exactly. This is an area we're constantly optimizing within the broader concept of tech transfer, particularly through the lens of manufacturability.

While it might be a somewhat contentious term, "manufacturability" serves as an excellent catch-all for what tech transfer aims to achieve in a modern context. It's highly beneficial to consider manufacturing concepts as early as possible – ideally, right at the sharp end of discovery. Doing so can significantly accelerate time to market.

> ***Within drug discovery, specifically in the pharmaceutical industry, a high percentage of drug candidates, often in the high 60s, fail at Phase 2 of clinical trials. This is after years of investment, millions of dollars, and immense time.***

Crucially, some of these failures aren't due to side effects or fundamental chemical/biological non-viability. Instead, they fail because they simply cannot be produced in a commercial setting.

Imagine at the very beginning, during target selection and validation, you identify a promising candidate that effectively cures an illness or shows high efficacy. But then you realize it's incredibly toxic to handle, prohibitively expensive, or requires freezing to -50°C.

Immediately, questions arise: Is this truly viable for a manufacturing plant?

By tagging these "manufacturability properties" as early as you're running experiments, you introduce critical checkpoints. Someone in a tech transfer cross-check, even in the early stages, could identify a candidate as 'unmanufacturable' and suggest cutting losses then and there.

Instead of spending years and millions to reach Phase 2 only to arrive at the same conclusion, you can stop after, say, an initial $1 million investment. I'm oversimplifying, of course, but anything that introduces this level of high-quality, attributed information and knowledge upfront can only help.

Currently, these critical decisions often happen too late, right at the very end of the process, by which point significant resources have already been expended. Integrating this proactive assessment into how people work from the outset represents a massive and necessary step change for the industry.

Q.5 ***What are the fundamental reasons for things to go wrong in a tech transfer process – can you explain this with a practical situation that you have witnessed over the years?***

Sure, for example, a common issue is the 'walls' that get put up between different parts of a biopharmaceutical company's process development. Traditionally, we have about four or five areas: cell line development, upstream processing, downstream processing, formulations, and fill and finish (which is the release of the product).

I have heard many, many times, and still hear frequently, even just between upstream and downstream, comments like: 'Oh no, that's upstream,' 'No, that's downstream,' or 'That has nothing to do with me; I'm not interested in that.'

> ***They should be highly interested in and invested in what's happening in other areas. These are two adjacent functions; they directly pass material samples from one to the next.***

These are colleagues who see and work with each other, yet in their minds, they are completely different operating units. In some cases, they actively do not care what happens in the other area.

If we consider extending this all the way from the sharp end of discovery to the manufacturing floor at the very end, which is a separate entity, you constantly run the risk – and it happens all the time – of people having no consideration or concept for what happens downstream or upstream of them.

Therefore, anything that can be done to foster a greater appreciation for other functions, or to establish a consistent naming convention across the organization, or to harmonize and understand processes that move from one stage to the next (including potentially having tooling that facilitates this), represents a huge area for potential improvements to optimize tech transfer.

This all leads back to something we were discussing earlier: the importance of doing it in-process rather than retrospectively.

Q.6

That's well explained. It's almost like having a manufacturing expert embedded from the very beginning to assess the viability, isn't it?

Yeah, and you know what? Think about those 'back to the floor' shows on TV, where the boss goes and sees how their teams really work. Something like that could bring massive benefits, not just for tech transfer to manufacturing, but everywhere.

When someone from one part of the business gets to work in another, it's like they're doing a crucial reality check. They can say, "This all looks good, but we missed something here. The information I'm getting doesn't cover it. I'm bringing it up now because it could save us a lot later."

Q.7

Do companies have a dedicated position for tech transfer? Based on your interactions with different stakeholders and companies, how is it typically structured? Is there someone specifically assigned to these tasks, or does it end up being an 'orphaned' project no one really wants to touch? Basically, where does this fit in the org chart?

We can evolve this to a "process, data, organization, and technology" (PDOT) framework. I'll make the distinction between "people" and "organization" shortly. If we start with the process, understanding it is crucial for success.

> ***You cannot move something from one person or group to another without a clear understanding of what's being done.***

You need a well-defined, documented, and transferable script, steps, and process flow. Without this, it's like starting all over again. If you tell someone what's been done and what they need to do, but don't provide clear instructions and confidence in the process, their natural human instinct is to say, "I'll do it myself." This, however, leads to a loss of valuable information surrounding the process.

Therefore, we must prioritize the process. If it's done exceptionally well, with all the necessary information and artifacts, it will be more readily accepted.

However, the process is just one piece. Moving on to the organization, which was more central to your question, it's not just about the individuals but how the entire organization operates. We must consider the people, geography, company culture, organizational setup, and strategy.

All of these contribute to a more effective knowledge-sharing strategy; you must consider them together. Is it currently one person handling this? Sometimes. Is it sometimes not? Yes. Should it become part of the organizational thinking? Absolutely.

> ***If you have alignment on as many aspects of the organization as possible, the rate and degree of change can be massively accelerated.***
> ***This represents a significant shift in how people need to think.***

An important aspect is integrating business value into the thinking around processes and organization. If you're moving towards an operating model that allows the business to operate more effectively, more profitably, more agile, or whatever your objective may be, people will then see that everyone is working towards a common goal.

While it might sound slightly cliché, ultimately, if this can be achieved to generate products and therapies faster, and deliver them to patients earlier and more effectively, then everyone in this space is essentially achieving their fundamental goal.

Q.8 ***Can you give me a ballpark figure as to what percentage of organizations as per your estimate have implemented PDOT framework?***

Having it completely sorted? I would confidently say 0%.

However, when we talk about organizations making inroads and doing great things in some or a number of these areas, it's difficult to put an exact number on it, but it's probably lower than you'd think.

Considering all aspects, you're likely only looking at 20% to 30% of organizations truly adopting this type of thinking.

This is, in fact, a symptom of the life sciences industry itself. We discussed the PDOT framework, and it's the 'T' – technology – where people tend to gravitate. Technology offers a solution; it's the 'solutionizing' aspect.

If you're a scientist, your brain is almost hardwired to try and find solutions – that's what we do, solve problems.

> ***So, it's very easy to jump to technology, thinking, "Oh, this piece of technology solves a problem; it gives me what I'm looking for." But if your processes aren't correctly defined and optimized, and if your organization isn't bought into what you're trying to achieve, technology can be useless.***

Many organizations are now realizing this: to successfully implement a data strategy linked to technology, you must also have robust processes and well-managed data.

It's a significant shift, I would say. Now, more organizations are probably focusing on the 'D' (data) and 'T' (technology), which might be higher percentages. But when you consider the 'P' (process) and 'O' (organization), those numbers are much lower. So, as a collective, the overall integration across all four elements remains quite low.

Q.9

So, what kind of role would a CSO (Chief Scientific Officer) play in this? Does Tech Transfer process fall under the CSO's domain, or the CIO's (Chief Information Officer's) domain?

Yes, it largely depends on the organization's culture and operating model; it could indeed fall under the purview of either, or both.

> ***However, I hold a specific viewpoint: the CSO (Chief Scientific Officer) should be instrumental in any activity within our domain.***

Everything we undertake should ultimately aim for a scientific business outcome. While this outcome can certainly be linked to efficiencies, time savings, or cost reductions, a far more transformative impact would be, for example, developing a cell therapy believed to cure spina bifida.

We might be close, but if the therapy is currently unviable due to prohibitive cost, preventing patient access, then cost becomes a major blocker.

In such a scenario, we must re-evaluate the science used to generate that therapy. What can we do scientifically to make it a viable treatment for patients?

This necessitates looking at the science, the scientific model, and even the cell model itself in a completely different light, integrating it with data, technology, and our organizational framework. This crucial drive must come from the CSO.

Q.10

For all-encompassing projects like Tech Transfer, which spans from early discovery to manufacturing, do you think a separate, dedicated team is needed from the outset? This team could be either internal or external.

The idea is for them to embed people across the value chain to ensure data and processes are captured effectively. This would allow specialists – like chemists focusing on chemistry or molecular biologists on biology – to concentrate on their core tasks.

Meanwhile, someone else would shadow them in the background, capturing data and parameters. This overall view, perhaps at a dashboard level, would confirm things are on track and flag potential pitfalls.

I know this sounds futuristic, but is anything like this already happening in the industry? If not, do you think it's an approach companies should consider?

You've hit on a crucial point: scientists already have demanding day jobs. This is precisely why the life sciences industry isn't further along in its digital transformation. New activities are constantly being shoehorned into existing core roles, which isn't sustainable.

> ***The ideal scenario would involve a separate, dedicated team specifically measured on delivering incremental and transformational change for areas like tech transfer. However, this requires significant investment, and whether an organization commits to it depends on its culture, size, and various other factors.***

In an ideal world, the impact of such a team would be considered substantial enough to warrant the investment. For me, this isn't merely a top-down or bottom-up decision; it requires a C-level mandate.

A Chief Scientific Officer (CSO) or Chief Information Officer (CIO), for instance, needs to declare: "This is our objective for business outcome – perhaps better science – and to achieve it, we are prepared to mobilize this specific type of structure." This top-level commitment is essential to drive such a fundamental shift.

Navigating AI Regulations in GxP: A Comparative Look at EU AI Act, EU Annex 22 & FDA AI Guidance

As artificial intelligence increasingly transforms the life sciences landscape, regulatory clarity is more and more critical — particularly in GxP-regulated environments. The incorporation of AI into drug development, production, and quality systems requires not just innovation but careful control to guarantee patient safety and data integrity.

As global regulators move to set limits and expectations, comprehension of the subtleties of new AI frameworks is no longer a nicety — it's mandatory for compliance professionals and technology planners both.

This article presents a comprehensive comparative analysis of the three most important AI regulation realized so far — **EU AI Act (finalized), EU GMP Annex 22 (draft), and FDA AI Draft Guidance – Considerations for the Use of Artificial Intelligence to Support Regulatory Decision – Making for Drug and Biological Products (draft)** — with a key focus on their significance to GxP-regulated environments and implementation-viewpoint insights.

Navigating AI Regulations in GxP: A Comparative Look at EU AI Act, EU Annex 22 & FDA AI Guidance

S. No.	Implementation Area	EU AI Act	EU GMP Annex 22	FDA AI Draft Guidance
01	Intended Use Definition	Acknowledge documented intended use and register in EU database for high-risk AI systems.	Detail intended use, including characteristics and limitations of input data. SME approval prior to testing.	Define the question of interest and context of use (COU). Clarify role of model and scope in decision-making.
02	Risk Classification & Model Type	Categorize AI systems as high-risk on the grounds of impact on health, safety, and basic rights. Emotion recognition in work environment and predictive profiling are the prohibited practices.	Static deterministic models permitted in mission-critical GMP applications only. Dynamic/adaptive models and probabilistic outputs are not permitted.	Measure model risk by model influence × decision consequence. Utilize risk matrix for establishment of credibility assessment rigor.
03	Testing & Validation	High-risk AI products need conformity assessment (third-party or internal). CE marking and post-market surveillance are required.	Establish test measures (e.g., F1 score, accuracy, sensitivity, specificity). Acceptance levels need to be at least as good as replaced process. Record test plan, deviations, and keep all records.	Provide performance measures with confidence limits. Specify evaluation techniques, test data independence, and consistency with observed data.
04	Data Governance	Training/validation/testing data should be high quality, representative, and bias-reduced. Special categories of personal data can be processed under strong protections.	Use representative, stratified, validated test data. Generative data should be avoided unless warranted. Data independence and audit trails must be ensured.	Data has to be fit-for-use: applicable, trustworthy, traceable. Document data management practices and bias reduction.
05	Explainability & Transparency	User instructions need to contain system capabilities, limitations, risks, and control mechanisms. Transparency needed in synthetic content and biometric systems.	Feature attribution with SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), or heatmaps. Explain why features are relevant.	Write down model architecture, features, training reasons. Add uncertainty quantification and explainability tools.
06	Human Oversight	Mandatory human oversight for high-risk AI. Biometric systems require dual human verification unless legally exempt.	HITL required for non-critical applications. Operator training and performance monitoring mandatory.	Human-AI team performance must be evaluated. Oversight roles and responsibilities must be defined.
07	Lifecycle Management & Change Control	Change control and post-market surveillance necessary. Significant changes initiate reassessment.	Adopt change and configuration control. Keep track of performance drift and sample space of input. Retest if there are changes.	Lifecycle maintenance necessary. Retrain/revalidate in case of change in performance. Document changes according to regulatory guidelines.
08	Regulatory Engagement	Interact with AI Office and federal authorities. Apply regulatory sandboxes for testing and innovation.	No formal engagement processes established; follows Annex 11 principles.	Early interaction highly recommended. Apply Pre-IND, CID, ISTAND, and other FDA initiatives.
09	Documentation & Traceability	Keep technical documentation, EU declaration of conformity, and register systems on EU database.	Keep all test documentation, access control logs, and audit trails.	Prepare credibility assessment report. Cover model development, evaluation, and deviations.
10	Confidence & Thresholds	Specify thresholds for high-risk AI forecasts. Employ confidence scores to ascertain reliability.	Store confidence scores. Utilize thresholds to mark undecided results.	Estimate uncertainty and confidence levels. Incorporate into performance measures.
11	Personnel Qualification & Training	Providers and deployers must ensure that personnel who take part in the operation of AI systems are AI literate and trained.	All personnel working on AI lifecycle should have their roles defined and appropriate qualifications.	Personnel working on model development and monitoring should be trained and qualified.

Navigating AI Regulations in GxP: A Comparative Look at EU AI Act, EU Annex 22 & FDA AI Guidance

S. No.	Implementation Area	EU AI Act	EU GMP Annex 22	FDA AI Draft Guidance
12	SME & Innovation Support	SMEs and start-ups receive priority access to sandboxes, simplified QMS, reduced fees, and standard templates.	Not specifically mentioned.	Not specifically mentioned.
13	Real-World Testing	Real-world testing allowed under close control; requires testing plan, consent, supervision, and registration.	Not applicable; testing must be controlled and documented.	Real-world monitoring of performance encouraged as part of lifecycle maintenance.
14	Biometric & Emotion Recognition Systems	Emotion recognition in educational/work environment not allowed; judicial approval and double human endorsement required for biometric systems.	Not applicable; emotion recognition and biometric categorization excluded.	If used, justification and tested for bias and risk required.
15	Fundamental Rights Impact Assessment	Compulsory for public deployers and private deployers of high-risk AI in high-concern domains (e.g., healthcare, education, law enforcement). Optional.	Optional.	Risk-based credibility testing involves ethics.
16	Adversarial Testing & Robustness	Suppliers of general-purpose AI models with systemic risk are obliged to perform adversarial testing (internal or external) before market launch.	Not clearly mentioned but robustness is implied through validation and performance tracking.	Sponsors ought to test model robustness and resilience to input variation and overfitting.
17	Cybersecurity & Model Protection	Cybersecurity controls required for high-risk and systemic-risk models, including model leakage protection, access control, and tampering.	Configuration control and detection of unauthorized change required.	Sponsors must ensure model integrity and security throughout lifecycle, especially in production.
18	Environmental Sustainability & Ethical Design	Voluntary codes of conduct promoted for energy-efficient AI, inclusive design, and ethical development.	Not mentioned.	Ethics are part of credibility assessment but not specifically linked to sustainability.
19	Stakeholder Participation & Inclusive Development	Encourages participation of civil society, academia, and various development teams in AI system design.	Requires SME, QA, IT, and data scientist collaboration.	Sponsors need to have subject matter experts included for model development and risk evaluation.
20	Voluntary Codes of Conduct for Non-High-Risk AI	Non-high-risk AI system providers and deployers should voluntarily adopt high-risk requirements (e.g., transparency, oversight, documentation).	Principles may be applied to non-critical GMP uses with HITL.	No official voluntary scheme, but sponsors may apply credibility principles to non-regulatory AI application.
21	Transition Periods & Legacy Systems	Market-placed AI systems before Aug 2026 must be compliant if significantly modified. Public sector systems must comply by Dec 2030.	New systems only; legacy systems may potentially need revalidation if AI is introduced.	No transition period is provided, but changes during the lifecycle must be documented and justified.
22	Serious Incident Reporting	Providers shall report serious events (e.g., harm to health, disruption of infrastructure, infringement of rights) to the competent authorities.	Deviations and failures shall be logged and investigated.	Sponsors shall track and report side effects associated with the application of AI models.

Final Insights for GxP Professionals

EU GMP Annex 22:	Provides straightforward validation guidelines for static AI models deployed in GMP production environments. It emphasizes documentation, explainability, and human oversight to ensure product quality and patient safety.
FDA Draft Guidance:	Provides a framework for credibility to make AI models credible in regulatory submissions. It applies a risk-based methodology with context of use and early collaboration with the FDA.
EU AI Act:	Requires lawfulness for high-risk and general-purpose AI, with emphasis on safety, transparency, and regulation. It brings strict obligations to providers, such as risk management and technical documentation.

Conclusion:

Yet the regulatory landscape for AI is far from settled. While the EU AI Act is finalized, the EU GMP Annex 22 and the FDA AI Draft Guidance are still in draft form. Both are open to revisions on the basis of industry input, advancements in technology, and shifting risk views.

Organizations therefore need to remain adaptable and up-to-date — because draft today may become policy tomorrow.

References:

1. The EU AI Act – *https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689*

2. EU GMP Annex 22: Artificial Intelligence – *https://health.ec.europa.eu/document/download/5f38a92d-bb8e-4264-8898-ea076e926db6_en?filename=mp_vol4_chap4_annex22_consultation_guideline_en.pdf*

3. FDA: Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products – *https://www.fda.gov/regulatory-information/search-fda-guidance-documents/considerations-use-artificial-intelligence-support-regulatory-decision-making-drug-and-biological*

Finding Our Google Maps Moment in Translational Genomics

The growth of large, well-defined biobanks like UK Biobank, FinnGen, and the All of Us Research Program has changed what is possible in the field of genetics research.

Researchers now have access to huge genetic and phenotypic datasets that cover entire populations and can show the genetic structure of disease in ways that have never been possible before. The promise is clear: enable discoveries, make drug development better, improve patient stratification, and eventually make precision medicine a part of everyday healthcare.

However, even though there is a lot of data, the path from biobank to biological insight is still very fragmented. In Translational Genomics, researchers today are in a strange situation – they have more data than ever, but the tools they have to find, combine, and use this data often don't work well.

Sandor Szalma, *Scientific Advisor at Zifo*

Metaphorically speaking, bridging large-scale genomic data to biological insights is like driving through a dense city – each path may lead to a different destination, whether it's a Polygenic Risk Score, a Biomarker, a Drug Target, or a Pharmacogenomic Guide.

The challenge lies in choosing the optimal route and interpreting the signs along the way.

Our current tools are like paper road maps – helpful, but not very flexible or connected. This is where the “Google Maps” metaphor makes sense.

Translational genomics needs a system that is dynamic, interactive, and layered, just like the map app lets us zoom in and out, see traffic and adapt the route, find alternative routes, add stops, and share our journeys.

A system that not only shows where associations are but also guides researchers from population-scale signals to interpretable biological insights.

From Atlases to GPS: Why Current Tools Fall Short

GWAS and related analyses have been the workhorse of human genetics, producing summary statistics that capture associations between genetic variants and traits. You can now query, visualise, and share these results thanks to public resources like the GWAS Catalog and open-source tools like PheWeb, Open Targets, and GWAS Atlas.

Manhattan plots and QQ plots, for example, are now well-known places on the map of discovery. Of course, the most sophisticated amongst these is Open Targets which integrates genetic associations data with a vast array of other biological evidence types (e.g., functional genomics, expression data, pathways, animal models, drug data) and a suite of tools such as fine-mapping, colocalization, and locus-to-gene (L2G) prioritization using machine learning to support in silico target triaging.

The software is available for download and local implementation, but it needs dedicated effort, and customizations can interfere with the evolution of the package by the Open Targets Consortium roadmap.

Hence, these tools aren't enough. At best, they are static snapshots that are often optimised for one dataset and are not often made with scalability, collaboration, or integrative analysis in mind.

> ***A researcher who wants to go from a GWAS hit to a possible drug target may have to use a lot of different platforms, download datasets, run command-line scripts, and connect the dots between genetic signals, biological pathways, and phenotypes by hand.***

The trip takes a long time, and the path is often unclear.

Biobanks: The New Cities of Genomics

Think about the biobank landscape itself to get an idea of how big the problem is. These massive datasets are like huge new cities: they have a lot of information, but it's easy to get lost in them if you don't have the right tools. The UK Biobank has genomic and longitudinal phenotypic data on 500,000 people. FinnGen combines genetic information with health records that go back decades. All of Us wants to sign up more than a million people from different backgrounds in the U.S. They make up an unprecedented map of human health and disease.

Privacy restrictions often make it hard to get to individual-level data, so summary statistics are usually the best way to do large-scale analysis.

> ***Summary statistics are powerful, but they need advanced downstream analysis – fine-mapping, colocalization, variant annotation, and Mendelian Randomization (MR) – to turn associations into biological meaning.***

These steps become roadblocks when workflows aren't integrated.

What a Google Maps for Genomics Would Look Like

So, what would it mean to have a Translational Genomics version of Google Maps? A few ideas stand out:

Zoom In, Zoom Out:	This feature lets you customise your visualisation beyond basic Manhattan/QQ plots. You can compare multiple traits, explore different regions, and add functional annotations on top of each other.
Interactive and Intuitive:	Tools that let you explore, test hypotheses, and see things in real time, not just in static plots.
User-friendly Upload and Management:	Of any GWAS summary statistics (private or public) including user management.

Integrated Downstream Analysis Workflows:	The platform integrates with pipelines for algorithms like MR, fine-mapping, and variant annotation , so you don't need to know a lot about the command line or download data for each step.
Ready for Collaboration:	Shared workspaces where research teams can mark up and talk about their findings in the same way that maps today let people share routes and get live updates.
Scalable and Extensible:	A platform that can add custom analysis and visualisation tools and interactive performance does not degrade significantly by increasing the size of the variants and phenotypes/studies.

Such a system would not only help researchers navigate today's complexity but also accelerate the journey from biobank-scale data to meaningful biological discovery – guiding scientists to see not just where the roads are, but which ones are most promising for translational insights and therapeutic discovery.

The Road Ahead: Why This Is Important Right Now

It is very important to build this next-generation infrastructure as soon as possible. Drug discovery pipelines are relying more on human genetics to derisk targets to improve the probability of technical success.

Robust interpretation of genomic data across populations is essential for patient stratification and risk prediction. And for precision medicine to work, we need to find a way to connect large-scale research with personalised care.

The promise of biobanks could be put on hold if navigation tools don't get better.

> ***Data will keep piling up, but insights will stay hidden behind technical problems, broken workflows, and separate systems.***

The opportunity is clear: we can unlock the full potential of biobank-scale research by thinking of new ways to manage, analyse, and visualize genomic data.

It's Time For a Google Maps Moment

Translational genomics is at a turning point. We have the data, the computing power, and the ambition. What we need now is a platform that is integrated, intuitive, and collaborative – one that can take us from biobank-scale data to discovery-ready insights.

A “Google Maps” for genomics could change the way we understand the complexity of human health in the same way that GPS changed how we get around cities.

The goal is clear: precision medicine for everyone.

It’s time to make the map that will take us there.

Foundational Models in Single-Cell Omics

A New Era of Biological Understanding

The field of single-cell omics has been revolutionized by the advent of foundational models, ushering in a new era of biological understanding and analysis.

These powerful machine learning tools are transforming how we interpret and utilize the vast amounts of data generated by single-cell technologies, offering unprecedented insights into cellular biology and gene expression patterns.

What Are Foundational Models in Single-Cell Omics?

Foundational models in single-cell omics are large-scale machine learning models, typically based on transformer architectures, that are pretrained on massive datasets of single-cell and spatial transcriptomics data.[1]

> ***These models aim to learn a unified cell representation that captures complex relationships between genes and cells across various tissues and conditions.***

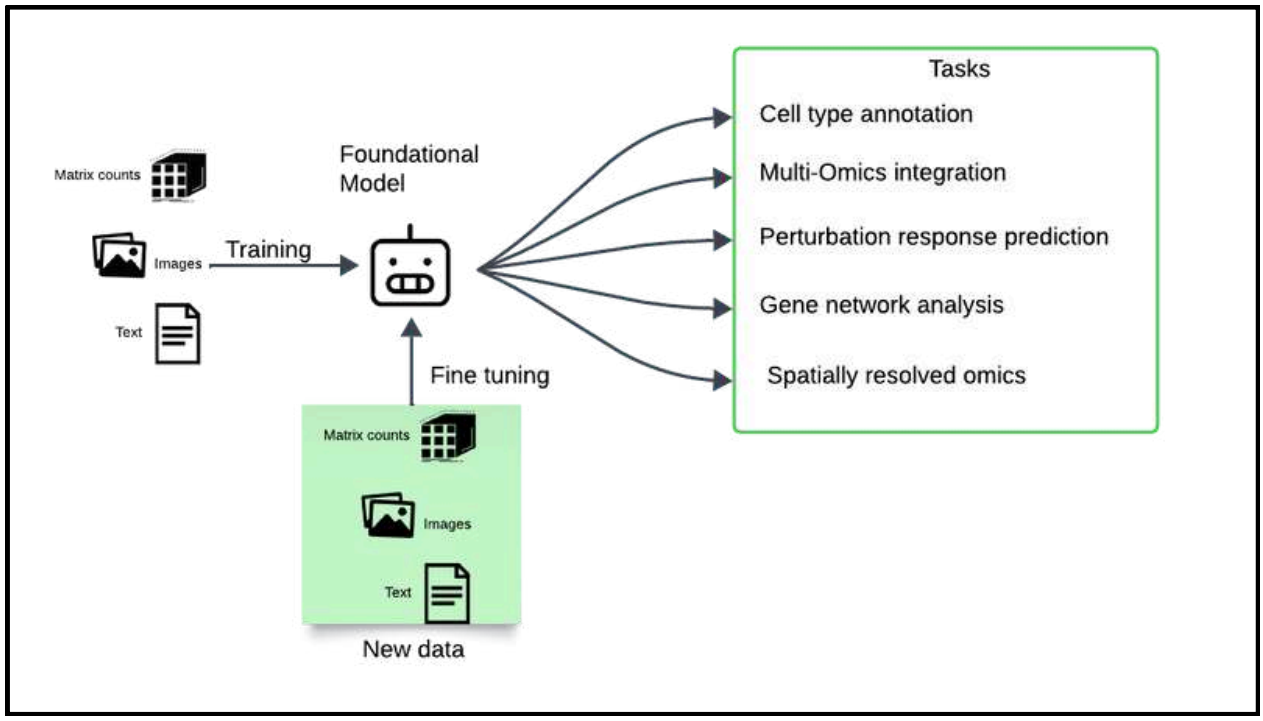


Figure 1:
Foundational models in omics analysis: A snapshot of their diverse applications and basic steps.

A deep dive into foundational models reveals their technical marvels. Transformer architectures, originally designed for natural language processing, have been adapted for single-cell data analysis.

> ***In these models, each cell is treated as a "sentence," with its genes representing "words."***

The models use special tokens to denote cell identity and position, with gene expression values often encoded as rank values. This approach allows the model to capture the complex interplay of gene expression patterns within a cell.

The self-attention mechanism in transformers enables the model to consider the relationships between all genes simultaneously, making it particularly well-suited for capturing the intricate dependencies in biological systems.

Transformer architectures in single-cell omics models have been further refined to handle the unique challenges of biological data. The self-attention mechanism has been adapted to capture gene-gene interactions, treating genes as tokens and computing attention scores between them.

This allows the model to identify influential genes for predicting others' expression, effectively modeling gene interaction networks. Some models, like scMoFormer, employ multiple transformers to handle different data modalities, with a cross-modality aggregation component bridging these transformers.

To address computational challenges, linearized transformers have been implemented to reduce complexity when dealing with large numbers of cells.

Additionally, these models often incorporate domain-specific knowledge about genes and proteins, enhancing their biological relevance. The tokenization of gene expression data, where each cell is represented as a sequence of gene tokens ordered by relative expression levels, helps preserve gene-gene relationships while mitigating technical batch effects.

> ***This sophisticated approach enables foundational models in single-cell omics to capture complex patterns across various cell types and conditions, leading to improved performance in tasks such as cell type annotation, gene regulatory network inference, and prediction of cellular responses to perturbations.***

Key Characteristics of These Models Include:

1. Large-scale Pretraining	Models like scFoundation and Nicheformer are trained on tens of millions of cells from diverse tissues and organisms.[2]
2. Multimodal integration:	Some models, such as scGPT and Nicheformer, combine data from both dissociated single-cell and spatial transcriptomics or proteomics technologies.[3]
3. Flexible Architecture:	Most models utilize transformer-based architectures, allowing them to capture complex contextual relationships among genes.[3]
4. Transfer Learning Capabilities:	These models can be fine-tuned or used for zero-shot learning on various downstream tasks.[3]

Use Cases

Use Case	Description	Models
Cell Type Annotation	Models can accurately classify cells into different types based on their gene expression profiles	scGPT, Geneformer, scaLR
Multi Batch Integration	Foundational models excel at integrating data from multiple experimental batches, reducing technical variability	COSMOS, scGPT, Geneformer
Perturbation Response Prediction	These models can predict how cells will respond to various perturbations, such as drug treatments or genetic modifications	scGPT, Geneformer
Gene Network Inference	By capturing complex relationships between genes, these models can help to infer gene regulatory networks	scGPT, Geneformer
Spatial Analysis	Predict spatial context and composition, enabling the transfer of rich spatial information to scRNA-seq datasets	Nicheformer, COSMOS

Challenges

1. Computational Resources:	These models can be very resource-intensive, often requiring multiple GPUs to train and run.[5]
2. Interpretability and Rxplainable AI:	Foundational models, while state-of-the-art, might fall victim to the “black box” phenomenon. Developing explainable AI methods is crucial for deeper biological insights.
3. Data Quality and Bias:	As with all machine learning models, the quality of output depends on the quality of input data. These models require heavy preprocessing and their robustness depends on the diversity of the training data.
4. Training Data Diversity:	The quality and diversity of the training data are crucial for developing robust and unbiased models.

Emerging Trends

1. Graph-based Models:	Integrating graph-based approaches with foundational models shows promise for capturing cellular heterogeneity and molecular patterns more effectively. Previous deep learning and graph-based approaches like DeepMAPS have shown great success in identifying gene networks [7]. Incorporating foundational models can enhance these approaches to provide more comprehensive representations of cellular interactions and gene regulatory networks.
2. Extensive Feature Selection:	Models like scaLR drastically reduce computational resources by selecting only the most important features for analysis [6].

3. Multimodal integration:	Future models are likely to incorporate even more diverse data types, including proteomics and epigenomics data [3].
4. Spatial Awareness:	Models like Nicheformer are paving the way for spatially aware representations of cellular variation at scale [4].

Conclusion

Foundational models mark a pivotal advancement in single-cell omics analysis. These sophisticated tools, built on extensive datasets and cutting-edge machine learning, offer unparalleled insights into cellular biology.

As this field progresses, we anticipate the emergence of increasingly powerful and adaptable models. These innovations will significantly enhance our comprehension of intricate biological systems and catalyze breakthroughs in biomedical research and drug development.

References

1. https://frontlinegenomics.com/single-cell-foundational-models-the-next-big-thing/

2. Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., Song, L. Large-scale foundation model on single-cell transcriptomics. Nat Methods. 2024 Aug;21(8):1481-1491. doi: 10.1038/s41592-024-02305-7. Epub 2024 Jun 6. PMID: 38844628.

3. Cui, H., Wang, C., Maan, H. et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 21, 1470–1480 (2024). https://doi.org/10.1038/s41592-024-02201-0

4. Schaar, A. C., Tejada-Lapuerta, A., Palla, G., Gutgesell, R., Halle, L., Minaeva, M., Vornholz, L., Dony, L., Drummer, F., Bahrami, M., & Theis, F. J. (2024). Nicheformer: a foundation model for single-cell and spatial omics. bioRxiv. https://doi.org/10.1101/2024.04.15.589472

5. Ma, Q., Jiang, Y., Cheng, H. et al. Harnessing the deep learning power of foundation models in single-cell omics. Nat Rev Mol Cell Biol 25, 593–594 (2024). https://doi.org/10.1038/s41580-024-00756-6

6. Jogani, S., Pol, A.S., Prajapati, M., Samal, A., Bhatia, K., Parmar, J., Patel, U., Shah, F., Vyas, N., Gupta, S. (2024). scaLR: A low-resource deep neural network-based platform for single cell analysis and biomarker discovery. bioRxiv: https://doi.org/10.1101/2024.09.19.613226.

7. Ma, A., Wang, X., Li, J. et al. Single-cell biological network inference using a heterogeneous graph transformer. Nat Commun 14, 964 (2023). https://doi.org/10.1038/s41467-023-36559-0

Early Days for AI but Scientific Data Management Gains Momentum

Early this year, Zifo conducted the “Data Readiness Survey,” which saw participation of scientists and informaticians from over 30 global companies, shows that scientifically driven companies (Pharma, Biotech, Chemicals, F&B, FMCG, Agrotech etc.) are beginning to seriously invest in AI and ML across the entire R&D, Manufacturing and Trials value chain.

A lot of the initial effort is going into the basics – making sure data is of good enough quality i.e. there are proper standards, and that everyone’s on the same page with things like metadata, ontologies, master data etc.

Read on for more

DATA READINESS SURVEY REPORT:

Early Days for AI but Scientific Data Management Gains Momentum

Authors

Paul Denny-Gouldson, Chief Scientific Officer, Zifo

Sakthi Prasad T, Content Director, Zifo

Data Compilation: Karthick Suresh, Analyst

Introduction

The scientific world is in the grips of AI fever, so to speak. Science driven companies are now starting to deeply appreciate the power of data for research and innovation; however, it's still a long road ahead when it comes to data standardization and ensuring ease of accessibility to support the use of AI.

Zifo's recently concluded "Data Readiness Survey," which saw participation of scientists and informaticians from over 30 global companies, shows that scientifically driven companies (Pharma, Biotech, Chemicals, F&B, FMCG, Agrotech etc.) are beginning to seriously invest in AI and ML across the entire R&D, Manufacturing and Trials value chain, and a lot of the initial effort is going into the basics -- making sure data is of good enough quality i.e. there are proper standards, and that everyone's on the same page with things like metadata, ontologies, master data etc.

However, the survey results show that both AI and ML initiatives are at very early stages and have a long road ahead before reaching maturity.

One of the big hurdles facing scientists and informaticians is the issue of data silos -- to fix this would require proper standards and alignment on elements such as metadata, ontologies, and master data. AI/ML initiatives will suffer from a 'garbage in, garbage out' problem without fixing the issues plaguing "data silos". Additionally, scientists and informaticians are concerned about measuring the value of their work, specifically how to quantify the impact of their endeavours on scientific outcomes.

The second big issue facing scientists and informaticians is getting all their lab instruments to talk to each other and exchange data seamlessly. There is a plethora of instruments, each with its own way of connecting, and sometimes the ageing infrastructure doesn't permit seamless data exchange. Fixing that often means a big investment and making sure everything is secure, which encompasses a whole set of other challenges.

While everyone is aiming for better products impacts in the end, the immediate wins Scientists and Informaticians are seeing with AI are aspects such as efficiency, cost savings, potential to speed up discoveries – "innovation", and getting a deeper understanding of the science.

But there are worries about privacy, especially with all the new generative AI tools popping up. Scientists and Informaticians are understandably cautious about where their data end up and how it's being used; that's why more

organizations seek to bring these technologies in-house to keep a tighter grip on proprietary data.

Ultimately, getting the most out of AI and scientific data will depend on how well companies have managed their data in the past, how they tackle these integration issues, and whether they focus on smaller, targeted AI applications that make a real difference.

The collected responses offer a valuable perspective on the current state of technological integration within research and development, manufacturing and trials environments, highlighting the varying degrees of preparedness across different organisations.

The survey data illuminates the obstacles currently encountered in the seamless integration of data to support artificial intelligence, while also identifying the key perceived benefits and concerns associated with these advancements. Ultimately, this overview provides a concise snapshot of the current landscape, underscoring both the progress achieved and the considerable work remaining to fully leverage the transformative potential of data and artificial intelligence within the scientific community.

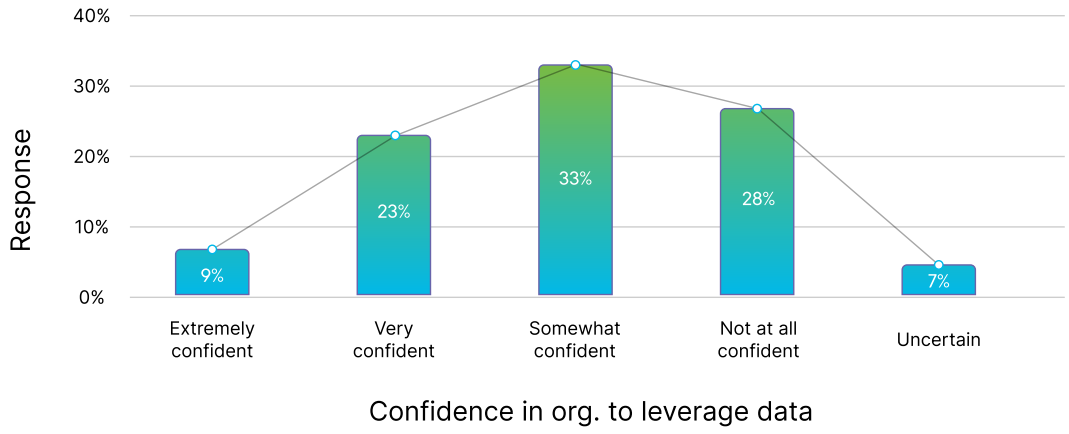
Zifo's Chief Scientific Officer (CSO), Paul Denny-Gouldson, has offered his insights on each of the findings. We have also included select quotes from survey respondents.

Please read the report and let us know your thoughts and opinions: info@zifornd.com

— Team Zifo

Survey Findings

1. Given the vast amounts of scientific data generated daily within your organization, how confident are you in your organization's ability to effectively leverage this data to drive AI programs and projects?



Approximately one third (32%) express "high confidence" (Extremely Confident + Very Confident), while a similar portion (35%) are either not confident at all or uncertain. The third group (33%) feels "somewhat confident," indicating a mixed perspective on the organization's capabilities. A notable 28% lack confidence entirely, highlighting potential challenges in data utilization for AI initiatives.

Overall, the responses suggest a need to address barriers and build greater confidence in effectively leveraging scientific data for AI.

Respondent Comments

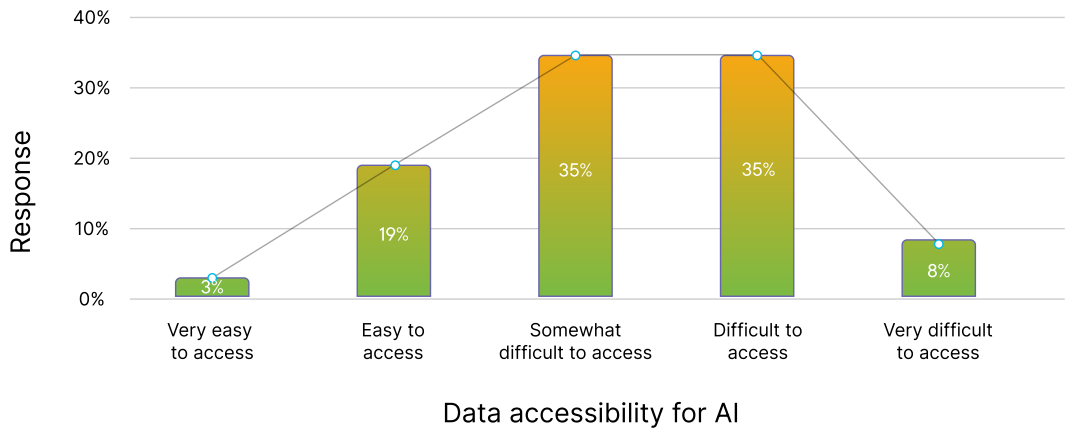
- "Data is not harmonized in terms of storage or metadata".
- "Since the organization is large and diverse, and each department is driven with a focus to crank out results with available technologies, it is difficult to gauge the ability of each individual department's ability to use AI driven programs to enhance productivity and efficiency".
- "Our company has an internal AI program, but it is very limited".

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *"Scientists are increasingly using data to drive their experimental design, instrument setup, and overall scientific workflow. They not only generate data within their own labs but also frequently need to share it with or receive it from other parts of the scientific process. This data exchange can involve instruments, master data systems, registration systems, data warehouses, or*

> *other data repositories, which are then used by others. Therefore, ensuring data quality, trustworthiness, and adherence to FAIR principles (Findable, Accessible, Interoperable, Reusable) is crucial. The level of "fairness" required can vary depending on the data's value and intended use, such as for advanced analytics or AI. Data readiness is paramount, along with understanding how the data is consumed. Digital literacy among scientists is also important for them to grasp the importance and usage of their data.*
>
> *The informatics team's role is to facilitate data availability, ensure quality, and make it consumable across the organization. The challenge lies in establishing the necessary data foundations without overwhelming scientists with excessive data preparation tasks, balancing this with the costs and time of implementation. IT serves as a support function to enable scientific progress and must adapt quickly to evolving scientific needs. While established, routine lab work is easier to manage, the dynamic nature of scientific research requires flexibility".*

2. In your experience, how easily accessible is the scientific data required for AI initiatives within your organization?



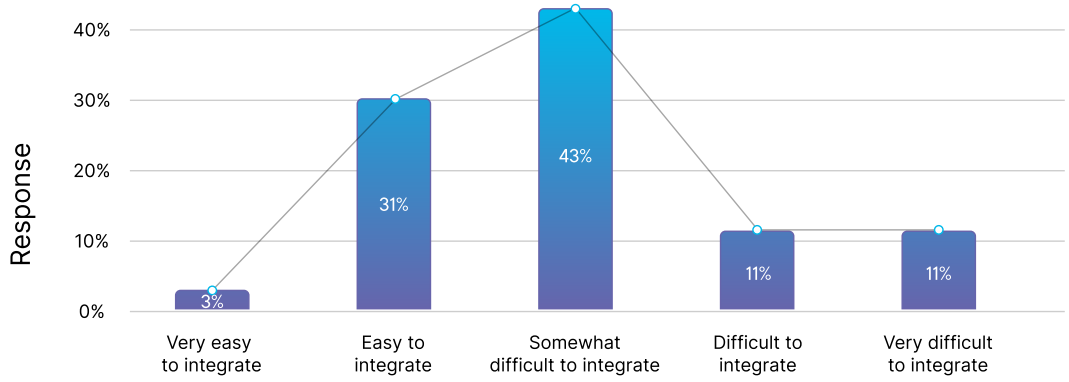
A substantial majority (70%) of respondents report that accessing scientific data for AI initiatives is either difficult or somewhat difficult within the organization. Conversely, only a small fraction (22%) find the required data to be easily accessible. The equal percentages for "Difficult" and "Somewhat difficult" access (both 35%) indicate a widespread and significant challenge.

Overall, the data strongly suggests that improving data accessibility is a critical area for supporting AI programs within the organization.

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *"The current state of AI and machine learning adoption and integration within scientific organizations is still in its early stages. A key focus is on understanding and educating scientists about the necessity of high-quality data that is well-documented with metadata. This ensures the data can be effectively utilized for algorithm development and machine learning initiatives. While AI can assist in data cleaning, the primary challenge lies in establishing a foundation of clean, well-described data. The speed at which this landscape changes over time will be a key indicator of progress. If minimal improvement in data accessibility for AI is observed in the near future, it might suggest that current strategies are insufficient or that this is a more protracted issue.*
>
> *Given the ongoing nature of scientific research, solutions must be implemented alongside active projects. Unlike a "greenfield" scenario, the existing complexities and change management aspects within established organizations ("brownfield") make this a significant change management and value demonstration challenge. Organizations and IT departments need to collaborate with the scientific community to evolve existing systems, rather than implementing radical overhauls. This evolution requires substantial investment".*

3. How would you describe the ease of data pipelining, integration and exchange of lab instruments data in your organization?

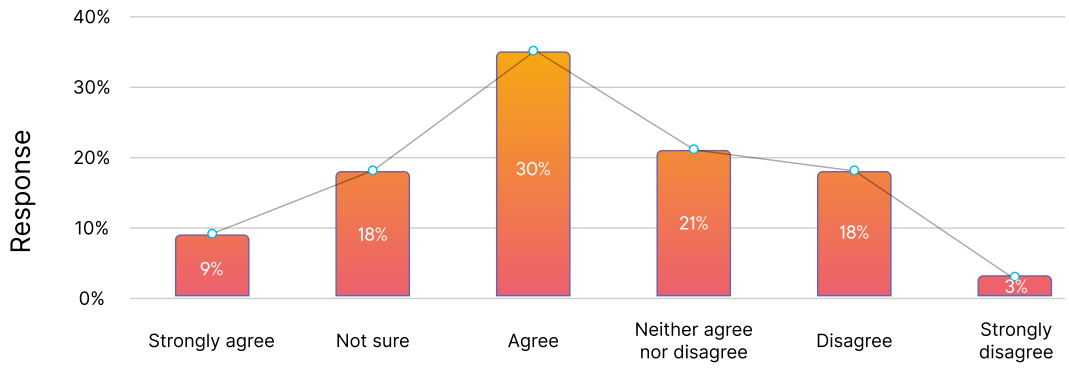


Most organizations find data pipelining, integration, and exchange of lab instrument data to be challenging. Specifically, 43% find it "Somewhat difficult to integrate," and a further 11% report it as "Difficult to integrate," while another 11% find it "Very difficult to integrate." In contrast, a smaller portion find it "Easy to integrate" (31%), and a very small fraction consider it "Very easy to integrate" (3%).

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *"The difficulties in integrating lab instrument data stem from several factors. These include the need for configuration and maintenance, the complexities of connecting to lab execution environments like ELNs or LIMS, and the intricate, often multidimensional, nature of the data itself. The diverse and heterogeneous instrument landscape, while adhering to connectivity standards like RS 232 or whatever it is, historically lacked the necessary physical infrastructure (cabling) for seamless data transfer. Addressing this often requires significant investment in infrastructure upgrades and ensuring network security, especially in regulated environments.*
>
> *Furthermore, the cost-benefit analysis of full integration can sometimes favor manual or paper-based systems, though this can impact scientist satisfaction.*
>
> *Establishing a 'single view of truth' for data is fundamental for progress, requiring instruments to communicate and data to be centrally stored in a standardized format. The importance of this varies depending on the stage of research, development, or manufacturing, with raw data being crucial in regulated environments. While archiving is essential for regulatory compliance, transforming archived data into a usable data source for analysis requires additional effort to meet FAIR principles.*
>
> *A key challenge is determining which data will hold future value, particularly in research. Leaders often question how to identify and prioritize data for long-term investment when future relevance is uncertain. This highlights the need to balance current data integration efforts with the potential future value of the data being captured".*

4. Does your organization have standardized data formats and ontologies (for eg. FAIR principles) that enable ease of data retrieval?



Does org. have standardized data formats? (eg FAIR)

A significant portion of respondents (39%) agree or strongly agree that their organization has standardized data formats and ontologies facilitating data retrieval. However, a considerable 39% are either unsure or disagree, indicating a lack of standardization in some organizations. About 21% remain neutral on this aspect. This suggests a mixed landscape regarding the adoption of standardized data practices for easier data access.

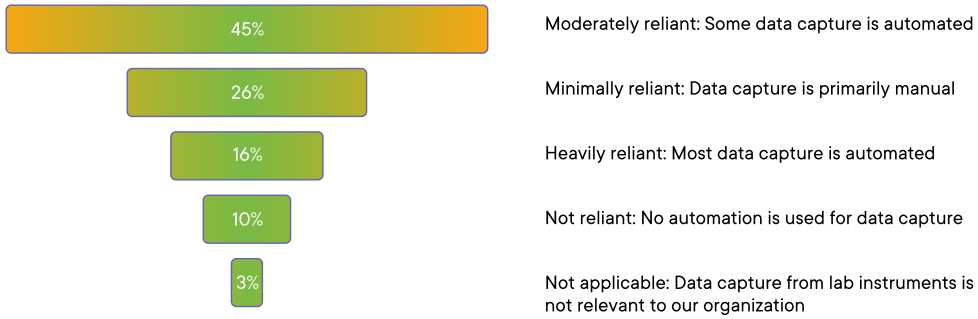
Respondent Comments

- "Some departments work with FAIR principle, while others are trying to implement it".
- "Work in progress"

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *"Standardized data format does depend on the organization and how focused they have been in the last 10 years or so when it comes to putting up infrastructure to capture data and data reuse. So, you see a big spread of organizations where some have not been spending a lot on data infrastructure, while a few did. And for those who haven't invested in harnessing their data, it is now becoming a problem because of the requirement to do all the AI machine learning, which is dependent on data reuse, ontologies and so on. Data Management is fundamental to ensuring data reuse and data retrieval, because that is the lifeblood of what enables FAIR data."*

5. To what extent does your organization rely on automation for data capture from lab instruments?



Based on the survey results, a significant portion of organizations (45%) are moderately reliant on automation for data capture from lab instruments, indicating a blend of automated and manual processes. A substantial number (26%) still primarily rely on manual data capture, highlighting a lower level of automation in these organizations. Conversely, a smaller but notable percentage (16%) heavily automates their data capture, while a minority (10%) uses no automation at all. Finally, a small fraction (3%) indicates that data capture from lab instruments is not relevant to their operations.

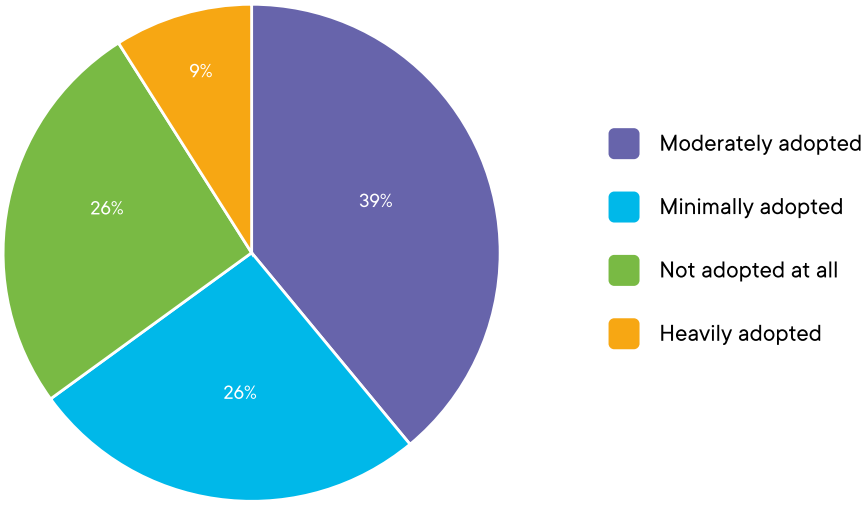
Respondent Comment

- "Would be nice to have more automation"

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *"The readiness for data reuse and AI/ML adoption varies significantly across organizations, largely reflecting their historical investment in data management over the past decade. Organizations that haven't prioritized data infrastructure are now facing challenges due to the increasing need for AI, machine learning, and broader data utilization. Essential data management practices, including standardized data, master data management, and ontologies, are fundamental for ensuring data's 'fairness' (Findable, Accessible, Interoperable, Reusable) and facilitating effective data retrieval.*
>
> *While some organizations report having these crucial elements in place, a substantial either express uncertainty or disagree, indicating a widespread deficit in foundational data management capabilities. This suggests that the primary factor influencing data readiness is likely the individual company's approach to data management rather than the specific sector (research, development, manufacturing or trials) it operates within. Many organizations still lack the core data management infrastructure necessary to fully leverage current data-driven technologies."*

6. To what extent has your organization adopted AI in its R&D operations?



The survey reveals that most organizations have adopted AI to some extent in their R&D operations. A significant 39% have moderately adopted AI, indicating its integration into some aspects of their R&D processes. Furthermore, a smaller but still considerable 26% have minimally adopted AI, suggesting initial or limited implementation. Notably, an equal percentage (26%) have not adopted AI at all in their R&D operations, while a smaller 10% have heavily adopted AI, signifying its widespread use.

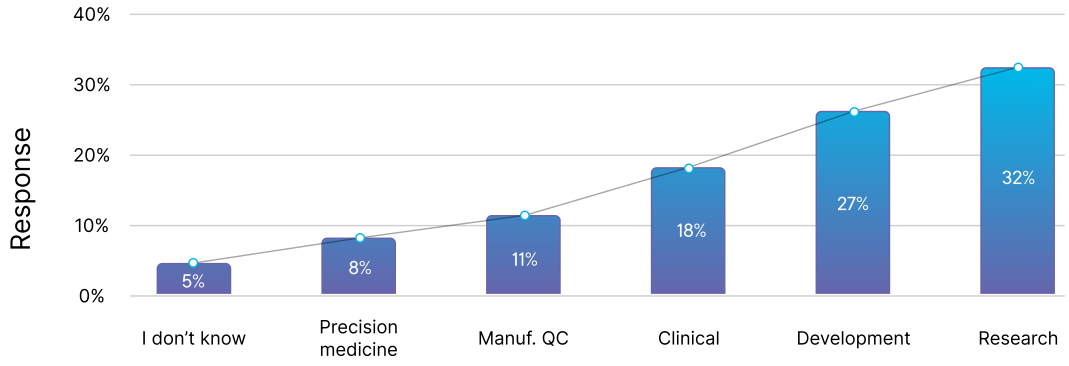
Respondent Comments

- “Getting more and more adopted”.
- “Work in progress”
- “Just starting with artificial intelligence”

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *“The level of automation in data management, including the use of ontologies and master data management principles, aligns with previous observations, with a majority indicating some level of implementation. Automation becomes particularly beneficial in manufacturing and regulated environments by minimizing human intervention and reducing the risk of data quality issues associated with manual data entry.*
>
> *However, the decision to automate is also influenced by user needs and the frequency of the task. For instance, in research settings with diverse activities, the effort required to automate a less frequent process might outweigh the benefits, leading to a preference for manual methods. Conversely, in high-throughput research or workflows heavily reliant on robotics, full automation is often prioritized to maximize efficiency and enable continuous operation, facilitating a higher volume of experiments than manual approaches would allow.”*

7. To the best of your knowledge, which areas of your organisation are looking at using AI? (Multiple Selections Allowed)



Areas of organisation looking in to AI

Organizations are primarily exploring AI adoption in Research (32%) and Development (27%) functions, indicating a strong interest in leveraging AI for innovation and product creation. Clinical applications are also being considered by a notable portion (18%), while Manufacturing QC and Precision Medicine show emerging interest (11% and 8% respectively). A small percentage of respondents (5%) are unsure about specific areas of AI exploration within their organization.

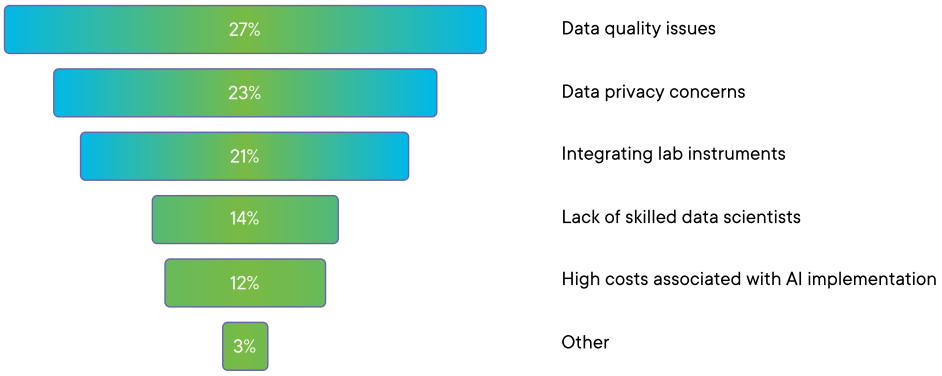
Respondent Comments

- “Barely looking”
- “Medical Affairs”
- “Interpreting huge amount chromatographic data, impurity profiling comparisons, and report generating”.
- “There’s interest across the board, but I can’t speak for some of these areas”.
- “All”

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *“The current trend in leveraging AI in R&D and related fields is shifting towards targeted, specific applications that augment human decision-making rather than fully automating it. Initially, the focus will likely be on AI assisting scientists in identifying data trends and providing insights, while human experts retain responsibility for final decisions. This approach is particularly relevant in research, where the primary need is for generating new ideas without immediate regulatory constraints. Agentic AI is beginning its journey, but this is where the decisions are made automatically based on inputs – there are various safety, regulatory and ethical concerns that need to be addressed before this is widely adopted.*
>
> *As projects progress towards development and clinical stages, the influence of regulatory considerations increases. Market feedback indicates a preference for numerous small, focused, AI interventions across a given process and value chain, rather than large-scale, all-encompassing AI solutions. These smaller integrations aim to incrementally improve speed, precision, and insights, enabling faster and more informed decision-making. While large-scale AI projects garner significant media attention, the true value is increasingly seen in integrating AI and machine learning into routine workflows across the entire value chain”.*

8. According to you, what is the biggest challenge your organization faces in leveraging scientific data for AI? (Multiple Selections Allowed)



Data quality issues are the most significant challenge (27%), followed closely by data privacy concerns (23%). Integrating lab instruments is also a major hurdle (21%). A lack of skilled data scientists (14%) and high costs associated with AI implementation (12%) are also notable challenges. A small percentage of organizations (3%) cited other challenges.

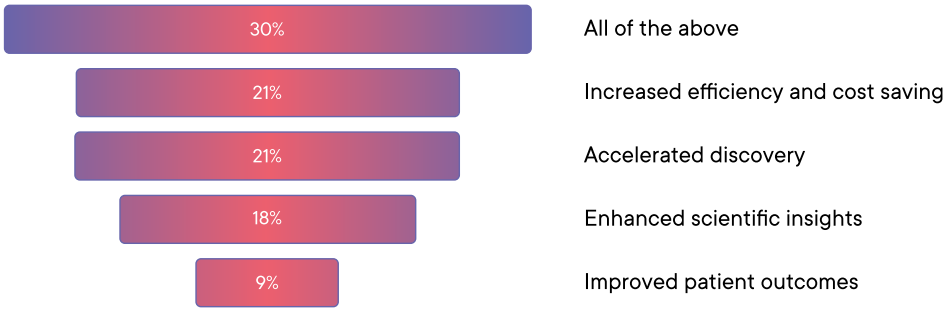
Respondent Comments

- "GxP"
- "Lack of mindset"
- "Most data management systems focus on lab notebooks. I run the HPC system, and ELNs do not help with the petabytes of unstructured data a typical HPC system has, so there's a gap between neat capture of data from the instrument (which is well automated) and storing of fair data products of analysed data (which we have well covered). In between the processing of the instrument data and other HPC analysis is not well covered".

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *"The prominence of data privacy concerns as a major challenge in leveraging scientific data for AI (ranking second) reflects a fundamental human instinct to protect one's data. Concerns arise from the potential for misuse or lack of understanding regarding how data will be utilized, particularly in the context of data reuse. The example of public generative AI tools like ChatGPT highlights these anxieties, as input data can inadvertently become part of the public learning model, posing significant confidentiality risks. This has prompted organizations to rapidly adopt in-house versions of such technologies to better manage and secure sensitive information within their own networks".*

9. What do you see as the greatest potential benefit of AI in biotech and pharma? (Multiple Selections Allowed)



A significant portion of respondents (30%) believe that “All the above” options represent the greatest potential benefit, indicating a holistic view of AI’s positive impact. Among the individual benefits, “Accelerated discovery” and “Increased efficiency and cost savings” are seen as highly beneficial (21% each). “Enhanced scientific insights” is also considered a major advantage (18%), while “Improved patient outcomes” is also recognized, though slightly less frequently (9%).

Paul Denny-Gouldson, Chief Scientific Officer, Zifo Technologies

> *“While improved patient outcomes should ideally be the end goal, the survey results suggest that respondents are also strongly focused on other tangible benefits of AI. These include increased efficiency and cost savings, accelerated discovery, and enhanced scientific insights. These intermediate goals are often seen as direct pathways to achieving better patient outcomes, as improvements in these areas contribute to the development of better products and processes.*
>
> *This perspective highlights the practical application of AI, where measurable outcomes like optimized processes, faster discoveries, and deeper data understanding are readily apparent and contribute to the overarching objective of improved patient care. The effectiveness of AI in achieving these benefits is most evident when applied in a targeted manner with a clear understanding of the use case and the expected value”.*

Looking Ahead

Data standardization and seamless data exchange across R&D, Manufacturing and Trials value chain is critical for scientifically driven industries such as Pharma, Biotech, Chemicals, F&B, FMCG, and Agrotech. If one were to put a label on the current era, we could call it “The Age of Data Management”, which will then eventually lead to “The Age of AI”.

About Zifo

Science led, people driven, technology centered.

Zifo is the leading global enabler of AI and data driven enterprise informatics for science driven organizations. With extensive solutions and services expertise spanning research, development, manufacturing, and clinical domains, we serve a diverse range of industries, including Pharma, Biotech, Chemicals, Food and Beverage, Oil & Gas, and FMCG. Trusted by over 190 science-focused organizations worldwide, Zifo is the partner of choice for advancing digital scientific innovation.

Visit: www.zifornd.com | Email: info@zifornd.com |


AD - 372
NB - 358
N - 4.9
27.337
34.852
77.259
29.625
88.273
34.682
L - 8.2
CB - 559
2703
3270
7843 - 8954 - 9854
1101 / 8943.6743