DATAHOW

Accelerating and Improving Bioprocess Development with Machine Learning Solutions

Alessandro Butté, Ph.D., and Michael Sokolov, Ph.D., DataHow

In an industry where precision and efficiency are paramount, DataHow has emerged as a pioneer in leveraging advanced machine learning techniques to optimize process development, manage risks, and support data-driven decision making. Bridging machine learning from big data to the far smaller bioprocess data, DataHow's core innovation lies in its hybrid modeling technology, which combines process data with engineering knowledge to enhance process development and manufacturing robustness.

In this Q&A, two of DataHow's founders, Chief Executive Officer Alessandro Butté, Ph.D., and Chief Operating Officer Michael Sokolov, Ph.D., discuss DataHow's journey from concept to implementation, highlighting the transformative potential of their solutions in accelerating process development, reducing errors, and facilitating a more agile response to the dynamic demands of pharmaceutical production, in a conversation with *Pharma's Almanac* Editor in Chief David Alvaro, Ph.D.

David Alvaro (DA): To begin, as two of DataHow's four founders, can you give us a concise history of the company's origins?

Alessandro Butté (AB): My background is in academia – I've been in the university environment for more than 20 years. At a certain point, I decided that what I was truly passionate about was solving practical problems around what my colleagues called technology, and not typically in a nice way. I made a transition into industry when I entered into a collaboration with Lonza – one of the largest pharmaceutical CDMOs in the world – to explore the use of modeling techniques to support quality by design.

At that point in time, when a CDMO engaged with a new client to produce product for a clinical study or commercial phase, they essentially had to start everything from scratch – a blank page, as though they had

never developed such processes and had no useful data to leverage. I found that very frustrating, because it was incredibly inefficient, new learnings were constantly being completely lost, and there was inherently a lot of uncertainty surrounding the processes because they were almost entirely based on the quality of the scientist in charge of development of the project.

To me, the clearest solution to these issues involved a tool like machine learning, which could enable processing of a huge amount of data and finding rational paths that could be used to support process development work. At that time, the main challenge was that machine learning has always been associated with big data. If you're talking about statistics, the more relevant data you have, the more powerful the results. However, pharma typically involves very small data. The real challenge lay in determining how to use such powerful tools with only a small number of experiments, maybe just 10 to 20 or so in the path from start to manufacturing. The answer was hybrid modelling, which is the core of our technology. Hybrid models are very complicated mathematically, but quite simple in concept: they involve combining two sources of knowledge – in this case, process data and engineering process knowledge – and combining mechanistic and machine learning models to reach an outcome that reduces the number of experiments (or the amount of data) needed to optimize a process, while adapting to the specifics of the considered process. Mathematically, this involves constraining the space of the solutions that a machine learning tool can find using prior knowledge from equations and adaptive learning of some components of these equations as a dynamic non-linear function of the process control conditions.

Michael Sokolov (MS): To put it another way, we identified a gap that needed to be closed. Machine learning was a technique that was exploding in some other fields that are spoiled with more data, so our goal was to figure out how to leverage it in an environment in which every data point inherently comes at a very large labor cost, in spite of the considerable complexity of bioprocesses to be solved.

AB: The beginning of DataHow's journey was determining whether it was possible decrease the number of experiments to a level that competes with the average number of experiments used to develop a process today. Today our journey has transitioned to following our main vision – supporting pharmaceutical companies, CMOs, and so on to improve their manufacturing data, especially process quality data, to make development way faster and more robust, to decrease a lot of the errors and failures in manufacturing, and to accelerate process development and allow pharmaceutical companies to handle larger pipelines because they need less resources to develop a process.

MS: It is important to note that the vision takes different forms depending on the processes or the underlying modality being explored. For well-established bioprocesses, such as the production of therapeutic proteins through platform processes, this technology is very likely to be of great help to accelerate programs, reduce costs, and transform how people are operating on a two-digit percentage basis: cutting costs and timelines by maybe 30–70%. However, in the new modality space, where processes are not yet well understood or established, we foresee the technology having an enabling effect – it simply might not be possible to bring a certain therapy to the market at all, or at the speed required for patient needs, without a digital technology playing an integral role. That applies to things like cell and gene therapies, but also food tech, cultivated meat, and so on.

DA: To realize that vision, did you intentionally build a team with different expertise and contrasting viewpoints and priorities?

AB: We have always aimed to bring together very different points of views on the technology, and in some cases on our strategy and tactics. The team comes from a range of backgrounds: some people are more academic, others are more industrial, and so on, and hence also very different personal experiences. However, we have very much focused tactically on the key concerns of the sector in which I was working before, simply because you have to have a focus. In that sector, we already had a lot of expertise, not to mention contacts we could speak with to hash out ideas. When we speak with clients, we have been able to offer advice drawn from our experience, our knowledge of the science and the data, and our understanding of the business. We can aggregate all these different perspectives in a coherent strategy to support our clients' goals.

The real challenge lay in determining how to use such powerful tools with only a small number of experiments, maybe just 10 to 20 or so in the path from start to manufacturing.

MS: On the big vision, I think we have always been very uniquely aligned. However, we feel that it is critical to continuously realign on more incremental details based on the feedback we receive from clients. The constant prototyping and exploration of solutions with clients led us to understand those needs better, especially across different segments. There is big pharma versus CDMOs and small biotechs. While they all converge on bioprocess, they have different expectations – some are looking for optimization, and others for enabling technologies. On the other hand, we had to understand who the potential users might be in-house and how we could best help customers embrace the technology as part of their organizational digital transformation.

AB: At the same time, the way we are perceived by customers is continuously changing. In the beginning, we were more experts brought in to consult, whereas today we are solution providers and even software providers, which is reflected in a radical shift in the discussions that we have with clients, who start playing a more active role in that digital transformation.

DA: I'm sure we could spend hours on the nuances of your technology, but could you give me a concise explanation of the key principles and their importance in achieving your goals and those of your customers?

AB: Our technology is based on three main pillars. The first, as I mentioned earlier, is hybrid modeling: machine learning for small data, which unlocks the ability to autonomously learn from process data.

The next pillar is a direct consequence of the application of machine learning, which we call transfer learning. In cases where the end processes for a given drug have been fully developed and all the data from these processes are available, machine learning can allow you to extract common knowledge from those data that then can be specifically readopted for the development of a new process. In many cases, this means you don't have to perform some of the new experiments for that process development because you can simply transfer what you have seen in the past and perform experiments only to create new or verify information.

The third pillar is optimizing design of experiments and other activities to support the end user in decision making. Today, all the models are deterministic or, technically, "classical statistics," in which you provide input into the model, and it outputs a number representing its best knowledge about that. In contrast, our work tools are based on Bayesian statistics, which means that every time that we get a prediction from our models, it is not only a number but also a probability distribution. That enables us to integrate risk considerations into our decisions, depending on how much data we possess and how ambitious the decision objective is about a given process under certain conditions.

As a result, a distinctive feature of our approach to developing processes is that we typically do so based on utility. We can aggregate different considerations in our tools that range from constraints on how likely it is that we meet all the quality constraints or improve the productivity of a process, how expensive it is to run certain processes, or how likely a given new experiment is to truly create new knowledge. The user can combine all these different aspects and their corresponding probabilities and risks to come up with a very efficient way to develop processes or to simply manage risks in biomanufacturing.

MS: Another aspect of our value proposition is that all of this is packaged as a user-friendly cloud solution, which facilitates collaboration and enables a team with only very limited experience in modeling (or none) to collaborate on the creation of the predictive model, which is then the engine to answer all their practical questions. Additionally, the tool is fully customized to the needs of the pharma industry. Having worked with several tens of different companies, we have a comprehensive understanding of the key questions they would like answered. With our tool, the final step is not the creation of the model, which is the case for many other software solutions – it's the decision derived from the model, which follows very practical needs: get more product, understand the process better, understand how to design and scale up the process, and so on.

This collaborative cloud architecture – combined with the very customized way the software is established in terms of workflow – allows us to democratize the use of machine learning across an organization that is conservative and not digital native. The magic happens in the background, but the solution is a user-friendly tool that serves as a bridge to our vision, and machine learning moves from just being a buzzword to a technology used every day – a commodity to create consistent value from the routinely measured data.

This requires that the technology be customized to the problem to be solved and to the users who need it. If you compare the current, third version of our software with the initial zero version, you can see how much more customer centric it has become. We began with what we thought the customer needed but have updated that to what different customers have said they need. Beyond the continuously generalized and diversified software itself, the added value for the users has helped us to evolve our perspective.

AB: Ultimately, the definition of key terms like digital twins can vary significantly from field to field and even individual background to background. An engineer probably does not see a digital twin in the same way a data scientist would define it. But in the end, the digital twin is a tool that allows the vast majority of the stakeholders within a pharmaceutical company to interact with the prior knowledge about the processes to act on ongoing process without needing to understand the underlying algorithmic details on the digital twin.

DA: Can you discuss the process of a customer adopting your technology and integrating it into their existing process development and manufacturing systems?

AB: We could probably discuss this for two days from 20 different perspectives, but in short, adoption is currently a relatively long and painful path. On one hand, we are developing processes that could not be developed otherwise. In a sense, especially at the beginning of a journey, we provide a very incremental improvement.

We have certain established tools. The challenge is not that the scientists we engage with are unable to understand our tools, but that these tools are not yet well accepted by the broader scientific community, particularly by regulators. Changing tools poses challenges for a pharmaceutical com-

Machine learning has the potential to radically change the way we approach every concept, from how we develop a process to how we manage the quality of drugs in a broader sense.

pany because they have to restart a lot of their discussions with regulators, who will challenge the approach and underlying tools in depth. With this type of technology, things often really depend on the first adopter, who takes up the challenge of clarifying all the problems for everybody; from that point on, it is a downhill journey. Five years ago, hybrid modeling started to become a topic in such discussions with regulators, and now we see them becoming a more regular interaction point alongside other machine learning techniques that are clearly superior to the old way of using simple statical tools only.

Another challenge is that most of our customers today are focused on process development – in other words, short- to intermediate-term improvements – while the greatest improvements the technology can achieve manifest over the middle to long term. Ultimately, machine learning has the potential to radically change the way we approach every concept, from how we develop a process to how we manage the quality of drugs in a broader sense. Unfortunately, this is very much in the future, but as you can imagine, it is complicating adoption. People tend to be a bit ultra-focused on what is happening tomorrow and the next week rather than what's over the horizon.

DA: Do you see different responses depending on the nature of the customer you are speaking with? I'd imagine that a small biopharma might be more willing to take risks than a big pharma company, but they might also be far more focused on the near term.

AB: The answer is totally counterintuitive. You would expect the early adopters to be companies really focused on manufacturing, like CMOs, rather than larger pharma companies where manufacturing is just one of many activities. Additionally, you'd expect, as you suggested, that small companies that are more flexible in their procedures would jump on the technology faster than big companies.

In both cases, however, what we have seen is the opposite. In the first case, I think that pharma companies are more interested than CMOs for two reasons. First, pharma is much more willing to invest in R&D, whereas CMOs need to be extremely efficient, with any innovation activity being perceived as a waste of time. The second reason is that it is ultimately difficult for CMOs to adopt innovative technologies without being backed by pharma, because if their pharma customers are not interested in using hybrid models to develop bioprocesses, that's the end of the discussion.

In the end, the activation energy barrier that you have to overcome to adopt the technology is directly proportional to your degree of digitalization or your readiness to digitalize. The more digitalized you are, the faster you can harvest results from our tools. Larger companies are way more digitalized than smaller companies, who value flexibility over a very standardized way of running things.

MS: In many cases, the key enabler of acceptance for a potential customer is an internal believer in the technology. The believer or influencer may be someone at C-level, but we have also had cases where the believer was a scientist or manager. But in many cases, the identification of a single champion who can open doors has been our best means of acceleration. This believer can be someone who really gets the technology, but it can also just be someone who has seen what we have done for others and become convinced of the value. Then this believer needs to convince a critical mass of people – from CEO to the end users or the other way around – that there is a business case or a clear need. We are still learning how best to segment things into different paths and customize our sales cycle to maximize success regardless of our path into an organization.

It's also always helpful to have references and tangible results, which can be something created for others. As soon as our software was tangible, discussions became much easier because we can run a demo and show what is possible. But as a thought leader, that's generally not the case because you're running in front of the industry with the original vision, which then becomes a prototype, which then becomes more and more of a solution. This critical mass of people believing in the tool and creating a community across organizations was a key enabler for us along the way.

That doesn't apply only to our solution; in general, companies who embraced a digital maturity vision combining different digital solutions as part of their assets already started to measure a return on investments last year, whereas the others are following.

DA: Can you expand on how you aim to make the solutions more accessible and user-friendly for a much more broad and potentially data-naïve audience who may want the outputs without understanding how they were achieved?

AB: In the end, all the stakeholders associated with manufacturing or the development of manufacturing processes have to interact. Today, without a tool like a digital twin, they have to interact from different perspectives on the processes. The scientist has a perspective, maybe focused on optimization and robustness. The technician has their perspective in terms of organizing the experiments and the data and collecting the results in an organic way. But then there is the QA person, who wants to understand the risks and the regulatory side of things and who is running risk analysis, validations, and so forth. There is the production team that has to control or simply schedule processes. This is the next challenge of our tool: to address all these stakeholders with a platform solution containing specific features that are designed to support each of their individual perspectives.

MS: Another important angle to all of this is the need for a clear mindset change perspective, where we can play an educational role. We need to have different storylines ready for someone with a more stubborn or conservative mindset versus someone who is very open. We are quite active in teaching – less about only what our tool can do but rather how new methodologies compare with old methodologies. Alessandro teaches at university to prepare young chemical engineers to see the value of machine learning,

and we run a few courses each year, for which we have had 250–300 industry participants from all the major pharma companies. Independent of the solution they choose, they can learn what to expect from the technology and how to receive answers on practical bioprocessing questions.

In addition to this crucial educational component, we need to help organizations overcome an additional barrier: despite the conviction that it's the right tool to use, the ideal users are often too busy in the lab to have time for adoption. This requires another shift to allow experimentally focused people to be in front of the computer for sufficient time to learn and use the technology to enable them to improve their work in the lab. If you want to have a return on investment on a digital solution, you need to allow your team to use it at least a certain amount of hours per week.

DA: In your ongoing R&D efforts, are you primarily focused on training the system on new data and creating more user-specific applications, or is there yet work to be done to make the fundamental machine learning technology "smarter?"

AB: That is a very good question, because we went through several different phases. At the beginning, we had to learn how to implement our technology, so we focused on making the hybrid models perform increasingly sophisticated tasks. Today, we are more in a phase of simplifying things. In a sense, we are going backwards – not in terms of results but going back to focus on decreasing the barrier to adoption. For that, the tools have to be very simple. Ultimately, we aren't arguing that we provide the best hybrid of machine learning; we are essentially the only company providing that. Instead, we are contrasting ourselves with the state-of-the-art technology, which is way less efficient but widely adopted.

Additionally, we are taking a few concrete actions, mostly increasing the spectrum of tools that we provide to pharma in two dimensions: different unit operations to cover the entire manufacturing process, possibly including formulation; and modalities, going from therapeutic proteins to mRNA, cell therapies, and so on.

We have a few other innovative initiatives in the works. One is exploring the ability of generative AI to aggregate all the historical data from process development, all the results and so on, and create an R&D report describing what has been done, how the process is running, all the tests, the history, and so forth, for filing purposes and presentations of these activities, which can be extremely time-consuming. We could also standardize how such reports are created.

DA: Since the company has already evolved a lot, from a more advisory role into solution providers, do you foresee further evolution of DataHow as the technology evolves and your relationships deepen over the coming years?

AB: The scalable parts of the company will always focus on the software. But like many similar companies, the service part will play a growing role. For our customers to get the most out of technology, they have to be supported.

But the most straightforward direction where we will be heading is into manufacturing. Eventually, all these activities and all this knowledge will be transferred to manufacturing to concretely support the full range of the activities for drug production, regulatory, and so on. We will also be involved in enabling technology providers to integrate our technology with what they are producing. For example, a bioreactor producer could integrate our base technology into their software to support the management of the data coming from that platform and normal activities. The same would be true for a company producing sensors – we could aggregate all this knowledge together so they could come up with a package of sensors that are able to capture knowledge and support manufacturing activities.

MS: We want to become a very established provider of this technology in the field. Of course, scalability will come from the software. But our entanglement around changes in mindset and the support of a growing base will help us to stay in touch with where the industry is going, and we will have the advantage of being the first mover in that direction. There are many doors we want to go through in manufacturing. But maintaining long-term relations with the customer is a critical goal for us, and a big part of that will be diversification of the software beyond the launch version. We want to have a full platform solution covering small and large scales and all sorts of unit operations.

Building this platform step by step will allow us to be a very established provider in that space with confirmation that other players are going our way and switching to hybrid model machine learning. That affirms that we are on the right track but reminds us that we need to move quickly and partner smartly.

ABOUT THE AUTHORS

Alessandro Butté, Ph.D.
Chief Executive Officer

Alessandro Butté received his MSc. at Politecnico di Milano (Italy) and his Ph.D. at ETH Zurich (Switzerland) in chemical engineering. After a postdoc at the Georgia Institute of Technology (USA), he joined ETH Zurich as senior scientist. In 2008, he joined Lonza as head of downstream technologies in the small molecules and peptides sector and, later, as project manager. In 2013, he rejoined ETH as lecturer and, in 2017, he co-founded DataHow AG, where he is serving as CEO. He is the author of more than 90 papers on international peer-reviewed journals and several patents. In 2015, he completed an executive MBA at the University of St. Gallen.

Michael Sokolov, Ph.D.
Chief Operating Officer

Dr. Michael Sokolov is co-founder and COO of DataHow AG, a spin-off company from ETH Zurich specialized on process data analytics and modeling with a particular focus on the biopharmaceutical and chemical domains. His main activities are centered on managing data analytics and software projects with global pharma accounts, as well as coordinating all operational and financial activities of DataHow. He also holds a lecturer position for statistics for chemical engineers at ETH.