THE ROLE OF AI AND ML in Efficiency and Innovation

By Chloe Lang and Timo Schmidberger, PhD

The integration of artificial intelligence (AI) and machine learning (ML) into bioprocess development represents a rapid shift in the way discovery, development, optimization, and production of biological products are approached.

Bioprocesses, which involve the use of living cells or enzymes to manufacture drugs, vaccines, and other high-value compounds, are inherently complex and data-intensive. Other fields in pharma, such as drug discovery or medical devices, seem to be leading the implementation of AI/ML methods. There is currently a list of 950 AI- and ML-enabled devices provided by the US Food and Drug Administration (FDA) [1]. The increased awareness and accessibility of AI and ML technologies also offers unprecedented opportunities to harness this complexity and turn vast amounts of data into actionable insights that can then lead to more efficient and cost-effective bioprocesses.

ALGORITHM GUIDANCE

It is important to acknowledge that many of these concepts, especially for ML, are currently established in industry as chemometric or multivariate data analysis (MVDA) techniques [2]. MVDA is a discipline within ML that can use both supervised and nonsupervised algorithms. Although these techniques have been used for many years, it does not mean they are less relevant or less effective than other AI/ML algorithms.

In fact, for certain applications and use cases, they provide advantages in explainability. Choosing which algorithm to use should be assessed on a case-by-case basis and based on factors such as the available data, complexity of the problem, and risk assessment. There is existing guidance in place from the United States Pharmacopeia [3], FDA [4], and European Medicines Agency (EMA) [5] providing clear starting points for further utilization of AI/ML applications.

HEALTH AUTHORITIES

A variety of health authorities have been proactive in encouraging the adoption of AI and ML. For example, in 2014, the US FDA Center for Drug Evaluation and Research (CDER) established the Emerging Technology Program to work collaboratively with companies to support the use of advanced manufacturing. In 2019, the FDA asked for feedback on the then proposed regulatory framework for modifications to the "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan," which was eventually published in January 2021. The US FDA also published the guideline "Good Machine Learning Practice for Medical Device Development: Guiding Principles," published in October 2021 [6].

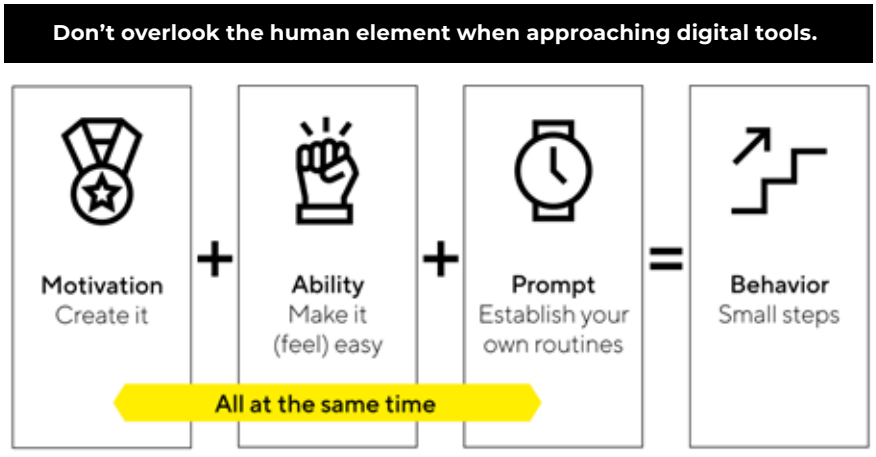
The benefits of using AI/ML are also acknowledged by the US FDA [7] and the EMA [8] to establish better guidance for AI/ML use in the whole development journey from molecule selection to drug product. International organizations, including the International Society for Pharmaceutical Engineering (ISPE) and BioPhorum, are publishing more holistic guidelines on requirements and steps to build AI/ML-based models in a regulated environment [9–12].

CHALLENGES

However, the successful implementation of AI and ML in bioprocess development is not without its challenges. One significant hurdle is the tendency for data scientists to work in silos, focusing on narrow aspects of the data science domain without a holistic view of the entire system. Typically, that means they are focused more on the data itself or the algorithms and less on the process science. Siloed thinking can limit the potential of AI and ML applications, as the true power of these technologies lies in their ability to integrate and analyze data across different stages and scales of the bioprocess. Developing AI expertise internally in organizations is a complex task. Creating the diverse teams needed to spearhead this change is a challenge. Moreover, the adoption of AI is often carried out in isolation. Typically, AI-driven processes are executed separately from the core scientific and daily operations, leading to a situation where AI applications are not seamlessly integrated with standard daily procedures [13].

The field of process science increasingly intersects with data science, as practitioners leverage advanced large language model (LLM) tools, such as ChatGPT, to develop AI and ML solutions tailored to specific problems. Although this approach seems to effectively democratize data analysis, enabling broader access and utilization, it also carries inherent risks, such as the suitability of the information utilized to produce an output from LLMs. Plus, process scientists may lack a comprehensive understanding of the underlying methodologies.

Figure 1: The benefits of AI/ML algorithms should encompass key aspects to enhance end user engagement.



Gaps in expertise, which can lead to misguided decisions, were identified as a current industry challenge in a 2022 report issued by the Association of the British Pharmaceutical Industry (ABPI). "This is a huge growth area for healthcare data science, and demand currently outstrips supply. A significant proportion of strong data science candidates have limited life sciences experience, and this must be taught on the job for them to become effective," said a respondent of a survey cited in the report [14].

AN INTEGRATED APPROACH

To fully leverage AI and ML in bioprocess development, it is crucial for AI/ML tools to be adopted with a more integrated approach. This means moving beyond the confines of specific domains of data science and making the tools accessible to interdisciplinary teams, including bioprocess engineers and other domain experts. By doing so, data scientists would ensure that AI/ML tools (models and algorithms) they develop are working as intended, and the bioprocess scientists would ensure the tool can be utilized in a meaningful and actionable way.

Teams involved in such projects will vary depending on the application and complexity. It is important to identify the stakeholders, which could include engineers developing a specific application to improve a control strategy, automate reporting, or create a feedback model incorporating a large amount of data (e.g., from research and development, production, or supply chain) to identify learnings that can be leveraged for the next product creation.

The integration of such tools and applications can have impacts at all levels of operations of an organization, which should be carefully considered to ensure time and resources are deployed in an effective manner. Generating data or applications does not always generate more value, and creating a company culture to embrace changes and adopt new tools or workflows can be challenging.

The Fogg Behavior Model [15] highlights this challenge and recommends key aspects for consideration. Introducing new tools or workflow requires introducing a new behavior to someone's workday. Behavior is driven by three ingredients working simultaneously: motivation, ability, and prompts (see Figure 1). This is important because you could have the best tool in the world, but if there is no motivation to use it, it is overly complicated for the end user, or there are no prompts or routine to incorporate its use, it will bring no value. Considering potential hindrances early in the process and encouraging necessary behavior changes, can lead to smooth rollouts that meet milestones..

Incorporating data analytics into the toolkit of process scientists and connecting company-specific applications from data scientists within their organization is a step toward creating more interdisciplinary collaboration. This article introduces the concept of working with both commercially available software platforms combined with open source to facilitate the rapid configuration of data analytic workflows. These workflows, crafted by data scientists, are intended for seamless adoption and utilization by process scientists, thereby bridging the gap between complex data analysis and practical application in the field.

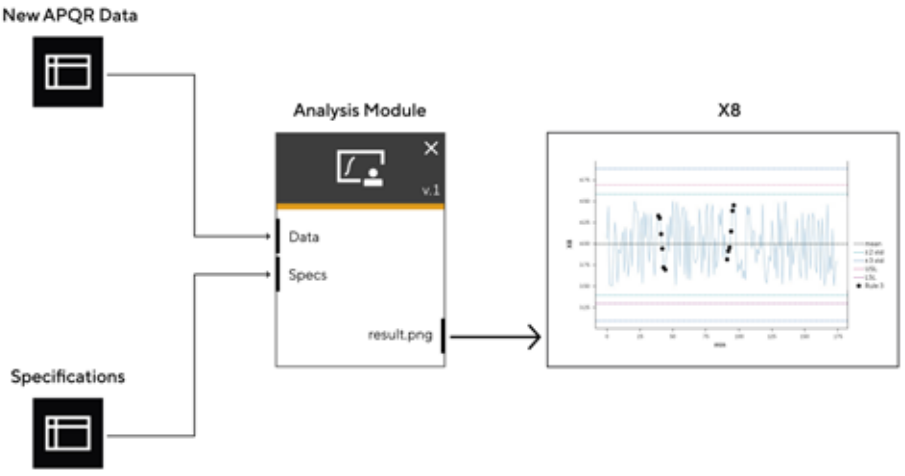
CASE I: QUALITY AND PERFORMANCE ASSESSMENT

A safe and efficacious product is the overarching goal for both manufacturers and authorities. To enable this goal, many guidelines and regulations are implemented and supported by internal company procedures. One central requirement for commercial products according to "Q7 Good Manufacturing Practice Guidance for Active Pharmaceutical Ingredients Guidance for Industry," updated by the FDA in 2020, is that "Regular quality-reviews of APIs should be conducted with the objective of verifying the consistency of the process. Such reviews should normally be conducted and documented annually…" [16].

Accordingly, this guidance may be systematically documented within the annual product quality review (APQR).

A standard workflow would involve a cross-functional team comprising a process scientist, an analytical expert from quality control (QC), or a manufacturing specialist who is responsible

Figure 2: This figure demonstrates workflow to assess data in terms of APQR. Breaches of Nelson Rules are indicated by black dots.



for aggregating data from various sources. This collated data is then submitted to the statistical department where professionals conduct statistical analyses in accordance with the company's standard operating procedures. The completed analyses would then be shared with the originating department. Subsequently, the manufacturing or QC team undertakes a thorough review and interpretation of the outcomes to ensure compliance with quality standards to inform continuous improvement initiatives or open investigations for noncompliance findings.

So how can that workflow be simplified so that the analysis can be done directly by manufacturing or the QC department? A workflow builder offers exactly this possibility to generate workflows to execute repetitive tasks and deploy the workflow to colleagues. How would a workflow be created? First, a data connection to the well-defined and qualified source data needs to be established. This should enable a secure connection to all relevant data (such as defined operating ranges, critical quality attributes [CQAs], and performance indicators) and increase the overall data integrity compliance for the application.

The statistical department is tasked with developing a module tailored for the statistical analysis and reporting of this data. It is crucial to ensure data format is standardized to facilitate seamless processing. A common methodology to support this analysis are Nelson Rules which are typically applied to control charts to identify datapoints that fall outside of specification as well as trends that are against random noise.

In this example, the statistical analysis would be to apply four different Nelson Rules to data (see Figure 2). Data points that violate one or more Nelson Rules should be flagged because they need to be investigated. The statistical model developed should encompass not only the statistical evaluation but also a results component. In the context of our example, this would include a graphical representation, such as a plot, to identify instances where Nelson Rules may have been breached.

The foremost purpose of the APQR is to quickly assess if the production process was under control or within specified limits in the previous reporting period. In the workflow presented previously, the manufacturing or QC employee requests to run the workflow which would be followed by a generated report. Another large advantage to automating reporting workflows is compliance with regulatory standards, as the steps in the process will require clear traceability to qualified systems. This reduces the possibility for manual errors and ensures compliance through automation and data integrity standards, ultimately resulting in higher operation efficiency from more effective resource utilization.

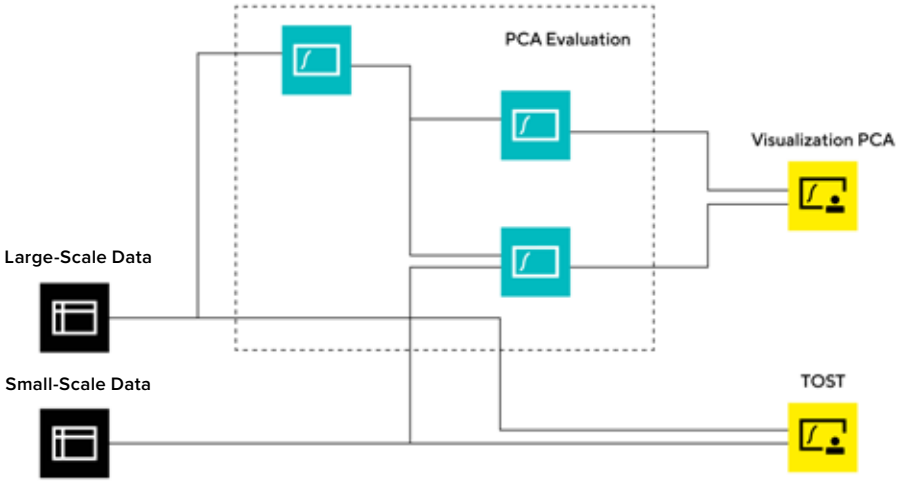
There are a variety of further use cases that can benefit from automated workflows, for example, the APQR result can be used for continuous process improvement. To achieve this, a second workflow can be designed for the manufacturing science and technology (MSAT) department using the very same data and data connection, but running data mining tools or digital twin kind of models to identify potential improvements to the process.

This is just one example of an application where robotic process automation (RPA) can be utilized. RPA has significant potential in the pharmaceutical industry, with benefits in both operational and compliance aspects. RPA can remove repetitive tasks, speed up decision-making, reduce the risk of errors in data collection and analysis, and expedite reporting to enable process optimization [17].

CASE 2: SCALE COMPARABILITY

The second use case has a slightly more complex structure but is equally important. Scaling in bioprocessing is a critical phase in the development and commercialization of biopharmaceutical products. It involves the careful transition of biological processes from the laboratory bench or pilot scale to larger-scale production systems. This step is essential for meeting the demand for therapeutic proteins, vaccines, and other biologics. On the other hand, there might be cases where scaling processes downward is

Figure 3: Workflow for scale comparison



needed to support troubleshooting or life cycle management of commercial processes.

Effective scaling is not merely a matter of increasing the size of the bioreactors and other equipment. It requires a deep understanding of the biological systems involved and the impact of scale on process parameters, such as mixing, oxygen transfer, and nutrient supply. Although physical properties like tip speed can be easily calculated, others such as the volumetric mass transfer coefficient (kLa) are more challenging to derive. One fundamental problem remains within bioprocessing—just because most physical considerations are similar across scales, the biological response of the living cells in the reactor might be different.

At the end of any scaling, it is important to assess if the scaling was successful or not, meaning the biological performance, including the quality of the product, is equivalent across scales. Typically, different statistical methods are used to check if quality is preserved across scales or that certain key performance indicators (KPIs) are met. The toolbox of statistical tools applied ranks from univariate comparison such as t-tests or equivalence testing to multivariate approaches to assess time series data.

Scaling is required for most commercially manufactured products, meaning that scale comparison and assessment is a repetitive task. Here again a generic workflow can be of great value to standardize and automate the evaluation. Scaling can be applied in both directions (scale-up and -down). This example, highlighted in Figure 3, shows the workflow for scale-down.

Workflows should address whether scales are equivalent. The first step is to decide what methods are needed to show the scale comparability. In this case, a principal component analysis PCA for time series data and two one-sided t-tests (TOST) for key performance indicators (KPIs) and critical quality attributes (CQAs) were selected. Typically, the data science department could then start to program the relevant modules. An alternative approach to build the models or statistical evaluation could be the use of self-programming tools like ChatGPT (which will be highlighted in case 3). Once the statistical modules are finished, the workflow can be assembled and tested. The workflow consists of three parts: data connection (black boxes in Figure 3), MVDA (teal-colored boxes in Figure 3), and assessment of results (yellow boxes in Figure 3).

The outcome of the workflow is shown in Figure 4. Figure 4A depicts the result of the PCA analysis of the time series data. First, the large-scale data is compressed using PCA. As a next step, the distance to origin for each large-scale batch is determined and two standard deviation 2SD limits calculated (shaded blue area). In the plot shown, the small-scale data is compared to limits based on the large-scale data. Batch 56 and batch 57 are outside and are displayed in detail in the plot next. Figure 4B shows boxplots of the harvest titer as a representation of a KPI for both large-scale and small-scale data. The more KPIs and CQAs, the more boxplots need to be generated. Figure 4C shows the result of the TOST for the harvest titer. The more KPIs and CQAs, the more TOSTs need to be performed.

As a final step, the process scientist would simply update the small-scale data and run the process. The first step in improving the small-scale model is in understanding if the scales are comparable. It requires the right tool set and an understanding of how to use data analytics for the assessment. A workflow like the one shown in this section can help get to a reliable assessment quicker to the right end user (the bioprocess subject matter expert [SME]). In case the evaluation demonstrated comparability, no further actions are needed other than the summary of the results and a formal report on small-scale validation.

However, in case the result of the workflow is nonequivalence between the scales, additional work needs to be done. Typically, the next stage is to understand what the scaling parameter culprit for nonequivalence could be across scale. Solving this problem could lead to another workflow that uses AI/ML with a high degree of explainability. There are examples demonstrating how design of

Figure 4: Results of Scale Comparison: A) PCA analysis results of the time series data B) Boxplots of the harvest titer, representing a KPI for both large-scale and small-scale C) TOST results for the harvest titer

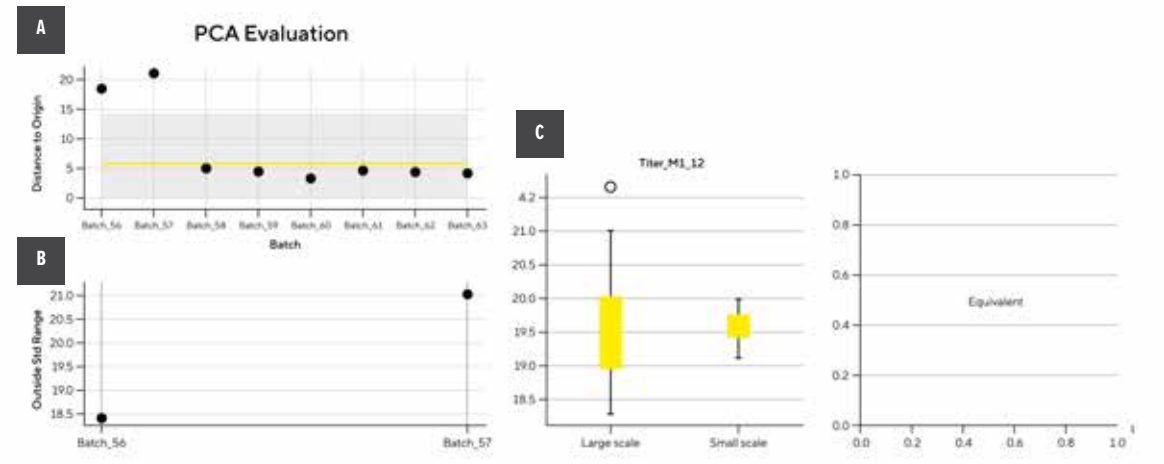
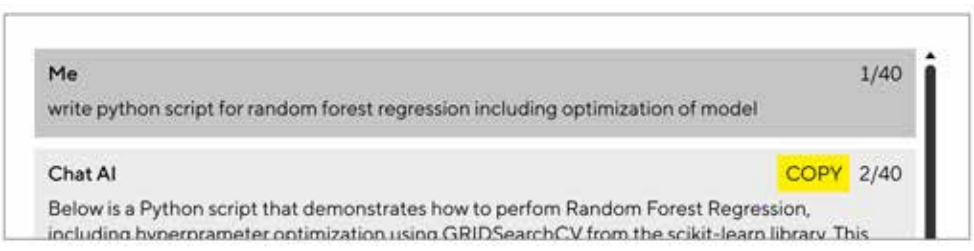


Figure 5: ChatAI workflow



experiment in combination with MVDA can help identify possible adjustments for more comparable biological performance between scales [18].

Another approach to detect and understand nonequivalence can be the use of hybrid models (also called digital twins). Both techniques can be based on AI/ML algorithms trying to represent the biological behavior within a process model. These models can be used to investigate different process settings and assess which ones will result to more comparable process [19]. Once a new scale-down approach was found (independent of the algorithm used), the equivalence testing (e.g., with the workflow shown here) must be repeated.

CASE 3: DATA MINING TOOL BASED ON RANDOM FOREST

A common and important task in bioprocess development is mining historical data to utilize existing knowledge to establish a starting point for any new bioprocessing development activities. Typically, and following quality by design (QbD) principles, this activity starts with risk assessment to identify the most important parameters to be investigated in design of experiment (DOE) approaches.

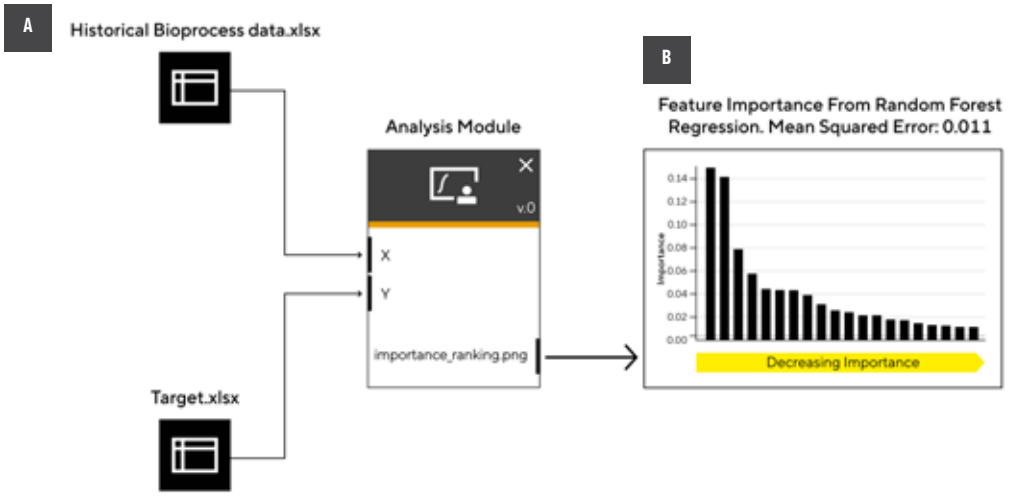
Part of the risk assessment involves looking over dozens or more data spreadsheets to identify trends and behaviors. However, looking over the historical data set can be a complex and time-consuming exercise and if not done properly, there is the risk of overlooking important parameters and starting the development journey from the wrong starting point. Powerful data mining tools such as principal component analysis or random forest can be helpful tools to identify trends and patterns in data; however, due to the limited expertise of process scientists to create these models, these techniques are not routinely utlized.

To provide the process scientist with a summary of the most important and influential parameters as a starting point for follow-up work and to reduce the time required to analyze dozens of data sheets, a random forest algorithm can be implemented. For this example, a random forest algorithm to identify the strongest influencing parameter(s) on the harvest titer of this bioprocess is required.

In the third use case, we want to demonstrate how problem-solving can be boosted by using chatbots for the programming task. In our case, the script's development timeline was reduced from several hours to less than 5 minutes. The process scientist used an internal company AI chatbot (see Figure 5) to develop a robust method with model optimization and feature extraction.

The resulting Python script contained the data import interface, a grid search for testing key parameters for the random forest, and a graphical output to visualize and rank the most important

Figure 6: Result data mining with random forest.



parameters. The resulting script was then embedded in an envelope script to be uploaded into a workflow builder.

Next the different building blocks are assembled, as indicated in Figure 6A. Figure 6B shows the top 20 variables in decreasing order of importance. This would be the basis for process science to identify which parameters to address first for bioprocess development.

This same data mining strategy can be reused or shared with colleagues for the next new product or target assessment and only the respective input files need to be exchanged. It should be noted that the output of the models and reliability of the data can only be as good as the data input. Data mining remains a crucial tool to set a solid foundation for any further development activity right from the start. This solid foundation not only helps the SME profoundly understand the dependencies with the process but opens more opportunities like DOEs, digital twin development, or further (self)-learning algorithms as the process matures.

TYING IT ALL TOGETHER

The ultimate value in the implementation of AI and ML to create efficiency and drive innovation will come with full connectivity and utilization of data created across the entire value chain. For this to be realized, the importance of well-structured, digitized, compliant, and accessible data should not be underestimated. The three use cases outlined previously could also bring greater value to an organization if they were approached in a broader, holistic way. Case 1 highlighted the use of RPA that can be used for compiling data, automating workflows, and reporting. Case 2 highlighted the use of cross-scale analysis, and Case 3 highlighted the use of data mining.

Combining these techniques and applications builds towards a future vision of fully integrated systems and processes that streamline product development, submission documents, process control settings, and operational information. This integration allows manufacturing protocols to be informed by real-time data, enables RPA reporting for batch comparisons, and consolidates relevant data for real-time release. Additionally, it could track market trends for informed production decisions, creating a feedback loop that supports future product development through data mining.

CONCLUSION

The integration of AI and ML methodologies is increasingly revolutionizing operational processes, enhancing efficiency, and expediting decision-making. This technological advancement holds the promise of accelerating the delivery of superior-quality pharmaceuticals to patients, thereby improving healthcare outcomes.

Despite the transformative potential of AI/ML techniques, they are not without their challenges. To begin with, there is the requirement for a well-structured and unified database to enable effective contextualization and analysis of data. The proper application of these technologies is paramount to avoid misuse and to ensure the integrity of the outcomes they produce.

The risk of misuse of AI/ML increases with chatbots for coding, as it offers the possibility to generate complex codes without really understanding the principles or even pitfalls of the approach chosen. But a profound comprehension of the underlying principles of the employed AI/ML methodologies is essential, as is a thorough understanding of their development and validation processes. This ensures that the results yielded are not only reliable, but also safe for drawing conclusions.

A common obstacle in the implementation of AI/ML solutions is the knowledge gap between data scientists and domain experts. Data scientists may lack the specialized expertise required to interpret results within a particular domain, whereas domain experts may not possess the technical expertise to construct and evaluate AI/ML models.

The proposed strategy aims to bridge this divide by fostering a symbiotic relationship between data scientists and domain experts. By designing and testing specific AI/ML workflows,

data scientists can tailor these tools for use by domain experts. This collaborative model not only ensures that workflows are user-friendly and domain-specific, but also encourages ongoing cooperation throughout the initial scope definition phase. Such a partnership is instrumental in aligning the technical capabilities of AI/ML with the nuanced requirements of domain expertise, ultimately leading to more effective and innovative solutions.

References

1. US Food and Drug Administration. "Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices." 7 August 2024. Accessed 22 March 2024. www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
2. Rathore, A. S., S. Nikita, G. Thakur, and S. Mishra. "Artificial Intelligence and Machine Learning Applications in Biopharmaceutical Manufacturing." *Trends in Biotechnology* 41, no. 4 (April 2023):497–510. doi:10.1016/j.tibtech.2022.08.007
3. US Pharmacopeial Convention. *USP<1039> Chemometrics.* Rockville, MD: US Pharmacopeial Convention, 2024. https://doi.usp.org/USPNF/USPNF_M2345_02_01.html
4. US Food and Drug Administration. "Guidance for Industry. PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance." October 2004. https://www.fda.gov/media/71012/download
5. European Medicines Agency. "Guideline on the Use of Near-Infrared Spectroscopy by the Pharmaceutical Industry and the Data Requirements for New Submissions and Variations." 12 June 2014. https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-use-near-infrared-spectroscopy-pharmaceutical-industry-and-data-requirements-new-submissions-and-variations_en.pdf
6. US Food and Drug Administration, Health Canada, and the United Kingdom's Medicines and Healthcare Products Regulatory Agency. "Good Machine Learning Practice for Medical Device Development: Guiding Principles." October 2021. www.fda.gov/media/153486/download
7. US Food and Drug Administration. "Artificial Intelligence and Machine Learning (AI/ML) for Drug Development." 18 March 2024. www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development
8. European Medicines Agency. "Reflection Paper on the Use of Artificial Intelligence (AI) in the Medicinal Product Lifecycle." 13 July 2023. www.ema.europa.eu/en/documents/scientific-guideline/draft-reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf
9. BioPhorum. "BioPhorum Digital Plant Maturity Model (DPMM) 3.0 Introduction." 15 January 2024. Accessed 27 March 2024. www.biophorum.com/webinar/biophorum-digital-plant-maturity-model-dpmm-3-0-introduction
10. Altrabsheh, E., M. Heitmann, P. Steinmüller, and B. P. Vinco. "The Road to Explainable AI in GXP-Regulated Areas." *Pharmaceutical Engineering* 43, no. 1 (January/February 2023):24–32. https://ispe.org/pharmaceutical-engineering/january-february-2023/road-explainable-ai-gxp-regulated-areas
11. Erdmann, N., R. Blumenthal, I. Baumann, and M. Kaufmann. "AI Maturity Model for GxP Application: A Foundation for AI Validation." *Pharmaceutical Engineering* 42, no. 2 (March/April 2022):18–24. https://ispe.org/pharmaceutical-engineering/march-april-2022/ai-maturity-model-gxp-application-foundation-ai
12. Blumenthal, R., N. Erdmann, M. Heitmann, A.-L. Lemettinen, and B. Stockton. "Machine Learning Risk and Control Framework." *Pharmaceutical Engineering* (January/February 2024). https://ispe.org/pharmaceutical-engineering/january-february-2024/machine-learning-risk-and-control-framework
13. McKinsey and Company. "AI in Biopharma Research: A Time to Focus and Scale." 10 October 2022. www.mckinsey.com/industries/life-sciences/our-insights/ai-in-biopharma-research-a-time-to-focus-and-scale
14. Association of the British Pharmaceutical Industry. "Bridging the Skills Gap in the Biopharmaceutical Industry: Maintaining the UK's Leading Position in life sciences." January 2022. www.abpi.org.uk/media/ya2fjboi/bridging-the-skills-gap-jan-2022.pdf
15. Stanford University Behavior Design Lab. "Fogg Behavior Model." Accessed 22 March 2024. https://behaviordesign.stanford.edu/resources/fogg-behavior-model
16. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. "Q7: Good Manufacturing Practice for Active Pharmaceutical Ingredients." Published November 2000, updated 14 April 2020. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/q7-good-manufacturing-practice-guidance-active-pharmaceutical-ingredients-guidance-industry.
17. First Gulf Solutions. "Revolutionizing Pharmaceuticals: How Robotic Process Automation (RPA) Is Enhancing Efficiency and Compliance." 31 October 2023. Accessed 21 March 2024. https://www.linkedin.com/pulse/revolutionizing-pharmaceuticals-how-robotic-process-0pzxf/
18. Sartorius. "Simplifying the Scaling Process Between Bioreactors and Using Small-Scale Models as a Starting Point for Large-Scale Multivariate Control." 2022. Accessed 30 January 2024. www.sartorius.com/en/pr/webinar-simplifying-the-scaling-process-between-bioreactors#id-946980
19. Sartorius. "Methods to Investigate and Identify Scale Differences for Bioprocessing Using SIMCA and Cell Insights by Umetrics." 2023. Accessed 30 January 2024. https://register.gotowebinar.com/recording/3612183103112406272

About the authors

Chloe Lang leads a team of application engineers for Europe, the Middle East, and Africa and Asia-Pacific regions at Sartorius Digital Solutions. With eight years of service at Sartorius, Chloe has specialized in the implementation of advanced data analytics, focusing on enabling organizations to effectively use these tools. Before her tenure at Sartorius, Chloe was part of the Process Analytical Technology (PAT) group for continuous manufacturing at Novartis, in Basel, Switzerland. Currently based in Sweden, Chloe continues to contribute her expertise and leadership to the field of data analytics, driving innovation and efficiency in manufacturing processes. She joined ISPE in 2017.

Timo Schmidberger, PhD, began his career at Novartis in 2008, specializing in mammalian cell culture process development. Over nine years, he excelled in scaling, process transfer, characterization, and statistical evaluation, eventually leading an in-process control QC laboratory. In this role, he provided analytical support to three production lines. In 2018, Timo joined Sartorius, leveraging his extensive bioprocessing and GMP knowledge. At Sartorius, he is known for his expertise in data analytics, including ML and hybrid model development for bioprocesses. Timo's contributions advance data science in biotechnology.

