

2 September 2021

Beyond building predictive models: TwinOps in biomanufacturing

Elisa Canzani¹, Sander W. Timmer

1. Cognizant Technology Solutions

Abstract

On the wave of more and more manufacturers embracing the pervasive mission to build digital twins, also biopharmaceutical industry envisions a significant paradigm shift of digitalisation towards an intelligent factory where bioprocesses continuously learn from data to optimise and control productivity. While extensive efforts are made to build and combine the best mechanistic and data-driven models, there has not been a complete digital twin application in pharma. One of the main reasons is that production deployment becomes more complex regarding the possible impact such digital technologies could have on vaccine products and ultimately on patients. To address current technical challenges and fill regulatory gaps, this paper explores some best practices for TwinOps in biomanufacturing -- from experiment to GxP validation -- and discusses approaches to oversight and compliance that could work with these best practices towards building bioprocess digital twins at scale.

Keywords

Hybrid models, automation, bioengineering, computing and processing, continuous validation, digital twin applications, gxp compliance, production deployment, regulated industry, twinops

Beyond building predictive models: TwinOps in biomanufacturing

Elisa Canzani

Cognizant Technology Solutions

Sander Timmer

GlaxoSmithKline Biologicals SA

Abstract— On the wave of more and more manufacturers embracing the pervasive mission to build digital twins, also biopharmaceutical industry envisions a significant paradigm shift of digitalisation towards an intelligent factory where bioprocesses continuously learn from data to optimise and control productivity. While extensive efforts are made to build and combine the best mechanistic and data-driven models, there has not been a complete digital twin application in pharma. One of the main reasons is that production deployment becomes more complex regarding the possible impact such digital technologies could have on vaccine products and ultimately on patients. To address current technical challenges and fill regulatory gaps, this paper explores some best practices for TwinOps in biomanufacturing – from experiment to GxP validation – and discusses approaches to oversight and compliance that could work with these best practices towards building bioprocess digital twins at scale.

■ **BIOPROCESS DIGITAL TWINS** are virtual representations of the complete bioprocess chain - from raw materials to drug products - aiming to predict critical quality attributes (CQAs) and performance attributes (PAs) across all upstream, downstream, and secondary operations. The ultimate goal is to solve all bioprocessing tasks by providing next-best-actions, confidence of predictions, and financial impacts. Many manufacturing industries already have twins with a certain degree of integration

with the control of physical systems [1,2]; however, medical and biotech companies are struggling for productionizing digital twin applications [3].

Most of the published research in life science R&D has been focusing on simulations and “hybrid models” [4] derived by combining expert knowledge and data-driven analysis that leverages experimental data and Process Analytics Technology (PAT) data [5]. Advanced machine-learning (ML), or artificial intelligence (AI), techniques are being also employed

to alleviate the challenges of sparse and small data sets through, e.g., data augmentation, ensemble learning [6], and self-supervised learning (SSL) [7]. Could we already name “digital twins” such predictive models? [8]. Beside the off-line use of hybrid models, standards for integration and validation of twin technologies are key enablers towards a GxP-compliant digital twin product. In a few words, GxP compliance refers to a series of “Good Practices” - defined by the Food and Drugs Administration (FDA) or the European Medicines Agency (EMA) - that guide quality in regulated industries.[9].

Wu et al. [10] states that most of FDA approved AI devices have been tested and validated only on retrospective studies, while none of the high-risk devices were evaluated on prospective studies.

While no doubts are on digital twins’ value in reducing costs, increasing compliance, and optimising production of drug products [7], less clear is the DevOps architecture required to build, deploy, and maintain digital twins in strongly regulated industries such as life science [8].

Of interest is the concept of TwinOps [11], that combines practices of DevOps and model-based engineering to tackle the physical system modeling aspect. In biomanufacturing, TwinOps would ensure high-quality target profile of the final drug product, thus obeying to the Quality by Design (QbD) paradigm [4,5]. Minerva et. al. [12] propose a generic architectural model for digital twins to use in different industries. Cloud-based digital twin platforms are available on the market [13]. The concern is whether they suit complex data and model requirements along with strict regulations characterising Pharma 4.0. To the best of our knowledge, best practices for orchestration, validation, and compliance checks of automated data and model pipelines for an effective TwinOps lifecycle are in their very early stage in the pharmaceutical industry.

To address these strong regulatory needs from a technology perspective, this paper outlines TwinOps requirements and tailored considerations for digital twin applications in biomanufacturing. Concluding that proper TwinOps will bring bioprocess digital twins as well as the whole Pharma 4.0 to the next level.

1. TWIN-OPS LIFECYCLE

For DevOps and automation, context is key: there is no “one size fits all” [14]. High complexity of

bioprocesses and strict GxP regulations steer TwinOps lifecycle customisation in pharma (Fig. 1).

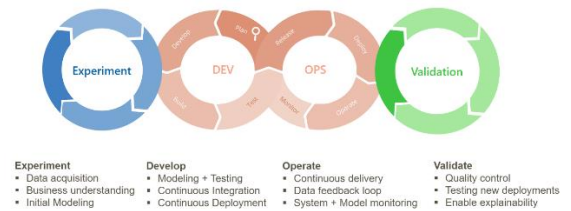


Figure 1. Twin-Ops lifecycle from experiment to validation.

Experiment

The experimental phase starts with understanding the physical process, i.e., complex dynamics of fed-batch cell cultures in bioreactors, to enable its mapping with the digital product. Lack of correlation between process definition and product definition leads to twin solutions based on “ideal definitions” of the process turning into twin’s limited performance in production [15].

During experimentation, first data acquisition from past process samples or through Design of Experiments (DoE) techniques allows parameter fitting and validation of initial modeling with actual measurements.

DevOps

Many companies are facing the challenge of unlocking the value of data through data science technologies [16,17]. This is because there is no adoption of DevOps concepts since initial stages of data acquisition and the modeling exercise. In biomanufacturing, making data-driven ML components, simulators, and mechanistic models work together is not trivial. TwinOps are DevOps practices extended to hybrid models, their integration, and validation.

TwinOps must ensure management and integrity of training data (from offline database, DoE, simulations, real-time sensor, and PAT data), needs for model re-training with different requirements according to twin product components (e.g., state estimator, optimiser, controller of the bioprocess), as well as concerns of model transparency and explainability. In fact, proper DevOps is also key enabler for the validation phase.

Challenges in orchestrating and deploying data ingestion pipelines from multiple sources and testing model releases in an automated fashion using

Continuous Integration and Continuous Delivery (CI/CD) pipelines are discussed in Section 2.

Validation

Validating data and models is a mandatory step for deploying applications in regulated industries such as life science. Digital twin products must be GxP compliant. This means, before production releases, new generated data and models must go through an ‘approval gate’ to check GxP compliance of our bioprocess digital twin application.

The use of TwinOps practices plays a crucial role in overcoming quality control and validation challenges. It enables explainability through automated reporting, monitoring scripts for data and model drifts, and pipelines to score models against predefined business KPIs.

2. BIOPROCESS DIGITAL TWINS IN PRODUCTION

Deploying twins for practical use is an industrial concern that goes unaddressed in research papers. With respect to the level of integration of digital twins, three kinds of twins have been identified in literature: ‘claimed twin’, ‘described twin’, and ‘actual twin’. Half of literatures described ‘digital model’ or ‘digital shadow’, although the authors claimed a digital twin in their papers [2].

A major issue is the lack of clarity and specificity of digital twin concepts regarding different industry fields. Adopting a domain-specific perspective, we propose a breakdown of complexities arising from bioprocess digital twins in production.

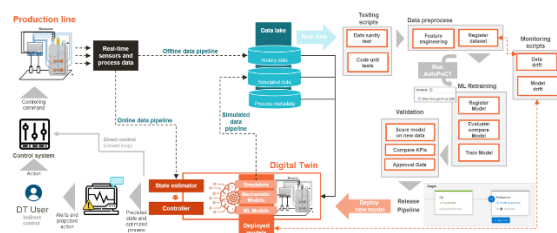


Figure 2. DevOps architecture for a bioprocess digital twin with control loops and component interactions through data and model pipelines ensuring orchestration and compliance of new deployments.

To fix ideas, Figure 2 depicts a bioreactor along the production line that the digital twin user (DT user) can control through an industrial PC (IPC) system,

performing actions suggested by the twin (in case of indirect control). Having manual control is a temporary solution we implement to demonstrate reliability of the twin and accumulate data before going to a fully automated process. The ultimate state is a fully integrate digital twin that continuously gathers and delivers data back to where it is needed in the value chain using a closed-loop monitoring and decision-making process to directly control the bioreactor.

In general, the two widely adopted components of a bioprocess digital twin are:

- a *state estimator*, which gets current conditions from sensors and process parameters to predict the process outcome;
- and a *controller*, which is triggered in case the process outcome is predicted not optimal with respect to production target.

Once triggered, the controller uses an optimiser to evaluate possible actions and perform the best one - accounting for physical process constraints and business criteria - towards steering production back to optimal target. How to build such predictive capabilities and orchestrate components of the bioprocess digital twin?

Data ingestion pipelines

There are two main types of dynamic source of data: data from real-time sensors and data from PLC. Such data is usually collected every few seconds or minutes and ingested from PAT software into the data lake. Sensor and control data history mainly comprises of in-spec events, i.e., samples around the optimal production target and control parameters set at production specifications. This means that if we retrain the state estimator with production data, it gets more and more precise around the optimal while it loses confidence when quality is out of specification ranges.

However, the main goal of the twin is controlling the bioprocess in case of out of spec events. How to keep the production at target in case of a temperature perturbation? How to handle a pipe’s loss of pressure? Initial data collection done via DoE. Since lab experiments are prohibitively expensive, experimental data has small size and is seldomly refreshed. Given the manual-entry nature of this data, a data format validation pipeline as first sanity check is recommended to handle future version storage into the data lake. Experimental data primarily serves for fit and validation of mechanistic models and simulators developed to replicate the bioprocess. Such complex

simulation models are then used to generate larger datasets of simulated out of spec events, constituting the main data source to train twin's ML components. Automated ingestion of simulated data into data lake often results a challenging task since it requires support for third-party simulation software.

Hybrid model pipelines

Main characteristic of bioprocess digital twins is the adoption of the so-called “hybrid models” [4], which help replicating complex dynamics, e.g., of cell metabolism within the bioreactor, through expert-based models while leveraging ML to learn and timely predict such dynamic patterns over time.

For off-line use, dynamic simulators serve to replicate abnormal events along the bioprocess chain and explore the solution space (process outcome) by varying process parameters. Simulated data from the mechanistic model is used to train the twin reacting on-line when the bioprocess is out of control. Such sophisticated models are often computationally slow, thus their use for real-time control is limited. ML models are employed to run on top of mechanistic models to overcome computational challenges and confer the twin adequate predictive power with respect to the actual bioprocess.

Making hybrid models work altogether require challenging re-train, validation, and deployment cycles to be performed by multiple automated CI/CD pipelines. For example, what happens when a new mechanistic model is built? We need a validation pipeline to assess simulation accuracy against process data. Qualified simulation data can thus be used to re-train the ML components, which are re-deployed in case of higher performance compared to current deployed ML models. It worth noticing that automating the deployment of the new simulation model itself might suffer limitations depending on the simulation software or source code compatibility with the twin platform.

Continuous validation

As biomanufacturing processes tends to encounter fluctuations after production starts, validating model outcomes in a continuous way is essential for having a compliant twin. By implementation of automated testing of data and models we can ensure each batch stays the limits of a golden batch, of which variability is twin-controlled. However, many described twins in the BioPharma space lack actual feedback loops to production systems. All events during production

must be recorded, the information analysed, and main findings pushed back to TwinOps in an automated feedback loop.

Testing required for regulatory approval is intensive, and hardcoded data for unit tests might not be sufficient. Testing protocols must include tests by ML models to detect anomalies, as well as tests by human to demonstrate how well the twin applies human factor engineering all along.

When developing, testing, and manufacturing GxP-regulated products, most important GxP targets are the integrity, reproducibility, and traceability of data and models. Consequently, implementing continuous validation of GxP controls turns into achieving continuous compliance of digital twins.

Automated twins

In biomanufacturing, the end-to-end process goes from raw material to the vaccines themselves. This means, we are looking at 1000s of models for our ambitious bioprocess digital twin that are all intertwined. To make this happen, one must heavily rely on automated modeling approaches that use a mixture of SSL methods [7], hyperparameter tuning, and model searches for given defined repeatable patterns. This way, we can with limited effort run a kind of automated proof-of-concept (“Auto PoC” in Figure 2) to determine if data quality and volume is of sufficient nature to create a baseline model. This baseline model can then be deployed to the production platform without human interaction, as all validation and reporting steps are automated as well. In one hand, such level of automation ensures to stay fully compliant with almost no effort needed. On the other hand, the “Auto PoC” step also has the advantage that it can be reiterated every time new or more data comes in, to help deciding on model re-training. Based on our experience in implementing TwinOps along the ongoing developments of bioprocess digital twins, Table 1 provides an estimation of feasible automation targets by model type from a technology perspective. Increasing twin automation clearly means envisioning a “twin core product” which is more real-world data-centric than based on time consuming mechanistic models that needs SME, and simulated data not reflecting real world variance and noise.

Table 1. Estimation of feasible automation targets for the TwinOps platform in biomanufacturing.

	Mechanistic models	Hybrid models	Data-driven models	Auto PoC
Data sources	Low	Medium	High	High
Data pipelines	Low	Medium	High	High
Modelling	Low	Medium	High	Fully
Validation	Low	Medium	High	Fully
Deployment	Medium	Medium	High	Fully
Monitoring	High	High	Fully	Fully

3. CONCLUSION AND WAY FORWARD

The novelty of this paper lies in adopting a technology-focused lens to analyse digital twin requirements for the biomanufacturing industry. We believe this shift from providing general twin concepts of previous studies to a domain-specific perspective is necessary to break down complexity of bioprocessing operations.

Future work is needed to further prove the use of automation and DevOps through well-defined use-cases along the bioprocess chain, such as fermentation in bioreactors, downstream purification steps, as well formulation and filling to ensure potency, long shelf-lives and stability of drug products.

In fact, adopting TwinOps practices since day zero when developing bioprocess digital twins for any given use-case will enable shaping the “twin core product”. In tech terms, the core product aims at gathering domain-specific procedures and templates for continuous GxP compliance, holding across other use-cases. Considerations and best practices for compliant production deployment encompassed by this article – to a certain extent and level of details – will help moving the whole Pharma 4.0 forward.

Beside building predictive models, we encourage biotech companies to put more efforts on identifying technical gaps and unblocking the use of DevOps and automation practices for achieving challenging regulatory targets towards compliant digital twin products.

ACKNOWLEDGMENT

We thank Sandrine Dessoy, digital innovation and automation lead in R&D at GSK Vaccines, for reviewing this work.

Conflict of Interest

This work was sponsored by GlaxoSmithKline Biologicals SA. Sander Timmer is employee of the GSK group of companies. Elisa Canzani is an employee of Cognizant Technology Solutions, a technical consultancy firm contracted by GlaxoSmithKline Biologicals SA in the context of this work.

REFERENCES

1. F. Tao, H. Zhang, A. Liu and A. Y. C. Nee, “Digital Twin in Industry: State-of-the-Art,” in *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2405-2415, April 2019, doi: 10.1109/TII.2018.2873186.
2. Liu, M., Fang, S., Dong, H. and Xu, C. Review of digital twin about concepts, technologies, and industrial applications. *Journal of Manufacturing Systems*, 2020.
3. Chen, Y., Yang, O., Sampat, C., Bhalode, P., Ramachandran, R., & Ierapetritou, M. (2020). Digital Twins in Pharmaceutical and Biopharmaceutical Manufacturing: A Literature Review. *Processes*, 8(9), 1088.
4. Benjamins S, Dhunoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine*. 2020 Sep 11;3(1):1-8.
5. Zobel-Roos S, Schmidt A, Mestmäcker F, Mouellef M, Huter M, Uhlenbrock L, Kornecki M, Lohmann L, Ditz R, Strube J. Accelerating biologics manufacturing by modeling or: is approval under the QbD and PAT approaches demanded by authorities acceptable without a digital-twin? *Processes*. 2019 Feb;7(2):94.
6. Oyetunde T, Liu D, Martin HG, Tang YJ. Machine learning framework for assessment of microbial factory performance. *PloS one*. 2019 Jan 15;14(1):e0210558.
7. A. Saeed, F. D. Salim, T. Ozcelebi and J. Lukkien, "Federated Self-Supervised Learning of Multisensor Representations for Embedded Intelligence," in *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1030-1040, 15 Jan.15, 2021, doi: 10.1109/JIOT.2020.3009358..
8. Portela RM, Varsakelis C, Richelle A, Giannelos N, Pence J, Dessoy S, von Stosch M. When Is an In Silico Representation a Digital Twin? A Biopharmaceutical Industry Approach to the Digital Twin Concept. Springer, 1-21, 2020.
9. Deloitte. Unravelling complexity: The challenge of compliance in the life sciences supply chain. Deloitte centre for health solutions. April 2017.
10. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, DE and Zou, J., 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, pp.1-3.
11. Hugues, J., Hristosov, A., Hudak, J.J. and Yankel, J., 2020, October. TwinOps-DevOps meets model-based

engineering and digital twins for the engineering of CPS. In Proceedings of the 23rd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings (pp. 1-5).

12. G. R. Minerva, G. M. Lee and N. Crespi, "Digital Twin in the IoT Context: A Survey on Technical Features, Scenarios, and Architectural Models," in Proceedings of the IEEE, vol. 108, no. 10, pp. 1785-1824, Oct. 2020, doi: 10.1109/JPROC.2020.2998530.
13. Microsoft. "Enhance innovation and agility with Digital Twins." Whitepaper, 2020. Url:https://download.microsoft.com/download/6/5/2/652a751d-d733-4668-84e4-764cfec34a70/Manufacturing_Digital_Twins_eBook_Final.pdf
14. I. Ozkaya, "Are DevOps and Automation Our Next Silver Bullet?" in IEEE Software, vol. 36, no. 04, pp. 3-95, 2019.doi: 10.1109/MS.2019.2910943
15. J. Wu, Y. Yang, X. Cheng, H. Zuo and Z. Cheng, "The Development of Digital Twin Technology Review," 2020 Chinese Automation Congress (CAC), IEEE Publisher, pp. 4901-4906, 2020, doi: 10.1109/CAC51589.2020.9327756.
16. C. Ebert, J. Heidrich, S. Martinez-Fernandez and A. Trendowicz, "Data Science: Technologies for Better Software" in IEEE Software, vol. 36, no. 06, pp. 66-72, 2019.doi: 10.1109/MS.2019.2933681
17. Sun, C., Shrivastava, A., Singh, S., and Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision. p. 843-852. 2017.

Sciences team. He holds a Ph.D. from the University of Cambridge and EMBL – European Bioinformatics Institute focusing in Big Data Biology and Genomics. Contact him at sander.x.timmer@gsk.com

Elisa Canzani, PhD, is lead data scientist at Cognizant Technology Solutions and has considerable experience in delivery end-to-end AI/ML solutions across a wide range of industry domains. Elisa has an international academic background as EU Marie Curie research fellow and holds a Ph.D. in Computer Science from the German Army Force University in Munich, in collaboration with Airbus Group and Stanford University. She has earned bachelor and master's degrees in Applied Mathematics at University of Camerino, in Italy. Contact her at elisacanzani@cognizant.com, or elisa.x.canzani@gsk.com.

Sander Timmer, PhD, is Director of Analytics and AI at GSK Vaccines, leading strategic efforts to utilise data and AI/ML to create a data driven biopharma. Sander has over a decade of Data Science experience, from writing neural nets to leading teams of data scientists and data engineers delivering production AI/ML systems. Before GSK, Sander was a Data & AI Industry Solutions Manager in Microsoft worldwide Health and Life