

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Golzarijalal, M;Aickelin, U;Otte, E

Title:

Addressing Small Data Challenges in Biopharmaceutical Development and Manufacturing: A Mini Review of Multi-Fidelity Techniques

Date:

2026

Citation:

Golzarijalal, M., Aickelin, U. & Otte, E. (2026). Addressing Small Data Challenges in Biopharmaceutical Development and Manufacturing: A Mini Review of Multi-Fidelity Techniques. *Biotechnology and Bioengineering*, 123 (6), pp.1-16. <https://doi.org/10.1002/bit.70213>.

Persistent Link:

<https://hdl.handle.net/11343/368723>

License:

[CC BY](#)

MINI REVIEW OPEN ACCESS

Addressing Small Data Challenges in Biopharmaceutical Development and Manufacturing: A Mini Review of Multi-Fidelity Techniques

Mohammad Golzarizjalal^{1,2} | Uwe Aickelin^{1,2} | Ellen Otte^{2,3}

¹School of Computing and Information Systems, The University of Melbourne, Melbourne, Victoria, Australia | ²The ARC Digital Bioprocess Development Hub, The University of Melbourne, Melbourne, Victoria, Australia | ³CSL Innovation, Melbourne, Victoria, Australia

Correspondence: Mohammad Golzarizjalal (m.golzarizjalal@unimelb.edu.au)

Received: 14 November 2025 | **Revised:** 10 April 2026 | **Accepted:** 11 April 2026

Funding: Australian Research Council, Grant/Award Number: IH210100051

Keywords: biopharmaceutical processes | Gaussian processes | graph neural networks | model interpretability | multi-fidelity modeling | neural networks

ABSTRACT

The growing demand for biopharmaceutical products reflects their effectiveness in medical treatments. However, developing new biopharmaceuticals remains a major bottleneck, often taking up to a decade before market approval. Machine learning (ML) models have the potential to accelerate this process, but their success depends on access to large and diverse data sets for training. Multi-fidelity ML techniques offer a promising solution by integrating abundant, low-cost, and less accurate low-fidelity (LF) data with limited, expensive, and more accurate high-fidelity (HF) data. In this framework, LF data capture global system trends, while HF data refine and align model predictions with the available ground truth. Such integration can substantially reduce development costs and timelines by minimizing the need to acquire HF data, for example, through extensive experimental campaigns. This work reviews developments in surrogate modeling within the biopharmaceutical context, including Gaussian processes, neural networks, and physics-informed approaches. It also provides practical recommendations for identifying appropriate LF and HF data. Existing research has primarily focused on upstream processing and drug discovery, highlighting opportunities to extend these methods to other stages, like downstream processing. While Gaussian processes and neural networks remain the most frequently used models, emerging architectures such as Transformer and diffusion models present promising directions for future research.

1 | Introduction

Recently, there has been a growing demand for biopharmaceutical manufacturing products such as monoclonal antibodies (mAbs). This is due to their ability in targeting and binding to specific antigens, making them a predominant treatment option for several conditions, including cancer, infectious diseases, and inflammation (Jin et al. 2022; Li et al. 2025). This advantage has caused the biopharmaceutical market to substantially expand over the past decade, constituting more than 20% of new drugs approved each year (Khuat et al. 2024), with an expected market projection of about \$566 by 2032 (Mikulic 2024). The long development time

for new biopharmaceuticals, that is, 10–15 years from discovery to development, emphasizes the need for predictive models that can speed up product development and improve process control, while minimizing the high costs associated with experimental trials.

Machine learning (ML) algorithms have gained significant interest for monitoring, control, and optimization of biopharmaceutical processes (Rathore et al. 2023). These algorithms can learn from historical data obtained from existing process knowledge (Pham et al. 2023; Rathore et al. 2023), facilitating the achievement of the quality-by-design (QbD) concept and offering a more economical and robust alternative to the traditional quality-by-testing

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Biotechnology and Bioengineering* published by Wiley Periodicals LLC.

approach (Hubert et al. 2014). Furthermore, ML can accelerate the timeline from antibody discovery to clinical application by predicting developability properties, such as viscosity and hydrophobicity, which allows researchers early in the discovery phase to identify and prioritize sequences that are less likely to pose chemistry, manufacturing, and control challenges (Li et al. 2025). While supervised ML has shown promise, training a reliable and robust ML model typically requires large enough data sets that are representative of the entire range of process variations (Banner et al. 2021).

However, such large data sets with data variability are often not available in the biopharmaceutical industry due to the common small data set problem (or so-called Low- N , where N is the number of historical batches; Tulsyan et al. 2018). This small data set issue can arise from several factors, including high costs of experimentation (Banner et al. 2021) and relocation of products to different manufacturing facilities to meet production demands and manage product life cycles (Tulsyan et al. 2018). Furthermore, data heterogeneity, which can result from mixing online measurements (e.g., glucose) with offline measurements (e.g., amino acids) (Hong et al. 2018), and data fragmentation, often arising from storing data separately in isolated systems, as well as low-frequency measurements and inconsistent record-keeping (Strüßmann et al. 2025), can further complicate data collection and lead to incomplete data sets. Despite recent advancements in biopharmaceutical data analytics, innovative digital strategies are still needed to overcome these data-related challenges and enhance model reliability.

Multi-fidelity surrogate models have recently shown promise in the design of systems across various engineering fields, including food storage (Golzarjalal, Ong, et al. 2025) and aerospace applications (Brevault et al. 2020), particularly where only small data sets are available. These models can optimally benefit from low-fidelity (LF) data, which are abundant, inexpensive, but less accurate sources, and combine them with high-fidelity (HF) data, which are more accurate but often costly and time-consuming to obtain. For example, bioreactor CFD simulations provide LF shear rate data compared to the HF results of physical experiments. LF data capture broad global trends, while HF data correct and refine these trends, aligning surrogate model predictions more closely with the available ground truth (Kuya et al. 2011). This approach facilitates the optimal use of resources, enhancing predictive capabilities of the model and providing a cost-effective method to address the complexities of biopharmaceutical development. Furthermore, by leveraging LF data, multi-fidelity optimization can identify optimal process conditions with fewer HF data (e.g., expensive experiments) than methods that rely solely on HF data (Peherstorfer et al. 2018).

Transfer learning is a method with an objective similar to multi-fidelity methods, which has recently gained interest for its ability to leverage pre-existing knowledge, often gained from large amounts of LF data, to improve learning efficiency in new and data-constrained environments (Barón Díaz et al. 2026). However, a clear distinction between these methodologies is established in the literature. Multi-fidelity models are typically trained on a combination of LF and HF data, often referred to in some references as multi-fidelity data fusion models (Zhang et al. 2024). In contrast, transfer learning models generally utilize LF and HF data sequentially,

involving an initial pre-training phase on LF data followed by fine-tuning on HF data (Zhang et al. 2024).

It is important to distinguish multi-fidelity models from other broad-spectrum architectures such as multi-modal models or foundation models. While multi-modal models process and integrate diverse data types (e.g., text, images, and physiological sequential data) (Wang et al. 2024) and foundation models, such as ChatGPT, are trained via self-supervised learning from massive, unlabeled data sets for general-purpose tasks (Jung 2023), multi-fidelity models specifically focus on integrating data sources of the same type but with varying levels of accuracy and cost.

This is the first paper to review the capabilities of multi-fidelity techniques across biopharmaceutical development and manufacturing life cycle, including drug discovery, upstream and downstream processing, and formulation. This review surveys the data time scale and types, as well as data acquisition methods. It also aims at summarizing and evaluating different modeling and optimization approaches reported in the literature. Moreover, research gaps are identified, and possible future directions are suggested to address the gaps. It also provides practical recommendations for industry practitioners on how to implement multi-fidelity techniques by selecting the best models or data sources that fit their available resources. For more technical details on multi-fidelity algorithms or how to derive them, readers can refer to existing reviews, such as (Giselle Fernández-Godino 2023) for modeling trends and (Peherstorfer et al. 2018) for uncertainty propagation, inference, and optimization.

2 | Data Acquisition and Heterogeneity in Biopharmaceutical Development and Manufacturing

Experimental data across the various stages of the biopharmaceutical development and manufacturing life cycle can be obtained over significantly different time scales. Some process parameters are measured in real-time using hardware-based online sensors at high frequencies, at intervals as frequent as every second (Mu'azzam et al. 2024), while others are obtained at a much lower frequency, often once per day, requiring manual sampling by skilled labor and performing offline analysis (Mu'azzam et al. 2024; Peng et al. 2026). Online process analytical technology (PAT) tools can provide spectral data, such as that from Raman or Near-Infrared (NIR) spectroscopy, providing real-time data for cell density and metabolite concentrations; however, these need a calibration model to derive meaningful measurements from raw data (Peng et al. 2026). The Combination of the online and offline data can lead to heterogeneous data sets within a specific phase and campaign. Table 1 offers a list of parameters typically recorded online or offline during biopharmaceutical development and manufacturing.

Compared to other stages, online data collection is widely used in upstream processing during bioreactor runs to monitor variables such as pH, temperature, and dissolved O_2 and CO_2 concentration (Khuat et al. 2024). While online measurements are standard for pH, conductivity, and mass flow rate in downstream processing, online monitoring of critical quality

TABLE 1 | Data types and measurement frequencies in the biopharmaceutical development and manufacturing life cycle.

Stage	Classification of representative parameters by data type and measurement frequency	References
Drug discovery ^a	Continuous: Binding affinity, IC ₅₀ , half-life, molecular weight, and docking scores Discrete: Number of hydrogen bonds, number of hits, categorical variables (e.g., active/inactive or toxic/nontoxic compounds), clinical outcomes (e.g., adverse drug reactions and patient subgroups)	Carracedo-Reboredo et al. (2021), Shareef et al. (2024)
Upstream	Continuous (online): Temperature, pH, dissolved O ₂ and CO ₂ , biomass capacitance, and spectral data (e.g., from Raman or NIR) Continuous (offline): Viable cell density, concentrations of nutrients, metabolites, and protein Discrete (offline): Feeding schedule, phase labels (e.g., exponential and production phases)	Khuat et al. (2024)
Downstream	Continuous (online): pH, conductivity, temperature, pressure, flow rate, Ultraviolet and Raman spectral data to measure mAb concentration and mAb aggregates, and Raman for measuring membrane fouling Continuous (offline): Elution conditions, product concentration, impurity profiles (e.g., host cell protein levels), aggregate levels, conductivity, and pH Discrete (offline): Number of chromatography columns, number of column cycles, and number of proteins identified in host cell proteins	Dürauer et al. (2024), Feidl et al. (2019), Roch and Mandenius (2016), Virtanen et al. (2017)
Formulation	Continuous (online): Temperature and relative humidity controlled during stability testing, Raman spectral data for measuring excipients during buffer exchange, microfluidic droplet rheometer to capture protein aggregation Continuous (offline): Protein aggregation rates, viscosity, bioavailability, chemical stability, and osmolality Discrete (offline): Number of freeze-thaw cycles, type, and number of excipients	Dash and Rathore (2021), Pavčnik et al. (2025), Vaskó et al. (2025), Zürcher et al. (2024)

^aThe online/offline framework is less applicable in drug discovery as it focuses on molecular characterization rather than temporal process monitoring.

attributes (CQAs), such as product concentration or impurity levels, is more limited due to strict requirements for high sensitivity, accuracy, and fast response times (Dürauer et al. 2024; Roch and Mandenius 2016). Nevertheless, some studies have demonstrated the use of Raman spectroscopy for measuring mAb concentration in chromatography (Feidl et al. 2019) and for monitoring membrane fouling during filtration (Virtanen et al. 2017). Similarly, the formulation step has limited applications for online monitoring of CQAs, such as the use of Raman for measuring excipients during buffer exchange (Vaskó et al. 2025). Also, since the focus in drug discovery shifts from monitoring temporal process variables to characterizing molecular and biochemical properties, the framework of online and offline measurements is generally less applicable.

Another source of data heterogeneity arises from different data types. Most process parameters are continuous (Khuat et al. 2024), such as cell densities (upstream), product concentration (downstream), and aggregation rate (formulation), while others are discrete, such as feeding schedules in upstream cell culture (Charaniya et al. 2008) and the number of chromatography columns in downstream units (Davis et al. 2021). Table 1

lists continuous and discrete variables across different biopharmaceutical development and manufacturing stages.

Several processing steps must be taken to prepare data for reliable ML training (Mu'azzam et al. 2024). To address variations in sampling intervals between online and offline measurements, temporal heterogeneity can be addressed by methods such as dynamic time warping (Ündey et al. 2002). Furthermore, data type heterogeneity can be managed by employing encoding methods (Khuat et al. 2024) or by applying algorithms that can learn from mixed-attribute data, such as fuzzy neural networks (NNs) (Khuat and Gabrys 2023) or classification and regression trees (Breiman et al. 2017). Beyond these specific challenges, generic pre-processing techniques are needed to further refine the data sets, including data cleaning, de-noising, missing value imputation, normalization (or standardization), and transformation (Mu'azzam et al. 2024). Also, dimensionality reduction, often conducted via techniques such as principal component analysis, is frequently applied to simplify models and improve predictive performance (Gargalo et al. 2020).

While these processing techniques tackle the heterogeneity challenge, the small data problem persists in biopharmaceutical

data sets, as data processing alone cannot fully resolve data scarcity. Although online PAT tools generate massive volumes of real-time information across thousands of features, the number of independent experiments remains limited mainly due to the high costs of experimentation (Banner et al. 2021). Therefore, the sample size often falls below the rule of thumb, where the number of samples should be 10–100 times greater than the number of features (Khuat et al. 2024). Furthermore, as discussed earlier, technology transfer between different facilities further constrains data availability (Tulsyan et al. 2018). Together, these factors pose challenges for the reliable training of ML models and make conventional approaches prone to overfitting. Multi-fidelity techniques can offer a promising solution to the limited data challenge, which are discussed in the following sections.

3 | Modeling and Optimization Methods

While other ML techniques, such as random forest (RF) and support vector regression, can be integrated into multi-fidelity frameworks, this review focuses on Gaussian process (GP) models, NNs, and physics-based models. These three types of surrogate models are most employed to build multi-fidelity frameworks in biopharmaceutical processing and are reviewed in the following sections (Table 2). Furthermore, a comparative analysis of their key strengths and limitations is provided in Section 3.4.

3.1 | GP

GP models are the most widely used algorithms in multi-fidelity techniques for engineering design problems, mainly because of their ability to predict confidence intervals for predictions (Petsagkourakis et al. 2021). The earliest approach was the linear autoregressive model (AR1), which performs well when a linear correlation exists between LF and HF data (Kennedy 2000). To address the limitations of linearity, the nonlinear autoregressive GP model (NARGP) was later proposed, which relates LF and HF data using a space-dependent nonlinear transformation (Le Gratiet and Garnier 2014). However, both autoregressive and nonlinear autoregressive GP models require a nested input structure (i.e., training data of lower fidelities must include data from higher fidelities) and employ a disjointed architecture (Figure 1). In these models, for each fidelity, a GP is fitted separately, and once a lower fidelity GP is built, it is not updated further, increasing model complexity and risk of overfitting (Brevault et al. 2020; Cutajar et al. 2019). More recently, deep GP models have been introduced, which can remove these limitations (Cutajar et al. 2019). Deep GPs employ a layered GP architecture trained in a coupled manner (e.g., $f_1^{1,2,3}$ in Figure 1, trained on inputs from fidelities 1, 2, and 3), enabling them to capture nonlinear correlations across fidelities without the disjointed structure of nonlinear autoregressive GPs (Figure 1). The drawback, however, is that deep GPs require higher computational cost and are more challenging to train, because all fidelity layers must be optimized simultaneously through an iterative process called variational inference, which is significantly more complex than the closed-form exact solutions used in AR1 and NARGP models (Bone et al. 2025; Brevault et al. 2020).

Furthermore, the number of parameters to be optimized increases from a few dozen in AR1 or NARGP models to several hundred or even thousands in Deep GPs, further adding to the computational burden (Brevault et al. 2020). This higher parameter count also diminishes the interpretability of Deep GPs (Augusto De Souza et al. 2023).

Deep (RMSEs = 0.118–0.136) and nonlinear autoregressive (RMSEs = 0.113–0.133) GPs performed similarly well in predicting final mAb titers (at Day 14) in CHO cell culture, both outperforming a GP model trained only on HF data (RMSEs = 0.151) (Heidari et al. 2026). The models were trained on 65 HF datapoints from limited laboratory bioreactor runs and 20,000 LF datapoints, selected from a public in silico repository of fed-batch CHO simulations (Golzarjalal, Aickelin, et al. 2025). Also, different batches of LF data were selected based on their similarity to HF data, showing that higher similarity (or correlation) between LF and HF data sets improved predictive accuracy of the models.

Multi-fidelity GP models can assist process optimization. A study used an autoregressive model as the surrogate function of the Bayesian optimization method in real-time batch-to-batch optimization of a photobioreactor (Petsagkourakis et al. 2021). The study aimed at maximizing phycocyanin production after 240 h by adjusting the light intensity and nitrate inflow rate. Thirty LF datapoints were obtained from a simplified model, which ignored light inhibition, to tune the GP hyperparameters. These were combined with 13 HF datapoints, obtained from a dynamic model based on Monod kinetics, to train the initial multi-fidelity GPs. The proposed optimization framework explored through an adaptive trust region, which expanded when predictions from autoregressive surrogate model matched those of HF GP model (which simulated true plant) and shrunk when autoregressive model predictions were poorer. Moreover, the study also considered some physical constraints (on product-to-biomass ratio, nitrate concentration, light intensity, and nitrate inflow) to retain only feasible points. This adaptive framework converged faster and with fewer constraint violations compared to a method without a trust region.

In another study (Cosenza et al. 2022), it was shown that multi-information-source Bayesian optimization could efficiently optimize a 14-component culture medium for Murine C2C12 cells by fusing two quick and simpler but less valuable assays (i.e., AlamarBlue and LIVE stain) with more costly but reliable measurements (multi-passage trypan blue cell counts). The method achieved media formulations that supported 181% higher cell numbers than a standard commercial medium, using 38% fewer experiments than a Box–Behnken design-of-experiments workflow.

Another study applied multi-fidelity Bayesian optimization to identify new histone deacetylase inhibitors (McDonald et al. 2025). A GP surrogate model using Morgan fingerprints and a Tanimoto kernel was employed, as this configuration outperformed RF and natural gradient boosting models. An iterative strategy was implemented, where initially an ML algorithm docked more than 3500 candidate molecules (LF). Then, 120 of these candidates were automatically synthesized and screened to obtain percent inhibition measurements, which are less costly and moderately correlated to potency (medium-fidelity). Finally, five of the candidates were selected for manual

TABLE 2 | Summary of multi-fidelity models applied in biopharmaceutical development and manufacturing life cycle, including their specific applications, LF and HF data types, and sample sizes, and the availability of associated code or data sets.

Model type	Application	LF and HF samples	Code/Data availability	Reference
Nonlinear autoregressive GP and deep GP	Predicting mAb production in fed-batch CHO Cell culture	LF: 20,000 in silico fed-batch 250-mL bioreactor runs HF: Experimental data from 65 fed-batch 250-mL bioreactor runs	Code NA ^a In silico data from (Golzarjalal, Aickelin, et al. 2025)	Heidari et al. (2026)
Bayesian optimization using a linear autoregressive GP as the surrogate model	Batch-to-batch optimization of a photobioreactor	LF: 30 datapoints from a simplified Monod kinetics, ignoring the light inhibition HF: 13 datapoints from Monod kinetics	NA	Petsagkourakis et al. (2021)
Multi-information-source Bayesian optimization	Optimization of culture media for cellular agriculture (Murine C2C12 cells)	LF: measurement from AlamarBlue and LIVE stain HF: multi-passage trypan blue cell counts	Code and data available at https://github.com/ZacharyCosenza/GradStuff_Cosenza.git	Cosenza et al. (2022)
Multi-fidelity Bayesian optimization using a GP surrogate model with Morgan fingerprints and a Tanimoto kernel	Molecular property optimization; Integration of multi-fidelity Bayesian optimization with an autonomous lab to search for new histone deacetylase inhibitors	LF: Docking scores for 3500 candidate molecules found from DiffDock simulation Medium-fidelity: Percent inhibition for 120 molecules HF: Dose-response curves for 5 molecules	Code and data available at https://zenodo.org/records/13983042	McDonald et al. (2025)
Multi-fidelity Bayesian optimization using a GP (from the class of linear models of coregonalization kernels) as the surrogate model	Originally for syngas fermentation, but relevant to optimizing the operating conditions of bioreactors	LF: ideal-mixing model (runtime < 2 min) HF: computational fluid dynamics model (runtime ~5 days)	NA	Eskandari et al. (2023)
Deep feed-forward NN surrogate model implemented in an active learning algorithm	Optimizing binding affinities for generative drug discovery	LF: total binding energy prediction plus 15 other interaction descriptors for 100,000 compounds from AutoDock4 Medium-fidelity: 91–102 laboratory-measured affinity values HF: A single binding energy prediction from molecular dynamics simulations	NA	Eckmann et al. (2024)

(Continues)

TABLE 2 | (Continued)

Model type	Application	LF and HF samples	Code/Data availability	Reference
(1) An all-in-one framework to train a random RF, SVM, and Graph NN. (2) a multilevel framework by first training a Graph NN on LF data, then adding graph embeddings of LF data to HF data. Next, HF data were used to train RF, SVM, and Graph NN models.	Predicting confirmatory-level (dose-response) compound activity at high-throughput screening during early-stage drug discovery	LF: bioaffinity measurements for hundreds of thousands of compounds at one concentration (single dose) HF: bioaffinity measurements for a few thousand compounds at multiple concentrations (dose response)	The code to assemble MF-PCBA data sets is available at https://github.com/davidbuterez/mf-pcba	Buterez et al. (2023)
(1) Graph NNs with sum and neural graph readout functions to build LF models in a multilevel framework. (2) Graph NN with sum readout functions in an all-in-one framework	Predicting molecular properties for drug discovery tasks	LF: bioaffinity measurements for hundreds of thousands of compounds at one concentration (single dose) HF: bioaffinity measurements for a few thousand compounds at multiple concentrations (dose response)	Code available at https://github.com/davidbuterez/multi-fidelity-gnns-for-drug-discovery-and-quantum-mechanics Data (except for industrial proprietary data) available in supplementary materials and prior works.	Buterez et al. (2024)
Multi-fidelity trust-region optimization	Maximizing the feed throughput of a moving bed chromatography	LF: using a coarser spatial discretization for solving first principles PDEs LF: obtained from reduced-order modeling based on Proper Orthogonal Decomposition HF: built from solving first principles PDEs using the finite element method	NA	Li et al. (2014)

^aNot available.

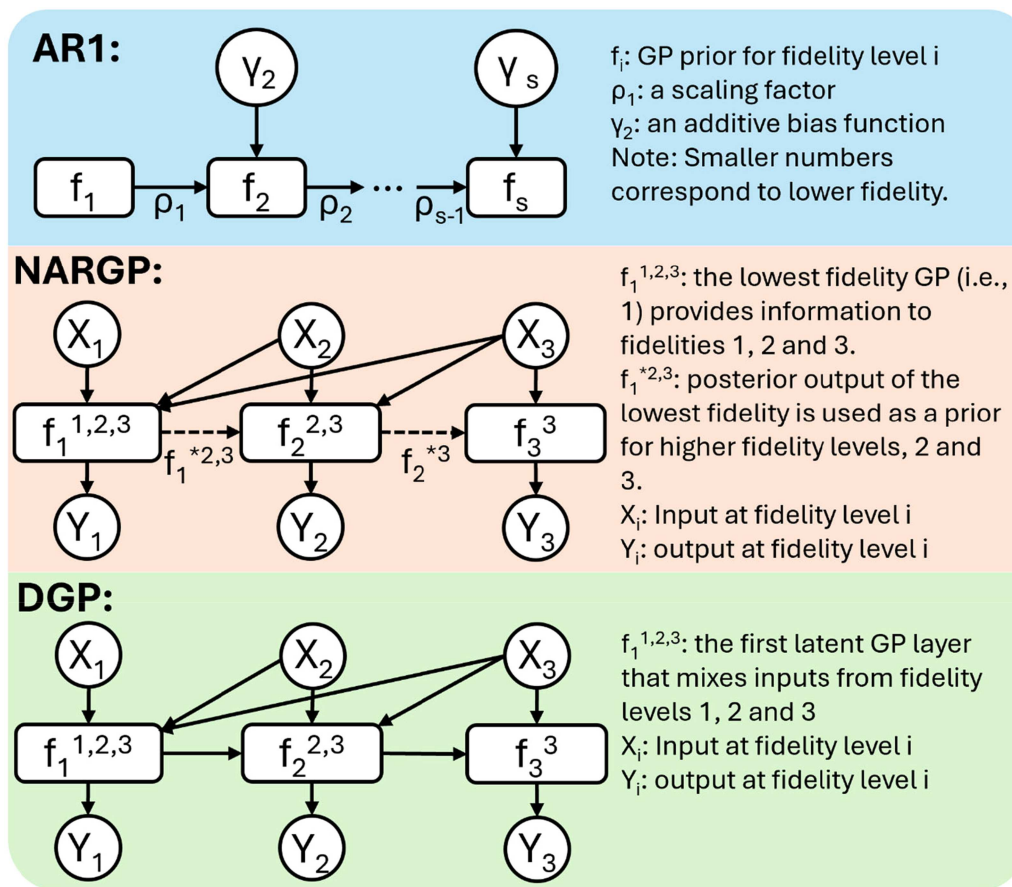


FIGURE 1 | Schematic views for Gaussian process-based surrogate models for a multi-fidelity model with 3 (NARGP and DGP) or more (AR1) fidelity levels. AR1, NARGP, and DGP represent the linear autoregressive model, the nonlinear autoregressive Gaussian process model, and the deep Gaussian process, respectively.

evaluation using dose-response IC_{50} assays (HF), which represent the ground truth but at a higher experimental cost. This multi-fidelity approach outperformed other methods, namely experimental funnels, transfer learning, and Bayesian optimization, solely with HF data, especially when search spaces were diverse. When there was a strong correlation between fidelities, experimental funnel approaches performed best. For weak or no correlations, Bayesian optimization using only HF data was most efficient.

Eskandari et al. (2023) applied multi-fidelity Bayesian optimization to maximize syngas conversion rate in an industrial-scale bioreactor, by combining a fast, LF ideal-mixing model (runtime < 2 min) with an HF computational fluid dynamics (CFD) model (runtime ~5 days). As the surrogate model, they implemented a GP from the class of linear models of coregionalization kernels, tailored specifically for multi-information source optimization problems. Their approach adaptively balanced accuracy and computational cost by deciding whether each new sample should come from the LF or the HF model. This optimization technique accelerated the search for optimal reactor conditions, that is, inlet gas flow rate, pressure, and biomass concentration, reducing reliance on computationally expensive CFD runs while still converging to conditions that maximized gas-to-ethanol conversion. While this example is not specific to biopharmaceutical manufacturing, the underlying principles are highly applicable. For instance, in predicting the oxygen mass transfer coefficient

($k_L a$), an LF approach could employ an empirical model (Xing et al. 2009), whereas an HF approach could utilize complex CFD models that account for hydrodynamics and mass transfer variables (Wutz et al. 2018).

3.2 | NNs

Recently, the ability of NNs in efficiently handling a large number of features as well as better generalizability for more complex nonlinear input spaces has brought them into attention as the surrogate function for multi-fidelity models (Ciarlatani and Gorlé 2025). Deep NNs, specifically feed-forward and Graph NNs, have shown promise in multi-fidelity modeling and optimization of drug discovery studies. One architecture is to use both LF and HF at the same time in training of a single NN (i.e., all-in-one NN; Figure 2) while another approach uses a separate network for each fidelity (multilevel NN; Figure 2) (Guo et al. 2022). Notably, an overlap exists between the multilevel NN approach and transfer learning NN models regarding sequential training, initially on LF data, followed by training on HF data. However, it has been argued that the second step of multilevel NN methods requires an entirely different network, thereby increasing the total number of parameters that should be trained, whereas transfer learning fine-tunes the parameters of the same pre-trained model, requiring a smaller set of parameters (Zhang et al. 2024).

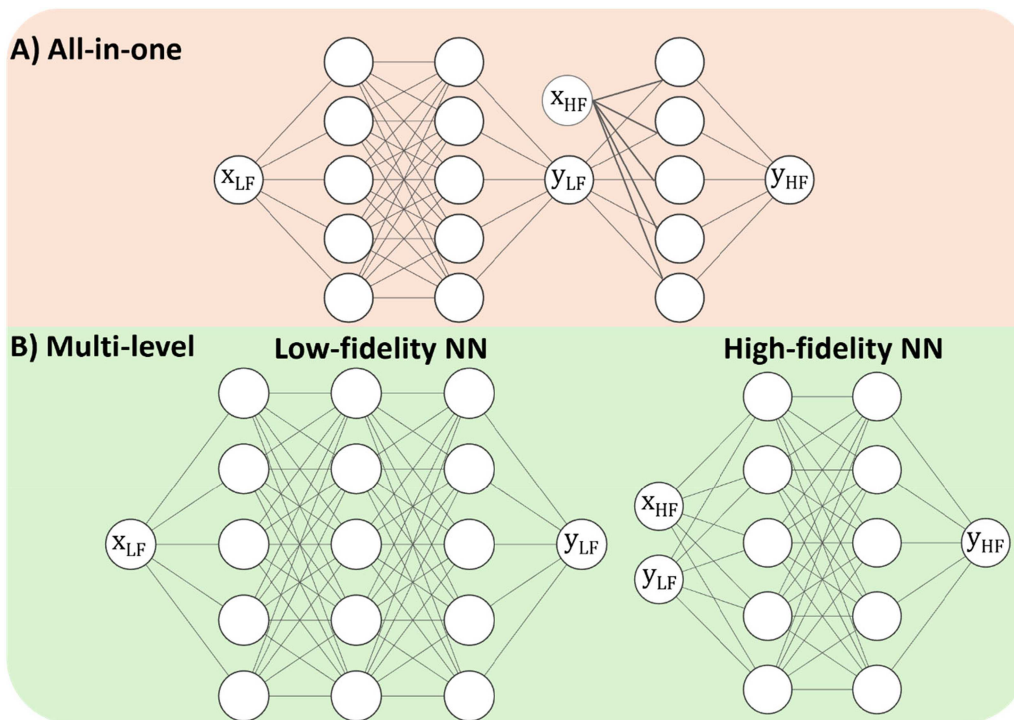


FIGURE 2 | Schematic representation of neural network architectures used in multi-fidelity modeling. (A) The all-in-one architecture jointly learns LF and HF relationships, using the LF output (y_{LF}) as an intermediate latent variable for predicting the HF output (y_{HF}). (B) In the multilevel architecture, two NNs are trained sequentially. First, an NN is trained on LF inputs (x_{LF}) to learn y_{LF} ; then, a second NN is trained using both y_{LF} and HF inputs (x_{HF}) to predict y_{HF} .

A deep feed-forward NN surrogate model implemented in an active learning algorithm could help to find optimum compound binding affinities for generative drug discovery in a cost-efficient manner. A study has shown this for two different proteins of BRD4(2) and c-MET (Eckmann et al. 2024). The authors leveraged data from three simulators for estimating binding affinities to target proteins: 100,000 from AutoDock4 (LF), 91–102 experimental data (medium-fidelity), and only 1 absolute binding free energy or ABFE (HF). A multi-fidelity model with a single deep NN for all fidelities generated more promising compounds for both proteins (i.e., with lower binding energies) across 20 generated compounds than single-fidelity models trained only on AutoDock4 or ABFE data.

Graph NNs, with a geometric deep NN architecture, have drawn attention in multi-fidelity modeling tasks in drug discovery, as they can learn from complex non-Euclidean data structures, where information is represented as nodes (objects) and edges (connections). Graph NNs were used in a study that curated 60 multi-fidelity data sets from two different steps of screening drug compounds available on PubChem BioAssay (Buterez et al. 2023). LF data were obtained from the primary screening step, providing an abundance of bioaffinity measurements (for hundreds of thousands of compounds) at one concentration (i.e., single dose), which were noisy but cheap and easily accessible. HF data came from a further confirmatory step, containing a much smaller subset of compounds (a few thousand compounds), which were tested at multiple concentrations (i.e., dose response), providing an accurate measurement of the compound's strength, making them slow and costly. An all-in-one multi-fidelity framework, implemented as direct label augmentation, combined both LF and HF data to train an

RF, a support vector machine (SVM), and a graph NN model to predict dose response data (HF). Additionally, a secondary multilevel framework was proposed, which utilized graph embeddings from a Graph NN pre-trained on LF data as additional features to train RF, SVM, and Graph NN models on HF data. Generally, both multi-fidelity frameworks outperformed the baselines, that is, RF, SVM, and Graph NN models trained solely on HF data, with the all-in-one approach showing superior performance, improving the R^2 of the baselines by 33%–40% on average (Buterez et al. 2023).

A further drug discovery study, employing the same types of LF and HF data sets as previously described (Buterez et al. 2023), predicted molecular properties by employing Graph NNs with both sum and neural readout functions within two frameworks: (1) a multilevel multi-fidelity framework that trained each fidelity independently and incorporated LF outputs as features for HF inputs, and (2) a transfer learning approach involving pre-training on LF data followed by fine-tuning on HF data (Buterez et al. 2024). Additionally, a third all-in-one multi-fidelity framework was also evaluated, which combined LF and HF data to train a Graph NN with sum readouts. It was shown that compared to the baseline models trained only on HF data, building LF models with sum readouts generally did not improve performance, while using neural readouts for LF models with both multilevel and transfer learning methods consistently enhanced baseline predictions, as evident by 20%–40% decrease in mean absolute error. While the all-in-one framework also improved baseline predictions by 10–40%, the transfer learning method with neural readouts generally outperformed both multi-fidelity frameworks. The superior performance of neural readouts was because they generated latent

spaces where active and inactive compounds were clearly separated, whereas sum readouts typically produced scattered, less informative embeddings (Buterez et al. 2024).

While the multi-fidelity frameworks proposed by Buterez et al. (2023, 2024) were originally developed for small-molecule discovery, the underlying methodological principles are highly relevant to large-molecule bioprocess development. In both domains, a major bottleneck for training robust ML models is the lack of large HF experimental data sets, which can be addressed by leveraging inexpensive LF data.

3.3 | Physics-Based Models

Surrogate models for multi-fidelity frameworks could also be built from physics-based models rather than GP or NN models mentioned above. Li et al. (2014) optimized moving-bed chromatography using simpler physics-driven models. They built a HF model by solving first principles partial differential equations (PDEs) using the finite element method, which yielded accurate predictions but was computationally expensive. Two different, simpler LF surrogate models were built, one using a coarser spatial discretization for solving PDEs, and another from reduced-order modeling based on Proper Orthogonal Decomposition. The surrogate models were separately embedded in a multi-fidelity trust-region optimization, in which LF predictions guided the search and HF evaluations guaranteed convergence to the true optimum of the original HF problem. This strategy restricted the optimization within a local

trust region, the radius of which shrank or expanded based on the closeness of predicted improvements from LF models to actual improvements observed in HF models. Considering the problem of maximizing the feed throughput while maintaining the product purity, both multi-fidelity optimizations reached the same optimum as an optimization led only by the HF model. However, the former methods significantly reduced computation time, with coarser and reduced order surrogates taking ~64 and ~55 min, respectively, while HF optimization took ~123 min (Li et al. 2014).

3.4 | Comparative Analysis of Surrogate Models

Both GPs and NNs could be used as multi-fidelity surrogate functions, and the choice between them depends on the specific application. GPs provide a probabilistic framework that naturally quantifies prediction uncertainty, making them particularly suitable for Bayesian optimization, where confidence intervals guide adaptive sampling. GPs work well with limited data sets and low-dimensional problems. NNs, on the other hand, are better suited for high-dimensional systems, but they lack inherent uncertainty quantification and, compared to GPs, need larger amounts of HF data to achieve accurate predictions (Ciarlatani and Gorlé 2025). In cases with the scarcity of HF data, all-in-one and multilevel multi-fidelity frameworks, such as those proposed by Buterez et al. (2023), can enable knowledge transfer from abundant LF data to limited HF data, improving predictive accuracy of NN surrogate models. Table 3

TABLE 3 | Key strengths and primary limitations of different surrogate models in multi-fidelity bioprocessing.

Model type	Key strengths	Primary limitations	Reference
Gaussian processes (GPs)	- Uncertainty quantification - Work well with limited data sets	Struggle with large number of features and datapoints	Augusto De Souza et al. (2023), Brevault et al. (2020), Ciarlatani and Gorlé (2025)
Auto-regressive GPs	- Easy to implement - Higher interpretability than deep models	- Limited to linear correlations between LF and HF data - Require nested input structure^a - Have a disjointed architecture	
Nonlinear auto-regressive GPs	- Capture nonlinear correlations between LF and HF data - Higher interpretability than deep models	- Require nested input structure - Have a disjointed architecture	
Deep GPs	- Capture nonlinear correlations with a layered architecture - Low interpretability	High computational cost due to variational inference approximations and many hyperparameters	Ciarlatani and Gorlé (2025), Tazi et al. (2023), Zhou et al. (2020)
Deep neural networks (NNs)	- Can handle high dimensionality - Automated feature extraction - Low interpretability - Graph NNs are excellent for complex non-Euclidean data structures, such as drug discovery	- Lack inherent uncertainty quantification - Require large amounts of HF data	
Physics-based models	Can generate mechanistically consistent and physically interpretable results	Require substantial domain knowledge	

^aTraining data of lower fidelities must include data from higher fidelities.

summarizes the strengths and limitations of different surrogate models applied in multi-fidelity bioprocessing.

Model interpretability is a critical factor in selecting a surrogate model, as it enables the identification of cause-and-effect relationships between inputs and outputs (Love et al. 2023). Shallow GPs, such as AR1 and NARGP, provide higher interpretability as they are trained by a small number of parameters (e.g., length scales of variations) that directly capture the underlying characteristics of the data (Tazi et al. 2023). In contrast, deep GPs and NNs are less interpretable, as they involve a much larger number of parameters that often lack a direct physical meaning (Augusto De Souza et al. 2023; Tazi et al. 2023).

Mechanistic models (or governing equations) of a process generally offer higher interpretability compared to GP and NN-based surrogates, providing mechanically consistent and physically meaningful results. These models can be used as either HF or LF models within a trust-region optimization framework, depending on their predictive accuracy and computational cost. However, their development typically requires substantial domain expertise. An example of such frameworks is the chromatography optimization, where LF models were derived by simplifying a detailed full-order HF model (Li et al. 2014).

One way to increase the interpretability of deep NNs is to integrate physical laws, typically in the form of PDEs, into their loss functions. This approach defines a class of models often referred to as physics-informed neural networks (PINNs) (Raissi et al. 2019). Regazzoni et al. (2021) proposed a multi-fidelity framework with an NN serving as an LF model, trained on extensive data sets (100–10,000 datapoints) derived from numerical approximations. The HF model was a PINN trained on limited, potentially noisy measurements (2–32 datapoints) sampled from the exact analytical solutions of two benchmark problems, that is, steady-state and time-dependent advection-diffusion-reaction PDEs. The HF model refined the initial guess provided by the LF model while incorporating PDE residual into its loss function, penalizing the network when it violated the underlying physics. Results showed that the proposed multi-fidelity PINN framework provided faster convergence and more accurate predictions compared to a standard PINN method trained solely on HF data. Notably, the framework was significantly more robust to noise in HF data, achieving error reductions up to two orders of magnitude lower than standard PINNs (Regazzoni et al. 2021). This robustness is particularly relevant to the types of measurements observed in bioprocessing (Khuat et al. 2024), suggesting that implementing PINN surrogate models can enhance model interpretability and, consequently, generalizability (Barón Díaz et al. 2026).

Another approach to improving model interpretability is the application of explainable AI (XAI) techniques, particularly post hoc methods, that are applied after a model has been trained (Miró-Nicolau et al. 2024). XAI aims to provide clear reasoning for a model's functioning, ensuring that results are understandable by end-users (Love et al. 2023). These methods increase the transparency of complex ML models, which is vital in a multi-fidelity framework to trace the specific contributions of both LF and HF data.

Among various post hoc XAI methods, SHapley Additive ex-Planation (SHAP) is a prominent method that quantifies interpretability by attributing the marginal impact of input features to

every individual prediction (Lundberg and Lee 2017). Zhang et al. (2025) implemented a deep SHAP methodology within a multi-fidelity Bayesian NN framework. While the study was conducted in the electrical engineering domain, specifically focusing on distribution system state estimation, the underlying approach to data fusion and interpretability can be highly relevant to bioprocessing. In this study, a deep NN was trained on abundant LF pseudo-measurements. Subsequently, a Bayesian NN layer was built utilizing limited, real-time HF measurements, integrated with outputs of the LF layer. The proposed framework outperformed single-fidelity and multi-fidelity baselines. Crucially, the application of deep SHAP rendered the individual impacts of LF and HF data traceable. For example, at specific nodes, it was observed that local HF data provided a dominant positive contribution to the final prediction. Furthermore, neighboring HF data contributed positively, whereas neighboring LF pseudo-measurements had a negative influence on the model's predictions (Zhang et al. 2025). These results align with physical principles, as HF data is inherently accurate while LF data tends to be noisier and susceptible to systematic bias. Such XAI methods can make multi-fidelity models more transparent and interpretable to humans.

When selecting a surrogate model for optimization, it is important to recognize the trade-off between model accuracy and computational performance. A highly accurate model may guarantee convergence to the true HF optimum, but can become impractical due to higher computational cost. Conversely, a model with a poor performance may lead to suboptimal convergence or even failure to identify the true optimum. Therefore, it is recommended to perform a sensitivity analysis to determine the simplest model that remains sufficiently close to the true optimum, without compromising the optimization performance. For instance, in the case of chromatography optimization, a sensitivity analysis identified the most suitable LF model as one using two finite elements and two collocation points per column (Li et al. 2014). Using a more accurate LF model did not further improve the optimal solution but increased computational time by 23%. On the other hand, using an LF model with a coarser mesh reduced computational time by 5% but failed to converge to the true optimal point.

4 | Practical Considerations for Selecting HF and LF Data Sources

HF data in the biopharmaceutical industry typically includes online and offline measurements from wet-lab experiments (Figure 3). These data directly reflect on success rates in drug discovery, as well as CQAs and key performance indicators throughout the processing stages, as they represent the most reliable, and often the most costly, sources of information. Examples include experimentally measured variables across different stages, such as drug discovery (e.g., binding affinity), cell culture (e.g., cell density and protein concentration), downstream purification (e.g., protein yield, purity, and aggregate levels), and formulation (e.g., drug stability and bioavailability).

Time and costs of experimentation are the primary factors limiting the availability of experimental HF data in biopharmaceutical development. Specifically, offline measurements are more limited as they are time-consuming and incur high costs for skilled labor and consumables (Tulsyan et al. 2020).

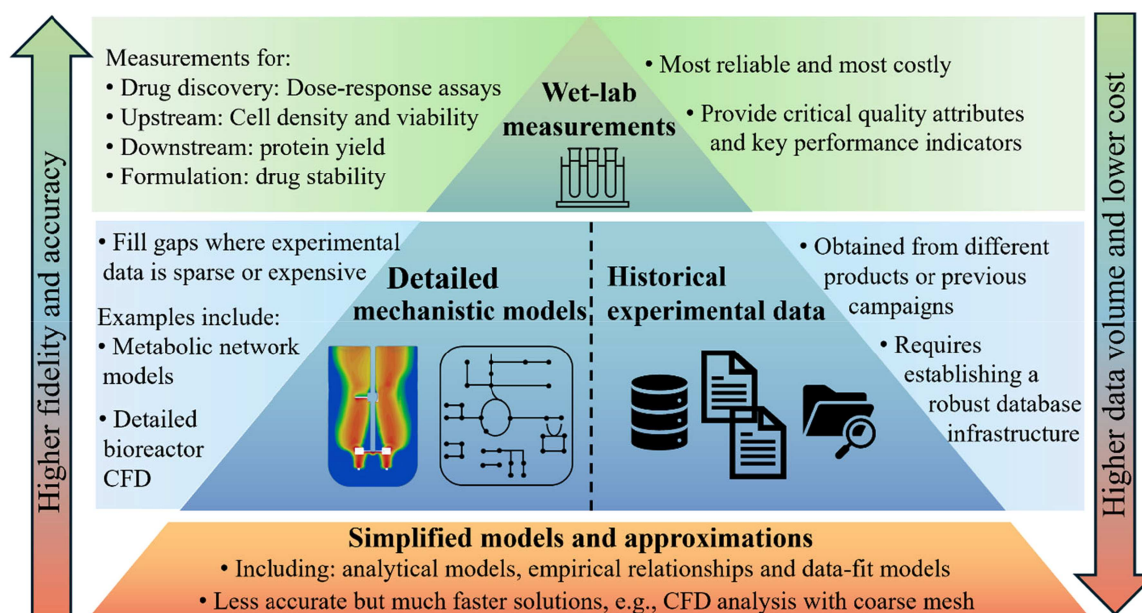


FIGURE 3 | Overview of high-fidelity (HF) and low-fidelity (LF) data sources in the biopharmaceutical industry, illustrating the accuracy, data volume, and acquisition cost. The bioreactor CFD image is reprinted from Kiviat et al. (2025) under the CC BY 4.0 license.

Conversely, data collected via online PAT tools can be more abundant due to higher acquisition frequencies, with upstream processing adopting these tools more broadly than other stages (Dürauer et al. 2024; Khuat et al. 2024). Downstream processing, in contrast, has fewer PAT implementations due to factors such as requiring significantly faster response times compared to upstream, and therefore often lacks comparable real-time data (Dürauer et al. 2024; Gerzon et al. 2022; Rathore and Kapoor 2015; Sathiyapriyan et al. 2025). Furthermore, experimental HF data related to processing are rarely publicly accessible, as companies typically maintain this information as proprietary (Khuat et al. 2024). Rare examples of open data include a 7-year anonymized data set for upstream CHO cell line development (Gangadharan et al. 2021), and a data set in the formulation stage, which includes experimentally measured aggregation rates for 20 preclinical and clinical-stage mAbs, as well as viscosities for a total of 47 antibodies, comprising the 20 initial mAbs plus an additional 27 commercial mAbs. (Lai et al. 2022). This scarcity in processing data contrasts with drug discovery, where several open databases provide biophysical, structural, and sequence data for proteins and antibodies. For instance, the Therapeutic Structural Antibody Database offers biophysical properties of approved mAbs (Raybould et al. 2020). A comprehensive list of such discovery-focused databases can be found elsewhere (Khuat et al. 2024).

LF data in the biopharmaceutical industry can be obtained from various sources, depending on the availability of corresponding HF data. One valuable source is historical experimental data, either from a different product or from previous campaigns that have become outdated due to minor changes in the production process, such as variations in cell culture conditions or bioreactor configurations (Figure 3). Therefore, establishing a robust database infrastructure to organize and preserve such historical data is essential for enabling biopharmaceutical companies to effectively leverage these resources in future model development and optimization efforts.

However, in practical applications, significant process drift (or batch-to-batch variation) often occurs, and an overreliance on historical data risks degrading model performance. For instance, a current cell culture may utilize growth media that have aged or accumulated degradants and impurities compared to the fresh lots used in historical campaigns (Rathore et al. 2018). Such discrepancies in raw material quality can lead to substantial variations in product yield. In such cases, historical data may provide misleading trends for optimizing feed media and feeding strategies. To mitigate this risk, historical data sets should be reviewed carefully for each project and down-weighted to reduce their influence on the model predictions, depending on their relevance to the current process.

Another important source of LF data is simulated data generated from advanced mechanistic models, which can help fill gaps in HF data sets where experimental measurements are expensive or sampling either over time or across the experimental space is limited (Figure 3). Such models include CFD for both upstream and downstream processes, genome-scale metabolic network models for cell culture, transport-dispersion models integrated with Langmuir-based binding models for chromatography, Arrhenius-based kinetic models for formulation, and molecular docking simulations for drug discovery studies. For instance, Heidari et al. (2026) used 20,000 LF samples from a genome-scale metabolic model to supplement 65 HF (cell culture) experimental data.

When no experimental data are available, HF data can be obtained from a detailed mechanistic model, and simpler models can be used as LF sources (Figure 3). These simpler models include analytical models, empirical relationships, and data-fit models (Peherstorfer et al. 2018). Although these models typically provide less accurate or lower-quality results, they can solve problems much faster than detailed simulations (e.g., CFD). Moreover, simplified mechanistic models, for example, CFD simulations with coarser meshes, can likewise serve as LF data sources.

While the potential of simulated data as both LF and HF sources is becoming increasingly evident in the biopharmaceutical domain, the legal and ethical implications of its use must be carefully considered (Nisevic et al. 2025). This is particularly critical in clinical trials involving simulated patient profiling, where data sensitivity is high. These concerns are reflected in EU legal frameworks such as the General Data Protection Regulation, which determines obligations based on the contextual risk of individual identifiability rather than the technical method used to generate the data. In this context, the legal and ethical standing of simulated data depends largely on its generation and application, requiring a balance between privacy and utility (El Emam 2020). For example, simulated data that closely resembles real data may be more useful for analytics but, at the same time, might increase the risk of indirectly revealing information about real individuals. Therefore, ongoing ethical oversight is required to ensure that simulated data complies with data protection principles within national and international regulatory frameworks (Nisevic et al. 2025).

Multi-fidelity methods can incorporate multiple LF sources simultaneously. This approach has previously been adopted in predicting binding affinities during drug discovery (Table 2) (Eckmann et al. 2024). For example, when data are available from laboratory experiments (HF), detailed simulations (medium-fidelity), and data-fit models (LF), they can be integrated by assigning appropriate weights to each fidelity level based on its accuracy and computational cost. Furthermore, it is critical to assess the correlation between LF and HF data to evaluate how well LF data represents HF data. If the correlation is weak, multi-fidelity models risk learning misleading patterns, which can lead to erroneous predictions (Andrés-Thió et al. 2024; Perdikaris et al. 2017).

For multi-fidelity models to reach industrial maturity within an automated laboratory, a core pillar of Pharma 4.0 (Phiri et al. 2025), they must be developed according to Good Automated Manufacturing Practice (GAMP 5) guidelines. This ensures a safe, compliant design and implementation of automated systems (Pedro et al. 2023). For example, McDonald et al. (2025) performed drug discovery in an autonomous platform using multi-fidelity Bayesian optimization. Within this framework, they implemented a tiered validation strategy: initial performance was verified as the algorithm iteratively selected candidates from docking simulations (LF) for physical synthesis and single-point inhibition measurements (medium-fidelity). A subsequent validation step was achieved when these candidates exhibited high potency in dose-response IC_{50} assays (HF). By systematically verifying model logic and validating predictions against physical experimental benchmarks, this approach could identify reliable, high-potency compounds and demonstrate the potential for alignment with the stringent quality and safety principles of GAMP 5 (Pedro et al. 2023). However, further research is required to fully integrate these multi-fidelity techniques with comprehensive GAMP 5 life cycle documentation and FDA safety approval guidelines.

5 | Conclusion and Outlook

There is currently a growing interest in multi-fidelity techniques; however, their full potential has yet to be realized in the biomanufacturing field. Existing studies have primarily focused

on upstream processing and drug discovery. Future work could investigate other process stages, particularly downstream processing. The formulation step has also faced data scarcity challenges when building ML models. To address this, Lai et al. (2022) predicted therapeutic antibody aggregation rates and viscosity at high concentrations by combining data from 27 commercial mAbs with 20 preclinical and clinical-stage mAbs to train conventional ML models, such as linear regression, support vector regression, and k-nearest neighbors. However, the integration of data within a multi-fidelity framework remains unexplored for formulation. In this context, LF data could be sourced from existing literature containing incomplete sequence information or different solution conditions (e.g., pH and protein concentrations) (Lai et al. 2022), viscosity fitting models (Armstrong et al. 2024), and molecular simulation approaches (Cruz et al. 2025). These LF sources could then supplement HF data sets, for example, expensive experimental data sets for high-concentration antibodies (Lai et al. 2022), within a multi-fidelity framework.

Another aspect that can be further explored is the selection of surrogate models. While many existing frameworks rely on GPs or Graph NNs, there is still potential in further exploring deep learning structures such as convolutional and recurrent NNs. Additionally, the adoption of more advanced surrogate models, including Transformer and diffusion architectures, could be further explored.

Convolutional NNs (CNNs) represent a class of deep NN architectures that have shown great potential in solving complex engineering problems. A multi-fidelity CNN integrates LF and HF data by treating multi-fidelity information similarly to image data, utilizing local receptive fields and convolution to capture implicit relationships between any HF datum and all available LF data (Chen et al. 2022). Despite their potential, CNNs have not been widely used in building multi-fidelity techniques in bioprocessing. However, they have been used in transfer learning studies. For example, Zhang et al. (2023) pre-trained an ensemble of kinetically interpretable NNs on in vitro kinetic data containing 14,000 targets, which were then integrated as intermediary layers into a deeper CNN backbone and fine-tuned on complex in vivo data to predict off-target gene-editing probabilities. This transfer learning method significantly improved the performance of optimized CNN baselines by 150%–440%. CNNs as surrogates for multi-fidelity models have also been applied in broader engineering applications. For instance, a multi-fidelity data aggregation model using CNN effectively combined 255 LF datapoints from a coarse mesh (50×50) with 12 HF datapoints from a fine mesh (300×300) to perform stress analysis on random microstructures (Chen et al. 2022). This model achieved an RMSE of 1.2%, significantly outperforming a single-fidelity model trained only on HF data, which yielded an RMSE of 38.9%. Future research could implement similar transfer learning or multi-fidelity architectures in various stages of bioprocessing. For example, in modeling Raman spectral data in cell culture (Min et al. 2023) or classification of subvisible particles in protein formulation (Calderon et al. 2018), areas where single-fidelity CNNs have already proven their utility.

Recurrent NNs (RNNs) are another class of deep NNs particularly effective for processing sequential information, such as time-series data (Lu et al. 2025). Long short-term memory

(LSTM) networks are a type of RNN capable of retaining information over long periods, which makes them suitable for bioprocess dynamic modeling (Ramos et al. 2024). Multi-fidelity models incorporating LSTM surrogates can relate LF and HF data by capturing complex temporal patterns and parametric dependencies (Conti et al. 2023). For instance, three multi-fidelity LSTM paradigms, with distinct architectures and training sequences, were used to estimate the temporal evolution of drag and lift coefficients in fluid flow past a cylinder. The multi-fidelity LSTM models, using 19 LF datapoints and 6-10 HF datapoints, significantly outperformed a single-fidelity NN trained solely on HF data, as well as a multi-fidelity model with a feed-forward NN surrogate model (Conti et al. 2023). In another study on wastewater treatment plants, Koksall and Aydin (2025) showed that LSTMs within a transfer learning framework can leverage LF data even when the source data set from a related plant (600 datapoints) is more limited than the HF data from a target plant (1100 datapoints). Specifically, the proposed transfer learning LSTM approach, involving pre-training on source data followed by fine-tuning on target data, improved the prediction of dissolved oxygen concentration by 47% compared to a standard LSTM trained only on target data during a 200-day validation period (Koksall and Aydin 2025). Despite their potential, the application of multi-fidelity LSTM surrogate models in bioprocessing remains largely unexplored, representing an opportunity for future investigation.

Transformer models, a type of NN architecture, have recently been applied as surrogate models in multi-fidelity problems. For instance, an encoder-decoder Transformer was applied to time-series data of a 2D channel flow around a cylinder, where each fidelity was separately encoded into a compressed hidden form (Da Mommio et al. 2025). Transformer layers then combined these representations into a single universal embedding, which a simple decoder used to predict the velocity field. A key advantage of this method is the combinatorial masking strategy, which makes the model robust to missing inputs and flexible to any combination of fidelity data. This can be applied in the biopharmaceutical context to integrate heterogeneous data sets, such as mechanistic simulations with experimental data from multiple scales (e.g., 250 mL, 5 L, and 200 L), to build robust digital twins of cell culture processes, enabling accurate predictions of product yield even when some data sources are unavailable or only partially available. Despite their advantages, Transformers require high computational costs, which can make it difficult to train a robust model. To mitigate this challenge, researchers have proposed various techniques, including sparse, deformable, and locality-aware attention mechanisms (Khoshnevis and Amirkhani 2026).

Diffusion models, a class of NNs best known for generating HF images and videos, have recently been used in multi-fidelity frameworks. Diffusion-based surrogate models learn physics-based information from both LF and HF data by starting from random noise and progressively denoising it to generate realistic samples that reflect the system's physical behavior. For instance, Shi et al. (2025) trained a denoising diffusion network using paired HF (128 × 128 grid) and LF (32 × 32 grid) CFD simulations of buoyancy-driven fume motion. The diffusion-based surrogate outperformed a classical GP surrogate model, which struggles with high-dimensional inputs (e.g., 128 × 128 = 16k variables), whereas the neural architecture of diffusion models could capture complex

spatial dependencies. In a biopharmaceutical context, for example, a similar approach could be applied by leveraging coarser LF chromatography simulations that capture only linear isotherms to predict outcomes of fine-grid HF models that also incorporate mass transfer resistance, axial dispersion, and competitive Langmuir isotherms.

Finally, in the age of artificial intelligence, where the importance of data is becoming increasingly evident, it is essential to understand available data resources and how to effectively leverage them for process optimization. Multi-fidelity techniques offer promise in the biopharmaceutical industry by leveraging cheap LF data to predict expensive HF results. These techniques can help identify optimal process conditions with fewer HF evaluations, thereby reducing development costs and timelines. By doing so, these methods can accelerate the delivery of life-saving medicines to society and enable yield improvements or process adjustments that significantly enhance manufacturing efficiency and cost-effectiveness.

Author Contributions

Mohammad Golzarjalal: conceptualization, writing – original draft, conducting literature search, visualization. **Uwe Aickelin:** conceptualization, writing – review and editing, funding acquisition, supervision. **Ellen Otte:** conceptualization, writing – review and editing.

Acknowledgments

This research was supported under the Australian Research Council's Industrial Transformation Research Program (ITRP) funding scheme (project number IH210100051r). The ARC Digital Bioprocess Development Hub is a collaboration between The University of Melbourne, University of Technology Sydney, RMIT University, CSL Innovation Pty Ltd, Cytiva (Global Life Science Solutions Australia Pty Ltd) and Patheon Biologics Australia Pty Ltd. Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australasian University Librarians.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no data sets were generated or analyzed during the current study.

References

- Andrés-Thió, N., M. A. Muñoz, and K. Smith-Miles. 2024. "Characterising Harmful Data Sources When Constructing Multi-Fidelity Surrogate Models." *Artificial Intelligence* 336: 104207.
- Armstrong, G. B., A. Roche, W. Lewis, and Z. Rattray. 2024. "Reconciling Predicted and Measured Viscosity Parameters in High Concentration Therapeutic Antibody Solutions." *mAbs* 16: 2438172.
- Augusto De Souza, D., A. Nikitin, and S. T. John, et al. 2023. "Thin and Deep Gaussian Processes." <https://arxiv.org/abs/2310.11527>.
- Banner, M., H. Alosert, C. Spencer, et al. 2021. "A Decade in Review: Use of Data Analytics Within the Biopharmaceutical Sector." *Current Opinion in Chemical Engineering* 34: 100758.
- Barón Díaz, D., A. Drommershausen, A. Grünberger, and D. Holtmann. 2026. "Transfer Learning Approaches in Bioprocess Engineering: Opportunities and Challenges." *Biotechnology and Bioengineering* : e70186. <https://doi.org/10.1002/bit.70186>.

- Bone, V., C. van der Heide, K. Mackle, I. Jahn, P. M. Dower, and C. Manzie. 2025. "Gradient-Enhanced Deep Gaussian Processes for Multifidelity Modeling." *Journal of Computational Physics* 520: 113474.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 2017. *Classification and Regression Trees*. Routledge.
- Brevault, L., M. Balesdent, and A. Hebbal. 2020. "Overview of Gaussian Process Based Multi-Fidelity Techniques With Variable Relationship Between Fidelities, Application to Aerospace Systems." *Aerospace Science and Technology* 107: 106339.
- Buterez, D., J. P. Janet, S. J. Kiddle, and P. Liò. 2023. "MF-PCBA: Multifidelity High-Throughput Screening Benchmarks for Drug Discovery and Machine Learning." *Journal of Chemical Information and Modeling* 63: 2667–2678. <https://doi.org/10.1021/acs.jcim.2c01569>.
- Buterez, D., J. P. Janet, S. J. Kiddle, D. Oglic, and P. Liò. 2024. "Transfer Learning With Graph Neural Networks for Improved Molecular Property Prediction in the Multi-Fidelity Setting." *Nature Communications* 15: 1517.
- Calderon, C. P., A. L. Daniels, and T. W. Randolph. 2018. "Deep Convolutional Neural Network Analysis of Flow Imaging Microscopy Data to Classify Subvisible Particles in Protein Formulations." *Journal of Pharmaceutical Sciences* 107: 999–1008.
- Carracedo-Reboredo, P., J. Liñares-Blanco, N. Rodríguez-Fernández, et al. 2021. "A Review on Machine Learning Approaches and Trends in Drug Discovery." *Computational and Structural Biotechnology Journal* 19: 4538–4558.
- Charaniya, S., W.-S. Hu, and G. Karypis. 2008. "Mining Bioprocess Data: Opportunities and Challenges." *Trends in Biotechnology* 26: 690–699.
- Chen, J., Y. Gao, and Y. Liu. 2022. "Multi-Fidelity Data Aggregation Using Convolutional Neural Networks." *Computer Methods in Applied Mechanics and Engineering* 391: 114490.
- Ciarlatani, M. F., and C. Górlé. 2025. "A Neural Network-Based Multi-Fidelity Modeling Approach for Large-Eddy Simulations With Application to Wind Loading Predictions." *Engineering Structures* 343: 120780.
- Conti, P., M. Guo, A. Manzoni, and J. S. Hesthaven. 2023. "Multi-Fidelity Surrogate Modeling Using Long Short-Term Memory Networks." *Computer Methods in Applied Mechanics and Engineering* 404: 115811.
- Cosenza, Z., R. Astudillo, P. I. Frazier, K. Baar, and D. E. Block. 2022. "Multi-Information Source Bayesian Optimization of Culture Media for Cellular Agriculture." *Biotechnology and Bioengineering* 119: 2447–2458.
- Cruz, M. A., M. Blanco, and I. Ekladios. 2025. "Mechanistic and Predictive Formulation Development for Viscosity Mitigation of High-Concentration Biotherapeutics." *mAbs* 17: 2550757.
- Cutajar, K., M. Pullin, A. Damianou, N. Lawrence, and J. González. 2019. "Deep Gaussian Processes for Multi-fidelity Modeling." <http://arxiv.org/abs/1903.07320>.
- Dash, R., and A. S. Rathore. 2021. "Freeze Thaw and Lyophilization Induced Alteration in mAb Therapeutics: Trastuzumab as a Case Study." *Journal of Pharmaceutical and Biomedical Analysis* 201: 114122.
- Davis, R. R., F. Suber, I. Heller, B. Yang, and J. Martinez. 2021. "Improving mAb Capture Productivity on Batch and Continuous Downstream Processing Using Nanofiber Prisma Adsorbents." *Journal of Biotechnology* 336: 50–55.
- Dürauer, A., A. Jungbauer, and T. Scharl. 2024. "Sensors and Chemometrics in Downstream Processing." *Biotechnology and Bioengineering* 121: 2347–2364.
- Eckmann, P., D. Wu, G. Heinzelmann, M. K. Gilson, and R. Yu. 2024. "MFBInd: A Multi-Fidelity Approach for Evaluating Drug Compounds in Practical Generative Modeling." <http://arxiv.org/abs/2402.10387>.
- El Emam, K. 2020. "Seven Ways to Evaluate the Utility of Synthetic Data." *IEEE Security & Privacy* 18: 56–59.
- Eskandari, M., L. Puiman, and J. Zeitler. 2023. "Multi-Fidelity Bayesian Optimisation of Syngas Fermentation Simulators." <http://arxiv.org/abs/2311.05776>.
- Feidl, F., S. Garbellini, and S. Vogg, et al. 2019. "A New Flow Cell and Chemometric Protocol for Implementing In-Line Raman Spectroscopy in Chromatography." *Biotechnology Progress* 35: e2847.
- Gangadharan, N., D. Sewell, R. Turner, et al. 2021. "Data Intelligence for Process Performance Prediction in Biologics Manufacturing." *Computers & Chemical Engineering* 146: 107226.
- Gargalo, C. L., S. C. de las Heras, and M. N. Jones, et al. 2020. "Towards the Development of Digital Twins for the Bio-Manufacturing Industry." *Advances in Biochemical Engineering/Biotechnology* 176: 1–34.
- Gerzon, G., Y. Sheng, and M. Kirkitadze. 2022. "Process Analytical Technologies—Advances in Bioprocess Integration and Future Perspectives." *Journal of Pharmaceutical and Biomedical Analysis* 207: 114379.
- Giselle Fernández-Godino, M. 2023. "Review of Multi-Fidelity Models." *Advances in Computational Science and Engineering* 1: 351–400.
- Golzarijalal, M., U. Aickelin, Q. C. T. Duong, E. Otte, and S. Gras. 2025. "400,000 In-Silico CHO-K1 Fed-Batch Runs Across Varying Media Concentrations, Feeding Strategies and Initial Seeding Densities [Data Set]." Figshare.
- Golzarijalal, M., L. Ong, U. Aickelin, D. J. E. Harvie, and S. L. Gras. 2025. "The Effect of Freezing and Thawing on Mozzarella Cheese: Insights From Industrial-Scale Experiments and Mathematical and Digital Analysis." *Food and Bioprocess Technology* 18: 5634–5653. <https://doi.org/10.1007/s11947-025-03787-9>.
- Le Gratiet, L., and J. Garnier. 2014. "Recursive Co-Kriging Model for Design of Computer Experiments With Multiple Levels of Fidelity." *International Journal for Uncertainty Quantification* 4: 365–386.
- Guo, M., A. Manzoni, M. Amendt, P. Conti, and J. S. Hesthaven. 2022. "Multi-Fidelity Regression Using Artificial Neural Networks: Efficient Approximation of Parameter-Dependent Output Quantities." *Computer Methods in Applied Mechanics and Engineering* 389: 114378.
- Heidari, M., L. Luo, M. Golzarjalal, E. Otte, and U. Aickelin. 2026. "Efficient Selection of Low-Fidelity Data for Multi-fidelity Surrogate Models." In *Lecture Notes in Computer Science*, edited by Y. Mei, C. Qian, Q. Bai, B. Xue and S. Khanna, vol. 16452, 475–491. Springer. https://doi.org/10.1007/978-981-95-7072-0_32.
- Hong, M. S., K. A. Severson, M. Jiang, A. E. Lu, J. C. Love, and R. D. Braatz. 2018. "Challenges and Opportunities in Biopharmaceutical Manufacturing Control." *Computers & Chemical Engineering* 110: 106–114.
- Hubert, C., P. Lebrun, S. Houari, E. Ziemons, E. Rozet, and P. Hubert. 2014. "Improvement of a Stability-Indicating Method by Quality-by-Design Versus Quality-by-Testing: A Case of a Learning Process." *Journal of Pharmaceutical and Biomedical Analysis* 88: 401–409.
- Jin, S., Y. Sun, X. Liang, et al. 2022. "Emerging New Therapeutic Antibody Derivatives for Cancer Treatment." *Signal Transduction and Targeted Therapy* 7: 39.
- Jung, K.-H. 2023. "Uncover This Tech Term: Foundation Model." *Korean Journal of Radiology* 24: 1038.
- Kennedy, M. 2000. "Predicting the Output From a Complex Computer Code When Fast Approximations Are Available." *Biometrika* 87: 1–13.
- Khoshnevis, S. A., and A. Amirkhani. 2026. "Tracking With Attention: A Review of Transformer-Based Object Tracking." *Engineering Science and Technology, an International Journal* 73: 102263.
- Khuat, T. T., R. Bassett, E. Otte, A. Grevis-James, and B. Gabrys. 2024. "Applications of Machine Learning in Antibody Discovery, Process Development, Manufacturing and Formulation: Current Trends, Challenges, and Opportunities." *Computers & Chemical Engineering* 182: 108585.

- Khuat, T. T., and B. Gabrys. 2023. "An Online Learning Algorithm for a Neuro-Fuzzy Classifier With Mixed-Attribute Data." *Applied Soft Computing* 137: 110152.
- Kiviat, K., D. E. Block, and H. Manikantan. 2025. "Simulation of Bioreactor Flow Environments for Large-Scale Animal Cell Culture Conditions Relevant to Cultivated Meat Production." *Biochemical Engineering Journal* 220: 109753.
- Koksai, E. S., and E. Aydin. 2025. "A Hybrid Approach of Transfer Learning and Physics-Informed Modelling: Improving Dissolved Oxygen Concentration Prediction in an Industrial Wastewater Treatment Plant." *Chemical Engineering Science* 304: 121088.
- Kuya, Y., K. Takeda, X. Zhang, and A. I. J. Forrester. 2011. "Multi-fidelity Surrogate Modeling of Experimental and Computational Aerodynamic Data Sets." *AIAA Journal* 49: 289–298.
- Lai, P. K., A. Gallegos, N. Mody, H. A. Sathish, and B. L. Trout. 2022. "Machine Learning Prediction of Antibody Aggregation and Viscosity for High Concentration Formulation Development of Protein Therapeutics." *mAbs* 14: 2026208.
- Li, B., S. Luo, and W. Wang, et al. 2025. "PROPERMAB: An Integrative Framework for In Silico Prediction of Antibody Developability Using Machine Learning." *mAbs* 17: 2474521.
- Li, S., L. Feng, P. Benner, and A. Seidel-Morgenstern. 2014. "Using Surrogate Models for Efficient Optimization of Simulated Moving Bed Chromatography." *Computers & Chemical Engineering* 67: 121–132.
- Love, P. E. D., W. Fang, J. Matthews, S. Porter, H. Luo, and L. Ding. 2023. "Explainable Artificial Intelligence (XAI): Precepts, Models, and Opportunities for Research in Construction." *Advanced Engineering Informatics* 57: 102024.
- Lu, Y.-A., Y. Fukae, W.-S. Hu, and Q. Zhang. 2025. "Recurrent Neural Networks for Forecasting Time-Varying Process Behavior in Mammalian Cell Culture." *Industrial & Engineering Chemistry Research* 64: 9048–9058.
- Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, edited by L. Guyon, U. Luxburg, S. Bengio, H. Wallach, R. Fergus, and S. Vishwanathan, Vol. 30, 4768–4777. Curran Associates Inc.
- McDonald, M. A., B. A. Koscher, R. B. Canty, J. Zhang, A. Ning, and K. F. Jensen. 2025. "Bayesian Optimization over Multiple Experimental Fidelities Accelerates Automated Discovery of Drug Molecules." *ACS Central Science* 11: 346–356.
- Mikulic, M. 2024. "Biopharmaceuticals Market Size Worldwide Forecast 2022–2032." <https://www.statista.com/statistics/1293077/global-biopharmaceuticals-market-size/#statisticContainer>.
- Min, R., Z. Wang, Y. Zhuang, and X. Yi. 2023. "Application of Semi-Supervised Convolutional Neural Network Regression Model Based on Data Augmentation and Process Spectral Labeling in Raman Predictive Modeling of Cell Culture Processes." *Biochemical Engineering Journal* 191: 108774.
- Miró-Nicolau, M., A. Jaume-i-Capó, and G. Moyà-Alcover. 2024. "Assessing Fidelity in XAI Post-Hoc Techniques: A Comparative Study With Ground Truth Explanations Datasets." *Artificial Intelligence* 335: 104179.
- Da Mommio, M., A. Manzoni, and P. Conti. 2025. "MELTER: A Transformer Architecture for Multi-Fidelity Problems."
- Mu'azzam, K., F. V. Santos da Silva, J. Murtagh, and M. J. Sousa Gallagher. 2024. "A Roadmap for Model-Based Bioprocess Development." *Biotechnology Advances* 73: 108378.
- Nisevic, M., D. Milojevic, and D. Spajic. 2025. "Synthetic Data in Medicine: Legal and Ethical Considerations for Patient Profiling." *Computational and Structural Biotechnology Journal* 28: 190–198.
- Pavčnik, L., M. Prunk, T. Trdan Lušin, and R. Roškar. 2025. "Accelerated Predictive Stability Testing: Accelerating Registration Phase and Application of Reduced Designs for Shelf-Life Determination of Parenteral Drug Product." *Pharmaceutics* 17: 160.
- Pedro, F., F. Veiga, and F. Mascarenhas-Melo. 2023. "Impact of GAMP 5, Data Integrity and QbD on Quality Assurance in the Pharmaceutical Industry: How Obvious Is It?" *Drug Discovery Today* 28: 103759.
- Peherstorfer, B., K. Willcox, and M. Gunzburger. 2018. "Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization." *SIAM Review* 60: 550–591.
- Peng, J., T. T. Khuat, K. Musial, and B. Gabrys. 2026. "Machine Learning Methods for Small Data and Upstream Bioprocessing Applications: A Comprehensive Review." *Biotechnology Advances* 87: 108749.
- Perdikaris, P., M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. 2017. "Nonlinear Information Fusion Algorithms for Data-Efficient Multi-Fidelity Modelling." *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473: 20160751.
- Petsagkourakis, P., B. Chachuat, and E. Antonio Del Rio-Chanona. 2021. "Safe Real-Time Optimization using Multi-Fidelity Gaussian Processes." In *Proceedings of the IEEE Conference on Decision and Control*, Vol. 2021, 6734–6741. Institute of Electrical and Electronics Engineers Inc.
- Pham, T. D., C. Manapragada, Y. Sun, R. Bassett, and U. Aickelin. 2023. "A Scoping Review of Supervised Learning Modelling and Data-Driven Optimisation in Monoclonal Antibody Process Development." *Digital Chemical Engineering* 7: 100080.
- Phiri, V. J., I. Battas, A. Semmar, H. Medromi, and F. Moutaouakkil. 2025. "Towards Enterprise-Wide Pharma 4.0 Adoption." *Scientific African* 28: e02771.
- Raissi, M., P. Perdikaris, and G. E. Karniadakis. 2019. "Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations." *Journal of Computational Physics* 378: 686–707.
- Ramos, J. R. C., J. Pinto, G. Poiars-Oliveira, L. Peeters, P. Dumas, and R. Oliveira. 2024. "Deep Hybrid Modeling of a HEK293 Process: Combining Long Short-Term Memory Networks With First Principles Equations." *Biotechnology and Bioengineering* 121: 1554–1568.
- Rathore, A. S., and G. Kapoor. 2015. "Application of Process Analytical Technology for Downstream Purification of Biotherapeutics." *Journal of Chemical Technology & Biotechnology* 90: 228–236.
- Rathore, A. S., D. Kumar, and N. Kateja. 2018. "Role of Raw Materials in Biopharmaceutical Manufacturing: Risk Analysis and Fingerprinting." *Current Opinion in Biotechnology* 53: 99–105.
- Rathore, A. S., S. Nikita, G. Thakur, and S. Mishra. 2023. "Artificial Intelligence and Machine Learning Applications in Biopharmaceutical Manufacturing." *Trends in Biotechnology* 41: 497–510.
- Raybould, M. I. J., C. Marks, A. P. Lewis, et al. 2020. "Thera-SAbDab: The Therapeutic Structural Antibody Database." *Nucleic Acids Research* 48: D383–D388.
- Regazzoni, F., S. Pagani, A. Cosenza, A. Lombardi, and A. Quarteroni. 2021. "A Physics-Informed Multi-Fidelity Approach for the Estimation of Differential Equations Parameters in Low-Data or Large-Noise Regimes." *Rendiconti Lincei, Matematica e Applicazioni* 32: 437–470.
- Roch, P., and C.-F. Mandenius. 2016. "On-Line Monitoring of Downstream Bioprocesses." *Current Opinion in Chemical Engineering* 14: 112–120.
- Sathiyapriyan, P., S. Mukherjee, and T. Vogel, et al. 2025. "Current PAT Landscape in the Downstream Processing of Biopharmaceuticals." *Analytical Science Advances* 6: e70013.
- Shareef, U., A. Altaf, M. Ahmed, et al. 2024. "A Comprehensive Review of Discovery and Development of Drugs Discovered From 2020–2022." *Saudi Pharmaceutical Journal* 32: 101913.

- Shi, N., H. Yan, S. Guo, and R. Al Kontar. 2025. "Diffusion-Based Surrogate Modeling and Multi-Fidelity Calibration." *IEEE Transactions on Automation Science and Engineering* 22: 18217–18230.
- Strüssmann, B. R., A. R. de Queiroz, and L. Hvam. 2025. "Data-Driven Yield Improvement in Upstream Bioprocessing of Monoclonal Antibodies: A Machine Learning Case Study." *Processes* 13: 3394.
- Tazi, K., J. A. Lin, and R. Viljoen, et al. 2023. "Beyond Intuition, a Framework for Applying GPs to Real-World Data." <https://arxiv.org/pdf/2307.03093>.
- Tulsyan, A., C. Garvin, and C. Ündey. 2018. "Advances in Industrial Biopharmaceutical Batch Process Monitoring: Machine-Learning Methods for Small Data Problems." *Biotechnology and Bioengineering* 115: 1915–1924. https://doi.org/10.1002/bit.26605?saml_referrer.
- Tulsyan, A., T. Wang, G. Schorner, H. Khodabandehlou, M. Coufal, and C. Undey. 2020. "Automatic Real-Time Calibration, Assessment, and Maintenance of Generic Raman Models for Online Monitoring of Cell Culture Processes." *Biotechnology and Bioengineering* 117: 406–416.
- Ündey, C., B. A. Williams, and A. Çınar. 2002. "Monitoring of Batch Pharmaceutical Fermentations: Data Synchronization, Landmark Alignment, and Real-Time Monitoring." *IFAC Proceedings Volumes* 35: 271–276.
- Vaskó, D., D. K. Hajdú, E. Márkus, et al. 2025. "Implementation of Inline Raman Spectroscopy for Buffer Exchange Monitoring in Monoclonal Antibody Formulation." *International Journal of Pharmaceutics* 684: 126129.
- Virtanen, T., S.-P. Reinikainen, M. Kögler, M. Mänttari, T. Viitala, and M. Kallioinen. 2017. "Real-Time Fouling Monitoring With Raman Spectroscopy." *Journal of Membrane Science* 525: 312–319.
- Wang, J., H. Jiang, and Y. Liu, et al. 2024. "A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks." <http://arxiv.org/abs/2408.01319>.
- Wutz, J., R. Steiner, K. Assfalg, and T. Wucherpennig. 2018. "Establishment of a CFD-Based $k_L a$ Model in Microtiter Plates to Support CHO Cell Culture Scale-Up During Clone Selection." *Biotechnology Progress* 34: 1120–1128.
- Xing, Z., B. M. Kenty, Z. J. Li, and S. S. Lee. 2009. "Scale-Up Analysis for a CHO Cell Culture Process in Large-Scale Bioreactors." *Biotechnology and Bioengineering* 103: 733–746.
- Zhang, J., J. Zhao, G. Cheng, A. Rouhani, and X. Chen. 2025. "Explainable Multi-Fidelity Bayesian Neural Network for Distribution System State Estimation." *Applied Energy* 392: 125972.
- Zhang, Z., A. R. Lamson, M. Shelley, and O. Troyanskaya. 2023. "Interpretable Neural Architecture Search and Transfer Learning for Understanding CRISPR–Cas9 Off-Target Enzymatic Reactions." *Nature Computational Science* 3: 1056–1066.
- Zhang, Z., Q. Ye, D. Yang, N. Wang, and G. Meng. 2024. "A Multi-Fidelity Transfer Learning Strategy Based on Multi-Channel Fusion." *Journal of Computational Physics* 506: 112952.
- Zhou, J., G. Cui, S. Hu, et al. 2020. "Graph Neural Networks: A Review of Methods and Applications." *AI Open* 1: 57–81.
- Zürcher, D., K. Wuchner, and P. Arosio. 2024. "Real-Time Observation of Protein Aggregation at Liquid-Liquid Interfaces in a Microfluidic Device." *Small* 20: e2401085.