Trends in Biotechnology

CellPress

Review

Machine learning in bioprocess development: from promise to practice

Laura M. Helleckes,[1,2] Johannes Hemmerich,[1,6] Wolfgang Wiechert,[1,2] Eric von Lieres,[1,2] and Alexander Grünberger[3,4,5,*]

Fostered by novel analytical techniques, digitalization, and automation, modern bioprocess development provides large amounts of heterogeneous experimental data, containing valuable process information. In this context, data-driven methods like machine learning (ML) approaches have great potential to rationally explore large design spaces while exploiting experimental facilities most efficiently. Herein we demonstrate how ML methods have been applied so far in bioprocess development, especially in strain engineering and selection, bioprocess optimization, scale-up, monitoring, and control of bioprocesses. For each topic, we will highlight successful application cases, current challenges, and point out domains that can potentially benefit from technology transfer and further progress in the field of ML.

Highlights

Bioprocess development requires identification of robust design spaces for specific bioproducts and involves efficient strain selection, bioprocess optimization, scale-up, and optimal control strategies for robust industrial production.

Beyond multivariate data analysis, deep learning, reinforcement learning, and other novel ML techniques start to complement and replace traditional data analysis approaches to accelerate screening, optimization, and control procedures.

Transfer learning is emerging as a means to leverage the potential of historic data to guide novel production processes.

No single algorithmic solution will be suitable for all aspects of bioprocess development. Instead, a flexible combination of various techniques is required to enhance the whole development pipeline.

Fast impact is expected in autonomous strain selection and the optimization of bioprocess parameters. The application of ML for scale-up has a high impact but needs further development.

Machine learning (ML) in biotechnology: state of the art

ML has become the most important discipline of artificial intelligence (AI) in terms of practical application. ML deals with algorithms and programs that learn to solve certain tasks based on data, where performance increases with experience (i.e., available data) [1]. More precisely, ML aims at finding suitable, mostly empirical models to describe datasets, learning from labeled samples or by identifying inherent patterns (see Box 1 for central paradigms). The vast spectrum of ML methods (Boxes 2 and 3) is particularly useful when large amounts of data are available and/or when datasets are too complex to be analyzed by sets of predefined rules (see the explanation of expert systems in Box 1). Other applications of ML aim at finding so-called **surrogate models** (see Glossary), in which ML models are used as approximations for mechanistic models that are costly or hard to evaluate [2].

In recent years, the life sciences have started looking into available ML methods and researchers began to assess which of these methods are suitable to tackle current challenges [3]. Thus, biology and biotechnology became influenced by recent advances in ML. This is reflected by many reviews, for example, ML in protein function prediction [4], multi-omics data analysis [5], developmental biology [6], biological network analysis [7], metabolic engineering [8], and biochemical engineering [9].

Generally, the biotechnological pipeline from target molecule to final product covers four essential stages: (i) target identification and molecule design, (ii) biocatalyst design, (iii) **bioprocess development**, as well as (iv) industrial-scale production. The first two stages are mainly addressed by molecular biotechnology and bioinformatics; in recent years, both fields were heavily influenced by technological progress on the experimental side (e.g., omics technologies) as well as increased computational power [10]. The resulting availability of big data and compute resources enabled the rise of ML, which nowadays is state-of-the-art. A notable, recent breakthrough of ML is AlphaFold [11], a deep learning (Box 2) program that predicts the 3D structure of proteins

[1]Institute for Bio- and Geosciences (IBG-1), Forschungszentrum Jülich GmbH, 52428 Jülich, Germany
[2]RWTH Aachen University, Templergraben 55, 52062 Aachen, Germany
[3]Multiscale Bioengineering, Technical Faculty, Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany
[4]Center for Biotechnology (CeBiTec), Bielefeld University, Universitätsstr. 25, 33615 Bielefeld, Germany
[5]Institute of Process Engineering in Life Sciences, Section III: Microsystems in Bioprocess Engineering, Karlsruhe Institute of Technology, Fritz-Haber-Weg 2, 76131, Karlsruhe, Germany

from sequence data. Since ML is abundant and diverse for the first two stages (i.e., molecule prediction and biocatalyst design), a thorough review is out-of-scope for this paper. The reader is instead referred to existing reviews (e.g., [7,12–15]).

The third stage of the biotechnological production pipeline, bioprocess development, focuses on increasing the production capacity for the target molecule by means of strain selection, process optimization, and scale-up. During this stage, high-throughput screening (HTS) experiments are typically performed to assess the performance of selected clones [16]. Furthermore, optimal cultivation parameters need to be identified from a huge design space. However, traditional analytical methods such as mass spectrometry will often not match the rate of experimentation and thus, analysis subsequently becomes a bottleneck [17,18]. This challenge can potentially be addressed

[6]Current address: DSM Food Specialities B.V., Alexander Fleminglaan 1, 2613 AX Delft, The Netherlands

*Correspondence: alexander.gruenberger@kit.edu (A. Grünberger).

Box 1. Machine learning (ML): paradigms and general challenges

Artificial intelligence (AI) is a broad field of science with a constantly changing definition [150]. One of the founders of the discipline, John McCarthy, defined it as 'the science and engineering of making intelligent machines' (http://www-formal.stanford.edu/jmc/whatisai.pdf). In practice, ML was the most important field of AI in recent years [151]. In contrast to expert systems, which are knowledge-based systems that emulate human decision-making by rules of reasoning [152], the focus of ML is to improve performance based on training data, which can be later applied to test data.

ML can be further divided into different subfields and central paradigms. A common distinction is between supervised, unsupervised, and reinforcement learning. The first one is using labeled data during training, meaning that the data provides the expected output. The goal is to learn functions that can describe relationships between input and output variables [153]. In contrast, unsupervised learning uses unlabeled input data; instead of providing labels in training, this branch of methods is focused on identifying patterns that are inherent to data [154]. Finally, reinforcement learning (Box 2) uses a different approach. Here, an agent acts in an environment, where a policy is learned to maximize the long-term reward [155]. In contrast to labeled data in supervised learning, the input signals for the reward are often delayed and batch-wise optimization is thus frequent [156].

To illustrate several common challenges and principles of ML, a regression task is visualized in Figure I. Regression is a typical supervised problem in which the relation between an observed, dependent variable and an independent variable is targeted. Essentially, a function (red) is learned that captures the trajectory (Figure IA). Different ML methods assume different functions (e.g., a linear function in **linear regression** or a normal distribution in GP regression; Box 3). The parameters of these functions are learned using labeled data (black dots), in this case response measurements with known input signals. Uncertainty quantification can be used to quantify how certain the estimation is (blue bands), depending on the input signal (Figure IA). The more data available, the lower the uncertainty in the parameter estimation. However, if supervised learning is performed on small sets of training data, overfitting is a common problem (Figure IB). In this case, the learned function is too specific for the training data and not able to sufficiently adapt for test datasets. In contrast, a function might be too simple to describe complex datasets, which is called underfit.

The challenge to describe training data well without overfitting is reflected in the concept of generalization, which is describing the ability of a model to represent unseen data that is originating from the same distribution as the training data. This topic is, for example, addressed in the field of transfer learning (Box 2). Here, common structures between models are identified to reuse trained models (parts) and make efficient use of historic data for new problems; this way, transfer learning can also help to reduce the generalization error [157]. This field of ML is particularly useful since it makes biotechnological processes, which are often small-data problems, accessible to otherwise big-data ML methods.

Overall, ML provides particularly useful methods when datasets are quite large or too complex for human analysis. ML is also beneficial to predict the behavior of biological systems for which domain knowledge is lacking. The current interest and investments in the AI sector foster rapid progress in method development and industry is acknowledging the growing asset of data [24], thus posing a demand for novel technologies. With advances in deep learning, reinforcement learning, and transfer learning [158], enabled by higher computational power, improved storage, and lower costs [159], bioprocess development is currently at the verge of a new, data-driven era. An overview about the most important methods for this review is given in Box 3. The current interest and investments in the AI sector foster rapid progress in method development and industry is acknowledging the growing asset of data [24], thus posing a demand for novel technologies. With advances in deep learning, reinforcement learning, and transfer learning [158], enabled by higher computational power, improved storage, and lower costs [159], bioprocess development is currently at the verge of a new, data-driven era. An overview about the most important methods for this review is given in Box 3.

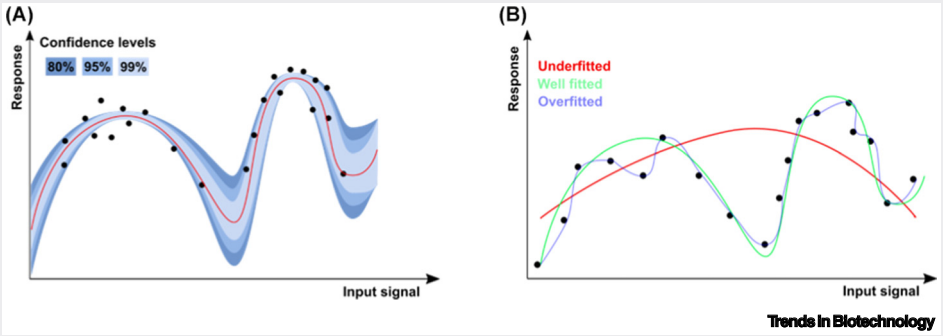


Figure I. Regression with uncertainty quantification (A) and the overfitting problem (B). (A) Typical machine learning (ML) task of fitting a model (red) to observed data (black dots) for a response in dependency of the input signal. Uncertainty (blue bands) is lower where more data is observed. (B) The number of parameters in a nonlinear model indicates its complexity. If it is much smaller than the number of observations, the model often fails to describe the training data (underfit model). If the number of parameters is much bigger than the observations, the model will be unable to generalize beyond the training set. The solution for the underfitting case is straightforward: increase the complexity of the model (number of parameters). In case of a neural network, this could, for example, mean an additional hidden layer or more neurons per layer. One solution for the overfitting case is reducing the model parameters. However, if the number of inputs/features is high, many parameters might be required for a good fit and training can be computationally demanding.

by using ML to rank samples by their predicted information content and accordingly schedule their analysis while a (high-throughput) platform can already perform further experiments.

Since biological and process parameters are correlated, iterative experimentation and data evaluation is needed to feed back information and insights from screening to strain design. This approach is reflected in the **design-build-test-learn (DBTL)** cycle [19], which is sometimes only referring to synthetic biology, but can be also applied to the stage of bioprocess development [20]. In the context of DBTL, all steps can be enhanced by ML, in particular to suggest informative designs for the next round of experimentation [21].

Box 2. Machine learning (ML): overview of concepts

- Supervised/unsupervised learning: see Box 1 in the main text.
- Deep learning: this field of ML extracts high-level, often hierarchical, features from raw input by learning how to represent them. Deep neural networks comprise a multitude of layers with (often nonlinear) activation functions, the composition of which is able to model nonlinear dependencies [10,160,161]. Deep learning methods are suitable for complex input such as image data [8].
- Transfer learning: a model developed for a specific task is reused as the starting point for a model on a second task. Initial efforts in developing a model structure are not needed anymore for working on the second task. Therefore, less training resources are needed. Key for transfer learning is to then control and handle the corresponding uncertainty [157].
- Reinforcement learning (RL): trial-and-error approach in which an agent acts in an environment choosing certain actions according to a policy that needs to be learned [155]. There are two main approaches: value- and policy-based methods. Value-based approaches try to estimate the value of all actions and states by a function to find the optimal policy; the other type, policy-based algorithms, try to learn the optimal policy directly from a policy space [162].
- Ensemble learning: instead of using a single model to learn, several (different) models are trained on the same dataset. Typically, their combined (or averaged) predictions yield a higher accuracy compared with a single algorithm, but at the cost of higher computing demand [163]. An example are random forests, which are an ensemble of many decision trees (Box 3).

Glossary

Backpropagation: in the context of artificial neural networks, backpropagation refers to a method that allows calculation of the gradient of the loss function, which is used for training the network (i.e., identifying its weights using data) [170].
Bayesian statistics: field of statistics covering methods with a Bayesian interpretation of probability, which, according to Bayes theorem, includes prior beliefs in an event. Methods cover, among others, Bayesian inference of parameters, statistical modeling, Bayesian optimization (for sequential design), and Bayesian networks.
Bioprocess development: bioprocess development aims for the identification of a robust design space for the production of a specific bioproduct with desired yield and purity. It requires experiments and data analysis to understand the interaction of parameters within the specific bioprocess.
Constraint-based modeling (COBRA): subfield of systems biology that takes into account the underlying physical, enzymatic, and topological constraints of a phenotype in a metabolic network [214]. Methods include flux balance analysis (FBA) [46], minimization of metabolic adjustment (MOMA) [47], or minimal cut sets (MCS) [48].
Critical process parameters (CPP): parameters of a (bio-)pharmaceutical production process that have been shown to affect the critical quality attributes of the final product. The parameter values have to be monitored and kept in proven ranges to not affect the corresponding critical quality attributes in a negative way.
Critical quality attributes: attributes of a (bio-)pharmaceutical product that determine its quality and for which certain value ranges have to be met in order to release the product.
Design-build-test-learn (DBTL) cycle: a loop used recursively to obtain a design that satisfies the desired specifications. In bioengineering, the DBTL cycle makes use of synthetic biology to engineer biomanufacturing solutions for industrial application.
Design of experiments (DoE): in design of experiments, the goal is to create an empirical model that describes how a process responds to changes in influential factors. It is often performed in two stages: (i) screening for identification

Box 3. Machine learning (ML): overview of methods and applications

ML provides researchers with a huge set of methods for data analysis. In the following we will shortly introduce the most important ones. The reader is referred to several recent reviews, which give an overview and short description of ML methods in the fields of biology and biotechnology [164,165]. For details on each method the reader is referred to the included references.

Methods

- Artificial neural network (ANN): network of nodes (neurons), which are connected by edges. Weights on each connection of a neuron influence the output, which is calculated by the weighted inputs. Neurons are typically structured into an input layer, an output layer, and hidden layers in-between [166].
- Recurrent neural network (RNN): RNNs can operate on ordered data like time-series data or sentences, where the sequence of the individual data points is important. Compared with feedforward neural networks, neurons can share connections and parameters with the same or previous layers. Thereby, information from prior inputs influences current inputs and outputs [167].
- Convolutional neural network (CNN): type of neural networks that is often applied in image processing. The name originates from the additional convolutional layers, which have the purpose to abstract the input by applying filter matrices on it. It is designed to automatically and adaptively learn spatial hierarchies of features (e.g., shapes in hand-written text) [168].
- Bayesian neural network (BNN): instead of identifying a single optimal set of parameters to define a neural network, BNNs determine probability distributions for each parameter, thus representing an infinite number of models that can describe the data [169]. This approach allows to quantify the uncertainty in parameters of a neural network [124].
- Autoencoder: autoencoders have the purpose of mapping high-dimensional inputs to a new feature representation (encoding) in a way that the input can be (approximately) reconstructed from the representation [170]. Although variational autoencoders (VAE) share basic architectural features with autoencoders, their purpose and mathematical formulation differ significantly. VAEs use a variational Bayesian model formulation and are applied as generative models, comparable with the application of GANs [171].
- Random forests: this technique is an ensemble learning method for classification and regression from many decision trees. The latter classify data by continuously splitting it according to certain features [8]; thereby, limited prediction power of an individual tree is overcome by the joint prediction from a forest of such trees [172].
- Support vector machines: algorithm that learns by example to assign labels to objects by separating those into two groups. Separation of groups is achieved by a hyperplane that maximizes its distance to the majority of all elements in the respective groups [173,174]. If applied in regression, the technique is referred to as support vector regression (SVR).
- Gaussian processes (GP): probability distributions over random functions that approximate sets of data points. The name originates from the fact that a subset of the random variables in a GP can always be described by a multivariate Gaussian distribution. GPs are often applied to learn (multidimensional) functional relationships from iteratively generated data [175]. Associated uncertainty of predictions can guide iterative experimental design towards experimental conditions with possible high reward.
- Generative adversarial networks (GANs): generative modeling is an unsupervised learning task that involves automatically discovering and learning the regularities or patterns in input data. It is done such that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset [176].
- Embedding vectors: concept taken from natural language processing (NLP), where it is often referred to as word embedding. The goal in the context of NLP is to quantify similarity of words (e.g., semantic similarity). The method uses vector spaces, thus assigning real numbers to each feature [177]. Modern word embeddings in NLP are based on learning weights of neural networks [178].
- Policy gradient algorithms: a policy-based reinforcement learning technique that relies on optimizing parametrized policies with respect to the expected return (long-term cumulative reward) using gradient descent [162].
- Q-learning algorithm: a model-free, value-based reinforcement learning algorithm. The Q value refers to the expected reward of playing an action at a certain state following a specific policy [179].

Applications

- Statistical process control (SPC): method of controlling any process based on monitoring temporal evolution of statistics derived from measurements that were demonstrated to be representative for process performance. Based on historical data, specification limits are determined which indicate whether the process runs in-spec or if action must be taken, depending on different patterns observed in the control charts.
- Model predictive control (MPC): a control concept to predict the future behavior of the controlled system by online assessment of current data. MPC computes an optimal control input while ensuring satisfaction of given system constraints [180]. In contrast to SPC, which evaluates (validated) limits of summary statistics derived from the process, MPC uses predictive models to forecast the temporal evolution.
- (Text) data mining: this ML-powered technology uses natural language processing to examine large volumes of documents for extraction and structuring of contained information. Use cases are, for example, to discover new information that helps answering research questions [181,182].

of influential factors, and (ii) prediction of response surfaces to identify optimal operation conditions [215]. Intensified DoE (iDoE) is an adaptation in which several set points for influential factors are tested as intra-experiment variations, thus reducing the overall number of experiments [216].

Digital twins: detailed, virtual representations of production systems, where feedback between model and physical systems is characteristic [217]. Applications include real-time monitoring of manufacturing processes and fault detection [116].

Flux balance analysis (FBA): a mathematical method to solve stoichiometric metabolic networks for steady state flux solutions. Applications include identification of targets for metabolic engineering and media design.

Hybrid modeling: combines data-driven model components with mechanistic, *a priori* knowledge into one superior model. In bioprocess development, such models often comprise systems of ODEs/DAEs and Gaussian processes or a type of artificial neural network.

Linear regression: a linear approach for modeling the relationship between a response and one or more explanatory variables. Nonlinear regression is a form of regression analysis in which observational data are modeled by a function that is a nonlinear combination of the model parameters and depends on one or more independent variables.

Natural language processing: field of AI that is concerned with automatically analyzing and representing human language; as such it is closely related to computer science and linguistics [218]. Applications include speech recognition, machine translation, and synthesis of language [219].

Process analytical technology (PAT): system for designing, analyzing, and controlling (pharmaceutical) manufacturing processes through measurements of critical quality and performance attributes (https://www.fda.gov/media/71012/download).

Surrogate models: approximations that are used when the desired output of a system is expensive or hard to simulate [2]. An example is physics-informed neural networks (PINNs) that can, for example, be used to replace costly simulations in computational fluid dynamics [102].

The fourth and final stage of the biotechnological production pipeline is concerned with reproducible and robust operation, for example, by controlling raw materials [22] of industrial-scale bioprocesses. Research at this stage is determining the long-term stable and consistent operation of a production process, for example, by process intensification at large scale to increase manufacturing capacity [23]. Methods and results at this stage are often proprietary and therefore scarcely available in public literature [24]. In market and business analyses, necessary productivity ranges are determined to meet bioprocess economics, a prerequisite for the fourth stage.

In the light of increased automation, data availability, and data exchange, data-driven bioprocess development will accelerate the time-to-market of bioproducts [25]. Moreover, the large cash flow in the AI sector will boost the integration of novel methods to the development pipeline [24]. In this review we will thus mainly discuss the application of ML in bioprocess development, particularly in upstream processes. ML is also advancing in downstream processing, where corresponding techniques are developed for specific technologies such as chromatography [26–31], but also for complex purification pipelines of specific products such as antibodies [32] or inclusion bodies [33]. Since available literature is highly diverse and vast enough to be covered in a separate review, ML for downstream processing is not discussed here. Where possible, ML trends in commercial bioprocesses [e.g., for **process analytical technology (PAT)**, **digital twins**, or model predictive control (MPC)] are included. Overall, four main topics are addressed:

1. Strain selection and engineering,
2. Bioprocess optimization,
3. Scale-up of bioprocesses,
4. Process monitoring and control.

For each area, we shed light on exemplary studies of the past years that push ML applications forward. Furthermore, we discuss potentials and challenges for future usage of ML methods. This leads to an overall discussion about ML as an enabling technology for bioprocess development.

Choosing between numerous candidates: strain engineering and selection

One central step before bioprocess development takes place is the selection of a biocatalyst or microorganism for production. State-of-the-art experimental methods for HTS exist to identify potent biocatalysts (e.g., by quantitative phenotyping of strain libraries) [16,34,35]. The current bottleneck is thus automated data processing and algorithm-driven decision making to select biocatalysts with the highest potential for commercial production.

Recent advances in ML provide a number of techniques to foster biochemical engineering of strains [8]. As a major challenge, the diversity of biocatalysts leads to a broad range of possible tasks, for example, design and selection of bacterial production strains, predicting production in different cell-free systems, or engineering mammalian cell lines. The latter poses many additional challenges such as clonal variation [36] and large-scale studies are needed to generate mechanistic understanding, which is so far required for non-ML methods [37]. To maintain focus, we review strain selection here. For insights into ML approaches for enzyme and biocatalyst engineering, the reader is referred to [38,39].

In the past decades, stoichiometric as well as kinetic genome-scale models have been used for both metabolic engineering and bioprocess development [40–42]. Besides genetic design, such models can give insights into suitable carbon sources, media design, or bioreactor parameters [43]. For many years, quantitative predictions for metabolic engineering have been made using

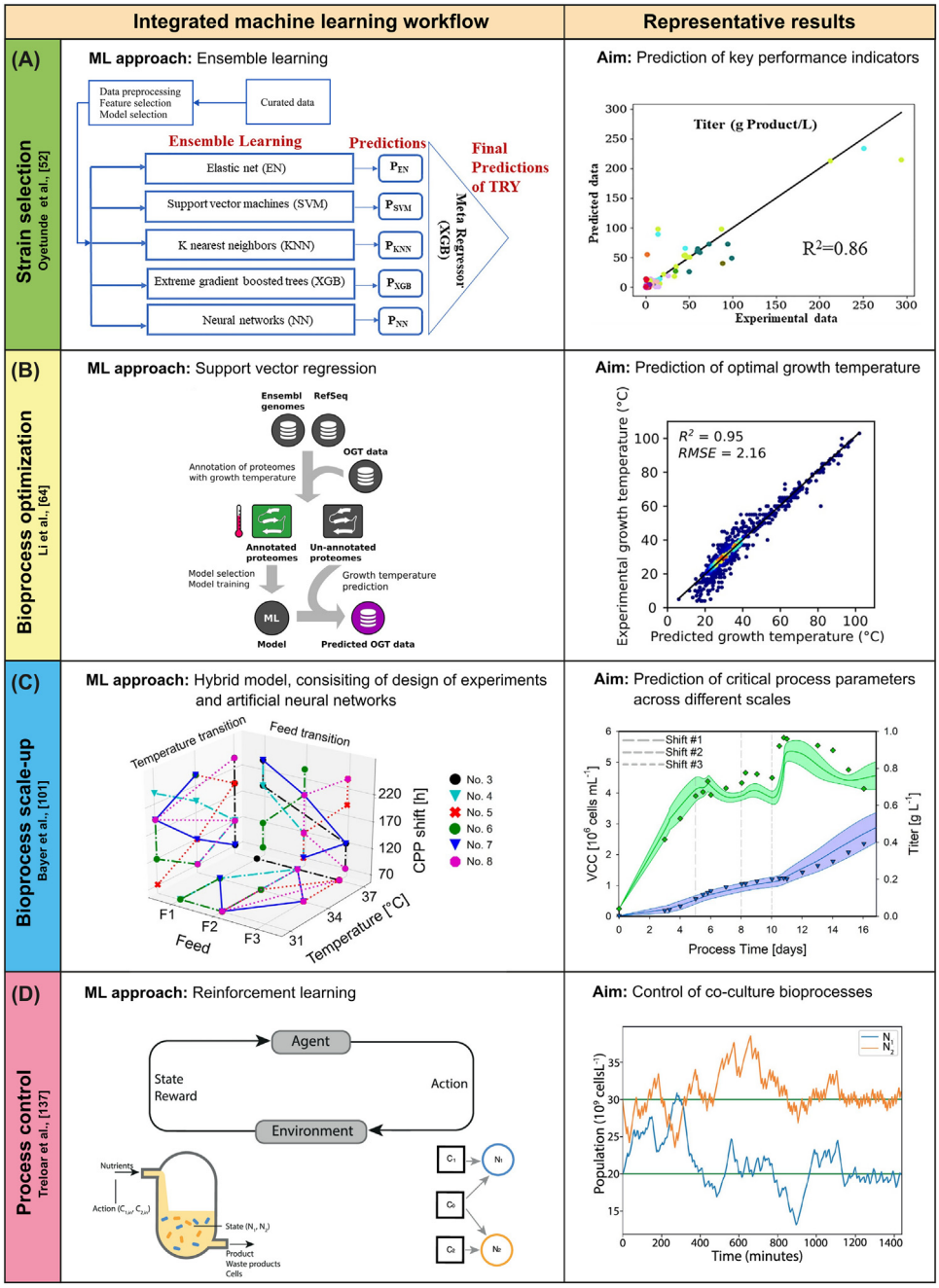
constraint-based modeling (COBRA) of genome-scale metabolic networks [44,45]. Methods of the COBRA toolbox like **flux balance analysis (FBA)** [46], minimization of metabolic adjustment (MOMA) [47], or minimal cut sets (MCS) [48] generally aim to optimize fluxes in a biological network (i.e., metabolism) to improve productivity by, for example, reducing side product formation or eliminating competing metabolic pathways. Resolving metabolic pathways and determining corresponding fluxes is experimentally demanding. As a result, FBA is largely limited by understanding of the underlying network structure [45]. Within the COBRA toolbox, FBA is probably the most popular method to find steady-state flux solutions.

In contrast, data-driven ML algorithms allow for analysis of large, complex (multi-)omics datasets, which can be generated in high throughput [10,49]. Different applications of ML for genome-scale models are emerging. On the one hand, ML is used to complement the typical modeling pipeline for constraint-based models, namely in the steps of gene annotation, gap filling, and integration of multi-omics data [50]. On the other hand, novel approaches of **hybrid modeling** have been proposed, as well as ML methods that replace the mechanistic, genome-scale models completely.

King *et al.* applied literature mining (Box 3) to create a database of *Escherichia coli* strain variants and their byproduct streams. They demonstrated how the database can be used to validate different genome-scale models [51]. Oyetunde *et al*. [52] combined simulated data from a genome-scale model with a manually curated dataset of bioprocess data from different *E. coli* strains as input data to predict production metrics for various products. As ML methods, they applied a combination of principal component analysis (PCA) and ensemble learning (Box 2), a strategy where different ML algorithms are combined to learn more efficiently (Figure 1A). The approach led to decent predictions of production titers, rates, and yields (TRY) under varying process and pathway conditions, thus demonstrating the potential of integrating both large datasets and mechanistic knowledge from the genome-scale model.

In a similar approach, Zhang *et al*. [53] first used a genome-scale model to identify relevant genes for metabolic engineering of tryptophan production strains to then apply ensemble learning on biosensor data generated with promoter libraries of the suggested genes. In contrast to Oyetunde *et al.* [52], this model does not use the genome-scale model predictions as training data, but only to identify genomic targets for the promoter libraries. The developed ML models were used to predict combinations of promoters and genes outside the training dataset, thus augmenting the experimentally tested designs. Although this approach led to even further improved variants, the authors observed a reduced prediction performance of the algorithm compared with the original tested library when constructing those designs in the wet lab. This lack of extrapolation is a commonly known problem of ML approaches. Since ML methods rely on training datasets to learn, exposing a trained model to data that lies outside the trained parameter ranges requires assumptions that are often difficult to make.

Among other tools, Zhang *et al*. [53] made use of the recently developed Automated Recommendation Tool for synthetic biology [54]. The toolbox combines the scikit-learn framework [55] with a **Bayesian statistics** approach of ensemble models, (i.e., several models are trained for prediction and to provide uncertainty quantification). Moreover, the toolbox is flexible with regards to experimental requirements, such as with the DBTL cycle, for which it can provide recommendations on strain design for the next iteration. This tool is an interesting starting point to further apply ML for synthetic biology [54,56] and should be expanded by the community.



Trends in Biotechnology

Figure 1. Overview of key studies of machine learning (ML) approaches within bioprocessing application for the four fields. (A) Left: ML pipeline. Ensemble learning using stacked regressors. (A) Right: Prediction of production metrics on the example of titer. Unbroken lines are shown on the diagonal that represent where all the points would lie for perfect prediction. Colored dots represent different products, ranging from fatty acids to amino acids. Reprinted from [52] with permission. (B) Left: Schematic overview of building an ML model to predict optimal growth temperature (OGT) for cells. (B) Right: Performance of the final support vector regression (SVR) model trained on dipeptide data. The correlation between predicted OGTs and those present in the original annotated dataset was evaluated. Colors indicate the density of the points. Reprinted from [64] with permission. Copyright 2019 American Chemical Society. (C) Left: Design space of the

(Figure legend continued at the bottom of the next page.)

As a further example of ML applications for genome-scale models, a recent preprint introduced artificial metabolic networks (AMN) [57], a concept where fluxes are predicted with a recurrent neural network (RNN) (Box 3). Here, FBA predictions are used to train the AMN, which can in turn replace the genome-scale model in the application phase. Since the AMN learns through **backpropagation**, it can be used to predict uptake rates based on external concentrations. Further methods of combining ML with genome-scale metabolic models (e.g., fluxomic analysis) have been recently reviewed [50,58,59].

Regarding kinetic models, detailed genome-scale models are often under-determined, meaning that the large number of kinetic parameters cannot be estimated from experimental data [60]. A remaining challenge thus is that these under-determined mathematical systems allow a multitude of parameter combinations that can equally well describe experimental measurements. However, many frameworks that determine the spaces of possible parameters propose a number of models that inherently contradict the experimentally observed physiology. To overcome this challenge, the reconstruction of kinetic models using deep learning (REKINDLE) framework was recently suggested, in which generative adversarial networks (GANs) (Box 3) are used to obtain mechanistic, kinetic models with biologically feasible dynamics [61]. The GAN architecture was essentially used to train one network in generating feasible kinetic models of a certain class, while the second network learned to distinguish between the training data and the generated data. This way, the generator is trained to produce results that are indistinguishable from the training input. The authors successfully validated the framework, which was shown to outperform existing frameworks regarding computation time. Moreover, they showed that via transfer learning, trained models can be reused to also predict kinetic models for physiologies with lower data availability.

Guiding strain optimization, Sabzevari *et al*. [62] applied a multi-agent reinforcement learning algorithm to both experimental data and data from a genome-scale kinetic model to tune metabolic enzyme levels. The algorithm outperforms another ML approach, namely Bayesian optimization on Gaussian processes (GPs) (Box 3), as well as a random search approach. Moreover, the multi-agent reinforcement learning approach allows for including parallelized experiments, which is important to make efficient use of modern HTS.

Overall, most studies on strain engineering and selection so far only focus on model host organisms and only few studies looked into transferability to other host systems [63]. A promising tool in this context could be transfer learning (Box 2); however, this approach usually requires large amounts of data to train the initial model [9], which are frequently not available. In the field of strain modeling, transfer learning is still underexplored. We thus identify predictions for non-model organisms as well as providing comparable results for a variety of different strains as major bottlenecks. Most likely, no single ML algorithm will solve this issue; instead, a suite of ML algorithms is needed to cover such a complex task in the future.

intensified design of experiments (iDoE) performed on the 15-l scale. Critical process parameter transitions for iDoE are displayed as additional planes (z-axis). The individual iDoE bioprocesses are represented by different colors and symbols. (C) Right: Prediction of large-scale critical process parameters in an iDoE with a hybrid model trained on data from shaker-scale experiments. The model estimations for the viable cell concentration (VCC, green lines) and product titer (blue lines) are indicated along with the respective confidence interval (shaded area). The analytical measurements are given for the VCC (green squares) and the product titer (blue triangles). Reprinted from [101] with permission. (D) Left: ML pipeline using reinforcement learning (Box 2) for the control of bioprocess cocultures in a simulated chemostat. The agent adds nutrient sources C1 and C2 to control the coculture composition. (D) Right: Coculture population in a chemostat. During the exploration phase, in which the agent learns the policy, the population levels (N1; N2) vary and random actions are taken; as the exploration rate decreases (shift towards exploitation), they move to the target values (green lines). Reprinted from [137] with permission. Abbreviations: R^2, coefficient of determination; RMSE, root-mean-square error. See [64].

Raising and stabilizing TRY: bioprocess optimization

During bioprocess development and optimization, lab-scale bioprocesses are used to improve TRY by identifying optimal physico-chemical parameters for cultivation. In this context, different ML techniques are used.

Aiming at the application of microorganisms and enzymes at extreme temperatures, Li *et al*. [64] (Figure 1B) developed a support vector machine (SVM) (Box 3) regression model to predict the optimal temperature for enzymatic activity, using optimal growth temperature and amino acid sequence information as input features. Another common ML approach for bioprocess optimization is GP regression. Use-cases include optimization of pigment production in algae [65,66] and tuning of media composition for protein production in *Corynebacterium glutamicum* [67].

Finally, artificial neural networks (ANNs) (Box 2) are frequently applied for a range of applications (e.g. optimizing media composition for wheat germ [68] or for pigment production in cyanobacteria [69]). Other studies optimize fermentation parameters; Pappu *et al*. [70], for example, investigated temperature, fermentation time, pH, k_La, biomass, and glycerol as influential parameters for xylitol production in the yeast *Debaryomyces nepalensis*. Ebrahimpour *et al*. [71] optimized production of a thermostable lipase in a *Geobacillus* strain with growth temperature, medium volume, inoculum size, agitation rate, incubation period, and initial pH as input variables. Finally, some studies look into the complex interaction of media composition and fermentation parameters (e.g., in bioethanol production with *Saccharomyces cerevisiae* [72] or growth of cell lines for therapy [73]).

Aiming at the transfer of knowledge between different bioprocesses, Rogers *et al.* simulated dynamical behavior in biochemical processes for different organisms via transfer learning, in this case by partially preserving layers between different ANNs [74]. Hutter and coworkers [75] combined GP regression with transfer learning, more precisely embedding vectors (Box 2), a technique that is used in **natural language processing** to quantify similarity between words [76]. Both approaches show how historic data can be used to predict dynamics for new products, which is beneficial for bioprocess optimization.

Video and image data (e.g., of cell morphology) is a rich source of information for bioprocess analysis and control [77,78]. Here, microfluidic systems, in combination with life-cell imaging, have been pioneering the image analysis methods, among others, for HTS of strains and to get improved understanding of cellular behavior at bioprocess-relevant cultivation conditions [79,80]. Deep learning techniques (Box 2) are well-suited to process such complex raw data from images in an automated fashion, thus laying the foundation for microfluidic-assisted, high-throughput bioprocess development [81,82]. Recent examples include prediction of growth and dynamics in microfluidic single-cell cultivation [83,84] and microfluidic droplet reactors, where multi-layer ANNs were used to predict performance of flow-focusing droplet generators [85].

Other applications in process optimization include the use of microscopic image data for spatio-temporal analysis of biofilms [86] and algae cultivation [87]. The latter requires complex management of light conditions and growth patterns (e.g., to avoid mutual shading during cultivation) [88,89]. Here, Long *et al*. [87] used SVM regression to predict light distribution patterns from microscopy images, which provide insight into the growth behavior and could ultimately help to develop novel cultivation designs.

Finally, we see the advance of ML in automated flowsheet synthesis in chemical engineering, for which, for example, hierarchical reinforcement learning and graph neural networks have been

successfully applied [90,91]. Although not yet demonstrated for bioprocesses, such techniques have great potential to accelerate bioprocess development.

Challenges at the verge to commercial scale: bioprocess scale-up

Having selected strains and optimized process conditions at laboratory scale, a bioprocess needs to be transferred to industrial production scale. The production scale is typically subject to increased variability of materials and feed streams, more complex hydrodynamics, and decreased spatial homogeneity [92]. Physical scale-down simulators (i.e., networks of purposefully heterogeneous laboratory devices) can be used to select strains and optimize process conditions under industrially relevant conditions [92,93]. Detailed measurements in industrial equipment are often intricate and prohibitively expensive. Model-based scale-up and -down simulators can help with closing this gap [93,94] and ensure industrial feasibility of the designed processes [95]. These models can facilitate the transfer of scale-independent knowledge and information, mostly related to the catalyst, while correcting the influence of scale-dependent mechanisms, mostly related to transport phenomena. Certain challenges arise when ML models (e.g., of cell metabolism) that have been trained at laboratory scale need to extrapolate beyond the experienced environmental conditions when applied at production scale.

Scale-up modeling involves various interconnected relationships between bioreactor design parameters, process parameters, and hydrodynamic characteristics, which are almost impossible to theoretically describe and computationally trace using mechanistic models [96]. Scale-up has thus been studied with multivariate data analysis, where methods such as PCA, SVM regression, and partial least square are overlapping with ML (e.g., [97,98]). Recent ML approaches such as RNNs [99] or decision trees [100], as well as hybrid models [101], are starting to emerge for incorporation and transfer of information between different scales.

In their study, Bayer *et al*. [101] investigated hybrid models (differential equations/ANNs) for **design of experiments (DoE)** and intensified DoE (iDoE) experiments with mammalian cells at different scales; more precisely, experiments were run in shaker-scale, bolus fed-batch experiments, as well as continuously fed 15-l scale bioreactors. The authors found that a hybrid model trained on shaker-scale DoE data performed well on test data of the 15-l reactor, especially for the iDoE data (Figure 1C). Moreover, a second model was trained on 15-l iDoE experiments, in which intra-experiment variations are made. This model also performed reasonably well on data from 15-l static runs (without variation), indicating that the iDoE concept can be generalized for modeling of mammalian cells. Compared with other studies, which often retrain the whole model for different scales, thus requiring large amounts of data, the results of Bayer *et al*. [101] are promising since they show the possibility of generalizing scale-up models across scales, meaning that a model trained for small scale is still valid at large scale without retraining.

Overall, we consider ML to have great potential in the discovery of nontraditional scale-up criteria, for example, by correlating validated mechanistic models describing bioprocess performance at laboratory and production scales. Beyond this application, ML models can also be used as surrogate models for complex scale-up models (e.g. by replacing costly simulations in computational fluid dynamics) [102]. Though literature in the field of ML learning for bioprocess scale-up is still scarce, we anticipate that methods will be evolving quickly, potentially using the field of chemical engineering as a blueprint (e.g., [103]).

Monitoring and controlling bioprocesses: ML in PAT

In the final phases of bioprocess development, the transition to commercial production is targeted. Here, process monitoring and control are important steps, especially if the desired

product needs to meet complex (pharmaceutical) regulations by the FDA or European Medicines Agency [104]. The need to record process data in a structured way to prove consistency for regulatory approval offers the opportunity to apply ML techniques. To support effective and efficient monitoring of **critical process parameters (CPPs)** in bio(pharmaceutical) processes, the FDA introduced the PAT initiative [105]. Since controlling CPPs is pivotal to ensure validated ranges of **critical quality attributes** of the product, PAT aims to establish regulatory approved monitoring capabilities for in-process controls, thus ensuring sufficient quality of the final product while the bioprocess is running.

Soft sensors, which are of high interest in the PAT framework [106], use mathematical models (software) to make real-time predictions of a system, similar to hardware sensors [107]. Soft sensors can provide information about process variables that cannot be measured reliably or at all by mapping their prediction to frequent online data [108]; the corresponding models have to be constantly updated to fit the online process data best. Soft sensor models are structured in three different classes as well as any combination thereof: mechanistic models, multivariate statistics, and AI/ ML [109]. In the context of this review, we focus on the latter; however, several general reviews on soft sensors for bioprocessing exist [104,110,111]. Soft sensors are also important building blocks for digital twins, which predictively describe the production process behavior [112,113]. A key feature of digital twins is bidirectional data exchange between a physical process and its model twin [114]. By analyzing the systems behavior *in silico*, further experimentation is guided efficiently towards process validation and qualification since the corresponding DBTL cycle iterations can be run faster [115,116].

Many ML-based approaches for soft sensors rely on ANNs (Box 3) or SVMs (Box 3, Table 1) [117]. Successful examples include ANN-based soft sensors for erythromycin production [118], or biomass estimation in plant cell cultures [119], as well as the description of L-lysine fermentation process data using a multi-output least squares SVM regressor (Box 3) [120]. Recent advances in the field make use of deep learning (Box 2) instead [121]. For example, Gopakumar *et al*. [122] demonstrated that their deep soft sensor outperforms traditional SVM approaches for

Table 1. Application of machine learning (ML) algorithms during bioprocess development

ML algorithm	Stage			
	Strain selection	Bioprocess optimization	Scale-up	Process control
ANN		[68–73,85,183–190]	[101,191]	[118–120,122,131,137,192–194]
RNN	[57]		[99]	[133,134,195]
CNN	[52,196]	[84]		
Unsupervised feature representation (e.g., autoencoders)				[122,123]
Trees/random forests	[52]	[197,198]	[100]	[199]
SVM/SVR	[52]	[64,87]	[97]	[108,120,194]
Gaussian processes		[65–67,75]		[132,194]
GANs/variational autoencoders	[61]	[83]		
Graph-based neural networks		[91]		
Transfer learning	[200]	[74,75,201]		[127]
Reinforcement learning	[62]	[91,202]		[134–138,203]
Ensemble learning	[52–54,204,205]	[206–208]		[209,210]
Text data mining	[51,211]	[212,213]		
Hybrid modeling		[66,75,183,185,208]		[101]

nonlinear systems, shown for crucial parameters in two fermentation processes (streptokinase and penicillin). Interestingly, Yao *et al*. [123] developed a soft sensor that combines unsupervised learning for feature extraction with a semisupervised classification approach. Finally, Mowbray *et al*. [124] incorporated uncertainty quantification and nonlinearities in their soft-sensors by using probabilistic ML methods such as Bayesian neural networks (Box 3). All three studies highlight the potential of transferring ML technologies to the field of process monitoring to outperform conventional approaches.

However, similar to other applications, particularly in process scale-up, ML approaches for soft sensors suffer from the problem of transferability. In particular, the training dataset and the actual system variables have to share the same feature space [125], making it hard to transfer models to new plants. Moreover, older plants are prone to changing process conditions, which require adaptation in the models as well [126]. These challenges require novel techniques of transfer learning, which are currently at the starting point of their application in bioprocessing [125]. We believe that adaptation to new use-cases will be crucial for ML techniques to truly enhance process monitoring in biotechnology. Apart from the question of transferability, another choice for soft sensor development is quality versus quantity of data, which is not sufficiently investigated so far (see Outstanding questions).

As an example from chemical engineering, Li *et al*. [127] implemented fault detection for a continuous stirred tank reactor and a plant-wide pulp mill by means of deep and transfer learning. Using simulated instead of measured data to train an ML algorithm inevitably leads to model-process mismatch since no mechanistic model can perfectly describe the real process. To overcome this challenge and use simulated data to train their convolutional neural network (CNN) (Box 3), the authors applied transfer learning for domain adaptation, meaning that measured data from other processes is used as well to increase prediction performance. Similar approaches could also enhance process monitoring in biotechnology.

Towards controlling bioprocesses, ML can have a high impact regarding MPC (Box 3). In principle, MPC is a methodology that uses three components: a model to predict system outputs, an objective function, as well as a control law [128]. As an advantage, MPC can optimize performance of a system while considering constraints [129]. Here, ML is particularly useful for complex nonlinear systems or systems for which little process understanding exists [130].

Nagy [131] used a detailed, mechanistic process model of a yeast fermentation, including biomass, media concentration, and oxygen, to generate training data for an ANN. The ANN was then shown to efficiently replace the mechanistic model in MPC. As for other surrogate models, this approach has the advantage that ML models are often much faster to solve than their mechanistic counterparts, which is instrumental for MPC and other real-time applications. In another study, Masampally [132] implemented a cascade structure of GP regression submodels, which can predict biomass concentration in a fed-batch reactor; the cascaded model was also validated for process control in a closed-loop environment. Statistical process control (SPC) (Box 3) is well-established in various industries, working with control charts that indicate whether a process is running in-specification or not. As an example for application of ML, a long short-term memory network, which is a variation of RNNs, was used to learn which raw data correlated to important control chart patterns [133].

In addition to MPC and SPC, recent approaches make use of reinforcement learning [134–138]. Petsagkourakis *et al*. [134] applied a policy gradient algorithm (Box 3) to update a control policy in a batch-to-batch learning approach using true plant data. A surrogate model for the system was

used in two simulated case studies to avoid a large number of costly evaluations of the true system for training. In both case studies, the approach outperformed nonlinear MPC, thus posing an interesting starting point for further applications on real plants. In another study using reinforcement learning, Treloar *et al*. [137] applied different variants of the Q-learning algorithm (Box 3) to control microbial cocultures via two different auxotrophic nutrients (Figure 1D). Using data from a chemostat model, which was applied for five parallel reactors, the authors showed that a control policy could be learned within a 24-h experiment. For long sample-and-hold intervals, the strategy outperformed a classic PI controller, thus giving a promising outlook for future applications on industrial cocultures.

While ML is starting to enhance classical process monitoring and control, further applications such as predicting running production costs and attrition rates due to changing resources are still lacking at this stage. Transfer of knowledge, however, can save both time and costs and thus, de-risking can be achieved by reduced trial-and-error approaches [24]. Nonetheless, availability of high-quality data from a significant amount of bioprocess development campaigns is a major challenge since sharing such valuable data in public is not likely to happen. However, concepts such as FAIR [139,140], which are currently widely discussed and recommended, could foster greater availability of high-quality data.

Additionally, full process data is often only available for production runs within specification since a predictably failing run would mean high loss of resources. This results in an imbalance of datasets available for ML [141,142]. Hence, knowledge transfer models seem realistic only as a company-internal project because of very likely nondisclosure of corresponding, valuable data.

Opportunities for ML in bioprocess development

ML approaches in bioprocess development show promising results, especially in the areas of strain selection, bioprocess optimization, and control. For the first two areas, this is increasingly facilitated by the availability of high-quality data from a wide search space of possible process parameters, easily acquired from established HTS. For bioprocess scale-up, applications and ML models that span various scales are scarce. However, efficient models across scales would have a high impact, particularly by ensuring that HTS data at small scale are representative for industrial scale. Since the understanding of parameter variation during scale-up is still limited, data-driven ML methods are promising. To address these challenges in the future, transfer learning (i.e., reusing data from other conditions) and increasing well-annotated data should be focused on (see Outstanding questions).

Information content of data significantly differs, from highly diverse and broad information at small-scale strain selection to very specific process information at commercial-scale process control. As a consequence, we are convinced that no specific ML method can cover all areas, but rather emphasize the demand for an umbrella of ML methods that can be flexibly combined for each bioprocess. This is also reflected in Table 1, which shows that different areas of bioprocess development are covered unequally and that application of ML methods selectively clusters in certain areas. In particular, we found that ANNs are widely used across the bioprocess development fields and that reinforcement learning is dominantly applied in process control. The development of novel ML models in other fields, for example, physics-informed neural networks, which function as surrogate models for complex mechanistic process models, are currently expanding the available toolbox. Especially in the context of digital twins, which have the purpose to digitally mimic the real system, combining different ML models with mechanistic models for other process modules is promising. This, however, introduces more mathematical and computational challenges for bioprocess modeling, which still need to be tackled.

Nowadays, the biotechnological pipeline is often rather linear, meaning that a funnel exists from early-stage screening to process validation at larger scale, in which the design space is narrowed down after each stage [35]. As a consequence, design iterations are currently mostly taking place at individual stages, such as narrowing down the number of strains at small scale. Here, ML methods help to balance exploration (searching new parameter combinations) versus exploitation (suggesting the best parameters using the so-far available data) during iterative experimentation. Eventually, they provide processed and abstracted information, enabling the user to make smarter choices. Especially at commercial scale, decisions may result in significant cost and resource demands, so that human responsibility and accountability are so far not delegated to a decision-making algorithm.

Ultimately, however, current developments lay the foundation for integrative ML approaches covering more than one stage of the biotechnological pipeline. One example are so-called algorithmic idea generators (i.e., AI algorithms that provide suggestions for biologically and economically feasible bioproducts and robust processes from scratch). This, however, requires efficient technology transfer to bioprocess applications, availability of large, high-quality datasets (e.g., [143]) and the commitment to well-curated databases, including detailed metadata of successful as well as failed bioprocess development experiments. Revisiting the DBTL concept, the ideal approach would be an iterative cycle over the whole biotechnological pipeline, meaning that feedback loops between all stages are established. Ultimately, the funnel of the biotechnological pipeline would need to be replaced by early-stage iterations (e.g., between process validation at large scale and strain selection/bioprocess optimization at small scale). To realize this vision, ML and AI algorithms ultimately need to foster autonomous decision-making instead of replacing modules in the current linear pipeline. Here, we see potential to significantly improve efficiency in bioprocess development and even change the current mode of operation.

Concluding remarks and future perspectives

In this paper, we reviewed the advent of ML for bioprocess development, where ML methods are increasingly established as a standard in the data analysis toolbox. However, we see potential to transition from individual tools to frameworks that cover the whole process pipeline. At this point, committing to open-source methodology and databases is required for fast progress [144], meaning that a change in mindset is needed to make data and software publicly available. Indeed, this change is happening while corresponding impacts are considered thoroughly [145–149]. These advances will enable unleashing the variety of ML algorithms to better explore the rich amount of data that remains underexplored nowadays. We do think that a continuous transition towards ML-driven bioprocess development is happening in our discipline. Exciting times lie ahead, giving rise to a new generation of engineers and scientists who can make use of the vast amount of collected yet unanalyzed data, thus generating new strategies for bioprocess development. To both sides, machine learners and bioprocess engineers, we want to express the need for collaboration, expanded networks, and joint training to elevate bioprocess development to a new level.

Outstanding questions

In the current biotechnological DBTL pipeline, strain selection and optimization of physico-chemical parameters is often performed sequentially. These stages are, however, highly interdependent. Can ML induce a shift in the conventional, linear pipeline towards an integrative, circular approach, including feedback mechanisms between the two stages?

Can proof-of-concept ML studies that have been conducted for well-established bioprocesses be transferred to (non-model) production hosts and bioprocesses?

To realize the aforementioned goals, thorough data collection and annotation with metadata, corresponding data formats and databases, interfaces for automation, and other technical requirements need to be established. How can we realize and motivate this change? Are the FAIR data principles sufficient and how can this transition be implemented in the mindset of bioprocess engineers and data scientists?

Publishing and using negative results as training data can be game-changing for improved ML models in bioprocessing. How can we motivate and reward publication of negative results?

Which data is more suitable for ML-based soft sensors: sparse and highly informative data that is expensive to acquire (e.g., infrequent mass spectrometry data) versus big amounts of data that are less informative but can be retrieved more cheaply (e.g., online spectroscopy)?

Uncertainty quantification of ML results and interpretability of ML procedures are important for assessing model quality and revealing biological meaning/drawing meaningful conclusions. How can we foster the integration of uncertainty quantification in biotechnology to increase our process understanding?

Acknowledgment

The authors want to thank the “Microbial Bioprocess Lab — A Helmholtz Innovation Lab”, part of the Enabling Spaces Program “Helmholtz Innovation Labs” of the German Helmholtz Association, for financial support. This work was performed as part of the Helmholtz School for Data Science in Life, Earth and Energy (HDS-LEE).

Declaration of interests

No interests are declared.

References

1. Mitchell, T. *et al.* (1990) Machine learning. *Annu. Rev. Comput. Sci.* 4, 417–433
2. Ender, T.R. and Balestrini-Robinson, S. (2015) Surrogate modeling. In *Modeling and Simulation in the Systems Engineering Life Cycle* (Loper, M.L., ed.), pp. 201–216, Springer
3. Miller, T.H. *et al.* (2018) Machine learning for environmental toxicology: a call for integration and innovation. *Environ. Sci. Technol.* 52, 12953–12955
4. Bonetta, R. and Valentino, G. (2020) Machine learning techniques for protein function prediction. *Proteins* 88, 397–413
5. Reel, P.S. *et al.* (2021) Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol. Adv.* 49, 107739
6. Villoutreix, P. (2021) What machine learning can do for developmental biology. *Development* 148, dev188474
7. Muzio, G. *et al.* (2021) Biological network analysis with deep learning. *Brief. Bioinform.* 22, 1515–1530
8. Volk, M.J. *et al.* (2020) Biosystems design by machine learning. *ACS Synth. Biol.* 9, 1514–1533
9. Mowbray, M. *et al.* (2021) Machine learning for biochemical engineering: a review. *Biochem. Eng. J.* 172, 108054
10. Angermueller, C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.* 12, 878
11. Jumper, J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
12. Walters, W.P. and Barzilay, R. (2021) Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* 54, 263–270
13. Butler, K.T. *et al.* (2018) Machine learning for molecular and materials science. *Nature* 559, 547–555
14. Ding, Y. *et al.* (2021) Machine learning approaches for predicting biomolecule-disease associations. *Brief. Funct. Genomics* 20, 273–287
15. Graves, J. *et al.* (2020) A review of deep learning methods for antibodies. *Antibodies (Basel)* 9, 12
16. Leavell, M.D. *et al.* (2020) High-throughput screening for improved microbial cell factories, perspective and promise. *Curr. Opin. Biotechnol.* 62, 22–28
17. Silva, T.C. *et al.* (2021) Automation and miniaturization: enabling tools for fast, high-throughput process development in integrated continuous biomanufacturing. *J. Chem. Technol. Biotechnol.* 97, 2365–2375
18. Wasalathanthri, D.P. *et al.* (2021) Process analytics 4.0: a paradigm shift in rapid analytics for biologics development. *Biotechnol. Prog.* 37, e3177
19. Carbonell, P. *et al.* (2018) An automated design-build-test-learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol.* 1, 66–69
20. Opgenorth, P. *et al.* (2019) Lessons from two design-build-test-learn cycles of dodecanol production in *Escherichia coli* aided by machine learning. *ACS Synth. Biol.* 8, 1337–1351
21. Liao, X. *et al.* (2022) Artificial intelligence: a solution to involution of design-build-test-learn cycle. *Curr. Opin. Biotechnol.* 75, 102712
22. Dickens, J. *et al.* (2018) Biopharmaceutical raw material variation and control. *Curr. Opin. Chem. Eng.* 22, 236–243
23. Jordan, M. *et al.* (2018) Intensification of large-scale cell culture processes. *Curr. Opin. Chem. Eng.* 22, 253–257
24. von Stosch, M. *et al.* (2021) A roadmap to AI-driven in silico process development: bioprocessing 4.0 in practice. *Curr. Opin. Chem. Eng.* 33, 100692
25. Artico, F. *et al.* (2022) The future of artificial intelligence for the BioTech big data landscape. *Curr. Opin. Biotechnol.* 76, 102714
26. Joshi, V.S. *et al.* (2017) Optimization of ion exchange sigmoidal gradients using hybrid models: implementation of quality by design in analytical method development. *J. Chromatogr. A* 1491, 145–152
27. Wang, G. *et al.* (2017) Root cause investigation of deviations in protein chromatography based on mechanistic models and artificial neural networks. *J. Chromatogr. A* 1515, 146–153
28. Brestrich, N. *et al.* (2018) Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression. *Chem. Eng. Sci.* 176, 157–164
29. Risum, A.B. and Bro, R. (2019) Using deep learning to evaluate peaks in chromatographic data. *Talanta* 204, 255–260
30. Kensert, A. *et al.* (2021) Deep Q-learning for the selection of optimal isocratic scouting runs in liquid chromatography. *J. Chromatogr. A* 1638, 461900
31. Vaskevicius, M. *et al.* (2021) Prediction of chromatography conditions for purification in organic synthesis using deep learning. *Molecules* 26, 2474
32. Liu, S. and Papageorgiou, L.G. (2019) Optimal antibody purification strategies using data-driven models. *Engineering* 5, 1077–1092
33. Walther, C. *et al.* (2022) Smart process development: application of machine-learning and integrated process modeling for inclusion body purification processes. *Biotechnol. Prog.* 38, e3249
34. Wehrs, M. *et al.* (2020) You get what you screen for: on the value of fermentation characterization in high-throughput strain improvements in industrial settings. *J. Ind. Microbiol. Biotechnol.* 47, 913–927
35. Hemmerich, J. *et al.* (2018) Microbioreactor systems for accelerated bioprocess development. *Biotechnol. J.* 13, e1700141
36. Grav, L.M. *et al.* (2018) Minimizing clonal variation during mammalian cell line engineering for improved systems biology data generation. *ACS Synth. Biol.* 7, 2148–2159
37. McKinley, K.L. and Cheeseman, I.M. (2017) Large-scale analysis of CRISPR/Cas9 cell-cycle knockouts reveals the diversity of p53-dependent responses to cell-cycle defects. *Dev. Cell* 40, 405–420
38. Mazurenko, S. *et al.* (2019) Machine learning in enzyme engineering. *ACS Catal.* 10, 1210–1223
39. Siedhoff, N.E. *et al.* (2020) Machine learning-assisted enzyme engineering. *Methods Enzymol.* 643, 281–315
40. Gu, C. *et al.* (2019) Current status and applications of genome-scale metabolic models. *Genome Biol.* 20, 121
41. Srinivasan, S. *et al.* (2015) Constructing kinetic models of metabolism at genome-scales: a review. *Biotechnol. J.* 10, 1345–1359
42. Almquist, J. *et al.* (2014) Kinetic models in industrial biotechnology - improving cell factory performance. *Metab. Eng.* 24, 38–60
43. Stalidzans, E. *et al.* (2018) Model-based metabolism design: constraints for kinetic and stoichiometric models. *Biochem. Soc. Trans.* 46, 261–267
44. Heirendt, L. *et al.* (2019) Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702
45. Oyetunde, T. *et al.* (2018) Leveraging knowledge engineering and machine learning for microbial bio-manufacturing. *Biotechnol. Adv.* 36, 1308–1315
46. Orth, J.D. *et al.* (2010) What is flux balance analysis? *Nat. Biotechnol.* 28, 245–248
47. Segre, D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15112–15117
48. Schneider, P. *et al.* (2020) An extended and generalized framework for the calculation of metabolic intervention strategies based on minimal cut sets. *PLoS Comput. Biol.* 16, e1008110
49. Mishra, B. *et al.* (2019) Systems biology and machine learning in plant-pathogen interactions. *Mol. Plant-Microbe Interact.* 32, 45–55
50. Rana, P. *et al.* (2020) Recent advances on constraint-based models by integrating machine learning. *Curr. Opin. Biotechnol.* 64, 85–91
51. King, Z.A. *et al.* (2017) Literature mining supports a next-generation modeling approach to predict cellular byproduct secretion. *Metab. Eng.* 39, 220–227
52. Oyetunde, T. *et al.* (2019) Machine learning framework for assessment of microbial factory performance. *PLoS One* 14, e0210558
53. Zhang, J. *et al.* (2020) Combining mechanistic and machine learning models for predictive engineering and optimization of tryptophan metabolism. *Nat. Commun.* 11, 4880

54. Radivojevic, T. *et al.* (2020) A machine learning automated recommendation tool for synthetic biology. *Nat. Commun.* 11, 4879
55. Pedregosa, F. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830
56. Carbonell, P. *et al.* (2019) Opportunities at the intersection of synthetic biology, machine learning, and automation. *ACS Synth. Biol.* 8, 1474–1477
57. Faure, L. *et al.* (2022) Artificial metabolic networks: enabling neural computation with metabolic networks. *bioRxiv* Published online January 11, 2022. https://doi.org/10.1101/2022.01.09.475487
58. Zampieri, G. *et al.* (2019) Machine and deep learning meet genome-scale metabolic modeling. *PLoS Comput. Biol.* 15, e1007084
59. Antonakoudis, A. *et al.* (2020) The era of big data: genome-scale modelling meets machine learning. *Comput. Struct. Biotechnol. J.* 18, 3287–3300
60. van Rosmalen, R.P. *et al.* (2021) Model reduction of genome-scale metabolic models as a basis for targeted kinetic models. *Metab. Eng.* 64, 74–84
61. Choudhury, S. *et al.* (2022) Reconstructing kinetic models for dynamical studies of metabolism using generative adversarial networks. *Nat. Mach. Intell.* 4, 710–719
62. Sabzevari, M. *et al.* (2022) Strain design optimization using reinforcement learning. *PLoS Comput. Biol.* 18, e1010177
63. Wu, S.G. *et al.* (2016) Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming. *PLoS Comput. Biol.* 12, e1004838
64. Li, G. *et al.* (2019) Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. *ACS Synth. Biol.* 8, 1411–1420
65. Bradford, E. *et al.* (2018) Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes. *Comput. Chem. Eng.* 118, 143–158
66. Vega-Ramon, F. *et al.* (2021) Kinetic and hybrid modeling for yeast astaxanthin production under uncertainty. *Biotechnol. Bioeng.* 118, 4854–4866
67. Freier, L. *et al.* (2016) Framework for Kriging-based iterative experimental analysis and design: optimization of secretory protein production in *Corynebacterium glutamicum*. *Eng. Life Sci.* 16, 538–549
68. Zheng, Z.Y. *et al.* (2017) Artificial neural network - genetic algorithm to optimize wheat germ fermentation condition: application to the production of two anti-tumor benzoquinones. *Food Chem.* 227, 264–270
69. del Rio-Chanona, E.A. *et al.* (2016) Dynamic modeling and optimization of cyanobacterial C-phycocyanin production process by artificial neural network. *Algal Res.* 13, 7–15
70. Pappu, S.M.J. and Gummadi, S.N. (2017) Artificial neural network and regression coupled genetic algorithm to optimize parameters for enhanced xylitol production by *Debaryomyces nepalensis* in bioreactor. *Biochem. Eng. J.* 120, 136–145
71. Ebrahimpour, A. *et al.* (2008) A modeling study by response surface methodology and artificial neural network on culture parameters optimization for thermostable lipase production from a newly isolated thermophilic *Geobacillus* sp. strain ARM. *BMC Biotechnol.* 8, 96
72. Sebayang, A.H. *et al.* (2017) Optimization of bioethanol production from sorghum grains using artificial neural networks integrated with ant colony. *Ind. Crop. Prod.* 97, 146–155
73. Rodriguez-Granrose, D. *et al.* (2021) Design of experiment (DOE) applied to artificial neural network architecture enables rapid bioprocess improvement. *Bioprocess Biosyst. Eng.* 44, 1301–1308
74. Rogers, A.W. *et al.* (2022) A transfer learning approach for predictive modeling of bioprocesses using small data. *Biotechnol. Bioeng.* 119, 411–422
75. Hutter, C. *et al.* (2021) Knowledge transfer across cell lines using hybrid Gaussian process models with entity embedding vectors. *Biotechnol. Bioeng.* 118, 4389–4401
76. Wang, Y. *et al.* (2018) A comparison of word embeddings for the biomedical natural language processing. *J. Biomed. Inform.* 87, 12–20
77. Bluma, A. *et al.* (2010) In-situ imaging sensors for bioprocess monitoring: state of the art. *Anal. Bioanal. Chem.* 398, 2429–2438
78. Marba-Ardebol, A.M. *et al.* (2019) In situ microscopy for real-time determination of single-cell morphology in bioprocesses. *J. Vis. Exp.* Published online December 5, 2019. https://doi.org/10.3791/57823
79. Grunberger, A. *et al.* (2014) Single-cell microfluidics: opportunity for bioprocess development. *Curr. Opin. Biotechnol.* 29, 15–23
80. Du, G. *et al.* (2016) Microfluidics for cell-based high throughput screening platforms - a review. *Anal. Chim. Acta* 903, 36–50
81. Riordon, J. *et al.* (2019) Deep learning with microfluidics for biotechnology. *Trends Biotechnol.* 37, 310–324
82. Galan, E.A. *et al.* (2020) Intelligent microfluidics: the convergence of machine learning and microfluidics in materials science and biomedicine. *Matter* 3, 1893–1922
83. Stallmann, D. *et al.* (2021) Towards an automatic analysis of CHO-K1 suspension growth in microfluidic single-cell cultivation. *Bioinformatics* 37, 3632–3639
84. O'Connor, O.M. *et al.* (2022) DeLTA 2.0: A deep learning pipeline for quantifying single-cell spatial and temporal dynamics. *PLoS Comput. Biol.* 18, e1009797
85. Lashkaripour, A. *et al.* (2021) Machine learning enables design automation of microfluidic flow-focusing droplet generation. *Nat. Commun.* 12, 25
86. Hartmann, R. *et al.* (2020) BiofilmQ, a software tool for quantiative image analysis of microbial biofilm communities. *Nat. Microbiol.* 6, 151–156
87. Long, B. *et al.* (2022) Machine learning-informed and synthetic biology-enabled semi-continuous algal cultivation to unleash renewable fuel productivity. *Nat. Commun.* 13, 541
88. Lee, C.-G. (1999) Calculation of light penetration depth in photobioreactors. *Biotechnol. Bioprocess Eng.* 4, 78–81
89. Wang, J. *et al.* (2015) The difference in effective light penetration may explain the superiority in photosynthetic efficiency of attached cultivation over the conventional open pond for microalgae. *Biotechnol. Biofuels* 8, 49
90. Göttl, Q. *et al.* (2021) Automated flowsheet synthesis using hierarchical reinforcement learning: proof of concept. *Chem. Ing. Tech.* 93, 2010–2018
91. Stops, L. *et al.* (2022) Flowsheet synthesis through hierarchical reinforcement learning and graph neural networks. *arXiv* Published online July 25, 2022. https://doi.org/10.48550/arXiv.2207.12051
92. Takors, R. (2012) Scale-up of microbial processes: impacts, tools and open questions. *J. Biotechnol.* 160, 3–9
93. Neubauer, P. and Junne, S. (2016) Scale-up and scale-down methodologies for bioreactors. In *Bioreactors* (Mandenius, C.-F., ed.), pp. 323–354, Wiley-VCH Verlag
94. Delvigne, F. *et al.* (2017) Bioprocess scale-up/down as integrative enabling technology: from fluid mechanics to systems biology and beyond. *Microb. Biotechnol.* 10, 1267–1274
95. Wang, G. *et al.* (2018) Comparative performance of different scale-down simulators of substrate gradients in *Penicillium chrysogenum* cultures: the need of a biological systems response analysis. *Microb. Biotechnol.* 11, 486–497
96. Karimi Alavijeh, M. *et al.* (2022) Digitally enabled approaches for the scale up of mammalian cell bioreactors. *Chem. Eng. Technol.* 4, 100040
97. Le, H. *et al.* (2012) Multivariate analysis of cell culture bioprocess data–lactate consumption as process indicator. *J. Biotechnol.* 162, 210–223
98. Facco, P. *et al.* (2020) Using data analytics to accelerate biopharmaceutical process scale-up. *Biochem. Eng. J.* 164, 107791
99. Smiatek, J. *et al.* (2021) Generic and specific recurrent neural network models: applications for large and small scale biopharmaceutical upstream processes. *Biotechnol. Rep. (Amst.)* 31, e00640
100. Sokolov, M. *et al.* (2018) Sequential multivariate cell culture modeling at multiple scales supports systematic shaping of a monoclonal antibody toward a quality target. *Biotechnol. J.* 13, e1700461

101. Bayer, B. *et al.* (2021) Model transferability and reduced experimental burden in cell culture process development facilitated by hybrid modeling and intensified design of experiments. *Front. Bioeng. Biotechnol.* 9, 740215
102. Cai, S. *et al.* (2022) Physics-informed neural networks (PINNs) for fluid mechanics: a review. *Acta Mech. Sinica* 37, 1727–1738
103. Mowbray, M. *et al.* (2022) Industrial data science – a review of machine learning applications for chemical and process industries. *React. Chem. Eng.* 7, 1471–1509
104. Luttmann, R. *et al.* (2012) Soft sensors in bioprocessing: a status report and recommendations. *Biotechnol. J.* 7, 1040–1048
105. Gerzon, G. *et al.* (2022) Process analytical technologies - advances in bioprocess integration and future perspectives. *J. Pharm. Biomed. Anal.* 207, 114379
106. Narayanan, H. *et al.* (2020) Bioprocessing in the digital age: the role of process models. *Biotechnol. J.* 15, e1900172
107. Kadlec, P. *et al.* (2009) Data-driven soft sensors in the process industry. *Comput. Chem. Eng.* 33, 795–814
108. Desai, K. *et al.* (2006) Soft-sensor development for fed-batch bioreactors using support vector regression. *Biochem. Eng. J.* 27, 225–239
109. Fortuna, L. *et al.,* eds (2007) *Soft Sensors for Monitoring and Control of Industrial Processes*, Springer
110. Randek, J. and Mandenius, C.F. (2018) On-line soft sensing in upstream bioprocessing. *Crit. Rev. Biotechnol.* 38, 106–121
111. Zhu, X. *et al.* (2020) Modern soft-sensing modeling methods for fermentation processes. *Sensors (Basel)* 20, 1771
112. Schmidt, A. *et al.* (2022) Process analytical technology as key-enabler for digital twins in continuous biomanufacturing. *J. Chem. Technol. Biotechnol.* 97, 2336–2346
113. Chen, Y. *et al.* (2020) Digital twins in pharmaceutical and biopharmaceutical manufacturing: a literature review. *Processes* 8, 1088
114. Hartmann, F.S.F. *et al.* (2022) Digital models in biotechnology: towards multi-scale integration and implementation. *Biotechnol. Adv.* 60, 108015
115. Portela, R.M.C. *et al.* (2021) When is an in silico representation a digital twin? A biopharmaceutical industry approach to the digital twin concept. *Adv. Biochem. Eng. Biotechnol.* 176, 35–55
116. Zobel-Roos, S. *et al.* (2021) Digital Twins in Biomanufacturing. *Adv. Biochem. Eng. Biotechnol.* 176, 181–262
117. Sun, Q. and Ge, Z. (2021) A survey on deep learning for data-driven soft sensors. *IEEE Trans. Industr. Inform.* 17, 5853–5866
118. Dai, X. *et al.* (2006) "Assumed inherent sensor" inversion based ANN dynamic soft-sensing method and its application in erythromycin fermentation process. *Comput. Chem. Eng.* 30, 1203–1225
119. Albiol, J. *et al.* (1995) Biomass estimation in plant cell cultures: a neural network approach. *Biotechnol. Prog.* 11, 88–92
120. Wang, B. *et al.* (2020) Soft-sensor modeling for L-lysine fermentation process based on hybrid ICS-MLSSVM. *Sci. Rep.* 10, 11630
121. Graziani, S. and Xibilia, M.G. (2020) Deep learning for soft sensor design. In *Development and Analysis of Deep Learning Architectures* (Pedrycz, W. and Chen, S.-M., eds), pp. 31–59, Springer
122. Gopakumar, V. *et al.* (2018) A deep learning based data driven soft sensor for bioprocesses. *Biochem. Eng. J.* 136, 28–39
123. Yao, L. and Ge, Z. (2018) Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. *IEEE Trans. Ind. Electron.* 65, 1490–1498
124. Mowbray, M. *et al.* (2022) Probabilistic machine learning based soft-sensors for product quality prediction in batch processes. *Chemom. Intell. Lab. Syst.* 228, 104616
125. Curreri, F. *et al.* (2021) Soft sensor transferability: a survey. *Appl. Sci.* 11, 7710
126. Kadlec, P. *et al.* (2011) Review of adaptation mechanisms for data-driven soft sensors. *Comput. Chem. Eng.* 35, 1–24
127. Li, W. *et al.* (2020) Transfer learning for process fault diagnosis: knowledge transfer from simulation to physical processes. *Comput. Chem. Eng.* 139, 106904
128. Camacho, E.F. and Alba, C.B. (2013) *Model Predictive Control*, Springer
129. Hewing, L. *et al.* (2020) Learning-based model predictive control: toward safe learning in control. *Annu. Rev. Control Robot. Auton. Syst.* 3, 269–296
130. Chee, E. *et al.* (2021) An integrated approach for machine-learning-based system identification of dynamical systems under control: application towards the model predictive control of a highly nonlinear reactor system. *Front. Chem. Sci. Eng.* 16, 237–250
131. Nagy, Z.K. (2007) Model based control of a yeast fermentation bioreactor using optimally designed artificial neural networks. *Chem. Eng. J.* 127, 95–109
132. Masampally, V.S. *et al.* (2018) Cascade Gaussian Process Regression Framework for Biomass Prediction in a Fed-batch Reactor. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*
133. Zan, T. *et al.* (2019) Statistical process control with intelligence based on the deep learning model. *Appl. Sci.* 10, 308
134. Petsagkourakis, P. *et al.* (2020) Reinforcement learning for batch bioprocess optimization. *Comput. Chem. Eng.* 133, 106649
135. Xie, H. *et al.* (2020) Model Predictive Control Guided Reinforcement Learning Control Scheme. In *2020 International Joint Conference on Neural Networks (IJCNN)*
136. Hedrick, E. *et al.* (2022) Reinforcement learning for online adaptation of model predictive controllers: application to a selective catalytic reduction unit. *Comput. Chem. Eng.* 160, 107727
137. Treloar, N.J. *et al.* (2020) Deep reinforcement learning for the control of microbial co-cultures in bioreactors. *PLoS Comput. Biol.* 16, e1007783
138. Oh, T.H. *et al.* (2022) Integration of reinforcement learning and model predictive control to optimize semi-batch bioreactor. *AIChE J.* 68, 6
139. Rehnert, M. and Takors, R. (2022) FAIR research data management as community approach in bioengineering. *Eng. Life Sci.* Published online April 28, 2022. https://doi.org/10.1002/elsc.202200005
140. Wilkinson, M.D. *et al.* (2016) The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018
141. Farid, S.S. *et al.* (2020) Benchmarking biopharmaceutical process development and manufacturing cost contributions to R&D. *MAbs* 12, 1754999
142. Faulon, J.L. and Faure, L. (2021) In silico, in vitro, and in vivo machine learning in synthetic biology and metabolic engineering. *Curr. Opin. Chem. Biol.* 65, 85–92
143. O'Brien, C.M. *et al.* (2021) A hybrid mechanistic-empirical model for *in silico* mammalian cell bioprocess simulation. *Metab. Eng.* 66, 31–40
144. Udaondo, Z. (2022) Big data and computational advancements for next generation of microbial biotechnology. *Microb. Biotechnol.* 15, 107–109
145. Giovani, B. (2017) Open data for research and strategic monitoring in the pharmaceutical and biotech industry. *Data Sci. J.* 16, 18
146. Gitter, D.M. (2008) Resolving the open source paradox in biotechnology: a proposal for a revised open source policy for publicly funded genomic databases. *Comput. Law Secur. Rev.* 24, 529–539
147. Sayers, E.W. *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 50, D20–D26
148. Oliveira, A.L. (2019) Biotechnology, big data and artificial intelligence. *Biotechnol. J.* 14, e1800613
149. Harrow, J. *et al.* (2021) ELIXIR-EXCELERATE: establishing Europe's data infrastructure for the life science research of the future. *EMBO J.* 40, e107409
150. Kok, J.N. and Unesco (2009) *Artificial Intelligence*, Eolss Publishers Company
151. Alpaydin, E. (2020) *Introduction to Machine Learning* (4th edn), MIT Press
152. Buchanan, B.G. and Smith, R.G. (1988) Fundamentals of expert systems. *Annu. Rev. Comput. Sci.* 3, 23–58
153. Cunningham, P. *et al.* (2008) Supervised learning. In *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval* (Cord, M. and Cunningham, P., eds), pp. 21–49, Springer

154. Ghahramani, Z. (2004) Unsupervised learning. In *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (Bousquet, O. *et al.*, eds), pp. 72–112, Springer
155. Kaelbling, L.P. *et al.* (1996) Reinforcement learning: a survey. *J. Artif. Intell. Res.* 4, 237–285
156. Sutton, R.S. (1992) Introduction: the challenge of reinforcement learning. In *Reinforcement Learning* (Sutton, R.S., ed.), pp. 1–3, Springer
157. Weiss, K. *et al.* (2016) A survey of transfer learning. *J. Big Data* 3, 9
158. Hua, J. *et al.* (2021) Learning for a robot: deep reinforcement learning, imitation learning, transfer learning. *Sensors (Basel)* 21, 1278
159. Mahmud, M. *et al.* (2018) Applications of deep learning and reinforcement learning to biological data. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 2063–2079
160. Voulodimos, A. *et al.* (2018) Deep learning for computer vision: a brief review. *Comput. Intell. Neurosci.* 2018, 7068349
161. Ching, T. *et al.* (2018) Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15, 20170387
162. Bennett, D. *et al.* (2021) Value-free reinforcement learning: policy optimization as a minimal model of operant behavior. *Curr. Opin. Behav. Sci.* 41, 114–121
163. Zhou, Z.-H. (2021) Ensemble learning. In *Mach Learn*, pp. 181–210, Springer
164. Lawson, C.E. *et al.* (2021) Machine learning for metabolic engineering: a review. *Metab. Eng.* 63, 34–60
165. Greener, J.G. *et al.* (2022) A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 23, 40–55
166. Wang, S.-C. (2003) Artificial neural network. In *Interdisciplinary Computing in Java Programming*, pp. 81–100, Springer
167. Dhruv, P. and Naskar, S. (2020) *Image Classification Using Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN): A Review*, Springer
168. Gu, J. *et al.* (2018) Recent advances in convolutional neural networks. *Pattern Recogn.* 77, 354–377
169. Izmailov, P. *et al.* (2021) What are Bayesian neural network posteriors really like? In *Proceedings of the 38th International Conference on Machine Learning* (Marina, M. and Tong, Z., eds), ICML
170. Goodfellow, I. *et al.* (2016) *Deep Learning*, MIT Press
171. Connor, M. *et al.* (2021) Variational autoencoder with learned latent structure. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (Arindam, B. and Kenji, F., eds), AAAI Press
172. Basu, S. *et al.* (2018) Iterative random forests to discover predictive and stable high-order interactions. *Proc. Natl. Acad. Sci. U. S. A.* 115, 1943–1948
173. Noble, W.S. (2006) What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567
174. Yang, Z.R. (2004) Biological applications of support vector machines. *Brief. Bioinform.* 5, 328–338
175. di Sciascio, F. and Amicarelli, A.N. (2008) Biomass estimation in batch biotechnological processes by Bayesian Gaussian process regression. *Comput. Chem. Eng.* 32, 3264–3273
176. Lan, L. *et al.* (2020) Generative adversarial networks and its applications in biomedical informatics. *Front. Public Health* 8, 164
177. Jiao, Q. and Zhang, S. (2021) A brief survey of word embedding and its recent development. In *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference*, IAEAC
178. Bengio, S. *et al.* (2009) Group Sparse Coding. In *Advances in Neural Information Processing Systems* (22) (Bengio, Y. *et al.*, eds), pp. 82–89, Curran Associates, Inc.https://proceedings.neurips.cc/paper/2009/file/3b3dbaf68507998acd6a5a5254ab2d76-Paper.pdf
179. Watkins, C.J.C.H. and Dayan, P. (1992) Q-learning. *Mach. Learn.* 8, 279–292
180. Schwenzer, M. *et al.* (2021) Review on model predictive control: an engineering perspective. *Int. J. Adv. Manuf. Technol.* 117, 1327–1349
181. Altman, R.B. *et al.* (2008) Text mining for biology–the way forward: opinions from leading scientists. *Genome Biol.* 9, S7
182. Jensen, L.J. *et al.* (2006) Literature mining for the biologist: from information retrieval to biological discovery. *Nat. Rev. Genet.* 7, 119–129
183. Pinto, J. *et al.* (2022) A general deep hybrid model for bioreactor systems: combining first principles with deep neural networks. *Comput. Chem. Eng.* 165, 107952
184. Nelofer, R. *et al.* (2012) Comparison of the estimation capabilities of response surface methodology and artificial neural network for the optimization of recombinant lipase production by *E. coli* BL21. *J. Ind. Microbiol. Biotechnol.* 39, 243–254
185. Wang, Y. *et al.* (2020) Optimization of dark fermentation for biohydrogen production using a hybrid artificial neural network (ANN) and response surface methodology (RSM) approach. *Environ. Prog. Sustain. Energy* 40, 2
186. Unni, S. *et al.* (2019) Artificial neural network-genetic algorithm (ANN-GA) based medium optimization for the production of human interferon gamma (hIFN-γ) in *Kluyveromyces lactis* cell factory. *Can. J. Chem. Eng.* 97, 843–858
187. Tavasoli, T. *et al.* (2019) A robust feeding control strategy adjusted and optimized by a neural network for enhancing of alpha 1-antitrypsin production in *Pichia pastoris*. *Biochem. Eng. J.* 144, 18–27
188. Zhang, L. *et al.* (2020) Modeling and optimization of microbial lipid fermentation from cellulosic ethanol wastewater by *Rhodotorula glutinis* based on the support vector machine. *Bioresour. Technol.* 301, 122781
189. Dong, C. and Chen, J. (2019) Optimization of process parameters for anaerobic fermentation of corn stalk based on least squares support vector machine. *Bioresour. Technol.* 271, 174–181
190. Kennedy, M.J. and Spooner, N.R. (1996) Using fuzzy logic to design fermentation media: a comparison to neural networks and factorial design. *Biotechnol. Tech.* 10, 47–52
191. Brunner, M. *et al.* (2017) Investigation of the interactions of critical scale-up parameters (pH, pO2 and pCO2) on CHO batch performance and critical quality attributes. *Bioprocess Biosyst. Eng.* 40, 251–263
192. Holubar, P. (2002) Advanced controlling of anaerobic digestion by means of hierarchical neural networks. *Water Res.* 36, 2582–2588
193. Glassey, J. *et al.* (1994) Enhanced supervision of recombinant *E. coli* fermentation via artificial neural networks. *Process Biochem.* 29, 387–398
194. Shokry, A. *et al.* (2018) Data-driven soft-sensors for online monitoring of batch processes with different initial conditions. *Comput. Chem. Eng.* 118, 159–179
195. Wong, W. *et al.* (2018) Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing. *Math* 6, 6110242
196. Barberi, G. *et al.* (2021) Anticipated cell lines selection in bioprocess scale-up through machine learning on metabolomics dynamics. *IFAC-PapersOnLine* 54, 85–90
197. Poth, M. *et al.* (2022) Extensive evaluation of machine learning models and data preprocessings for Raman modeling in bioprocessing. *J. Raman Spectrosc.* 53, 1580–1591
198. Hassan, S. *et al.* (2013) Bioprocess data mining using regularized regression and random forests. *BMC Syst. Biol.* 7, S5
199. Shrivastava, R. *et al.* (2017) Application and evaluation of random forest classifier technique for fault detection in bioreactor operation. *Chem. Eng. Commun.* 204, 591–598
200. Probst, D. *et al.* (2022) Biocatalysed synthesis planning using data-driven learning. *Nat. Commun.* 13, 964
201. Kotidis, P. and Kontoravdi, C. (2020) Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab. Eng. Commun.* 10, e00131
202. Nikita, S. *et al.* (2021) Reinforcement learning based optimization of process chromatography for continuous processing of biopharmaceuticals. *Chem. Eng. Sci.* 230, 116171
203. Pan, E. *et al.* (2021) Constrained Q-learning for batch process optimization. *IFAC-PapersOnLine* 54, 492–497
204. Heidari Baladehi, M. *et al.* (2021) Culture-free identification and metabolic profiling of microalgal single cells via ensemble learning of ramanomes. *Anal. Chem.* 93, 8872–8880

205. Czajka, J.J. *et al.* (2021) Integrated knowledge mining, genome-scale modeling, and machine learning for predicting *Yarrowia lipolytica* bioproduction. *Metab. Eng.* 67, 227–236
206. Mowbray, M. *et al.* (2022) Ensemble learning for bioprocess dynamic modelling and prediction. *Biotech. Bioeng.* Published online May 4, 2020. https://doi.org/10.22541/au.158456506.69710259
207. Liu, Y. and Gunawan, R. (2017) Bioprocess optimization under uncertainty using ensemble modeling. *J. Biotechnol.* 244, 34–44
208. Pinto, J. *et al.* (2019) A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development. *Bioprocess Biosyst. Eng.* 42, 1853–1865
209. Tokuyama, K. *et al.* (2020) Data science-based modeling of the lysine fermentation process. *J. Biosci. Bioeng.* 130, 409–415
210. Agarwal, A. *et al.* (2019) 110th Anniversary: ensemble-based machine learning for industrial fermenter classification and foaming control. *Ind. Eng. Chem. Res.* 58, 16719–16729
211. Mante, J. *et al.* (2019) A heuristic approach to handling missing data in biologics manufacturing databases. *Bioprocess Biosyst. Eng.* 42, 657–663
212. Zhang, T. *et al.* (2019) Pattern recognition in chemical process flowsheets. *AIChE J.* 65, 592–603
213. Coşgun, A. *et al.* (2022) Analysis of lipid production from *Yarrowia lipolytica* for renewable fuel production by machine learning. *Fuel* 315, 122817
214. Resendis-Antonio, O. (2013) Constraint-based modeling. In *Encyclopedia of Systems Biology* (Dubitzky, W. *et al.*, eds), pp. 494–498, Springer
215. Kumar, V. *et al.* (2014) Design of experiments applications in bioprocessing: concepts and approach. *Biotechnol. Prog.* 30, 86–99
216. von Stosch, M. and Willis, M.J. (2017) Intensified design of experiments for upstream bioreactors. *Eng. Life Sci.* 17, 1173–1184
217. Garetti, M. *et al.* (2012) Life cycle simulation for the design of product–service systems. *Comput. Ind.* 63, 361–369
218. Chowdhary, K.R. (2020) Natural language processing. In *Fundamentals of Artificial Intelligence*, pp. 603–649, Springer
219. Hirschberg, J. and Manning, C.D. (2015) Advances in natural language processing. *Science* 349, 261–266