Biochemical Engineering Journal 172 (2021) 108054

Contents lists available at ScienceDirect

Biochemical Engineering Journal

journal homepage: www.elsevier.com/locate/bej

# Machine learning for biochemical engineering: A review

Max Mowbray [a,1], Thomas Savage [a,b,1], Chufan Wu [a], Ziqi Song [a], Bovinille Anye Cho [a], Ehecatl A. Del Rio-Chanona [c], Dongda Zhang [a,c,*]

[a] Department of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester, M1 3BU, UK
[b] University of Cambridge, Department of Chemical Engineering and Biotechnology, Philippa Fawcett Drive, Cambridge, CB3 0AS, UK
[c] Department of Chemical Engineering, Imperial College London, South Kensington, SW7 2AZ, UK

ARTICLE INFO

*Keywords:*
Machine learning
Data-driven modelling
Biochemical engineering
Industrial biotechnology
Digitalisation
Digital twin

ABSTRACT

The field of machine learning is comprised of techniques, which have proven powerful approaches to knowledge discovery and construction of 'digital twins' in the highly dimensional, nonlinear and stochastic domains common to biochemical engineering. We review the use of machine learning within biochemical engineering over the last 20 years. The most prevalent machine learning methods are demystified, and their impact across individual biochemical engineering subfields is outlined. In doing so we provide insights into the true benefits of each technique, and obstacles for their wider deployment. Finally, core challenges into the application of machine learning in biochemical engineering are thoroughly discussed, and further insight into adoption of innovative hybrid modelling and transfer learning strategies for development of new digital biotechnologies is provided.

## 1. Introduction

Biochemical engineering is a research area that closely links with several global research themes and strategic national priorities such as energy, sustainability, and healthcare. It covers all aspects of the development of biological processes ranging from raw materials preparation to bioproduct synthesis and biowaste valorisation, with diverse applications primarily relevant to industrial biotechnology, environmental biotechnology, and medical and healthcare biotechnology. Biochemical engineering seamlessly integrates both natural science subjects such as chemistry and biology with engineering subfields such as chemical engineering and process systems engineering. Given the breadth of this research area, biochemical engineering has been divided by the scientific community into different fields such as biocatalysis, biomaterials, systems and synthetic biology, bioreactor engineering, tissue and protein engineering, metabolic engineering, and environmental biotechnology, to name a few.

In particular, the word 'engineering' indicates that biochemical engineering research is not only concerned with natural science discovery, but is also focused on the engineering aspect, which aims to translate knowledge from science to practical industrial implementations. Stemming from the domain of chemical engineering, one of the earliest engineering sectors and a product of the first industrial revolution, biochemical engineering has enjoyed unprecedented development and had a tremendous impact on modern industry and society. With the UK's biochemical engineering sector contributing to over 5 million jobs, the UK's Bioeconomy is currently worth £220 billion and is expected to double by 2030 [1]. Since 2011, it has been widely accepted that current industry is transitioning into the 4th Industrial Revolution. This conceptual framework for industry was first proposed by the German government and aims to computerise and automate manufacturing processes. Therefore, digitalising biochemical systems and constructing industrially focused mathematical modelling tools have become key drivers for current biochemical engineering research and innovation.

However, a number of severe challenges have obstructed the realisation of this task. Traditionally, mathematical models are constructed based on first-principle knowledge. At this moment, there have been many mechanistic and physics-based empirical models developed to simulate the kinetics of microbial biomass growth and product synthesis [2,3]; to estimate intracellular metabolic flux [4]; to calculate fluid dynamics and thermodynamics for bioreactor scale-up and bioseparation unit design [5]; and to design new proteins and enzymes [6]. Justified by these successes, continuous development of these mechanistic models will last for decades to come due to the necessity of in-depth human understanding of the underlying system. Similarly, the pharmaceutical and fermentation industries still use the conventional

* Corresponding author at: Department of Chemical Engineering and Analytical Science, The University of Manchester, Oxford Road, Manchester, M1 3BU, UK.
*E-mail address:* dongda.zhang@manchester.ac.uk (D. Zhang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.bej.2021.108054
Received 7 February 2021; Received in revised form 8 April 2021; Accepted 5 May 2021
Available online 7 May 2021

'scale-up' approach, which consumes substantial time and labour resources to translate a technology from lab to industrial operations [7]. Nonetheless, it is understood that within the 4th Industrial Revolution, more time efficient approaches such as the 'scale-down' simulation approach must be adopted to save time and resources for technology development, and experiments conducted at lab/bench-scale must be well designed such that they can be used to accurately predict industrial-scale performance [8].

Most importantly, distinct from the petrochemical industry, which was initially driven by a single economic index, the bloom of biochemical industry is predominantly driven by the more pressing demand for economic and environmental sustainability. As a result, the creation and selection of industrial biotechnology will be determined by more diverse metrics. This directly results in the 'whole-system' approach and 'reverse-design' for biomaterials and bioprocess development, which aims to synchronously combine all the information from the molecular design level to the large-scale manufacturing and life cycle assessment level for decision-making. In return, this introduces more mathematical and computational challenges for bioprocess multiscale modelling. Despite the advanced mathematical theories (e.g. for optimisation, control, and statistical analysis) and high-performance computing facilities which have been developed, the current research community and industry still lacks robust, efficient, flexible, and high-fidelity modelling techniques that can accurately quantify complex biochemical engineering domain knowledge into mathematical formulae.

Within the last decade, there has been a significant shift from physical modelling into data-driven modelling, applying machine learning techniques and the large amount of data accumulated in the biochemical industry. The combination of machine learning algorithms and informative datasets has demonstrated great potential in the discovery of complex relationships, as expressed implicitly in the data [9–11]. This attractive trait has greatly appealed to biochemical engineers and scientists within the research community considering that biochemical processes are some of the most complex systems in the real world. The complexity of a bioprocess is mostly governed by the generally poor definition of the biochemical system, including: the use of complex raw materials and macromolecules released from the cells; lack of understanding of the biological mechanisms; difficult-to-quantify heterogeneity in the microbial population, due to genetic and environmental variations; and the multiphase nature of bioreactor systems. A quick search from the Web of Science shows that there have been 610 research articles published with combined use of 'machine learning' and 'biochemical' keywords during the last 30 years, with an average annual growth rate of up to 50 %. Evidently, the real number of publications could be much higher than this if a more thorough search is carried out (e.g. using specific names of machine learning algorithms or biochemical engineering sub-fields as a key word).

As a result, it will be intriguing and timely to understand how machine learning has shaped the current field of biochemical engineering research. Therefore, this review will primarily target biochemical engineers who are interested in the application of machine learning into biochemical engineering. The objectives of this review are to: (i) briefly explain the algorithmic concepts of several machine learning methods that have been (or will be) widely used in the biochemical engineering area; (ii) give a handful of examples to demonstrate how machine learning has been implemented into different biochemical engineering studies; and (iii) highlight the current challenges of machine learning and its future prospects in biochemical engineering. In order to ensure this review is up-to-date, a thorough literature search was conducted selecting over 200 research articles primarily published from 2000 to 2020. By analysing these publications, we found that at this moment the most popular machine learning methods are artificial neural networks (43 %), followed by ensemble learning (13 %), multivariate statistical analysis (9%), support vector machines (7%), and Gaussian processes (6%). It is necessary to emphasise that: (i) the popularity of these machine learning methods only indicates the trend within the last 20 years, which may be different from previously published reviews (e.g. multivariate statistical analysis was the predominant method used in the early 1990s); and (ii) specific machine learning algorithms have been grouped into different categories of machine learning methods for convenience. For instance, both partial least squares and principal component analysis have been grouped into the multivariate statistical analysis method. The timelines regarding the application of these major machine learning methods are presented in Figs. 1–5.

The structure of this review is presented as follows. In Section 2, we will focus on the question '*which machine learning algorithms have been used in biochemical engineering?*'. We will demystify 6 algorithms including artificial neural networks, principal component analysis, partial least squares, support vector machines, decision tree based ensemble learning methods, and reinforcement learning. In Section 3, we will answer the question '*how are these methods applied into different biochemical engineering fields?*'. In particular, we will illustrate several recent case studies across 7 biochemical engineering fields including: bioreactor engineering, biodevices and biosensors, biomaterials engineering, bioresources and bioenergy, environmental bioengineering, metabolic engineering, and cell culture and protein engineering. Finally, in Section 4 and 5, we will highlight the current challenges and future prospects of using novel machine learning methods and digital twins for biochemical engineering applications. We will also provide an in-depth discussion about how to integrate machine learning techniques with domain knowledge and physical models to improve their accuracy and allow for their further application in industrial bioprocess and biotechnology development.

## 2. Elucidation of machine learning algorithms

### 2.1. Multivariate statistical analysis

Multivariate statistical analysis (MSA) includes a range of supervised and unsupervised machine learning algorithms that have been extensively used since the 1990s, as shown in Fig. 6. Some of the most well known include principal component analysis, partial least squares, and polynomial regression. Arguably, most of these algorithms were popularised in the early-mid 1990s before other machine learning methods were widely used in engineering research. Although at this moment MSA algorithms has been well developed and are known as 'old' techniques, they still remain as the most commonly used machine learning tools for the bioprocess industry. In this section, we will mainly focus on two MSA algorithms, principal component analysis and partial least squares.

#### 2.1.1. Principal component analysis

Principal component analysis (PCA) is a dimensionality reduction algorithm. This is achieved by creating new uncorrelated variables (i.e. latent variables) and maximising their variance. PCA has been previously used to analyse spectroscopic data to identify chemical compounds [12] and discriminate cell phenotypes [13], to discover the hidden patterns between different reactor operating factors for biogas [14] and biofuel [15] production, and to detect process abnormality for large scale cell culturing systems [16].

The mathematical theory of PCA is that for each dataset $\boldsymbol{X}_{i\in N} = (x_{i1}, x_{i2}, \ldots, x_{iJ})^T$ there are $J$ different variables $(x_1, x_2, \ldots, x_J)$, the aim of PCA is to construct $K < J$ new variables $\boldsymbol{Z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})^T$ which can best represent the 'most significant' information of each dataset by linearly combining their original variables shown as Eq. (2.1).

$$z_{ik} = a_{k1} \cdot x_{i1} + a_{k2} \cdot x_{i2} + \ldots + a_{kJ} \cdot x_{iJ} \tag{2.1}$$

where $x_{ij}$ refers to the original variable $x_j$ in the $i$ th datapoint, and $z_{ik}$ refers to the latent variable $z_k$ in the $i$ th datapoint. Here, the original variable $x_j$ can be either a measurement at a specific wavelength in a spectroscopic dataset or a process parameter (e.g. pH, temperature,

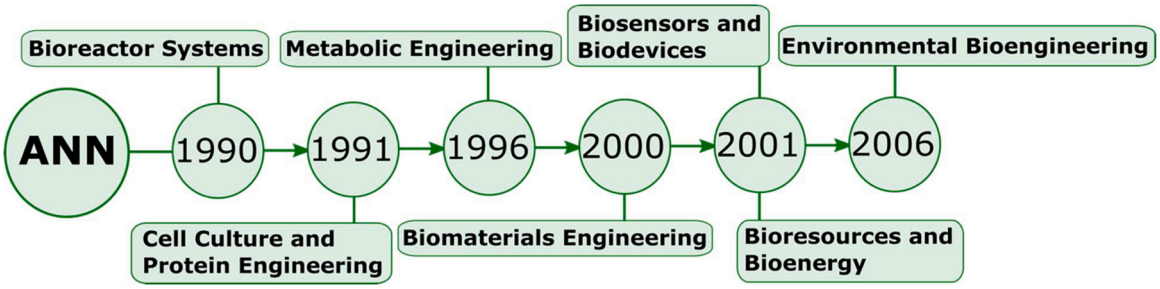


**Fig. 1.** Timeline of artificial neural networks applications in biochemical engineering.

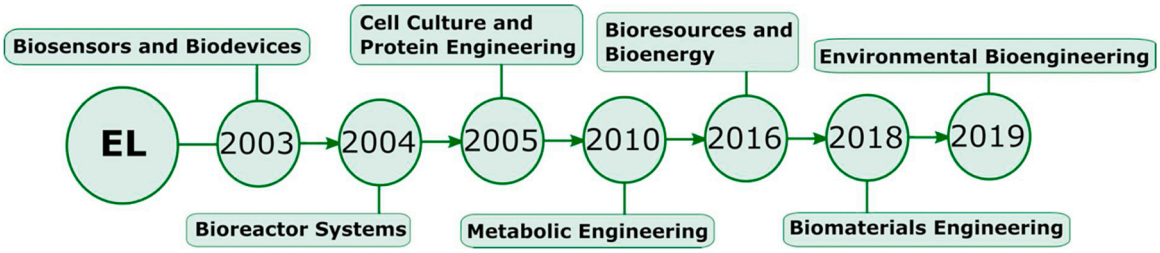


**Fig. 2.** Timeline of ensemble learning methods applications in biochemical engineering.

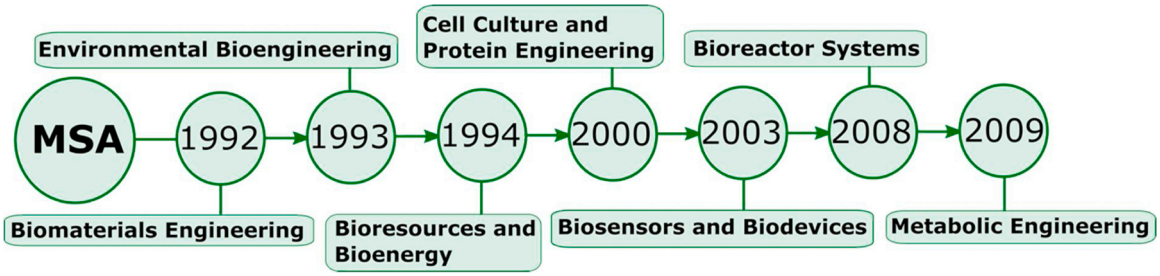


**Fig. 3.** Timeline of multivariate statistical analysis applications in biochemical engineering.

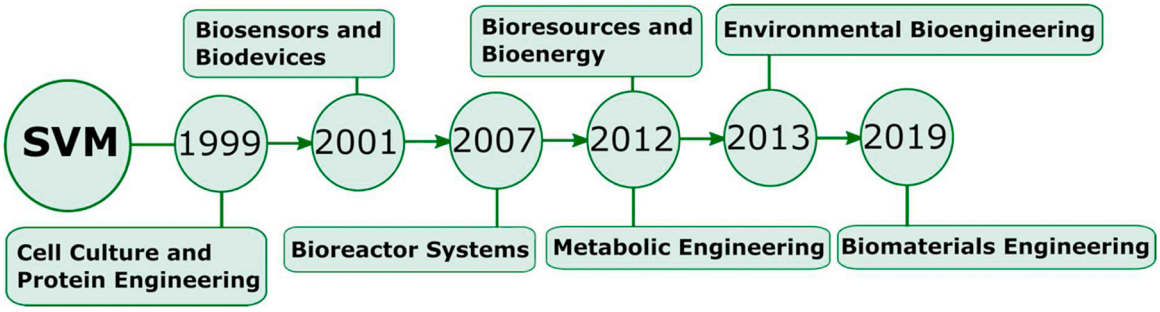


**Fig. 4.** Timeline of support vector machines applications in biochemical engineering.

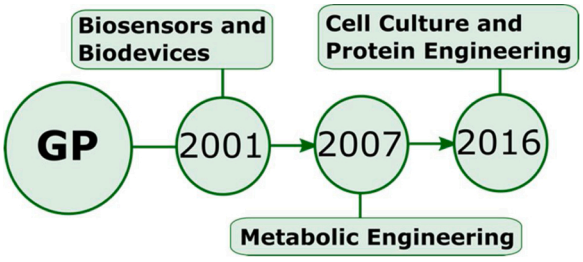


**Fig. 5.** Timeline of Gaussian processes applications in biochemical engineering.

flowrate) in a bioreactor operation dataset, depending on the case study.

The vector $\boldsymbol{a}_k = (a_{k1}, a_{k2}, \ldots, a_{kJ})$ is calculated by maximising the variance of latent variable $z_k$ shown in Eqs. (2.2)–(2.4) [17]. This is based on the assumption that the variance of latent variables can represent the 'hidden' pattern within the original datasets and hence is the best way to distinguish different datasets. When applying PCA, most studies only calculate the first two principal components (shown in Fig. 7) as they typically hold more information than further principal components [14,13,12]. The original datasets are then labelled by their respective principal component pairs (1st PC, 2nd PC) and are plotted to discover potential relationships.

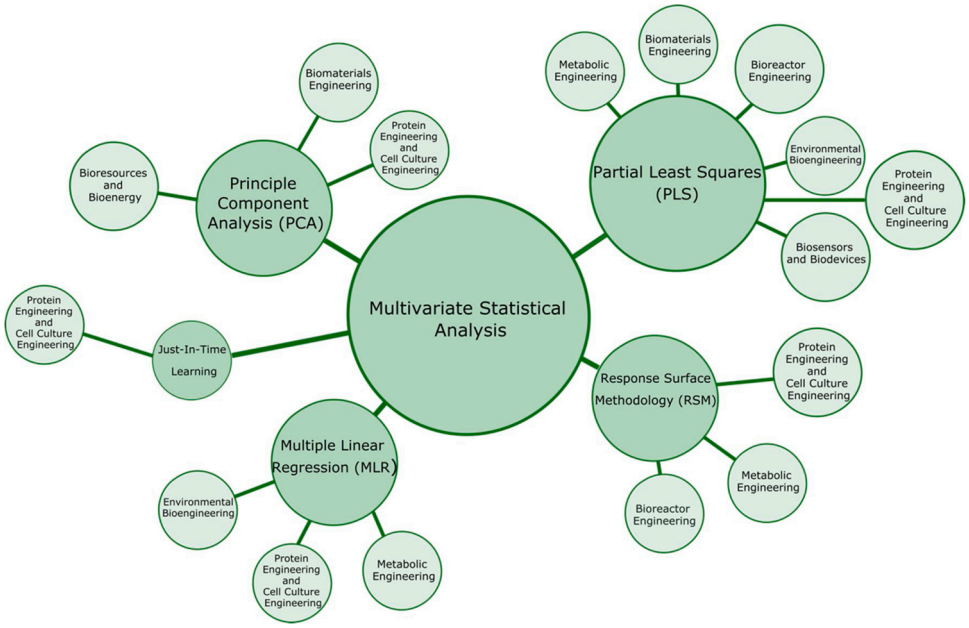


**Fig. 6.** Outline of multivariate statistical analysis papers within biochemical engineering.

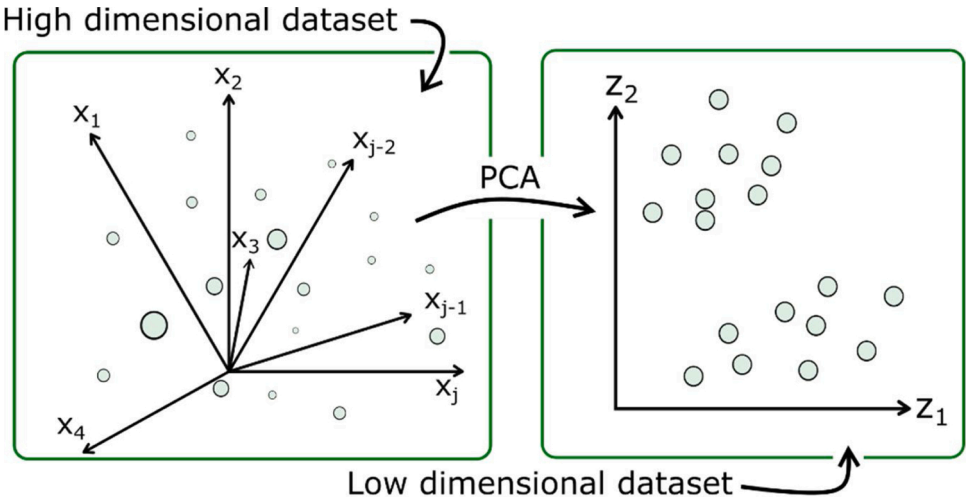


**Fig. 7.** PCA of a high-dimensional dataset to a 2-dimensional dataset.

$$\max \sum_{i=1}^{N} (z_{ik})^2 = z_{1k}^2 + z_{2k}^2 + \ldots + z_{ik}^2 \tag{2.2}$$

$$s.t.\ \ a_{k1}^2 + a_{k2}^2 + \ldots + a_{kJ}^2 = 1 \tag{2.3}$$

$$\boldsymbol{a_k} \cdot \boldsymbol{a_m^T} = 0,\ m = 1, 2, \ldots, k-1 \tag{2.4}$$

### 2.1.2. Partial least squares

Partial least squares (PLS) is a similar dimensionality reduction algorithm but can directly find the relationships between an input dataset $\boldsymbol{X_i} \in \mathbb{R}^J$ and a corresponding output dataset $\boldsymbol{Y_i} \in \mathbb{R}^K$, with $K \leq J$ [18]. It aims to establish a linear correlation between the input and output variables within their latent space. In other words, PLS transforms the original variables $\boldsymbol{X_i}$ and $\boldsymbol{Y_i}$ into their respective latent variables $\boldsymbol{Z_i} \in \mathbb{R}^p$ and $\boldsymbol{Q_i} \in \mathbb{R}^l$, with $p \leq J$ and $l \leq K$, and then seeks a most probable linear correlation between $\boldsymbol{Z_i}$ and $\boldsymbol{Q_i}$. The equations to calculate $\boldsymbol{Z_i}$ and $\boldsymbol{Q_i}$ are similar to PCA, thus not repeated here.

Although in theory PLS is a multiple-input multiple-output (MIMO) model, in practice it is mainly used as a multiple-input single-output (MISO) model. Therefore, if there are several output variables to be correlated with input variables, it is common to construct several independent PLS models in parallel. The number of latent input variables used to correlate the output variable is also flexible and is usually decided by the user. In addition, through the use of data unfolding, the multiway partial least squares (MPLS) algorithm has also been developed to extend the use of PLS for time-series event simulation [18].

PLS has been previously used for bioreactor state estimation and control [19,20] and bioprocess monitoring [21]. It also finds a number of applications in identifying relationships within different metabolic reaction fluxes [22] and mapping amino acid sequences with protein functions [23]. As PCA and PLS are developed based on linear regression, they are easy to construct and are particularly efficient for high dimensional data analysis. However, as most of the biochemical reaction systems are highly nonlinear and low dimensional (total number of measurable state variables and operating parameters ~10), the accuracy of PCA and PLS is often found to be low when simulating these processes. As a result, they have been mostly replaced by artificial neural networks

since the early 2000s. Amongst the 200 selected research papers published between 2000 and 2020, PCA and PLS only account for 9% of the machine learning applications. Moreover, with the development of neural network based dimensionality reduction techniques such as autoencoders, which have already been applied into protein engineering [24] and biosensors design [25], it is envisaged that the use of linear regression based PCA and PLS might be less popular in the future.

### 2.2. Support vector machines

Support vector machines (SVMs) are a well-established technique and have been applied to both classification and regression tasks within the biochemical engineering domain since as far back as the late 1990s and early 2000s. SVMs enable the extraction of complex nonlinear relationships as typically expressed within bio-applications. This is shown in Fig. 8. They have been used to predict metabolic fluxes [26], to model bioreactor membrane fouling for sewage treatment [27] and to classify nucleotides for nanopore sensing [28]. However, it is worth noting that SVMs provide this analysis outside of the framework provided by Bayesian reasoning, as discussed in Section 2.4. Therefore, SVMs do not express any inherent notion of uncertainty. As a result, methods for constructing an approximate posterior distribution have been proposed for SVM classification and we direct the interested reader appropriately for more information [29].

In the following, we will discuss the mathematical formalisation of SVM binary classification with reference to a histopathological example. Consider a dataset $D = \{\boldsymbol{x}_i, y_i\}_{i=1:N}$, where $\boldsymbol{x} \in \mathbb{R}^{n_x}$ is a feature vector characteristic of a single datapoint and $y \in \{-1, 1\}$ is a corresponding class label. Here, the feature vector $\boldsymbol{x}$ contains information specific to the datapoint (or sample). In the scope of pathological diagnosis the feature vector $\boldsymbol{x}$ may be a spectral measurement of a tissue sample. Each dimension within $\boldsymbol{x}$ provides structural and chemical information about the sample. It is of interest to construct a predictive model to classify datapoints or samples, e.g. whether the tissue is either 'healthy' or 'diseased', and use the model thereafter to help guide decision-making for further patient treatment.

Explicitly, we can imagine that the dataset $D$ occupies some higher dimensional feature space, with $n_x$ dimensions. Binary classification with SVM finds a hyperplane, which spatially separates the datapoints $\{\boldsymbol{x}_i\}_{i=1:N}$ based on the corresponding labels $\{y_i\}_{i=1:N}$. SVM proceeds by constructing a model $f(\boldsymbol{x})$ as a linear combination of kernel functions. For the sake of this analysis, a kernel function is a mathematical mapping from the original feature space $\boldsymbol{x}$ to a higher dimensional space (in actuality, kernels possess specific mathematical properties) [30]. When a linear kernel is applied within the model, the hyperplane is expressed as Eq. (2.5):

$$\{\boldsymbol{x} : f(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{w} + b = 0\} \tag{2.5}$$

where $\boldsymbol{w} \in \mathbb{R}^{n_x}$ and $b \in \mathbb{R}$ are the weights and bias of the model $f(\boldsymbol{x})$, respectively [31]. The weights define an orthogonal vector to the hyperplane, which separates our two classes. SVM proceeds to form a quadratic programming (QP) problem to maximise the distance from the hyperplane to the 'support vectors' [31] illustrated in Fig. 9 by solving the following optimisation problem as detailed in Eqs. (2.6) and (2.7):

$$\min_{\mathbf{w},b} \boldsymbol{w}^T\boldsymbol{w} \tag{2.6}$$

$$s.t.\ \forall_i\ (\boldsymbol{w}^T\boldsymbol{x}_i + b)y_i \geq 1 \tag{2.7}$$

Clearly, the kernel function in Eq. (2.5) results in a classification model where the classes are linearly separable in the original feature space. If this is not possible, other kernels may be introduced to project the feature vectors $\boldsymbol{x}$ into a space in which the classes become linearly separable. Common examples of kernels include the radial basis function (RBF), polynomial functions, and sigmoid functions [31]. In recent years, a number of research developments have ironed out early limitations of SVMs, particularly with respect to the use of medium sized datasets, enabling the construction of powerful predictive models. This has been facilitated through the development of novel numerical approaches to the optimisation and explains the use of SVMs as a

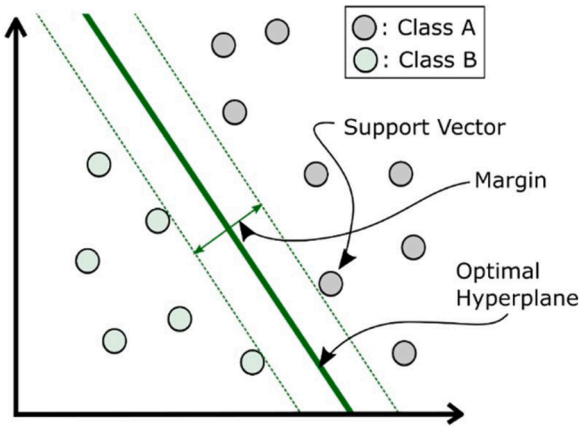


**Fig. 9.** Demonstration of the concepts common to SVM binary classification.

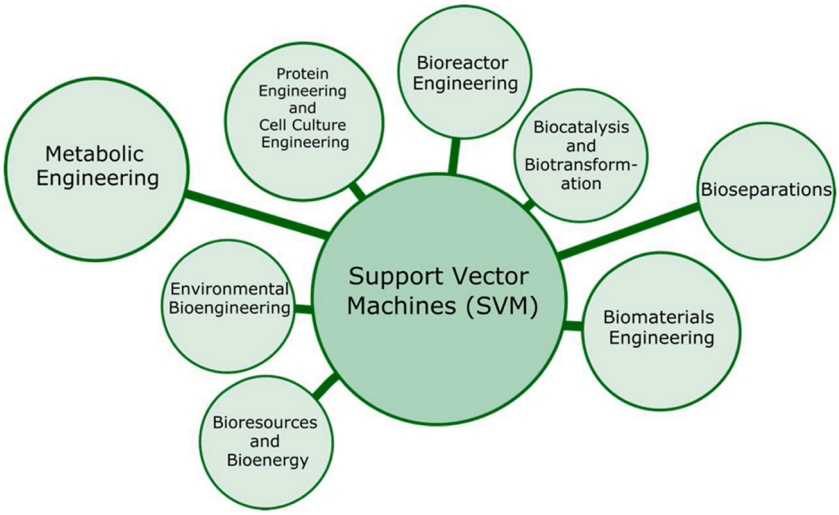


**Fig. 8.** Outline of support vector machine papers within biochemical engineering.

benchmark in many applications [32]. However, SVMs are inherently limited when applied to large datasets. This is because the use of a kernel could demand large memory and is computationally intensive. Explicitly, the memory demands are $O(N^2)$ and computation demands $O(N^2 n_x)$ operations [33], where $N$ and $n_x$ are the number of datapoints and variables, respectively. Increasingly, industry and professional practice requires methods with ease of model interpretability combined with the ability to leverage large datasets. Also, SVMs do not inherently provide an uncertainty associated with prediction. This combination of factors will ultimately obstruct its use in many settings.

### 2.3. Artificial neural networks

Neural networks are an extremely common regressive machine learning technique, finding use throughout biochemical engineering since the early 1990s with a strong recent resurgence catalysed by the advent of deep learning in 2015. To the best of our knowledge, some earliest uses of artificial neural networks (ANNs) in biochemical engineering were in 1990 when the method was used for the on-line prediction of fermentation variables [34], with ANNs applied to characterise amino acid sequences the following year [35], a problem that was recently deemed 'solved' 29 years later [36]. In other biochemical domains neural networks have originally found use for the prediction of freezing time of food products [37], bioprocess simulation [38,39], and for modelling the bioremediation of contaminated groundwater [40]. It is a testament to the continued success and innovation of ANNs that 129 out of the 200 studies reviewed in this paper made use of some form of neural network shown in Fig. 10.

The method is supervised and therefore receives a set of input data, for example observable quantities such as pH or substrate composition, and returns predictions of a specified quantity for example, biomass concentration or material property. Neural networks are highly flexible due to their large amount of parameters and are able to provably approximate any function [41]. By providing this specific flexible model structure and a set of input and output data, commonly referred to as training data, the parameters of the neural network can be iteratively changed such that the inputs correspond to their correct outputs and incrementally ensuring estimates become closer and closer to the training data. Subsequently, when a new set of inputs is shown to the network, predictions of the unobserved output can be made as the neural network has 'learned' from the training data about reasonable outputs for a given input, as is the case for any regressive interpolation.

Mathematically, a series of linear combinations specified by weights and nonlinear equations specified by the activation function allows for the transformation from an input vector $\boldsymbol{x} \in \mathbb{R}^{n_x}$ to an output vector $\boldsymbol{y} \in \mathbb{R}^{n_y}$, with intermediate variables at layer $i$ denoted as $z_i \in \mathbb{R}^{n_{z_i}}$. The transformation of inputs over a single layer is calculated as Eq. (2.8):

$$z_i = f(\boldsymbol{W}_i^T \boldsymbol{x}) \tag{2.8}$$

where $\boldsymbol{W}_i \in \mathbb{R}^{n_x \times n_{z_i}}$ represents the weight matrix of layer $i$, and $f(\cdot)$ is the activation function providing nonlinearity to the network. $z_i$ can be a vector of any length $n_{z_i}$, with its length specifying the number of nodes in a hidden layer. Some commonly used activation functions are the ReLU (rectified linear unit) function (Eq. (2.9)) or the sigmoid activation function (Eq. (2.10)) providing a piecewise linear, and nonlinear transformation respectively:

$$f_{ReLU}(\boldsymbol{W}_i^T \boldsymbol{x}) = \max(\boldsymbol{W}_i^T \boldsymbol{x}, 0) \tag{2.9}$$

$$f_{sigmoid}(\boldsymbol{W}_i^T \boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{W}_i^T \boldsymbol{x}}} \tag{2.10}$$

As stated, the structure of an ANN has no physical meaning, disregarding origins in modelling the interaction between neurons, and is specified as such to provide a flexible regression framework in order to predict the often complex relation between inputs and outputs. In order to vary the weights to approximate an underlying function, ANNs most commonly take advantage of the backpropagation algorithm. Simply put, the derivative of the error between the training output and predicted output with respect to the weights of the network is calculated,

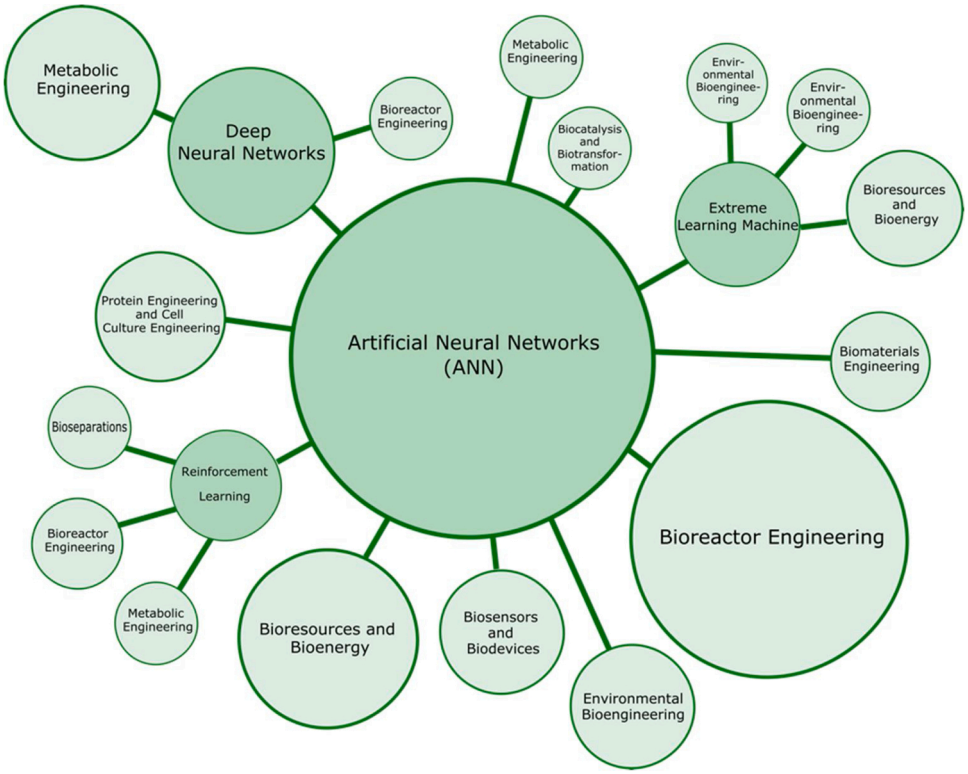


**Fig. 10.** Outline of artificial neural network papers within Biochemical Engineering.

enabling gradient-based optimisation solvers to minimise this error also known as the loss. An interested reader should refer to [42].

The size of the input and output layer are problem specific however, there is no specification as to the size or structure of hidden layers within a neural network and as such a result a number of creative network structures have been proposed and shown in Fig. 11, each with advantages in specific scenarios. Notable watershed structures include Convolutional Neural Networks (CNNs), which enable a matrix or tensor of inputs such as an image, and have been used in bioreactor systems to optimise algal biofuel production [5], to classify the 3D structure of enzymes in metabolic engineering [43], and most recently to predict protein structures to remarkable accuracy [44]. Another notable structure is the class of Recurrent Neural Networks (RNNs), which have advantages in modelling series data utilising what can be seen of as an 'internal memory' and have been used to recognise promoter sequences in the field of genetic engineering [45], statistically summarise amino-acid sequences in protein engineering [46], and for the calibration of soft-sensors in the biosensor domain [47]. Undoubtedly the biggest structural development has been the advent of Deep Neural Networks (DNNs) [48] where it was discovered that a large number of hidden layers can facilitate extremely complicated underlying functions to be modelled, largely due to the sheer number of parameters. The majority of previously mentioned applications also incorporate a 'deep' aspect insofar as they incorporate a large number of hidden layers, and this concept can be generalised to most neural network structures.

A key challenge in ensuring the success of neural networks is hyper-parameter selection, in other words the determination of the correct number of hidden layers, neurons per hidden layer, and activation function. Different approaches have been proposed ranging from basic random search of hyper-parameters to more advanced techniques such as Bayesian optimisation. Regardless of the method, correct hyper-parameter selection can ensure the success of neural network modelling, and is an important aspect that is often left unaccounted for. Identifying optimal hyper-parameters is also vital if a neural network is used for feature selection and dimensionality reduction, such as the use of auto-encoders to de-noise RNA sequence datasets [49] and learn latent representations of genomic data [50].

### 2.4. Gaussian processes

Gaussian processes (GPs) are a probabilistic machine learning method. As a result, predictions returned are probability distributions as opposed to scalar values. Commonly placed under the class of interpolating machine learning techniques, with no assumed measurement noise a GP will fit exactly to the dataset. Broadly, predictions are made based on a weighted sum of the existing output data, weighted by the predictions distance from the existing data in input space. The probability distributions returned are particularly useful as they provide an insight into the uncertainty of a prediction. As shown in Fig. 12, GPs have been used to ensure constraints are adhered to probabilistically in bioreactor control [51], estimate a distribution of biomass concentration

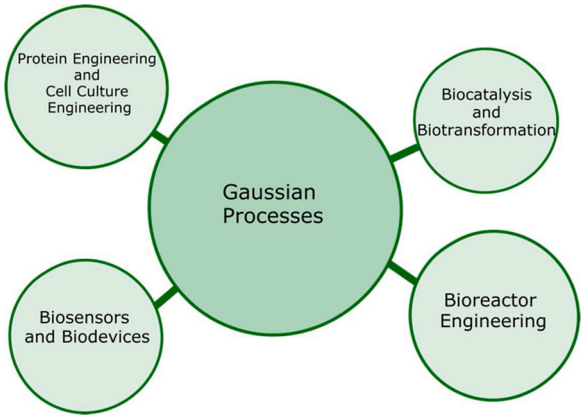


**Fig. 12.** Outline of Gaussian process papers within biochemical engineering.

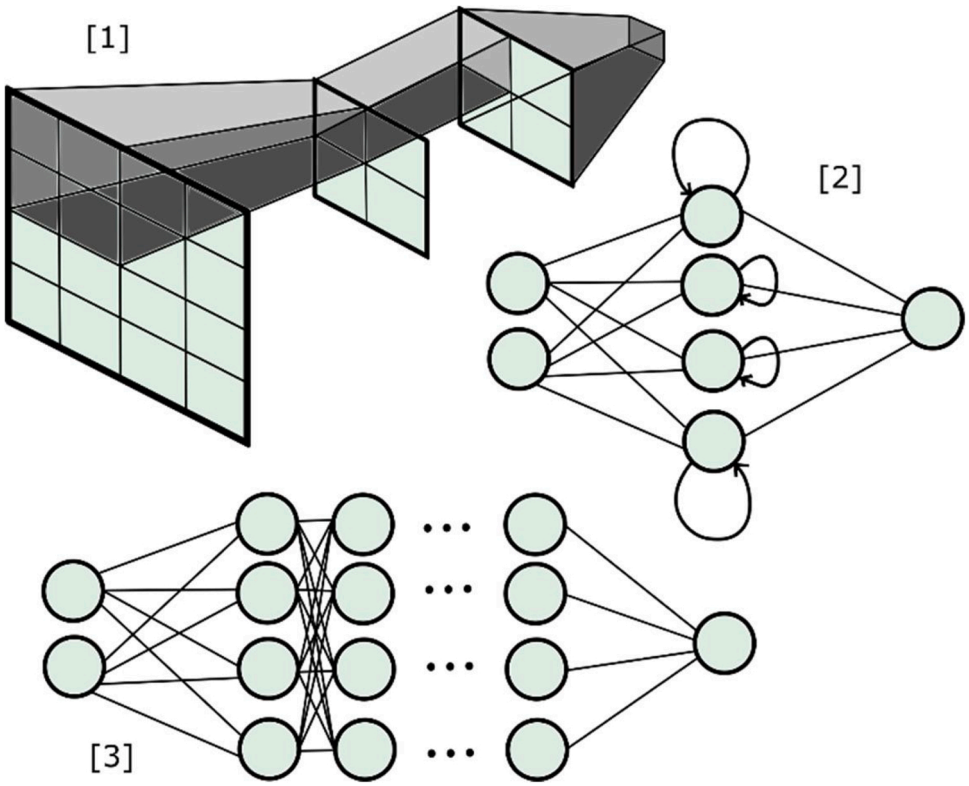


**Fig. 11.** Different neural network structures: [1] convolutional neural network [2], recurrent neural network [3], deep neural network.

in batch bioprocesses [52], and aid the creation of directed evolution mutants in protein engineering [53].

Generally a GP is a probability distribution over functions, in other words, a distribution that as opposed to returning single values instead returns functions. This distribution is then conditioned on the training data using Bayesian reasoning resulting in a posterior predictive distribution. For the interested reader [30] provides a more detailed explanation. A GP prediction of inputs $\boldsymbol{x} \in \mathbb{R}^{m\times d}$ is therefore seen as a sample from a Gaussian process conditioned on training data $\boldsymbol{x}^* \in \mathbb{R}^{n\times d}$ with corresponding mean function $m(\cdot)$ and covariance (kernel) function $k(\cdot, \cdot)$ shown in Eq. (2.11).

$$f_{GP}(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^*)) \tag{2.11}$$

The prior distribution over functions is specified by the kernel function and allows for information known prior to modelling to be incorporated. Kernel functions must be positive definite with examples including the most commonly used squared-exponential (Eq. (2.12)) or periodic function (Eq. (2.12)), illustrated as Fig. 13:

$$k_{SE}(\boldsymbol{x}, \boldsymbol{x}^*) = \sigma^2 \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}^*)^T(\boldsymbol{x}-\boldsymbol{x}^*)}{2l^2}\right) \tag{2.12}$$

$$k_{periodic}(\boldsymbol{x}, \boldsymbol{x}^*) = \sigma^2 \exp\left(-\frac{2\sin^2(\pi|\boldsymbol{x}-\boldsymbol{x}^*|/p)}{l^2}\right) \tag{2.13}$$

where $\sigma \in \mathbb{R}, l \in \mathbb{R}$ and $p \in \mathbb{R}$ are hyper-parameters that control the properties of the underlying distribution over functions, and are optimised to maximise the maximum likelihood of the GP model. Both the lengthscale parameter $l$ and periodicity $p$ can also be specified to vary over the input dimensions, in which case the expressions of Eqs. (2.12) and (2.13) are modified. This is known as automatic relevance determination (ARD) – for more information see [30].

The main bottleneck in Gaussian process prediction is the inversion of a covariance matrix, an operation which computationally scales with the number of data points. Thus, alternatives such as sparse Gaussian processes have been developed that either approximate the posterior predictive distribution or the precision matrix, which scale with exponentially larger datasets [54]. Though it should be noted that these methods, based on the papers surveyed, are yet to be exploited in the biochemical engineering domain due to an emphasis on the performance of GPs with small amounts of data [51,55].

### 2.5. Ensemble learning

Ensemble learning is a general term for supervised learning methods that combine predictions of multiple inducers to make a decision. An inducer, also termed a basic learner, is a simple machine learning algorithm that approximates the relationship between the input and output data. The often improved predictive performance of ensemble learning methods is achieved by avoiding overfitting, thus decreasing the risk of obtaining locally optimal models and extending the search place to achieve a better fit [56,57]. According to the relationships between each individual inducer, ensemble learning methods can be divided into dependent frameworks and independent frameworks [58]. In a dependent framework, the output of an inducer affects the construction of the next inducer, whilst in an independent framework each inducer is constructed independently from the others.

Random Forests (RF) are an example of an independent framework. As probably one of the most popular ensemble learning methods used in biochemical engineering, they have shown high accuracy and simplicity in implementation. RFs have been previously used to predict the molecular signature of biological pathogens [59], to optimise fermentation for Rifamycin B production [60] and to estimate the cetane number of different fatty acid methyl esters [61] for biodiesel synthesis. Recently, they have been employed to predict amino acid types for biomaterials engineering applications [62] and to predict temporal bioavailability changes of complex chemical mixtures in contaminated soils [63]. In contrast, Gradient Boosting (GB) falls under the domain of dependent framework-based ensemble learning methods, in which the construction of each inducer is dependent on its predecessor that has previously been trained. Gradient boosting has been used in the bioseparation domain to predict liquid-liquid phase separation [64] and also in the biomaterials engineering domain to predict mechanical functionality of protein networks [65].

Summarised in Fig. 14, decision tree-based ensemble learning methods have been applied in different biochemical engineering studies. Given that RFs are the more widely used method based on the literature surveyed in this review, we will use the RF as an example to briefly illustrate the concept of ensemble learning. Specifically, we will use the case study in [63] to explain how an RF is used to predict the bioavailability and toxicity of complex chemical mixtures. The inducers that make up an RF are decision trees, hence the term 'forest'. Decision trees are a commonly used classification algorithm utilising piecewise linear models in order to make predictions. Decision trees take on a flowchart-like structure whereby a series of if-else statements are applied to the inputs to produce an output prediction. As a single decision tree is not efficient in regression unless unrealistically large, RF uses a number of decision trees (hence ensemble learning) $\{T_1(X), \ldots, T_A(X)\}$, whereby $A$ is the number of trees, to enhance the overall model accuracy. In the case study of [63], the input $X = \{x_1, \ldots, x_n\}$ is an $n$-dimensional vector of properties associated with type of soil, and the output of each decision tree is toxicity and bioavailable concentration at any time $t$.

Once the prediction of each decision tree is completed, these results are averaged to obtain the final prediction of the RF. The GB method also follows a similar approach, however, in this case the decision trees are constructed in sequence rather than in parallel, as each decision tree is used to rectify the prediction error of its predecessor. The structure of an RF and a GB is presented in Fig. 15. Similar to ANNs, the accuracy of decision tree-based ensemble learning algorithms are sensitive to several hyper-parameters such as the number of trees, maximum depth of each tree (or maximum number of leaves per tree) and learning rate.

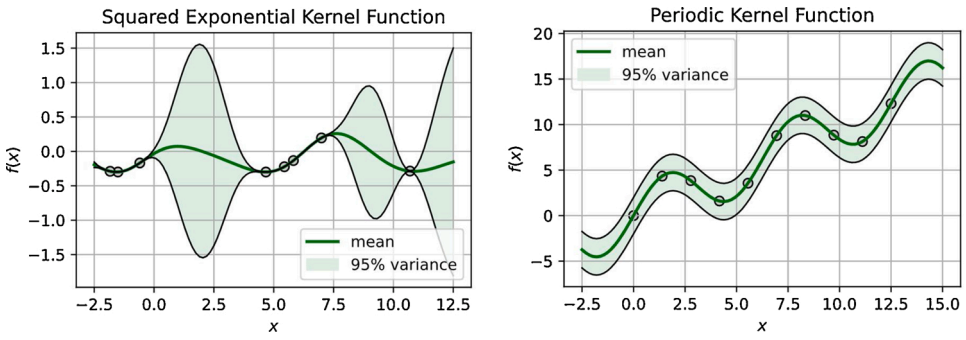


**Fig. 13.** Demonstration of the effect of different Gaussian process kernel functions.

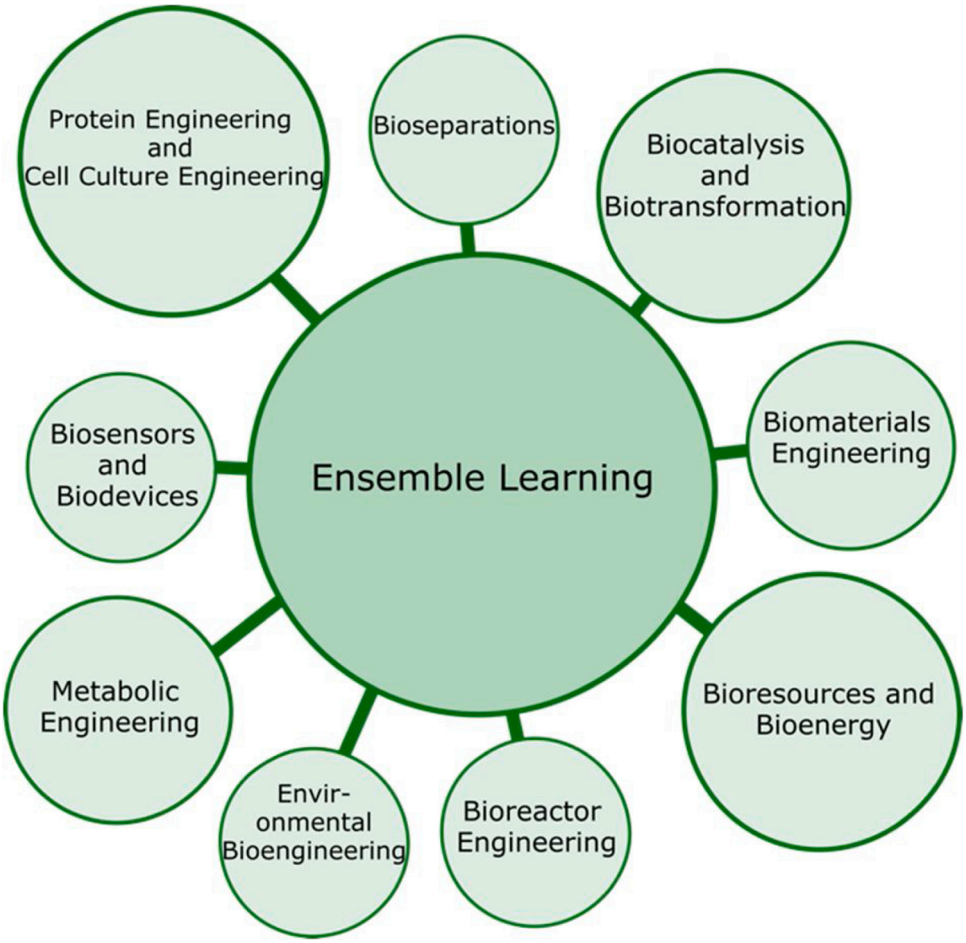


**Fig. 14.** Outline of ensemble learning papers within biochemical engineering.

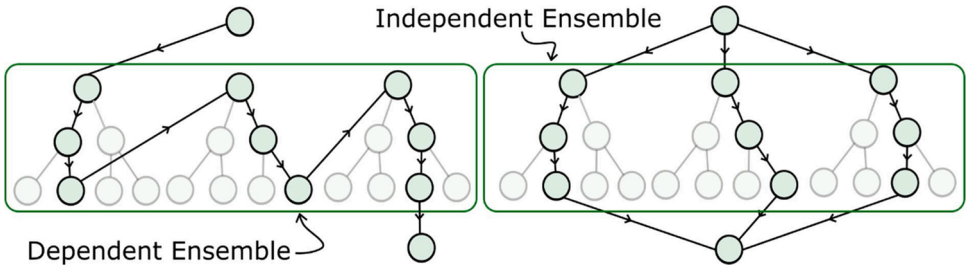


**Fig. 15.** Gradient boosting (left) and Random forest (right).

Although the training of ensemble learning methods can be parallelised efficiently, with typically small variance and strong generalisation capabilities, these methods still have several inherent challenges. For example, RFs are prone to overfitting particularly when applied to small or low dimensional datasets. GB usually requires many trees (greater than 1000), which can potentially consume a lot of memory and increase computational cost [58]. Within the 200 research articles collected in this literature review, more than 30 % of the research articles have applied RF with a primary focus on protein engineering and cell culture engineering. The application of RF in some other fields such as bioreactor engineering and metabolic engineering however, is relatively rare. Considering the structure between different ensemble learning methods, it is often suggested that RFs should be mainly used for classification purposes [66] and GB should be used for regression purposes [67]. Given that the use of GB for biochemical engineering studies (e.g. data regression for bioprocess modelling) is still at its early stage, one can expect that these ensemble learning algorithms will also become more popular in the future.

### 2.6. *Reinforcement learning*

Reinforcement learning (RL) is different from supervised and unsupervised learning. Instead RL underpins a trial and error approach to learning an optimal input sequence (e.g. control inputs, or symbols used to encode the chemical structures), which is conditioned upon observations of an underlying system. The learning of this optimal input sequence is facilitated by the assignment of a reward or penalty to each input within the sequence as trialled. RL has been widely celebrated within game-based domains [68], but is yet to observe wide deployment within biochemical engineering. Nonetheless, it is an area of increasing academic focus and has been adopted in some biochemical engineering applications. To our knowledge, its earliest use in biochemical

engineering was to adjust the key parameters of kinetic models for the metabolic activity of microbial cells in the context of metabolic engineering [69]. More recently, RL has been used for the optimisation of a phycocyanin fed-batch process [70,71] and a continuous cation exchange chromatography system for separation of charged species [72].

To give an overview of the mathematical theory behind RL, we use the case study from [70]. This study is a dynamic optimisation of a fed-batch bioprocess (a microalgal phycocyanin production process). There are two types of variables: the process states, $\boldsymbol{x}$ (e.g. concentrations of biomass, substrate, and bioproduct) and the control inputs, $\boldsymbol{u}$ (e.g. nitrate inflow rate and light intensity). A reward function, $r(\boldsymbol{x},\boldsymbol{u})$ is formulated to provide a scalar feedback signal about the quality of control inputs conditionally selected with respect to the process objective (maximise total phycocyanin production). The aim of RL is to learn a control policy, $\pi$, which is conditional upon observations of the system state, $\boldsymbol{x}$ (such that $\pi := \pi(\boldsymbol{u}|\boldsymbol{x})$), in order to maximise the total accumulated reward. Often, this control policy is parameterised by an ANN. Once the RL training process is 'complete', the policy can predict the approximate optimal control $\boldsymbol{u}^* = \boldsymbol{\pi}^*(\boldsymbol{u}|\boldsymbol{x})$ given an observation of the current process state, $\boldsymbol{x}$ at the current time step. Following the policy thereafter will maximise the cumulative reward observed in process operation. This definition expresses inherent analogy to the optimal control problem and is expressed more formally by Eqs. (2.14) and (2.15).

$$J = E_{\pi}\left[\sum_{t=1}^{T} r_t \,\middle|\, \boldsymbol{x}_0 \sim p(\boldsymbol{x}_0)\right] \tag{2.14}$$

$$\pi^* = argmax_{\pi}\, J \tag{2.15}$$

where $J$ denotes the expectation of total accumulated reward from a discrete number of control interactions $T$ under policy $\pi$, given an initial state distribution $p(\boldsymbol{x}_0)$. This objective is defined in terms of expectation given the stochasticity of the underlying system and policy. The general RL framework is shown as Fig. 16.

The landscape of RL is divided into model-free and model-based algorithms. Model-based RL attempts to model the system and find an optimal policy (e.g. optimal control actions) based on a process model. This shares strong similarities with model predictive control (MPC), which is widely used for the chemical industry. Whereas, model-free RL can learn an optimal policy from the system directly. As such, model-free RL is a promising approach to control and optimisation of systems, which are explicitly difficult to model. Model-free RL can be broadly categorised into policy optimisation methods and value-based methods. Typical policy optimisation methods directly optimise or learn a policy and implicitly learn the state-action value function. The specific RL algorithm employed in [70] is a policy gradient algorithm, REINFORCE, which is often used in RL problems that necessitate use of continuous control variables, $\boldsymbol{u}$. Value-based algorithms, typically represented by Q-learning, explicitly learn and optimise the state-action value function and generate the optimal policy by acting greedily with respect to it i.e. choosing the control corresponding to the maximum $Q_{\pi}(\boldsymbol{x},\boldsymbol{u})$ value (state-action value). There are also hybrid algorithms, such as actor-critic methods, which combine policy optimisation methods and value-based methods. Although RL has shown success in game-based control benchmarks, such as AlphaGo [73], it still has plenty of unexplored potential and yet to be applied in biochemical engineering. As a result, a more comprehensive discussion about the potential of and challenges of RL will be presented in Section 4.

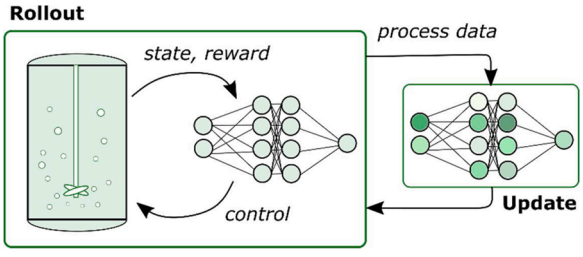


**Fig. 16.** Schematic of a generic reinforcement learning framework.

## 3. Applications of machine learning in different subfields

### 3.1. Bioreactor engineering

Machine learning has been extensively applied to approximate the intrinsic complexities between bioreactor operating factors (e.g. temperature, pH, substrate and contaminants concentrations, agitation and mixing times, nutrient inflow rate) and key cellular metabolisms (nutrient uptake, product synthesis). This enables (i) the estimation and prediction of important state variables (e.g. cell biomass concentration), (ii) improvements of the bioreactor's performance and efficiency [74,75], and (iii) monitoring the process conditions for control and safety purposes [76].

Amongst different machine learning methods, ANNs and GPs are the most used algorithms for prediction tasks (e.g. state estimation) in bioreactor systems. For example, [77] applied different ANNs for reactor online prediction. Specifically, they investigated performance for one and multi-step ahead prediction of consumed sugar and lysine production in an industrial fed-batch fermentation by *Brevibacterium flavum*. Their architecture made use of five inputs (accumulated $CO_2$, respiratory quotient, consumed sugar in current and previous time steps) and predicted three outputs (produced lysine at the current and future time steps) with a coefficient of determination $R^2$ of 0.997. Similarly, both real-time and offline predictions with ANNs were demonstrated by [78] for long-term dynamic simulation of microalgal lutein photo-production by *Desmodesmus* sp. Here the authors estimated the dynamic system's rate of change at the next time step by feeding the ANN essential process information at the current time step. This improved the ANN's accuracy in predicting biomass, nutrient, and lutein concentrations. Notably, these findings applied offline when making multi-step ahead predictions, provided with only the initial conditions. In addition, GPs analytically express associated aleatoric and epistemic uncertainty. For instance, [79] predicted biomass, nitrate and lutein concentrations with 99 % confidence regions. They proposed an iterative method, which enabled propagation of model uncertainty for the full time horizon, facilitating both one and multi-step ahead prediction. Aside from the uncertainty estimation, the results were comparable to those of an ANN.

Although the aforementioned algorithms present great potentials for bioprocess modelling, their extrapolative predictions outside of the training data sets are often limited due to the lack of internal model structure. First-principle knowledge in the form of simplified differential equations (e.g. mass and energy balances) have provided mechanistic structure to ANN in the form as hybrid models by [80] for state estimation and prediction in a fed-batch stirred tank bioreactor. The fermentation process variables (e.g. biomass concentration at time $t$) acts as the input to the ANN. The ANN then provides an estimate to the current parameter (e.g. specific growth rate at time $t$) that serves an input to the first principle component (e.g. a simplified kinetic model). This component then produces the output (e.g. biomass concertation at time $t+1$) as fermentation process variables at the end of each sampling time. This adaptation reduces the potential error source as the hybrid ANN clearly identifies the contribution of each part (i.e. ANN and mechanistic model) in the overall prediction. Predictions conditional upon full and partial state measurements were reported to be better than those of pure ANNs and required significantly fewer training examples. Similar findings were reported by [81] who incorporated the output of a fuzzy expert system as input to the ANN within a hybrid ANN-fuzzy expert system.

To improve the bioreactor's output and efficiency via optimisation, researchers have commonly used ANN and GP based algorithms. For

example, in [82] the authors trained an ANN under nitrate limiting conditions and used operational space search to optimise nitrate replenishment plans for phycocyanin production in a 12-day fed-batch process. Their optimised conditions yielded a higher biomass concentration and increased phycocyanin production to 74.9 % and 85.6 % compared to previous results, respectively. In addition, the optimisation results of ANNs have been compared against those of other multivariate statistical methods like Response Surface Methodology (RSM). For instance, [83] reported a superior result with ANN ($R^2 = 0.989$) in comparison to RSM ($R^2 = 0.95$) for the optimisation of culture parameters for thermostable lipase production from a newly isolated thermophilic *Geobacillus* sp. strain. Their findings were consistent with those of [84] as a relatively lower average error was reported for ANN (6.5 %) as to RSM (20 %) in the media optimisation for fermentative production of scleroglucan. The typical approach to operational optimisation in these applications is the combination of ANNs with derivative-free optimisation techniques such as genetic algorithm (GA) and particle swarm algorithms (PSO). As an example, a hybrid ANN-GA was used by [85] to optimise the fermentation medium for the production of the enzyme hydantoinase by *Agrobacterium radiobacter*. The authors achieved a utility 11–14 % higher than the solution identified by polynomial regression. Similarly, [86] optimised the concentrations of 6 media components for xylitol fermentation by *Candida mogii*. They obtained xylitol yield which was improved by 15 %, compared to the previous result.

Unlike optimising solely the cellular productivity (e.g. maximum lutein production) with process data obtained from an existing bioreactor, deep neural networks have been recently adopted to efficiently optimise both bioreactor configurations and operating conditions jointly [87] achieved this with multi-objective optimisation and a Convolutional Neural Network (CNN). Firstly, an integrated biokinetic and computational fluid dynamic (CFD) model was constructed to simulate hydrodynamics, biomass growth and bioproduct synthesis in the upper and lower bounds of the photobioreactor's design variables. This subsequently became the training and validating data sets for CNN, which later replaced the rigorous biokinetic-CFD model for fast optimisation and interpolation of new bioreactor design and operating conditions. This strategy has yielded an increase in final biomass concentration and bisabolene production by up to 15 %.

Furthermore, there has also been a recent increase in the number of machine learning algorithms applied for predictive control and management of the complex bioprocesses and biosystems. This is partly due to the strong nonlinear dynamics, which has rendered linear controllers ineffective and presented adaptive and nonlinear controllers as an excellent choice. For instance, [88] implemented three bioreactor temperature control strategies: (i) linear model predictive control (LMPC), (ii) proportional-integral-derivative (PID) control, and (iii) neural network model based predictive control (NNMPC), during a continuous yeast fermentation. The NNMPC showed superiority over LMPC and PID for set-point changes and noisy temperature measurements during the 500 h of process control. A similar approach was also proposed by [89] for the management and operation of large scale sucrose-based $H_2$ production in bioreactors. In addition to embedding a single ANN into the model predictive control (MPC) framework as mentioned above, several ANNs composed in a hierarchical manner were implemented by [90] for advance control of methane production in an anaerobic digester. The fermentation pH, volatile fatty acid (VFAs) and biogas (methane) were modelled by 3 different ANNs with the outputs of 2 ANNs (pH and VFAs) being included as inputs into the 3rd ANN (biogas). They achieved results with methane composition constantly above 60 % and its control being stable for 341 days of fermentation.

However, the ANN-NMPC based frameworks mentioned in these studies are unable to quantity model uncertainty and model-plant mismatch, which are highly important for robust optimisation and large scale bioreactor operation. By embedding GPs into an MPC framework, the authors [51] introduced a new GP-based NMPC algorithm for finite horizon control of biochemical processes by employing concepts common to stochastic MPC. The authors satisfy joint chance constraints, by performing constraint tightening. Constraints are tightened using explicit back-offs (which can be thought of as a distance) to 'back' the nominal (or average) system evolution away from the constraint. This constraint tightening procedure allows for process uncertainties and guarantees high probability constraint satisfaction. The method showed faster online evaluation in comparison to the existing methods, and its efficiency was well demonstrated using a semi-batch bioprocess case study.

As can be concluded, significant achievements have been made within the last decades to apply machine learning techniques for bioreactor dynamic modelling, online and offline optimisation, and real-time control. However, most of these studies only focused on lab scale bioreactors, with few of them being implemented to pilot and industrial scale bioreactors. As a result, it is possible that future research will move towards the use of machine learning for bioreactor upscaling, real-time visualisation, and design of high efficiency new industrial bioreaction systems.

### 3.2. Biodevices and biosensors

There are three primary areas of focus in this application of machine learning: the optimisation and control of microbial fuel cells, the development of microfluidic and soft-sensors, and chemometric analysis for data collected from analytical measurements and techniques.

Firstly, we will examine the application of machine learning for microbial fuel cells (MFCs). MFCs have developed excitement within the academic community given their exploitation of microorganisms' anaerobic metabolic processes for the generation of electricity. Despite their potential as a sustainable option for power generation, a number of challenges exist both in the context of optimal design and operation of MFCs. As a result, MFCs are yet to observe economic maturation. These challenges underpin a lack of understanding of the multiscale physical processes that govern MFC performance and power generation [91]. Therefore, extensive works have demonstrated the application of data-driven approaches to both design and operation of MFCs. There are a number of key design parameters (cell configuration, selection of electrode material, buffer capacity, selection of cell membrane) which influence the performance of MFCs. Similarly, a series of operational factors also affect MFCs' performance. These include, but are not limited to cell temperature, ionic concentration of working solution, initial pH and nitrogen concentration.

The diversity of the factors described is undoubtedly characteristic of biochemical engineering problems and highlights the troubles associated with identification of mechanistic models. In [92], the authors investigate the use of ANN for the prediction of power density and current density for a novel membrane-less MFC under given operational conditions. In effect, the prediction of these factors enables construction of a model, which may be used for further design of experiments and operational optimisation. Similarly, in [93,94], comparable approaches to prediction of MFC performance are deployed, with the use of a relevance vector machine (RVM) – the Bayesian relative of the SVM – and, SVM and ANN, respectively. Such models provide a point prediction of the system with data generated from polarisation curve tests of the MFC. In addition [95], has proposed an machine learning based control scheme for the operational control and optimisation of MFCs. The framework is composed of two stages, an offline learning stage and an online control stage. In the offline learning stage, the authors abstract optimal control inputs to the system based on an assumed model of the system. The authors then learn an machine learning model of the optimal control trajectory in order to enact 'fast' control and optimisation of the real system online.

The second area, which also significantly benefits from the translation of machine learning approaches, is microfluidics and

microfluidics-based soft sensing. The advent of microfluidics provides promise for many biochemical engineering subfields. The constituent techniques may be translated to a diverse set of applications such as: rapid biomolecular diagnostics, which covers areas including proteomics [96], DNA and RNA sequencing [97], and lipidomics [98]; as well as single-cell analysis [99] and wearable soft-sensors. A major benefit of microfluidics approaches is the potential for highly parallelised experimentation, and consequently data acquisition. The availability of large datasets naturally lends itself to deep learning. The potential application of deep learning to microfluidics is well recognised and for further insight we refer the reader to this recent review [100]. However, in the following we expand a recent publication for the interested reader. In [47], the authors present a deep learning approach to calculating the pressure and location of contact on a wearable piezoelectric soft-sensor. Typically, these sensors exhibit a high degree of nonlinearity as well as hysteresis making it difficult to interpret the information they generate. In this work, the authors propose an RNN to learn a temporally dependent representation of the analogue voltage signal, which is generated from the sensor. This representation is then passed to two separate layers for respective prediction of the location and pressure of the contact on the sensor. These three separate components formally constitute the Hierarchical Recurrent Sensing Network proposed by the paper. In validation of the approach, the method displayed high predictive accuracies. Further developments of this nature will play part in advancing wearable soft sensing technology. It is expected that these technologies will also facilitate the arrival of highly personalised approaches to healthcare in conjunction with cloud based computing and the internet of things (IoT) approach [101].

Finally, we consider the impact of machine learning in chemometric analyses of information generated by analytical methods. Traditionally, PCA and PLS have dominated in chemometrics, but more recently more competent methods have been developed in order to account for the often nonlinear relationships observed within informative signals, as well as low signal-to-noise ratios. Low signal-to-noise ratios may be inherent to the sensing method. However, they are often imposed upon measurement acquisition in order to preserve the structure and function of the analyte or tissue being measured. This is particularly true of *in-vivo* diagnostics that utilise the informative nature of light i.e. spectroscopy. In [102], the authors present a deep learning based framework for the denoising and reconstruction of low signal-to-noise ratio Raman spectroscopy signals acquired from low power *in-vivo* measurements. Typically, specific acquisition sequences are enforced to avoid damage of tissue. These are often lengthy and ultimately still provide risk to tissue. The method proposed by the authors is centred on the use of two CNNs, one of which provides signal denoising and the other provides reconstruction of super-resolution Raman spectra. In short, the authors report that the framework facilitates a 160-fold decrease in imaging time and a 9-fold improvement upon conventional Raman denoising methods. Given the data demand of the method, the authors also demonstrate the use of transfer learning to translate the CNN models from one dataset to another.

In a further work, recently reported by [103,104], a method (SKiNET) was proposed based on self-organising maps (SOM) and an approach to feature extraction termed the SOM discriminant index (SOMDI). SKiNET enables simultaneous classification of spectral scans and extraction of the spectral features or peaks most specific to each class. This enables interpretation of the reasons for classification prediction and provides insight into the biochemical nature of class differentiation. This is a powerful concept, especially within healthcare where interpretability of the prediction is necessary for subsequent treatment. Initially, SKiNET was demonstrated for differentiation between spectral scans of different tissues in the porcine eye [103], but has since been used to differentiate between severity of traumatic brain injury (TBI) in a cortical impact mouse model [104]. In the former application, SKiNET was benchmarked against methods such as SVM, PLS-DA and a shallow ANN, and was reported to achieve state-of-the-art performance. The works discussed in this section demonstrate the increasing impact of machine learning, particularly deep learning methods, in the scope of healthcare diagnostic technologies and highlight the potential of machine learning to uncover the biochemical basis for disease investigation.

### 3.3. Biomaterials engineering

The field of biomaterials engineering is motivated by three broad research objectives: the development of new biomaterials, which have novel or improved function over those that exist; the efficient design and manufacture of existing biomaterials; and the quantification of structure-function relationships. In many cases, these drivers for research overlap and commonly contribute towards the economic maturation of biomaterials.

Initially focussing on the creation of novel biomaterials, recently there has been increasing interest in the development of biocompatible high entropy alloys (HEAs), primarily for their unique mechanical, electrical and magnetic properties, which make them suitable for applications in domains ranging from aerospace to medical implants [105]. HEAs are typically synthesised via liquid-phase methods and therefore, accurate phase prediction is required for appropriate selection of the combining elements. In [106], the authors present and adopt the use of three machine learning methods: K-nearest neighbours (KNN), SVM and ANN to learn a mapping which classifies a material as any one of three phases – solid solution, intermetallic compound, and a mixture of the previous two. The accuracy of ANN, SVM and KNN methods were found to achieve 74.3 %, 64.2 % and 68.6 % on a test set, respectively. Similarly, the development of nanopore-based sensing devices for rapid biomolecular diagnostic technology is of great interest [107] presents a computational simulation of a blockage event across a Molybdenum disulphide (MoS2) nanopore. Logistic regression, RF and KNN models are constructed to classify amino acids based on features of their respective translocation events. It was found that the models had a predictive accuracy of 72.5 %, 94.6 %, and 99.6 %, respectively.

In addition, in tissue engineering, it is often difficult to obtain samples of the target organ or cell, and as a result, cells of similar phenotypes are often used instead [108]. In [13], the authors present a methodology for the identification of cells of similar phenotype via the use of Raman spectroscopy in conjunction with a combined PCA and linear discriminant analysis (LDA) model. Specifically, the authors demonstrate the ability to differentiate between diseased and healthy cells *in-situ* for the tissue engineering of bone, achieving in the region of 70 % and 90 % accuracy.

Next, we highlight how machine learning can be applied to aid the design and manufacture of biomaterials. In [109] a machine learning pipeline was constructed incorporating image feature analysis to locate relationships between the quantitative metrics of cell structure and shape, and a classification of the substrate upon which they were grown. Specifically, SVM are deployed to learn a classification model with predictive accuracies in the region of up 90 %, successfully demonstrating machine learning based design of biomaterials' structures and substrates. Similarly, [110] highlights how machine learning can help to inform the design of methods for administering biomaterials presenting an ANN based design of a custom syringe for the optimal injectability of microparticle based drug formulations. Explicitly, the authors present an ANN based mapping from two dimensionless parameters, particle and needle geometry, to the percentage of microparticles injected. As a result, the authors employ the Buckingham $\pi$ theorem as a form of feature engineering, highlighting the importance of translating existing engineering principles to machine learning. In conjunction with rigorous CFD modelling and use of computer-aided design software, the authors were able to design a product, which enhanced the injectability of the drug formulation 6-fold when benchmarked against a commercially available alternative. Furthermore, 3D printing is a potential avenue for manufacture of biomaterials and enables rapid design and

construction of prototypes. In [111] the authors demonstrate the use of a CNN to improve the quality of light-based 3D printed structures for tissue engineering applications which are prone to undesired light scattering events. The authors conclude that the inverse model enabled printing of high fidelity and quality structures, relative to the base case.

Machine learning has also found use for the quantification of often highly complex structure-activity or structure-function relationships. For example, in [112], the authors present the use of a Bayesian ANN for the prediction of the degree of nonspecific protein binding of a surface-modified gold nanoparticle and benchmark the model against a linear quantitative structure-activity relationship (QSAR) model. It is shown that the Bayesian ANN provides a greater degree of accuracy in prediction. Similar structure-activity modelling is conducted in [113]. Here, the authors demonstrate the use of a PLS model to predict drug transport properties given descriptors of the drug molecule itself. The work reports high statistical accuracy across all models constructed and demonstrates the identification of quantitative structure-activity relationships. What is more, in [114], an ANN is used to model the relationship between physical metrology (topology, linker length, and largest cavity diameter, among other features) of a metal-organic framework (MOF) and its bulk modulus. Upon introduction of biocompatible ligands into the structure, MOFs find application in a wide variety of bio-domains [115]. It is found that the use of ANNs enables high accuracy prediction of structure properties and reinforces the potential application of machine learning to aid in the design of biomaterials.

The final example we will discuss, is the work reported by [65] which applies ensemble learning methods. Here, the authors investigate the mechanical functionality of protein structures by constructing data-driven models, which provide a mapping from structural features of the protein network to functional properties. Explicitly, gradient boosting (GB) regression is deployed to find a mapping from 26 features descriptive of the structure of a protein network to the mechanical response of the structure when subject to an external loading. The GB model is benchmarked against a random forest (RF) model and was found to perform well, achieving a mean percentage error in prediction of approximately 11 %.

### 3.4. Bioresource and bioenergy

Non-renewable resources such as fossil fuels are limited, polluting and destructive towards the environment. Interest in the use of the renewable alternatives, such as the biofuels and bioenergy have increased in recent years. Accordingly, the modelling and optimisation of biofuel and bioenergy processes are important to support the bioeconomy and ensure they have a place in future sustainable energy systems. Machine learning has found many applications in the biofuels and bioenergy domain primarily for research into the effect of using biofuels in traditional combustion engines as well as for the production of biofuels.

The effect of using biofuels within existing combustion engines has been an important research subject potentially enabling fewer exhaust emissions, lower fuel consumption, and widening the engine operational range. However, engine performance experiments are time-consuming, and the engine must be recalibrated between using different blends of biofuels. Recently, ANNs have been proposed to deal with these challenges [116] and [117] modelled an engine combusting mixed biodiesel and diesel fuel by a traditional ANN to predict the fuel engine performance indicators: power, torque, specific fuel consumption and exhaust emissions based on the engine speed and percentage of biofuel blending with the fossil fuel. The prediction results of ANN models in these works had high correlation coefficients $R^2$ of 0.9689 and 0.9994, respectively. In practice, the biofuel ratio is a sensitive operating parameter, which significantly affects engine performance. A higher biofuel ratio can cause higher fuel consumption but lower hydrocarbon, carbon monoxide and particulate matter emission. This results in a multiple-objective optimisation problem balancing fuel consumption and exhaust emission. To solve this problem, extreme learning machine (ELM), a special type of ANN which calculates output weights using a generalised inverse algorithm instead of the backpropagation algorithm (used in a traditional ANN) has been applied.

The ELM is known to have a faster learning speed and better generalisation capacity compared to a traditional ANN [118]. A kernel-based extreme learning machine (K-ELM) was introduced in [118] as an alternative to traditional ANNs for the prediction of the fuel engine performance. The authors also employed another kernel-based method called least-squares support vector machine (LS-SVM) as a baseline to show the effectiveness of the K-ELM model. The accuracy of the K-ELM model was found to be similar or sometimes better than that of the LS-SVM model, but the training and testing times of the K-ELM model were 8 and 16 times faster than that of LS-SVM model. To further reduce the model construction time cost for online process prediction, a new version of ELM named the sparse Bayesian extreme learning machine (SBELM) was developed to model an online fuel engine in [119]. Through verification, the SBELM model was proven to have the same accuracy compared to the conventional ELM model, but takes less training and testing time, and has fewer model parameters making it more favourable as an indicator for the prediction of online fuel engine performance when using biofuel mixtures.

When considering biofuel production, modelling challenges often arise from identifying key operating parameters such as biofuel extraction rate and operating temperature, as well as the interdependence between operating parameters. In recent publications, ANN-based regression techniques and decision tree-based classification methods have been widely used to address these challenges [120] adopted an ANN to model biogas and methane productions in the anaerobic treatment of wastewater molasses. The accuracy of predictions for both biogas and methane production by the ANN model were greater than 90 % [121] tested different types of ANN models for predicting the methane percentage in biogas which was recovered from landfill upon injection of liquid organic waste. The ELM learning method was found to have an outstanding prediction performance with the highest correlation coefficient $R^2$ of 0.88 and lowest mean absolute percentage error (MAPE) of 2.1 % [122] also built accurate ANNs to predict the effect of operating parameters (temperature, moisture content, ash content) on hydrogen yield in a biomass gasification process. In addition, the decision tree-based classifier has been used in [123] to explore the optimal combination of operating parameters for algal biomass growth and lipid synthesis. The authors divided the training data into high or low class of the biomass productivity and lipid content by the median of them. The overall classification accuracies for the validation of high biomass productivity and high lipid content experiments are 89.6 % and 77.4 %, respectively.

In summary, machine learning has been applied to help reduce the waste emission and fuel consumption of combustion engines as well as locating critical operating parameters to maximise performance of a number of biofuel production processes. ANN-based methods are the most commonly used machine learning algorithms in previous bioresources and bioenergy studies, however other machine learning methods such as decision trees have also exhibited increasing popularity in this field, indicating their future in production and utilisation of renewable resources.

### 3.5. Environmental bioengineering

Machine learning, in particular ANN, has also been applied towards addressing serious ecological issues in soils and waterways in 21st century, relating mainly to bioremediation and biological wastewater treatment. For example, in the case of groundwater remediation whereby *in-situ* bioremediation of hydrocarbon contaminants (BTEX) is modelled using computationally expensive partial differential equations (PDEs), ANNs have been applied as a surrogate model to reduce the computational burden and improve predictions of BTEX concentration

levels [124] used the PDE-based BIOPLUME III (a rigorous simulator) to generate data over a three-year bioremediation period, then trained an ANN with input and output as pumping rates and maximum BTEX concentration, respectively. They reported a strong correlation with $R^2 = 0.998$ between the BIOPLUME III output data and the trained ANN. In another study by [125], ANNs were used to model the complex non-linear relationships between soil properties and polycyclic aromatic hydrocarbons (PAHs) dissipation in contaminated soils. Their architecture made use of 8 inputs, each a separate soil property with the trained ANN showing excellent generalisation ($R^2 = 0.948$) and prompt predictions of the PAHs concentrations in soils. This approach was found to be advantageous in reducing the frequency of chemical analysis for on-site bioremediation monitoring purposes, supporting rapid decision-making for bioremediation end points.

The forecasting capability of ANNs in this domain has importantly been contrasted against other machine learning methods. As examples, [126] investigated the ability and performance of six methods in predicting the temporal bioavailability changes of 16 PAHs in contaminated soils amended with compost. Through cross-validation, the overall performance of these six algorithms were ranked with an RBF (radial basis function) found the most accurate and linear regression the least. Regardless, all of these algorithms satisfactorily predicted the PAHs concentrations, with the superiority of SVR over MLP consistent with the findings of [127]. Similarly, a comparison between ANNs and RFs in predicting the temporal bioavailability changes and toxicity of complex chemical mixtures in contaminated soils amended with compost or biochar was investigated by [63]. The results showed that both ANNs and RFs were able to model the bioavailability of various contaminants though RFs performed slightly better in predicting the toxicity as well as providing a relative importance of each input feature.

With respect to hybrid ANNs, ANNs and fuzzy logic framework learning algorithms termed adaptive network-based fuzzy inference system (ANFIS) have been applied to predict the effluent quality of large scale anaerobic wastewater treatment plants (WWTP) [128] used ANFIS to estimate the effluent chemical oxygen demand of a sugar factory WWTP to acceptable standards ($R^2 = 0.8354$) based solely on 5 on-line process inputs. In a similar fashion, the ANFIS algorithm was referred to as a data mining technique by [124] when applied for the prediction of methane production from sludge treatment data which was collected from Des Moines Wastewater Reclamation Facility, the United States. The forecasting power of ANFIS was also compared to that of SVMs by [129] in the prediction of a biological WWTP removal efficiency of Kjeldahl Nitrogen from wastewater. Although the predictions of both algorithms were satisfactory and reasonably comparable with $R^2$ of 0.85 and 0.91 respectively, the SVM in addition defined the interrelationships between various wastewater quality variables. For this reason, SVM was recommended over ANFIS for potential application in evaluating the efficiency of aerobic biological processes in WWTP. In addition, the number of wastewater quality variables constituting the network's input varies from study to study resulting in the dimension of the input vector being large with some components of the vectors being highly correlated and thus redundant. Hence, [130] used PCA to reduce the dimension of an input vector by eliminating the principal components with least contribution to variance in the dataset. This hybridisation improved ANN's mapping abilities and led to satisfactory predictions of the chemical oxygen demand removal efficiency of a reactor for the treated real cotton textile wastewater diluted with domestic wastewater.

Furthermore, [131] combined an ANN with a genetic algorithm for the optimisation of groundwater remediation management. Although the achieved results were comparable to that of conventional simulation-optimisation techniques, the ANN surrogate optimisation approach was reported to be advantageous due to a lower computational burden originating from the parallel processing of flow and transport simulations with the optimisation solver, and the ability to somewhat reuse these simulations as they are parameterised within the ANN. Similar findings were reported by [40]. In another study using a similar procedure, SVM was used for the optimal design of an *in-situ* bioremediation system which was optimised using particle swarm optimisation [132]. The rigorous BIOFDM simulations were replaced with an SVM as a surrogate model. The method yielded an optimal transient pumping strategy resulting in an overall reduction in the pumping cost by almost 20 %. Finally, machine learning has been used to determine the organic carbon and nitrogen content in organic waste and compost. For example, in [133] the researchers constructed PLS to analyse the infrared spectroscopy of waste samples. Two techniques: attenuation total reflectance (PLS-ATR-MIR) and diffuse reflectance (PLS-DR-NIR), were found to show good correlation ability (respective $R^2$ of 0.76 and 0.82), with the former providing better prediction performance.

### 3.6. Metabolic engineering

In the field of metabolic engineering, completing missing information for metabolic reaction network reconstruction is unquestionably one of the top priorities. Given the large size of available metabolic databases, different machine learning methods have been used to address this challenge. For example, [134] adopted naïve Bayes, decision trees, logistic regression, and ensemble methods to predict the existence of potential metabolic pathways. This was achieved by screening over 5000 pathways from different databases to select a number of common pathway features, which were subsequently sent to the machine learning algorithms to predict which metabolic pathways are present in an organism based on the organism's annotated genome. The accuracy of the machine learning methods was validated to be higher than 90 % [135] proposed a matrix factorisation and integer least squares optimisation based algorithm to estimate the missing stoichiometric information for metabolic network model refinement. The algorithm was proposed to fill the gaps of several incomplete stoichiometric matrices based on a universal reactions database, and was found to yield a 60 % accuracy, more than twice that of other existing tools [136] also proposed a GP based automated enzyme search tool to estimate if an enzyme can catalyse a specific metabolic reaction. The accuracy of this GP based classifier was compared against an SVM classifier and was found to be high. This enzyme search tool was also demonstrated to be useful in designing novel pathways for synthetic biology applications.

In addition, machine learning has also found application in manipulating gene expression levels and product synthesis by identifying essential and influential enzymes and genes [137]. For example [138], proposed an ANN based design of experiments framework to maximise cell-free protein synthesis by optimising variations of DNA sequences and amino acids concentrations, yielding 350 % greater production than standard conditions [139] exploited PCA to identify key enzymes, of which expression must be adjusted in order to maximise heterologous terpene production. This was achieved by using PCA to analyse a collection of proteomics data and target molecule production data. By applying the conclusion drawn from PCA, a 40 % improvement in the production of different terpenes was observed. In addition [140], developed an SVM based classifier to predict essential genes in the metabolism of an *E. coli* strain by learning from both reaction-gene and flux-reaction pairs. This method was proved to be superior to previous methods with higher sensitivity and specificity. Furthermore, these machine learning algorithms have also been combined with biological network topological features to predict essential genes and proteins. A more detailed summary can be found at [141].

Within the remit of metabolic engineering, perhaps the most extensive application of machine learning methods has been to quantify the complex interactions across omics ranging from genomics to fluxomics as well as extracellular microbial growth kinetics and culture environment. For instance, [22] proposed a hybrid metabolic flux analysis framework to simulate metabolic networks in the absence of sufficient mechanistic detail. This framework includes a first-principle metabolic flux analysis to simulate the sub-network with well-defined

stoichiometry, and a PLS model to estimate the flux distribution within the unknown or ill-defined sub-network by directly using experimental measurements. A similar idea was also presented by [142] where different machine learning algorithms including SVM, k-nearest neighbours, and decision trees were used to mine relationships between environmental and genetic factors with metabolic fluxes. In reporting the relative performance of these algorithms, the authors find that SVM was the most appropriate one, although this conclusion may not necessarily hold true due to significant recent improvements in the performance of ANNs and RFs.

However, most surprisingly, such a 'hybrid' approach can be traced back to 1990s when [143] developed a hybrid ANN-first principles framework for cell metabolism modelling. This could be the first application of reinforcement learning to the field of metabolic engineering. In this work, the authors constructed a mechanistic model based on mass and energy balance equations to simulate metabolic activity and cell growth kinetics. The ANN was used to estimate some of the mechanistic model parameters. As these parameters could not be measured (thus their true value was unknown), reinforcement learning was used to construct the ANN by increasing the model accuracy iteratively (e.g. selecting the negative of model prediction error as the reward function). However, within the last decade, construction of ANN embedded hybrid models for bioprocess simulation has been well studied. It is now understood that this ANN training problem can be addressed via the chain rule of differentiation [144]. Nonetheless, this was a very challenging task 30 years ago due to the lack of computational power and efficient mathematical algorithms. Hence, the use of reinforcement learning was deemed appropriate and innovative.

Finally, with the rapid development of deep learning algorithms, a range of deep neural networks such as CNN and RNN have been successfully adopted in metabolic engineering to address other types of challenging tasks which are often related to metabolic and molecular database augmentation and development. These draw parallels to natural language processing and image recognition as the underlying system is not primarily governed by quantitative physical correlations. Within 2018–2020, there have been several publications using deep neural networks for molecular genetics open reading frames (ORF) identification and annotation, genomes editing, enzyme and pathway design, and omics data processing [145]. The large number of parameters within a deep learning neural network is particularly advantageous in these cases as the physical relationships between the network's input and output are often hidden within high dimensional data space (e.g. DNA sequences), thus most of the 'shallow' machine learning methods are not suited to this type of task. A comprehensive literature review on these topics can be found at [145].

### 3.7. Cell culture and protein engineering

Machine learning applications within this section can be broadly divided into three sub-areas: protein engineering, enzymatic catalysis, and cell culture engineering.

Firstly, we will focus on the development of protein engineering through machine learning. As early as the 1990s, ANN were used to analyse molecular sequences [35]. In this work, an ANN was deployed in various amino acid and nucleotide sequence analysis. Despite achieving a highest prediction accuracy of 64.3 %, which is only about 10 % more accurate than human-designed systems, it was the first indication that machine learning can be a viable alternative for knowledge and pattern discovery. It also successfully proved the concept of applying machine learning to real-world problems in the bioengineering area. Recently, the ground-breaking result of [44] was presented. By adopting a CNN to predict the distances between pairs of residues, the shape of a protein was constructed from an amino-acid sequence with unprecedented accuracy. The resulting system, named AlphaFold, achieved outstanding accuracy, even for sequences with fewer homologous sequences. In the annual 'critical assessment of methods of protein structure prediction' (CASP13) competition, Alphafold achieved a score 68.3 compared to the next highest score 48.3 [146]. The following year at CASP14 Alphafold performed even better, achieving a score of 90 out of 100 in two-thirds of the structure prediction tasks, causing many figures to deem the protein structure prediction problem 'solved' [147].

Similarly, protein function prediction from protein sequences, protein structures and protein-protein interactions is another promising domain for the inclusion of machine learning. In [148], the authors use a CNN to learn a classification model, named $SDN_2GO$ that extracts features from protein sequences and protein-protein interaction networks to make predictions of gene ontology (GO) terms. The number of nodes in the input, output and hidden layers were set according to different species and GO sub-ontology. The model achieved a favourable F-max value (maximum variance ratio) for each sub-ontology of GO. In addition to ANNs, RFs and SVM are also commonly applied in protein engineering. In [149] SVMs were deployed as a classifiers for sequence feature based prediction of protein stability changes upon amino acid substitutions. The overall predictive accuracy of this classifier was 84.6 %. Furthermore, [150] applied a RF and a GP trained on a protein and ligand information database to predict the protein-ligand binding affinity. The predictive results of the RF model and GP model gave correlation coefficients ($R^2$) of 0.76 and 0.74 respectively, demonstrating the potential of machine learning to identify the binding site for the interaction of protein and ligand based on their structural, physicochemical, and coordinate features.

Secondly, we briefly move to the impact of machine learning in the area of enzymatic catalysis. In [151] enzyme family classes were predicted by adopting SVM based on discrete wavelet transform. A high accuracy prediction of 91.9 % was achieved for the identification of the six enzyme families. [152] applied five different machine learning methods: linear regression, PLS, elastic net, RFs and deep neural networks to predict enzyme turnover numbers. After training the models, the authors demonstrated expected results that the linear modelling techniques (linear regression, PLS and elastic net) have larger cross-validation errors than the nonlinear modelling techniques (RFs and DNNs) for enzyme turnover numbers prediction. The $R^2$ values of the DNN were 0.76 and 0.31 for maximal enzymatic effective turnover rate and turnover numbers, respectively. Finally, we give an example about how machine learning is applied in cell culture engineering area for experimental design and optimisation [153] proposed an ANN for protein glycosylation prediction and applied it into four recombinant glycoproteins produced in Chinese hamster ovary (CHO) cells, two fusion proteins and two monoclonal antibodies. The output of ANN model is glycan distribution profile of the protein, with the authors indicating that the site-specific glycoform distributions of up to eighteen glycan species can be accurately predicted with an average absolute error of only 1.1 %.

In general, according to the previous discussion, the application of machine learning in cell culture and protein engineering is primarily for classification tasks. The prediction accuracies of machine learning methods are often higher than mathematical models constructed through mechanistic means. In addition, many machine learning techniques like SVMs are capable of handling high dimensional data efficiently. However, one of the fundamental limitations of machine learning is the requirement of a large amount of training data. Applying machine learning to small datasets for applications in this area will be a significant challenge in the future. Another challenge is the interpretation of the results [154]. It has to be considered that not only the format or illustration of the output, which is relevant to the interpretation, but also the specifications of the selected algorithm, the parameter settings and data.

## 4. Challenges and prospects for machine learning in biochemical engineering

Whilst machine learning has led to significant contributions across

almost all of the fields within biochemical engineering, there are still some fundamental challenges obstructing its future application. There are also several novel machine learning techniques that have exhibited great potential for biochemical engineering studies and are yet to be fully applied.

### 4.1. Uncertainty quantification

Compared to chemical processes, biochemical processes are generally less reproducible due to their complex and stochastic metabolic reaction networks and enzymatic reaction kinetics. As a result, models used to simulate a biochemical process must have not only high accuracy, but also low uncertainty. Nevertheless, so far only few studies have incorporated uncertainty analysis into machine learning model construction. Indeed, this is not a trivial task and often involves sophisticated statistical and mathematical theories (e.g. calculating the second derivatives of a model). At present, it is not an exaggeration to say that bioreactor engineering is almost the only area within which uncertainty estimation is commonly included during model construction. However, most of these models are mechanistic models as their theory of kinetic model parameter estimation is relatively well established [155]. To estimate the uncertainty of a data-driven model, mainly two approaches can be used.

The first is to select a machine learning method that can automatically estimate uncertainty. This type of machine learning model is often a Bayesian learning based algorithm e.g. Gaussian processes and Bayesian neural networks. Some of their applications have been introduced before. However, these methods are mainly suited to simulate systems with a small amounts of data. Compared to artificial neural networks, the application of Bayesian learning based algorithms in biochemical engineering is less common and is still at an early stage. The second approach is to use bootstrapping. This is applicable to most machine learning algorithms but it can only approximate rather than estimate the true model uncertainty. The main concept behind this approach is to train several machine learning models with the same structure via bootstrapping and then calculate the variation between these models. A recent example of using bootstrapping to estimate the uncertainty of an ANN for bioprocess dynamic simulation can be found in [156]. There are also other similar methods such as the dropout method that has been used to approximate the uncertainty of deep learning neural networks [157]. In summary, uncertainty analysis should be addressed in future studies as it is crucial to assist in many decision-making tasks, including data-driven model structure identification, robust bioprocess optimisation, metabolic pathways prediction, and design of experiments.

### 4.2. Data quantity and quality

The most common application of machine learning is to simulate systems with substantial amounts of data. However, most studies within biochemical engineering deal with small sized datasets. Although intuitively this was a challenge for machine learning model construction, in practice there have been different data augmentation strategies proposed to solve this problem. The most straightforward one is to add ‘random’ noise into the original data. Due to the inherent stochasticity of bioprocesses, it can be assumed that any data point may deviate within a proper range from the original and still capture the underlying physics of the process. This ‘noise’ range could come from deviations of experimental measurements or empirical knowledge. Thus, a large amount of synthetic data can be generated from an often small set of experimental data. This dataset augmentation strategy has been found to be efficient for machine learning model construction for bioprocess modelling as well as soft-sensing [78,55]. It is important to highlight that the reason why a machine learning model requires a large amount of data is because of the large number of model parameters (e.g. hundreds to thousands of parameters in a neural network). From a mathematical perspective, there must be enough training data points to avoid overfitting. The nature of training data (e.g. experimental or synthetic), on the other hand, is irrelevant to machine learning model parameter estimation, as long as the data is physically correct.

Parallels to this viewpoint are observed in the artificial image generation techniques, which have created excitement in the machine learning community and public domain. Specifically, these methods use generative adversarial networks (GANs) to artificially synthesise new images from an existing dataset [158]. Recently, GANs have been translated to augment different data types including time-series datasets [159], providing a statistical alternative to the use of process knowledge for decision regarding the noise range. This could provide a useful technique when system stochasticity is difficult to quantify or even approximate, although to our knowledge such an approach has not yet been implemented within the field. Another strategy which can also embed process prior knowledge into data augmentation is to build several simple first-principle models (e.g. classical kinetic models) to fit the original experimental datasets separately [160]. We can then use these first-principle models to simulate a number of bioprocesses conducted under similar conditions to the original experiments. In this way, synthetic data can be generated. Compared to the random noise approach, this method can offer additional physical insight into synthetic data and enhance the accuracy of the machine learning model; however, it is also more time consuming as even minimising the mismatch between a simple kinetic model and experimental data could be difficult. Thus, this approach has not been commonly used.

In terms of other fields such as metabolic engineering and protein engineering, there has been comprehensive databases made available to disseminate large volumes of data for machine learning model construction. However, one of the main issues in this case is the compatibility of different databases. Many databases lack biological data standards to share information, and there may be many missing records within each database. This is a common issue encountered by different scientific domains for chemical synthesis, molecular design, and process upscaling. A promising approach for the sharing of data within and cross-domains is the creation of ontologies that contribute towards the semantic web of data. Similar to the World-Wide-Web of documents, the semantic web stores the connections between data. These ontologies describe the concepts (nodes) and relations (connections) between items of data allowing for their interactions to be understood and allowing for the relationships between data across domains to be modelled. Whilst this approach has found use in chemical processes [161,162], as well as the medical domain [163], the authors are yet to be made aware of an investigation as to the advantages of representing bioprocess data using an ontology. This potentially may not only provide a FAIR (findable, accessible, interpretable, reusable) way of representing data for machine learning research in biochemical engineering, but once data has been represented in this way techniques such as the embedding of knowledge graphs may be applicable to discover new bioprocess data.

Another issue of data quality is that many high-dimensional datasets are sparse, meaning that they do not contain enough information for each individual dimension (e.g. time dimension for dynamic system simulation or some feature dimensions for de novo protein design) for machine learning model construction. To solve this issue, dimensionality reduction and feature selection must be adopted to reduce the data into a lower dimension [164]. Inevitably, this will lead to the loss of essential system knowledge for machine learning model construction, thus undermining model accuracy [145]. Data augmentation in this case is not effective due to the sparsity of original datasets; the only possible way to resolve this issue is to use rigorous first-principle models if possible to resolve the knowledge gaps.

### 4.3. Integrating bioprocess knowledge

Regardless of the type of machine learning method, the nature of data-driven modelling dictates a lack of physical insight and is therefore

incapable of accurately extrapolating bioprocess behaviours. To expand the use of machine learning for future research, it is extremely important to integrate machine learning with known process mechanisms. At present, there are two major strategies to accomplish this task. The first which is also the most extensively investigated approach is to embed machine learning models into first-principle physical models [144]. These models are often termed 'hybrid', and the machine learning models are either used to simulate parameters of first-principle models or are inserted as part of a first-principle model (e.g. constraints) for process prediction [165]. Some applications of this strategy include: the estimation of kinetic model parameters for bioreactor dynamic modelling [166]; quantification of complex relationships between culture conditions and metabolic fluxes for metabolic flux analysis [22]; rectifying mismatch between mechanistic model and experimental data for high-value bioproducts optimisation [167]; and, integration with mass balance equations for soft-sensor development [168]. These hybrid models often show higher predictive power and data efficiency compared to purely physical or data-driven models, and are robust to small datasets with low quality (e.g. noisy data).

Although the concept of hybrid modelling is applicable to all fields within biochemical engineering, the use of this strategy depends on the availability of physical models for a specific application. For instance, although there have been many physical models developed for bioprocess kinetic modelling [2,169], biodevice design [170], and metabolic reaction network construction [171], there are few physical models proposed in other areas such as cell culture engineering, protein engineering and biomaterials engineering. As a result, hybrid models are not yet applicable to these areas. In addition, training a machine learning model encapsulated within a physical model is also more challenging than building a machine learning model alone [165]. At present, most of the hybrid models only use artificial neural networks as the principle data-driven aspect [166]. However, given the importance of model uncertainty estimation, it is envisaged that other advanced machine learning algorithms such as GPs and ensemble learning methods will also have bright futures in hybrid model construction.

The second strategy is to use synthetic data to develop a machine learning model. However, in this case the synthetic data is generated from rigorous mechanistic models which are highly accurate but could be computationally expensive. For example, computational fluid dynamics and molecular dynamics are high fidelity first-principle models which have been used for bioreactor modelling and protein structure simulation, respectively. Nonetheless, these models are time consuming to run and cannot be used for real-time applications (e.g. fast screening of protein structure or real-time bioreactor control). To resolve this challenge, we can first use these rigorous models to generate synthetic data and then train a machine learning model to substitute them for practical application. This strategy has been studied in different domains and hence has different names, for example, in process systems engineering it is referred as surrogate modelling; whilst in data science it is termed emulation.

Previously, a CNN has been constructed to learn from a CFD model and was used to optimise the configuration and operating conditions of a pilot scale bioreactor for biofuel production [87]. An RF has also been built to learn from a mechanical protein network model and was applied to predict the mechanical behaviour of new protein networks [65]. Surrogate modelling can greatly reduce the computational time cost from months or years down to hours or seconds, but it is also restricted by the availability of rigorous physical models. As a result, it is promising to resolve some industrial challenges (e.g. visualising and optimising large scale manufacturing plants) but is not effective if the system's underlying relationship cannot be accurately quantified by mechanistic models.

### 4.4. Novel machine learning algorithms

Several 'novel' machine learning algorithms such as deep learning, reinforcement learning, and transfer learning have also been recently adopted within chemical and biochemical process control [172], optimisation [70], fault detection [173], and metabolic pathways optimisation [174]. Although the theory of these machine learning algorithms has been relatively well established during the last decade, they are still considered 'novel' algorithms since their application in biochemical engineering has not been reported until the last few years.

#### 4.4.1. Deep learning

Deep learning is particularly efficient to extract information from high dimensional datasets (e.g. spectroscopic data, DNA sequence data) and tensoral datasets such as images [48]. It has been primarily used for audio and speech recognition, natural language processing, medical image analysis, and bioinformatics [175]. The large number of parameters inside a deep learning model can effectively quantify the 'qualitative' relations between input variables and system response (e.g. enzyme commission number, protein features, gene expression activity) [145]. As a result, they are an excellent complement to hybrid and surrogate modelling techniques.

So far, almost all the deep learning studies only focus on different types of neural networks [175]. However, other deep learning techniques should also be given enough attention. For example, deep Gaussian processes (also termed composite Gaussian processes) are particularly effective for multi-fidelity data simulation, where the dataset consists of a large portion of low accuracy data and a small portion of high accuracy data [176]. Most industrial bioprocess data falls within this category. Similarly, ensemble learning methods such as random forests can also be considered as a deep learning version of decision trees. The structure of ensemble learning can enable the estimation of model uncertainty and can be advantageous for classification tasks commonly seen in metabolic engineering, biomaterials engineering, and environmental bioengineering [31]. A downside of deep learning is its risk in overfitting due to its large number of model parameters and difficulty in justifying the fitness of a qualitative input-output correlation. Hence, an efficient hyperparameter optimisation framework and thorough model validation procedure must be developed and implemented [177].

#### 4.4.2. Reinforcement learning

Reinforcement learning, specifically model-free reinforcement learning, is an approach to learning an optimal control policy under uncertain or stochastic process dynamics. At present, there have been several excellent reviews for this technique [178–180]. At a high level, model-free RL relies on performing Monte Carlo realisations of different control policies under the process dynamics to determine which one is optimal, instead of relying upon explicit knowledge of the process dynamics. This relative independence of the control solution from explicit knowledge of the process dynamics is particularly appealing in biochemical engineering. Typically, the systems of concern are nonlinear and often exhibit aleatoric uncertainty, which poses difficulty to structural and practical identification of mechanistic models for subsequent model-based optimisation and design tasks. As such, RL poses a potential data-driven control solution [70,181] or optimal design method [182] for a wide variety of biochemical engineering applications.

However, the application of model-free RL to biochemical engineering is faced with a number of challenges. A primary one is the safe guarantee of operational constraints. A general RL solution policy maximises an objective of expected reward but does not provide an explicit mechanism for constraint satisfaction, thus introducing a risk on bioprocess operation as microbial and mammalian cells are highly sensitive to the culture environments. Some methods have been proposed to address this issue [70,181], although few achieve high probability constraint satisfaction. Another challenge is associated with the data efficiency of RL algorithms. Model-free RL algorithms rely on samples or observations of system response under a control policy.

Given the typical stochasticity of the systems concerned, learning an optimal policy requires many iterations of trial-and-error experiments. Acquisition of large amounts of real-world data is expensive and infeasible in biochemical engineering. Therefore, in practice a process model (either data-driven or mechanistic) must be constructed to initialise a relatively accurate RL policy through offline learning. A novel framework to address this challenge is apprenticeship learning via inverse reinforcement learning, where an initialisation of RL (i.e. apprenticeship learning model) can be developed based on historical process data [183]. A brief introduction of this approach is presented in the next section.

Despite these challenges, RL presents a high potential research area. Provided it continues to receive an increasing interest from chemical engineering, process systems engineering, and biochemical engineering, there is no reason progress cannot be made to successfully adopt RL into diverse biochemical engineering applications.

#### 4.4.3. Transfer, continual and imitation learning

Transfer Learning (TL) and Continual Learning (CL) are different approaches to preserving and transferring knowledge, when provided with new data, by modifying model parameters or structure built from existing data. TL enables the translation of knowledge from a previous inference task to a new one - where the data source is different. It has observed deployment outside of the field of RL in biosensing [102]. Knowledge is typically translated to a new task via the parameters and topology of the data-driven model. Conversely, CL underpins learning on new data as it is received from an ongoing process. This new data is assumed to be from the same source, and so CL is concerned with maximising the information abstracted from new data, without 'forgetting' knowledge learned previously [184].

Whilst TL and CL are concerned with handling the dynamic nature and diversity of a new process, the final concept we raise here looks to extract expert knowledge from existing process datasets. Specifically, imitation learning is a technique to extracting a control policy from an existing dataset. As reminder, a control policy is the selection of a control input to a system, conditional upon an observation of that system. Existing industrial datasets, which are the result of an interaction between an ongoing process and some kind of decision-making element, ultimately contain information about experts' experience (e.g. decisions made by bioprocess engineers for bioreactor control). Imitation learning aims to extract and quantify this experience. There are a number of approaches to imitation learning - here we briefly discuss behavioural cloning (BC) and apprenticeship learning via inverse reinforcement learning (IRL). BC synthesises a control policy by learning input-output mappings expressed by a dataset. The inputs in this case may be the system state, and the outputs may be the control inputs selected by the existing control policy. Supervised learning is then deployed in order to identify a model, which best represents the control policy expressed by the process data [185]. Although a very intuitive procedure, BC is subject to the same pitfalls of pure data-driven modelling. In real bioprocess applications, this can be particularly problematic if control inputs are predicted for system states never seen before.

It is accepted that IRL, however, does not fall foul of this problem to the same extent. Instead of directly learning input-output mappings, IRL aims to abstract a reward function (e.g. a bioprocess engineer's empirical knowledge), where the existing control policy is considered optimal [186]. This enables the learning of an RL based policy (hence apprenticeship learning), which provides a parameterisation of the existing control policy expressed in data [183]. However, this typically demands a model (could be data-driven) of the real underlying process [187] as discussed in the previous section. Despite this requirement, IRL provides promise to robustly extract expert domain knowledge contained in existing datasets. For more information on imitation learning we refer the reader to this excellent review [188].

Previously, we have highlighted two works [189,102], where the authors deployed TL effectively. However, we are not aware of many other works within the field, which have demonstrated the concepts (e.g. continual and imitation learning) discussed in this section. This could be partly due to the nature of biochemical engineering research, where access to large or evolving datasets is rare and therefore, it is likely that much of the developments made in these areas will be driven directly by the bioprocess industry.

## 5. Conclusions

Fundamentally, a model is constructed to serve three purposes: to assist knowledge discovery, to predict system behaviour, and to save time. Irrelevant to the type of a machine learning algorithm, for biochemical engineering studies, it is always the knowledge behind the data rather than the data itself that we are looking for. No matter which machine learning algorithm is used for bioprocess data classification or regression, if its result cannot provide additional physical insight or facilitate domain knowledge generation, then it will have no bright future in biochemical engineering research. Given the data-driven nature of machine learning, purely using it to reveal the hidden knowledge of an unknown bioprocess could be wishful thinking. For example, most dimensionality reduction methods are unsupervised learning, thus the identified latent space may have no physical correlation with respect to the system response. As a result, increasing the prediction accuracy and reducing the uncertainty of machine learning through integration of process knowledge via different means is the most critical step to enable its continuous application in future studies. Hence, along with exploring other advanced machine learning methods, constructing novel physical models particularly for the fields which are currently lacking theoretical understanding must be also given high priority.

It is also worth noting that without proper documentation of machine learning models their results can often be difficult to reproduce. Code-level optimisation, meaning the fine-tuning of aspects of the model such as hyperparameters by hand, means that results can differ wildly between seemingly similar models or training procedures. Producing rigorous and transparent descriptions of machine learning models and allowing for results to be reproduced is essential for convincing the non-academic biochemical engineering community as to their benefits. This will enable their adoption within industrial settings, thus motivating further research in the field supporting both the academic and industrial community.

From a historical point of view, it is worth highlighting that the concept of integrating data-driven models with physical models is not new. This concept has existed since the late 19th Century and has been widely used in chemical engineering. If we look at an undergraduate thermodynamics text book, we can find that the compressibility coefficient is used to modify the ideal gas law (first-principle model) to account for the real gas behaviour. This physical model parameter can be calculated by the Virial equation, which is a polynomial regression model (a data-driven model) used to quantify the temperature effect on gas compressibility. In fact, the popularisation of using machine learning methods for hybrid and surrogate model construction is largely attributed to the rapid development of computing facilities rather than themselves being an 'innovative' modelling concept. In addition, to build an accurate machine learning model, we must also understand the critical role of essential mathematical theories (e.g. optimisation theory, multivariate statistics) for model training and validation.

Finally, whilst we are excited about the recent exhibitions of deep learning techniques for omics data processing, enzyme annotation, metabolic pathways prediction, bioprocess control and optimisation, and biosystem visualisation, we should be aware that deep learning methods are not always appropriate for all types of biochemical engineering applications without cautious justification. Combined with physical models, conventional 'shallow' learning methods have shown high accuracy and great success in a range of applications, and they are more flexibly integrated with physical models. The 'deepness' of a machine learning algorithm shall not be used as the only metric to indicate

its efficiency, novelty and suitability for any biochemical engineering case studies. Whilst deep learning algorithms can enjoy their superiority in analysing high-dimensional and large datasets, they are also more prone to overfitting and are still constructed without physical knowledge. As most biochemical processes are strongly governed by physical mechanisms, in many cases the use of 'shallow' learning methods is already accurate enough.

As a result, two principles should come to mind when applying machine learning models to biochemical engineering research. The first is George Box's famous saying that "all models are wrong, but some are useful". Machine learning techniques can be extremely useful when used correctly but as previously mentioned they are not magic, and the size or technicality of a model is not indicative of its power or usefulness, which motivates the second principle: that of Ockham's Razor stating "the simplest model is often the right one". It may be tempting to adapt an existing complicated machine learning model or simply increase the number of model parameters in order to get a better fitting result; however, this serves to dilute the quality of research and reduce its potential impact. Overall, in view of the excitement surrounding the field of machine learning it is important to acknowledge that all techniques have their benefits, with no one class of model considered to be universally better than another. It is thought that if the principles discussed are adhered to, the emerging relationships between shallow, deep, mechanistic and hybrid modelling will continue to be fruitful and complementary in nature. A smart combination of these modelling tools will also enable the design of advanced digital twins that can greatly accelerate the development of a biotechnology from lab-scale reaction and pathways identification to industrial-scale operation and optimisation.

## Declaration of Competing Interest

The authors report no declarations of interest.

## References


[1] R. Harrington, Growing the Bioeconomy, HM Government, 2019.
[2] I. Harun, E.A. Del Rio-Chanona, J.L. Wagner, K.J. Lauersen, D. Zhang, K. Hellgardt, Photocatalytic production of bisabolene from green microalgae mutant: process analysis and kinetic modeling, Ind. Eng. Chem. Res. 57 (8) (2018) 10336–10344.
[3] L. Mears, S.M. Stocks, M.O. Albaek, G. Sin, K.V. Gernaey, Mechanistic fermentation models for process design, monitoring, and control, Trends Biotechnol. 35 (10) (2017) 914–924.
[4] M.R. Antoniewicz, Methods and advances in metabolic flux analysis: a mini-review, J. Ind. Microbiol. Biotechnol. 42 (2015) 317–325.
[5] E.A.D. Rio-Chanona, J.L. Wagner, H. Ali, F. Fiorelli, D. Zhang, K. Hellgardt, Deep learning-based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design, Aiche J. 65 (3) (2018) 915–923.
[6] J. Damborsky, J. Brezovsky, Computational tools for designing and engineering enzymes, Biocataly. Biotrans. Bioinorg. Chem. 19 (4) (2014) 8–16.
[7] N.K. Tripathi, A. Shrivastava, Scale up of biopharmaceuticals production, Chapter 4, in: A.M. Grumezescu (Ed.), Nanoscale Fabrication, Optimization, Scale-Up and Biological Aspects of Pharmaceutical Nanotechnology, William Andrew Publishing, 2018, pp. 133–172.
[8] F. Delvigne, H. Noorman, Scale-up/Scale-down of microbial bioprocesses: a modern light on an old issue, Microb. Biotechnol. 10 (7) (2017) 685–687.
[9] A. Antonakoudis, R. Barbosa, P. Kotidis, C. Kontoravdi, The era of big data: genome-scale modelling meets machine learning, Comput. Struct. Biotechnol. J. 18 (2020) 3287–3300.
[10] P. Natarajan, R. Moghadam, S. Jagannathan, Online deep neural network - based feedback control of a Lutein bioprocess, J. Process Control 98 (2021) 41–51.
[11] B. Sandui, O.A. Ramirez Calderon, O.M. Abdeldayem, J. Lazaro-Gil, E.R. Rene, U. Sambuu, Applications of machine learning algorithms for biological wastewater treatment: updates and perspectives, Clean Technol. Environ. Policy 23 (2021) 127–143.
[12] K. J.Voorhees, S.J. DeLuca, A. Noguerola, Identification of chemical biomarker compounds in bacteria and other biomaterials by pyrolysis—tandem mass spectrometry, J. Anal. Appl. Pyrolysis 24 (1) (1992) 1–21.
[13] I. Notingher, G. Jell, U. Lohbauer, V. Salih, L.L. Hench, In situ non-invasive spectral discrimination between bone cell phenotypes used in tissue engineering, J. Cell. Biochem. 92 (6) (2004) 1180–1192.
[14] V. V.Nair, H. Dhar, S. Kumar, A.K. Thalla, S. Mukherjee, J. W.C.Wong, Artificial neural network based modeling to evaluate methane yield from biogas in a laboratory-scale anaerobic bioreactor, Bioresour. Technol. 217 (2016) 90–99.
[15] S. Marklund, E. Wikström, G. Löfvenius, I. Fängmark, C. Rappe, Emissions of polychlorinated compounds in combustion of biofuel, Chemosphere 28 (10) (1994) 1895–1904.
[16] J. C.Gunther, D. E.Seborg, J. S.Conner, Fault detection and diagnosis in industrial fed-batch cell culture, IFAC Proc. Vol. 39 (2) (2006) 203–208.
[17] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A 374 (2065) (2016).
[18] M.D.C. Nicoletti, L.C. Jain, Computational Intelligence Techniques for Bioprocess Modelling, Supervision and Control, Springer, 2009.
[19] C. A.Duran-Villalobos, S. Goldrick, B. Lennox, Multivariate statistical process control of an industrial-scale fed-batch simulator, Comput. Chem. Eng. 132 (2020).
[20] S. Stubbs, J. Zhang, J. Morris, BioProcess performance monitoring using multiway interval partial least squares, Chapter 10, in: R. Singh, Z. Yuan (Eds.), Process Systems Engineering for Pharmaceutical Manufacturing, Computer Aided Chemical Engineering, 2018, pp. 243–259.
[21] R. S.Freirea, M.M. Ferreira, N. Durán, L. T.Kubota, Dual amperometric biosensor device for analysis of binary mixtures of phenols by multivariate calibration using partial least squares, Anal. Chim. Acta 485 (2) (2003) 263–269.
[22] N. Carinhas, V. Bernal, A.P. Teixeira, M.J. Carrondo, P.M. Alves, R. Oliveira, Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products, BMC Syst. Biol. 5 (2011).
[23] K.K. Yang, Z. Wu, F.H. Arnold, Machine-learning-guided directed evolution for protein engineering, Nat. Methods 16 (2019) 687–694.
[24] J.G. Greener, L. Moffat, D.T. Jones, Design of metalloproteins and novel protein folds using variational autoencoders, Sci. Rep. 8 (2018).
[25] X. Zhu, K.U. Rehman, B. Wang, M. Shahzad, Modern soft-sensing modeling methods for fermentation processes, Sensors 20 (6) (2020).
[26] S.G. Wu, Y. Wang, W. Jiang, T. Oyetunde, R. Yao, X. Zhang, K. Shimizu, Y. J. Tang, F.S. Bao, Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming, PLoS Comput. Biol. (2016).
[27] M. Gao, J. Tian, J. Li, The Study of Membrane Fouling Modeling Method Based on Support Vector Machine for Sewage Treatment Membrane Bioreactor, Harbin, China, 2007.
[28] S.W.-H. Wenonah Vercoutere, H. Olsen, D. Deamer, D. Haussler, M. Akeson, Rapid discrimination among individual DNA hairpin molecules at single-nucleotide resolution using an ion channel, Nat. Biotechnol. 19 (2001) 248–252.
[29] H.-T. Lin, C.-J. Lin, R.C. Weng, A note on Platt's probabilistic outputs for support vector machines, Machine Learning volume 68 (2007) 267–276.
[30] Carl Edward Rasmussen, K.I. Christopher, Williams, Gaussian Processes for Machine Learning, MIT Press, 2005.
[31] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, 2 ed., Springer, 2009.
[32] J. Platt, Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, 1998 [Online]. Available: https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/. [Accessed 17 01 2021].
[33] S. Si, C.-J. Hsieh, I.S. Dhillon, Memory efficient kernel approximation, J. Mach. Learn. Res. 18 (20) (2017) 1–32.
[34] J. Thibault, V.V. Breusegem, A. Chéruy, On-line prediction of fermentation variables using neural networks, Biotechnol. Bioeng. 36 (10) (1990) 1041–1048.
[35] P.K. Chan, Machine Learning in Molecular Biology Sequence Analysis, Columbia University Computer Science Technical Reports, 1991.
[36] E. Callaway, It Will Change Everything': DeepMind's AI Makes Gigantic Leap in Solving Protein Structures, 2020 [Online]. Available: https://www.nature.com/articles/d41586-020-03348-4. [Accessed 12 01 2021].
[37] G. Mittal, J. Zhang, Prediction of freezing time for food products using a neural network, Food Res. Int. 33 (7) (2000) 557–562.
[38] B. Guo, D. Li, C. Cheng, Z.-a. Lu, Y. Shen, Simulation of biomass gasification with a hybrid neural network model, Bioresour. Technol. 76 (2) (2001) 77–83.
[39] M. Quaglio, M. Roberts, M.S. Bin Jaapar, E.S. Fraga, V. Dua, F. Galvanin, An artificial neural network approach to recognise kinetic models from experimental data, Comput. Chem. Eng. 135 (2020).
[40] R.K. Prasad, S. Mathur, In-Situ Bioremediation of Contaminated Groundwater Using Artificial Neural Network, World Environmental and Water Resources Congress, 2006.
[41] G. Cybenko, Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals and Systems volume 2 (1989) 303–314.
[42] P. Werbos, Backpropagation through time: what it does and how to do it, Proc. Ieee 78 (10) (1990) 1550–1560.
[43] A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, E. I. Zacharaki, EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation, Bioinform. Genomics (2018).
[44] A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710.
[45] M. Oubounyt, Z. Louadi, H. Tayara, K.T. Chong, DeePromoter: robust promoter predictor using deep learning, Front. Genet. 10 (2019) 286.
[46] E.C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, G.M. Church, Unified rational protein engineering with sequence-based deep representation learning, Nat. Methods 16 (2019) 1315–1322.
[47] S. Han, T. Kim, D. Kim, Y.-L. Park, S. Jo, Use of deep learning for characterization of microfluidic Soft sensors, IEEE Robot. Autom. Lett. 3 (2) (2018) 873–880.
[48] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[49] G. Eraslan, L.M. Simon, M. Mircea, N.S. Mueller, F.J. Theis, Single-cell RNA-seq denoising using a deep count autoencoder, Nat. Commun. 10 (2019).

[50] H.-I.H. Chen, Y.-C. Chiu, T. Zhang, S. Zhang, Y. Hueng, Y. Chen, GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization, BMC Syst. Biol. 12 (2018) 142.
[51] E. Bradford, L. Imsland, D. Zhang, E.Ad.R. Chanona, Stochastic data-driven model predictive control using gaussian processes, Comput. Chem. Eng. 139 (2020).
[52] Fd. Sciascio, A.N. Amicarelli, Biomass estimation in batch biotechnological processes by Bayesian Gaussian process regression, Comput. Chem. Eng. 32 (12) (2008).
[53] M.O.H.N. Yutaka Saito, T. Niide, T. Kameda, K. Tsuda, M. Umetsu, Machine-learning-Guided mutagenesis for directed evolution of fluorescent proteins, ACS Synth. Biol. 7 (9) (2018) 2012–2022.
[54] H. Liu, Y.-S. Ong, X. Shen, J. Cai, When gaussian process meets big data: a review of scalable GPs, IEEE Trans. Neural Netw. Learn. Syst. 31 (11) (2020).
[55] A. Tulsyan, C. Garvin, C. Undey, Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems, Biotechnol. Bioeng. 115 (8) (2018) 1915–1924.
[56] T.G. Dietterich, Ensemble learning. The Handbook of Brain Theory and Neural Networks, 2002, pp. 110–125.
[57] R. Polikar, Ensemble based systems in decision making, IEEE Circuits Syst. Mag. 6 (3) (2006) 21–45.
[58] O. Sagi, L. Rokach, Ensemble Learning: A Survey, Vol. 1249, WIREs Data Mining and Knowledge Discovery, 2018.
[59] G.T. Ooi, S.H. Paik, E.W. Ferguson, Molecular Signature of Biological Pathogens, SUN BIOMEDICAL TECHNOLOGIES RIDGECREST, CA, 2003.
[60] P.M. Bapat, P.P. Wangikan, Optimization of rifamycin B fermentation in shake flasks via a machine-learning-based approach, Biotechnol. Bioeng. 86 (2) (2004) 201–208.
[61] S.M.R. Miraboutalebi, P. Kazemi, P. Bahrami, Fatty Acid Methyl Ester (FAME) composition used for estimation of biodiesel cetane number employing random forest and artificial neural networks: a new approach, Fuel 166 (2016) 143–151.
[62] A.B. Farimani, M. Heiranian, N.R. Aluru, Identification of amino acids with sensitive nanoporous MoS2: towards machine learning-based prediction, Npj 2d Mater. Appl. 2 (14) (2018).
[63] S. Cipullo, B. Snapir, G. Prpich, P. Campo, F. Coulon, Prediction of bioavailability and toxicity of complex chemical mixtures through machine learning models, Chemosphere (2019) 388–395.
[64] T. Sun, Q. Li, Y. Xu, Z. Zhang, L. Lai, J. Pei, Prediction of Liquid-Liquid Phase Separation Proteins Using Machine Learning, 2020.
[65] P. Asgharzadeh, A.I. Birkhold, Z. Triverdi, B. Özdemir, R. Reski, O. Röhrle, A nanoFE simulation-based surrogate machine learning model to predict mechanical functionality of protein networks from live confocal imaging, Comput. Struct. Biotechnol. J. 18 (2020) 2774–2788.
[66] C. Zhang, Y. Ma, Ensemble Machine Learning, Springer, 2012.
[67] A.C. Müller, S. Guido, Introduction to Machine Learning With Python, O-Reilly Media, 2016.
[68] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing Atari With Deep Reinforcement Learning, 2013.
[69] P.-C. Fu, J. Barford, A hybrid neural network—first principles approach for modelling of cell metabolism, Comput. Chem. Eng. 20 (6–7) (1996) 951–958.
[70] P. Petsagkourakis, I.O. Sandoval, E. Bradford, F. Galvanin, D. Zhang, E.A. del Rio-Chanona, "Chance constrained policy optimization for process control and optimization, arXiv (2020).
[71] Y. Ma, D.A. Norena-Caro, A.J. Adams, T.B. Brentzel, J.A. Romagnoli, M. G. Benton, Machine-learning-based simulation and fed-batch control of cyanobacterial-phycocyanin production in Plectonema by artificial neural network and deep reinforcement learning, Comput. Chem. Eng. (2020).
[72] S. Nikita, A. Tiwari, D. Sonawa, T. Kodamana, A.S. Rathore, Reinforcement learning based optimization of process chromatography for continuous processing of biopharmaceuticals, Chem. Eng. Sci. 230 (2021).
[73] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G.V.D. Driessche, J. Schrittwiser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, Mastering the Game of Go With Deep Neural Networks and Tree Search, Vol. 529, 2016, pp. 484–489.
[74] D. Solle, B. Hitzmann, C. Herwig, M. Pereira Remelhe, S. Ulonska, L. Wuerth, A. Prata, T. Steckenreiter, Between the Poles of Data-Driven and Mechanistic Modeling for Process Operation, Chemie Ingenieur Tech. 89 (5) (2017).
[75] S. Subashchandrabose, L. Wang, V. Kadiyala, R. Naidu, M. Mallavarapu, Interactive effects of PAHs and heavy metal mixtures on oxidative stress in Chlorella sp. MM3 as determined by artificial neural network and genetic algorithm, Algal Res. 21 (2017) 203–212.
[76] A. Zhang, K. Zhu, X. Zhuang, L. Liao, S. Huang, A robust soft sensor to monitor 1,3-propanediol fermentation process by Clostridium butyricum based on artificial neural network, Biotechnol. Bioeng. 117 (11) (2020).
[77] Y.-H. Zhu, T. Rajalahti, S. Linko, Application of neural networks to lysine production, Chem. Eng. J. Biochem. Eng. J. 62 (3) (1996) 207–214.
[78] E.Ad. Rio-Chanona, F. Fiorelli, D. Zhang, N.R. Ahmed, K. Jing, N. Shah, An efficient model construction strategy to simulate microalgal lutein photo-production dynamic process, Biotechnol. Bioeng. 114 (11) (2017) 2518–2527.
[79] E. Bradford, A.M. Schweidtmann, D. Zhang, K. Jing, E.Ad. Rio-Chanona, Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes, Comput. Chem. Eng. 118 (2018) 143–158.
[80] D.C. Psichogios, L.H. Ungar, A hybrid neural network-first principles approach to process modeling, AIChE J. 38 (10) (1992) 1499–1511.
[81] J. Schubert, R. Simutis, M.D.I. Havlik, A. Lübbert, Bioprocess optimization and control: application of hybrid modelling, J. Biotechnol. 35 (1) (1994) 51–68.
[82] E.Ad. Rio-Chanona, E. Manirafasha, D. Zhang, Q. Yue, K. Jing, Dynamic modeling and optimization of cyanobacterial C-phycocyanin production process by artificial neural network, Algal Res. 13 (2016) 7–15.
[83] A. Ebrahimpour, R.N.Z.R.A. Rahman, D.H.E. Ch`ng, M. Basri, A.B. Salleh, 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 (2008).
[84] K.M. Desai, S.A. Survase, P.S. Saudagar, S. Lele, R.S. Singhal, Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: case study of fermentative production of scleroglucan, Biochem. Eng. J. 41 (3) (2008).
[85] Y. Nagata, K.H. Chu, Optimization of a fermentation medium using neural networks and genetic algorithms, Biotechnol. Lett. 25 (2003) 1837–1842.
[86] F. Baishan, C. Hongwen, X. Xiaolan, W. Ning, H. Zongding, Using genetic algorithms coupling neural networks in a study of xylitol production: medium optimisation, Process. Biochem. 38 (7) (2003) 979–985.
[87] E.A. Del Rio-Chanona, J.L. Wagner, H. Ali, F. Fiorelli, D. Zhang, K. Hellgardt, Deep learning-based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design, AIChE J. 65 (3) (2018) 915–923.
[88] Z.K. Nagy, Model based control of a yeast fermentation bioreactor using optimally designed artificial neural networks, Chem. Eng. J. 127 (1–3) (2007) 95–109.
[89] N.B. Özkaya, A. Visa, C.-Y. Lin, J.A. Puhakka, O. Yli-Harja, An artificial neural network based model for predicting H2 production rates in a sucrose-based bioreactor system, Int. J. Math. Phys. Eng. Sci. (2007).
[90] P. Holubar, L. Zani, M. Hager, W. Fröschl, Z. Radak, R. Braun, Advanced controlling of anaerobic digestion by means of hierarchical neural networks, Water Res. 36 (10) (2002) 2583–2588.
[91] B.E. Logan, Microbial Fuel Cells, John Wiley & Sons, Inc., 2007.
[92] A. Tardast, M. Rahimnejad, G. Najafpour, A. Ghoreyshi, G.C. Premier, G. Bakeri, S.-E. Oh, Use of artificial neural network for the prediction of bioelectricity production in a membrane less microbial fuel cell, Fuel 117 (A) (2014) 697–703.
[93] F. Fang, G.-L. Zang, M. Sun, H.-Q. Yu, Optimizing multi-variables of microbial fuel cell for electricity generation with an integrated modeling and experimental approach, Appl. Energy 110 (2013) 98–103.
[94] A. Garg, V. Vijayaraghavan, S. Mahapatra, K. Tai, C.H. Wong, Performance evaluation of microbial fuel cell by artificial intelligence methods, Expert Syst. Appl. 41 (4) (2014) 1389–1399.
[95] K. Wang, Q. Wang, Intelligent explicit model predictive control based on machine learning for microbial desalination cells, Proc. Inst. Mech. Eng. Part I J. Syst. Control. Eng. 233 (7) (2019) 751–763.
[96] G. Alterovitz, R.M. Benson, M. Ramoni, Automation in Proteomics and Genomics: An Engineering Case-Based Approach, Wiley, 2009.
[97] J. Rofeh, S. Schankweiler, D. Morton, S. Mortezaei, L. Qiang, J. Gundlach, J. Fisher, L. Theogarajan, Microfluidic block copolymer membrane arrays for nanopore DNA sequencing, Appl. Phys. Lett. 114 (21) (2019).
[98] X. Han, K. Yang, R.W. Gross, Microfluidics-based electrospray ionization enhances the intrasource separation of lipid classes and extends identification of individual molecular species through multi-dimensional mass spectrometry, Rapid Commun. Mass Spectrom. 22 (13) (2008) 2115–2124.
[99] D. Anselmetti, Single Cell Analysis: Technologies and Applications, 1 ed., Wiley-Blackwell, 2009.
[100] J. Riordon, D. Sovilj, S. Sanner, D. Sinton, E.W. Young, Deep learning with microfluidics for biotechnology, Trends Biotechnol. 37 (3) (2019) 310–324.
[101] P. Bachtiger, C.M. Plymen, P.A. Pabari, J.P. Howard, Z.I. Whinnett, F. Opoku, S. Janernig, A.A. Faisal, D.P. Francis, N.S. Peters, Artificial intelligence, data sensors and interconnectivity: future opportunities for heart failure, Card. Fail. Rev. (2020).
[102] C.C. Horgan, M. Jensen, A. Nagelkerke, J.-P. St-Pierre, T. Vercauteren, M. M. Stevens, M.S. Bergholt, High-throughput molecular imaging via deep learning enabled Raman spectroscopy, arXiv 9 (2020), 2009.13318 [physics].
[103] C. Banbury, R. Mason, I. Styles, N. Eisenstein, M. Clancy, A. Belli, A. Logan, P. G. oppenheimer, Development of the self optimising Kohonen Index Network (sKiNet) for Raman spectroscopy based Detection of Anatomical eye tissue, Sci. Rep. 9 (2019).
[104] C. Banbury, I. Styles, N. Eisenstein, E.R. Zanier, G. Vegliante, A. Belli, A. Logan, P. G. Oppenheimer, Spectroscopic detection of traumatic brain injury severity and biochemistry from the retina, Biomed. Opt. Express 11 (11) (2020) 6249–6261.
[105] N. Ma, S. Liu, W. Liu, L. Xie, D. Wei, L. Wang, L. Li, B. Zhao, Y. Wang, Research progress of titanium-based high entropy alloy: methods, properties, and applications, Front. Bioeng. Biotechnol. 8 (11) (2020) 603522.
[106] W. Huang, P. Martin, H.L. Zhuang, Machine-learning phase prediction of high-entropy alloys, Acta Mater. 169 (5) (2019) 225–236.
[107] A. Barati Farimani, M. Heiranian, N.R. Aluru, Identification of amino acids with sensitive nanoporous MoS2: towards machine learning-based prediction, Npj 2d Mater. Appl. 2 (12) (2018) 14.
[108] S. Caddeo, M. Boffito, S. Sartori, Tissue engineering approaches in the design of healthy and pathological in vitro tissue models, Front. Bioeng. Biotechnol. 5 (7) (2017) 40.
[109] F. Tourlomousis, C. Jia, T. Karydis, A. Mershin, H. Wang, D.M. Kalyon, R. C. Chang, Machine learning metrology of cell confinement in melt electrowritten three-dimensional biomaterial substrates, Microsyst. Nanoeng. 5 (12) (2019) 15.
[110] M. Sarmadi, A.M. Behrens, K.J. McHugh, H.T.M. Contreras, Z.L. Tochka, X. Lu, R. Langer, A. Jaklenec, Modeling, design, and machine learning-based framework for optimal injectability of microparticle-based drug formulations,", Science Advances 6 (7) (2020) eabb6594.

[111] S. You, J. Guan, J. Alido, H.H. Hwang, R. Yu, L. Kwe, H. Su, S. Chen, Mitigating scattering effects in light-based three-dimensional printing using machine learning, J. Manuf. Sci. Eng. 142 (8) (2020) 081002.
[112] D.A. Winkler, F.R. Burden, B. Yan, R. Weissleder, C. Tassa, S. Shaw, V.C. Epa, Modelling and predicting the biological effects of nanomaterials, SAR QSAR Environ. Res. 25 (2) (2014) 161–172.
[113] T. Österberg, U. Norinder, Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics, J. Chem. Inf. Comput. Sci. 40 (11) (2000) 1408–1411.
[114] P.Z. Moghadam, S.M.J. Rogge, A. Li, C.-M. Chow, J. Wieme, N. Moharrami, M. Aragones-Anglada, G. Conduit, D.A. Gomez-Gualdron, V. Van Speybroeck, D. Fairen-Jimenez, Structure-mechanical stability relations of metal-organic frameworks via machine learning, Matter 1 (7) (2019) 219–234.
[115] H.-S. Wang, Y.-H. Wang, Y. Ding, Development of biological metal–organic frameworks designed for biomedical applications: from bio-sensing/bio-imaging to disease treatment, Nanoscale Advances 2 (2020) 3788–3797.
[116] B. Ghobadian, H. Rahimi, A. Nikbakht, G. Najafi, T. Yusaf, Diesel engine performance and exhaust emission analysis using waste cooking biodiesel fuel with an artificial neural network, Renew. Energy 34 (4) (2009) 976–982.
[117] H. Oğuz, I. Sarıtas, H.E. Baydan, Prediction of diesel engine performance using biofuels with artificial neural network, Expert Syst. Appl. 37 (9) (2010) 6579–6586.
[118] P.K. Wong, K.I. Wong, C.M. Vong, C.S. Cheung, Modeling and optimization of biodiesel engine performance using kernel-based extreme learning machine and cuckoo search, Renew. Energy 74 (2015) 640–647.
[119] K.I. Wong, C.M. Vong, P.K. Wong, J. Luo, Sparse Bayesian extreme learning machine and its application to biofuel engine performance prediction, Neurocomputing 149 (A) (2015) 397–404.
[120] K. Yetilmezsoy, F.I. Turkdogan, I. Temizel, A. Gunay, Development of ann-based models to predict biogas and methane productions in anaerobic treatment of molasses, Int. J. Green Energy 10 (9) (2012) 885–907.
[121] S.K. Behera, S.K. Meher, H.-S. Park, Artificial neural network model for predicting methane percentage in biogas recovered from a landfill upon injection of liquid organic waste, Clean Technol. Environ. Policy 17 (2014) 443–453.
[122] J. George, P. Arun, C. Muraleedharan, Assessment of producer gas composition in air gasification of biomass using artificial neural network model, Int. J. Hydrogen Energy 43 (20) (2018) 9558–9568.
[123] A. Coşgun, M.E. Günay, R. Yıldırıma, Exploring the critical factors of algal biomass and lipid production for renewable fuel production by machine learning, Renew. Energy 163 (2021) 1299–1317.
[124] A. Kusiak, X. Wei, Prediction of methane production in wastewater treatment facility: a data-mining approach, Ann. Oper. Res. 216 (2014) 71–81.
[125] H. Bao, J. Wang, J. Li, H. Zhang, F. Wu, Effects of corn straw on dissipation of polycyclic aromatic hydrocarbons and potential application of backpropagation artificial neural network prediction model for PAHs bioremediation, Ecotoxicol. Environ. Saf. 186 (2019).
[126] G. Wu, C. Kechavarzi, X. Li, S. Wu, S.J. Pollard, H. Sui, F. Coulon, Machine learning models for predicting PAHs bioavailability in compost amended soils, Chem. Eng. J. 223 (2013) 747–754.
[127] B. Yadav, S. Ch, S. Mathur, J. Adamowski, Estimation of in-situ bioremediation system cost using a hybrid Extreme Learning Machine (ELM)-particle swarm optimization approach, J. Hydrol. 543 (B) (2016) 373–385.
[128] A. Perendeci, S. Arslan, S.S. Çelebi, A. Tanyolaç, Prediction of effluent quality of an anaerobic treatment plant under unsteady state through ANFIS modeling with on-line input variables, Chem. Eng. J. 145 (1) (2008) 78–85.
[129] D. Manu, A.K. Thalla, Artificial intelligence models for predicting the performance of biological wastewater treatment plant in the removal of Kjeldahl Nitrogen from wastewater, Appl. Water Sci. 7 (2017) 3783–3791.
[130] K. Yetilmezsoy, Z. Sapci-Zengin, Stochastic modeling applications for the prediction of COD removal efficiency of UASB reactors treating diluted real cotton textile wastewater, Stoch. Environ. Res. Risk Assess. 23 (2009) 13–26.
[131] L.L. Rogers, F.U. Dowla, Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling, Water Resour. Res. 30 (2) (1994) 457–481.
[132] S. Ch, D. Kumar, R.K. Prasad, S. Mathur, Optimal design of an in-situ bioremediation system using support vector machine and particle swarm optimization, J. Contam. Hydrol. 151 (2013) 105–116.
[133] M. Sisouane, M. Cascant, S. Tahiri, S. Garrigues, M. el Krati, G. Boutchich, M. Cervera, M. Guardia, Prediction of organic carbon and total nitrogen contents in organic wastes and their composts by Infrared spectroscopy and partial least square regression, Talanta 167 (2017) 352–358.
[134] J.M. Dale, L. Popescu, P.D. Karp, Machine learning methods for metabolic pathway prediction, BMC Bioinformatics 11 (2010).
[135] T. Oyetunde, M. Zhang, Y. Chen, Y. Tang, C. Lo, BoostGAPFILL: improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods, Bioinformatics 33 (4) (2017) 601–611.
[136] J. Mellor, I. Grigoras, P. Carbonell, J.-L. Faulon, Semisupervised gaussian process for automated enzyme search, ACS Synth. Biol. 5 (6) (2016) 518–528.
[137] S. Yong Teng, G. Yong Yew, K. Sukacova, P. Loke Show, V. Masa, J. Chang, Microalgae with artificial intelligence: a digitalized perspective on genetics, systems and products, Biotechnol. Adv. (2020).
[138] F. Caschera, M.A. Bedau, A. Buchanan, J. Cawse, Dd. Lucrezia, G. Gazzola, M. M. Hanczyc, N.H. Packard, Coping with complexity: Machine learning optimization of cell-free protein synthesis, Biotechnol. Bioeng. 108 (9) (2011).
[139] J. Alonso-Gutierrez, E.-M. Kim, T. S.Batth, N. Cho, Q. Hu, L.J. G.Chan, C. J. Petzold, N.J. Hillson, P.D. Adams, J.D. Keasling, H.G. Martin, T.S. Lee, Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering, Metab. Eng. 28 (2015) 123–133.
[140] S. Nandi, A. Subramanian, R.R. Sarkar, An integrative machine learning strategy for improved prediction of essential genes in Escherichia coli metabolism using flux-coupled features, Mol. Biosyst. (8) (2017).
[141] X. Zhang, V.M.L. Acencio, N. Lemke, Predicting essential genes and proteins based on machine learning and network topological features: a comprehensive review, Front. Physiol. (2016).
[142] S.G. Wu, Y. Wang, W. Jiang, T. Oyetunde, R. Yao, X. Zhang, K. Shimizu, Y.J. Tang, F.S. Bao, Rapid prediction of bacterial heterotrophic fluxomics using machine learning and constraint programming, PLoS Comput. Biol. 12 (4) (2016).
[143] P.-C. Fu, J. Barford, A hybrid neural network—first principles approach for modelling of cell metabolism, Comput. Chem. Eng. 20 (6–7) (1996) 951–958.
[144] M. Stosch, R. Oliveira, J. Peres, S.Fd. Azevedo, Hybrid semi-parametric modeling in process systems engineering: past, present and future, Comput. Chem. Eng. 60 (2014) 86–101.
[145] C.E. Lawson, J.M. Martí, T. Radivojevic, S.V.R. Jonnalagadda, R. Gentz, N. J. Hillson, S. Peisert, J. Kim, B. A.Simmons, C.J. Petzold, S.W. Singer, A. Mukhopadhyay, D. Tanjore, G.J. Dunn, H.G. Martin, Machine learning for metabolic engineering: a review, Metab. Eng. 63 (2021) 34–60.
[146] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A.W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S.C. P. Kohli, D.T. Jones, Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13), Proteins (2019) 1141–1148.
[147] K. Noble, Artificial Intelligence Solution to a 50-year-old Science Challenge Could 'revolutionise' Medical Research, 30 11 [Online]. Available:, 2020 https://predictioncenter.org/casp14/doc/CASP14_press_release.html.
[148] Y. Cai, J. Wang, L. Deng, SDN2GO: an integrated deep learning model for protein function prediction, Front. Bioeng. Biotechnol. (2020).
[149] S. Teng, A.K. Srivastava, L. Wang, Sequence feature-based prediction of protein stability changes upon amino acid substitutions, BMC Genomics (2010).
[150] I. Kundu, G. Paul, R. Banerjee, A machine learning approach towards the prediction of protein–ligand binding affinity based on fundamental molecular properties, RSC Adv. (22) (2018).
[151] J.-D. Qiu, J.-H. Huang, S.-P. Shi, R.-P. Liang, Using the concept of Chou's pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform, Protein Pept. Lett. 17 (6) (2010) 715–722.
[152] D. Heckmann, C.J. Lloyd, N. Mih, Y. Ha, D.C. Zielinski, Z.B. Haiman, A. A. Desouki, M.J. Lercher, B.O. Palsson, Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models, Nat. Commun. 9 (2018).
[153] P. Kotidis, C. Kontoravdi, Harnessing the potential of artificial neural networks for predicting protein glycosylation, Metab. Eng. Commun. (2020) e00131.
[154] T. Wuest, D. Weimer, C. Irgens, K.-D. Thoben, Machine learning in manufacturing: advantages, challenges, and applications, Prod. Manuf. Res. 4 (2016).
[155] A. Gábor, J.R. Banga, Robust and efficient parameter estimation in dynamic models of biological systems, BMC Syst. Biol. 9 (12) (2015) 74.
[156] J. Pinto, C. Rodrigues de Azevedo, R. Oliveira, M. von Stosch, A bootstrap-aggregated hybrid semi-parametric modeling framework for bioprocess development, Bioprocess Biosyst. Eng. 42 (11) (2019) 1853–1865.
[157] Y. Gal, G. Zoubin, Dropout as a bayesian approximation: representing model uncertainty in deep learning. International Conference on Machine Learning (ICML), 2016.
[158] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, arXiv 6 (2014). :1406.2661 [cs, stat].
[159] J. Yoon, D. Jarrett, M. van der Schaar, Time-series Generative Adversarial Networks, 2019.
[160] D. Zhang, E.A. Del Rio-Chanona, P. Petsagkourakis, J. Wagner, Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization, Biotechnol. Bioeng. 116 (11) (2019) 2919–2930.
[161] A. Devanand, G. Karmakar, N. Krdzavac, R. Rigo-Mariani, Y.S. Foo Eddy, I. A. Karimi, M. Kraft, OntoPowSys: a power system ontology for cross domain interactions in an Eco Industrial park, Energy and AI 1 (8) (2020) 100008.
[162] F. Farazi, M. Salamanca, S. Mosbach, J. Akroyd, A. Eibeck, L.K. Aditya, A. Chadzynski, K. Pan, X. Zhou, S. Zhang, M.Q. Lim, M. Kraft, Knowledge graph approach to combustion chemistry and interoperability, ACS Omega 5 (7) (2020) 18342–18348.
[163] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, D. Sontag, Learning a health knowledge graph from electronic medical records, Sci. Rep. 7 (2017) 1–11.
[164] D. Feldman, M. Volkov, D. Rus, Dimensionality reduction of massive sparse datasets using coresets, Adv. Neural Inf. Process. Syst. 29 (2016) 2766–2774.
[165] R. Oliveira, Combining first principles modelling and artificial neural networks: a general framework, ESCAPE 13 28 (5) (2004) 755–766.
[166] M. Von Stosch, R. Oliveira, J. Peres, S.F. de Azevedo, Hybrid semi-parametric modeling in process systems engineering: past, present and future, Comput. Chem. Eng. 60 (2014) 86–101.
[167] D. Zhang, T. Savage, B. Anye Cho, Combining model structure identification and hybrid modelling for photo-production process predictive simulation and optimisation, Biotechnol. Bioeng. 117 (11) (2020) 3356–3367.
[168] V. Brunner, M. Siegl, D. Geier, T. Becker, Biomass soft sensor for a Pichia pastoris fed-batch process based on phase detection and hybrid modeling, Biotechnol. Bioeng. 117 (9) (2020) 2749–2759.

[169] P. Xu, Analytical solution for a hybrid Logistic-Monod cell growth model in batch and continuous stirred tank reactor culture, Biotechnol. Bioeng. 117 (3) (2019).
[170] M. Esfandyari, M.A. Fanaei, R. Gheshlaghi, M. Akhavan Mahdavi, Mathematical modeling of two-chamber batch microbial fuel cell with pure culture of Shewanella, Chem. Eng. Res. Des. 117 (1) (2017) 34–42.
[171] M.A. Oberhardt, B.Ø. Palsson, J.A. Papin, Applications of genome-scale metabolic reconstructions, Mol. Syst. Biol. 5 (1) (2009) 320.
[172] R. Nian, J. Liu, B. Huang, A review On reinforcement learning: Introduction and applications in industrial process control, Comput. Chem. Eng. 139 (8) (2020) 106886.
[173] W. Li, S. Gu, X. Zhang, T. Chen, Transfer learning for process fault diagnosis: knowledge transfer from simulation to physical processes, Comput. Chem. Eng. 139 (8) (2020) 106904.
[174] B.J. Kotopka, C.D. Smolke, Model-driven generation of artificial yeast promoters, Nat. Commun. 11 (12) (2020) 2113.
[175] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, M. Dehmer, An introductory review of deep learning for prediction models with big data, Frontiers in Artificial Intelligence 3 (2) (2020).
[176] K. Cutajar, M. Pullin, A. Damianou, N. Lawrence, J. González, Deep Gaussian Processes for Multi-fidelity Modeling, 3, 2019.
[177] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014).
[178] L. Buşoniu, T. de Bruin, D. Tolić, J. Kober, I. Palunko, Reinforcement learning for control: performance, stability, and deep approximators, Annu. Rev. Control 46 (1) (2018) 8–28.
[179] J. Lee, J. Shin, M. Realff, Machine learning: overview of the recent progresses and implications for the process systems engineering field, Comput. Chem. Eng. 114 (2018) 111–121.
[180] J. Shin, T. Badgwell, K.-H. Liu, J. Lee, Reinforcement Learning – overview of recent progress and implications for process control, Comput. Chem. Eng. 127 (2019) 282–294.
[181] E. Pan, P. Petsagkourakis, M. Mowbray, D. Zhang, A. del Rio-Chanona, Constrained Model-Free Reinforcement Learning for Process Optimization, 2020.
[182] M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de novo drug design, Sci. Adv. 4 (2018) p. eaap7885.
[183] M. Mowbray, R. Smith, E.A. Del Rio-Chanona, D. Zhang, Using process data to generate an optimal control policy via apprenticeship and reinforcement learning, AIChE J. (2021), https://doi.org/10.1002/aic.17306.
[184] J. Schwarz, J. Luketina, W. Czarnecki, A. Grabska-Barwinska, Y. Whye Teh, R. Pascanu, R. Hadsell, Progress & compress: a scalable framework for continual learning, Proceedings of the 35th International Conference on Machine Learning (2018).
[185] M. Bain, C. Sammut, A Framework for Behavioural Cloning, 1995.
[186] P. Abbeel, A. Ng, Apprenticeship learning via inverse reinforcement learning, Proceedings of the 21 st International Conference on Machine Learning, Banff (2004).
[187] B. Ziebart, A. Maas, A. Bagnell, A. Dey, Maximum entropy inverse reinforcement learning, Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3, Chicago (2008).
[188] N. Ab Azar, A. Shahmansoorian, M. Davoudi, From inverse optimal control to inverse reinforcement learning: a historical review, Annu. Rev. Control 50 (2020) 119–138.
[189] P. Petsagkourakis, I.O. Sandoval, E. Bradford, D. Zhang, E. del Rio Chanona, Reinforcement Learning for Batch Bioprocess Optimization, 2019.