Review

Bridging the Gap between Mechanistic Biological Models and Machine Learning Surrogates

Ioana Gherman[1], Zahraa S. Abdallah[1,*], Wei Pang[2,*], Thomas Gorochowski[3,*], Claire Grierson[3,*], Lucia Marucci[1,*]

[*] *Senior authors*

1. *Department of Engineering Mathematics, University of Bristol*

2. *School of Mathematical and Computer Sciences Heriot – Watt University*

3. *School of Biological Sciences University of Bristol*

Abstract

Mechanistic models have been used for centuries to describe complex interconnected processes, including biological ones. As the scope of these models has widened, so have their computational demands. This complexity can limit their suitability when running many simulations or when real-time results are required. Surrogate machine learning models can be used to approximate the behaviour of complex mech-anistic models, and once built, their computational demands are several orders of magnitude lower. This paper provides an overview of the relevant literature, both from an applicability and a theoretical per-spective. For the latter, the paper focuses on the design and training of the underlying machine learning models. Application-wise, we show how machine learning surrogates have been used to approximate dif-ferent mechanistic models. We present a perspective on how these approaches can be applied to models representing biological processes with potential industrial applications (e.g., metabolism and whole-cell modelling) and show why surrogate machine learning models may hold the key to making the simulation of complex biological models possible using a typical desktop computer.

Keywords: systems biology; machine learning; surrogate model

1 Introduction

Mathematical mechanistic models have been used for centuries to understand and represent the natural laws that shape the world around us. Initially, the focus was on modelling specific phenomena and the mechanics underpinning them. Later, the direction shifted to combining and improving models such that they could better represent more complex and inter-connected processes (Fuller et al., 2020; Caputo et al., 2019). As these mathematical models mimic reality more closely, they also become increasingly complex. This brings a number of challenges, among which their computational demand is one of the most pressing ones. Simulations of complex mechanistic models can take hours or days to run, making them unfeasible for real-time decision-making or sensitivity analysis. This can make users reluctant to utilise them, despite their predictive power.

Here, we show how the high computational demand of some mechanistic models can be alleviated by using machine learning (ML) surrogates as a proxy. ML surrogates, also known as emulators or metamodels, are simple models that approximate the behaviour of a mechanistic one. The process of training and using an ML surrogate is shown in Figure 1. To create a surrogate, it is necessary to decide what output needs to be predicted and which inputs of the mechanistic model will be varied. Based on this, several simulations of the mechanistic model can then be run to create input-output pairs for training the ML model. In terms of the computational demand, the speed of the training phase relies on the ML model used and the number of iterations needed to obtain satisfactory accuracy. Once trained, a surrogate can be used in all future simulations, since getting ML predictions is typically faster compared to simulating the original mechanistic model.

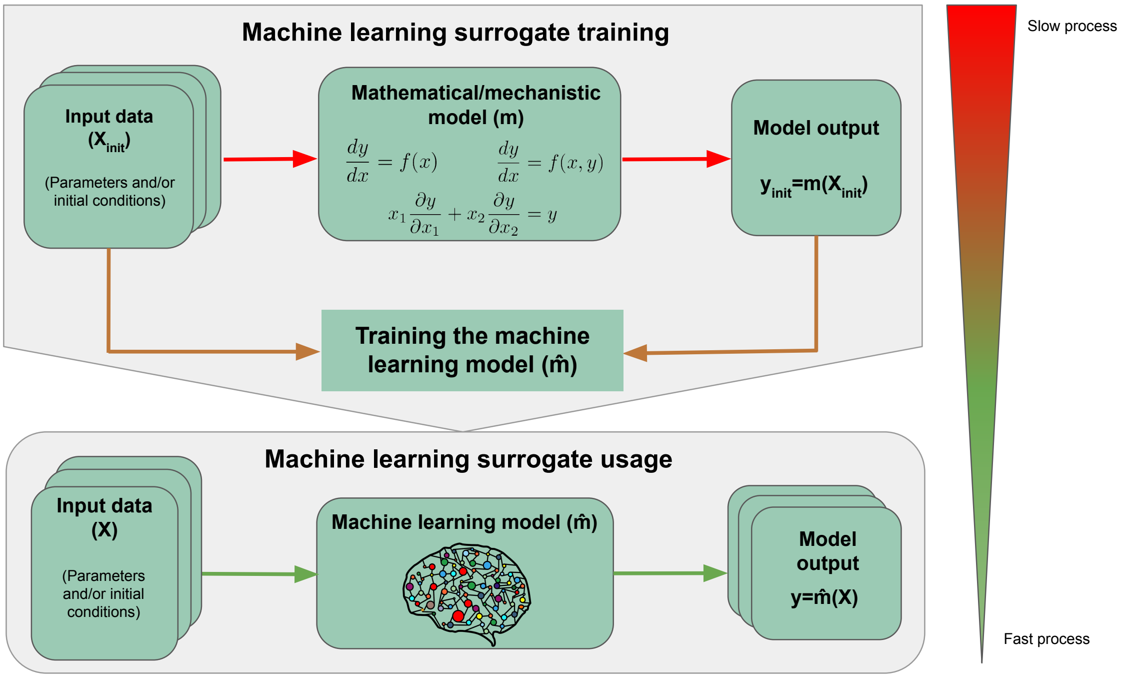
Machine learning surrogate training
Input data
(X_{init})
(Parameters and/or initial conditions)
Mathematical/mechanistic model (m)
$\frac{dy}{dx} = f(x)$
$\frac{dy}{dx} = f(x,y)$
$x_1\frac{\partial y}{\partial x_1} + x_2\frac{\partial y}{\partial x_2} = y$
Model output
$y_{init}=m(X_{init})$
Training the machine learning model (m̂)
Slow process
Machine learning surrogate usage
Input data
(X)
(Parameters and/or initial conditions)
Machine learning model (m̂)
Model output
y=m̂(X)
Fast process

Figure 1: Schematic representation for training and using an ML-based surrogate model. The mechanistic model is simulated (the top process connected by red arrows) to obtain the input-output pairs that are used to train the ML surrogate. This training stage (the middle process connected by orange arrows) is an average process in terms of speed. Its complexity will depend on the ML algorithm used, the complexity of the data pre-processing steps and the training iterations needed to obtain a satisfactory accuracy. Once this is achieved, the ML model can be used for all future predictions, effectively approximating the mechanistic model while running several orders of magnitude faster. This final (fast) process is represented by the green arrows at the bottom of the figure.

The improvement in computational speed that ML surrogate models achieve is particularly useful when predictions are needed in real-time (Liang et al., 2018a; Dabiri et al., 2019; Stolfi and Castiglione, 2021) or when large numbers of simulations have to be run, for instance, to explore a model's parameter space (Davies et al., 2019; Noè et al., 2019). It is important to acknowledge that simplified models can also lead to further improvements of the original ones. For example, different types of surrogate models have been used in the literature to analyse the uncertainty in the structure and the predictions of mechanistic models (Doherty and Christensen, 2011; Matott and Rabideau, 2008). Also, while building the surrogate, it is possible to gain further understanding of the model's relationship between inputs, parameters and outputs, discovering for example insensitive parameters (Young and Ratto, 2011).

Biological processes such as gene expression (Ay and Arnosti, 2011; Rué and Garcia-Ojalvo, 2013), metabolism (Gombert and Nielsen, 2000; Gu et al., 2019), signalling (Bray et al., 1998), cell growth (Shu and Shuler, 1989), and the cell cycle (Goldbeter, 1991; Tyson, 1991) have been modelled mechanistically for decades. It is possible to split biological mechanistic models into two paradigms. The first one classifies them by scale (Motta and Pappalardo, 2013); considering the cell as the 'unit' for measurement, it is possible to create models at the sub-cellular, cellular or macroscopic level. Sub-cellular models describe the evolution of individual physical and biochemical states of a cell (Helms, 2018). Cellular-level models describe the interactions among different molecules and processes within cells, and macroscopic-level models describe processes involved in groups of cells (Motta and Pappalardo, 2013). The second paradigm classifies biological models based on the mathematical formalism that they use (Soheilypour and Mofrad, 2018). Biological processes are commonly modelled using Ordinary Differential Equations (ODEs) (Wong et al., 2011; Lee et al., 2010; Yi et al., 2003), Partial Differential Equations (PDEs) (Cao et al., 2016; Yi et al., 2007; Cootes et al., 1995), Agent-Based Modelling (An et al., 2017; Soheilypour and Mofrad, 2018), cellular automata (Ermentrout and Edelstein-Keshet, 1993; Xu et al., 2007), stoichiometric matrices (Smolders et al., 1994; Taymaz-Nikerel et al., 2010), stochastic techniques (for example stochastic differential equations, SDEs) (Lee et al., 2010), or rule-based methods (Hwang et al., 2009). Details of each modelling approach are addressed by Motta and Pappalardo (2013) and Soheilypour and Mofrad (2018).

The main aim of this review is to bridge the gap between computationally demanding mechanistic models that describe biological systems at different cellular levels and the potential use of ML surrogates. First, in section 2, we will review the performance of different ML-based surrogate models, while analysing their advantages and disadvantages when applied to ODE, SDE and PDE-based mechanistic models (Lu and Ricciuto, 2019; Lu et al., 2018; Gong et al., 2015). Then, in section 3, we will discuss the benefits of using surrogate ML models in general, their limitations and the future avenues for improving these models and making them more usable by scientists from different fields and communities. Finally, in section 4, we will present how ML surrogate models can be relevant to approximate mechanistic models in the context of biotechnological industrial applications.

2 Machine learning as a surrogate in systems biology

ML-based surrogates were used to approximate mechanistic models of biological systems based on ODEs, SDEs (Wang et al., 2019; Renardy et al., 2018) and PDEs (Davies et al., 2019; Noè et al., 2019; Longobardi et al., 2020; Liu et al., 2019; Noè et al., 2016). These applications will be summarised below, focusing on the methodology and results of each study. Table 1 presents a summary of the surrogate ML models used for biological applications and their performance relative to the original mechanistic model they approximate. An overview of relevant methodological studies that apply ML-based surrogate modelling to mechanistic models from other engineering disciplines is presented in Table 2. The details of how these results were obtained are explained in the following paragraphs of this section.

Original model description	**Surrogate algorithm**	**Surrogate accuracy**	**Improvement in computational time**	**Reference**
SDE model of the MYC/E2F pathway (Wong et al., 2011; Lee et al., 2010).	LSTM	R^2 0.925-0.998	-	(Wang et al., 2019)
Heterotrimetric G-protein of budding yeast (Yi et al., 2003).	Orthogonal polynomial basis from the generalized polynomial chaos (gPC).	MAE 2.5 x 10^{-2}	20% reduction in CPU time.	(Renardy et al., 2018)
Pattern formation in *E. coli* (Cao et al., 2016).	LSTM.	R^2 0.987-0.99	30,000 fold acceleration.	(Wang et al., 2019)
Pheromone-induced cell polarization in Budding yeast (Yi et al., 2007).	Orthogonal polynomial basis from the generalized polynomial chaos.	MAE 0.14	180-fold reduction.	(Renardy et al., 2018)
Statistical model for aorta shapes (Cootes et al., 1995; Heimann and Meinzer, 2009; Liang et al., 2017).	PCA, Bidirectional neural network, Feedforward neural network.	Avg. MAE 0.533 mm	4 orders of magnitude.	(Liang et al., 2018b)
Risk for ascending aortic aneurysm (Liang et al., 2017).	PCA, Bidirectional neural network, Feedforward neural network.	Avg. MAE: 1.366 KPa	4 orders of magnitude.	(Liang et al., 2018a)
Stress analysis of arterial walls under atherosclerosis (Madani et al., 2019).	Feedforward neural network.	Test error 9.86%	-	(Madani et al., 2019)
Normal left ventricle (Dabiri et al., 2018; Baillargeon et al., 2015; Sack et al., 2018).	XGBoost or Cubist.	MAE for volume 1.495, MAE for pressure 1.544	2-3 orders of magnitude.	(Dabiri et al., 2019)
Human left ventricle model (Davies et al., 2019; Wang et al., 2013; Gao et al., 2015).	Gaussian process.	MSE 0.0001	3 orders of magnitude.	(Davies et al., 2019)
Human left ventricle (Wang et al., 2013; Gao et al., 2014).	K-Nearest Neighbour, XGBoost, Multilayer Perceptron.	R^2 0.999 (for the XGBoost and Multilayer Perceptron).	3-4 orders of magnitude.	(Cai et al., 2021)
Physiology models: Small and HumMod (Hester et al., 2011).	SVM regression.	Average error for Small: 0.05±2.47 and for HumMod: -0.3 +/- 3.94.	6 orders of magnitude.	(Pruett and Hester, 2016)

Table 1: Summary of the performance and methodologies of the ML surrogates of the systems biology models described in Section 2. The accuracy of the models is reported in terms of Mean Average Error (MAE), Mean Squared Error (MSE), coefficient of determination (R^2) and average error.

2.1 ML surrogates of ODE and SDE systems

Dynamical systems that evolve only in one dimension are usually modelled using ODEs or SDEs. ML surrogates were successfully applied to approximate systems biology models based on both types of techniques.

Original model description	Surrogate algorithm	Surrogate accuracy	Improvement in computational time	Reference
Simplified land model in the Energy Exascale Earth System Model (Lu and Ricciuto, 2019)	Singular value decomposition and neural network	MSE loss: 0.02	3-4 orders of magnitude	(Lu and Ricciuto, 2019)
Stochastic analysis of time-dependent PDEs solved using Monte Carlo method (Nikolopoulos et al., 2022)	Convolutional autoencoder and feed forward neural network	MSE loss: 0.03	81 times	(Nikolopoulos et al., 2022)

Table 2: Summary of the performance and methodologies of the ML surrogate models that describe engineering processes with methodologies that can be extended to surrogates of biological models. Accuracy is reported as Mean Squared Error (MSE).

For example, Renardy et al. (2018) built a surrogate based on an orthogonal polynomial basis from the generalised polynomial chaos (gPC) using the least square approximation for the heterotrimeric G-protein cycle of budding yeast. Once trained, the surrogate was used to compare its outputs to experimental data, with results showing high consistency between the two, a mean absolute error (MAE) of $2.5 * 10^{-2}$, as well as a 20% reduction in CPU time. The authors noted that this speed-up in computational time might not be high enough to balance the time invested in building the surrogate, suggesting that it is important to approximate *a priori* the expected improvements in computational time that a surrogate might bring.

Wang et al. (2019) used a surrogate model based on a Long Short-Term Memory (LSTM) deep neural network to replicate the behaviour of an SDEs model describing the MYC transduction pathways with E2F regulator (MYC/E2F) in cell-cycle progression. Apart from the high accuracy of the surrogate model and the improvement in computational time (Table 1), this analysis shows how surrogate ML models can be used to replicate stochastic systems. Although different runs of the mechanistic model using the same parameters will produce different concentration levels for each molecule, the distribution of these concentrations is deterministic for a sufficiently large number of runs. This suggests that each combination of parameters leads to a unique distribution for each molecule, corresponding to the spatial output of the SDE model, which can be predicted by the surrogate neural network.

2.2 ML surrogates of PDE systems

Complex dynamical systems that evolve in two or more dimensions are often modelled using PDEs. Traditionally, these models are solved numerically using Finite Element Analysis (FEA) methods (Segerlind, 1991). In this section, we will review the applicability of ML surrogates to mathematical models described by PDEs for molecular biology processes (Wang et al., 2019; Renardy et al., 2018) and biomedical systems (Noè et al., 2019; Davies et al., 2019; Cai et al., 2021; Di Achille et al., 2018; Longobardi et al., 2020; Noè et al., 2016).

2.2.1 Applications in molecular biology

In molecular biology, surrogate models based on LSTM neural networks (Hochreiter and Schmidhuber, 1997; Wang et al., 2019) were built to predict pattern formation in *E.coli* programmed by a synthetic gene circuit (Cao et al., 2016) represented as the spatial distribution of different molecules. The LSTM took as input the parameters of the mechanistic model and was trained to predict two outputs: the logarithm of the peak value of the profile of different molecules and their normalised profile. The authors reported a 30,000-fold computational acceleration (Wang et al., 2019), the LSTM being successfully used to identify new patterns by screening 10^8 parameter sets in 12 days (compared to thousands of years which is how long it would have taken for the PDE model to achieve this). To improve the robustness of the ML model, the authors also proposed a reliability metric based on a voting system across different neural networks trained in parallel. This was an important addition to surrogate modelling that can prove particularly useful since in the cases when the surrogate is uncertain about a prediction, the mechanistic model can be run.

Renardy et al. (2018) presented a technique based on polynomial surrogates using a Legendre polynomial basis that was applied to a spatial model of pheromone-induced cell polarization of budding yeast. Once the polynomial surrogate was fit, it was used to compute parameter sensitivities and perform rapid Bayesian parameter inference. Using the surrogate, it was possible to run simulations that would take approximately 200 years to run using the mechanistic model. Furthermore, the surrogate facilitated the convergence for the distribution of 15 parameters in only a few hours using Bayesian inference.

2.2.2 Applications in organ modelling and physiology

Biomedical engineering is a field where surrogate models have been built extensively over the past decade, with a particular focus on biophysical models of the heart. Modelling myocardial properties that can help in making real-time clinical decisions could contribute to understanding and treating heart diseases (Gao et al., 2017). Several mathematical models of the myocardium could be used for these aims. However, most are restricted by their high computational demand. Several studies suggest that these limitations can be addressed by implementing surrogate models based on different ML algorithms (Liang et al., 2018b,a; Madani et al., 2019; Dabiri et al., 2019; Longobardi et al., 2020; Noè et al., 2019; Cai et al., 2021; Di Achille et al., 2018; Noè et al., 2016).

For example, Liang et al. (2018b) built a ML surrogate of the FEA method to estimate the zero-pressure geometry of the human thoracic aorta. The input (i.e. a pair of shapes) and output were first encoded as a set of scalars using Principal Component Analysis (PCA). Then, the non-linear mapping between the encoded input and output was performed using a feed-forward fully connected neural network. Lastly, the output was decoded again using PCA. It was shown that ML surrogates could enable real-time applications of the model, with prediction time under one second and an average mean absolute error of 0.533 mm. A similar approach was used by Liang et al. (2018a), the main difference being that here the model takes one shape as input, whilst Liang et al. (2018b) used a pair of inputs for the ML pipeline.

Two deep learning approaches were tested to predict the point-wise distribution of stress on the arterial walls under atherosclerosis in (Madani et al., 2019). The inputs of the surrogate model were parameters describing the geometry and arterial pressure, and the outputs were point-wise stress distributions. The performance of a feed-forward neural network was compared against that of a convolutional neural network, with the first outperforming the second. Similarly to other studies (Dabiri et al., 2019; Longobardi et al., 2020), the authors performed a features' importance analysis by adjusting one input feature at a time and studying the impact of these changes on the accuracy of the deep learning model. This approach revealed expected correlations between arterial pressure and stress, but also less obvious ones such as the fact that lipid pool information had more impact on maximum stress compared to calcium deposits. This suggests that, besides their predictive power, ML surrogates can also unravel some dynamics of the system that have not been studied previously.

Cai et al. (2021) also used simulations of the LV diastolic filling with the aim of estimating model parameters. Features were first projected into a lower dimensional space, and three different ML models (K-nearest neighbour, XGBoost and a multilayer perceptron) were tested to assess how well they learn the pressure-volume and pressure-strain relationships. The computational cost of simulations is reduced by 3-4 orders of magnitude when using the ML surrogate. Davies et al. (2019) used two interpolation methods and two loss functions to estimate the material properties of a healthy volunteer's left ventricle using only non-invasive data. Minimizing the loss between the biomechanical model's output and the emulator produced an estimate of the unknown parameters that have to be fitted. Two loss functions were used: the Euclidean loss function which assumes that the outputs are independent, and the Mahalanobis distance-based loss function which allows for correlation across outputs. The best results were achieved by the emulation of the output of the biomechanical model using local Gaussian Process interpolation and the Euclidean loss function. The reported mean square error (MSE) was 0.0001, and the computational time was reduced by approximately three orders of magnitude, from weeks to a quarter of an hour.

Another proof-of-concept for the usability of surrogate modelling assessed the applicability of ML models to emulate two physiology mathematical models, Small and HumMod (Pruett and Hester, 2016; Hester et al., 2011). Support Vector Machine (SVM) regression models were used to map the parameter samples to the drop in mean arterial pressure. The accuracy of these surrogates was calculated with respect to the drop in mean arterial pressure observed after running the original mathematical model. Further error analysis showed that there was no significant difference between the performance of the ML model and the mechanistic one. The authors also compared the time complexity of the two approaches and showed that the ML model could make predictions approximately six orders of magnitude faster than both dynamical models.

Besides the improvements in computational demand, the studies presented in this section also address other important modelling aspects such as building surrogates of stochastic models, implementing reliability metrics (Wang et al., 2019), performing parameter sensitivity and inference (Renardy et al., 2018) and analysing feature importance (Madani et al., 2019). Furthermore, since biological systems are often high dimensional, dimensionality reduction is another important modelling aspect that has been combined with surrogate-based ML models in studies describing both biological(Liang et al., 2018b,a; Cai et al., 2021) and other engineering systems (Lu and Ricciuto, 2019; Nikolopoulos et al., 2022).

The studies above show that there is no consensus regarding the best ML algorithms to use. Based on the results from Tables 1 and 2 we see that neural networks and decision tree-based methods perform very well. However, they need a significant amount of data to be trained, meaning that more simulations of the mechanistic model need to be run. On the other hand, algorithms such as Gaussian Processes have the advantage of higher interpretability and the capacity to estimate the uncertainty in the predictions. Therefore, choosing the right algorithm remains at the latitude of the user based on the problem to be solved, data availability and the mechanistic model to be emulated. To help with such decisions, in Section 3 we will review some technical aspects that can help the creators of ML-based surrogates to design optimal models.

3 Building, training and using ML surrogates

Each of the studies presented in Section 2 has its own set of limitations. Some of these are domain-specific and rely heavily on the knowledge of domain experts. For example, some methods cannot be used in a clinical setting yet because they were only trained on myocardial models coming from a patient with specific characteristics, and hence they do not include inter-patient variability. Other limitations and design matters are more general, being common across surrogate models, regardless of their application area. These technical aspects addressing the design and training of machine learning surrogates are discussed below.

3.1 Active learning

Given that surrogates are built to avoid running expensive simulations many times, it is important to minimise the number of simulations needed to train the ML model while keeping them as informative as possible. In active learning, a model can choose the data that it will learn from next, based on the information it gained from previous training examples. A summary of the process combining surrogate modelling and active learning is explained in Figure 2. Active learning has been applied together with surrogate models in a few engineering studies, where the original mechanistic models were based on PDEs (Pestourie et al., 2020; Lye et al., 2021). Pestourie et al. (2020) built a neural network-based surrogate for the PDE model representing the Maxwell equations for composite materials, and used an active learning algorithm that selects new training points from the parameter space where the estimated model error was higher. This error was recalculated after the training set was updated with new data obtained by running the simulations using the mechanistic model. The active learning approach was compared to a baseline, where the training set was randomly sampled from the mechanistic model's parameter space. The active learning surrogate matched the numerical integration result more closely, using one order of magnitude less training data compared to the surrogate trained on randomly sampled points.

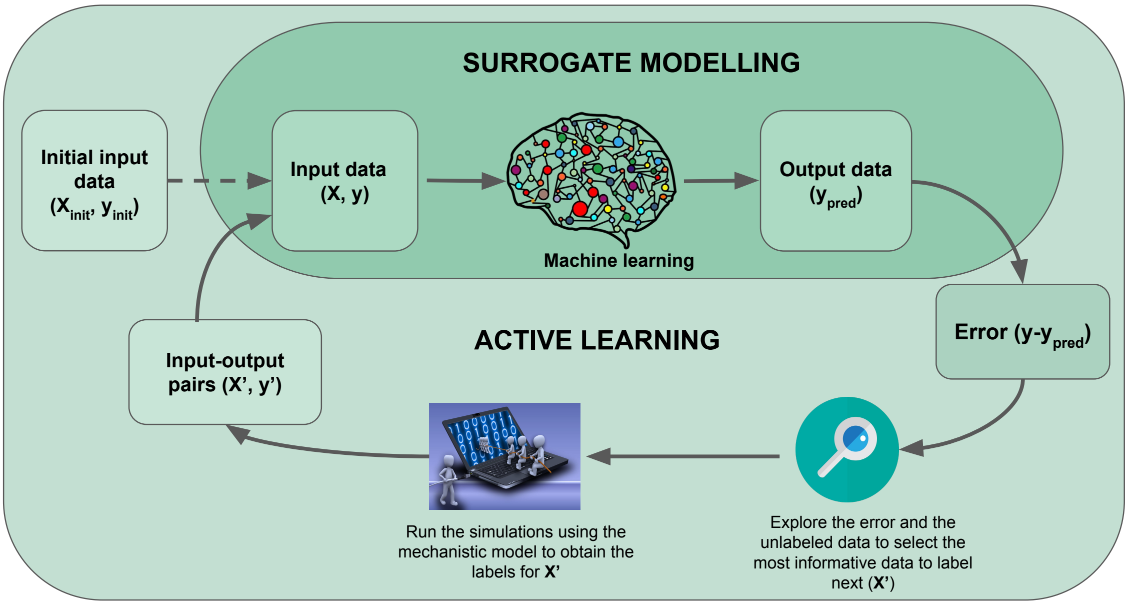
SURROGATE MODELLING
Initial input data (X_{init}, y_{init})
Input data (X, y)
Machine learning
Output data (y_{pred})
Error (y-y_{pred})
ACTIVE LEARNING
Input-output pairs (X', y')
Run the simulations using the mechanistic model to obtain the labels for X'
Explore the error and the unlabeled data to select the most informative data to label next (X')

Figure 2: Schematic representation of how active learning and ML surrogates can work together. Initially, a ML model is trained on a set of data generated by some initial simulations of a mechanistic model (X_{init}, y_{init}), which are equivalent to (X, y) for this initial step. The ML model is used to make predictions (y_{pred}). The estimated error between the prediction of the mechanistic model (y) and that of the ML model (y_{pred}) is used to select a subset from all the possible input data that hasn't been used to make predictions using the mechanistic model in the past (X'). The mechanistic model is run using X' as input to obtain a new set of input-output pairs (X, y), equivalent to the newly generated (X', y'), that when included in the ML process are expected to reduce the estimated error $(y - y_{pred})$.

Lye et al. (2021) used deep learning surrogates and active learning to solve the constrained optimisation problem of three systems: optimal control problem for a nonlinear ODE, parameter identification for the heat equation, and shape optimisation of airfoils subject to the Euler equations. The algorithm presented is called Iterative Surrogate Model Optimization (ISMO), where a deep neural network queries a standard optimisation algorithm, quasi-Newton approximation, to provide training examples that will minimise its error. The ISMO algorithm outperformed the purely deep neural network surrogate in terms of error decay and robustness to parameter change and the standard optimisation algorithm for aerodynamic shape optimisation by more than an order of magnitude (Lye et al., 2021). Balaprakash et al. (2013) presented strategies for building surrogate models with active learning using iterative parallel computations on single-core, multi-core and multi-node architectures. Understanding how this can be done can further speed up the modelling process.

3.2 Designing the ML model

ML-based surrogates could outperform the mechanistic model that was used to train them when tested against experimental observations. This effect can be further enhanced if simulated data is combined with experimental one to train the ML-based surrogate. To the best of our knowledge, this has not been done before. However, augmenting experimental data with a simulated one coming from a mechanistic model was done by Costello and Martin (2018), and in the training phase, this had a positive impact on the model's performance.

Another constraint comes from the fact that often the surrogate might not be able to make predictions for all the outputs of a high-dimensional mechanistic model. However, depending on the size of these outputs, it is possible to train multi-output models, which use the same input variables to predict several outputs. The feasibility of this approach depends on the dimensionality of the mechanistic model. Xu et al. (2019) reviewed different methods and challenges regarding multi-output ML approaches. The survey focused on assessing the algorithms based on volume, velocity, variety and veracity, all being important characteristics of models of biological systems. Another approach to address the high dimensionality of mechanistic models' outputs has been to apply dimensionality reduction techniques to them (Lu and Ricciuto, 2019; Nikolopoulos et al., 2022; Cai et al., 2021).

Since most of the mechanistic models presented here describe the dynamics of systems, it is expected that an increasing number of future models will be based on time-series data. With the rise of computer power and deep learning techniques, there has been a lot of progress in the field of time-series analysis and prediction. Several reviews have been published recently to outline the state of the art when it comes to time-series forecasting algorithms (Tealab, 2018; Torres et al., 2021; Deb et al., 2017) and time-series classification algorithms (Bagnall et al., 2017; Ruiz et al., 2021; Fawaz et al., 2019), some with the aim to make them interpretable (Assaf and Schumann, 2019; Selvaraju et al., 2017; Nguyen et al., 2020).

3.3 Model usability

To train the ML surrogates, the user needs to be able to run the mechanistic model and train an ML model. Depending on the usability of the original model, its complexity and the user's knowledge regarding ML, this could take more time than running the computationally expensive mechanistic model (Renardy et al., 2018). This suggests that it would be highly beneficial to include a reproducibility metric as part of surrogate modelling studies. For the initial training of the surrogate model to be as efficient as possible, the mechanistic models should be easy to use when it comes to running the simulations for creating the input-output pairs needed for the surrogate. We believe this should already be the case for most models, but we want to emphasize the need for clear instructions on how simulations should be run under different conditions. Furthermore, the training of the ML model should be accessible to non-experts. Once the training data is available, this is already possible using tools such as AutoML (Guyon et al., 2019) or TPOT (Olson et al., 2016a,b; Le et al., 2020).

The other important aspect that needs to be addressed when we discuss usability and reproducibility is how easy it is to deploy the already built surrogate, to re-train it or to slightly modify its scope. We believe that as surrogate modelling begins to be widely used and becomes part of experimental or clinical pipelines, it is important to think about its tunability. To address this, it is important that the code used to build the surrogates is publicly available, well structured and the ML pipeline is presented clearly.

It is also important to consider why surrogate ML models are able to predict the dynamics and state of different systems. This is particularly interesting since in the past 50 years the research community has focused on building mechanistic models rather than ML-based models. We believe that it is expected for the ML surrogates to work when approximating a deterministic non-linear mechanistic model, especially when given *sufficient training data*. The modern machine/deep learning algorithms running on modern computers are able to learn the non-linear dependencies between all different kinds of input-output pairs, including images, text (Goodfellow et al., 2016) and other unstructured data (Feldman et al., 2007). Therefore, there should not be reasons why they fail when predicting input-output pairs of mechanistic models.

Different design approaches may be considered when the mechanistic models to be emulated are not fully deterministic. Surrogate models have been used to approximate stochastic mechanistic models, and it has been shown that if sufficient simulations are run, the distribution of the output of these models is approximately deterministic (Wang et al., 2019). Another approach for building surrogates of stochastic models is to include the random seed which was used for the simulations as an input when training the ML model (Angione et al., 2022).

3.4 Interpretability

Surrogate ML models can also help explain some unstudied dynamics of dynamical system (Madani et al., 2019; Dabiri et al., 2019; Longobardi et al., 2020). With the recent progress in the area of explainable artificial intelligence, once predictions are made, it becomes possible to interpret for example whether all input data impacts the prediction (Arrieta et al., 2020). Furthermore, it is possible to understand which features influence the prediction the most and quantify this impact. For example, whether an increase in the value of one feature changes the prediction and in the case of regression models whether the prediction is generally increased or decreased. Such methods could outline some behaviours of the system that were not previously known, especially when experimental data are used to train as well. In general, to make sure the results are robust, it can be helpful to apply different explainability methods and compare their results.

Sections 2 and 3 have described the way surrogate machine learning models have been used in the literature and how the design and usability of such models can be enhanced. Using the information acquired from these sections, we propose future avenues for applying surrogate machine learning models to industrially relevant biotechnology modelling.

4 Further applications of ML-based surrogates in biotechnology

Given the potential of metabolic and whole cell models in designing novel renewable biofuels (Beller et al., 2015; Chubukov et al., 2016) and drugs (Ajikumar et al., 2010), as well as their versatility for minimal genome design (Wang and Maranas, 2018; Rees-Garbutt et al., 2020), we further present our vision regarding the applicability of surrogate modelling for these types of mechanistic models. Metabolism is among the most complex processes taking place in a cell. Genome-scale metabolic models include all the known information about the metabolism of an organism, such as genes, enzymes, reactions and metabolites (Passi et al., 2022). These models can be used not only to predict metabolic fluxes but also to understand genotype-phenotype interactions. In addition, they can have a significant impact on understanding strain development for the production of bio-based materials and chemicals, drug targeting, predictions of enzyme function and modelling interactions among different cells (Gu et al., 2019). Given the system-level complexity of these techniques, the models often end up containing thousands of genes, metabolites and reactions that interact with each other. This entails a great computational demand, with some models running for up to 7 hours, especially when a protein expression network is included (Yang et al., 2019). Metabolic kinetic models are even more computationally expensive since they predict the temporal behaviour of the process and they combine multi-omics data sets with reaction network models (Islam et al., 2021). This computational problem is amplified when complex organisms are modelled or several simulations have to be run.

Often, metabolic models are used as part of a Design Build Test Learn (DBTL) pipeline (Nielsen and Keasling, 2016), corresponding to multiple combinations of inputs (Figure 3). This frequently involves a significant number of trial and error experiments, suggesting that ML surrogates of metabolic models would be particularly useful for such cases. For example, the ML surrogates can be trained on the initial

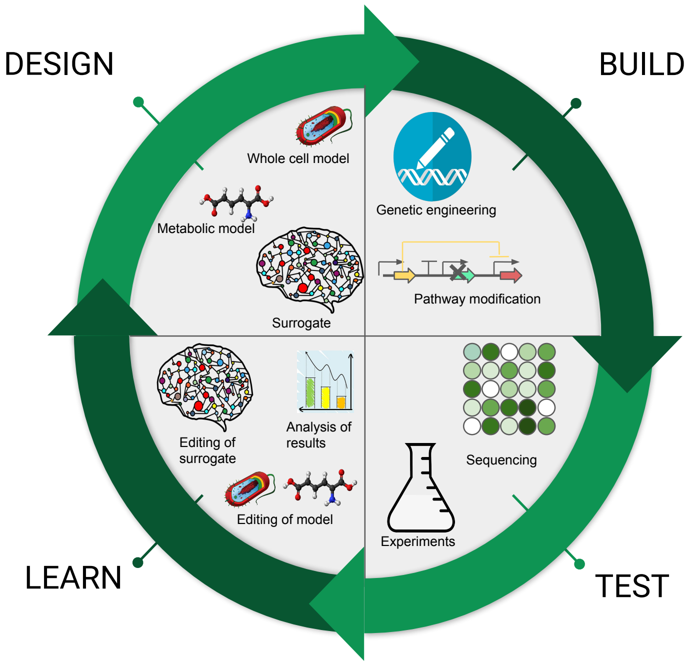

Figure 3: An example of the design-build-test-learn (DBTL) pipeline, where the metabolic or whole cell models can be replaced by surrogate models.

state of the input variables of the mechanistic model and/or the parameters of the model, with the target variable being the desired phenotype to be predicted (a specific titer, rate, yield, or product). Once the training phase is completed, the surrogate can be used to approximate the original metabolic model. One of the challenges that may occur when implementing this framework is caused by the high dimensionality of metabolic models. This suggests that if the initial conditions are used as input for the ML model, it will be necessary to run several simulations to cover most of the variables' space. According to Kuo and Sloan (2005) and Lawson et al. (2021), at least 4-5 times as many simulations as the number of variables are needed to avoid the curse of dimensionality. In the cases when this is possible, it can also lead to the discovery of interesting dynamics of the system since an explainable ML model can show which parameters and variables are influencing the phenotype the most. However, in other cases, running a high number of mechanistic simulations might defeat the purpose of building a surrogate model. For such situations it is possible to reduce the dimension of the input by applying different dimensionality reduction techniques (Espadoto et al., 2019), or by keeping only the variables that are known to influence the desired phenotype (Stolfi and Castiglione, 2021).

Whole-cell models are mathematical models that include and link all the well-annotated genes and processes of a cell. Two such models have been published to date, one for *Mycoplasma genitalium* (Karr et al., 2012), another one for *Escherichia coli* (Macklin et al., 2020), and more are underway (Karr, 2019). The completeness of these models makes them particularly powerful since, when used in a DBTL pipeline, they facilitate the study and design of interactions among different cellular processes (i.e., something that a metabolic model alone cannot achieve). Similarly to metabolic models, whole-cell models have already been used for *in-silico* minimal genome design (Rees-Garbutt et al., 2020), and we anticipate that they will change the paradigm for metabolic engineering and development of microbial chassis (Marucci et al., 2020). These models add some extra levels of complexity to the genome-scale metabolic models, and therefore are even more computationally expensive (Macklin et al., 2014), with a simulation time of 20 minutes - 24 hours per cell (on a desktop computer). This makes applications that involve multiple cells growing over multiple generations prohibitively expensive due to their high computational time.

ML surrogate models can represent a strategy to address this challenge. For example, the input to the ML model can be defined as a subset of the initial conditions and parameters of the model, and the output as the phenotype to be predicted. This can be a continuous variable such as the growth rate of a cell, the production, titer, or yield of different metabolites, or binary indicating for example whether a cell divides or not. In the case of metabolic models, for example, whole-cell models can have thousands of candidate variables that can be used as input to the ML model. As mentioned before, these can be reduced using dimensionality reduction techniques (Espadoto et al., 2019) or prior knowledge about the process under investigation (Stolfi and Castiglione, 2021).

5 Conclusion

There is a growing number of studies in the literature showing how ML surrogates can be used to emulate mechanistic models of biological processes, both at the molecular and macroscopic levels. These show that, besides the performance of the surrogate models in terms of accuracy against numerical integration and improvement in computational speed, it is also beneficial to consider other design aspects. First, it is important to assess whether the mechanistic model is complex enough to invest the time in building a surrogate (Renardy et al., 2018). The design of the protocol for obtaining the training data of the ML surrogate should consider aspects such as stochasticity and whether active learning could bring any additional value (Wang et al., 2019; Angione et al., 2022; Pestourie et al., 2020; Lye et al., 2021). Other aspects such as dimensionality reduction of the inputs and/or outputs of the ML surrogate (Liang et al., 2018a,b; Cai et al., 2021; Lu and Ricciuto, 2019; Nikolopoulos et al., 2022), and parameter sensitivity analysis (Renardy et al., 2018) can help to optimise the performance of the model, but also to unravel some information about the dynamics of the system.

6 Funding

L.M. was funded by the Engineering and Physical Sciences Research Council (EPSRC, EP/S01876X/1) and Biotechnology and Biological Sciences Research Council (BBSRC, Bristol Centre for Synthetic Biology, BB/L01386X/1; BrisEngBio, BB/W013959/1). T.E.G. was funded by the Royal Society University Research Fellowship (UF160357) and Turing Fellowship from The Alan Turing Institute (EPSRC grant EP/N510129/1). IG is supported by an EPSRC Doctoral Training Studentship.

References

P. K. Ajikumar, W.-H. Xiao, K. E. Tyo, Y. Wang, F. Simeon, E. Leonard, O. Mucha, T. H. Phon, B. Pfeifer, and G. Stephanopoulos. Isoprenoid pathway optimization for taxol precursor overproduction in escherichia coli. *Science*, 330(6000):70–74, 2010.

G. An, B. Fitzpatrick, S. Christley, P. Federico, A. Kanarek, R. M. Neilan, M. Oremland, R. Salinas, R. Laubenbacher, and S. Lenhart. Optimization and control of agent-based models in biology: a perspective. *Bulletin of mathematical biology*, 79(1):63–87, 2017.

C. Angione, E. Silverman, and E. Yaneske. Using machine learning as a surrogate model for agent-based simulations. *Plos one*, 17(2):e0263150, 2022.

A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

R. Assaf and A. Schumann. Explainable deep neural networks for multivariate time series predictions. In *IJCAI*, pages 6488–6490, 2019.

A. Ay and D. N. Arnosti. Mathematical modeling of gene expression: a guide for the perplexed biologist. *Critical reviews in biochemistry and molecular biology*, 46(2):137–151, 2011.

A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data mining and knowledge discovery*, 31(3):606–660, 2017.

B. Baillargeon, I. Costa, J. R. Leach, L. C. Lee, M. Genet, A. Toutain, J. F. Wenk, M. K. Rausch, N. Rebelo, G. Acevedo-Bolton, et al. Human cardiac function simulator for the optimal design of a novel annuloplasty ring with a sub-valvular element for correction of ischemic mitral regurgitation. *Cardiovascular engineering and technology*, 6(2):105–116, 2015.

P. Balaprakash, R. B. Gramacy, and S. M. Wild. Active-learning-based surrogate models for empirical performance tuning. In *2013 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 1–8. IEEE, 2013.

H. R. Beller, T. S. Lee, and L. Katz. Natural products as biofuels and bio-based chemicals: fatty acids and isoprenoids. *Natural product reports*, 32(10):1508–1526, 2015.

D. Bray, M. D. Levin, and C. J. Morton-Firth. Receptor clustering as a cellular mechanism to control sensitivity. *Nature*, 393(6680):85–88, 1998.

L. Cai, L. Ren, Y. Wang, W. Xie, G. Zhu, and H. Gao. Surrogate models based on machine learning methods for parameter estimation of left ventricular myocardium. *Royal Society open science*, 8(1):201121, 2021.

Y. Cao, M. D. Ryser, S. Payne, B. Li, C. V. Rao, and L. You. Collective space-sensing coordinates pattern scaling in engineered bacteria. *Cell*, 165(3):620–630, 2016.

F. Caputo, A. Greco, M. Fera, and R. Macchiaroli. Digital twins to enhance the integration of ergonomics in the workplace design. *International Journal of Industrial Ergonomics*, 71:20–31, 2019.

V. Chubukov, A. Mukhopadhyay, C. J. Petzold, J. D. Keasling, and H. G. Martín. Synthetic and systems biology for microbial production of commodity chemicals. *NPJ systems biology and applications*, 2(1): 1–11, 2016.

T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

Z. Costello and H. G. Martin. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ systems biology and applications*, 4(1):1–14, 2018.

Y. Dabiri, K. L. Sack, S. Shaul, P. P. Sengupta, and J. M. Guccione. Relationship of transmural variations in myofiber contractility to left ventricular ejection fraction: implications for modeling heart failure phenotype with preserved ejection fraction. *Frontiers in physiology*, 9:1003, 2018.

Y. Dabiri, A. Van der Velden, K. L. Sack, J. S. Choy, G. S. Kassab, and J. M. Guccione. Prediction of left ventricular mechanics using machine learning. *Frontiers in physics*, 7:117, 2019.

V. Davies, U. Noè, A. Lazarus, H. Gao, B. Macdonald, C. Berry, X. Luo, and D. Husmeier. Fast parameter inference in a biomechanical model of the left ventricle by using statistical emulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(5):1555–1576, 2019.

C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah. A review on time series forecasting techniques for building energy consumption. *Renewable and Sustainable Energy Reviews*, 74:902–924, 2017.

P. Di Achille, A. Harouni, S. Khamzin, O. Solovyova, J. J. Rice, and V. Gurev. Gaussian process regressions for inverse problems and parameter searches in models of ventricular mechanics. *Frontiers in physiology*, 9:1002, 2018.

J. Doherty and S. Christensen. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. *Water Resources Research*, 47(12), 2011.

G. B. Ermentrout and L. Edelstein-Keshet. Cellular automata approaches to biological modeling. *Journal of theoretical Biology*, 160(1):97–133, 1993.

M. Espadoto, R. M. Martins, A. Kerren, N. S. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE transactions on visualization and computer graphics*, 27(3):2153–2173, 2019.

H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. Deep learning for time series classification: a review. *Data mining and knowledge discovery*, 33(4):917–963, 2019.

R. Feldman, J. Sanger, et al. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

A. Fuller, Z. Fan, C. Day, and C. Barlow. Digital twin: Enabling technologies, challenges and open research. *IEEE access*, 8:108952–108971, 2020.

H. Gao, H. Wang, C. Berry, X. Luo, and B. E. Griffith. Quasi-static image-based immersed boundary-finite element model of left ventricle under diastolic loading. *International journal for numerical methods in biomedical engineering*, 30(11):1199–1222, 2014.

H. Gao, W. Li, L. Cai, C. Berry, and X. Luo. Parameter estimation in a holzapfel–ogden law for healthy myocardium. *Journal of engineering mathematics*, 95(1):231–248, 2015.

H. Gao, K. Mangion, D. Carrick, D. Husmeier, X. Luo, and C. Berry. Estimating prognosis in patients with acute myocardial infarction using personalized computational heart models. *Scientific reports*, 7(1): 1–14, 2017.

A. Goldbeter. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proceedings of the National Academy of Sciences*, 88(20):9107–9111, 1991.

A. K. Gombert and J. Nielsen. Mathematical modelling of metabolism. *Current opinion in biotechnology*, 11(2):180–186, 2000.

W. Gong, Q. Duan, J. Li, C. Wang, Z. Di, Y. Dai, A. Ye, and C. Miao. Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrology and Earth System Sciences*, 19(5): 2409–2425, 2015.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

C. Gu, G. B. Kim, W. J. Kim, H. U. Kim, and S. Y. Lee. Current status and applications of genome-scale metabolic models. *Genome biology*, 20(1):1–18, 2019.

I. Guyon, L. Sun-Hosoya, M. Boullé, H. J. Escalante, S. Escalera, Z. Liu, D. Jajetic, B. Ray, M. Saeed, M. Sebag, A. Statnikov, W. Tu, and E. Viegas. Analysis of the automl challenge series 2015-2018. In *AutoML*, Springer series on Challenges in Machine Learning, 2019. URL `https://www.automl.org/wp-content/uploads/2018/09/chapter10-challenge.pdf`.

T. Heimann and H.-P. Meinzer. Statistical shape models for 3d medical image segmentation: a review. *Medical image analysis*, 13(4):543–563, 2009.

V. Helms. *Principles of computational cell biology: from protein complexes to cellular networks*. John Wiley & Sons, 2018.

R. Hester, A. Brown, L. Husband, R. Iliescu, W. A. Pruett, R. L. Summers, and T. Coleman. Hummod: a modeling environment for the simulation of integrative human physiology. *Frontiers in physiology*, 2: 12, 2011.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

M. Hwang, M. Garbey, S. A. Berceli, and R. Tran-Son-Tay. Rule-based simulation of multi-cellular biological systems—a review of modeling techniques. *Cellular and molecular bioengineering*, 2(3):285–294, 2009.

M. M. Islam, W. L. Schroeder, and R. Saha. Kinetic modeling of metabolism: Present and future. *Current Opinion in Systems Biology*, 26:72–78, 2021.

J. Karr. Models: Comprehensive computational models of individual cells, Mar 2019. URL `https://www.wholecell.org/models/`.

J. R. Karr, J. C. Sanghvi, D. N. Macklin, M. V. Gutschow, J. M. Jacobs, B. Bolival Jr, N. Assad-Garcia, J. I. Glass, and M. W. Covert. A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401, 2012.

F. Y. Kuo and I. H. Sloan. Lifting the curse of dimensionality. *Notices of the AMS*, 52(11):1320–1328, 2005.

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, et al. Machine learning for metabolic engineering: A review. *Metabolic Engineering*, 63:34–60, 2021.

T. T. Le, W. Fu, and J. H. Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256, 2020.

T. J. Lee, G. Yao, D. C. Bennett, J. R. Nevins, and L. You. Stochastic e2f activation and reconciliation of phenomenological cell-cycle models. *PLoS biology*, 8(9):e1000488, 2010.

L. Liang, M. Liu, C. Martin, J. A. Elefteriades, and W. Sun. A machine learning approach to investigate the relationship between shape features and numerically predicted risk of ascending aortic aneurysm. *Biomechanics and modeling in mechanobiology*, 16(5):1519–1533, 2017.

L. Liang, M. Liu, C. Martin, and W. Sun. A deep learning approach to estimate stress distribution: a fast and accurate surrogate of finite-element analysis. *Journal of The Royal Society Interface*, 15(138): 20170844, 2018a.

L. Liang, M. Liu, C. Martin, and W. Sun. A machine learning approach as a surrogate of finite element analysis–based inverse method to estimate the zero-pressure geometry of human thoracic aorta. *International journal for numerical methods in biomedical engineering*, 34(8):e3103, 2018b.

M. Liu, L. Liang, and W. Sun. Estimation of in vivo constitutive parameters of the aortic wall using a machine learning approach. *Computer methods in applied mechanics and engineering*, 347:201–217, 2019.

S. Longobardi, A. Lewalle, S. Coveney, I. Sjaastad, E. K. Espe, W. E. Louch, C. J. Musante, A. Sher, and S. A. Niederer. Predicting left ventricular contractile function via gaussian process emulation in aortic-banded rats. *Philosophical Transactions of the Royal Society A*, 378(2173):20190334, 2020.

D. Lu and D. Ricciuto. Efficient surrogate modeling methods for large-scale earth system models based on machine-learning techniques. *Geoscientific Model Development*, 12(5):1791–1807, 2019.

D. Lu, D. Ricciuto, M. Stoyanov, and L. Gu. Calibration of the e3sm land model using surrogate-based global optimization. *Journal of Advances in Modeling Earth Systems*, 10(6):1337–1356, 2018.

K. O. Lye, S. Mishra, D. Ray, and P. Chandrashekar. Iterative surrogate model optimization (ismo): An active learning algorithm for pde constrained optimization with deep neural networks. *Computer Methods in Applied Mechanics and Engineering*, 374:113575, 2021.

D. N. Macklin, N. A. Ruggero, and M. W. Covert. The future of whole-cell modeling. *Current opinion in biotechnology*, 28:111–115, 2014.

D. N. Macklin, T. A. Ahn-Horst, H. Choi, N. A. Ruggero, J. Carrera, J. C. Mason, G. Sun, E. Agmon, M. M. DeFelice, I. Maayan, et al. Simultaneous cross-evaluation of heterogeneous e. coli datasets via mechanistic simulation. *Science*, 369(6502), 2020.

A. Madani, A. Bakhaty, J. Kim, Y. Mubarak, and M. R. Mofrad. Bridging finite element and machine learning modeling: stress prediction of arterial walls in atherosclerosis. *Journal of biomechanical engineering*, 141 (8), 2019.

L. Marucci, M. Barberis, J. Karr, O. Ray, P. R. Race, M. de Souza Andrade, C. Grierson, S. A. Hoffmann, S. Landon, E. Rech, et al. Computer-aided whole-cell design: taking a holistic approach by integrating synthetic with systems biology. *Frontiers in Bioengineering and Biotechnology*, 8:942, 2020.

doi:10.20944/preprints202209.0455.v1

L. S. Matott and A. J. Rabideau. Calibration of complex subsurface reaction models using a surrogate-model approach. *Advances in Water Resources*, 31(12):1697–1707, 2008.

S. Motta and F. Pappalardo. Mathematical modeling of biological systems. *Briefings in Bioinformatics*, 14 (4):411–422, 2013.

T. T. Nguyen, T. Le Nguyen, and G. Ifrim. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 77–94. Springer, 2020.

J. Nielsen and J. D. Keasling. Engineering cellular metabolism. *Cell*, 164(6):1185–1197, 2016.

S. Nikolopoulos, I. Kalogeris, and V. Papadopoulos. Non-intrusive surrogate modeling for parametrized time-dependent partial differential equations using convolutional autoencoders. *Engineering Applications of Artificial Intelligence*, 109:104652, 2022.

U. Noè, W. Chen, M. Filippone, N. Hill, and D. Husmeier. Inference in a partial differential equations model of pulmonary arterial and venous blood circulation using statistical emulation. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 184–198. Springer, 2016.

U. Noè, A. Lazarus, H. Gao, V. Davies, B. Macdonald, K. Mangion, C. Berry, X. Luo, and D. Husmeier. Gaussian process emulation to accelerate parameter estimation in a mechanical model of the left ventricle: a critical step towards clinical end-user relevance. *Journal of the Royal Society Interface*, 16(156): 20190114, 2019.

R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, pages 485–492, New York, NY, USA, 2016a. ACM. ISBN 978-1-4503-4206-3. doi: 10.1145/2908812.2908918. URL http://doi.acm.org/10.1145/2908812.2908918.

R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, and J. H. Moore. *Applications of Evolutionary Computation: 19th European Conference, EvoApplications 2016, Porto, Portugal, March 30 – April 1, 2016, Proceedings, Part I*, chapter Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, pages 123–137. Springer International Publishing, 2016b.

A. Passi, J. D. Tibocha-Bonilla, M. Kumar, D. Tec-Campos, K. Zengler, and C. Zuniga. Genome-scale metabolic modeling enables in-depth understanding of big data. *Metabolites*, 12(1):14, 2022.

R. Pestourie, Y. Mroueh, T. V. Nguyen, P. Das, and S. G. Johnson. Active learning of deep surrogates for pdes: application to metasurface design. *npj Computational Materials*, 6(1):1–7, 2020.

W. A. Pruett and R. L. Hester. The creation of surrogate models for fast estimation of complex model outcomes. *PloS one*, 11(6):e0156574, 2016.

J. Rees-Garbutt, O. Chalkley, S. Landon, O. Purcell, L. Marucci, and C. Grierson. Designing minimal genomes using whole-cell models. *Nature communications*, 11(1):1–12, 2020.

M. Renardy, T.-M. Yi, D. Xiu, and C.-S. Chou. Parameter uncertainty quantification using surrogate models applied to a spatial model of yeast mating polarization. *PLoS computational biology*, 14(5):e1006181, 2018.

P. Rué and J. Garcia-Ojalvo. Modeling gene expression in time and space. *Annual review of biophysics*, 42: 605–627, 2013.

A. P. Ruiz, M. Flynn, J. Large, M. Middlehurst, and A. Bagnall. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 35(2):401–449, 2021.

K. L. Sack, E. Aliotta, D. B. Ennis, J. S. Choy, G. S. Kassab, J. M. Guccione, and T. Franz. Construction and validation of subject-specific biventricular finite-element models of healthy and failing swine hearts from high-resolution dt-mri. *Frontiers in Physiology*, 9:539, 2018.

L. J. Segerlind. *Applied finite element analysis*. John Wiley & Sons, 1991.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

J. Shu and M. Shuler. A mathematical model for the growth of a single cell of e. coli on a glucose/glutamine/ammonium medium. *Biotechnology and bioengineering*, 33(9):1117–1126, 1989.

G. Smolders, J. Van der Meij, M. Van Loosdrecht, and J. Heijnen. Model of the anaerobic metabolism of the biological phosphorus removal process: stoichiometry and ph influence. *Biotechnology and bioengineering*, 43(6):461–470, 1994.

M. Soheilypour and M. R. Mofrad. Agent-based modeling in molecular systems biology. *BioEssays*, 40(7): 1800020, 2018.

P. Stolfi and F. Castiglione. Emulating complex simulations by machine learning methods. *BMC bioinformatics*, 22(14):1–14, 2021.

H. Taymaz-Nikerel, A. E. Borujeni, P. J. Verheijen, J. J. Heijnen, and W. M. van Gulik. Genome-derived minimal metabolic models for escherichia coli mg1655 with estimated in vivo respiratory atp stoichiometry. *Biotechnology and bioengineering*, 107(2):369–381, 2010.

A. Tealab. Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, 3(2):334–340, 2018.

J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso. Deep learning for time series forecasting: A survey. *Big Data*, 9(1):3–21, 2021.

J. J. Tyson. Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences*, 88(16):7328–7332, 1991.

H. Wang, H. Gao, X. Luo, C. Berry, B. Griffith, R. Ogden, and T. Wang. Structure-based finite strain modelling of the human left ventricle in diastole. *International journal for numerical methods in biomedical engineering*, 29(1):83–103, 2013.

L. Wang and C. D. Maranas. Mingenome: an in silico top-down approach for the synthesis of minimized genomes. *ACS synthetic biology*, 7(2):462–473, 2018.

S. Wang, K. Fan, N. Luo, Y. Cao, F. Wu, C. Zhang, K. A. Heller, and L. You. Massive computational acceleration by using neural networks to emulate mechanism-based biological models. *Nature communications*, 10(1):1–9, 2019.

J. V. Wong, G. Yao, J. R. Nevins, and L. You. Viral-mediated noisy gene expression reveals biphasic e2f1 response to myc. *Molecular cell*, 41(3):275–285, 2011.

D. Xu, Y. Shi, I. W. Tsang, Y.-S. Ong, C. Gong, and X. Shen. Survey on multi-output learning. *IEEE transactions on neural networks and learning systems*, 31(7):2409–2429, 2019.

X. Xu, L. Chen, and P. He. A novel ant clustering algorithm based on cellular automata. *Web Intelligence and Agent Systems: An International Journal*, 5(1):1–14, 2007.

L. Yang, A. Ebrahim, C. J. Lloyd, M. A. Saunders, and B. O. Palsson. Dynamicme: dynamic simulation and refinement of integrated models of metabolism and protein expression. *BMC systems biology*, 13(1): 1–15, 2019.

T.-M. Yi, H. Kitano, and M. I. Simon. A quantitative characterization of the yeast heterotrimeric g protein cycle. *Proceedings of the National Academy of Sciences*, 100(19):10764–10769, 2003.

T.-M. Yi, S. Chen, C.-S. Chou, and Q. Nie. Modeling yeast cell polarization induced by pheromone gradients. *Journal of Statistical Physics*, 128(1):193–207, 2007.

P. C. Young and M. Ratto. Statistical emulation of large linear dynamic models. *Technometrics*, 53(1): 29–43, 2011.