# Systems & Control Transactions

# Integrated hybrid modelling of lignin bioconversion

**Sidharth Laxminarayan[a], Lily Cheung[a], and Fani Boukouvala[a]\***

[a] Georgia Institute of technology, School of Chemical and Biomolecular Engineering, Atlanta, GA, USA 30331
\* Corresponding Author: faniboukouvala@chbe.gatech.edu

## ABSTRACT

Global biomanufacturing is projected to expand rapidly in the coming decade due to advancements in DNA sequencing and manipulation. However, the complexity of cellular behaviour introduces difficulty in modelling and optimizing biomanufacturing processes. Phenomenological models that represent the physics of the system in empirical equations suffer from poor robustness, while their machine learning (ML) counterparts suffer from poor extrapolative capability. On the other hand, hybrid models allow us to leverage both physical constraints and the flexibility of ML. This work describes a new approach for hybrid modeling that integrates the time-variant parameter estimation and ML model training into a singular step. We implement this approach on a proposed scheme for the cell-mediated conversion of a lignin derivative into a bioplastic precursor and show that our integrated hybrid model outperforms the traditional two-step hybrid, phenomenological, and ML model counterparts. Lastly, we demonstrate how to execute an interpretability analysis on the ML component of the integrated hybrid model to reveal new physical insights that are then used to further improve model performance.

**Keywords**: Biosystems, Machine Learning, Dynamic Modelling, Lignin Valorization

## Motivation & Background

Growing need and sustainability concerns are increasing the strain on the traditional manufacturing pipelines for chemicals, drugs, and materials [1]. This has prompted increased interest on bio-centric manufacturing, as bioprocesses leverage the capability of microbial, plant, and animal cells to produce complex commercially relevant chemicals while enjoying a reduced carbon footprint, moderate reactor conditions, and high stereoselectivity [1]. Previously inaccessible renewable feedstocks can now be converted into bulk chemicals with the aid of metabolic strain engineering. Of particular interest is the soil bacterium *Pseudomonas putida*, which can valorize lignocellulosic biomass (Figure 1). Studies have demonstrated that *P. putida* can convert lignin into cis,cis-muconic acid (MA) [2]. MA can then be utilized to create polymers like Nylon and PET [2]. While the bacterial conversion of lignin into MA is documented, the kinetics is not well understood.

Cells are extremely complex systems with numerous reaction pathways, intermediates and products. Factoring in inter and intra-cellular relationships, it becomes intractable to simulate each cell and track metabolic dynamics [2]. This problem is countered using phenomenological models based on biological intuition that neglect information regarding biomass heterogeneity and metabolite influences on empirical parameters [3]. On the other hand, purely machine learning (ML) models have been shown to capture the nonlinear complexity of bioprocesses but have poor extrapolation and interpretability [3]. Additionally, experimental datasets are often sparse and noisy, resulting in poor model development and calibration, especially for ML models.
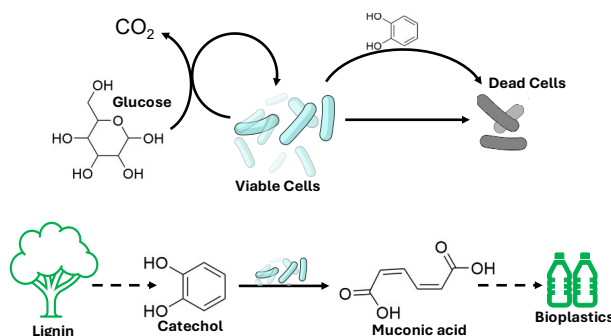


**Figure 1.** Conversion of lignocellulosic biomass into bioplastics by *P. putida*.

Hybrid models have been proposed to leverage the physical constraints of phenomenological models and the flexibility of ML ones [3]. This work explores an embedded hybrid model wherein ML models are utilized to capture the complexity of certain parameters "embedded" within the overarching ODE structure [4]. The hybrid modelling approach aids in capturing the complex relationships between external metabolites and bacteria physiology. Moreover, the embedded approach lends a better framework for interpretability.

To test our hybrid model, we outline a ground truth model that simulates the dynamics of the conversion of catechol, a lignin derivative, to MA by P. putida. The ground truth model also includes glucose as the representative carbon source for bacterial growth and the cell toxicity of catechol. A sparse and noisy dataset is then generated which mimics traditional experimental datasets.

This work will explore two scenarios for constructing hybrid models: A sequential method and an integrated method, which differ on how the parameters of the models are determined. Furthermore, we compare the extrapolative performance of the hybrid model to its phenomenological and ML counterparts. Finally, an interpretability study is performed on the ML components of the hybrid model to gain physical insight. Utilizing Captum python package, an Integrated gradients (IG) analysis is performed which gives insight into how changes in the input affect changes in output for the ML component [5].

## Methods

### Ground truth model

A system of ODEs is constructed to model the dynamics of *P. putida* consuming glucose and converting catechol to MA in a batch process. The system is constructed based on biological principles and previous characterization of *P. putida* [2] The following assumptions are made: (i) the growth rate of viable cells ($X_v$) follows a logistic-Monod equations aimed at better recapitulating the lag phase of batch cultures (Eq. 1 and 2); (ii) glucose ($C_g$) promotes cell division while catechol ($C_c$) inhibits it (Eq. 2); (iii) the cell death rate is composed of a zero-order, catechol-independent component and a catechol-dependent component represented with a Monod-style rate equations (Eq. 2); (iv) dead cells ($X_d$) are not degraded and accumulate in the culture and contribute to the total biomass ($X_t$) (Eq. 4 and 5); (v) glucose is consumed for cell growth and to neutralize the toxic effect of catechol (Eq. 6); (vi) the generation of MA ($C_m$) from catechol has non-growth associated and growth associated components (Eq. 7); and (vii) all the catechol consumed is directly converted to MA by the viable cells and no byproducts are formed (Eq. 8). Note that the following equation can be modified for batch and semi-batch by

changing the spike glucose input ($C_g^{IN}$), spike catechol input ($C_c^{IN}$) and spike time-period ($\tau$).

$$\frac{dX_v}{dt} = (\mu_{gr} - \mu_{dt})X_v \tag{1}$$

$$\mu_{gr} = \left(1 - \frac{X_v}{Y_{xg}C_g^0}\right)\left(\frac{\mu_g C_g}{C_g + K_g}\right) \tag{2}$$

$$\mu_{dt} = k_d + \left(\frac{\mu_c C_c}{C_c + K_c}\right) \tag{3}$$

$$\frac{dX_d}{dt} = \mu_{dt}X_v \tag{4}$$

$$X_t = X_v + X_d \tag{5}$$

$$\frac{dC_g}{dt} = -\left(\frac{\mu_{gr}}{Y_{xg}} + \frac{m_k C_c}{C_c + k_{cm}}\right)X_v + C_g^{IN}f(t) \tag{6}$$

$$\frac{dC_m}{dt} = Y_{mx}(\mu_{gr} + \beta)X_v\frac{C_c}{C_c + k\_cr} \tag{7}$$

$$\frac{dC_c}{dt} = -\frac{dC_m}{dt} + C_c^{IN}f(t) \tag{8}$$

$$f(t) = \sum_{k=0}^{Z}\delta(t - \tau k) \tag{9}$$

Training data is generated by assuming batch operation. A 72-hour batch process is simulated with a 10-point discretization. To better reflect limitations in data acquisition commonly experienced in biomanufacturing, only data on the total biomass, glucose, catechol and MA concentrations are used for training (Fig. 2). For instance, only total biomass is easily measured during culturing (e.g., by measuring changes on optical density) while distinguishing viable and dead cells is more challenging. Glucose, catechol and MA are generally quantified with either chromatographic or biochemistry analyzers, which are expensive and time consuming to operate thereby limiting the amount of data collected.
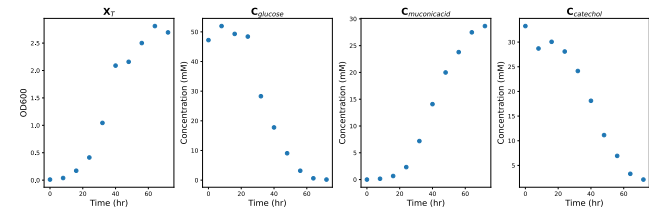


**Figure 2.** Ground truth model of the batch MA production bioprocess.

### Phenomenological model construction

This model reflects the traditional reactor kinetics style system of ODEs that are generally constructed to capture bioprocess dynamics. The data acquisition limitations highlighted in the previous section along with relevant literature informed the model construction. Combining dynamic observations of (Fig. 2) and literature studies [2], the following ODEs for the phenomenological model are created: (i) The total biomass growth is tied positively to glucose and inhibited by catechol via Monod kinetics (Eq.11); (ii) The MA production is observed to

occur in both the growth and stationary phases (Eq. 13); (iii) The MA concentration increased at the same rate as the catechol one decreased (Eq. 14)

$$\frac{dX_t}{dt} = \mu_{gr}X_t \tag{10}$$

$$\mu_{PM} = \left(\frac{\mu_g C_g}{C_g + K_g}\right) - \left(\frac{\mu_c C_c}{C_c + K_c}\right) \tag{11}$$

$$\frac{dC_g}{dt} = -\left(\frac{\mu_{gr}}{Y_{xg}}\right)X_t + C_g^{IN}f(t) \tag{12}$$

$$\frac{dC_m}{dt} = Y_{mx}(\mu_{gr} + \beta)X_t\frac{C_c}{C_c + k\_cr} \tag{13}$$

$$\frac{dC_c}{dt} = -\frac{dC_m}{dt} + C_c^{IN}f(t) \tag{14}$$

$$f(t) = \sum_{k=0}^{Z}\delta(t - \tau k) \tag{15}$$

The spike input of glucose and catechol are set to zero for the training data. Note how the growth of the biomass in the phenomenological model closely mirrors the ground truth model.

## Hybrid model structure – Sequential vs Integrated

The construction of the hybrid model entrails the following steps: (a) time-variant parameter estimations (TVPE), (b) ML model parameter regression and (c) ODE-ML model integration. The integrated method differs from the sequential one as it performs steps (a) and (b) together. The equations shown here reflect the hybrid construction methodology. The observations outlined in the previous section are used to inform the overarching ODE structure. An assumption is made that the growth rate did not have a straightforward relationship between glucose and catechol. An artificial neural network (ANN) is used to model the growth rate which used glucose, catechol, MA and total biomass as inputs. Equations (16-19) highlight the hybrid model structure.

$$\frac{dX_t}{dt} = \mu_{ML}X_t \tag{16}$$

$$\mu_{ML} = ANN(X_t, C_g, C_c, C_m) \tag{17}$$

$$\frac{dC_g}{dt} = -\left(\frac{\mu_{ML}}{Y_{xg}}\right)X_t + C_g^{IN}f(t) \tag{18}$$

$$\frac{dC_m}{dt} = Y_{mx}(\mu_{ML} + \beta)X_t\frac{C_c}{C_c + k\_cr} \tag{19}$$

Equations 14, 15 are included in the hybrid model to track the dynamics of catechol and the spike input frequency. The sequential hybrid modelling structure initially calculates the TVPE and then trains a ML model outside. Notice that within the initial optimization framework the TVPE is calculated as a discrete value at different time steps. A multi-objective optimization problem is formulated with a state space MSE and a smoothing index (sum of second derivatives) for the time variant parameters. The formulation shown below shows the sequential hybrid model's TVPE.

$$\min_{\theta^{TV},\theta_{TIV}} w_M MSE + w_s SM \tag{20}$$

$$s.t.\frac{dC_{si}}{dt} = f_{ODE}(C_{xi}, \theta_{TIV}, \theta_i^{TV}), \forall x = [0,S]\forall i = [0,N] \tag{21}$$

$$MSE = \frac{1}{NtS}\sum_{i=1}^{N}\sum_{j=0}^{t}\sum_{s=0}^{S}(C_{sji}^{pred} - C_{sji}^{act})^2 \tag{22}$$

$$SM = \frac{1}{Nt}\sum_{i}^{N}\sum_{j}^{t-2}(\theta_{ij+1}^{TV} - 2\theta_{ij}^{TV} + \theta_{ij-1}^{TV})^2 \tag{23}$$

$$\theta_{it}^{TV}, \theta_{TIV} \geq 0 \tag{24}$$

$N$ represents the number of experimental datasets, $t$ is the number of time points in each dataset and $S$ is the number of dynamic species. $f_{ODE}$ represent the hybrid system of ODE captured by Eqs (14-19). $\theta_{TIV}$ are the time invariant parameters in Eqs (14-19) while $\theta^{TV}$ represents time variant parameters. The integrated hybrid modelling structure simultaneously estimates the growth parameter while also training the ANN model to capture the parameter behavior. The optimization formulation shown below details how the ML model is trained and the TVPE is performed.

$$\min_{W,b,\theta_{TIV}} \frac{1}{NtS}\sum_{i=1}^{N}\sum_{j=0}^{t}\sum_{s=0}^{S}(C_{sji}^{pred} - C_{sji}^{act})^2 \tag{25}$$

$$s.t.\frac{dC_{si}}{dt} = f_{ODE}(C_{xi}, \theta_{TIV}, \theta_i^{TV}), \forall x = [0,S]\forall i = [0,N] \tag{26}$$

$$\theta_i^{TV} = f_{ML}(W, b, C_{xi}) \forall i = 0 \text{ to } N \tag{27}$$

$$\theta_{TIV} \geq 0, W, b \subset [-1,1] \tag{28}$$

$f_{ML}$ represents the ANN whose weights and biases ($W$ & $b$) are used to model the physics of the $\theta^{TV}$. The sequential hybrid modelling structure ignores the second equality constrain shown in Eq. (27) which captures the relationship between the time variant parameter and the ML model. Note that both the integrated and sequential method has hyperparameters which need to be tuned. For the phenomenological, integrated and sequential scenarios, the IPOPT multistart solver with fifty different initializations is used to solve the Pyomo DAE formulation [5].

## Black box model construction

The black box model is constructed using an ANN surrogate. The model inputs are the initial condition and time. A preprocessing step is performed where the inputs and outputs are scaled using a Z-score normalization. A rigours 5-fold cross validation and grid search is performed for tuning the hyperparameters of the system. The activation function, the number of hidden layers, the number of nodes, and the strength of the L2 regularization are all varied.

$$(X_t^t, C_g^t, C_m^t, C_c^t) = ANN(X_t^0, C_g^0, C_m^0, C_c^0, t) \tag{29}$$

## Extrapolative performance comparison

The initial data is generated by performing a Latin hypercube sampling of three initial conditions in the following ranges of initial glucose and catechol concentration: [50 mM, 70 mM] and [25 mM, 35 mM] with a constant total biomass concentration of 0.01 OD600. The performance of all models is evaluated utilizing an initial condition extrapolation for a fed-batch bioprocess. Note how in the training data the spike inputs are set to zero, however for the performance testing the spike inputs are respectively set to 10 mM and 3 mM for glucose and catechol with a time-period of 25 hrs. In the extrapolative scenarios, the glucose and catechol initial concentrations are varied from [20 mM, 100 mM] and [5 mM, 55 mM], respectively. Heatmaps and distribution curves of $R^2$ values are generated to compare the performance of different models and highlight performance in edge-cases.

### Interpretability analysis

Utilizing the Captum software, the ANN growth rate in Eq. 17 is analyzed and some of the model behaviours are explained [6]. A primary attribution analysis called Integrated gradients (IG) analysis is performed which tells the contribution of each input to the output of the model. The higher the attribution scores the more the input contributes to the model output [6]. Positive attribution means the output increase with the increase of the input and vice versa [6]. The attribution is calculated utilizing an integrated gradients technique.

### Improved hybrid model

Based on some insight from the ML interpretability analysis, new physical insights are integrated into the hybrid model to improve its overall performance. The physical insight added to the hybrid model is the Monod-kinetics relationship between biomass growth and glucose. The following ODE equations shown details the improved hybrid model. The improved hybrid model utilizes an ANN to model the fraction of viable biomass in the reactor. Equations 14, 15 are included in the improved hybrid model to track the dynamics of catechol and the spike input frequency.

$$\frac{dX_t}{dt} = \mu f_v X_t \tag{30}$$

$$f_v = ANN(X_t, C_g, C_c, C_m) \tag{31}$$

$$\mu = \left(\frac{\mu_g C_g}{C_g + k_g}\right) \tag{32}$$

$$\frac{dC_g}{dt} = -\left(\frac{\mu}{Y_{xg}}\right) f_v X_t + C_g^{IN} f(t) \tag{33}$$

$$\frac{dC_m}{dt} = Y_{mx}(\mu + \beta) f_v X_t \frac{C_c}{C_c + k\_cr} \tag{34}$$

## Results and Discussion

### Sequential vs Integrated hybrid modeling

The hyperparameter tuning for the integrated model

is a 3-fold cross validation to determine the ANN architecture. The hyperparameter tuning for the sequential model consist of two-steps: (i) a 3-fold cross validation to determine the smoothing weight ($w_s$) and (ii) a 5-fold cross validation to determine the ANN architecture. To simplify the hyperparameter search, the MSE weight ($w_M$) is set to a constant of unity. The performance of the sequential model is highest at a smoothing weight of 10. Following this, a grid search is done to determine the ANN architecture by varying the activation function, number of layers and nodes and L2 regularization weight.

The integrated model outperforms the sequential one across all points of extrapolation despite negligible computational cost differences (Figure 3). The heatmap can be divided into three broad regions: (I) high glucose, (II) low-glucose and low-catechol, and (III) low-glucose and high-catechol. The integrated models performs better than the sequential model in region (I) and (III) but suffers in region (II). In fact, the integrated model performs its best predictive performance in region (I) given the strong effect of high glucose concentrations on the growth rate. However, in region (II) the low glucose compounded with low catechol concentrations decreases the fraction of viable biomass which decreases growth rate. The integrated model overestimates the growth rate in this region. These results agree with our ground truth model which has a stunted growth rate at low glucose and low catechol concentrations, thereby negatively affecting the biomass accumulation. The CPU seconds of the integrated and sequential models showed marginal differences, 1.53 s and 0.95 s, respectively. Overall, while the integrated approach has advantages in performance, its cost will increase as the size of the ANN increases.
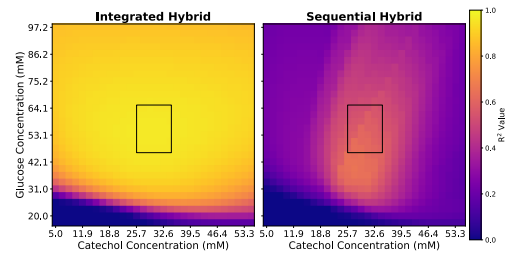


**Figure 3.** Extrapolative performance heatmap comparing the integrated and squential models.

### Hybrid model extrapolative performance

The previous section shows the performance of the integrated hybrid model to be better than the sequential hybrid one. Thus, the following sections will exclusively utilize the integrated hybrid model for comparison with the phenomenological and black box models. On average, the hybrid model outperforms both the phenomenological and back box models in extrapolation

(Figure 4). The black box model had extremely poor extrapolative performance; while it captured the intricacies of the training data, it struggled to extrapolate the behaviour. The hybrid model outperforms the phenomeno-logical model in regions (I & III). However, it is extrapolative performance suffers slightly in regions (II) compared to its phenomenological counterpart, given that the latter has the physical insight which allows it to anticipate the contribution of catechol in low glucose environments.

Next, we sought to test versatility of the hybrid model. The glucose consumption in the ground truth model (Eq. 6) has two contributing terms: biomass growth and maintenance. The maintenance is tied to concentration of the catechol concentration. The hybrid model and the phenomenological model are formulated with the assumption that the maintenance term is not a huge contributor to glucose consumption. The versatility of the hybrid modelling structure is tested by increasing the contribution of the glucose consumption to match the growth-related glucose consumption. Despite the change in the physics of the ground truth model, Eq. 10-19 are not changed. The hybrid model still outperforms both the phenomenological and black box model in all the regions (Figure 5), showing that the versatility of the embedded ANN can capture the intricacies of the growth rate while also accounting for the additional strain of maintenance. The hybrid model's performance suffers in region (III); however, a trend observed to occur in the phenomenological model as well. The higher impact of catechol reduces the fraction of viable cells thereby overestimating the metabolic activity of the biomass to convert catechol into MA. This analysis demonstrates the flexibility of the integrated hybrid modelling structure to adapt to changes in the underlying physics of the system.
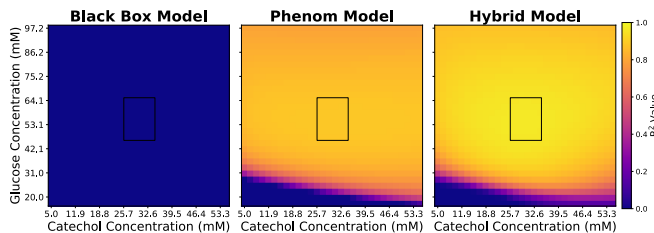

**Figure 4.** Extrapolative performance heatmap comparing the ML, phenomenological and hybrid models
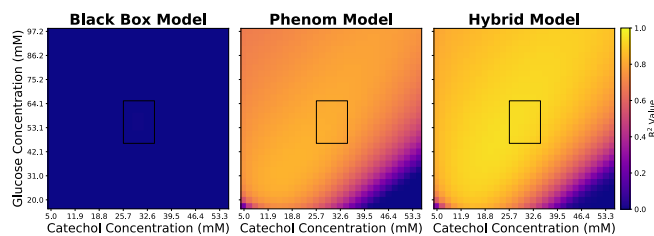

**Figure 5.** Extrapolative performance boxplot comparing the ML, phenomenological and hybrid models for greater contribution of maintenance coefficient effect on glucose

## Hybrid model interpretability

The hybrid model still has an embedded ANN piece within its structure and understanding the physics that is captured by the ANN component could aid in model refinement and experimental design. The bar graph shows the attribution score for the ANN component of the hybrid model. Note, the primary contributor to the model's predictive performance is the glucose and catechol concentrations. The glucose has a positive effect while catechol has a negative effect. This aligns with the physics that is present in the ground truth model. When we further analyze the attribution score of glucose as the concentration changes, further biophysical insights are caught. Fig. 6 looks at the glucose attribution score as the normalized glucose and catechol concentration changes. An important trend is how the attribution versus glucose plot mirrors the relationship between growth rate and substrate for Monod-style growth kinetics. Similarly note that catechol attribution score increases in magnitude as the concentration increases.
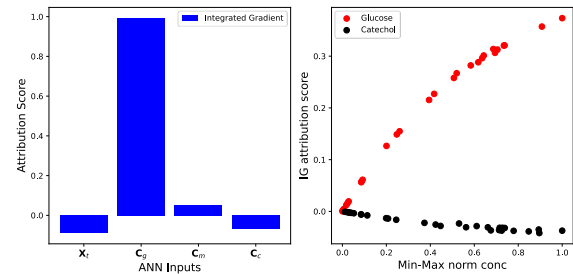

**Figure 6.** Primary attribution scores for the ML model in the hybrid ODE for training data. Model attribution score for changes in glucose and catechol concentration

The attribution score analysis presented above can function as first step in constructing more accurate empirical models to capture unexplained phenomenological behaviour. With the aid of the ML model interpretability, we could implement a build-train-improve cycle to incorporate and improve the physics captured by the hybrid model. We expect such an approach to be successful when the physics captured by the ANN is dominated by a single input. However, ANNs with multiple equally contributing inputs and multiple outputs might present difficulties to the interpretability analysis.

## Improved hybrid models with increased physics embedding

Based on the interpretability analysis, a Monod style relationship between glucose and growth rate is determined and incorporated into the improved hybrid model (Eq. 28-32). The improved hybrid model utilizes the ANN flexibility to capture the fraction of viable biomass in the system, which is the only entity metabolizing both glucose and catechol. Consequently, the aim of the

improved hybrid is capturing this fact even though only total biomass (viable plus dead cells) is used for training. As seen on the performance distributions plots below, the improved hybrid models outperformed the original hybrid model due to an increased accuracy in low-glucose high-catechol environments. The improved model performs better in regions where the original hybrid one struggles. Note the thinner and taller distribution peak of the improved model in high R2 value range. Since some burden is alleviated from the ANN component by incorporating learned physics insight, the black-box part of the improved hybrid model has less impact on the generalizability of the overall prediction. This two-step approach to hybrid modeling highlights how the black-box components, coupled by quantitative interpretability analysis could expedite model development. If important dynamics are not capture neither in the partial physics nor the data, the analysis might not lead to an improved hybrid model.
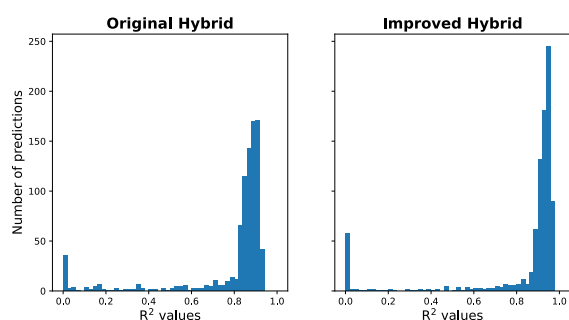


**Figure 7.** Extrapolative performance distributions for original hybrid and improved hybrid models.

## Conclusion

Here we outline the implementation of a hybrid modelling strategy to capture the dynamics of the bioconversion of catechol into MA, a bioplastics precursor. The biomass growth in the hybrid model is captured utilizing an ANN. A new hybrid model construction strategy is introduced which integrated the TVPE and the ML model training into a single step. The versatility, extrapolative capability, and sensitivity to initial conditions of the hybrid model is assessed by increasing the influence of various physical phenomena in the ground truth model. Our results show that the integrated hybrid model outperforms its sequential counterpart as well as commonly used phenomenological and ML models. An attribution analysis is performed to lend interpretability to the ANN piece of the hybrid model, revealing buried ground truth behaviour, which is then used to guide a model improvement cycle. The ultimate improved hybrid model reduced the computational load of the ANN by incorporating previously discovered biophysical insights.

## Digital supplementary material

The GitHub page has the code, data and figures shown in this paper: Escape35_SL_code.

## References

1. Clomburg, J. M.; Crumbley, A. M.; Gonzalez, R. Industrial Biomanufacturing: The Future of Chemical Production. Science 2017, 355 (6320). https://doi.org/10.1126/science.aag0804.
2. Mokwatlo, S. C.; Klein, B. C.; Benavides, P. T.; Tan, E. C. D.; Kneucker, C. M.; Ling, C.; Singer, C. A.; M. A.; Ramirez, K. J.; Johnson, C. W.; Beckham, G. T.; Salvachúa, D. Bioprocess Development and Scale-up for Cis,Cis-Muconic Acid Production from Glucose and Xylose by Pseudomonas Putida. Green Chem. 2024, 26 (19), 10152–10167. https://doi.org/10.1039/D4GC03424D.
3. Tsopanoglou, A.; Jiménez del Val, I. Moving towards an Era of Hybrid Modelling: Advantages and Challenges of Coupling Mechanistic and Data-Driven Models for Upstream Pharmaceutical Bioprocesses. Current Opinion in Chemical Engineering 2021, 32, 100691. https://doi.org/10.1016/j.coche.2021.100691.
4. Bradley, W.; Kim, J.; Kilwein, Z.; Blakely, L.; Eydenberg, M.; Jalvin, J.; Laird, C.; Boukouvala, F. Perspectives on the Integration between First Principles and Data-Driven Modeling. Computers & Chemical Engineering 2022, 166, 107898. https://doi.org/10.1016/j.compchemeng.2022.107898.
5. Nicholson, B.; Siirola, J. D.; Watson, J.-P.; Zavala, V. M.; Biegler, L. T. Pyomo.Dae: A Modeling and Automatic Discretization Framework for Optimization with Differential and Algebraic Equations. Math. Prog. Comp. 2018, 10 (2), 187–223. https://doi.org/10.1007/s12532-017-0127-0.
6. Kokhlikyan, N.; Miglani, V.; Martin, M.; Wang, E.; Alsallakh, B.; Reynolds, J.; Melnikov, A.; Kliushkina, N.; Araya, C.; Yan, S.; Reblitz-Richardson, O. Captum: A Unified and Generic Model Interpretability Library for PyTorch. arXiv September 16, 2020. https://doi.org/10.48550/arXiv.2009.07896