Process Biochemistry 143 (2024) 174–185

Contents lists available at ScienceDirect

Process Biochemistry

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

Experimentally implemented dynamic optogenetic optimization of ATPase expression using knowledge-based and Gaussian-process-supported models

Sebastián Espinel-Ríos [a,1], Gerrich Behrendt [a,1], Jasmin Bauer [a], Bruno Morabito [b], Johannes Pohlodek [c], Andrea Schütze [a], Rolf Findeisen [c], Katja Bettenbrock [a], Steffen Klamt [a,*]

[a] Analysis and Redesign of Biological Networks Group, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, Magdeburg, 39106, Germany
[b] Yokogawa Insilico Biotechnology GmbH, Meitnerstraße 9, Stuttgart, 70563, Germany
[c] Control and Cyber-Physical Systems Laboratory, Technical University of Darmstadt, Landgraf-Georg-Straße 4, Darmstadt, 64283, Germany

ARTICLE INFO

Keywords:
Dynamic metabolic control
Optimization
Modeling
Gaussian processes
Gene expression
Optogenetics

ABSTRACT

Optogenetic modulation of adenosine triphosphatase (ATPase) expression represents a novel approach to maximize bioprocess efficiency by leveraging enforced adenosine triphosphate (ATP) turnover. In this study, we experimentally implement a model-based open-loop optimization scheme for optogenetic modulation of the expression of ATPase. Increasing the intracellular concentration of ATPase, and thus the level of ATP turnover, in bioprocesses with product synthesis coupled with ATP generation, can lead to increased substrate uptrake and product formation. Previous simulation studies formulated optimal control problems using dynamic constraint-based models to find optimal light inputs in fermentations with optogenetically mediated ATPase expression. However, using these models poses challenges due to resulting bilevel optimizations and complex parameterization. Here, we outline a simplified unsegregated and quasi-unstructured kinetic modeling approach that reduces the number of dynamic states and leads to single-level optimizations. The models can be augmented with Gaussian processes to compensate for model uncertainties. We implement optimal control constrained by knowledge-based and hybrid models for optogenetic ATPase expression in *Escherichia coli* with lactate as the main product. To do so, we genetically engineer *E. coli* to obtain optogenetic expression of ATPase using the CcaS/CcaR system. This represents the first experimental implementation of model-based optimization of ATPase expression in bioprocesses.

1. Introduction

Global challenges such as climate change, depletion of non-renewable fossil resources, and a growing population are driving the search for sustainable production systems. Microbial cell factories are (engineered) microorganisms capable of synthesizing valuable metabolites from renewable resources. They show potential to substitute, e.g., petrochemical production of chemicals, materials, and fuels with bio-based and sustainable alternatives [1,2]. It is often necessary to optimize both the production processes as well as the cell's metabolism to achieve product titers, yields, and volumetric productivities that ensure profitability [3,4]. The product yield determines how much substrate is needed to produce a certain amount of product. Volumetric productivity is the rate of product formation per culture volume and determines how fast production occurs.

In *traditional* genetic and metabolic engineering, the cell is rewired to optimize the *steady-state* metabolic flux distribution, often toward maximizing the product yield under a specific cultivation environment. Maximizing the product yield diverges resources from biomass synthesis, thus frequently decreasing the volumetric productivity [5]. Furthermore, the gene expression of enzymes involved in production pathways is constitutive in several cases, i.e., always active and happening at a constant rate [6–8]. Thus, from a control engineering perspective, traditional metabolic engineering follows a *static* approach. That is, the cell metabolism is not *a priori* engineered to actively respond to external or internal signals for steering gene expression and metabolism toward desired production *modes*. Of course, the cell's metabolism naturally reacts to several external/internal signals; this is vital for microorganisms. However, in *static* metabolic engineering, these signals are not dynamically exploited for the optimization of

* Corresponding author.
E-mail address: klamt@mpi-magdeburg.mpg.de (S. Klamt).
[1] These authors contributed equally.

https://doi.org/10.1016/j.procbio.2024.04.032
Received 17 January 2024; Received in revised form 19 March 2024; Accepted 23 April 2024
Available online 30 April 2024

metabolism. Since the cell has a *static* flux distribution for engineered pathways, these pathways cannot adapt to changing conditions, hence lacking operational flexibility.

Alternatively, cells can be engineered to express metabolism-relevant proteins such as enzymes in an *inducible* and *dynamic* fashion, following a dynamic metabolic engineering approach [8–11,12]. This idea can be exploited for manipulating metabolic fluxes using external inputs for online process optimization and control [13,14]. Dynamic metabolic control can enable, e.g., optimal shift between growth and production metabolic *modes*. It can also help to minimize the metabolic burden associated with the constitutive expression of metabolic enzymes/pathways. Dynamic adenosine triphosphate (ATP) turnover is a promising metabolic control strategy. In bioprocesses where the product synthesis is stoichiometrically coupled with ATP formation, enforcement of ATP loss can lead to an increase in product formation and substrate uptake (cf. e.g. [15,16,14,17–20]).

In previous *simulation-based* studies [13,14], we proposed to put the F_1-subunit of the adenosine triphosphatase (ATPase) enzyme, responsible for the hydrolysis of ATP into ADP, under the regulation of an optogenetic gene expression system, i.e., making it inducible by light. This would enable one to influence the intracellular amount of ATPase by manipulating light and thereby the level of ATP *wasting*. In this work, we interchangeably use the term ATPase to refer to the F_1-subunit of that enzyme complex. Light can actuate on biological systems in a precise spatiotemporal, orthogonal, and reversible way, which is convenient for external control of gene expression [21,22]. Finding the optimal light trajectories for process optimization imposes, however, several challenges. For instance, for dynamic ATP turnover applications, one needs to carefully fine-tune the level of the F_1-ATPase, and thus the ATP turnover, to avoid driving the cell into unstable states [23].

In [13,14], we showed a *model-based* optimal control problem to find the optimal light trajectories for enhanced product yield via optogenetic modulation of the ATPase. To model the system, we outlined a dynamic constraint-based model that integrates the dynamics of metabolic reactions, the light-inducible genetic actuator, and resource allocation phenomena. The model considers extracellular and intracellular metabolites, as well as intracellular biomass components. The extracellular metabolites and the intracellular biomass components are the dynamic states, while the intracellular metabolites are assumed in a quasi-steady state. The latter model is formulated as an optimization problem subject to constraints and can be considered an extended version of dynamic enzyme-cost flux balance analysis [24].

Although the model in [13,14] offers deep insight into metabolism and resource allocation, it can be *too complex* for experimental implementations. It can be technically challenging to gather measurements of the dynamic states to parameterize and validate such a model due to, e.g., a lack of adequate sensors and the unavailability of analytical technologies. Although one could in principle use *soft* sensors (cf. e.g. [25,26]), their suitability depends on the number of states that can be measured and the quality of the underlying mathematical models, which are sometimes limited.

The modeling and optimization framework in [13,14] is also computationally expensive. Because the underlying dynamic model is formulated as an optimization problem, the resulting model-based optimal control problem turns out to be a *bilevel optimization*. Solving bilevel optimizations is not trivial. One often needs to make assumptions on the relation between the upper- and lower-level optimization problems, e.g., an *optimistic* or *pessimistic* relation (cf. [27–29] for more details). As done in [13,14], we can reformulate the bilevel optimization into a single-level optimization, following an optimistic approach, by substituting the lower-level problem by its corresponding Karush-Kuhn-Tucker conditions. This results in a mathematical program with complementarity constraints which becomes non-convex given the non-linearity of the complementarity constraints, hence, in general, difficult to solve. The Lagrange multiplier or dual variables, coming from the Karush-Kuhn-Tucker conditions, further increase the size of the optimization problem as they become additional optimization variables.

Due to the previous reasons, a more straightforward modeling approach for fermentations with optogenetic control of the ATPase is pertinent. For simplicity of experimental validation and implementation, we seek to minimize the number of dynamic states, without sacrificing the model predictability for model-based optimal control of metabolism. To this end, we model only the most relevant extracellular states and the optogenetically controlled intracellular enzyme. In situations of significant model uncertainty, we outline the use of machine learning to *learn* the *error* of the *a priori* known dynamic equations, thus rendering a *hybrid* model. Experiments in biotechnology are, nevertheless, often expensive and time-consuming; thus, large and high-quality training datasets for machine learning are typically scarce. Here, we focus on Gaussian processes [30] because they can offer good predictability even if small-to-medium datasets are available (cf. e.g. [31–33]), as is the case in this work.

We formulate a single-level model-based dynamic optimization problem using the *simplified* system model. This circumvents bilevel optimization schemes, facilitating the numerics and computational effort. As a major contribution, we experimentally validate our modeling and optimization framework using the anaerobic lactate fermentation of an engineered *Escherichia coli* with optogenetic control of the ATPase, the same biological system considered in [13,14]. Remark that the latter references are simulation-based studies, while in this work also experimental validation is shown.

A scheme of the dynamic optimization control strategy is presented in Fig. 1. Note that the control strategy could be in principle performed both in an open-loop or closed-loop manner. For simplicity of experimental implementation, this study focuses exclusively on open-loop control. Closed-loop control would require real-time sensors for process monitoring, which were not available to us. The process inputs can in principle encompass both *intracellular* and *extracellular* degrees of freedom, e.g., light intensity to modulate ATPase expression (*intracellular*) and substrate feeding rate in case of fed-batch fermentations (*extracellular*). This study specifically focuses on batch processes, thus no feeding rate is considered. Additionally, the model utilized for constraining the dynamic optimization can be solely based on knowledge or of a hybrid nature incorporating machine-learning parts such as Gaussian processes. In this study, we consider both of these model alternatives.

The remainder of this paper is structured as follows. First, we offer an overview of the biological system and optogenetic setup, which will be the basis for introducing our modeling framework and dynamic optimization problem. Afterwards, we outline specific model assumptions for the lactate fermentation case with optogenetic control of the ATPase. Finally, we present the experimental results that support the proposed modeling and optimization framework.

2. Materials and methods

2.1. Overview of the biological system and experimental setup

Under anaerobic fermentation of glucose, lactate production is net ATP positive. As a basis to showcase the proposed model-based optimization strategy of the optogenetically regulated ATPase, we consider an *E. coli* strain with blocked ethanol and acetate pathways (cf. Fig. 2-A) [15]. Therefore, lactate synthesis becomes the main fermentation pathway to achieve redox balance, making lactate production suitable for enforced ATP turnover [14–17].

We aim to control gene expression of the ATPase using light as an external input. To do so, we utilize the two-component system known as CcaS/CcaR (chromatic acclimation sensor/regulator) to establish control over ATPase F_1-subunit (*atpAGD*) expression in *E. coli*. The CcaS/CcaR system, originated from cyanobacteria, allows for the regulation of gene expression by changing the red-to-green light ratio to which the cells are exposed (cf. e.g. [34–36]). Green light causes CcaS to

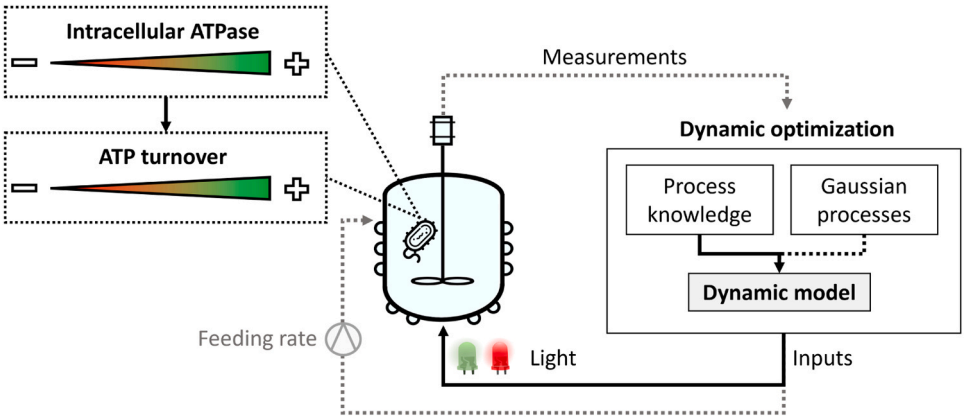

Fig. 1. Open-loop control of ATP turnover via optogenetic modulation of the ATPase in batch processes. In gray we show other potential configurations such as closed-loop control and fed-batch fermentations, although these fall out of the scope of this study.

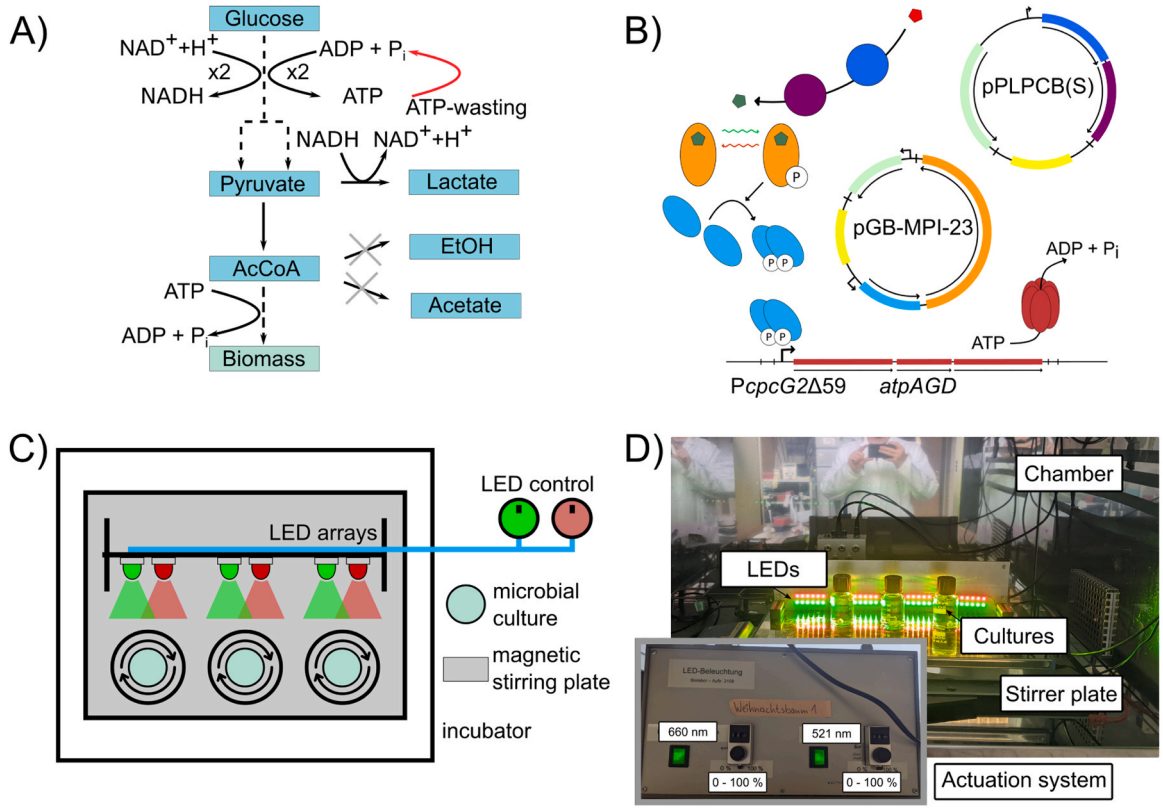

Fig. 2. A) Simplified core metabolism of the microorganism used in this study, i.e., *E. coli* sGB015 with enforced ATP wasting. It shows the conversion of glucose into lactate. Relevant redox and energy co-factors are presented. B) The light-inducible ATP wasting is managed by three heterologous genetic elements in *E. coli* sGB015: pPLPCB(S), pGB-MPI-23 and the chromosomal insertion of PcpcG2Δ59-*atpAGD*-rrnBT1. Plasmids are shown circular, while the chromosome is shown linear. Note that genes (italicized) and their related proteins (non-italicized) are shown with the same color. pPLPCB(S) expresses *ho1* (dark blue) and *pcyA* (purple), thereby enabling the conversion of heme (red pentagon) into phycocyanobilin (green pentagon), the chromophore necessary for CcaS to detect light. The expression of *ccaS* (orange) and *ccaR* (pale blue) is enabled through pGB-MPI-23. CcaS autophosphorylates after a conformational change induced by green light with the photon-protein interaction enabled through phycocyanobilin. Afterward, CcaS phosphorylates CcaR (phosphate group P is represented by circles), leading to CcaR dimerization and functioning as a transcription factor for PcpcG2Δ59 on the chromosome, thereby initiating the expression of *atpAGD* (red). Promoters are presented as arrows on the plasmids and the chromosome. Open reading frames are shown as arrows next to their related genes. Terminators are marked as black perpendicular lines. Origins of replication are shown in yellow and antibiotic resistances in pale green. C) Scheme of the fermentation setup with the green and red light delivery system based on LEDs. D) Photograph of the actual setup shown in C). The temperature regulation chamber, magnetic stirrer plate, LED arrays, and actuation system are labeled. The actuation system is shown with the two tunable regulators for determining the current that is delivered to the green and red LED arrays.

autophosphorylate and then phosphorylate CcaR. CcaR dimerizes when phosphorylated, becoming an activating transcription factor. Red light causes CcaS to dephosphorylate and gene expression is repressed.

To construct our biological system, we inserted an *atpAGD* expression cassette into the chromosome of *E. coli* in which all genes are regulated by the optimized promoter of *cpcG2* [37]. Subsequently, we introduced plasmid pPLPCB(S) for the production of phycocyanobilin, a cofactor necessary for photo-sensing, as well as plasmid pGB-MPI-23 to

express the CcaS/CcaR proteins. This resulted in the final strain *E. coli* sGB015 where ATPase expression is regulated by green and red light (cf. Fig. 2-B). Refer to Section 2.1.1 for the detailed genetic engineering procedure.

Anaerobic fermentation experiments with *E. coli* sGB015 subjected to optogenetic manipulation were carried out at 37 ˚C in a Certomat BS-1 incubator (B. Braun Biotech International) (cf. Fig. 2-C). Red and green light outputs were generated through light-emitting diode (LED) arrays (Osram OSLON SSL LED green $\lambda_{peak} = 521$ nm; Osram OSLON SSL LED red $\lambda_{peak} = 660$ nm), whereby the emitted light was regulated via the supplied current. The corresponding photon flux density in μmol m^{-2} s^{-1} units was determined using an ULM-500 Universal Light Meter (Heinz Walz GmbH). All experiments were performed using three biological replicates. For more details on the fermentation procedure and analytical measurements, refer to Section 2.1.2.

2.1.1. Genetic engineering procedure

The strain KBM10111s (= MG1655 Δ *adhE* Δ *ackA-pta*) [15] was used as a basis to create an *E. coli* strain that produces lactate as the main fermentation product and portrays a light-inducible expression of *atpAGD*. First, pGB-MPI-035, a modified version of pSKA397 [34], was transformed into KBM10111s together with pTNS3 [38] for Tn7-based insertion of *atpAGD* regulated by PcpcG2Δ59 [37], downstream of *glmS* [39]. This resulted in sGB013 (= MG1655 Δ *adhE* Δ *ackA-pta* Tn7::*cat*-PcpcG2Δ59-*atpAGD*-rrnBT1). The chloramphenicol resistance cassette of this strain was removed though FLP activity by transformation with pCP20 [40], resulting in sGB014 (= MG1655 Δ *adhE* Δ *ackA-pta* Tn7::PcpcG2Δ59-*atpAGD*-rrnBT1). The genes for the light-regulated transcription system, *ccaS/ccaR* controlled through PccaR, were introduced in a *sfgfp*-deficient variant of pSKA413 [34] (pGB- MPI-23). Furthermore, the *Synechocystis* PCC6803 genes *ho1* and *pcyA*, enabling the conversion of heme into phycocyanobilin, were introduced by transformation with pPLPCB(S) [41]. Both plasmids were transformed into sGB014, resulting in the final strain used for all experiments, i.e., *E. coli* sGB015 (= MG1655 Δ *adhE* Δ *ackA -pta* Tn7::PcpcG2Δ59-*atpAGD*-rrnBT1 pPLPCB(S) pGB-MPI-23). Sequence files for all genetic elements created in this work can be found in genebank format through the Edmond repository: https://doi.org/10.17617/3.H5GT8I.

2.1.2. Fermentation experiments and analytical measurements

Single colonies of *E. coli* sGB015 were used to start 10 ml aerobic cultures with LB_0 (10 g/l tryptone, 5 g/l yeast extract, 5 g/l NaCl) in 100 ml shake flasks with baffles at 37 ˚C and 200 rpm. Next, precultures with standard defined medium (4 g/l glucose, 34 mM NaH_2PO_4, 64 mM K_2HPO_4, 20 mM $(NH_4)_2SO_4$, 9.52 mM $NaHCO_3$, 1 μM $Fe(SO_4)_4$, 300 μM $MgSO_4$, 1 μM $ZnCl_2$, 10 μM $CaCl_2$) [42] with 150 μg/ml spectinomycin and 25 μg/ml chloramphenicol were inoculated and grown overnight under red light in 50 ml Schott flasks with 25 ml culture volume at 37 ˚C and 180 rpm. The main fermentation experiments were inoculated from the latter preculture in fresh standard defined medium, cultivated in 50 ml flasks with 50 ml culture volume. For sampling purposes, the vessels were briefly transferred to a Whitley A25 anaerobic workstation (Meintrup DWS Laborgeräte GmbH) with an oxygen-free atmosphere (80 % N_2, 10 % CO_2, 10 % H_2). During cultivation, the flasks were tightly closed to prevent gas exchange.

Glucose concentrations were measured with the HK assay kit (Megazyme Ltd.). Lactate was measured by reversed phase HPLC utilizing an Inertsil ODS-3 column (5 μm, RP-18 100A, 250 × 4.6 mm) (GL Sciences Inc.) with a flow rate of 1.0 ml/min, using a running buffer consisting of 0.1 M $NH_4H_2PO_4$ at pH 2.6 and 40 ˚C. The injection volume was 10 μl and detection was performed with a UV-DAD detector at 210 nm.

2.2. Modeling of fermentations with optogenetic ATPase expression

We outline the general structure of our modeling approach for fermentations with optogenetic regulation of the ATPase expression. For clarity of presentation, we keep for the moment the notation general, while in Section 3.1 we elaborate on the specific model assumptions for the lactate fermentation case study. Note that we use bold fonts for vectors and matrices, and non-bold fonts for scalar variables and parameters.

2.2.1. General model formulation

We consider the dynamics of biomass $B \in \mathbb{R}$, *rate-limiting* external substrates $\boldsymbol{s} \in \mathbb{R}^{n_s}$, *rate-limiting* (by)products and products of interest $\boldsymbol{p} \in \mathbb{R}^{n_p}$, as well as the intracellular ATPase $E \in \mathbb{R}$. While B, $\boldsymbol{s}$, and $\boldsymbol{p}$ are expressed in mass per culture volume, E is expressed in mass of ATPase per mass of cells. We treat the biomass as a homogeneous population of cells, hence the model is *unsegregated* and the biomass is modeled as a single component. Since we lump up intracellular metabolism, except for the dynamics of the ATPase expression, the model is also *quasi-unstructured*. The dynamic input $u \in \mathbb{R}$ is the green light photon flux density. Based on these considerations, the process dynamics read

$$\frac{dB}{dt} = \boldsymbol{S_B}\boldsymbol{r}(\boldsymbol{x}(t), u(t), \boldsymbol{\theta}) + \boldsymbol{Q_B}\boldsymbol{w}(\boldsymbol{x}(t), u(t), \boldsymbol{\tau}), \tag{1a}$$

$$\frac{dE}{dt} = \boldsymbol{S_E}\boldsymbol{r}(\boldsymbol{x}(t), u(t), \boldsymbol{\theta}) + \boldsymbol{Q_E}\boldsymbol{w}(\boldsymbol{x}(t), u(t), \boldsymbol{\tau}), \tag{1b}$$

$$\frac{d\boldsymbol{s}}{dt} = \boldsymbol{S_s}\boldsymbol{r}(\boldsymbol{x}(t), u(t), \boldsymbol{\theta}) + \boldsymbol{Q_s}\boldsymbol{w}(\boldsymbol{x}(t), u(t), \boldsymbol{\tau}), \tag{1c}$$

$$\frac{d\boldsymbol{p}}{dt} = \boldsymbol{S_p}\boldsymbol{r}(\boldsymbol{x}(t), u(t), \boldsymbol{\theta}) + \boldsymbol{Q_p}\boldsymbol{w}(\boldsymbol{x}(t), u(t), \boldsymbol{\tau}), \tag{1d}$$

$$\boldsymbol{x} := [B, E, \boldsymbol{s}^\top, \boldsymbol{p}^\top]^\top, \tag{1e}$$

$$\boldsymbol{x}(t_0) = \boldsymbol{x_0}, \tag{1f}$$

where $t \in [t_0, t_f] \subset \mathbb{R}_{\geq 0}$ is the process time, t_0 the initial process time, and t_f the final process time.

In the previous equations, $\boldsymbol{r} : \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_r}$ is a vector-valued function that comprises the reaction rates of the process, e.g., production, consumption, degradation, and dilution rates. $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ are parameters of the reaction rates. $\boldsymbol{S_B} \in \mathbb{R}^{1\times n_r}$, $\boldsymbol{S_E} \in \mathbb{R}^{1\times n_r}$, $\boldsymbol{S_s} \in \mathbb{R}^{n_s\times n_r}$, $\boldsymbol{S_p} \in \mathbb{R}^{n_p\times n_r}$ map the coefficients of the reaction rates to the differential equations. The previous terms comprise the *knowledge-based* part of the model. Note that modeling E allows one to capture possible time delays in the extracellular rates arising from the lumped transcription/translation dynamics of the ATPase. Furthermore, we consider in the dynamic equations model uncertainty due to, e.g., oversimplified or wrong model assumptions resulting from the lack of mechanistic description of the intracellular metabolism. The model *error* is defined by the vector-valued function $\boldsymbol{w} : \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_\tau} \to \mathbb{R}^{n_w}$ and it aims at capturing the dynamics neglected or misrepresented by the knowledge-based part of the dynamic equations. $\boldsymbol{\tau} \in \mathbb{R}^{n_\tau}$ comprises the parameters of $\boldsymbol{w}$. $\boldsymbol{Q_B} \in \mathbb{R}^{1\times n_w}$, $\boldsymbol{Q_E} \in \mathbb{R}^{1\times n_w}$, $\boldsymbol{Q_s} \in \mathbb{R}^{n_s\times n_w}$, $\boldsymbol{Q_p} \in \mathbb{R}^{n_p\times n_w}$ map the functions describing the model error to the corresponding differential equations. Hereafter, we will omit the time dependency of the variables when clear from the context.

Naturally, if the knowledge-based part of the dynamic equations describes the real system sufficiently well, i.e., without considering the model error $\boldsymbol{w}$, one can simply neglect $\boldsymbol{w}$ from the final model. However, in cases of significant model-plant mismatch, we propose to create *hybrid* models where we *learn* $\boldsymbol{w}$ from process data with machine learning, particularly using Gaussian processes.

2.2.2. Gaussian processes

Gaussian processes are classified as probabilistic machine-learning methods, whereby the predictions contain a measurement of the prediction uncertainty. Gaussian processes render a *probabilistic distribution*

over functions, as opposed to other approaches such as conventional neural networks where the predictions are deterministic. In this section, we present a general overview of Gaussian processes; we refer the reader to [30,43–45] for more information.

Let $l \in \mathbb{R}$ be the label (regression output) of one Gaussian process regressor and $\boldsymbol{v} \in \mathbb{R}^{n_v}$ the corresponding features (regression inputs). The Gaussian process regressor aims to model an unknown function $h : \mathbb{R}^{n_v} \to \mathbb{R}$ using noisy observations l of $h(\boldsymbol{v})$

$$l = h(\boldsymbol{v}) + \epsilon, \tag{2}$$

where ϵ is Gaussian distributed measurement noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ with zero mean and variance σ_n^2.

Let us define $\boldsymbol{V} \in \mathbb{R}^{n_v \times n_d}$ as the matrix of the supplied training inputs and $\boldsymbol{L} \in \mathbb{R}^{1 \times n_d}$ as the training outputs, where n_d is the number of training datasets. Furthermore, let $\boldsymbol{v}_i \in \mathbb{R}^{n_v \times 1}$ and $\boldsymbol{v}_j \in \mathbb{R}^{n_v \times 1}$ be two arbitrary input vectors.

In Gaussian processes, we assume that the labels are normally distributed

$$h(\boldsymbol{v}) \sim \mathcal{N}(m(\boldsymbol{v}), k(\boldsymbol{v}, \boldsymbol{v})), \tag{3}$$

where $m : \mathbb{R}^{n_v} \to \mathbb{R}$ is the mean function and $k : \mathbb{R}^{n_v} \times \mathbb{R}^{n_v} \to \mathbb{R}$ is the kernel or covariance *function*. Overall, the kernel function describes the neighborhood or similarity between data points in the feature space.

Gaussian processes start from a *prior distribution* of functions, characterized by a prior mean function and a prior covariance function, i.e., prior to observing data. The prior distribution is determined by the choice of the kernel function, which is chosen to be infinitely differentiable, smooth, and continuous. Many kernel functions are possible; in this work, we use the squared-exponential kernel function

$$k(\boldsymbol{v}_i, \boldsymbol{v}_j | \boldsymbol{\tau}) = \sigma^2 \exp\left(\frac{-(\boldsymbol{v}_i - \boldsymbol{v}_j)^\top (\boldsymbol{v}_i - \boldsymbol{v}_j)}{2d^2}\right), \tag{4}$$

where $\sigma^2 \in \mathbb{R}$ is the signal variance, $d \in \mathbb{R}$ is the length-scale, both of which are hyperparameters of the kernel function.

Furthermore, we obtain the covariance matrix $\boldsymbol{K} \in \mathbb{R}^{n_d \times n_d}$ based on the chosen kernel function and the supplied training data

$$\boldsymbol{K} = \begin{bmatrix} k(\boldsymbol{v}_1, \boldsymbol{v}_1) & \cdots & k(\boldsymbol{v}_1, \boldsymbol{v}_{n_d}) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{v}_{n_d}, \boldsymbol{v}_1) & \cdots & k(\boldsymbol{v}_{n_d}, \boldsymbol{v}_{n_d}) \end{bmatrix}, \tag{5}$$

which captures the relationship between the features.

We optimize the hyperparameters of the kernel function by maximizing the log marginal likelihood, i.e., $\boldsymbol{\tau}^* = \arg\max_{\boldsymbol{\tau}} \log p(\boldsymbol{L}|\boldsymbol{V}, \boldsymbol{\tau})$, with

$$\log(p(\boldsymbol{L}|\boldsymbol{V}, \boldsymbol{\tau})) = -\frac{1}{2}\boldsymbol{L}^T(\boldsymbol{K} + \sigma_n^2 I)^{-1}\boldsymbol{L} - \frac{1}{2}\log\left(|\boldsymbol{K} + \sigma_n^2 \boldsymbol{I}|\right) - \frac{n_d}{2}\log(2\pi), \tag{6}$$

where $\boldsymbol{\tau} := [\sigma^2, d, \sigma_n^2]$ and $\boldsymbol{I}$ is the identity matrix of appropriate size.

Finally, the conditional posterior of the Gaussian process with optimal hyperparameters follows a normal distribution for a test input vector $\boldsymbol{v}^* \in \mathbb{R}^{n_v \times 1}$, i.e., $p(\bar{h}(\boldsymbol{v}^*)|\boldsymbol{V}, \boldsymbol{L}) \sim \mathcal{N}(\bar{h}, \Sigma)$, with predictive mean $\bar{h}$ and variance Σ

$$\bar{h}(\boldsymbol{v}^*) = \widetilde{\boldsymbol{k}}^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \boldsymbol{L}, \tag{7a}$$

$$\Sigma(\boldsymbol{v}^*) = k(\boldsymbol{v}^*, \boldsymbol{v}^*) - \widetilde{\boldsymbol{k}}^\top (\boldsymbol{K} + \sigma_n^2 \boldsymbol{I})^{-1} \widetilde{\boldsymbol{k}}, \tag{7b}$$

$$\widetilde{\boldsymbol{k}} := [k(\boldsymbol{v}_1, \boldsymbol{v}^*), \ldots, k(\boldsymbol{v}_{n_d}, \boldsymbol{v}^*)]^\top. \tag{7c}$$

Since we aim to capture with Gaussian processes the model-plant mismatch $\boldsymbol{w}$ of differential equations, then $h(\boldsymbol{v}) := w_i(\boldsymbol{v})$, where w_i is the model error of a differential equation i. We consider as many Gaussian process regressors as n_w, i.e., multi-input single-output Gaussian processes. The features of the Gaussian process can be, e.g., appropriate model states and inputs.

2.3. Open-loop dynamic optimization of the optogenetically modulated ATPase expression

We propose to find the optimal light input trajectories that maximize the efficiency of the process with optogenetic control of the ATPase by solving the following optimal control problem

$$\max_{u(\cdot),\, \boldsymbol{x_0}} J(\cdot), \tag{8a}$$

$$\text{s.t. Eqs.(1a)} - \text{(1f)}, \tag{8b}$$

$$0 \leq \boldsymbol{g}(\boldsymbol{x}, u, \boldsymbol{\theta}, \boldsymbol{\tau}), \tag{8c}$$

where $J(\cdot)$ is the cost function that captures the efficiency of the process and $\boldsymbol{g} : \mathbb{R}^{n_x} \times \mathbb{R} \times \mathbb{R}^{n_\theta} \times \mathbb{R}^{n_\tau} \to \mathbb{R}^{n_g}$ are additional process constraints. The decision variables of the optimization problem can include in principle the dynamic input, in this case, green light photon flux density, and the initial state concentrations, e.g., initial substrate concentration. In other words, we cover both *static* and *dynamic* degrees of freedom in the formulation above. The cost function in (8a) can be the product volumetric productivity, titer, yield, or an explicit function that captures the economic profit of the process. The constraints in Eq. (8c) can include, e.g., safety, economic, or technical constraints.

2.4. Numerical methods

The parameter estimation procedure for the knowledge-based part of the model was performed in COPASI using the particle swarm algorithm [46]. We used HILO-MPC [43], a Python toolbox for machine-learning-supported optimal control, for training the Gaussian process regressors and solving the open-loop optimization problems.

3. Results and discussion

3.1. Model of the lactate fermentation with optogenetic control of the ATPase

Having described the general modeling and optimization framework in Sections 2.2 and 2.3, we proceed to describe the proposed model for the lactate fermentation case study in batch mode. We consider the following dynamic states: glucose $s_G \in \mathbb{R}$, lactate $p_L \in \mathbb{R}$, *E. coli*'s biomass $B_c \in \mathbb{R}$, and the intracellular ATPase $E \in \mathbb{R}$. Therefore, in the case study, $\boldsymbol{s} := s_G$, $\boldsymbol{p} := p_L$, and $B := B_c$. Compared to the dynamic constraint-based model proposed in [13,14], this represents a significant reduction in the number of dynamic states, i.e., from 23 to only 4 states. The control input is the green light photon flux density $u_l \in \mathbb{R}$, hence $\boldsymbol{u} := u_l$.

The proposed model follows

$$\frac{\mathrm{d}s_G}{\mathrm{d}t} = -q_G(s_G, E, \boldsymbol{\theta})B_c + w_G(s_G, B_c, p_L, E, u_l, \boldsymbol{\tau}), \tag{9a}$$

$$\frac{\mathrm{d}B_c}{\mathrm{d}t} = \mu(s_G, E, \boldsymbol{\theta})B_c + w_c(s_G, B_c, p_L, E, u_l, \boldsymbol{\tau}), \tag{9b}$$

$$\frac{\mathrm{d}p_L}{\mathrm{d}t} = q_L(s_G, E, \boldsymbol{\theta})B_c + w_L(s_G, B_c, p_L, E, u_l, \boldsymbol{\tau}), \tag{9c}$$

$$\frac{\mathrm{d}E}{\mathrm{d}t} = q_E(u_l, \boldsymbol{\theta}) - d_E(E, \boldsymbol{\theta}), \tag{9d}$$

$$s_G(t_0) = s_{G_0}, B_c(t_0) = B_{c_0}, p_L(t_0) = p_{L_0}, E(t_0) = E_0, \tag{9e}$$

where q_G, μ, q_L, q_E, and d_E are known kinetic functions with appropriate parameters, and w_G, w_c, and w_L are Gaussian-process regression functions with appropriate parameters describing the model error. In Eqs.

(9a)-(9d), we assume only two rate-limiting components in the kinetic functions, namely the substrate glucose and the light-inducible intracellular ATPase. We neglect dilution or degradation effects in equations (9a)-(9c).

Measuring the extracellular concentrations to parameterize the model is relatively straightforward, however, measuring the intracellular ATPase concentration remains a challenge. When available and affordable, proteomics [47] can be an option to measure the intracellular ATPase, but it is generally time-consuming and requires dedicated sample preparation. Note that, although proteomics is effective offline, implementing it for real-time monitoring remains a challenge. Alternatively, biosensors could, in principle, be used to estimate the intracellular ATPase, for example, by attaching fluorescent protein tags to the enzyme [48]. However, in some cases, this can lead to protein malfunction [49], and the intrinsic folding and maturation of fluorescent proteins [48] may cause delays in the output reading. One could argue that if the ATPase cannot be measured, it may still be estimated with soft sensors as shown in [14,26]. Yet, for soft sensors, we require a suitable validated mathematical model, which, in the context of this study, is unavailable. For these reasons, we did not implement here a biosensor or a soft sensor for ATPase determination, but it can be the focus of future studies, in particular within feedback control schemes where real-time monitoring is desirable.

Since in our experimental setup we cannot measure the intracellular ATPase, we regard E as a *virtual variable* expressed in virtual units (VU) per gram of biomass and we assume no model uncertainty in Eq. (9d). For simplicity, we assume that the dynamics of the intracellular ATPase are a function of only the ATPase concentration and the light input (cf. Eq. (9d)). Note that the inoculum/preculture preparation in the fermentations follows a well-standardized protocol (cf. Section 2.1.2), e.g., the preculture was always grown under red light conditions, i.e., without ATPase induction. Therefore, we arbitrarily set $E(t_0) = 0$ VU/g in all fermentation experiments. We neglect the exact biological meaning of Eq. (9d) and associated parameters as long as they help to describe well the dynamics of the extracellular concentrations. As mentioned before, Eq. (9d) allows coupling the input-dependent intracellular state of the cell (micro-scale variable) to the extracellular species (macro-scale variables). By doing so, one can capture time delays in the macro-scale fermentation dynamics arising from the lumped transcription and translation of the intracellular ATPase.

We model the kinetic rates as follows

$$q_G(s_G, E) = q_{G_{\max}}\left(\frac{s_G}{s_G + k_G}\right)\left(1 + \frac{E^{n_1}}{E^{n_1} + k_{\mathrm{GV}}^{n_1}}\right), \tag{10a}$$

$$\mu(s_G, E) = Y_{\mathrm{BG}}(q_G(s_G, E) - m_G)\left(1 - \frac{E^{n_2}}{E^{n_2} + k_{\mathrm{BV}}^{n_2}}\right), \tag{10b}$$

$$q_L(s_G, E) = (Y_{\mathrm{LB}}\mu(s_G, E) + m_L)\left(1 + \frac{E^{n_3}}{E^{n_3} + k_{\mathrm{LV}}^{n_3}}\right), \tag{10c}$$

$$q_E(u_l) = q_{E_0} + q_{E_{\max}}\frac{u_l^{n_4}}{u_l^{n_4} + k_u^{n_4}}, \tag{10d}$$

$$d_E(E) = k_d E, \tag{10e}$$

where $\boldsymbol{\theta} := [k_{\mathrm{BV}}, k_G, k_{\mathrm{GV}}, k_{\mathrm{LV}}, m_G, m_L, n_1, n_2, n_3, q_{G_{\max}}, Y_{\mathrm{BG}}, Y_{\mathrm{LB}}, q_{E_0}, q_{E_{\max}}, n_4, k_u, k_d]^\top$ comprises the 17 parameters of the assumed *known* kinetic functions.

If we assume no enforced ATP turnover, i.e., $E(t) = 0$, the specific substrate uptake rate (Eq. (10a)) follows conventional Monod-type kinetics [50], the specific growth rate (Eq. (10b)) follows the Pirt's equation for substrate distribution [51], and the specific lactate production rate (Eq. (10c)) follows the Leudeking-Piret's equation for catabolic products [52]. We include Hill-type activation terms [53] in Eqs. (10a)-(10c) to account for either an increase (+) or decrease (-) in the specific rates as a result of the enforced ATP turnover, i.e., for $E(t) > 0$. The production of the ATPase is activated by green light following the Hill function [53], Eq. (10d). Note that we assume homogeneous light penetration in the bioreactor. Finally, we assume an average lumped dilution/degradation rate of ATPase based on first-order kinetics, Eq. (10e).

3.2. Model validation

To validate the proposed model, we ran five fermentation experiments under different constant green light input values, namely 0, 175, 349, 524, and 873 μmol m^{-2} s^{-1}. Samples were taken equidistantly every hour. As can be seen in the blue bars in Fig. 3, the yield of lactate on glucose for the entire batch $Y_{\mathrm{LG,batch}}$ increases with increasing light input, up to around 1 g/g, i.e., the maximum theoretical yield. Note that there cannot be biomass generation if the lactate yield is at its maximum theoretical value. Thus, there must have been a slight overestimation of the actual lactate yield, in particular for the experiment with constant $u_l = 873$ μmol m^{-2} s^{-1}, very likely arising from a systematic measurement error of lactate concentration. Similarly, an increasing light input translates into decreasing biomass on glucose yield $Y_{\mathrm{BG,batch}}$ and lactate volumetric productivity $r_{\mathrm{L,batch}}$ averaged over the batch. This is the expected behavior since higher light photon flux density is linked to higher ATPase induction level, thus higher ATPase accumulation and ATP wasting [13,14].

Based on the batch experiments presented in Fig. 4, we carried out a parameter estimation procedure considering the data from all the batch experiments simultaneously. The resulting optimized *nominal* parameter values are $k_{\mathrm{BV}} = 2.605 \times 10^{-4}$ VU/g, $k_G = 5.340 \times 10^{-7}$ g/l, $k_{\mathrm{GV}} = 1.053 \times 10^{-6}$ VU/g, $k_{\mathrm{LV}} = 1.002 \times 10$ VU/g, $m_G = 1.232 \times 10^{-6}$ g/g/h, $m_L = 1.910$ g/g/h, $n_1 = 1.000 \times 10^{-2}$, $n_2 = 1.028 \times 10^{-1}$, $n_3 = 1.000 \times 10$, $q_{G_{\max}} = 1.731$ g/g/h, $Y_{\mathrm{BG}} = 1.083 \times 10^{-1}$ g/g, $Y_{\mathrm{LB}} = 2.204$ g/g, $q_{E_0} = 1.000 \times 10^{-6}$ VU/g/h, $q_{E_{\max}} = 1.000 \times 10$ VU/g/h, $n_4 = 4.718$, $k_u = 3.729 \times 10^2$ μmol m^{-2} s^{-1}, $k_d = 0.988$ 1/h. Note that the latter parameters were estimated considering only the knowledge-based part of the model, i.e., without learning the model-plant mismatch with Gaussian processes. We call this model the nominal model. Our goal with the parameter estimation was not to provide unique parameters rendering perfect fitting, but rather to approximate the behavior of the dynamic system such that, if necessary, the model fitting and predictability could be enhanced with Gaussian processes.

Overall, the nominal model fits well the experimental data for all the tested light inputs in Fig. 4. Despite the good fitting of the nominal model, we still implemented -as a proof of concept- Gaussian process regressors to learn the remaining model-plant mismatch. That is, we learned the model error w_i of equations (9a)-(9c). As input features for each of the Gaussian processes, we used the initial condition of the system at a sampling time t_k and the applied input over a time interval of size $t_{k+1} - t_k$. The output label of each of the Gaussian processes was the corresponding model-plant error w_i, whose function value was approximated as

$$w_i(t_k) = \frac{x_e(t_{k+1}) - x_e(t_k)}{t_{k+1} - t_k} - \frac{x_m(t_{k+1}) - x_m(t_k)}{t_{k+1} - t_k}, \tag{11}$$

where $x_e(t_k)$ and $x_e(t_{k+1})$ are the measured experimental state values at sampling times t_k and t_{k+1}, respectively. Similarly, $x_m(t_k)$ and $x_m(t_{k+1})$ are the predicted state values at times t_k and t_{k+1} for the knowledge-based part of the model. At each sampling time, $x_m(t_k) := x_e(t_k)$. The knowledge-based part of the model uses the parameters previously estimated for the nominal model; that is, the training of the model error via the Gaussian processes occurs after the parameter estimation. Each multi-input single-output Gaussian process was trained with 39 data entries, with each entry comprising values of the input features and output label.

Eq. (11) intrinsically assumes that the true dynamic equation

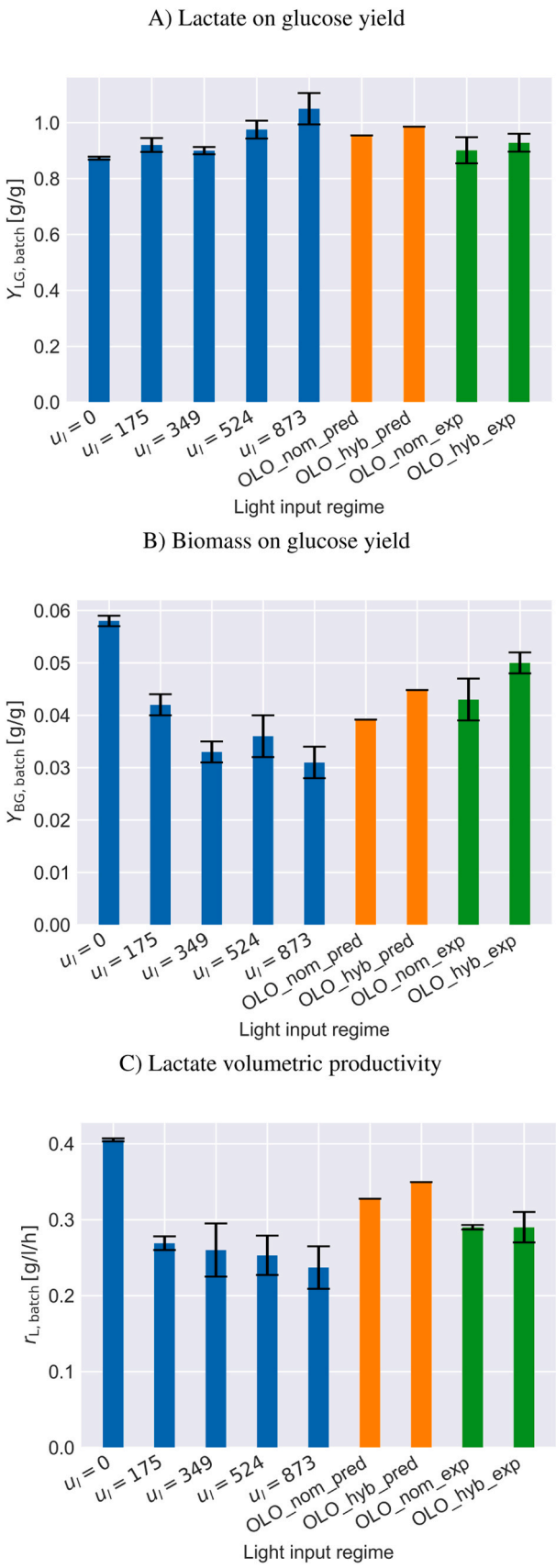

Fig. 3. A) Average lactate on glucose yield $Y_{LG,batch}$, B) biomass on glucose yield $Y_{BG,batch}$, and C) lactate volumetric productivity $r_{L,batch}$. Blue bars: experiments used to model the system. Orange bars: predicted open-loop optimization results using the nominal (OLO_nom_pred) and hybrid (OLO_hyb_pred) models. Green bars: actual experimental results for the open-loop optimizations using the nominal (OLO_nom_exp) and hybrid (OLO_hyb_exp) models. These metrics correspond to average batch results, i.e., considering initial and final points. The units of u_l are μmol m^{-2} s^{-1}.

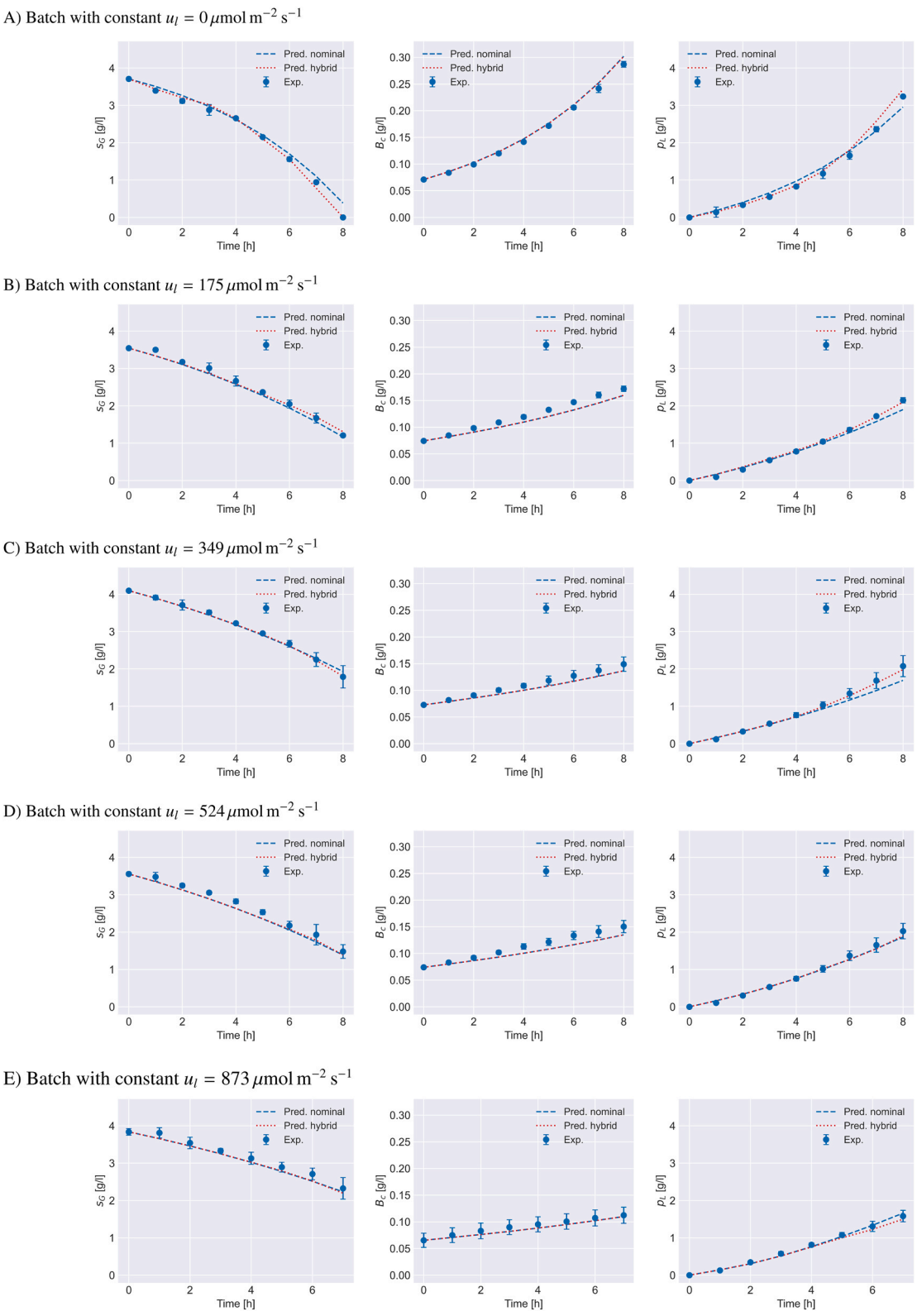

Fig. 4. Fitting of the nominal and hybrid models to the experimental data under different constant light inputs.

governing the experimental observations is the sum of the dynamic equation described by the assumed known model plus another differential equation describing the error or missing part. Furthermore, Eq. (11) considers that the right-hand side of the differential equations at a time t_k, over an interval $t_{k+1} - t_k$, can be approximated as the difference in the state values with respect to the length of the corresponding time interval. In other words, w_i can be regarded as the approximation of the right-hand side of the differential equation capturing the model error. Given the relatively short and equidistant sampling time in the experiments of 1 h, we deemed the outlined approach as a good approximation of the model error. Note that discrete versions of the model are in principle also possible. We refer to the model combining the *knowledge-*

based part with the Gaussian process regressors as the *hybrid* model. As expected, the hybrid model can fit slightly better the experimental data, particularly for the glucose and lactate profiles over the mid-end term of the fermentation experiments (cf. Fig. 4).

3.3. Implementation of open-loop optogenetic control of ATPase

We formulated optimal control problems, as described in Section 2.3, constrained by the nominal and hybrid models. The idea was to show the applicability of both models for optimal control of the optogenetically controlled ATPase expression. Specifically, we maximized the final batch lactate concentration in a time frame of eight hours, by applying piece-wise constant inputs of green light every hour (*dynamic* degree of freedom) and determining the optimal initial glucose concentration (*static* degree of freedom). Discretizing the input renders the optimization problem finite-dimensional and practical to solve since finding u_l as a function would otherwise make the problem infinite-dimensional. The input was constrained to the values used for fitting/training the model (cf. Eq. (12c)) and we demanded the optimizer to deplete all glucose by the end of the fermentation (cf. Eq. (12d)). Furthermore, we constrained the optimizer to achieve a user-defined product on glucose yield over the batch (cf. Eq. (12e)) $\widetilde{Y}_{\mathrm{LG,batch}}$. We also included a box constraint for the initial glucose concentration (cf. Eq. (12f)) with zero and $s_{G_{\max}} = 5$ g/l as the lower and upper bounds, respectively.

The resulting optimization problem reads

$$\max_{u_l(\cdot),\, s_G(t_0)} p_L(t_f), \tag{12a}$$

$$\text{s.t. Eqs.(9a)} - \text{(9d)}, \tag{12b}$$

$$0 \leq u_l \leq 873, \tag{12c}$$

$$s_G(t_f) = 0, \tag{12d}$$

$$\frac{p_L(t_f) - p_L(t_0)}{s_G(t_0) - s_G(t_f)} = \widetilde{Y}_{\mathrm{LG,batch}}, \tag{12e}$$

$$0 < s_G(t_0) \leq s_{G_{\max}}. \tag{12f}$$

The optimization problem in (12) thus maximizes the volumetric productivity in the fixed time frame under given product yield, light, initial substrate, and substrate consumption constraints. Remark that we neglect the model error $\boldsymbol{w}$ from the Eqs. (9a)-(9d) when we use the *nominal* model in the optimization. When we use instead the *hybrid* model, we then include the Gaussian-process-based model error $\boldsymbol{w}$ in Eqs. (9a)-(9d). For the optimization based on the nominal model, we arbitrarily set $\widetilde{Y}_{\mathrm{LG,batch}}$ to 0.954 g/g, while this was set to 0.986 g/g for the optimization based on the hybrid model. The goal was not to compare one-to-one the performance of the optimization using the two models but rather to show the flexibility of our approach to handle both knowledge-based and hybrid Gaussian-supported models. Furthermore, we also wanted to highlight the fact that modulating ATPase expression dynamically can be exploited to adjust the batch-to-batch fermentation performance in terms of product yield and productivity. Hence, the different selected values for $\widetilde{Y}_{\mathrm{LG,batch}}$.

Intuitively, as long as $\widetilde{Y}_{\mathrm{LG,batch}}$ is larger than the one achievable by the cell without ATPase induction (cf. Fig. 3, constant $u_l = 0$), the optimizer is expected to utilize the light-mediated ATP turnover mechanism to increase the product yield. Therefore, formulating an optimal control problem as done in (12) is a way to obtain trade-offs between enhancement of product yield and drop in volumetric productivity in the context of dynamic ATP turnover as discussed in [13,16]. The predicted open-loop optimization results based on the nominal and hybrid models are presented in Fig. 5. For the sake of rigor, the input trajectory generated by the optimizer was subsequently used to simulate the expected dynamics using the same system model employed by the optimizer. It is worth noting that there was no difference between the state dynamics of the model used by the optimizer upon convergence to a solution and those of the simulated plant given the optimized trajectory.

In both open-loop optimizations, the predicted input follows a two-stage profile, with a first phase at 0 μmol m^{-2} s^{-1} (no ATPase induction), followed by a second phase at 873 μmol m^{-2} s^{-1} (induction at the maximum green light photon flux density value). The main difference between the two optimization problems is the time at which the second phase is triggered, i.e., 3 h for the optimization based on the nominal model and 4 h for the optimization based on the hybrid model. As expected, in both cases the optimizer predicts a slight decrease in the biomass growth rate and a slight increase in the glucose uptake rate and lactate production rate for the second fermentation phase. The predicted optimal initial glucose concentration determined by the optimization based on the nominal model is 2.745 g/l, while this is 2.834 g/l for the optimization based on the hybrid model. As demanded, the optimizer predicts full consumption of glucose by the end of the fermentations.

The predicted yields of lactate on glucose over the batch for both fermentations are as demanded, i.e., 0.954 g/g for the optimization based on the nominal model and 0.986 g/g for the optimization based on the hybrid model (Fig. 3, orange bars). Compared to the scenario without ATPase induction (Fig. 3, $u_l = 0$), this represents a predicted increase in product yield of 9 % and 13 %, respectively. The predicted increase in batch product yield is at the expense of a decrease in the biomass on glucose yield by 33 % and 23 %, respectively. This also correlates to a decrease in volumetric productivity by 19 % and 14 %, respectively. As can be seen, the hybrid model predicts a less pronounced drop in biomass on glucose yield and volumetric productivity, even though the demanded increase in product yield in the optimization with the hybrid model is higher than in the optimization with the nominal model. In other words, the hybrid model seems to be more optimistic, which is in line with the fittings observed in Fig. 4, where the hybrid model predicts slightly higher lactate formation rates (in particular for treatments at 0, 175, 349 μmol m^{-2}s^{-1}).

We experimentally validated the predicted open-loop optimizations. As can be seen in Fig. 5, the experiments follow the overall trends as predicted by the optimal control problems. There is, nevertheless, some model-plant mismatch, in particular for the lactate concentration profile in the optimization based on the hybrid model. One should also notice that there is a slight mismatch between the target and the experimental initial concentrations, which is natural due to human error (cf. e.g. the glucose initial concentration in Fig. 5, optimization based on the hybrid model). The final lactate concentrations for both optimizations are within the same range, i.e., 2.324 ± 0.020 and 2.317 ± 0.018 g/l for the optimizations based on the nominal and hybrid models, respectively. Compared to the scenario without ATPase induction ($u_l = 0$ constant), the increase in lactate on glucose yield over the entire batch for the optimizations based on the nominal and hybrid models is, in average, 3 % and 6 %, respectively (cf. Fig. 3, green bars). In both optimizations, this is lower than the demanded $\widetilde{Y}_{\mathrm{LG,batch}}$ values; however, if we take into account the reported standard deviations, the experimental results are still close to the predicted values.

It is worth noting that the predicted input, following an off-on trajectory, differs from the more gradual trajectories found in previous simulation studies involving bilevel optimizations using dynamic constraint-based models (cf. [13,14]). This difference can be explained by the fact that the dynamic constraint-based models mentioned earlier are constrained by redox and energy balances, as well as resource allocation phenomena, which are neglected in our simplified modeling approach. Also, constraint-based models as in [13,14] assume a cell's dynamic (*evolutionary*) objective function, which, when formulated within a (bilevel) optimization scheme may bias the outcome of the optimization. Overall, we deem the predicted inputs in this work reasonable since they follow the widely discussed two-stage

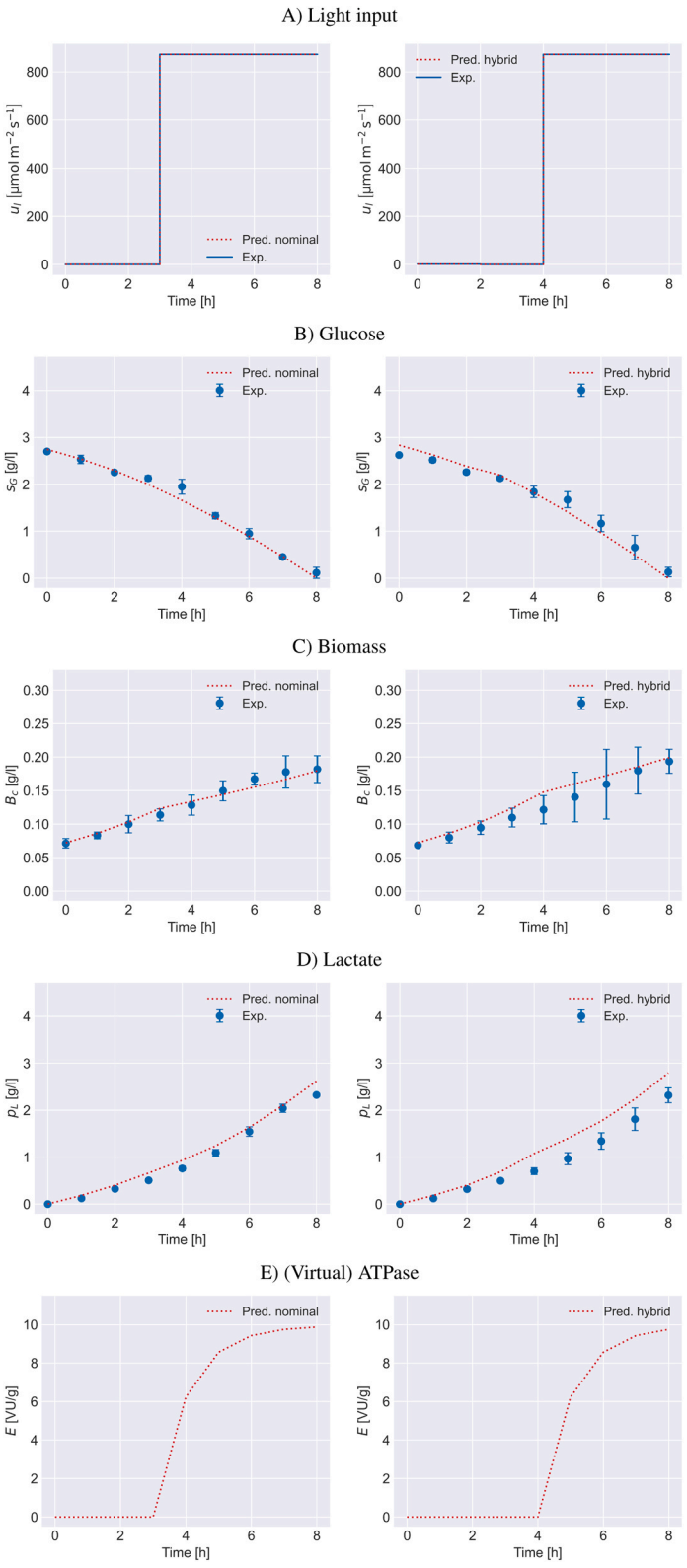

Fig. 5. Results of the open-loop optimization prediction and experimental implementation. The optimizations based on the nominal and hybrid models are shown on the left and right sides, respectively.

fermentation approaches for bioprocess optimization [10,12,23]. Even though the experimental comparison between different modeling and model-based optimization approaches was not within the scope of this study, it is still of interest for future research. For example, one could compare the use of a validated constraint-based dynamic model and a macro-kinetic-like model as the one outlined here to elucidate whether a higher modeling complexity also translates into enhanced optimization performance.

The fact that the optimizations were performed in an open-loop fashion, i.e., without taking correcting actions online, hinders the full

potential of model-based optimization. To address possible system uncertainty, as observed in the validation experiments, one could, e.g., implement feedback control schemes such as model predictive control. In model predictive control, the optimization is updated with the initial conditions of the plant and resolved at every sampling point (cf. e.g. [13,14,25,44,54–57]). The repetitive solution of the optimization problem leads to closed-loop control. Closed-loop control when considered in the frame of metabolic systems with external induction of gene expression is often referred to as *metabolic cybergenetics* [14,57,58]. As mentioned before, we did not have at hand real-time sensors for monitoring the fermentation dynamics, which hindered our possibility of implementing feedback control. Closing the control loop would be the natural next step once the monitoring aspect is addressed.

Other modeling techniques for metabolic systems with external induction of gene expression are also worth considering for future experimental implementations of model-based control. One interesting approach is the use of flux balance analysis to inform machine-learning surrogates to be embedded into the reaction rates of macro-kinetic models, effectively creating hybrid physics-informed models [57,59]. In this case, valuable information captured by metabolic networks, such as intrinsic metabolic trade-offs, redox and energy balances, etc., can be transferred into structurally simpler models. This strategy could help to reduce the gap between unstructured kinetic models and structured models based on flux balance analysis. The use of Gaussian processes to learn the model error, as described in this study, could still be employed to enhance the predictability of these hybrid physics-informed models.

In the hybrid model, we only considered the predicted mean of the Gaussian processes (Eq. (7a)). In addition, the input features of the Gaussian processes were intrinsically regarded as deterministic. As shown in Fig. 4, this was sufficient for fitting the error of the assumed known model given the experimental data. Having said that, future research endeavors will focus on exploiting the property of Gaussian processes for quantifying uncertainty (cf. e.g. Eq. (7b)). For instance, one could treat the input features of the Gaussian processes as normally distributed random variables [60] or estimate the state uncertainty via Monte Carlo sampling from the distributions governed by the Gaussian processes [33]. Being able to quantify the uncertainty of the hybrid model can unlock stochastic control that is robust in probability, capable of dealing with chance constraints [60–63,44]. As such, stochastic control approaches can be of significant relevance when dealing with uncertain bioprocesses. Furthermore, as suggested by [64], the predicted variance of Gaussian processes can be in principle incorporated into the objective function of optimization problems to balance exploration and exploitation properties. That is, one could actively explore areas of high model uncertainty while still exploiting the system. This could facilitate efficient model adaptation and enhance the performance of process optimization as a result of increasing model certainty.

4. Conclusion

We have proposed and experimentally validated a (Gaussian-process-supported) model-based optimization strategy for open-loop optogenetic control of the ATPase to maximize bioprocesses production efficiency through dynamic enforced ATP turnover. We have outlined a simplified modeling framework, i.e., an unsegregated and quasi-unstructured kinetic modeling approach, that captures relevant process dynamics. This facilitates model parameterization and simplifies model-based optimization compared to previously proposed dynamic constraint-based models. We have also considered hybrid models combining knowledge-based and Gaussian-process-supported components for modeling and optimal control.

For the experimental implementation, we have engineered *E. coli* to carry the CcaS/CcaR system to achieve optogenetic control of the ATPase. This engineered *E. coli* produces lactate as the main fermentation product under anaerobic conditions. Following optimal control problems constrained by knowledge-based and hybrid models, we have maximized lactate concentration while aiming at a target product yield and depleting all available glucose. However, there is still some model-plant mismatch which limits the full potential of the presented approach. Further work includes the experimental implementation of model predictive control schemes coupled with soft sensors in the context of metabolic cybergenetic systems. Overall, the presented model-based open-loop optimization strategy, validated with experiments, outlines an example of a simple and structured way to maximize production efficiency in optogenetically regulated metabolic processes.

Data accessibility

The data that support the findings of this study are available from the corresponding author upon reasonable request.

CRediT authorship contribution statement

Sebastian Espinel-Rios: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Gerrich Behrendt:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation. **Jasmin Bauer:** Investigation, Validation, Writing – original draft, Writing – review & editing. **Bruno Morabito:** Methodology, Writing – original draft, Writing – review & editing. **Johannes Pohlodek:** Methodology, Writing – original draft, Writing – review & editing. **Andrea Schütze:** Investigation, Validation, Writing – review & editing. **Rolf Findeisen:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Katja Bettenbrock:** Writing – review & editing, Writing – original draft, Conceptualization, Funding acquisition, Resources, Supervision. **Steffen Klamt:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the International Max Planck Research School for Advanced Methods in Process and Systems Engineering (IMPRS ProEng) and by the state of Saxony-Anhalt within the SmartProSys initiative. We would like to thank Reiner Könning for his support with the instrumentation of the light-delivery system.

References

[1] D. Yang, J.S. Cho, K.R. Choi, H.U. Kim, S.Y. Lee, Systems metabolic engineering as an enabling technology in accomplishing sustainable development goals, Micro Biotechnol. 10 (5) (2017) 1254–1258.
[2] J.S. Cho, G.B. Kim, H. Eun, C.W. Moon, S.Y. Lee, Designing microbial cell factories for the production of chemicals, JACS Au 2 (8) (2022) 1781–1799.
[3] J.M. Woodley, Towards the sustainable production of bulk-chemicals using biotechnology, N. Biotechnol. 59 (2020) 59–64.
[4] J.A. Lee, H.U. Kim, J.-G. Na, Y.-S. Ko, J.S. Cho, S.Y. Lee, Factors affecting the competitiveness of bacterial fermentation, Trends Biotechnol. (2022). S0167779922002761.
[5] D. Banerjee, A. Mukhopadhyay, Perspectives in growth production trade-off in microbial bioproduction, RSC Sustain. (2023), 10.1039.D2SU00066K.
[6] W.J. Holtz, J.D. Keasling, Engineering static and dynamic control of synthetic pathways, Cell 140 (1) (2010) 19–23.
[7] F. Zhang, J.M. Carothers, J.D. Keasling, Design of a dynamic sensor-regulator system for production of chemicals and fuels derived from fatty acids, Nat. Biotechnol. 30 (4) (2012) 354–359.

[8] I.M. Brockman, K.L.J. Prather, Dynamic metabolic engineering: new strategies for developing responsive cell factories, J. Biotechnol. 10 (9) (2015) 1360–1369.
[9] N. Venayak, N. Anesiadis, W.R. Cluett, R. Mahadevan, Engineering metabolism through dynamic control, Curr. Opin. Biotechnol. 34 (2015) 142–152.
[10] M.A. Lalwani, E.M. Zhao, J.L. Avalos, Current and future modalities of dynamic control in metabolic engineering, Curr. Opin. Biotechnol. 52 (2018) 56–65.
[11] C.J. Hartline, A.C. Schmitz, Y. Han, F. Zhang, Dynamic control in metabolic engineering: theories, tools, and applications, Metab. Eng. 63 (2021) 126–140.
[12] J.M. Burg, C.B. Cooper, Z. Ye, B.R. Reed, E.A. Moreb, M.D. Lynch, Large-scale bioprocess competitiveness: the potential of dynamic metabolic control in two-stage fermentations, Curr. Opin. Chem. Eng. 14 (2016) 121–136.
[13] S. Espinel-Ríos, B. Morabito, J. Pohlodek, K. Bettenbrock, S. Klamt, R. Findeisen, Optimal control and dynamic modulation of the ATPase gene expression for enforced ATP wasting in batch fermentations, IFAC Pap. 55 (7) (2022) 174–180.
[14] S. Espinel-Ríos, B. Morabito, J. Pohlodek, K. Bettenbrock, S. Klamt, R. Findeisen, Toward a modeling, optimization, and predictive control framework for fed-batch metabolic cybergenetics, Biotechnol. Bioeng. 121 (1) (2024) 366–379.
[15] O. Hädicke, K. Bettenbrock, S. Klamt, Enforced ATP futile cycling increases specific productivity and yield of anaerobic lactate production in *Escherichia coli*: ATP wasting to improve yield and productivity, Biotechnol. Bioeng. 112 (10) (2015) 2195–2199.
[16] S. Espinel-Ríos, K. Bettenbrock, S. Klamt, R. Findeisen, Maximizing batch fermentation efficiency by constrained model-based optimization and predictive control of adenosine triphosphate turnover, AIChE J. 68 (4) (2022) e17555.
[17] J. Wichmann, G. Behrendt, S. Boecker, S. Klamt, Characterizing and utilizing oxygen-dependent promoters for efficient dynamic metabolic engineering, Metab. Eng. 77 (2023) 199–207.
[18] S. Boecker, A. Zahoor, T. Schramm, H. Link, S. Klamt, Broadening the scope of enforced ATP wasting as a tool for metabolic engineering in *Escherichia coli*, Biotechnol. J. 14 (9) (2019) 1800438.
[19] S. Boecker, B.-J. Harder, R. Kutscha, S. Pflügl, S. Klamt, Increasing, ATP turnover boosts productivity of 2,3-butanediol synthesis in *Escherichia coli*, Micro Cell Factor. 20 (1) (2021) 63.
[20] A. Zahoor, K. Messerschmidt, S. Boecker, S. Klamt, ATPase-based implementation of enforced ATP wasting in *Saccharomyces cerevisiae* for improved ethanol production, Biotechnol. Biofuels 13 (1) (2020) 185.
[21] S. Pouzet, A. Banderas, M. LeBec, T. Lautier, G. Truan, P. Hersen, The promise of optogenetics for bioproduction: dynamic control strategies and scale-up instruments, Bioengineering 7 (4) (2020) 151.
[22] A. Baumschlager, M. Khammash, Synthetic biological approaches for optogenetics and tools for transcriptional light-control in bacteria, Adv. Biol. 5 (5) (2021) 2000256.
[23] S. Klamt, R. Mahadevan, O. Hädicke, When do two-stage processes outperform one-stage processes? Biotechnol. J. 13 (2) (2018) 1700539.
[24] S. Waldherr, D.A. Oyarzún, A. Bockmayr, Dynamic optimization of metabolic networks coupled with gene expression, J. Theor. Biol. 365 (2015) 469–485.
[25] B. Jabarivelisdeh, L. Carius, R. Findeisen, S. Waldherr, Adaptive predictive control of bioprocesses with constraint-based modeling and estimation, Comput. Chem. Eng. 135 (2020) 106744.
[26] S. Espinel-Ríos, B. Morabito, K. Bettenbrock, S. Klamt, R. Findeisen, Soft sensor for monitoring dynamic changes in cell composition, IFAC Pap. 55 (23) (2022) 98–103.
[27] S. Dempe, S. Franke, Solution of bilevel optimization problems using the KKT approach, Optimization 68 (8) (2019) 1471–1489.
[28] S. Dempe, Bilevel optimization: theory, algorithms, applications and a bibliography, In: S. Dempe, A. Zemkoho, (Eds.), Bilevel Optimization, Vol. 161, Springer International Publishing, Cham, 2020, 581-672, series Title: Springer Optimization and Its Applications.
[29] H.B. Carœ, Bilevel optimization with application in energy (PhD thesis), University of Copenhagen, Faculty of Science, Department of Mathematical Sciences, Copenhagen (2018).
[30] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, Adaptive Computation And Machine Learning, oCLC, MIT Press, Cambridge, Mass, 2006.
[31] M.N. Cruz-Bournazou, H. Narayanan, A. Fagnani, A. Butté, Hybrid Gaussian process models for continuous time series in bolus fed-batch cultures, IFAC Pap. 55 (7) (2022) 204–209.
[32] Y. Sun, W. Nathan-Roberts, T.D. Pham, E. Otte, U. Aickelin, Multi-Fidelity Gaussian Process for Biomanufacturing Process Modeling with Small Data, arXiv: 2211.14493 (2022).
[33] A.W. Rogers, Z. Song, F.V. Ramon, K. Jing, D. Zhang, Investigating 'greyness' of hybrid model for bioprocess predictive modelling, Biochem Eng. J. 190 (2023) 108761.
[34] A. Milias-Argeitis, M. Rullan, S.K. Aoki, P. Buchmann, M. Khammash, Automated optogenetic feedback control for precise and robust regulation of gene expression and cell growth, Nat. Commun. 7 (1) (2016) 12546.
[35] E.J. Olson, L.A. Hartsough, B.P. Landry, R. Shroff, J.J. Tabor, Characterizing bacterial gene circuit dynamics with optically programmed gene expression signals, Nat. Methods 11 (4) (2014) 449–455.
[36] S. Senoo, S.T. Tandar, S. Kitamura, Y. Toya, H. Shimizu, Light-inducible flux control of triosephosphate isomerase on glycolysis in *Escherichia coli*, Biotechnol. Bioeng. 116 (12) (2019) 3292–3300.
[37] S.R. Schmidl, R.U. Sheth, A. Wu, J.J. Tabor, Refactoring and optimization of light-switchable *Escherichia coli* two-component systems, ACS Synth. Biol. 3 (11) (2014) 820–831.
[38] K.H. Choi, T. Mima, Y. Casart, D. Rholl, A. Kumar, I.R. Beacham, H.P. Schweizer, Genetic tools for select-agent-compliant manipulation of *Burkholderia pseudomallei*, Appl. Environ. Microbiol 74 (4) (2008) 1064–1075.
[39] K.-H. Choi, J.B. Gaynor, K.G. White, C. Lopez, C.M. Bosio, R.R. Karkhoff-Schweizer, H.P. Schweizer, A. Tn7-based, broad-range bacterial cloning and expression system, Nat. Methods 2 (6) (2005) 443–448.
[40] P.P. Cherepanov, W. Wackernagel, Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant, Gene 158 (1) (1995) 9–14.
[41] J.J. Tabor, A. Levskaya, C.A. Voigt, Multichromatic control of gene expression in *Escherichia coli*, J. Mol. Biol. 405 (2) (2011) 315–324.
[42] S. Tanaka, S.A. Lerner, E.C. Lin, Replacement of a phosphoenolpyruvate-dependent phosphotransferase by a nicotinamide adenine dinucleotide-linked dehydrogenase for the utilization of mannitol, J. Bacteriol. 93 (2) (1967) 642–648.
[43] J. Pohlodek, B. Morabito, C. Schlauch, P. Zometa, R. Findeisen, Flexible development and evaluation of machine-learning-supported optimal control and estimation methods via HILO-MPC, Int. J. Robust Nonlinear Control (2024), https://doi.org/10.1002/rnc.7275 in press.
[44] B. Morabito, J. Pohlodek, L. Kranert, S. Espinel-Ríos, R. Findeisen, Efficient and simple Gaussian process supported stochastic model predictive control for bioreactors using HILO-MPC, IFAC Pap. 55 (7) (2022) 922–927.
[45] A. Himmel, J. Matschek, R. Kok, B. Morabito, H.H. Nguyen,and R. Findeisen. Machine learning for control of (bio)chemical manufacturing systems. In: Artificial Intelligence in Manufacturing,181-240, (2024), Elsevier. https://doi.org/10.1016/B978-0-323-99134-6.00009-8.
[46] S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, U. Kummer, COPASI—a COmplex PAthway simulator, Bioinformatics 22 (24) (2006) 3067–3074.
[47] S.R. Shuken, An introduction to mass spectrometry-based proteomics, J. Proteome Res. 22 (7) (2023) 2151–2171.
[48] K. Thorn, Genetically encoded fluorescent tags, Mol. Biol. Cell 28 (7) (2017) 848–857.
[49] U. Weill, G. Krieger, Z. Avihou, R. Milo, M. Schuldiner, D. Davidi, Assessment of GFP tag position on protein localization and growth fitness in yeast, J. Mol. Biol. 431 (3) (2019) 636–641.
[50] J.J. Heijnen, B. Romein, Derivation of kinetic equations for growth on single substrates based on general properties of a simple metabolic network, Biotechnol. Prog. 11 (6) (1995) 712–716.
[51] S.J. Pirt, The maintenance energy of bacteria in growing cultures, Proc. R. Soc. B 163 (991) (1965) 224–231.
[52] R. Luedeking, E.L. Piret, A kinetic study of the lactic acid fermentation. Batch process at controlled pH, J. Biochem. Technol. Eng. 1 (4) (1959) 393–412.
[53] A.V. Hill, The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves, J. Physiol. 40 (1910) iv–vii.
[54] B. Jabarivelisdeh, S. Waldherr, Optimization of bioprocess productivity based on metabolic-genetic network models with bilevel dynamic programming, Biotechnol. Bioeng. 115 (7) (2018) 1829–1841.
[55] B. Morabito, A. Kienle, R. Findeisen, L. Carius, Multi-mode model predictive control and estimation for uncertain biotechnological processes Towards risk-aware machine learning supported model predictive control and open-loop optimization for repetitive processes, IFAC Pap. 52 (1) (2019) 709–714.
[56] B. Morabito, J. Pohlodek, J. Matschek, A. Savchenko, L. Carius, R. Findeisen, Towards risk-aware machine learning supported model predictive control and open-loop optimization for repetitive processes, IFAC Pap. 54 (6) (2021) 321–328.
[57] S. Espinel-Ríos, J.L. Avalos, Hybrid Physics-informed Metabolic Cybergenetics: Process Rates Augmented with Machine-learning Surrogates Informed by Flux Balance analysis, arXiv:2401.00670 (2024).
[58] C. Carrasco-López, S.A. García-Echauri, T. Kichuk, J.L. Avalos, Optogenetics and biosensors set the stage for metabolic cybergenetics, Curr. Opin. Biotechnol. 65 (2020) 296–309.
[59] S. Espinel-Ríos, J.L. Avalos, Linking Intra- and Extra-cellular Metabolic Domains Via Neural-network Surrogates for Dynamic Metabolic Control, arXiv:2310.17179 (2023).
[60] E. Bradford, A.M. Schweidtmann, D. Zhang, K. Jing, E.A. del Rio-Chanona, Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes, Comput. Chem. Eng. 118 (2018) 143–158.
[61] L. Hewing, K.P. Wabersich, M. Menner, M.N. Zeilinger, Learning-based model predictive control: toward safe learning in control, Annu Rev. Control Robot Auton. Syst. 3 (1) (2020) 269–296.
[62] E. Bradford, L. Imsland, D. Zhang, E.A. Del RioChanona, Stochastic data-driven model predictive control using Gaussian processes, Comput. Chem. Eng. 139 (2020) 106844.
[63] L. Hewing, J. Kabzan, M.N. Zeilinger, Cautious model predictive control using Gaussian process regression, IEEE Trans. Control Syst. Technol. 28 (6) (2020) 2736–2743.
[64] S. Espinel-Ríos, R. Kok, S. Klamt, J.L. Avalos, R. Findeisen, Batch-to-batch optimization with model adaptation leveraging Gaussian processes: the case of optogenetically assisted microbial consortia, In: 23rd International Conference on Control, Automation and Systems (ICCAS), IEEE, 2023, 1292-1297.