Quantification of proteomic and metabolic burdens predicts growth retardation and overflow metabolism in recombinant *Escherichia coli*

Hong Zeng[1], Aidong Yang[1]*

[1]Department of Engineering Science, University of Oxford, Parks Road, Oxford, OX1 3PJ, UK

***Corresponding Author** Aidong Yang; email: aidong.yang@eng.ox.ac.uk; telephone: +44 1865 2 73094.

Abstract

Escherichia coli has been the host organism most frequently investigated for efficient recombinant protein production. However, the production of a foreign protein in recombinant *E. coli* often leads to growth deterioration and elevated secretion of acetic acid. Such observed phenomena have been widely linked with cell stress responses and metabolic burdens originated particularly from the increased energy demand. In this work, flux balance analysis (FBA) and dynamic flux balance analysis (DFBA) were applied to investigate the observed growth physiology of recombinant *E. coli*, incorporating the proteome allocation theory and an adjustable maintenance energy level (ATPM) to capture the proteomic and energetic burdens introduced by recombinant protein synthesis. Model predictions of biomass growth, substrate consumption, acetate excretion and protein production with two different strains were in good agreement with the experimental data, indicating that the constraint on the available proteomic resource and the change in ATPM might be important contributors governing the growth physiology of recombinant strains. The modelling framework developed in this work, currently with several limitations to overcome, offers a starting point for the development of a practical, model-based tool to guide metabolic engineering decisions for boosting recombinant protein production.

KEYWORDS

Escherichia coli, recombinant protein production, overflow metabolism, genome-scale model, proteome allocation constraint, dynamic flux balance analysis

1 | INTRODUCTION

Expression of recombinant proteins in microbes or higher organisms has been considered as an efficient approach to protein production [1–3]. Among various host cells, *Escherichia coli* is the most widely used organism due to its well-characterized genetics and physiology, ability to achieve high cell density, fast growth and high-yield nature [4–7]. However, the production of a foreign protein often triggers local (reaction-specific) and global (system-level) cell stress responses that lead to growth retardation, production of undesirable/toxic by-products such as acetic acid (known as overflow metabolism) and eventually low productivity of targeted proteins [8–10]. Such stress responses are considered to originate from the increased cellular metabolic burden associated with plasmid maintenance [11], protein folding, secretion and degradation of misfolded protein [12–14]. In particular, metabolic burdens often manifest as the increase in energy demand or maintenance energy requirement [15–17]. Furthermore, a recent study shows that the formation of acetate (as an example of the overflow metabolism) in *E. coli* is a result of the coordination of energy demand with carbon influx given constrained proteomic resources; the overproduction of "useless" proteins amplifies such phenomenon by reducing the proteome fraction available for energy biogenesis and biomass synthesis [18].

In this work, we refer to the proteome fraction occupied by the expression of recombinant protein as proteomic burden, and attempt to use a constraint-based model (CBM) [19] to investigate quantitatively the role of proteomic and metabolic burdens in the observed growth deterioration and acetate formation associated with recombinant protein production. Several studies using CBMs have previously attempted to predict the impact of gene mutations (gene knockout or overexpression) on the improvement of recombinant protein production for *Pichia pastoris* [20–22], *Synechococcus elongatus* [23] and *Lactococcus lactis* [24]. The manipulations with *in silico* determined targeted genes were tested experimentally to verify the effectiveness of the model prediction. In addition to bacteria and algae, CBMs has been applied to recombinant eukaryotic cell hosts for protein production, where genome-scale metabolic models (GEMs) were used to identify targets for genetic modification, improve cellular metabolic capabilities, design media supplementation, and interpret high-throughput omics data [25]. Another relevant work [26] used a CBM for recombinant *E. coli* which fixed the growth rate, substrate uptake rate

and acetate production rate to the experimentally measured values and maximized recombinant protein production, to study the possible changes and preference of the central metabolic pathways. In all these past studies, the way how the production of recombinant proteins is incorporated in a CBM is typically by the insertion of the protein synthesis reaction or a set of heterologous pathways associated with the protein formation, substrate transport and energy generation, to account for the mass and energy consumption.

To our knowledge, there is currently a lack of work in the quantitative prediction of the observed reduced growth and increased acetate formation in recombinant protein production via CBMs. In this study, we use flux balance analysis (FBA) and dynamic flux balance analysis (DFBA) to investigate the important contributors in predicting such growth physiology of two recombinant *E. coli* strains, namely green fluorescence protein (GFP) recombinant *E. coli* W3110 and maltose binding protein-glucose isomerase (GI-malE) fusion recombinant *E. coli* XL1. Metabolic burdens are quantified by (i) the local effect due to the extra mass and energy requirements in the synthesis of foreign proteins and (ii) the global effect on the increased maintenance energy caused by plasmid maintenance, protein secretion, folding, degradation and other possible disturbance associated with the recombinant proteins. Furthermore, the recently proposed proteome allocation theory (PAT) [18] is integrated into the stoichiometry model to investigate the role of proteomic burden. This approach shares the same principle of cellular resource allocation [27] with several existing constraint-based models including FBAwMC [28,29], RBA [30–33], ME-Model [34] and its extension [35]. The main target of these existing models is to quantitatively predict the maximum cellular growth rate, while the overflow metabolism has mostly been captured qualitatively. The more recently presented model named CAFBA [36] is able to accurately predict both cell growth and acetate production. However, this and some of the earlier models (e.g. FBAwMC) formulate their resource allocation constraint by referring to the costs of individual metabolic reactions, which in principle would mean that a large number of parameters are to be introduced. Building on the PAT proposed by Basan *et al.* which depicts a constraint of proteomic allocation between two energy pathways and the biomass synthesis sector (as opposed to individual reactions), the current work is aimed to adopt a modelling approach that involves a small number of

parameters which potentially can be estimated by fitting the model to experimental data. Overall, we aim to predict the observed growth retardation and elevated acetate production for recombinant *E. coli* using such a practical FBA/DFBA framework, which in the future may be developed as a predictive tool to guide metabolic engineering decisions.

2 | METHODS

2.1 | Representing proteomic and metabolic burdens

2.1.1 | The proteomic constraint

Based on the recently developed theory which suggests that the overflow metabolism in *Escherichia coli* originates from the discrepancy in the proteomic efficiency between different energy biogenesis pathways, namely (oxidative) fermentation and respiration [18] and a subsequent modelling framework termed Constrained Allocation Flux Balance Analysis [36], we introduce the following equation to represent the constraint on the utilization of proteomic resources:

$$w_f v_f + w_r v_r + b\lambda \leq \phi_{max} \quad (1)$$

where λ is the growth rate; v_f (v_r) is a flux representing the overall activity of the fermentation (respiration) pathway. Specifically, we choose the enzymatic reaction enolase (ENO) located in the lower part of the glycolysis to be v_f, while specifying the entrance flux of the tricarboxylic acid (TCA) cycle citrate synthase (CS) as v_r. ENO and CS have been selected in this work since they are able to correctly capture the increasing trend of fermentation and the decreasing trend of respiration in the overflow region. It is worth noting that, rigorously speaking, these two fluxes do not correspond to "pure" fermentation or respiration fluxes, although in the case of ENO it contains predominantly the fermentation flux in the overflow region. Conceptually, a flux in the TCA cycle after the drawing of metabolites for biomass synthesis could be a better indicator than CS on the respiration side. In fact, replacing CS with AKGDH (2-oxogluterate dehydrogenase), which is one of such fluxes, was tested in this work, which showed results comparable to those from using CS (see Supplementary Material, Table S2 and Figures S4-S6 for details). In the rest of this paper, we present results obtained with the ENO-CS combination. ϕ_{max} is the maximum fraction of proteome attainable to energy biogenesis sectors

(fermentation and respiration) and biomass synthesis, previously determined to be approximately 0.484 [37]. Compared to the protein-constrained GEMs that consider enzymatic activities at the level of individual intracellular reactions [38], Equation 1 is much less detailed and offers a coarse-grain constraint at the pathway level. On the other hand, it directly reflects the difference in proteomic efficiencies of the two energy pathways. Besides, it contains a small number of parameters which, as shown below, can easily be estimated from cell culture data.

For simplicity, ϕ_{max} was divided on both sides of Equation 1 to give the normalized form of the proteomic constraint.

$$w_f^* v_f + w_r^* v_r + b^* \lambda \leq 1 \tag{2}$$

where $w_f^* \equiv w_f/\phi_{max}$, $w_r^* \equiv w_r/\phi_{max}$ and $b^* \equiv b/\phi_{max}$. Equation 2 does not contain the influence of recombinant protein production and is referred to as the wild-type proteomic constraint in this study. w_f^*, w_r^* and b^* are referred to as the wild-type proteomic parameters.

2.1.2 | Introducing the recombinant protein sector into the proteomic constraint

The overproduction of "useless" proteins has been shown to amplify the rate of acetate formation by reducing the fraction of proteome available for energy production and biomass synthesis [18]. Here, we consider the production of a foreign protein as an additional sector that consumes the cellular proteomic resources, and add an extra term to the wild-type proteomic constraint:

$$w_f^* v_f + w_r^* v_r + b^* \lambda + \phi_{recP}^* \leq 1 \tag{3}$$

where $\phi_{recP}^* \equiv \frac{\phi_{recP}}{\phi_{max}}$, ϕ_{recP} represents the fraction of proteome occupied for the synthesis and maintenance of the recombinant protein, considered as a (constant) property of the specific recombinant strain. It is evident that the presence of ϕ_{recP}^* means that the fraction of proteome accessible by biomass growth and energy pathways is reduced to $1 - \phi_{recP}^*$. Equation 3 is referred to as the generic proteomic constraint, with ϕ_{recP}^* being zero for the wild type and non-zero for recombinant strains.

2.1.3 | Maintenance ATP as an indicator of the metabolic burden

In GEMs, maintenance ATP (ATPM) is generally defined as an energy 'drain' flux that accounts for the energy requirement resulted from non-growth-associated cellular activities [39,40]. In a GEM, it is a defined flux in addition to the growth-associated energy cost embedded in the biomass synthesis reaction, with lower/upper bounds. The level of ATPM for wild-type *E. coli* ranges from 3.15-8.39 mmol ATP gDW^{-1} h^{-1} subjected to the different versions of reconstruction and different strains (see BiGG database, http://bigg.ucsd.edu/).

It is commonly accepted that the production of a foreign protein introduces an extra metabolic burden to the host cell [16,41]. Energetically, this metabolic burden manifests as the increase in cellular energy demand [15,42] or elevated maintenance energy [12,26]. An early study suggested that the synthesis of the recombinant protein alone would not be able to account fully for the extra demand of ATP [43]. Recent work validated this hypothesis by showing that the higher maintenance energy mainly results from the refolding and/or secretion stress triggered by the recombinant protein [12,13] or is due to the degradation of misfolded protein [14].

The aforementioned studies warrant the investigation on the impact of ATPM on the model prediction of the growth physiology of the recombinant *E. coli*. In a GEM, this entails treating ATPM as an adjustable parameter (as opposed to fixing it to the default value), in addition to the inclusion of the synthesis reaction of the recombinant protein into the GEM which accounts for the energy requirement of protein synthesis.

2.2 | Constraint-based metabolic modelling

2.2.1 | The genome-scale model

Genome-scale metabolic model *i*AF1260 [40] was used for all simulations in this study. In the adopted model, the hydrogen production reaction FHL and the oxidative stress response reactions CAT, SPODM and SPODMpp were closed to prevent infeasible state under aerobic-glucose conditions [40]. The glucose dehydrogenase reaction (GLCDpp) was switched off since it is only active if pyrroloquinoline quinone (PQQ) is supplied to the system (see EcoCyc [44] entry on this enzyme). The

D-glucose (galactose):proton symport reaction (GLCt2pp) was also considered to be off as it is functional when 2-deoxy-D-galactose is the specific substrate (see MetaCyc [45] entry on this enzyme). Reactions describing the synthesis of two recombinant proteins, GFP and GI-malE fusion protein were added to the original stoichiometry model. The protein reactions comprise the amino acid building blocks as well as the energy required for protein synthesis. Further details are provided in the Supplementary Material Section 1, including Equations S1-S2; and Table S1.

2.2.2 | Flux balance analysis

Flux balance analysis (FBA) [46] was used to determine the optimal flux distribution under different growth conditions, with a set of constraints:

$$\max f_{obj}, \text{ subject to}$$

$$\text{(i) } S\mathbf{v} = 0 \tag{9}$$

$$\text{(ii) } v^L \leq \mathbf{v} \leq v^U \tag{5}$$

$$\text{(iii) } w_f^* v_f + w_r^* v_r + b^* \lambda + \phi_{recP}^* \leq 1 \tag{11}$$

where f_{obj} is the assumed cellular objective. S is the stoichiometric matrix; **v** is a column vector comprising the reactions/fluxes included in the (modified) *i*AF1260 model; v^L and v^U represent the minimum and maximum attainable fluxes, respectively; λ denotes the growth rate; w_f^* , w_r^*, b^* and ϕ_{recP}^* are the (normalized) proteomic parameters as introduced earlier in Eqs. (1)-(4); v_f is a single flux (in this case, ENO) chosen to represent the overall activity of the fermentation pathway and v_r is the representative flux of the respiration pathway (in this case, CS). Together, v_f and v_r indicate the cellular modulation of energy flux (fermentation vs. respiration).

The prediction of the extent of overflow metabolism (rates of acetate production) was determined by applying a proper set of values for w_f^* , w_r^*, b^* and ϕ_{recP}^* to the third constraint of FBA (generic proteomic constraint). Implementation details for introducing the proteomic constraint into the FBA model are given in the Supplementary Material, Equations S3-S4. The proteomic-constraint-based FBA was run under aerobic-glucose conditions, with the objective function set to either maximize the growth

rate or minimize glucose consumption, depending on the nature of the input data available, as explained below. The optimal flux distribution was solved via COBRA toolbox [47] with Gurobi 6.0 as the linear programming (LP) solver.

2.2.3 | The dynamic framework

To facilitate the estimation of key model parameters using the literature data available in the form of records of batch cultures, dynamic flux balance analysis (DFBA) [48] was adopted to connect the dynamic batch growth behavior with the intracellular flux distribution. The DFBA framework comprises two parts: the microscale cellular metabolic flux distribution (determined by FBA) and the change in the macroscale extracellular environment modelled by ordinary differential equations (ODEs). Dynamic simulation is achieved by the exchange of information between the two parts at a number of time intervals that divide the total simulated time duration: the FBA model is solved at the beginning of every time interval to give steady state flux distribution; the values of the fluxes of interest are then passed to the ODEs to update the outer environment state. This procedure is repeated until the entire time duration has been simulated.

Maximizing biomass growth was taken as the objective function in all DFBA simulations. The glucose uptake rate v_{glc} was assumed to follow the Michaelis-Menten kinetics [49] for the simulation of GFP-bearing *E. coli* where the data of glucose concentration during a batch were available and were used to determine a proper set of kinetic parameters $v_{glc,max}$ and K_g in Eq, (6a); whereas in the GI-malE case, v_{glc} was directly set to a function of time (Equation 6b) obtained from regression using the specific uptake rates during a batch which were directly available in the data source. The detailed derivation of Equation 6b is provided in Supplementary Material, Equations S5-S8c.

$$v_{glc,GFP} = v_{glc,max}\frac{G}{K_g+G} \tag{6a}$$

$$v_{glc,GI-malE} = f(t) \tag{6b}$$

The dynamic extracellular environment was modelled by the following ODE system:

$$\frac{dX}{dt} = \lambda X \tag{7}$$

$$\frac{dG}{dt} = -v_{glc}X \tag{8}$$

$$\frac{dA}{dt} = v_{ac}X \tag{9}$$

$$\frac{dP}{dt} = v_{recP}X \tag{10}$$

where X, G, A and P are the concentrations of biomass, glucose, acetate and recombinant protein, respectively. The growth rate (λ), the acetate exchange flux (v_{ac}) was resolved via FBA while the recombinant protein synthesis flux (v_{recP}) was fixed to the experimentally measured values. Derivations of v_{recP} are shown in Supplementary Material, Equations S9-S16. Equations 7-10 were solved by stiff ODE solver ode15s in Matlab.

2.2.4 | Estimation of proteomic parameters and ATPM

To obtain the proteomic parameters w_f^*, w_r^* and b^* for wild-type *E. coli* W3110, given the fact that the original experimental data was in the form of steady state rates of acetate production at different growth rates, we fixed the growth rates to the experimental values and used the minimization of the glucose uptake rate as the objective function for FBA. An optimization solver (the genetic algorithm toolbox in MATLAB) coupled with FBA was used to find the optimum values of w_f^*, w_r^* and b^* that give the best fit between predicted acetate production rates and the corresponding experimental measurements at different growth rates ATPM was assumed to remain as the default one (obtained from *E. coli* MG1655), because MG1655 and W3110 are genetically very similar [50,51]. Next, the three proteomic parameter values were taken into the DFBA model that incorporated the generic proteomic constraint (Equation 4) assuming that the wild-type and the recombinant strain share the identical set of w_f^*, w_r^* and b^*. The basis for this assumption is that (i) w_f^* and w_r^* represent the intrinsic proteomic cost/enzymatic requirement of fermentation and respiration reactions, which we considered unchanged between the wild type and the recombinant type, and (ii) b^* represents the proteome demand of biomass synthesis, which should not change if the assumed composition of biomass reaction does not vary under different conditions. The DFBA model was capable of simulating the batch behavior, hence allowing the simulation results to be compared with the experimentally measured batch data. The optimization

solver was now coupled with DFBA to determine the best values of ϕ^*_{recP} and (adjusted) ATPM for the GFP-bearing W3110 by minimizing the difference between predicted and measured batch data, including the accumulated concentrations of biomass, acetate, glucose and recombinant protein versus culturing time. During the analysis, glucose uptake rate was determined by Michaelis-Menten kinetics (Equation 6a and Table 1); recombinant protein production rate (v_{recP}) was fixed to the experimentally measured values. The cellular objective was set to maximizing growth rate.

In the case of GI-malE-bearing *E. coli*, batch experimental data was available for both un-induced and induced strains. Therefore, the DFBA model was adopted for both cases, coupled with the optimization solver for parameter estimation, following the same approach as for the case of GFP-bearing W3110. Firstly, w_f^* , w_r^*, b^* were estimated using the batch data of the un-induced strain. As we did not find evidence suggesting wild type MG1655 and XL1 possess similar level of ATPM, nor the value of maintenance energy has been reported specifically for XL1, ATPM was co-estimated with the proteomic parameters in this case. The lower and upper bound of the estimated ATPM were set to 3.15 and 8.39 referring to the known range of maintenance energy of various *E. coli* strains (BiGG database, http://bigg.ucsd.edu/). Thereafter, estimated values of the three proteomic parameters were taken into DFBA for the induced strain, which was coupled with the optimization solver to further estimate ϕ^*_{recP} and to re-estimate ATPM using the batch data of the induced strain. Similar to the GFP case, rates of glucose uptake and recombinant protein production were fixed to the experimental values; biomass growth maximization was the assumed cellular objective.

3 | RESULTS

Values of all the model parameters, either estimated in this work or adopted from the literature, are presented in Table 1.

Table 1. List of model parameters used in simulations

Parameters	W3110-wt	W3110-GFP	XL1-GI-malE-uninduced	XL1-GI-malE-induced
w_f^*	0.05775†	Same as wt	0.08134†	Same as uninduced type
w_r^*	0.1847†		0.7900†	
b^*	0.1†		0.4546†	
ϕ_{recP}^*	-	0.1333†	-	0.5033†
ϕ_{GFP}	-	0.06454†	-	-
ϕ_{GI}	-	-	-	0.2436†
ϕ_{max}	0.484‡	0.484‡	0.484‡	0.484‡
ATPM (mmol ATP/gDW-h)	8.39§	32.6†	3.670†	3.759†
$v_{glc,max}$ (mmol/gDW-h)	-	4.690	-	-
K_g (g/L)	-	1	-	-

† estimated in this study

‡ obtained from Scott et al., 2010

§ default value in *i*AF1260

3.1 | Modelling wild-type and GFP-bearing recombinant *E. coli* W3110

The first case study is based on an experimental work studying the expression of green fluorescent protein (GFP) in recombinant *E. coli*, with the parental strain being W3110 [52]. In order to model the growth physiology of the GFP – bearing recombinant *E. coli*, we started from finding the proteomic parameters for the wild-type W3110 (W3110-wt). The best experimental datasets we found were from [53,54]. The original data can be readily converted to steady state rates of acetate production versus growth rates [36], which were subsequently used as the input of FBA for the estimation of wild-type proteomic parameters. It is worth noting that the corresponding bacterial strain of the experimental data available was actually *E. coli* MG1655. However, MG1655 and W3110 are both *E. coli* K-12 derived wild-type strains and are genetically similar [50,51] thus we used the growth data obtained from MG1655 to approximate the behavior of W3110-wt. Figure 1 shows that with the wild-type proteomic constraint in place, the predicted rates of acetate production are in good agreement with the experimental data.

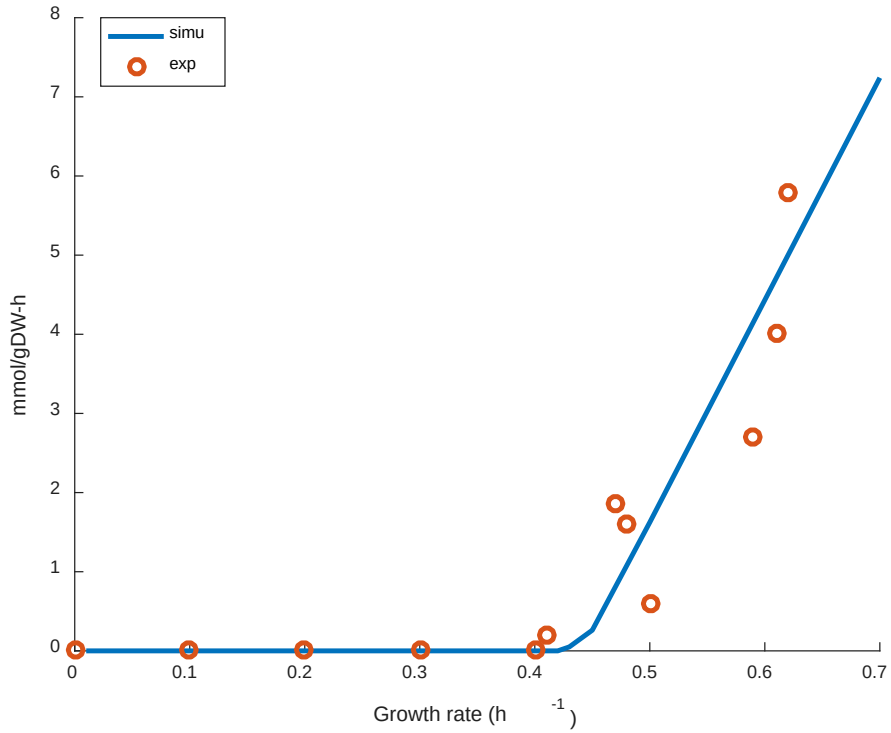

Figure 1. Comparison between model prediction and experimental data of the acetate production for wild-type *E. coli* W3110 (W3110-wt). Experimental data were obtained from Fig.3 of Mori et al., 2016 which were originally reported in Nanchen et al., 2006 and Vemuri et al., 2006 for wild-type *E.coli* MG1655. W3110 was assumed to possess the same overflow pattern with a close strain, MG1655. Abbreviations are: simu, simulation results; exp, experimental data.

We then move on to the targeted GFP-bearing recombinant *E. coli* W3110, denoted as W3110-GFP. It is worth noting that, the GFP synthesis reaction embedded in the GEM already includes the mass and energy needed for protein synthesis. What we considered in this section, i.e. ϕ^*_{recP} and ATPM, are the additional proteomic and metabolic burdens due to the production of the recombinant protein. As stated earlier (see Methods), the production of recombinant protein can provoke perturbations in (i) cellular proteome (manifested by adding ϕ^*_{recP} to the wild type proteomic constraint) and (ii) the cellular maintenance energy (modification in ATPM). To gauge the individual importance and sufficiency of each of the two parameters in predicting the growth physiology of the recombinant strain, we first fixed the ATPM to the default value (8.39 mmol ATP gDW^{-1} h^{-1}) and estimated ϕ^*_{recP} (w_f^*, w_r^* and b^* were kept identical to the wild-type) using the experimental batch growth data of W3110-GFP coupled with DFBA. In a parallel test, we estimated ATPM alone without adding ϕ^*_{recP} to the proteomic constraint. Results of these two tests are shown in Figure 2 A-D and E-H, respectively. Although adjusting ATPM

alone performed slightly better than merely adopting ϕ^*_{recP}, neither of the two individual parameters could render satisfactory model prediction.

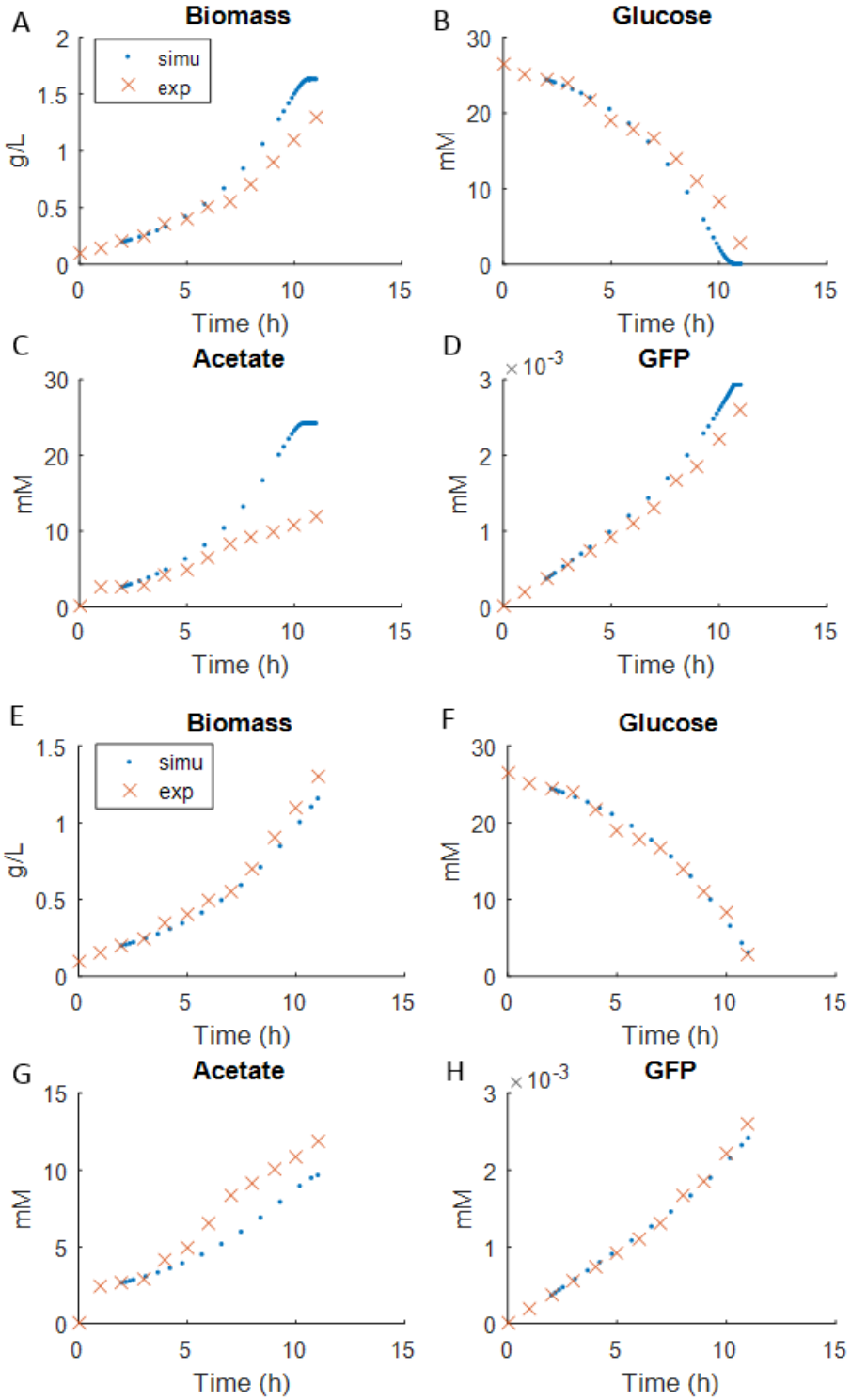

Figure 2. Comparison between model prediction and experimental data of the growth physiology of a batch culture of GFP-bearing *E. coli* W3110 (W3110-GFP). A-D. Model with estimated ϕ^*_{recP} alone, best fit $\phi^*_{recP} = 0.6266$, $\phi_{GFP} = 0.3033$. E-H. Model with changing ATPM alone, best fit ATPM = 40 mmol/gDW-h. Experimental data were obtained from Fig. 3 A-D in Lara et al., 2006. Abbreviations are: simu, simulation results; exp, experimental data; GFP, green fluorescence protein.

Next, we estimated ϕ^*_{recP} and ATPM simultaneously to see the how the combination of the two perturbations affects the accuracy of the model. Sensitivity analysis was carried out which confirmed that the predictions of biomass growth and metabolites production are sensitive to both parameters; see the Supplementary Material, Figures S1-S2 for details. This time, with a $\phi^*_{recP} = 0.133$ and an approximate three times increase in the default ATPM (see Table 1), model predictions of the accumulated concentrations of biomass, acetate, glucose and recombinant protein for W3110-GFP agreed well with the experimental data (Figure 3 A-D). This result indicates that the dual usage of the proteomic constraint and the updated ATPM is key to the accurate prediction of the growth physiology of W3110-GFP using CBMs. It is worth noting that, the estimated ATPM for W3110-GFP (32.6 mmol ATP/gDW-h) is in-line the range of the values reported previously for other recombinant *E.coli* strains (40.2-45.3 mmol ATP/gDW-h in [15] and 20 mmol ATP/gDW-h in [55]).

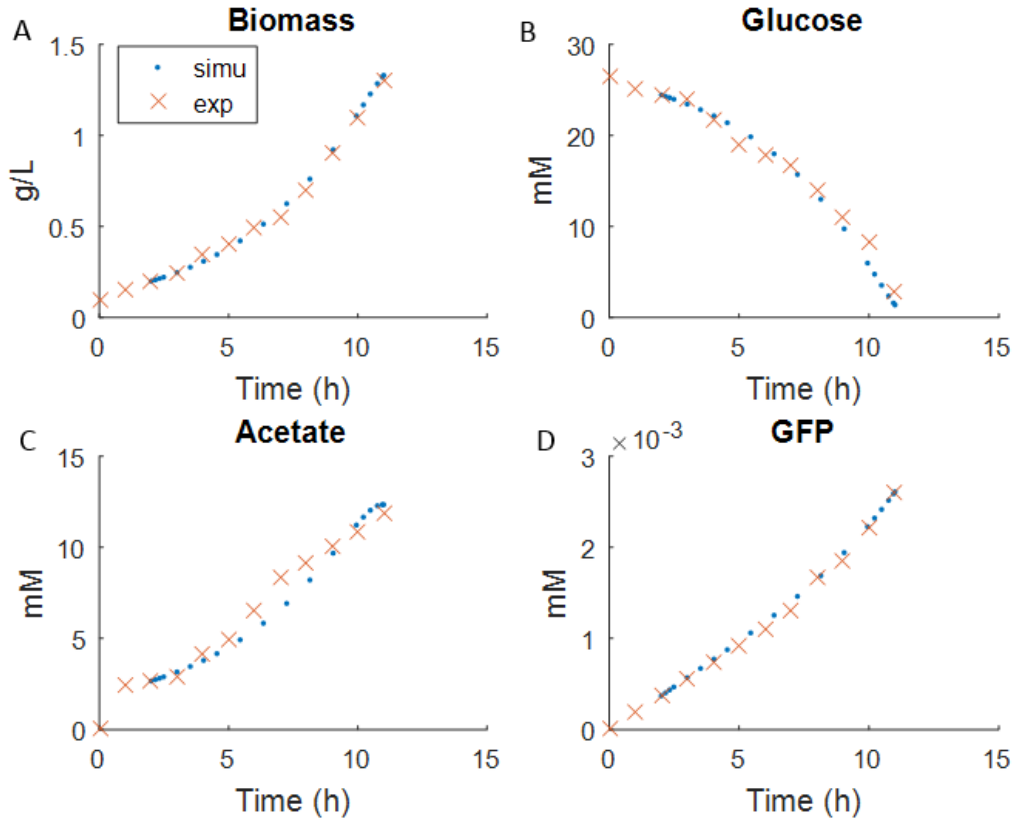

Figure 3. Comparison between model prediction and experimental data of the growth physiology of a batch culture of GFP-bearing *E. coli* W3110 (W3110-GFP). A-D. Concentrations of biomass, glucose, acetate and recombinant protein during a batch fermentation. Both ϕ^*_{recP} and ATPM were adopted to the model. The estimated values of ϕ^*_{recP} and ATPM used in simulations are given in Table 1. Experimental data were obtained from Fig. 3 A-D in Lara et al., 2006. Abbreviations are: simu, simulation results; exp, experimental data; GFP, green fluorescence protein.

3.2 | Modelling GI-malE-bearing recombinant *E. coli* XL1

The second case tested is on the production of GI-malE fusion protein expressed in recombinant *E. coli* XL1 [26], referred to as XL1-GI-malE. Similar to the GFP case, we started with determining w_f^*, w_r^* and b^* for the control strain XL1-GI-malE-uninducued (no protein production) via DFBA coupled with experimentally obtained dynamic batch growth data. As there was no basis for assuming the validity of the default ATPM (obtained from MG1655) still stands for XL1, it was co-estimated with the proteomic parameters. As for the induced XL1 (induction occurred at 10h; GI-malE production began at 10h), we estimated ϕ_{recP}^* and ATPM simultaneously with the lower bound of ATPM set to the value of the uninduced strain. Sensitivity analysis was also carried out to confirm that the model predictions are sensitive to both ϕ_{recP}^* and ATPM; see the Supplementary Material, Figure S3 for details. Interestingly, the estimated values of ATPM for induced and un-induced strains are rather close, indicating that the change of ATPM between protein production and no protein production strains in this second case plays a less significant role compared to the first case.

The simulation results with the estimated parameters are shown in Figure 4. The starting point of the simulation for the induced strain was set to 10 h in accordance to the induction time in the experiment. The original data source covered both the growth phase and the following zero-growth phase; the latter was not used here because our model was constructed to investigate the growth-related overflow metabolism and therefore acetate production at zero growth rate was beyond the scope of the current model. With the generic proteomic constraint in place, our model succeeded in capturing the overall growth physiology of both XL1-GI-malE-uninduced and XL1-GI-malE-induced. More specifically, model predictions of the biomass concentration and the production of the recombinant protein are in good agreement with the experimental results. As for acetate, simulation results are consistent with the first two experimental data points after the induction for XL1-GI-malE-induced; note that the original experimental dataset is lack of acetate data for the 15-30 h period. It should also be mentioned that the conversion of the original data available for this case (Supplementary Material, Equations S5-S8c, S12-S16) was far from straightforward to derive the dataset usable in FBA/DFBA, which might have undermined the reliability of the data that was eventually used for parameter estimation. Nevertheless,

the results clearly show that the model adopted here was able to predict that significant changes in biomass growth and acetate excretion occur due to recombinant protein production.

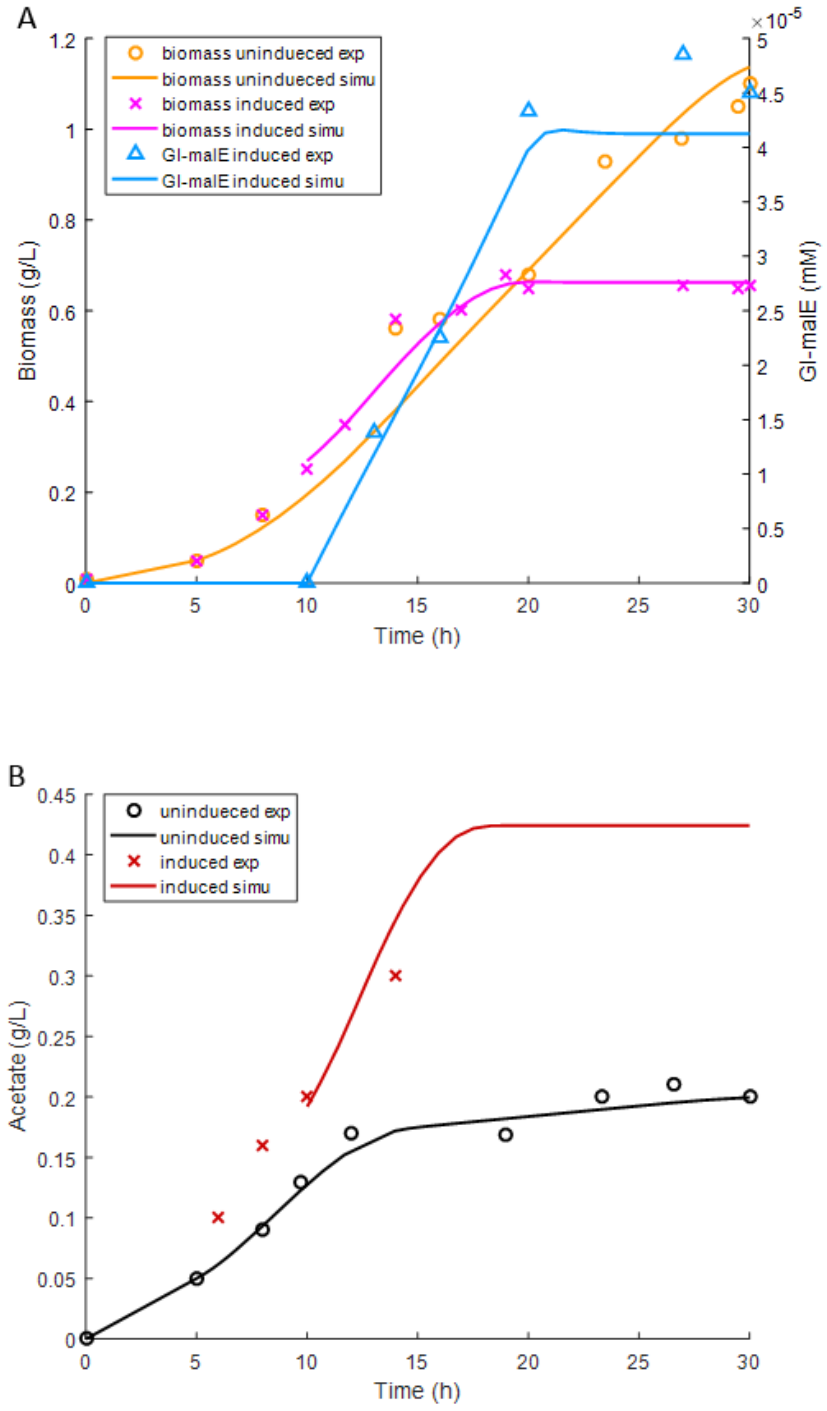

Figure 4. Comparison between model prediction and experimental data of the growth physiology of batch cultures of GI-malE-bearing recombinant *E. coli* XL1 (XL1-GI-malE-induced and -uninduced). A. Change of biomass concentration for un-induced and induced strains and recombinant protein GI-malE production for the induced strain. B. Change of acetate concentration for un-induced and induced strains. Experimental data shown in the figure were derived from Özkan et al., 2005. Abbreviations are: simu, simulation results; exp, experimental data; GI-malE, maltose binding protein-glucose isomerase fusion protein.

4 | DISCUSSION

4.1 | Comparison of proteomic parameters between the two cases

It is interesting to contrast the estimated proteomic parameters between the two recombinant strains, particularly b^* and ϕ^*_{recP}. While W3110 (wild-type) possess a relatively low b^*, that of XL1 (un-induced) is about four times larger. In the Methods section, we have indicated that b^* represents the proteomic cost of biosynthesis pathways. From the definition and in accordance with a previous study [37], the higher the b^*, the lower efficiency in the synthesis of biomass building blocks and thus the lower the maximal growth rate the strain can achieve. This is consistent with the fact that the experimentally determined maximum growth rate of W3110 is higher than that of XL1. The estimated ϕ^*_{recP} of GI-malE-bearing XL1 is also higher than GFP-bearing W3110. As stated earlier, ϕ^*_{recP} denotes the portion of the total proteome occupied for the production of the recombinant protein. GI-malE fusion protein has a higher molecular weight of 87.2 kDa while the molecular weight of GFP is about 27 kDa, which at least partially explains the discrepancy between the estimated ϕ^*_{recP} of the two cases.

4.2 | Effect of recombinant protein production on ATPM

A comparison between a recombinant strain with its wild-type (or the un-induced one) shows that ATPM variation due to the additional protein synthesis is strain dependent. For W3110, the result of modelling shows that the production of GFP leads to the dramatic increase in the maintenance energy (from 8.39 to 32.6 mmol ATP/gDW-h), whereas for XL1, no significant change in ATPM was found between the induced- and uninduced-case. This discrepancy leads to the hypothesis that the introduction of GFP to W3110 provokes strong perturbation in the metabolic network and at the same time causes a higher level of maintenance activities for e.g. sustaining cellular homeostasis. In contrast, the production of GI-malE fusion protein appears to severely disturb the host primarily through reducing the portion of proteome available to energy biogenesis and biomass synthesis, but with a rather minor change in the maintenance energy (as suggested by the large value of ϕ^*_{recP} and the small increase of ATPM, see Table 1).Similar pattern is also observed in *LacZ*-overexpressing *E. coli*, where the overflow

phenomenon was predicted accurately by considering only the proteomic disturbance, without altering the assumed cellular energy demand [18].

4.3 | Towards a predictive tool to support engineering design for recombinant protein production

In this work, DFBA has been used to establish the impact of proteomic burden (ϕ^*_{recP}) and maintenance burden (ATPM) on the batch cultures of recombinant *E. coli*. We further envisage the potential of using our model as a predictive tool to guide metabolic engineering decisions, e.g. to boost the batch productivity of the recombinant protein of interest. As mentioned in Section 1, the growth physiology predicted by our model in the production of a foreign protein, including reduction in cellular growth rate and increase in the production of un-wanted metabolites (acetate in particular) represents undesirable effects for the batch culture of recombinant strains, yet these effects typically accompany enhanced protein expression. This conflict points to the need for engineering optimization in both design (e.g. selection of an optimal protein expression level) and operation (e.g. determination of the timing of induction during a batch). To this end, it is desirable to develop a predictive model to guide these engineering decisions. The DFBA modelling approach from this work is not yet able to fulfil this purpose, due to several important limitations: (i) with respect to the design of the recombinant strain, a connection is yet to be made between the engineering of the expression system and the attainable protein expression level (quantified by e.g. specific protein production rate or mass fraction of the recombinant protein), and it is still unclear how the latter is precisely related to the proteomic burden, denoted by ϕ^*_{recP}; (ii) the relationship between ATPM of a recombinant strain and its protein expression level is to be developed; and (iii) similarly how the protein expression level would affect the substrate uptake kinetics is still unknown. Nevertheless, as an illustration of the potential usage of a fully predictive model, we have run simulations using the production of GFP-by recombinant *E.coli* as an example to show how in principle a future model could help find the optimal expression level of the recombinant protein, represented by ϕ^*_{recP} for the sake of this illustration. We assume that the glucose update kinetics remains the same as the case of W3110-GFP, and that ATPM either remains as a constant (taking the value estimated for W3110-GFP) or is linearly proportional to ϕ^*_{recP}. ϕ^*_{recP} was fixed to a

specific level at each simulation with maximizing growth rate as the cellular objective. All simulations started at the same initial glucose concentration and were kept running until glucose was depleted. The final protein concentration was recorded for each batch simulation. Further details of the approach can be found in the Supplementary Material, Equations S17-S18. The result presented in Figure 5 shows that an optimal protein expression level could be predicted, and the position of the optimal point can be affected by the variation of the extra maintenance burden (ATPM).

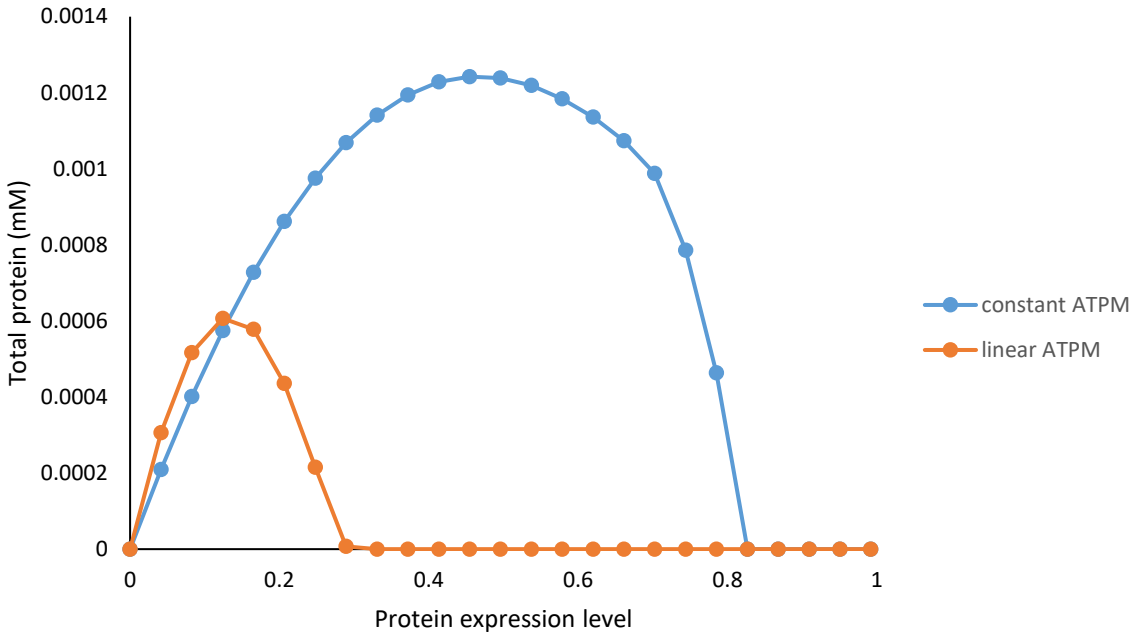

Figure 5. Illustrative model application for predicting the optimal protein expression level at constant maintenance energy and linear maintenance energy. x-axis shows the normalised level of protein expression, denoted by ϕ^*_{recP}; y-axis is the protein concentration at the end of the batch. Abbreviation: ATPM, maintenance energy.

5 | CONCLUSIONS

We have shown the effectiveness of the combined use of a proteome allocation constraint and adjustment of maintenance energy (ATPM) in predicting the retarded growth and the increased acetate production for recombinant *E. coli* via constraint-based models (CBMs). To our knowledge, this is the first time a CBM, and DFBA in particular, is used for capturing such observed growth physiology for recombinant strains. Through two test cases, the inclusion of the proteomic burden for synthesizing a recombinant protein appears to be essential for the model prediction, whereas the significance of ATPM adjustment may be strain-dependent. This work quantitatively depicts the proteomic and metabolic

burdens of recombinant protein expression in terms of the proportion of the proteomic resource occupied by the extra protein production and the increase in maintenance energy, respectively. Further development of the current modelling framework to overcome several limitations has the potential to offer a starting point for the development of a practical, model-based tool, which complements other types of tools to guide design and operational decisions for efficient protein production.

ACKNOWLEDGEMENT

H.Z. is supported by the China Scholarship Council through a PhD scholarship.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Ahmad, M., Hirz, M., Pichler, H. & Schwab, H. Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production. *Appl. Microbiol. Biotechnol.* **98**, 5301–5317 (2014).
2. Kim, J. Y., Kim, Y.-G. & Lee, G. M. CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Appl. Microbiol. Biotechnol.* **93**, 917–930 (2012).
3. Demain, A. L. & Vaishnav, P. Production of recombinant proteins by microbes and higher organisms. *Biotechnol. Adv.* **27**, 297–306 (2009).
4. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in Escherichia coli: advances and challenges. *Front. Microbiol.* **5**, 172 (2014).
5. Choi, J. H., Keum, K. C. & Lee, S. Y. Production of recombinant proteins by high cell density culture of Escherichia coli. *Chem. Eng. Sci.* **61**, 876–885 (2006).
6. Chen, R. Bacterial expression systems for recombinant protein production: E. coli and beyond. *Biotechnol. Adv.* **30**, 1102–1107 (2012).
7. Huang, C.-J., Lin, H. & Yang, X. Industrial production of recombinant therapeutics in Escherichia coli and its recent advancements. *J. Ind. Microbiol. Biotechnol.* **39**, 383–399 (2012).
8. Mahalik, S., Sharma, A. K. & Mukherjee, K. J. Genome engineering for improved recombinant protein expression in Escherichia coli. *Microb. Cell Fact.* **13**, 177 (2014).
9. Zhou, K., Qiao, K., Edgar, S. & Stephanopoulos, G. Distributing a metabolic pathway among a microbial consortium enhances production of natural products. *Nat. Biotechnol.* **33**, 377–383 (2015).
10. Eiteman, M. A. & Altman, E. Overcoming acetate in Escherichia coli recombinant protein fermentations. *Trends in Biotechnology* **24**, 530–536 (2006).
11. Corchero, J. L. & Villaverde, A. Plasmid maintenance in Escherichia coli recombinant cultures is dramatically, steadily, and specifically influenced by features of the encoded proteins.

Biotechnol. Bioeng. **58**, 625–632 (1998).

12. Jordà, J. *et al.* Metabolic flux profiling of recombinant protein secreting Pichia pastoris growing on glucose:methanol mixtures. *Microb. Cell Fact.* **11**, 57 (2012).

13. Heyland, J., Fu, J., Blank, L. M. & Schmid, A. Quantitative physiology of Pichia pastoris during glucose-limited high-cell density fed-batch cultivation for recombinant protein production. *Biotechnol. Bioeng.* **107**, 357–368 (2010).

14. van Rensburg, E., den Haan, R., Smith, J., van Zyl, W. H. & Görgens, J. F. The metabolic burden of cellulase expression by recombinant Saccharomyces cerevisiae Y294 in aerobic batch culture. *Appl. Microbiol. Biotechnol.* **96**, 197–209 (2012).

15. Heyland, J., Blank, L. M. & Schmid, A. Quantification of metabolic limitations during recombinant protein production in Escherichia coli. *J. Biotechnol.* **155**, 178–184 (2011).

16. Wu, G. *et al.* Metabolic Burden: Cornerstones in Synthetic Biology and Metabolic Engineering Applications. *Trends Biotechnol.* **34**, 652–664 (2016).

17. Bhattacharya, S. K. & Dubey, A. K. Metabolic burden as reflected by maintenance coefficient of recombinant Escherichia coli overexpressing target gene. *Biotechnol. Lett.* **17**, 1155–1160 (1995).

18. Basan, M. *et al.* Overflow metabolism in Escherichia coli results from efficient proteome allocation. *Nature* **528**, 99–104 (2015).

19. Becker, S. A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc* **2**, (2007).

20. Saitua, F., Torres, P., Pérez-Correa, J. R. & Agosin, E. Dynamic genome-scale metabolic modeling of the yeast Pichia pastoris. *BMC Syst. Biol.* **11**, 27 (2017).

21. Nocon, J. *et al.* Model based engineering of Pichia pastoris central metabolism enhances recombinant protein production. *Metab. Eng.* **24**, 129–138 (2014).

22. Sohn, S. B. *et al.* Genome-scale metabolic model of methylotrophic yeast Pichia pastoris and its use for in silico analysis of heterologous protein production. *Biotechnol. J.* **5**, 705–715 (2010).

23. Hirokawa, Y., Matsuo, S., Hamada, H., Matsuda, F. & Hanai, T. Metabolic engineering of Synechococcus elongatus PCC 7942 for improvement of 1,3-propanediol and glycerol production based on in silico simulation of metabolic flux distribution. *Microb. Cell Fact.* **16**, 212 (2017).

24. Oddone, G. M., Mills, D. A. & Block, D. E. A dynamic, genome-scale flux model of Lactococcus lactis to increase specific recombinant protein expression. *Metab. Eng.* **11**, 367–381 (2009).

25. Gutierrez, J. M. & Lewis, N. E. Optimizing eukaryotic cell hosts for protein production through systems biotechnology and genome-scale modeling. *Biotechnol. J.* **10**, 939–949 (2015).

26. Özkan, P., Sariyar, B., Ütkür, F. Ö., Akman, U. & Hortaçsu, A. Metabolic flux analysis of recombinant protein overproduction in Escherichia coli. *Biochem. Eng. J.* **22**, 167–195 (2005).

27. Molenaar, D., van Berlo, R., de Ridder, D. & Teusink, B. Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol* **5**, (2009).

28. Beg, Q. K. *et al.* Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *P Natl Acad Sci USA* **104**, (2007).

29. Vazquez, A. *et al.* Impact of the solvent capacity constraint on E. coli metabolism. *BMC Syst Biol* **2**, (2008).

30. Goelzer, A., Fromion, V. & Scorletti, G. Cell design in bacteria as a convex optimization problem. *Automatica* **47**, 1210–1218 (2011).

31. Goelzer, A. & Fromion, V. Bacterial growth rate reflects a bottleneck in resource allocation. *Biochim. Biophys. Acta (BBA)-General Subj.* **1810**, 978–988 (2011).

32. Goelzer, A. & Fromion, V. Resource allocation in living organisms. *Biochem. Soc. Trans.* BST20160436 (2017).

33. Goelzer, A. *et al.* Quantitative prediction of genome-wide resource allocation in bacteria. *Metab. Eng.* **32**, 232–243 (2015).

34. O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. O. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693–693 (2014).

35. Zhuang, K., Vemuri, G. N. & Mahadevan, R. Economics of membrane occupancy and respiro-fermentation. *Mol Syst Biol* **7**, (2011).

36. Mori, M., Hwa, T., Martin, O. C., De Martino, A. & Marinari, E. Constrained Allocation Flux Balance Analysis. *PLoS Comput. Biol.* **12**, (2016).

37. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science (80-.).* **330**, 1099–1102 (2010).

38. Nilsson, A., Nielsen, J. & Palsson, B. O. Metabolic models of protein allocation call for the kinetome. *Cell Syst.* **5**, 538–541 (2017).

39. Pirt, S. J. The maintenance energy of bacteria in growing cultures. *Proc. R. Soc. Lond. B* **163**, 224–231 (1965).

40. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**, (2007).

41. Liu, M. *et al.* Metabolic engineering of Escherichia coli to improve recombinant protein production. *Appl. Microbiol. Biotechnol.* **99**, 10367–10377 (2015).

42. Glick, B. R. Metabolic load and heterologous gene expression. *Biotechnol. Adv.* **13**, 247–261 (1995).

43. Bailey, J. E. Critical limitations in biological production of chemicals: process or genetic solutions? *FEMS Microbiol. Rev.* **16**, 271–276 (1995).

44. Keseler, I. M. *et al.* EcoCyc: a comprehensive database of Escherichia coli biology . *Nucleic Acids Res.* **39**, D583–D590 (2011).

45. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46**, D633–D639 (2018).

46. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248 (2010).

47. Schellenberger, J. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. *Nat. Protoc.* **6**, 1290 (2011).

48. Mahadevan, R., Edwards, J. S. & Doyle, F. J. Dynamic Flux Balance Analysis of Diauxic Growth in Escherichia coli. *Biophys. J.* **83**, 1331–1340 (2002).

49. Hjersted, J. L. & Henson, M. a. Steady-state and dynamic flux balance analysis of ethanol production by Saccharomyces cerevisiae. *IET Syst. Biol.* **3**, 167–179 (2009).

50. Hayashi, K. *et al.* Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol. Syst. Biol.* **2**, (2006).

51. Jensen, K. F. The Escherichia coli K-12" wild types" W3110 and MG1655 have an rph frameshift mutation that leads to pyrimidine starvation due to low pyrE expression levels. *J. Bacteriol.* **175**, 3401–3407 (1993).

52. Lara, A. R. *et al.* Engineering Escherichia coli to improve culture performance and reduce formation of by-products during recombinant protein production under transient intermittent anaerobic conditions. *Biotechnol. Bioeng.* **94**, 1164–1175 (2006).

53. Nanchen, A., Schicker, A. & Sauer, U. Nonlinear dependency of intracellular fluxes on growth rate in miniaturized continuous cultures of Escherichia coli. *Appl. Environ. Microbiol.* **72**, 1164–1172 (2006).

54. Vemuri, G. N., Altman, E., Sangurdekar, D. P., Khodursky, A. B. & Eiteman, M. A. Overflow metabolism in Escherichia coli during steady-state growth: transcriptional regulation and effect of the redox ratio. *Appl Env. Microbiol* **72**, (2006).

55. Weber, J., Hoffmann, F. & Rinas, U. Metabolic adaptation of Escherichia coli during temperature-induced recombinant protein production: 2. Redirection of metabolic fluxes. *Biotechnol. Bioeng.* **80**, 320–330 (2002).