Derivation of a biomass proxy for dynamic analysis of whole genome metabolic models

Timothy Self[1], David Gilbert[1], and Monika Heiner[1,2]

[1] Brunel University London, UK
timself101@hotmail.com, david.gilbert@brunel.ac.uk
[2] Brandenburg Technical University, Cottbus, Germany
monika.heiner@{b-tu.de|brunel.ac.uk}

Abstract. A whole genome metabolic model (GEM) is essentially a reconstruction of a network of enzyme-enabled chemical reactions representing the metabolism of an organism, based on information present in its genome. Such models have been designed so that flux balance analysis (FBA) can be applied in order to analyse metabolism under steady state. For this purpose, a biomass function is added to these models as an overall indicator of the model's viability.

Our objective is to develop dynamic models based on these FBA models in order to observe new and complex behaviours, including transient behaviour. There is however a major challenge in that the biomass function does not operate under dynamic simulation. An appropriate biomass function would enable the estimation under dynamic simulation of the growth of both wild-type and genetically modified bacteria under different, possibly dynamically changing growth conditions.

Using data analytics techniques, we have developed a dynamic biomass function which acts as a faithful proxy for the FBA equivalent for a reduced GEM for *E. coli*. This involved consolidating data for reaction rates and metabolite concentrations generated under dynamic simulation with gold standard target data for biomass obtained by steady state analysis using FBA. It also led to a number of interesting insights regarding biomass fluxes for pairs of conditions. These findings were reproduced in our dynamic proxy function.

1 Introduction

A large amount of publicly available information, regarding whole genome metabolic reaction networks in e.g. *Escherichia coli* (*E. coli*), has been encoded as constraint-based flux-balance analysis (FBA) models. This forms a very useful resource, especially when combined with genome information, as in the BiGG collection [13]. Our overall aim is to build on this knowledge to make whole genome metabolic models (GEMs) available for dynamic simulation in order to be able to observe new and complex behaviours including, for example, under dynamically changing growth conditions. In previous work we have reported our methodology to convert FBA models into dynamic models [7], as the first steps that we have already made in this direction.

Constraint-based FBA models are designed to analyse metabolism activity under steady state. For this purpose a biomass function is added, implemented as an abstract reaction over metabolites and serving as an overall indicator of the model's

viability. However this artificial function is very complex and highly tuned in that it comprises many substrates and products, with a wide range of specific non-integer stoichiometries [22]. This tuned complexity means that we have found it impossible to directly use the existing FBA biomass function as an indicator of viability in the simulation of dynamic models.

The work reported in this paper describes a data analytics approach to derive a proxy biomass function for dynamic GEMs, relying on averaged stochastic simulation traces of both metabolite concentrations and reaction rates. This proxy has been developed to be both highly robust and accurate with respect to a wide variety of growth conditions. Such a biomass proxy will enable the estimation by dynamic simulation of the growth of both wild-type and genetically modified bacteria under different growth conditions.

Our contributions include: the development of a well-defined general method, organised as a workflow which provides guidance to derive a biomass function for any GEM. We demonstrate our method for the well-established reduced *E. coli* core model for the K12 strain [21] available in SBML format. Our workflow exploits a number of well recognised data analytics methods, including regression analysis and machine learning. The gold standard FBA data on which our work is based was generated using the *Cobra* software [25], and the dynamic simulation data was generated with the stochastic simulation algorithm Delta Leaping [23] using the *Marcie* software [11]. For this purpose, we converted the SBML model into a stochastic Petri net by help of the *Snoopy* software [10]. The robustness of our results was ensured by the use of a large number of observations generated by single and combined growth conditions. An additional unexpected result was the observation of the non-linear additive effects of certain paired growth conditions which were found in the FBA results and faithfully preserved in the predictions of our biomass proxy.

This paper is organised as follows. In the next section we review some related work, followed by a section on the data used, its generation and preparation. Next we describe the data analytics methods deployed and their application in our workflow. We then evaluate the key results, followed by conclusions and outlook. Some additional information is provided as Supplementary Materials, available at
`http://www-dssz.informatik.tu-cottbus.de/DSSZ/Software/Examples`.

2   Related work

A genome scale metabolic model (GEM) is essentially a reconstruction of a network of enzyme-enabled chemical reactions representing the metabolism of an organism, based on information present in its genome. It can be used to understand an organism's metabolic capabilities. The reconstruction involves a number of steps, including the functional annotation of the genome, identification of the associated reactions and determination of their stoichiometry, which is the relationship between the relative quantities of substances taking part in a reaction. It also involves determining the biomass composition, estimating energy requirements and defining model constraints [1]. A characteristic of these models is that although they describe the reactions in terms of substrates and products, they do not contain information on

reaction rate constants because these cannot be determined by the current reconstruction process.

GEMs have become an invaluable tool for analysing the properties and steady state behaviour of metabolic networks, and have been especially successful for *E. coli* [9]. The most recent model iJO1366 has been accepted as the reference for *E. coli* network reconstruction. It has provided valuable insights into the metabolism of *E. coli* and been used to formulate intervention strategies for targeted modifications of the metabolism for biotechnological applications. Bacterial GEMs can comprise about 5000 reactions and metabolites, and encode a huge variety of growth conditions. The BiGG public domain database contains 92 GEMs, of which 52 are for *E. coli* [13].

However, it has been argued that as the size and complexity of genome scale models increases, limitations are placed on popular modelling techniques, such as constraint-based modelling and kinetic modelling [9]. A similar argument was put forward by [4] as a justification for developing a network reduction algorithm to derive smaller models by unbiased stoichiometric reduction, based on the view that the basic principles of an organism's metabolism can be studied more easily in smaller models. A number of reduced models have been proposed, including [21] and [9].

Flux balance analysis (FBA) is a constraint-based approach for analysing the flow of metabolites through a metabolic network by computing the reaction fluxes in the steady state. This enables the prediction of the growth rate of an organism or the production rate of biotechnologically important metabolites. An additional biomass objective function is added to compute an optimal network state and resulting flux distribution out of the set of feasible solutions. The growth rate as reflected by the steady state flux of the biomass function is constrained by the measured substrate uptake rates and by maintenance energy requirements [22].

The biomass function indirectly indicates how much certain reactions contribute to the phenotype. It does so by being represented as a pseudo (i.e. abstract and artificial) "biomass reaction" that drains substrate metabolites from the system at their relative stoichiometries to simulate biomass production The biomass reaction is based on experimental measurements of biomass components. This reaction is scaled so that the flux through it is equal to the exponential growth rate (μ) of the organism [20].

FBA has limitations as it is unable to predict metabolite concentrations because it does not use initial metabolite concentrations or kinetic parameters. The mathematical model incorporates the stoichiometric matrix and any biologically meaningful constraints over the flux ranges. Therefore it is only suitable for determining relative fluxes at steady state [21].

Dynamic simulation. The network described by the stoichiometric matrix can be equally read as a dynamic model to explore the temporal behaviour of the system by tracing how metabolite concentrations and reaction rates (fluxes) change over time [6, 7]. For this purpose the model has to be enriched by initial metabolite concentrations and kinetic reaction rates (kinetic laws and corresponding parameters), both initially estimated and ultimately determined by experimental observation.

There are three main approaches for dynamic simulation: qualitative, stochastic, and deterministic approaches. The most abstract representation of a biochemical network is qualitative. However, biochemical systems are inherently governed by stochas-

tic laws, though due to the computational resources required, continuous models are commonly used in place of stochastic models to approximate stochastic behaviour with a deterministic approach [8]. These approaches to dynamic simulation do not make any assumptions about steady state, unlike FBA and dynamic FBA [16], thus facilitating the analysis of the transient behaviour of the biological system.

The dynamic simulation of large and complex whole genome models has been a bottleneck in the past [26], which has presented considerable difficulties both for stochastic and deterministic methods [7]. However, stochastic simulation based on Delta Leaping [24], permits the efficient simulation of these very large GEMs, enabling the observation of new and complex behaviours [7].

There is also another limitation however, which is that the biomass function for constraint-based GEMs does not work correctly under the dynamic simulation of transient behaviour without quasi-steady state assumption, due to the complexity in terms of the number of variables and specificity in terms of the stoichiometries of the function. A systematic approach to the development of a proxy function that can be used to determine the amount of biomass produced is the main focus of the work presented here.

3   Data

Model. The research reported in this paper builds on the reduced *E. coli* core model for the K12 strain of Orth et al. [21] available in SBML format from `http://systemsbiology.ucsd.edu/Downloads/EcoliCore`. Its reactions and pathways have been chosen to represent the most well-known and widely studied metabolic pathways of *E. coli*.

The metabolic reconstruction of the model includes 54 unique metabolites in two compartments: cytosol and extracellular, and these metabolites may exist as SBML species in both, differentiated by appropriate tags. The cytosol contains 52 of these metabolites (of which 34 are uniquely cytosol species), and the extracellular compartment contains 20 metabolites, two of which are not found in the cytosol. By definition each of the 20 extracellular metabolites exists as two copies – 'boundary' and 'extracellular' – the latter type being used in the transport mechanism between extracellular and cytosol compartments, making 40 extracellular species. In total there are 52+40=92 species in the SBML specification.

The model has 94 reactions of which 46 are reversible, which can be categorised into 49 metabolic reactions, 25 transport reactions between compartments, and 20 exchange reactions [20]. Exchange reactions are always reversible and exist for each extracellular metabolite (boundary condition), the directions of which can be changed using the flux constraints. Additionally there is a biomass function implemented as an abstract (irreversible) reaction, which comprises 16 substrates and 7 products with stoichiometries varying from 0.0709 to 59.81, see Table 4 in the Supplementary Materials.

The model can be configured to investigate the effect of different growth conditions using the 20 extracellular species. Of these, 14 are carbon source growth conditions including formate; we follow standard practice to ignore formate due to viability

issues [20] leaving 13 carbon sources that we considered. Five of the remaining 6 boundary conditions correspond to the ingredients of a minimal growth medium based on M9 [17], namely CO_2, H+, H_2O, D-Glucose, ammonium, and phosphate. Finally, oxygen is also a boundary condition. Each of the carbon sources can be considered both aerobically and anaerobically, making in total $2 \cdot 13 = 26$ single growth conditions while ignoring formate.

For the purpose of simulation, we convert the SBML model into a stochastic Petri net (SPN), which is done with the Petri net editor and simulator *Snoopy* [10]. This involves four adjustments.

- As required for any discrete dynamic modelling approach, reversible reactions are modelled by two opposite transitions representing the two directions a reversible reaction can occur.
- Metabolites which have been declared as boundary conditions are associated with additional source and sink transitions (called boundary transitions), mimicking the FBA assumption of appropriate in/outflow. This transforms a place-bordered net into a transition-bordered net, if all boundary places (i.e., source/sink places) have been declared as boundary conditions.
- Reaction rates are assigned to all transitions following the mass-action pattern with uniform kinetic parameters of 1.
- The initial concentration is set to zero, except for those 12 metabolites involved in mass conservation (P-invariants), computed with *Charlie* [12], which were all set to the same initial amount, e.g. 10.

The Petri net model (ignoring the biomass function) comprises 180 transitions $(94 + 46 + 2 \cdot 20$ boundary transitions) and 92 places; it is shown in Fig. 6 in the Supplementary Materials. In addition, the biomass function is present, but never active.

Datasets. "Gold standard" target data for biomass was generated using FBA with the *Cobra* software [25]. Time-series data for reaction activity and metabolite concentrations were generated under dynamic simulation with the approximative stochastic simulation algorithm Delta Leaping using the *Marcie* software [11] and recorded for all species (places) and reactions (transitions) for 1,000 time points averaged over 10,000 runs. The average of the last 200 time points was calculated for each of the reactions and metabolites in the dynamic time series data for use in regression analysis, based on the assumption that this best represented the steady state. The rates of the forward and backward transitions of reversible reactions were combined and appropriate new variables introduced with suffix FwRe. Redundant variables were removed, e.g. boundary transitions introduced by the conversion of SBML into SPN, and also the original biomass function. These data preparation steps are summarised in Fig. 1.

This was initially done for the 26 different single growth conditions. However, the number of predictors (reactions and metabolites) in our model is large in relation to the number of observations (growth conditions), and regression analysis is more accurate for larger numbers of observations. Furthermore, analysis based on more conditions helps to reduce the likelihood of overfitting and allows more predictor variables to be included in the regression equation. Given a certain number of observations, there is an upper limit to the complexity of the model that can be derived

with any acceptable degree of uncertainty [2]. Also there is a broadly linear relationship between the sample size (number of observations) and the number of predictors included in a multiple linear regression model used for prediction [14].

Therefore, additional observations were generated using pairwise combinations of 13 carbon sources, both aerobic and anaerobic, yielding $2 \cdot (13^2 - 13)/2 = 156$ pairwise conditions. This also enabled us to investigate the effect of paired combinations of carbon sources. Finally in order to further enhance the effectiveness of the regression analysis, we combined the 156 pairwise observations with the 26 single condition observations to create a combined dataset of 182 observations (growth conditions) with 300 variables (metabolites and reactions) and 1 dependant variable (Biomass proxy).
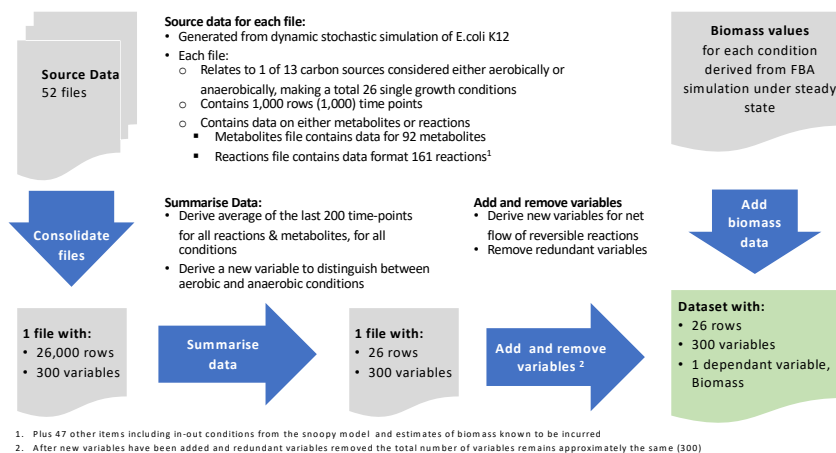

Fig. 1. Summary of data preparation steps. A dataset was created for analysis with 300 variables and 26 single growth conditions, later extended by 156 pairwise conditions.

4   Data analytics methods

In terms of data analytics we wish to derive a mathematical function to correctly predict w.r.t. FBA results the biomass, to be precise: the steady state flux of the biomass reaction, for different growth conditions based on the various metabolite concentrations and reaction rates in the steady state as determined by simulation of a dynamic model. In other words, the expected result is a proxy function which predicts the FBA value in the steady state. Our analysis is based on the assumption that a steady state exists for that model.

The overall workflow is illustrated in Fig. 2, and essential steps are explained below. A more detailed workflow protocol is provided in the Supplementary Materials.

We use the term regression analysis to refer to the analysis of the relationships between a dependent variable (which in this paper is biomass) and the predictor variables (which in this paper are the metabolites and reactions).
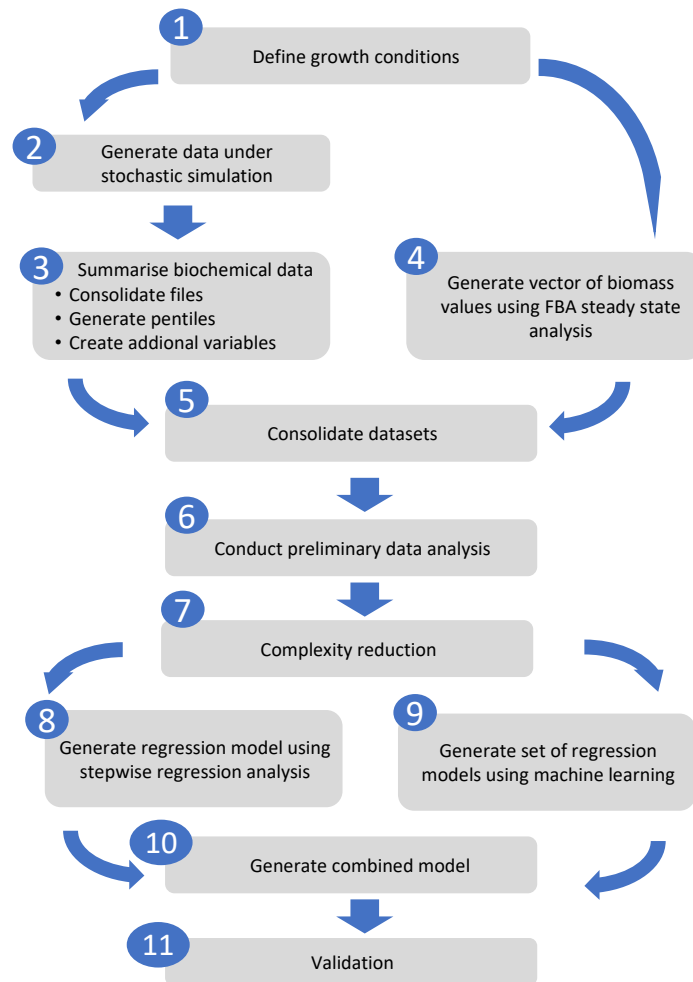
Fig. 2. Workflow of key analytical steps in the development of the proxy function. Steps 1–5 are covered in Section 3, steps 6–10 in Section 4 and step 11 in Section 5.

Preliminary data analysis generated two important observations that shaped the approach for regression analysis.

(i) The biomass values of anaerobic conditions follow "zero inflated distribution" (in which a large portion of values were either zero or close to zero), whereas biomass production for aerobic conditions resembled a normal distribution, as illustrated in Fig. 3. This finding led to the creation of a dichotomous (binary) variable to distinguish between the two sets of conditions.

(ii) There are a large number of independent variables or predictors with the potential to lead to too much complexity. Many of these variables were highly correlated

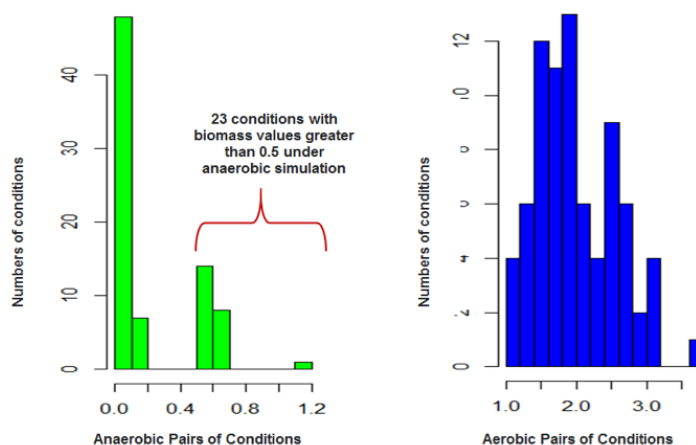with each other, leading to collinearity which can cause inaccurate predictors in the regression equation.

Fig. 3. Histograms of biomass for pairs of conditions; anaerobic conditions (left) and aerobic conditions (right).

Complexity reduction. The large number of predictors (metabolite concentrations and reaction fluxes) had the potential to generate overwhelming complexity. Further investigation into these variables revealed that a large number of them were highly correlated with each other and this helped to reduce some of the complexity and to highlight the risk of collinearity in the regression model. Some variables were in fact found to be perfectly correlated, given that they had Pearson correlation coefficients of 1. This was because they related to the same metabolite at different stages in transportation (or different compartments) as represented by the underlying biology model. So, the concentrations of metabolite did not change irrespective of whether it was outside the *E. coli* bacterium or passing through the outer part of the *E. coli* bacterium.

Clustering techniques were used to identify groups of highly correlated variables. Fig. 4 provides an illustration of hierarchical variable clustering using complete linkage for 37 key variables. Note that only a limited number of variables were used as including all 300 would not be visually effective.

We applied two approaches to regression analysis — stepwise regression and a machine learning based algorithmic approach.

Stepwise regression analysis. The decision was taken to initially develop a multiple linear regression model to predict biomass in preference to employing machine learning algorithms, due to the additional insight that statistical methods can offer in terms of inference or interpretability of the parameters (which in this case is the underlying biochemistry) as opposed to simply looking at prediction. The methodol-
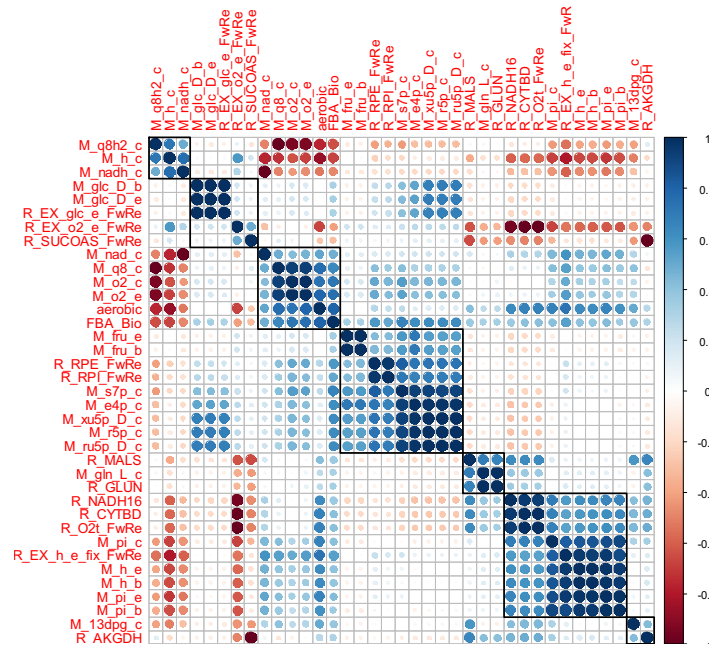
Fig. 4. Variable correlation matrix with hierarchical clustering based on complete linkage, using 37 key variables from the initial dataset of 26 single conditions.

ogy and approach applied to the regression analysis was strongly influenced by the findings identified in the preliminary analysis described above.

An initial regression equation was derived by applying a process of stepwise regression to the small dataset of 26 single conditions. Variables were included in the stepwise regression process based on the earlier work carried-out around correlation and clustering analysis. The validity of using correlation analysis could be questioned because a linear combination of a few dependent variables that are only weakly correlated with the dependent variable may have larger correlation with the dependent variable than a linear combination of a few strongly correlated variables. However, it should be pointed out that an alternative, more formulaic approach to variable selection was applied when running the automated feature selection as outlined in the section below.

In spite of the fact that a dataset with only 26 conditions imposed limits on the scope of the regression analysis, the results were promising and served as a starting point for analysis based on the larger dataset. A process of stepwise regression followed in which different terms were successively added and removed. A regression equation with an adjusted r-squared value of 0.91 was obtained, providing a fair amount of explanatory power.

A comparison of datasets generated for paired, as opposed to single conditions identified a drop in the adjusted r-squared value from 0.92 to 0.83 when the same

regression model was applied to the dataset for paired conditions. This finding led to the creation of a second dichotomous variable '$Pair$' to distinguish between paired condition and single conditions.

A procedure known as StepAIC (available in the Mass package in R) was then applied and the results obtained were used to help validate this model. Interaction terms were then added to reflect the combined effect of the predictors (metabolite concentrations and reaction fluxes) and the dichotomous variable created to distinguish between aerobic and anaerobic conditions. The inclusion of such terms in the regression model led to a significant improvement and an adjusted r-squared of 0.976 was obtained, illustrating the strong explanatory power. Furthermore, all of the coefficients and the overall model were shown to be highly statistically significant

A machine learning based algorithmic approach to regression. One of the main challenges identified in the preliminary analysis was the need to manage the complexity created by the large number of variables (300), which is a characteristic of many modern datasets. Kursa and Rudnicki identified two main issues with large datasets. Firstly, the decrease in accuracy that can occur when too many variables are included, known as the *minimal optimal problem*. Secondly, the challenges in finding all relevant variables as opposed to just the non-redundant ones, which is known the *all-relevant problem* [15]. This is of particular importance when one wishes to understand the mechanisms related to the subject of interest, as opposed to purely building a black box predictive model. Kursa and Rudnick have developed *Boruta*, a package in R [3] for variable selection, which includes a variable selection algorithm (also called Boruta) to address the all-relevant problem. The algorithm employs a wrapper approach which is built around a random forest classifier. In a wrapper approach, the classifier (in this case a random forest classifier) is used as a black box to return output, which is used to evaluate the importance of variables. Random forest is an ensemble method used in machine learning in which classification is performed by voting on (or taking the average of) multiple unbiased weak classifiers - decision trees. These trees are independently developed on different samples of the training dataset [15].

Diagnostics terminology. In the following we first explain the terminology of the methods that we have used, followed by their application in our approach.

Akaike's information criterion (AIC) is a diagnostic used in regression, which takes into account how well the model fits the data while adjusting for the ability of that model to fit any dataset. It seeks to strike a balance between goodness of fit and parsimony and assigns a penalty based on the number of predictors to guard against overfitting. It is defined as

$$AIC = -2 \cdot ln(L) + 2 \cdot p,$$

where L is a measure of the log likelihood and p is the number of variables in the model [18].

Bayesian information criterion (BIC) is a Bayesian extension of AIC with

$$BIC = -2 \cdot ln(L) + p \cdot ln(n),$$

where L is a measure of the log likelihood and p is the number of variables in the model as above. It is known to be a more conservative measure than AIC in the sense

that it assigns a stronger penalty as more predictors are added to the model. Like AIC, the lower the value of BIC the better.

Note that AIC and BIC are used to determine the relative quality of different statistical models based on the same dataset. They cannot be used to compare models generated from different datasets

Cross-validation is an evaluation technique, which is used to assess the accuracy of results obtained from training data on test data. In cross-validation, the number of folds 'k' is defined in advance. The data is then split equally into 'k' folds. Each fold in turn is used for testing and the remainder used for training. This procedure is repeated 'k' times so that at the end every instance has been used exactly once for testing [27]. The cross-validation residual is then derived by calculating the difference between the prediction using the 'refit' regression model and the actuals for the test dataset. Witten et al. claim that in extensive tests on numerous different datasets, with different learning techniques, 10-fold cross validation is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up [27].

The following steps were applied in our algorithmic approach to perform regression analysis in order to derive a proxy function for biomass.

(i) The Boruta package in R was used to identify 80 important independent variables from a total of 300.

(ii) Collinearity was then removed by eliminating variables with a variance inflation factor (VIF) higher than 4, using a routine developed with the 'car' package in R [5]. Collinearity refers to strong correlations between independent variables. It can result in biased coefficients in the regression equation, which means it is difficult to assess the impact of the independent variables on the dependent variable. VIF is an excellent measure of the collinearity of the ith independent variable with the other independent variables in the model, according to O'Brien [19]. He also argues against the need to apply low VIF thresholds as was the case here. In fairness a higher threshold could have been used, however as we will see below, the algorithm used to test all the combinations of linear regression models is extremely resource intensive and only a limited number could be employed.

(iii) A matrix was created to store all of the potential subsets of predictors.

(iv) Training and validation samples were created.

(v) Linear models were generated, using the 14 most important variables resulting in the creation of 16,384 (2^{14}) different models. Due to the exponential complexity of the problem we confined our analysis to a maximum of 14 variables, which took about 2 hours to run.

(vi) Key diagnostics are captured for all models including: r-squared, adjusted r-squared, p-values, AIC and BIC and k-fold cross-validation mean squared error.

Combining results from stepwise regression with the machine learning based algorithmic approach. The results obtained from this algorithmic approach were inferior to the results obtained through stepwise regression. However, there were 12 predictors that appeared in the top *algorithmic models* that were also absent from the stepwise regression model, which were reviewed in order to determine whether any improvement could be made to the results of the stepwise regression analysis.

After another process of stepwise regression, two additional predicators were included and a regression model to estimate the biomass was developed leading to an improvement in the adjusted r-squared from 0.976 to 0.979:

$$
\begin{aligned}
Biomass \approx \\
& - 14.2113 \\
& + 2.1133 \cdot M\_fru\_b + 2.1744 \cdot M\_glc\_D\_b + 4.5078 \cdot M\_o2\_b \\
& + 13.4913 \cdot R\_GLUN \\
& + Aerobic\,(0.7191 \cdot Pair - 0.1056 \cdot M\_h\_b \\
& \quad\quad + 1.8578 \cdot M\_fru\_b + 1.8466 \cdot M\_glc\_D\_b - 3.4306 \cdot M\_o2\_c \\
& \quad\quad + 0.8033 \cdot R\_RPI - 3.5964 \cdot R\_SUCOAS\_FwRe).
\end{aligned}
\tag{1}
$$

See Table 1 for an explanation of all variables used in the function. Note that unlike the original biomass function (compare Table 4 in the Supplementary Materials), reaction rates as well as metabolite concentrations are involved.

Table 1. Variables occurring in the biomass proxy, see Equation (1). *Aerobic* represents a dichotomous variable which was added to distinguish between aerobic and anaerobic conditions; likewise for *Pair*. Prefixes: M – metabolite, R – reaction; suffixes: b – boundary condition, c – cytosol, $FwRe$ – combined rate of forward and backward direction of a reversible reaction.

short name	explanation
Aerobic	dichotomous variable
Pair	dichotomous variable
M_fru_b	Fructose
M_glc_D_b	D-Glucose
M_o2_b, *M_o2_c*	Oxygen
M_h_b	Hydrogen
R_GLUN	glutaminase
R_RPI	ribose-5-phosphate isomerase, forward reaction
R_SUCOAS_FwRe	succinyl-CoA synthetase (ADP-forming)

Validation of the proxy function was undertaken. The standard diagnostics were reviewed, which included but were not limited to the following.

(i) The adjusted r-squared value of 0.979 was very high. The adjusted r-squared being the preferred measure of explanatory power as it is more conservative than the r-squared value and has been adjusted for the number of predictors in the regression model.

(ii) The p-value for the F-statistic was a lot less than 0.1% (0.001), meaning that it is highly statistical significant and that there is strong evidence of a relationship between the dependent and independent variables.

(iii) All the p-values for the coefficients were statistically significant at the 0.1% (0.001) level, meaning that there is evidence that the coefficients are significant.

Finally, the reassuring results were obtained from 10-fold cross-validation. The 10 dashed lines in Fig. 5, which relate to the best fit lines for the 10 respective folds in cross-validation do not vary significantly and are parallel and close together, as would be expected in a good model. The overall mean square value, i.e. the mean squared difference between the predicted value and the actual value, is a commonly used diagnostic in cross-validation and is 0.0265 for this data.

Further analysis was undertaken to ascertain whether the regression model meets assumptions for linear regression in order to determine whether it can be used for inference in addition to prediction. Some modest violations were identified with regard to homoscedasticity and some collinearity was also identified, but it was demonstrated that this could be effectively addressed by removing two of the variables from the regression equation with only a modest drop in the adjusted r-squared value from 0.979 to 0.965.

5 Evaluation of key results

Preliminary data analysis generated a number of critical insights that helped to guide the approach towards the regression analysis.

(i) First, it was found that biomass production for the different anaerobic conditions followed what can be described as a 'zero inflated distribution' (in which a large portion of values were either zero or close to zero), whereas biomass production for aerobic conditions resembles a normal distribution, as illustrated in Fig. 3.

(ii) The large number of predictors (metabolite concentrations and reaction fluxes) had the potential to generate overwhelming complexity. Further investigation into these variables revealed that a large number of them where highly correlated with each other and this together with clustering analysis helped to reduce the number of dimensions and to highlight the risk of collinearity in the regression model.

(iii) The initial dataset with only 26 single conditions imposed restrictions on the scope of the regression modelling, as there were not enough conditions to incorporate all the key predictors without a risk of overfitting. This led to the generation of additional data for pairs of conditions. The benefits of obtaining this data were twofold, firstly it improved the regression model by allowing for the inclusion of more predictors, without the same risk of overfitting. Secondly, it led to some interesting insights around biomass values for pairs of conditions which will be discussed below.

Key insights from the analysis of biomass for pairs of conditions. Not only did the additional data on 156 pairs of conditions help to improve the predictive power of the regression model, but it led to some serendipitous findings. First, pairs of aerobic conditions always have biomass values that are between 1% and 7% larger than the sum of the two single conditions as illustrated in Table 2.

Secondly, it was also shown that acetaldehyde, which does not produce biomass anaerobically as a single condition, produced biomass when paired with a number of other conditions that do not produce biomass anaerobically as illustrated in Table 3.

Table 2. Comparing the sum of aerobic single conditions with pair of conditions.

Condition 1 Name	Biomass	Condition 2 Name	Biomass	Condition 1+2	Paired value	Biomass Total	Increase %
Ethanol	0.70	Glutamine	1.16	1.86	1.97	0.104	5.6%
Ethanol	0.70	Fumarate	0.79	1.49	1.58	0.097	6.5%
Ethanol	0.70	Malate	0.79	1.49	1.58	0.097	6.5%
Fructose	1.79	Glutamine	1.16	2.95	3.05	0.096	3.2%
Glucose	1.79	Glutamine	1.16	2.95	3.05	0.096	3.2%
Ethanol	0.70	Glutamate	1.24	1.94	2.04	0.096	4.9%
Glutamine	1.16	Lactate	0.74	1.90	2.00	0.092	4.9%
Ethanol	0.70	Auccinate	0.84	1.54	1.63	0.092	5.9%
Acetaldehyde	0.61	Glutamine	1.16	1.77	1.86	0.091	5.1%
Fructose	1.79	Glutamate	1.24	3.03	3.12	0.090	3.0%
Glucose	1.79	Glutamate	1.24	3.03	3.12	0.090	3.0%
Acetaldehyde	0.61	Fumarate	0.79	1.39	1.48	0.090	6.5%
Acetaldehyde	0.61	Malate	0.79	1.39	1.48	0.090	6.5%
Acetaldehyde	0.61	Glutamate	1.24	1.85	1.94	0.090	4.8%
Acetaldehyde	0.61	Auccinate	0.84	1.45	1.53	0.087	6.0%
Glutamate	1.24	Lactate	0.74	1.98	2.07	0.085	4.3%
Ethanol	0.70	Fructose	1.79	2.49	2.57	0.083	3.3%
Ethanol	0.70	Glucose	1.79	2.49	2.57	0.083	3.3%
Fructose	1.79	fumarate	0.79	2.58	2.66	0.083	3.2%
Fructose	1.79	Malate	0.79	2.58	2.66	0.083	3.2%

Table 3. FBA values for anaerobic paired conditions.

Acetaldehyde paired with	FBA value for paired condition
Fumarate	0.145
Malate	0.145
Lactate	0.117
2-oxoglutarate	0.068
Glutamate	0.045
Glutamine	0.040
All other conditions	< 0.01

Approach towards development of a proxy function to predict biomass using multiple linear regression. Two separate approaches were used in relation to the predictive modelling. The first was the traditional statistical approach of stepwise regression. The second was to use feature selection algorithms to select variables together with an automated process to iterate through all the different combinations of the variables. Interestingly, the stepwise regression model outperformed the model generated through the algorithmic approach. This scenario was unexpected, but analysis showed that the single most important factor in improving the predictive power of the regression model was the inclusion of interaction terms to reflect the combined effect of the predictors (metabolite concentrations and reaction fluxes) together with the dichotomous variable created to distinguish between aerobic and anaerobic conditions, that were not included in the automated algorithmic approach. The lesson here is that one should not overlook the importance of the preliminary data analysis in helping shape the approach toward predictive model building.

Elements of the automated approach involving automated feature selection and regression model building did however help to improve the final stepwise regression model, see Equation (1), with an adjusted r-squared of 0.98.
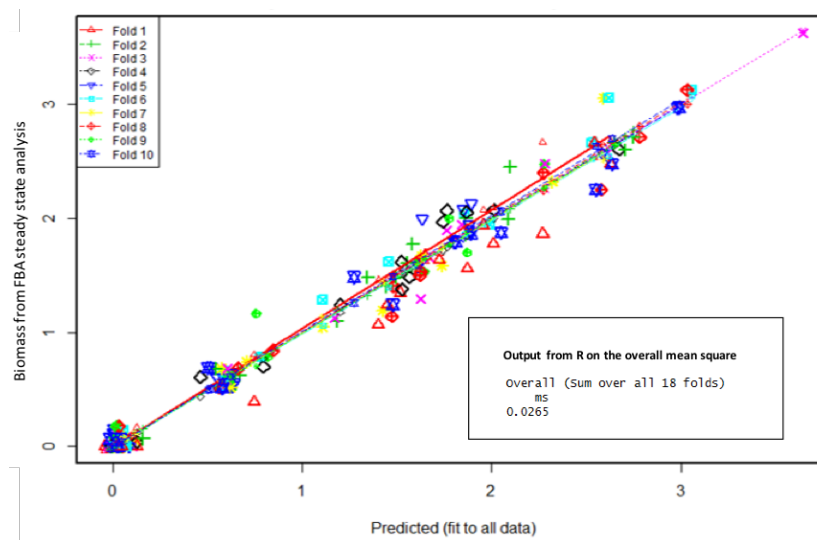

Fig. 5. Cross-validation output for the final regression model. Small symbols show predicted values; large symbols represent actuals. The 10 dashed lines relate to the best fit line for the respective folds.

Interpretation of the biomass proxy. The methodology that we employed to derive the biomass proxy – incorporating regression analysis and machine learning – was by its very nature designed to derive a robust and accurate proxy along the lines of Occam's razor, without overt regard to biological interpretation.

Our starting point was an FBA model, lacking appropriate kinetic data. Thus, to be able to demonstrate our approach, we assumed mass-action rates with uniform kinetic parameters for all reactions. Our workflow embodies a general approach which works for any kinetic parameters; their choice, however, may influence the final outcome of the derived proxy function. Also note that the result obtained is not unique, because there are many highly correlated variables — some were even perfectly correlated, given that they had Pearson correlation coefficients of 1. The representative for an equivalence class of pairwise highly correlated variables (above an appropriate threshold) is selected according to predictive power and collinearity.

In other words, our function given in Equation (1) is inherently not explanatory, but mimics the calculation flux value of the FBA biomass function (given in Table 4 in the Supplementary Materials). Moreover, the proxy function is not a pseudo-reaction in the way that the FBA one is, but is merely a function over a subset of the observables, both metabolite concentrations and reaction rates. It is for this reason that a mere syntactic comparison between the two is not meaningful, along the lines of comparing apples and pears, and it is the predictive power of the proxy which is of interest.

Reproducibility. Supplementary Materials can be found on our website `http://www-dssz.informatik.tu-cottbus.de/DSSZ/Software/Examples`, where we provide the original SBML model and its *Snoopy* version in ANDL format, which can be easily configured and simulated for the various growth conditions using the script provided. All the data analytics methods used are well recognised. Please also note, only public tools were used; thus all results presented are reproducible. Further data is also available in the form of additional tables and figures.

6   Conclusions

The research reported here describes a workflow to derive a dynamic biomass function which acts as a robust and accurate proxy for the FBA equivalent. The application of the method was illustrated for a reduced GEM for *E. coli*. Data generated by stochastic simulation for growth under a wide variety of conditions was used to develop a proxy function to predict biomass in the dynamic model, using data analytics techniques. This involved consolidating data for reactions and metabolites generated under dynamic simulation with gold standard target data for biomass generated under steady state analysis using a state of the art FBA solver.

The complexity generated by the large number of potential predictors (metabolites and reactions) was addressed through correlation and clustering analysis. In addition, the limited number of conditions in the initial dataset led to the need to generate more data using pairs of conditions. This not only improved the regression model by allowing for the inclusion of more predictors without the risk of overfitting, but led to a number of interesting insights regarding biomass for pairs of conditions. Namely, that pairs of aerobic conditions always have a biomass value that is between 1% and 7% larger than the sum of the two single conditions. In addition, it was shown that acetaldehyde, which does not produce biomass anaerobically as a single condition, produced biomass when paired with a number of other conditions that do

not produce biomass anaerobically. These findings were faithfully reproduced in our dynamic proxy function.

Our workflow operates with any sets of kinetic data [7], and the biomass proxy results may be refined as more precise kinetic parameters become available.

Outlook. In further work we want to semi-automate the workflow developed and apply it to unreduced GEMs. Because of our unexpected finding that regression outperforms machine learning, we also plan to modify the algorithmic machine learning approach in such a way that we incorporate interactive terms that combine the effect of the predictors together with the dichotomous variables which distinguish between aerobic and anaerobic environments.

We also intend to investigate whether the biomass proxy function will correctly predict biomass in transient states before a steady state is reached. This would permit us to explore the effects of dynamic changes in growth conditions – for example during the process whereby a carbon source is gradually exhausted, or the availability of carbon sources in the environment fluctuates up and down, or the oxygen available is gradually used up. It would also be interesting to investigate how an active biomass function could be included in a dynamic model in order to retain its recycling properties as well as the draining of biomass components, possibly by decomposition into parts, or by employing non-mass action kinetics. This would enable us to investigate the dynamic evolution of the biomass function, i.e. to analyse more realistically at what time points the system becomes biologically non-viable under certain conditions. This would allow us to address the interpretation of the proxy function compared with the re-engineered biomass function in the context of the simulation of dynamic GEMs, i.e. whether the proxy function derived by regression analysis and machine learning can be not only predictive but also explanatory with regard to the behaviour of large-scale metabolism.

References

1. Baart, G.J., Martens, D.E.: Genome-scale metabolic models: reconstruction and analysis. In: Neisseria meningitidis, pp. 107–126. Springer (2012)
2. Babyak, M.A.: What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine **66**(3), 411–421 (2004)
3. Chavent, M., Kuentz, V., Liquet, B., Saracco, L.: Clustofvar: an R package for the clustering of variables. arXiv preprint arXiv:1112.0295 (2011)
4. Erdrich, P., Steuer, R., Klamt, S.: An algorithm for the reduction of genome-scale metabolic network models to meaningful core models. BMC systems biology **9**(1) (2015)
5. Fox, J., Friendly, G.G., Graves, S., Heiberger, R., Monette, G., Nilsson, H., Ripley, B., Weisberg, S., Fox, M.J., Suggests, M.: The car package. R Foundation for Statistical Computing (2007)
6. Gilbert, D., Fuß, H., Gu, X., Orton, R., Robinson, S., Vyshemirsky, V., Kurth, M., Downes, C., Dubitzky, W.: Computational methodologies for modelling, analysis and simulation of signalling networks. Briefings in Bioinformatics **7**(4), 339–353 (2006)

7. Gilbert, D., Heiner, M., Jayaweera, Y., Rohr, C.: Towards dynamic genome-scale models. Briefings in Bioinformatics (2017)
8. Gilbert, D., Heiner, M., Lehrack, S.: A unifying framework for modelling and analysing biochemical pathways using Petri nets. In: Proc. CMSB, pp. 200–216. LNCS/LNBI 4695, Springer (2007)
9. Hädicke, O., Klamt, S.: Ecolicore2: a reference network model of the central metabolism of Escherichia coli and relationships to its genome-scale parent model. Scientific Reports **7**, 39647 (2017)
10. Heiner, M., Herajy, M., Liu, F., Rohr, C., Schwarick, M.: Snoopy – a unifying Petri net tool. In: Proc. Petri Nets, *LNCS*, vol. 7347, pp. 398–407. Springer (2012)
11. Heiner, M., Rohr, C., Schwarick, M.: MARCIE - Model checking And Reachability analysis done effiCIEntly. In: Proc. Petri Nets, *LNCS*, vol. 7927, pp. 389–399. Springer (2013)
12. Heiner, M., Schwarick, M., Wegener, J.: Charlie – an extensible Petri net analysis tool. In: Proc. Petri Nets, *LNCS*, vol. 9115, pp. 200–211. Springer (2015)
13. King, Z., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J., Ebrahim, A., Palsson, B., Lewis, N.: BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic acids research **44**(D1), D515–D522 (2016)
14. Knofczynski, G., Hadavas, P., Hoffman, L.: Effects of implementing projects in an elementary statistics class. Journal of Mathematical Sciences & Mathematics Education **2**(2) (2007)
15. Kursa, M.B., Rudnicki, W.R., et al.: Feature selection with the Boruta package. J Stat Softw **36**(11), 1–13 (2010)
16. Mahadevan, R., Edwards, J.S., Doyle III, F.J.: Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical journal **83**(3), 1331–1340 (2002)
17. Mamiatis, T., Fritsch, E., Sambrook, J., Engel, J.: Molecular cloning–A laboratory manual. New York: Cold Spring Harbor Laboratory. 1982, 545 S. (1985)
18. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis, vol. 821. John Wiley & Sons (2012)
19. O'Brien, R.M.: A caution regarding rules of thumb for variance inflation factors. Quality & Quantity **41**(5), 673–690 (2007)
20. Orth, J.: Systems biology analysis of Escherichia coli for discovery and metabolic engineering. Ph.D. thesis, University Of California, San Diego (2012)
21. Orth, J., Fleming, R., Palsson, B.: Reconstruction and use of microbial metabolic networks: the core Escherichia coli metabolic model as an educational guide. EcoSal Plus **4**(1) (2010)
22. Palsson, B.: Systems biology: constraint-based reconstruction and analysis. Cambridge University Press (2015)
23. Rohr, C.: Simulative analysis of coloured extended stochastic Petri nets. Ph.D. thesis, BTU Cottbus, Dep. of CS (2017)
24. Rohr, C.: Discrete-Time Leap Method For Stochastic Simulation. Fundamenta Informaticae **160**(1-2), 181–198 (2018). DOI 10.3233/FI-2018-1680
25. Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., Feist, A., Zielinski, D., Bordbar, A., Lewis, N., Rahmanian, S., et al.: Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2. 0. Nature protocols **6**(9), 1290–1307 (2011). DOI 10.1038/nprot.2011.308
26. Smallbone, K., Mendes, P.: Large-scale metabolic models: From reconstruction to differential equations. Industrial Biotechnology **9**(4), 179–184 (2013)
27. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)