

## Article

# Construction and Analysis of an Enzyme-constrained Metabolic Model of *Corynebacterium glutamicum*

Jinhui Niu <sup>1,2,3</sup>, Zhitao Mao <sup>2,3,\*</sup>, Yufeng Mao <sup>2,3</sup>, Ke Wu <sup>2,3</sup>, Zhenkun Shi <sup>2,3</sup>, Qianqian Yuan <sup>2,3</sup>, Jingyi Cai <sup>2,3</sup> and Hongwu Ma <sup>1,2,3,\*</sup>

<sup>1</sup> School of Life Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China

<sup>2</sup> Biodesign Center, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China

<sup>3</sup> National Technology Innovation Center of Synthetic Biology, Tianjin, China

\* Correspondence: mao\_zt@tib.cas.cn; ma\_hw@tib.cas.cn

**Abstract:** Genome-scale metabolic model (GEM) is a powerful tool for interpreting and predicting cellular phenotypes under various environmental and genetic perturbations. However, GEM only considers stoichiometric constraints, and the simulated growth and product yield values will show a monotonic linear increase with increasing substrate uptake rate, which deviates from the experimentally measured values. Recently, the integration of enzymatic constraints into stoichiometry-based GEMs was proven to be effective in making novel discoveries and predicting new engineering targets. Here we present the first genome-scale enzyme-constrained model (eciCW773) for *Corynebacterium glutamicum* reconstructed by integrating enzyme kinetic data from various sources using ECMpy workflow based on the high-quality GEM of *C. glutamicum* (obtained by modifying the iCW773 model). The enzyme-constrained model improved the prediction of phenotypes and simulated overflow metabolism, while also recapitulating the trade-off between biomass yield and enzyme usage efficiency. Finally, we used eciCW773 to identify several gene modification targets for L-lysine production, most of which agree with previously reported genes. This study shows that incorporating enzyme kinetic information into the GEM enhances the cellular phenotypes prediction of *C. glutamicum*, which can help identify key enzymes and thus provide reliable guidance for metabolic engineering.

**Keywords:** enzyme-constrained model; *Corynebacterium glutamicum*; metabolic engineering

## 1. Introduction

*Corynebacterium glutamicum* is widely known as an excellent producer of amino acids [1]. Recent advances in metabolic engineering and synthetic biology have expanded the scope of chemicals that can be produced from *C. glutamicum*, but it remains difficult to synthesize these compounds on an industrially relevant scale [2]. Genome-scale metabolic models (GEMs) are a proven tool for the prediction of cellular behaviour and the discovery of potential engineering targets [3]. Several GEMs of *C. glutamicum* have been developed (Fig. S1), and used to guide the production of high-value compounds such as glutaric acid [4], anthocyanins [5] and L-glutamate family amino acids [6]. The most widely used model of *C. glutamicum* is iCW773, constructed in 2017 [7], which accurately predicts the growth of cells cultured under different conditions. Although the quality of *C. glutamicum* models has improved in the last decade, they mostly only consider reaction stoichiometries and do not accurately depict the real situation inside the cell [8]. For example, metabolic overflow is a phenomenon in which incomplete oxidation of glucose to ethanol or acetate occurs in microorganisms in the presence of sufficient substrate, which has been recognized for a long time and frequently occurs in microbial cultures [9]. It has been shown that the limitation of intracellular protein resources is the cause of the metabolic overflow phenomenon [10], which cannot be properly explained by models only considering reaction stoichiometries.

With the accumulation of enzyme kinetic data and the availability of high-throughput omics data, it has become possible to integrate these data into models to add constraints on individual reactions or aggregate constraints on enzyme resources [11, 12]. In 2007, Zhang et al. constructed the FBAwMC model by introducing both a crowding coefficient and cell volume constraint to limit the space occupied by enzymes [13]. Subsequently, the researchers developed other protein resource integration methods [14], which referred to as enzyme-constrained genome-scale models (ecGEMs), including MOMENT [15], GECKO [16], AutoPACMEN [17] and ECMpy [18]. The GECKO method was reported in 2017 and was applied to construct an enzyme-constrained model of *Saccharomyces cerevisiae* by adding many pseudo-metabolites to represent enzyme utilization and including  $k_{cat}$  values to expand the stoichiometric matrix. Notably, this enzyme-constrained model could accurately predict several metabolic phenotypes. Recently, Domenzain et al. upgrade GECKO to 2.0 to enhance models with enzyme and proteomics constraints for any organism with a compatible GEM reconstruction, which also proposed an automated calibration process for enzyme kinetic parameters and developed conventional algorithms based on ecGEMs (e.g., flux variability analysis) [19]. Inspired by MOMENT and GECKO, in 2020 Bekiaris et al. proposed a simpler method for enzyme-constrained model construction, called AutoPACMEN [17], which could automatically download kinetic parameters of enzymes from the BRENDA [20] and SABIO-RK [21] databases. It was used to construct an enzyme-constrained model of *Escherichia coli* which only introduced one pseudo-reaction and pseudo-metabolite into the stoichiometric matrix. Different from GECKO and AutoPACMEN, ECMpy simply adds a constraint on the total amount of enzymes and does not require modification of the stoichiometric matrix, while providing higher prediction accuracy for the simulation of the *E. coli* growth rate [18]. Currently, ecGEMs have been constructed for many species, such as *E. coli* [18, 22], *S. cerevisiae* [16], *Yarrowia lipolytica* [19], *Aspergillus niger* [23], and *Bacillus subtilis* [24].

In the enzyme-constrained models,  $k_{cat}$  and molecular weight (MW) of an enzyme set constraints on the fluxes of the reactions catalyzed by that enzyme. Previous studies have made efforts to automatically acquire  $k_{cat}$  values from databases and fill missing values using methods like machine learning [25]. In contrast, few studies paid attention to molecular weight. It may seem straightforward to obtain the MW of a protein from databases like UniProt. However, the MW values from these databases are for monomers and many enzymes contain two or more subunits. For example, 6-phosphogluconate dehydrogenase encoded by gene *Cgl1452* is a homodimer and therefore MW is 105.2 kDa instead of 52.6.2 kDa for the monomer. There are also many enzymes consisting of subunits encoded by different genes which are represented as 'and' GPR relationship in GEMs. However, what is missing in the GPR relationships is the number of each subunit in the protein complex. For example, Succinyl-CoA synthetase is a heterotetramer containing two alpha subunits (encoded by *Cgl2565* in *C. glutamicum* with an MW of 30.26 kDa) and two beta subunits (encoded by *Cgl2566* with an MW of 41.76 kDa). Therefore, the MW of this enzyme complex should be 144.04 kDa ( $2 \times 30.26 + 2 \times 41.76$ ) instead of 72.02 kDa. Such quantitative information on enzyme subunit composition is difficult to obtain from databases and often missing in the published GEMs, leading to incorrect MW values which affect the prediction accuracy of enzyme-constrained models.

In this study, we first systematically corrected the GPR relationships in the iCW773 model based on GPRuler [26] and protein homology similarity comparisons, and extended GPRuler to allow access to the quantitative subunit composition of each protein in the model. Then, we gathered the enzyme kinetics data of *C. glutamicum* using AutoPACMEN and constructed the enzyme-constrained model eciCW773 based on the ECMpy workflow. We further comprehensively evaluated eciCW773 and confirmed that it had a better prediction accuracy of phenotypes than iCW773 and could simulate a variety of biological phenomena. Finally, we applied eciCW773 to metabolic engineering and discovered potential targets for increasing the production of L-lysine.

## 2. Materials and Methods

### 2.1. Model Calibration

We obtained the iCW773 model, which has 773 genes, 1207 reactions, and 950 metabolites, from the supplemental data of Zhang et al. [9], and converted it to XML format. To meet the requirements of the AutoPACMEN and ECMpy processes for metabolic network format input, we modified the gene, reaction and metabolite information in the model as follows:

(1) Metabolite correction: '(e)' to '\_e', '-D' to '\_\_D', '-L' to '\_\_L', '-R' to '\_\_R' and other '-' to '\_'.

(2) Reaction correction: '-' to '\_' in reactions beginning with 'EX' and '-' to '\_' in other reactions.

(3) Adding UniProt ID information to the annotation, which is the basis for obtaining kinetic parameters.

### 2.2. Correction of GPR relationship

We found some errors in the GPR relationships in iCW773 during the analysis and corrected these errors using two methods. First, modified GPRuler was used to identify more "and" relationships. GPRuler identify the 'and' relationships based on the protein complex information extracted from databases such as UniProt [27] and Complex Portal [28]. However, the original terms ('subunit' and 'chain') used to identify complexes in GPRuler were extracted based on human and yeast protein description information and did not cover all *C. glutamicum* protein complexes. For example, the protein name of P06557 (encoding by *Cgl3029*) is Anthranilate synthase component 1, which will not be identified as a subunit forming an 'and' relationship with another subunit using the original terms. Therefore, we updated the terms by carefully checking the words used in UniProt to describe possible protein complex formation (e.g., 'component', 'binding protein', and 'assembly factor', etc. see Table S1 for the full list) to obtain more 'and' relationships in *C. glutamicum*. We also simplified the GPRuler process by parsing UniProt data directly to obtain the corresponding GPR relationships without running the processes for gene identification, reaction identification, and gene filter.

We also observed that some 'and' relationships in iCW773 were not identified by GPRuler and could be wrong. To address this, we developed a semi-automated process based on protein similarity to determine the correct relationship. We calculated the protein sequence similarity for the remaining 'and' relationships in iCW773 and revised the relationship to 'or' if similarity exists between protein sequences as proteins with similarity are more likely to be isoenzymes rather than forming protein complexes. We then manually checked the gene annotation information in databases (BioCyc [29] and KEGG [30]) to ensure the correction is right.

### 2.3. Acquisition of quantitative subunit composition

As discussed in the Introduction section, quantitative subunit information of an enzyme is essential to correctly determine its MW but is missing in the GPR relationships in the models. We have manually collected the subunit number of each protein in our previous approach for constructing the enzyme-constrained model of *E. coli* eciML1515 [18]. Here we used a new automatic method to acquire the quantitative subunit information by extending GPRuler to resolve the subunit number of a protein based on information in the 'Interaction information' section in UniProt. For example, Q8NMK2 is described in UniProt as 'Homodimer', so its subunit number is 2. We created a word list to parse the description information (e.g., Homodimer; Heterotrimer; Tetramer of two alpha and two beta chains) and translated them into corresponding subunit numbers (Table S2).

### 2.4. Construction of eciCW773

After model correction, we split the reversible reactions in the model into two irreversible reactions because of different  $k_{cat}$  values for different catalytic directions. We also

split the isozyme-catalyzed reaction into multiple individual enzyme-catalyzed reactions. Then the molecular weight for each protein monomer was obtained automatically from Uniprot using AutoPACMEN and combined with the protein subunit composition data obtained in section 2.3 to calculate the molecular weight for each enzyme using equation  $MW = \sum_{j=1}^m N_j * MW_j$ , where  $m$  is the number of different subunits in the enzyme complex and  $N_j$  is the number of  $j$ th subunits in the complex. We further obtained the kinetic parameters of the enzymes mainly from BRENDA and SABIO-RK, using AutoPACMEN. In addition, we calculated the mass fraction of total cellular enzymes (Eq. 4) using published RNA-Seq data of wild-type *C. glutamicum* ATCC 13032 grown on glucose [31, 32]. Finally, we used the ECMpy process to construct eciCW773 (see Fig. 1 for details), which can be mathematically represented as follows:

$$Z = \max\{C^T * v\} \quad (1)$$

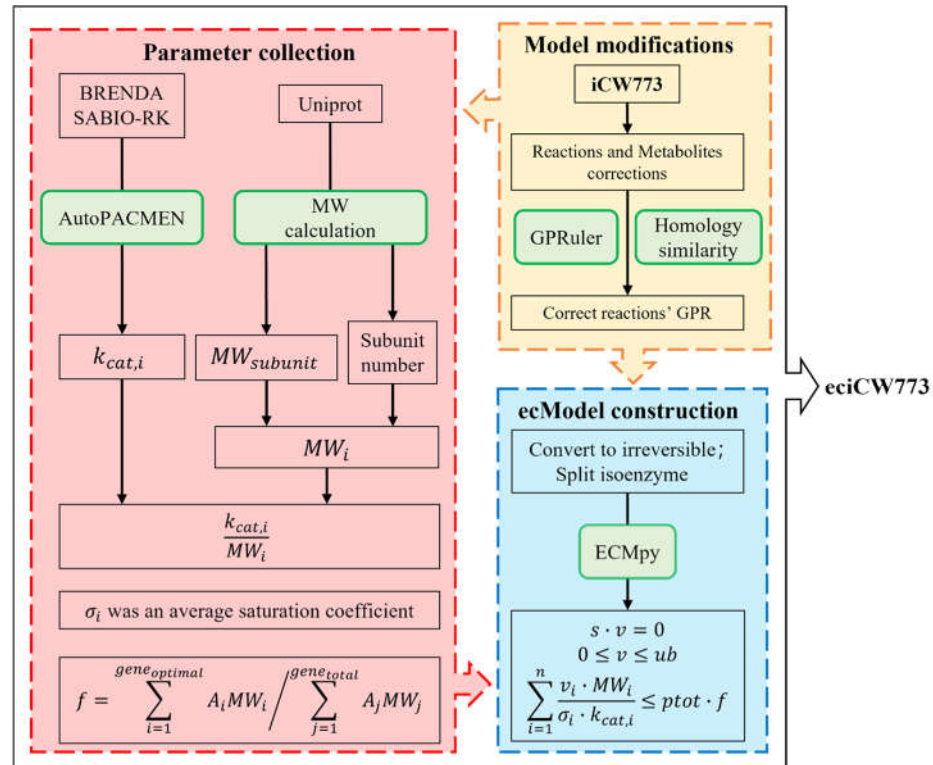
$$S * v = 0 \quad (2)$$

$$lb \leq v \leq ub \quad (3)$$

$$f = \sum_{i=1}^{n(gene_{model})} A_i MW_i / \sum_{j=1}^{n(gene_{total})} A_j MW_j \quad (4)$$

$$\sum_{i=1}^n \frac{v_i * MW_i}{\sigma_i * k_{cat,i}} \leq P_{total} * f \quad (5)$$

where  $C^T$  is the transposed vector of the integer coefficient of each flux in the objective function  $Z$ ;  $S$  is the stoichiometric matrix;  $lb$  and  $ub$  are the lower and upper bounds of the reaction fluxes in the system, respectively;  $k_{cat,i}$  is the turnover number of enzymes that catalyze reaction  $i$ ;  $MW_i$  denotes the molecular weight of enzyme  $i$ ;  $\sigma_i$  is the saturation coefficient for enzyme  $i$ , whereby we use an average value of 0.5 for all the enzymes [18];  $P_{total}$  of 0.56 is the average protein content in most microbial cells [15];  $f$  is the total mass fraction of all cellular enzymes in our ecGEM.



**Figure 1.** Workflow for the construction of eciCW773.

### 2.5. Calibration of the Original $k_{cat}$ Values

Generally, the initial enzyme-constrained model was unable to accurately predict the experimental value of the maximal growth rate, requiring an adjustment of the original  $k_{cat}$  values, like GECKO, AutoPACMEN and ECMpy. Since no suitable  $^{13}\text{C}$  data were found, the correction of eciCW773 was performed using only method 1 proposed by ECMpy, which is based on enzyme usage. For enzymes that require calibration, the EC number was obtained and its corresponding  $k_{cat}$  value was substituted by the highest value in the BRENDA and SABIO-RK databases for the given enzyme class. This iterative correction process was continued until the experimental value or the predefined number of iterations was reached, as described in GECKO 2.0 [19].

### 2.6. Comparative flux variability analysis

We provided a fair comparison of flux variability range distributions between iCW773 and eciCW773 for a protein-limited regime (e.g.,  $0.479 \text{ h}^{-1}$ ), using the computational procedure of GECKO 2.0. For reactions containing isozymes, we used the maximal value of the corresponding flux variability range (Eq. 6). For reversible reactions from ecModel, the corresponding flux variability ranges were solved using Eq. 7.

$$FV_i = \max(v_{i,isoj}^{max} - v_{i,isoj}^{min}), j \in m \quad (6)$$

$$FV_i = (v_i^{max} - v_i^{min}) - (v_{i,REV}^{max} - v_{i,REV}^{min}) \quad (7)$$

### 2.7. Phenotype phase plane (PhPP) analysis

PhPP analysis is a powerful tool that provides a global view of how optimal growth rates are affected by changes in two environmental variables such as the carbon and oxygen uptake rate [33, 34]. We performed PhPP analysis on iCW773 and eciCW773 to predict the biomass-specific growth rates by varying the glucose and oxygen uptake rates, as described in the literature [23]. Therefore, we varied the exchange reaction fluxes of glucose



in the range of 0-10 mmol/gDCW/h and oxygen in the range of 0-10 mmol/gDCW/h, with the objective set to maximize the biomass production rate, after which the results were analyzed by parsimonious FBA (pFBA) calculations [35].

### 2.8. Simulation of overflow metabolism

We explored the overflow metabolism of *C. glutamicum* using eciCW773 by varying the substrate uptake rate from 1 to 6.3 mmol/gDCW/h (the model reaches its maximum growth rate at 0.479 mmol/gDCW/h) and solving for the pFBA to maximize the biomass. To observe the trade-off phenomenon in unrestrained growth, we set glucose as the carbon source, and varied the substrate uptake rate from 1 to 6.3 mmol/gDCW/h to obtain the trade-off between the biomass yield (Eq. 8) and enzyme usage efficiency (Eq. 9).  $E_{min}$  was calculated using the minimum enzyme amount algorithm of ECMpy [18].

$$\text{biomass yield} = \frac{v_{\text{biomass}}}{v_{\text{glucose}} * MW_{\text{glucose}}} \quad (8)$$

$$\text{enzyme usage efficiency} = \frac{v_{\text{biomass}}}{E_{min}} \quad (9)$$

### 2.9. Prediction of metabolic engineering targets

Enzyme cost can be used to identify key enzymes in the product synthesis pathway. For example, Ye et al. calculated the enzyme cost for L-lysine biosynthesis by fixing a low biomass growth rate (0.1 h<sup>-1</sup>) using the enzyme-constrained model of *E. coli* and improved L-lysine production by optimizing the expression of the proteins in the top 20 of the enzyme cost [22]. But this approach can only identify overexpressed targets in the pathway, we extended this approach to explore enhanced and weakened metabolic engineering targets. First, we determined the protein cost differences in reactions between two scenarios: high growth low product generation (HGLP, growth rate was set at 0.46 h<sup>-1</sup>) and low growth high product generation (LGHP, growth rate set at 0.1 h<sup>-1</sup>). Subsequently, we calculated the cost of each reaction (Eq. 10) for both pathways [18]. Finally, we calculated the fold changes of enzyme cost and those with a fold change greater than 1.5 were chosen as potential targets for metabolic engineering (Eq. 11-12).

$$\text{Enzyme cost}_i = \frac{v_i * MW_i}{\sigma_i * k_{cat,i}} \quad (10)$$

$$\text{Enhance target} = \left\{ \text{Enzyme} \left| \frac{\text{Enzyme cost}_{LGHP}}{\text{Enzyme cost}_{HGLP}} \geq 1.5 \right. \right\} \quad (11)$$

$$\text{Weaken target} = \left\{ \text{Enzyme} \left| \frac{\text{Enzyme cost}_{HGLP}}{\text{Enzyme cost}_{LGHP}} \geq 1.5 \right. \right\} \quad (12)$$

## 3. Results

### 3.1. GPR Correction of iCW773

There are 1207 reactions in the iCW773 model, 96 of which are "and" relationships. We obtained protein composition information for a total of 112 reactions using GPRuler, of which 24 had GPR relationships consistent with the model. The 88 reactions with inconsistent GPR relationships were manually checked and corrected using information from database like UniProt, Biocyc, KEGG and literature (Table S3). They can be classified into three categories. First, the GPR relationship in the model is correct, so there is no need to replace (14 reactions). Second, the prediction results of GPRuler are correct and can be used to replace the GPR relationship in the model directly (12 reactions). For example, the GPR relationship of succinyl-CoA synthetase is 'Cgl2565 or Cgl2566' in the model. However, in UniProt Succinyl-CoA synthetase is described as a heterotetramer containing two

alpha and two beta subunits. The beta subunit (Cgl2566) provides nucleotide specificity of the enzyme and binds the substrate succinate, while the binding sites for coenzyme A and phosphate are found in the alpha subunit (Cgl2565). Third, both the prediction results of GPRuler and the GPR relationship in the model are wrong and should be manually corrected using Uniprot, Biocyc, KEGG or literature (62 reactions). For example, the gene composition of the Adenosylcobalamin 5'-phosphate synthase (ADOCBLS) is given as 'Cgl0245 and Cgl2201' in the model, while from GPRuler the GPR relationship was obtained as 'Cgl0245 and Cgl0246'. Whereas Cgl0245 and Cgl0246 do form a protein complex, its function is Lipid II isoglutaminyl synthase (glutamine-hydrolyzing) rather than Adenosylcobalamin 5'-phosphate synthase. Therefore, we corrected the GPR relationship for this reaction to 'Cgl2201' as Cgl2201 is the true adenosylcobalamin-5'-phosphate synthase based on information from Uniprot.

In addition, we observed that 58 "and" relationships in the model were not identified by GPRuler. To determine the correctness of these relationships, we calculated the protein sequence similarity for all gene pairs which have 'and' relationships in the model. The results showed that no similarity between protein pairs in the 41 reactions (Table S4). However, the remaining 17 reactions with an 'and' relationship have high similarities (greater than 20%) between the corresponding proteins, which are more likely to be isozymes. For example, the GPR relationship for Methylisocitrate lyase (MCITL2) is 'Cgl0658 and Cgl0695', which has 81% sequence similarity. By further searching the KEGG database for verification, we found that Cgl0695 and Cgl0658 have the same protein name, catalyze the same reaction and map the same EC number but not any information implying they are subunits of a protein complex. They are more likely to be two isoenzymes and therefore the correct GPR relationship for Methylisocitrate lyase should be 'Cgl0658 or Cgl0695'. After manual investigation, we modified 'and' relationships to 'or' relationships for 17 reactions in the model (Table S4).

### 3.2. ecModel Calibration

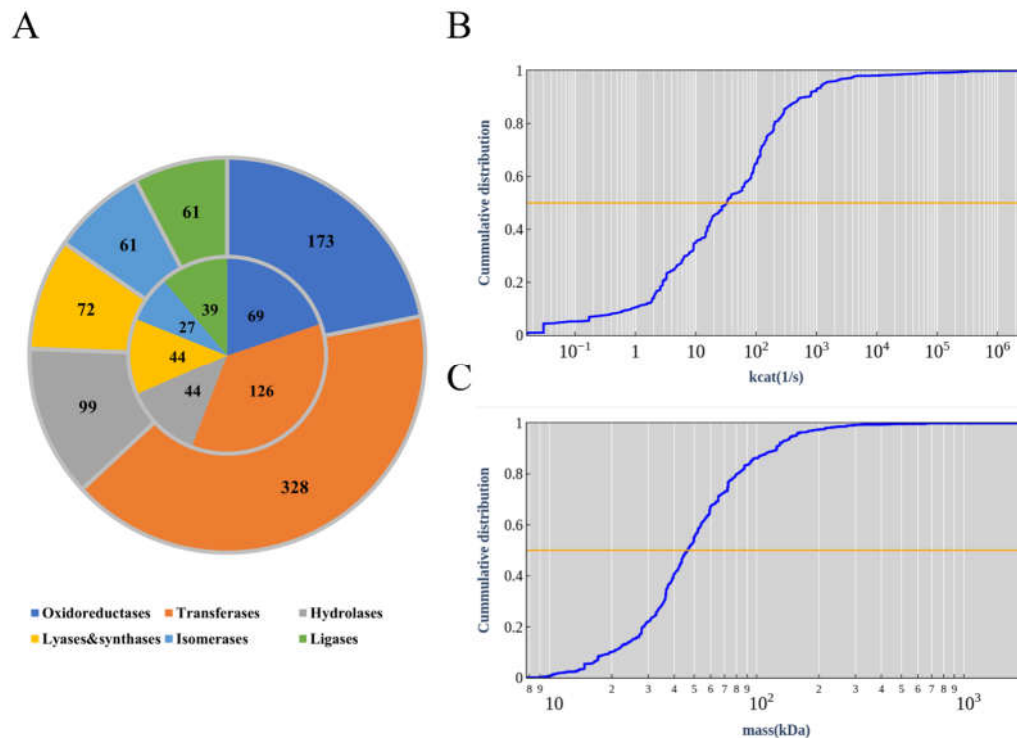
After that, we split each reversible reaction into independent forward and reverse reactions, including 333 reactions. The isozyme-catalyzed reactions were divided into multiple reactions (append num in reaction ID, e.g., ACOATA\_num1), and a total of 202 reactions were split into 571 reactions. After this step, the number of reactions in the model was expanded to 1850. The growth rate predicted by the initial ecModel at a glucose uptake rate of 5.05 mmol/gDCW/h was low compared to the experimental value (0.12 h<sup>-1</sup> vs. 0.45 h<sup>-1</sup>) [36]. We calibrated the initial ecModel based on the enzyme usage, and after 13 rounds of calibration and modification of 10 reactions, the simulated growth rate of 0.423 h<sup>-1</sup> was close to the experimental value (Table S5). Although the growth rate was corrected, the abnormal flux in the TCA cycle was mainly caused by the exceptionally large molecular weight of the pyruvate dehydrogenase complex consisting of 24 subunits, which was corrected using the maximal  $k_{cat}$  of this enzyme in the databases (Fig. S2). After this modification, the growth rate reached 0.454 h<sup>-1</sup>, which was consistent with the experimental value.

### 3.3. Basic information of eciCW773

There were 794 reactions in the modified ecModel with available  $k_{cat}$  data of enzymes with EC numbers, accounting for 42.92% of the total reactions, which were divided into 6 major categories, most of which were transferases (Fig. 2A, outer ring). These 794 reactions were catalyzed by a total of 349 enzymes (different EC numbers) (Fig. 2A, inner ring), and the  $k_{cat}$  values spanned 7 orders of magnitude with a median value of 33.3 s<sup>-1</sup> (Fig. 2B). There are 1107 enzymes present in the eciCW773, while non-monomeric enzymes occupy 30.81% (Fig. S3). The distribution of subunit composition in *C. glutamicum* is slightly different to that of *E. coli* [18], in that the proteins containing more than 2 subunits are generally low in *C. glutamicum* (Fig. S3). The biological reason for the difference is not clear.

and might be an interesting topic for further research. Finally, the molecular weights of the enzymes spanned a range from 7 to 2000 kDa (Fig. 2C).

Since the protein abundance of *C. glutamicum* could not be found in PAXdb [37], we obtained the abundance of each protein from the published data [31, 32]. An  $f$  value of 0.46 was obtained by calculating the mass fraction of enzymes that could be expressed by the genes in the model, which was a representation of a percentage of the total protein constraints in *C. glutamicum*.



**Figure 2.** Basic information of eciCW773. (A) Enzyme classification. The outer ring shows the 646 reactions with available enzyme kinetic data, which were divided into six categories. The inner ring shows the 304 different enzymes that catalyze these reactions according to the EC numbers, which can also be divided into six categories. (B) Cumulative distribution of  $k_{cat}$  values. (C) Cumulative distribution of molecular weights.

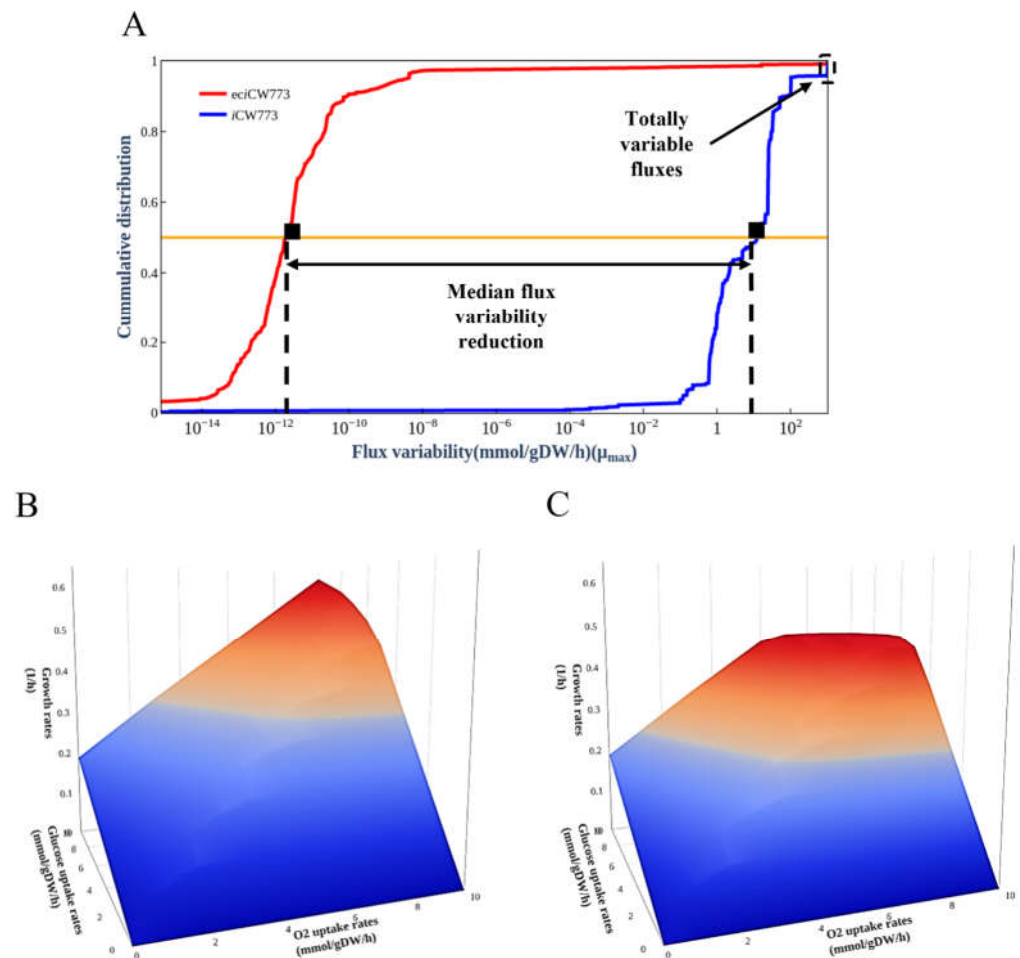
### 3.4. eciCW773 reduces the solution space

A major challenge for GEMs is how to obtain a biologically meaningful flux distribution, since there are alternate optimal solutions in which the same maximal objective can be achieved through different flux distributions [38]. This limitation can be overcome by integrating experimentally measured exchange fluxes as constraints [19]. Previous studies have demonstrated that enzyme constraint models can significantly reduce the solution space of fluxes [18, 19]. We compared the cumulative distributions of the flux variability ranges of iCW773 and eciCW773, which revealed that the median flux variability range at a high growth rate is significantly reduced by 10 orders of magnitude after introducing enzyme constraints (Fig. 3A). The cumulative distribution also showed a decrease in the number of reactions with completely variable fluxes, which may represent undesirable futile cycles caused by a lack of information regarding their thermodynamic or enzymatic costs [19]. At high growth rates, completely variable fluxes accounted for 4% of the active reactions in iCW773, in contrast to the complete absence of this extreme range of variability in eciCW773.

With the subsequent increase of carbon source and oxygen fluxes, PhPP analysis showed that the growth rate of iCW773 increased linearly with increasing carbon source consumption, which is inconsistent with the experimental observations, while eciCW773 significantly reduced the solution space (Fig. 3BC). All these results demonstrate that



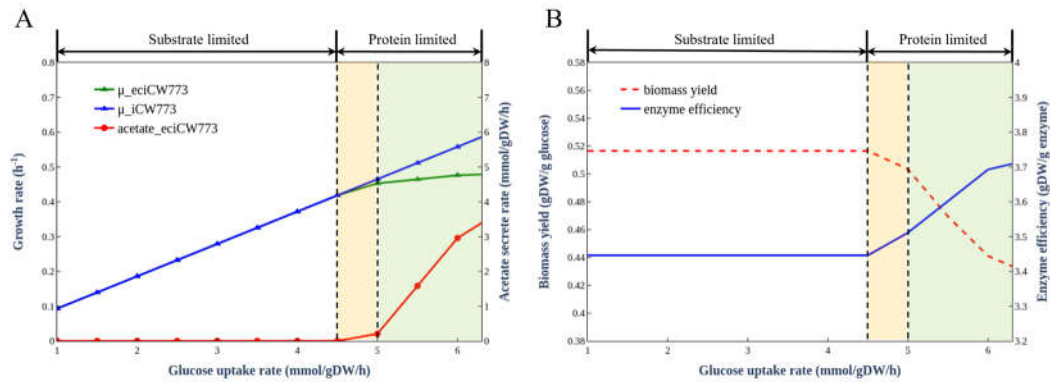
incorporating more information and constraints into a GEM can improve the predictive accuracy of the model and enable it to simulate a more realistic cellular phenotype.



**Figure 3.** The solution space of iCW773 and eciCW773. (A) Cumulative distribution of flux variability of eciCW773 and iCW773 at high growth rates. Growth rates at various glucose and oxygen uptake rates were simulated using iCW773 (B) and eciCW773 (C).

### 3.5. Simulation of overflow metabolism

In previous studies, ecModels were used to simulate overflow metabolism in *S. cerevisiae* [16] and *E. coli* [18]. As shown in Fig. 4A, eciCW773 was also able to precisely simulate the overflow metabolism phenomenon at a glucose uptake rate of 4.5 mmol/gDCW/h, which could not be reproduced in iCW773. When overflow metabolism occurs, the microorganism must activate a fermentation pathway with low energy production efficiency but high enzyme efficiency to maintain growth, and this pathway will cause a portion of the substrate to be converted into by-products [39], resulting in carbon loss and a rapid decrease of the biomass yield, illustrating a trade-off between enzyme efficiency and growth rate (Fig. 4B). This metabolic process can be divided into a substrate-limited stage, overflow switching stage, and overflow stage. In the first stage, the glucose uptake rate is low (less than 4.5 mmol/gDCW/h) and has a linear relationship with the growth rate, which is consistent with iCW773 (Fig. 4A). When the substrate supply gradually increased (between 4.5 and 5 mmol/gDCW/h), the cell growth is limited, and metabolism is switched to a more enzymatically efficient pathway. Finally, when the overflow metabolism phenomenon occurs (greater than 5 mmol/gDCW/h), the by-product pathway of acetate production switches to higher enzymatic efficiency, consuming more substrates and resulting in lower biomass yield, which is consistent with empirical models of microbial growth [40].



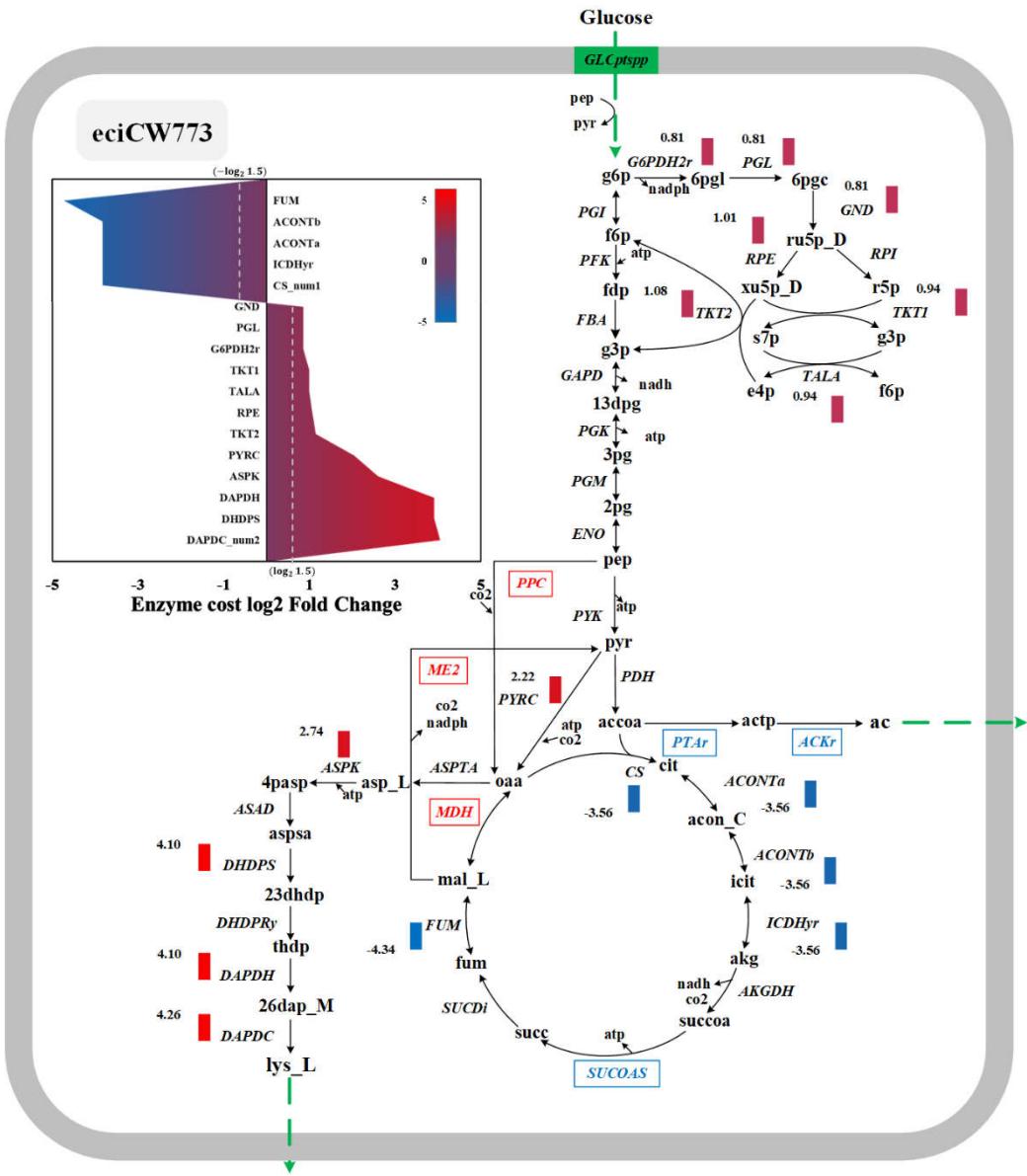
**Figure 4.** Simulation of overflow metabolism. (A) Comparison of in silico overflow metabolism between *iCW773* and *eciCW773*. (B) Trade-off phenomenon simulated by *eciCW773*.

### 3.6. Exploration of the targets based on enzyme cost

We compared the pathway characteristics at LGHP and HGLP to analyze the differences in the product synthesis pathways and found 23 reactions in which the change of enzyme cost was higher (Fig. 5). We first noticed the pathway changes in the L-lysine synthesis pathway between the two conditions, some reactions (e.g., ME2, PPC and MDH) were more favorable for L-lysine production (with increased flux) and some others (e.g., PTAr, ACKr and SUCOAS) are in contrast (Fig. 5 red and blue boxes). For example, malate dehydrogenase (ME2) catalyzes the conversion of malate to pyruvate to regenerate pyruvate depleted by pyruvate carboxylase (PYRC)[41]. PYRC catalyzed the conversion of pyruvate to oxaloacetate, a known precursor of L-lysine synthesis [40]. ME2, MDH and PYRC, therefore, form a cycle for NADPH regeneration, which can provide more NADPH for the L-lysine synthesis pathway. What's more, the pyruvate-oxaloacetate (OAA) supply has been considered a bottleneck for L-lysine production, while overexpression and point mutation of phosphoenolpyruvate carboxylase (PPC) has been applied to increase OAA availability [42-44].

At the same time, some reactions were more amenable for high-level growth and should be downregulated for L-lysine production. For example, decreased flux via succinyl-CoA synthetase (SUCOAS) was reported to increase the L-lysine yield and maintain optimal cell growth at the same time [45]. In addition, as we mentioned above, when enzyme constraints occurred, the acetate overflow phenomenon can be captured by *eciCW773*. It was reported that the production of organic acids such as acetate or lactate may reduce the yield of biological products [39]. Thus, less flux toward acetate overflow reactions (PTAr and ACKr) may result in more L-lysine production.

After analyzing these pathway changes, we further analyzed the enzyme costs of the L-lysine synthetic pathway calculated from the enzyme-constrained model. To explore the variability of enzyme costs, we calculated the log<sub>2</sub> fold changes of enzyme costs between LGHP and HGLP as shown in Fig. 5. We found that the upregulated values were mainly focused on the pathway of L-lysine synthesis from aspartate (DAPDC, DAPDH, DHSPS, etc.) and the pentose phosphate (PP) pathway (TKT1, TALA, GND, etc.), which was in agreement with the literature [41, 46, 47]. In addition, lower fluxes in TCA may lead to more fluxes for lysine production [48-50]. For example, Jan et al. showed that reduced citrate synthase (CS) activity leads to a strong accumulation of L-lysine [50].



**Figure 5.** Simulation of differential pathways under high and low L-lysine production conditions. The red box indicates reactions that are more beneficial for L-lysine production, and the blue box indicates reactions that are more beneficial for growth. Red bars indicate enhanced targets for L-lysine production in *C. glutamicum*. Blue bars indicate weak targets that may increase L-lysine production in *C. glutamicum*.

#### 4. Discussion

We constructed the enzyme-constrained model eciCW773 based on the iCW773 model of *C. glutamicum* using the upgraded workflow (Fig. 1). First, we updated the workflow of ECMpy to automatically acquire  $k_{cat}$  values from databases and fill missing values using AutoPACMEN. In the construction of the ecModel of *E. coli* using ECMpy, we have emphasized the impact of the subunit composition of proteins on the accuracy of model simulations, but it was done by manually correcting GPR relationships and collecting the subunit number of each protein [18]. In this study, we achieved the semi-automated correction of GPR relationships using GPRuler and protein homology similarity, and the automated acquisition of the quantitative subunit composition data based on UniProt. The GEMs and most of the ecGEMs neglect the quantitative subunit composition information of non-monomeric enzymes, so the introduction of the number of subunits in the GPR relationship might also be necessary for future model reconstruction.

The eciCW773 model exhibited a better simulation accuracy of strain behavior than the original iCW773. The growth rate increased linearly with the substrate uptake in the classical model, while the enzyme constraint narrowed the solution space, resulting in a model prediction that is closer to the real experimental observations. Due to growth rate limitations and enzyme resource constraints, eciCW773 was better able to predict the phenomenon of overflow metabolism, which was absent from the original model. This indicates that enzyme restriction is the primary driver behind enzyme protein redistribution and corresponding metabolic flux changes, which was consistent with previous studies [51]. Thus, our study not only corroborates the hypothesis that effective proteome reassignment is an important principle of metabolic regulation, but also shows how simple physicochemical constraints can be integrated into a GEM to improve its predictive power. Our model also made predictions based on the enzyme cost, offering a more intuitive reproduction of metabolic engineering strategies than the original GEM. According to the results of simulations using eciCW773, some potential targets in the glycolysis pathway should be considered for improving L-lysine production in the future, which may generate more energy and phosphoenolpyruvate, thus redistributing more flux toward the L-lysine synthesis pathway and cell growth.

Even though the eciCW773 model offers better phenotype predictions, it still has several shortcomings. First, the GEM of *C. glutamicum* is still evolving and there is still a lack of annotated information in some areas. Recently, Feierabend et al. reconstructed the iCGB21FR model using iEZ482 as a reference [52], which added seventeen different databases that are cross-referenced in the model's annotations and reached a high MEMOTE score (Fig. S1)[53]. This new model expands the number of reactions in the GEM of *C. glutamicum* to 1892. Of course, our process for constructing the enzyme-constrained model of *C. glutamicum* is generic, and another enzyme-constrained model can be constructed by simply replacing the initial model. Besides, the quality of the enzyme-constrained model depends on the input data on enzyme kinetics and intracellular protein abundance [16]. Unfortunately, there is little data on both for *C. glutamicum*. There are three ways to improve the coverage of enzyme kinetic parameters in the model: 1) directly populate unknown reactions with mean or median values of enzyme kinetic parameters from other reactions [16-18]; 2) expand the EC number annotation information of model reactions using EC number prediction tools [54, 55]; and 3) directly predict reactions with unknown parameters based on existing enzyme kinetic parameters via machine learning or deep learning approaches [25, 56]. In addition, the kinetic data are mainly sourced from the BRENDA and SABIO-RK databases, which mostly collect in vitro measurements that differ somewhat from the in vivo data. The improvement of parameter accuracy and coverage will increase the prediction efficiency and reduce the cost of result evaluation, which will help construct high-quality metabolic models of species such as *C. glutamicum*. Finally, although ecGEM has improved predictive power compared to traditional GEMs, biological systems are also subject to other constraints in addition to enzyme resources, and the construction of multi-constraint models (e.g., ETGEM [57] and ETFL [58]) will certainly provide new prospects for systems biology research.

## 5. Conclusions

In this study, we constructed an enzyme-constrained genome-scale metabolic model of *C. glutamicum* (eciCW773) by integrating various enzymatic parameters at the entire network level. The results show that constraints on enzyme resources can simulate strain growth limitations and recapitulate metabolic overflow phenomena, resulting in more realistic pathway predictions, which can be used to identify key enzymes to provide metabolic engineering targets for creating cell factories to produce valuable chemicals.

**Supplementary Materials:** Figure S1: Model development of *Corynebacterium glutamicum*; Figure S2: Pathway changes after changing the  $k_{cat}$  of PDH; Figure S3: Number of subunits per complex (only consider complex with two or more subunits) in *C. glutamicum*; Table S1: Keywords of protein complex formation in UniProt; Table S2: Correspondence between subunit descriptions and true



numbers in UniProt; Table S3: GPR modifications based on GPRuler results; Table S4: GPR Modifications by homology similarity; Table S5: Reactions for  $k_{cat}$  calibration.

**Author Contributions:** Conceptualization, J.N., Z.M. and H.M.; data curation, J.N. and Z.M.; formal analysis, J.N. and K.W.; funding acquisition, H.M.; methodology, J.N., Z.M. and H.M.; project administration, Z.M. and H.M.; software, J.N., Z.M. and K.W.; validation, Y.M., K.W., Z.S., Q.Y. and J.C.; writing original draft, J.N., Z.M., Y.M. and H.M.; writing—review and editing, J.N., Z.M., Y.M. and H.M. All authors have read and agreed to the publication of the final manuscript.

**Funding:** This research was funded by the National Key Research and Development Program of China (2018YFA0900300), Tianjin Synthetic Biotechnology Innovation Capacity Improvement Project (TSBICIP-PTJS-001), International Partnership Program of Chinese Academy of Sciences (153D31KYSB20170121), the National Natural Science Foundation of China (NSFC-32101186, 21908239).

**Data Availability Statement:** The scripts and datasets generated and/or analyzed in the study can be found at: <https://github.com/tibbdc/eciCW773>.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Wolf S, Becker J, Tsuge Y, Kawaguchi H, Kondo A, Marienhagen J, Bott M, Wendisch VF, Wittmann C: **Advances in metabolic engineering of *Corynebacterium glutamicum* to produce high-value active ingredients for food, feed, human health, and well-being.** *Essays in Biochemistry* 2021, **65**:197-212.
2. Becker J, Wittmann C: **Advanced biotechnology: Metabolically engineered cells for the bio-based production of chemicals and fuels, materials, and health-care products.** *Angewandte Chemie - International Edition* 2015, **54**:3328-3350.
3. Gudmundsson S, Nogales J: **Recent advances in model-assisted metabolic engineering.** *Current Opinion in Systems Biology* 2021, **28**:100392.
4. Han T, Kim GB, Lee SY: **Glutaric acid production by systems metabolic engineering of an L-lysine-overproducing *Corynebacterium glutamicum*.** *Proceedings of the National Academy of Sciences of the United States of America* 2020, **117**:30328-30334.
5. Zha J, Zang Y, Mattozzi M, Plassmeier J, Gupta M, Wu X, Clarkson S, Koffas MAG: **Metabolic engineering of *Corynebacterium glutamicum* for anthocyanin production.** *Microbial Cell Factories* 2018, **17**.
6. Sheng Q, Wu XY, Xu X, Tan X, Li Z, Zhang B: **Production of L-glutamate family amino acids in *Corynebacterium glutamicum*: Physiological mechanism, genetic modulation, and prospects.** *Synthetic and Systems Biotechnology* 2021, **6**:302-325.
7. Zhang Y, Cai J, Shang X, Wang B, Liu S, Chai X, Tan T, Zhang Y, Wen T: **A new genome-scale metabolic model of *Corynebacterium glutamicum* and its application.** *Biotechnology for Biofuels* 2017, **10**:1-16.
8. Orth JD, Thiele I, Palsson BO: **What is flux balance analysis?** *Nature Biotechnology* 2010, **28**:245-248.
9. Crabtree HG: **Observations on the carbohydrate metabolism of tumours.** *Biochemical Journal* 1929, **23**:536-545.
10. Berkhout J, Bosdriesz E, Nikerel E, Molenaar D, de Ridder D, Teusink B, Bruggeman FJ: **How biochemical constraints of cellular growth shape evolutionary adaptations in metabolism.** *Genetics* 2013, **194**:505-512.
11. Liu JK, O'Brien EJ, Lerman JA, Zengler K, Palsson BO, Feist AM: **Reconstruction and modeling protein translocation and compartmentalization in *Escherichia coli* at the genome-scale.** *BMC Systems Biology* 2014, **8**.
12. Nilsson A, Nielsen J, Palsson BO: **Metabolic Models of Protein Allocation Call for the Kinetome.** *Cell Systems* 2017, **5**:538-541.
13. Beg QK, Vazquez A, Ernst J, De Menezes MA, Bar-Joseph Z, Barabási AL, Oltvai ZN: **Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**:12663-12668.
14. Chen Y, Nielsen J: **Mathematical modeling of proteome constraints within metabolism.** *Current Opinion in Systems Biology*, vol. 25. pp. 50-56: Elsevier; 2021: 50-56.
15. Adadi R, Volkmer B, Milo R, Heinemann M, Shlomi T: **Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters.** *PLoS Computational Biology* 2012, **8**.
16. Sanchez BJ, Zhang C, Nilsson A, Lahtvee P-J, Kerkhoven EJ, Nielsen J: **Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints.** *Molecular Systems Biology* 2017, **13**.
17. Bekiaris PS, Klamt S: **Automatic construction of metabolic models with enzyme constraints.** *BMC Bioinformatics* 2020, **21**:1-13.
18. Mao Z, Zhao X, Yang X, Zhang P, Du J, Yuan Q, Ma H: **ECMPy, a Simplified Workflow for Constructing Enzymatic Constrained Metabolic Network Model.** *Biomolecules* 2022, Vol 12, Page 65 2022, **12**:65.
19. Domenzain I, Sánchez B, Anton M, Kerkhoven EJ, Millán-Oropeza A, Henry C, Siewers V, Morrissey JP, Sonnenschein N, Nielsen J: **Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0.** *Nature Communications* 2022, **13**:3766.
20. Jeske L, Placzek S, Schomburg I, Chang A, Schomburg D: **BRENDA in 2019: A European ELIXIR core data resource.** *Nucleic Acids Research* 2019, **47**:D542-D549.
21. Wittig U, Rey M, Weidemann A, Kania R, Müller W: **SABIO-RK: An updated resource for manually curated biochemical reaction kinetics.** *Nucleic Acids Research* 2018, **46**:D656-D660.



22. Ye C, Luo Q, Guo L, Gao C, Xu N, Zhang L, Liu L, Chen X: **Improving lysine production through construction of an *Escherichia coli* enzyme-constrained model.** *Biotechnology and Bioengineering* 2020, **117**:3533-3544.
23. Zhou J, Zhuang Y, Xia J: **Integration of enzyme constraints in a genome-scale metabolic model of *Aspergillus niger* improves phenotype predictions.** *Microbial Cell Factories* 2021, **20**.
24. Massaiu I, Pasotti L, Sonnenschein N, Rama E, Cavaletti M, Magni P, Calvio C, Herrgard MJ: **Integration of enzymatic data in *Bacillus subtilis* genome-scale metabolic model improves phenotype predictions and enables in silico design of poly--glutamic acid production strains.** *Microbial Cell Factories* 2019, **18**.
25. Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, Kerkhoven EJ, Nielsen J: **Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction.** *Nature Catalysis* 2022.
26. Di Filippo M, Damiani C, Pescini D: **GPRuler: Metabolic gene-protein-reaction rules automatic reconstruction.** *PLoS Computational Biology* 2021, **17**:e1009550.
27. UniProt C: **UniProt: the universal protein knowledgebase in 2021.** *Nucleic Acids Research* 2021, **49**:D480-D489.
28. Meldal BHM, Orchard S: **Searching and Extracting Data from the EMBL-EBI Complex Portal.** In *Protein Complex Assembly: Methods and Protocols*. Edited by Marsh JA. New York, NY: Springer New York; 2018: 377-390.
29. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, Keseler IM, Krummenacker M, Midford PE, Ong Q, et al: **The BioCyc collection of microbial genomes and metabolic pathways.** *Briefings in Bioinformatics* 2017, **20**:1085-1093.
30. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M: **KEGG: integrating viruses and cellular organisms.** *Nucleic Acids Research* 2020, **49**:D545-D551.
31. Walter T, Veldmann KH, Götter S, Busche T, Rückert C, Kashkooli AB, Paulus J, Cankar K, Wendisch VF: **Physiological response of *Corynebacterium glutamicum* to indole.** *Microorganisms* 2020, **8**:1-23.
32. Keppel M, Hünnefeld M, Filipchuk A, Viets U, Davoudi CF, Krüger A, Mack C, Pfeifer E, Polen T, Baumgart M, et al: **HrrSA orchestrates a systemic response to heme and determines prioritization of terminal cytochrome oxidase expression.** *Nucleic Acids Research* 2020, **48**:6547-6562.
33. Edwards JS, Ramakrishna R, Palsson BO: **Characterizing the metabolic phenotype: A phenotype phase plane analysis.** *Biotechnology and Bioengineering* 2002, **77**:27-36.
34. Duarte NC, Palsson BØ, Fu P: **Integrated analysis of metabolic phenotypes in *Saccharomyces cerevisiae*.** *BMC Genomics* 2004, **5**:63.
35. Lewis NE, Hixson KK, Conrad TM, Lerman JA, Charusanti P, Polpitiya AD, Adkins JN, Schramm G, Purvine SO, Lopez-Ferrer D, et al: **Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models.** *Molecular Systems Biology* 2010, **6**.
36. Wang Z, Liu J, Chen L, Zeng A-P, Solem C, Jensen PR: **Alterations in the transcription factors GntR1 and RamA enhance the growth and central metabolism of *Corynebacterium glutamicum*.** *Metabolic Engineering* 2018, **48**:1-12.
37. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C: **Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines.** *Proteomics* 2015, **15**:3163-3168.
38. Mahadevan R, Schilling CH: **The effects of alternate optimal solutions in constraint-based genome-scale metabolic models.** *Metabolic Engineering* 2003, **5**:264-276.
39. Péquignot C, Dussap CG, Pons A, Gros JB: **Intra- and extracellular concentrations of glutamate, lactate and acetate during growth of *Corynebacterium glutamicum* on different media.** *Journal of Industrial Microbiology and Biotechnology* 1997, **18**:312-318.
40. Koch AL: **Microbial Physiology and Ecology of Slow Growth.** *Microbiology and Molecular Biology Reviews* 1997, **61**:305-318.
41. Becker J, Zelder O, Häfner S, Schröder H, Wittmann C: **From zero to hero-Design-based systems metabolic engineering of *Corynebacterium glutamicum* for l-lysine production.** *Metabolic Engineering* 2011, **13**:159-168.
42. O'Regan M, Thierbach G, Bachmann B, Villeval D, Lepage P, Viret J-F, Lemoine Y: **Cloning and nucleotide sequence of the phosphoenolpyruvate carboxylase-coding gene of *Corynebacterium glutamicum* ATCC13032.** *Gene* 1989, **77**:237-251.
43. Cremer J, Eggeling L, Sahm H: **Control of the Lysine Biosynthesis Sequence in *Corynebacterium glutamicum* as Analyzed by Overexpression of the Individual Corresponding Genes.** *Applied and Environmental Microbiology* 1991, **57**:1746-1752.
44. Chen Z, Bommarreddy RR, Frank D, Rappert S, Zeng AP: **Deregulation of feedback inhibition of phosphoenolpyruvate carboxylase for improved lysine production in *Corynebacterium glutamicum*.** *Applied and Environmental Microbiology* 2014, **80**:1388-1393.
45. Shang X, Chai X, Lu X, Li Y, Zhang Y, Wang G, Zhang C, Liu S, Zhang Y, Ma J, Wen T: **Native promoters of *Corynebacterium glutamicum* and its application in l-lysine production.** *Biotechnology Letters* 2018, **40**:383-391.
46. Hoffmann SL, Kohlstedt M, Jungmann L, Hutter M, Poblete-Castro I, Becker J, Wittmann C: **Cascaded valorization of brown seaweed to produce l-lysine and value-added products using *Corynebacterium glutamicum* streamlined by systems metabolic engineering.** *Metabolic Engineering* 2021, **67**:293-307.
47. Xu J, Zhang J, Guo Y, Zhang W: **Genetically modifying aspartate aminotransferase and aspartate ammonia-lyase affects metabolite accumulation in l-lysine producing strain derived from *Corynebacterium glutamicum* ATCC13032.** *Journal of Molecular Catalysis B: Enzymatic* 2015, **113**:82-89.
48. Zhou L-B, Zeng A-P: **Exploring Lysine Riboswitch for Metabolic Flux Control and Improvement of l-Lysine Synthesis in *Corynebacterium glutamicum*.** *ACS Synthetic Biology* 2015, **4**:729-734.
49. Xu J, Han M, Zhang J, Guo Y, Qian H, Zhang W: **Improvement of l-lysine production combines with minimization of by-products synthesis in *Corynebacterium glutamicum*.** *Journal of Chemical Technology & Biotechnology* 2014, **89**:1924-1933.

- 
50. van Ooyen J, Noack S, Bott M, Reth A, Eggeling L: **Improved L-lysine production with *Corynebacterium glutamicum* and systemic insight into citrate synthase flux and activity.** *Biotechnology and Bioengineering* 2012, **109**:2070-2081.
  51. Nilsson A, Nielsen J: **Metabolic Trade-offs in Yeast are Caused by F1F0-ATP synthase.** *Scientific Reports* 2016, **6**.
  52. Feierabend M, Renz A, Zelle E, Nöh K, Wiechert W, Dräger A: **High-Quality Genome-Scale Reconstruction of *Corynebacterium glutamicum* ATCC 13032.** *Frontiers in Microbiology* 2021, **12**:3432.
  53. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, et al: **MEMOTE for standardized genome-scale metabolic model testing.** *Nature Biotechnology* 2020, **38**:272-276.
  54. Ryu JY, Kim HU, Lee SY: **Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers.** *Proceedings of the National Academy of Sciences of the United States of America* 2019, **116**:13996-14001.
  55. Shi Z, Yuan Q, Wang R, Li H, Liao X, Ma H: **ECRECer: Enzyme Commission Number Recommendation and Benchmarking based on Multiagent Dual-core Learning.** 2022.
  56. Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, Desouki AA, Lercher MJ, Palsson BO: **Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models.** *Nature Communications* 2018, **9**.
  57. Yang X, Mao Z, Zhao X, Wang R, Zhang P, Cai J, Xue C, Ma H: **Integrating thermodynamic and enzymatic constraints into genome-scale metabolic models.** *Metabolic Engineering* 2021, **67**:133-144.
  58. Salvy P, Hatzimanikatis V: **The ETFL formulation allows multi-omics integration in thermodynamics-compliant metabolism and expression models.** *Nature Communications* 2020, **11**.