PARROT: Prediction of enzyme abundances using protein-constrained metabolic models

Mauricio Alexander de Moura Ferreira[a], Wendel Batista da Silveira[a], Zoran Nikoloski [b,c,*]

* Corresponding author.

[a] *Department of Microbiology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil*

[b] *Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany*

[c] *Systems Biology and Mathematical Modelling, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany*

E-mail: zoran.nikoloski@uni-potsdam.de (Zoran Nikoloski, corresponding author)

Abstract

Protein allocation determines the activity of cellular pathways and affects growth across all organisms. Therefore, different experimental and machine learning approaches have been developed to quantify and predict protein abundances, respectively. Yet, despite advances in protein quantification, it remains challenging to predict condition-specific allocation of enzymes in metabolic networks. Here we propose a family of constrained-based approaches, termed PARROT, to predict enzyme allocations based on the principle of minimizing the enzyme allocation adjustment using protein-constrained metabolic models. To this end, PARROT variants model the minimization of enzyme reallocation using four different (combinations of) distance functions. We demonstrate that the PARROT variant that minimizes the Manhattan distance of enzyme allocations outperforms existing approaches based on the parsimonious distribution of fluxes or enzymes for both *Escherichia coli* and *Saccharomyces cerevisiae*. Further, we show that the combined minimization of flux and enzyme allocation adjustment leads to inconsistent predictions. Together, our findings indicate that minimization of resource, rather than flux, redistribution is a governing principle determining steady-state pathway activity for microorganism grown in suboptimal conditions.

KEYWORDS: Metabolic modelling; Metabolic engineering; Quantitative proteomics; Systems biology

Introduction

Constraint-based approaches have been employed to simulate and predict phenotypes based on genome-scale metabolic models (GEMs) [1]. While already useful for predicting a wide range of phenotypes, the predictive performance of GEMs has been further improved by integrating protein constraints, such as: enzyme catalytic rates and the allocation of enzyme abundances across reactions [2,3]. These protein-constrained GEMs (pcGEMs) have been used to predict complex phenotypes, such as the overflow metabolism, in which fermentation predominates over respiration when microorganisms grow in high sugar concentrations [3,4], and diauxic growth, when multiple carbon sources are available and the microbial growth presents two or more growth phases [5]. The models also allow for the incorporation of proteomics data, and thus provide a framework for multi-omics data analysis and integration [3,6].

The parameters included in pcGEMs are: (i) the enzyme turnover numbers, k_{cat}, a first-order rate constant with the unit of s^{-1}, that describes the limiting rate of reactions catalysed by enzymes when these are fully occupied at their saturation point; and (ii) enzyme abundances (in mmol/gDW), obtained from quantitative proteomics experiments. Values of k_{cat} can be measured from biochemical assays or estimated from computational methods based on constraint-based and data-driven approaches [7], while enzyme abundances are obtained from absolute proteomics measurements. More specifically, they are obtained from peptide intensity-based quantification or spectral counting [8]. However, proteomics experiments for absolute quantification are still difficult to perform, given the challenges put forward by the diversity of physicochemical properties of protein [9], lack of standards and problems in reproducibility [10], and overall inaccessibility given the high costs of equipment and supplies [11].

Computational methods have also been developed to predict protein abundance, mostly based on data-driven models. These models often explore the central dogma of molecular biology by assessing the relationship between transcription and protein biosynthesis. Notable approaches to estimate protein abundance include the joint learning approach devised by Li et al [12], where an ensemble model was constructed by combining different supervised learning algorithms, outperforming competing approaches in the NCI-CPTAC DREAM Proteogenomics Challenge. Another approach, developed by Terai and Asai [13], uses features such as the accessibility around the Shine-Dalgarno sequence, minimum free energy of the mRNA molecule, Viterbi score, and inside-outside score. Further, Ferreira et al. [14] explored codon usage bias information to train an AdaBoost regression model, achieving higher correlations than previous approaches without the usage of transcriptomics data.

Aside from machine learning models, constraint-based approaches have also been used to predict protein abundance. Using approaches such as MOMENT [2] or GECKO [3], it is possible to calculate the optimal concentration of enzymes necessary to carry the provided flux with the provided catalytic rate, given the relationship:

$$v_j \leq k_{cat}^{ij} \cdot [E_i] \qquad (1)$$

where v_j is the metabolic flux of reaction j, $[E_i]$ is the concentration of an enzyme i, and k_{cat}^{ij} is the catalytic rate of an enzyme i catalyzing a reaction j. This allows for deriving k_{cat}^{ij} values given the other two are available. This relationship was explored by Heckmann et al. [15] by using pcGEMs to predict enzyme concentrations given catalytic rates predicted computationally, achieving a 43% lower root mean squared error.

Assuming that pcGEMs that integrate proteomics data predict flux distributions that reflect the corresponding metabolic state, we ask whether the reverse operation could be employed to predict proteomics data that match a given physiological state. Moreover, as cells are exposed to stresses or changing environmental conditions, the optimal growth state is disturbed, leading to a suboptimal growth state in which gene expression, regulatory pathways and metabolic flux are changed in adjusting the cell to this new physiological condition [16]. Despite the aforementioned advances in predicting protein abundances, the problem of predicting enzyme allocation under suboptimal growth conditions remains largely unexplored. Here we propose PARROT (Figure 1), for **P**rotein allocation **A**djustment fo**R** suboptimal envi**RO**nmen**T**s, a family of constraint-based approaches for prediction of protein abundances for suboptimal conditions using protein abundances measured in a reference, optimal state. Our proposed approach is inspired by Minimization of Metabolic Adjustment (MOMA) [17], which minimizes the distance between a reference state and a gene knock-out state while ensuring cell survival in the later. We show that PARROT predicted enzyme concentrations in very good agreement with experimental data and outperformed competing methods for minimizing flux distributions. Therefore, PARROT can be used to parameterize pcGEMs for unseen, suboptimal conditions from which metabolic phenotypes can further be analysed.

Methods

The principle of minimizing the change in enzyme usage between a suboptimal and reference state

To find the enzyme distribution vector that matches the enzyme usage of a cell growing in suboptimal growth conditions, we propose PARROT, an approach that minimizes the distance between a reference enzyme allocation $\mathbf{E_{ref}}$ and a suboptimal growth enzyme allocation $\mathbf{E_s}$ (Figure 1). This is consistent with observations that micro-organisms minimize expenditures to perform a growth and associated flux state [18]. We define and compare four different objectives to model the distance between enzyme allocations in suboptimal and reference states: (i) the Manhattan distance; (ii) the Euclidean distance; (iii) the weighted sum of the Manhattan distance between enzyme allocations and the Manhattan distance between flux distributions; (iv) the weighted sum of the Euclidean distance between enzyme allocations and the Euclidean distance between flux distributions. The first can be formulated as a linear optimization problem (LP1), specified as follows:

$$\min\left\|\frac{\mathbf{E_{ref}}}{\mathrm{E_{ref}^{tot}}}-\frac{\mathbf{E_s}}{\mathrm{E_s^{tot}}}\right\|_1 \qquad (2)$$

$$\text{s.t. } \mathbf{Nv}=\mathbf{0} \qquad (3)$$

$$v_{s,min}\le v_s\le v_{s,max} \quad (4)$$

$$v_s\le k_{cat}\cdot[E_s] \qquad (5)$$

$$\sum E_s=E_s^{tot} \qquad (6)$$

$$v_{bio}=\mu, \qquad (7)$$

where E_{ref}^{tot} and E_s^{tot} represent the total enzyme usage in the model for the reference and suboptimal states, respectively; $\mathbf{N}$ is the stoichiometric matrix; $\mathbf{v}$ is the flux distribution vector; v_{bio} is the flux through the biomass pseudo-reaction; and μ is the specific growth rate, determined from measurements in the suboptimal state. The other objectives are captured by the following:

$$\text{QP1: }\left\|\frac{\mathbf{E_{ref}}}{\mathrm{E_{ref}^{tot}}}-\frac{\mathbf{E_s}}{\mathrm{E_s^{tot}}}\right\|_2, \quad (8)$$

$$\text{LP2: }\left\|\frac{\mathbf{E_{ref}}}{\mathrm{E_{ref}^{tot}}}-\frac{\mathbf{E_s}}{\mathrm{E_s^{tot}}}\right\|_1+\lambda\|\mathbf{v_{ref}}-\mathbf{v_s}\|_1, (9)$$

$$\text{QP2: }\left\|\frac{\mathbf{E_{ref}}}{\mathrm{E_{ref}^{tot}}}-\frac{\mathbf{E_s}}{\mathrm{E_s^{tot}}}\right\|_2+\lambda\|\mathbf{v_{ref}}-\mathbf{v_s}\|_2 .(10)$$

where the parameter λ is a weighting factor chosen by inspecting the difference between the norms of enzyme allocation and the flux distributions. We solved the corresponding problems under the same constraints as in Eq. 2. We implemented and solved the problems in MATLAB (The MathWorks Inc., Natick, Massachusetts) using the COBRA Toolbox [19] and the Gurobi solver v9.1.1 [20]. The implementation of PARROT can be found in the GitHub repository: https://github.com/mauricioamf/PARROT.

Experimental data and simulation constraints

To test the variants of the proposed approach, PARROT, we used the pcGEMs of *Saccharomyces cerevisiae*, ecYeast8 [21], and *Escherichia coli*, eciML1515 [22]. We employed quantitative proteomics measurements for both species performed in a number of growth conditions, ranging from optimal growth in standard physiological conditions to stress conditions, alternative nutrient usage and chemostat cultivation.

For *S. cerevisiae*, we used the protein measurements from Chen and Nielsen [23] for 19 different growth conditions, which were collected from four studies [24–27]. These included proteomics measurements in yeast growing in ethanol, osmolarity, and high temperature stresses [24]; yeast growing in chemostats with reducing nitrogen availability [25]; and yeast growing in chemostats limited by the nitrogen source in increasing dilution rates and in chemostats with alternative nitrogen sources [27]. We also used measurements of nutrient uptake rates, growth rates and protein content from these studies to constrain the batch model, which does not consider protein measurements and rely on the protein pool constraint.

For *E. coli*, we used the proteomics data for 20 different growth conditions collected in [28] from three different studies [29–31]. These include batch cultivations of *E. coli* growing with different carbon sources and a glucose-limited chemostat culture, with dilution rates ranging from 0.12 h^{-1} to 0.5 h^{-1} performed by Schmidt et al. [31], a second chemostat limited by glucose at dilution rates ranging from 0.11 h^{-1} to 0.49 h^{-1} [29], and a third chemostat limited by glucose at dilutions rates ranging from 0.21 h^{-1} to 0.51 h^{-1} [30]. Similar to *S. cerevisiae*, the batch model was constrained with the nutrient uptake rates, growth rates and protein content measured in the studies where the protein measurements were taken. For both species, we excluded the conditions that did not have measured uptake rates, growth rates, or protein content. In addition, we excluded the temperature stress conditions from Lahtvee et al. [24], as temperature can severely impact the function of enzymes [32], and temperature stress responses entail changes beyond metabolic flux redistribution [16].

Pre-processing of protein measurements for the reference state

From the protein measurements obtained from Davidi et al. [28] and Chen and Nielsen [23] we separated the measurements according to each experiment performed in the original studies. From each experiment we selected the control sample to represent the reference state in our approach PARROT. We corrected the protein measurements for the reference state measurements by integrating the values into the pcGEMs ecYeast8 and eciML1515 for *S. cerevisiae* and *E. coli*, respectively, using the GECKO Toolbox 2 [22]. The GECKO Toolbox 2 identifies the enzyme usage values that most limit growth and flexibilises the values to prevent over-constraining the model. We then used for the $\mathbf{E_{ref}}$ vector of each experiment the values for flexibilised proteins along with values for proteins that were unchanged.

Assessment of predicted enzyme usage distributions

The protein measurements, $\mathbf{E_s^{exp}}$, for the suboptimal growth conditions obtained from Davidi et al. [28] and Chen and Nielsen [23] were not used directly in simulations. These experimental measurements were instead employed to calculate a baseline to which predictions of $\mathbf{E_s}$ were compared. Assuming that simulations performed with pcGEMs use only the optimal concentration of enzymes necessary to carry a given metabolic flux, the model-allocated protein usage would underestimate the *in vivo* enzyme concentrations. To allow for a fair comparison, we devised a baseline by integrating the experimental proteomics measurements of each experiment into the pcGEMs using the GECKO Toolbox 2 in which we minimized the total enzyme allocation given the following optimization problem:

$$\min \left\| \mathbf{E_s^{exp}} \right\|_1 \quad (11)$$

$$\text{s.t.} \ \mathbf{Nv} = \mathbf{0} \quad (12)$$

$$\mathbf{v_{s,min}} \leq \mathbf{v_s} \leq \mathbf{v_{s,max}} \quad (13)$$

$$v_{s,j} \leq k_{cat}^{ij} \cdot \left[E_s^{exp,i}\right] \quad (14)$$

$$\sum E_s^{exp} = E_s^{exp,tot} \qquad (15)$$

$$v_{bio} = \mu. \qquad (16)$$

The resulting enzyme usage distribution, $\mathbf{E_s^{exp}}$, was then defined as the baseline for each sample of each proteomics experiment. We compared the predicted $\mathbf{E_s}$ values from the four variants of PARROT to $\mathbf{E_s^{exp}}$ by calculating the Pearson correlations of each sample. Further, we calculated the root-median square error (RMdSE) to measure the difference between predicted and baseline values. For assessing both correlations and the RMdSE, we log10-transformed the values for the predictions and the baseline.

We also performed a robustness analysis to check the effect of using the minimization of the second norm in constructing a baseline. In addition, we compared the predictions of our approaches to those obtained using an extension of parsimonious enzyme usage FBA (pFBA) [33] to consider enzyme constraints. To this end, for each sample of each experiment, we defined the optimization problem as:

$$\min \sum_{j=1}^{m} v_{j,s,irrev} \qquad (17)$$

$$\text{s.t. } \mathbf{N_{s,irrev}} \cdot \mathbf{v_{s,irrev}} = \mathbf{0} \qquad (18)$$

$$0 \leq \mathbf{v_{s,irrev}} \leq \mathbf{v_{s,irrev,max}} \qquad (19)$$

$$v_{s,irrev,j} \leq k_{cat}^{ij} \cdot \left[E_{s,i}\right] \qquad (20)$$

$$\sum E_s = E_s^{tot} \qquad (21)$$

$$v_{bio} = \mu, \qquad (22)$$

where $v_{j,s,irrev}$ corresponds to the flux distribution of an irreversible model in a non-optimal growth condition. We also assessed a modified version of pFBA with enzyme constraints with the following objective:

$$\min \sum_{j=1}^{m} E_{s,i} \cdot k_{cat}^{ij} \cdot \qquad (23)$$

For pFBA and the modified implementation, we applied the same constraints on nutrient uptake rates and growth rates as for the four approaches assessed previously, and calculated the Pearson correlations and the RMdSE. Lastly, as a negative control to benchmark the performance of PARROT, we equated $E_{s,i}$ to k_{cat}^{ij}, meaning that k_{cat} values we used directly as the enzyme usage. We calculated the correlation values and RMdSE for all assessed optimization problems and compared them to the predictions of pFBA and its modified implementation using a Pairwise Wilcoxon rank sum test with Bonferroni correction.

Assessment of optimal values for the λ weighting factor

To systematically assess the impact of different lambda values, we optimised the LP2 and QP2 variants using λ values ranging from 0 (no fluxes used) to 1 (fluxes and enzyme usages equally considered). Additionally, we optimised the LP2 and QP2 variants using λ values ranging from 0.1 to 1 in order to make sure fluxes are always used for the objective function. In both scenarios, we calculated the Pearson correlation to the baseline for each λ value. We determined the optimal λ value as the value that outputs predictions with the highest Pearson correlation when compared to the first norm baseline.

Results

PARROT successfully captures protein allocation changes in yeast

We used PARROT to predict the enzyme usage distribution for 19 growth conditions under constraints provided by experimental data. First, we built a baseline for comparison with predictions from PARROT (Figure 1). To this end, we integrated the experimental proteomics measurements obtained from Lahtvee et al. [24], Yu et al. [25], Di Bartolomeo et al. [26], and Yu et al. [27] (Table S1) in the ecYeast8 model and minimized the enzyme allocation (Methods). The resulting allocation of enzymes $\mathbf{E}_{\mathbf{s}}^{\mathbf{exp}}$ included 286 to 336 enzymes with abundance in all considered conditions. For the reference condition, we used the experimental proteomics measurements from optimal (control) growth conditions in the respective four groups of experiments, after flexibilization following GECKO 2.0 (see Methods) (Table S1). The number of enzymes contained in $\mathbf{E}_{\mathbf{ref}}$ ranged from 533 to 744, depending on the investigated control sample.

With the resulting enzyme allocation at the reference and the baseline of a suboptimal condition, $\mathbf{E}_{\mathbf{ref}}$ and $\mathbf{E}_{\mathbf{s}}^{\mathbf{exp}}$, we used the four variants of PARROT (see Methods) to predict the enzyme allocation, $\mathbf{E_s}$, for the suboptimal condition. The number of enzymes contained in the predicted $\mathbf{E_s}$ ranged from 18 to 336 over the considered experiments. When comparing the median of the calculated Pearson correlations between the baseline and predicted enzyme allocation correlations, we found that all PARROT variants achieved a higher median correlation when compared to pFBA and its modified implementation, except for the minimization of the Euclidean distances considering fluxes (Figure 2a, see QP2, Methods). We also evaluated the RMdSE between predictions and the baselines, and observed that the minimization of the Euclidean distance considering fluxes (QP2, Methods) resulted in a median error comparable to pFBA and its modified implementation, EsKcat (see Methods) (Figure 2b). Further, all PARROT variants outperformed the null model, where k_{cat} values are used directly as the enzyme usage. Taken together, the results demonstrated that PARROT achieved good predictive performance based on the data from *S. cerevisiae*.

Different variants of PARROT outperformed contending methods for *E. coli*

To verify if the conclusions from PARROT hold in another unicellular model organism, we applied it to predict enzyme allocation $\mathbf{E_s}$ in suboptimal conditions for *E. coli* given constraints provided by growth experiments. As in the case of *S. cerevisiae*, we built a baseline for comparison with the predictions obtained from PARROT by integrating the experimental proteomics measurements from Valgepea et al. [29], Peebo et al. [30] and Schmidt et al. [31] (Table S2) in the eciML1515 model, and minimized the total enzyme allocation (see Methods). The resulting $\mathbf{E_s^{exp}}$ included protein allocation for 164 to 176 enzymes. Further, as reference condition we considered the control samples or the chemostat measurements with the smallest dilution rate (Table S2). The number of enzymes contained in $\mathbf{E_{ref}}$ ranged from 152 to 188 depending on the control experimented used.

The prediction of $\mathbf{E_s}$ distributions and their assessment were similar to *S. cerevisiae*, with the number of predicted values ranging from 19 to 141. After performing a comparison of Pearson correlations between variants of PARROT, pFBA and its modified implementation, EsKcat (see Methods), we found that different variants outperformed pFBA. Both minimizations of the Manhattan distance, with or without metabolic fluxes (LP1 and LP2, Methods), exhibited significantly higher median correlations compared to pFBA (p-value = $1.24 \cdot 10^{-13}$ and $6.2 \cdot 10^{-14}$ for Pearson correlations respectively, pairwise Wilcoxon rank sum test) (Figure 3a). Another variant with a significant difference to pFBA was the minimization of the Euclidean distance of enzyme usages (QP1, Methods). Regarding the RMdSE, the minimization of the weighted sum of the Euclidean distance of enzyme usage and Euclidean distance of flux distributions outperformed the other PARROT variants. As with *S. cerevisiae*, PARROT outperformed the null model in all comparisons. These findings demonstrated that PARROT is applicable with data from another microorganism without decrease in performance.

Robustness analysis shows the consistency of prediction from PARROT

To further evaluate the predictions made by PARROT, we investigated how the usage of a baseline constructed by minimizing the second norm of the vector $\mathbf{E_s^{exp}}$ impacts the comparisons. To this end, we repeated all comparisons as performed for a baseline constructed by minimizing the first norm, using the predicted $\mathbf{E_s}$ obtained by the PARROT variants. Importantly, the results were consistent between the two baseline approaches. For *S. cerevisiae*, the minimization of the weighted sum of the Manhattan distance of enzyme usage and Manhattan distance of flux distributions (LP2, Methods) was the variant that achieved the highest mean Pearson correlations than pFBA and its modified implementation (Figure S1).

For the RMdSE, all PARROT variants had errors comparable to the positive controls (Figure S2). As observed for comparisons using the first norm baseline, all PARROT variants outperformed the null model.

The comparisons performed using predictions obtained for *E. coli* were also consistent with different variants of PARROT that outperformed pFBA. Considering the Pearson correlations, the minimization of the Manhattan distance (LP1, Methods) and the minimization of the weighted sum of the Manhattan distance of enzyme usage and Manhattan distance of flux distributions (LP2, Methods) also had the highest median correlations and were significantly different to pFBA. Likewise, these PARROT variants also had a significant difference to EsKcat, the modified implementation of pFBA (Figure S3). The comparison of RMdSE values were also consistent, as the errors were comparable to the positive controls (Figure S4). Altogether, these results highlight the robustness of estimations of $\mathbf{E_s}$ obtained from PARROT.

Proteome-aware minimalization is more relevant than minimization of flux distances

We assessed the impact of different λ values ranging from 0 (no fluxes used) to 1 (fluxes and enzyme usages equally considered). We also considered a scenario of λ values ranging from 0.1 to 1 in order to probe different solutions where metabolic fluxes are always considered. We considered λ value to be optimal if it resulted in the highest Pearson correlation to the baseline. In the first scenario, for both *S. cerevisiae* and *E. coli* the most frequent optimal λ was 0, with decreasing correlation values as λ values increased (Figure 4a, 4c). In the second scenario, the optimal λ values were more equally distributed, with *S. cerevisiae* having a higher frequency of lower values (Figure 4b). For *E. coli*, lower λ values were also frequent, while also having a λ of 1 slighly more frequent than a λ of 0.2 (Figure 5d). Taken together, these results indicate that the problem of minimizing enzyme usage contributes more to predictions than minimizing metabolic fluxes.

Discussion

Here we proposed a family of constraint-based approaches, termed PARROT, that address the problem of predicting reallocation of protein abundance from an optimal condition to a suboptimal condition. PARROT is based on the principle that organisms tend to minimally adjust cellular physiology between growth conditions to make effective use of resources [18]. The predictions of enzyme allocation generated by PARROT rely on quantitative proteomics data for a reference condition. The resulting optimization problems constructed are thus similar to MOMA, which depends on a model representing a wild-type strain to predict a minimally adjusted flux distribution for a mutant strain.

By comparing the predictions to a baseline constructed with experimental proteomics measurements for suboptimal conditions, we found that PARROT predicted protein abundances with very good agreement with the baseline. In addition, we demonstrated that these predictions were consistent and robust to how the baseline is constructed. The performance of PARROT also holds for two model organisms, *S. cerevisiae* and *E. coli*, highlighting the general application of the principle of minimal protein adjustment on which the predictions are based.

From the different variants of PARROT, the minimization of the Manhattan distance (LP1) and the minimization of the weighted sum of the Manhattan distance of enzyme usage and Manhattan distance of flux distributions (LP2) were the best contenders. The variants QP1 and QP2 – that minimizes Euclidean distances instead of Manhattan distances – resulted in good but also inconsistent performance between *S. cerevisiae* and *E. coli*. This agrees with the fact that the first norm distance is the natural metric for enzyme abundances in the cell, because a change in enzyme concentration requires ribosomal activity that scales linearly with the enzyme abundance [34].

The baseline approach devised to assess the predictions allows for a fair comparison between the predicted enzyme usage distribution and the experimental protein abundance values. In constraining the pcGEMs using the proteomics measurements, the experimental values are first readjusted to match the enzyme levels that actually carry flux in the model, since more protein is produced than actually needed by the cell [35]. This, however, implies that the predicted values are not directly comparable to experimental proteomics values, which affect the determined measures of performance. By adjusting the experimental values to levels that are compatible with what is actually employed to carry metabolic flux, we could more adequately assess the correlation with enzyme allocation predicted from the pcGEMs, albeit losing the direct correspondence to experimental data.

The parameter λ is a factor that weights the usage of metabolic fluxes for the optimisation problem. By varying this value between 0 and 1, we could assess how much the minimization of metabolic fluxes contributes to the problem of predicting enzyme usage. A λ value of 0 would render the variants LP2 and QP2 equivalent to LP1 and QP1, respectively, as metabolic flux would be neglected in the optimal solutions. A λ value of 1, in the other hand, renders LP2 and QP2 as equivalent to using a pcGEM with the canonical implementation of MOMA, which considers all fluxes equally. When the two PARROT variants are free to vary λ between 0 and 1, there is a strong preference for lower λ values. When constraining λ to a value between 0.1 and 1, higher values of λ are present but still not more prevalent than lower values of λ. This suggests that the joint minimization of fluxes and enzymes is not a principle of flux redistribution, and the principle is guided by minimization of resource redistribution, as best captured by LP1 and QP1, and by LP2 and QP2 with low values of λ. Thus, by being proteome-aware, PARROT is better suited for simulations using pcGEMs than the quadratic and linear implementations of MOMA, given that higher participation of metabolic fluxes lowers the overall predictive performance. Altogether, we demonstrated that minimizing the readjustment of enzyme resource allocation is one principle underpinning microbial adjustment to a suboptimal condition. Thus, PARROT may allow for study and engineering of microbial cell factories, as these are often under suboptimal growth conditions in industrial settings [36].

Despite the advantages of using a baseline, predictions of enzyme levels using Eq. (1) still underestimates protein abundance, leading to a disparity between predictions and *in vivo* concentrations. This remaining portion of proteins, termed the "proteome reserve", is useful for the cell to quickly adapt to unstable environments, being an evolutionary conserved strategy [37]. It is important to highlight, though, that this reasoning does not assume that cells are operating at the saturation point for all metabolites, but rather that enzymes are used inefficiently. If enzymes are operating near V_{max}, then enzymes would be the only cellular components that exert control on metabolic fluxes. As noted by Hackett et al. [38], however, is that cell overexpress enzymes and uses metabolite concentrations to control metabolic flux. This falls in line with the evolutionary conservation of protein stoichiometries at the pathway level as demonstrated by Lalanne et al. [39]. Although it is still not understood how preferred enzyme stoichiometry is determined, it was observed that the preferred range of enzyme stoichiometry follows a narrow distribution among pathways in Gram-positive and -negative bacteria, likely a result of evolutionary conservation or convergence. As suggested in the study, protein biosynthesis and consequently its usage is bound to a cost-benefit trade-off, where the optimal level of enzymes is balanced with the need for a buffer zone in case of changing environments. Similar to our approach, the works of Mori et al. [37] and Lalanne et al. [39] deals with proteome reallocation in a suboptimal growth condition. However, the first deals with proteome sectors, while the latter concerns with pathway-centric stoichiometries. Our approach thus differs as we consider protein reallocation for each enzyme individually.

Nevertheless, other approaches for estimating *in vivo* protein concentrations would still need to overcome the underestimating capacity of pcGEMs, especially by considering the proteome reserve. These approaches could include features such as cellular machinery beyond enzymes that participate in metabolism, or by integrating constraint-based approaches with data-driven approaches.

CRediT authorship contribution statement

Mauricio Ferreira: Conceptualization, Methodology, Software, Investigation, Validation, Writing - Original Draft, Writing - Review & Editing. **Wendel Silveira**: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Zoran Nikoloski**: Conceptualization, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Competing interests

The authors have declared no competing interests.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. We thank Marius Arend, Philipp Wendering and Eduardo Almeida for their critical discussion and comments on this study.

Data availability

All data and code are publicly available in the GitHub repository: (https://github.com/mauricioamf/PARROT)

ORCID

0000-0002-6545-6813 (Mauricio Ferreira)
0000-0001-7869-8144 (Wendel Silveira)
0000-0003-2671-6763 (Zoran Nikoloski)

References

[1] N.D. Price, J.A. Papin, C.H. Schilling, B.O. Palsson, Genome-scale microbial in silico models: The constraints-based approach, Trends in Biotechnology. 21 (2003) 162–169. https://doi.org/10.1016/S0167-7799(03)00030-1.

[2] R. Adadi, B. Volkmer, R. Milo, M. Heinemann, T. Shlomi, Prediction of Microbial Growth Rate versus Biomass Yield by a Metabolic Network with Kinetic Parameters, PLoS Computational Biology. 8 (2012) e1002575. https://doi.org/10.1371/journal.pcbi.1002575.

[3] B.J. Sánchez, C. Zhang, A. Nilsson, P. Lahtvee, E.J. Kerkhoven, J. Nielsen, Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating

enzymatic constraints, Mol Syst Biol. 13 (2017) 935. https://doi.org/10.15252/msb.20167411.

[4] M. Basan, S. Hui, H. Okano, Z. Zhang, Y. Shen, J.R. Williamson, T. Hwa, Overflow metabolism in Escherichia coli results from efficient proteome allocation, Nature. 528 (2015) 99–104. https://doi.org/10.1038/nature15765.

[5] Q.K. Beg, A. Vazquez, J. Ernst, M.A. De Menezes, Z. Bar-Joseph, A.L. Barabási, Z.N. Oltvai, 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. 104 (2007) 12663–12668. https://doi.org/10.1073/pnas.0609845104.

[6] P.S. Bekiaris, S. Klamt, Automatic construction of metabolic models with enzyme constraints, BMC Bioinformatics. 21 (2020) 1–13. https://doi.org/10.1186/S12859-019-3329-9/TABLES/2.

[7] M.A. de M. Ferreira, W.B. da Silveira, Z. Nikoloski, Protein constraints in genome-scale metabolic models: data integration, parameter estimation, and prediction of metabolic phenotypes, Authorea Preprints. (2022). https://doi.org/10.22541/AU.166082043.36599845/V1.

[8] C. Lindemann, N. Thomanek, F. Hundt, T. Lerari, H.E. Meyer, D. Wolters, K. Marcus, Strategies in relative and absolute quantitative mass spectrometry based proteomics, Biological Chemistry. 398 (2017) 687–699. https://doi.org/10.1515/HSZ-2017-0104/ASSET/GRAPHIC/J_HSZ-2017-0104_FIG_005.JPG.

[9] A. Otto, D. Becher, F. Schmidt, Quantitative proteomics in the field of microbiology, Proteomics. 14 (2014) 547–565. https://doi.org/10.1002/pmic.201300403.

[10] F. Calderón-Celis, J.R. Encinar, A. Sanz-Medel, Standardization approaches in absolute quantitative proteomics with mass spectrometry, Mass Spectrom Rev. 37 (2018) 715–737. https://doi.org/10.1002/mas.21542.

[11] A. Swiatly, S. Plewa, J. Matysiak, Z.J. Kokot, Mass spectrometry-based proteomics techniques and their application in ovarian cancer research, Journal of Ovarian Research. 11 (2018) 1–13. https://doi.org/10.1186/s13048-018-0460-6.

[12] H. Li, O. Siddiqui, H. Zhang, Y. Guan, Joint learning improves protein abundance prediction in cancers, BMC Biol. 17 (2019) 1–14. https://doi.org/10.1186/S12915-019-0730-9/FIGURES/6.

[13] G. Terai, K. Asai, Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility, Nucleic Acids Res. 48 (2020). https://doi.org/10.1093/nar/gkaa481.

[14] M. Ferreira, R. Ventorim, E. Almeida, S. Silveira, W. Silveira, Protein Abundance Prediction Through Machine Learning Methods, J Mol Biol. 433 (2021) 167267. https://doi.org/10.1016/J.JMB.2021.167267.

[15] D. Heckmann, C.J. Lloyd, N. Mih, Y. Ha, D.C. Zielinski, Z.B. Haiman, A.A. Desouki, M.J. Lercher, B.O. Palsson, Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models, Nat Commun. 9 (2018) 1–10. https://doi.org/10.1038/s41467-018-07652-6.

[16] P.-J. Lahtvee, R. Kumar, B.M. Hallstrom, J. Nielsen, Adaptation to different types of stress converge on mitochondrial metabolism, Molecular Biology of the Cell. 27 (2016) 2505–2514. https://doi.org/10.1091/mbc.E16-03-0187.

[17] D. Segrè, D. Vitkup, G.M. Church, Analysis of optimality in natural and perturbed metabolic networks, Proceedings of the National Academy of Sciences of the United States of America. 99 (2002) 15112–15117. https://doi.org/10.1073/PNAS.232349399/SUPPL_FILE/3493SUPPLINKS.HTML.

[18] A. Goelzer, J. Muntel, V. Chubukov, M. Jules, E. Prestel, R. Nölker, M. Mariadassou, S. Aymerich, M. Hecker, P. Noirot, D. Becher, V. Fromion, Quantitative prediction of genome-wide resource allocation in bacteria, Metab Eng. 32 (2015) 232–243. https://doi.org/10.1016/J.YMBEN.2015.10.003.

[19] L. Heirendt, S. Arreckx, T. Pfau, S.N. Mendoza, A. Richelle, A. Heinken, H.S. Haraldsdóttir, J. Wachowiak, S.M. Keating, V. Vlasov, S. Magnusdóttir, C.Y. Ng, G. Preciat, A. Žagare, S.H.J. Chan, M.K. Aurich, C.M. Clancy, J. Modamio, J.T. Sauls, A. Noronha, A. Bordbar, B. Cousins, D.C. El Assal, L. V. Valcarcel, I. Apaolaza, S. Ghaderi, M. Ahookhosh, M. Ben Guebila, A. Kostromins, N. Sompairac, H.M. Le, D. Ma, Y. Sun, L. Wang, J.T. Yurkovich, M.A.P. Oliveira, P.T. Vuong, L.P. El Assal, I. Kuperstein, A. Zinovyev, H.S. Hinton, W.A. Bryant, F.J. Aragón Artacho, F.J. Planes, E. Stalidzans, A. Maass, S. Vempala, M. Hucka, M.A. Saunders, C.D. Maranas, N.E. Lewis, T. Sauter, B. Palsson, I. Thiele, R.M.T. Fleming, Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0, Nat Protoc. 14 (2019) 639–702. https://doi.org/10.1038/s41596-018-0098-2.

[20] L. Gurobi Optimization, Gurobi Optimizer Reference Manual, (2020).

[21] H. Lu, F. Li, B.J. Sánchez, Z. Zhu, G. Li, I. Domenzain, S. Marcišauskas, P.M. Anton, D. Lappa, C. Lieven, M.E. Beber, N. Sonnenschein, E.J. Kerkhoven, J. Nielsen, A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism, Nat Commun. 10 (2019). https://doi.org/10.1038/s41467-019-11581-3.

[22] I. Domenzain, B. Sánchez, M. Anton, E.J. Kerkhoven, A. Millán-Oropeza, C. Henry, V. Siewers, J.P. Morrissey, N. Sonnenschein, J. Nielsen, Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0, Nat Commun. 13 (2022) 1–13. https://doi.org/10.1038/s41467-022-31421-1.

[23] Y. Chen, J. Nielsen, In vitro turnover numbers do not reflect in vivo activities of yeast enzymes, Proc Natl Acad Sci U S A. 118 (2021) e2108391118. https://doi.org/10.1073/PNAS.2108391118/SUPPL_FILE/PNAS.2108391118.SD08.XLSX.

[24] P.J. Lahtvee, B.J. Sánchez, A. Smialowska, S. Kasvandik, I.E. Elsemman, F. Gatto, J. Nielsen, Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast, Cell Syst. 4 (2017) 495-504.e5. https://doi.org/10.1016/j.cels.2017.03.003.

[25] R. Yu, K. Campbell, R. Pereira, J. Björkeroth, Q. Qi, E. Vorontsov, C. Sihlbom, J. Nielsen, Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast, Nat Commun. 11 (2020) 1–12. https://doi.org/10.1038/s41467-020-15749-0.

[26] F. Di Bartolomeo, C. Malina, K. Campbell, M. Mormino, J. Fuchs, E. Vorontsov, C.M. Gustafsson, J. Nielsen, Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift, Proc Natl Acad Sci U S A. 117 (2020) 7524–7535. https://doi.org/10.1073/PNAS.1918216117/SUPPL_FILE/PNAS.1918216117.SD07.XLSX.

[27] R. Yu, E. Vorontsov, C. Sihlbom, J. Nielsen, Quantifying absolute gene expression profiles reveals distinct regulation of central carbon metabolism genes in yeast, Elife. 10 (2021). https://doi.org/10.7554/ELIFE.65722.

[28] D. Davidi, E. Noor, W. Liebermeister, A. Bar-Even, A. Flamholz, K. Tummler, U. Barenholz, M. Goldenfeld, T. Shlomi, R. Milo, Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements, Proc Natl Acad Sci U S A. 113 (2016) 3401–3406. https://doi.org/10.1073/pnas.1514240113.

[29] K. Valgepea, K. Adamberg, A. Seiman, R. Vilu, Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins, Mol Biosyst. 9 (2013) 2344–2358. https://doi.org/10.1039/C3MB70119K.

[30] K. Peebo, K. Valgepea, A. Maser, R. Nahku, K. Adamberg, R. Vilu, Proteome reallocation in Escherichia coli with increasing specific growth rate, Mol Biosyst. 11 (2015) 1184–1193. https://doi.org/10.1039/c4mb00721b.

[31] A. Schmidt, K. Kochanowski, S. Vedelaar, E. Ahrné, B. Volkmer, L. Callipo, K. Knoops, M. Bauer, R. Aebersold, M. Heinemann, The quantitative and condition-dependent Escherichia coli proteome, Nat Biotechnol. 34 (2016) 104–110. https://doi.org/10.1038/nbt.3418.

[32] G. Li, Y. Hu, Jan Zrimec, H. Luo, H. Wang, A. Zelezniak, B. Ji, J. Nielsen, Bayesian genome scale modelling identifies thermal determinants of yeast metabolism, Nat Commun. 12 (2021) 1–12. https://doi.org/10.1038/s41467-020-20338-2.

[33] N.E. Lewis, K.K. Hixson, T.M. Conrad, J.A. Lerman, P. Charusanti, A.D. Polpitiya, J.N. Adkins, G. Schramm, S.O. Purvine, D. Lopez-Ferrer, K.K. Weitz, R. Eils, R. König, R.D. Smith, B. Palsson, Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models, Molecular Systems Biology. 6 (2010). https://doi.org/10.1038/msb.2010.47.

[34] T. von der Haar, A quantitative estimation of the global translational activity in logarithmically growing yeast cells, BMC Syst Biol. 2 (2008) 1–14. https://doi.org/10.1186/1752-0509-2-87/FIGURES/7.

[35] E.J. O'Brien, J. Utrilla, B.O. Palsson, Quantification and Classification of E. coli Proteome Utilization and Unused Protein Costs across Environments, PLOS Computational Biology. 12 (2016) e1004998. https://doi.org/10.1371/JOURNAL.PCBI.1004998.

[36] Q. Deparis, A. Claes, M.R. Foulquié-Moreno, J.M. Thevelein, Engineering tolerance to industrially relevant stress factors in yeast cell factories, FEMS Yeast Research. 17 (2017) 1–35. https://doi.org/10.1093/femsyr/fox036.

[37] M. Mori, S. Schink, D.W. Erickson, U. Gerland, T. Hwa, Quantifying the benefit of a proteome reserve in fluctuating environments, Nat Commun. 8 (2017) 1–8. https://doi.org/10.1038/s41467-017-01242-8.

[38] S.R. Hackett, V.R.T. Zanotelli, W. Xu, J. Goya, J.O. Park, D.H. Perlman, P.A. Gibney, D. Botstein, J.D. Storey, J.D. Rabinowitz, Systems-level analysis of mechanisms regulating yeast metabolic flux, Science. 354 (2016). https://doi.org/10.1126/SCIENCE.AAF2786.

[39] J.B. Lalanne, J.C. Taggart, M.S. Guo, L. Herzel, A. Schieler, G.W. Li, Evolutionary Convergence of Pathway-Specific Enzyme Expression Stoichiometry, Cell. 173 (2018) 749-761.e38. https://doi.org/10.1016/J.CELL.2018.03.007.

Figures

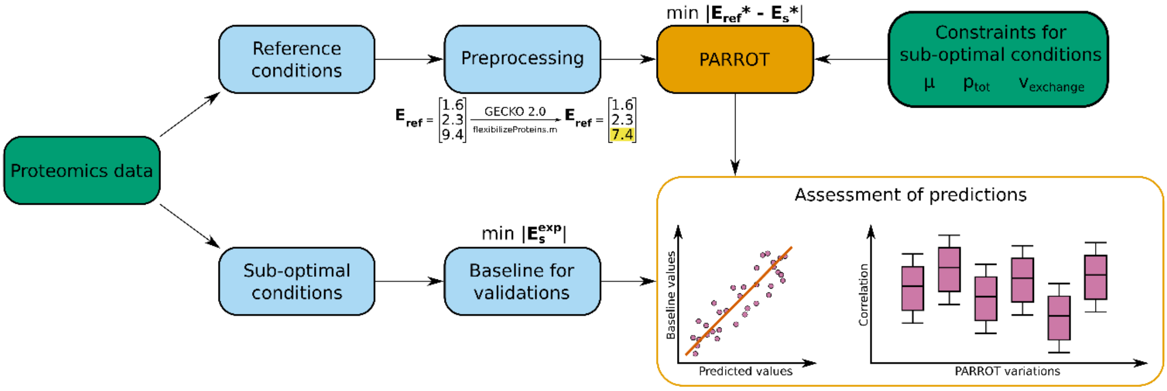
Proteomics data
Reference conditions
Preprocessing
$\mathbf{E}_{\mathbf{ref}} = \begin{bmatrix}1.6\\2.3\\9.4\end{bmatrix} \xrightarrow[\text{flexibilizeProteins.m}]{\text{GECKO 2.0}} \mathbf{E}_{\mathbf{ref}} = \begin{bmatrix}1.6\\2.3\\7.4\end{bmatrix}$
min $|\mathbf{E}_{\mathbf{ref}}^* - \mathbf{E}_{\mathbf{s}}^*|$
PARROT
Constraints for sub-optimal conditions
μ p_{tot} $v_{exchange}$
Sub-optimal conditions
min $|\mathbf{E}_{\mathbf{s}}^{\mathbf{exp}}|$
Baseline for validations
Assessment of predictions
Baseline values
Predicted values
Correlation
PARROT variations

Figure 1. Workflow of PARROT to predict enzyme usage for suboptimal growth conditions.

PARROT uses experimental proteomics data from an optimal growth condition as a reference point, and experimental physiological parameters from a suboptimal growth condition in a protein-constrained model. The proteomics data from the reference state is pre-processed by integrating the data in a pcGEM using the GECKO Toolbox 2 and flexibilising its values. The proteomics data from the suboptimal state is used to generate a baseline, which is in turn used for comparison with predictions from the PARROT variants.

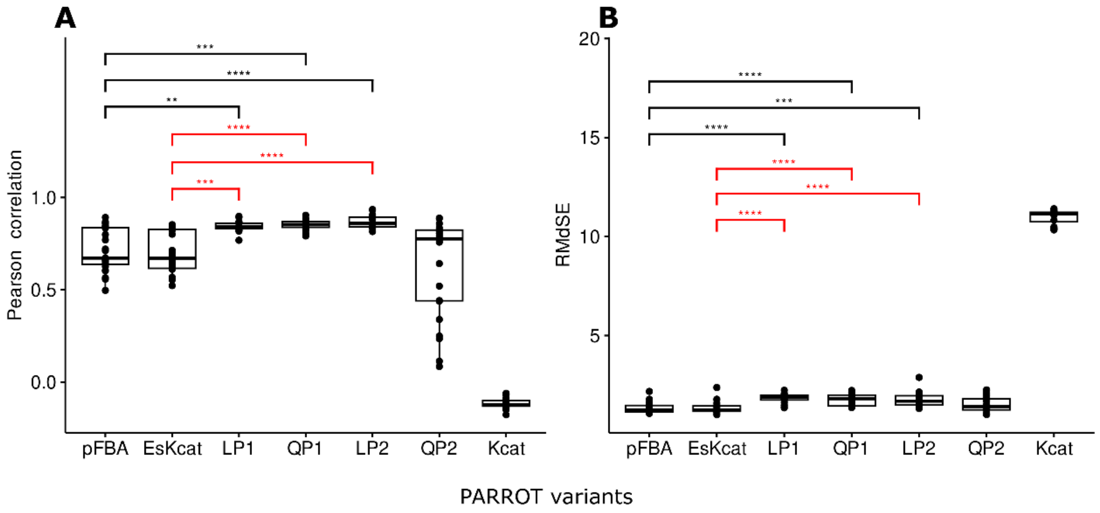

Figure 2. Comparative performance analysis of PARROT with proteomics data from *S. cerevisiae.*

All protein abundance values were log10-transformed prior to comparisons. **a**. Pearson correlation calculated between predicted enzyme distribution and the baseline obtained from minimizing the first norm of the experimental enzyme usage distribution. The four variants of PARROT are denoted as LP1 (Manhattan distance of enzyme distributions), LP2 (weighted Manhattan distance, considering flux and enzyme distributions), QP1 (Euclidean distance of enzyme distributions), and QP2 (weighted Euclidean distance of flux and enzyme distributions). The performance of PARROT was compared to pFBA and its modified version EsKcat (first norm of enzyme usage), see Methods. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < $1\cdot10^{-5}$, *** p-value < $2\cdot10^{-4}$, ** p-value < $5\cdot10^{-4}$. **b**. Assessment of model performance based on the root median squared error (RMdSE). A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value < $9\cdot10^{-6}$, *** p-value < $2\cdot10^{-5}$. Black significance bar indicates comparisons to pFBA. Red significance bar indicates comparison to EsKcat.

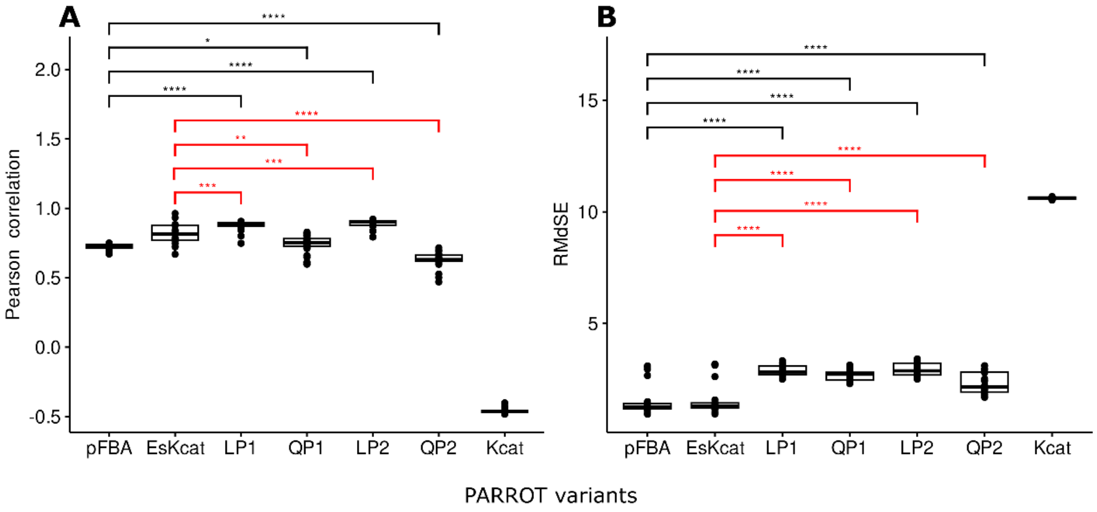

Figure 3. Comparative performance analysis of PARROT with proteomics data from *E. coli.*

All protein abundance values were log10-transformed prior to comparisons. **a**. Pearson correlation calculated between predicted enzyme usage distribution and the baseline obtained from minimizing the first norm of the experimental enzyme usage distribution. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value $< 2 \cdot 10^{-11}$, *** p-value $< 2 \cdot 10^{-4}$, ** p-value $< 6 \cdot 10^{-3}$, * p-value $< 3 \cdot 10^{-2}$. **b**. Assessment of model performance based on the RMdSE in *E. coli*. A pairwise Wilcoxon rank sum assesses the statistical significance: **** p-value $< 1 \cdot 10^{-5}$. Black significance bar indicates comparisons to pFBA. Red significance bar indicates comparison to EsKcat.

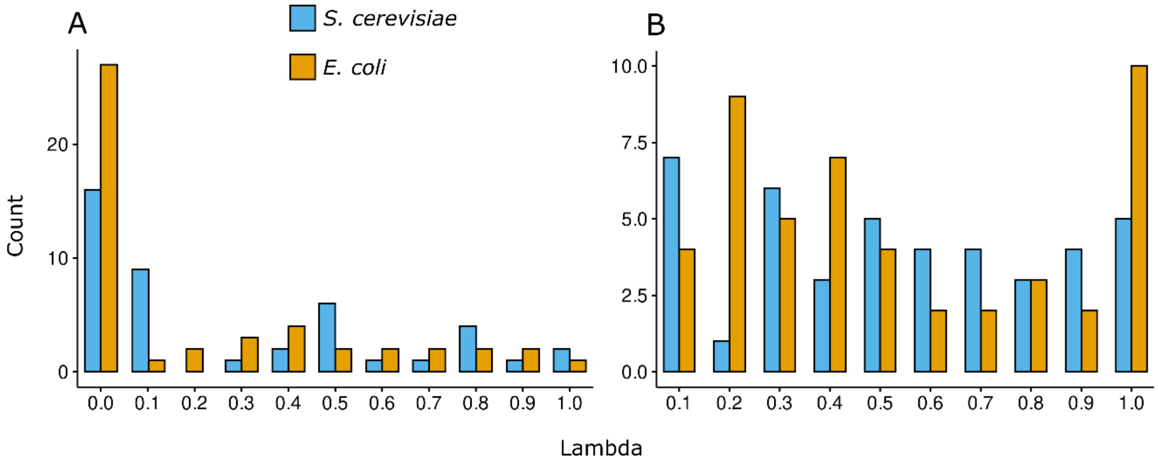

Figure 4. Optimal λ values across conditions and PARROT variants.

The optima λ value was determined by optimising the LP2 and QP2 variants and finding the value that outputs predictions with the highest Pearson correlation when compared to the baseline. Blue bars correspond to *S. cerevisiae*, and orange bars correspond to *E. coli*. **a**. Number of occurrences of an optimal λ value in a range of 0 to 1. Note that a λ value of zero means that no fluxes are used for the objective, being equivalent to the LP1 and LP2 variants. **b**. Number of occurrences of an optimal λ value in a range of 0.1 to 1. In this scenario, fluxes are always used for the objective.