Main Manuscript for

Accurate prediction of *in vivo* protein abundances by coupling constraint-based modelling and machine learning.

Maurício Alexander de Moura Ferreira[1], Philipp Wendering[2,3], Marius Arend[2,3], Wendel Batista da Silveira[1], Zoran Nikoloski[2,3,*]

1 Department of Microbiology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

2 Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany

3 Systems Biology and Mathematical Modelling, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

*Corresponding author: Zoran Nikoloski.

Email: zoran.nikoloski@uni-potsdam.de

Author Contributions: MAMF, PW, MA, WBS and ZN designed research. MAMF performed research and analyzed data. MAMF and ZN wrote the manuscript. MAMF, PW, MA, WBS and ZN revised the manuscript.

Competing Interest Statement: The authors declare no conflict of interest.

Classification: Biological Sciences, Systems Biology

Keywords: Protein allocation, Environmental effects, Multi-model framework

This PDF file includes:

Main Text
Figures 1 to 4

Abstract

Quantification of how different environmental cues affect protein allocation can provide important insights for understanding cell physiology. While absolute quantification of proteins can be obtained by resource-intensive mass-spectrometry-based technologies, prediction of protein abundances offers another way to obtain insights into protein allocation. Yet, building machine learning models of high accuracy across diverse, sub-optimal conditions remains notoriously difficult due to the dynamic nature of protein allocation. Here we present CAMEL, a framework that couples constraint-based modelling with machine learning to predict protein abundance for any environmental condition. This is achieved by building machine learning models that leverage static features, derived from protein sequences, and condition-dependent features predicted from protein-constrained metabolic models. Our findings demonstrate that CAMEL results in excellent prediction of protein allocation in *E. coli* (average Pearson correlation of at least 0.9), and moderate performance in *S. cerevisiae* (average Pearson correlation of at least 0.5). Therefore, CAMEL outperformed contending approaches without using molecular read-outs from unseen conditions and provides a valuable tool for using protein allocation in biotechnological applications.

Significance Statement

Proteins perform diverse functions, and their allocation to different pathways affects the physiology of cellular systems. However, quantification of protein allocation is a challenging task, and methods for accurate prediction of protein abundance offer an alternative for characterizing the relationship between protein expression and cellular physiology. Yet, prediction of protein allocation remains notoriously difficult, particularly under suboptimal conditions, with existing attempts relying on other molecular read-outs from the respective environments. Here, we present an approach that couples machine learning with advances in constraint-based modelling to accurately predict *in vivo* protein abundances. The usage of environment-specific features, predicted from enzyme-constrained metabolic models, along with static protein features, renders the approach readily applicable in variety of biotechnologically relevant scenarios simulated by these models.

Main Text

Introduction

Proteomics technologies facilitate system-wide profiling of the identity as well as changes in abundance, distribution, modifications, and interactions of proteins that drive different cellular processes (1). As cells are dynamic systems, these characteristics of the proteome change in response to different environmental cues to facilitate the system's adaptation to different physiological states (2, 3). As a result, high-throughput quantitative proteomics technologies, based on mass spectrometry (MS) techniques, have provided valuable information for various biotechnological and medical applications (4, 5). However, the measurement of absolute protein abundance still poses great challenges, due to physicochemical properties that interfere with ionization efficiency in mass spectrometry (6, 7) and the lack of a standardized approach, affecting the reproducibility of resulting data (8).

Another approach for protein quantification relies on predicting protein abundances by using a variety of machine learning approaches using features that are more facile to measure or quantify.

For instance, with gene expression data as predictors, Torres-García et al. (9) trained gradient-boosted trees to predict protein abundances in *Desulfovibrio vulgaris*, resulting in prediction accuracies, quantified by the coefficient of determination, R^2, ranging from 0.39 to 0.58. Similarly, Mehdi et al. (10) used Bayesian networks to predict protein abundance for *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* using gene expression data, yielding Pearson's correlation coefficients ranging from 0.50 to 0.71. Mergner et al. (11) used stepwise and LASSO regression models to predict protein abundances in *Arabidopsis thaliana* based on gene expression as well as sequence-derived features, achieving Pearson correlation in the range from 0.61 to 0.79 across different tissues. Codon usage metrics have also been used to predict protein abundance in *Saccharomyces cerevisiae*, with the best model achieving an R^2 value of 0.74 (12). In addition, structural features of mRNA molecules have been used to predict protein abundances in *Escherichia coli*, achieving a Spearman correlation coefficient of 0.71. However, these machine learning models have been developed using quantitative proteomics and transcriptomics data from optimal growth conditions and their performance in sub-optimal growth conditions remains unexplored. In addition, given the dynamic nature of the proteome, it is expected that the predictions across different conditions from machine learning models based on environment-invariant features (e.g., codon usage and structure-related), are poor.

The advent of protein-constrained genome-scale metabolic models (pcGEMs) (13) has facilitated not only the prediction of protein abundance but also the usage of proteomics data to predict metabolic and physiological traits (14–16). This is achieved by parameterizing pcGEMs with data on enzyme turnover numbers, providing the basis to link steady-state fluxes with enzyme abundances. For instance, a pcGEM of *E. coli* was used by Adadi et al. (17) to predict protein abundance, achieving Pearson correlation coefficient of 0.84 with gene expression data of *E. coli* grown on a glucose minimal medium. However, given that comparisons were made with gene expression data, it remains unexplored how these predictions fare against measured proteomics data. Further, Heckmann et al. (18) trained machine learning approaches to predict enzyme catalytic efficiency, namely *in vitro* (k_{cat}) and *in vivo* values (k_{max}^{vivo}), that were later used to predict protein abundances with the approach of Adadi et al. (17). Comparing the predicted protein abundances to experimental data from Schmidt et al. (19), they found that using k_{max}^{vivo} values resulted in a 43% lower root mean squared error (RMSE) in comparison to k_{cat} values. However, when comparing the prediction to measured proteomics data, the obtained R^2 values were poor to modest, ranging from 0.11 to 0.68. As a result, we still lack an understanding of the factors that affect the quality of the protein abundance predictions. Moreover, like the machine learning approaches above, the existing predictions of protein abundance based on pcGEMs have only been assessed under optimal conditions. Therefore, the problem of predicting protein abundance under sub-optimal growth conditions is particularly difficult since many mechanisms affecting protein abundance (e.g., codon usage (20), RNA levels (21), protein interactions (11), 3' or 5' UTR motifs (22)) are not included in pcGEMs.

To address these issues, we coupled constraint-based modelling with machine learning to predict *in vivo* protein abundances, leading to a multi-model framework termed CAMEL (Coupled Approach of MEtabolic modelling and machine Learning) that leverages the strengths of both approaches. First, CAMEL predicts the protein abundance in different conditions using a constraint-based approach. Next, these protein abundance predictions are employed together with experimentally measured protein abundances to calculate protein reserve ratios. CAMEL then trains machine learning (ML) models to predict the protein reserve ratios using different static features, derived

from the protein sequence, and condition-dependent features, obtained from the pcGEM. Lastly, the predicted protein reserve ratios along with the predicted protein abundances from the pcGEM are used to calculate *in vivo* protein abundances. Once trained, the models of CAMEL can be used to make predictions by using features predicted by the pcGEM instead of data on molecular read-outs from unseen conditions. Our results showed that the CAMEL approach outperformed contending methods on data from two model organisms, *E. coli* and *S. cerevisiae*, highlighting the advantages of a coupling of constraint-based modelling and machine learning approaches. In addition, these applications of CAMEL point to particular proteins and metabolic pathways for which protein abundance is difficult to predict, requiring the consideration of additional features to improve prediction performance in future developments.

Results and Discussion

Machine learning accurately predicts protein reserve ratios

Our first observation is that the predicted abundance, $E_{s,i}$, of an enzyme i in condition s from pcGEMs is usually smaller than the measured *in vivo* abundance $E_{s,i}^{exp}$. This is the case since predictions on $E_{s,i}$ must match the corresponding flux, as $v_j = k_{cat}^{i,j} \cdot E_{s,i}$, and $E_{s,i}$ cannot exceed the measured protein abundance. We term this discrepancy between the measured and predicted protein abundance for an enzyme i in condition s the protein reserve ratio, $\phi_{s,i}$, which is calculated by $\frac{E_{s,i}}{E_{s,i}^{exp}}$.

Since $\phi_{s,i}$ cannot be predicted from pcGEMs alone, we rely on machine learning to train models for ϕ_i, given data on measured protein abundance and predicted protein usage from pcGEMs over multiple conditions (Figure 1). We refer to the collection $\boldsymbol{\phi}$ of reserve ratios over all proteins present in a pcGEM and measured by quantitative proteomic techniques as the distribution of **protein reserve ratios**. Since pcGEMs can be parameterized with *in vitro* measured or with *in vivo* estimated turnover numbers, denoted by k_{cat} and k_{max}^{vivo}, respectively, we performed a comparative analysis of protein reserve ratios in the two parameterization scenarios. Our first observation is that proteins differ with respect to their distributions of reserve ratios, which are dependent on the parameterization of the pcGEMs (Figure 2). For example, the protein reserve ratios of phosphoglycerate kinase in *E. coli* exhibit a narrow distribution for the pcGEM parameterization based on k_{cat}, while the distribution is bimodal for the parameterization with k_{max}^{vivo}. In contrast, the phosphoglycerate kinase in *S. cerevisiae* has the same narrow distribution for both parameterizations. We also observed that the distributions of protein reserve ratios for all investigated enzymes have a heavy tail, due to very large reserve ratios in few conditions.

Next, we used TPOT (23) to optimize machine learning pipelines to predict $\boldsymbol{\phi}$ based on two types of features—static and condition-dependent (Table S1). Static features were obtained from codon usage metrics, while condition-dependent features included enzyme usage and fluxes of the catalyzed reactions predicted by the pcGEM in the two parameterization scenarios (see Material and Methods). Our results indicated that the optimized pipelines for all considered cases exhibited similar structures, composed of two parts: starting with a stacked ensemble of linear and tree-based algorithms and then following with an Extreme Gradient Boosting (XGBoost) algorithm (24), except

for the *S. cerevisiae* model using k_{max}^{vivo} values, whose pipeline included the Extra-trees Regressor (Table S2). Using the trained and optimized pipelines, we then predicted $\boldsymbol{\phi}$ for the validation data set, which had not been used during model training and optimization. For *E. coli*, the machine learning models using either the k_{cat} or k_{max}^{vivo} values showed excellent performance (R_{adj}^2 > 0.9), while the model for *S. cerevisiae* showed worse performance (R_{adj}^2≈ 0.7) (Figure 3). By comparing the R_{adj}^2 values, we found that for *E. coli*, the pipelines with condition-dependent features derived by using k_{max}^{vivo} values outperformed those based on k_{cat} values by 6.7% (Figure 3). For yeast, however, we did not observe a difference in the overall performance of the machine learning models between the two parameterizations.

The reasons for the lower predictive performance for yeast in comparison to *E. coli* could be due to several factors: First, from a modelling perspective, there is more and better-curated knowledge about the metabolism of *E. coli* compared to that of *S. cerevisiae* (25). In addition, the metabolism of yeast is compartmentalized and includes mechanisms for enzyme regulation not accounted for in pcGEMs of any organism (e.g., organellar enzyme pools, post-translational modifications, diffusion effects). Second, from a data perspective, a lack of variability in the protein measurements, resulting from the consideration of samples from similar conditions, can hamper the training of machine learning models. Putting the two perspectives together, the eciML1515 model (1259 enzymes, (15)) includes a larger number of enzymes in comparison to the ecYeast8 model (965 enzymes, (26)), allowing to capture enzymes that may vary more across stress conditions. Lastly, the accuracy of the k_{cat} values in the BRENDA database (27) can also affect the quality of the predictions, given the challenges in measuring them experimentally.

Importance of fluxes for predicting protein reserve ratios

We were also interested in identifying the extent to which fluxes for reactions associated with a protein affect the predictions of the corresponding protein reserve ratios. To this end, we removed these condition-dependent features from the training data set and re-trained the previously obtained TPOT-optimized pipelines using the reduced training data set. We found that the removal of fluxes resulted in reduced predictive performance in all considered scenarios. For instance, the obtained R_{adj}^2 values for *E. coli* were at least 15.4% lower using k_{cat} values and 4.5% lower using k_{max}^{vivo}. For *S. cerevisiae*, there was a 10.8% decrease for the model using k_{cat} values and an 8.1% decrease using k_{max}^{vivo} values (Figure 3). These findings demonstrated that the usage of fluxes as condition-dependent features improved model performance in both organisms, particularly in the case when pcGEMs were parametrized with k_{cat} values. In addition, this shows that the usage of fluxes as features may, to a certain extent, mitigate the effects of model parameterization.

Proteins with particular prediction error of protein reserve ratio are enriched in different pathways

Next, we were interested if there are any distinguishing characteristics between the two groups of proteins -- with low or high relative error in predictions. We considered a protein to show low prediction error if the relative error between the predicted and measured protein reserve ratio was lower than 50%, and a high prediction error if the relative error was higher than 100% in each model

(noting that varying the definitions with lower, or respectively higher, values have a negligible effect on the conclusions).

First, we checked whether the two groups of proteins were enriched in any specific GO terms (28, 29). For *E. coli*, we found that proteins with low prediction error were enriched in biosynthesis of secondary metabolites, biosynthesis of cofactors, terpenoid biosynthesis and riboflavin metabolism for models using either k_{cat} and k_{max}^{vivo} values (Figure S1). For proteins with high prediction error, enriched GO terms shared between models using either k_{cat} or k_{max}^{vivo} values included biosynthesis of secondary metabolites, carbon metabolism, fatty acid metabolism and degradation (Figure S1). The five proteins with the highest relative error (Table S3) for both model parameterizations were part of the same metabolic pathways, namely: glycolysis, fatty acid biosynthesis, beta-oxidation, and cell wall biogenesis.

Similarly, for *S. cerevisiae*, the proteins with low prediction error for the model using k_{cat} and k_{max}^{vivo} values were enriched in the biosynthesis of secondary metabolites, biosynthesis of amino acids, and carbon metabolism. Likewise, proteins with high prediction error were enriched in biosynthesis of secondary metabolites, biosynthesis of amino acids, carbon metabolism and biosynthesis of cofactors for the models using k_{cat} and k_{max}^{vivo} values (Figure S2). The GO term biosynthesis of secondary metabolites encompasses enzymes catalyzing a broad set of reactions, that may explain its significant enrichment in all tested protein sets. The five proteins with the highest relative error (Table S4) showed overlap between the two model parameterizations and were part of purine metabolism, hexose metabolism, and porphyrin-containing compound metabolism. These comparisons demonstrated that the protein reserve ratios did not exhibit distinct patterns based on which metabolic pathway(s) in which the proteins are involved.

To further compare the two groups of proteins, we examined whether the differences in the relative error of predictions were associated with the coefficient of variation (CV) of protein reserve ratios calculated from the models using either k_{cat} and k_{max}^{vivo} values. For all compared scenarios, many proteins with low CV also had high relative errors. However, we did not identify a statistically significant difference in CV between proteins with low and high relative errors in all scenarios except the *E. coli* model using k_{max}^{vivo} values (Pairwise Wilcoxon rank sum test with Bonferroni correction, Figure S3). Taken together, these results indicated that proteins with low or high prediction errors are used in similar metabolic contexts, suggesting that other physiological factors might contribute to the predictive performance of their abundances.

CAMEL accurately predicts *in vivo* protein abundances

Next, we used the predicted distributions of protein reserve ratios, $\widehat{\boldsymbol{\phi}}$, to estimate protein abundance, E_s', which is given by $E_{s,i}' = \frac{E_{s,i}}{\widehat{\phi_{s,i}}}$, where $E_{s,i}$ is the condition-specific protein abundance predicted from the pcGEM. The obtained protein abundance estimates were compared to experimental measurements (30, 31) by calculating the Pearson correlation coefficient of $\log_{10}$-transformed protein abundances for each growth condition separately (shown in Tables S5-S8) as well as the average across all conditions (Figure 4). For *E. coli*, we found a high and significant average value for the Pearson correlation coefficient of 0.987 (p-values < 2.5e-5) over the considered conditions using the models for $\widehat{\boldsymbol{\phi}}$ based on the model parameterized with k_{cat} values, and 0.994 (p-values < 3.1e-11) using the model parameterized with k_{max}^{vivo} values. In *S. cerevisiae*,

using the k_{cat} parameterization, we found a smaller, yet significant average Pearson correlation of 0.483 (p-values < 0.05 for all significant correlations) over the considered conditions; using the pcGEM with k_{max}^{vivo} values resulted in an average Pearson's correlation of 0.513 (p-values < 0.03 for all significant correlations).

Here, too, we took a closer look at proteins with very high relative errors and how their CV compare to proteins with very low relative errors. In *E. coli*, the proteins with the highest relative error showed values larger than 1000% (Table S9). These proteins were part of the chorismate biosynthesis, fatty acid beta-oxidation, pentose phosphate pathway and peptidoglycan biosynthesis when predictions were based on the pcGEM with k_{cat} values. In the case when k_{max}^{vivo} parameterization was employed, these proteins were part of fatty acid biosynthesis, pentose phosphate pathway and purine metabolism. For *S. cerevisiae*, the proteins with the highest errors showed more extreme values than in the case of *E. coli* (Table S10). For both pcGEM parameterizations, the proteins with high error are related to amino acid biosynthesis. Regarding the CV, we observed no significant difference between proteins with very high or very low errors (Figure S4) (pairwise Wilcoxon rank sum test with Bonferroni correction), except for the yeast pcGEM with k_{max}^{vivo} values (p-value < 0.027).

Previous attempts for predicting protein abundance using constraint-based models achieved good predictive performances, but they underestimated *in vivo* protein abundance. The predictions performed by Adadi et al. (17) using the MOMENT approach achieved a Pearson's correlation of 0.84 between predicted protein contents and gene expression data of *E. coli* grown on a glucose minimal medium. We showed that CAMEL with k_{max}^{vivo} parameterized pcGEM resulted in Pearson correlations of over 0.9 for *E. coli* over different conditions, demonstrating considerably improved performance by using the protein reserve ratios to predict protein usage values compatible to the *in vivo* protein usage. In addition, CAMEL does not require the use of any additional molecular read-outs, thus saving valuable resources for making predictions in unseen conditions. We note that the performance of CAMEL is comparable to that of Resource Balance Analysis (RBA) model of *Bacillus subtilis*, achieving an R^2 value of 0.94 (32); however, predictions from RBA models consider more information about transcription, protein translation, translocation, compartmentalization, folding and thermostability, which are difficult to parameterize even in prokaryotic model organisms. CAMEL, on the other hand, depends on pcGEMs, quantitative proteomics data and coding sequences for generating its predictions.

Attempts to predict protein abundance using machine learning have previously achieved good predictive performances but rely on either transcriptomics data or sequence-derived features, which have low correlation to protein content, and are invariable to changes in environmental conditions, respectively (12, 33). For instance, the model constructed by Terai and Asai (22), based on three different algorithms, was trained to predict protein abundance in *E. coli* using sequence-derived features and mRNA structure information. The model performance was assessed by Spearman correlation, which ranged from 0.55 to 0.71. Predictions for *S. cerevisiae* were carried out by the Bayesian network constructed by Mehdi et al. (10), which combined transcriptomics measurements and sequence-derived features, using data from *S. cerevisiae* and *Schizosaccharomyces pombe*. The predictions were evaluated by Spearman correlation, which ranged from 0.61 to 0.77. Ferreira et al. (12) generated a predictive model using only codon usage metrics as features, which achieved a Spearman correlation of 0.74. Although the resulting correlations were higher than those from CAMEL, both studies relied only on data obtained from

optimal growth conditions. However, CAMEL can be effectively used to make predictions of protein abundance in sub-optimal scenarios, which is a notoriously difficult endeavor (34, 35). Given that predicted protein abundance based on the protein reserve ratios are in good agreement with experimental values, these results indicated that the coupling of pcGEMs and machine learning can accurately predict the protein allocation *in vivo*.

By coupling constraint-based approaches and machine learning approaches, we demonstrated that our proposed approach, CAMEL, generated accurate predictions of the protein reserve ratios. In addition, we showed that in combination with constraint-based models, CAMEL leverages these machine learning models to obtain a prediction of protein abundance that overcome notable limitations inherent to pcGEMs (e.g., over-accumulation of enzymes, uncertainty in k_{cat} values). As its most notable merit, CAMEL predicts protein abundance by relying on physiological data, used in the prediction of fluxes, rather than requiring additional omics measurements such as gene expression. Therefore, CAMEL is readily applicable to any condition given that measurements of growth rates and exchange fluxes are available, widening the possibilities for its biotechnological applications (e.g., design of strains). In addition, as our findings demonstrated, CAMEL results in good predictions for protein abundance for sub-optimal conditions, opening the possibility for using these predictions as features to model other complex physiological traits.

Materials and Methods

Data set construction

We obtained the experimental enzyme abundances for *E. coli* over 31 growth conditions from three different studies (19, 36, 37), as reported in Davidi et al. (30). The growth conditions in these experiments range from alternative carbon substrates (e.g., acetate, glucosamine, glucose, glycerol, mannose, pyruvate, and xylose) to chemostats with increasing dilution rates, with glucose as a carbon source. For *S. cerevisiae*, we used the enzyme abundance values from the data set of Chen and Nielsen (31), which contains quantitative proteomics measurements for 30 growth conditions obtained from four studies (31, 38–40), including: optimal batch growth, with glucose as carbon source, nitrogen- and glucose-limited chemostats with increasing dilution rates, stress states achieved with high ethanol and osmolarity, and reduced nutrient availability.

Condition-dependent features are predicted from the pcGEMs and include the enzyme usage and the metabolic fluxes associated with each enzyme i. These features are obtained by solving a linear problem to minimize excess enzyme usage (Eqs. (1)-(8)), detailed below. To predict the enzyme usage distributions, we used the pcGEMs eciML1515 (15) and ecYeast8 (26), from *E. coli* and *S. cerevisiae*, respectively, both without integrated enzyme measurements (batch model). To this end, we minimized excess enzyme usage by solving the following linear programming (LP) problem, termed LP1:

$$\min_{\mathrm{v,E}} E_{s,tot} - \sum_{i=1}^{p} MW_i \cdot E_{s,i} \qquad (1)$$

s. t.

$$\mathbf{N} \cdot \mathbf{v} = \mathbf{0} \qquad (2)$$

$$v_{min,j} \leq v_j \leq v_{max,j}, 1 \leq j \leq r \qquad (3)$$

$$v_{uptake,s} = v_{uptake,s}^{exp} \qquad (4)$$

$$v_{secretion,s} = v_{secretion,s}^{exp} \quad (5)$$
$$v_{bio} = \mu_s \quad (6)$$
$$v_j \leq k_{cat}^{ij} \cdot E_{s,i}, 1 \leq j \leq r, 1 \leq i \leq p \quad (7)$$
$$E_{s,tot} = P_{s,tot}^{exp} \cdot f \cdot \sigma \quad (8)$$

where $E_{s,i}$ is the enzyme i, ranging from 1 to the number of proteins p in the model ($1 \leq i \leq p$), in condition s, MW_i is the molecular weight of an enzyme i, $\mathbf{N}$ is the stoichiometric matrix, $\mathbf{v}$ is the flux distribution vector, v_j is the metabolic flux through reaction j, ranging from 1 to the number of reactions r in the model ($1 \leq j \leq r$), v_{bio} is flux through the biomass reaction, and μ is the specific growth rate, E_{tot} is the total enzyme content, P_{tot}^{exp} is the total protein content, f is the mass fraction of all measured proteins included in the model, and σ is the average *in vivo* enzyme saturation. The constraints applied to the problem are based on the available data about specific growth rates, secretion, and nutrient uptake rates, protein usage and total protein content.

For both organisms, we constrained the models given the measurements of nutrient uptake rates and specific growth rates using data from Davidi et al. (30) and Chen and Nielsen (31). We excluded the experimental conditions that lacked physiological data (e.g., nutrient uptake rates, specific growth rates, or protein content). In addition, we opted to exclude the temperature stress conditions from Lahtvee et al. (38), as temperature stress triggers responses that entail cellular changes that can impact the function of enzymes (41) and require more tailored modelling approaches (42). This filtering resulted in data from 20 conditions for *E. coli*, and 19 conditions for *S. cerevisiae*. For conditions for which the models were overconstrained, we flexibilized the constraints on the uptake and secretion rates by increments of 1% until the measured specific growth rate could be achieved.

We were also interested in whether using pcGEMs integrated with *in vivo* catalytic rates (k_{max}^{vivo}) instead of *in vitro* values (k_{cat}) would impact the performance of the ML models. To this end, we exchanged the integrated k_{cat} values with the k_{max}^{vivo} values, obtained from the compilation by (42) of estimated catalytic rates using pFBA from the Davidi et al. (30) and Chen and Nielsen (31) studies. For both *E. coli* and *S. cerevisiae*, we concatenated the predictions generated from the considered growth conditions into a single data set for each organism. The resulting data set for *E. coli* included 2256 enzymes for the pcGEM using k_{cat} values, and a data set with 2246 enzymes for the pcGEM k_{max}^{vivo} values. For *S. cerevisiae*, the resulting data set included 4596 enzymes for the pcGEM using k_{cat} values, and a data set with 4590 enzymes for the pcGEM k_{max}^{vivo} values.

Training and assessment of machine learning models

To train the machine learning models, we used the Tree-based Pipeline Optimization Tool (TPOT), an automated machine learning tool that optimizes machine learning pipelines to predict the target variable using genetic programming (23). We configured TPOT for regression problems since we aimed to obtain quantitative predictions of the protein reserve ratio, ϕ_i, for each enzyme $i, 1 \leq i \leq p$. For all conditions in each species, we iterated TPOT over 100 generations to identify the optimized pipeline. We used a population size of 50, which is the number of the best pipelines that are predicted by TPOT in one generation that are then carried to next generation, but with randomly altered parameters in their constituent ML algorithms. We separated 80% of the constructed data sets for training and 20% for validation. The training data subset was used through all steps, while the validation data subsets were kept out of the model during TPOT optimization to prevent data

leakage and/or training bias. For each training run, we further applied 10-fold cross-validation to the training data subset. All condition-dependent features were $\log_{10}$-transformed before training.

For performance assessment, we employed the adjusted coefficient of determination score (R^2_{adj}). It is often not possible to retrieve feature importance from TPOT-optimized pipelines, since many of the scikit-learn functions selected by TPOT do not support ranking of feature importance. As a result, we assessed how the condition-dependent features impact the predictions by removing them from the data sets and assessing how the TPOT-optimized pipelines of each data set perform. To further assess the prediction results, we used ϕ and E_s to obtain predictions about *in vivo* enzyme abundances $E_s^{exp\prime}$ which were compared with the experimental value using the Pearson's correlation coefficient. We performed GO enrichment analysis using the clusterProfiler R package (43, 44). The implementation of CAMEL can be found in the GitHub repository: https://github.com/mauricioamf/CAMEL.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001, and by the German Academic Exchange Service (DAAD). P.W. and Z.N. acknowledge the support from the Research Focus Area "Evolutionary Systems Biology" of the University of Potsdam. M.A. and Z.N. acknowledge the support of the Max Planck Society. M.A. and Z.N were supported by the European Union's Horizon 2020 research and innovation programme, project PlantaSYST (SGA-CSA No. 739582 under FPA No. 664620).

References

1. O. T. Schubert, H. L. Röst, B. C. Collins, G. Rosenberger, R. Aebersold, Quantitative proteomics: challenges and opportunities in basic and applied research. *Nature Protocols 2017 12:7* **12**, 1289–1294 (2017).
2. J. Nielsen, Yeast Systems Biology: Model Organism and Cell Factory. *Biotechnol J* **14**, 1–9 (2019).
3. Y. Liu, *et al.*, Proteomics: a powerful tool to study plant responses to biotic stress. *Plant Methods 2019 15:1* **15**, 1–20 (2019).
4. D. Kültz, Evolution of cellular stress response mechanisms. *J Exp Zool A Ecol Integr Physiol* **333**, 359–378 (2020).
5. J. R. Lill, W. R. Mathews, C. M. Rose, M. Schirle, Proteomics in the pharmaceutical and biotechnology industry: a look to the next decade. *https://doi.org/10.1080/14789450.2021.1962300* **18**, 503–526 (2021).
6. N. Pappireddi, L. Martin, M. Wühr, A Review on Quantitative Multiplexed Proteomics. *ChemBioChem* **20**, 1210–1224 (2019).
7. A. Otto, D. Becher, F. Schmidt, Quantitative proteomics in the field of microbiology. *Proteomics* **14**, 547–565 (2014).
8. F. Calderón-Celis, J. R. Encinar, A. Sanz-Medel, Standardization approaches in absolute quantitative proteomics with mass spectrometry. *Mass Spectrom Rev* **37**, 715–737 (2018).
9. W. Torres-García, W. Zhang, G. C. Runger, R. H. Johnson, D. R. Meldrum, Integrative analysis of transcriptomic and proteomic data of Desulfovibrio vulgaris: a non-linear model to predict abundance of undetected proteins. *Bioinformatics* **25**, 1905–1914 (2009).
10. A. M. Mehdi, R. Patrick, T. L. Bailey, M. Boden, Predicting the Dynamics of Protein Abundance. *Molecular & Cellular Proteomics* **13**, 1330–1340 (2014).

11. J. Mergner, *et al.*, Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* **579**, 409–414 (2020).
12. M. Ferreira, R. Ventorim, E. Almeida, S. Silveira, W. Silveira, Protein Abundance Prediction Through Machine Learning Methods. *J Mol Biol* **433**, 167267 (2021).
13. Q. K. Beg, *et al.*, Intracellular crowding defines the mode and sequence of substrate uptake by Escherichia coli and constrains its metabolic activity. *Proc Natl Acad Sci U S A* **104**, 12663–12668 (2007).
14. B. J. Sánchez, *et al.*, Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. *Mol Syst Biol* **13**, 935 (2017).
15. I. Domenzain, *et al.*, Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *Nat Commun* **13**, 1–13 (2022).
16. P. S. Bekiaris, S. Klamt, Automatic construction of metabolic models with enzyme constraints. *BMC Bioinformatics* **21**, 1–13 (2020).
17. 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 Comput Biol* **8**, e1002575 (2012).
18. D. Heckmann, *et al.*, Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nat Commun* **9**, 1–10 (2018).
19. A. Schmidt, *et al.*, The quantitative and condition-dependent Escherichia coli proteome. *Nat Biotechnol* **34**, 104–110 (2016).
20. E. M. Novoa, I. Jungreis, O. Jaillon, M. Kellis, T. Leitner, Elucidation of Codon Usage Signatures across the Domains of Life. *Mol Biol Evol* **36**, 2328–2339 (2019).
21. B. Eraslan, *et al.*, Quantification and discovery of sequence determinants of protein-per-mRNA amount in 29 human tissues. *Mol Syst Biol* **15**, e8513 (2019).
22. G. Terai, K. Asai, Improving the prediction accuracy of protein abundance in Escherichia coli using mRNA accessibility. *Nucleic Acids Res* **48**, e81–e81 (2020).
23. T. T. Le, W. Fu, J. H. Moore, Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics* **36**, 250–256 (2020).
24. T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-Augu**, 785–794 (2016).
25. D. B. Bernstein, B. Akkas, M. N. Price, A. P. Arkin, Critical assessment of E. coli genome-scale metabolic model with high-throughput mutant fitness data. *bioRxiv*, 2023.01.05.522875 (2023).
26. H. Lu, *et al.*, A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat Commun* **10** (2019).
27. A. Chang, *et al.*, BRENDA, the ELIXIR core data resource in 2021: New developments and updates. *Nucleic Acids Res* **49**, D498–D508 (2021).
28. M. Ashburner, *et al.*, Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25 (2000).
29. S. Carbon, *et al.*, The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res* **49**, D325–D334 (2021).
30. D. Davidi, *et al.*, Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro kcat measurements. *Proc Natl Acad Sci U S A* **113**, 3401–3406 (2016).
31. 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**, e2108391118 (2021).

32. A. Goelzer, *et al.*, Quantitative prediction of genome-wide resource allocation in bacteria. *Metab Eng* **32**, 232–243 (2015).
33. H. Li, O. Siddiqui, H. Zhang, Y. Guan, Joint learning improves protein abundance prediction in cancers. *BMC Biol* **17**, 1–14 (2019).
34. S. R. Upadhya, C. J. Ryan, Experimental reproducibility limits the correlation between mRNA and protein abundances in tumor proteomic profiles. *Cell Reports Methods* **2**, 100288 (2022).
35. C. Buccitelli, M. Selbach, mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics 2020 21:10* **21**, 630–644 (2020).
36. K. Valgepea, K. Adamberg, A. Seiman, R. Vilu, Escherichia coli achieves faster growth by increasing catalytic and translation rates of proteins. *Mol Biosyst* **9**, 2344–2358 (2013).
37. K. Peebo, *et al.*, Proteome reallocation in Escherichia coli with increasing specific growth rate. *Mol Biosyst* **11**, 1184–1193 (2015).
38. P. J. Lahtvee, *et al.*, Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst* **4**, 495-504.e5 (2017).
39. R. Yu, *et al.*, Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. *Nat Commun* **11**, 1–12 (2020).
40. F. Di Bartolomeo, *et al.*, Absolute yeast mitochondrial proteome quantification reveals trade-off between biosynthesis and energy generation during diauxic shift. *Proc Natl Acad Sci U S A* **117**, 7524–7535 (2020).
41. G. Li, *et al.*, Bayesian genome scale modelling identifies thermal determinants of yeast metabolism. *Nat Commun* **12**, 1–12 (2021).
42. P. Wendering, M. Arend, Z. Razaghi-Moghadam, Z. Nikoloski, Data integration across conditions improves turnover number estimates and metabolic predictions. *Nature Communications 2023 14:1* **14**, 1–12 (2023).
43. G. Yu, L. G. Wang, Y. Han, Q. Y. He, ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
44. T. Wu, *et al.*, clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).

Figures and Tables

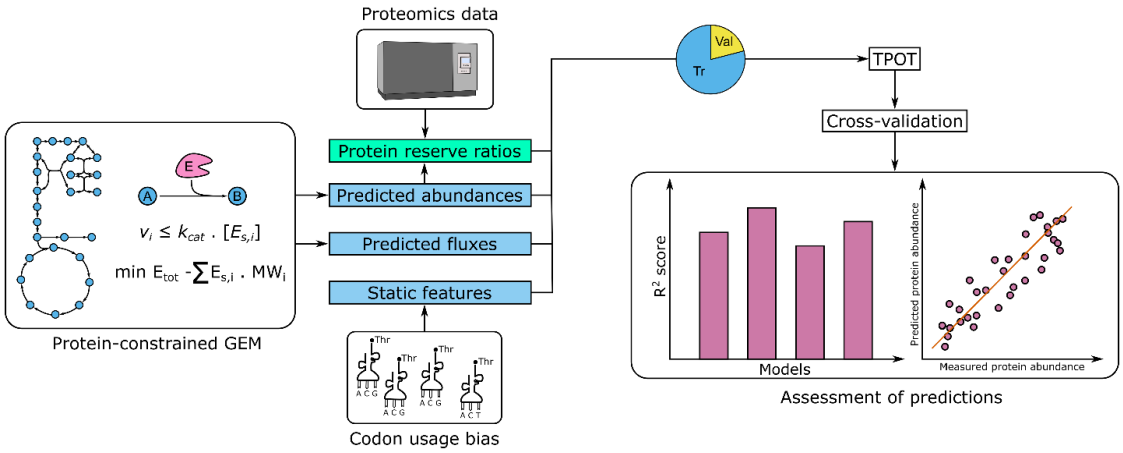

Figure 1. Schematic overview of CAMEL, a constraint-based approach to predict protein abundance. CAMEL uses two types of features, marked as blue boxes: static, obtained from codon usage metrics, and condition-dependent, given by metabolic flux and enzyme usage predicted by constraint-based modelling with pcGEMs. The predicted enzyme usage is employed together with experimental proteomics data to calculate protein reserve ratios that are then used as the target variable for machine learning (green box). The predictive model is obtained by selecting the optimal ensemble of machine learning approaches using TPOT. The performance is assessed by using cross-validation based on regression metrics. Abbreviations: Tr – training; Val – validation.

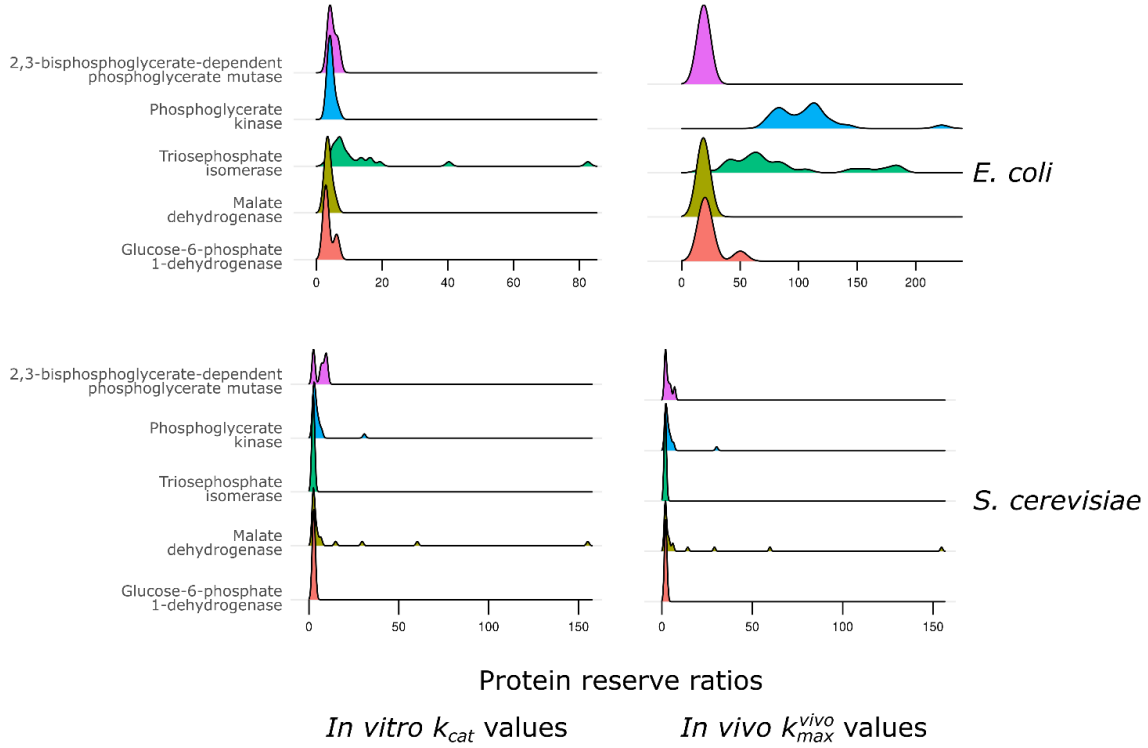

Figure 2. Distribution of protein reserve ratios. The selected proteins take part in central metabolic pathways such as glycolysis, pentose phosphate pathway and citric acid cycle, and are thus present in both organisms. Protein reserve ratios are unitless values.

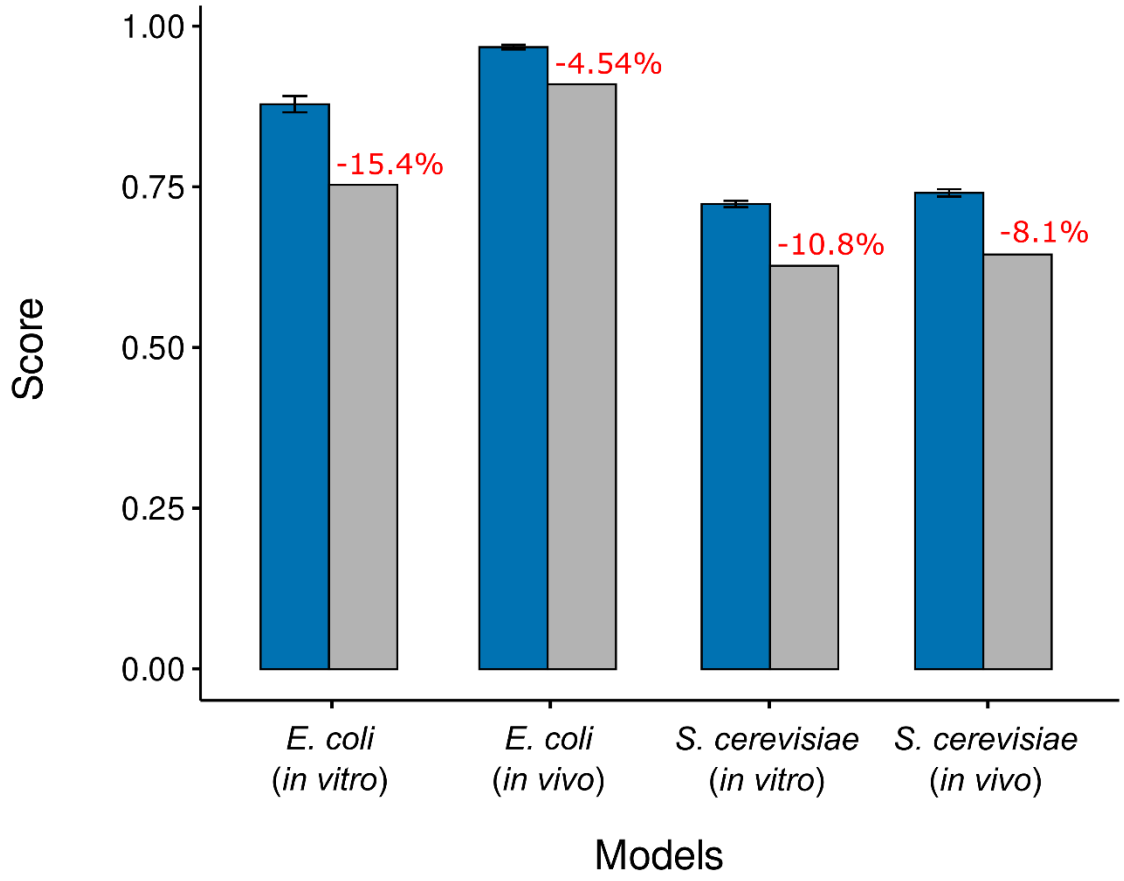

Figure 3. Performance of models from optimized pipelines. The figure shows a comparison of R^2_{adj} scores for the optimized pipelines based on all features (blue) or excluding condition-dependent features (grey) with either *in vitro* k_{cat} or k_{max}^{vivo} values. Negative percentages (red) indicate a decrease in predictive performance compared to scores obtained from the models trained using all features. Error bars indicate R^2_{adj} scores obtained from 10-fold cross-validation.

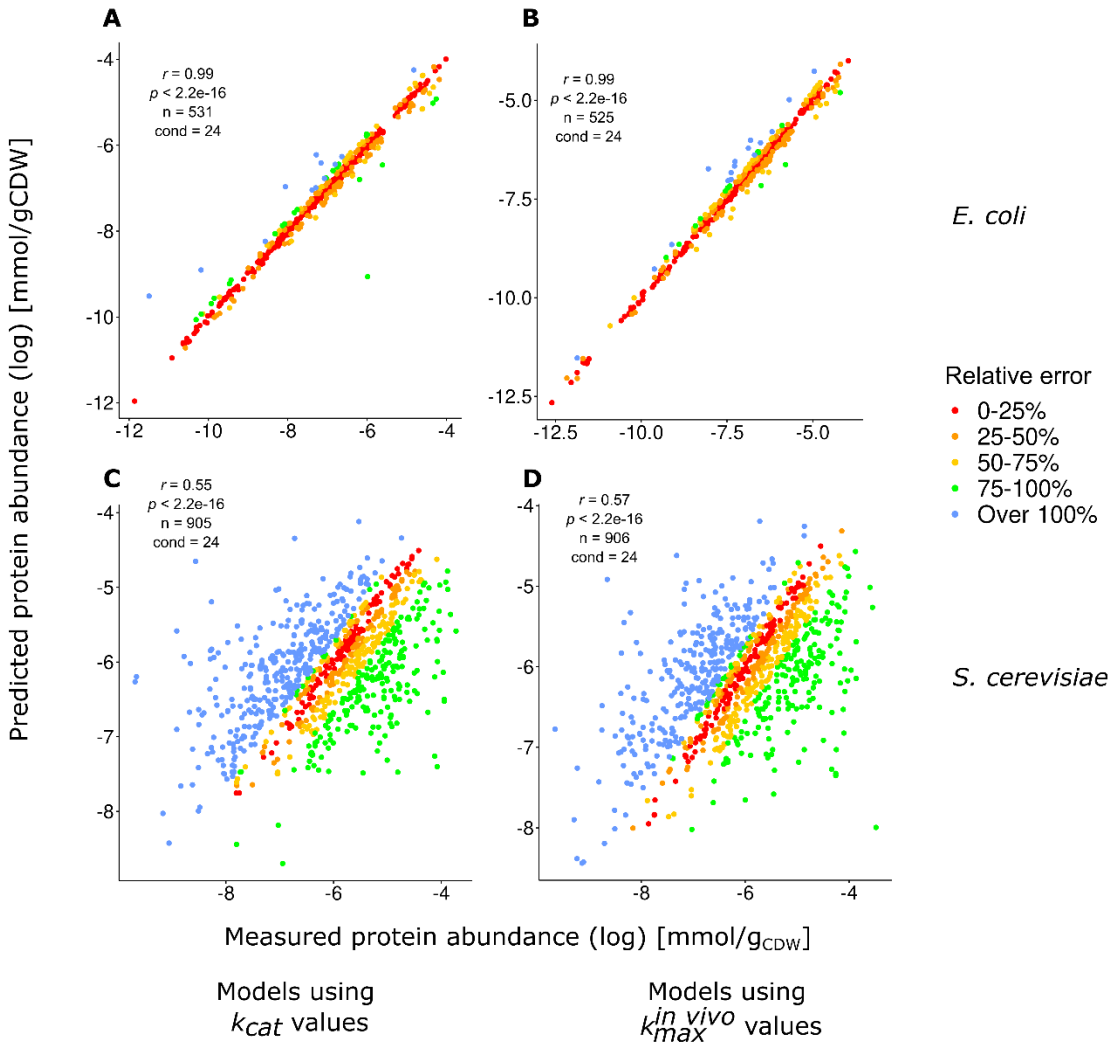

Figure 4. Comparison of measured and predicted protein abundances. The colour scheme indicates the relative error when comparing the measured (E_s^{exp}) and predicted protein abundances ($E_s^{exp'}$). Relative error was calculated using protein abundances in nominal scale. Panel (A) shows predictions using the *E. coli* model with *in vitro* k_{cat} values (A), and in vivo k_{max}^{vivo} values (B). The panels (C) and (D), depict the prediction performance of the *S. cerevisiae* model with k_{cat} values (C) and k_{max}^{vivo} values (D).