Estimates of *in vivo* turnover numbers by simultaneously considering data from multiple conditions improve metabolic predictions

Philipp Wendering[1†], Marius Arend[1,2†], Zahra Razaghi-Moghadamkashani[2], Zoran Nikoloski[1,2*]

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

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

* Zoran Nikoloski

[†] authors contributed equally

Email: nikoloski@mpimp-golm.mpg.de

Author Contributions: PW, MA performed research and analyzed data, PW contributed code for PRESTO approach, MA assessed model performance and performed statistical analysis, PW, MA, ZR, ZN designed research, PW, MA, ZN wrote the paper.

Competing Interest Statement: The authors declare no competing interest.

Abstract

Turnover numbers characterize a key property of enzymes, and their usage in constraint-based metabolic modeling is expected to increase prediction accuracy of diverse cellular phenotypes. *In vivo* turnover numbers can be obtained by ranking of estimates obtained by integrating reaction rate and enzyme abundance measurements from individual experiments; yet, their contribution to improving predictions of condition-specific cellular phenotypes remains elusive. Here we show that available *in vitro* and *in vivo* turnover numbers lead to poor prediction of condition-specific growth rates with protein-constrained models of *Escherichia coli* and *Saccharomyces cerevisiae,* particularly in the ultimate test scenario when protein abundances are integrated in the model. We then demonstrate that *in vitro* turnover numbers can be corrected via a constraint-based approach that simultaneously leverages heterogeneous physiological data from multiple experiments. We find that the resulting estimates of *in vivo* turnover numbers lead to improved prediction of condition-specific growth rates, particularly when protein abundances are used as constraints, and are more precise than the available *in vitro* turnover numbers. Therefore, our approach provides the means to decrease the bias of *in vivo* turnover numbers and paves the way towards cataloguing *in vivo* kcatomes of other organisms.

Significance Statement

Integration of turnover numbers in protein-constrained metabolic models has provided insights in enzyme allocation and can improve prediction of metabolic phenotypes. While *in vivo* turnover numbers are estimated by ranking the outcomes from the integration of condition-specific proteomics data and reaction rates, simultaneous consideration of physiological read-outs over multiple conditions has not been considered, yet. Here we designed and tested a constraint-based approach that leverages heterogeneous data to improve estimates of *in vivo* turnover numbers in model unicellular organisms. We showed that the resulting turnover numbers are more precise than *in vitro* estimates from public databases and lead to improved prediction of condition-specific growth. The approach is readily applicable to non-model organisms and paves the way to documenting their kcatomes.

Main Text

Introduction

Genome-scale metabolic models (GEMs) together with advances in constrained-based modeling have led to improved understanding of how cellular resources are used to fulfill different cellular tasks (1–3). Recent advances are largely propelled by the development of protein-constrained GEMs (pcGEMs) in which the catalytic capacities of individual enzymes are linked to the allocation of enzyme abundances (4). Such models have led to more accurate predictions of maximum specific growth rates on different carbon sources (5–7), flux distributions (7), and other complex phenotypes (8) in *Escherichia coli* and *Saccharomyces cerevisiae*. However, the development of pcGEMs critically depends on integration of organism-specific enzyme turnover numbers, k_{cat} (9), comprising the kcatome of an organism.

Measuring the kcatome of an organism based on *in vitro* characterization is limited due to impossibility to purify specific enzymes, lack of availability of substrates, and knowledge of required cofactors, such that their relevance for studies of *in vivo* phenotypes remains questionable (10, 11). Proxies for turnover numbers, referred to as maximal *in vivo* catalytic rates, can be obtained by combining constrained-based approaches for flux prediction with measurements of protein abundance under different growth conditions or genetic modifications (12–14). The results from this approach, that entails ranking of condition-specific estimates that use individual data sets, have shown that the proxies for *in vivo* turnover numbers generally concur with *in vitro* k_{cat} values in *E. coli* (12). However, applications with data from *S. cerevisiae* (13, 15) and *A. thaliana* (16) indicated that these proxies for *in vivo* turnover numbers do not reflect *in vitro* measurements. Another approach to estimate the kcatome relies exclusively on machine and deep learning methods that use variety of features of enzymes (e.g. network-based, structure-based, and biochemical) (17–19), resulting in predictive models that can explain up to 70% of the variance in turnover numbers obtained *in vitro*.

The estimates of turnover numbers are integrated into metabolic models by different constraint-based approaches that have been grouped into coarse-grained (e.g. MOMENT (5), sMOMENT (20), eMOMENT (21), and GECKO (7)) and fine-grained (e.g. resource balance analysis (1, 2) and ME-models (3)). Of these, GECKO (7) has been adopted in several recent studies due to the elegantly structured formulation of the protein constraints. In addition, GECKO allows for integration of protein contents and correction factors that account for mass fraction of enzymes (f) included in the model as well as average *in vivo* saturation (σ) of all enzymes, facilitating the development of condition-specific models.

While data-driven estimation of *in vivo* turnover numbers improves the coverage of k_{cat} values in pcGEMs, the available estimates usually lead to over-constrained models when combined with

measurements of total protein content. In addition, the bias in the estimates of k_{cat} values can be determined in the ultimate test, in which they are used alongside constraints from enzyme abundance to predict growth—scenario that has not yet been examined. Here, we propose PRESTO (for protein-abundance-based correction of turnover numbers)—a scalable constraint-based approach to correct turnover numbers by matching predictions from pcGEMs with measurements of cellular phenotypes – simultaneously – over multiple conditions. As a constraint-based approach, PRESTO facilitates the investigation of variability of the proposed corrections. We show that predictions of growth by pcGEMs of *S. cerevisiae* with turnover numbers corrected by PRESTO are more accurate than those based on the models that include k_{cat} values corrected based on a contending heuristic that relies on enzyme control coefficients (22). We also demonstrate that same conclusions hold when enzyme abundances are integrated into the *E. coli* model using PRESTO. Therefore, PRESTO paves the way to broaden the applicability of pcGEMs for organisms with biotechnological applications and to arrive at genotype-specific estimates of the kcatome.

Results and Discussion

PRESTO – protein-abundance-based correction of turnover numbers

For a given data set of protein abundances over a set of conditions, the enzymes with turnover numbers in a pcGEM can be partitioned into three groups. For instance, a data set of protein abundances that was recently used to estimate *in vivo* turnover numbers in *S. cerevisiae* (15) includes 45%, 41%, and 14% measured over all, at least one (but not all), and none of the 27 used conditions, respectively. Therefore, there is then a different data support for correcting the k_{cat} values of these classes of proteins. PRESTO relies on solving a linear program that minimizes a weighted linear combination of the average relative error for predicted specific growth rates and the correction of the initial turnover numbers integrated in the pcGEM model (Fig. 1, Methods). It further employs K-fold cross validation (here, K = 3) with 10 repetitions while ensuring steady state and integrating protein constraints for proteins measured over all conditions (Fig 1, Methods). The training set of conditions is used to generate a single set of corrected k_{cat} values, that provide estimates of *in vivo* turnover numbers, by using the respective protein abundances. The resulting corrected k_{cat} values are in turn used to determine the relative error of predicted specific growth rate for each condition in the test set using flux balance analysis with the pcGEM, while only applying a constraint on total protein content and measured uptake rates. The relative error of predicted specific growth rate along with the sum of introduced corrections are lastly used to select the value for the tuning parameter λ in the objective function of PRESTO, as done in machine learning approaches that rely on regularization.

PRESTO outperforms a contending heuristic for correction of turnover numbers in *S. cerevisiae*

To determine the performance of PRESTO and compare it to that of contending heuristics, we used a data set comprising protein abundances and exchange fluxes from 27 diverse conditions, as supported by the principal component analysis (Fig S1). Application of PRESTO with a pcGEM of *S. cerevisiae* with initial *in vitro* turnover numbers obtained from BRENDA resulted in a mean relative error of 0.68 from the cross-validation procedure, yielding a correction of on average 213 turnover numbers (Fig S2a). For the *S. cerevisiae* pcGEM, we found a value of 10^{-7} for the parameter λ in the PRESTO objective provides the optimal trade-off between both the relative error and the sum of introduced corrections (see Methods). Moreover, we observed a high overlap between the sets of proteins with corrected turnover numbers in the cross-validation (average Jaccard distance of 0.07 (Fig S2b, c), suggesting that the integrated data from different conditions point at a specific subset of enzymes that need to be corrected to improve performance of growth prediction.

Unlike PRESTO, GECKO implements a heuristic for correction of turnover that is based on the control coefficient of a protein in each condition (Fig S3) (22). The control coefficient of a protein is determined by increasing the turnover number by 1000-fold and scoring the effect on the predicted specific growth rate. The proteins are then ranked in a decreasing order of their control coefficients, and the turnover number of the first enzyme in the list is changed to the maximum value found in BRENDA for this enzyme. This procedure is repeated with the remaining enzymes until the pcGEM predicts a growth rate that is at most 10% smaller than the measured specific growth rate for that condition or no additional k_{cat} value that strongly constraints the solution can be found. In contrast to this procedure, PRESTO corrects at once the turnover numbers of multiple enzymes that are measured in all investigated conditions by simultaneously leveraging the data from the different conditions. As a result, rather than deriving condition-specific corrected k_{cat} values, which are difficult to use in making predictions for unseen scenarios, PRESTO results in a single set of corrected k_{cat} values.

Next, we compared the performance of PRESTO with the heuristic implemented in GECKO in three modeling scenarios that consider: (i) only condition-specific total protein content, (ii) both total protein content and uptake constraints, and (iii) additional constraints from abundances of enzymes measured in all conditions. For corrections of turnover number from PRESTO, we observed that the relative error spans the range from 0.15 to 0.88 in the least constrained scenario (i) (Fig 2a) and from 0.67 to 0.98 in the most constrained scenario (iii) (Fig 2c). In contrast, the relative error with the corrections of turnover numbers from the GECKO heuristic is in the range from 0.94 to 1.00 in scenario (iii) (Fig 2c). In addition, in scenario (iii), the relative error in the case of the GECKO

heuristic for each condition is larger than the relative error of the PRESTO predicted growth rate (Fig 2c). We observed that predictions from flux balance analysis, considering enzyme abundances, without a constraint on the total protein content, led to average relative error of 0.69 with k_{cat} values corrected according to PRESTO and 0.99 with k_{cat} values corrected according to GECKO (Table S1A). Further, considering the constraint on the total protein content resulted in average relative errors of 0.87 and 0.99 with k_{cat} values corrected according to PRESTO and GECKO, respectively.

We also performed sensitivity analysis by investigating a smaller value, of 10^{-10}, for the weighting factor λ used in the PRESTO objective. We found that when the weighting factor is 10^{-10} (at which the total corrections of the initial k_{cat} values plateaus), the relative errors from PRESTO can be further decreased to 0.69 with the constraint on the total protein content, with no effects on the other findings (Fig S2a). We also note that the relative error lies in the range from 0.35 to 0.80 over the considered weighting factors in the range from 10^{-14} to 10^{-1}. Together, these results demonstrated that k_{cat} values corrected according to PRESTO provide better model performance than the values obtained by the contending heuristic in the case of *S. cerevisiae*.

PRESTO provides precise corrections and estimates of *in vivo* turnover numbers

In the following, we investigated the precision of the corrected k_{cat} values from the application of PRESTO to data and pcGEM model of *S. cerevisiae*. To this end, we determined the range that the correction of the k_{cat} value of each enzyme can take, while fixing the relative error in specific growth rate and total corrections from the optimum of PRESTO (Methods). Moreover, we complemented this analysis by sampling corrected k_{cat} values that achieve the optimum of PRESTO with two values of the weighting factor λ of 10^{-7} and 10^{-10}.

In the case of the corrected k_{cat} values for *S. cerevisiae* with a weighting factor of 10^{-7}, we found that the k_{cat} values with the largest corrections are more precisely determined (Fig S4). In addition, the corrections show an average Euclidean distance to the respective mean of 4.88 s^{-1}, indicating that the corrected values are more precise than the values in BRENDA, that on average show a value of 27.54 s^{-1} to the mean (Fig S5). Importantly, while k_{cat} values with smaller correction showed larger variability, the 25 and 75 percentiles of the sampled corrections for 42 enzymes are concentrated around those resulting from PRESTO. Repeating the analysis with a weighting factor of 10^{-10} showed that the larger total corrections of the initial k_{cat} values resulted in also larger variability for the corrections over all k_{cat} (Fig S6). Here, too, for 62 enzymes the 25 and 75 percentiles of the sampled corrections are concentrated around those resulting from PRESTO. Therefore, we concluded that the corrections from PRESTO are precise and can be used in downstream analyses.

Pathways of carbon metabolism are enriched in enzymes with corrected turnover numbers in pcGEM model of *S. cerevisiae*

To check which metabolic processes are more likely to require correction of *in vitro* k_{cat} values, we next conducted an enrichment analysis based on the KEGG pathway terms linked to corrected k_{cat} values (Methods). The most prominent pathway in this analysis was the synthesis of secondary metabolites, particularly the synthesis of cofactors and terpenoids (Fig 3a). However, several terms linked to central carbon metabolism, such as: the tricarboxylic acid cycle (i.e. citrate cycle) and oxidative phosphorylation, were also significantly enriched. Interestingly, amino acid synthesis was the only term linked to nitrogen metabolism that came up in this analysis, although many pathways of nitrogen metabolism were among the tested terms. This analysis suggested that particularly *in vitro* turnover numbers in carbon metabolism need to be corrected, due to underestimation in *in vitro* assays.

Next, we aimed to identify the extent to which the corrected k_{cat} values differ between PRESTO and the GECKO approach. To this end, we determined the intersection of enzymes with k_{cat} values corrected manually (7), by PRESTO, and by the GECKO heuristic. For this comparison, we considered the union of all condition-specific corrected k_{cat} values from the GECKO approach. With the weighing factor $\lambda = 10^{-7}$, PRESTO adapted the k_{cat} values of 48% of enzymes corrected by the GECKO heuristic (Fig 3b, Table S1B). We did not find a significant Spearman correlation ($\rho_S = 0.19$, $P = 0.385$) between the log-transformed k_{cat} values in this intersection (Fig 3c), owing to the different principles employed in the two procedures. In addition, the intersection between enzymes with manually corrected values and those corrected by the GECKO heuristic is higher than with PRESTO. This is expected since the manual curation partly aimed at correcting the most constraining turnover numbers (7).

We also compared the k_{cat} values adjusted by GECKO against values for the maximum apparent catalytic rate obtained by parsimonious FBA (pFBA) using the same proteomics data (15) (Fig S7 a, b). We confirmed the low correspondence ($\rho_S = 0.23$) between the k_{cat} values obtained from BRENDA, included in the GECKO model without manual modifications, and the maximum apparent catalytic rates. As expected, the correspondence of the maximum apparent catalytic rates to the turnover numbers corrected based on PRESTO was higher ($\rho_S = 0.34$). To investigate how turnover numbers obtained from pFBA perform as model parameters, we also generated a pcGEM in which BRENDA values were substituted by the maximum apparent catalytic rates from (15), whenever available. In scenarios without enzyme abundance values, this model performed worse than that including the k_{cat} values corrected by PRESTO as well as the model combining the maximum of all condition-specific GECKO corrections (Fig S8 a, b). In the enzyme abundance constrained scenario, the model with turnover numbers obtained from pFBA performed slightly

better than GECKO but still only achieved a minimum relative error of 0.93, which is larger than 0.71 resulting from PRESTO (Fig S8 c). These results demonstrated the value of the PRESTO in combining the genome-scale coverage of BRENDA with *in vivo* proteomics chemostat measurements to obtain less biased estimates of *in vivo* k_{cat} values.

Application of PRESTO with protein-constrained model of *E. coli* metabolism

To demonstrate the applicability of PRESTO across species, we deployed it with a pcGEM of *E. coli* (eciML1515) (22, 23). To this end, we used a large dataset comprising 31 different growth conditions (12, 24–26). Due to the lack of data on nutrient exchange rates, the same GAM value (i.e., 75.55 $\frac{mmol}{gDWh}$) was used across all conditions. Similarly, we used the same value for total protein content since condition-specific measurements were not available (see Methods).

By applying a three-fold cross-validation, we found the optimal value for the λ parameter to be 10^{-5} (Fig S9a). This value was associated with an average relative error of 1.95 (overall average: 3.32) and 73 corrected turnover numbers, while on average 156 k_{cat} values were corrected across all explored values for λ. On average, the Jaccard distance between cross-validation folds was 0.13 (Fig S9b), while the average Jaccard distance between unique sets of enzymes with corrected turnover numbers for each λ parameter was three-fold larger (0.4, Fig S9c). Thus, the corrected k_{cat} values among cross-validation folds for each λ are more similar (maximum Jaccard distance of 0.29). Moreover, the union of the set of enzymes with corrected k_{cat} values can remain similar over a range of chosen λ parameters up to four orders of magnitude (Fig S9c), demonstrating the robustness of the method.

The performance of PRESTO was assessed and compared to GECKO using scenarios (i) and (iii) since no condition-specific uptake rates were available. With default uptake rates, the relative error for predicted growth ranged between 0.01 and 8.56 in the less constrained scenario (i). Further, we obtained relative errors between 0.01 and 0.88 for the more constrained scenario (iii), when using the k_{cat} values corrected by PRESTO (Fig 4 a, b). In contrast, when using the k_{cat} values from the GECKO approach, the relative error was in the range between 0.01 and the 4.89 for scenario (i) and between 0.89 and 0.99 for scenario (iii). In this scenario, too, we observed that the relative error using k_{cat} values corrected by GECKO was consistently larger than the relative error resulting from the single set of corrected k_{cat} values obtained by PRESTO (Fig 4a, b).

The sum of introduced k_{cat} corrections reached a plateau at 10^{-11} for the weighting factor λ in the PRESTO objective. We found that the relative validation error at this value was 5.26, which is 2.7-fold larger than the relative error obtained using the optimal λ. Hence, when only the pool constraint is applied, allowing for more and larger corrections results in a decrease of the overall relative error

in PRESTO, at the cost of predicting growth that is consistently higher than the observed. This observation is in line with the small number of corrections introduced by the GECKO approach, where only the pool constraint is considered. We conclude that the prediction performance of the eciML1515 model was improved by using turnover numbers corrected by PRESTO only when enzyme abundances are integrated.

To assess the precision of the introduced k_{cat} corrections, we performed variability analysis and sampling (see Methods) of the introduced corrections to the initial k_{cat} values for two values of the weighting factor λ, namely 10^{-5} and 10^{-11}. We observed that the 25 and 75 percentiles enclose a narrow interval around the values resulting from PRESTO (Fig S10) and are thus not evenly distributed across the respective interval determined by the variability analysis. We further noted that here, the predictions of smaller δ are generally more precise than the large corrections ($\delta \geq p_{50}$), which span 2.12 orders of magnitude (small δ (< p_{50}): 1.83, Fig S10). However, we also observed that the precision decreased when more corrections were allowed in PRESTO. This further justified our choice for the optimal parameter λ, which results in a lower number of 73 corrections compared to 204 at $\lambda = 10^{-11}$, and moreover guarantees more precise estimates (Fig S11). In conclusion, the application of PRESTO is not limited to a single species but presents a versatile tool for the correction of turnover numbers across species.

In contrast to the observations made in *S. cerevisiae* we found that a model parameterized with the turnover numbers estimated by pFBA (12) outperformed both PRESTO and GECKO in the modelling scenario where no enzyme abundance constrains are taken into account (Fig S8d). This is due to the fact, that pFBA, in contrast to PRESTO and GECKO, allows for the decrease of *in vitro* k_{cat} values, in turn leading to more accurate predictions. However, in the scenario with enzyme abundance constrains PRESTO predicts growth rates closer to the experimental observation in 87% of the conditions (Fig S8e). Thus, in this scenario the integration of information from different modelling conditions achieved in PRESTO serves to obtain *in vivo* k_{cat} value that perform better than the k_{app}^{max} approach applied by (12).

The low number of corrections introduced by GECKO leads to an overlap of only 3 enzymes (75%) whose k_{cat} values were also corrected by PRESTO (Table S1C, Fig S12a). The pathway enrichment analysis for PRESTO corrections at $\lambda = 10^{-5}$ identifies amino acid and secondary metabolite synthesis as significantly enriched terms among the enzymes with corrected turnover numbers (Fig S12b). These results argue for a systematic underestimation of *in vivo* turnover numbers in these pathways compared to *in vitro* data, irrespective of the investigated organism. However, the lower order KEGG pathway terms enriched in *E. coli* do not overlap with the ones found in *S. cerevisiae*. Here, fatty acid metabolism and the synthesis of hydrophobic amino acids

are among the pathways requiring correction of turnover numbers. This is due to the differences in modeling the pathways in lipid metabolism for the two organisms.

Conclusion

Characterization of enzyme parameters that can inform models of reaction rates is key to expanding and further propelling the usage of metabolic models in diverse biotechnological applications. While the generation of pcGEMs has facilitated the integration of more biophysically relevant constraints, it necessitates access to estimates of turnover numbers as key enzyme parameters. The bias in the available *in vitro* and *in vivo* turnover numbers can readily be assessed by considering the accuracy of growth predictions based on the integration of enzyme abundances. Indeed, we showed that condition-specific growth rates cannot be reliably predicted with pcGEMs of *S. cerevisiae* and *E. coli* when available *in vitro* and *in vivo* estimates of turnover numbers are used. While these estimates can be adjusted by considering recently proposed Bayesian statistical learning (18), this approach has not considered the total enzyme abundance information from proteomics measurements as a constraint.

To resolve this issue, we proposed PRESTO, a constraint-based approach that simultaneously considers heterogeneous physiological read-outs and enzyme abundance measurements to correct *in vitro* turnover numbers. Through a series of comparative analyses, we demonstrated that the *in vivo* estimates of turnover numbers from PRESTO ultimately increase the prediction accuracy of condition-specific growth for the two organisms when enzyme abundance data are integrated in the corresponding pcGEMs. We also showed that the maximal *in vivo* catalytic rates, obtained by ranking of condition-specific estimates that use proteomics and fluxomics data, are more highly (but modestly) correlated to estimates from PRESTO than to *in vitro* turnover numbers. Owing to the constraint-based formulation of PRESTO, we also determined the precision of the *in vivo* estimates of turnover numbers. Previous studies have shown that even for the well-studied model organism *Sacharomyces cerevisiae*, only 52% of enzyme turnover numbers in the pcGEM can be obtained from organism-specific *in vitro* measurements (22). Using organism unspecific k_{cat} values for parameterization and correction of pcGEMs, as done in the GECKO pipeline, assumes that enzyme kinetic properties are comparable within one EC number class (27, 28). However, we did not identify clear differences between EC classes, down to the second digit, when considering the distribution of k_{cat} similarities within EC classes (Fig13). Indeed, it has been reported that EC class plays only a minor in the prediction of turnover numbers (19) and show stronger similarity with concordant GO categories (29). . Interestingly, our findings show that the *in vivo* turnover numbers obtained from PRESTO are more centered around the means of the of the EC classes for both studied organisms (Fig. S5). Together, these findings demonstrated PRESTO can be readily used to decrease the bias of *in vitro* and *in vivo* estimates of turnover numbers. This paves the way for

employing the outcome of PRESTO and future extensions towards effectively predicting the kcatome from available protein sequences.

Materials and Methods

Experimental data

S. cerevisiae We made use of a dataset gathered by (15) from four different studies (30–33), which included protein abundance data as well as measured growth or dilution rates and nutrient exchange fluxes. Exchange fluxes missing in certain conditions were set to 1000 mmol/gDW/h if the nutrient was present in the used culture media. We further augmented this data set by adding total protein content measurements from the original studies. For subsequent analyses, we used the maximum abundance of each protein over all replicates per experimental condition. Similarly, we used the average value for growth rates and nutrient exchange rates. Since no measurement of total cellular protein mass was available for the two conditions evaluated in the Di Bartolomeo study (33), we used the maximum protein content measured across the remaining conditions for these conditions (i.e., $0.67\ g/gDW$). Moreover, we excluded three temperature stress conditions (i.e., Lahtvee2017_Temp33, Lahtvee2017_Temp36, Lahtvee2017_Temp38) from the analysis since temperature can have a large effect on the catalytic activity of an enzyme. Gene names in the proteomics dataset were translated to UniProt identifiers using the batch retrieval service of the UniProt REST API (34).

E. coli We used a dataset comprising 31 experimental conditions, which was gathered by Davidi and colleagues and augmented by Xu et al. (12, 14) from three publications (24–26). Here, too, we used the maximum protein abundance over all replicates. Due to the absence of total protein content measurements in two of the original studies, we relied on the maximum protein content measured in the Valgepea study (i.e., $0.61\ g/gDW$) to be used for all conditions. Since precise data on nutrient uptake rates were only given for a few conditions, we assigned a default upper bound of 1000 mmol/gDW/h to all nutrients contained in the M9 medium with additional carbon sources as specified. Gene identifiers were translated to UniProt similar as for *S. cerevisiae*.

Model preparation

The proposed approach aims at parsimonious correction of turnover values in genome-scale enzyme-constraint metabolic models using measured protein abundances. Therefore, it is important to consider differential association between enzymes and reactions, i.e., isozymes, enzyme complexes, and promiscuous enzymes. We decided to use the GECKO formalism (7), which deals with these problems elegantly by directly encoding the required information into the stoichiometric matrix. The genome-scale metabolic models for *S. cerevisiae* (YeastGEM v.8.5.0)

and *E. coli* (iML1515) were obtained from the yeast-GEM and ecModels GitHub repository, respectively [(22, 35), https://github.com/SysBioChalmers ; accessed on 22.08.2021]. For subsequent steps, functions of the COBRA v3.0 toolbox (36) and GECKO2.0 toolbox (22) were employed, of which several functions were adapted for our purposes.

To arrive at raw enzyme-constraint models for both organisms, the GECKO2.0 model enhancement pipeline was adapted to allow the k_{cat} correction procedure to be omitted. Moreover, any manual corrections of turnover numbers were excluded from model generation. In the process of adapting the raw pcGEM to the respective experimental conditions for both organisms, the GAM value per condition was fitted using the *scaleBioMass* function of GECKO2.0, based solely on the condition-specific nutrient exchange rates, and returning the minimum 9 $\frac{mmol}{gDWh}$) or maximum (161 $\frac{mmol}{gDWh}$) interval boundary if reached (only *S. cerevisiae*). Furthermore, we omitted enzyme abundances, which were not measured across all experiments as the approach proposed here only works for enzymes with measured abundances ($E_{measured}$).

PRESTO approach

In the design of PRESTO, we modified the enzyme mass-balance constraints of the augmented stoichiometric matrix, created by GECKO, from

$$-\frac{1}{k_{cat}^{ij}} v_i + e_i = 0 \tag{1}$$

to inequality constraints that use the measured protein abundance directly and further assume a single turnover number per enzyme i over all catalysed reactions $\left(k_{cat_i}^{min} = \min k_{cat_i}.\right)$:

$$\sum_{r \in GPR(E_i)} v_r \leq k_{cat_i}^{min} \cdot [E_i] \,. \tag{2}$$

We justify making this assumption based on our observation that most enzymes in the *S. cerevisiae* model are associated with no more than four reactions. Further, the vast majority of enzymes are assigned a single unique turnover number even though they catalyse multiple reactions (Fig S14).

We then introduced a correction factor δ, which is added to each k_{cat} if the protein abundances for the underlying enzyme were available:

$$\sum_{r \in GPR(E_i)} v_r \leq \left(k_{cat_i}^{min} + \delta_i\right) E_i,\ \forall i \in E_{measured} \,. \tag{3}$$

The value for each δ was constrained by the fold change ε with respect to $k_{cat_i}^{min}$ and a cut-off value K^{max}, which denotes the maximum allowed k_{cat} value.

To find a biologically relevant minimal set of adaptations with respect to the sum of δ, we minimized the weighted sum of the average relative error, ω, between measured (μ^{exp}) and predicted growth rates (v_{bio}) over all experimental conditions C, and the average δ:

$$\min_{v,\delta,\omega} \frac{1}{|C|} \sum_{j \in conditions} |\omega_j| + \frac{\lambda}{|E_{measured}|} \sum_{i \in E_{measured}} \delta_i \ . \tag{4}$$

Finally, the linear programming formulation of the k_{cat} correction in PRESTO is the following:

$$\min_{v,\delta,\omega} \frac{1}{|C|} \sum_{j \in conditions} |\omega_j| + \frac{\lambda}{|E_{measured}|} \sum_{i \in E_{measured}} \delta_i$$

subject to

$$\mathbf{N}\mathbf{v_j} = \mathbf{0},\ 1 \leq j \leq |C| \tag{5}$$

$$\sum_{r \in GPR(E_i)} v_r \leq \left(k_{cat,i}^{min} + \delta_i\right) E_{ij},\ \forall i \in E_{measured}$$

$$\mathbf{v_j^{min}} \leq \mathbf{v_j} \leq \mathbf{v_j^{max}} \tag{6}$$

$$\delta_i \leq (\varepsilon - 1) \cdot k_{cat,i}^{min} \tag{7}$$

$$k_{cat,i}^{min} + \delta_i \leq K^{max} \tag{8}$$

$$v_{bio,j} \cdot \omega_j \geq \mu_j^{exp} - v_{bio,j} \tag{9}$$

$$v_{bio,j} \cdot \omega_j \geq v_{bio,j} - \mu_j^{exp} \tag{10}$$

$$\omega \leq \theta,\ \delta \geq 0\ .$$

The value for ω was bound from above by a value θ, which was set to 0.6. Further, ε was set to 10^5 and the value for K^{max} was set to 57,500,000 s^{-1}.

The parameter λ controls the trade-off between both minimization objectives (see Eq. 4). As λ is unknown and may also be condition- and model-specific, it was fitted using a 3-fold cross-validation scheme, which was repeated for 10 iterations. To this end, we scanned a log-scale interval between 10^{-14} and 10^{-1}. In each iteration, we performed k_{cat} corrections on two folds of condition-specific

models and validated the obtained corrections on the remaining fold of condition-specific models. The relative errors (e_r) and the sum of δ (i.e., Δ) were then used to calculate the scores s_λ, which helped us choose the optimal value for λ:

$$s_\lambda = \frac{1}{10}\sum_{\tau=1}^{10} \frac{e_{r,\lambda,\tau} - e_{r,\lambda,\tau}^{min}}{e_{r,\lambda,\tau}^{max} - e_{r,\lambda,\tau}^{min}} \cdot \frac{\log_{10}\frac{\Delta_{\lambda,\tau}}{\Delta_{\lambda,\tau}^{min}}}{\log_{10}\frac{\Delta_{\lambda,\tau}^{max}}{\Delta_{\lambda,\tau}^{min}}}. \tag{11}$$

In addition to the optimal λ, we also compared our results to a second λ, where the sum of δ reached a plateau ($\lambda = 10^{-10}$ for *S. cerevisiae* and $\lambda = 10^{-11}$ for *E. coli*, Figs S2a and S8a).

Variability analysis for δ

While PRESTO considers multiple experimental conditions to find a set of universal corrections for k_{cat} values, it does not provide an exhaustive view over all possible solutions to this problem. To assess the precision of the corrections, we first performed a variability analysis for δ to find the minimum and maximum possible values. To guarantee that a solution of equal quality is found with respect to the previously determined sum of δ and the relative errors to experimentally measured growth rates (i.e., ω^{opt}), corresponding constraints were added to arrive at the following linear programming problem:

$$\min/\max_{v,\delta,\omega} \delta_i$$

s.t.

$$\mathbf{N}\mathbf{v_j} = \mathbf{0},\ 1 \le j \le |C|$$

$$\sum_{r \in GPR(E_i)} v_r \le \left(k_{cat,i}^{min} + \delta_i\right)E_{ij},\ \forall i \in E_{measured}$$

$$\mathbf{v_j^{min}} \le \mathbf{v_j} \le \mathbf{v_j^{max}}$$

$$\delta_i \le (\varepsilon - 1) \cdot k_{cat,i}^{min}$$

$$k_{cat,i}^{min} + \delta_i \le K^{max}$$

$$v_{bio,j} \cdot \omega_j \ge \mu_j^{exp} - v_{bio,j}$$

$$v_{bio,j} \cdot \omega_j \ge v_{bio,j} - \mu_j^{exp}$$

$$0.99 \cdot \omega_j^{opt} \leq \omega_j \leq 1.01 \cdot \omega^{opt} \tag{12}$$

$$\Delta - 10^{-3} \leq \sum_{i \in E_{measured}} \delta_i \leq \Delta + 10^{-3} \tag{13}$$

$$\omega \leq \theta,\ \delta \geq 0\ .$$

As the distribution within the obtained min/max intervals can be skewed, we sampled 10,000 points within the obtained intervals. For uniform random sampling, we created random vectors of corrections δ^* within the determined intervals and projected them onto the solution space by minimizing the distance of δ to the respective random vector. Therefore, we updated the objective of the program above:

$$\min_{v,\delta,\omega} \sum_{i \in E_{measured}} |\delta_i - \delta_i^*|\ . \tag{14}$$

To ensure reproducibility and compatibility with the COBRA toolbox (36), we solved all optimization problems using the *optimizeCbModel* of the COBRA toolbox. Within this environment, we used the Gurobi solver v9.1.1 (37) but we note that any other supported solver can also be used. As we observed numerical instability of the problems in some cases, we decreased the feasibility tolerance (i.e., *feasTol* parameter) for the COBRA solver to 10^{-9} for all predictions.

Validation of corrected models

We used the adapted GECKO pipeline (fitting a condition specific GAM; excluding manual k_{cat} adaptions) to obtain models with k_{cat} values adapted according to the objective control coefficient heuristic. We note that, when no manual modifications were introduced to the *Saccharomyces cerevisiae* models, the k_{cat} adaption of the GECKO pipeline would stop because no objective control coefficient above the threshold of 0.001 could be found and corrected models would still be below the predicted growth error tolerance of 10%. To compare the predictive performance of PRESTO and GECKO corrected models, the models were adapted with the same condition-specific GAM, biomass reaction and total protein content, $P_{tot.}$ Additionally, PRESTO models where constrained using the same condition-specific saturation rate, σ and enzyme mass fraction, f, as obtained from the GECKO pipeline. In contrast to the GECKO formulation, we did not subtract the mass of measured enzymes from the total protein pool constraint but instead introduced the measured protein concentration as upper bound on the enzyme usage reaction, E_i, in the respective scenario. This formulation still guarantees that the mass of all used enzymes is lower or equal to the approximated cellular protein pool according to

$$\sum_i E_{i,j} \cdot M_i \leq P_{tot,j} \cdot f \cdot \sigma_j, \tag{15}$$

where M is the respective molecular weight of the protein. By considering measured and unmeasured enzymes in (15) we do not have to change *f* and use the same factor as for the scenario where no protein abundance measures are used (7). Maximum growth was predicted in three different constraint scenarios: (i) using only the protein pool constrain and default uptake rates (1000 mmol/gDW/h), (ii) using the pool constrain and experimentally measured uptake rates, (iii) using the previous constraints plus the absolute enzyme abundance.

The two studies which generated *in vivo* k_{cat} estimates from pFBA (12, 15)calculated a single value per reaction irrespective of the presence of isoenzymes. Thus, to parameterize the raw pcGEM (containing only uncorrected BRENDA values) we substituted the k_{cat} values of all isoenzyme reactions with the same estimate provided in the study. Reactions catalyzed by complexes where not corrected. Since PRESTO and the pFBA studies provide a single condition independent model, we generated a condition independent GECKO model by following the maximum over all conditions approach: For the comparisons in Fig S7 the condition wise GECKO models were aggregated into a single union model in which for each reaction the maximum k_{cat} value was used.

Pathway enrichment analysis

The KEGG pathway terms (38), associated with each enzyme that was measured in all conditions, were acquired using the KEGG REST API. The one-sided p-value, p, for significant enrichment of a pathway term among the enzymes with corrected k_{cat} values was calculated using the hypergeometric density distribution:

$$p(x) = 1 - \sum_{i=1}^{x-1} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}} . \tag{16}$$

Only KEGG pathway terms associated to at least two corrected enzymes were taken into consideration. The p-values associated with all tested pathway terms were corrected for a false discovery rate of 0.05 using the Benjamini-Hochberg correction (39).

Funding

P.W. and Z.N. would like to thank the Research Focus Group “Evolutionary Systems Biology” of University of Potsdam for funding. Z.N., M.A., and Z.R. would like to thank the Max Planck Society for funding. Z.R. was supported by the European Union’s Horizon 2020 research and innovation programme grant 862201 (to Z.N.) (this publication reflects only the author’s view and the Commission is not responsible for any use that may be made of the information it contains).

Code Availability

All code that was used to generate the results of this study, including the PRESTO method, are available at https://github.com/pwendering/PRESTO.

References

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

2. J. A. Lerman, *et al.*, In silico method for modelling metabolism and gene product expression at genome scale. *Nat. Commun.* **3**, 929 (2012).

3. E. J. O'Brien, J. A. Lerman, R. L. Chang, D. R. Hyduke, B. Ø. Palsson, Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. *Mol. Syst. Biol.* **9**, 693 (2013).

4. Y. Chen, J. Nielsen, Mathematical modeling of proteome constraints within metabolism. *Curr. Opin. Syst. Biol.* **25**, 50–56 (2021).

5. 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).

6. 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.* **104**, 12663–12668 (2007).

7. 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).

8. C. Malina, R. Yu, J. Bjorkeroth, E. J. Kerkhoven, J. Nielsen, Adaptations in metabolism and protein translation give rise to the Crabtree effect in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2112836118 (2021).

9. A. Nilsson, J. Nielsen, B. O. Palsson, Metabolic Models of Protein Allocation Call for the Kinetome. *Cell Syst.* **5**, 538–541 (2017).

10. K. van Eunen, B. M. Bakker, The importance and challenges of *in vivo*-like enzyme kinetics. *Perspect. Sci.* **1**, 126–130 (2014).

11. P. Labhsetwar, M. C. R. Melo, J. A. Cole, Z. Luthey-Schulten, Population FBA predicts metabolic phenotypes in yeast. *PLoS Comput. Biol.* **13**, e1005728 (2017).

12. D. Davidi, *et al.*, Global characterization of in vivo enzyme catalytic rates and their correspondence to *in vitro* k_{cat} measurements. *Proc. Natl. Acad. Sci.* **113**, 3401–3406 (2016).

13. D. Heckmann, *et al.*, Kinetic profiling of metabolic specialists demonstrates stability and consistency of *in vivo* enzyme turnover numbers. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 23182–23190 (2020).

14. R. Xu, Z. Razaghi-Moghadam, Z. Nikoloski, Maximization of non-idle enzymes improves the coverage of the estimated maximal in vivo enzyme catalytic rates in Escherichia coli. *Bioinformatics* **37**, 3848–3855 (2021).

15. 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**, 2108391118 (2021).

16. A. Küken, K. Gennermann, Z. Nikoloski, Characterization of maximal enzyme catalytic rates in central metabolism of *Arabidopsis thaliana*. *Plant J.* **103**, 2168–2177 (2020).

17. P. Zikmanis, I. Kampenusa, Relationships between kinetic constants and the amino acid composition of enzymes from the yeast Saccharomyces cerevisiae glycolysis pathway. *Eurasip J. Bioinforma. Syst. Biol.* **2012** (2012).

18. F. Li, *et al.*, Deep learning based kcat prediction enables improved enzyme constrained model reconstruction. *bioRxiv [Preprint]* (2021) https:/doi.org/10.1101/2021.08.06.455417 (March 18, 2022).

19. 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).

20. P. S. Bekiaris, S. Klamt, Automatic construction of metabolic models with enzyme constraints. *BMC Bioinformatics* **21**, 19 (2020).

21. P. Wendering, Z. Nikoloski, Genome-Scale Modeling Specifies the Metabolic Capabilities of *Rhizophagus irregularis*. *mSystems* **7**, e01216-21 (2022).

22. I. Domenzain, *et al.*, Reconstruction of a catalogue of genome-scale metabolic models with enzymatic constraints using GECKO 2.0. *bioRxiv [Preprint]*, 2021.03.05.433259 (2021).

23. J. M. Monk, *et al.*, iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* **35**, 904–908 (2017).

24. 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).

25. K. Peebo, *et al.*, Proteome reallocation in *Escherichia coli* with increasing specific growth rate. *Mol. Biosyst.* **11**, 1184–1193 (2015).

26. A. Schmidt, *et al.*, The quantitative and condition-dependent *Escherichia coli* proteome. *Nat. Biotechnol.* **34**, 104–110 (2016).

27. A. Bar-Even, *et al.*, The Moderately Efficient Enzyme: Evolutionary and Physicochemical Trends Shaping Enzyme Parameters. *Biochemistry* **50**, 4402–4410 (2011).

28. D. Davidi, L. M. Longo, J. Jabłońska, R. Milo, D. S. Tawfik, A Bird's-Eye View of Enzyme Evolution: Chemical, Physicochemical, and Physiological Considerations. *Chem. Rev.* **118**, 8786–8797 (2018).

29. Z. Mao, H. Ma, iMTBGO: An Algorithm for Integrating Metabolic Networks with Transcriptomes Based on Gene Ontology Analysis. *Curr. Genomics* **20**, 252–259 (2019).

30. 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).

31. R. Yu, *et al.*, Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. *Nat. Commun.* **11**, 1881 (2020).

32. 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**, e65722 (2021).

33. 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).

34. A. Bateman, UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

35. H. Lu, *et al.*, A consensus *S. cerevisiae* metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. *Nat. Commun.* **10**, 3586 (2019).

36. L. Heirendt, *et al.*, Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* **14**, 639–702 (2019).

37. L. Gurobi Optimization, Gurobi Optimizer Reference Manual (2022).

38. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

39. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).

40. L. Jeske, S. Placzek, I. Schomburg, A. Chang, D. Schomburg, BRENDA in 2019: a European ELEXIR core data resource. *Nucleic Acids Res.* **47**, D542–D549 (2019).

Figures

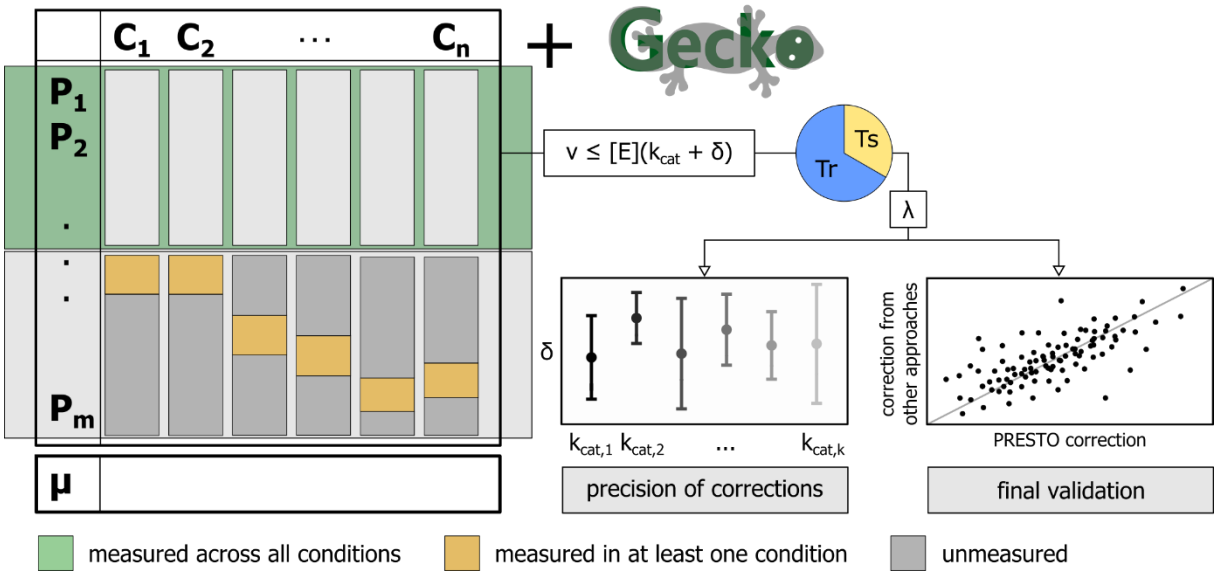

Fig 1. Schematic overview of the PRESTO approach for k_{cat} correction. The approach uses a GECKO-formatted pcGEM (7) for an organism containing turnover numbers from BRENDA (40). Using available data from n experimental conditions, n condition-specific models are generated using nutrient uptake rates and protein contents. PRESTO then uses data on abundances for the enzymes measured across the n investigated conditions and solves a linear program that minimizes a weighted sum of two objectives, the relative error to measured growth rates and the sum of positive k_{cat} corrections, δ. The optimal weighting factor, λ, which modulates the trade-off between the two objectives, is then determined by cross-validation, choosing the parameter, which is associated with the lowest average relative error. Using the optimal value for λ, PRESTO combines all models for the experimental conditions to find a k_{cat} correction for each enzyme with measured abundance. Last, the precision of δ values are assessed by variability analysis as well as by sampling and corrected k_{cat} values are validated by comparing them to values obtained from other approaches.

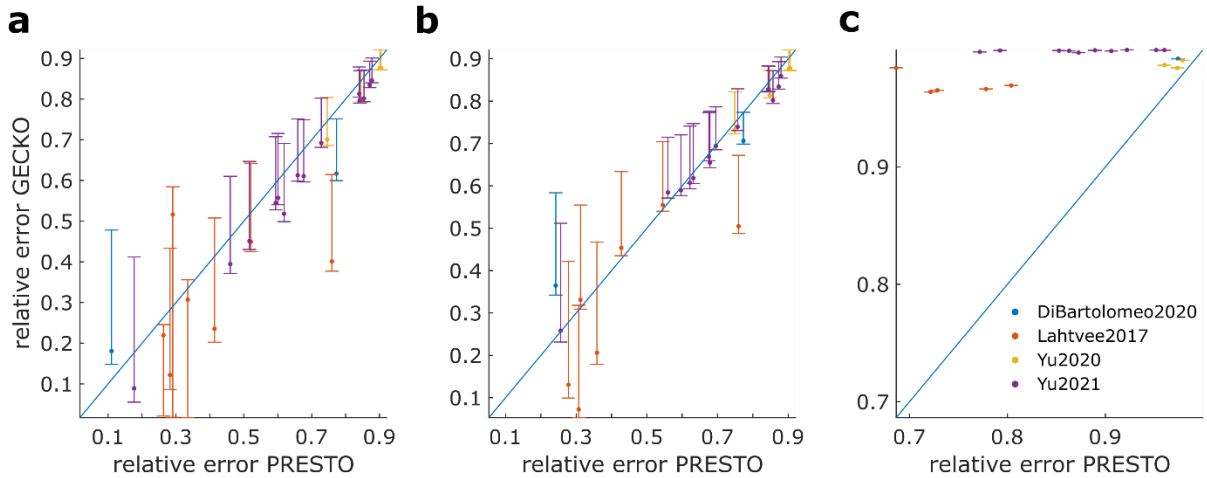

Fig 2. Comparison of predicted growth of *S. cerevisiae* from pcGEMs with k_{cat} corrections from GECKO and PRESTO. Condition-specific pcGEMs with corrected k_{cat} values generated using GECKO and were used to predict the growth rate in each condition. The error bars along the y-axis denote the range of relative errors obtained from the models resulting from the corrected (condition-specific) k_{cat} values based on GECKO. The dot in the interval indicates the mean. The x-axis shows the relative error of predicted growth rate from the PRESTO model ($\lambda = 10^{-7}$) by using the single set of corrected k_{cat} values in the respective pcGEM. **(a)** Only the measured total protein pool was used to constrain the solution and condition-specific uptake rates were bounded by 1000 $\frac{mmol}{h\ gDW}$; **(b)** measured uptake rates were also considered; **(c)** abundances of enzymes measured in all conditions were used as additional constraints. The colour code indicates the study from which experimental measurements where obtained (see legend). The compared pcGEMs in each condition used the same respective biomass coefficients, GAM, σ, and $Ptot$ values (Methods).

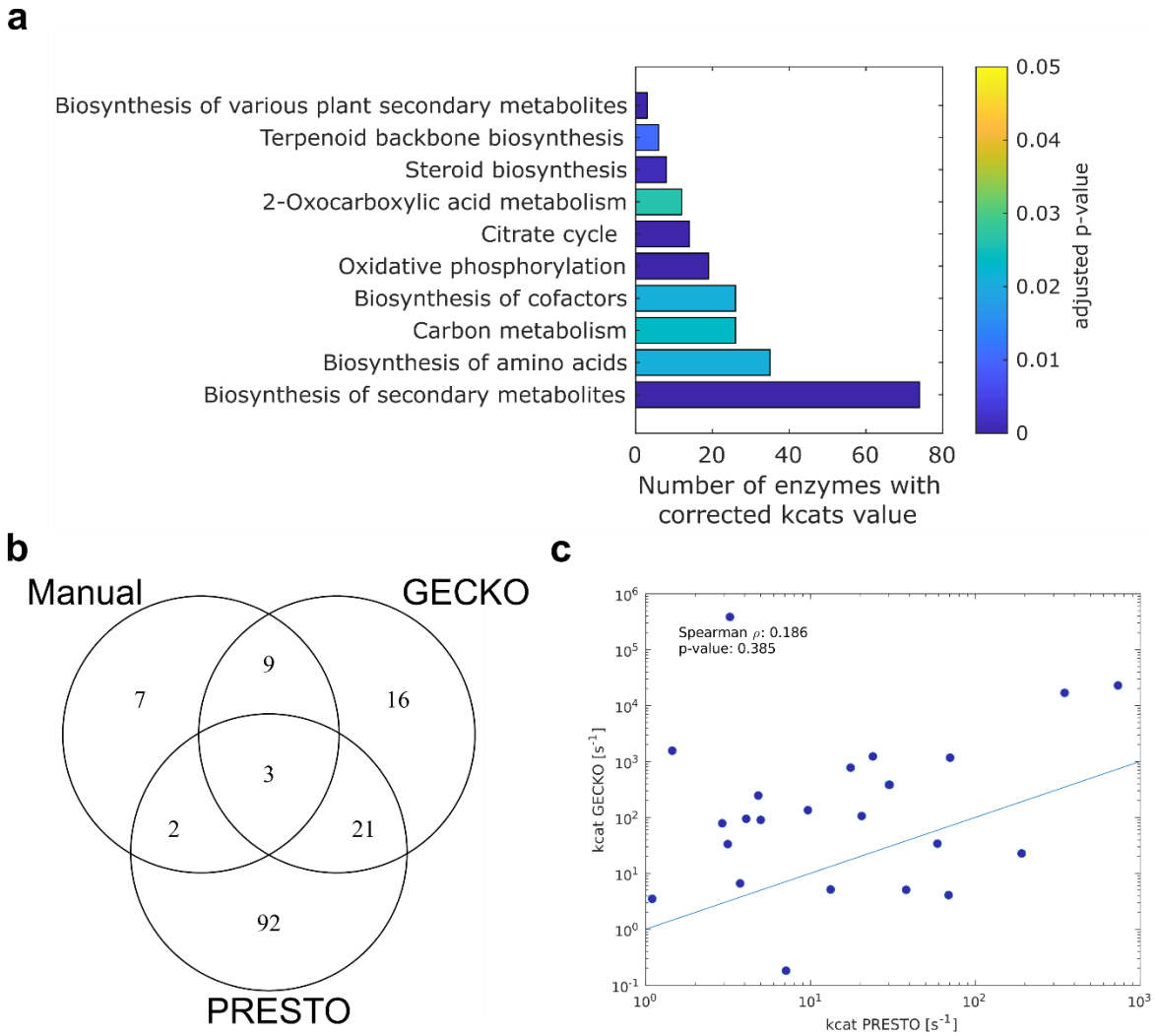

Fig 3. Comparison of enzymes with corrected k_{cat} values by both GECKO and PRESTO. (a) KEGG Pathway terms significantly enriched in the set of enzymes corrected by PRESTO ($\lambda = 10^{-7}$) in the *S. cerevisiae* model. The x-axis gives the number of corrected enzymes linked to the given term. The one-sided p-value was calculated using the hypergeometric density distribution and was corrected using the Benjamini-Hochberg procedure (39). **(b)** Venn diagram showing the overlap of enzymes whose k_{cat} values were manually corrected (7) ("Manual"), automatically corrected by the GECKO heuristic in any of the conditions ("GECKO"), or corrected by PRESTO ("PRESTO"). **(c)** Log-transformed k_{cat} values corrected using both the GECKO heuristic and PRESTO are not associated (Spearman correlation coefficient of 0.186, p-value = 0.385).

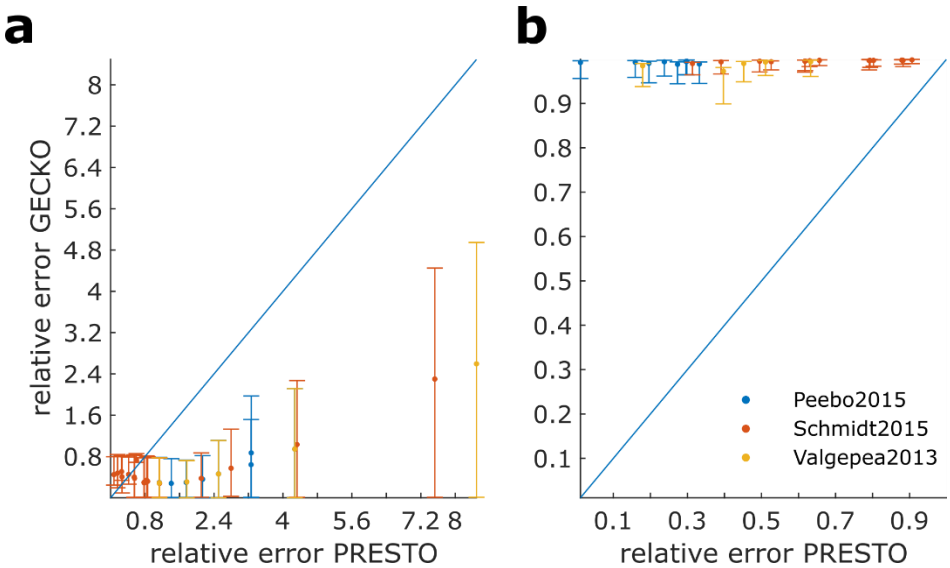

Fig 4. Comparison of predicted growth of *E. coli* from pcGEMs with k_{cat} corrections from GECKO and PRESTO. Condition-specific pcGEMs with corrected k_{cat} values generated using GECKO were used to predict the growth rate in each condition. The bars along the y-axis denote the range of relative errors obtained from the models resulting from the corrected (condition-specific) k_{cat} values based on GECKO. The dot in the interval indicates the mean. The x-axis denotes the relative error of predicted growth rate from the PRESTO model ($\lambda = 10^{-5}$) by using the single set of corrected k_{cat} values in the respective pcGEM. **(a)** Only the measured total protein pool was used to constrain the solution and condition-specific uptake rates were bounded by 1000 $\frac{mmol}{h\ gDW}$; **(b)** abundances of enzymes measured in all conditions were used as additional constraints. The colour code indicates the study from which experimental measurements where obtained (see legend). The compared pcGEMs in each condition used the same respective biomass coefficients, GAM, σ, and P_{tot} values (Methods).