Proteome Regulation Patterns Determine *Escherichia coli* Wild-Type and Mutant Phenotypes

© Tobias B. Alter,a* Lars M. Blank,a Birgitta E. Eberta,b

aInstitute of Applied Microbiology (iAMB), Aachen Biology and Biotechnology (ABBt), RWTH Aachen University, Aachen, Germany
bAustralian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, Brisbane, Australia

**ABSTRACT**

It is generally recognized that proteins constitute the key cellular component in shaping microbial phenotypes. Due to limited cellular resources and space, optimal allocation of proteins is crucial for microbes to facilitate maximum proliferation rates while allowing a flexible response to environmental changes. To account for the growth condition-dependent proteome in the constraint-based metabolic modeling of *Escherichia coli*, we consolidated a coarse-grained protein allocation approach with the explicit consideration of enzymatic constraints on reaction fluxes. Besides representing physiologically relevant wild-type phenotypes and flux distributions, the resulting protein allocation model (PAM) advances the predictability of the metabolic responses to genetic perturbations. A main driver of mutant phenotypes was ascribed to inherited regulation patterns in protein distribution among metabolic enzymes. Moreover, the PAM correctly reflected metabolic responses to an augmented protein burden imposed by the heterologous expression of green fluorescent protein. In summary, we were able to model the effects of important and frequently applied metabolic engineering approaches on microbial metabolism. Therefore, we want to promote the integration of protein allocation constraints into classical constraint-based models to foster their predictive capabilities and application for strain analysis and engineering purposes.

**IMPORTANCE**

Predictive metabolic models are important, e.g., for generating biological knowledge and designing microbes with superior performance for target compound production. Yet today’s whole-cell models either show insufficient predictive capabilities or are computationally too expensive to be applied to metabolic engineering purposes. By linking the inherent genotype-phenotype relationship to a complete representation of the proteome, the PAM advances the accuracy of simulated phenotypes and intracellular flux distributions of *E. coli*. Being equally computationally lightweight as classical stoichiometric models and allowing for the application of established *in silico* tools, the PAM and related simulation approaches will foster the use of a model-driven metabolic research. Applications range from the investigation of mechanisms of microbial evolution to the determination of optimal strain design strategies in metabolic engineering, thus supporting basic scientists and engineers alike.

**KEYWORDS**

constraint-based modeling, enzyme kinetics, metabolic engineering, protein allocation, transcriptional control, *Escherichia coli*

For many decades, metabolic models have been developed to describe, unravel, and understand the drivers of microbial phenotypes. In their simplest forms, these models quantitatively connect observable phenomena such as carbon source consumption and biomass formation, leading to seminal empirical growth laws such as the Monod equation (1). In general, coarse-grained models aid in explaining the dependencies between intracellular processes and corresponding phenotypes (2–7). Constraint-based modeling techniques facilitate the prediction of growth rates and by-product secretion, as well as the investigation of metabolic flux distributions solely based on
the stoichiometry of biochemical reaction networks and an appropriate cellular objective function (8–12). The development of genome-scale constraint-based models (GEM) fostered the investigation of fundamental biological phenomena (9, 13, 14), the systematic analysis of complex omics data sets (15–18) and the suggestion of favorable genetic perturbations for the overproduction of desired chemicals (19–23).

The utilization of GEMs has made valuable contributions to systems biological analyses and metabolic engineering. However, the GEM’s predictive capabilities of cellular phenotypes strongly rely on ad hoc capacity bounds on key reactions (15, 24), without which basic phenomena such as overflow metabolism are not observable in silico. The consideration of additional cellular processes and properties in metabolic reconstructions resolved these predictive insufficiencies of GEMs. In this manner, macromolecular expression (ME) models couple metabolism to gene expression by linking enzyme concentrations to metabolic reactions and accounting for the transcriptional and translational processes leading to enzyme expression (25). ME models simultaneously simulate maximum growth and substrate uptake rates and the underlying responses on the mRNA level, as well as the corresponding gene expression profiles at metabolic steady state (26). Thus, ME models facilitate holistic insights into intracellular processes and how they are affected by environmental, biochemical, or genetic perturbations (27–29), while reliably informing about corresponding flux distributions. As advantageous ME models are for correct predictions at the flux or phenotypic level, their detail and complexity can be cumbersome for future applications in strain design approaches. Many other approaches exist that add various resource or spatial constraints to classical stoichiometric models to investigate otherwise undisclosed growth laws (30–36). However, most of them are either similarly computationally expensive or demand a tremendous number of mostly unknown parameters.

Cellular protein allocation and its regulation have previously been suggested as the main drivers of metabolic phenomena and a key process behind bacterial growth laws (3, 5–7, 37–39). The incorporation of protein constraints in GEMs exploits the principles of protein allocation as a fundamental growth law. It simultaneously allows for the use of established, tractable, and intuitive constraint-based modeling methods. By dividing the limited proteome into three growth-variant sectors representing (i) ribosomal proteins, (ii) anabolic enzymes, and (iii) catabolic enzymes and one invariant housekeeping protein sector, the constrained allocation flux balance analysis (CAFBA) framework computes the optimal partitioning between these protein sectors and the correspondingly weighted flux rates to reach maximum growth (37, 39). Thus, CAFBA accounts for the trade-off between a limited protein availability for biosynthesis and growth and enables a quantitative prediction of pathway usages under various conditions, particularly suboptimal growth yields during overflow metabolism. The integration of enzyme kinetics into a GEM of Saccharomyces cerevisiae and Escherichia coli in the form of explicit enzymatic constraints on flux rates facilitated the utilization of proteomics data (40). The respective GECKO framework (GEM with enzymatic constraints using kinetic and omics data) gives detailed insights into metabolic realizations based on proteome measurements and predicts growth phenomena even without augmented protein concentration data.

Here, we introduce an approach that consolidates protein allocation and enzymatic constraints on metabolic fluxes of an E. coli GEM. The resulting protein allocation model (PAM) represents the major protein sectors and their various ratios with changing growth conditions (41). The PAM predicts experimentally observed phenotypes and intracellular flux distributions at maximum growth and on various growth media. Besides the wild-type behavior, the PAM’s predictive ability is demonstrated for strains harboring gene deletions or expressing heterologous proteins. In line with previous studies, we emphasize the fundamental role of protein allocation in steering microbial metabolism. Therefore, the PAM framework is envisioned as a practical approach to improve the rational design of microbial production strains, as it enables rapid dry-lab screening of design and cultivation strategies with improved reliability.
RESULTS

Accounting for the total proteome in genome-scale metabolic models. A key challenge for microbes is distributing the limited proteins to the intracellular processes in a way that allows for maximum growth under a given environmental condition. While the concentration of ribosomes efficiently meets the demand of protein biosynthesis under mildly nutrient-limited to unlimited conditions (5, 42, 43), the protein household for energy and biomass precursor production generally contains unutilized and underutilized enzymes (44) to allow for a flexible response to environmental changes. Three major protein sectors cover this condition-dependent proteome and represent (i) translational protein, including ribosomes, (ii) metabolically active enzymes, and (iii) un- as well as underutilized enzymes (3). We will refer to the latter as the unused enzyme sector for simplicity. The fourth protein sector covers the housekeeping proteins, whose abundance is constant under any growth condition. Consequently, this sector cannot cause condition-dependent phenotypes. The biomass synthesis equation covers its constant demand, but the protein allocation toward the housekeeping sector does not need to be modeled separately.

To quantitatively account for the condition-dependent protein allocation, we modeled and added each relevant protein sector independently to the E. coli K-12 MG1655 GEM iML1515 (45) (Fig. 1). We refer to the resulting model as the protein allocation model (PAM). The mass concentration, $\phi$, of each sector is a linear function of one or more inherited variables or fluxes of iML1515. They have been partly fitted to experimental proteomic data as depicted in Table 1 and are described in detail in Materials and Methods.

The active enzyme sector. The active enzymes sector covers the protein demand of all enzymatic reactions of the iML1515 model in a GECKO fashion (40). GECKO introduces...
enzyme mass balances to classical stoichiometric models and couples the manifestation of metabolic fluxes to enzyme abundances. That is, each modeled enzyme concentration within this sector is linearly dependent on the flux rate of the catalyzed metabolic reaction. The linear relations are based on a simplified rate law of reversible Michaelis-Menten reactions, thus on simple mass action kinetics, which neglects the nonlinear reversibility, enzyme saturation, and regulation factors (38). With this simplification, enzymes are assumed to operate at their maximal velocity, at which the turnover number $k_{cat}$ describes the relation between enzyme concentration and flux rate. Hence, in silico metabolic fluxes are only limited by the enzyme’s maximum capacity. While this formulation can lead to biologically more feasible solutions (40), enzymes do not generally operate at their maximum velocity (46, 47). Therefore, in the PAM context, the turnover number describes an enzyme’s maximum capacity within a metabolic network operating under unlimited growth conditions, that is, under maximum growth and substrate uptake rate. The additional protein burden due to an incompletely used capacity of enzymes, e.g., under carbon-limited growth conditions, is accounted for by the unused enzyme sector introduced in the following section.

The unused enzyme sector. Particularly under nutrient-limited conditions, microbial cells put enzymes on hold to react quickly to changing environmental conditions (27, 44). This behavior contrasts with the general assumption that cells make efficient use of limited resources but reflects a necessary trade-off between maximal resource efficiency and quick adaptability. However, reversible enzyme kinetics prohibit the simultaneous manifestation of maximum velocity for all enzymes since the substrate of one enzyme is the product of another, inflicting opposing driving forces on both enzymes (38). As shown by metabolomics analyses, enzymes operate at substrate concentrations in their $K_M$ range under most physiological situations, thus well below their maximum velocity, and are considered underutilized (46). Moreover, unutilized enzymes of inactive, mostly catabolic metabolic pathways may be expressed to allow for swift metabolic adjustments upon changes in the environmental conditions or substrate availability.

With a detailed analysis of experimental proteomic data of *E. coli* (41), O’Brien et al. (44) showed the existence of such under- and unutilized enzymes, which we summarized in the PAM as the unused enzyme sector. More specifically, O’Brien augmented an *E. coli* ME model with a comprehensive proteomic data set covering all enzymes of the iML1515 model (Fig. 2A) and assessed the environment-specific proteome utilization in silico. They showed that the mass concentration of the active enzyme sector decreases with increasing growth rates, despite the increased metabolic activity and rising protein requirements. Conversely, enzymes, which are not catalytically active, accumulate the stronger the carbon limitation and the lower the growth rate. This phenomenon is regulated by the cAMP signaling pathway, which senses the carbon influx (4). We linked the resulting negative feedback loop between protein expression and substrate uptake caused by the cAMP signaling pathway (Fig. 1) to the functional representation of the unused enzyme sector (see Materials and Methods for details). As a result, protein allocation toward the unused enzyme sector decreases with increasing substrate uptake rate and diminishes at the maximum substrate uptake rate. In this way, the PAM allows for the adaption of substrate-specific allocation characteristics of the unused enzyme sector.

The translational protein sector. The quantitative description of the translational protein sector was empirically derived from global *E. coli* proteome measurements.

### Table 1

<table>
<thead>
<tr>
<th>Protein mass concn</th>
<th>Symbol</th>
<th>Linear dependency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active enzymes</td>
<td>$\phi_{AE}$</td>
<td>Flux rates $\nu$ of metabolic reactions</td>
</tr>
<tr>
<td>Unused enzymes</td>
<td>$\phi_{UE}$</td>
<td>Substrate uptake rate $\nu_s$</td>
</tr>
<tr>
<td>Translational protein</td>
<td>$\phi_T$</td>
<td>Growth rate $\mu$</td>
</tr>
<tr>
<td>Total protein</td>
<td>$\phi_P$</td>
<td>$\phi_{AE}, \phi_{UE}, \phi_T$</td>
</tr>
</tbody>
</table>

*Parameters within these linear dependencies and their (data) sources are listed in Table 2.*
The experimental proteomic data set was taken from Schmidt et al. (41), who investigated the quantitative dependencies between protein allocation and growth for a wide range of conditions, substrates, and different \textit{E. coli} strains. We processed the measured protein mass data (fg cell$^{-1}$) to arrive at model-relevant mass units relative to the cell dry weight (cdw) (g g$^{-1}$cdw). We found that the translational protein mass concentration correlates linearly with the growth rate and modeled this protein sector accordingly (Fig. 2B). This linear growth dependency depicts a resource-efficient regulation of the translational apparatus to justly provide the proteins necessary for maintaining maximal division rates in a particular environmental condition, as reported previously (48, 49).

The total protein concentration constraint. Under most conditions, the total protein content of an \textit{E. coli} cell is approximately 0.55 g g$^{-1}$cdw (50). A proteomic analysis of various \textit{E. coli} strains grown under different conditions (41) covered a constant protein mass concentration of 0.32 g g$^{-1}$cdw (58% of the total protein content) across conditions. Therefore, a constant 81%, or 0.26 g g$^{-1}$cdw ($\phi_{P,c}$), is allocated to the three protein sectors, that is, the active enzymes, unused enzymes, and translational protein sectors (Fig. 2C). The unassigned proteins of the proteomic data set and those not covered by the experimental data were assigned to the housekeeping protein sector. The protein fraction $\phi_{P,c}$ of 0.26 g g$^{-1}$cdw considered by the PAM is in agreement with the fact that the growth-dependent part of the proteome constitutes roughly half of \textit{E. coli}'s total protein content (3–5, 37, 39, 51). Note that the metabolic cost of the total protein biosynthesis in terms of energy and amino acid demand is fully covered by the biomass equation inherited from iML1515. The proteins within the data sets of Schmidt et al. (41) not covered by the PAM are mainly poorly characterized proteins (51% of all the protein fractions within the proteomic data set not covered by the PAM) or are assigned to transcription (8%), replication (6%), and posttranslational modification (6%). In turn, approximately one-third of the gene products considered by the PAM have not been quantified by Schmidt et al. (41). Those proteins are mainly inner and outer membrane proteins (58% of all gene products in the PAM not covered by the proteomic data set) or allocated to the glycerophospholipid (9%) and alternate carbon (8%) metabolism.

As experimental proteomic data suggest a consistent protein mass concentration across conditions, we modeled and added the invariable sum of active enzymes, unused enzymes, and translation protein to the PAM:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{Figure2.png}
\caption{Protein mass data of distinct protein sectors under diverse growth conditions taken from Schmidt et al. (41). (A) All proteins represented by the iML1515 GEM and found in the proteomic data set, comprising the active and unused enzymes sectors. (B) Translational sector covering proteins assigned to the COG (Clusters of Orthologous Groups) class “translation, ribosomal structure, and biogenesis,” which are not included in the iML1515. The black lines in panels A and B are linear fits of the data points resulting in the shown equations and coefficients of determination $R^2$. (C) Sum of the protein mass concentrations shown in panels A and B. The horizontal black line marks the 81% of measured protein mass used to constrain protein availability in the PAM. Glucose chemostat and batch experiments are highlighted in blue and red, respectively. Note that protein concentrations relative to the cell dry weight (cdw) were related to the total protein concentration of 0.32 g g$^{-1}$cdw measured by Schmidt et al. (41).}
\end{figure}
Here, $\phi_{P,c}$, $\phi_{T}$, $\phi_{AE}$, and $\phi_{UE}$ are the variable protein mass fractions of the translational protein, active enzymes, and unused enzyme sector, respectively. $\phi_{P,c}$ is the constant total protein mass concentration. Equation 1 introduces a concise representation of the capabilities and limits of $E. coli$’s metabolism and its condition-dependent protein allocation to stoichiometric modeling approaches. Due to the linear nature of the added protein allocation and enzyme activity constraints, classical flux balance analysis (FBA) and other established constraint-based modeling approaches can be applied to the PAM without a significant loss in computation speed compared to purely stoichiometric models.

The protein allocation model predicts wild-type phenotypes. To benchmark the PAM’s predictive capabilities, we simulated the wild-type phenotypic behavior on glucose minimal medium and compared the results to extensive literature data (52–59). The maximum glucose uptake rate, a parameter for the unused enzyme sector, was set to $8.9 \text{ mmol g}^{-1}\text{ cdw h}^{-1}$, which supported an observed maximal growth rate of $0.65 \text{ h}^{-1}$ (52). The simulated data derived with this constraint is in good agreement with experimentally observed phenotypes for a range of carbon-limited conditions and depicts significant improvements compared to the purely stoichiometric iML1515 model (Fig. 3). In particular, the acetate secretion trend correctly mirrors the metabolic overflow characteristics of $E. coli$ above glucose uptake rates of $4.3 \text{ mmol g}^{-1}\text{ cdw h}^{-1}$ (Fig. 3B). Despite reasonable projections of growth, acetate secretion, and oxygen uptake rates, the PAM as well as the iML1515 model overestimate carbon dioxide secretion rates, pointing to potential inconsistencies in the biomass synthesis equation used in both models.

The phenotypes simulated by the PAM are also mirrored in the fluxes through central metabolic pathways as shown in Fig. 3E. The distribution of the carbon flux among the metabolic pathways qualitatively follows previous findings (53). Under carbon-
limited and fully respiratory conditions, flux rates through the Embden-Meyerhof-Parnas (EMP) pathway, the tricarboxylic acid (TCA) cycle, the pentose phosphate (PP) pathway, and the glyoxylate shunt (GLYXS) scale linearly with the glucose uptake rate. The GLYXS activity reflects the need for anaplerosis to compensate for the drainage of TCA cycle intermediates under slow growth conditions, which has been experimentally demonstrated (60).

In the vicinity of the turning point from a fully respiratory to partially fermentative metabolism, the activity of the GLYXS diminishes completely. Beyond glucose uptake rates of 4.3 mmol g\textsubscript{cdw}^{-1} h\textsuperscript{-1}, limitations in the global protein household impede exclusive ATP production via respiration. As a consequence, the carbon flux is partly diverted from the NADH-yielding, and thus respiration-fueling, TCA cycle toward acetate. Simultaneously, the split between the EMP and PP pathways starts to increase with elevated glucose uptake rates in favor of glycolysis. While the NADPH supply through the PP pathway follows the growth-rate-dependent demand for biomass synthesis, the glycolytic flux (EMP) is accelerated to compensate for the carbon loss toward acetate.

The PAM’s stoichiometric and protein allocation constraints support feasible growth states beyond the assumed maximum glucose uptake rate (Fig. 3). However, the increase in growth rate is at the expense of biomass yield, which becomes evident by a drastic decrease of the flux through the EMP pathway and the TCA cycle and a progressively increasing acetate secretion rate. These simulation results do not necessarily indicate an underestimated maximum glucose uptake rate since PAM parameterization stays feasible even for maximum glucose uptake rates far beyond physiologically relevant values (data not shown). In fact, this suggests that the main metabolically limiting factor for \textit{E. coli} is neither stoichiometry nor protein allocation. As we will show in the next section, maximal growth rates may be attributed to a limited, maximum molar protein synthesis rate, thus indicating transcriptional restrictions as the determining growth-limiting factor.

Computing maximum substrate uptake rates. Since glucose is the preferred carbon source of \textit{E. coli}, its metabolism and regulation are adapted for an effective glucose utilization, which led us to assume that there are no unused enzymes under glucose-excess conditions. However, it is reasonable to assume that such a state of adaption does not hold for alternative carbon sources. Thus, experimentally observed substrate uptake rates may not reflect growth states with (near-)zero unused enzymes. To enable the PAM to simulate growth on alternative carbon sources without the need for cultivation data, we approximated maximum substrate uptake rates assuming a maximum total protein synthesis rate $N_{P,\text{max}}$. $N_{P,\text{max}}$ represents a global cellular constraint for protein biosynthesis, applicable for any substrate. It manifests at the maximum substrate uptake rate and therefore at maximum growth at which the unused enzyme sector approaches zero. In the PAM, $N_{P,\text{max}}$ represents the sum of molar synthesis rates of proteins from the active enzymes and the translational sectors.

Based on the PAM and phenotypic data of \textit{E. coli} grown on several substrates (61), we found that an $N_{P,\text{max}}$ of 2.04 $\mu$mol g\textsubscript{cdw}^{-1} h\textsuperscript{-1} defines maximum substrate uptake rates (Table S1). For glucose as the sole carbon source, an uptake rate of 9.82 mmol g\textsubscript{cdw}^{-1} h\textsuperscript{-1} is predicted accordingly. Interestingly, experimental values suggest that the maximum growth rate on acetate is approached for an uptake rate of 19.6 mmol g\textsubscript{cdw}^{-1} h\textsuperscript{-1} and a protein synthesis rate below $N_{P,\text{max}}$. This phenomenon indicates that protein allocation limits the maximum growth on acetate. Acetate is metabolized through the TCA cycle and the GLYXS (61). Energy-yielding routes other than respiration are biochemically impossible since substrate phosphorylation via the protein-efficient fermentation pathway results in acetate formation and, hence, a metabolic cycle with zero net carbon uptake. As a consequence, ATP generation solely relies on protein-costly respiration. We assume that the allocation of protein to the respiratory pathway is ultimately limited by the total protein concentration $\phi_{P,c}$. Thus, ATP supply and cofactor regeneration become the limiting factors before the maximum molar protein synthesis capability is reached.

Computing maximum substrate uptake rates. Since glucose is the preferred carbon source of \textit{E. coli}, its metabolism and regulation are adapted for an effective glucose utilization, which led us to assume that there are no unused enzymes under glucose-excess conditions. However, it is reasonable to assume that such a state of adaption does not hold for alternative carbon sources. Thus, experimentally observed substrate uptake rates may not reflect growth states with (near-)zero unused enzymes. To enable the PAM to simulate growth on alternative carbon sources without the need for cultivation data, we approximated maximum substrate uptake rates assuming a maximum total protein synthesis rate $N_{P,\text{max}}$. $N_{P,\text{max}}$ represents a global cellular constraint for protein biosynthesis, applicable for any substrate. It manifests at the maximum substrate uptake rate and therefore at maximum growth at which the unused enzyme sector approaches zero. In the PAM, $N_{P,\text{max}}$ represents the sum of molar synthesis rates of proteins from the active enzymes and the translational sectors.

Based on the PAM and phenotypic data of \textit{E. coli} grown on several substrates (61), we found that an $N_{P,\text{max}}$ of 2.04 $\mu$mol g\textsubscript{cdw}^{-1} h\textsuperscript{-1} defines maximum substrate uptake rates (Table S1). For glucose as the sole carbon source, an uptake rate of 9.82 mmol g\textsubscript{cdw}^{-1} h\textsuperscript{-1} is predicted accordingly. Interestingly, experimental values suggest that the maximum growth rate on acetate is approached for an uptake rate of 19.6 mmol g\textsubscript{cdw}^{-1} h\textsuperscript{-1} and a protein synthesis rate below $N_{P,\text{max}}$. This phenomenon indicates that protein allocation limits the maximum growth on acetate. Acetate is metabolized through the TCA cycle and the GLYXS (61). Energy-yielding routes other than respiration are biochemically impossible since substrate phosphorylation via the protein-efficient fermentation pathway results in acetate formation and, hence, a metabolic cycle with zero net carbon uptake. As a consequence, ATP generation solely relies on protein-costly respiration. We assume that the allocation of protein to the respiratory pathway is ultimately limited by the total protein concentration $\phi_{P,c}$. Thus, ATP supply and cofactor regeneration become the limiting factors before the maximum molar protein synthesis capability is reached.
By parameterizing the unused enzyme sector with the substrate-specific maximum substrate uptake rates, PAM simulations of phenotypes of *E. coli* grown on alternative carbon sources showed a good correlation with experimentally observed data (Fig. 4). The overestimation of growth and acetate secretion rates at the determined maximum substrate uptake rates (Fig. S1) indicates an unused potential to adapt *E. coli* to these alternative carbon sources. This assumption is, for instance, supported by a study by Fong et al. (62), in which the glycerol uptake rate was evolved to around 15 mmol g\textsubscript{cdw}^{-1} h^{-1}, which is close to the PAM's prediction. Interestingly, the shift from glucose to glycerol at the beginning of the evolutionary adaption of *E. coli* caused immediate, significant transcriptional responses. Regulatory mechanisms appeared to significantly affect the expression of genes associated with catabolism, such as *cra* and *crp*, presumably to nonspecifically scavenge and prepare for alternate substrates. During the adaptive evolution process, metabolism was focused on the available substrate, glycerol, and the basal transcriptional state was approached again. Thus, optimal glycerol uptake and utilization were mediated by changes in transcriptional regulation resulting in the adaption of catabolism to this sole carbon source. Moreover, flagellar and motility gene expression decreased during the adaptation, probably increasing the protein availability for growth-related processes.

**Prediction of *E. coli* flux distributions.** By exploiting the optimality principles of microbial growth, GEMs give quantitative insights into the intracellular flux distribution and pathway usage, based purely on stoichiometric constraints. For a particular environmental condition of interest, the prediction accuracy of the stoichiometric GEM generally scales with the amount and accuracy of experimental data introduced to the model in the form of flux constraints. With a minimum need for such data, i.e., the substrate uptake rate, the PAM allows for an accurate blueprint of the intracellular metabolic processes. A comparison with fluxomics data from multiple studies shows that flux distributions of the central carbon metabolism of *E. coli* grown on a minimal glucose medium are well predicted by the PAM, which is indicated by Pearson correlation coefficients up to 0.97 (Fig. 5). Here, the PAM was constrained with the measured glucose uptake rates, and the unused enzyme sector was parameterized according to the
methodically determined maximum glucose uptake rate, which, in turn, was a direct
corollary of the computational model output (cf. the previous section). Thus, protein allocation and enzymatic con-
straints alone enhance constraint-based modeling as a metabolic prediction tool, mak-
ing this modeling approach particularly useful under data scarcity. The relatively large
discrepancies between simulated and experimental
flux data for the pyruvate kinase (53, 63) (Fig. 5B and C) result from the neglect of the phosphotransferase system (PTS)
for glucose uptake in the
flux analysis models. The PTS system transfers
the phosphate group from phosphoenolpyruvate (PEP) to glucose during its import,
thereby producing glucose-6-phosphate and pyruvate. The omission of this alternative
uptake system necessitates an elevated pyruvate kinase
flux to balance PEP. In tradi-
tional
flux analysis, glucose uptake via the PTS and the ABC trans-
porter cannot be dissected, as they do not affect the
isotope patterns of metabo-
lites. It is therefore common to incorporate only one pathway in these models.

We further used the methodically determined maximum substrate uptake rates and
constrained the PAM with measured uptake rates to predict flux distributions for non-
glucose carbon sources (Fig. S2). Here, the prediction capabilities were more diverse,
resulting in high correlations for acetate, galactose, and succinate ($r > 0.92$) but inter-
mediate to weak predictions, e.g., for fructose or gluconate ($r > 0.65$). In the case of
fructose consumption, the Entner-Doudoroff (ED) pathway was experimentally
observed to be the main catabolic route, whereas the simulated carbon flux was
exclusively channeled toward pyruvate through the pentose phosphate (PP) path-
way. The flux split between the PP and ED pathways is highly sensitive to the ratio
in the protein demands of both pathways (data not shown). Consequently, inconsis-
tencies in the applied $k_{\text{cat}}$ values, particularly of backward reactions, but also
the substrate-dependent differences in the biomass compositions (13) may cause these
observed discrepancies. Further in-depth investigations are needed to validate one
or the other assumption.

**PAM explains the growth defect upon heterologous protein expression.** Enzyme
overproduction or the expression of heterologous genes is a common strategy in
many biotechnological disciplines. General purposes are introducing novel cellular
functionalities, flux enforcement through a specific pathway, or investigation on cellular
processes via reporter proteins. In any case, the introduced pull of proteins from
the limited native protein household and the resulting metabolic burden inevitably
cause a growth defect.

To investigate and evaluate the response of the PAM to such an induced protein

---

**FIG 5** PAM predictions of intracellular fluxes of the central carbon metabolism of *E. coli* grown on a glucose
minimal medium. The glucose uptake rate was constrained with experimentally determined values. The unused
enzymes sector was parameterized according to the computationally determined maximum value of $9.82 \text{ mmol} \ g_{\text{cdw}}^{-1} \ h^{-1}$ based on a maximum total protein synthesis rate $N_P$ of $2.04 \mu\text{mol} \ g_{\text{cdw}}^{-1} \ h^{-1}$. (A to C) The predictions
are compared with experimental flux data from panels A (61), B (63), and C (53). The goodness of the
correlations was computed based on the Pearson correlation coefficient $r$ and the corresponding $P$ value.
demand, we simulated growth for a range of expression levels of an enhanced green fluorescent protein (eGFP) and compared relative growth rates with experimental data from Bienick et al. (64). For the standard concentration of total condition-dependent protein $\phi_{PC}$ of 0.26 g $g_{cdw}{-1}$, the simulated relative growth defect is much more pronounced than to experimentally observed growth for intracellular eGFP concentrations beyond 0.03 g $g_{cdw}{-1}$ (Fig. 6A). The use of an E. coli TUNER strain, a derivative of the genome-reduced BL21 strain optimized for protein expression, in this study, explains this discrepancy. In this strain, the protein requirements for the growth rate-independent housekeeping sector are reduced, and more resources are available for the metabolic, condition-dependent protein sectors, which can be reflected in the PAM by increasing the total protein concentration $\phi_{PC}$. Accordingly, increasing the protein availability by expanding $\phi_{PC}$ about 20% attenuates the detrimental effects of eGFP expression on growth and results in an excellent reproduction of experimentally observed phenotypes (Fig. 6A).

Interestingly, the relation between eGFP expression strength and growth predicted by the PAM is nonlinear, which contrasts with a previous theoretical postulation (64). The nonlinearity arises from a combined effect of a protein drain from the translational and the active metabolic enzyme sectors. The enforced protein drain toward eGFP causes a decline in the ribosome concentration, resulting in a reduced translation rate. Under the assumption of a constant total protein concentration, the protein allocation to the metabolic sectors producing biomass precursors, amino acids, and energy decreases, which, in a vicious cycle, further decelerates the protein production rate and growth. This intracellular tug-of-war for proteins is intrinsically manifested in the PAM, leading to outperformance in predicting protein overexpression phenotypes over too simplified coarse-grained models.

The PAM also discloses relevant effects of the protein drain on the pathway flux level. For an increasing eGFP expression strength and the accompanied protein deficiency, central carbon fluxes are progressively diverted to fermentation pathways (acetate secretion) and eventually to the ED pathway. Both routes are more protein efficient but yield fewer energy equivalents per substrate molecule than respiration or the EMP pathway (39, 65, 66). The PAM confirms this diversion of ATP generation from respiratory to fermentation pathways. The proportion of ATP generated by respiration, more precisely by the ATP synthase, in the total cellular ATP generation rate

---

**Fig 6** (A) Simulated growth rates relative to the maximum growth rate are shown for a range of intracellular eGFP concentrations using the PAM with a standard (red line) and increased (blue line) total condition-dependent protein concentration $\phi_{PC}$ (0.26 g $g_{cdw}{-1}$ and 0.31 g $g_{cdw}{-1}$, respectively). Experimental data are taken from reference 64. P values are derived from Student’s t-test and indicate how well model predictions explain the experimental data. The dashed lines illustrate the predicted range of values based on theoretical considerations by Bienick et al. (64). (B) Simulated fluxes through central metabolic pathways are shown for a range of intracellular eGFP concentrations and a $\phi_{PC}$ of 0.31 g $g_{cdw}{-1}$. 

---

March/April 2021 Volume 6 Issue 2 e00625-20 msystems.asm.org
significantly decreases with an increasing eGFP expression (Fig. S3). Thus, the excelled protein burden enforces a shift toward protein-efficient, substrate-level phosphorylation in *E. coli*.

**Limitations in the protein allocation of single enzymes lead to gene deletion mutant phenotypes.** Alongside the over- and heterologous expression of genes, rearrangement of metabolic networks and flux distributions by gene deletions is a core instrument in metabolic engineering. In recent years, many computational strain design methods have been developed to accelerate and rationalize the engineering of microbial cell factories. However, in contrast to the vast number of model-driven strain design and optimization methods (67), constraint-based methods have often proven unreliable in predicting phenotypes of gene deletion mutant strains (GMSs). While GMSs have been shown to evolve toward FBA-predicted phenotypes (68), observed growth defects and intracellular fluxes of nonevolved GMSs cannot be explained by stoichiometry and a cellular growth objective alone (69–71).

First, we tested the impact of 10 single-gene deletions on the PAM’s FBA results by parameterizing the unused enzyme sector with the methodically determined maximum glucose uptake rate of 9.82 mmol g<sub>CDW</sub> h<sup>−1</sup>. The calculated phenotypes did not significantly differ from the unperturbed wild-type solutions and, hence, did not compare to experimental data (70, 72) (Fig. S4). Only after constraining the glucose uptake rate to the measured values did the simulated growth and acetate secretion rates correlate well with the experiments (Fig. 7). Furthermore, a blind test in the form of FBA simulations augmented with *in vivo* glucose uptake rates of the GMS, but with an intact target gene, yielded the same phenotypes (Fig. S5). These results led us to conclude that the main driver for the observed growth defects of GMSs is a naturally orchestrated catabolite uptake repression induced by the respective network perturbations. A gene disruption resulting in a growth phenotype implicates a decrease in the levels of one or more amino acid precursors since tight proteome coordination (4) hampers enzyme allocation toward alternative precursor synthesis routes. According to the coarse-grained model of You et al. (4), which considers the general mechanism of rRNA transcription (73), low amino acid levels stall ribosomal protein synthesis via ppGpp. Moreover, low precursor levels, such as oxaloacetate or α-ketoglutarate, lead to an increased cAMP synthesis and elevated CRP-cAMP levels, which foster the expression of catabolic enzymes for alternative substrates adding to growth reduction. Elevated cAMP concentrations, which were experimentally observed in multiple GMSs (74), indicate increased CRP-cAMP activities and support this view on gene deletion triggering...
the integral feedback of metabolic control. Therefore, we argue that the degree of metabolic rewiring in GMSs necessary to overcome growth defects is often low. Indeed, restoration of growth can be achieved by evolutionary adaptation processes that alter these regulatory patterns (74).

The assumption that a regulatory substrate uptake inhibition shapes GMS’ phenotypes raised the question if the PAM can quantitatively predict substrate uptake modes. To tackle this question, we recalled computational frameworks such as minimization of metabolic adjustment (MOMA) (75), regulatory on/off minimization (ROOM) (76), and relative change (RELATCH) (69), which determine the metabolic impact of gene knockout strains. The common principle behind all three methods is the minimization of the metabolic response to genetic perturbations due to an unchanged regulatory system that forces the GMS’ flux distribution toward the original steady state. In the context of protein allocation, the minimal response principle can be translated as follows: upon a network perturbation, a GMS establishes a substrate uptake rate so that increases in protein allocation toward single enzymatic reactions are minimal compared to genetically unperturbed strains. The cellular objective is to allow for maximum metabolic activity in the face of knockout-induced flux rerouting and a hampered reallocation of protein due to a strict (wild-type) regulatory regime.

We applied the minimal response principle to the PAM for mutant phenotype predictions, and we simulated growth optimal flux distributions for a range of substrate uptake rates (cf. Materials and Methods for a detailed description). For each flux distribution, the difference in the synthesis rate \( \Delta N_e \) between a reference, wild-type state at maximum growth and a mutant state is calculated for each enzyme within the PAM. The GMS’s maximum substrate uptake rate is determined from the flux distribution for which \( \Delta N_e \) meets a defined upper bound \( \Delta N_{e}^{\text{crit}} \) for at least one enzyme within that flux distribution. In doing so, we assign a certain flexibility to the overexpression capacity of single enzymes, which may be attributed to the activation of underutilized enzyme capacities. Moreover, if \( \Delta N_{e}^{\text{crit}} \) is met for one enzyme, flux rerouting to circumvent the saturated enzyme reaction is impossible. Refusing substantial metabolic rewiring conforms to experimental observations that unevolved GMSs do generally not show a large amount of metabolic adjustment to the metabolic network perturbations (68, 75). The strict upper bound \( \Delta N_{e}^{\text{crit}} \) reflects this metabolic rigidity on changes of enzyme synthesis rates in GMSs compared to the parental wild-type strain.

Figure 8 shows the prediction results in comparison to experimentally determined phenotypic data (70–72) for an \( \Delta N_{e}^{\text{crit}} \) of 16 nmol \( q_{\text{wild}}^{-1} \cdot h^{-1} \). \( \Delta N_{e}^{\text{crit}} \) was derived from minimizing the discrepancy between simulated and experimentally observed phenotypes for various single-gene deletion mutants (cf. Materials and Methods for a detailed description). With one exception, there is a general agreement between predicted and experimentally determined phenotypes (Pearson correlation: \( r = 0.81, P = 7.76 \times 10^{-11} \)). By disregarding the \( \Delta rpe \) deletion mutant, the Pearson correlation improves to \( r = 0.97 \) with a \( P \) value of \( 1.5 \times 10^{-23} \). For the omitted outlier, simulations show a near-wild-type behavior, whereas growth is significantly reduced in the experiment. The \( \text{in vivo} \) observed growth defect of the \( \Delta rpe \) GMS is peculiar since knockouts of enzymes adjacent to ribulose-phosphate 3-epimerase, such as \( gnd \), do not show a similar phenotype. Even though the gene products of \( gnd \) and \( rpe \) catalyze consecutive reactions transforming \( \delta \)-phospho-\( \delta \)-glucanate to \( \delta \)-xylulose 5-phosphate, an important intermediate in the PP pathway, deletion of \( gnd \) has a significantly smaller effect on growth, as does the deletion of \( rpe \). Moreover, the experimentally determined phenotype of a \( \Delta rpe \) GMS from reference 77 was similar to the wild-type \( E. coli \) strain and thus obscures a plausible judgment of the simulated phenotype.

The considerable agreement between experimental and simulated phenotypic data also applies to intracellular flux data. We computed flux distributions for single-gene deletions that result in an experimentally proven inactivation of the corresponding reaction (i.e., we excluded single-gene deletions of isozymes) (71). The predicted flux responses correlated well with experimentally determined flux data (Fig. 9), noticeable...
on Pearson’s correlation coefficients of $r > 0.93$, for all tested single deletions, except for the ΔtpiA mutant ($r = 0.67$). For the ΔtpiA GMS, the experimentally observed rerouting of the glycolytic flux through the methylglyoxal pathway toward pyruvate (71, 78) contrasts with the PAM results, which suggests the activation of the ED pathway surpasses the blocked glycolysis. The simulated behavior can be traced back to an underestimated protein demand of the ED pathway, as predictions were improved after decreasing the $k_{cat}$ values of the two central ED pathway reaction steps phosphogluconate dehydratase and 2-dehydro-3-deoxyphosphogluconate aldolase to 2.5% of the original values. After reevaluating the maximum glucose uptake rate, the simulations showed a diversion of the glycolytic flux into the methylglyoxal pathway and acetate secretion (Fig. S6). This selective adaption of turnover numbers, and thus of protein costs for the corresponding reactions, resulted in a significant convergence of simulated flux data toward measured values for the ΔtpiA and also the Δrpe GMS ($r$ of 0.90 and 0.99, respectively) (Fig. S7). However, at the same time, flux predictions for the Δgnd mutant were deteriorated ($r = 0.95$). Here, flux was shifted from the ED pathway toward glycolysis, which contradicts experimental data. Nevertheless, the $k_{cat}$ value adaption only slightly lessened the PAM’s overall predictive capabilities (Fig. S7) and thus generally points to the need for fine-tuning kinetic parameters to improve the flux split between pathways across different conditions.

**DISCUSSION**

Proteins are the major molecular class in cells, and because they are the catalyst for global cellular functionalities, the importance of the mutual connection between microbial metabolic behaviors and protein allocation is broadly accepted (3, 5–7, 37–39). Based on existing techniques for implementing protein allocation and enzymatic constraints in constraint-based modeling (37, 40), we implemented a methodology to account for the total condition-dependent proteome in an *E. coli* GEM. Besides the integration of basic enzyme kinetics in the form of turnover numbers to model the concentrations of active enzymes, the resulting PAM considers simple, linear relations between microbial growth and the translational protein as well as the unused enzyme sector to describe 81% of the total protein mass concentration.

The PAM’s accurate predictions confirmed the prominent role of protein allocation in shaping microbial metabolism. Nevertheless, protein allocation itself appears to be
regulated by biochemical limits. Genetic sequencing results of *E. coli* strains that have undergone extensive adaptive laboratory evolution (ALE) (79) suggest a causal link between transcription limitations and maximum cell proliferation rates. The adapted strains, all exhibiting a fitness increase of up to 1.6-fold, showed mutations in the *rpoB* or *rpoC* gene leading to single amino acid substitutions in the β/β′ subunit of the RNA polymerase. This subunit is part of the enzyme’s active center. Thus, the observed mutations globally affect transcription (79). We introduced the total molar synthesis rate of proteins as a proxy for the transcription capability of *E. coli* since the PAM does not represent any transcriptional processes. We found that maximum substrate uptake rates, and consequently maximum growth rates, are dictated by an ultimately limited total molar synthesis rate of condition-dependent proteins. Based on our study, we presume that a reduction of transcriptional limitations in evolved *E. coli* strains allows for an increased protein allocation toward the metabolically active enzymes and translational protein sector. A detailed proteomics study is necessary to uncover changing protein allocation principles in growth-optimized microbial strains and to test this hypothesis with the PAM.

By recognizing transcriptional limitations as hard constraints for the microbial metabolism and recalling that the presented PAM predictions represent growth optimal flux states, we deduce a cellular principle, which is supported by previous findings (39); particularly under substrate-limited conditions, *E. coli* regulates central cellular processes to maintain a Pareto-optimum between growth rate and the ability to adapt its metabolism flexibly to changing environmental conditions. The degree of flexibility is directly linked to the amount of allocated unused enzymes, which is hard-coded in the

![Comparison between experimentally determined and simulated flux distributions of GMSs without considering observed glucose uptake rates as model constraints. Predicted intracellular flux rates are plotted against experimentally determined values taken from Long and Antoniewicz (71). All data were normalized with corresponding glucose uptake rates. Predictions were made applying the PAM and constraining the unused enzyme sector according to a maximum glucose uptake rate of 9.82 mmol g\(^{-1}\)cdw h\(^{-1}\). The maximum overexpression capacity of single enzymes \(\Delta N^\text{crit}\) was set to 16 nmol g\(^{-1}\)cdw h\(^{-1}\).](http://msystems.asm.org/)
PAM. Interestingly, the cell retains a strictly substate uptake-orientated allocation of protein to the unused enzyme sector even when protein becomes a metabolically limiting factor, apparent from the onset of overflow metabolism. This preservation of flexibility could allow for an evolutionary advantage in the original environmental niche. However, it may be a promising target for engineering *E. coli* or any microbe toward a high-performance cell factory.

Recently, systems metabolic engineering was emphasized as an integral part of the development and optimization of microbial cell factories (67, 80–83). Thus, we want to put the PAM forth to highlight the advantages of considering total condition-dependent proteome allocation for constraint-based modeling techniques that favor more accurate mutant strain predictions. We showed sound predictions of growth upon a range of overexpression levels of a nonenzymatic protein. The predictive capability for the overexpression of enzymes participating in metabolic reactions, e.g., in heterologously expressed pathways, still needs to be verified. However, the strict regulation of protein allocation, represented by the functional description of translational protein and unused enzyme sectors in the PAM, appeared to shape metabolic responses and is insensitive to genetic interferences. We also confirmed a rather inflexible protein allocation behavior for gene deletion mutant strains. Coherent prediction results were obtained by allowing only small divergences from a wild-type expression rate of single enzymes. Hence, in addition to the restrictions mediated by CRP-cAMP, we propose a link to transcriptional limitations, similar to our observation of maximum growth of wild-type strains mentioned above. This hypothesis is supported by ALE experiments in which adaption of single-gene deletion mutants frequently generated mutations affecting the regulation of global and pathway-specific transcription (74). These mutations possibly eliminate transcriptional hurdles and (partly) restore growth rates.

In summary, we want to stress the importance of protein allocation constraints in GEMs for the systematic constraint-based reconstruction and analysis (COBRA) and the design of microbial metabolism without having to sacrifice computational speed or applicability of established COBRA methods (84). However, the limited availability and credibility of enzymatic kinetic data, particularly of turnover numbers, still poses a significant obstacle in providing PAMs for any microorganism. Therefore, we join the call to establish a thorough kcatome as part of an accessible, genome-wide kinetome (85).

**MATERIALS AND METHODS**

Formulating the protein allocation model. The protein allocation model (PAM) is based on the *E. coli* K-12 MG1655 GEM iML1515 reconstruction (45). It includes linear representations of the protein mass concentrations of the active enzymes, the unused enzymes, and the translational protein sector as additional constraints. The sum of the protein mass concentrations of all sectors is kept constant according to Equation 1. The modeled protein sectors are described in detail in the following.

The translational protein sector. In agreement with previous studies, a linear correlation between the translational protein sector $\phi_T$ and the growth rate $\mu$ (5, 41) was implemented. Therefore, the inverse of the maximum ribosomal elongation rate $w_\mathrm{T}$ and a measure $\phi_{T,0}$ for an increasing overcapacity of ribosomes with a decreasing growth rate (43, 86) were used as the slope and intercept, respectively:

$$\phi_T = \phi_{T,0} + w_T \cdot \mu \quad (2)$$

Both parameters $w_T$ and $\phi_{T,0}$ were determined for *E. coli* by fitting Equation 2 to measured, cross-conditional concentration data of the translational protein sector (41) (Fig. 2B), resulting in values of 50.0 mg g$^{-1}$ (19% of the total protein mass concentration $\phi_P$ of 0.26 g g$^{-1}$) and 36.8 mg h g$^{-1}$ for $\phi_{T,0}$ and $w_T$, respectively. The relative value for $\phi_{T,0}$ differs from the visual intercept in Fig. 2B, since protein mass concentrations in Fig. 2 were related to the total protein concentration of 0.32 g g$^{-1}$ measured by Schmidt et al. (41).

The unused enzyme sector. An evolutionary characteristic of microbes facing unforeseeable changes in environmental conditions is the synthesis of unused proteins. This protein hedging empowers the cell to quickly increase central carbon metabolism fluxes upon a sudden increase in substrate availability or to immediately catalyze a new carbon source (44). Enzyme overabundance also facilitates metabolic robustness against genetic perturbations (87). However, these proteins reserve significantly reduce the growth rate, discussed in depth by O’Brien et al. (44) and Scott et al. (3). Protein expression and (over)allocation is generally coordinated by the cyclic AMP (cAMP)-dependent signaling pathway via the cAMP-activated global transcriptional receptor protein (CRP) (4, 88) (Fig. 1). The CRP-cAMP complex enhances the transcription of over 100 genes by attaching near or at their promoter...
regions, thereby mediating the binding of RNA polymerase for transcription initiation \((89)\). The target genes are mainly associated with catabolism, and thus, CRP-cAMP aids in stimulating the carbon influx leading to the accumulation of precursors for amino acids and stimulation of protein synthesis. Amino acid precursors, such as oxaloacetate or \(\alpha\)-ketoacids, inhibit the cAMP synthesis, thus lowering the CRP-cAMP level and closing the negative feedback loop between protein precursors and substrate uptake. We refer to You et al. \((4)\) for respective insights into the cAMP signaling pathway. In this way, CRP-cAMP indirectly coordinates the global protein allocation, including the allocation of unused enzymes, even though the transcription of only a small fraction of genes is directly affected. Considering the cAMP-controlled signaling pathway as a blueprint for protein synthesis regulation, we modeled the unused enzyme sector as a negative linear function of the substrate uptake flux, which we mathematically expressed as

\[
\phi_{\text{UE}} = \phi_{\text{UE,0}} - w_{\text{UE}} \nu_{\text{s}} \tag{3}
\]

where \(\phi_{\text{UE,0}}\) is the unused enzyme concentration at zero substrate uptake, and \(\nu_{\text{s}}\) is the substrate uptake rate. \(w_{\text{UE}}\) relates the decrease of the unused enzymes’ concentration to the increase in \(\nu_{\text{s}}\). \(\phi_{\text{UE,0}}\) was determined from ME model simulations of un- and underutilized enzymes \((44)\), resulting in a value of 0.17 g g\(^{-1}\) equivalent to 65% of the total protein mass concentration \(\phi_{\text{P,c}}\). At zero substrate uptake rate \(\nu_{\text{s}} = 0\), and therefore at zero growth, the sum of unused enzymes and translational protein \(\phi_{\text{UE,0}} + \phi_{\text{P,c}}\) is 0.22 g g\(^{-1}\). Thus, near \(\nu_{\text{s}} = 0\) the PAM explains only a fraction (84%) of the total protein mass concentration \(\phi_{\text{P,c}}\) which is, however, assumed to be constant under any given condition. We expect this discrepancy to be caused by deviations from an otherwise constant total protein household in severely starving cells or specific protein requirements during non-growth conditions. The fraction not explained by the PAM is compensated for by nonphysiological, random activation of metabolic fluxes when calculating growth-optimal solutions near substrate uptake rates of zero. Thus, the PAM may produce feasible solutions at low metabolic activity but hardly represent physiologically relevant flux states.

The slope \(w_{\text{UE}}\) is a measure of the increase in enzyme usage efficiency with increasing substrate uptake rate. It is individually assigned for each substrate under the assumption that the unused enzyme concentration is zero at the maximum substrate uptake rate. Thus, \(w_{\text{UE}}\) is calculated from

\[
w_{\text{UE}} = \frac{\phi_{\text{UE,0}}}{\nu_{\text{s}} \text{max}} \tag{4}
\]

The maximum substrate uptake rate \(\nu_{\text{s}} \text{max}\) needs to be provided as an observable but can be directly inferred from the PAM by assuming an upper limit in the transcriptional capacity, as shown in “Determination of Maximum Substrate Uptake Rates.”

**The active enzyme sector.** To account for the enzyme demand of metabolic fluxes, we integrated enzyme mass balances for all relevant metabolic reactions in the stoichiometric matrix of iML1515 according to the GECKO framework \((40)\), as explained in the following. The employed enzyme mass balance formulation (Equation 5) relates the flux \(\nu_{\text{e}}\) to the minimally required concentration \(\rho_{\text{e}}\) of the enzyme catalyzing reaction \(\text{e}\) by the enzyme’s turnover number \(k_{\text{cat},\text{e}}\). Since the PAM accounts for the total proteome by mass, we transformed molar to mass concentrations using the molar mass \(M_{\text{e}}\) of each enzyme in the GEM.

\[
\rho_{\text{e}} = \frac{\nu_{\text{e}} M_{\text{e}}}{k_{\text{cat},\text{e}}} \tag{5}
\]

Eventually, the sum of the concentration of all \(E\) enzymes constitutes the metabolically active enzyme sector \(\phi_{\text{AE}}\) and is expressed as

\[
\phi_{\text{AE}} = \sum_{\text{e}} \rho_{\text{e}} \tag{6}
\]

Parameterization of the active enzyme sector is important to facilitate a meaningful relation between fluxes and enzyme concentrations. For the PAM, we determined \(k_{\text{cat}}\) values for 2,843 enzymes of the iML1515 model from queries of the BRENDA \((90)\) database following the protocol of Sánchez et al. \((40)\). Since simulated maximum growth rates were unreasonably low when applying the initial \(k_{\text{cat}}\) set, the \(k_{\text{cat}}\) data set was manually curated to allow for the computation of reasonable phenotypes (cf. Data Set S1, sheet 1 for the final data set). Therefore, enzymes were ranked according to the sensitivity of FBA-derived, maximal growth rates toward their \(k_{\text{cat}}\) values. Approximately 150 \(k_{\text{cat}}\) parameters that were found to be most influential on computed growth rates were reevaluated. For most of the underlying enzymes, no specific \(k_{\text{cat}}\) entry could be found in BRENDA, and \(k_{\text{cat}}\) values had been taken from close enzyme classes during the automated database queries. This \(k_{\text{cat}}\) approximation was corrected by manually querying SABIO-RK \((91)\), UniProt \((92)\), and primary literature for more sound values regarding growth optimality. If no specific data could be found for an enzyme, \(k_{\text{cat}}\) values from adjacent enzyme classes were manually selected and assigned to the PAM.

The manual curation effort transformed the primary in vivo \(k_{\text{cat}}\) estimates to effective (or apparent) in vivo turnover numbers \(k_{\text{app}}\) \((48, 93)\), relating enzyme concentrations to metabolic fluxes under unlimited growth conditions. Recently, measurements of \(k_{\text{app}},\text{max}\) the maximum \(k_{\text{app}}\) values across conditions, were extrapolated to genome scale using machine learning models informed with biochemical and
Protein Allocation Model of E. coli

enzymatic data (94). Interestingly, although the distribution of \( k_{\text{app,max}} \) values is significantly different from our curated set (Fig. S8), simulations using the PAM parameterized with either or both \( k_{\text{app,max}} \) sets yielded comparable phenotypes (data not shown). Thus, it appears that the ratio of \( k_{\text{app}} \) values between different reactions or pathways is an essential factor for the predictive capability of PAMs in general.

**Processing of experimental proteomic data from literature.** The experimental proteomic data set generated by Schmidt et al. (41) was used in this study to parameterize the translational and the unused enzymes sector of the PAM. To fit the units of the data set to the model variables, the protein mass per cell measured by Schmidt was converted to dry cell weight-based concentration values:

\[
\phi_p = \frac{M_p}{V_{\text{cdw}} \cdot \mu} \tag{7}
\]

Here, \( \phi_p \) is the intracellular concentration (g g\(_{\text{cdw}}^{-1}\)), and \( M_p \) is the experimentally determined mass per cell of a protein or a protein sector. \( V_{\text{cdw}} \cdot \mu \) is the growth rate-dependent cellular volume, which was thoroughly determined by Vollmer and Heinemann (95). The cell dry weight (cdw) concentration per cell volume \( c_{\text{cdw}} \) was assumed to be 268.36 g\(_{\text{cdw}}\) liter\(^{-1}\) (50). Refer to Data Set S1, sheets 2 and 3 for the complete data set.

UniProt identifiers (92) and the COG classification (Clusters of Orthologous Groups) were used to match the proteins and protein sectors of the PAM to the proteomic data.

**Solving the protein allocation model problem.** All flux solutions and corresponding phenotypes in this work represent growth-optimal solutions of the following, classical FBA-problem amended with additional protein allocation and enzymatic constraints:

\[
\begin{align*}
\text{max} & \quad \mu \\
\text{s.t. Classical FBA constraints} & \quad \Sigma \mu_i \leq \mu_{\text{ub}} \quad \forall i \in N \\
\text{Protein allocation constraints} & \quad \mu_i \geq 0 \\
& \quad \phi_T = \phi_{T,0} + \nu_T \cdot \mu \\
& \quad \phi_{UE} = \phi_{UE,0} + \nu_{UE} \cdot \mu \\
& \quad \phi_{P,c} = \phi_T + \phi_{UE} + \sum_{j} \rho_j \\
\end{align*}
\]

Here, \( S \) is the stoichiometric matrix of the original GEM, and \( \mu_i \) is the flux variable of reaction \( i \) from the metabolic reaction pool \( N \). Note that each reversible reaction in the original GEM is split into irreversible forward and backward reactions to only allow for positive flux values; hence, the lower and upper flux bounds \( \nu_i^{\text{lb}} \) and \( \nu_i^{\text{ub}} \) are equal to or greater than zero. The additional protein allocation and enzymatic constraints comprise the mass concentrations \( \rho_j \) for each considered enzyme or enzyme complex, as well as the mass concentrations of the ribosomal and unused enzyme sector \( \phi_T \) and \( \phi_{UE} \), respectively. All protein allocation constraints in Equation 8 were added to the stoichiometric matrix \( S \) of the GEM ML1515 representing the E. coli K-12 MG1655 strain (45).

Each reaction \( e \) from pool \( E \), comprising all reactions linked to one or more genes via a gene-protein-reaction (GPR) relation, is assigned to precisely one protein with a unique turnover number \( k_{\text{cat}} \) and molar mass \( M_e \). In case a reaction \( e \) is catalyzed by an enzyme complex (multiple genes are connected via logical OR operators in the GPR relation), the molar mass \( M_e \) is the sum of molar masses of the participating gene products. If two or more enzymes can catalyze the same reaction independently from each other (multiple genes are connected via logical AND operators in the GPR relation), the molar mass \( M_e \) is kept constant according to Equation 1 and comprises the sum of the translational sector \( \phi_T \), unused enzyme sector \( \phi_{UE} \), and active enzyme sector \( \phi_{P,c} \). An overview of the applied parameters is given in Table 2.

**Determination of maximum substrate uptake rates.** By applying the PAM with parameterized protein sectors and minimal medium constraints, maximum uptake rates for single substrates were
determined assuming a maximally allowable total protein synthesis rate $N_{p,\text{max}}$. $N_p$ represents the sum of molar synthesis rates of proteins from the active enzymes, unused enzymes, and the translational sector. At the maximum substrate uptake rate, and therefore at maximum growth, the unused enzyme sector is zero. Thus, the maximum total protein synthesis rate $N_{p,\text{max}}$ is

$$N_{p,\text{max}} = \mu \left( \frac{\phi_2}{M_k} + \sum_i \frac{\phi_i}{k_{\text{cat},i}} \right)$$  \hspace{1cm} (9)$$

Here, $M_k$ is the sum of the molar masses of all 21 ribosomal subunits of $3.5 \times 10^4$ g mol$^{-1}$, and $r_e$ is the flux of each enzymatically catalyzed reaction $e$.

We hypothesized that there is a unique maximum total protein synthesis rate $N_{p,\text{max}}$ which determines maximum uptake rates for any given substrate. Knowledge of $N_{p,\text{max}}$ would allow the determination of maximum substrate uptake rates and the parameterization of the unused enzyme sector (cf. Equations 3 and 4) solely based on model simulations. A correspondingly parameterized PAM should therefore be able to correctly predict phenotypes for substrate-limited conditions.

To test this hypothesis, we simulated $N_{p,\text{max}}$ values for substrates considered by Gerosa et al. (61) using PAMs parameterized with a wide range of maximum substrate uptake rates. More specifically, the unused enzyme sector was parameterized at each of the tested maximum substrate uptake rates according to Equation 4. Growth rates were simulated by optimizing the parameterized PAMs for growth and constraining the substrate uptake rate to the experimentally determined value. These simulated growth rates were then compared with the observed values by computing the absolute difference. At an $N_{p,\text{max}}$ of $2.04 \mu$mol g$^{-1}$ h$^{-1}$, the sum of absolute differences between simulated and measured growth rates was minimal. For each of the considered substrates, maximum uptake rates (parameter of the PAM) leading to $N_{p,\text{max}}$ by applying the experimentally observed substrate uptake rate (constraint) were extracted and are shown in Table S1.

**eGFP overexpression.** Expression of eGFP was simulated by introducing an additional column to the stoichiometric matrix of the PAM representing a protein with a mass of $2.8 \times 10^4$ g mol$^{-1}$. Thus, the expression strength of eGFP (g h$^{-1}$ g$^{-1}$ cdw) is controlled by the respective model variable describing the protein’s intracellular concentration and the growth rate. To identify the total protein concentration that represents the elevated protein availability in the *E. coli* TUNER strain used by Bienick et al. (64), the correlation between eGFP concentrations and growth rates relative to the wild-type was computed with the PAM for a range of total protein concentration values $P,c$.

For each $\phi_{p,e}$, the summed difference between simulated and measured relative growth rates was calculated as a measure for the agreement between simulation and experiment. Student’s t-test was applied to determine the $\phi_{p,e}$, yielding the best possible correlation between measurements and simulations.

**Phenotype determination of gene deletion mutants.** The deletion of a gene was simulated by identifying all reactions connected to this gene via the model-inherent GPR rules and setting the respective upper bounds to zero. To predict maximum substrate uptake rates of GMSs, enzyme synthesis rates were calculated from growth-optimal flux distributions for a wide range of substrate uptake rates. The synthesis rate $N_e$ of an enzyme $e$ was computed from flux distributions by

$$N_e = p_e \cdot \mu$$  \hspace{1cm} (10)$$

where $\mu$ and $p_e$ are the growth rate and the molar concentration of enzyme $e$, both being optimization variables to the PAM.

For each tested substrate uptake rate, the maximum difference in the computed enzyme synthesis rate $\Delta N_{e,\text{max}}$ between the GMS and a reference state representing a wild-type strain grown under substrate-limited conditions was identified among all modeled enzymes. By scanning $\Delta N_{e,\text{max}}$ values from high to low substrate uptake rate, the maximum substrate uptake rate for a GMS is found as soon as $\Delta N_{e,\text{max}}$ meets a critical value $\Delta N_{e,\text{crit}}$ (also termed maximum overexpression capacity). Thus, the corresponding metabolic mode supports a maximally achievable growth rate under a limited flexibility to change or reallocate protein among metabolic pathways and their single reactions. We assumed the level of restriction in flexibility $\Delta N_{e,\text{crit}}$ to be the same in any GMS according to phenotypic data of GMS from Long et al. (70), Long and Antoniewicz (71), and Fong et al. (72). We found that a $\Delta N_{e,\text{crit}}$ of $16.0 \mu$mol g$^{-1}$ h$^{-1}$ leads to a minimal sum of errors between predicted and observed growth, substrate uptake, and metabolic flux rates.
uptake, and acetate secretion rates. The corresponding comparisons of experimentally determined and predicted phenotypes as well as flux distributions are shown in Fig. 8 and 9, respectively.

**Implementation.** All conducted simulations, model reconstructions, and data analyses were performed in MATLAB 2018a on a Windows 7 machine with 16 GB of RAM and an AMD FX-8350 eight-core (at 4.00 GHz) processor. COBRA toolbox functions (84) and the Gurobi Optimizer (8.0.0; Gurobi Optimization, Inc.) were utilized to process and solve the metabolic models. All MATLAB functions necessary to handle and build a protein allocation model (PAM) from a COBRA format-based, stoichiometric reconstruction are provided on GitHub (https://github.com/Spherotob/PAM_public).

**SUPPLEMENTAL MATERIAL**

Supplemental material is available online only.

**DATA SET S1**, XLSX file, 0.3 MB.
**FIG S1**, EPS file, 0.3 MB.
**FIG S2**, EPS file, 0.3 MB.
**FIG S3**, EPS file, 0.1 MB.
**FIG S4**, EPS file, 0.2 MB.
**FIG S5**, EPS file, 0.2 MB.
**FIG S6**, EPS file, 0.1 MB.
**FIG S7**, EPS file, 0.4 MB.
**FIG S8**, EPS file, 0.5 MB.
**TABLE S1**, DOCX file, 0.02 MB.

**ACKNOWLEDGMENTS**

T.B.A. was partially funded by the Excellence Initiative of the German federal and state governments (grant ID PFL5015). B.E.E. acknowledges funding through the CSIRO-UQ Synthetic Biology Alliance. The laboratory of L.M.B. received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy within the Clusters of Excellence “236 TMFB” and “2186 The Fuel Science Center.” The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

**REFERENCES**


25. Lerman JA, Hyduke DR, Latif H, Portnoy VA, Lewis NE, Orth JD, Schrimpe-erichia coli


33. Alter et al.


Protein Allocation Model of Escherichia coli


