nature COMMUNICATIONS

ARTICLE

Received 20 Apr 2016 | Accepted 3 Nov 2016 | Published 20 Dec 2016

DOI: 10.1038/ncomms13806 OPEN

A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains

Ali Khodayari[1] & Costas D. Maranas[1]

Kinetic models of metabolism at a genome scale that faithfully recapitulate the effect of multiple genetic interventions would be transformative in our ability to reliably design novel overproducing microbial strains. Here, we introduce k-ecoli457, a genome-scale kinetic model of *Escherichia coli* metabolism that satisfies fluxomic data for wild-type and 25 mutant strains under different substrates and growth conditions. The k-ecoli457 model contains 457 model reactions, 337 metabolites and 295 substrate-level regulatory interactions. Parameterization is carried out using a genetic algorithm by simultaneously imposing all available fluxomic data (about 30 measured fluxes per mutant). The Pearson correlation coefficient between experimental data and predicted product yields for 320 engineered strains spanning 24 product metabolites is 0.84. This is substantially higher than that using flux balance analysis, minimization of metabolic adjustment or maximization of product yield exhibiting systematic errors with correlation coefficients of, respectively, 0.18, 0.37 and 0.47 (k-ecoli457 is available for download at http://www.maranasgroup.com).

[1] Department of Chemical Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA. Correspondence and requests for materials should be addressed to C.D.M. (email: costas@psu.edu).

Rapid and pervasive advances in genome editing[1] and copying[2] techniques have dramatically reduced the time and cost[3] for microbial strain construction. This has enabled probing multiple and often complex intervention strategies in pursuit of an overproducing cellular phenotype. Existing computational strain design tools operating at a genome scale[4–6] primarily rely on a stoichiometric description of metabolism. Despite many successes[7–9], designed mutants often fail[10,11] as stoichiometric models do not directly account for enzyme level, metabolite concentration and substrate-level regulatory barriers. Kinetic models provide a computational vehicle for capturing these effects; however, they have been mostly descriptive in nature, parameterized for a single metabolic phenotype[12] and are thus incapable of predicting the often systemic effect of metabolic perturbations. A number of efforts have been proposed that standardize kinetic expressions through approximation and provide a tractable parameterization workflow even at a genome scale (for example, a log-lin approach for *Escherichia coli*[13] and a lin-log approach for yeast[14]). However, by design, these approaches are aimed at providing accurate production in the vicinity of the reference state. Kinetic models that satisfy multiple flux data sets[15] have so far been limited in their metabolic scope capturing mostly central metabolism[15,16].

Recent approaches first deconstruct complex kinetic expressions into their elementary steps[17] and then reassemble them into kinetic expressions that capture all known substrate-level regulations. Uncertainty in assigning unique kinetic parameter values is circumvented by creating not only a single kinetic model, but rather an ensemble of kinetic models through parameter sampling[13]. The ensemble modelling (EM) approach[18], for example, relies on the sampling of reaction reversibilities and enzyme fractions[19] to create an ensemble of models. The original ensemble is subsequently 'whittled down' by successively rejecting model parameterizations inconsistent with concentration or flux data for knockout mutants[18]. This approach ultimately leads to an empty ensemble if a large number of flux data for knockout mutants are imposed as required model output. A more efficient way to sample promising kinetic parameter space is achieved through integration of the EM procedure with a machine-learning inspired genetic algorithm (GA) that exchange the best reaction parameterizations across all models in the ensemble through the recombination operation[15]. This procedure was used to develop a core kinetic model (core model) of *E. coli* metabolism by seamlessly integrating flux data for a wild-type and seven mutant strains under aerobic conditions with glucose as the carbon substrate. The model was accurate in recapitulating genetic perturbations as long as growth conditions remained aerobic and the list of deleted genes was in the neighbourhood of the ones used during model parameterization[15]. In fact, a follow-up study revealed that the core model did worse than flux balance analysis (FBA) in predicting fluxes for fermentative (that is, anaerobic) growth[11]. These observations reaffirmed that, unlike FBA, kinetic model predictive ability is determined by the completeness of the modelled regulatory interactions and scope of imposed metabolic flux data in response to gene knockouts. Consequently, by increasing the set of flux data for mutants distributed across a wide range of pathways and augmenting the set of regulatory interactions, prediction fidelity is successively improved[11,15] while subsuming the benefits afforded by a genome-scale level description. These benefits include a detailed description of biomass and global inventory of all metabolites and cofactors. These observations motivated this study, where both the scope of flux data sets and the complexity of regulatory interactions are significantly increased over previous efforts in an attempt to endow the kinetic model with sufficient detail to enable the reliable prediction of perturbed metabolic phenotypes.

The k-ecoli457 genome-scale kinetic model of *E. coli* metabolism was parameterized by combining a machine-learning algorithm[15] and the EM formalism[18]. Model parameterization is performed by minimizing discrepancies between model predictions and experimentally measured steady-state flux distributions for 25 mutant strains including 21 genetically perturbed strains with glucose as the carbon substrate (19 under aerobic[20] and two under anaerobic conditions[21]), and four with different carbon substrates under aerobic conditions (three with pyruvate[22] and one with acetate[23]). Model predictions were tested against multiple experimentally measured data sets that were not used during model parameterization. These included (i) 898 steady-state metabolite concentrations for 20 of the mutant strains[20–23], (ii) 234 Michaelis–Menten constants (185 K_m and 49 k_{cat} values) extracted from BRENDA[24] and EcoCyc[25] and (iii) 320 literature reported product yields for designed strains covering 24 different bioproducts. Comparisons revealed that 66% of the predicted metabolite concentrations as well as 51 and 63% of the estimated K_m and k_{cat} values, respectively, are within the experimentally reported ranges. This level of agreement of k-ecoli457 with experimental data exceeded the metrics reached by the core model[15], despite the significantly increased scope of the model and coverage of fewer studied pathways. A primary reason for this prediction fidelity is that by directly imposing the biomass measurements in the model, the flux of over 30 reactions (coupled to biomass) in fatty acid and amino acid synthesis pathways is resolved. Notably, the average relative error of k-ecoli457 predictions for the product yield in 129 out of 320 designed strains is within 20% of the measured values. Stoichiometric model based techniques such as FBA, minimization of metabolic adjustment (MOMA) or maximization of product yield were within 20% of the experimentally reported yield for only 16, 18 and 65 of the designed strains, respectively. Overall, the predicted product yields by k-ecoli457 achieve significantly higher value of correlation with experimental data (that is, Pearson's correlation coefficient of 0.84) than FBA, MOMA or maximization of product yield (that is, 0.18, 0.37 and 0.47, respectively). These results quantitatively demonstrate that k-ecoli457 can reliably be used to predict genetically perturbed *E. coli* phenotypes under different growth conditions with a substantially higher accuracy than any other earlier modelling effort.

Results

k-ecoli457 construction and predictions. The k-ecoli457 model contains 457 reactions and 337 metabolites and includes all reactions from the largest previously published mechanistic kinetic model of *E. coli*[15] encompassing representations for most pathways from the genome-scale *i*AF1260 model (see Supplementary Data 1 for details). In general, out of 2,390 reactions in the *i*AF1260 model of *E. coli*, 1,603 reactions do not carry any flux, as they are either inactive (that is, 720 reactions) upon imposing the flux data of the reference (wild-type) strain or blocked (that is, 883 reactions)[26]. From the reactions that can carry flux, k-ecoli457 does not cover lipopolysaccharide and murein biosynthesis pathways as the molecular composition of some of these polymers is unknown and they only contribute $<6\%$ of the overall biomass[27]. In addition, 295 substrate-level regulatory interactions mined from enzyme biochemistry repositories such as BRENDA[24] and EcoCyc[25] were integrated in the model (see Fig. 1). In summary, k-ecoli457 is more than three times larger than that of the previously developed largest

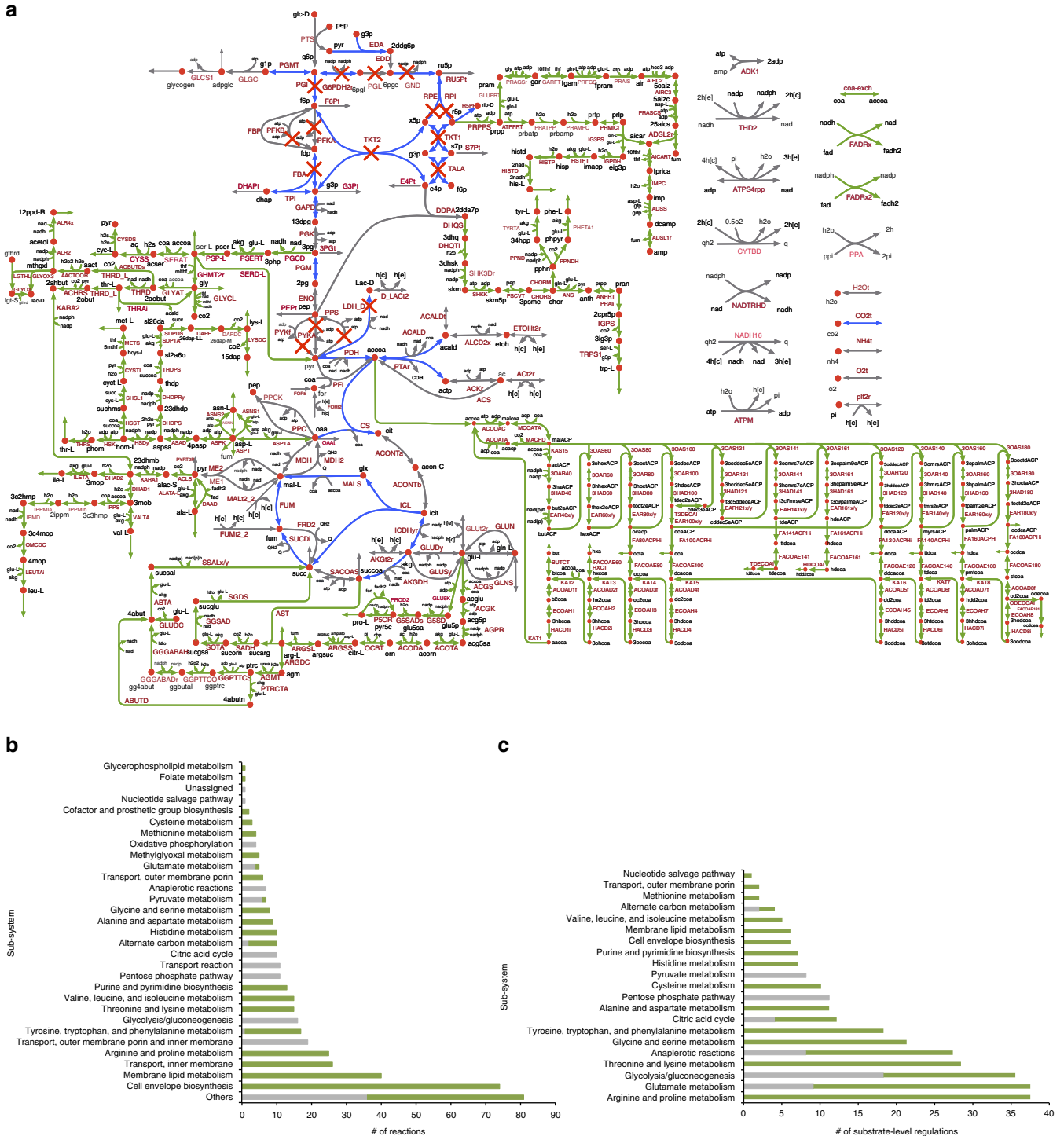

Figure 1 | A representation of k-ecoli457 and the core model. (**a**) A pictorial representation of the k-ecoli457 model of *E. coli* metabolism. Red X's denote the location of reaction deletions in the mutant data sets. Reactions in the previously developed core model[15] are shown in grey (no flux data) and blue (with flux data) while the additional reactions in k-ecoli457 are shown in green (no flux data). (**b**) Sub-system classification of reactions in the k-ecoli457 model. (**c**) Sub-system classification of the integrated regulatory interactions. Grey bars denote the content of the core model[15], while green bars denote the additional reactions/regulations included in k-ecoli457.

core kinetic model (core model)[15] (that is, 138 reactions, 93 metabolites and 60 substrate-level regulatory interactions).

For the k-ecoli457 model parameterization, an initial ensemble of elementary kinetic models that converge to the steady-state flux distribution of the reference (wild-type) strain was constructed (see Fig. 2a). A two-step optimization procedure was used to parameterize k-ecoli457. The first step identified the equivalent Michaelis–Menten constants (that is, K_m and v_{max}) using the experimentally measured flux data for the nineteen mutant strains grown aerobically with glucose. This is achieved by identifying the best combination of the sampled kinetic parameters in the ensemble (see Fig. 2b and Supplementary Methods, GA

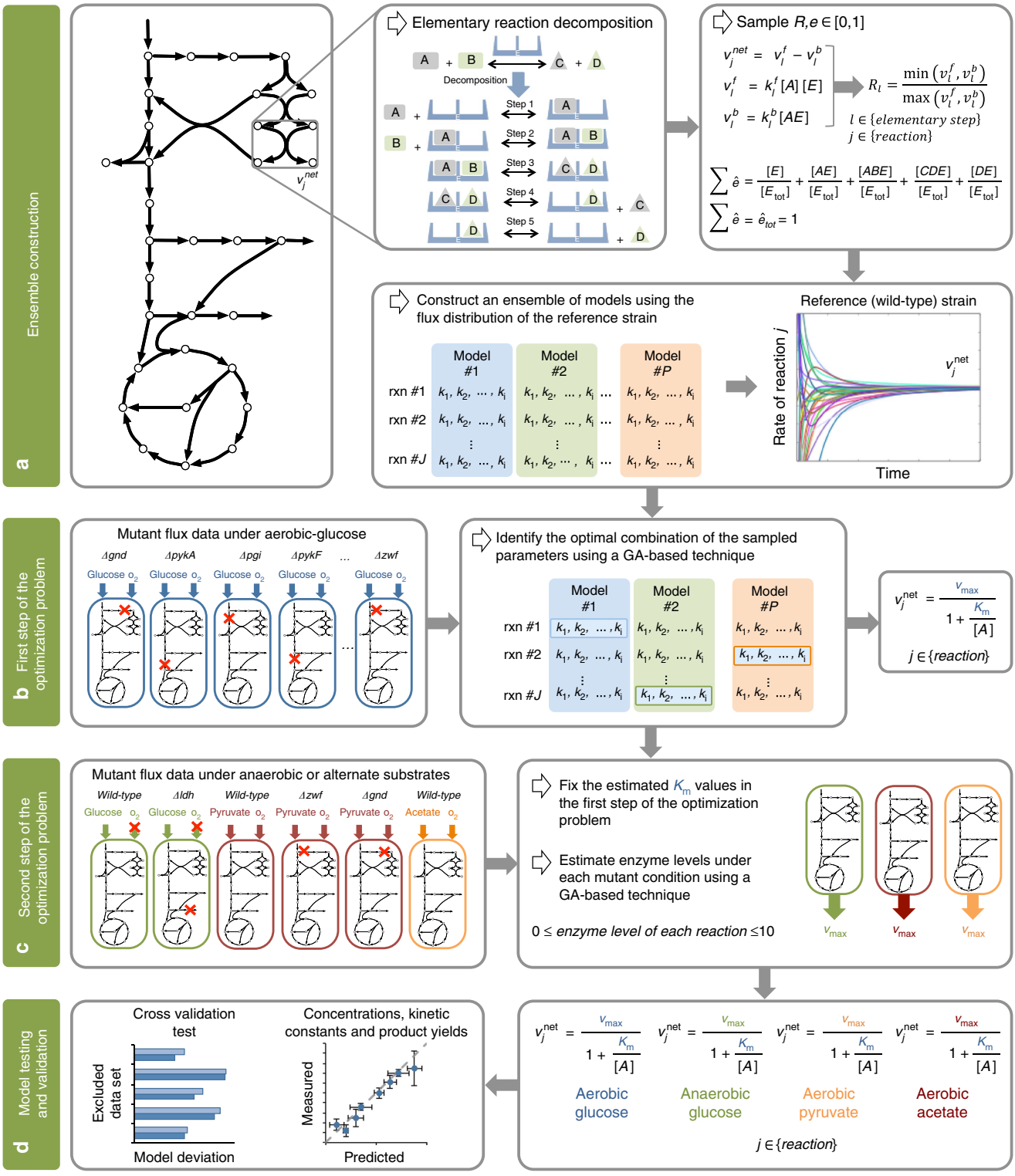

Figure 2 | A schematic representation of k-ecoli457 construction procedure. (**a**) Reactions are decomposed into their elementary steps and an ensemble of *P* models is constructed by sampling enzyme fractions and reaction reversibilities. (**b**) The first step of the optimization problem identifies the optimal combination of the sampled parameters by minimizing k-ecoli457 prediction deviation from the measured flux data in the nineteen mutant strains grown aerobically with glucose. (**c**) The second step of the optimization problem estimates the level of enzymes under the other three mutant conditions by minimizing k-ecoli457 prediction deviation from the measured flux data in the remaining six mutant strains grown anaerobically with glucose, aerobically with pyruvate and aerobically with acetate. (**d**) Model predictions are tested and validated using cross-validation analysis and experimentally measured metabolite concentrations, Michaelis–Menten constants and product yields.

implementation). Next, the estimated parameters (that is, K_m) were fixed and the levels of enzymes (that is, v_{max}) were estimated under the other three growth conditions (that is, anaerobically with glucose, aerobically with pyruvate and aerobically with acetate), separately, by solving the second step of the optimization procedure (see Fig. 2c). In general, the parameterized k-ecoli457 model performs well by attaining the predicted fluxes within the experimental ranges for 61% of the reactions with measured flux data in the mutant strains (see Supplementary Data 2 for the predicted and measured flux data sets). In particular, we observed that k-ecoli457 predictions are often (that is, 73%) completely within the experimental error ranges for the reactions eliminated in the mutant strains, including those in the glycolysis and pentose phosphate (PP) pathways (that is, blue coloured reactions crossed in Fig. 1). This is because the metabolic response upon elimination of the reaction quantifies the extent of the reaction's effect on metabolism and assesses the plasticity of the reaction loss in other mutation scenarios. For the remaining reactions, k-ecoli457 predictions are between 2 and 3 s.d. of the reported experimental ranges for 79 and 88% of the reactions, respectively (see Supplementary Fig. 2).

We also performed leave-one-out and leave-two-out cross-validation analyses to assess the robustness of the estimated parameters in k-ecoli457 compared with that of the core model (see Supplementary Methods, cross-validation analysis). A comparison showed that core model predictions exhibited higher deviations (that is, by about threefold) compared with those of k-ecoli457 in predicting metabolic fluxes of the cross-validated mutant (that is, predictions within 5% of the original model in k-ecoli457 versus 15% in the core model). This implies that k-ecoli457 parameterization is significantly more robust compared with that of the core model. In general, the analyses revealed that a diverse set of mutant experimental data for different pathways under different growth conditions is required to achieve a robust model parameterization[15]. For example, we observed that the failure of model cross-validation for mutant Δ*pgi* by the previously published core model[15] was resolved in k-ecoli457 through the presence of three additional mutant flux data sets (that is, Δ*pfkA*, Δ*pfkB* and Δ*fbaB*) that filled in information lost by removing the Δ*pgi* data set. Unsurprisingly, model parameterization was not as robust for the two mutant strains under anaerobic conditions (that is, wild-type and Δ*ldh*) as alluded by higher prediction deviations from experimental data upon removal of one of the two data sets (that is, a 14% increase in average scaled relative deviation). Because fluxes of the fermentative products (that is, formate, lactate, acetate and ethanol) are significantly different between the two data sets, both data sets are needed as they cannot complement information upon the loss of one of the two. These discrepancies propagate to some extent in other parts of the network, however, model prediction is still adequate for the rest of reactions (that is, an average 8% deviation from the experimental ranges).

Comparison of the estimated elementary kinetic parameters. We first compared the estimated values for each individual elementary kinetic parameter in k-ecoli457 with those in the previously published core model[15] by defining a confidence range of 10% from the estimated parameter values (see Supplementary Data 2 for the estimated parameters). The comparison revealed significant differences, as for 90% of the elementary kinetic parameters there was no overlap between the estimated ranges. The majority (that is, 63%) of the elementary kinetic parameters that changed in value significantly belong to reactions utilizing cofactors (that is, atp, adp, amp, nad(h), nadp(h)). This is due to the fact that the core model[15] could not accurately track the concentration of cofactors as only a limited number of reactions that contribute to the cofactor balance were included. For example, amino acid and membrane lipid metabolism reactions were absent in the core model[15]. Thus, errors in the estimation of cofactor concentrations propagated to the corresponding elementary kinetic parameter estimates.

Comparison of predicted Michaelis–Menten parameters. To assess the accuracy of the estimated elementary parameters, we compared the corresponding Michaelis–Menten constants with experimental values from BRENDA[24] and EcoCyc[25]. In accordance with the EM approach, the elementary kinetic parameters and therefore the Michaelis–Menten constants, are scaled by the corresponding metabolite concentrations in the reference (that is, wild-type) strain. We used the experimentally reported concentration data for the wild-type strain[20,28] to rescale the estimated Michaelis–Menten constants. We extracted 234 measured Michaelis–Menten constants including 185 K_m and 49 k_{cat} values for the reactions present in k-ecoli457 (see Supplementary Data 2 for the estimated and measured ranges). In general, the results showed that 51% of the estimated K_m values and 63% of the k_{cat} values in k-ecoli457 overlap with the reported experimental value ranges (see Fig. 3a,b). For the parameters shared by both the core model and k-ecoli457, the majority (that is, 77% for K_m and 86% for k_{cat}) of the predictions within the confidence ranges in the core model were predicted again within ranges by k-ecoli457. Notably, the computed Michaelis–Menten constants in k-ecoli457 for amino acid and pyruvate metabolism exhibited significantly higher agreement with experimental data compared with the values in the core model[15] (see Fig. 3c,d). In addition, we performed the same comparisons for only those parameters whose confidence ranges did not exceed two orders of magnitude which included 120 K_m and 14 k_{cat} parameters. Comparisons revealed that for this set 36% of the estimated K_m and 29% of the estimated k_{cat} values are within the experimentally reported ranges. Overall, these analyses demonstrate the importance of integrating pathways distant to central metabolism into the model as they can affect the quality of parameterization through the pools of shared metabolites such as cofactors. This motivates the development of flux data sets in response to genetic perturbations distal from central metabolism pathways to improve overall model parameterization. It is also important to note that the large magnitude in the confidence ranges for both measured and computed k_{cat} values detracts from the confidence for some of the estimated k_{cat} values.

Comparison of predicted concentrations against unused data. Similar to the Michaelis–Menten constants, the predicted normalized steady-state metabolite concentrations of the mutant strains were first recast as actual concentrations. In total, we extracted 898 experimentally measured concentrations spanning ~45 metabolites in twenty mutant strains (a total of 294 concentrations in the core model[15]). The comparison showed that 66% of the predicted metabolite concentration ranges overlap with the experimentally reported ranges (see Fig. 4a and Supplementary Data 2 for the predicted and measured concentration data sets). We note that for the metabolite concentrations present in both the core and k-ecoli457 models, the majority of concentrations (that is, 80%) that were within the confidence ranges in the core model[15] were again predicted within the ranges by k-ecoli457. In particular, we observed k-ecoli457 achieved higher accuracy for metabolites in pathways challenged with more mutant flux data (that is, pyruvate metabolism, the tricarboxylic acid (TCA) cycle and the PP pathway). For example, the predicted concentration ranges of the commonly

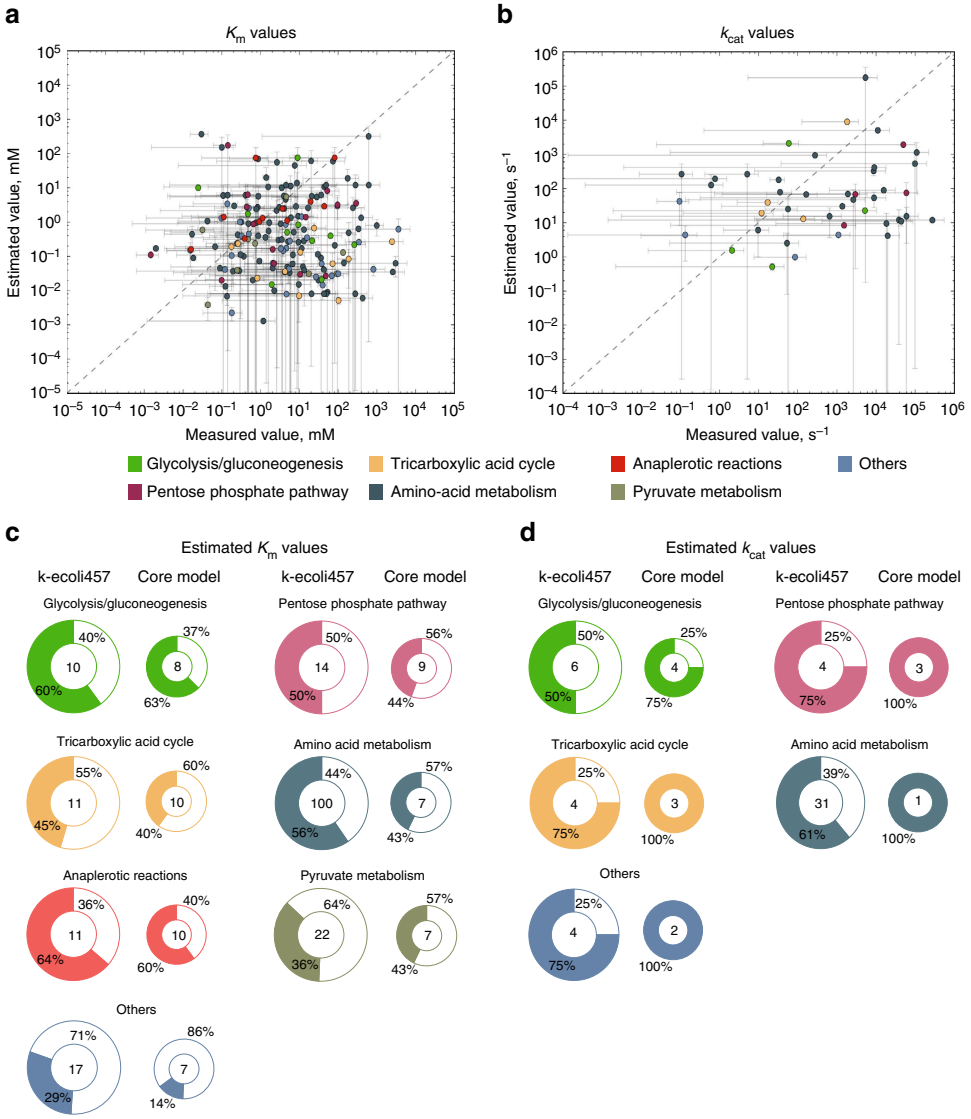

Figure 3 | Estimated and measured kinetic parameters. Comparison of the computed (**a**) Michaelis constant K_m and (**b**) turnover number k_{cat} in k-ecoli457 and the values reported in BRENDA[24] and EcoCyc[25]. The error bars denote 1 s.d. confidence interval for the corresponding parameter, the dots represent the mean value of the measured versus computed parameters and dashed lines show equal predicted and measured parameters. Comparison of the computed (**c**) K_m and (**d**) k_{cat} in k-ecoli457 (left) and the core model (right)[15]. The numbers within the circle plots indicate the number of measured parameters and the darker sectors represent the portion of the parameters whose estimated ranges overlap experimentally reported ranges.

measured metabolites acetate, lactate, succinate, malate, fumarate, ribose-5-phosphate and xylulose 5-phosphate by k-ecoli457 showed overlap with experimental ranges in the 20 mutant strains for 80% of the measured concentrations, while the core model predictions showed overlap for only 31% of the measured concentrations (see Fig. 4c). This highlights the efficacy of parameter estimation for reactions directly affected by the mutations in the training data sets. In addition, energy and cofactor regulatory information, whenever available in BRENDA or EcoCyc (that is, 64 regulations), were included to account for the substrate-level effect of redox and energy regulation in k-ecoli457. In general, we observed more accurate predictions for the concentration of cofactors (that is, nad(h), atp, amp and adp) with k-ecoli457 compared with the core model[15] (82% versus 43% of the predicted concentrations overlap with experimental ranges, respectively). This is because a more complete integration of cofactor utilizing pathways in k-ecoli457 led to more accurate component balances, therefore, yielding more accurate concentration predictions. The accuracy of prediction, however, was often limited to the metabolites in central metabolism. For example, a high-prediction deviation was observed for metabolites in amino acid metabolism in Δ*pgi*, Δ*gnd*, Δ*zwf*, Δ*rpiA*, Δ*rpiB* and wild-type strain grown with acetate (see Fig. 4b). This is a manifestation of the lack of flux data sets for the associated pathways during model parameterization. In addition, the important transcriptional regulation of cellular growth on the oxidative section of the PP pathway[29] was not captured in k-ecoli457. This led to erroneous predictions for the aromatic amino acid concentrations in five knockout mutants. Even with missing experimental data and regulatory interactions, we note that 59% of the predicted concentrations for the metabolites in amino acid metabolism are still well within the experimental measurement error ranges. In fact, this tight coupling of the metabolite concentration and kinetic parameters through reaction kinetics leads to accurate predictions by providing mutual backup for missing information. In addition, a comparison between the

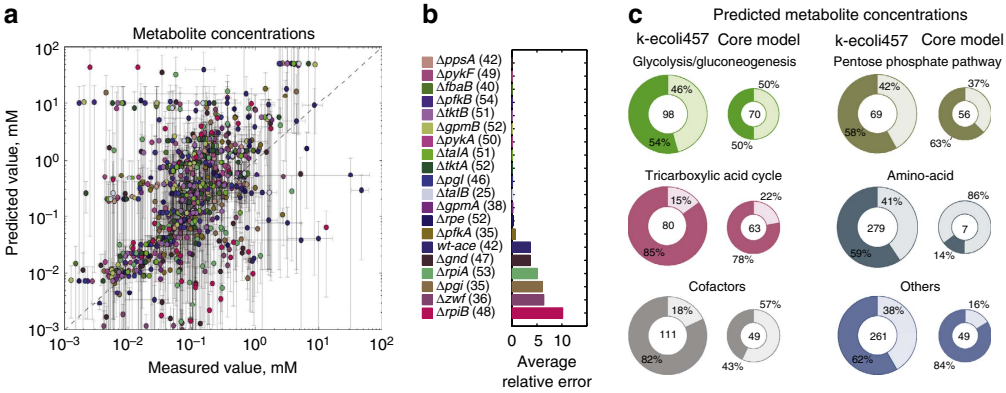

Figure 4 | Predicted and measured metabolite concentrations. (**a**) Comparison of the predicted steady-state metabolite concentration in k-ecoli457 and the experimentally reported values[20,23]. The error bars denote 1 s.d. confidence interval for the corresponding concentration, the dots represent the mean value of the measured versus predicted parameters and dashed line shows equal predicted and measured concentrations. (**b**) The average relative error of the predicted metabolite concentrations in the twenty mutant strains. The number next to each mutant represents the number of metabolites with measured concentration. (**c**) Comparison of the predicted steady-state metabolite concentration in k-ecoli457 (left) and the predicted values in the core model (right)[15]. The numbers within the circle plots indicate the total number of metabolites with measured concentration in the twenty mutant strains and the darker sectors represent the portion of the metabolites whose predicted concentration ranges overlap experimentally reported ranges.

predicted and measured concentrations with confidence ranges no greater than two orders of magnitude (that is, 786 concentrations) revealed that 63% of the estimated concentrations are still within the experimentally reported ranges.

Predicted product yield for 320 overproducing mutants. We extracted experimentally reported yields of overproduction for 320 different engineered *E. coli* strains for a diverse range of bioproducts from 47 separate studies and compared them with k-ecoli457, FBA, MOMA and maximization of product yield predictions (see Methods for calculation). The target products included 24 different chemicals under both aerobic (184 mutants) and anaerobic (136 mutants) conditions spanning biofuels (ethanol, butanol and isobutanol), pharmaceuticals and nutraceuticals (artemisinin and naringenin), polymer precursors (hydroxybutyrate, hydroxystyrene, butanediol, acetate, formate, glucaric acid and styrene) and commodity and specialty chemicals (lycopene, indigo, malate, fumarate, succinic, cinnamic and muconic acids) (see Methods and Supplementary Data 2).

The Pearson's correlation coefficients between the experimental data and product yield predictions by k-ecoli457, FBA, MOMA and maximization of product yield, respectively, are 0.84, 0.18, 0.37 and 0.47 ($P<10^{-3}$). A comparison also revealed that the product yield in 129 out of 320 designed strains are predicted within 20% of the experimental measurements by k-ecoli457, whereas FBA, MOMA and maximization of product yield predicted the product yield for only 16, 18 and 65 of the designed strains, respectively, with the same accuracy. In general, under aerobic conditions k-ecoli457 closely reproduces the yields for products that branch out from the PP pathway (for example, naringenin and muconic acid), the TCA cycle (for example, malate, succinate, fumarate, threonine and butanol) and pyruvate metabolism (for example, lactate, isobutanol and 2,3-butanediol). Both FBA and MOMA tend to underestimate their yields as more carbon is diverted towards biomass (see Fig. 5a) with the exception of naringenin and threonine where FBA overestimates their yield as it is blind to substrate-level regulatory interactions. For the naringenin engineered strain, FBA does not capture two strong feedback inhibitions of tyrosine and phenylalanine on the chorismate pathway[30], thus pooling additional arabino-heptonate 7-phosphate (2dda7p) towards the flavanone pathway. Likewise, regulatory inhibitions in glycolysis, chorismate and the PP pathways limit flux towards erythrose 4-phosphate, phosphoenolpyruvate and shikimate. The same inability to capture multiple regulatory interactions in the aspartate, lysine, asparagine, methionine and isoleucine pathways that share precursor oxaloacetate with the threonine pathway leads to an over prediction for the threonine yield by FBA. Maximization of product yield not surprisingly typically overestimated experimental product yield. Product yields under anaerobic conditions were again tracked better by k-ecoli457 as both FBA and MOMA often underestimate them (65 and 82% of strains, respectively) by directing all carbon flux towards biomass (see Fig. 5b). We find that reliable product yield prediction by k-ecoli457 requires both the change in the enzyme level under anaerobic conditions and the activation of relevant substrate-level regulations. For example, for lactate, malate and succinate overproducing mutants, the experimentally reported high activities of the glyoxylate shunt[31] and reductive section of the TCA cycle[32] were correctly captured by the updated enzyme levels under anaerobic conditions (see Supplementary Data 2).

k-ecoli457 often underestimates the yield of acetate under aerobic conditions as well as ethanol and formate under anaerobic conditions (see Fig. 5c,d) due to inadequate training flux data sets involving mutants of the respective pathways. For example, only two mutant flux data sets under anaerobic conditions (that is, a wild-type and Δ*ldh*) were available for model parameterization implying insufficient training data as affirmed in the cross-validation analysis (see Supplementary Methods, cross-validation analysis). Another important factor that is often overlooked is the lack of flux data sets that span multiple growth phases or dilution rates, and not just early growth phase or low dilution rates. In particular, acetate production is poorly captured by k-ecoli457 as the fluxomic information for all nineteen mutant strains under aerobic glucose conditions were measured under low dilution rates (that is, $0.2\,h^{-1}$), while product yield data were often reported in higher dilution rates. These observations identify the limits of prediction and pinpoint needed flux measurements under different dilution rates, growth phases, alternate substrates and anaerobic conditions using both batch[33] and continuous cultures[16]. Nonetheless, despite of data scarcity, k-ecoli457 predictions remain substantially better than FBA, MOMA and maximization of product yield. In general, we observed that both FBA and MOMA predictions demonstrated reasonable agreement with experimental

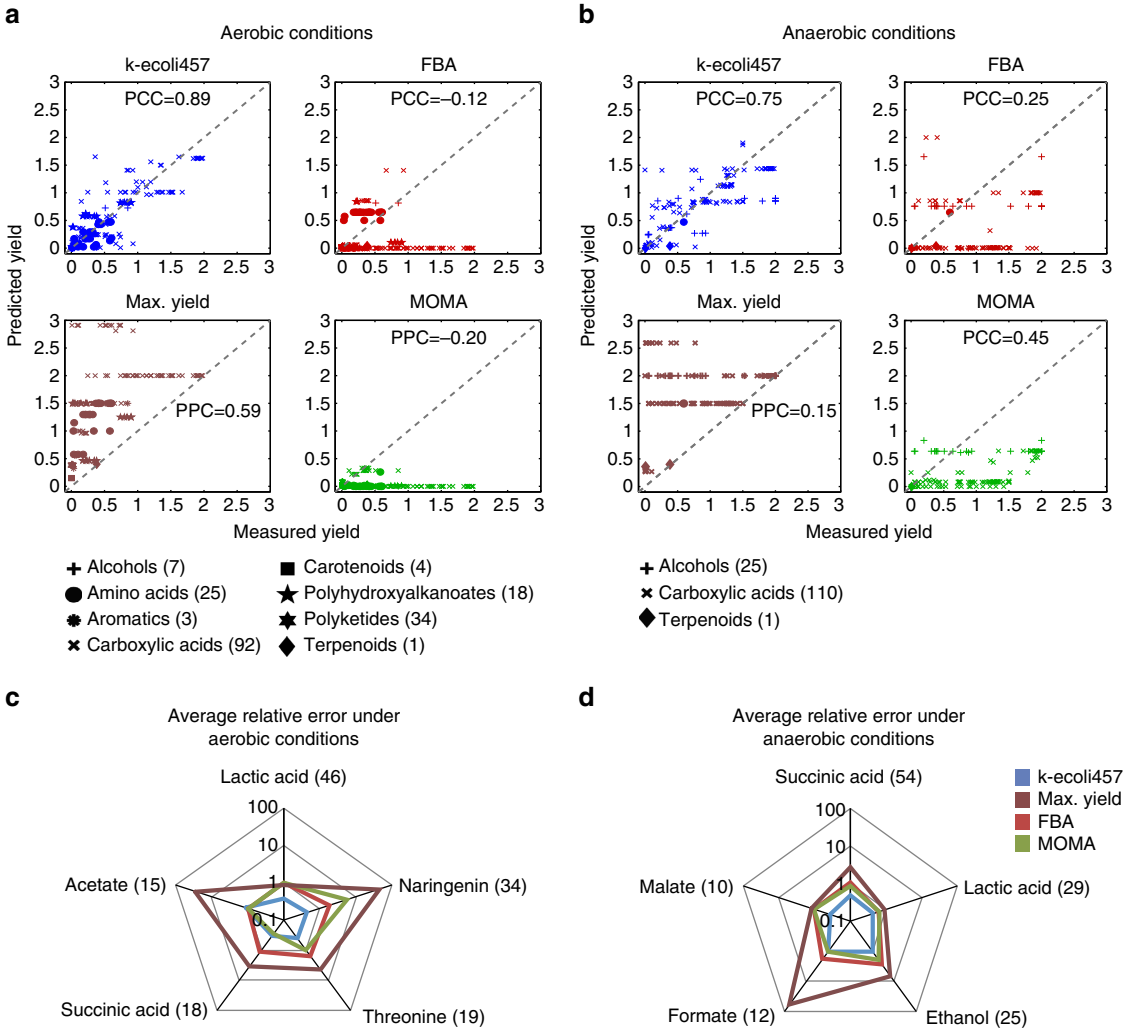

Figure 5 | Predicted and measured product yields. Comparison of the experimentally reported yield of overproduction for 320 different engineered *E. coli* strains and k-ecoli457, FBA, MOMA and maximization of product yield predictions under (**a**) aerobic and (**b**) anaerobic conditions. PCC is the Pearson's correlation coefficient between the predicted product yield values and experimental data. Dashed lines show equal predicted and measured yields and the number next to each category represents the total number of designed strains extracted from literature. Average relative error is displayed between k-ecoli457 predictions and experimental data under (**c**) aerobic and (**d**) anaerobic conditions for the chemicals with at least ten data points. The number next to each chemical shows the number of designed strains extracted from literature. The reported average relative error was calculated as the average relative difference between the measured and predicted product yield values (product mol per glucose mol).

data only whenever the targeted product is energy/redox coupled with biomass and measured product yield value scaled by the maximum theoretical yield was relatively low (that is, <0.2). The maximization of product yield better reflected the measured product yields whenever the scaled product yield value was relatively high (that is, higher than 0.9). Whenever the designed mutant's product yield did not significantly differ from wild-type, MOMA performed better than FBA (for example, lycopene and styrene under aerobic conditions).

Discussion

Here, we developed k-ecoli457, a kinetic model of *E. coli* metabolism that approaches genome-scale coverage (457 reactions and 337 metabolites). Comparisons of k-ecoli457 with a previously constructed core model[15] revealed significant improvement in prediction accuracy despite the significantly expanded model scope and the corresponding paucity of fluxomic data for distal pathways. We found that the global inventory of highly participating metabolites (that is, cofactors), the large number of resolved reaction fluxes coupled to the biomass measurement and the complete description of the proportions of metabolites sequestered within biomass contributed to the prediction improvements in k-ecoli457. A comparison between predicted fluxes, however, revealed that the average relative error of k-ecoli457 when applied to only the core reactions is higher compared with the core model (3 versus 10%). This is because the core model has been tuned exclusively for these reactions whereas k-ecoli457 must describe four times more fluxes in (25 versus 7) experimental data sets. Cofactor concentrations in k-ecoli457 now participate in hundreds of reactions making it very difficult to pinpoint a unique value that matches all experimental data. Significant uncertainty in the experimental data sets across multiple pathways also contributes to the inability to perfectly match the core reaction set. Despite these challenges, 61% of the predicted fluxes by k-ecoli457 (78% in the core model) are within 1 s.d. of the experimental data. In addition, the agreement of the experimental yields (see Fig. 5) provides additional confidence for the robustness of the developed model. Prediction deficiencies remained for pathways lacking metabolic flux data sets in

response to genetic perturbations (for example, membrane lipid metabolism and ED pathways). For example, we observed that even after the inclusion of additional flux data sets for k-ecoli457 parameterization compared with that of core model, the activity of the ED pathway was not properly captured. As the majority of the training flux data sets (that is, 22 out of 25) had an inactive ED pathway, the k-ecoli457 model predicted the same. While simple inclusion of additional data sets with nonzero ED flux may have rectified this limitation, this *a posteriori* correction may not be a fair representation of the proposed model and methodology. These limitations are likely to be ameliorated as expanded metabolomic (for example, MetaboLights[34]) and fluxomic (for example, CeCaFDB[35]) data sets are becoming increasingly available. Given data sets that span the metabolic capabilities of *E. coli*, the proposed machine-learning inspired parameterization strategy demonstrated that it is indeed possible to train a single model to predict the genetic and environmentally perturbed phenotypes with fidelity. In the same spirit, the same multi-data set parameterization concepts can be leveraged for applications of kinetic models in personalized healthcare[36], biomarker identification[37], drug discovery[38] and modelling of microbial communities[39].

Remaining challenges not addressed in this effort include allowing for substrate(s) uptake rates to become an output of the kinetic model. In k-ecoli457 all mutant fluxes in the training data sets were scaled with the corresponding substrate uptake rate. Given substrate uptake rates data sets[24,40] for different mutations and growth conditions, a kinetic formalism that describes carbon uptake could be constructed and parameterized largely independent of internal reactions. In addition, large-scale metabolomic data sets for absolute or even relative concentrations[34] can directly be ported in the machine-learning algorithm to further constrain model parameter values. This was not attempted here as we chose to treat metabolomic data *a posteriori* as a model consistency check. In addition, the assembled compilation of experimental product yields for 320 designed strains could serve as a starting point for more comprehensive compilations[41] that will help to fairly assess follow-up efforts aimed at improving the accuracy and coverage of k-ecoli457. Kinetic model parameterization using such comprehensive data sets, however, must be carefully interpreted. For example, there exists substantial evidence for the presence of pathway channelling[42–44] in metabolism as a mechanism for increasing the local concentration of metabolites and thus boost reaction rates (for example, channelling of glycolysis intermediates in *E. coli*[45]). As a result, if the relevant metabolite participates in other reactions, then the kinetic model will simply pool all the 'local' concentrations of the metabolite within a single 'average' concentration. This difference between local and average concentrations will propagate in the value of k_{cat} so as the reaction flux value is matched. This means that the values of the estimated metabolite concentrations and k_{cat} values may not always reflect *in vivo* kinetics but rather represent cell-averaged values. In addition, many other factors such as growth stage of the strain or even experimental group carrying out the fluxomic analysis can affect the quality and reproducibility of the flux data sets. These factors can lead to different flux data sets for exactly the same genotype (for example, different flux distributions for wild-type strain in ref. 20 versus ref. 46 or the different effect of *pfkA* and *pfkB* knockouts on glycolytic activity in ref. 20 versus ref. 47). Including conflicting flux data sets would cause significant problems in parameter estimation as the model will try to match the average values between the two data sets that are likely not physiologically relevant. Accounting for such variations remains an important topic to be addressed in follow-up studies.

Looking beyond substrate-level regulation, the increasing availability of data sets that provide genome-wide collections of interactions between mRNAs[48], proteins[49] and metabolites[50] has dramatically expanded our knowledge of transcriptional, (post)translational and substrate-level interactions. Although the relative contribution of each of these regulatory layers is likely to be context dependent[51], systematic implementation of regulatory events across multiple layers will ultimately be needed. Successful implementation of allosteric modification[52] and substrate-level regulation[11] with elementary kinetic mechanisms described in this paper establish a foundation for including additional layers. Developing efficient methods for reducing the complexity of models into more manageable ones without any information loss would also increase usability and community acceptance. This will also reduce computation complexity of integrating kinetic information into computational strain design protocols[4,11]. In particular, we have recently integrated 36 reactions with a kinetic description in the core model for strain design using the k-OptForce procedure[11]. Moving towards strain design with a full kinetic representation will ultimately require advances in solution techniques and accelerated ways of reaching steady-state fluxes and concentrations.

Methods

Model scope and initial ensemble construction. A metabolic model of *E. coli* metabolism composed of 457 reactions and 337 metabolites was constructed (see Fig. 1a and Supplementary Data 1). This model accounts for all reactions from the genome-scale *i*AF1260 model[27] that carry flux under the experimental conditions of the flux measurements except lipopolysaccharide and murein biosynthesis. These include reactions in glycolysis/gluconeogenesis, the PP pathway, the TCA cycle, anaplerotic reactions, amino acid synthesis/degradation, fatty acid oxidation/synthesis and a number of reactions in other parts of the metabolism, such as co-factor and alternative carbon metabolism, membrane lipid and cell envelope synthesis and oxidative phosphorylation pathways (see Fig. 1b). In addition, 295 regulatory interactions were extracted from BRENDA[24] and EcoCyc[25] and included in the model by introducing new regulatory reactions (see Fig. 1c) (see Supplementary Data 1 for a comprehensive list)[53]. A simplified version of the biomass equation for *E. coli* described in (ref. 54) was also integrated into the model including all amino acids as well as other biomass constituent precursors in the central metabolism. This biomass equation was only used to 'drain' biomass precursors from the pathways absent in k-ecoli457 (see Fig. 1).

Following the EM procedure[19], all metabolic reactions and regulatory interactions in the model were first decomposed into their elementary steps and a mass action kinetic was developed for each elementary reaction. Elementary kinetic parameters were next expressed in terms of elementary reaction parameters, reaction reversibilities R and enzyme fractions $\hat{e}$, both bounded between zero and one. Reaction reversibility is defined as the elementary rate of the backward reaction divided by the forward reaction for each elementary step ($R = \left(v^b/v^f\right)^{sign(V_{net})}$ where V_{net} is the net flux of the reaction). Enzyme fraction is defined as the level of unbound enzyme e normalized by the total pool of the specific enzyme ($\hat{e} = e/e_{tot}$ and $\hat{e}_{tot} = \sum \hat{e} = 1$). Metabolite concentrations were also normalized using the reference (wild-type) strain values. Elementary kinetic representations and scaled metabolite concentrations provide a convenient scaling between zero and one for parameter sampling for enzyme fractions and reaction reversibilities consistent with thermodynamic principles[19]. We used experimentally measured flux data for a wild-type *E. coli* strain growing aerobically with glucose[20] as the reference strain for the ensemble scaling and construction.

The reference flux distribution was obtained by first maximizing biomass yield (per 100 mmol g DW^{-1} h^{-1} of glucose uptake and 200 mmol g DW^{-1} h^{-1} of oxygen uptake) in the *i*AF1260 model using FBA while imposing the experimental flux measurements (for 43 reactions)[20] as constraints. Next, flux variability analysis[55] was used to identify the flux ranges for reactions without experimental measurements upon fixing the biomass flux at the value obtained in the first step. The obtained flux ranges were then used to constrain the reactions without available measurements. A feasible flux distribution was then obtained by imposing the experimental data along with flux variability analysis-driven flux ranges in our model by maximizing biomass yield. This flux distribution was then used to anchor steady-state fluxes during the uniform sampling of the model parameters and elementary kinetic ensemble construction. We used an ensemble size of $2^{17} = 131{,}072$ models, as no improvement in model predictions and convergence to the optimal solution was achieved for a larger ensemble size.

Multiple flux data sets for model parameterization. Model parameterization was carried out using steady-state flux data for 25 mutant strains of *E. coli* under

aerobic as well as anaerobic conditions with multiple carbon substrates. In particular, the flux data sets include nineteen single knockout mutant strains growing under aerobic conditions (that is, Δ*pgi,* Δ*pykA,* Δ*pykF,* Δ*ppsA,* Δ*gnd,* Δ*zwf,* Δ*rpe,* Δ*pfkA,* Δ*pfkB,* Δ*fbaB,* Δ*gpmA,* Δ*gpmB,* Δ*pgl,* Δ*rpiA,* Δ*rpiB,* Δ*talA,* Δ*talB,* Δ*tktA* and Δ*tktB*)[20] and two strains growing under anaerobic conditions (that is, a wild-type and Δ*ldh*) with glucose as the carbon substrate[21], three strains growing under aerobic conditions with pyruvate as the carbon substrate (that is, wild-type, Δ*zwf* and Δ*gnd*)[22] and one strain growing under aerobic conditions with acetate as the carbon substrate (that is, wild-type)[23]. In total, we extracted flux measurements for ~30 reactions in each mutant strain including the major intracellular reactions in glycolysis, the PP pathway and the TCA cycle (see Fig. 1a)[20]. In addition, we performed a coupling analysis with the *E. coli* biomass reaction[56] to augment the set of flux data beyond those in central metabolism. We found that between 33 and 36 reactions (depending on the specific mutant) were fully coupled with biomass production and were added to the list of measured fluxes. These reactions were, not surprisingly, located in biomass constituent biosynthesis pathways, such as cell envelope metabolism and amino acid synthesis pathways (that is, threonine, lysine, methionine, valine, leucine, isoleucine, tyrosine, tryptophan, phenylalanine and histidine metabolism) (see Supplementary Data 1).

Confidence intervals of the estimated and measured parameters. Confidence intervals were provided only for the measured fluxes of the wild-type strain in the training data sets[20]. Consequently, we used the same confidence intervals reported for each reaction in the wild-type strain to construct a flux range (that is, 1 s.d. confidence interval) around the reported values in the mutant strains. Likewise, for the metabolites with no reported confidence range in the mutant strains, we used the same confidence intervals reported for each metabolite in the wild-type strain to construct a concentration range (that is, 1 s.d. confidence interval). In addition, for the experimentally measured Michaelis–Menten constants (that is, K_m and k_{cat}) with multiple reported values in BRENDA[24] and EcoCyc[25], confidence intervals were calculated using 1 s.d. from the mean of the reported parameter values. For parameters with only a single measured value we used a range of 10% from the reported parameter values as confidence intervals (see Supplementary Data 2). In addition, the predicted normalized metabolite concentrations and Michaelis–Menten constants were scaled by the corresponding metabolite concentration confidence ranges in the reference (wild-type) strain to convert them into the actual ranges (see Supplementary Methods, conversion of estimated elementary kinetic parameters). To reduce the uncertainties of the base concentrations, we extracted seven additional measured concentration data sets for the wild-type strain in addition to the reference strain data and used their confidence range intersections as the base ranges[16,28,33,57–60] (see Supplementary Table 2). For metabolites with no concentration data in the wild-type strain (that is, 12%) we used the reported ranges in the *i*AF1260 model of *E. coli*[27]. Next, the agreement quality is gauged as an overlap between the reported and predicted confidence ranges.

Estimation of k-ecoli457 parameters. Model parameterization was performed in two stages: (a) estimation of the elementary kinetic parameters and isozyme activity using the experimentally measured flux data for the single knockout mutants growing under aerobic minimal glucose conditions (that is, a total of nineteen flux data sets); and (b) estimation of enzyme levels under minimal glucose fermentative (anaerobic) conditions and alternate carbon substrates pyruvate and acetate (that is, a total of six flux data sets).

Estimation of elementary kinetic parameters and isozyme activity. A GA implementation was used that minimized the deviation of the model predictions from the available flux measurements for the nineteen mutant strains grown aerobically with glucose[15]. In essence, this optimization based approach identifies the optimal combination of the sampled parameters and consequently the equivalent K_m and v_{max} values in Michaelis–Menten description by permuting parameterizations for different reactions across models in the ensemble. The consequence of an enzyme knockout is captured by fixing the total pool of the normalized enzyme level $\hat{e}_{tot}$ to zero (that is, $\hat{e}_{tot} = \sum \hat{e} = 0$) for the deleted reaction, while $\hat{e}_{tot}$ is assumed to remain unchanged for the remaining reactions. For the reactions catalysed by isozymes, deletion of one isozyme reduces (total enzyme) $\hat{e}_{tot}$ to a level between zero and the reference level (that is, $0 \leq \hat{e}_{tot} \leq 1$), as it is not clear the extent at which the remaining isozyme can complement for the lost activity. Therefore, we allowed the minimization of the model predictions with experimental measurements to arrive at the best value for $\hat{e}_{tot}$ while maintaining the sum of $\hat{e}_{tot}$ for all the isozymes to one. In total, we allowed the normalized enzyme levels $\hat{e}_{tot}$for seven reactions catalysed by 12 isozymes to vary (that is, Δ*pfkA,* Δ*pfkB,* Δ*fbaB,* Δ*gpmA,* Δ*gpmB,* Δ*pgl,* Δ*rpiA,* Δ*rpiB,* Δ*talA,* Δ*talB,* Δ*tktA* and Δ*tktB* under aerobic conditions with glucose as the carbon substrate) during step one (see Supplementary Methods, GA implementation).

Estimation of enzyme levels. Upon fixing the estimated elementary kinetic parameters (that is, K_m), we next estimated $\hat{e}_{tot}$ for the mutant strains growing anaerobically and those with pyruvate or acetate as the carbon substrate, separately. We note that this is equivalent to estimating v_{max} ($= \hat{e}_{tot} k_{cat}$) in Michaelis–Menten description, while fixing K_m to the estimated values (see Supplementary Methods, GA implementation). This is because we anticipated that only the enzyme levels are likely to change significantly when growth is switched from aerobic to anaerobic or to an alternate carbon substrate[61–63]. Therefore, the $\hat{e}_{tot}$ for all reactions in the model were allowed to vary from zero (that is, deletion) to a 10-fold overexpression (that is, $0 \leq \hat{e}_{tot} \leq 10$) for mutant strains growing (i) anaerobically with glucose (that is, two flux data sets), (ii) aerobically with pyruvate as the carbon substrate (that is, three flux data sets) and (iii) aerobically with acetate (that is, one flux data set). The total number of enzymes whose total normalized pool was allowed to vary was kept at a minimum to avoid model overparameterization (see Supplementary Methods, enzyme level changes).

Evaluation of predicted yields for overproducing mutants. We implemented enzyme level changes by allowing the total pool of the normalized enzyme to vary between a 10-fold downregulation and the wild-type level ($0.1 \leq \hat{e}_{tot} \leq 1$) for reported enzyme downregulations. For enzyme upregulations, the normalized enzyme level was set between the wild-type level and a 10-fold upregulation ($1 \leq \hat{e}_{tot} \leq 10$). This approximation was used because the level of enzyme change is often not available in the relevant literature. Gene deletions were implemented by setting the $\hat{e}_{tot}$ of the encoded enzyme to zero.

Least squares model parameterization and yield analysis. The objective function of the parameterization problem z is to minimize the relative deviation of k-ecoli457 predicted flux v_j from the experimental data v_j^{exp} for reaction j with measured data N across all the mutant strains M. For each reaction with measured flux data, the average relative error is scaled by its coefficient of variation CV_j to capture the reported uncertainty in the experimental data (average scaled relative deviation)[64]. As a result, the reactions with tighter confidence intervals have a larger contribution in the objective function.

$$z = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{j=1}^{N}\left(\frac{1}{CV_j}\left|\frac{v_j - v_j^{exp}}{v_j^{exp}}\right|\right)$$

Product yield y is defined as the rate of reaction that produces the product per rate of uptake reaction (product mol per glucose mol).

$$y = \frac{v^{product}}{v^{glucoseuptake}}$$

For comparisons, the majority of experimental studies reported only the measured product yield with no confidence range. Consequently, for each product yield the error was calculated based on relative deviation between the measured y^{exp} and predicted y^{pre} values.

$$z = \left|\frac{y^{pre} - y^{exp}}{y^{exp}}\right|$$

All parameterization problems were solved using a GA implementation. All calculations, including the ensemble construction and the GA problems, were implemented in MATLAB (MathWorks Inc.) and solved in parallel on three nodes of Intel Xeon E5-2670 v2 with 2.5 GHz (20 cores per node and 256 GB RAM) on the Penn State Lion-X clusters. The k-ecoli457 parameterization took ~58,800 CPU-hour. Integrating k-ecoli457 (that is, solving the system of ordinary differential equations) for a given metabolic condition also takes up to a few minutes. The estimated elementary kinetic parameters as well as stoichiometry matrix of the model are available in Supplementary Data 1 and the executable k-ecoli457 is available for download in Supplementary Software 1 and at http://www.maranasgroup.com as a MAT-file (MathWorks Inc.).

Code availability. The authors declare that the code supporting the findings of this study is available within the article's Supplementary Information files (Supplementary Software 1) and at http://www.maranasgroup.com as a MAT-file (MathWorks Inc.).

Data availability. All data generated or analysed during this study are included in this published article and its Supplementary Information files.

References

1. Cheng, J. K. & Alper, H. S. The genome editing toolbox: a spectrum of approaches for targeted modification. *Curr. Opin. Biotechnol.* **30,** 87–94 (2014).
2. Casini, A., Storch, M., Baldwin, G. S. & Ellis, T. Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.* **16,** 568–576 (2015).
3. Strotksaya, A., Semenova, E., Savitskaya, E. & Severinov, K. Rapid multiplex creation of *Escherichia coli* strains capable of interfering with phage infection through CRISPR. *Methods Mol. Biol.* **1311,** 147–159 (2015).
4. Segre, D., Vitkup, D. & Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl Acad. Sci. USA* **99,** 15112–15117 (2002).

5. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84,** 647–657 (2003).
6. Ranganathan, S., Suthers, P. F. & Maranas, C. D. OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* **6,** e1000744 (2010).
7. Cardenas, J. & Da Silva, N. A. Metabolic engineering of *Saccharomyces cerevisiae* for the production of triacetic acid lactone. *Metab. Eng.* **25,** 194–203 (2014).
8. Xu, P., Ranganathan, S., Fowler, Z. L., Maranas, C. D. & Koffas, M. A. Genome-scale metabolic network modeling results in minimal interventions that cooperatively force carbon flux towards malonyl-CoA. *Metab. Eng.* **13,** 578–587 (2011).
9. Lin, F. *et al.* Improving fatty acid availability for bio-hydrocarbon production in *Escherichia coli* by metabolic engineering. *PLoS ONE* **8,** e78595 (2013).
10. Cautha, S. C. *et al.* in *12th IFAC Symposium on Computer Applications in Biotechnology* 221–226 (Mumbai, India, 2013).
11. Khodayari, A., Chowdhury, A. & Maranas, C. D. Succinate overproduction: a case study of computational strain design using a comprehensive *Escherichia coli* kinetic model. *Front. Bioeng. Biotechnol.* **2,** 76 (2014).
12. Chowdhury, A., Khodayari, A. & Maranas, C. D. Improving prediction fidelity of cellular metabolism with kinetic descriptions. *Curr. Opin. Biotechnol.* **36,** 57–64 (2015).
13. Chakrabarti, A., Miskovic, L., Soh, K. C. & Hatzimanikatis, V. Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. *Biotechnol. J.* **8,** 1043–1057 (2013).
14. Smallbone, K., Simeonidis, E., Swainston, N. & Mendes, P. Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst. Biol.* **4,** 6 (2010).
15. Khodayari, A., Zomorrodi, A. R., Liao, J. C. & Maranas, C. D. A kinetic model of *Escherichia coli* core metabolism satisfying multiple sets of mutant flux data. *Metab. Eng.* **25,** 50–62 (2014).
16. Chassagnole, C., Noisommit-Rizzi, N., Schmid, J. W., Mauch, K. & Reuss, M. Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnol. Bioeng.* **79,** 53–73 (2002).
17. Miskovic, L. & Hatzimanikatis, V. Modeling of uncertainties in biochemical reactions. *Biotechnol. Bioeng.* **108,** 413–423 (2011).
18. Tan, Y. & Liao, J. C. Metabolic ensemble modeling for strain engineers. *Biotechnol. J.* **7,** 343–353 (2012).
19. Tran, L. M., Rizk, M. L. & Liao, J. C. Ensemble modeling of metabolic networks. *Biophys. J.* **95,** 5606–5617 (2008).
20. Ishii, N. *et al.* Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. *Science* **316,** 593–597 (2007).
21. Kabir, M. M., Ho, P. Y. & Shimizu, K. Effect of ldhA gene deletion on the metabolism of *Escherichia coli* based on gene expression, enzyme activities, intracellular metabolite concentrations, and metabolic flux distribution. *Biochem. Eng. J.* **26,** 1–11 (2005).
22. Zhao, J., Baba, T., Mori, H. & Shimizu, K. Global metabolic response of *Escherichia coli* to gnd or zwf gene-knockout, based on 13C-labeling experiments and the measurement of enzyme activities. *Appl. Microbiol. Biotechnol.* **64,** 91–98 (2004).
23. Zhao, J. & Shimizu, K. Metabolic flux analysis of *Escherichia coli* K12 grown on 13C-labeled acetate and glucose using GC-MS and powerful flux calculation method. *J. Biotechnol.* **101,** 101–117 (2003).
24. Schomburg, I. *et al.* BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Res.* **41,** D764–D772 (2013).
25. Keseler, I. M. *et al.* EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* **41,** D605–D612 (2013).
26. Gopalakrishnan, S. & Maranas, C. D. 13C metabolic flux analysis at a genome-scale. *Metab. Eng.* **32,** 12–22 (2015).
27. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).
28. Bennett, B. D. *et al.* Absolute metabolite concentrations and implied enzyme active site occupancy in *Escherichia coli*. *Nat. Chem. Biol.* **5,** 593–599 (2009).
29. Sprenger, G. A. Genetics of pentose-phosphate pathway enzymes of *Escherichia coli* K-12. *Arch. Microbiol.* **164,** 324–330 (1995).
30. Koopman, F. *et al. De novo* production of the flavonoid naringenin in engineered *Saccharomyces cerevisiae*. *Microb. Cell Fact.* **11,** 155 (2012).
31. Choudhary, M. K., Yoon, J. M., Gonzalez, R. & Shanks, J. V. Re-examination of metabolic fluxes in *Escherichia coli* during anaerobic fermentation of glucose using C-13 labeling experiments and 2-dimensional nuclear magnetic resonance (NMR) spectroscopy. *Biotechnol. Bioproc. E* **16,** 419–437 (2011).
32. Salmon, K. *et al.* Global gene expression profiling in *Escherichia coli* K12. The effects of oxygen availability and FNR. *J. Biol. Chem.* **278,** 29837–29855 (2003).
33. Kadir, T. A., Mannan, A. A., Kierzek, A. M., McFadden, J. & Shimizu, K. Modeling and simulation of the main metabolism in *Escherichia coli* and its several single-gene knockout mutants with experimental verification. *Microb. Cell Fact.* **9,** 88 (2010).
34. Salek, R. M. *et al.* The MetaboLights repository: curation challenges in metabolomics. *Database (Oxford)* **2013,** bat029 (2013).
35. Zhang, Z. *et al.* CeCaFDB: a curated database for the documentation, visualization and comparative analysis of central carbon metabolic flux distributions explored by 13C-fluxomics. *Nucleic Acids Res.* **43,** D549–D557 (2015).
36. Bordbar, A. *et al.* Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. *Cell Syst.* **1,** 283–292 (2015).
37. Cisek, K., Krochmal, M., Klein, J. & Mischak, H. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrol. Dial. Transplant*. doi:10.1093/ndt/gfv364 (2015).
38. Cece-Esencan, E. N. *et al.* Software-aided cytochrome P450 reaction phenotyping and kinetic analysis in early drug discovery. *Rapid Commun. Mass Spectrom.* **30,** 301–310 (2016).
39. Zomorrodi, A. R., Islam, M. M. & Maranas, C. D. d-OptCom: dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synth. Biol.* **3,** 247–257 (2014).
40. Wittig, U. *et al.* SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Res.* **40,** D790–D796 (2012).
41. Winkler, J. D., Halweg-Edwards, A. L. & Gill, R. T. The LASER database: formalizing design rules for metabolic engineering. *Metab. Eng. Commun.* **2,** 30–38 (2015).
42. Digel, M., Ehehalt, R., Stremmel, W. & Fullekrug, J. Acyl-CoA synthetases: fatty acid uptake and metabolic channeling. *Mol. Cell. Biochem.* **326,** 23–28 (2009).
43. Noor, E. *et al.* Pathway thermodynamics highlights kinetic obstacles in central metabolism. *PLoS Comput. Biol.* **10,** e1003483 (2014).
44. Bar-Even, A., Flamholz, A., Noor, E. & Milo, R. Thermodynamic constraints shape the structure of carbon fixation pathways. *Biochim. Biophys. Acta* **1817,** 1646–1659 (2012).
45. Shearer, G., Lee, J. C., Koo, J. A. & Kohl, D. H. Quantitative estimation of channeling from early glycolytic intermediates to CO in intact *Escherichia coli*. *FEBS J.* **272,** 3260–3269 (2005).
46. Park, J. O. *et al.* Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. *Nat. Chem. Biol.* **12,** 482–489 (2016).
47. Kotlarz, D., Garreau, H. & Buc, H. Regulation of the amount and of the activity of phosphofructokinases and pyruvate kinases in *Escherichia coli*. *Biochim. Biophys. Acta* **381,** 257–268 (1975).
48. Lewis, B. A. *et al.* PRIDB: a protein–RNA interface database. *Nucleic Acids Res.* **39**(suppl 1): D277–D282 (2011).
49. Basse, M. J., Betzi, S., Morelli, X. & Roche, P. 2P2Idb v2: update of a structural database dedicated to orthosteric modulation of protein-protein interactions. *Database (Oxford)* **2016,** pii: baw007 (2016).
50. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44,** D471–D480 (2015).
51. Gerosa, L. *et al.* Pseudo-transition analysis identifies the key regulators of dynamic metabolic adaptations from steady-state data. *Cell Syst.* **1,** 270–282 (2015).
52. Saa, P. & Nielsen, L. K. A general framework for thermodynamically consistent parameterization and efficient sampling of enzymatic reactions. *PLoS Comput. Biol.* **11,** e1004195 (2015).
53. Cornish-Bowden, A. in *Fundamentals of Enzyme Kinetics* 4th edn (Wiley, 2012).
54. Leighty, R. W. & Antoniewicz, M. R. Parallel labeling experiments with [U-13C]glucose validate *E. coli* metabolic network model for 13C metabolic flux analysis. *Metab. Eng.* **14,** 533–541 (2012).
55. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5,** 264–276 (2003).
56. Burgard, A. P., Nikolaev, E. V., Schilling, C. H. & Maranas, C. D. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14,** 301–312 (2004).
57. Kabir, M. M. & Shimizu, K. Metabolic regulation analysis of icd-gene knockout Escherichia coli based on 2D electrophoresis with MALDI-TOF mass spectrometry and enzyme activity measurements. *Appl. Microbiol. Biotechnol.* **65,** 84–96 (2004).
58. Peng, L. & Shimizu, K. Effect of fadR gene knockout on the metabolism of *Escherichia coli* based on analyses of protein expressions, enzyme activities and

intracellular metabolite concentrations. *Enzyme Microb. Technol.* **38,** 512–520 (2006).
59. Yang, C., Hua, Q., Baba, T., Mori, H. & Shimizu, K. Analysis of *Escherichia coli* anaplerotic metabolism and its regulation mechanisms from the metabolic responses to altered dilution rates and phosphoenolpyruvate carboxykinase knockout. *Biotechnol. Bioeng.* **84,** 129–144 (2003).
60. Hoque, M. A., Siddiquee, K. A. Z. & Shimizu, K. Metabolic control analysis of gene-knockout *Escherichia coli* based on the inverse flux analysis with experimental verification. *Biochem. Eng. J.* **19,** 53–59 (2004).
61. Haverkorn van Rijsewijk, B. R., Nanchen, A., Nallet, S., Kleijn, R. J. & Sauer, U. Large-scale 13C-flux analysis reveals distinct transcriptional control of respiratory and fermentative metabolism in *Escherichia coli*. *Mol. Syst. Biol.* **7,** 477 (2011).
62. Fendt, S. M. *et al.* Unraveling condition-dependent networks of transcription factors that control metabolic pathway activity in yeast. *Mol. Syst. Biol.* **6,** 432 (2010).
63. Kochanowski, K., Sauer, U. & Chubukov, V. Somewhat in control--the role of transcription in regulating microbial metabolic fluxes. *Curr. Opin. Biotechnol.* **24,** 987–993 (2013).
64. Abdi, H. Coefficient of variation. *Encyclopedia of Research Design* **1,** 169–171 (2010).

Acknowledgements

We gratefully acknowledge funding from the DOE (http://www.energy.gov/) grant no. DE-SC0012377. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Anthony P. Burgard, Margaret Simons, Ali R. Zomorrodi, Akhil Kumar, Chiam Yu Ng, Ratul Chowdhury, Satyakam Dash and Anupam Chowdhury for their valuable inputs during the preparation of the manuscript. We thank Chuck Gilbert from Penn State Institute for CyberScience (ICS) for his assistant in running the codes on the Penn State Lion-X clusters.

Author contributions

Experiments were conceived and designed by C.D.M. and A.K. A.K. performed the experiments. A.K. and C.D.M. analysed the data. A.K. and C.D.M. contributed reagents/materials/analysis tools. A.K. and C.D.M wrote the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Khodayari, A. & Maranas, C. D. A genome-scale *Escherichia coli* kinetic metabolic model k-ecoli457 satisfying flux data for multiple mutant strains. *Nat. Commun.* **7,** 13806 doi: 10.1038/ncomms13806 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.