

# 1 Literature mining supports a next-generation modeling 2 approach to predict cellular byproduct secretion

3 Zachary A. King<sup>1</sup>, Edward J. O'Brien<sup>1</sup>, Adam M. Feist<sup>1,2</sup>, Bernhard O. Palsson<sup>1,2,3,\*</sup>

4 1. Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La  
5 Jolla, CA 92093, USA

6 2. Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemi-  
7 torvet, Building 220, DK-2800 Kongens Lyngby, Denmark

8 3. Department of Pediatrics, University of California, San Diego, 9500 Gilman Drive, La Jolla,  
9 CA 92093, USA

10 \* Corresponding author. E-mail: palsson@ucsd.edu

11 **The metabolic byproducts secreted by growing cells can be easily measured and provide**  
12 **a window into the state of a cell; they have been essential to the development of microbiol-**  
13 **ogy<sup>1</sup>, cancer biology<sup>2</sup>, and biotechnology<sup>3</sup>. Progress in computational modeling of cells has**  
14 **made it possible to predict metabolic byproduct secretion with bottom-up reconstructions of**  
15 **metabolic networks. However, owing to a lack of data, it has not been possible to validate**  
16 **these predictions across a wide range of strains and conditions. Through literature mining,**  
17 **we were able to generate a database of *Escherichia coli* strains and their experimentally mea-**  
18 **sured byproduct secretions. We simulated these strains in six historical genome-scale models**

19 **of *E. coli*, and we report that the predictive power of the models has increased as they have**  
20 **expanded in size and scope. Next-generation models of metabolism and gene expression are**  
21 **even more capable than previous models, but parameterization poses new challenges.**

## 22 **1 Introduction**

23 All cells secrete metabolic byproducts in the course of growing and producing energy, and these  
24 byproducts play important roles in the study of biological systems. Byproducts are a readout of the  
25 cellular state; lactate excretion, for instance, is characteristic of tumor cell growth<sup>2,4</sup>. Byproducts  
26 can be engineered for bioproduction of commodity chemicals and biofuels<sup>5-7</sup>. And byproducts of  
27 yeast fermentation – including ethanol – are responsible for the most popular beverages in human  
28 history<sup>8</sup>. With the critical roles played by metabolic byproducts in disease and biotechnology, it  
29 is of great interest to be able to predict the byproducts that a cell will secrete under a specific  
30 condition. However, no published study has assessed whether existing computational methods are  
31 able to predict metabolic byproducts for a range of strains and conditions.

32 Computational models have been shown to correctly predict byproduct secretion under com-  
33 mon laboratory conditions. During aerobic growth, the model bacterium *Escherichia coli* oxidizes  
34 substrate molecules to secrete CO<sub>2</sub> and water; during anaerobic fermentation, *E. coli* secretes  
35 mixed-acid fermentation products (ethanol, acetate, formate, D-lactate, and succinate)<sup>9</sup>. Genome-  
36 scale models (GEMs) and constraint-based reconstruction and analysis (COBRA) methods rely on

37 knowledge of the metabolic network and mass-balance during steady state growth to predict the  
38 optimal distribution of metabolic flux for growth<sup>10</sup>. GEMs have been shown to be able to predict  
39 *E. coli* byproduct secretions in certain cases<sup>11,12</sup>. In the context of GEMs, the byproducts that must  
40 be secreted for optimal growth are called *growth-coupled*, and computational methods have been  
41 developed to predict and engineer growth-coupled chemical production<sup>13–15</sup>. However, few exper-  
42 imental studies have followed from the computational method development (among them:<sup>12,16</sup>), so  
43 it is unclear how these methods would scale up to a wide variety of strains and conditions.

44 Next-generation GEMs of metabolism and gene expression (called ME-models<sup>17–19</sup>) are now  
45 available; ME-models predict the composition of the entire proteome of a cell. In contrast, GEMs  
46 of metabolism (M-models) predict only the reaction fluxes in a metabolic network<sup>18</sup>. One new  
47 capability of ME-models is the ability to predict the bacterial Warburg effect, the tendency of bac-  
48 teria to secrete acetate during aerobic growth in the presence of excess substrate<sup>4,20</sup>. In ME-models,  
49 the limitations of ribosome efficiency lead to low-yield metabolic approaches like acetate secre-  
50 tion<sup>17</sup>. The same effect can be seen in smaller-scale growth models and is supported by phenotypic  
51 data<sup>4,20</sup>. Whether ME-models can correctly predict byproduct secretion for other conditions is not  
52 currently known.

53 High-quality genotypic and phenotypic data are required to test any model predictions, and  
54 such data have not been available for the study of byproduct secretion. The present study takes  
55 a novel approach by mining the research literature for examples of engineered strains of *E. coli*  
56 with diverse byproduct secretion mixtures. We collected 73 papers reporting a total of 89 strains of

57 *E. coli* that have a wide range of gene knockouts, heterologous pathways, and growth conditions,  
58 and we simulated these paired genotype-phenotype data in 6 historical GEMs of *E. coli*, includ-  
59 ing the next-generation ME-model. We find that GEMs have been improving in their ability to  
60 recapitulate measured byproducts from experimental studies as the models have increased in size  
61 and scope. We explore the possible reasons for incorrect predictions and provide insights into the  
62 challenges of simulating byproduct secretion for any growing cell.

## 63 **2 Results**

64 **Literature mining provides a diverse set of strains and phenotypes.** An impressive body of  
65 data on *E. coli* byproduct secretion can be found in the peer reviewed literature (Fig. 1). We gen-  
66 erated a bibliomic database using a workflow for identifying relevant papers, extracting data, and  
67 performing quality assessment (Fig. S1). Each paper in the database reported a strain design of  
68 *E. coli* in which the fermentation pathways were engineered to force the cell to secrete a target  
69 molecule (Fig. 2). The bibliomic database includes the gene knockouts, heterologous pathway  
70 descriptions, substrate conditions, oxygen availability, and the parent cell line for each strain (Sup-  
71plementary Data 1). It is difficult to extract and normalize quantitative measures of byproduct  
72 secretion from the literature. Instead, we recorded the molecule that was targeted for overproduc-  
73 tion in the study, and we confirmed that this byproduct was the major secretion product in each  
74 case (see Methods). The bibliomic database contains 73 papers and 89 strains of *E. coli*; this  
75 is approximately 20% of all papers on metabolic engineering of *E. coli* collected in the LASER  
76 database<sup>21</sup>.

77 The strains in the bibliomic database were simulated in six GEMs of *E. coli* (Table 1). The  
78 models have increased in size and complexity over the past decade; they include five M-models and  
79 one ME-model that includes 1,683 genes and accounts for 80% of the proteome by mass<sup>17,18</sup>. Gene  
80 knockouts, heterologous pathways, and environmental conditions from the bibliomic database  
81 were recreated in each of the GEMs. For each strain, flux balance analysis (FBA)<sup>22</sup> was used  
82 to find the predicted growth rate and the growth-coupled yield, the carbon yield of a compound  
83 at the maximum growth rate. The analysis began with two comparisons between the bibliomic  
84 database and the simulations: (1) whether the strain grew in a given environment and (2) whether  
85 the simulation predicted growth-coupled secretion of the target byproduct from the study.

86 The predictive power of GEMs has generally increased over time, with the increasing size  
87 and scope of the models. New GEMs provide better predictions of growth-coupled secretion com-  
88 pared to their predecessors (“Model accuracy” in Fig. 3). In order to understand the reasons for  
89 this trend, we designed a computational approach to categorize cases of incorrect prediction. Ex-  
90 haustive search and parameter sampling were employed in the M- and ME-models, respectively,  
91 to determine what changes to the modeling approach might lead to *in silico* secretion of the tar-  
92 get byproduct (see Methods). These categories provide insights into the general challenges of  
93 modeling byproduct secretion.

94 **Genome-scale models do not differentiate between isozymes.** Isozymes are common in metabolic  
95 networks, and they are represented in M-models, but their diverse regulatory and catalytic prop-  
96 erties lead to a broad and complex set of challenges for metabolic modeling. Reactions are often

97 catalyzed by a major isozyme that is responsible for most catalysis, while minor isozymes are also  
98 present in the cell but have a smaller role (they may not be expressed or have less-favorable ki-  
99 netics)<sup>23</sup>; recent progress in studying enzyme promiscuity and underground metabolism suggests  
100 that isozymes are even more widespread than previously thought<sup>24</sup>. Many experimental studies  
101 report gene knockouts of major isozymes that decrease the activity of the associated reaction sig-  
102 nificantly, enough so that the minor isozymes can be ignored (e.g. removing *ldhA* and ignoring  
103 *dld*<sup>25–27</sup>). However, M-models do not distinguish between major and minor isozymes, so these  
104 cases are incorrectly predicted in the model; the minor isozyme catalyzes the reaction *in silico*,  
105 and the *in silico* gene knockout of the major isozyme has no effect. Therefore, to simulate byprod-  
106 uct secretion for real-world experiments, it was necessary to employ a “greedy knockout” strategy  
107 in which all reactions associated with a gene knockout are disabled, even if minor isozymes might  
108 be present (Fig. 4a).

109 There are exceptions where greedy knockouts are not appropriate. For example, the alanine  
110 racemase activity of isozymes *alrR* and *dadX* is necessary for *in silico* growth, so applying the  
111 greedy knockout strategy to the reported strain that has a knockout of *alrR* leads to a prediction of  
112 cell death<sup>28</sup>. In other words, this strain can not be correctly simulated by M-models with or without  
113 the greedy knockout strategy. This issue can only be addressed through continued development of  
114 genome-scale modeling methods to address regulation, kinetics, allosteric inhibition, and the many  
115 biophysical properties that differentiate isozymes. Furthermore, ME-models can potentially select  
116 the appropriate enzyme based on protein cost, but ME-models do not include regulatory effects that  
117 often are responsible for the distinction between major and minor isozymes, so greedy knockouts

118 are still generally required. In this study, the greedy knockout approach was sufficient to correctly  
119 simulate most of the gene knockouts in the bibliomic database.

120 **Larger models solve false predictions of cell death.** Every strain in the bibliomic database was  
121 able to grow in the published experimental studies, but many simulations of these strains in early  
122 GEMs resulted in predictions of no growth (defined as *in silico* specific growth rate less than  
123  $0.005 \text{ hr}^{-1}$ ). These incorrect predictions have decreased as the GEMs have increased in size and  
124 scope (“Experimental KO(s) are lethal *in silico*” in Fig. 3). In most cases, the reason for the  
125 improved prediction is that the more comprehensive GEMs include a pathway that can rescue an  
126 essential cellular function when another important pathway is disabled by gene knockouts. In  
127 the five *E. coli* M-models, the lethal genotypes were analyzed by exhaustively searching for the  
128 minimal combinations of reactions that lead to *in silico* cell death (Fig. S2).

129 The biggest improvement in modeling the strains in the bibliomic database can be attributed  
130 to a single reaction. The models *iJR904* and *iAF1260* incorrectly predict that fumarate reduc-  
131 tase (FRD, *frd*) is essential under anaerobic conditions, and 63% of the designs in the bibliomic  
132 database include a knockout in the *frd* operon (see the large jump from *iAF1260* to *iAF1260b*  
133 in Fig. 3a). These incorrect predictions were corrected in *iAF1260b* and later GEMs with the  
134 inclusion of a new reaction (DHORDfum) that rescues growth when FRD is removed (Fig. 4b).  
135 However, there is no experimental evidence to support the presence of the DHORDfum reaction.  
136 So why does this reaction exist in the models, and why does it improve predictions?

137 One explanation is that the DHORDfum reaction does not take place in the cell, and, in-

138 instead, succinate dehydrogenase (SUCDi, *sdh*) acts in the reverse direction to rescue conversion of  
139 fumarate to succinate; this has actually been shown experimentally<sup>29</sup>. Thus, the evidence supports  
140 removing DHORDfum from the models and making SUCDi reversible. However, this change in-  
141 troduces the challenges associated with modeling isozymes for the activity catalyzed by *frd* and  
142 *sdh*, so the presence of DHORDfum has served as a convenient hack for modeling *E. coli*.

143 **Simulations suggest that some strains have room to evolve.** When the experimental observa-  
144 tions of byproduct secretion disagree with predictions, another possible explanation is that the  
145 experimental strain could evolve to grow faster by adopting the byproduct secretion strategy pre-  
146 dicted by the model (“Target byproduct is not growth-coupled” in Fig. 33). FBA simulations  
147 predict the metabolic state of a cell that is operating close to optimal growth; GEMs are power-  
148 ful for predicting cellular behavior precisely because fast growing cells often adopt a near-optimal  
149 strategy for growth<sup>30,31</sup>. Thus, some of the disagreement between observation and prediction might  
150 be caused, not by model errors, but rather by an assumption of the modeling approach (the opti-  
151 mality assumption). This hypothesis can be tested through laboratory evolution by passing the  
152 strain repeatedly<sup>12</sup>. (The process is also called serial passage, metabolic evolution, growth rescue,  
153 or adaptive laboratory evolution (ALE).) Laboratory evolution was used in 14 studies (19 strains)  
154 in the bibliomic database to improve byproduct secretion, and the predictive power of the model  
155 is greater for these cases than for the bibliomic database in general (Fig. S3). This supports the  
156 hypothesis that FBA predicts byproduct secretions that are not correct for the reported strains but  
157 would be correct if the strains were evolved through growth selection.

158 **Next-generation ME-models improve predictions but require parameterization.** ME-models  
159 expand upon M-models by explicitly accounting for all of the biochemical reactions in the gene  
160 expression machinery of the cell (including transcription and translation)<sup>17,18</sup>. To include protein  
161 production in the ME-model, one must estimate the turnover rate of each enzyme ( $k_{eff}$ ) that de-  
162 termines how many active proteins must be present to convert one set of reactants to products in  
163 a given time. ME-model simulations used a set of experimentally validated kinetic parameters  
164 from a recent study<sup>32</sup>. For high-flux reactions, the  $k_{eff}$ s were shown to be consistent across four  
165 growth conditions. However, it is still possible for  $k_{eff}$ s to change between conditions, depending  
166 on metabolite concentrations and other variables (they range between 0 and  $k_{cat}$ ). Therefore, we  
167 sampled  $k_{eff}$ s in the ME-model to generate an ensemble of models for each strain that was not  
168 growth-coupled with default parameters (see Methods). We found that 26 / 41 strains in this set  
169 could be growth-coupled in the ME-model with at least one model in the ensemble, including 9 / 11  
170 designs for succinate production (Fig. 4e). Addressing kinetic parameters will have to be a part  
171 of ME-model development going forward, and this should lead to better predictions of byproduct  
172 secretion.

173 The protein costs associated with metabolic pathways in the ME-model also solve another  
174 failure mode in M-models: alternative optimal solutions. Alternative optimal solutions occur in M-  
175 models when two metabolic states lead to the same growth rate, and this common failure mode has  
176 been solved with next-generation ME-models (“Alternative optimal growth-coupled solutions” in  
177 Fig. 3)<sup>33</sup>. In ME-model simulations, each pathway has specific enzyme costs that must be precisely  
178 allocated using cellular resources. Therefore, pathways with the same metabolic contribution to

179 cellular growth (e.g. same ATP production and redox balance) that are equivalent in the M-model  
180 have different proteomic costs in the ME-model. In all cases, this failure mode of M-models  
181 disappear in ME-model predictions (with one example provided in Fig. 4d).

182 In addition to removing alternative optimal solutions, the proteomic pathway costs in the  
183 ME-model can address challenges of encoding reversibility in the M-model. As an example, the  
184 production of isobutanol using a 2-keto acid based pathway was recently demonstrated<sup>34,35</sup>, and  
185 the optimal *in silico* phenotype of this production strain varies between models of *E. coli* (Fig. 3b).  
186 *iAF1260b* correctly predicts the production of isobutanol as the optimal fermentation product; in  
187 contrast, *iJO1366* predicts that hexanoic acid, a 6-carbon intermediate in the  $\beta$ -oxidation cycle, is  
188 the preferred product. This difference can be traced to the thermodynamic reversibility of the thi-  
189 olase reaction in the second round of the reversed  $\beta$ -oxidation cycle – it is irreversible in *iAF1260*  
190 (KAT2) and reversible in *iJO1366* (ACACT2r) (Fig. 4c). The reversibility in *iJO1366* is in line  
191 with experimental evidence<sup>36</sup>, but it also leads to the seemingly incorrect prediction of hexanoic  
192 acid secretion. The ME-model suggests that the incorrect prediction of hexanoic acid secretion  
193 by *iJO1366* is not so much a matter of thermodynamics as a matter of pathway length and thus  
194 proteomic cost. When the cost of producing enzymes for metabolic pathways is incorporated into  
195 genome-scale models, long pathways like the hexanoic acid production route through  $\beta$ -oxidation  
196 carry a greater cost than the shorter 2-keto acid route to isobutanol. This case shows the power  
197 of a constraint-based modeling approach: Properly encoding reversibility in M-models has been a  
198 long-standing challenge, so the ME-model applies a completely different constraint (pathway cost)  
199 that makes the reversibility of  $\beta$ -oxidation unimportant for correct predictions.

### 200 3 Discussion

201 As cellular models become larger and more complicated, the datasets used to validate them must  
202 also grow. This study presents a novel approach to model validation based on literature mining.  
203 In spite of the uneven quality of literature data, this approach was capable of generating important  
204 insights into the abilities of GEMs to predict byproduct secretion. Higher-quality data would  
205 enable an even more thorough model validation, and there is a great need in systems biology  
206 for standardizing genotype-phenotype datasets. Standards for storing phenotypic data have been  
207 discussed<sup>37,38</sup>, and it is essential that progress be made.

208 There are a few challenges that will have to be addressed to scale these methods to larger and  
209 more complicated systems. First, many data points in the bibliomic database cannot be modeled in  
210 existing GEMs. For instance, regulatory knockouts are not in the scope of M- and ME-models, so  
211 they were ignored in this study. The correct predictions of strains in the bibliomic database draw  
212 largely from the concept of redox balance in the cell (NAD(P)H produced during glycolysis must  
213 be consumed by fermentation pathways), and extending prediction of byproduct secretion to other  
214 applications where redox balance is not the driving phenomenon may require further development  
215 of the modeling methods. However, constraint-based modeling methods are generally extensi-  
216 ble, as we have seen with the development and implementation of ME-models. Exploration of  
217 constraint based approaches to other subsystems – including protein structures, membrane translo-  
218 cation, and regulation – are under way<sup>39</sup>.

219 Second, strains modeled using GEMs and FBA must be operating close to an optimal growth

220 state. Understanding the byproduct secretion of strains that are not growing rapidly will require  
221 research into other objective functions that could make the models predictive for strains that are  
222 not optimizing for growth<sup>40,41</sup>. On the other hand, the optimality assumption of FBA offers an  
223 advantage: GEMs and laboratory evolution can be used together for systematic optimization of  
224 microorganisms<sup>12,16</sup>.

225 Finally, the extension of these methods to larger and more complex organisms, such as tumor  
226 cells, will require rigorous development and assessment of GEMs. This study provides an example  
227 of validating model predictions using genotype-phenotype data mined from the literature. The  
228 collection of these data will need to be scaled up to validate larger and more complex models. All  
229 cells have the same basic features that include gene expression, metabolism, and, by necessity,  
230 byproduct secretion; with targeted validation studies, we can feel increasingly confident in our  
231 ability to model and understand them.

## 232 **4 Methods**

233 **Literature mining.** A literature mining search was performed to identify all papers reporting the  
234 construction of a cell factory strain of *E. coli* for the production of a fermentation product. A  
235 workflow was developed (Fig. S1), hundreds of papers were collected, and 73 were included in the  
236 bibliomic database based on their matching the following criteria:

- 237 • Utilized a strain of *E. coli*.

238 • Modified the strain for production of a native or heterologous metabolite.

239 • Removed alternative fermentation pathways using gene knockouts.

240 Metadata were collected from each paper, including the target production molecule, whether  
241 simulations were performed to identify knockouts, the parent *E. coli* strain, the genetic additions  
242 and deletions, the aerobicity and carbon sources during fermentation experiments, whether lab-  
243 oratory evolution was performed, and (when possible) the measured fermentation profile of the  
244 engineered strain.

245 A single target molecule was selected for each experiment, even though in some cases a  
246 mixture of products was reported. When papers reported mixtures of hydrogen or formate with a  
247 coproduct, the coproduct was considered the target molecule.

248 **Simulations.** To simulate reported designs, the gene knockouts were implemented *in silico* using  
249 a “greedy knockout” strategy. For each gene that was knocked out experimentally, all reactions  
250 associated with that gene in the metabolic model are turned off. The alternative strategy is to  
251 evaluate the gene-protein-reaction (GPR) rules for each reaction in turn, to determine whether the  
252 reaction is turned off or remains unchanged; however, as discussed in the text, only the “greedy  
253 knockout” approach was able to correctly simulate strains in the bibliomic database.

254 For all non-native genes reported in the papers, pathways were reconstructed by creating  
255 *in silico* reactions corresponding to the genes used in these experiments. For transport reactions,  
256 transport was assumed to be non-energy-coupled unless otherwise specified in the *iJO1366* recon-

257 struction or in the literature.

258 Polymer production must be considered separately from ordinary metabolite secretions. To  
259 simulate these strains, the production of the monomer was optimized. It is unclear whether poly-  
260 mers such as polylactic acid (PLA) would be growth coupled. The PHA synthase is not energy  
261 coupled<sup>42</sup>, so an equilibrium between monomer and polymer would probably be achieved in the  
262 optimal state (this has been shown for soluble heteroglycans<sup>43</sup>). However, by upregulating the  
263 PHA synthase in a strain optimized for monomer production, one can use the growth-coupling  
264 effect to perform much of the strain optimization. Thus, growth-coupling of the monomer is of  
265 interest.

266 Five M-models and one ME-model of *E. coli* K-12 MG1655 were used for the simulations in  
267 this work. The M-models were collected from the BiGG Models database<sup>44</sup>, and they were used as  
268 reported in their respective publications (Table 1). As described previously, the *iJO1366* oxidative  
269 stress reactions CAT, SPODM, and SPODMpp and the FHL reaction were constrained to zero<sup>45</sup>.  
270 A new software implementation of the ME model *iOL1650-ME* was used. Pathway diagrams were  
271 generated using Escher<sup>46</sup>, and COBRA simulations were performed with COBRApy<sup>47</sup>.

272 For M-model simulations, the substrate uptake rates (SURs) for the solitary carbon substrates  
273 in each simulation were constrained to a maximum uptake rate of 10 mmol gDW<sup>-1</sup> hr<sup>-1</sup>. The  
274 oxygen uptake rates were constrained to 0 for anaerobic conditions and 20 mmol gDW<sup>-1</sup> hr<sup>-1</sup> for  
275 aerobic conditions. For ME-model simulations, SURs were left unbounded and the ME-model  
276 optimization procedure chose optimal SURs. If LB or yeast extract was present in the medium,

277 the simulations were still performed with an *in silico* minimal media based on the assumption  
278 that cells will preferentially consume glucose before more-complex carbon sources; however, if  
279 this approximation led to a lethal phenotype in *iJO1366*, then supplementations known to exist  
280 in rich media were added to alleviate the lethal phenotype. Microaerobic designs were assumed  
281 to be anaerobic because it has been observed that even under aerobic conditions the anaerobic  
282 physiology contributes to fermentation<sup>48</sup>.

283 FBA was used to find the maximum and minimum secretion of each metabolite in the net-  
284 work when the growth rate is near its maximum (within 0.01%)<sup>22</sup>. The key outputs of these sim-  
285 ulations are *predicted growth rate* – the flux through the biomass objective function – and the  
286 *growth-coupled yield* – the minimum carbon flux through the target molecule exchange reaction at  
287 the maximum growth rate

288 **Parameter sampling.** Parameter sampling in the ME-model was employed to determine the sen-  
289 sitivity of ME-model simulations to  $k_{eff}$  values. For each sampling simulation, an ensemble of  
290 200 models was generated with  $k_{eff}$  values selected randomly from a lognormal distribution of  
291 possible  $k_{cat}$ s. The distribution was determined from a collection of all  $k_{cat}$ s in the BRENDA  
292 enzyme database ( $\mu = 2.48$  and  $\sigma = 3.29$ )<sup>49</sup>.

293 **Failure model categorization.** Growth-coupling was defined as secretion of the target molecule  
294 with greater than 15% carbon yield or, for hydrogen production, greater than 2 mmol gDW<sup>-1</sup> hr<sup>-1</sup>.  
295 Lethal phenotypes were defined as having an *in silico* growth rate below 0.005 hr<sup>-1</sup>. Alternative  
296 optima were identified by finding designs whose maximum secretions were above the threshold

297 for growth coupling but whose minimum secretions were below this threshold.

## 298 **5 Acknowledgments**

299 We would like to thank Gabriela I. Guzmán and Joshua A. Lerman for their guidance and sug-  
300 gestions during the course of this project. Funding for this work was provided by the National  
301 Science Foundation Graduate Research Fellowship [DGE-1144086 to Z.A.K.] and by the Novo  
302 Nordisk Foundation through the Center for Biosustainability at the Technical University of Den-  
303 mark [NNF16CC0021858].

## 304 **6 Author contributions**

305 Z.A.K. ran the analysis. Z.A.K., E.J.O., A.F.M., and B.O.P. designed the study and wrote the  
306 manuscript.

## 307 **7 Competing financial interests**

308 The authors declare no competing financial interests.

## 309 **8 References**

310 1. Barnett, J. A. Beginnings of microbiology and biochemistry: the contribution of yeast re-  
312 search. *Microbiology* **149**, 557–567 (2003).

- 313 2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674  
314 (2011).
- 315 3. Demain, A. L. History of industrial biotechnology. In Soetaert, W. & Vandamme, E. J. (eds.)  
316 *Industrial Biotechnology: Sustainable Growth and Economic Success*, chap. 1, 17–77 (Wiley-  
317 VCH, Weinheim, Germany, 2010).
- 318 4. Basan, M. *et al.* Overflow metabolism in bacteria results from efficient proteome allocation  
319 for energy biogenesis. *Nature* (2015).
- 320 5. Lee, S. Y. & Kim, H. U. Systems strategies for developing industrial microbial strains. *Nat.*  
321 *Biotechnol.* **33**, 1061–1072 (2015).
- 322 6. Zhang, F., Rodriguez, S. & Keasling, J. D. Metabolic engineering of microbial pathways for  
323 advanced biofuels production. *Curr. Opin. Biotechnol.* **22**, 775–783 (2011).
- 324 7. Chubukov, V., Mukhopadhyay, A., Petzold, C. J., Keasling, J. D. & Martín, H. G. Synthetic  
325 and systems biology for microbial production of commodity chemicals. *npj Systems Biology*  
326 *and Applications* **2**, 16009 (2016).
- 327 8. Piškur, J., Rozpedowska, E., Polakova, S., Merico, A. & Compagno, C. How did saccha-  
328 romyces evolve to become a good brewer? *Trends Genet.* **22**, 183–186 (2006).
- 329 9. David P. Clark. The fermentation pathways of *Escherichia coli*. *FEMS Microbiol. Rev.* **63**,  
330 223–234 (1989).

- 331 10. Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. Constraint-based models predict  
332 metabolic and associated cellular functions. *Nat. Rev. Genet.* **15**, 107–120 (2014).
- 333 11. Varma, A., Boesch, B. W. & Palsson, B. O. Stoichiometric interpretation of *Escherichia coli*  
334 glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* **59**, 2465–2473  
335 (1993).
- 336 12. Fong, S. S. *et al.* In silico design and adaptive evolution of *Escherichia coli* for production of  
337 lactic acid. *Biotechnol. Bioeng.* **91**, 643–648 (2005).
- 338 13. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework  
339 for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.*  
340 **84**, 647–657 (2003).
- 341 14. Feist, A. M. *et al.* Model-driven evaluation of the production potential for growth-coupled  
342 products of *Escherichia coli*. *Metab. Eng.* **12**, 173–186 (2010).
- 343 15. Lewis, N. E., Nagarajan, H. & Palsson, B. Ø. Constraining the metabolic genotype-phenotype  
344 relationship using a phylogeny of *in silico* methods. *Nat. Rev. Microbiol.* **10**, 291–305 (2012).
- 345 16. Yim, H. *et al.* Metabolic engineering of *Escherichia coli* for direct production of 1,4-  
346 butanediol. *Nat. Chem. Biol.* **7**, 445–452 (2011).
- 347 17. O’Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R. & Palsson, B. Ø. Genome-scale  
348 models of metabolism and gene expression extend and refine growth phenotype prediction.  
349 *Mol. Syst. Biol.* **9**, 693 (2013).

- 350 18. O'Brien, E. J. & Palsson, B. O. Computing the functional proteome: recent progress and  
351 future prospects for genome-scale models. *Curr. Opin. Biotechnol.* **34**, 125–134 (2015).
- 352 19. Lerman, J. a. *et al.* In silico method for modelling metabolism and gene product expression at  
353 genome scale. *Nat. Commun.* **3**, 929 (2012).
- 354 20. Molenaar, D., van Berlo, R., de Ridder, D. & Teusink, B. Shifts in growth strategies reflect  
355 tradeoffs in cellular economics. *Mol. Syst. Biol.* **5**, 323 (2009).
- 356 21. Winkler, J. D., Halweg-Edwards, A. L. & Gill, R. T. The LASER database: Formalizing  
357 design rules for metabolic engineering. *Metabolic Engineering Communications* **2**, 30–38  
358 (2015).
- 359 22. Orth, J. D., Thiele, I. & Palsson, B. Ø. What is flux balance analysis? *Nat. Biotechnol.* **28**,  
360 245–248 (2010).
- 361 23. Nakahigashi, K. *et al.* Systematic phenome analysis of *Escherichia coli* multiple-knockout  
362 mutants reveals hidden reactions in central carbon metabolism. *Mol. Syst. Biol.* **5**, 306 (2009).
- 363 24. Guzmán, G. I. *et al.* Model-driven discovery of underground metabolic functions in *Es-*  
364 *cherichia coli*. *Proceedings of the National Academy of Sciences* **112**, 929–934 (2015).
- 365 25. Trinh, C. T., Li, J., Blanch, H. W. & Clark, D. S. Redesigning *Escherichia coli* metabolism  
366 for anaerobic production of isobutanol. *Appl. Environ. Microbiol.* **77**, 4894–4904 (2011).
- 367 26. Stols, L. & Donnelly, M. I. Production of succinic acid through overexpression of NAD<sup>+</sup>-

- 368 dependent malic enzyme in an *Escherichia coli* mutant. *Appl. Environ. Microbiol.* **63**, 2695–  
369 2701 (1997).
- 370 27. Zhou, S., Shanmugam, K. T. & Ingram, L. O. Functional replacement of the *Escherichia coli*  
371 d(-)-lactate dehydrogenase gene (*ldha*) with the l-(+)-lactate dehydrogenase gene (*ldhl*) from  
372 *pediococcus acidilactici*. *Appl. Environ. Microbiol.* **69**, 2237 (2003).
- 373 28. Zhang, X., Jantama, K., Moore, J. C., Shanmugam, K. T. & Ingram, L. O. Production of  
374 l-alanine by metabolically engineered *escherichia coli*. *Appl. Microbiol. Biotechnol.* **77**, 355–  
375 366 (2007).
- 376 29. Maklashina, E., Berthold, D. a. & Cecchini, G. Anaerobic expression of *Escherichia coli* suc-  
377 cinate dehydrogenase: functional replacement of fumarate reductase in the respiratory chain  
378 during anaerobic growth. *J. Bacteriol.* **180**, 5989–5996 (1998).
- 379 30. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive  
380 evolution to achieve in silico predicted optimal growth. *Nature* **420**, 20–23 (2002).
- 381 31. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. In silico predictions of *Escherichia coli*  
382 metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* **19**, 125–130  
383 (2001).
- 384 32. Ali Ebrahim, Elizabeth Brunk, Justin Tan, Edward J. O'Brien, Donghyuk Kim, Richard Szu-  
385 bin, Joshua A. Lerman, Anna Lechner, Anand Sastry, Aarash Bordbar, Adam M. Feist, Bern-  
386 hard O. Palsson. Multi-omic data integration enables discovery of hidden biological regulari-  
387 ties In revision.

- 388 33. Lewis, N. E. *et al.* Omic data from evolved *E. coli* are consistent with computed optimal  
389 growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).
- 390 34. Atsumi, S. *et al.* Engineering the isobutanol biosynthetic pathway in escherichia coli by com-  
391 parison of three aldehyde reductase/alcohol dehydrogenase genes. *Appl. Microbiol. Biotech-*  
392 *nol.* **85**, 651–657 (2010).
- 393 35. Atsumi, S., Hanai, T. & Liao, J. C. Non-fermentative pathways for synthesis of branched-chain  
394 higher alcohols as biofuels. *Nature* **451**, 86–89 (2008).
- 395 36. Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the  
396  $\beta$ -oxidation cycle for the synthesis of fuels and chemicals. *Nature* **476**, 355–359 (2011).
- 397 37. McMurry, J. *et al.* Navigating the phenotype frontier: The monarch initiative (2016).
- 398 38. Check Hayden, E. Synthetic biologists seek standards for nascent field. *Nature News* **520**, 141  
399 (2015).
- 400 39. King, Z. A., Lloyd, C. J., Feist, A. M. & Palsson, B. O. Next-generation genome-scale models  
401 for metabolic engineering. *Curr. Opin. Biotechnol.* **35**, 23–29 (2015).
- 402 40. Zhao, Q., Stettner, A. I., Reznik, E., Paschalidis, I. C. & Segrè, D. Mapping the landscape of  
403 metabolic goals of a cell. *Genome Biol.* **17**, 109 (2016).
- 404 41. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predict-  
405 ing intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119 (2007).
- 406 42. Lee, S. Y. Bacterial polyhydroxyalkanoates. *Biotechnol. Bioeng.* **49**, 1–14 (1996).

- 407 43. Kartal, O., Mahlow, S., Skupin, A. & Ebenhöf, O. Carbohydrate-active enzymes exemplify  
408 entropic principles in metabolism. *Mol. Syst. Biol.* **7**, 542 (2011).
- 409 44. King, Z. A. *et al.* BiGG models: A platform for integrating, standardizing and sharing genome-  
410 scale models. *Nucleic Acids Res.* **44**, D515–22 (2016).
- 411 45. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of *Escherichia coli*  
412 metabolism—2011. *Mol. Syst. Biol.* **7**, 535 (2011).
- 413 46. King, Z. A. *et al.* Escher: A web application for building, sharing, and embedding data-rich  
414 visualizations of biological pathways. *PLoS Comput. Biol.* **11**, e1004321 (2015).
- 415 47. Ebrahim, A., Lerman, J. A., Palsson, B. O. & Hyduke, D. R. COBRApy: COstraints-Based  
416 reconstruction and analysis for python. *BMC Syst. Biol.* **7**, 74 (2013).
- 417 48. Ingram, L. O., Conway, T., Clark, D. P., Sewell, G. W. & Preston, J. F. Genetic engineering of  
418 ethanol production in *Escherichia coli*. *Appl. Environ. Microbiol.* **53**, 2420–2425 (1987).
- 419 49. Bar-Even, A., Noor, E. & Savir, Y. The moderately efficient enzyme: Evolutionary and physic-  
420 ochemical trends shaping enzyme parameters. *Biochemistry* 4402–4410 (2011).
- 421 50. Palsson, B. Ø. *Systems Biology: Properties of Reconstructed Networks* (Cambridge University  
422 Press, Cambridge, UK, 2006).
- 423 51. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. Ø. An expanded genome-scale model of  
424 *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**, R54 (2003).

- 425 52. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12  
426 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**,  
427 121 (2007).
- 428 53. Blankschien, M. D., Clomburg, J. M. & Gonzalez, R. Metabolic engineering of *Escherichia*  
429 *coli* for the production of succinate from glycerol. *Metab. Eng.* **12**, 409–419 (2010).
- 430 54. Donnelly, M. I., Millard, C. S., Clark, D. P., Chen, M. J. & Rathke, J. W. A novel fermentation  
431 pathway in an *Escherichia coli* mutant producing succinic acid, acetic acid, and ethanol. *Appl.*  
432 *Biochem. Biotechnol.* **70**, 187–198 (1998).
- 433 55. Lee, S. J. *et al.* Metabolic engineering of *Escherichia coli* for enhanced production of succinic  
434 acid , based on genome comparison and in silico gene knockout simulation. *Appl. Environ.*  
435 *Microbiol.* **71**, 7880 (2005).
- 436 56. Ma, J. *et al.* Enhancement of succinate production by metabolically engineered *Escherichia*  
437 *coli* with co-expression of nicotinic acid phosphoribosyltransferase and pyruvate carboxylase.  
438 *Appl. Microbiol. Biotechnol.* (2013).
- 439 57. Sánchez, A. M., Bennett, G. N. & San, K.-Y. Novel pathway engineering design of the anaero-  
440 bic central metabolic pathway in *Escherichia coli* to increase succinate yield and productivity.  
441 *Metab. Eng.* **7**, 229–239 (2005).
- 442 58. Sánchez, A. M., Bennett, G. N. & San, K.-Y. Efficient succinic acid production from glucose  
443 through overexpression of pyruvate carboxylase in an *Escherichia coli* alcohol dehydrogenase  
444 and lactate dehydrogenase mutant. *Biotechnol. Prog.* **21**, 358–365 (2005).

- 445 59. Singh, A., Cher Soh, K., Hatzimanikatis, V. & Gill, R. T. Manipulating redox and ATP bal-  
446 ancing for improved production of succinate in *E. coli*. *Metab. Eng.* **13**, 76–81 (2011).
- 447 60. Stols, L., Kulkarni, G., Harris, B. G. & Donnelly, M. I. Expression of ascaris suum malic  
448 enzyme in a mutant escherichia coli allows production of succinic acid from glucose. *Appl.*  
449 *Biochem. Biotechnol.* **63-65**, 153–158 (1997).
- 450 61. Vemuri, G. N., Eiteman, M. A. & Altman, E. Effects of growth mode and pyruvate carboxylase  
451 on succinic acid production by metabolically engineered strains of *Escherichia coli*. *Appl.*  
452 *Environ. Microbiol.* **68**, 1715–1727 (2002).
- 453 62. Zhang, X., Shanmugam, K. T. & Ingram, L. O. Fermentation of glycerol to succinate by  
454 metabolically engineered strains of *Escherichia coli*. *Appl. Environ. Microbiol.* **76**, 2397–2401  
455 (2010).

456 **9 Tables**

Table 1: The increasing size and scope of genome-scale models of *E. coli*.

Model	Genes	Reactions	Metabolites/Components	Year (Reference)
Core model	137	95	72	2006 <sup>50</sup>
<i>iJR904</i>	904	1075	761	2003 <sup>51</sup>
<i>iAF1260</i>	1,260	2,382	1,668	2007 <sup>52</sup>
<i>iAF1260b</i>	1,260	2,388	1,668	2010 <sup>14</sup>
<i>iJO1366</i>	1,366	2,583	1,805	2011 <sup>45</sup>
<i>iOL1650-ME</i>	1,683	12,009	6,563	2013 <sup>17</sup>



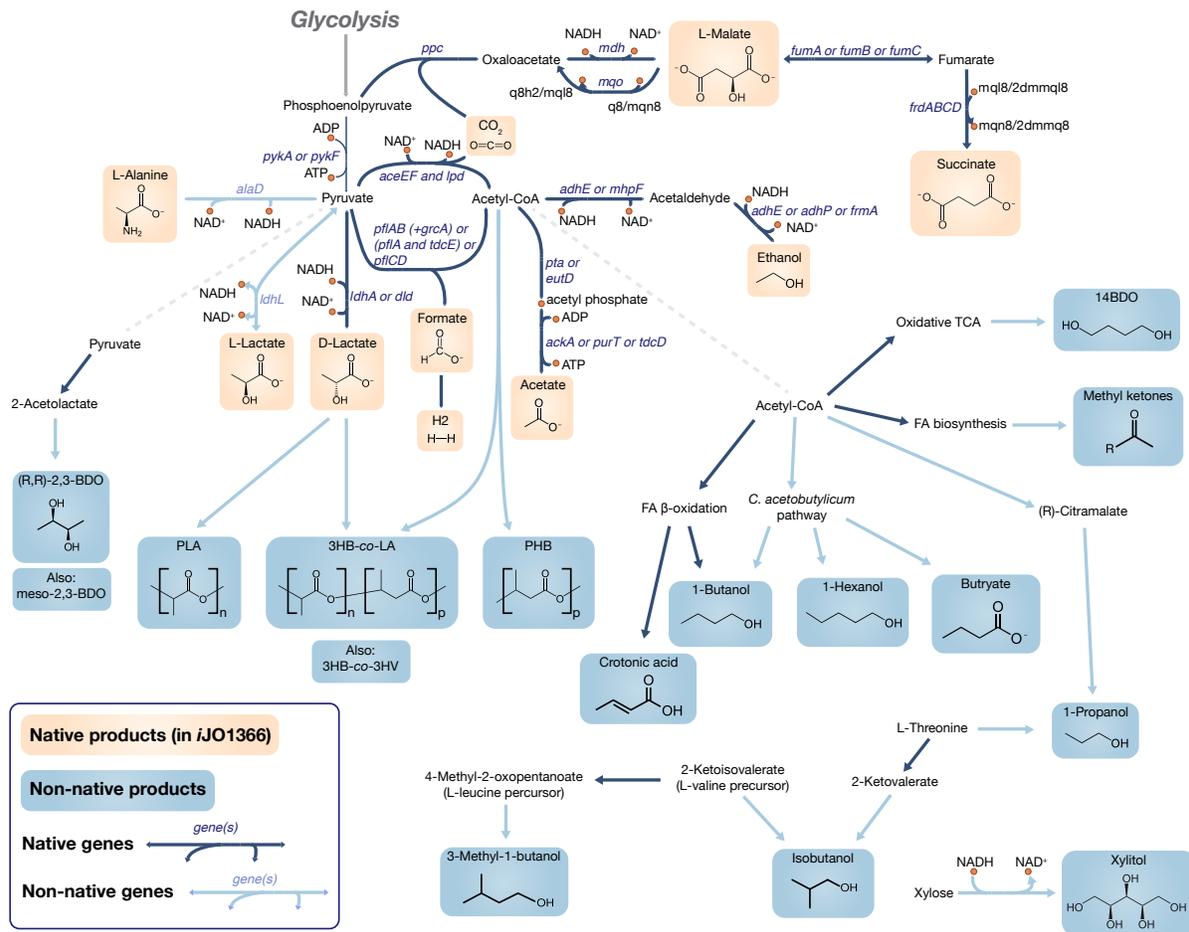
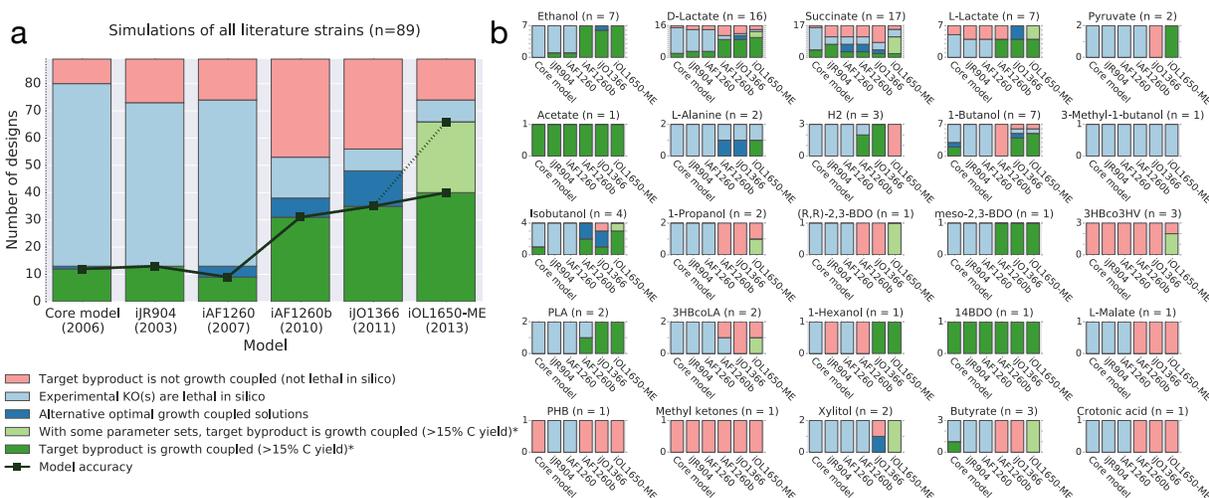
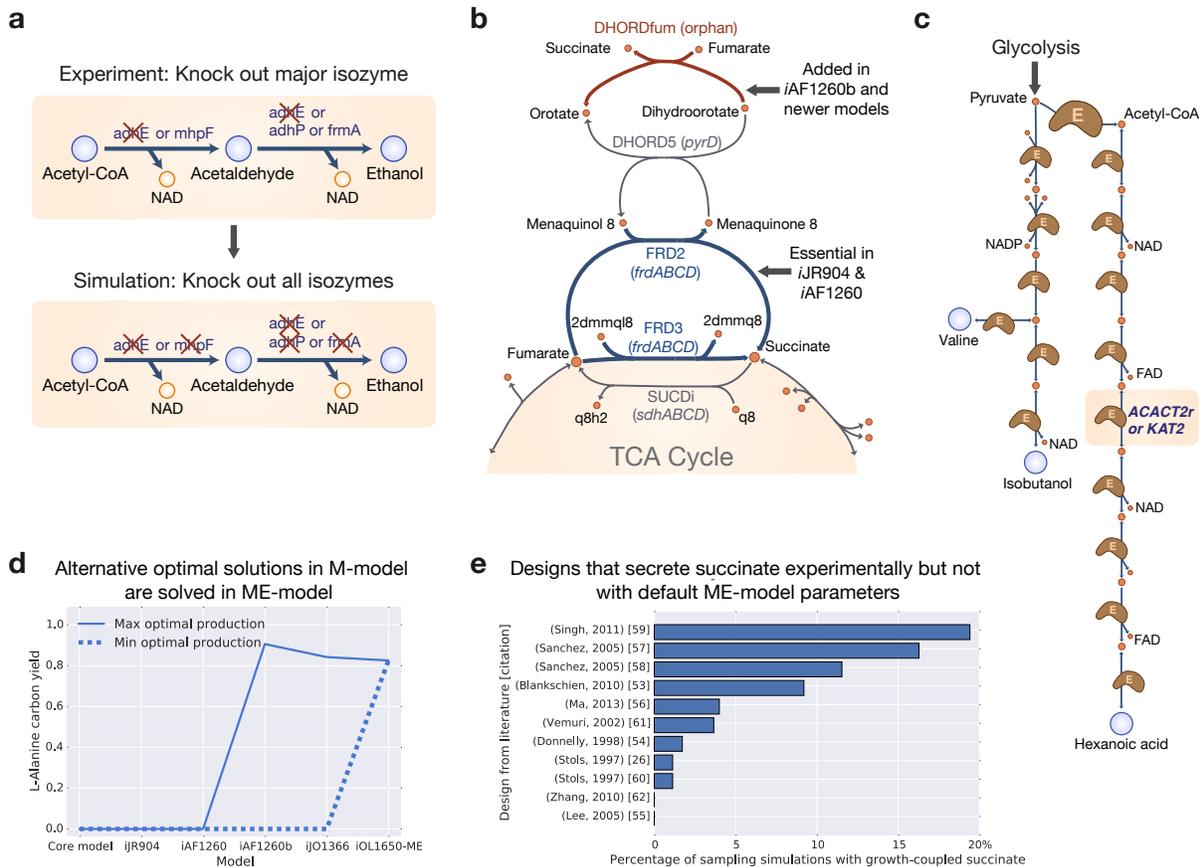


Figure 2: **The engineered fermentation pathways in *E. coli*.** All the engineering pathways in the bibliomic database are shown, along with their metabolic precursors. Native products (yellow) are those that appear in the genome-scale model *iJO1366*. Native pathways in *iJO1366* (dark blue arrows) and non-native pathways (light blue arrows) are also differentiated.



**Figure 3: Simulations of the bibliomic dataset in *E. coli* GEMs.** (a) The 89 strains in the bibliomic database were simulated in six GEMs of *E. coli*, and the incorrect predictions were categorized to suggest a reasons for the errors. The solid line signifies that the experimentally observed target byproduct is growth-coupled in the model. The dashed line represents the possibility of improving predictions in the ME-model by correctly determining the kinetic parameters ( $k_{effs}$ ). (b) The categories separated according to the target molecule.



**Figure 4: Comparing simulations with experiments.** All modeling approaches have failure modes, and comparing model predictions to experimental results allows these failure modes to be analyzed. (a) A “greedy knockout” strategy is necessary to contend with major and minor isozymes that are difficult to simulate in GEMs. (b) The genes in the *frd* operon are responsible for most of the incorrect predictions of cell death in *iJR904* and *iAF1260*. This error was fixed in *iAF1260b* and later models with the addition of the reaction DHORDfum. (c) For an isobutanol design, the ME model correctly predicts isobutanol secretion in preference to hexanoic acid secretion because the hexanoic acid pathway has greater protein cost<sup>34,35</sup>. (d) Alternative optimal phenotypes appear in M-models when two pathways have equivalent stoichiometries, as in this example for L-alanine secretion. ME-models explicitly account for the cost of producing pathway enzymes, so the shorter L-alanine production pathway is optimal in ME-models. (e) Succinate secretion is difficult to predict using existing GEMs, but an ensemble of ME-models with sampled kinetic parameters demonstrates that for certain parameter sets succinate secretion is correctly predicted.

458 **11 Extended Data**

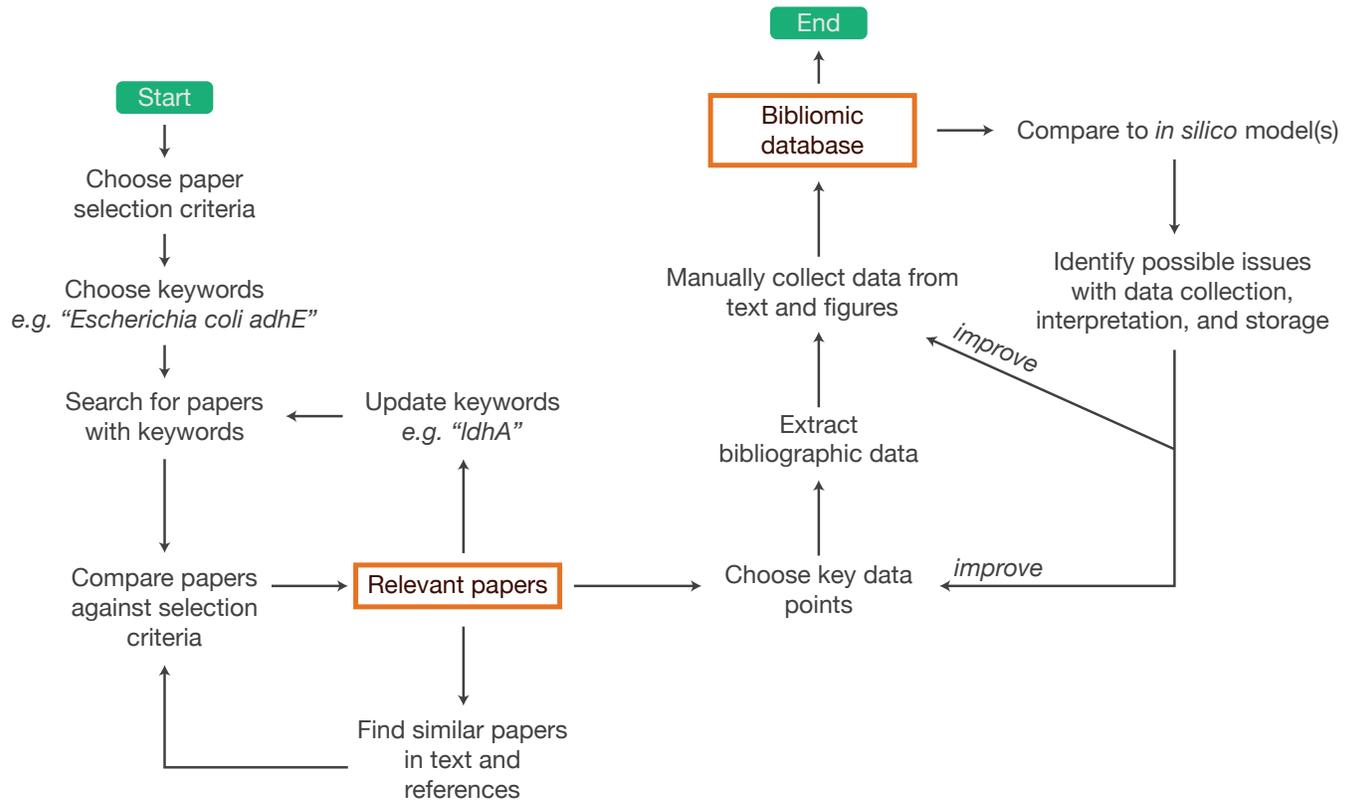


Figure S1: A workflow for generating a bibliomic dataset through literature mining.

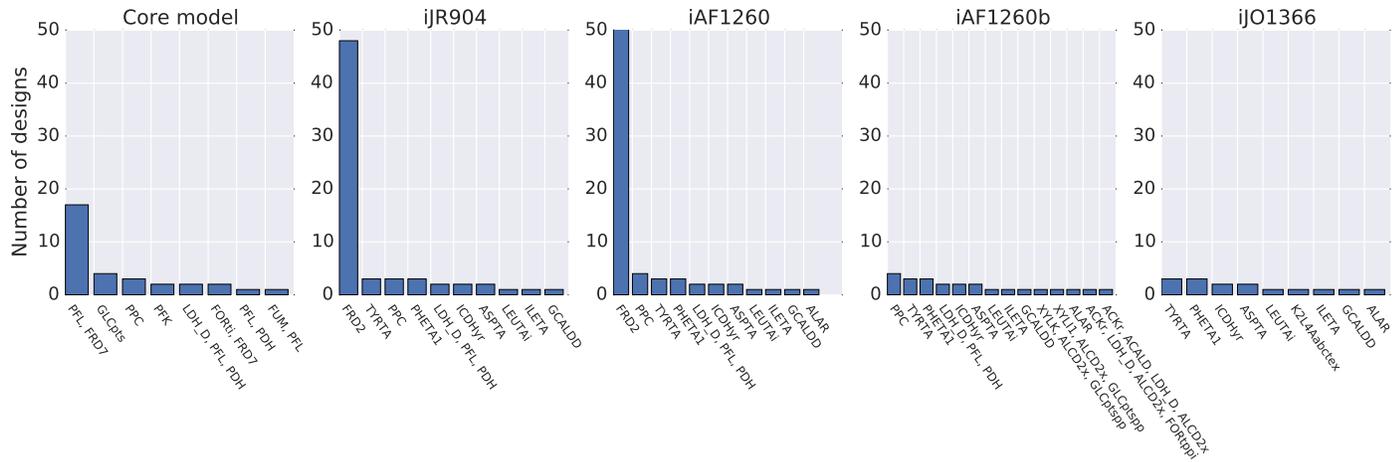


Figure S2: The combinations of reaction knockouts that are lethal in the M-models after simulating the bibliomic database.

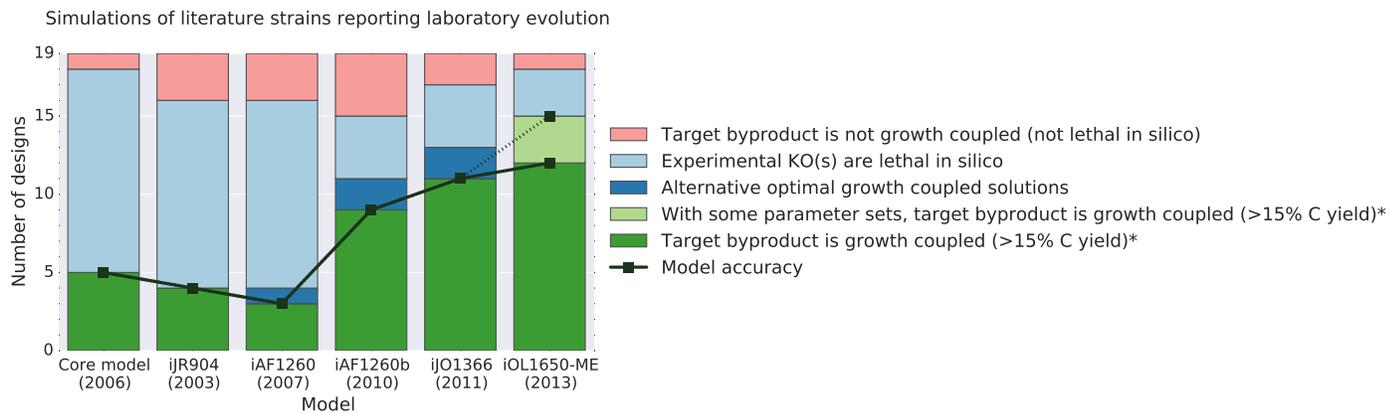


Figure S3: Comparing simulations with experiments for studies reporting laboratory evolution.

459 **12 Supplementary Data**

460 Supplementary Data 1: The complete bibliomic database (tab-separated text file).