

1 Harnessing natural modularity of cellular metabolism to
2 design a modular chassis cell for a diverse class of
3 products by using goal attainment optimization

4 Sergio Garcia^{1,2} and Cong T. Trinh^{1,2,*}

5 ¹*Department of Chemical and Biomolecular Engineering, The University of*
6 *Tennessee, Knoxville, TN, United States*

7 ²*Center for Bioenergy Innovation, Oak Ridge National Laboratory Oak*
8 *Ridge, TN, United States*

9 ^{*}*Corresponding author: 1512 Middle Drive, DO432, Department of*
10 *Chemical and Biomolecular Engineering, University of Tennessee,*
11 *Knoxville, TN 37996, United States. Tel: 865-974-2181. Email:*
12 *ctrinh@utk.edu.*

13 **Abstract**

14 Living cells optimize their fitness against constantly changing environments to sur-
15 vive. Goal attainment optimization is a mathematical framework to describe the si-
16 multaneous optimization of multiple conflicting objectives that must all reach a perfor-
17 mance above a threshold or goal. In this study, we applied goal attainment optimization
18 to harness natural modularity of cellular metabolism to design a modular chassis cell for
19 optimal production of a diverse class of products, where each goal corresponds to the
20 minimum biosynthesis requirements (e.g., yields and rates) of a target product. This
21 modular cell design approach enables rapid generation of optimal production strains
22 that can be assembled from a modular cell and various exchangeable production mod-
23 ules and hence accelerates the prohibitively slow and costly strain design process. We
24 formulated the modular cell design problem as a blended or goal attainment mixed
25 integer linear program, using mass-balance metabolic models as biological constraints.

26 By applying the modular cell design framework for a genome-scale metabolic model
27 of *Escherichia coli*, we demonstrated that a library of biochemically diverse products
28 could be effectively synthesized at high yields and rates from a modular (chassis) cell
29 with only a few genetic manipulations. Flux analysis revealed this broad modular-
30 ity phenotype is supported by the natural modularity and flexible flux capacity of core
31 metabolic pathways. Overall, we envision the developed modular cell design framework
32 provides a powerful tool for synthetic biology and metabolic engineering applications
33 such as industrial biocatalysis to effectively produce fuels, chemicals, and therapeutics
34 from renewable and sustainable feedstocks, bioremediation, and biosensing.

35 **Keywords**— Biocatalysis, modular cell, ModCell, modular design, metabolic network mod-
36 eling, constraint-based modeling, multi-objective optimization, mixed integer linear programming,
37 goal programming, Benders decomposition.

38 1 Introduction

39 Microbial metabolism can be engineered to produce a large space of molecules from renewable
40 and sustainable feedstocks.¹ Currently, only a handful of fuels and chemicals out of the many
41 possible molecules offered by nature are industrially produced by microbial conversion, mainly
42 because the strain engineering process is too laborious and expensive.² Thus, innovative technologies
43 enabling rapid and economically-feasible strain engineering are needed to harness a large space of
44 industrially-relevant biochemicals.¹⁻³ To tackle this challenge, the principles of modular design that
45 have shown great success in traditional engineering disciplines can be adapted to construct modular
46 cell biocatalysts in a plug-and-play fashion with minimal strain optimization cycles.⁴

47 Multi-objective optimization is a powerful mathematical framework widely applied in engi-
48 neering disciplines to tackle the optimal design of a complex system with multiple conflicting
49 objectives.^{5,6} This framework has recently been exploited for not only explaining the modularity of
50 natural biological systems that enable cellular robustness and adaptability⁷⁻¹¹ but also implement-
51 ing modular engineering design.¹² Using multi-objective optimization, microbial metabolism can
52 be redirected to generate modular production strains that are systematically assembled from an

53 engineered modular cell and exchangeable production modules, each of which synthesizes a target
54 molecule.¹³ This modular cell (ModCell) design approach, known as ModCell2, uses the principles
55 of mass balance and thermodynamics of biochemical reaction networks to predict metabolic fluxes
56 upon genetic manipulations.^{13,14} Based on such flux predictions, a multi-objective optimization
57 problem is then formulated and solved with a multi-objective evolutionary algorithm (MOEA)^{15,16}
58 to yield a sample of the Pareto front (i.e., the set of optimal solutions to the problem with minimal
59 trade-offs among objectives) that a designer can explore genetic manipulation targets for modular
60 cell engineering.

61 In this study, we developed ModCell2-MILP, a ModCell2-based formulation to be compatible
62 with mixed integer linear programming (MILP) algorithms. This framework presents a significant
63 advancement from ModCell2 in solving the multi-objective strain design problem for modular cell
64 engineering. Specifically, ModCell2-MILP is developed to (i) guarantee optimal solutions, (ii) com-
65 pletely enumerate alternative solutions of a target design, and (iii) describe practical engineering
66 goals more directly (e.g., design of a modular cell where all production modules lead to a prod-
67 uct yield above 50% of the theoretical maximum). By applying ModCell2-MILP to analyze the
68 genome-scale metabolic network of *Escherichia coli*, we could identify a universal modular cell that
69 is compatible with a diverse class of production modules. Finally, we shed light on the underlying
70 features of the universal modularity phenotype by systematically analyzing feasible flux distribu-
71 tions of all modular production strains. We anticipate ModCell2-MILP can provide a powerful tool
72 for not only elucidating natural and synthetic metabolic modularity but also rationally designing
73 modular production strains for novel synthetic biology and metabolic engineering applications.

74 **2 Methods**

75 **2.1 Modular cell design**

76 **2.1.1 Design principles**

77 ModCell design enables rapid assembly of production strains with desirable phenotypes from a
78 modular (chassis) cell.^{4,13,17} More specifically, a modular cell contains core metabolic pathways

79 shared among production modules (Figure 1a). The chassis interfaces with the modules through
80 enzymatic and genetic synthesis machinery and precursor metabolites (Figure 1b). Modules con-
81 tain auxiliary regulatory and metabolic pathways (Figure 1c) that enable a desired phenotype for
82 optimal biosynthesis of a target molecule, for example, weak growth coupled to product formation
83 (*wGCP*), where a positive correlation between growth and product synthesis rates is enforced (Fig-
84 ure 1d).^{13,18,19} The design objective phenotypes are determined from cellular growth and product
85 synthesis rates based on steady-state, mass-balance metabolic models.²⁰ A modular cell is said to
86 be *compatible* with a module if the design objective of the resulting production strain is above a
87 specified threshold. The different biochemical nature of production modules to synthesize target
88 metabolites can make the design objectives compete with each other and also the cellular objec-
89 tives (e.g., biomass formation) compete with the engineering objectives (e.g., product formation),
90 turning the ModCell design problem into a multi-objective and multi-level optimization problem.

91 2.1.2 Multi-objective optimization formulation

The modular cell design problem is stated as a general multi-objective optimization problem of the form:

$$\max_x F(x) = (f_1(x), f_2(x), \dots, f_K(x))^T \quad \text{s.t. } x \in X \quad (1)$$

where f_k is the desirable phenotype for production module k , x are the problem variables including binary design variables corresponding to genetic manipulations, and X is the set of constraints including mass balance of metabolism. Optimal solutions for the multi-objective optimization problem (1) are defined using the concept of domination: A vector $a = (a_1, \dots, a_K)^T$ *dominates* another vector $b = (b_1, \dots, b_K)^T$, denoted as $a \prec b$, if and only if $a_i \geq b_i \forall i \in \{1, 2, \dots, K\}$ and $a_i \neq b_i$ for at least one i . A feasible solution $x^* \in X$ of the multi-objective optimization problem is called a Pareto optimal solution if and only if there does not exist a vector $x' \in X$ such that $F(x') \prec F(x^*)$. The set of all Pareto optimal solutions is called Pareto set:

$$PS := \{x \in X : \nexists x' \in X, F(x') \prec F(x)\} \quad (2)$$

The projection of the Pareto set in the objective space is denoted as Pareto front:

$$PF := \{F(x) : x \in PS\} \quad (3)$$

92 Different feasible points in PS (i.e., different genetic manipulations) which map to a single point
93 in PF (i.e., the same phenotype) are denoted *alternative solutions*.

94 The design variables x in ModCell2 correspond to chassis reaction deletions, that remove un-
95 desired metabolic functions, and module reaction insertions, that allow to identify optimal module
96 configurations without extensive prior knowledge of the product synthesis pathway. The con-
97 straint set X is comprised of two types: (i) flux simulation constraints (e.g., mass balance, reaction
98 reversibility, and flux bound) that allow to predict fluxes in the design objectives upon genetic
99 manipulations, and (ii) implementation constraints that involve the maximum number of reaction
100 deletions in the chassis (denoted by α) and the maximum number of module reaction insertions per
101 module (denoted by β). The following sections describe the problem formulation in detail using
102 the definitions compiled in Section 5.

103 2.1.3 Design objectives

Design objectives, f_k , that correspond to specific metabolic phenotypes within the space of feasible
steady-state reaction fluxes, Π_{km} , of production network k (i.e., the combination of the chassis
network with the production module k) and metabolic state m , are defined as follows:

$$\Pi_{km}(e_{jk}) := \{v_{jkm} \in \mathbb{R} : \quad (4)$$

$$\sum_{j \in \mathcal{J}_k} S_{ijk} v_{jkm} = 0 \quad \forall i \in \mathcal{I}_k \quad (5)$$

$$l_{jkm} e_{jk} \leq v_{jkm} \leq e_{jk} u_{jkm} \quad \forall j \in \mathcal{J}_k \quad (6)$$

104 Here, v_{jkm} is the rate (mmol/gCDW/hr) of reaction j in production network k under metabolic state
105 m . Constraint (5) enforces mass balance for all metabolites according to reaction stoichiometry
106 given by the coefficients S_{ijk} , and constraint (6) imposes bounds, l_{jkm} and u_{jkm} , for the metabolic
107 fluxes according to reaction reversibility, experimentally measured values, and specified metabolic

108 state. The binary variable e_{jk} is used in the overall optimization problem to indicate whether
 109 reaction j in production network k is removed and thus cannot carry any flux. Two metabolic
 110 states m are considered, growth and non-growth, denoted μ and $\bar{\mu}$, respectively. These states
 111 are differentiated by their flux bounds l_{jkm} and u_{jkm} . For growth state, the lower bound of the
 112 biomass formation reaction that represents cell division, $v_{Xk\mu}$, is set to a minimum value of γ , i.e.,
 113 $l_{Xk\mu} = \gamma$ ($\forall k \in \mathcal{K}$), while there is no upper limit to growth, i.e., $u_{Xk\mu} = \infty$ ($\forall k \in \mathcal{K}$). On the other
 114 hand, for the non-growth state both bounds are set to 0, i.e., $l_{Xk\bar{\mu}} = 0$ and $u_{Xk\bar{\mu}} = 0$ ($\forall k \in \mathcal{K}$).

Given the feasible metabolic flux space, Π_{km} , the following design objectives, based on the product synthesis rate reaction, v_{Pkm} , are of interest:

$$f_k^{wGCP} = \frac{v_{Pk\mu}}{v_{Pk\mu}^{max}} \in [0, 1], \quad \forall k \in \mathcal{K} \quad (7)$$

$$f_k^{lsGCP} = b_\mu \frac{v_{Pk\mu}}{v_{Pk\mu}^{max}} + b_{\bar{\mu}} \frac{v_{Pk\bar{\mu}}}{v_{Pk\bar{\mu}}^{max}} \in [0, b_\mu + b_{\bar{\mu}}], \quad \forall k \in \mathcal{K} \quad (8)$$

$$f_k^{NGP} = \frac{v_{Pk\bar{\mu}}}{v_{Pk\bar{\mu}}^{max}} \in [0, 1], \quad \forall k \in \mathcal{K} \quad (9)$$

The product synthesis fluxes, including $v_{Pk\mu}$, $v_{Pk\mu}^{max}$, $v_{Pk\bar{\mu}}$, and $v_{Pk\bar{\mu}}^{max}$, are computed by solving the following linear programming problems:

$$v_{Pk\mu} \in \arg \max \{v_{Xk\mu} - \epsilon v_{Pk\mu} : v_{k\mu} \in \Pi_{k\mu}(e_{jk})\} \quad (10)$$

$$v_{Pk\mu}^{max} \in \arg \max \{v_{Pk\mu} : v_{k\mu} \in \Pi_{k\mu}(e_{jk} = 1, \forall j \in \mathcal{J}_k)\} \quad (11)$$

$$v_{Pk\bar{\mu}} \in \arg \min \{v_{Pk\bar{\mu}} : v_{k\bar{\mu}} \in \Pi_{k\bar{\mu}}(e_{jk})\} \quad (12)$$

$$v_{Pk\bar{\mu}}^{max} \in \arg \max \{v_{Pk\bar{\mu}} : v_{k\bar{\mu}} \in \Pi_{k\bar{\mu}}(e_{jk} = 1, \forall j \in \mathcal{J}_k)\} \quad (13)$$

115 The maximum product synthesis fluxes (11) and (13) used for objective scaling are only calculated
 116 once by not using any deleted reactions ($e_{jk} = 1$), while the target phenotype fluxes (10) and (12)
 117 are functions of the deleted reactions e_{jk} . The design objectives, $wGCP$ (7), $lsGCP$ (8), and NGP
 118 (9), were previously proposed¹³ and briefly described here. The weak growth coupled to product
 119 formation objective ($wGCP$) (7) seeks to maximize the minimum product rate at the maximum
 120 cellular growth, which is accomplished by a titled objective function²¹ (10). The linearized strong

121 growth coupled to product formation (*lsGCP*) (8) objective seeks to maximize the minimum prod-
 122 uct synthesis rate at the non-growth state $v_{Pk\bar{\mu}}$ in addition to the goal of *wGCP*. Finally, the
 123 non-growth production (*NGP*) (9) objective seeks to optimize the minimum product synthesis rate
 124 during the non-growth state.

125 2.1.4 Design constraints

All the constraints of the modular cell design problem are gathered as follows:

$$\Omega := \{f'_k \in \mathbb{R}, y_j, z_{jk}, d_{jk}, w_k, e_{jk} \in \{0, 1\} : \quad (14)$$

$$\sum_{j \in \mathcal{C}} (1 - y_j) \leq \alpha \quad (15)$$

$$\sum_{j \in \mathcal{C} - \mathcal{N}_k} z_{jk} \leq \beta_k \quad \forall k \in \mathcal{K} \quad (16)$$

$$z_{jk} \leq 1 - y_j \quad \forall j \in \mathcal{C} - \mathcal{N}_k, k \in \mathcal{K} \quad (17)$$

$$d_{jk} = y_j \vee z_{jk} \quad \forall j \in \mathcal{C}, k \in \mathcal{K} \quad (18)$$

$$f'_k = f_k w_k \quad \forall k \in \mathcal{K} \quad (19)$$

$$e_{jk} = (d_{jk} \wedge w_k) \vee \neg w_k \quad \forall j \in \mathcal{C}, k \in \mathcal{K} \quad (20)$$

$$w_k \leq M^w f_k \quad \forall k \in \mathcal{K} \quad (21)$$

$$v_{Pkm} \in \Psi_{km}(e_{jk}) \quad \forall k \in \mathcal{K}, m \in \mathcal{M} \quad (22)$$

126 Constraints (15)-(18) are formulated for practical limitations and features of the modular cell.
 127 Specifically, the two variables that represent design choices for genetic manipulations include: (i)
 128 y_j that takes a value of 0 if reaction j is deleted in the chassis (and consequently in all production
 129 networks) and 1 otherwise and (ii) z_{jk} that takes a value of 1 if reaction j is inserted in production
 130 network k . The maximum number of reaction deletions, is limited by α through constraint (15)
 131 while the maximum number of module reactions in each module β_k is imposed by (16). Constraint
 132 (16) excludes non-candidate reactions \mathcal{N}_k (since $j \in \mathcal{C} - \mathcal{N}_k$) so that endogenous module reactions
 133 can be fixed (i.e., $z_{jk} = 1$), according to problem-specific knowledge. Constraint (17) ensures that
 134 only reactions deleted in the chassis can be inserted back to the modules. Constraint (18) indicates

135 that reaction j is deleted in production network k if the reaction is deleted in the chassis and not
136 added as an endogenous module reaction. The designer can gradually increase α and β_k to obtain
137 solutions with higher performance.

138 Constraints (19)-(21) are introduced for modeling purposes. The indicator variable, w_k , is
139 introduced to allow for certain production networks to be ignored from the final solution. Without
140 w_k , the whole multi-objective problem becomes infeasible if a set of deletions renders one of the
141 production networks infeasible (e.g., its minimum growth rate cannot be accomplished). However,
142 in practice it is acceptable for some modules not to work with the chassis cell. If $w_k = 0$, the
143 objective value $f'_k = 0$ (19) and reaction deletions do not apply to network k since $e_{jk} = 1$ (20); if
144 $w_k = 1$, $f'_k = f_k$ and $e_{jk} = d_{jk}$, where f_k is any of the design objectives presented earlier (7)-(9).
145 The use of w_k is likely to introduce symmetry (i.e., alternative integer solutions with no practical
146 meaning) due to cases where $f_k = 0$ for a given k while the associated production network remains
147 feasible, allowing w_k to take a value of 0 or 1. This symmetry is removed by enforcing w_k to be 0
148 if $f_k = 0$ (21).

149 Finally, constraint (22) indicates that the fluxes featured in the design objectives, v_{Pkm} , are
150 contained in the polytope Ψ_{km} . The space of v_{Pkm} is originally defined as an optimization problem
151 (10)-(13), thus representing a non-linear constraint and turning the ModCell design problem into
152 a bilevel optimization problem. These inner optimization problems are linearized, leading to Ψ_{km}
153 as described in Section 2.1.6.

154 2.1.5 Linearization of logical expressions

155 The logical expressions in Ω are replaced by the following linear constraints in the final problem
156 formulation:

$d_{jk} = y_j \vee z_{jk}$ corresponds to:

$$d_{jk} \leq y_j + z_{jk} \quad (23)$$

$$d_{jk} \geq y_j \quad (24)$$

$$d_{jk} \geq z_{jk} \quad (25)$$

$$0 \leq d_{jk} \leq 1 \quad (26)$$

$f'_k = f_k w_k$ corresponds to:

$$f'_k \leq w_k M^{obj} \quad (27)$$

$$f'_k \leq f_k - (1 - w_k) M^{obj} \quad (28)$$

$$f'_k \leq f_k \quad (29)$$

$$0 \leq f'_k \leq M^{obj} \quad (30)$$

$e_{jk} = (d_{jk} \wedge w_k) \vee \neg w_k$, given $r_{jk} = d_{jk} \wedge w_k$, corresponds to:

$$e_{jk} = r_{jk} + 1 - w_k \quad (31)$$

$$r_{jk} \leq w_k \quad (32)$$

$$r_{jk} \leq d_{jk} \quad (33)$$

$$r_{jk} \geq w_k + d_{jk} - 1 \quad (34)$$

$$0 \leq r_{jk} \leq 1 \quad (35)$$

157 2.1.6 Linearization of inner optimization problems

Non-linear constraints expressed as linear programming problems can be linearized using basic mathematical programming theory. Consider the following canonical linear program, with primal variables $x \in \mathbb{R}^n$ and its dual variables $u \in \mathbb{R}^m$:

$$\max \{c^\top x : Ax \leq b, x \geq 0\} \quad (36)$$

$$\min \{b^\top u : A^\top u \geq c, u \geq 0\} \quad (37)$$

the strong duality theorem states that the objective functions of primal (36) and dual (37) are equal at their optima, $c^\top x^* = b^\top y^*$. Thus the optimal solution to the primal problem is described by the following linear constraints:

$$x^* \in \{x \in \mathbb{R}^n : \quad (38)$$

$$Ax \leq b \quad (39)$$

$$A^T u \geq c \quad (40)$$

$$c^T x = b^T u \quad (41)$$

$$x, u \geq 0 \quad (42)$$

Using the strong duality theorem as presented by Maranas and Zomorodi,²² the inner optimization problems (22) are linearized as follows:

$$\Psi_{km}(e_{jk}) := \{v_{jkm} \in \mathbb{R} : \quad (43)$$

$$\sum_{j \in \mathcal{J}_k} S_{ijk} v_{jkm} = 0 \quad \forall i \in \mathcal{I}_k \quad (44)$$

$$l_{jkm} e_{jk} \leq v_{jkm} \leq e_{jk} u_{jkm} \quad \forall j \in \mathcal{J}_k \quad (45)$$

$$\sum_{i \in \mathcal{I}_k} \lambda_{ikm} S_{ijk} - \mu_{jkm}^l + \mu_{jkm}^u = c_{jkm} \quad \forall j \in \mathcal{J}_k \quad (46)$$

$$\lambda_{ikm} \in \mathbb{R} \quad \forall i \in \mathcal{I}_k \quad (47)$$

$$0 \leq \mu_{jkm}^l \leq M \quad \forall j \in \mathcal{J}_k \quad (48)$$

$$0 \leq \mu_{jkm}^u \leq M \quad \forall j \in \mathcal{J}_k \quad (49)$$

$$\begin{aligned} \sum_{j \in \mathcal{J}_k} c_{jkm} v_{jkm} = & - \sum_{j \in \mathcal{J}_k - \mathcal{C}} (l_{jkm} \mu_{jkm}^l) + \sum_{j \in \mathcal{J}_k - \mathcal{C}} (u_{jkm} \mu_{jkm}^u) \\ & - \sum_{j \in \mathcal{C}} (l_{jkm} p_{jkm}^l) + \sum_{j \in \mathcal{C}} (u_{jkm} p_{jkm}^u) \end{aligned} \quad (50)$$

$$p_{jkm}^l \leq e_{jk} M \quad \forall j \in \mathcal{C} \quad (51)$$

$$\mu_{jkm}^l - (1 - e_{jk}) M \leq p_{jkm}^l \leq \mu_{jkm}^l \quad \forall j \in \mathcal{C} \quad (52)$$

$$0 \leq p_{jkm}^l \leq M \quad \forall j \in \mathcal{C} \quad (53)$$

$$p_{jkm}^u \leq e_{jk} M \quad \forall j \in \mathcal{C} \quad (54)$$

$$\mu_{jkm}^u - (1 - e_{jk}) M \leq p_{jkm}^u \leq \mu_{jkm}^u \quad \forall j \in \mathcal{C} \quad (55)$$

$$0 \leq p_{jkm}^u \leq M \quad \forall j \in \mathcal{C} \quad (56)$$

158 Constraints (44)-(45) correspond to the primal metabolic network problem and were introduced
 159 earlier in Π_{km} . Constraints (46)-(49) correspond to the dual problem. We use the dual variables,

160 λ_{ikm} , for the primal mass balance constraints (44), together with μ_{jkm}^l and μ_{jkm}^u for the primal flux
161 bound inequalities (45) involving lower and upper reaction bounds respectively. Constraints (47)-
162 (49) emphasize the domain of the dual variables, with M being a large value above the expected
163 value of any dual variable. Constraints (50)-(56) correspond to the strong duality equality. The left
164 hand side of the strong duality equality (50) features the objectives presented in (10) for $m = \mu$ and
165 (12) for $m = \bar{\mu}$. On the right hand side, products of binary and continuous variables appear, thus
166 requiring linearization variables p_{jkm}^l and p_{jkm}^u . Constraints (51)-(56) ensure that $p_{jkm}^l = e_{jk}\mu_{jkm}^l$
167 and $p_{jkm}^u = e_{jk}\mu_{jkm}^u$.

168 2.1.7 Conversion of a multi-objective problem into a single-objective problem

169 The multi-objective optimization problem (1) is now described entirely in terms of linear constraints
170 through Ω . However, to make the formulation compatible with MILP solver algorithms, the objec-
171 tive function vector, f' , must be expressed as a scalar. To accomplish this without loss of relevant
172 information, we employed blended and goal attainment formulations.

173 2.1.8 Blended formulation

In the blended formulation,²³ all objectives are summed as follows:

$$\max \sum_{k \in \mathcal{K}} a_k f'_k \quad \text{s.t. } f' \in \Omega \quad (57)$$

174 where a_k is a scalar weighting factor associated with the design objective of product k . Different
175 Pareto optimal solutions can be obtained by varying these weights. The blended formulation always
176 provides Pareto optimal solutions as long as $a_k > 0$ ($\forall k \in K$). In practice, the product priority,
177 a_k , can be determined by criteria such as product market value or “pathway readiness level” (i.e.,
178 certain pathways are easier to engineer than others).

179 2.1.9 Goal attainment formulation

In the goal attainment problem,²³ a target value is defined for each objective:

$$\min \sum_{k \in \mathcal{K}} (a_k^+ \delta_k^+ + a_k^- \delta_k^-) \quad (58)$$

s. t.

$$f'_k + \delta_k^+ - \delta_k^- = g_k \quad \forall k \in \mathcal{K} \quad (59)$$

$$\delta_k^+, \delta_k^- \geq 0 \quad \forall k \in \mathcal{K} \quad (60)$$

$$f' \in \Omega \quad (61)$$

180 The problem seeks to minimize the variables δ_k^+ and δ_k^- that represent the deficiency and excess of
181 the objective f'_k from the target value g_k , respectively. Weighting parameters a_k^+ and a_k^- correspond
182 to different types of discrepancy to be minimized. In general, when it is important to meet the
183 target value without exceeding it, we set $a_k^+ = a_k^- = 1$; however, when the design objective is
184 required to be greater or equal than the target value, we set $a_k^+ = 1$ and $a_k^- = 0$, effectively
185 converting (59) into $f'_k + \delta_k^+ \geq g_k$. Solutions to the goal attainment problem are not guaranteed to
186 be Pareto optimal, even if all demands g_k are met. To address this issue, the blended problem (57)
187 can be solved where the objectives are constrained to be equal or greater than the values found
188 by solving the goal attainment problem. In practice, the goal attainment formulation corresponds
189 to the identification of the modular cell *compatible* with the largest number of modules. Here, a
190 module k is said to be *compatible* if $f'_k \geq g_k$.

191 2.2 Implementation

192 2.2.1 Metabolic models

193 We used two parent models from which production networks were built, including: i) a core
194 metabolic model of *E. coli*¹⁷ to develop the ModCell2-MILP algorithm and compare with previous
195 ModCell2 results,¹³ and ii) the iML1515 genome-scale metabolic model of *E. coli*²⁴ for biosynthesis
196 of a library of endogenous and heterologous metabolites, including 4 organic acids, 6 alcohols, and

197 10 esters (Table 2).^{25–34} These models were configured as in the previous ModCell2 study¹³, briefly:
198 Anaerobic conditions were imposed by setting oxygen exchange fluxes to be 0, and the glucose up-
199 take rate was constrained to be at most 10 mmol/gCDW/h. When using the genome-scale model
200 iML1515 to simulate *wGCP* designs, only the commonly observed fermentative products (acetate,
201 CO₂, ethanol, formate, lactate, succinate) were allowed for secretion as described elsewhere.³⁵

202 2.2.2 ModCell2-MILP simulations

203 ModCell2-MILP was implemented using Pyomo,³⁶ an algebraic modeling language embedded in the
204 Python programming language. All simulations were performed on a computer with an Intel Core
205 i7-3770 processor, 32 GB of random access memory, and the Arch Linux operative system. The
206 implementation and scripts used to generate the results of this manuscript are available as part of the
207 ModCell2 package via Supplementary Material 2 and <https://github.com/trinhlab/modcell2>.

208 2.2.3 Optimization solver configuration

209 The Pyomo³⁶ implementation of ModCell2-MILP was solved with IBM Ilog Cplex 12.8.0. To
210 avoid incorrect solutions associated with numerical issues the following Cplex parameters were
211 changed from their default values: (i) *numerical emphasis* was set to “true”, (ii) *integrality tolerance*
212 was lowered to 10⁻⁷, and (iii) the *MIP pool relative gap* was increased to 10⁻⁴ for enumerating
213 alternative solutions. Alternative solutions were enumerated using the Cplex “populate” procedure.

214 2.3 Analysis methods

215 2.3.1 Reference flux distribution

The reference flux distribution, $\frac{v_{jk}^*}{|v_{S_k}^*|}$, is determined by solving the following quadratic program based on the parsimonious enzyme usage hypothesis:^{37,38}

$$\begin{aligned} \min_{v_{jk}} \quad & \sum_{j \in \mathcal{J}_k} v_{jk}^2 & (62) \\ \text{s.t.} \quad & \end{aligned}$$

$$\sum_{j \in \mathcal{J}_k} S_{ijk} v_{jkm} = 0 \quad \forall i \in \mathcal{I}_k \quad (63)$$

$$l_{jk} \leq v_{jk} \leq u_{jk} \quad \forall j \in \mathcal{J}_k \quad (64)$$

$$v_{Xk} = \text{MaxDesignBio} \quad (65)$$

216 Constraint (63) corresponds to mass balance for the metabolic network. Constraint (64) corre-
217 sponds to reaction bounds, including reaction deletions found in the modular cell design problem.
218 Constraint (65) fixes the biomass formation rate, v_{Xk} , to the maximum reachable by the design.
219 This value (MaxDesignBio) is obtained by maximizing v_{Xk} subject to (63) and (64). The refer-
220 ence flux distribution $\frac{v_{jk}^*}{|v_{S_k}^*|}$ represents the desired metabolic state of a *wGCP* designed production
221 network. This distribution, if feasible, is unique because the convex optimization problem is for-
222 mulated with a positive definite quadratic objective function (see Theorem 16.4 in Nocedal and
223 Wright³⁹).

224 2.3.2 Flux sampling

225 To determine an ensemble of flux distributions for a production network, we used the ACHR
226 algorithm⁴⁰ in the COBRA toolbox.⁴¹ Constraints for flux sampling simulation include the reaction
227 deletions and module reactions found in the ModCell design problem solution, a fixed substrate
228 uptake rate of -10 mmol glucose/gCDW/hr, and a minimum product synthesis flux of 50% of its
229 maximum value.

230 2.3.3 Metabolic map drawing

231 Drawings of metabolic map were performed using the Escher⁴² tool (<https://escher.github.io>)
232 that produces *svg* files. Coloring, highlighting candidate reactions, and other systematic adjust-
233 ments of metabolic maps were done with the Python-based *lxml* module. Additional editing for
234 visual enhancement was done with the Inkscape software.

235 3 Results

236 3.1 Performance and solution time optimization of ModCell2-MILP

237 3.1.1 ModCell2-MILP can not only reproduce the results of the original Mod- 238 Cell2 formulation but also find more alternative solutions

239 To evaluate ModCell2-MILP, we compared its performance with the previously developed Mod-
240 Cell2 platform¹³ that solves the optimization problem with multi-objective evolutionary algorithms
241 (MOEAs). As a basis of comparison, we used the same *E. coli* core metabolic model, maximum
242 number of deletion reactions α , and maximum number of module-specific reactions β_k for both Mod-
243 Cell2 and ModCell2-MILP. Due to fundamental differences in problem formulations for MOEA and
244 MILP, we used the *lsGCP* design objective for ModCell2-MILP with multiple weighting factors, a_k ,
245 specifically selected to reproduce previous results, in the blended formulation and the *sGCP* design
246 objective for ModCell2 (Supplementary Material 1). The results showed that ModCell2-MILP could
247 generate the same Pareto optimal designs like ModCell2. In addition, ModCell2-MILP enumerated
248 a larger number of alternative solutions than ModCell2. For example, the design named *sGCP-5-*
249 *0-6* generated by ModCell2 had 3 alternative solutions while ModCell2-MILP found 8 alternative
250 solutions. By increasing α to 8 and β to 2, we could identify a utopia design (i.e., one solution with
251 the maximum value for all objectives) with 192 alternative solutions, which significantly expands
252 the possibilities for experimental implementation.

253 3.1.2 Tuning MILP formulations significantly improves solution times

254 We considered three techniques that can improve solution times of ModCell2-MILP, including:

255 (i) *Fixing the network feasibility indicator w_k* . If all modules are expected to be compatible
256 with a final ModCell design (i.e., $f_k > 0, \forall k \in \mathcal{K}$), w_k is set to be 1 for all $k \in \mathcal{K}$ in order to avoid
257 computational efforts in finding non-optimal feasible solutions.

258 (ii) *Flux bound tightening*. Constraints of the form $e_{jkm}l_{jkm} \leq v_{jkm} \leq e_{jkm}u_{jkm}$ are known to
259 result in weak linear relaxations, i.e., feasible values of v_{jkm} are far from their bounds l_{jkm} and
260 u_{jkm} . To tighten the formulation by making continuous relaxations closer to the feasible integer

261 solution, smaller values of u_{jkm} and l_{jkm} are determined by solving a series of linear programs that
262 maximize and minimize each flux v_{jkm} in the parent production networks $\Pi_{km}(e_{jk} = 1, \forall j \in \mathcal{J}_k)$.

263 (iii) *Benders decomposition*. ModCell2-MILP has a separable structure compatible with Benders
264 decomposition^{43,44} that creates a master problem, using binary variables and associated constraints
265 (15)-(21), and sub-problems for each production network $\Psi_{km}(e_{jk})$ with fixed binary variables. This
266 decomposition implementation is automatically done by Cplex 12.8.

267 We evaluated these three techniques for tuning MILP formulations and used the core *E. coli*
268 model¹³ for the benchmark study. The results showed that flux bound tightening, fixed w_k , and
269 Benders decomposition could reduce the solution time to find solutions by 50%, 80%, and 95%,
270 respectively (Table 1). By combining these techniques, the solution time was shortened by 96%
271 from 63.3 s to 2.8 s. In subsequent studies, we used these three tuning techniques to solve the
272 ModCell design problem unless otherwise noted.

273 3.1.3 Choices of design parameters affect solution time

274 In designing a modular cell with ModCell2-MILP, the designer needs to specify the formulation
275 type (i.e., blended or goal attainment formulation), the target phenotype (e.g., *wGCP*, *lsGCP*, and
276 *NGP*), and the limits of deletion reactions (α) and endogenous module-specific reactions (β_k). We
277 evaluated the impact of these parameters on solution time using the *E. coli* core model (Figure
278 2). Regardless of the formulation type, increasing α and β led to harder problems and hence
279 required more solution time due to the exponentially increasing number of feasible solutions as
280 expected. The goal attainment formulation took longer time to solve for the *lsGCP* and *NGP*
281 design objectives, but about the same time for the *wGCP* design objective. Interestingly, the
282 overall difficulty of *wGCP* is higher than that of *lsGCP* in both the blended and goal attainment
283 formulations, despite *lsGCP* having approximately twice the number of constraints. Furthermore,
284 the *NGP* design objective could be solved most quickly, likely due to the narrower design space
285 associated with the no-growth associated production of target metabolites.

286 **3.2 Design of a universal modular cell for a genome-scale metabolic** 287 **model of *E. coli***

288 **3.2.1 Reduction of the candidate reaction deletion set enables ModCell2-MILP** 289 **to find modular cell designs for a large-scale metabolic network**

290 Finding genetic modifications towards a desired phenotype using mathematical optimization for
291 large-scale metabolic networks has been known to be a computationally expensive task due to
292 the combinatorial search space spanned by a large number of reaction deletion candidates in the
293 network.^{21,45} Preprocessing of metabolic networks to reduce reaction candidates is not only critical
294 but also practical for experimental implementation. Previous implementation of ModCell2 for the
295 latest genome-scale *E. coli* model (iML1515)²⁴ showed that the preprocessing step could reduce the
296 set of reaction candidates from 2,712 to 276. By using ModCell2 with the *wGCP* objective, an *E.*
297 *coli* modular cell was identified to be compatible with 17 out of 20 products with requirement of
298 only 4 reaction deletions.¹³ Since MOEA implemented in ModCell2 does not guarantee optimality,
299 here we aimed to evaluate the capability of ModCell2-MILP for handling a large-scale metabolic
300 network and identifying the Pareto optimality and potential alternative solutions.

301 We applied ModCell2-MILP to analyze the same iML1515 model with a set of 20 products using
302 the same design parameters (i.e., α and β_k) and the blended formulation with all objective weights
303 $a_k = 1$. The simulation shows that ModCell2-MILP could not solve the ModCell design problem to
304 optimality over 2 days of run time, likely due to the large number of candidate deletion reactions
305 still present in the genome-scale model. To address this problem, the set of candidate reactions
306 must be further reduced. Since only a small subset of all metabolic reactions in genome-scale
307 models tend to be deleted by strain design algorithms,^{13,21,46} we used a pool of *wGCP* designs with
308 $\alpha = 4, 5, 6$ and $\beta = 0, 1$ reported with ModCell2¹³ to identify relevant deletion candidates. From a
309 set of 601 designs found by ModCell2, only 33 out of 276 candidates reaction deletions were used
310 at least once. Hence, these 33 reactions were used to create a new, computationally-tractable set
311 of reaction candidates. This new set contains reactions mostly from the well-characterized central
312 metabolic pathways (Figure 3a) while the original set includes reactions in peripheral pathways
313 that lead to biomass synthesis. Interestingly, within these 33 reaction candidates, only a few are

314 used in most designs (Figure 3b), highlighting the importance of their removal in growth-coupled
315 production phenotypes. Reactions with high deletion frequencies mainly occur in high-flux central
316 metabolic pathways (Figure 3c), closely associated with cellular energetics and carbon precursors
317 that interface with the production modules (Figure 3d).

318 Using the reduced candidate reaction deletion set, ModCell2-MILP could find an optimal solu-
319 tion in ~ 30 min and enumerated all optimal solutions in ~ 8 hours. All the optimal solutions found
320 by ModCell2-MILP in this case were in agreement with those previously found in ModCell2.¹³

321 **3.2.2 ModCell2-MILP can identify a universal modular cell compatible with all** 322 **exchangeable production modules**

323 Based on the computationally-tractable candidate reaction deletion set, we next evaluated whether
324 the goal programming formulation could help identify a universal ModCell design that is compatible
325 with all modules. By screening for various α and β_k , we identified a universal modular cell that is
326 compatible with all production networks, corresponding to the defined minimum design objective
327 goal of 0.5 (i.e., 50% of the theoretical maximum product yield attained at the maximum growth
328 rate), $\alpha = 6$, and $\beta = 1$ (Figure 4a). Remarkably, most products greatly overcame this minimum
329 goal with yields above 90% of the theoretical maximum values (Figure 4b). All production networks
330 displayed a feasible metabolic space where an increase in product synthesis rate is needed to attain
331 faster growth rates (Figure 4c). This designed phenotype is useful for optimal pathway selection
332 using adaptive laboratory evolution^{47,48} and/or pathway libraries.⁴⁹

333 **3.3 Flexible metabolic flux capacity of *E. coli* core metabolism** 334 **enables the design of a universal modular cell**

335 **3.3.1 Endogenous modules responsible for metabolic flexibility of a universal** 336 **modular cell are identified by comparing flux distributions of production** 337 **networks**

338 The designed universal modular cell (Section 3.2.2) can theoretically adapt to the contrasting
339 metabolic requirements of all production modules (Table 2). To gain further insight into this unique

340 metabolic capability of the modular cell and its potential to be realized in practice, we analyzed
341 its *reference flux distributions* (Section 2.3.1) across the production networks. Reactions with the
342 highest flux changes across the production networks are likely critical for the proper operation of
343 the universal modular cell and might present potential bottlenecks. Such reactions were identified
344 by filtering their reference flux standard deviation (calculated across production networks) with an
345 *ad hoc* threshold of 0.2 (mol/substrate mol). Over 90% of the 535 active reactions, each of which
346 carries a non-zero flux in at least one production network, had standard deviation values below
347 the threshold, indicating highly conserved metabolic core pathways among production networks.
348 Only 9.5% of the active reactions presented a standard deviation magnitude above the threshold
349 (Figure 5a).

350 In our case study of designing a universal modular cell compatible with all 20 production mod-
351 ules, unbiased clustering analysis (Figure 5b) revealed the presence of four endogenous module
352 types in the core metabolism of *E. coli* that are activated to fit specific production modules (Fig-
353 ure 5c). In the context of chassis metabolism, an endogenous module corresponds to a reaction
354 or group of highly coupled reactions that become active to accomplish a certain metabolic func-
355 tion. The endogenous module classification can be understood in terms of location (i.e., proximity
356 in the metabolic network) and three metabolic functions. The first function is the direction of
357 carbon towards general precursor metabolites including (i) pyruvate and acetyl-CoA captured by
358 acetyl-CoA-associated modules and (ii) oxaloacetate, succinate, succinyl-CoA, and α -ketoglutarate
359 captured by TCA-associated modules. The second function is the direction of carbon from the pre-
360 cursor metabolites towards secretable molecules, captured by the upstream and TCA-associated
361 modules. The third function is the use of ATP- and NADP(H)-dependent pathways required
362 to maintain homeostasis, captured by the acetyl-CoA-associated and energetic modules. While
363 these functions are conceptually separable, their biochemical manifestation overlaps, i.e., specific
364 metabolic reactions or pathways can simultaneously fulfill several functions.

365 Each endogenous module can be viewed as an interface of the universal modular cell with
366 production modules that are exchangeable. The endogenous modules might become potential
367 metabolic bottlenecks in practice if they cannot satisfy the predicted fluxes, and thus might be
368 critical engineering targets when the associated production modules are used.

369 **Acetyl-CoA-associated endogenous modules.** This module type contains pyruvate for-
370 mate lyase (PFL) and pyruvate dehydrogenase enzyme complex (PDH) reactions that convert
371 pyruvate to acetyl-CoA. Intuitively, products derived from pyruvate, such as isobutanol, require
372 a low flux through PFL and PDH while those derived from acetyl-CoA require a high flux. Re-
373 markably, the redox states of production strains determine the ratios of PFL to PDH fluxes. For
374 example, the ethanol production network has a relatively high flux through PDH and a low flux
375 through PFL; however, for ethyl acetate that has a lower degree of reduction than ethanol (Table 2),
376 PFL with formate secretion is prioritized over PDH with NADH generation. Note that our model
377 did not include the regulatory restriction that PDH is inhibited in *E. coli* anaerobically because the
378 function of PDH is equivalent with the coupling of PFL and heterologous NADH-dependent for-
379 mate dehydrogenase (FDH) demonstrated experimentally for increased butanol^{30,50} and pentanol³²
380 production.

381 **Upstream modules.** This module type is formed by reactions located directly upstream of
382 a secretable metabolite, often associated with the target production module, and thus provides
383 the necessary precursor metabolite(s). Such reactions are commonly over-expressed in practice,
384 e.g., the ECOAH1-HACD1-ACACT1r endogenous module (comprising of 3-hydroxyacyl-CoA de-
385 hydratase, 3-hydroxyacyl-CoA dehydrogenase, and acetyl-CoA acetyl transferase) responsible for
386 generating butyryl-CoA and the ACLS-DHAD1-KARA1 endogenous module (comprising of aceto-
387 lactate synthase, dihydroxy-acid dehydratase, and keto-acid reductoisomerase) responsible for gen-
388 erating isobutyryl-CoA. These endogenous modules can also become active to form byproducts in
389 certain production networks, e.g., the PTAr-ACKr-ACT2rpp-ACtex endogenous module (compris-
390 ing of phosphate acetyl transferase, acetate kinase, and cytosolic and periplasmic acetate transport)
391 that not only carries the highest flux in the acetate production network but also becomes active in
392 the propanol-associated modules.

393 **TCA-associated endogenous modules** This module type has the same function as the
394 upstream endogenous modules but it is localized in the TCA (Krebs) cycle. Several products,
395 including adipic acid, 1,4-butanediol, propanol, pentanol, and their associated esters, are derived
396 from the TCA intermediates and interface with the universal modular cell via the TCA-derived

397 endogenous modules. The SUCOAS-MMM-MMCD endogenous module (comprising of succinyl-
398 CoA synthetase, Methylmalonyl-CoA mutase, methylmalonyl-CoA decarboxylase) must be acti-
399 vated to convert succinate into succinyl-CoA and then propanoyl-CoA. Remarkably, two routes are
400 present to synthesize fumarate from oxaloacetate, including the conventional MDH-FUM endoge-
401 nous module (comprising of malate dehydrogenase and fumarase) that consumes NADH and the
402 cyclic ASPTA-GLUDY-ASPT endogenous module (comprising of aspartate transaminase, gluta-
403 mate dehydrogenase, and L-aspartase) that consumes NADPH. These NADH/NADPH cofactors
404 are not interchangeable due to the deletion of the transhydrogenase THD2pp in the universal mod-
405 ular cell, so the isobutyl pentanoate and pentyl pentanoate modules, that are derived from the
406 ASPTA-GLUDY-ASPT endogenous module, also have a high NADPH requirement. Some pro-
407 duction networks, such as pyruvate and isobutyl acetate that are not based on the TCA-derived
408 endogenous modules, secrete succinate instead of ethanol and/or lactate to balance redox by using
409 the PPC-MDH-FUM-SUCctex endogenous module (comprising of phosphoenolpyruvate carboxy-
410 lase, malate dehydrogenase, fumarase, and succinate transport).

411 **Energetic modules** This module type primarily involves NAD(P)-dependent transhydroge-
412 nase (THD2pp) and ATP synthase (ATPS4rpp). Other reactions that allow coupling of phosphate-
413 and electron-transfer cofactors are also included. The reactions in this module help buffer the
414 diverse electron and ATP requirements of production networks. THD2pp is deleted in the chassis
415 but used as a module reaction in the isobutanol and acetate production networks. In the case of
416 isobutanol production, transhydrogenase expression has been demonstrated to increase the synthe-
417 sis of NADPH and thus isobutanol.⁵¹ Acetate has the smallest degree of reduction after pyruvate,
418 which results in redox imbalance that is compensated via formate secretion. In conjunction with
419 these mechanisms, ATPsynthase works in the reverse direction by hydrolyzing excess ATP. Other
420 production networks also use ATPS4rpp to eliminate excess ATP as observed, for example, in
421 the ethyl acetate production network. This strategy is consistent with ATP wasting approaches
422 recently demonstrated.⁵²

3.3.2 Comparison between simulated and measured intracellular fluxes reveals flexible metabolic flux capacity of *E. coli* to accommodate the required wide flux ranges

Flux analysis of the production networks suggests that the core metabolic reactions (Figure 5b) require a wide range of fluxes. To successfully implement this modular design in practice, we need to evaluate whether the metabolism of *E. coli* has the inherent metabolic flux capacity to accommodate the required fluxes of the designed universal modular cell when coupled with various exchangeable production modules. We compared the simulated reference flux distributions with a recent collection of 45 measured metabolic fluxes⁵³ that are collected from multiple studies across various conditions (e.g., growth under aerobic and anaerobic conditions, use of glucose or acetate or pyruvate as a carbon source) and genotypes (e.g., wild-type *E. coli* and mutants with single gene deletions).⁵⁴⁻⁵⁷ Note that this dataset provides a baseline for wild-type and relatively small deviations from that state (i.e., single gene deletion mutants), thus highly engineered strains (e.g., with three or more gene deletions) are likely to attain wider flux distributions.

Within the 23 reaction groups that constitute endogenous modules (Figure 5b), 8 reactions could be matched to this experimental dataset (Figure 5d). Remarkably, a highly consistent overlap of flux ranges was observed between the simulated and measured fluxes for malate dehydrogenase (MDH), pyruvate dehydrogenase (PDH), acetaldehyde dehydrogenase (ACALD), fumarase (FUM), and 2-dehydro-3-deoxy-phosphogluconate aldolase (EDA). For the cases of D-lactate dehydrogenase (LDH_D), and pyruvate secretion (EX_pyr_e) that are directly coupled with the biosynthesis of lactate and pyruvate, respectively, we observed the maximum simulated fluxes surpass the measured values, suggesting that further engineering of wild-type and single-gene deletion *E. coli* is needed to attain the required fluxes. Indeed, previous studies^{58,59} have been able to redirect metabolic fluxes in *E. coli* for yields of lactate and pyruvate above 75% of the theoretical maximum values by simultaneous elimination of competing fermentative pathways, including acetate ($\Delta ackA$), formate ($\Delta pflB$), and ethanol ($\Delta adhE$). The only remaining discrepancy between the simulated and measured fluxes is PPC. Studies, not included in the comparison data set, have reported up to 50% more PPC flux observed under aerobic conditions^{60,61}, which is still considerably below several of the

451 simulated fluxes. This result suggests that PPC can be a potential metabolic bottleneck in certain
452 production modules. One potential solution is to include in the affected production modules the
453 heterologous PPC from *Actinobacillus succinogenes* which has been successfully over-expressed in
454 *E. coli* for increased succinate production.⁶² Additionally, bacterial PPC activity can be increased
455 by elevating the acetyl-CoA pool.⁶³

456 **3.3.3 Random sampling of metabolic fluxes confirms the narrow operation** 457 **range of endogenous modules**

458 The calculated reference flux distributions represent the ideal metabolic states for each produc-
459 tion strain. However, other metabolic states might also exist. To address this uncertainty, we
460 performed randomized flux sampling^{40,41} for each production network under the constraint that
461 product synthesis rate has to be above 50% of the maximum value (Section 2.3.2). The results
462 show that the metabolic flux distributions for most reactions involved in the endogenous modules
463 (Figure 6a-u) are very narrow, except the two alternative pathways for ethanol synthesis, i.e., the
464 endogenous PDH-ACALD-ALCD2x route (comprising of pyruvate dehydrogenase, acetaldehyde
465 dehydrogenase, and alcohol dehydrogenase) (Figure 6t) route and the heterologous PDC-ALCD2x
466 route (comprising of pyruvate decarboxylase and alcohol dehydrogenase). The range of experimen-
467 tal and simulated fluxes are comparable, which is consistent with the results in Section 3.3.2. In
468 summary, even though reactions in the endogenous modules must have flexible metabolic flux ca-
469 pacities to enable a universal modular cell to be compatible with various exchangeable production
470 modules, they must also operate within in a narrow flux range when interfacing with a specific
471 production module.

472 **4 Conclusions**

473 Modular cell design seeks to accelerate strain development towards broader biotechnological appli-
474 cation of synthetic biology and metabolic engineering, similar to the proven advantages of modular
475 design in conventional engineering disciplines.⁴ In this study, we adapted the recently proposed¹³
476 multi-objective modular strain design method to a MILP computational framework that can guar-

477 antee Pareto optimal solutions, exhaustively search the space of alternative solutions, and specify
478 design goals such as module prioritization. Remarkably, the proposed method identified a universal
479 modular cell that harnesses the inherent modularity and flexibility of native *E. coli* metabolism^{64,65}
480 to properly interface with a variety of biochemically diverse heterologous pathways. This universal
481 design is predicted to display a growth-coupled to product formation phenotype for all pathways,
482 enabling its use as a platform for pathway optimization through high-throughput library selec-
483 tion or adaptation. The feasibility of this universal design strategy is found to be consistent with
484 experimental evidence of isolated metabolic engineering strategies towards target products and
485 measured intracellular flux ranges. We anticipate this is the first example of upcoming method-
486 ological developments in the multi-objective strain design approach, which will follow a path similar
487 to single-phenotype strain design algorithms⁶⁶ introduced in the early 2000s,¹⁸ including the ad-
488 dition of heterologous metabolic reactions from large biochemical databases⁶⁷ and up- and down-
489 regulation of genes in addition to knock-outs⁶⁸, as well as the use of alternative modeling paradigms
490 for flux prediction such as kinetic models⁶⁹ and ME-models.⁷⁰ Additionally, we anticipate that the
491 method developed in this study can be applied to exchangeable metabolic modules whose functions
492 can be expanded to bioremediation⁷¹ and biosensing^{72,73}.

493 Acknowledgments

494 This research was funded by the NSF CAREER Award (NSF#1553250) and the Center of Bioen-
495 ergy Innovation (CBI), U.S. Department of Energy Bioenergy Research Center supported by the
496 Office of Biological and Environmental Research in the DOE Office of Science. The funders had
497 no role in the study design, data collection and analysis, decision to publish, or preparation of the
498 manuscript.

499 5 Definitions

500 Sets

501 \mathcal{I}_k Metabolites in production network k .

502 \mathcal{J}_k Reactions in production network k .

- 503 \mathcal{K} Production networks that are derived from a combination of the parent metabolic network
 504 with the metabolic pathways associated with production modules. The parent metabolic
 505 network is the network of the host strain that is genetically manipulated to build a modular
 506 cell chassis.
- 507 \mathcal{M} Metabolic states that correspond to the growth phase, denoted μ , and the non-growth or
 508 stationary phase, denoted $\bar{\mu}$.
- 509 \mathcal{C} Candidate deletion reaction set. The removal of these reactions are applied to all production
 510 networks, $\mathcal{C} \subseteq \mathcal{J}^{\text{parent}} \subseteq \mathcal{J}_k, \forall k \in \mathcal{K}$.
- 511 \mathcal{N}_k Non-targeted deletion reaction set in production network k . This set arises from the use of
 512 fixed endogenous module reactions z_{jk} in certain production networks.

513 Binary variables

- 514 y_j Reaction deletion indicator that takes a value of 0 if reaction j is deleted in the chassis and
 515 1 otherwise.
- 516 z_{jk} Endogenous module reaction indicator that takes a value of 1 if reaction j is added back as
 517 module reaction in production network k and 0 otherwise.
- 518 d_{jk} Reaction activity indicator that takes a value of 0 if reaction j in production network k
 519 might not carry a flux and 0 otherwise, thus $d_{jk} = y_j \vee z_{jk}$. This variable is declared as
 520 a continuous and linear constraints enforce the OR relation and thus makes the variable
 521 binary.
- 522 w_k Production network feasibility indicator that takes a value of 0 if reaction deletions are
 523 ignored and the objective value is set to 0 for production network k , and a value of 1
 524 otherwise.
- 525 e_{jk} Reaction activity indicator adjusted to w_k that takes the value of d_{jk} if $w_k = 1$ and a value
 526 of 1 if $w_k = 0$, thus $e_{jk} = (d_{jk} \wedge w_k) \vee \neg w_k$.
- 527 r_{jk} Linearization variable, $r_{jk} = d_{jk} \vee w_k$.

528 Continuous variables

- 529 v_{jkm} Flux (mmol/gCDW/hr) of reaction j from network k at metabolic state m .
- 530 v_{Pkm} Flux (mmol/gCDW/hr) of product synthesis reaction from network k at metabolic state m .
- 531 v_{Xkm} Flux (mmol/gCDW/hr) of biomass synthesis reaction from network k at metabolic state m .
- 532 f_k General objective function for production network k that can be represented by f_k^{wGCP} ,
 533 f_k^{lsGCP} , or f_k^{NGP} .
- 534 f'_k Objective function adjusted by w_k such that $f'_k = f_k$ if $w_k = 1$ and $f'_k = 0$ otherwise.
- 535 δ_k^+ Amount required by the objective value f'_k to attain the target goal g_k , i.e.. $\delta_k^+ = g_k - f_k$ if
 536 $f'_k < g_k$.
- 537 δ_k^- Amount that the objective value f'_k surpasses the target goal g_k , i.e., $\delta_k^- = f'_k - g_k$ if $f'_k > g_k$.
- 538 λ_{ikm} Dual variable associated with mass balance constraint of metabolite i from production net-
 539 work k at growth state m .
- 540 μ_{jkm}^l Dual variable associated with the lower bound of reaction j from production network k at
 541 growth state m .
- 542 μ_{jkm}^u Dual variable associated with the upper bound of reaction j from production network k at
 543 growth state m .
- 544 p_{jkm}^l Linearization variable, $p_{jkm}^l = e_{jk} \mu_{jkm}^l$.
- 545 p_{jkm}^u Linearization variable, $p_{jkm}^u = e_{jk} \mu_{jkm}^u$.

546 Parameters

547	S_{ijk}	Stoichiometric coefficient of metabolite i in reaction j of production network k .
548	l_{jkm}	Lower bound for reaction j of production network k at metabolic state m .
549	u_{jkm}	Upper bound for reaction j of production network k at metabolic state m .
550	γ	Minimum biomass synthesis rate required for growth states. Note that in this study a
551		conservative value of 20% of the maximum predicted growth rate of the wild-type strain
552		was used to generate all results.
553	α	Maximum number of deleted reactions in the modular cell chassis.
554	β_k	Maximum number of endogenous module reactions in production network k .
555	ϵ	Small scalar used for tilting the biomass objective function, leading to the minimum product
556		rate available at the maximum growth rate. Note that in our study $\epsilon = 0.0001$ was used
557		to generate all results.
558	$b_\mu, b_{\bar{\mu}}$	Weights on the growth and non-growth objectives of f_k^{lsGCP} , respectively. Note that in our
559		study $b_\mu = 1$ and $b_{\bar{\mu}} = 10$ were used to generate all results.
560	a_k	Weighting factor applied to the objective function for production network k in the blended
561		formulation. Note that in our study $a_k = 1, \forall k \in \mathcal{K}$ was used unless otherwise noted.
562	g_k	Target value for objective f'_k in the goal programming formulation.
563	a_k^+	Weighting factor applied to δ_k^+ which emphasizes the importance of objective value f'_k to
564		avoid falling below the target value g_k . Note that in our study $a_k^+ = 1, \forall k \in \mathcal{K}$ was used
565		in all cases.
566	a_k^-	Weighting factor applied to δ_k^- which emphasizes the importance of the objective f'_k to
567		avoid surpassing the target value g_k . Note that in our study $a_k^- = 1, \forall k \in \mathcal{K}$ was chosen
568		everywhere except to determine the universal modular cell design, where $a_k^- = 0, \forall k \in \mathcal{K}$
569		was used.
570	M^w	Determines the minimum value of f_k that allows w_k to not be 0. A value of 10, corresponding
571		to $f_k \geq 0.01$ for $w_k \neq 0$, was used in all cases.
572	M^{obj}	Upper bound for each objective value. Note that in our study a value of 20 was set for all
573		cases.
574	M	Upper bound for dual variables. Note that in our study a value of 100 was set for all cases.

575 6 Reaction abbreviations

Identifier	Name
ACACT1r	Acetyl-CoA C-acetyltransferase
ACACT2rpp	Acetate reversible transport via proton symport (periplasm)
ACALD	Acetaldehyde dehydrogenase (acetylating)
ACKr	Acetate kinase
ACLS	Acetolactate synthase
ACtex	Acetate transport via diffusion (extracellular to periplasm)
ALCD2x	Alcohol dehydrogenase (ethanol)
ASPTA	Aspartate transaminase
ASPT	L-aspartase
ATPS4rpp	ATP synthase (four protons for one ATP) (periplasm)
DHAD1	Dihydroxy-acid dehydratase (2,3-dihydroxy-3-methylbutanoate)
ECOAH1	3-hydroxyacyl-CoA dehydratase (3-hydroxybutanoyl-CoA)
576 EDA	2-dehydro-3-deoxy-phosphogluconate aldolase
FUM	Fumarase
HACD1	3-hydroxyacyl-CoA dehydrogenase (acetoacetyl-CoA)
KARA1	Ketol-acid reductoisomerase (2,3-dihydroxy-3-methylbutanoate)
MDH	Malate dehydrogenase
MMCD	Methylmalonyl-CoA decarboxylase
MMM	Methylmalonyl-CoA mutase
PDH	Pyruvate dehydrogenase
PFL	Pyruvate formate lyase
PPC	Phosphoenolpyruvate carboxylase
PTAr	Phosphotransacetylase
SUCctex	Succinate transport via diffusion (extracellular to periplasm)
SUCOAS	Succinyl-CoA synthetase (ADP-forming)
THD2pp	NAD(P) transhydrogenase (periplasm)

577 References

- 578 1. Lee, S. Y. *et al.* A comprehensive metabolic map for production of bio-based chemicals.
579 *Nature Catalysis* **2**, 18 (2019).
- 580 2. Nielsen, J. & Keasling, J. D. Engineering Cellular Metabolism. *Cell* **164**, 1185–1197
581 (2016).
- 582 3. Trinh, C. T. & Mendoza, B. Modular cell design for rapid, efficient strain engineering
583 toward industrialization of biology. *Current Opinion in Chemical Engineering* **14**, 18–
584 25 (2016).
- 585 4. Garcia, S. & Trinh, C. T. Modular design: Implementing proven engineering principles
586 in biotechnology. *Biotechnology Advances* (2019).
- 587 5. Coello, C. A. C. & Lamont, G. B. *Applications of multi-objective evolutionary algo-*
588 *rithms* (World Scientific, Singapore, 2004).
- 589 6. Rangaiah, G. P. *Multi-objective optimization: techniques and applications in chemical*
590 *engineering* (World Scientific, Singapore, 2009).

- 591 7. Kitano, H. Biological robustness. *Nature Reviews Genetics* **5**, 826 (2004).
- 592 8. Kashtan, N., Noor, E. & Alon, U. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences* **104**, 13711–13716 (2007).
- 593
- 594 9. Clune, J., Mouret, J.-B. & Lipson, H. The evolutionary origins of modularity. *Proc. R. Soc. B* **280**, 20122863 (2013).
- 595
- 596 10. Shoal, O. *et al.* Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science*, 1217405 (2012).
- 597
- 598 11. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336**, 601–604 (2012).
- 599
- 600 12. Helmer, R., Yassine, A. & Meier, C. Systematic module and interface definition using component design structure matrix. *Journal of Engineering Design* **21**, 647–675 (2010).
- 601
- 602 13. Garcia, S. & Trinh, C. T. Multiobjective strain design: A framework for modular cell engineering. *Metabolic Engineering* **51** (2019).
- 603
- 604 14. Garcia, S. & Trinh, C. T. Comparison of Multi-Objective Evolutionary Algorithms to Solve the Modular Cell Design Problem for Novel Biocatalysis. *Processes* **7** (2019).
- 605
- 606 15. Zhou, A. *et al.* Multiobjective evolutionary algorithms: A survey of the state of the art. *Swarm and Evolutionary Computation* **1**, 32–49 (2011).
- 607
- 608 16. Deb, K., Pratap, A., Agarwal, S. & Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation* **6**, 182–197 (2002).
- 609
- 610
- 611 17. Trinh, C. T., Liu, Y. & Conner, D. J. Rational design of efficient modular cells. *Metabolic engineering* **32**, 220–231 (2015).
- 612
- 613 18. Burgard, A. P., Pharkya, P. & Maranas, C. D. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnology and bioengineering* **84**, 647–657 (2003).
- 614
- 615
- 616 19. Klamt, S. & Mahadevan, R. On the feasibility of growth-coupled product synthesis in microbial strains. *Metabolic engineering* **30**, 166–178 (2015).
- 617
- 618 20. Palsson, B. Ø. *Systems biology: constraint-based reconstruction and analysis* (Cambridge University Press, United Kingdom, 2015).
- 619
- 620 21. Feist, A. M. *et al.* Model-driven evaluation of the production potential for growth-coupled products of Escherichia coli. *Metabolic engineering* **12**, 173–186 (2010).
- 621
- 622 22. Maranas, C. D. & Zomorrodi, A. R. *Optimization Methods in Metabolic Networks* (John Wiley & Sons, Hoboken, New Jersey, 2016).
- 623
- 624 23. Marler, R. T. & Arora, J. S. Survey of multi-objective optimization methods for engineering. *Structural and multidisciplinary optimization* **26**, 369–395 (2004).
- 625
- 626 24. Monk, J. M. *et al.* iML1515, a knowledgebase that computes Escherichia coli traits. *Nature biotechnology* **35**, 904 (2017).
- 627
- 628 25. Akita, H., Nakashima, N. & Hoshino, T. Pyruvate production using engineered Escherichia coli. *AMB Express* **6**, 94 (2016).
- 629

- 630 26. Atsumi, S., Hanai, T. & Liao, J. C. Non-fermentative pathways for synthesis of branched-
631 chain higher alcohols as biofuels. *Nature* **451**, 86 (2008).
- 632 27. Layton, D. S. & Trinh, C. T. Engineering modular ester fermentative pathways in
633 *Escherichia coli*. *Metabolic Engineering* **26**, 77–88 (2014).
- 634 28. Niu, D. *et al.* Highly efficient L-lactate production using engineered *Escherichia coli*
635 with dissimilar temperature optima for L-lactate formation and cell growth. *Microbial*
636 *cell factories* **13**, 78 (2014).
- 637 29. Rodriguez, G. M., Tashiro, Y. & Atsumi, S. Expanding ester biosynthesis in *Escherichia*
638 *coli*. *Nature Chemical Biology* **10**, 259–265 (2014).
- 639 30. Shen, C. R. *et al.* Driving Forces Enable High-Titer Anaerobic 1-Butanol Synthesis in
640 *Escherichia coli*. *Applied and Environmental Microbiology* **77**, 2905–2915 (2011).
- 641 31. Trinh, C. T., Unrean, P. & Sreenc, F. Minimal *Escherichia coli* Cell for the Most Ef-
642 ficient Production of Ethanol from Hexoses and Pentoses. *Applied and Environmental*
643 *Microbiology* **74**, 3634–3643 (2008).
- 644 32. Tseng, H.-C. & Prather, K. L. Controlled biosynthesis of odd-chain fuels and chemicals
645 via engineered modular metabolic pathways. *Proceedings of the National Academy of*
646 *Sciences*, 201209002 (2012).
- 647 33. Yim, H. *et al.* Metabolic engineering of *Escherichia coli* for direct production of 1,
648 4-butanediol. *Nature chemical biology* **7**, 445–452 (2011).
- 649 34. Yu, J., Xia, X., Zhong, J. & Qian, Z. Direct biosynthesis of adipic acid from a synthetic
650 pathway in recombinant *Escherichia coli*. *Biotechnology and bioengineering* **111**, 2580–
651 2586 (2014).
- 652 35. Von Kamp, A. & Klamt, S. Growth-coupled overproduction is feasible for almost
653 all metabolites in five major production organisms. *Nature communications* **8**, 15956
654 (2017).
- 655 36. Hart, W. E. *et al.* *Pyomo — Optimization Modeling in Python* (Springer International
656 Publishing, Cham, 2017).
- 657 37. Machado, D. & Herrgård, M. Systematic evaluation of methods for integration of tran-
658 scriptomic data into constraint-based models of metabolism. *PLoS Comput Biol* **10**,
659 e1003580 (2014).
- 660 38. Lewis, N. E. *et al.* Omic data from evolved *E. coli* are consistent with computed optimal
661 growth from genome-scale models. *Molecular systems biology* **6**, 390 (2010).
- 662 39. Nocedal, J. & Wright, S. *Numerical optimization* (Springer Science & Business Media,
663 United States of America, 2006).
- 664 40. Kaufman, D. E. & Smith, R. L. Direction choice for accelerated convergence in hit-and-
665 run sampling. *Operations Research* **46**, 84–95 (1998).
- 666 41. Heirendt, L. *et al.* Creation and analysis of biochemical constraint-based models: the
667 COBRA Toolbox v3. 0. *arXiv preprint arXiv:1710.04038* (2017).

- 668 42. King, Z. A. *et al.* Escher: a web application for building, sharing, and embedding data-
669 rich visualizations of biological pathways. *PLoS computational biology* **11**, e1004321
670 (2015).
- 671 43. Geoffrion, A. M. Generalized benders decomposition. *Journal of optimization theory*
672 *and applications* **10**, 237–260 (1972).
- 673 44. Fischetti, M., Ljubić, I. & Sinnl, M. Benders decomposition without separability: A
674 computational study for capacitated facility location problems. *European Journal of*
675 *Operational Research* **253**, 557–569 (2016).
- 676 45. Von Kamp, A. & Klamt, S. Enumeration of Smallest Intervention Strategies in Genome-
677 Scale Metabolic Networks. *PLOS Computational Biology* **10**, 1–13 (2014).
- 678 46. King, Z. A., O’Brien, E. J., Feist, A. M. & Palsson, B. O. Literature mining supports
679 a next-generation modeling approach to predict cellular byproduct secretion. *Metabolic*
680 *Engineering* **39**, 220–227 (2017).
- 681 47. Fong, S. S. *et al.* In silico design and adaptive evolution of Escherichia coli for production
682 of lactic acid. *Biotechnology and bioengineering* **91**, 643–648 (2005).
- 683 48. Trinh, C. & Sreenc, F. Metabolic engineering of Escherichia coli for efficient conversion
684 of glycerol to ethanol. *Appl Environ Microbiol* **75**, 6696–6705 (2009).
- 685 49. Garst, A. D. *et al.* Genome-wide mapping of mutations at single-nucleotide resolution
686 for protein, metabolic and genome engineering. *Nature biotechnology* **35**, 48 (2017).
- 687 50. Nielsen, D. R. *et al.* Engineering alternative butanol production platforms in heterolo-
688 gous bacteria. *Metabolic engineering* **11**, 262–273 (2009).
- 689 51. Shi, A., Zhu, X., Lu, J., Zhang, X. & Ma, Y. Activating transhydrogenase and NAD
690 kinase in combination for improving isobutanol production. *Metabolic engineering* **16**,
691 1–10 (2013).
- 692 52. Hädicke, O., Bettenbrock, K. & Klamt, S. Enforced ATP futile cycling increases specific
693 productivity and yield of anaerobic lactate production in Escherichia coli. *Biotechnology*
694 *and bioengineering* **112**, 2195–2199 (2015).
- 695 53. Khodayari, A. & Maranas, C. D. A genome-scale Escherichia coli kinetic metabolic
696 model k-ecoli457 satisfying flux data for multiple mutant strains. *Nature Communica-*
697 *tions* **7** (2016).
- 698 54. Ishii, N. *et al.* Multiple high-throughput analyses monitor the response of E. coli to
699 perturbations. *Science* **316**, 593–597 (2007).
- 700 55. Kabir, M. M., Ho, P. Y. & Shimizu, K. Effect of ldhA gene deletion on the metabolism
701 of Escherichia coli based on gene expression, enzyme activities, intracellular metabolite
702 concentrations, and metabolic flux distribution. *Biochemical Engineering Journal* **26**,
703 1–11 (2005).
- 704 56. Zhao, J., Baba, T., Mori, H. & Shimizu, K. Global metabolic response of Escherichia coli
705 to gnd or zwf gene-knockout, based on 13 C-labeling experiments and the measurement
706 of enzyme activities. *Applied microbiology and biotechnology* **64**, 91–98 (2004).

- 707 57. Zhao, J. & Shimizu, K. Metabolic flux analysis of Escherichia coli K12 grown on ¹³C-
708 labeled acetate and glucose using GC-MS and powerful flux calculation method. *Journal*
709 *of biotechnology* **101**, 101–117 (2003).
- 710 58. Zhou, S., Causey, T., Hasona, A., Shanmugam, K. & Ingram, L. Production of optically
711 pure D-lactic acid in mineral salts medium by metabolically engineered Escherichia coli
712 W3110. *Appl. Environ. Microbiol.* **69**, 399–407 (2003).
- 713 59. Causey, T., Shanmugam, K., Yomano, L. & Ingram, L. Engineering Escherichia coli
714 for efficient conversion of glucose to pyruvate. *Proceedings of the National Academy of*
715 *Sciences* **101**, 2235–2240 (2004).
- 716 60. Peng, L., Arauzo-Bravo, M. J. & Shimizu, K. Metabolic flux analysis for a ppc mutant
717 Escherichia coli based on ¹³C-labelling experiments together with enzyme activity as-
718 says and intracellular metabolite measurements. *FEMS Microbiology Letters* **235**, 17–
719 23 (2004).
- 720 61. Siddiquee, K. A. Z., Arauzo-Bravo, M. & Shimizu, K. Metabolic flux analysis of pykF
721 gene knockout Escherichia coli based on ¹³C-labeling experiments together with mea-
722 surements of enzyme activities and intracellular metabolite concentrations. *Applied mi-*
723 *crobiology and biotechnology* **63**, 407–417 (2004).
- 724 62. Kim, P., Laivenieks, M., Vieille, C. & Zeikus, J. G. Effect of overexpression of Acti-
725 nobacillus succinogenes phosphoenolpyruvate carboxykinase on succinate production in
726 Escherichia coli. *Appl. Environ. Microbiol.* **70**, 1238–1241 (2004).
- 727 63. Lin, H., Vadali, R. V., Bennett, G. N. & San, K.-Y. Increasing the acetyl-CoA pool in
728 the presence of overexpressed phosphoenolpyruvate carboxylase or pyruvate carboxylase
729 enhances succinate production in Escherichia coli. *Biotechnology progress* **20**, 1599–1604
730 (2004).
- 731 64. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical
732 organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
- 733 65. Noor, E., Eden, E., Milo, R. & Alon, U. Central carbon metabolism as a minimal
734 biochemical walk between precursors for biomass and energy. *Molecular cell* **39**, 809–
735 820 (2010).
- 736 66. Machado, D. & Herrgård, M. J. Co-evolution of strain design methods based on flux
737 balance and elementary mode analysis. *Metabolic Engineering Communications* **2**, 85–
738 92 (2015).
- 739 67. Pharkya, P., Burgard, A. P. & Maranas, C. D. OptStrain: a computational framework
740 for redesign of microbial production systems. *Genome research* **14**, 2367–2376 (2004).
- 741 68. Pharkya, P. & Maranas, C. D. An optimization framework for identifying reaction
742 activation/inhibition or elimination candidates for overproduction in microbial systems.
743 *Metabolic engineering* **8**, 1–13 (2006).
- 744 69. Chowdhury, A., Zomorodi, A. R. & Maranas, C. D. k-OptForce: integrating kinetics
745 with flux balance analysis for strain design. *PLoS computational biology* **10**, e1003487
746 (2014).

- 747 70. Dinh, H. V., King, Z. A., Palsson, B. O. & Feist, A. M. Identification of growth-
748 coupled production strains considering protein costs and kinetic variability. *Metabolic*
749 *engineering communications* **7**, e00080 (2018).
- 750 71. De Lorenzo, V. Systems biology approaches to bioremediation. *Current opinion in*
751 *biotechnology* **19**, 579–589 (2008).
- 752 72. Meyer, A. J., Segall-Shapiro, T. H., Glassey, E., Zhang, J. & Voigt, C. A. Escherichia coli
753 “Marionette” strains with 12 highly optimized small-molecule sensors. *Nature chemical*
754 *biology*, 1 (2018).
- 755 73. Fernandez-Rodriguez, J., Moser, F., Song, M. & Voigt, C. A. Engineering RGB color
756 vision into Escherichia coli. *Nature chemical biology* **13**, 706 (2017).
- 757 74. King, Z. A. *et al.* BiGG Models: A platform for integrating, standardizing and sharing
758 genome-scale models. *Nucleic acids research* **44**, D515–D522 (2015).

759 Tables

Table 1: Solution time reduction by tuning the ModCell2-MILP formulation with Benders decomposition, bound tightening, and/or fixed network indicator ($w_k = 1, \forall k \in \mathcal{K}$). The simulations were performed in triplicates.

Feasibility indicator w_k fixed	Benders decomposition	Bounds tightened	Solution time (s)
No	No	No	63.3 ± 16.9
No	No	Yes	32.5 ± 10.2
No	Yes	No	3.6 ± 0.1
No	Yes	Yes	3.4 ± 0.4
Yes	No	No	13.8 ± 2.7
Yes	No	Yes	11.9 ± 1.7
Yes	Yes	No	2.7 ± 0.3
Yes	Yes	Yes	2.8 ± 0.1

Table 2: Overall production module stoichiometries, degree of reduction (DoR) of the final product (mol e^- / mol C), and metabolite secretion profiles from simulated reference flux distributions (mol C / mol C) of the universal modular cell design. Flux (mmol/gCDW/hr) abbreviations: r_p , product; r_{ac} , acetate; r_{co_2} , CO₂; r_{for} , formate; r_{succ} , succinate. Note that the negative CO₂ fluxes in pyruvate and acetate production networks indicate overall CO₂ uptake enabled by phosphoenolpyruvate carboxylase (PPC).

Overall reaction	DoR	r_p	r_{ac}	r_{co_2}	r_{for}	r_{succ}
pyr + nadh → ethanol accoa + 2 nadh → ethanol (native)	7.0	0.58	0.01	0.27	0.04	-
oaa + glu + 2 atp + 2 nadph + nadh → akG + propanol	6.7	0.31	0.36	0.07	0.18	-
2 accoa + 4 nadh → butanol	6.5	0.59	0.01	0.28	0.04	-
2 pyr + nadph + nadh → isobutanol	6.5	0.62	-	0.31	-	-
oaa + glu + accoa + 3 nadh + 2 atp + 2 nadph → akG + pentanol	6.4	0.50	0.21	0.24	0.03	-
succ + akG + atp + 4 nadh + accoa → ac + 1,4-butanediol	5.5	0.46	0.33	0.17	-	-
→ pyruvate	3.0	0.46	-	-0.16	-	0.66
pyr + nadh → D-lactate	3.7	0.91	-	-	-	-
accoa → atp + acetate	3.5	0.60	0.60	-0.30	0.61	-
accoa + succoa + 2 nadh → atp + adipic acid	4.0	0.82	0.05	0.04	0.06	-
accoa + pyr + nadh → ethyl acetate	5.0	0.63	-	-	0.32	-
accoa + oaa + glu + 2 atp + 2 nadph + nadh → akG + propyl acetate	5.2	0.41	0.30	-	0.24	-
accoa + 2 pyr + nadph + nadh → isobutyl acetate	5.3	0.36	-	0.02	0.06	0.52
2 accoa + 3 nadh + pyr → ethyl butanoate	5.3	0.61	-	0.09	0.23	-
2 accoa + 3 nadh + oaa + glu + 2 atp + 2 nadph → akG + propyl butanoate	5.4	0.68	0.03	0.23	0.04	-
4 accoa + 6 nadh → butyl butanoate	5.5	0.61	-	0.14	0.18	-
2 accoa + 3 nadh + 2 pyr + nadph → isobutyl butanoate	5.5	0.64	-	0.16	0.16	-
oaa + glu + accoa + 2 nadh + 2 atp + 2 nadph + pyr → akG + ethyl pentanoate	5.4	0.68	0.03	0.23	0.04	-
oaa + glu + accoa + 2 nadh + 2 atp + 3 nadph + 2 pyr → akG + isobutyl pentanoate	5.6	0.67	0.01	0.25	0.03	-
2 oaa + 2 glu + 2 accoa + 4 nadh + 4 atp + 4 nadph → 2 akG + pentyl pentanoate	5.6	0.53	0.22	0.20	0.02	-

760 Figures

Figure 1: Principles of modular cell design. (a) Modular cell chassis. (b) Interfaces. (c) Production modules. (d) Production strains. A modular cell is designed to provide the necessary precursors for biosynthesis pathway modules that are independently assembled with the modular cell to generate production strains exhibiting desirable phenotypes. The *wGCP* phenotype, one of the possible design objectives, enforces the coupling between the desirable product synthesis rate and the maximum cellular growth rate.

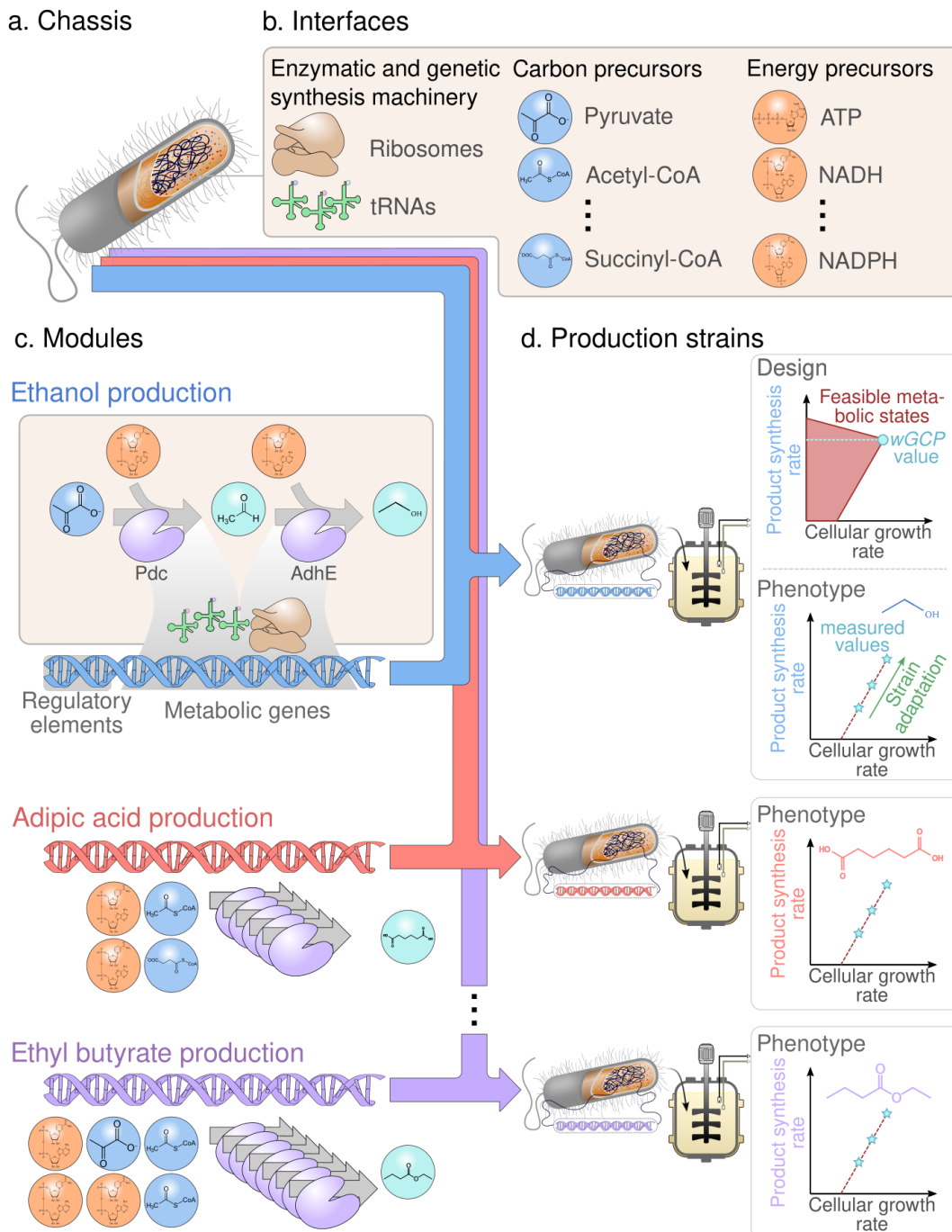
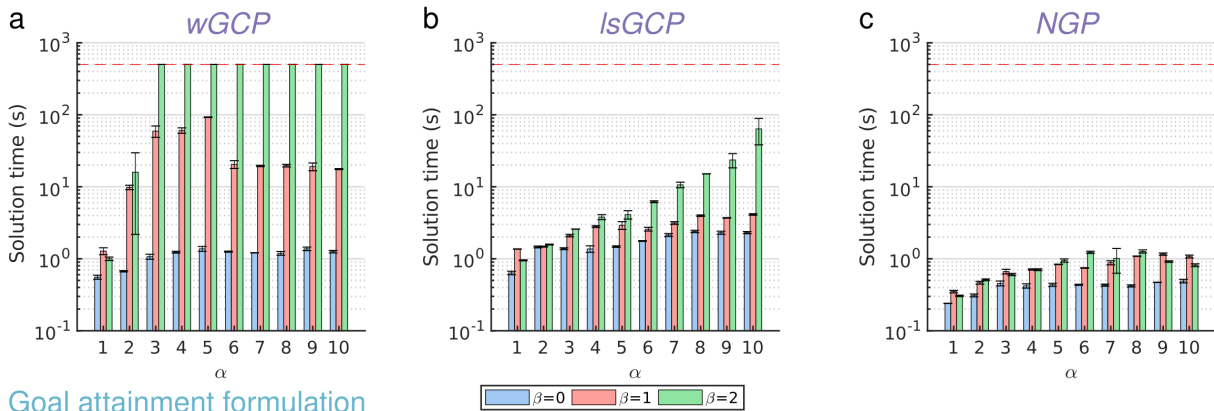


Figure 2: Effect of design parameters, including the target design objective (i.e., $wGCP$, $lsGCP$, and NGP) and the limits of deletion reactions α and endogenous module-specific reactions β_k , on computation time for solving the ModCell2-MILP problem with the blended (a-c) and goal attainment (d-f) formulations. A time limit of 500 seconds indicated by a red dashed line was used in all cases, but only reached by certain $wGCP$ and $lsGCP$ cases with $\beta \geq 2$. The simulations were performed in duplicates.

Blended formulation



Goal attainment formulation

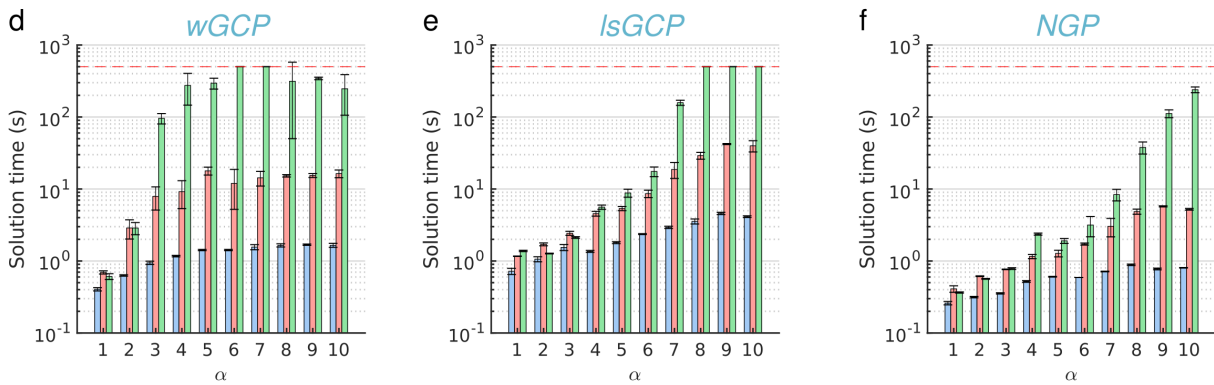


Figure 3: Analysis of reaction deletion candidates. (a) Subsystem distribution for the original set of 276 candidate reactions in the iML1515 model. Those subsystems that contain a reaction used in at least one design are colored. (b) Deletion frequency for the reduced set of 33 candidate reactions. The analysis is based on a pool of 601 *wGCP* designs from different α and β parameters whose Pareto fronts were previously determined with ModCell2.¹³ Bar colors indicate membership of these reactions to the subsystems. (c) Metabolic map of core metabolism. Key metabolites, including precursors for the 20 product modules (i.e., pyruvate, acetyl-CoA, succinyl-CoA, succinate, and α -ketoglutarate), are highlighted in green. Reactions are colored according to subsystem labels indicated in (a), reactions colored in light gray do not appear in any of the subsystems of (a), and reactions that are candidates for deletion, listed in (b), are labeled in red. (d) Link between major precursors and target products where colors are only used to facilitate visualization. Reaction and metabolite abbreviations correspond to BiGG⁷⁴ identifiers (<http://bigg.ucsd.edu/>).

Figure 3: (Caption previous page)

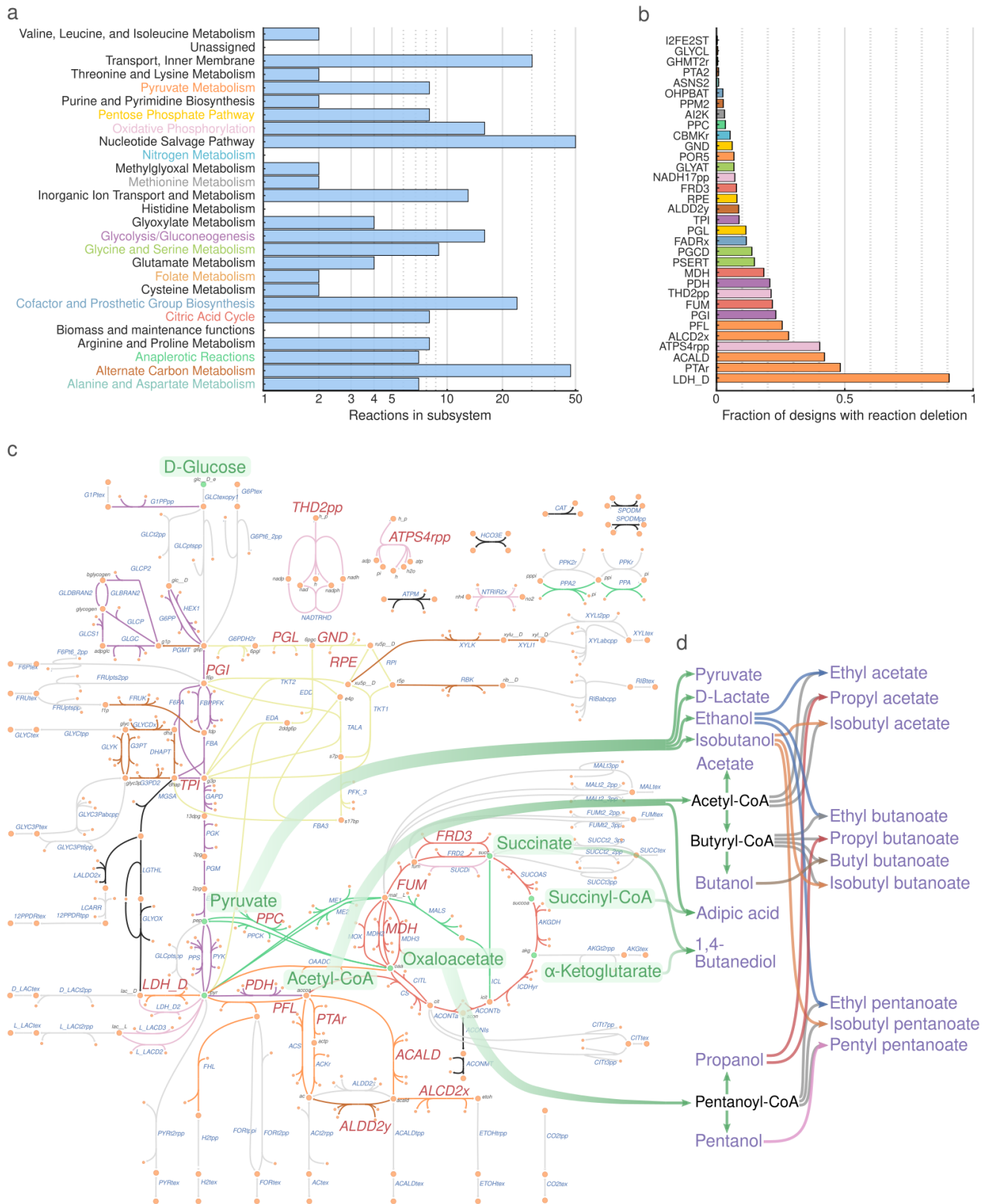


Figure 4: Identification of a universal modular cell compatible with all production modules using the *wGCP* design objective. (a) Goal programming solutions with increasing α and β values. The goal programming objective value (58) in the y-axis measures the difference between the performance of production strains and the target goal, i.e., $\sum_{k \in \{k \in \mathcal{K}: f'_k < g_k\}} (f'_k - g_k)$ where the target goal is set to be $g_k = 0.5$. The parameters $\alpha = 6$ and $\beta = 1$ are sufficient to identify a universal modular cell design meeting the required goal for all production networks. (b) Comparison between the yield performances of the designed modular production strains and maximum theoretical values. (c) The feasible flux spaces for the wild-type (gray) and designed modular production strains (crimson). Based on the *wGCP* design phenotype, to increase growth rate, each mutant must increase product synthesis rate. The genetic manipulations of this universal modular cell design are indicated in the metabolic map of Figure 5c.

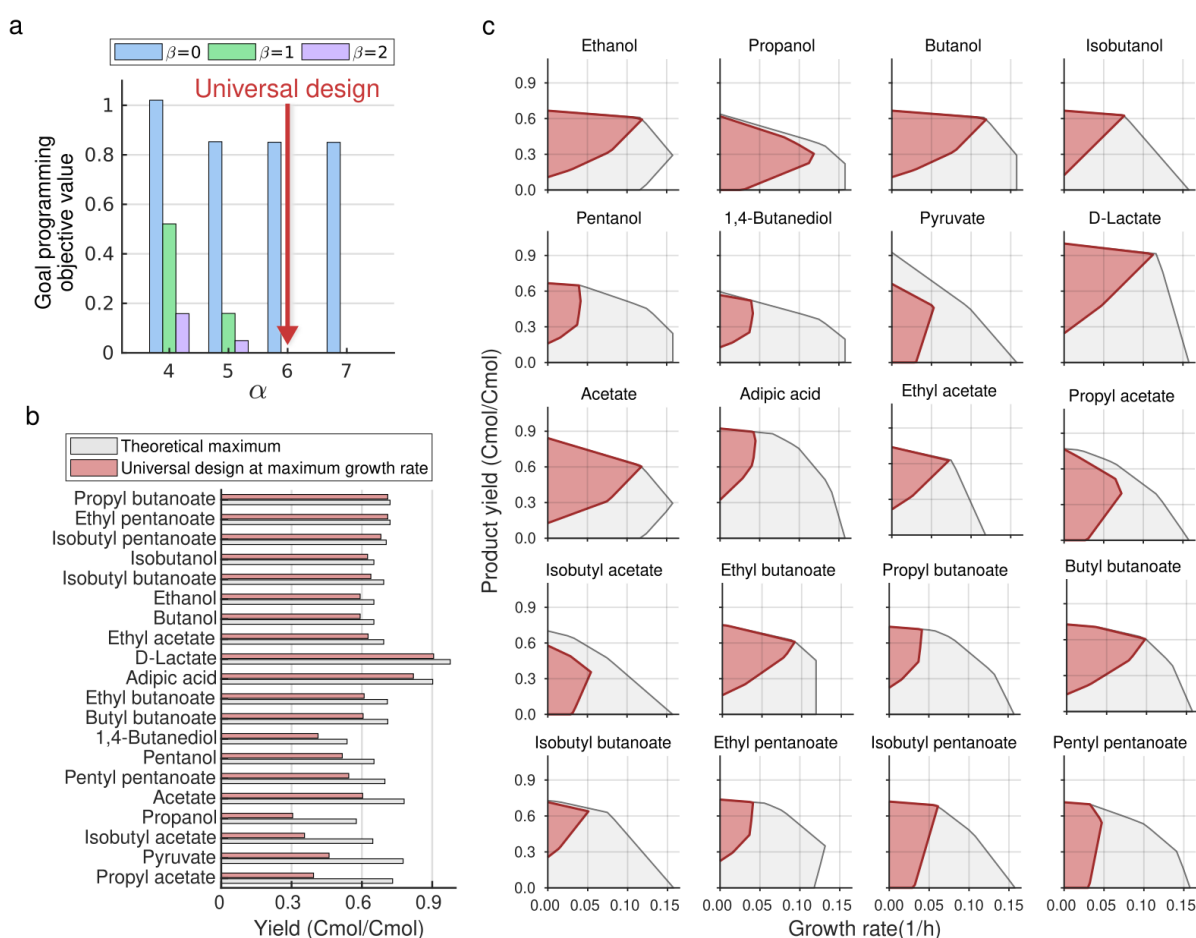


Figure 5: Flexible metabolic flux capacity of *E. coli* metabolism enables the universal modular cell design. (a) Standard deviation of each reaction flux across production networks. (b) Scaled fluxes of the 51 reactions with standard deviation magnitude above 0.2, excluding proton, water transport, and exchange reactions. A scaled flux for a reaction is determined by the reference flux distribution value divided by the maximum value of that reaction across all production networks. Thus, a scaled flux of 0 indicates a given reaction does not carry any flux, and a scaled flux of 1 indicates that this reaction carries the highest flux across production networks. Several columns have multiple reactions, separated by |, since they carry exactly the same flux. (c) Endogenous modules of the universal modular cell. The reactions colored in red are deleted in the chassis, but are used as module reactions in the production networks shown in the adjacent gray boxes. Metabolites in periplasmic and extracellular compartments have “_p” and “_e” suffixed to their abbreviations, respectively. Metabolite and reaction abbreviations follow BiGG⁷⁴ notation. (d) Comparison between simulated and measured fluxes. The solid lines within the “violins” correspond to samples. The simulated fluxes for the reversible reactions, including FUM, LDH, MDH, and ACALD, were multiplied by -1 to reflect their most common direction.

Figure 5: (Caption previous page)

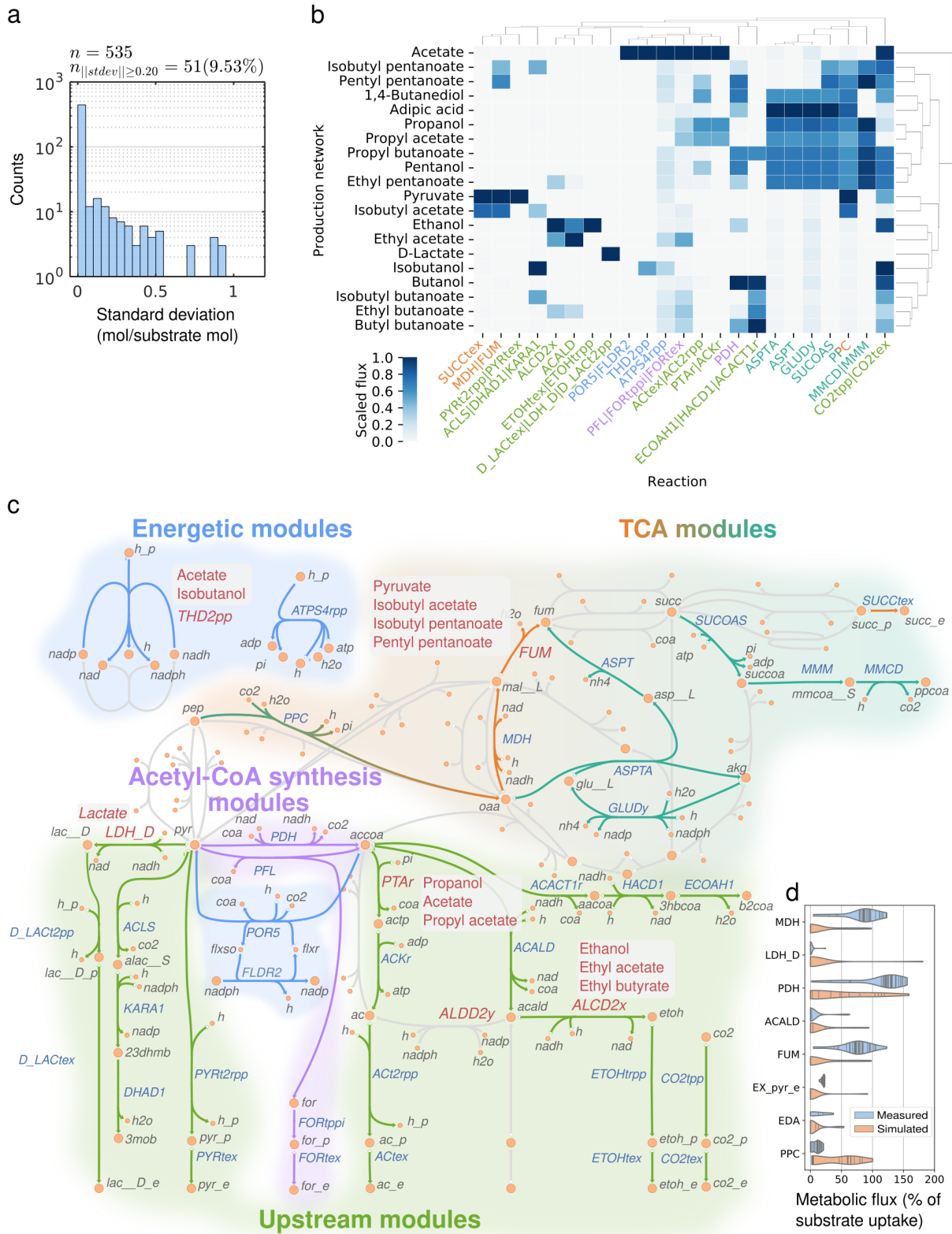
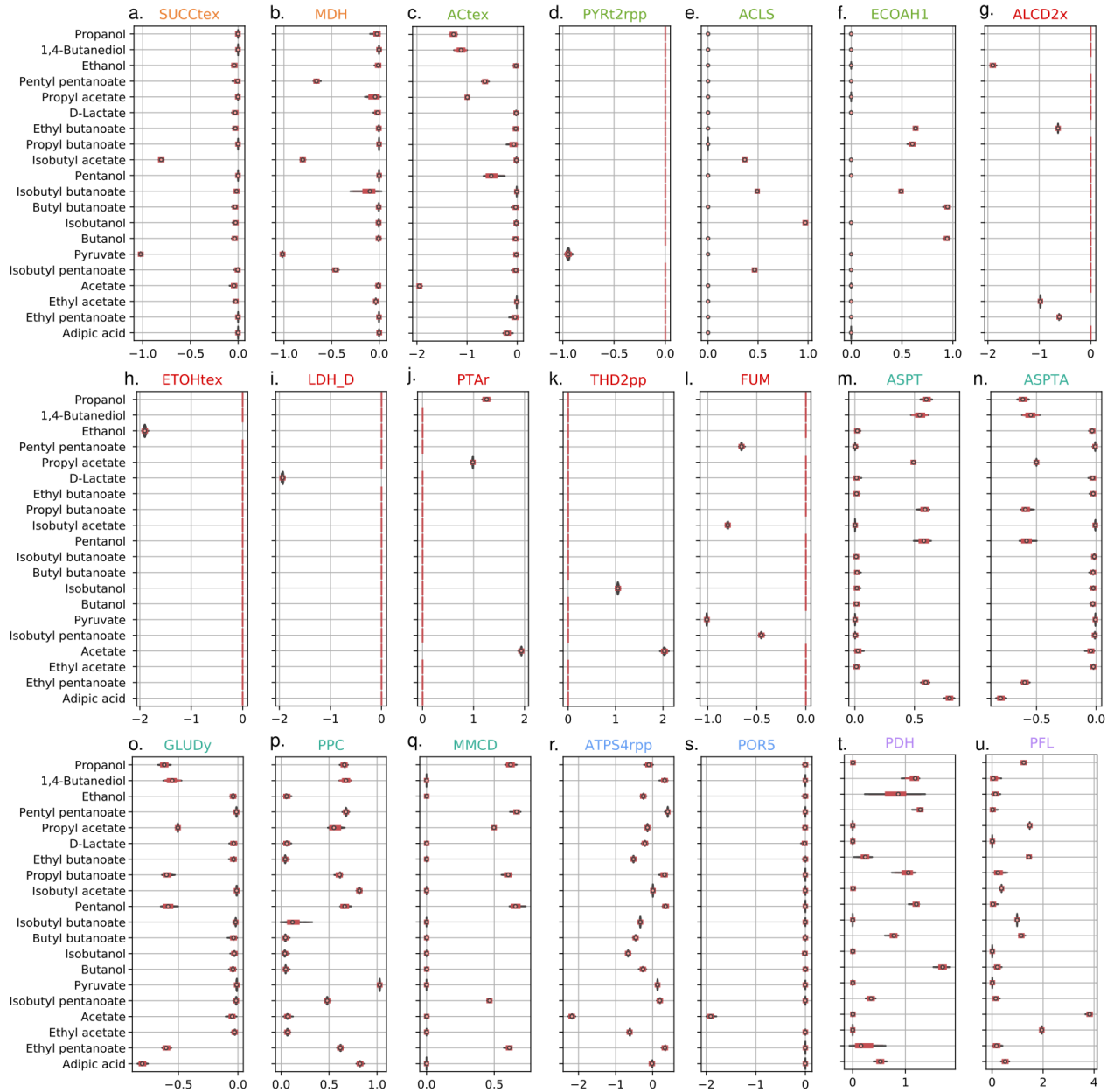


Figure 6: Violin plots of sampled flux distributions of the reactions of interest. Reaction colors are consistent with Figure 5. The flux of SUACOAS could not be sampled since this reaction is involved in a thermodynamically infeasible cycle.



761 **Supplementary Materials**

762 **Supplementary Material 1** Modular cell designs for *E. coli* core model.

763 **Supplementary Material 2** Computer programs used to generate the results of this study.