ACS SyntheticBiology
Research Article
Cite This: ACS Synth. Biol. XXXX, XXX, XXX−XXX
pubs.acs.org/synthbio

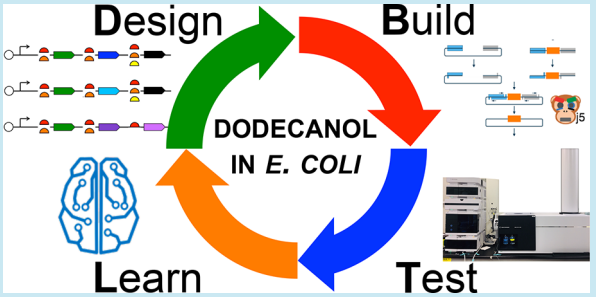
Lessons from Two Design−Build−Test−Learn Cycles of Dodecanol Production in *Escherichia coli* Aided by Machine Learning

Paul Opgenorth,[†,‡,⊗] Zak Costello,[†,‡,§,⊗] Takuya Okada,[∥] Garima Goyal,[†,‡,§] Yan Chen,[†,‡,§] Jennifer Gin,[†,‡,§] Veronica Benites,[†,‡,§] Markus de Raad,[⊥,#] Trent R. Northen,[†,⊥,#] Kai Deng,[¶] Samuel Deutsch,[#] Edward E. K. Baidoo,[†,‡,§] Christopher J. Petzold,[†,‡,§] Nathan J. Hillson,[†,‡,§,#] Hector Garcia Martin,[†,‡,§,▽] and Harry R. Beller*[,†,‡,⊗]

[†]Joint BioEnergy Institute (JBEI), Emeryville, California 94608, United States
[‡]Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States
[§]DOE Agile BioFoundry, Emeryville, California 94608, United States
[∥]Research Institute for Bioscience Product & Fine Chemicals, Ajinomoto Co., Inc., Kawasaki 210-8680, Japan
[⊥]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, United States
[#]DOE Joint Genome Institute, Walnut Creek, California 94598, United States
[¶]Sandia National Laboratories, Livermore, California 94550, United States
[▽]BCAM, Basque Center for Applied Mathematics, 48009 Bilbao, Spain

Supporting Information

ABSTRACT: The Design−Build−Test−Learn (DBTL) cycle, facilitated by exponentially improving capabilities in synthetic biology, is an increasingly adopted metabolic engineering framework that represents a more systematic and efficient approach to strain development than historical efforts in biofuels and biobased products. Here, we report on implementation of two DBTL cycles to optimize 1-dodecanol production from glucose using 60 engineered *Escherichia coli* MG1655 strains. The first DBTL cycle employed a simple strategy to learn efficiently from a relatively small number of strains (36), wherein only the choice of ribosome-binding sites and an acyl-ACP/acyl-CoA reductase were modulated in a single pathway operon including genes encoding a thioesterase (UcFatB1), an acyl-ACP/acyl-CoA reductase (Maqu_2507, Maqu_2220, or Acr1), and an acyl-CoA synthetase (FadD). Measured variables included concentrations of dodecanol and all proteins in the engineered pathway. We used the data produced in the first DBTL cycle to train several machine-learning algorithms and to suggest protein profiles for the second DBTL cycle that would increase production. These strategies resulted in a 21% increase in dodecanol titer in Cycle 2 (up to 0.83 g/L, which is more than 6-fold greater than previously reported batch values for minimal medium). Beyond specific lessons learned about optimizing dodecanol titer in *E. coli*, this study had findings of broader relevance across synthetic biology applications, such as the importance of sequencing checks on plasmids in production strains as well as in cloning strains, and the critical need for more accurate protein expression predictive tools.



KEYWORDS: *DBTL, machine learning, dodecanol, proteomics, synthetic biology*

Although biobased chemicals represent an appealing and more sustainable alternative to traditional petrochemicals, their widespread adoption has been partially stymied by the limited efficacy of strain development. Historically, strain development has not been a wholly systematic enterprise; instead, genes, their expression levels, and the chassis organism itself have often been tested on a trial-and-error basis, typically informed by biochemical intuition, in order to eventually settle on a strain with production metrics suitable for scaleup. This work typically takes millions of dollars and 3−10 years to complete: it took 150 person-years of effort to express the 16-enzyme artemisinin pathway[1] and 575 person-years of effort for DuPont's 1,3-propanediol pathway.[2] The need for these Herculean efforts represents a significant bottleneck in the biobased chemical production pipeline. However, the advent of exponentially improving capabilities in DNA synthesis, genome editing, and high-throughput screening coupled with machine-learning methods opens the door to disruptive new approaches to metabolic engineering.[3−6] These new approaches call for a systematic, product-independent metabolic engineering strat-

Received: January 17, 2019
Published: May 9, 2019

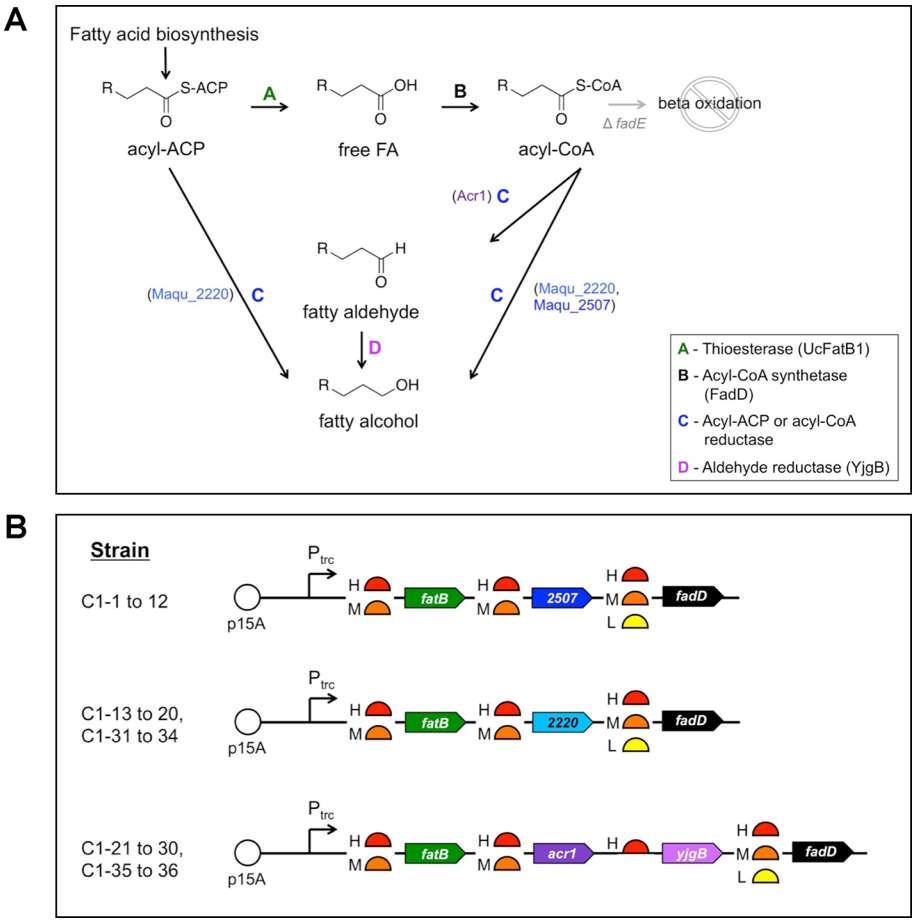


Figure 1. Key metabolic pathways (A) and Cycle 1 design strategies (B) considered for dodecanol biosynthesis in this study. Color coding of enzymes is the same in (A) and (B). Panel (B) shows the combinatorial strategy for 36 Cycle 1 plasmids, all of which have a p15A origin and P_{trc} promoter for a single operon; SBOL Visual[14] symbols are used to represent origins of replication, promoters, and RBSs. The constructs differ by the predicted strength of RBSs (Low, Medium, or High) and the choice of acyl-ACP or acyl-CoA reductase. The host strain in all cases was *E. coli* MG1655 Δ*fadE*. Abbreviated gene names include *fatB* (representing *UcFatB1*), 2507 (representing *Maqu_2507*), and 2220 (representing *Maqu_2220*).

egy that leverages engineering principles. One of those engineering principles is the Design−Build−Test−Learn (DBTL) cycle—a loop used recursively to obtain a design that satisfies the desired specifications. The DBTL cycle represents a framework that helps systematize metabolic engineering and increase its efficacy and generalizability. In this work, we aim to systematically optimize 1-dodecanol production in *E. coli* by performing two DBTL cycles that leverage proteomic data in an attempt to more efficiently guide the strain development process.

Like other medium-chain fatty alcohols, dodecanol is used in a number of commercial applications, including detergents, emulsifiers, lubricants, and cosmetics. The most straightforward way to produce dodecanol biochemically is by reduction of the C_{12} fatty acid, lauric acid, or its ACP (acyl carrier protein) or CoA (coenzyme A) derivatives. Figure 1A displays four enzymes (denoted as A, B, C, D) relevant to making fatty alcohols from fatty acids, or more specifically, from fatty acyl-ACPs or fatty acyl-CoAs.[7] Although technically one could bypass the thioesterase (A) and acyl-CoA synthetase (B) steps, this would lead to fatty alcohols with a chain-length distribution similar to that of the native fatty acids of the host organism, which in *E. coli* would typically maximize at C_{16} or C_{18}, not C_{12}. A thioesterase with a preference for C_{12} acyl-ACPs, such as UcFatB1 (or BTE) from *Umbellularia californica*, has been shown to bias *E. coli* fatty alcohol production to C_{12}.[8,9] Furthermore, acyl-ACP thioesterases have also been shown to effectively deregulate fatty acid production by hydrolyzing acyl-ACPs, which normally stringently regulate acetyl-CoA carboxylase and other key enzymes involved in bacterial fatty acid biosynthesis (ref 7 and references therein). This deregulation of fatty acid biosynthesis leads to markedly increased fatty acid titers. After release of C_{12} fatty acids by the thioesterase, the fatty acids should be activated as acyl-CoAs to make them appropriate substrates for reduction to fatty alcohols; this reaction is catalyzed by an acyl-CoA synthetase (B), which occurs as FadD in *E. coli*. Several enzymes (C, in Figure 1) have been shown to reduce acyl-CoAs to fatty aldehydes (*e.g.*, Acr1 from *Acinetobacter calcoaceticus* or Orf1594 from *Synechococcus elongatus* PCC 7942)[10,11] or directly to fatty alcohols (*e.g.*, Maqu_2220 and Maqu_2507 from *Marinobacter aquaeolei* VT8).[12] Finally, for acyl-CoA reductases that produce fatty aldehydes, an aldehyde reductase (D) is needed to produce the fatty alcohol; one such enzyme is YjgB from *E. coli*, which has been shown to be effective at this reduction.[13]

Over the past decade, a number of studies have reported *n*-fatty alcohol production in *E. coli*, including C_{12} and C_{14}

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

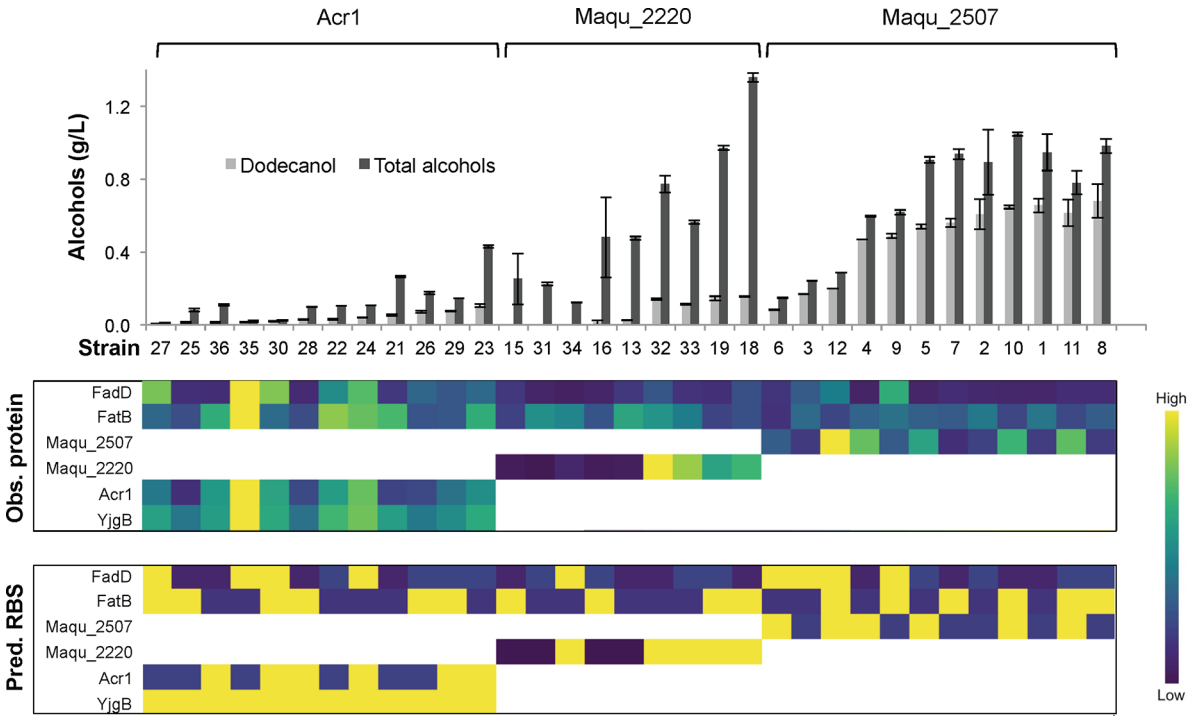


Figure 2. Cycle 1 results for alcohols and targeted proteomics, as well as predicted RBS strength. Means of dodecanol (light gray) and total alcohols (dark gray) are plotted as bars and error bars represent one standard deviation (*n* is typically 3 or 5, and in rare cases 2, biological replicates). Strain names are shown for strains C1−1 to C1−36, but the "C1" prefix has been omitted for brevity. The acyl-ACP/acyl-CoA reductase used in the strains is indicated above the histogram. The upper heat map represents targeted proteomic results observed for pathway enzymes, where the color scale represents normalization to the highest value for each protein. The lower heat map represents predicted Translation Initiation Rates (TIR) for RBSs, with the color scale normalized to the highest value for each protein.

alcohols (*e.g.*, refs 8, 9, 13, 15−18). These studies have used various combinations of the enzymes depicted in Figure 1, among others. Overall, the exploration of combinations of specific enzymes and modulation of their strength of expression (*via* plasmid or chromosomal copy number, promoter strength, or RBS strength) has thus far been limited to a relatively small number; typically, fewer than 10 or 15 strains in a study were compared side-by-side for medium-chain fatty alcohol production. Nonetheless, these studies have clearly demonstrated that modulation of these factors can be very important for fatty alcohol production.

In this study, we aimed to leverage the DBTL cycle and make a more systematic assessment of various enzyme combinations and expression strength to optimize *E. coli* for dodecanol production. We report on implementation of two DBTL cycles to optimize dodecanol production from glucose using 60 engineered *E. coli* MG1655 strains. The first DBTL cycle employed a simple strategy to learn efficiently from a relatively small number of strains (36), wherein only the choice of RBSs and an acyl-ACP/acyl-CoA reductase were modulated in a single pathway operon including genes encoding a thioesterase (UcFatB1), an acyl-ACP/acyl-CoA reductase, and an acyl-CoA synthetase (FadD) (Figure 1B). Measured variables included dodecanol and all proteins in the engineered pathway, which allowed for assessment of the accuracy of RBS strength calculation and the relationship of dodecanol titer to the ensemble composition of pathway proteins. We used the data produced in the first DBTL cycle to train several machine-learning algorithms and to suggest protein profiles for the second DBTL cycle that should increase production. These strategies resulted in a 21% increase in dodecanol titer (up to 0.83 ± 0.125 g/L) in Cycle 2. Beyond specific lessons learned about optimizing dodecanol titer in *E. coli*, this study produced findings of broader relevance across synthetic biology applications, such as the importance of sequencing checks on plasmids in production strains as well as in cloning strains, and the critical need for more accurate protein expression predictive tools.

RESULTS AND DISCUSSION

Cycle 1 Design, Build, And Test: Relationships among Fatty Alcohol Titer, Engineered Pathway Proteins, and RBS Strength. As discussed above, the design strategy for the 36 Cycle-1 strains was combinatorial and modulated between use of three acyl-CoA/acyl-ACP reductases (Maqu_2507, Maqu_2220, or Acr1) as well as different RBS strengths, determined with RBS calculation software,[19−21] for the pathway proteins (Figure 1B). The aim of the design was to have a small number of variables, yet exert sufficient control over key enzymes catalyzing the conversion of acyl-ACPs to dodecanol to effectively inform the machine-learning algorithms.

Certain findings were clear from Cycle 1 regarding the effects of acyl-CoA/acyl-ACP reductases on fatty alcohol titer (Figure 2): (1) the Maqu_2507 reductase performed much better than the other reductases tested with respect to dodecanol titer, and (2) the Maqu_2220 reductase performed the best with regard to total fatty alcohol titer, but not dodecanol titer. Nine of the 12 strains expressing Maqu_2507 (strains C1−1 to C1−12) had dodecanol titers of 0.47 to 0.68 g/L, which were far higher than titers for all 21 strains expressing Maqu_2220 and Acr1 (Figure 2). Some of these

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

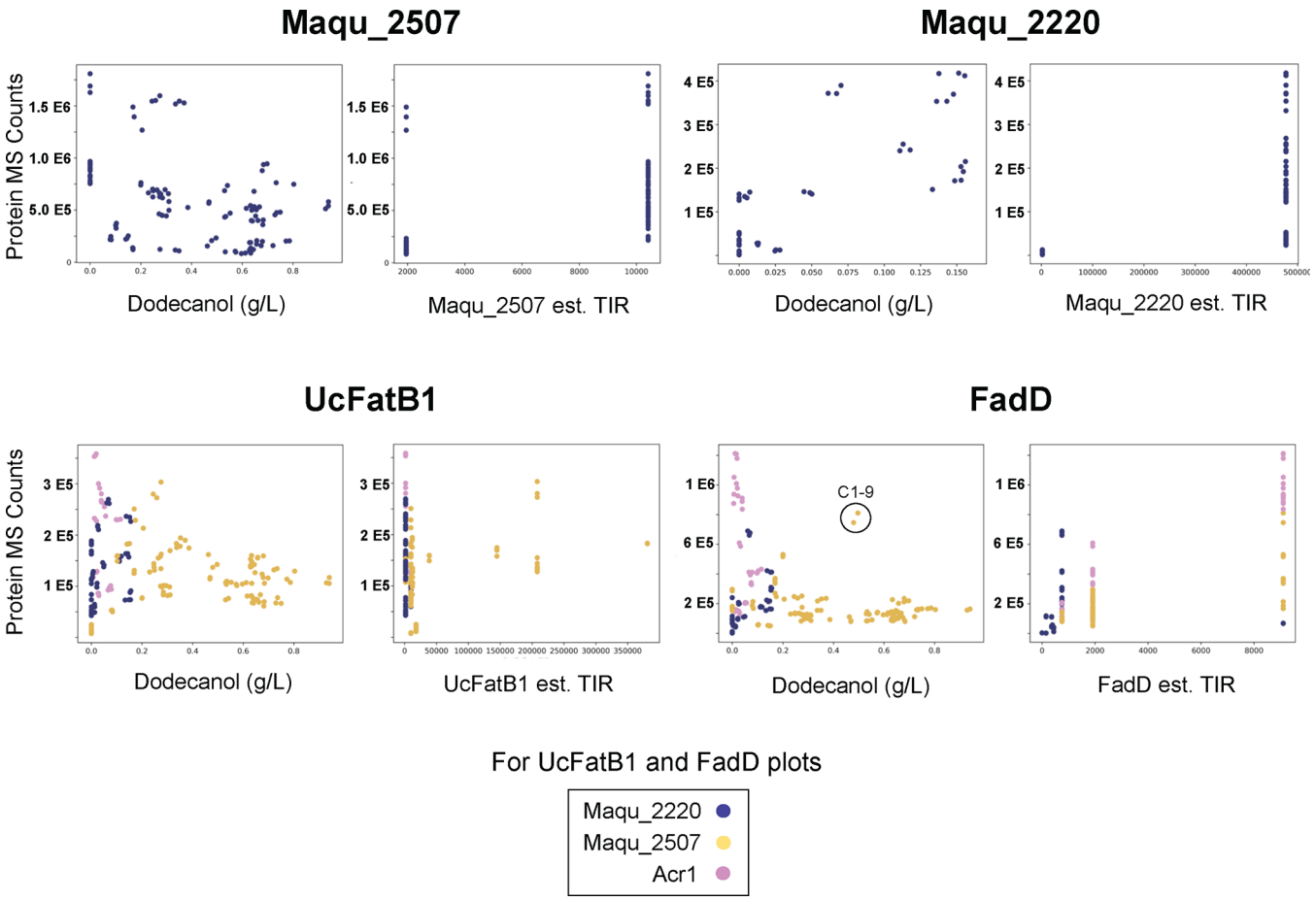
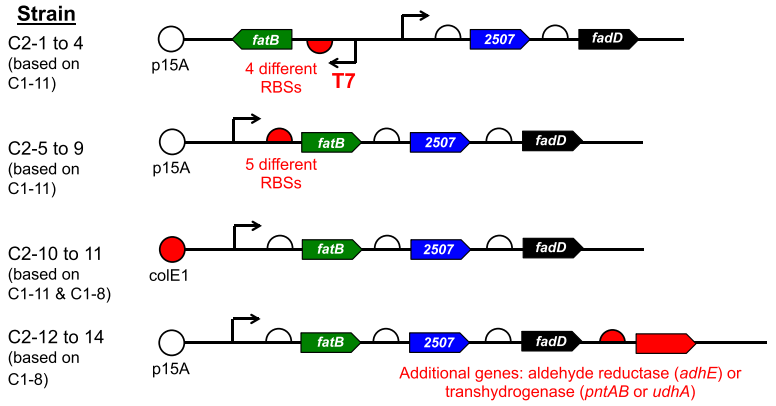


Figure 3. Scatterplots of targeted proteomics results for four pathway proteins (Maqu_2507, Maqu_2220, UcFatB1, and FadD) *vs* dodecanol titer and predicted RBS strength for combined Cycle 1 and Cycle 2 strains. Reductases are color-coded in plots for UcFatB1 and FadD. The highlighted C1−9 data points in the FadD *vs* dodecanol plot are discussed in the text. See Figure 1B for an illustration of the different predicted RBS strengths used in Cycle 1 for each of the pathway enzymes shown.

Maqu_2507 strains, such as C1−11, had both high dodecanol titers and a high proportion of dodecanol relative to total alcohols (*e.g.*, ∼79% for strain C1−11). In contrast, dodecanol titers never exceeded 0.15 g/L in Maqu_2220-expressing strains (C1−13 to −19 and C1−31 to −34), however, total alcohol titers reached ∼1.4 g/L (strain C1−18) and were the highest values observed in the study. Strains expressing the Acr1 acyl-CoA reductase in combination with the YjgB aldehyde reductase (C1−21 to −30, C1−35 to −36) had uniformly low alcohol titers, with only one strain exceeding 0.075 g/L dodecanol and 0.26 g/L total alcohols. Thus, the combination of the UcFatB1 thioesterase, which has been documented to preferentially hydrolyze C_{12}-acyl-ACP thioesters,[22] and the Maqu_2507 reductase, was the most favorable with respect to absolute and relative dodecanol titer. A possible explanation for relatively high total alcohol titers but low dodecanol titers in Maqu_2220-expressing strains relates to the relative activities of UcFatB1, Maqu_2220, and Maqu_2507 on acyl-ACPs in *E. coli*. If Maqu_2220 were as catalytically effective with acyl-ACPs as with acyl-CoAs, which is supported by *in vitro* studies,[23] and also competed effectively with UcFatB1 for acyl-ACPs (Figure 1A), then strains expressing Maqu_2220 would be more likely to have typical fatty acid chain lengths (C_{16}, C_{18}) than C_{12} as fatty alcohol precursors. Furthermore, if Maqu_2507 acted more effectively on acyl-CoAs than acyl-ACPs, and consequently did not compete well with UcFatB1 for acyl-ACPs, then it would allow for more effective acyl chain truncation at C_{12} by the thioesterase.

Overall, there were no easily discernible univariate relationships between dodecanol titer and observed amounts of the various pathway proteins (Figure 2). For example, just expressing more of UcFatB1, the acyl-CoA/acyl-ACP reductase (Acr1, Maqu_2220, or Maqu_2507), and/or FadD did not result in higher alcohol titer. Scatterplots of dodecanol concentrations *vs* pathway-protein amounts (Figure 3) reflect generally poor linear relationships, with the exception of dodecanol *vs* Maqu_2220, which did appear to be positively correlated ($r^2 = 0.60$, $p < 5 \times 10^{-11}$). An interesting trend was observed between FadD and dodecanol, where almost all dodecanol titers greater than 0.2 g/L were associated with relatively low FadD amounts. The unexplained exception to this trend was strain C1−9 (duplicate samples highlighted in Figure 3). The general trend of higher production with lower FadD could be explained by toxicity associated with *fadD* expression, which is reflected in a moderately strong correlation between FadD and glucose in the spent growth medium ($r^2 = 0.65$, $p < 2 \times 10^{-20}$) (Figure S1). As additional evidence that *fadD* expression was associated with toxicity, it is noteworthy that the three strains that could not be built out of the 36 designed Cycle 1 strains (C1−14, −17, and −20) all included the high-strength RBS for *fadD* and contained

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX



Protein target (fold relative to base strain)			Model type and predicted dodecanol titer (g/L)					Success probability (%)
FatB	FadD	Maqu_2507	Ensemble	Random Forest	Polynomial	Neural	TPOT	
12.1	1.49	1.35	1.49	0.43	4.1	1.5	0.55	52
8.98	1.49	1.32	1.32	0.49	3.4	1.3	0.55	49.7
6.60	1.49	1.12	1.09	0.49	2.5	1.2	0.55	46.7

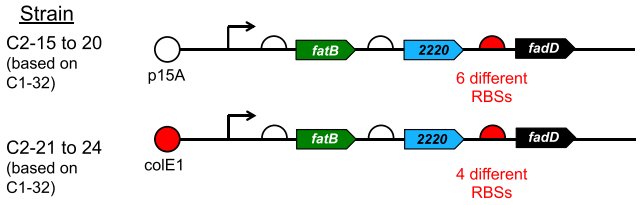
Figure 4. Cycle 2 design strategies for 14 Maqu_2507-expressing strains and the proteomic targets specified by models for dodecanol optimization. In the schematics for the four strategies shown, red is used to highlight variables that have been changed from the Cycle 1 base strain (C1−11 or C1−8). SBOL Visual symbols are used to represent origins of replication, promoters, and RBSs. Tabular information indicates targeted pathway protein expression levels relative to the base strain, along with corresponding dodecanol titers predicted by the four regressor models and ensemble model used. The protein highlighted in red was the primary target for the design strategy. Weight factors in the Ensemble model for the Random Forest, Polynomial Regressor, Neural Regressor, and TPOT models were, respectively, 0.352, 0.208, 0.260, and 0.180. Dodecanol titers for the two base strains, C1−11 and C1−8, were 0.61 ± 0.074 and 0.68 ± 0.092 g/L, respectively. Protein amounts for the base strains can be found in the EDD (Experiment Data Depot) database for this study.

mutations either in this RBS or in the *fadD* gene itself; further, in a preliminary single-gene construct (C1-E7) containing the high-strength RBS for *fadD*, the *fadD* gene was missing entirely (Table S1). Note that historical evidence that *fadD* overexpression could be toxic to strain MG1655[8] prompted us to include a "low-strength" RBS for *fadD* in Cycle 1 Design (Figure 1B). In light of the FadD-dodecanol trend, a similar but less pronounced relationship was observed between UcFatB1 and dodecanol (Figure 3). Here, too, a trend of increasing dodecanol titer with decreasing UcFatB1 was clear. It is possible that the thioesterase-catalyzed truncation of fatty acids at C_{12} and resultant skewing of phospholipid fatty acid chain-length distributions in the cell membrane created stress in the *E. coli* host.

Beyond the lack of strong positive univariate correlations between dodecanol concentrations and pathway protein amounts, there were also no strong correlations between observed pathway protein amounts and predicted expression based on RBS calculation software. This is apparent in a qualitative sense in Figure 2, in which the heat map patterns for observed and predicted protein expression differ substantially. A more quantitative assessment is given in scatterplots of observed pathway protein amounts *vs* predicted Translation Initiation Rates (TIRs[19,20]) (Figure 3) or EMOPEC (Empirical Model and Oligos for Protein Expression Changes) ratings[21] (Figure S2). To more rigorously evaluate the accuracy of RBS prediction software, we compared the strength predictions against protein expression levels while controlling for operon context and estimated plasmid copy number using partial correlation analysis (Figure S3). Using this analysis, it is possible to determine a correlation coefficient for two random variables while controlling for confounding variables. In Figure S3, we plot the relationship between the RBS strength and protein expression after controlling for the confounding variables. These modified variables are called "residuals". The residual EMOPEC calculation results were significantly correlated for 4 of 6 of the conditions tested and the residual TIRs were significantly correlated in 5 of 6 cases tested. However, the correlations, while significant, were frequently weak ($r < 0.5$) or inverse to expectations (*i.e.*, with negative slope).

Cycle 1 Learn: Machine-Learning and Model Building Based on Dodecanol and Pathway Protein Concentrations. Observed Cycle-1 pathway protein concentrations and dodecanol titers were used to predict protein levels expected to maximize dodecanol titer. The basic underlying assumption is that the expression levels of the proteins in the engineered pathway (*e.g.*, UcFatB1, Maqu_2507, and FadD) are sufficiently determinative of final production to guide metabolic engineering efforts, as has been shown previously for certain isoprenoid pathways, where proteomics-based modulation of a heterologous mevalonate pathway resulted in enhanced product titers.[24] For Cycle 1 Learn, predictions were segregated by the acyl-CoA/acyl-ACP reductase used and all replicates of all strains corresponding to each reductase were used as the training sets for modeling (*i.e.*, data for 48 replicates for Maqu_2507 and 24 replicates for Maqu_2220). Four machine-learning regression approaches from the scikit-learn library were applied to fit the data: (1) random forest,[25] (2) polynomial, (3) multilayer perceptron, and (4) the TPOT (Tree-Based Pipeline Optimization Tool) meta-learner.[26,27] For additional details on the first three model implementations, see the relevant scikit-learn documentation.[28] Each model f_i was evaluated using 5-fold cross validation and scored using the mean squared error between predicted and actual production e_i.

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX



Protein target (fold relative to base strain)			Model type and predicted dodecanol titer (g/L)					Success probability (%)
FatB	FadD	Maqu_2220	Ensemble	Random Forest	Polynomial	Neural	TPOT	
2.82	1/2.73	1.04	0.24	0.15	0.35	0.56	0.14	56
1.92	1/10.1	1.04	0.23	0.15	0.32	0.53	0.14	53.9

Figure 5. Cycle 2 design strategies for ten Maqu_2220-expressing strains and the proteomic targets specified by models for dodecanol optimization. In the schematics for the two strategies shown, red is used to highlight variables that have been changed from the Cycle 1 base strain (C1−32). Tabular information indicates targeted pathway protein expression levels relative to the base strain, along with corresponding dodecanol titers predicted by the four regressor models and ensemble model used. The protein highlighted in red was the primary target for the design strategy. Weight factors in the Ensemble model for the Random Forest, Polynomial Regressor, Neural Regressor, and TPOT models were, respectively, 0.330, 0.2105, 0.119, and 0.3395. Protein amounts for the base strains can be found in the EDD (Experiment Data Depot) database for this study.

Each model was chosen for its strengths in particular applications. A model's cross validation performance should be highest when it is most useful in making predictions on the test data set. The random forest algorithm is considered to perform well across a range of problem types; however, it does not extrapolate well.[29] The multilayer perceptron can extrapolate, but typically requires more data to fit its parameters than other approaches. The polynomial regressor was chosen as an informed expectation for the relationship between protein concentration and dodecanol titer to be easily approximated by a low-degree polynomial. The TPOT meta-learner is a meta-learning algorithm that looks for the best cross-validation performance over a range of models.[26,27] More background on the TPOT meta-learner can be found in Olson *et al.*[27] and the extensive information available in its online documentation. In order to make use of the strengths of each model, a composite ensemble model was created from a weighted sum of each of the four models described above (see Methods section). The ensemble model was searched for regions of the proteomic space that would maximize dodecanol titer. To find the best candidate protein amounts, the following optimization problem is then solved,

$$\underset{p}{\operatorname{argmax}} f_c(p)$$

where $f_c(p)$ is the ensemble model that maps protein amount to dodecanol titer. (Note that use of the term protein "amount" here refers not to absolute concentration, which was not measured with peptide standards, but rather to mass spectral counts, which should be linearly related to concentration in the observed range.) The problem is subject to the conditions that the protein amount returned must be positive for each element in the vector and no larger than 1.5 times the maximum observed amount for any given protein. This optimization problem was solved using a differential evolution optimizer implemented by scipy.[30] The optimizer attempts to maximize the likelihood of picking a good candidate target in the next round. This is performed in an iterative fashion, by picking one candidate at a time until the desired number of strain predictions is reached. After a strain is selected, a new constraint is added to the optimization problem, preventing the next strain from being within a chosen radius of the point, so as to promote diversity in suggested protein profiles.

Pathway proteomic targets predicted by the models to increase dodecanol titer are presented in tabular form for Maqu_2507 strains (Figure 4) and for Maqu_2220 strains (Figure 5). Proteomic targets were ranked based on probability of success (as determined by the estimated likelihood of the predicted strain exceeding the highest dodecanol titer observed in Cycle 1; see Methods) and only the top 3 (Figure 4) or top 2 (Figure 5) proteomic targets are presented. The Cycle-1 strains to be used as base strains for engineering to attain the proteomic targets were primarily C1−11 for Maqu_2507 (Figure 4) and C1−32 for Maqu_2220 (Figure 5). These base strains were chosen by finding a Cycle-1 strain that minimized the Euclidean distance to each target in proteomic space.

Cycle 2 Design and Build: Strategies for Optimization of Maqu_2507- and Maqu_2220-Expressing Strains from Cycle 1. The models trained using Cycle 1 dodecanol and proteomic data suggested different optimization strategies for strains utilizing the Maqu_2507 and Maqu_2220 reductases. These design strategies, represented schematically in Figures 4 and 5, were attempts to address the key protein expression targets specified by the models while still taking into account that resource constraints only allowed for 24 total strains in Cycle 2. Ideally, the design strategy for Cycle 2 would have entailed using RBSs to modulate engineered pathway protein expression toward the targets specified by the ensemble models (*i.e.*, the protein targets shown in the tables in Figures 4 and 5). This approach would have been systematic and consistent with the combinatorial RBS design used for the Cycle-1 training set. We did employ an RBS-based strategy; however, the lack of a strong correlation between our protein expression data and predicted RBS strength (Figure 3) presented a formidable challenge for designing Cycle-2 strains. To compensate for this uncertainty, we adopted a strategy that relied on making minimal changes to the dodecanol-pathway operon and maximizing attempts at correct RBS selection. In most cases, this meant that we only changed one RBS per strain while keeping the other two RBSs constant from the Cycle-1 base strain (*i.e.*, C1−11, C1−8, or C1−32); further, we devoted multiple strains to using different RBSs to hit the same protein expression target. For example, strains C2−5 to

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

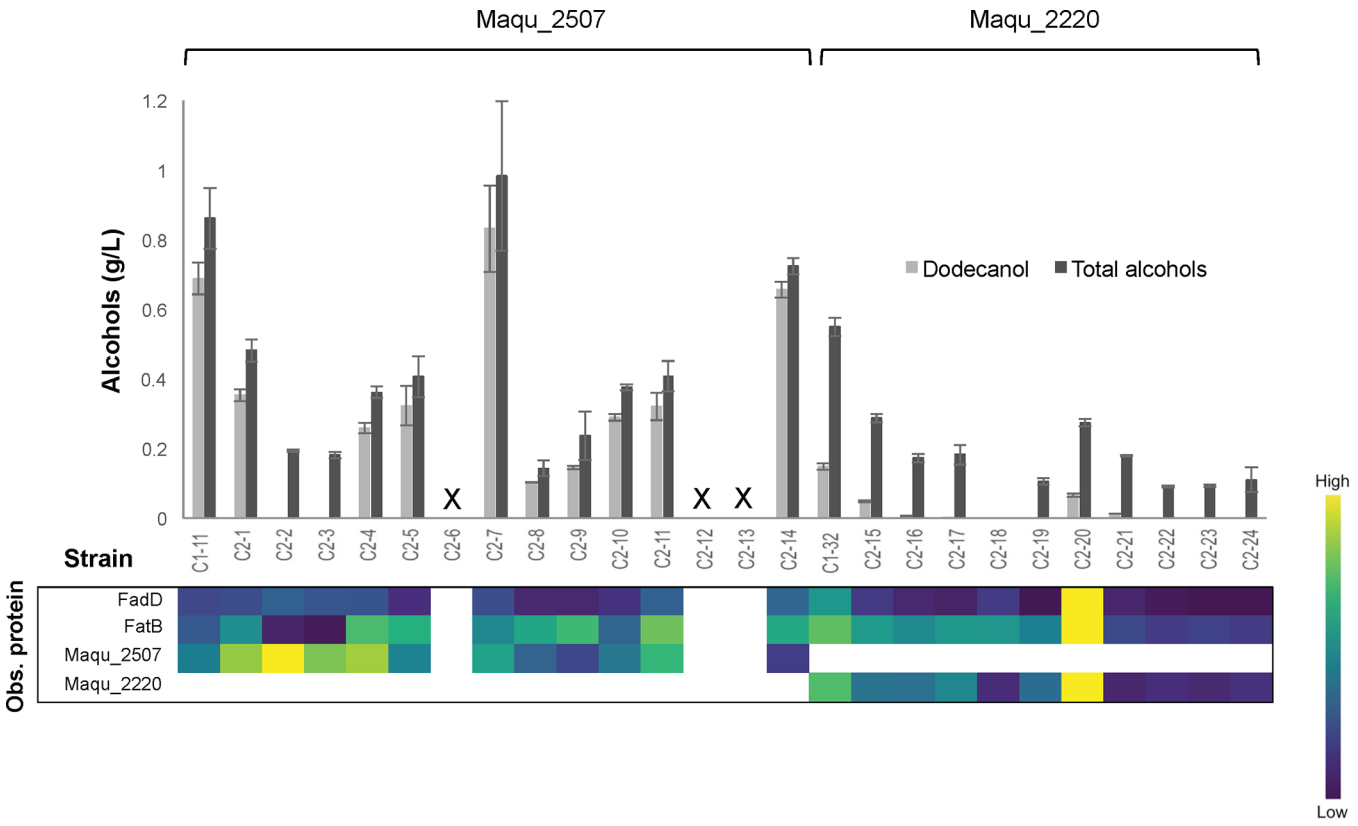


Figure 6. Cycle 2 results for alcohols and targeted proteomics. Means of dodecanol (light gray) and total alcohols (dark gray) are plotted as bars and error bars represent one standard deviation. Strain names are shown on the *x*-axis for Cycle-2 strains C2−1 to C2−24 along with Cycle-1 base strains C1−11 and C1−32 (these Cycle-1 strain results represent assays run concurrently with Cycle-2 strains). The acyl-ACP/acyl-CoA reductase used in the strains is indicated above the histogram. An X on the histogram *x*-axis indicates that the strain could not be constructed (see Table S1 for details). The heat map construction was performed as in Figure 2.

C2−9 (Figure 4) were all devoted to selecting an RBS that would enhance UcFatB1 expression 6.6- to 12-fold relative to base strain C1−11. As described below, we also used alternative strategies in an attempt to meet ensemble model protein targets, acknowledging that relying solely on RBS modification could be risky.

In the case of the Maqu_2507 strains for Cycle 2 (Figure 4), the top three dodecanol titers projected by the ensemble model specified a large increase in thioesterase (UcFatB1) expression (6.6- to 12-fold; Figure 4) and more modestly enhanced levels of FadD and Maqu_2507 (1.1- to 1.5-fold) relative to a high dodecanol-producing Cycle 1 strain (C1−11 or C1−8). Various design strategies (Figure 4) intended to greatly enhance UcFatB1 expression included an RBS-based strategy (strains C2−5 to C2−9) of using stronger predicted RBSs upstream of *UcFatB1*, and alternative strategies, such as driving UcFatB1 expression with a stronger promoter (T7 rather than P_{trc}) and using a higher copy-number origin of replication (colE1 rather than p15A).

Other strategies for Maqu_2507 strains based on metabolic engineering precedents rather than model projections included expressing transhydrogenases (PntAB or UdhA) to promote cofactor balance, which might have been perturbed by high fatty acid biosynthesis and corresponding consumption of NADPH by the key enzyme FabG, or β-ketoacyl-ACP reductase.[31−33] Such a combination of statistical methods and pathway/host engineering precedent has been used successfully in other DBTL studies to enhance flavonoid production.[3]

For the Maqu_2220 strains in Cycle 2 (Figure 5), the top two dodecanol titers projected by the ensemble model specified a substantial decrease in acyl-CoA synthetase (FadD) expression (2.7- to 10-fold; Figure 5) and somewhat enhanced levels of UcFatB1 (1.9- to 2.8-fold) relative to a moderate dodecanol-producing Cycle 1 strain (C1−32). Design strategies included using weaker predicted RBSs upstream of *fadD*, in some cases with a higher copy-number origin of replication (colE1 rather than p15A) to account for the possibility that the enhanced levels of UcFatB1 were also important for predicted performance.

While the Build phase of Cycle 2 was straightforward in principle, since we limited the changes from Cycle 1, in practice, building some Cycle-2 constructs containing *Maqu_2507* was very problematic (Table S1). For example, multiple deletions occurred in the plasmid for strain C2−6 after transformation into the MG1655 Δ*fadE* production host, even though the plasmid sequence was error-free in the DH10B cloning strain. In the end, we were never able to successfully construct strain C2−6. Similar problems occurred after transformation of the strain C2−9 plasmid into the production host, but eventually an error-free plasmid was obtained after numerous attempts. It is intriguing for strains C2−5 to C2−9 (Figure 4) that such different Build outcomes occurred among plasmids with such similar composition (differing only in the RBS for *UcFatB1*). It is also noteworthy that such plasmid mutation problems occurred only with Maqu_2507 strains and never with any Maqu_2220 strains.

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

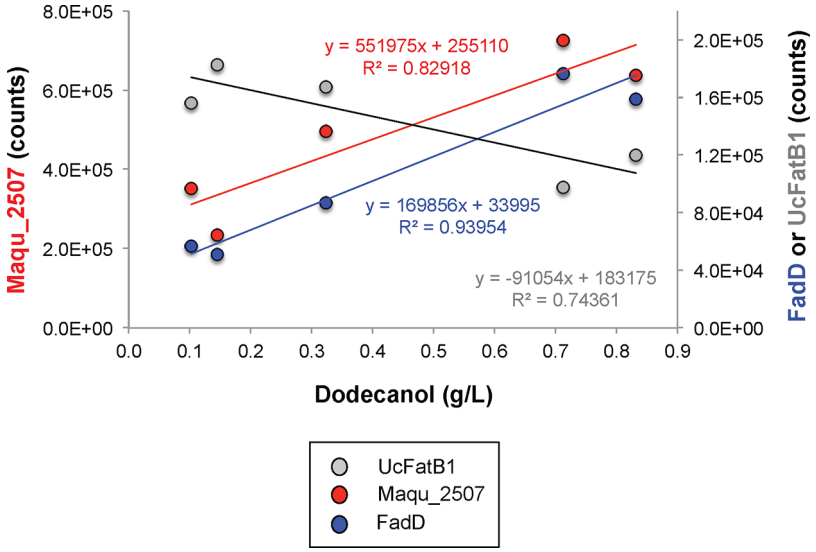


Figure 7. Pathway protein amounts (Maqu_2507, FadD, and UcFatB1) *versus* dodecanol titer for the subset of Cycle-2 strains C2−5, −7, −8, −9 and their base strain C1−11. Slopes, intercepts, and r^2 values for linear regressions are shown.

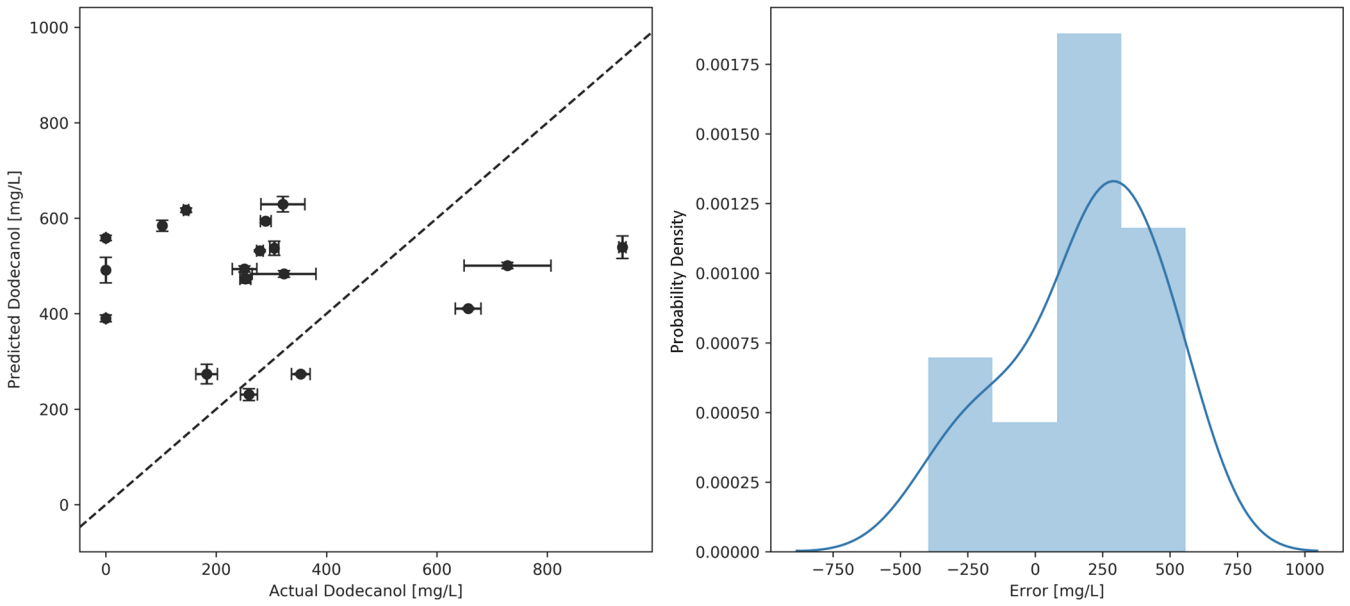


Figure 8. Predictions of ensemble model for Cycle-2 strains containing the Maqu_2507 reductase. (Left) Model predictions of dodecanol titer *vs* observed titer for 18 unique measurements of Cycle-2 strains (grouped by BioLector batch in which the strain was run, if sets of biological replicates were run in multiple batches). The dashed line represents perfect prediction ($x = y$). These predictions were obtained through 5-fold cross validation. The associated coefficient of determination (R^2) is −0.77. (Right) Probability density *vs* error in predicted dodecanol titer (the average error was 249 mg/L).

Cycle 2 Test and Learn: Strain Performance and Predictability. Fatty alcohol and pathway protein concentrations for Cycle-2 strains are presented in Figure 6. The best dodecanol-producing strain in Cycle 2, and in the entire study, was C2−7, which produced 0.83 ± 0.125 g/L of dodecanol (n = 6), a 21% increase in titer relative to the Cycle-1 base strain (C1−11; Figure 4); in different BioLector fermentation batches with triplicates of both strains, strain C2−7 averaged from 14 to 27% higher dodecanol titer than strain C1−11. The fraction of dodecanol to total alcohols for strain C2−7 was slightly better than that of its base strain (84.6% *vs* 79%). The dodecanol titer of strain C2−7 was ∼6.4-fold higher than the best published titer for batch conditions with minimal medium, ∼130 mg/L,[9] which was attained for a strain expressing similar pathway genes (*Maqu_2220*, *fadD*, and two thioesterase genes, including *UcFatB1*) but in a different configuration. However, the titer of strain C2−7 was only approximately half of the reported fed-batch titer of a strain expressing the same pathway genes (*Maqu_2507*, *fadD*, and *UcFatB1*) in a different configuration and with a different *E. coli* MG1655 background (*e.g.*, using chromosomal integration of *fadD* and *UcFatB1*, and with additional knockouts, such as Δ*ackA* and Δ*pta*).[16] Although the dodecanol titer of strain C2−7 (in batch) was considerably lower than that in the fed-batch study of Youngquist and co-workers, the productivity of C2−7 was 2- to 3-fold higher.

For Cycle-2 strains containing Maqu_2507, the best performing strain (C2−7) belonged to the tightly related

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

group C2−5 to C2−9. The design of these five strains was directly based on the operon of strain C1−11, with an alternate RBS in front of the gene encoding the UcFatB1 thioesterase (Figure 4). As this set of strains represented the closest match to the Cycle-1 Maqu_2507 constructs that served as the training set for machine-learning algorithms (*i.e.*, only one RBS change), and also encompassed a broad range of dodecanol titers (~0.1 to 0.83 g/L) and Build success/failure (discussed above), we focus our discussion on these strains. While strains C2−5 to C2−9 fell well short of the targeted UcFatB1 expression increase of 6.6−12-fold relative to the base strain (C1−11), there was still a substantial increase in their thioesterase expression (1.7- to 2.6-fold). Surprisingly, among these strains, UcFatB1 expression was not positively correlated with dodecanol titer; rather, it was inversely correlated (Figure 7). In contrast, FadD and Maqu_2507 expression were strongly (positively) correlated with dodecanol titer (r^2 values of 0.83 and 0.94, respectively; Figure 7). As a related observation, it is clear that an inverse polarity was occurring in the 3-gene pathway operon, whereby increasing strength in the RBS upstream of the first gene in the operon (*UcFatB1*) was correlated with decreasing expression of the downstream proteins FadD and Maqu_2507 (Figure S4). Thus, several aspects of the results for strains C2−5 to C2−9 were surprising in light of overall Cycle 1 trends and model predictions, for example: (a) the putatively toxic *fadD* was positively correlated with dodecanol titer and (b) the model-promoted increase in UcFatB1 expression was inversely correlated with dodecanol titer. One possible reason that the results for this small group of samples is not reflective of the more combinatorial space encompassed in Cycle 1 is that samples C2−5 to C2−9 represent a relatively narrow proteomic space, and local trends of the kind observed for strains C2−5 to C2−9 may have been obscured by a broader data set in Cycle 1. Notably, the addition of this small set of Cycle-2 strains (C2−5 to C2−9) to the original Cycle-1 training set for the machine-learning algorithms produced very different recommendations for protein profiles than those based solely on Cycle-1 data (compare Table S2 to Figure 4), indicating that the algorithms have not yet converged to a stable function relating protein profiles to dodecanol production. This is not surprising given the paucity of training data available in Cycle 1, and highlights the need to include more training samples in order to reach better predictions (Figure S5 shows how the prediction error decreases with the number of strains used for training). Regardless of sources of uncertainty, the average error of the ensemble model in predicting dodecanol titer was 0.25 g/L for all Cycle-2 strains expressing Maqu_2507 (Figure 8). The ensemble model predictions tended to cluster around a titer of *ca.* 0.3 to 0.6 g/L for Cycle 2 strains, which is not surprising since the titer range in the training set of Maqu_2507 strains was similarly constrained. Specifically for the best-performing strain (C2−7), the model prediction was 0.49 g/L. While cross-validation results for the ensemble model are presented in Figure 8, cross-validation results for each of the four contributing models (random forest, polynomial, multilayer perceptron, and TPOT) are presented in Figures S6, S7, S8, and S9.

Cycle-2 strains expressing the Maqu_2220 reductase had considerably lower titer than their Cycle-1 base strain, C1−32 (Figure 6). In these Cycle-2 strains, the design strategy entailed a substantial (63% to 90%) decrease in FadD expression (Figure 5); accordingly, actual FadD expression decreased from 73 to 99% in strains C2−15, −16, −17, and −19. However, surprisingly, decreasing the RBS for *fadD*, which is the last gene in the operon, dramatically decreased the expression of proteins encoded by the two upstream genes as well (*UcFatB1* and *Maqu_2220*) (Figure S10). This positive polarity for RBS strength is the opposite of what we observed for the Maqu_2507 constructs (Figure S4) and was unexpected because polar effects are typically observed for the first genes in an operon rather than the final one. Ultimately, the decrease in expression of all three proteins in the pathway operon, not just the targeted FadD, likely contributed to a decrease in dodecanol production (in many cases, dodecanol was not detected).

CONCLUSIONS

In summary, although we met with some success in this study, attaining a final dodecanol titer in Cycle 2 that is more than 6-fold greater than previously reported batch values, we encountered many challenges that detracted from a smooth increase in production guided by the machine-learning algorithms. While the details of these challenges are specific to this study, the larger issues that they represent may be widely applicable to other DBTL-based projects. These challenges included the following: (1) a very limited ability to predictably modulate protein expression with existing RBS prediction software (Figures 3, S2, S3, and S11), which severely constrained Design strategies; (2) nontarget effects of pathway proteins at both the Build and Test levels (*e.g.*, apparent FadD toxicity; Figures 3 and S1, Table S1), which highlighted the importance of sequencing checks on plasmids in *production strains* as well as in cloning strains; (3) whole-operonic effects of changing a single RBS (*e.g.*, polarity effects, Figures S4 and S10); (4) the masking of local data trends in the full data space (*e.g.*, comparing Figure 7 and Figure 3) when using training sets of closely related constructs instead of more heterogeneous and combinatorial constructs; and (5) the number of data points for training machine-learning algorithms is critical (Figure S5).

While there will always be a tension between relying on machine-learning models and constructing training data sets of sufficient scope to productively inform the models, carefully constructing those data sets can have a substantial impact on the performance of the models. In our case, expanding the data set to include a time course instead of just an end point might have allowed us to detect some of the dynamics of protein regulation that would have given us more confidence in these measurements, which in turn would have led to more accurate models and predictions without the construction of additional strains. Further, expanding our proteomic targets to include key enzymes in glycolysis and fatty acid biosynthesis could have yielded more systemic information on carbon flow through our entire pathway and would have given us a wider range of parameters for optimization. In addition, shotgun proteomic analysis could have complemented targeted proteomic analysis and might have provided more information on the functional proteome and, more generally, on sources of stress in the various engineered strains. Other changes in approach could also be beneficial. For example, while DNA construction costs are still limiting, cost-effectiveness could be boosted by using interference mechanisms (CRISPRi or RNA silencing). Transfer learning techniques[34] could be of great use here, as well. In any case, the observation that average prediction error decreased with increasing training set size

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(Figure S5) suggests that larger data set sizes would likely have improved model performance, enabling a more reliable test of our hypothesis that product titer can be predicted (or approximated) based upon pathway-protein expression.

In any global optimization, there will be a trade-off between exploiting what is known about a fitness landscape and exploring the larger space of possibilities. In this study, we completed two DBTL cycles to improve dodecanol titer, which organically split this study into an exploration-based combinatorial Design approach for Cycle 1 (Figure 1B) and an exploitative Design approach for Cycle 2 (Figures 4 and 5). Ideally, this DBTL process would be repeated a number of times, which would skew the relative weights of exploration *vs* exploitation in each subsequent iteration, with increasing weight being given to exploitation in later cycles. In this study, our fully exploitative Learn and Design approach to Cycle 2 (for Maqu_2507) targeted a very narrow proteomic space of increasing FatB 6.6- to 12-fold (Figure 4). Although a Cycle-2 strain did attain an improved dodecanol titer (21% increase), Cycle 2 data (*e.g.*, Figure 7, 8) suggests that the ensemble model did not capture the system dynamics well. Presumably, this is at least partially a result of sources of uncertainty discussed earlier, including a relatively small sample training set. In hindsight, it may have been preferable to give greater weight to exploration in our strain-picking algorithm, especially with a limited number of DBTL cycles, and dedicate some Cycle-2 strains to exploring underdetermined proteomic regions of our ensemble model. We could, for example, have used cross validation to explore parameter sensitivity in titer prediction (*e.g.*, determining which target proteins were associated with the largest uncertainty in predictions). Overall, this study highlights the challenges of determining the appropriate balance between contrasting approaches, such as exploitation *vs* exploration, and sample number *vs* machine-learning accuracy, which are going to be key parameters for optimizing machine learning and leveraging high-throughput DBTL systems for synthetic biology in the future.

Although the observed ∼20% increase in dodecanol titer in Cycle 2 was not a transformational result, comparable increases compounded over successive iterations of the DBTL cycle could lead to substantial gains in production: ∼250% for 5 cycles (1.2^5) and ∼620% for 10 cycles (1.2^{10}). For context, an 800% increase in production would approach 100% of maximum theoretical yield for dodecanol from glucose. We expect that exponential increases in our capability to synthesize DNA and characterize phenotype will make this scenario of inexpensive and fast DBTL cycles a reality in the near future. This approach will then provide a technique that can be applied systematically to any molecule, pathway, and host, without the need for an encyclopedic knowledge of its metabolism.

METHODS

Plasmid and Strain Construction. *Level 0 Constructs and Single-ORF Expression Constructs.* Design and Build activities were coordinated and facilitated with j5[35] DNA assembly design automation software and DeviceEditor[36] web-based bioCAD software. The DNA fragments for Level 0 constructs (used to build Level 1 constructs) and single-ORF expression constructs (for proteomic method development) were amplified using Q5 Hot Start High Fidelity 2X Master Mix (NEB, Ipswich, MA): 50-μL PCR reactions consisted of 0.5 μL (50 μM) of each forward and reverse primer, 4 μL of template (5 ng/μL), 25 μL of Q5 High Fidelity 2X Master Mix, and 20 μL reagent water. The following touchdown PCR thermocycling conditions were used: 98 °C for 30 s, then 10 cycles of {98 °C for 10 s, annealing at specified temperature for 30 s with a decrease in annealing temperature of 0.5 °C per cycle, 72 °C for 20 s/kb}, then 25 cycles of {98 °C for 10 s, annealing at the specified temperature for 30 s, 72 °C for 20 s/kb}, and final extension at 72 °C for 2 min. Following PCR amplification, residual (methylated) DNA template in each PCR reaction was DpnI digested and purified using a NIMBUS size selection robot (Hamilton, Reno, NV). Gel purified DNA fragments were cleaned up with AMPure magnetic beads (NEB) to remove the buffer that was used for elution by the NIMBUS robot; this step increased the DNA concentration by eluting in less reagent water to improve the Gibson assembly efficiency. Gibson assembly[37] was performed by mixing the DNA parts in an equimolar ratio, as specified in Table S3. Ten μL of Gibson assembly reaction consisted of 5 μL of Gibson Mix (NEB), an equimolar ratio of DNA parts, and reagent water, and was incubated at 50 °C for 1 h. Five μL of Gibson assembly reaction was transformed into chemically competent DH10B cells (Invitrogen, Carlsbad, CA). Eight colonies per construct were selected and grown overnight in 1 mL lysogeny broth (LB) with the appropriate antibiotic in a 96-well plate. Cell cultures were used for DNA sequencing on a MiSeq system (Illumina, Inc., San Diego, CA) for sequence verification, as described elsewhere.[38] Sequence-verified colonies from each Level 0 construct were further used for building Level 1 constructs. Single-ORF constructs were transformed into electrocompetent MG1655 Δ*fadE* cells and 4 colonies per construct were selected for MiSeq sequencing. Sequence-verified colonies from each expression construct were used for proteomics analysis. Detailed listings of oligonucleotides, templates, and PCR conditions are given in Table S3.

Level 1 Constructs for Dodecanol Production. Level 1 constructs were built by digesting Level 0 constructs with the Type II endonuclease BsaI and ligating them as specified in Tables S4A and S4B in a 10-μL Golden-Gate assembly reaction,[39,40] which consisted of equimolar DNA parts, 5 μL Golden-Gate mix (NEB), and reagent water. Reactions were then incubated at 37 °C for 1 h. Five μL of the Golden-Gate assembly was used for transformation into chemically competent DH10B cells. Eight colonies per construct were selected and grown overnight in 1 mL of LB with kanamycin in a 96-well plate. Cell cultures were used for DNA sequencing on a MiSeq system for sequence verification, as described elsewhere.[38]

Sequence-verified colonies from each Level 1 construct were grown in a 10-mL culture and mini-prepped (Qiagen, Hilden, Germany) for plasmid isolation. Plasmids were transformed into electrocompetent MG1655 Δ*fadE* cells (the production host for dodecanol). Differences between transformation results in the host strain *versus* the DH10B cloning strain were observed, presumably due to biological instability or toxicity (see text). Some of the Level 1 constructs that were sequence-verified in DH10B cells were found to have mutations in MG1655 Δ*fadE* cells. Although the Level 1 constructs were sequence-verified in the cloning strain, after transforming into the host strain, 8 colonies were selected again for sequencing on the MiSeq system.

A complete list of Level 0 and Level 1 plasmids and strains is provided in Table S5. Strains, plasmids, and their associated

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

information from Table S5 are available in the public instance of the JBEI registry[41] (https://public-registry.jbei.org/folders/385).

Production Runs in a BioLector Microbioreactor. Plasmids were transformed into the production strain, *E. coli* MG1655 Δ*fadE*, and colonies were picked, grown overnight in LB at 37 °C, and the plasmids were resequenced with a MiSeq system (Illumina). In preparation for the final production run, strains were acclimated overnight in M9-MOPS growth medium with 2% glucose at 30 °C. The dodecanol production experiments were run in triplicate on a BioLector microbioreactor (m2p-labs, Hauppauge, NY) with a 48-well flat-bottom plate. The cultures were inoculated with 50 μL of stationary-phase preculture, grown in M9-MOPS medium with 2% glucose, and were induced with 0.1 mM IPTG at the start of the incubation period. The total culture volume was 1 mL, with a 200 μL overlay of dodecane spiked with 500 μg (2.5 μg/μL) of the internal standard for fatty alcohol and aldehyde analysis, 1-dodecan-d_{25}-ol (98 atom % D; Sigma-Aldrich, St. Louis, MO). The BioLector was run at 1000 rpm at 30 °C and ambient chamber pressure. The total time for each run was 27 h, at which time the cultures were harvested by centrifugation at 20 817*g* for 4 min at 4 °C. The dodecane overlay was used for fatty alcohol and aldehyde analysis, the supernatant was sampled for glucose and short-chain acids (*e.g.*, acetate, pyruvate), and the cell pellet was used for targeted proteomics and selected metabolites. As an additional quality control measure, a long-chain alcohol-producing strain, JBEI-9017, studied by Haushalter *et al.*[17] was incubated in replicate in each BioLector batch and was subjected to most of the analyses described below.

Fatty Alcohols by Gas Chromatography−Mass Spectrometry (GC−MS). Dodecanol and other fatty alcohols captured in the dodecane overlay were analyzed by GC−MS. Before analysis, 1 μL of the dodecane overlay was diluted 100-fold with hexane. Electron ionization (EI) GC−MS analyses with a quadrupole mass spectrometer were performed with a model 7890A GC (Agilent Technologies, Santa Clara, CA) coupled to a HP-5ms fused silica capillary column (30-m length, 0.25-mm inner diameter, 0.25-μm film thickness; Agilent) and an HP 5975C series mass selective detector (Agilent); 1 μL injections were performed by a model 7683B autosampler (Agilent). The GC oven was programmed from 40 °C (held for 3 min) to 295 °C at 15 °C/min; the injection port temperature was 250 °C, and the transfer line temperature was 280 °C. The carrier gas, ultra high-purity helium, flowed at a constant rate of 1 mL/min. Injections were splitless, with the split turned on after 1 min. For full-scan data acquisition, the MS scanned from 50 to 600 atomic mass units at a rate of 2 scans per s. Internal standard quantification was performed. The internal standard, perdeuterated dodecanol, or 1-dodecan-d_{25}-ol (98 atom % D; Sigma-Aldrich), was spiked into the dodecane overlay such that a 1:100 dilution in hexane for GC−MS analysis would result in a 25 ng/μL concentration in the diluted extract. GC−MS standards (C_{12}, C_{14}, C_{16}, C_{18} fatty alcohols) at three concentration levels (5, 20, and 50 ng/μL) all contained 25 ng/μL of the deuterated internal standard.

Glucose and Short-Chain Organic Acids by High-Performance Liquid Chromatography (HPLC). Glucose and organic acids from cell cultures were measured by an 1100 Series HPLC system equipped with a 1200 Series refractive index detector (RID) (Agilent) and Aminex HPX-87H ion-exclusion column (300 mm length, 7.8 mm internal diameter; Bio-Rad Laboratories, Inc., Hercules, CA). One hundred-microliter aliquots of cell cultures were removed at various time points during production and filtered through a spin-cartridge with a 0.45-μm nylon membrane, and 5 μL of the filtrate was eluted through the column at 50 °C with 4 mM sulfuric acid at a flow rate of 600 μL/min for 25 min. Metabolites were quantified by using external standard calibration with authentic standards.

Targeted Proteomics by LC−MS/MS. *Proteomic Sample Preparation.* Cell lysis and protein precipitation were achieved by using a chloroform−methanol extraction as previously described.[42] The pellets were resuspended in 200 μL methanol. 50 μL chloroform and 150 μL water were added to each well. The samples were centrifuged for 1 min at maximum speed to induce the phase separation. The methanol and water layers were removed, then 300 μL methanol was added to each well. The samples were centrifuged for 1 min at maximum speed, then the chloroform and methanol layers were removed and the protein pellets were dried at room temperature for 5 min prior to resuspension in 100 mM ammonium bicarbonate with 20% methanol. The protein concentration of the samples was measured using the DC Protein Assay Kit (Bio-Rad, Hercules, CA) with bovine serum albumin used as a standard. 200 μg of protein from each sample was reduced by adding tris 2-(carboxyethyl)phosphine (TCEP) to a final concentration of 5 mM. Iodoacetamide was added to a final concentration of 10 mM to alkylate the protein samples. Trypsin was added at a ratio of 1:50 trypsin:total protein, and the samples were incubated for 16 h at 37 °C.

Liquid Chromatography−Mass Spectrometry. Peptides were analyzed using an Agilent 1290 liquid chromatography system coupled to an Agilent 6460 QQQ mass spectrometer (Agilent Technologies, Santa Clara, CA). The peptide samples (20 μg loaded on column) were separated on an Ascentis Express Peptide ES-C18 column (2.7 μm particle size, 160 Å pore size, 5 cm length × 2.1 mm i.d., coupled to a 5 mm × 2.1 mm i.d. guard column with similar particle and pore size; Sigma-Aldrich, St. Louis, MO), with the system operating at a flow rate of 0.400 mL/min and the column compartment at 60 °C. Peptides were eluted into the mass spectrometer *via* a gradient with an initial starting condition of 95% Buffer A (99.9% water, 0.1% formic acid) and 5% Buffer B (99.9% acetonitrile, 0.1% formic acid). Buffer B was held at 5% for 0.2 min, then increased to 35% B over 5.5 min. Buffer B was further increased to 80% of 0.3 min, where it was held for 2 min, then ramped back down to 5% B over 0.5 min, where it was held for 1.5 min to re-equilibrate the column to the initial starting condition. The data were acquired using Agilent MassHunter, version B.08.00 and the data files were processed by using Skyline version 4.1 (MacCoss Lab, University of Washington, Seattle, WA) with mProphet to refine peak quantification.

In Silico Selected Reaction Monitoring (SRM) Methods Selection. Skyline software was used for SRM screening, peak selection, method development, analysis, and data processing purposes. Selection criteria excluded peptides with Met/Cys residues, tryptic peptides followed by additional cut sites (KK/RR), and peptides with proline adjacent to K/R cut sites. All possible doubly charged peptides were screened for y-series ions to establish the peptide identity and the most sensitive transitions. Methods generated by Skyline were set for Agilent 6460QQQ and included instrument-specific collision energies that were exported to an instrument method file using a

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

template file that contained LC conditions as described above. To facilitate confident peptide selection, each target protein was highly overproduced in a separate *E. coli* strain and tested with the *in silico* SRM predictions from Skyline. Acquired SRM data were imported into Skyline, where peptides were manually curated into a subset meeting the criteria described above. Skyline methods, which contain proteotypic peptides and SRM information, are available from the Panorama knowledgebase[43] located at https://panoramaweb.org/DBTL-ML-for-dodencanol-production-in-e-coli.url. Protein quantification was based on the summed peak areas of the transitions for each peptide.

Fatty Aldehyde Analysis by Nanostructure-Initiator Mass Spectrometry (NIMS). Dodecanal and other fatty aldehydes captured in the dodecane overlay were analyzed using oxime bioconjugate chemistry and Nanostructure-Initiator Mass Spectrometry (NIMS). The synthesis and the subsequent oxime derivatization reactions with the *O*-alkyloxyamine fluorous tag were carried out as reported elsewhere.[44] Briefly, a 2 μL aliquot of the dodecane overlay was transferred into a vial containing 6 μL of 100 mM glycine acetate, pH 1.3, 3 μL of ethanol, 1 μL of *O*-alkyloxyamine fluorous tag [10 mM in 1:1 (v/v) water:methanol], and 0.13 μL of aniline. The mixture was incubated at room temperature for 16 h before NIMS analysis. For each sample, to 1 μL of the oxime reaction mixture, 8 μL water, 1 μL ethanol and 0.01 μL formic acid were added. Samples were printed onto a NIMS substrate (NIMS substrates were processed as described elsewhere[45]) using an ATS-100 acoustic transfer system (BioSera, San Diego, CA) with a sample deposition volume of 10 nL. Samples were printed in clusters of four replicates, with the microarray spot pitch (center-to-center distance) set at 900 μm. MS-based imaging was performed using a 5800 MALDI TOF/TOF (AB Sciex, Foster City, CA) mass spectrometer with laser intensity of 6000 over a mass range of 500−4000 Da. Each position accumulated 20 laser shots. The instrument was controlled using the MALDI-MSI 4800 Imaging Tool using a 75 μm step size. Average ion intensity of the conjugated fatty aldehydes were determined using the OMAAT tool.[46]

Fatty Acid and Acyl-CoA Analysis by LC−MS. Liquid chromatography separation conditions for fatty acyl-CoA measurements were described previously by Goh *et al.*[47] Fatty acid separation was conducted at 55 °C with a Kinetex XB-C18 column (100 mm length, 3 mm internal diameter, 2.6 mm particle size; Phenomenex, Inc., Torrance, CA) using a 1200 Series Rapid Resolution HPLC (high-performance liquid chromatography) system (Agilent Technologies, CA). The mobile phase was composed of water (solvent A) and methanol (solvent B) (HPLC grade, Honeywell Burdick & Jackson, CA). One to two microliters of samples were injected and separated with the following gradient: 60−98% B for 3.47 min, held at 98% B for 5.2 min, decreased to 60% B for 2.42 min, and held at 60% B for a further 2.78 min. The total analysis run time was 13.87 min. A flow rate of 0.42 mL/min was used throughout. The HPLC system was coupled to an Agilent Technologies 6545 quadrupole time-of-flight mass spectrometer (QTOF-MS). Electrospray ionization was conducted in the negative ion mode (*i.e.*, $[M - H]^-$) *via* the Agilent Jet Stream thermal gradient focusing technology, where the sheath gas flow rate and temperature were set to 12 L/min and 350 °C, respectively. Drying and nebulizing gases were set to 10 L/min and 25 lb/in^2, respectively, and a drying-gas temperature of 300 °C was used throughout. The fragmentor, skimmer, and OCT 1 RF Vpp voltages were set to 150, 50, and 170 V, respectively. Data acquisition and processing were conducted *via* the Agilent MassHunter software package.

Ensemble Model Construction. Four machine-learning models are used as an ensemble to improve prediction. The ensemble model is defined by

$$f_c(p) = \sum_{i=1}^{4} w_i f_i(p)$$

The weights were determined from the vector of model cross validation scores e using the equation

$$w = \text{softmax}(-\alpha e)$$

where alpha is a parameter that determines how much to weight differences in performance. In the extreme cases: $\alpha = 0$ results in even weightings regardless of performance, and, for large α approaching infinity, the best performing model is weighted as 1 while the rest of the weights are zero. So, for small α, the composite model is an average of all four models independent of performance, and for large α, only the best performing model is used for prediction. The softmax function is defined as

$$\text{softmax}(x) \doteq \frac{e^x}{\sum_{i=1}^{4} e^{x_j}}$$

Note that, in the above notation, all operations on vectors are performed element-wise. The composite model $f_c(p)$ represents the best estimated map from pathway protein levels to dodecanol titer. This ensemble model was used to make all predictions in both cycles of Learn.

Cross Validation of Machine-Learning Models. Each individual model and the ensemble model were evaluated using 5-fold cross validation and the results were scored using the mean squared error between predicted and actual production of dodecanol. The results of this analysis are shown in Figure 8 (for the ensemble model) and Figures S6−S9.

Calculation of Success Probability of Designed Strains. To guide the process of selecting which model-recommended strains to construct in Cycle 2, a naïve estimate of the probability that a recommended strain would exceed the maximum dodecanol titer observed in Cycle 1 was created. This estimate is called the "success probability". To find the success probability estimate for a given strain prediction, the mean, μ, and standard deviation, σ, of the 5-fold cross validated prediction error on the data set are first calculated. The model error is assumed to be normally distributed about these moments. If a strain is predicted to have dodecanol titer d and the best observed strain has titer $\hat{d}$, success probability is the probability that the normal distribution $\mathcal{N}(\mu + d, \sigma)$ takes a value greater than $\hat{d}$. Success probabilities were reported in Figures 4 and 5.

Partial Correlation Analysis. In order to evaluate the accuracy of the RBS calculation software used in this study, a partial correlation analysis was used to isolate the relationship between predicted RBS strength and observed protein expression for each protein in the engineered pathway. This analysis controlled for multiple confounding variables, including plasmid copy number and strength of the other RBSs in the operon.

More formally, partial correlation analysis can determine the correlation coefficient between random variables X and Y,

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

while controlling for a set of n confounding variables $\mathbf{Z} = \{Z_1, ..., Z_n\}$. Assume that we are working with a sample of m realizations drawn from each random variable with the ith realization denoted by x_i, y_i, and $\mathbf{z}_i$, respectively, with $x_i, y_i \in \mathbb{R}$ and $\mathbf{z}_i \in \mathbb{R}^n$. First, two linear regressions relating both the independent and dependent variables independently to the set of confounding variables $\mathbf{Z}$ are fit. This is realized mathematically with the equations

$$w_X = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{m} (\langle z_i, w\rangle - x_i)^2$$

$$w_Y = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{m} (\langle z_i, w\rangle - y_i)^2$$

Then, the residuals between the confounding variables and the independent and dependent variables are calculated to control for the influence of the confounding variables,

$$e_x = \{x_i - \langle \mathbf{z}_i, \mathbf{w}_x\rangle\}_{i=1}^{m}$$

$$e_y = \{y_i - \langle \mathbf{z}_i, \mathbf{w}_y\rangle\}_{i=1}^{m}$$

Now, the partial correlation coefficient and associated p-value are calculated simply by computing the Pearson correlation between the residuals e_x and e_y. Further details are available elsewhere.[48]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.9b00020.

Figures S1−S12 (PDF)
Table S1 (summary of failed constructs) (XLSX)
Table S2 (modified Cycle 2 design strategies) (XLSX)
Table S3 (Level 0 constructs and single-ORF expression constructs for Cycle 1), Table S4 (Level 1 constructs), Table S5 (list of plasmids and strains) (XLSX)
Table S6 (LC−MS results) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: hrbeller@lbl.gov.

ORCID

Markus de Raad: 0000-0001-8263-9198
Trent R. Northen: 0000-0001-8404-3259
Nathan J. Hillson: 0000-0002-9169-3978
Hector Garcia Martin: 0000-0002-4556-9685
Harry R. Beller: 0000-0001-9637-3650

Author Contributions

H.R.B., N.J.H., G.G., P.O., and S.D. conducted Design activities; G.G., N.J.H., P.O., and T.O. conducted Build activities; P.O. and T.O. conducted BioLector incubations and ancillary analyses (*e.g.*, glucose); P.O. and H.R.B. conducted GC−MS analyses; C.J.P., Y.C., and J.G. conducted proteomics analyses; M.dR., K.D., and T.N. conducted NIMS analyses; E.B. and V.B. conducted metabolite analyses; Z.C. and H.G.M. conducted Learn analyses (machine learning); H.R.B., Z.C., and P.O. analyzed the data. The manuscript was written by H.R.B., P.O., Z.C., and H.G.M., and all authors contributed to or refined the text.

Author Contributions

⊗P.O., Z.C., and H.R.B. contributed equally.

Notes

The authors declare no competing financial interest.

Associated information for the strains and plasmids listed in Table S5 are available in the public version of the JBEI registry (https://public-registry.jbei.org/folders/385). Chemical and proteomic data, predicted RBS strength (TIR and EMOPEC values), and links to synthetic biology parts characterization data from this study are available on the public version of the Experiment Data Depot (EDD[49]) site (https://public-edd.jbei.org/s/ajinomoto/). In addition, for proteomic data, all generated Skyline files are available in the Panorama Public repository at the following link: https://panoramaweb.org/DBTL-ML-for-dodencanol-production-in-e-coli.url. Proteomic data are also available *via* ProteomeXchange with identifier PXD012312. All code used to perform the analysis is publicly available under a BSD license at https://github.com/JBEI/Ajinomoto.

■ ACKNOWLEDGMENTS

This work was part of the DOE Joint BioEnergy Institute (https://www.jbei.org) and the DOE Joint Genome Institute (https://jgi.doe.gov) supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, and was part of the Agile BioFoundry (http://agilebiofoundry.org) supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. T.O. was supported by Ajinomoto Co., Inc., Tokyo, Japan, as part of a joint project between the Joint BioEnergy Institute and Ajinomoto Co., Inc. H.G.M. was also supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2017-0718. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

■ REFERENCES

(1) Keasling, J. D. (2014) *Testimony before the Subcommittee on Research and Technology; Committee on Science, Space, and Technology*, pp 1−4, U.S. House of Representatives, Washington, D.C.

(2) Hodgman, C. E., and Jewett, M. C. (2012) Cell-free synthetic biology: thinking outside the cell. *Metab. Eng. 14*, 261−269.

(3) Carbonell, P., Jervis, A. J., Robinson, C. J., Yan, C., Dunstan, M., Swainston, N., Vinaixa, M., Hollywood, K. A., Currin, A., Rattray, N. J. W., Taylor, S., Spiess, R., Sung, R., Williams, A. R., Fellows, D., Stanford, N. J., Mulherin, P., Le Feuvre, R., Barran, P., Goodacre, R., Turner, N. J., Goble, C., Chen, G. G., Kell, D. B., Micklefield, J., Breitling, R., Takano, E., Faulon, J. L., and Scrutton, N. S. (2018) An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol. 1*, 66.

(4) Xu, P., Rizzoni, E. A., Sul, S. Y., and Stephanopoulos, G. (2017) Improving Metabolic Pathway Efficiency by Statistical Model-Based

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

Multivariate Regulatory Metabolic Engineering. *ACS Synth. Biol. 6*, 148−158.
(5) Zhou, H., Vonk, B., Roubos, J. A., Bovenberg, R. A., and Voigt, C. A. (2015) Algorithmic co-optimization of genetic constructs and growth conditions: application to 6-ACA, a potential nylon-6 precursor. *Nucleic Acids Res. 43*, 10560−10570.
(6) Jervis, A. J., Carbonell, P., Vinaixa, M., Dunstan, M. S., Hollywood, K. A., Robinson, C. J., Rattray, N. J. W., Yan, C., Swainston, N., Currin, A., Sung, R., Toogood, H. S., Taylor, S., Faulon, J. L., Breitling, R., Takano, E., and Scrutton, N. S. (2019) Machine learning of designed translational control allows predictive pathway optimization in *Escherichia coli*. *ACS Synth. Biol. 8*, 127−136.
(7) Beller, H. R., Lee, T. S., and Katz, L. (2015) Natural products as biofuels and bio-based chemicals: fatty acids and isoprenoids. *Nat. Prod. Rep. 32*, 1508−1526.
(8) Steen, E. J., Kang, Y., Bokinsky, G., Hu, Z., Schirmer, A., McClure, A., Del Cardayre, S. B., and Keasling, J. D. (2010) Microbial production of fatty-acid-derived fuels and chemicals from plant biomass. *Nature 463*, 559−562.
(9) Liu, A., Tan, X., Yao, L., and Lu, X. (2013) Fatty alcohol production in engineered *E. coli* expressing *Marinobacter* fatty acyl-CoA reductases. *Appl. Microbiol. Biotechnol. 97*, 7061−7071.
(10) Reiser, S., and Somerville, C. (1997) Isolation of mutants of *Acinetobacter calcoaceticus* deficient in wax ester synthesis and complementation of one mutation with a gene encoding a fatty acyl coenzyme A reductase. *J. Bacteriol. 179*, 2969−2975.
(11) Schirmer, A., Rude, M. A., Li, X., Popova, E., and del Cardayre, S. B. (2010) Microbial biosynthesis of alkanes. *Science 329*, 559−562.
(12) Willis, R. M., Wahlen, B. D., Seefeldt, L. C., and Barney, B. M. (2011) Characterization of a fatty acyl-CoA reductase from *Marinobacter aquaeolei* VT8: a bacterial enzyme catalyzing the reduction of fatty acyl-CoA to fatty alcohol. *Biochemistry 50*, 10550−10558.
(13) Akhtar, M. K., Turner, N. J., and Jones, P. R. (2013) Carboxylic acid reductase is a versatile enzyme for the conversion of fatty acids into fuels and chemical commodities. *Proc. Natl. Acad. Sci. U. S. A. 110*, 87−92.
(14) Quinn, J. Y., Cox, R. S., 3rd, Adler, A., Beal, J., Bhatia, S., Cai, Y., Chen, J., Clancy, K., Galdzicki, M., Hillson, N. J., Le Novere, N., Maheshwari, A. J., McLaughlin, J. A., Myers, C. J., P, U., Pocock, M., Rodriguez, C., Soldatova, L., Stan, G. B., Swainston, N., Wipat, A., and Sauro, H. M. (2015) SBOL Visual: A Graphical Language for Genetic Designs. *PLoS Biol. 13*, e1002310.
(15) Zheng, Y. N., Li, L. L., Liu, Q., Yang, J. M., Wang, X. W., Liu, W., Xu, X., Liu, H., Zhao, G., and Xian, M. (2012) Optimization of fatty alcohol biosynthesis pathway for selectively enhanced production of C12/14 and C16/18 fatty alcohols in engineered *Escherichia coli*. *Microb. Cell Fact. 11*, 65.
(16) Youngquist, J. T., Schumacher, M. H., Rose, J. P., Raines, T. C., Politz, M. C., Copeland, M. F., and Pfleger, B. F. (2013) Production of medium chain length fatty alcohols from glucose in *Escherichia coli*. *Metab. Eng. 20*, 177−186.
(17) Haushalter, R. W., Groff, D., Deutsch, S., The, L., Chavkin, T. A., Brunner, S. F., Katz, L., and Keasling, J. D. (2015) Development of an orthogonal fatty acid biosynthesis system in *E. coli* for oleochemical production. *Metab. Eng. 30*, 1−6.
(18) Fatma, Z., Jawed, K., Mattam, A. J., and Yazdani, S. S. (2016) Identification of long chain specific aldehyde reductase and its use in enhanced fatty alcohol production in *E. coli*. *Metab. Eng. 37*, 35−45.
(19) Salis, H. M., Mirsky, E. A., and Voigt, C. A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol. 27*, 946−950.
(20) Espah Borujeni, A., Channarasappa, A. S., and Salis, H. M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res. 42*, 2646−2659.
(21) Bonde, M. T., Pedersen, M., Klausen, M. S., Jensen, S. I., Wulff, T., Harrison, S., Nielsen, A. T., Herrgard, M. J., and Sommer, M. O. (2016) Predictable tuning of protein expression in bacteria. *Nat. Methods 13*, 233−236.
(22) Pollard, M. R., Anderson, L., Fan, C., Hawkins, D. J., and Davies, H. M. (1991) A specific acyl-ACP thioesterase implicated in medium-chain fatty acid production in immature cotyledons of *Umbellularia californica*. *Arch. Biochem. Biophys. 284*, 306−312.
(23) Hofvander, P., Doan, T. T., and Hamberg, M. (2011) A prokaryotic acyl-CoA reductase performing reduction of fatty acyl-CoA to fatty alcohol. *FEBS Lett. 585*, 3538−3543.
(24) Alonso-Gutierrez, J., Kim, E. M., Batth, T. S., Cho, N., Hu, Q., Chan, L. J. G., Petzold, C. J., Hillson, N. J., Adams, P. D., Keasling, J. D., Garcia Martin, H., and Lee, T. S. (2015) Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng. 28*, 123−133.
(25) Breiman, L. (2001) Random Forests. *Machine Learning 45*, 5−32.
(26) Olson, R., Fu, W., Daniel, N., Grishma, J., and Raschka, S. (2017) rhiever/tpot: Sparse matrix support, early stopping, and checkpointing. *Zenodo*, DOI: 10.5281/zenodo.998172.
(27) Olson, R. S., Urbanowicz, R. J., Andrews, P. C., Lavender, N. A., Kidd, L. C., and Moore, J. H. (2016) Automating Biomedical Data Science Through Tree-Based Pipeline Optimization, In *Applications of Evolutionary Computation. EvoApplications 2016* (Squillero, G., and Burelli, P., Eds.), Springer.
(28) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011) Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 2825−2830.
(29) Caruana, R., Karampatziakis, N., and Yessenalina, A. (2008) An empirical evaluation of supervised learning in high dimensions, In *Proceedings of the 25th International Conference on Machine Learning*, pp 96−103, Helsinki, Finland.
(30) Oliphant, T. E. (2007) Python for scientific computing. *Comput. Sci. Eng. 9*, 10−20.
(31) Goh, E. B., Chen, Y., Petzold, C. J., Keasling, J. D., and Beller, H. R. (2018) Improving methyl ketone production in *Escherichia coli* by heterologous expression of NADH-dependent FabG. *Biotechnol. Bioeng. 115*, 1161−1172.
(32) Gonzalez, J. E., Long, C. P., and Antoniewicz, M. R. (2017) Comprehensive analysis of glucose and xylose metabolism in *Escherichia coli* under aerobic and anaerobic conditions by C-13 metabolic flux analysis. *Metab. Eng. 39*, 9−18.
(33) Hua, Q., Yang, C., Baba, T., Mori, H., and Shimizu, K. (2003) Responses of the central metabolism in *Escherichia coli* to phosphoglucose isomerase and glucose-6-phosphate dehydrogenase knockouts. *J. Bacteriol. 185*, 7053−7067.
(34) Pan, S. J., and Yang, Q. (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering 22*, 1345−1359.
(35) Hillson, N. J., Rosengarten, R. D., and Keasling, J. D. (2012) j5 DNA assembly design automation software. *ACS Synth. Biol. 1*, 14−21.
(36) Chen, J., Densmore, D., Ham, T. S., Keasling, J. D., and Hillson, N. J. (2012) DeviceEditor visual biological CAD canvas. *J. Biol. Eng. 6*, 1.
(37) Gibson, D. G., Young, L., Chuang, R. Y., Venter, J. C., Hutchison, C. A., 3rd, and Smith, H. O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods 6*, 343−345.
(38) Thompson, M. G., Sedaghatian, N., Barajas, J. F., Wehrs, M., Bailey, C. B., Kaplan, N., Hillson, N. J., Mukhopadhyay, A., and Keasling, J. D. (2018) Isolation and characterization of novel mutations in the pSC101 origin that increase copy number. *Sci. Rep. 8*, 1590.
(39) Engler, C., Gruetzner, R., Kandzia, R., and Marillonnet, S. (2009) Golden gate shuffling: a one-pot DNA shuffling method based on type IIs restriction enzymes. *PLoS One 4*, e5553.
(40) Engler, C., Kandzia, R., and Marillonnet, S. (2008) A one pot, one step, precision cloning method with high throughput capability. *PLoS One 3*, e3647.

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX

(41) Ham, T. S., Dmytriv, Z., Plahar, H., Chen, J., Hillson, N. J., and Keasling, J. D. (2012) Design, implementation and practice of JBEI-ICE: an open source biological part registry platform and tools. *Nucleic Acids Res. 40*, e141.
(42) Gonzalez Fernandez-Nino, S. M., Smith-Moritz, A. M., Chan, L. J., Adams, P. D., Heazlewood, J. L., and Petzold, C. J. (2015) Standard flow liquid chromatography for shotgun proteomics in bioenergy research. *Front. Bioeng. Biotechnol. 3*, 44.
(43) Sharma, V., Eckels, J., Taylor, G. K., Shulman, N. J., Stergachis, A. B., Joyner, S. A., Yan, P., Whiteaker, J. R., Halusa, G. N., Schilling, B., Gibson, B. W., Colangelo, C. M., Paulovich, A. G., Carr, S. A., Jaffe, J. D., MacCoss, M. J., and MacLean, B. (2014) Panorama: a targeted proteomics knowledge base. *J. Proteome Res. 13*, 4205−4210.
(44) Deng, K., Takasuka, T. E., Heins, R., Cheng, X., Bergeman, L. F., Shi, J., Aschenbrener, R., Deutsch, S., Singh, S., Sale, K. L., Simmons, B. A., Adams, P. D., Singh, A. K., Fox, B. G., and Northen, T. R. (2014) Rapid kinetic characterization of glycosyl hydrolases based on oxime derivatization and nanostructure-initiator mass spectrometry (NIMS). *ACS Chem. Biol. 9*, 1470−1479.
(45) Woo, H. K., Northen, T. R., Yanes, O., and Siuzdak, G. (2008) Nanostructure-initiator mass spectrometry: a protocol for preparing and applying NIMS surfaces for high-sensitivity mass analysis. *Nat. Protoc. 3*, 1341−1349.
(46) de Raad, M., de Rond, T., Rubel, O., Keasling, J. D., Northen, T. R., and Bowen, B. P. (2017) OpenMSI Arrayed Analysis Toolkit: Analyzing Spatially Defined Samples Using Mass Spectrometry Imaging. *Anal. Chem. 89*, 5818−5823.
(47) Goh, E. B., Baidoo, E. E., Burd, H., Lee, T. S., Keasling, J. D., and Beller, H. R. (2014) Substantial improvements in methyl ketone production in *E. coli* and insights on the pathway from *in vitro* studies. *Metab. Eng. 26*, 67−76.
(48) Wang, J. (2013) Partial Correlation Coefficient, In *Encyclopedia of Systems Biology* (Dubitzky, W., Wolkenhauer, O., Cho, K. H., and Yokota, H., Eds.), Springer, New York, NY.
(49) Morrell, W. C., Birkel, G. W., Forrer, M., Lopez, T., Backman, T. W. H., Dussault, M., Petzold, C. J., Baidoo, E. E. K., Costello, Z., Ando, D., Alonso-Gutierrez, J., George, K. W., Mukhopadhyay, A., Vaino, I., Keasling, J. D., Adams, P. D., Hillson, N. J., and Garcia Martin, H. (2017) The Experiment Data Depot: A Web-Based Software Tool for Biological Experimental Data Storage, Sharing, and Visualization. *ACS Synth. Biol. 6*, 2248−2259.

DOI: 10.1021/acssynbio.9b00020
ACS Synth. Biol. XXXX, XXX, XXX−XXX