ACS Synthetic Biology
Research Article
Cite This: ACS Synth. Biol. 2019, 8, 127−136
pubs.acs.org/synthbio

Machine Learning of Designed Translational Control Allows Predictive Pathway Optimization in *Escherichia coli*

Adrian J. Jervis,[†] Pablo Carbonell,[†] Maria Vinaixa,[†] Mark S. Dunstan,[†] Katherine A. Hollywood,[†] Christopher J. Robinson,[†] Nicholas J. W. Rattray,[§] Cunyu Yan,[†] Neil Swainston,[†] Andrew Currin,[†] Rehana Sung,[†] Helen Toogood,[†] Sandra Taylor,[†] Jean-Loup Faulon,[†,‡] Rainer Breitling,[†] Eriko Takano,[†] and Nigel S. Scrutton*,[†]

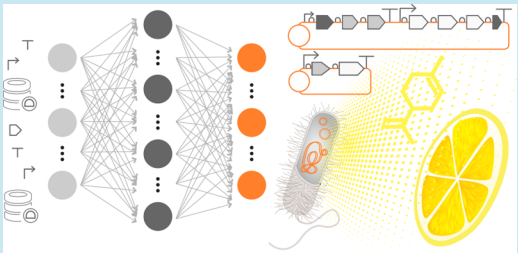
[†]Manchester Synthetic Biology Research Centre for Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology and School of Chemistry, University of Manchester, Manchester M1 7DN, United Kingdom

[‡]MICALIS, INRA-AgroParisTech, Domaine de Vilvert, 78352 Jouy en Josas Cedex, France

[§]Strathclyde Institute of Pharmacy and Biomedical Sciences, Strathclyde University, 161 Cathedral Street, Glasgow G4 0RE, United Kingdom

Supporting Information

ABSTRACT: The field of synthetic biology aims to make the design of biological systems predictable, shrinking the huge design space to practical numbers for testing. When designing microbial cell factories, most optimization efforts have focused on enzyme and strain selection/engineering, pathway regulation, and process development. *In silico* tools for the predictive design of bacterial ribosome binding sites (RBSs) and RBS libraries now allow translational tuning of biochemical pathways; however, methods for predicting optimal RBS combinations in multigene pathways are desirable. Here we present the implementation of machine learning algorithms to model the RBS sequence−phenotype relationship from representative subsets of large combinatorial RBS libraries allowing the accurate prediction of optimal high-producers. Applied to a recombinant monoterpenoid production pathway in *Escherichia coli*, our approach was able to boost production titers by over 60% when screening under 3% of a library. To facilitate library screening, a multiwell plate fermentation procedure was developed, allowing increased screening throughput with sufficient resolution to discriminate between high and low producers. High producers from one library did not translate during scale-up, but the reduced screening requirements allowed rapid rescreening at the larger scale. This methodology is potentially compatible with any biochemical pathway and provides a powerful tool toward predictive design of bacterial production chassis.



KEYWORDS: *ribosome binding site, pathway engineering, machine learning, terpenoids, translational tuning, synthetic biology*

Synthetic biology has made great strides in engineering biological systems over the past decade, especially in the engineering of microbial hosts for the production of specialty and fine chemicals. One key aim of synthetic biology is to make the design of biological systems predictable, but even relatively simple organisms such as bacteria are extremely complex, and understanding all the processes occurring in the cell at one time is difficult, making it hard to predict the outcomes of perturbations. In developing microbial production chassis efforts have predominantly concentrated on optimization of the desired biochemical pathway *via* enzyme selection,[1−3] at the transcriptional level (vectors, promoters, gene circuits),[4−8] by chassis selection,[4,8−10] and by engineering and process development.[9,11,12] More recently, a number of tools have become available to design bacterial ribosome binding sites (RBSs) and RBS libraries, presenting the opportunity to predictably control pathway flux at the translational level.[13−17] Several examples exist of studies using designed RBSs and even tuning whole pathways using RBS libraries.[7,16,38] However, when introducing the variant sequences at multiple genes in a pathway, the number of possible sequence permutations increases exponentially resulting in intractable numbers for building or screening. *In silico* tools are required to guide intelligent sampling combined with methods to learn and predict improved performance. Machine learning is now being applied to biological problems[18] and has been used to allow prediction of promoter sequences and activity,[19,20] genome annotation,[21] identification of DNA/RNA binding proteins,[22] and other applications. To date there have been no examples of methods to learn solely from RBS sequence data to allow the prediction of sequences with improved phenotype (*e.g.*, target molecule production titers). Machine learning is ideally placed to address this need.

Isoprenoids, especially terpenes, have been at the forefront of natural product production in microbial hosts,[9,23−25] and

Received: September 21, 2018
Published: December 18, 2018

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

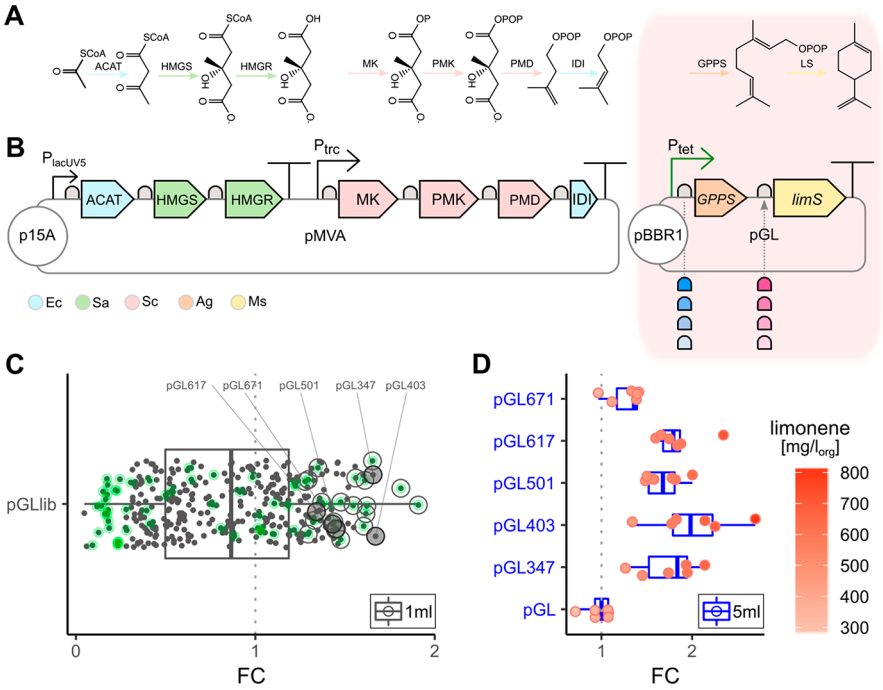


Figure 1. Translational tuning of GPPS and LimS genes encoded on plasmid pGL. (A) Schematic of the MVA pathway from acetyl-CoA (Ac-CoA) to (*S*)-limonene *via* the isoprenoid precursors IPP and DMAPP. (B) Schematic of the pMVA plasmid encoding the MVA pathway and pGL carrying the tr*Ag-gpps* and *trMs-limS* genes. The pGLlib library contains variable RBS sequences for the tr*Ag-gpps* and *trMs-limS* genes as indicated. The color code of each gene indicates the source organism (Table S1). (C) Library pGLlib1, in combination with pMVA, was screened for *in vivo* limonene production in DWPs as 1 mL cultures and limonene production is shown as fold-change (FC) relative to the original pGL production levels. Green highlight indicates clones that were sequenced; individual clones displayed in panel D are labeled. (D) Rescreening of high producing clones from pGLlib, grown in triplicate cultures and two biological repeats, in 5 mL individual cultures. The amount of limonene produced is shown by the intensity of the red color as indicated. Translational tuning of GPPS and LimS created a library of variants capable of a wide range of production titers and allowed the selection of single clones, which reproducibly produce higher titers than the template plasmid, pGL.

high titers (>1 g/L) of several compounds have been reported, including amorphadiene,[26] isoprene,[27] and lycopene.[28] All isoprenoids are synthesized from the isoprene precursors isopentenyl pyrophosphate (IPP) and dimethylallyl pyrophosphate (DMAPP) *via* either the mevalonate (MVA) or the 2-*C*-methyl-D-erythritol 4-phosphate (MEP) pathway (Figure 1A), but despite sharing these common precursors, the titers of monoterpenoids achieved to date have lagged behind those of other isoprenoids, including higher-order terpenes. The C_{10} monoterpenoids include many valuable compounds used as fragrances, flavors, and medicines, and have potential as novel jet fuels. They also offer sustainable routes to terpene-based materials.[29,30] *Escherichia coli* possesses an endogenous MEP pathway, but the highest monoterpene titers have been reported for (*S*)-limonene in strains containing a recombinant MVA pathway, along with heterologous geranyl pyrophosphate synthase (GPPS) and limonene synthase (LimS) enzymes (435 mg/L and 600 mg/L in refs 31, 32, respectively). Balanced regulation of several steps in recombinant terpenoid pathways have been shown to be important due to either toxicity of intermediates including 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA), DMAPP, IPP, farnesyl pyrophosphate (FPP), and GPP[4,33] or substrate inhibition as seen for mevalonate kinase.[34] Additionally, expression of plant enzymes in bacteria is often poor, predominantly due to inefficient folding leading to insoluble protein and low *in vivo* activity, this has been seen for both GPPS and monoterpenoid synthases.[9,32,35] Recombinant MVA pathways in bacteria have therefore undergone extensive optimization *via* enzyme selection,[1,12] genetic architecture refinement including promoter selection and plasmid copy number,[4,31] growth conditions,[36,12] and post-transcriptional control,[37] but they have not, to date, been subjected to translational tuning by engineering of the translation initiation rate (TIR) of each gene within a pathway using RBS libraries.

The aim of this study was to implement machine learning using a training set of RBS sequence data from a subset of a large combinatorial biochemical pathway library to allow prediction of high performance and avoid the requirement for resource-intensive exhaustive screening. This approach was applied to optimize a nine-step MVA pathway to the monoterpenoid (*S*)-limonene in *E. coli* by introducing RBS libraries throughout the pathway at key genes with a novel monoterpenoid screening pipeline. Initial tests to tune the final 2 genes of the pathway (using a 144 member library) validated the approach and identified a library variant with more than 2-fold improvement in production titer over the original pathway. Expansion of this approach to the remainder of the MVA pathway (a 5184 member library) showed that screening less than 3% of this library provided sufficient data to apply machine learning to successfully predict an enriched library of high-producers and improved production over 3-fold.

■ RESULTS AND DISCUSSION

Development of an Automated Screening Pipeline. The introduction of RBS libraries at multiple points within a genetic pathway typically results in an intractable number of potential permutations to screen. To increase our screening capabilities, we developed an *in vivo* monoterpenoid

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

Table 1. RBS Library Sequences of Genes in the Limonene Production Pathway[a]

Library	Gene	RBS library sequence	Variants	Predicted TIR range (a.u.)	
				Low	High
pGLlib	*trAg-gpps*	ACGATCTTA**AGTARRCGVGGA**AAATAATG	12	613	86522
	trMs-limS	AAACTAAGC**ATCTAAGRGSGV**TACTAATG	12	191	68922
pMVA2lib1	*Ef-mvaE*	AGATCTTTT**AAGGAVGDAACG**TACATATG	9	2037	74594
	Ef-mvaS	GCGACAAAA**KATGAGGMGTRC**AAAAAATG	8	1156	86990
	Sp-mvaK1	CCATTTAAC**ACGHGASGAGGM**ATACGATG	12	1203	79074
	Ec-idi	CGAGACGCC**AAATWGGGAGGH**GGGCGATG	6	2513	61387
pMVA2lib2	*Ef-mvaE*	AGATCTTTT**AAGGAGGKAACG**TACATATG	2	41554	74594
	Ef-mvaS	GCGACAAAA**KATGAGGAGTRC**AAAAAATG	4	24672	86990
	Sp-mvaK1	CCATTTAAC**ACGMGAGGAGGM**ATACGATG	4	20496	79074
	Ec-idi	CGAGACGCC**AAATTGGGAGGA**GGGCGATG	1	21803	21803

[a]TIR, Translation initiation rate. Bases allowed to vary in bold, degenerate bases in red, cognate start codon underlined.

production pipeline (Figure S1) employing a multiwell format compatible with robotics platforms. Screening for the *in vivo* production of monoterpenoids in microorganisms is complicated by both their volatility and toxicity. These issues were partially addressed by using a biphasic growth medium with an organic phase (dodecane) to capture the product and remove it from the culture broth.[34] A standard two-phase shake-flask method was scaled down to a sealed, 96-deepwell plate (DWP) format. This required an increase in the proportion of dodecane (40% of culture broth volume) to allow accurate addition and removal by robotics and a robust seal to prevent evaporation of both limonene and dodecane during culture. This allowed the growth of 96 individual 1 mL cultures with growth monitoring, induction of pathway expression, organic phase extraction and processing for GC-MS analysis, with each operation performed simultaneously for all 96 samples. Sample processing (extraction, dehydration and dilution) was designed to be compatible for direct analysis using a GC-QTOF-MS equipped with a 96-well plate autosampler, which allowed the screening of 96 samples in under 14 h and was coupled with automated data extraction scripts for the simultaneous analysis of large data sets. This pipeline meant that three plates (288 samples) could be screened every 60 h, from colony to processed data. Experiments could also be multiplexed across robotic platforms to further increase throughput if required.

The pipeline was benchmarked in *E. coli* DH10β using an existing limonene-production pathway[9] (Figure 1B) and TB media and compared to the performance in 5 mL shake cultures. In this system, a previously described IPTG-inducible MVA pathway[31] was encoded on one plasmid (pMVA), while the aTet-inducible *trAg-gpps* and *trMs-limS* genes (Table S1) were on a second plasmid driven, allowing differential regulation of the two modules. Increased throughput allowed screening of multiple inducer concentrations (25 conditions) with 3 biological replicates in a single plate (75 wells). There was a clear differentiation of limonene titers between induction conditions (Figure S2), with production up to 8.0 ($\pm$ 0.2SD) mg/L_{org} compared with 350 ($\pm$ 7.6SD) mg/L_{org} observed at the 5 mL scale as seen previously. Testing the pathway (pJBEI6410) and conditions employed by Alonso-Gutierrez and colleagues[31] (*E. coli* DH1, 25 mL EZ Rich defined media in sealed, baffled flasks), we were able to achieve only 77 mg/L_{org} ($\pm$ 7.5SD), significantly lower than the published titers. This highlights some of the difficulties in metabolic engineering; even relatively small changes in conditions or methodologies can result in large differences in the performance and behavior of microbial fermentation. In this case the differences observed between the 1 and 5 mL scale is most likely due to the aeration state (DWP cultures had a gas impermeable seal and reduced headspace) and increased organic:aqueous ratio. Despite the lower titers compared to shake cultures, the differentiation between limonene titers under different induction conditions during optimization of inducer concentrations (Figure S2) shows that the DWP-based screening pipeline is effective and has the necessary resolution to measure relative expression levels between clones within an experiment such as library screening.

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

Translational Tuning of GPPS and LimS. Functional expression of trAg-GPPS and trMs-LimS is known to be a bottleneck for microbial limonene production, and so these two genes were targeted for initial expression optimization at the translational level *via* their RBSs (Figure 1B). 12-member RBS libraries spanning predicted TIRs over 2 orders of magnitude were designed for both genes (total 144 possible combinations) using the RBS Library Calculator[25] (Table 1) and introduced to plasmid pGL to create a library, pGLlib. This library was cotransformed into *E. coli* DH10β cells with plasmid pMVA to create a library of limonene production clones. 360 individual colonies were screened (the probability that all variants are represented, p = 0.92) for limonene production and titers were observed over 2 orders of magnitude, the highest producer displaying over 2-fold improvement over the original pGL plasmid control (Figure 1C). The average production was 0.8-fold of the pGL control, with 41.9% of the library producing more limonene than pGL. Plasmid DNA from the top 24 producers was recovered and retransformed into fresh cells, followed by triplicate colony screening (Figure S3), with 22 of the 24 plasmids again producing more limonene than the pGL control. The five top producers were then tested at the 5 mL scale where all displayed enhanced production with the highest titer from clone pGL403 of 809 mg/L_{org} (Figure 1D), a 1.9-fold improvement over the starting plasmid pGL. This demonstrated that our DWP screening pipeline was effective at identifying plasmids with improved production at scale-up.

Modeling of the Sequence−Phenotype Relationship. Our data shows that RBS tuning is clearly an effective method to improve production titers, presumably through balanced enzyme expression improving pathway flux. However, carrying out this process on more than two genes increases the combinatorial search space to intractable numbers for screening even with high-throughput methods. To address this, we sought to learn from the sequence−phenotype data, to determine whether library performance could be predicted from screening a small subset of library members. 64 members from the pGLlib were selected for sequencing from clones covering the full range of production titers (Table S2). 11/12 and 12/12 of the potential RBS sequences were represented in at least one clone for *trAg-gpps* and *trMs-limS*, respectively, and 56/144 (39%) of all possible RBS combinations were observed. RBS sequences for *trMs-limS* displayed clear differences between high and low producers, whereas this distinction was less pronounced for *trAg-gpps* (Figure S4). This evidence supported the theory that the monoterpenoid synthase is the major bottleneck for limonene production. In contrast, trAg-GPPS expression did not appear to be a limiting factor. To investigate the relationship between the RBS sequence and limonene production, the data were used to train a supervised machine learning algorithm toward the development of a model that predicts limonene titers from RBS sequences. Every sequence was represented by a binary vector where each element corresponds to one of the possible bases in the RBS positions. The actual bases of each sequence were set to "1" and the rest of elements of the vector to "0" (one-hot encoding). Such representation allowed the model to learn the relationship between the relative frequencies of the variable base positions and the production titers (Figure 2). The resulting model accurately predicted the levels of production with a Q^2 = 0.87 (leave-one-out cross-validation; Figures S5, Table S3). There were some outliers in the predictions, but this may be due to biological/technical reasons (*e.g.*, sequence mutations outside of the RBS regions) or the relatively small sample set. There were also 6 predicted top producers that were not identified by sequencing, although the coverage of the library screen would mean these have probably been tested. The top experimentally tested producer (pGL403) was predicted to be the seventh best producer. The predictive success of the model trained on just a subset of the total library is encouraging and suggests that this approach could be extended to much larger library sizes allowing for a significant reduction in the extent of experimental screening.

Translational Tuning of an MVA Pathway. The pMVA pathway has already undergone extensive optimization, as already discussed. We therefore sought to construct new, untested, fully refactored MVA pathways to test the extent to which translational tuning could optimize pathway flux using a machine learning approach. A hybrid all-bacterial MVA pathway was designed and assembled into four genetic architectures similar to that of the original pMVA plasmid, with varying strong (P_{trc}) and weak (P_{lacUV5}) IPTG-inducible promoters (Figure S6). Each MVA pathway was tested in combination with pGL403 and all produced limonene but at lower titers than observed for pMVA (Figure S7), and so these plasmids were ideal for testing the impact of translational tuning. Of the four plasmids, pMVA2 and pMVA4 produced the similar high levels of limonene, whereas pMVA3 and pMVA5 were both poor producers. Between pMVA2 and pMVA4 we selected pMVA2 for further translational tuning with the rationale that it has a second promoter boosting the latter half of the pathway and would result in higher levels of transcript, which could potentially allow a wider range of TIRs for the transcribed genes.

Four of the six genes in pMVA2 were selected as targets for translational tuning to improve flux through the pathway and to alleviate toxicity. *Ef-mvaS* (HMGS), *Ef-mvaE* (ACAT/HMGR) *and Sp-mvaK1* (MK) are critical for commitment of acetyl-CoA into the MVA pathway and for control of flux around HMG-CoA/mevalonate (substrate inhibition, intermediate toxicity), whereas *Ec-idi* (IDI) is important for balancing the availability of IPP and DMAPP (flux and toxicity).

RBS libraries were designed for each of the four genes (Table 1) and built as a pMVA2 RBS library (pMVA2lib1; Figure 3) using the same process as for pGLlib. Rather than exhaustively screening the library just 156 colonies were screened. 17% of clones produced more limonene than pMVA2, up to 1.7-fold of that from pMVA2 (Figure 3). 39 individual clones, covering a range of limonene production from low to high, were sequenced and showed a good coverage for the expected variants at each RBS (9/9, 11/12, 7/8, and 6/6 possible variants observed for *Ef-mvaE*, *Ef-mvaS*, *Sp-mvaK1*, and *Ec-idi*, respectively) (Table S4), with clear enrichment patterns for certain RBS sequences when comparing high to low producers (Figure S8). These data suggested that our library could indeed provide a quality training set for building a machine-learning model. Validation of the predicted model showed a strong correlation between the sequence predictions and observed limonene production levels (Figure S9, Table S3). We then used the model to predict the response of all 5184 possible RBS library variants and analyzed the top 100 predicted RBS combinations (∼2% of the total library). We focused on nucleotide bases that appear at least in 20% of these 100 predicted best producers. These criteria drastically

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

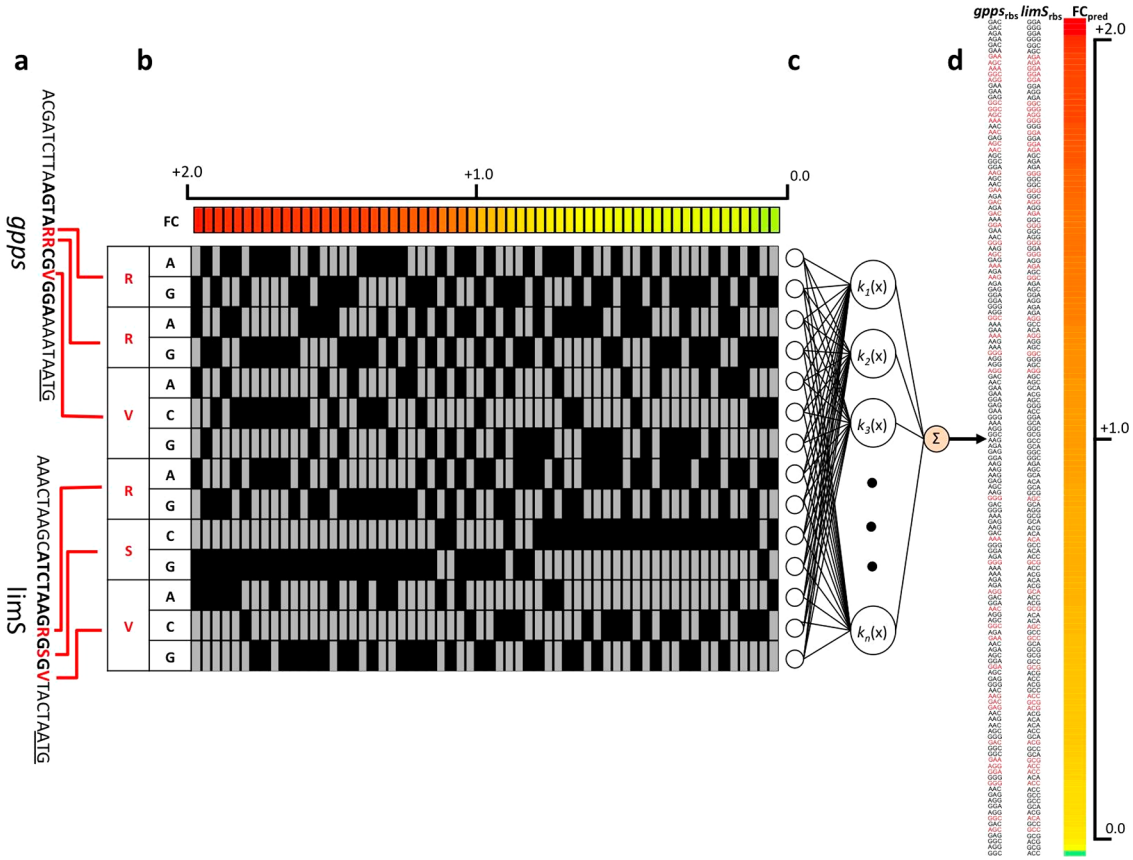


Figure 2. Predictive modeling process of the RBS libraries. Shown are (a) the two RBS libraries from pGLlib, (b) a visual representation of all combinations, (c) followed by the predictive model composed of an input layer, a middle layer, which in the case of a support vector regression consists of the kernel functions $k_i(x)$, and in the case of a feedforward neural network is the hidden layer with activation or transfer functions $k_i(x)$, and an output layer Σ, and (d) a predicted titer for each sequence as a fold-change relative to plasmid pGL. Sequenced and tested library members are shown in red.

reduced the number of RBS sequences associated with predicted high limonene production to just 2, 4, 4, and 1 variants for *Ef-mvaE*, *Ef-mvaS*, *Sp-mvaK1*, and *Ec-idi*, respectively, indicating that most of the top predictions were focused on a small subset of the library (32 combinations). The reduced library encompassing all 32 RBS combinations was selected as potential high producers (Table 1).

The 32-member predicted library (pMVA2lib2) was built and tested for limonene production in the presence of pGL403, and the results compared to those for pMVA2lib1 and the original pMVA2 plasmid. The new pMVA2lib2 library displayed a clear enrichment for high producers, with all functional clones displaying improved limonene production relative to pMVA2 (Figure 3B). In the highest producers, up to double the limonene production compared to that from pMVA2 was observed and included clones with higher production than the best members of pMVA2lib1. Plasmid DNA from the top 23 producers from pMVA2lib2, plus the top producer from the first library (pMVA2RBS041), were reintroduced to fresh cells. Triplicate screening confirmed their high production levels, with over 3-fold higher production relative to pMVA2 (Figure S10). However, pMVA2RBS041 was the second highest producer in this rescreening. This data clearly shows that the "machine learnt" library was successfully enriched in high producers of limonene but did not allow the discovery of any significantly higher producers than observed during screening of library pMVA2lib1, suggesting that a good representation of the whole library was observed.

The top five producers from library pMVA2lib1 were then grown at the 5 mL scale but only pMVA2RBS041 and pMVA2RBS314 retained limonene titers above the original pMVA2 plasmid (1.6-fold compared with over 3-fold higher production observed at the 1 mL scale), while the other library clones performed no better than pMVA2 (Figure 3B). This showed that, unlike for pGL, results from RBS tuning of the MVA pathway did not reliably translate during the scale-up from the DWP, 1 to 5 mL shake cultures. This suggests that the gene expression levels and cellular metabolism are significantly different under these two growth conditions, such that the requirements for balancing of the pathway flux are correspondingly different. Rescreening of library pMVA2-lib1 at the 5 mL scale may allow selection of clones which have improved production under these conditions. The number of clones that were screened from library pMVA2lib1 to successfully identify high producers at the 1 mL scale was tractable for manual screening, and so we rescreened this library at the 5 mL scale. The library spanned a range of limonene titers with 68% of clones producing more limonene than pMVA2 (compared to 17% of clones from the 1 mL scale screen of this library) with the highest single clone providing a 3.2-fold improvement (compared with 1.7-fold for the best producer at 1 mL scale). The top producers were rescreened and consistently produced high titers (2-fold or greater)

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

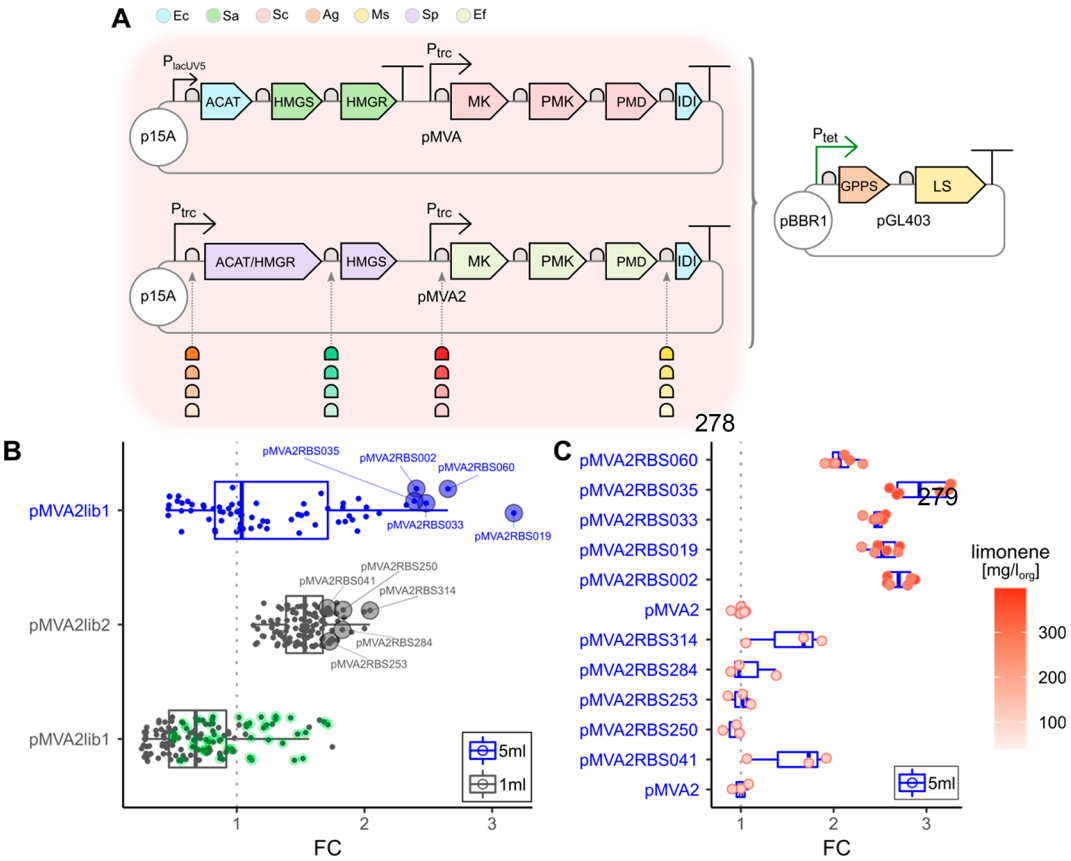


Figure 3. Translational tuning of an MVA pathway encoded on plasmid pMVA2. (A) Schematic of the pMVA and pMVA2 plasmids encoding the MVA pathway and pGL403, a translationally tuned variant from pGLlib. The pMVA2lib library contains variable RBS sequences for the four genes indicated. The color code of each gene indicates the source organism (Table S1). (B) Library pMVA2lib1 and learnt library pMVA2lib2 were screened for limonene production in either DWP as 1 mL cultures (gray) or individual tubes as 5 mL cultures (blue) and limonene production shown as fold-change (FC) relative to the original pMVA2 production levels (1 mL = 4.5 mg/L_{org}; 5 mL = 145 mg/L_{org}). Green highlight indicates clones which were sequenced, individual labeled clones are displayed in panel C. (C) Rescreening of high producing clones from pMVA2lib1/2 grown in triplicate in 5 mL individual cultures. High-producing variants of pMVA2 identified during 5 mL screening produced more limonene than those identified from 1 mL screening.

relative to pMVA2 (Figure 3C). Comparison of production titers of individual clones grown at both the 1 and 5 mL scales showed there was a positive association between the two scales, but the correlation is rather weak as indicated by a nonsignificant (p = 0.218) Pearson correlation coefficient r = 0.274 (Figure S11). A comparison of the RBS sequences enriched in the high and low limonene producers showed clear differences between the 1 and 5 mL screens, with the differences particularly pronounced for the *Sp-mvaK1* gene (encoding mevalonate kinase) in high producers (Figure S12). The activity of this enzyme is known to be key for optimizing pathway flux, due to the toxicity of its substrate and substrate inhibition. The importance of balancing expression of this gene clearly differs under the two different growth conditions.

The DWP-fermentation was not a reliable screening method for pMVA2 tuning, and so the key tuned pathways (pMVA/pGL403, pMVA2/pGL403, and pMVA2RBS035/pGL403) were further screened as 25 mL cultures in baffled flasks to check their robustness. Additionally, to allow comparison to previous limonene production pathways, plasmid pJBEI6410[31] was also included in the screen. The strain and the media were also varied to match the 1 and 5 mL experiments (DH10β/TB) or previous work[31] (DH1/EZ Rich) and significantly different titers were observed in each condition (Figure S13). The highest titers for each pathway were observed using *E. coli* DH10β grown in TB media but the same relative pattern of production titers between the pathways in each condition was maintained to match the previous screens. The pMVA/pGL403 combination consistently outperformed pathway pJBEI6410 by at least 1.5-fold with the highest titers of 1151 mg/L_{org} (± 45 SD) (in comparison to 593 mg/L_{org} (± 30 SD) from pJBEI6410) observed with DH10β grown in TB media.

■ CONCLUSIONS

Here we have presented, to our knowledge, the first example of machine learning applied to RBS DNA sequences for the predictive enrichment of high-producing recombinant biochemical pathways in bacteria. The modulation of pathway flux by balancing TIRs has previously been shown to be effective at improving production titers from microbial chassis; however, for multigene pathways the number of potential library variants becomes intractable for screening without a powerful high-throughput method (colorimetric, fluorescence-based *etc.*). Our dual approach using automated screening in combination with machine learning allows increased screening throughput with a significant reduction of the design space. Despite being faced with difficulties during scale-up, this method to learn from a small data set has allowed us to significantly improve production titers for our monoterpenoid production pathways by over 50%. The reduction in the required screening was

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

essential in this case because of the technical difficulties of screening limonene production due to the volatility of limonene. We see this as a generic approach that could be routinely applied as a tool for translational tuning of pathways for which it is difficult to develop high-throughput screens.

METHODS

Bacterial Strains and Media. *Escherichia coli* DH10β (New England Biolabs) was used for both cloning and (*S*)-limonene production experiments. Strains were maintained on Lysogeny Broth (LB) or LB agar containing ampicillin (100 μg/mL) and/or kanamycin (50 μg/mL).

Plasmid Design. Selected enzymes are listed in Table S1. Refactoring of the MVA pathway was achieved by codon optimization of each ORF for expression in *E. coli* (GeneArt) followed by scanning with the RBS Calculator (v2.0) Predict function.[13,39] Any potential mid-ORF translational start sites were removed if scoring >1000 arbitrary units (au). Subsequently the "Design" function was used to produce bespoke 5′ UTRs containing a RBS with a target translation rate of 15,000 au for each ORF, using the 20 bp immediately upstream as "Presequence" (except *Ef-mvaE*, which used the RBS from the destination vector). These blocks of ORFs with associated 5′ UTRs were then assembled *in silico* to produce the final design and the full construct was synthesized by GeneArt (Thermo Scientific). The MVA construct was designed to include an NdeI restriction site at the 5′ end (overlapping the *Ef-mvaE* start codon) and a T7 terminator and XhoI restriction site at the 3′ end. The pMVA pathway was then subcloned (NdeI-XhoI) into pBbA1k-RFP and pBbA5k-RFP[6] replacing the *rfp* gene to create plasmids pMVA4 and pMVA5, respectively (Table S6). A P_{trc} promoter element was introduced between the *mvaE* and *mvaK1* genes by InFusion (Takara) using primers mvaK1trc-F and mvaK1trc-R (Table S5). This linearized and reannealed the plasmid to create pMVA2 and pMVA3 (Table S6).

RBS Variant Library Design. RBS variant libraries were designed for each ORF in both the pMVA2 and pGL plasmids using the RBS Library Calculator v2.0 function.[13,39] For each ORF, a library of RBSs was designed across a translation initiation rate range of 1000 to 100 000 arbitrary units (au), with a target library size of "4−20". Sequence parameters were as follows: 20 bp "Presequence", and the existing 5′ UTR allowing bases −6 to −17 relative to the start codon defined as "N" (any base) as the "Initial RBS sequence". Each variant library is encoded by a single sequence containing degenerate nucleotides,[40] and libraries are listed in Table 1, along with the number of variants that each library contains.

Introduction of RBS Libraries. Libraries were introduced into plasmids by fragmenting each *via* PCR such that RBS sites were at the 5′ termini of the PCR products, followed by reassembly using the ligase cycling reaction (LCR).[40] PCR primers were designed to contain degenerate bases encoding the RBS libraries. The pGL library was designed as three parts corresponding to the *trAg-gpps*, *trMs-limS*, and vector backbone (GL-P1−3), which were amplified using primer pairs GPPSRBS-F/R, limSRBS-F/R, and pGL-F/R, and then annealed using bridging oligos brGL, brLP, and brPG (Table S6). The pMVA2 library was split into 8 parts (MVAlib1.1−3, MVAlib2.1−2, MVAlib3, and MVAlib4.1−2), amplified with primer pairs prefixed with MVAlib, then annealed with bridging oligos prefixed with brMVAlib (Table S5). LCR reactions were performed as previously described.[41] The pGL library was screened by colony PCR using primers pBbseq-R2 and tetprom to check insert size (95% correct colonies) followed by sequencing of the RBS sequences for 12 clones all showing library sequences. The pMVA2 library was screened by restriction analysis and showed >75% correct assembly.

Limonene Production. Limonene production experiments were performed at three scales; 1 mL cultures in DWPs, 5 mL cultures in 50 mL round-bottom centrifuge tubes as previously published.[9] Fresh overnight colonies from transformations were inoculated into Terrific Broth (TB) with 0.4% glucose (no glycerol) and appropriate antibiotics in 96-deep-well plates (using automated colony picking for plates), or in centrifuge tubes and grown at 37 °C with shaking at 1000 rpm (DWP) or 180 rpm (tubes), respectively. Once the optical absorbance at 600 nm (A_{600}) reached ∼0.2, the temperature was reduced to 30 °C. At A_{600} ∼ 0.6 cultures were induced by the addition of Isopropyl-β-D-thiogalactoside (IPTG; 25 μM) and anhydrotetracycline (aTet; 200 nM for plate growth, or 50 nM for 5 mL cultures), where appropriate, and overlaid with dodecane (40% for plate growth, or 20% for 5 mL cultures) then incubated for a further 24h at 30 °C with shaking. A_{600} readings were taken at the end of fermentation and organic phases were separated by centrifugation (3500*g*, 10 min). Organic phases were removed into fresh tubes and diluted 1:1 (1 mL) or 1:20 (5/25 mL) with dodecane containing *sec*-butylbenzene (final concentration 0.005%) and dried over anhydrous $MgSO_4$ before centrifuging again. Clarified organic phases were analyzed for limonene content. For plate-based production, all steps were automated using a Hamilton Star robotics platform fitted with both 8- and 96-head liquid handling capability with pressure sensing liquid detection allowing efficient and reproducible extraction of the organic phase from bacterial cell cultures. An integrated Cytomat storage unit and ClarioStar plate reader allowed online monitoring of cell cultures prior to induction and at harvest. To prevent evaporation of the organic phase all 96-well culture plates were sealed postinduction with an ALPS 3000 thermo-sealer.

For production in 250 mL, sealed, baffled Erlenmeyer flasks, production was carried out essentially as previously reported.[31] Fresh transformants were inoculated in either 5 mL EZ Rich defined media (Teknova, CA) with 1% glucose or TB with 0.4% glycerol and grown overnight with shaking at 200 r.p.m. overnight at 30 °C. Cultures were subcultured to OD_{600} = 0.1 and incubated at 30 °C until the culture reached OD_{600} = 1 whereupon they were induced with 25 μM IPTG ± 50 nM aTet and a 20% dodecane overlay. Cultures were further incubated for 72 h before sampling, as above for limonene quantification.

Limonene Quantification. Limonene within the organic phase was detected and analyzed on an Agilent Technologies 7200 accurate mass Q-TOF mass spectrometer coupled to a 7890B GC and equipped with a PAL RSI 85 autosampler. The sample (1 μL) was injected onto a VF-5 ms column (30 m × 250 μm × 0.25 μm; Agilent Technologies) with an inlet temperature of 280 °C and a split ratio of 100:1. Helium was used as the carrier gas with a flow rate of 1.5 mL/min and a pressure of 16.2 psi. The chromatography was programmed to begin at 100 °C with a hold time of 1 min, followed by an increase to 160 °C at a rate of 50 °C/min, then a subsequent increase to 325 °C at a rate of 120 °C/min and a final hold time of 1 min. The total runtime per analysis was 4.6 min. The MS was equipped with an electron impact ion source using 70

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

eV ionization and a fixed emission of 35 μA. The mass spectrum was collected for the range of 35−500 m/z with an acquisition rate of 5 spectra/s and an acquisition time of 200 ms/spectrum. The PAL RSI autosampler permitted the analysis of samples from 96-well plates and allowing for high-throughput analysis.

For all GC-MS analyses, *sec*-butylbenzene (Sigma) was used as an internal standard to allow for accurate quantification. 0.005% *sec*-butylbenzene was added to all samples and the quantification of limonene was calculated relative to the peak area of this internal standard. In addition an eight-point calibration curve was constructed in the range of 0−200 mg/L of (*S*)-limonene. A calibration curve was analyzed before and after the analysis of the samples and an average standard curve generated.

Q-TOF vendor binary files were converted to open source mzXML data format[42] using ProteoWizard msConvert.[43], Automated limonene quantitation was conducted using in-house scripts written in R to automatically extract relevant peak areas. Quantification of limonene was conducted by extrapolating the limonene/*sec*-butylbenzene peak area against the generated standard curve.

Reported titers are shown as either "fold-change" (FC) relative to a control, or as an absolute concentration in the organic phase (mg/L_{org}).

Machine Learning. RBS sequences for each ORF in a plasmid and their associated titers were compiled and used as a training set for machine learning. To that end, we used one-hot encoding of DNA sequences to form the input feature training set,[44] where each component in the vector corresponds to a different nucleotide base at a specific position in the RBS site with at least two variants in the training set. The training algorithm used for the pGL library was support vector regression with anovadot kernel using the kernlab R library.[45] For the pMVA2 library, an initial model was trained based on support vector regression with polydot kernel. Initial validation statistics of the models for the two libraries were performed based on permutation tests (1000 runs) by computing the coefficient of determination of prediction and of the 10-fold cross-validation. On the basis of the performance observed for the support vector regression model in this initial assessment, we trained a final model for the pMVA2 library based on a feedforward neural network of fully connected (Dense) layers with two hidden layers of 128 nodes each, using sigmoid and ReLU activation with default uniform initialization built using the Python deep learning package Keras.[46] Performance assessment of the final models was performed on the basis of the coefficient of determination of a leave-one-out cross-validation.

Source code and data are available at http://github.com/synbiochem/opt-mva/ under the MIT license.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acssynbio.8b00398.

Figures S1−S13; Tables S1, S3, S5, S6 (PDF)

Tables S2, S4 (XLSX)

Github software repository (ZIP)

Plasmid sequences (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +44 (0) 161 306 5152. E-mail: nigel.scrutton@manchester.ac.uk.

ORCID

Cunyu Yan: 0000-0002-3603-2421
Neil Swainston: 0000-0001-7020-1236
Helen Toogood: 0000-0003-4797-0293
Rainer Breitling: 0000-0001-7173-0922
Eriko Takano: 0000-0002-6791-3256
Nigel S. Scrutton: 0000-0002-4182-3500

Author Contributions

A.J.J., P.C., M.V., M.D., K.H., C.J.R., N.R., A.C., C.Y., and H.T. conceived the experimental design. A.J.J., M.V., M.D., K.H., C.J.R., N.R., C.Y., R.S., S.T. performed the experimental work. P.C. and N.S. developed and performed all machine learning and flux modeling. A.J.J., P.C., M.V., C.J.R., M.D., K.H., J.L.F., R.B., E.T., N.S.S. compiled and commented on the manuscript. A.J.J., R.B., E.T., N.S.S. supervised the project.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was funded by the UK Biotechnology and Biological Sciences Research Council (BBSRC BB/M017702/1, BB/L027593/1, BB/M000354/1, BB/J015512/1). J.L.F. acknowledges funding provided by French National Research Agency under grant ANR-15-CE1-0008. N.S.S. received funding as a Royal Society Wolfson Merit Award holder and is an Engineering and Physical Sciences Research Council (EPSRC; EP/J020192/1) Established Career Fellow.

■ REFERENCES

(1) Yoon, S. H., Lee, S. H., Das, A., Ryu, H. K., Jang, H. J., Kim, J. Y., Oh, D. K., Keasling, J. D., and Kim, S. W. (2009) Combinatorial expression of bacterial whole mevalonate pathway for the production of beta-carotene in E. coli. *J. Biotechnol. 140*, 218−226.

(2) Kim, B., Du, J., Eriksen, D. T., and Zhao, H. (2013) Combinatorial design of a highly efficient xylose-utilizing pathway in Saccharomyces cerevisiae for the production of cellulosic biofuels. *Appl. Environ. Microbiol. 79*, 931−941.

(3) Fehér, T., Planson, A. G., Carbonell, P., Fernández-Castané, A., Grigoras, I., Dariy, E., Perret, A., and Faulon, J. L. (2014) Validation of RetroPath, a computer-aided design tool for metabolic pathway engineering. *Biotechnol. J. 9*, 1446−1457.

(4) Martin, V. J., Pitera, D. J., Withers, S. T., Newman, J. D., and Keasling, J. D. (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol. 21*, 796−802.

(5) Ajikumar, P. K., Xiao, W. H., Tyo, K. E., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T. H., Pfeifer, B., and Stephanopoulos, G. (2010) Isoprenoid pathway optimization for Taxol precursor overproduction in Escherichia coli. *Science 330*, 70−74.

(6) Lee, T. S., Krupa, R. A., Zhang, F., Hajimorad, M., Holtz, W. J., Prasad, N., Lee, S. K., and Keasling, J. D. (2011) BglBrick vectors and datasheets: A synthetic biology platform for gene expression. *J. Biol. Eng. 5*, 12.

(7) Smanski, M. J., Bhatia, S., Zhao, D., Park, Y., Woodruff, L. B., Giannoukos, G., Ciulla, D., Busby, M., Calderon, J., Nicol, R., and Gordon, D. B. (2014) Functional optimization of gene clusters by combinatorial design and assembly. *Nat. Biotechnol. 32*, 1241−1249.

(8) 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.,

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

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., and Scrutton, N. S. (2018) An automated Design-Build-Test-Learn pipeline for enhanced microbial production of fine chemicals. *Commun. Biol. 1*, 66.

(9) Leferink, N. G., Jervis, A. J., Zebec, Z., Toogood, H. S., Hay, S., Takano, E., and Scrutton, N. S. (2016) A 'plug and play' platform for the production of diverse monoterpene hydrocarbon scaffolds in Escherichia coli. *ChemistrySelect. 1*, 1893−1896.

(10) Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., Herrgård, M. J., and Feist, A. M. (2016) Multi-omics Quantification of Species Variation of Escherichia coli Links Molecular Features with Strain Phenotypes. *Cell. Syst. 28*, 238−251.

(11) Cao, W., Ma, W., Wang, X., Zhang, B., Cao, X., Chen, K., Li, Y., and Ouyang, P. (2016) Enhanced pinocembrin production in Escherichia coli by regulating cinnamic acid metabolism. *Sci. Rep. 6*, 32640.

(12) Tsuruta, H., Paddon, C. J., Eng, D., Lenihan, J. R., Horning, T., Anthony, L. C., Regentin, R., Keasling, J. D., Renninger, N. S., and Newman, J. D. (2009) (2009) High-level production of amorpha-4, 11-diene, a precursor of the antimalarial agent artemisinin, in Escherichia coli. *PLoS One 4*, e4489.

(13) 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.

(14) Na, D., and Lee, D. (2010) RBSDesigner: software for designing synthetic ribosome binding sites that yields a desired level of protein expression. *Bioinformatics 26*, 2633−2634.

(15) Seo, S. W., Yang, J. S., Kim, I., Yang, J., Min, B. E., Kim, S., and Jung, G. Y. (2013) Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng. 15*, 67−74.

(16) Ng, C. Y., Farasat, I., Maranas, C. D., and Salis, H. M. (2015) Rational design of a synthetic Entner−Doudoroff pathway for improved and controllable NADPH regeneration. *Metab. Eng. 29*, 86−96.

(17) Jeschek, M., Gerngross, D., and Panke, S. (2016) Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. *Nat. Commun. 7*, 11163.

(18) Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018) Next-generation machine learning for biological networks. *Cell 173*, 1581−1592.

(19) Abbas, M. M., Mohie-Eldin, M. M., and Yasser, E. M. (2015) Assessing the effects of data selection and representation on the development of reliable E. coli sigma 70 promoter region predictors. *PLoS One 10*, e0119721.

(20) Meng, H., Ma, Y., Mai, G., Wang, Y., and Liu, C. (2017) Construction of precise support vector machine based models for predicting promoter strength. *Quant. Biol. 5*, 90−98.

(21) Yip, K. Y., Cheng, C., and Gerstein, M. (2013) Machine learning and genome annotation: a match meant to be? *Genome Biol. 14*, 205.

(22) Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol. 33*, 831−838.

(23) Paddon, C. J., Westfall, P. J., Pitera, D. J., Benjamin, K., Fisher, K., McPhee, D., Leavell, M. D., Tai, A., Main, A., Eng, D., and Polichuk, D. R. (2013) High-level semi-synthetic production of the potent antimalarial artemisinin. *Nature 496*, 528−532.

(24) George, K. W., Alonso-Gutierrez, J., Keasling, J. D., and Lee, T. S. (2015) Isoprenoid drugs, biofuels, and chemicals—artemisinin, farnesene, and beyond. In *Biotechnology of Isoprenoids*, pp 355−389, Springer International Publishing, Cham, Switzerland.

(25) Andersen-Ranberg, J., Kongstad, K. T., Nielsen, M. T., Jensen, N. B., Pateraki, I., Bach, S. S., Hamberger, B., Zerbe, P., Staerk, D., Bohlmann, J., and Møller, B. L. (2016) Expanding the landscape of diterpene structural diversity through stereochemically controlled combinatorial biosynthesis. *Angew. Chem., Int. Ed. 55*, 2142−2146.

(26) Westfall, P. J., Pitera, D. J., Lenihan, J. R., Eng, D., Woolard, F. X., Regentin, R., Horning, T., Tsuruta, H., Melis, D. J., Owens, A., and Fickes, S. (2012) Production of amorphadiene in yeast, and its conversion to dihydroartemisinic acid, precursor to the antimalarial agent artemisinin. *Proc. Natl. Acad. Sci. U. S. A. 109*, E111−E118.

(27) Lv, X., Wang, F., Zhou, P., Ye, L., Xie, W., Xu, H., and Yu, H. (2016) Dual regulation of cytoplasmic and mitochondrial acetyl-CoA utilization for improved isoprene production in Saccharomyces cerevisiae. *Nat. Commun. 7*, 12851.

(28) Xie, W., Lv, X., Ye, L., Zhou, P., and Yu, H. (2015) Construction of lycopene-overproducing Saccharomyces cerevisiae by combining directed evolution and metabolic engineering. *Metab. Eng. 30*, 69−78.

(29) Messiha, H. L., Ahmed, S. T., Karuppiah, V., Suardíaz, R., Ascue Avalos, G. A., Fey, N., Yeates, S., Toogood, H. S., Mulholland, A. J., and Scrutton, N. S. (2018) Biocatalytic routes to lactone monomers for polymer production. *Biochemistry 57* (13), 1997−2008.

(30) Thomsett, M. R., Storr, T. E., Monaghan, O. R., Stockman, R. A., and Howdle, S. M. (2016) Progress in the synthesis of sustainable polymers from terpenes and terpenoids. *Green Mater. 4* (3), 115−134.

(31) Alonso-Gutierrez, J., Chan, R., Batth, T. S., Adams, P. D., Keasling, J. D., Petzold, C. J., and Lee, T. S. (2013) Metabolic engineering of Escherichia coli for limonene and perillyl alcohol production. *Metab. Eng. 19*, 33−41.

(32) 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., and Martin, H. G. (2015) Principal component analysis of proteomics (PCAP) as a tool to direct metabolic engineering. *Metab. Eng. 28*, 123−133.

(33) Pitera, D. J., Paddon, C. J., Newman, J. D., and Keasling, J. D. (2007) Balancing a heterologous mevalonate pathway for improved isoprenoid production in Escherichia coli. *Metab. Eng. 9*, 193−207.

(34) Ma, S. M., Garcia, D. E., Redding-Johanson, A. M., Friedland, G. D., Chan, R., Batth, T. S., Haliburton, J. R., Chivian, D., Keasling, J. D., Petzold, C. J., and Lee, T. S. (2011) Optimization of a heterologous mevalonate pathway through the use of variant HMG-CoA reductases. *Metab. Eng. 13*, 588−597.

(35) Carter, O. A., Peters, R. J., and Croteau, R. (2003) Monoterpene biosynthesis pathway construction in Escherichia coli. *Phytochemistry 64*, 425−433.

(36) Newman, J. D., Marshall, J., Chang, M., Nowroozi, F., Paradise, E., Pitera, D., Newman, K. L., and Keasling, J. D. (2006) High level production of amorpha-4, 11-diene in a two-phase partitioning bioreactor of metabolically engineered Escherichia coli. *Biotechnol. Bioeng. 95*, 684−691.

(37) Pfleger, B. F., Pitera, D. J., Smolke, C. D., and Keasling, J. D. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. *Nat. Biotechnol. 24*, 1027−1032.

(38) Farasat, I., Kushwaha, M., Collens, J., Easterbrook, M., Guido, M., and Salis, H. M. (2014) Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria. *Mol. Syst. Biol. 10*, 731.

(39) 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.

(40) Cornish-Bowden, A. (1985) Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations. *Nucleic Acids Res. 13*, 3021−3030.

(41) Kok, S. D., Stanton, L. H., Slaby, T., Durot, M., Holmes, V. F., Patel, K. G., Platt, D., Shapland, E. B., Serber, Z., Dean, J., and Newman, J. D. (2014) Rapid and reliable DNA assembly *via* ligase cycling reaction. *ACS Synth. Biol. 3*, 97−106.

(42) Pedrioli, P. G., Eng, J. K., Hubley, R., Vogelzang, M., Deutsch, E. W., Raught, B., Pratt, B., Nilsson, E., Angeletti, R. H., Apweiler, R., and Cheung, K. (2004) A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol. 22* (11), 1459.

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136

(43) Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., Baker, T. A., Brusniak, M. Y., Paulse, C., Creasy, D., Flashner, L., Kani, K., Moulding, C., Seymour, S. L., Nuwaysir, L. M., Lefebvre, B., Kuhlmann, F., Roark, J., Rainer, P., Detlev, S., Hemenway, T., Huhmer, A., Langridge, J., Connolly, B., Chadick, T., Holly, K., Eckels, J., Deutsch, E. W., Moritz, R. L., Katz, J. E., Agus, D. B., MacCoss, M., Tabb, D. L., and Mallick, P. (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol. 30*, 918−920.

(44) Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016) Deep learning for computational biology. *Mol. Syst. Biol. 12*, 878.

(45) Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004) kernlab - An S4 Package for Kernel Methods in R. *J. Stat. Softw. 11*, 1−20.

(46) Chollet, F. (2015) Keras: The Python Deep Learning Library, https://github.com/fchollet/keras.

DOI: 10.1021/acssynbio.8b00398
ACS Synth. Biol. 2019, 8, 127−136