ACS Synthetic Biology

pubs.acs.org/synthbio

Research Article

PASIV: A Pooled Approach-Based Workflow to Overcome Toxicity-Induced Design of Experiments Failures and Inefficiencies

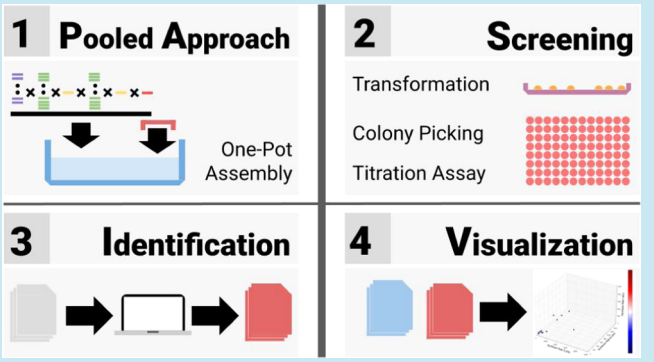
Alexis Casas, Matthieu Bultelle, Charles Motraghi, and Richard Kitney*

Cite This: *ACS Synth. Biol.* 2022, 11, 1272–1291

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present here a newly developed workflow—which we have called PASIV—designed to provide a solution to a practical problem with design of experiments (DoE) methodology: i.e., what can be done if the scoping phase of the DoE cycle is severely hampered by burden and toxicity issues (caused by either the metabolite or an intermediary), making it unreliable or impossible to proceed to the screening phase? PASIV—standing for pooled approach, screening, identification, and visualization—was designed so the (viable) region of interest can be made to appear through an interplay between biology and software. This was achieved by combining multiplex construction in a pooled approach (one-pot reaction) with a viability assay and with a range of bioinformatics tools (including a novel construct matching tool). PASIV was tested on the exemplar of the lycopene pathway—under stressful constitutive expression—yielding a region of interest with comparatively stronger producers.



KEYWORDS: *design of experiments, scoping, pooled approach, construct matching*

INTRODUCTION

Dealing with Large Combinatorial Spaces: Review. To optimize the output of a metabolic pathway, there are many design parameters that one can vary when engineering the pathway, such as the variation of its coding sequences, the engineering of expression levels by gene dosage, or transcriptional and translational engineering.[1] This results in a very large combination of design parameters. Combinatorial design can be achieved through different means, either brute force where all of the possible combinations are made, which is limited by the number of combinations and permutations of parts from the metabolic pathway, also known as the problem of combinatorial explosion or by applying combinatorial strategies.[2] When optimizing a metabolic pathway with a combinatorial approach, one of the main issues is scale and how to deal with large design spaces. For example, consider the violacein pathway (five genes)[3] and an operon design—yielding 10^5 possible combinations for even a small library of 10 RBS. Varying more components in the design (e.g., the promoter driving the operon or adding degradation tags to the enzyme to control their expression) expands the design space by further orders of magnitude and easily reaches millions of potential combinations. Such large spaces then become prohibitively expensive and difficult to investigate with standard construction and analysis methods.

It is desirable to gain as much information as possible from a minimal number of experiments.[4] In recent years, design of experiments (DoE) has become a popular method[5] and is used for different applications to design and optimize synthetic biology systems. DoE techniques are employed for protocol development such as optimizing media and culture conditions for a maximum yield of the metabolic pathway,[6,7] optimizing a transformation protocol,[8] or the optimization of a cell-free system.[9] DoE techniques have also been applied recently to the optimization of metabolic pathways. Cis-regulatory elements in pathways such as promoters and RBS can be selected to optimize their outputs. Recent advances in metabolic engineering have made use of DoE techniques for pathway simplification.[10] Others leverage the combination of DoE and high-throughput automation platform by developing an automated design–build–test–learn (DBTL) cycle[11] or utilizing full-featured DNA foundry platforms.[12]

The standard theoretical textbook DoE workflow as described by Gilman et al.[5] is grouped in three categories of iterative experiments: scoping, screening, and optimization. The scoping experiment is used to identify the region of interest. The screening experiments identify the most significant factors, and

Received: November 3, 2021
Published: March 9, 2022

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272–1291

finally, the optimization stage consists of fine-tuning the factors to optimize the response.

In this work, we focus on the scoping phase of the DoE cycle when applied to the optimization of the production of a metabolite over a large combinatorial design space of constructs. In particular, we focus on the case when finding an adequate region of space to explore, screen, and optimize over is difficult due to toxicity and burden issues. To this end, we have developed a novel scoping approach, which we called pooled approach, screening, identification, and visualization (PASIV).

Lycopene Exemplar. For the development of the novel approach, we have used a popular pathway in metabolic engineering: the lycopene pathway. To create the necessarily challenging conditions, we will consider its production under constitutive expression.

Lycopene, a naturally produced bright red pigment, is a carotenoid present in many plants and organisms[13] and is of high bioeconomic interest for the pharmaceutical (thanks to its antioxidant properties) and cosmetic and food industries (as a coloring agent).[14,15]

The demand in the industry for carotenoids such as lycopene in medical and pharmaceutical applications keeps increasing,[16−18] but its chemical synthesis is limited by high cost, low yield, and quality.[19] Microbial lycopene production is an alternative promising strategy and has been developed in various hosts such as yeast,[20,21] *Bacillus subtilis*[22] or *Yarrowia li polytica*,[23] and *Escherichia coli*.[16,24−26] The lycopene synthetic pathway (Figure 1A) comprises three successive enzymatic

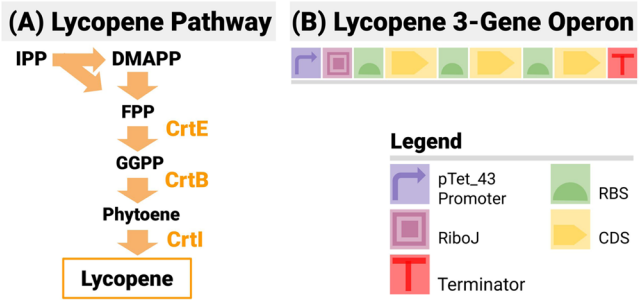


Figure 1. Three-gene lycopene pathway and its implementation with an operon design. (A) The lycopene pathway is made up of three enzymes (crtE, crtB, and crtI). (B) The design includes these three genes arranged in an operon pattern. All gene orders are permitted. A constitutive promoter drives the operon, and an insulating RiboJ is inserted post promoter.

reactions catalyzed by the enzymes crtE, crtB, and crtI.[27,28] Such simplicity, coupled with the fact that all of the associated regulatory mechanisms have been characterized in several strains, makes the lycopene pathway an appealing exemplar case study for optimization problems in metabolic engineering. Lycopene production can also be crippled by metabolic burden (strong production of lycopene and intermediaries take away resources otherwise used for cellular growth)[29] and toxicity—lycopene having been proven to be toxic to the cell[15,30] due to its accumulation in the cellular membrane.

The construct design used in this study will be the design already used in Exley et al.[31] in the context of the use of well-characterized parts for DoE workflows. The construct design (see Figure 1B) is based on the following principles:

- An operon design is used as it is a common design feature in nature and in synthetic biology. It also reduces the design space compared to a design where each gene has its own transcription unit and reduces the chances of homologous recombination if promoters are reused.
- A constitutive promoter drives the operon. Five constitutive promoters are chosen from the SynBIS library[32] to span a range from weak (K137085) to strong (Kelly's reference promoter J23101)[33] to very strong (BioFab promoter apFab32)[34]—see Table S1 for their relative strengths.
- An insulating element RiboJ[35] is inserted between the promoter and first UTR-RBS as a direct consequence of using the BASIC assembly method and its library parts.
- All three RBS in the operon can be varied, as per Blazeck et al.[36] and Salis et al.[37] For each position, three RBS parts from a subset of the Biolegio library (referred to as RBS1, 2, and 3 in the rest of this paper) are used.
- Finally, the design does not fix the gene order and allows for permutation instead—thus permitting a reinvestigation of previous studies such as Nishizaki et al.[38] with the carotenoid pathway in *E. coli* (and yeast), which demonstrated the influence of the said gene order. In the rest of this article, constructs are labeled according to their gene order (e.g., BEI is labeled 1)—see Table S2.

These original choices by Exley et al.[31] lead to a design space of 810 promoter/RBS/gene order combinations (6 permutations of the gene order, 5 promoters, and 3 RBS in 3 positions). The size of the design space (810 constructs) makes this combinatorial optimization problem an ideal test bed for their integration of automation and software, as it is small enough to build all combinations using standard modular plasmid construction methods such as BASIC assembly,[39] vehicle being large enough to employ DoE and for any developed to be capable of scaling to larger, more complex problems.

Failures and Failure Modes. The design space in our work follows that of Exley et al.[31] As a preliminary study for this work, the same initial DoE workflow as Exley et al.[31] was implemented.

An initial set of 8 constructs out of the 88 random constructs selected by DoE software in that study were assembled and transformed into *E. coli*. The transformants were plated and left for growth in an attempt to isolate some colonies for a further culture to extract and measure the produced lycopene. It was then observed that the growth of the transformed colonies was either nonexistent or very slow and the first colonies observed took as long as 72 h to grow. To further investigate the reason for the observed growth discrepancy, the colonies that grew were picked and sent for Sanger sequencing. Two sets of issues were identified. Sequencing results showed that many operon genes had either been deleted or mutated and construct parts could not be identified—showing a propensity of the cells to mutate or get rid of the operon. A significant number of nonexpressing lycopene colonies were observed growing on top of lycopene-expressing ones and outcompeting them on the Petri dish.

These failures can be attributed to two biological factors: metabolic burden (strong production of lycopene and intermediaries take away resources otherwise used for cellular growth)[29] and toxicity—lycopene having been proven to be toxic to the cell[15,30] due to its accumulation in the cellular membrane. Practically, these failures are compounded by two sources of technical failures—experimental failures and assembly failures.

In the context of a DoE cycle, burden and toxicity are major obstacles to an efficient, reliable, and reproducible investigation

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

of the design space; in the worst cases, they are stumbling blocks. In general, burden leads to discrepancy in cellular growth, which is itself a major hindrance to any effort to operate colony picking since colonies cannot be picked within the same time frame—thus reducing the throughput of the entire operation, whether conducted manually or with automation. In titration assays, growth discrepancy is also a significant source of variation in the data. Toxicity (and mutations) will restrict the viable portion of the design space—more or less severe depending on the capacity of the chassis to tolerate the said toxicity.[40] Practically, cells containing constructs lying outside the said viable region will fail to grow.

Debugging individual failures—i.e., identifying the reason for a failure to transform and grow a colony is, in general, cumbersome—for instance, transformation failures cannot be observed immediately after the act itself but only later in the process. Separating failure modes at population levels is much easier, however, since they are distributed very differently over the design space (as illustrated in Figure 2):

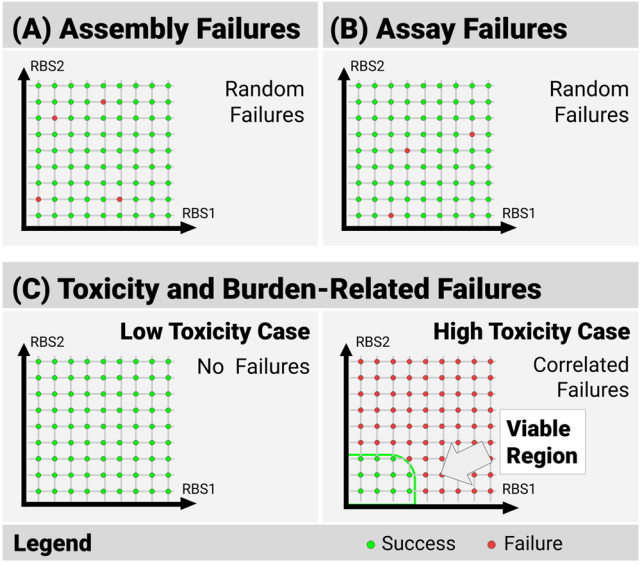


Figure 2. Failure modes affecting the scoping round. Three types of failures are considered in this work. (A) Randomly distributed assembly failures. (B) Assay failures (also randomly distributed). (C) Toxicity-induced failures (highly correlated).

- Assembly failures are sparsely occurring events (Figure 2A)—common modular assembly methods are very effective when short pathways are assembled as is the case for the lycopene exemplar and can be assumed independent from the construct (see the discussion on our solution PASIV).
- Assay failures are uniformly distributed: assay/experimental failures are independent of the construct; their frequency depends on the skills of the experimentalist (or the level of automation) and the protocol adopted (Figure 2B).
- Toxicity and burden issues are NOT independent of the construct. The constructs encode a set of enzymes, and the recorded output (viability, OD, concentration of a metabolite) depends on the concentration of enzymes that are produced. When toxicity and burden are too low to affect the constructs, the entire design space is, a priori, viable (Figure 2C, left). Conversely, when toxicity and burden are issues—only contiguous portions—the design space is viable (Figure 2C, right).

Bootstrapping the DoE Cycle. Typically, the scoping phase of a DoE cycle consists of the random draw of a comparatively small subset of the design space, to be followed by successive, targeted, draws based on data analysis of the collected data, and finally an optimization phase once a narrow-enough region of interest has been identified. Figure 3A shows what can be expected at the end of the scoping study

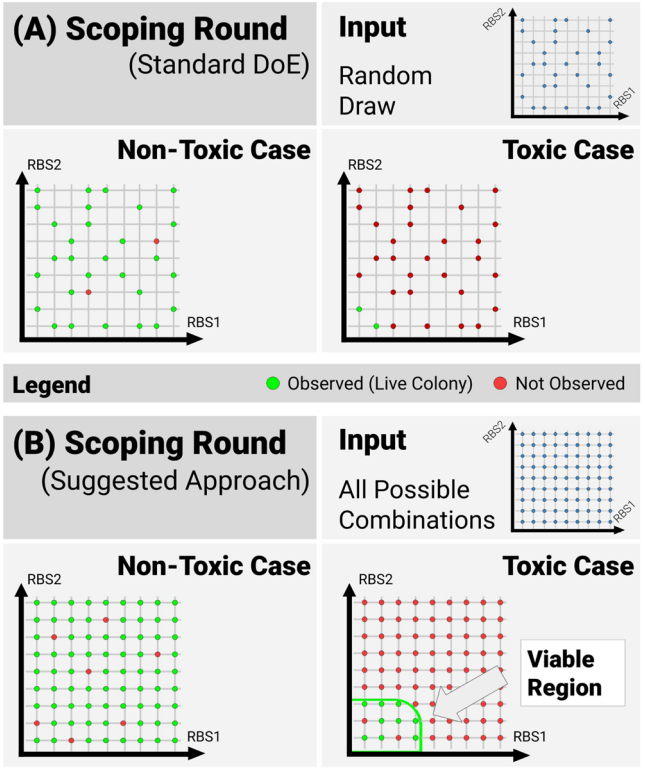


Figure 3. Expected outcome to a scoping round in the nontoxic and toxic cases and for two different sets of input constructs. (A) A subset generated by a random draw from the design space is used as input. (B) The whole space is used. In the toxic case, colonies will only grow for constructs located in the viable region. Only by using a dense input data set will the viable region reveal itself by continuity—this simple remark is the foundation of the PASIV method.

for two extreme cases—when the whole design space is unaffected by toxicity and burden issues (nontoxic case—Figure 3A, left) and when toxicity and burden issues are severe but in a small region of the design space (toxic case—Figure 3A, right). In the nontoxic case, it is possible to transform, grow, and assay for the vast majority of the initial constructs—failures can be solely attributed to technical issues—and enough data can be collected to proceed to the targeted rounds. The scoping round works as intended. Conversely, in the toxic case (for instance, the exemplar lycopene study), a few of the constructs that are initially drawn lie in this viable region. Failures may also be due to assembly or experimental issues rather than burden or toxicity. Overall, only a very few data can be collected at the end of this scoping phase—too few to analyze and proceed to the later phases of the DoE cycle. The scoping round fails.

Figure 3B shows the expected outcome of a scoping round if the whole design space was used as input. In the nontoxic case, most colonies corresponding to the input constructs will grow

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

and yield data—as before (the few) failures can be assigned to assay or assembly issues. In the toxic case, colonies will only grow for constructs located in the viable region. The viable region reveals itself by continuity when successful assays are mapped in the construct space. Using the whole space as input is of course totally impractical—and implementing a brute force approach is antithetic to the idea of DoE. It is possible and tractable, however, to randomly sample large sets using multiplex construction in a pooled approach (PA). The high density of the input data set will then still allow us to exploit the different distributions of the failure modes and identify the continuous regions.

We present in the next section a novel workflow—coined PASIV—that implements this idea and, through an interplay between biology (multiplex construction in a pooled approach) and software (a range of bioinformatics tools including a novel construct matching tool to overcome issues inherent to pooled approaches), offer a novel way to perform the scoping phase of the DoE cycle and identify the (viable) region of interest.

RESULTS AND DISCUSSION

PASIV Approach. The acronym PASIV stands for pooled approach, screening, identification, and visualization—the four distinct phases of the workflow. PASIV is also, as its name implies, a passive method, as the viable region of the design space is allowed to appear through the viability assay (unlike the more active standard approach of drawing random samples and inferring the regions of interest from them). The PASIV approach comprises four sequential phases (illustrated in Figure 4).

Phase 1: Construction with a Pooled Approach (PA). Instead of the targeted construction of a set of constructs (drawn at random), multiplexed construction via a pooled approach is implemented. The purpose of this stage is to build all possible constructs simultaneously in a fast and effective manner. Practically, it is expected to generate a large number of constructs from across the entire design space—with no obvious bias (large gaps, different relative frequencies). A pooled approach helps ensure that the same experimental conditions apply to all of the constructs that are built, thus minimizing one of the potential sources of errors previously listed.

The BASIC assembly method[39] is used for the modular assembly of the constructs in a one-pot reaction, where all varying parts (promoters, RBS, etc.) are mixed in equal quantities to achieve a purely combinatorial assembly. When several gene orders are investigated, it is necessary to operate several pools—each corresponding to a given gene order. In the lycopene exemplar, six pools that correspond to the six possible orders of gene combinations (crtE, crtB, and crtI) were created. BASIC assembly relies on standard 21 base-pair overhangs and 12 base-pair adapters for ligation and to drive assembly toward the specified constructs. The overhangs are highly specific, and their efficiency is assumed to be unaffected by the upstream and downstream sequences (to the best of the authors' knowledge, no results have been published to indicate such an effect at the time of publication of this paper).

Phase 2: A Viability Screening (S) Assay. The cells are transformed with the genetic pool in another one-pot reaction. The transformed cells are then plated and grown for the selection of viable candidates. Viable candidates are then picked up from the plate(s) and sent for sequencing. Selection should be at random and, in particular, should not be affected by the apparent production (as assessed by the color of the colonies)—

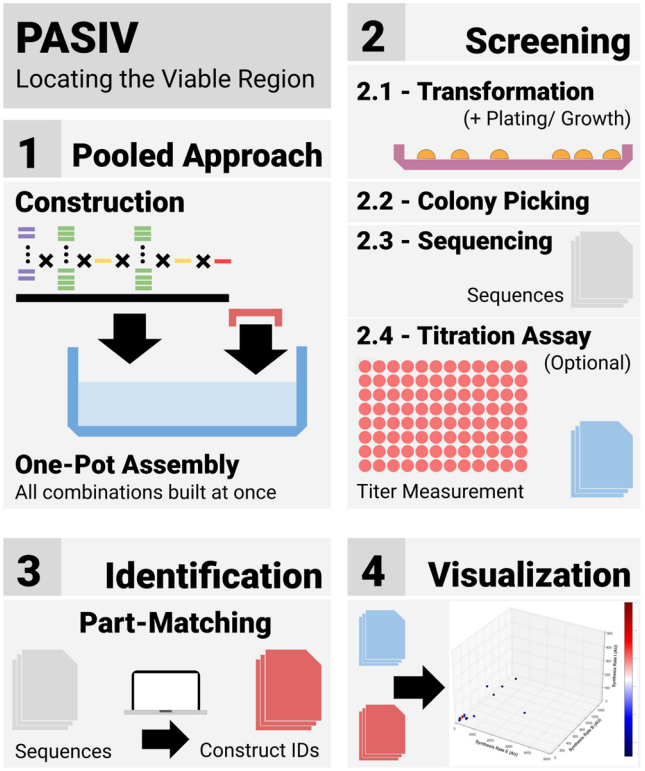


Figure 4. PASIV workflow—summary. PASIV is made of four successive phases. Phase 1, a construction step using modular assembly in a pooled approach and aiming at building all constructs at once. Phase 2, a viability assay consisting of a transformation and culture step followed by a colony picking step (the assay may be followed by a titration assay where metabolite production by the yet unidentified colonies is measured). Phase 3, an identification step where the genetic content of the picked colonies is identified. Phase 4, a visualization/analysis step that combines all collected information.

ideally, all colonies should be picked to reduce possible biases. A titration assay can then be run with the viable candidates to collect data on metabolite production. This second assay collects standard titration features such as the optical density (either as a time series or as an end point of assay) and the concentration of the metabolite of interest and intermediaries if available (also either as a time series or as an end point of assay).

Phase 3: An Identification (I) Phase. Since the cells have been transformed with a mix containing all possible plasmids, there is little prior knowledge of the genetic content of the picked colonies—except that it matches a given template, was assembled from a known set of components, and therefore lies in a, possibly large, space of possibilities. There is also no certainty that all constructs will feature among the picked colonies, while some may be represented several times. This identification phase is crucial to the overall success of the method—without it, no consequent data visualization and analysis are possible. Identification is achieved using a construct matching software called cMatch (see below) that was developed as part of PASIV—and that is now routinely used in-house for other applications such as quality control of the modular assembly. cMatch analyzes sequences and identifies the closest member of a design space corresponding to a given template and a set of component libraries—or returns an error if the sequences are of poor quality or too distant from the constructs in the design space.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

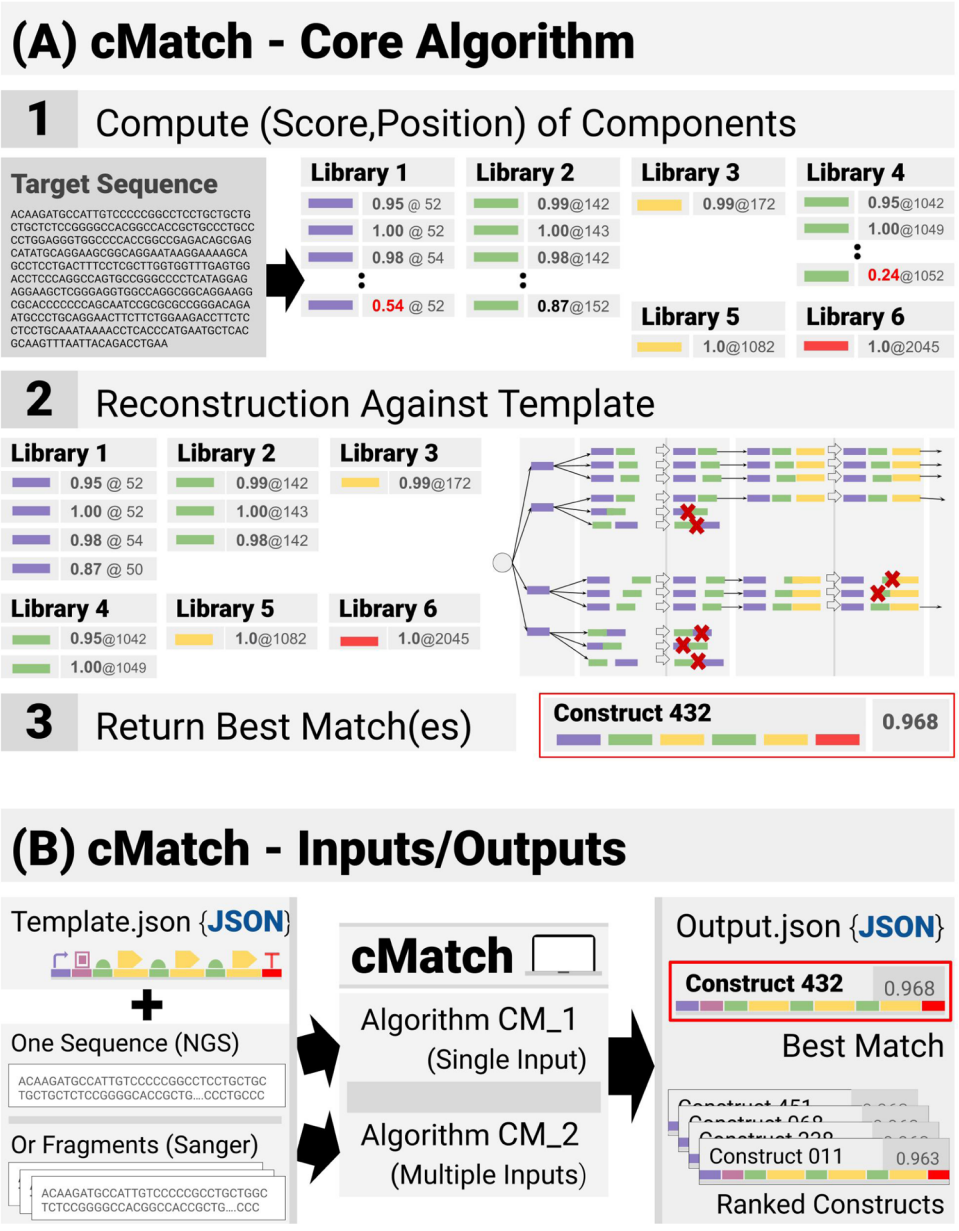


Figure 5. Main features of the construct matching software cMatch. (A) The core algorithm cMatch used for construct matching is made of three steps. Step 1: cMatch looks for all possible individual components in the input sequence. Poor matches (low scores, in red) are pruned out. Step 2: the list of all admissible constructs is generated. Dynamic reconstruction was implemented. A construct matching score is generated from the individual component matching scores for all possible constructs. Step 3: constructs with the highest scores are returned. (B) cMatch uses two types of inputs: a JSON template encoding the positional and combinatorics constraints for a given search space and the sequence data to analyze (single or multiple input). Single inputs are processed with the core algorithm CM_1 detailed above; multiple inputs by an extension CM_2. cMatch returns a JSON file listing the best match, as well as a ranking of all admissible constructs.

Phase 4: A Visualization (V) Phase. Finally, the collected data (colony content and assay measurements) are combined, analyzed, and visualized. The goal of this stage is the localization of the viable region—that is, the region least affected by toxicity and therefore eligible for titer optimization.

Within PASIV, metabolite production is treated with a black-box approach. The constructs in the design space are seen as an enzyme producer. Enzyme production is itself viewed as a multidimensional stimulus to the chassis, which is controlled by tuning construct components (promoter, RBS, etc.). The bacterial chassis has its own limited resources, feedback systems, including resistance to toxicity, and capacity for producing a metabolite of interest. No attempt is made to model these processes; instead, we rely on the data to estimate the dose–response relationship. After data combination, all constructs in the design space are assigned:

- A multiplicity order: the number of times the construct has been identified among the picked colonies—this is assumed strong evidence of whether the colony is viable and/or subject to burden issues.
- The data collected during the titration assay (if any). To distinguish between the different sources of biological noise, they are aggregated at two levels—by colony and for each colony by assay sample (see the Results and Discussion section). Titration data can also be used to further restrict the viable region.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

PASIV uses a minimal set of coordinates corresponding to estimates of the synthesis rates of all products encoded by the construct (the enzymes crtE, crtB, and crtI in the case of the lycopene example). These coordinates—called "modeling coordinates" in the rest of this work as they are based on the modeling of protein production—correspond to a representation of the stimulus amplitude. Modeling coordinates can be used as universal coordinates for constructs encoding the same functions (e.g., enzyme production)—and thus grant comparisons between constructs with different gene orders or based on different patterns (operon vs transcription units). They also allow for the deployment of the most common machine learning preprocessing techniques such as dimension reduction, distribution normalization/standardization, and outlier correction if needed (for instance, in high-dimensional problems). Crucially, metabolic burden and toxicity effects are correlated to the amplitude of the stimulus——the higher the stimulus, the higher the burden and in the case of lycopene, the more lycopene is produced, the higher the toxicity for the cells. This makes such a set of coordinates ideal to differentiate toxicity effects from construction and experimental issues.

Construct Identification with cMatch. PASIV uses software called cMatch that has been developed for the purpose of automating quality control in (high-throughput) workflows encountered in combinatorial pathway approaches to metabolic engineering. Thanks to the way it performs quality control (detailed below), cMatch can be applied to the more general problem of the identification of synthetic constructs lying in a specified design space without having to perform a brute force search for the best match in that design space.

cMatch deals with identification problems of plasmid-level complexity: constructs are expected to include a few genes (10 at most), while sequences to analyze will only be a few kbp-long (15 kbp at most). Furthermore, information will not be evenly distributed in the sequences. The biological functions encoded in the constructs will have sequences spanning several orders of magnitude (CDS are 1 kb or more, promoters around a 100 bp, UTRs around 50 bp or less, while degradation tags are often less than 10 bp long). At the same time, very short sequences (regulatory components, as well as short adjacent sequences to functional components) may massively affect the output to the construct[30,41,42] and will need to be identified with the utmost precision—in contrast to applications such as barcoding,[43] where short barcode sequences (with a space of possibilities in the billions) are used to identify constructs in a design space of several thousands/millions.

To deal with such features and constraints, cMatch adopts a very different strategy to recent annotation tools such as pLannotate,[44] which rely on BLAST[45] for matching and do not use any positional and combinatorial information. First, cMatch uses the highly validated Smith−Waterman algorithm,[46,47] instead of BLAST,[48] which is better suited to genome length sequences. This choice was born out of the need for a high level of precision and the relatively short length of the sequences to analyze. Second, cMatch heavily exploits positional and combinatorial information—which also happen to be the only pieces of information available on the genetic material when pooled construction workflows are used, namely:

- The template, describing their modular structure (component types, relative order, interfacing constraints) and providing the structural constraints for the space.
- For each component of the template, a library of admissible elements—providing a description of the combinatorial constraints for the space.

Rather than using a brute force approach to search the design space for the constructs that best match an input sequence, cMatch performs a three-level process. cMatch first searches for the individual components listed in the template and then generates all admissible combinations. Finally, the combinations are ranked. The matching (homology) scores generated at both levels of the process are not only used to rank the admissible combinations but also automate its decision-making and quantify the reliability of the prediction. cMatch's (core) algorithm practically proceeds, as shown in Figure 5A:

- Step 1—component matching: this looks for the individual components in the input sequence—looping over all libraries of components to match each of their elements to a subsequence of the input sequence—assigning a matching score (normalized by the length of the component) and a position (or several) in the process. Pruning follows: only matches with a score above a user-specified threshold are kept.
- Step 2—reconstruction and pruning: the list of all admissible constructs is generated. Rather than using a purely combinatorial approach with a combinatorial product of the results of the first step, followed by pruning against the template, an iterative reconstruction has been implemented. Combinatorial recombination proves extremely costly indeed when the input contains repetitions and elements are detected in several locations—expanding the number of combinations by several orders of magnitude. Instead, positional constraints are applied as early as possible to prune out entire branches of the reconstruction tree.
- Step 3—output: all admissible constructs are assigned a global score (the geometric mean of their component scores). The combinations with the highest scores are returned as best matches.

Focusing on the individual components of the constructs and then on their combination is an efficient way to investigate large combinatorial spaces and control combinatorial explosion, as it exploits the high level of similarity among constructs in the same design space to flatten the design space and reduce the number and complexity of the computations. Even with the simple example of the lycopene operon, the power of the approach is apparent. There are 810 admissible combinations in the design space (101 250 with the entire Biolegio library), but the 9 libraries only contain 18 elements (1 library of 5 promoters, 3 libraries with 3 RBS each, 3 libraries with 1 CDS each, and 1 library with 1 terminator)—54 with the entire Biolegio library. Positional constraints are powerful pieces of information in general. There are 9!/3! = 60 480 as many ways to arrange the components, i.e., almost 50 millions (and more than 5 billions with the entire Biolegio library) when their order is not imposed. In the context of reconstruction, the positional constraints can be applied very efficiently to accept/reject combinations since the relative positions of the components can be simply inferred from the positions of the components as returned during the first phase (component matching).

As it operates at both component and construct levels, cMatch performs a more thorough operation than simple sequence alignment. This operation, called "construct matching", returns

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

(and quantifies) a description of the modular structure of the construct, including:

- Structure matching: Does the construct match the design? Are all components present? Do their order and interfacing match the construct design?
- Quality of the matching: How close is the matching? Are there any close constructs? Are any components affected by mutations? Are there any insertions or deletions? Matching (homology) scores are generated for all components and constructs.

It is worth emphasizing that such a level of analysis is only possible thanks to the use of the template and knowledge of the elements in the component libraries. Likewise, the identification of a construct would also be intractable without these pieces of information.

cMatch uses two types of inputs: a JSON template encoding the positional and combinatorics constraints for a given search space and the sequence data to analyze (Figure 5B). Sequence data can be either single-input, long-read data (as with next-gen sequencing) or multiple-input, short-read data (typical of Sanger sequencing, which was the case in this study). Two algorithms, the core algorithm CM_1 detailed above and an extension CM_2 have been developed—each tailored for a type of input (CM_1 for single input, CM_2 for multiple input). cMatch returns a JSON file listing the best match, as well as a ranking of all of the constructs in the search space (or an error log), and all intermediary results.

cMatch has been implemented in Python 3.9 and is publicly available as an open-source package on the Kitney Lab Github page (https://github.com/kitneylab) under MIT license (https://choosealicense.com/licenses/mit/). The core functionalities are implemented as three different modules: matching.py, reconstruction.py, and extension.py, which, respectively, implement the core sequence, component libraries and component classes and their matching methods (calling biopython pairwise2 local alignment function), and the reconstruction and extension functions. All input and output files are in JSON (for simplicity) except the sequence files.

Application of PASIV to the Lycopene Exemplar. We now illustrate how the general PASIV workflow can be used in practice with the lycopene exemplar. Three of the four phases of the workflow were customized as follows (the third phase, identification, is completely independent of the application and need not be customized).

Phase 1: Pooled Approach. The BASIC assembly method[39] is used for the modular assembly of the constructs in a one-pot reaction. As previously mentioned, when several gene orders are investigated, it is necessary to operate several pools—one for each gene order. This is of course, primarily, to ensure no combination lying outside the design space (constructs with multiple repeats of some CDS and missing CDS) is constructed. In the context of lycopene, where toxicity and burden are the most limiting factors, it is crucial to the success of the methods as constructs missing one of the pathway enzymes will exhibit a significant growth advantage (no lycopene made) and will therefore be overwhelmingly represented at the screening stage.

Six different pools—each corresponding to one of the six possible gene combinations for crtE, crtB, and crtI—were created. For each of these pools:

- The other varying parts (promoters, RBS) are mixed in equal quantities.
- The cells are transformed with these genetic pools in a one-pot reaction.
- The transformed cells are plated and grown for the selection of viable candidates.

Phase 2: Screening of the Viable Candidates. Viable and lycopene-expressing colonies are selected and picked. Practically, the transformed cells that have grown on the Petri dish and that express the characteristic orange color hue phenotype of lycopene expression[49] are selected and colonies are picked and grown in liquid culture for 24 h for further analysis: lycopene extraction and measurement, and sequencing. Lycopene is extracted with dimethyl sulfoxide (DMSO) (see the Methods section) from the liquid culture of each of the viable candidates. OD_{600} and absorption at 471 nm, the characteristic wavelength of lycopene,[50] are measured, and the lycopene yield is estimated.

Phase 4: Derivation of the Modeling Coordinates. In the lycopene study, the production of each enzyme can be modeled with the standard constitutive expression model

$$dm/\mathrm{d}t = K_1 - d_\mathrm{m} \times m$$

$$\mathrm{d}E/\mathrm{d}t = K_2 \times m - d_\mathrm{E} \times E$$

where K_1 is the effective transcription rate, K_2 is the effective translation rate, d_m is the effective mRNA degradation rate, and d_E is the effective enzyme degradation rate. The synthesis rate is K_1K_2/d_m, and the corresponding steady-state concentration is $K_1K_2/d_\mathrm{m}d_\mathrm{E}$.

Since visualization, clustering and other operations conducted in phase 4, can be conducted up to a multiplicative constant, not all quantities need to be estimated and those who do can themselves be estimated up to a multiplicative constant.

The degradation rates of the three enzymes crtE, crtI, and crtB are constants of the problem. Thanks to RiboJ's protection of mRNA at the 5′ end, the mRNA degradation rate can also be assumed a constant of the problem—independent of the gene order, albeit hard to estimate. It was therefore decided to use K_1K_2, which is the product of the effective transcription rate and the effective translation rate for the effective synthesis rate, instead of the exact K_1K_2/d_m to represent the synthesis rate. All synthesis rates used in the rest of this paper are henceforth expressed in arbitrary units due to this simplification.

With the lycopene operon used in this study, the effective transcription rate K_1 for each gene is estimated from the characterization data of the leading promoter (Table S1). All values are expressed in relative promoter unit (RPU, a relative unit of strength common in promoter characterization,[33] which has been estimated at approximately 0.02 RNAP/s/promoter).[51] The use of the insulating element RiboJ[35] ensures that transcription and translation are decoupled from each other and these estimated promoter strengths could be used for the first gene of the operon. Effective transcription rates for the other positions were derived by multiplying its strength by a factor depending on its position down the operon, as suggested by Nishizaki[38] whose experimental results proved that "mRNA abundance decreased by roughly half from one gene to the next".

The decoupling of transcription and translation is an example of favorable compositional context—unfortunately, no such simplification is possible for the estimation of the effective translation rate. Context is the set of interrelated factors that modulate the operation of biological processes. These factors are traditionally grouped as composition-specific, environment-specific, and host-specific.[52] In general, the expression of functional components is affected by short adjacent upstream/

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

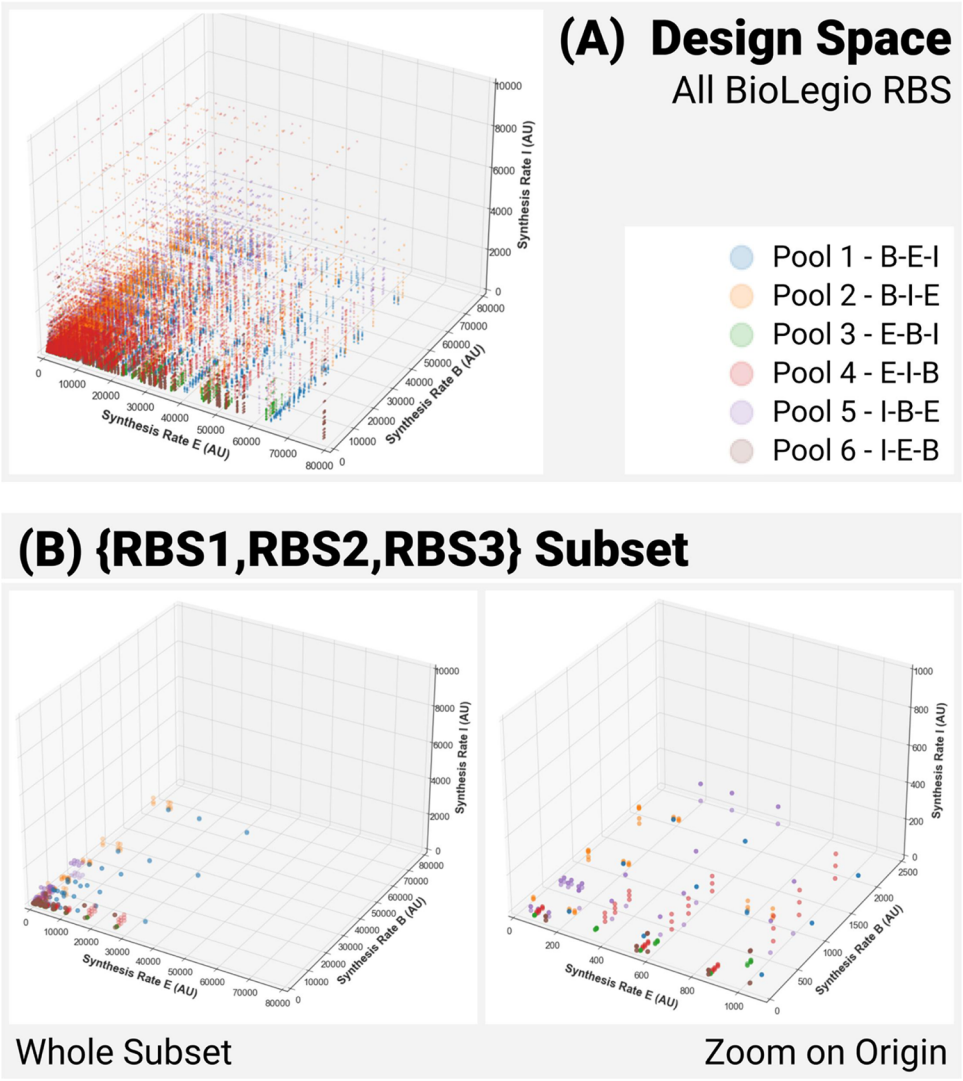


Figure 6. Design space generated from the entire RBS Biolegio library and the subset {RBS1, RBS2, RBS3}. Constructs are color-coded according to their gene order/pool. (A) The design space generated from the entire Biolegio library has 101 250 constructs and is heavily concentrated by the $E-B$ plane. (B) The design space generated from the subset {RBS1, RBS, RBS3} only includes 810 constructs ((B), left) and covers a region closer to the origin. A zoom onto the origin ((B), right) shows a dense sampling of this very weak region.

downstream sequences.[41,42,53] Calculator tools such as the RBS calculator (https://www.denovodna.com/)[37,54] now require 35 bp upstream and 60 bp downstream to estimate the translation rate of an RBS.

No attempt was made at experimentally characterizing the RBS in context. The effective translation rate K_2 was estimated with the RBS calculator instead for all of the RBS in the Biolegio library and for all possible contexts with the operon design (upstream and downstream sequences from the flanking genes or RiboJ—listed in the Supporting files Pool 1 - BEI; Pool 2 - BIE; Pool 3 - EBI; Pool 4 - EIB; Pool 5 - IBE; and Pool 6 - IEB). Corresponding estimations can be found in the Supplementary Information and provide a vivid illustration of the influence of context on RBS expression—not only do translation rates can vary by as much as 3 orders of magnitude indeed but the ranking orders between RBS in the library are also not conserved when context changes. Such unpredictability contrasts sharply with transcription, which varies with gene order in a predictable manner by a factor of up to 4.

Preliminary Investigation of the Design Space. The first immediate application of the set of coordinates was to conduct a preliminary investigation (Figure 6) of the design space generated by the operon design and for the entire BASIC Biolegio library and the chosen subset of {RBS1, RBS2, RBS3}. For all visualizations in this paper, the synthesis rates (synthesis(crtE), synthesis(crtB), synthesis(crtI)) are used as coordinates. We will refer henceforth to this set of coordinates as "synthesis coordinates" and use the shorthands E, B, and I for the enzyme names crtE, crtB, and crtI, respectively.

Using the entire 15-strong Biolegio RBS library yields a design space made of 101 250 constructs (5 promoters, 15 RBS in position 1, 15 RBS in position 2, 15 RBS in position 3, and 6 permutations for the gene order). The corresponding design space (Figure 6A) is highly anisotropic. Constructs exhibit a larger density close to the origin, along the E and B axes, and are also overwhelmingly located close to the $E-B$ plane—this is expected as it corresponds to low crtI production and is a direct consequence of the blocking effect the crtI coding region has on RBS upstream of it. This was verified by conducting a codon optimization of the crtI sequence (https://eu.idtdna.com/pages/tools/codon-optimization-tool) and comparing the estimated translation rates for both crtI and its codon-optimized

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

Table 1. Results of the Sequence Identification Step[a]

Name	Promoter	RBS E	RBS B	RBS I	Gene Order	Multiplicity
C_01	J23101	RBS1	RBS1	RBS2	I-E-B	1
C_02	J23101	RBS1	RBS1	RBS3	I-E-B	3
C_03	J23101	RBS1	RBS2	RBS2	E-I-B	1
C_04	J23106	RBS1	RBS1	RBS2	E-B-I	3
C_05	J23106	RBS1	RBS2	RBS1	E-B-I	1
C_06	J23106	RBS1	RBS2	RBS2	E-B-I	2
C_07	J23106	RBS1	RBS2	RBS3	E-B-I	1
C_08	J23106	RBS2	RBS3	RBS1	E-I-B	2
C_09	J23108	RBS1	RBS1	RBS2	E-B-I	1
C_10	J23108	RBS1	RBS2	RBS2	E-I-B	1
C_11	J23108	RBS1	RBS3	RBS2	E-B-I	2
C_12	K137085	RBS1	RBS1	RBS2	E-I-B	9
C_13	K137085	RBS1	RBS1	RBS2	I-E-B	1
C_14	K137085	RBS1	RBS1	RBS3	E-I-B	6
C_15	K137085	RBS1	RBS2	RBS2	B-E-I	1
C_16	K137085	RBS1	RBS2	RBS2	E-B-I	4
C_17	K137085	RBS1	RBS2	RBS2	E-I-B	9
C_18	K137085	RBS1	RBS2	RBS3	E-B-I	2
C_19	K137085	RBS1	RBS2	RBS3	E-I-B	12
C_20	K137085	RBS1	RBS3	RBS1	E-I-B	1
C_21	K137085	RBS1	RBS3	RBS2	E-I-B	5
C_22	K137085	RBS1	RBS3	RBS3	E-B-I	2
C_23	K137085	RBS1	RBS3	RBS3	E-I-B	3

[a]Constructs of higher multiplicity (4 and above) are colored red. The predominant pools, 3 and 4, are colored blue and purple, respectively. Finally, the RBS placed before crtE are also color-coded (green for RBS1, yellow for RBS2) to illustrate the predominance of RBS1 in that position.

sequence. Results (see the Supporting Information) show an increase by an order of magnitude with the codon-optimized sequence. A wide region by the origin ([0,25 000] × [0,25 000] × [0,1000] in $E-B-I$ coordinates) was densely sampled by the design space, while regions further and further from the origin were less and less so.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

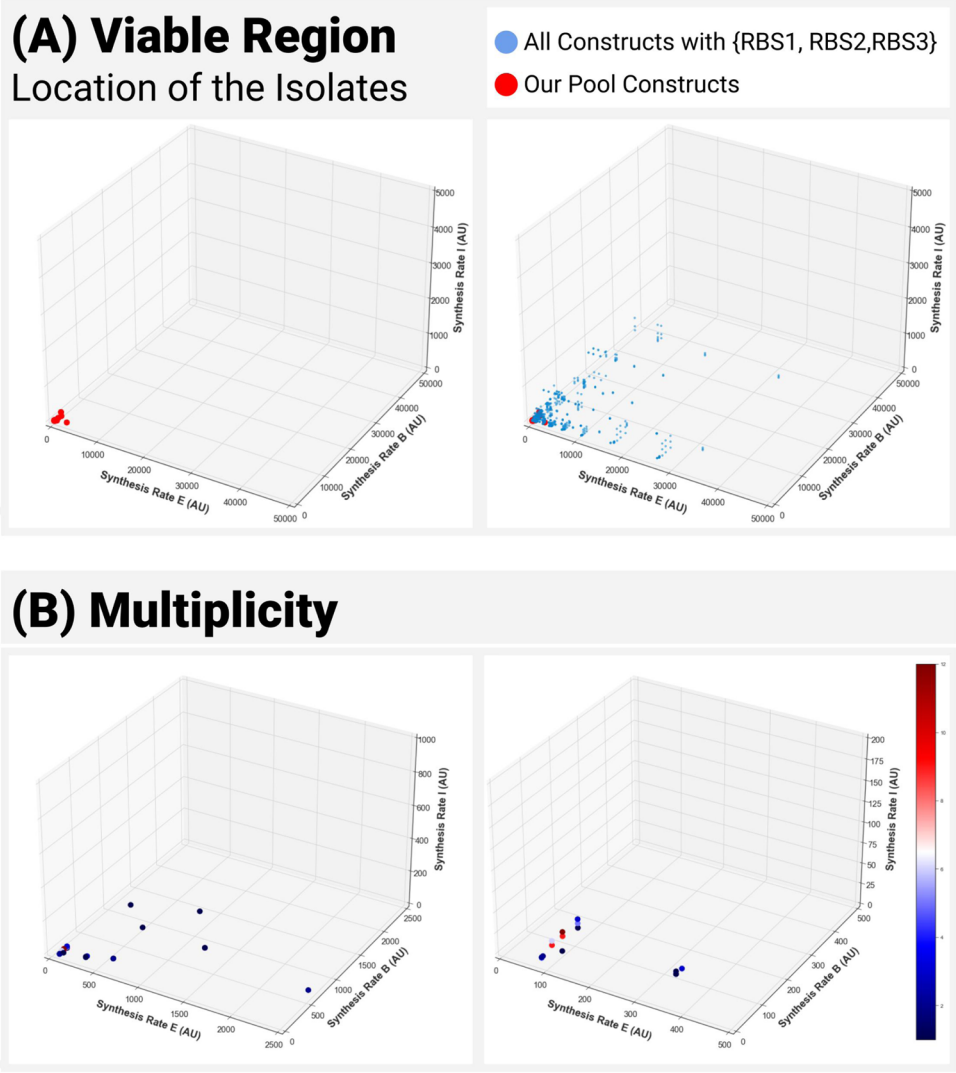


Figure 7. Location of the identified isolates in the *E*−*B*−*I* coordinate system. (A) Identified isolates (red dots) cluster very close to the origin—and only lie in a small portion of the available design space (blue dots). (B) Constructs are colored according to their multiplicity in the sequence data set—from blue (low multiplicity) to red (high multiplicity). A zoom on the region closest to the origin ((B), right) shows that constructs with the highest multiplicity coincide with the lowest available value of *E*.

Influence of the gene order can be glimpsed thanks to the color scheme used in the figure. Separate visualization of the constructs by gene order (see the Supporting Information) shows that constructs tend to be located in different sections of space depending on their gene order—suggesting a possible reason why some gene orders are associated with better production and yield: over-representation in production sweet spots. Separate visualization of the constructs by promoters (Supporting Information) shows that the dense regions are the results of not only the expression of the RBS in different contexts and modulated transcription but also the weighting by promoters of different strengths.

Figure 6B shows the 810-strong design space used in the constitutive study (5 promoters, 3 RBS in position 1, 3 RBS in position 2, 3 RBS in position 3, and 6 permutations for the gene order). The design space generated with this subset is much sparser as it is reduced by a factor of more than a 100 (810 instead of 101 325). It also exhibits a truncated coverage as coordinates cannot go above 30 000 along the *E*-axis, and coordinates do not exceed 1000 along the *I* axis (Figure 6B, left). A zoom on the region close to the origin (Figure 6B, right), corresponding to the weakest constructs, producing the enzymes in little quantity and thus expected to produce little lycopene, shows that the region remains densely (in the *E*−*B* plane) but only contains very low values along the *I* coordinates due to the blocking effect of crtI. Preliminary results obtained while attempting to reproduce Exley et al.[31] have shown how difficult it is to obtain results even for weak constructs; therefore, it was decided to stick to the original {RBS1, RBS2, RBS3} subset and not try to extend the design space with stronger (in context) RBS. Further post hoc justification for this choice of the reduced RBS library was supplied by the experimental results presented in the next section, showing that all constructs yielded by the pooled approach were located well within this region of weak constructs.

Screening Results. In phase 2 of the PASIV workflow, viable colonies are picked to be sequenced and cultured for a titration assay. Colony picking was conducted irrespective of the color of the colony. Only a few colonies were present—confirming previous attempts at cultivating the *E. coli* for lycopene production and how difficult it is. Overall, 99 colonies (from pools 1 to 6) were isolated (see Table S3). Pools 3 (EBI;

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

Table 2. Yield Results (in mg/g of Dry Cell Weight) for All Unique Constructs Drawn with the Pooled Approach[a]

Id	Mean Yield	Max Yield	New Yield	Multiplicity
C_01	26.10	31.85	26.10	1
C_02	58.03	80.69	68.92	3
C_03	1.29	2.42	1.29	1
C_04	4.83	9.66	9.00	3
C_05	0.94	1.31	0.94	1
C_06	9.86	20.08	17.80	2
C_07	7.63	8.26	7.63	1
C_08	7.19	34.78	11.52	2
C_09	2.84	4.49	2.84	1
C_10	8.25	11.29	8.25	1
C_11	9.17	25.31	17.62	2
C_12	22.73	86.07	57.09	9
C_13	3.44	4.39	3.44	1
C_14	17.54	40.57	30.64	6
C_15	20.85	22.82	20.85	1
C_16	25.05	65.71	38.78	4
C_17	16.54	67.63	58.40	9
C_18	9.49	11.52	10.54	2
C_19	24.48	73.78	65.71	12
C_20	5.80	6.58	5.80	1
C_21	23.23	55.33	44.69	5
C_22	63.18	108.62	80.86	2
C_23	15.38	38.28	30.97	3

[a]Constructs of higher multiplicity (4 and above) are colored red. The best performers in terms of yield are colored purple (max yield) and blue (mean yield of the replicates). A new measure of yield is denoted as "new yield"—high performers according to this metric are colored magenta.

37/99) and 4 (EIB; 54/99) were overwhelmingly represented among these isolates as these pools yielded the most colonies for picking. Pools 1, 2, and 6 yielded very few viable colonies (all were picked), while pool 5 yielded none.

All picked colonies were sent for Sanger sequencing (see the Methods section for details on the primers). Returned sequences were then analyzed with cMatch to identify each of the components of the constructs (promoters, RBS, CDS) and reconstruct them. Eighty-seven of these constructs were

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

successfully sequenced. Seventy-three out of the sequenced 87 isolates could be identified with sufficient reliability—that is, they met the minimum quality requirements for construct matching (each component could be matched with a sufficiently high score). Most of the identified isolates proved to be duplicates, and, in the end, only 23 constructs were unique. Such a low number was anticipated since preliminary experimental results hinted at a very limited viable region for the design space and toxicity and burden effects having very noticeable effects on visible outcomes (cell viability and lycopene production).

Table 1 lists these 23 distinct constructs, their components, and their multiplicity among the 73 reconstructed sequences. Instead of listing construct components in template order (promoter, RBS in position 1, CDS in position 1, etc.), they are ordered in (promoter, RBS_E, RBS_B_RBS_I, gene order) order, where RBS_E (respectively, RBS_B, RBS_I) is the RBS located in front of the CDS for crtE (respectively, crtB, crtI). This was done to make the results easier to compare across gene orders and also to match the idea of modeling coordinates, which is independent of the gene order. Results indicate that:

- K137085 (the weakest promoter) is the most commonly represented promoter (12/23) and is used in all of the constructs with high multiplicity (4 or more; all colored in red). This is consistent with a scenario where toxicity is the limiting factor.
- Conversely, apFab32 (the strongest promoter) is not represented at all—again consistent with toxicity as a limiting factor.
- The RBS placed before crtE is almost always the weakest RBS in context RBS1 (22/23, all colored in green), except in one case (RBS2, the second weakest one, in yellow)—again consistent with toxicity as a limiting factor.
- No construct uses the same RBS in all three positions at the same time. This observation contradicts projections by the EFM calculator (https://barricklab.org/django/efm/)[55,56] that repetitions of the RBS only has a minor effect on the overall RIP score (adding an extra 0.1), whereas the choice of promoter has a significantly larger effect (K137085 and its ATATATATATATATAT sequence being estimated to have add 50 points to the RIP score). A possible explanation is that the weight used in the EFM calculator corresponds to much more benign context than the toxic context our cells are under and therefore underestimates the possibility of recombination events when all RBS are identical.
- Pools 3 (EBI; 9/23; colored in purple) and 4 (EIB; 10/23; in blue) remain the most dominant pools (see Table S4 for the aggregated statistics). They also contain all of the higher-multiplicity constructs (hence the relative drop in prevalence compared to the colony results)—hinting at a growth and/or production advantage for some constructs in these pools.

Location and Multiplicity. When plotted in $E-B-I$ (synthesis) coordinates, the identified isolates were all found to cluster close to the origin (Figure 7A, left). All constructs inherited the low-I distribution from the design space. Comparison with the whole design space (Figure 7A, right) illustrates how concentrated the constructs are close to the origin, compared to the available space sampled by the design space.

When construct multiplicity is added to the plots (Figure 7B, left), it can be seen that all higher-multiplicity constructs are not only located close to the origin but also coincide with the lowest values for E-synthesis (Figure 7B, right)—illustrating results in Table S3, which show that these constructs used the weakest promoter and the weakest RBS in front of E. These findings are consistent with the toxicity explanation, where cells die due to lycopene-induced toxicity (accumulation in the membrane), and therefore, higher synthesis levels of enzymes lead to higher lycopene production and accumulation in the membrane. In the $E-B-I$ synthesis coordinates, all isolates were located in a box $[0,2500] \times [0,1500] \times [0,200]$. Highest-multiplicity isolates (exhibiting a growth advantage) were located in a smaller box $[0,100] \times [0,200] \times [0,25]$.

Analysis of the Titration Assay Data. Lycopene production and growth data were acquired as per the protocols described in the Methods section: optical density at 600 nm (standard optical density) and 471 nm, to be, respectively, converted into dry cell weight (DCW) and lycopene concentration. Corresponding yields (in mg/g of dry cell weight) were finally computed, as it is the most commonly used metric for metabolic performance.

When using PASIV, biological replicates are of two sorts. For each isolate that is picked up, several biological repeats originating from that colony are run as part of the titration assay (four in the exemplar study)—they are referred to as "type 1" in this section. Only after identification of their genetic material can the isolates, and associated repeats, be grouped according to their genetic material (the construct)—type 2 repeats. For type 1 repeats, the biological samples are from the same colony, and it is expected that measurements can be safely aggregated and that the statistics derived from these operations will be indicative of the performance of the colony (and the accuracy of the measuring process). For type 2 repeats, however, variations will also be indicative of the differences in terms of metabolic state for colonies with identical genetic material. This distinction between types 1 and 2 is especially relevant to the case of a constitutive design, placing significant duress on the cells (especially for the higher-producing constructs) and where assays cannot be synchronized as they would be with an inducible design—and thus liable to result in significantly different metabolic states for colonies containing identical genetic materials.

Analysis of the measurements was therefore conducted in two steps to separate sources of variation in the data (interisolate vs interconstruct) and aggregate data safely. Only analysis of type 2 data (aggregated by construct) is presented in this section; analysis of the type 1 data can be found in the Supporting Information.

Table 2 lists the results for the yield for all of the 23 distinct constructs. Three statistics are used to identify the best performers in terms of yield: mean of all of the replicates associated with the construct, maximum over these replicates, and a more robust statistic presented below to compensate for the possible variance in metabolic states. Both mean yield and maximum yield identify a similar list of strong performers—the discrepancies between both lists being caused by the multiplicity of the construct. Average yields for the constructs with the highest multiplicity (hence with the highest number of replicates) were often significantly lower than their max value (by as much as 80%)—indicating that genotypically identical constructs in the pool were liable to feature in disparate metabolic states. At the other end of the spectrum, construct C_01 with a multiplicity of only one returned similar values of the mean and max. The best performers are, in all but two cases,

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

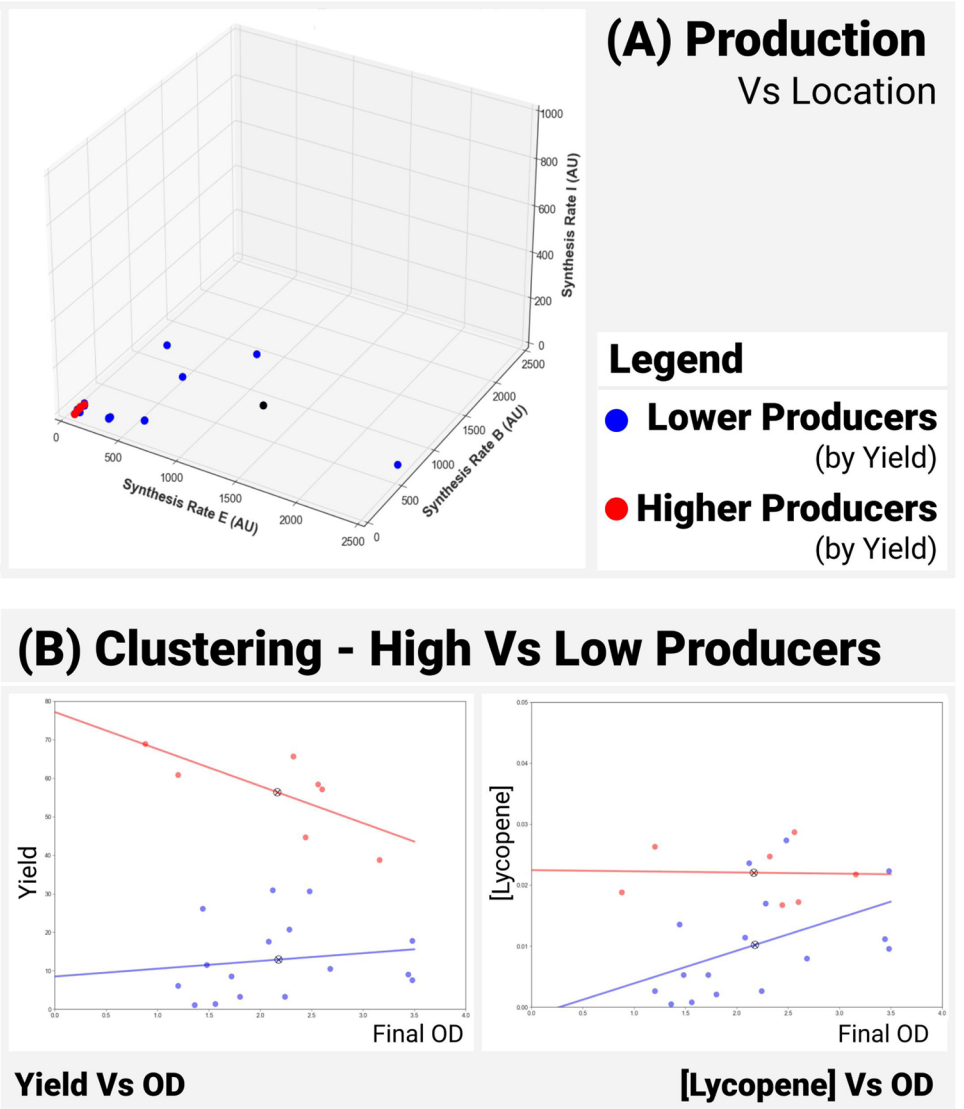


Figure 8. Separating the high and low producers. (A) Location of the high- and low-producer clusters in EBI coordinates. Red dots are used for the strong performers; blue dots are used for the weak ones. Strong performers are all located in the cluster closest to the origin and associated with higher multiplicity. (B) Result of *k*-means clustering ($n = 2$) in the yield vs OD plane (left) and the concentration vs OD plane (right). Red dots are used for the strong performers; blue dots are used for the weak ones. Linear regression was performed on both clusters in both cases.

among the constructs with higher multiplicity (located in the cluster closest to the origin). The best performer, C_22 (K137085, RBS1, RBS3, RBS3), is also located in the cluster. This was surprising, as constructs in the cluster (with some fitness advantage) were not expected to be strong producers (an activity imposing duress). The best yields corresponded to the strongest RBS (RBS2 and RBS3) placed in front of the final enzyme of the pathway (crtI)—hinting that better yields could be found with stronger RBS in front of crtI and that our original choice of RBS was too restrictive (stronger RBS placed in front of crtI should be tolerated and lead to better yields).

To decouple interisolate and interconstruct variation, the following statistic was added. The statistic is computed in two stages. First, for all isolates, the average is computed for all features (DCW, lycopene concentration, and yield). These average statistics are considered reliable indicators of the behavior of the isolate. Then, for constructs of multiplicity larger than one, the isolate, the most representative of the potential of a construct, is identified. The averaged features of that isolate are then assigned to the construct. In the present case, the isolate with the largest mean yield is selected as we are interested in lycopene production (mean lycopene concentration could also be used). Table 2 shows the value of this new yield statistic (alongside the mean and max) for all constructs. The new statistic does not penalize constructs with higher multiplicity as much as the mean of all repeats. Conversely, because it is based on some averaging, it does not reward outliers as much as the maximum.

The new statistic was used to separate the constructs into high and low producers. *k*-means clustering was used on the 23-construct data set for two target clusters and using all three dimensions (yield, concentration, and DCW). Figure 8A shows the location of both clusters in EBI coordinates. Red dots are used for the strong performers; blue dots are used for the weak ones. Red dots are all located in the cluster closest to the origin and associated with higher multiplicity—confirming that the cluster and nearby region should be investigated further if one wishes to find good, reliable performers. A black dot is used to

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

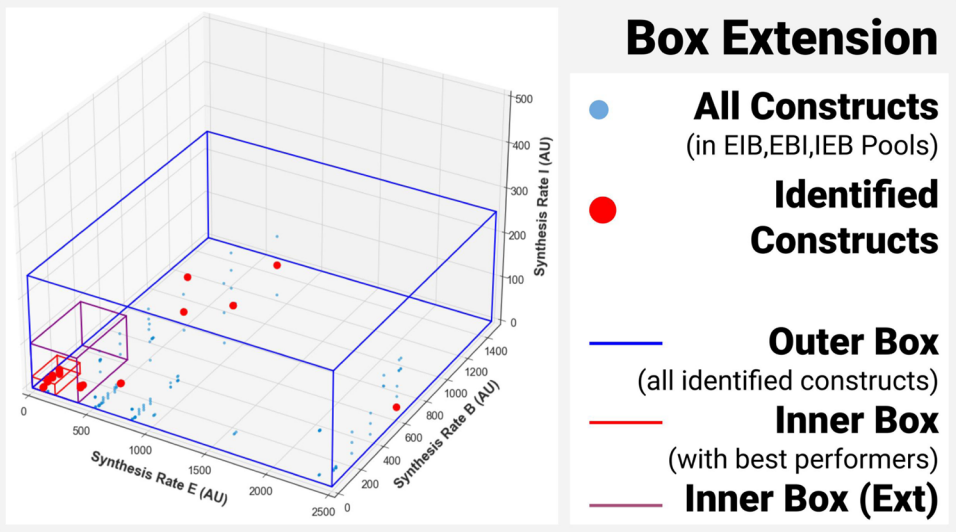


Figure 9. Concluding the scoping study. The higher-producing constructs feature in the red box—close to the origin. To compensate for possible errors in the process, the region is expanded—most noticeably in the *I* direction—yielding the purple box, which is considered the most promising and should be the subject of further investigation.

label construct C_02, which has a high yield but is also associated with low dry cell weight. As the construct is on its own, surrounded by blue dots (low producers), it either corresponds to an outlier (unlikely since it has a multiplicity of 3 in the pool) or is located in a region of the metabolic landscape (high yield, low OD/DCW) that is difficult to access via a pooled approach.

Figure 8B displays the result of the clustering in the yield vs OD plane (left) and concentration vs OD plane (right). Linear regression was performed on both clusters. In the yield vs OD plane, the high performers (red) show a clear downward trend (consistent with a classic growth−production trade-off), while the trend is almost flat for lower producers. In the concentration vs OD plane, no clear trend can be observed for the high performers (red).

Identification of a Region for Further Investigation. All viable constructs were located in the [0,2500] × [0,1500] × [0,200] box (in *E*−*B*−*I* coordinates), while the best performers were in a much smaller box [0,100] × [0,200] × [0,25] (Figure 9). The fact that higher producers were very heavily clustered inside the region, while lower producers were spread further from the origin, coupled with our expectations regarding the effect of toxicity and burden on production, and the results of the viability assay (very few viable colonies) convinced us not to look for further regions of interest in the lycopene case. Instead, it was decided to concentrate on the region containing the most promising isolates, as it clearly was the region to favor for the screening phase. In parallel, the larger region containing all isolates was to be retained for a possible separate screening, as it had been much more sparsely sampled.

Another important feature of the construct design is gene order. Although a safe approach would be to consider all possible orders, using fewer gene orders would simplify construction workflows. Practically, it was decided it would be advantageous to only use pools 3, 4, and 6 (EIB, EBI, and IEB) since

- Pools 3 and 4 yield the most constructs and some of the strongest performers.
- Pool 6 yields fewer constructs but the best performers.
- Other pools yield very few constructs and weaker ones than can be found in other pools.

Finally, as a way to account for possible errors in the process and to make sure that as little as possible of the promising higher-producing region is left out, the region was extended as follows:

- Along the *B* and *E* dimensions, the maximum value was doubled.
- Yield data for the stronger producers hint that it is advantageous to have stronger RBS than RBS3 in front of crtI; the maximum along the *I* dimension was therefore substantially increased from 25 to 100.

Analysis of the data has identified two overlapping regions of the design space: the blue box containing all of the identified constructs and a red box containing the higher-producing region (in red). To compensate for possible errors in the process, it was decided to expand the higher-producing region (in red)—most noticeably in the *I* direction; the resulting region (in purple) is considered the most promising and should be the subject of further investigation.

DISCUSSION

The paper has presented a novel workflow to bootstrap DoE cycles for combinatorial optimization problems affected by severe toxicity issues. More specifically, we have presented a novel workflow—which we have called PASIV—to perform the scoping phase of the DoE cycle and identify a suitable region of the design phase for the screening and optimization phases to be conducted on (or subregions of).

PASIV is based on a multiplex construction phase via a pooled approach in conjunction with a viability assay. This experimental phase is coupled to a software phase, where the isolated colonies are sequenced, their genetic content (construct) is identified, to finally yield an estimated location for one (or several) region(s) of interest. We believe that this form of interplay between biology and software methodology provides a flexible, time-efficient solution. Even at the relatively modest scale of this study, identification of the viable region of the design space would have been a cumbersome, lengthy, and unreliable process without these tools and methods.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

Although PASIV has been illustrated with lycopene (a very bright pigment with a short pathway), it is worth emphasizing that the workflow has not been developed for a broader range of applications. The modular assembly method, BASIC, is also known to perform well for longer pathways—thus extending the range of pathways PASIV can be applied to. The screening phase is also not dependent on the presence of a visual reporter (be it a colorful metabolite or through the addition of a dedicated sensor[57]). PASIV rests on a viability assay indeed—viable colonies are to be picked regardless of the visual stimulus they emit. The viability assay is followed by a titration assay, which is specific to the metabolite of interest. For instance, in the lycopene exemplar (see the Methods section), lycopene was first extracted with[50] DMSO from the liquid culture and its concentration was estimated from absorption reading at 471 nm.[50] Visual reporters could, in theory, be used to steer toward high-producing regions in the picking phase, but no attempts to do so were conducted in our work, as too few viable colonies were present on the plates and the authors were concerned with importing possible biases in the screening phase. As for the software phases of the workflow, they are agnostic to the chosen pathway.

In the lycopene exemplar, a homogeneous, contiguous region of space, containing strong performers could be identified. The region, as could be expected for a problem where toxicity plays a major role, corresponds to the weakest constructs. Equivalently, it was located close to the origin with the minimal set of coordinates (the enzyme synthesis rates) used by PASIV. Although extrapolating from the data yielded with PASIV should be done with suitable care, there are good reasons to consider the identified region for further investigation.

The identified region had an enviable set of properties. Collected data show a continuous, viable region of space that contains healthy and productive colonies: some constructs indeed yield colonies with high multiplicity, with a clear growth advantage and performing well. In the context of metabolic engineering, these constructs would be good choices as they are, most likely, reliable performers. Conversely, constructs that fail the pooled approach, while possibly capable of good production, are more likely to be subject to the toxicity effects and be unreliable performers.

Higher producers were also very heavily clustered inside the region, while lower producers were spread further from the origin. This fact, coupled with our expectations regarding the effect of toxicity and burden on production, and the results of the viability assay (very few viable colonies) convinced us not to look for further regions of interest in the lycopene case. Instead, the region containing the most promising isolates was expanded (as shown in the Results and Discussion section and further discussed below)—as it clearly was the region to favor for the screening phase. In parallel, the larger region containing all isolates was retained for a possible separate screening, as it contained more constructs but had been much more sparsely sampled. Metabolic landscapes are in most cases not as accommodating as the lycopene landscape. And it is, in general, not possible to tell with certainty if the whole of the viable region has been identified without several repeats of the workflow—especially for more complex pathways where several viable regions may exist. Even in such a case, advantages in using PASIV remain. The multiplex construction remains fast and simple as all components are mixed at once in a one-pot reaction—with identical experimental conditions. More importantly, each new round of PASIV will yield constructs that accumulate in the viable regions—which remains a better outcome than the outcome of random sampling rounds (with a lot of failures).

Turning the locations of the identified constructs into a region of interest for further investigation means performing a practical trade-off between efficiency (region should not be too large) and safety (region should be large enough)—and involves design decisions that are typical of a scoping study. On the one hand, preliminary analysis of the data can lead to some forms of restriction on the region of interest. In the lycopene exemplar, a reasonable case could be made for not using all six gene orders but only three instead. On the other hand, it is better to identify a region that is too large than too small, lest the best-producing constructs are missed. PASIV is inherently conservative: its goal is not to identify some optimal region, but a region worthy of further investigation instead. Also, due to the way the viability screening works, gaps between identified constructs are to be expected and should be filled—the viability assay will miss some viable constructs, even if they are represented on the plates, as there are practical limits to the number of colonies that can be picked. In the Results and Discussion section of this work, a very simple strategy was adopted: the $E-B-I$ coordinates were used to group all of the constructs that were identified into a single box. An extension of that box was also performed—practically the boundaries of the previous box were changed to give ourselves a margin of error for the entire PASIV process.

Although the development of a rigorous statistical framework to assess whether the viable region(s) have been identified already, or if more rounds are needed, is beyond the scope of this work, some elements of such a framework are worth discussing. Since the purpose of PASIV is the safe identification of viable regions and is followed by a targeted sampling of these regions, it is enough to ensure its outcome remains stable over a range of iterations. We suggest using an equivalent criterion, but more amenable to quantification, and turn PASIV's outcome into a simple predictive model (inside the box(es): predicted to be viable, outside: predicted not to be viable) and track its performances over a range of iterations of PASIV by comparing its predictions against the data collected during a new iteration. A stopping condition can easily be constructed from the standard classification performance metrics (recall and F1 since the scoping round aims to avoid false negatives). When there are several subregions, clustering metrics (silhouette score, Davies–Bouldin index, Calinski–Harabasz index) can also be used to quantify the evolution of the clustering results. In parallel, as more data are collected, estimation of the production landscape (mean, confidence interval) can be refined, as can the low-producer/high-producer clustering.

The benefits of conducting a proper titration assay in the second phase, and not just a viability assay, extend beyond collecting data that will be of use in the subsequent rounds of the DoE cycle: the collected data can be useful in identifying the location of the region of interest, as shown in the lycopene example. Clustering identified two clusters (one of lower producers, one of higher producers)—the stronger producers being located very close to the origin—and corresponded to the lowest values of E-synthesis. Also, the yield data hint that it is advantageous to have stronger RBS than RBS3 in front of crtI and that regions with strong producers should be expanded along the I dimension.

A benefit of using coordinates to define a region of interest is to enable resampling of the said region to make it less sparse—this would be of clear benefit to generate constructs even closer

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

to the *E*-axis. In practice, it means identifying more constructs using the specified operon design but outside the original design space as they use new components. This can be achieved, for instance, by extending the RBS libraries to include all of the members of the Biolegio collection and drawing from the constructs that fall in the newly identified region of interest. Likewise, more suitable constructs can be generated by adjusting the promoter in the operon and using weak promoters.

Finally, another benefit of operating at the region level (rather than individual construct level) and adding margins of error to the location of the region of interest is to be found in the final visualization step in PASIV. The visualization phase relies on a minimal set of coordinates related to enzyme synthesis and for these coordinates to be estimated for each element in the design space. Transcription values were derived from promoter characterization data in a similar context. For translation, it was decided to rely solely on a bioinformatic tool, the RBS calculator, to estimate the translation rates of all of the RBS involved in the study and for all possible upstream and downstream contexts. The decision was taken for several practical reasons. First, the RBS calculator is a popular tool that is consistently improved. Second, experimental characterization of the RBS in context (to determine translation rates) was cumbersome:

- First, common reporters such as GFP cannot be used outright since their first 60 bp differ from the crtE, crtI, and crtB enzymes. Modifying the reporters at their 5′ ends was rejected as too liable to modify the reporters' properties. Using a fusion protein was also rejected as it was liable to alter too much the metabolic burden (and could lead to folding problems).

- His-tag characterization of protein concentration was also investigated but protein purification considerations, concerns about folding, and concerns about potential overexpression of the recombinant proteins[58] led to its rejection.

Finally, although there are other methods—most notably targeted proteomics[59,60]—that are well suited to directly assess protein levels, they were rejected for the scoping phase, as these modern capacities are not widely accessible yet, and PASIV relies on growth information first (multiplicity and OD) and production information second. These methods should of course be considered for the subsequent targeted rounds of the DoE cycle.

Although it is tempting to use calculators, when possible, due to their simplicity of use, it is important to bear in mind that their use may lead to some features of some of the constructs being estimated with a significant margin of error—making individual construct predictions risky. While this is a potential issue in the screening and optimization phases (protein measurements should then be conducted), we argue that it is less of a problem in the scoping phase as conducted in PASIV. PASIV's purpose is the identification of a region that is worthy of further investigation. PASIV is not so much concerned with minimizing false positives (including constructs that should not be) than maximizing true positives (making sure the region covers the right area of the design space). This is reflected in the simple manner the region of interest is reconstructed from the data (clustering the identified isolates into boxes with added margins for errors).

CONCLUSIONS

The PASIV approach presented in this work offers a simple solution to a particular issue in DoE cycles: what can be done if the scoping phase is severely hampered by burden and toxicity issues. It is also a showcase for a systematic approach to synthetic biology based on a few principles:

- As far as possible, the biology should be allowed to behave naturally—the first two phases of PASIV harness biology, so the viable region reveals itself.
- Reproducibility, reliability, and robustness are key. In PASIV, they featured in the design of the workflow—to account for possible failure modes—as well as the conservative approach adopted to identify the viable region.
- Automation and software are to be deployed as much as possible to operate at scale and make it possible to reliably collect and process large amounts of data. PASIV would not work without its software identification phase (phase 3).
- Synthetic biology should move further into the direction of data science—and the biology be abstracted as much as possible. The visualization phase (phase 4) is but a small example of such an application.

As a consequence of these principles, several computational tools, as well as a set of problem-specific experimental methods, have been developed. Although the most obvious application area for these tools and methods is metabolic engineering and large combinatorial optimization problems, we are confident that they can be applied to a wider range of problems.

METHODS

BASIC DNA Assembly. BASIC linkers and part preparation were done using the standard BASIC part preparation protocol (https://www.basic-assembly.org/protocols). BASIC assemblies were also done following the standard BASIC assembly protocol in a 96-well standard plate either manually or with the Opentrons OT-2.

Preparation of BASIC Linkers and BASIC Bioparts. BASIC linkers were obtained from Biolegio. Synthesized genes were ordered as gBlocks from NEB. New linkers are prepared as follows:

1. Spin down the tubes with lyophilized linkers to ensure oligos are at the bottom of the tube.
2. Set the heating block to 95 °C.
3. Add 200 μL of the linker annealing buffer to each linker tube and leave it on the bench for 1 h.
4. Vortex the tubes and collect the liquid at the bottom of the tube with a quick centrifuge spin.
5. After the heating block reaches 95 °C, place the tubes into the block and slightly loosen the tube caps to allow for heat expansion.
6. After 5 min, switch off the heating block and tighten the tube caps again to avoid evaporation.
7. Allow the tubes to cool down to room temperature over at least 1 h in the heat block.
8. Collect the solution at the bottom of the tube with a quick centrifuge spin.
9. Linkers are stored at −20 °C. Bioparts for the BASIC assembly are provided in storage plasmids (pSEVA18). For each BASIC linker ligation reaction, 50 ng of plasmid per 1 kb of total plasmid size (including BASIC part and

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

storage backbone pSEVA18) is required. Usually, that amount of DNA is provided in 1 μL of a typical miniprep of biopart storage plasmids (200 ng/μL for a 4 kb plasmid). If polymerase chain reaction (PCR) products or gene fragments are used as reaction inputs, 50 ng per 1 kb linear DNA is required.

BASIC Reaction. For each BASIC linker ligation reaction, one PCR tube with 30 μL total volume was set up: dH_2O 17 μL, Promega T4 buffer (10×), 3 μL prefix linker, 1 μL suffix linker, 1 μL BASIC biopart 0.5−6 μL (50 ng per 1 kb total plasmid size), dH_2O was added to reach 28.5 μL volume, NEB *Bsa*I-HF v2 enzyme (R3733) 20 U/μL, 1 μL Promega T4 ligase (M1801), 1−3 U/μL 0.5 μL mix by pipetting up and down. After mixing, the tubes are placed in a PCR machine running the following program: (37 °C, 2 min) × 20 cycles (20 °C, 1 min), (37 °C 5 min), (80 °C, 20 min).

Magbead Purification. This was done using 0.5 mL of 70% EtOH per BASIC reaction and bringing the magnetic beads stored at 4 °C back into the homogeneous mix by shaking thoroughly. We used 96-well Falcon plates (Falcon 351177) in combination with an Ambion magnetic plate (AM10050) for quick magbead immobilization and easy pipetting access. Purification protocol was applied as follows:

1. Add 54 μL of magnetic beads into a 96-well Falcon plate (one well per BASIC reaction) and add 30 μL BASIC linker ligation from the PCR machine step, mix by pipetting 10 times.
2. Wait 5 min to allow DNA binding to magbeads.
3. Place the Falcon plate on the magnetic stand and wait for rings to form and the solution to clear.
4. Remove the solution with a 200 μL pipette tip from the center of each well.
5. Add 190 μL 70% EtOH to each well and wait for 30 s.
6. Remove the solution from each well (pipette set to 200 μL volume).
7. Add 190 μL 70% EtOH to each well and wait for 30 s.
8. Remove the solution from each well (pipette set to 200 μL volume).
9. Leave the plate to dry for 1−2 min.
10. Remove the Falcon plate from the magnet and resuspend magbeads in 32 μL dH_2O.
11. Wait for 1 min for DNA to elute.
12. Place the Falcon plate back on the magnetic stand and allow the ring to form and the solution to clear.
13. Pipette 30 μL of H_2O with eluted DNA into a fresh 1.5 mL Eppendorf tube for direct use in assembly or storage at −20 °C for up to 1 month.

Assembly Reaction. For each BASIC assembly, parts were combined with buffer in a PCR tube: dH_2O 2 μL, NEB CutSmart buffer 10 × 1 μL, linker ligated BASIC part crtE 1 μL, linker ligated BASIC part crtB 1 μL, Linker ligated BASIC part crtI 1 μL, dH_2O top up to 10 μL total volume. Assembly reaction is run in a PCR machine with the following program: 50 °C for 45 min followed by 4 °C on hold.

Transformation. Fifty microliters of chemically competent cells DH5alpha with high transformation efficiency (10^9 CFU/μg pUC19, for instance, NEB C2987I) was used to transform 5 μL of each BASIC assembly:

1. Chemically competent cells are stored at −80 °C.
2. Thaw the competent cells on ice (takes 5−10 min); 50 μL per BASIC assembly to be transformed.
3. Cool 5 μL of the BASIC DNA assembly in a 1.5 mL Eppendorf tube on ice.
4. Add 50 μL of competent cells to each precooled 5 μL BASIC reaction.
5. Incubate on ice for 20 min.
6. Apply heat shock in a 42 °C water bath for 45 s and place back on ice for 2 min.
7. Add 200 μL of the SOC medium to each tube and incubate shaking at 37 °C for 1 h recovery.
8. Spot or plate cells on agar plates with the appropriate antibiotics. Depending on the number of parts assembled and the transformation efficiency, 2−250 μL might be spotted or plated.
9. Incubate agar plates at 37 °C overnight, and the next day, pick the colony for assay or miniprep. We used a PCR machine, heat block (up to 95 °C) fitting 2 mL Eppendorf tubes, water bath (42 °C) for transformation, magnetic plate Ambion AM10050 (Thermo), 96-well U-bottom Falcon plate, Falcon 351177 (Thermo) Eppendorf tubes, magnetic beads Ampliclean (Nimagen), dH_2O, 70% EtOH, Biolegio BASIC linkers BBT-18100 (Biolegio), BASIC parts in storage plasmids (200 ng/μL), NEB *Bsa*I-HF v2 enzyme (R3733), 20 U/μL; includes CutSmart buffer R3733 (NEB), Promega T4 ligase (M1801) 1−3 U/μL; includes Promega T4 buffer M1801 (Promega), chemically competent cells (DH5alpha, 1×10^9 CFU/μg pUC19); includes SOC media C2987I (NEB).

Pooled Transformation. Instead of single RBS at a single location, we use three RBS while maintaining a relative proportion of RBS (relative to volume). The promoters, RBS, and genes were added to the transformation mix. The order of the linkers in the BASIC assembly allows the fixed ordering of the three genes (crtE, crtB, crtI), hence, the six different pools containing six different gene order combinations.

Cloning and Cultivation. DH5α (New England Biolabs) *E. coli* was used for cloning with a standard approach. *E. coli* cells were grown in Luria Broth (LB) medium supplemented with 25 μg/mL kanamycin (Kan). DH5α *E. coli* starter cells were grown at 37 °C. Cell cultures for lycopene production were grown at 28 °C (optimum growth temperature for lycopene production).

Extraction and Measurement of Lycopene. The 24 h growth plate has its OD_{600} measured on a plate reader (Clariostar). Fifty microliters from the 1000 μL total volume is added to 150 μL of LB in a new, flat-bottom plate; it is mixed thoroughly and then tested. The final OD value that the plate reader gives us is multiplied by 4 to compensate for this dilution. The remaining culture is spun at 4000 rpm for 15 min, at room temperature, until pellets appear. The supernatant is removed by turning the plate upside-down onto paper towels. The pellets are resuspended in 500 μL of DMSO, using a robot (OT-2) or a multichannel pipette. The plates are then incubated at 37 °C for 30 min at 900 rpm on a benchtop incubator, to extract the lycopene, while covered in aluminum foil to limit exposure to light. The plates are then spun down to generate pellets (15 min at 4000 rpm). Two hundred microliters of the lycopene-containing DMSO was moved to a flat-bottom plate to measure the OD at 471 nm (Clariostar).

Sequence Analysis and Part-Matching. To analyze the Sanger sequences and identify the best candidate parts of the constructs, we used the cMatch software. This allowed us to automate our workflow and was sorely needed to procedurally match against the 810 possible in silico constructs.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

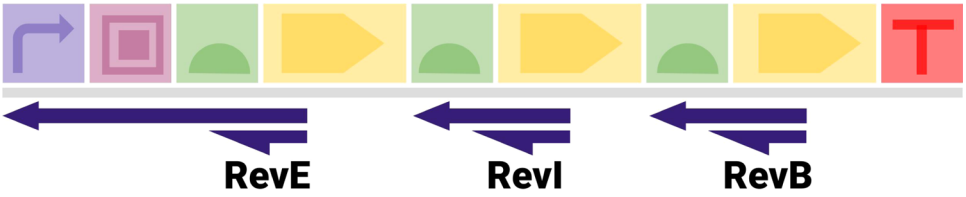


Figure 10. Lycopene operon with the three reverse sequencing primers for the three genes crtE, crtI, and crtB.

We sequenced with three distinct reverse primers starting at 1/3 in each of the three genes (crtB, crtI, and crtE) (Figure 10). The three subsequences obtained are approximately 800 base-pair long, which is long enough to let us determine the upstream parts. The multi-input variant CM_2 of the cMatch algorithm was used.

Yield Calculation. Cell growth (*E. coli* DH5α) was followed by measuring the optical density at 600 nm and correlated to cell dry weight (CDW) with a ratio of CDW/OD = 0.36.

Ratio of cell dry weight (g/L)/OD_{600}.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acssynbio.1c00562.

Design of the lycopene operon (Figure S1 and Tables S1 and S2); variation of RBS expression with its context (Figures S2−S4); design space—visualization—visualization by gene order (Figures S5−S10); visualization of whole design space (Figure S11) and subset (Figure S12); visualization by leading promoter (Figures S13 and S14); viability screening assay—results—multiplicity of the isolates (Table S3); multiplicity of the constructs (Table S4); titration assay analysis—level 1—measurements on the isolates (Figures S15−S19); measurements on the constructs (Figures S20−S25 and Tables S5 and S6) (PDF)

Translation rate for all of the RBS in the Biolegio library and for all possible contexts with the operon design (Pool 1 - BEI; Pool 2 - BIE; Pool 3, EBI; Pool 4 - EIB; Pool 5 - IBE; and Pool 6 - IEB) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Richard Kitney − *Department of Bioengineering, Imperial College London, London SW7 2BX, United Kingdom*; Email: r.kitney@imperial.ac.uk

Authors

Alexis Casas − *Department of Bioengineering, Imperial College London, London SW7 2BX, United Kingdom*; orcid.org/0000-0002-3301-0235

Matthieu Bultelle − *Department of Bioengineering, Imperial College London, London SW7 2BX, United Kingdom*

Charles Motraghi − *Department of Bioengineering, Imperial College London, London SW7 2BX, United Kingdom*

Complete contact information is available at: https://pubs.acs.org/10.1021/acssynbio.1c00562

Author Contributions

C.M. designed the pooled approach and conducted all wet-lab experiments. A.C. contributed to the wet-lab experiments. A.C. and M.B. developed the software listed in the manuscript and conducted all data analysis. R.K. managed the overall project and the research strategy, as well as specific aspects of the software development.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the project described in the paper by grants from the U.K. Government Department of Business, Energy and Industrial Strategy (via the U.K.'s National Physical Laboratory) NPL 2937881 and The Engineering and Physical Sciences Research Council (United Kingdom) EP/L011573/1 and EP/S001859/1.

■ REFERENCES

(1) Jeschek, M.; Gerngross, D.; Panke, S. Combinatorial Pathway Optimization for Streamlined Metabolic Engineering. *Curr. Opin. Biotechnol.* **2017**, *47*, 142−151.

(2) Naseri, G.; Koffas, M. A. G. Application of Combinatorial Optimization Strategies in Synthetic Biology. *Nat. Commun.* **2020**, *11*, No. 2446.

(3) Hoshino, T. Violacein and Related Tryptophan Metabolites Produced by *Chromobacterium violaceum*: Biosynthetic Mechanism and Pathway for Construction of Violacein Core. *Appl. Microbiol. Biotechnol.* **2011**, *91*, 1463−1475.

(4) Yadav, V. G.; De Mey, M.; Giaw Lim, C.; Kumaran Ajikumar, P.; Stephanopoulos, G. The Future of Metabolic Engineering and Synthetic Biology: Towards a Systematic Practice. *Metab. Eng.* **2012**, *14*, 233−241.

(5) Gilman, J.; et al. Statistical Design of Experiments for Synthetic Biology. *ACS Synth. Biol.* **2021**, *10*, 1−18.

(6) Singleton, C.; Gilman, J.; Rollit, J.; Zhang, K.; Parker, D. A.; Love, J. A Design of Experiments Approach for the Rapid Formulation of a Chemically Defined Medium for Metabolic Profiling of Industrially Important Microbes. *PLoS One* **2019**, *14*, No. e0218208.

(7) Azubuike, C. C.; Edwards, M. G.; Gatehouse, A. M. R.; Howard, T. P. Applying Statistical Design of Experiments To Understanding the Effect of Growth Medium Components on *Cupriavidus necator* H16 Growth. *Appl. Environ. Microbiol.* **2020**, *86*, No. e00705-20.

(8) Yildirim, S.; Thompson, M. G.; Jacobs, A. C.; Zurawski, D. V.; Kirkup, B. C. Evaluation of Parameters for High Efficiency Transformation of *Acinetobacter baumannii*. *Sci. Rep.* **2016**, *6*, No. 22110.

(9) Spice, A. J.; Aw, R.; Bracewell, D. G.; Polizzi, K. M. Improving the Reaction Mix of a Pichia Pastoris Cell-Free System Using a Design of Experiments Approach to Minimise Experimental Effort. *Synth. Syst. Biotechnol.* **2020**, *5*, 137−144.

(10) Xu, P.; Rizzoni, E. A.; Sul, S.-Y.; Stephanopoulos, G. Improving Metabolic Pathway Efficiency by Statistical Model-Based Multivariate Regulatory Metabolic Engineering. *ACS Synth. Biol.* **2017**, *6*, 148−158.

(11) 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.; Scrutton, N. S. An Automated

Design-Build-Test-Learn Pipeline for Enhanced Microbial Production of Fine Chemicals. *Commun. Biol.* **2018**, *1*, No. 66.
(12) Rajakumar, P. D.; Gowers, G.-O. F.; Suckling, L.; Foster, A.; Ellis, T.; Kitney, R. I.; McClymont, D. W.; Freemont, P. S. Rapid Prototyping Platform for *Saccharomyces cerevisiae* Using Computer-Aided Genetic Design Enabled by Parallel Software and Workcell Platform Development. *SLAS Technol.* **2019**, *24*, 291−297.
(13) Hernández-Almanza, A.; Montañez, J.; Martínez, G.; Aguilar-Jiménez, A.; Contreras-Esquivel, J. C.; Aguilar, C. N. Lycopene: Progress in Microbial Production. *Trends Food Sci. Technol.* **2016**, *56*, 142−148.
(14) Ciriminna, R.; Fidalgo, A.; Meneguzzo, F.; Ilharco, L. M.; Pagliaro, M. Lycopene: Emerging Production Methods and Applications of a Valued Carotenoid. *ACS Sustainable Chem. Eng.* **2016**, *4*, 643−650.
(15) Wang, C.; Zhao, S.; Shao, X.; Park, J.-B.; Jeong, S.-H.; Park, H.-J.; Kwak, W.-J.; Wei, G.; Kim, S.-W. Challenges and Tackles in Metabolic Engineering for Microbial Production of Carotenoids. *Microb. Cell Fact.* **2019**, *18*, No. 55.
(16) Kim, M. J.; Noh, M. H.; Woo, S.; Lim, H. G.; Jung, G. Y. Enhanced Lycopene Production in *Escherichia coli* by Expression of Two MEP Pathway Enzymes from Vibrio Sp. Dhg. *Catalysts* **2019**, *9*, No. 1003.
(17) Jung, J.; Lim, J. H.; Kim, S. Y.; Im, D.-K.; Seok, J. Y.; Lee, S.-J. V.; Oh, M.-K.; Jung, G. Y. Precise Precursor Rebalancing for Isoprenoids Production by Fine Control of GapA Expression in *Escherichia coli*. *Metab. Eng.* **2016**, *38*, 401−408.
(18) Yen, H.-W.; Palanisamy, G.; Su, G.-C. The Influences of Supplemental Vegetable Oils on the Growth and β-Carotene Accumulation of Oleaginous Yeast-*Rhodotorula glutinis*. *Biotechnol. Bioprocess Eng.* **2019**, *24*, 522−528.
(19) Sevgili, A.; Erkmen, O. Improved Lycopene Production from Different Substrates by Mated Fermentation of *Blakeslea trispora*. *Foods* **2019**, *8*, No. 120.
(20) Ma, T.; Shi, B.; Ye, Z.; Li, X.; Liu, M.; Chen, Y.; Xia, J.; Nielsen, J.; Deng, Z.; Liu, T. Lipid Engineering Combined with Systematic Metabolic Engineering of *Saccharomyces cerevisiae* for High-Yield Production of Lycopene. *Metab. Eng.* **2019**, *52*, 134−142.
(21) Yamano, S.; Ishii, T.; Nakagawa, M.; Ikenaga, H.; Misawa, N. Metabolic Engineering for Production of Beta-Carotene and Lycopene in *Saccharomyces cerevisiae*. *Biosci. Biotechnol. Biochem.* **1994**, *58*, 1112−1114.
(22) Hartz, P.; Milhim, M.; Trenkamp, S.; Bernhardt, R.; Hannemann, F. Characterization and Engineering of a Carotenoid Biosynthesis Operon from *Bacillus megaterium*. *Metab. Eng.* **2018**, *49*, 47−58.
(23) Schwartz, C.; Frogue, K.; Misa, J.; Wheeldon, I. Host and Pathway Engineering for Enhanced Lycopene Biosynthesis in *Yarrowia lipolytica*. *Front. Microbiol.* **2017**, *8*, No. 2233.
(24) Gallego-Jara, J.; de Diego, T.; del Real, Á.; Écija-Conesa, A.; Manjón, A.; Cánovas, M. Lycopene Overproduction and in Situ Extraction in Organic-Aqueous Culture Systems Using a Metabolically Engineered *Escherichia coli*. *AMB Express* **2015**, *5*, No. 65.
(25) Kim, Y.-S.; Lee, J.-H.; Kim, N.-H.; Yeom, S.-J.; Kim, S.-W.; Oh, D.-K. Increase of Lycopene Production by Supplementing Auxiliary Carbon Sources in Metabolically Engineered *Escherichia coli*. *Appl. Microbiol. Biotechnol.* **2011**, *90*, 489−497.
(26) Yoon, S.-H.; Kim, J.-E.; Lee, S.-H.; Park, H.-M.; Choi, M.-S.; Kim, J.-Y.; Lee, S.-H.; Shin, Y.-C.; Keasling, J. D.; Kim, S.-W. Engineering the Lycopene Synthetic Pathway in *E. coli* by Comparison of the Carotenoid Genes of *Pantoea agglomerans* and *Pantoea ananatis*. *Appl. Microbiol. Biotechnol.* **2007**, *74*, 131−139.
(27) Yoon, S.-H.; Lee, Y.-M.; Kim, J.-E.; Lee, S.-H.; Lee, J.-H.; Kim, J.-Y.; Jung, K.-H.; Shin, Y.-C.; Keasling, J. D.; Kim, S.-W. Enhanced Lycopene Production In *Escherichia coli* Engineered to Synthesize Isopentenyl Diphosphate and Dimethylallyl Diphosphate from Mevalonate. *Biotechnol. Bioeng.* **2006**, *94*, 1025−1032.
(28) Misawa, N.; Nakagawa, M.; Kobayashi, K.; Yamano, S.; Izawa, Y.; Nakamura, K.; Harashima, K. Elucidation of the Erwinia Uredovora Carotenoid Biosynthetic Pathway by Functional Analysis of Gene Products Expressed in *Escherichia coli*. *J. Bacteriol.* **1990**, *172*, 6704−6712.
(29) Borkowski, O.; Ceroni, F.; Stan, G.-B.; Ellis, T. Overloaded and Stressed: Whole-Cell Considerations for Bacterial Synthetic Biology. *Curr. Opin. Microbiol.* **2016**, *33*, 123−130.
(30) Taylor, G. M.; Heap, J. T. Combinatorial Metabolic Engineering Platform Enabling Stable Overproduction of Lycopene from Carbon Dioxide by Cyanobacteria. *bioRxiv* **2020**, No. 2020.03.11.983833.
(31) Exley, K.; Reynolds, C. R.; Suckling, L.; Chee, S. M.; Tsipa, A.; Freemont, P. S.; McClymont, D.; Kitney, R. I. Utilising Datasheets for the Informed Automated Design and Build of a Synthetic Metabolic Pathway. *J. Biol. Eng.* **2019**, *13*, No. 8.
(32) Bultelle, M.; de Murieta, I. S.; Kitney, R. In *Introducing SynBIS—The Synthetic Biology Information System*, IET/SynbiCITE Engineering Biology Conference; Institution of Engineering and Technology: London, U.K., 2016; pp 1−2.
(33) Kelly, J. R.; Rubin, A. J.; Davis, J. H.; Ajo-Franklin, C. M.; Cumbers, J.; Czar, M. J.; de Mora, K.; Glieberman, A. L.; Monie, D. D.; Endy, D. Measuring the Activity of BioBrick Promoters Using an in Vivo Reference Standard. *J. Biol. Eng.* **2009**, *3*, No. 4.
(34) Mutalik, V. K.; Guimaraes, J. C.; Cambray, G.; Lam, C.; Christoffersen, M. J.; Mai, Q.-A.; Tran, A. B.; Paull, M.; Keasling, J. D.; Arkin, A. P.; Endy, D. Precise and Reliable Gene Expression via Standard Transcription and Translation Initiation Elements. *Nat. Methods* **2013**, *10*, 354−360.
(35) Clifton, K. P.; Jones, E. M.; Paudel, S.; Marken, J. P.; Monette, C. E.; Halleran, A. D.; Epp, L.; Saha, M. S. The Genetic Insulator RiboJ Increases Expression of Insulated Genes. *J. Biol. Eng.* **2018**, *12*, No. 23.
(36) Blazeck, J.; Garg, R.; Reed, B.; Alper, H. S. Controlling Promoter Strength and Regulation in *Saccharomyces cerevisiae* Using Synthetic Hybrid Promoters. *Biotechnol. Bioeng.* **2012**, *109*, 2884−2895.
(37) Salis, H. M.; Mirsky, E. A.; Voigt, C. A. Automated Design of Synthetic Ribosome Binding Sites to Control Protein Expression. *Nat. Biotechnol.* **2009**, *27*, 946−950.
(38) Nishizaki, T.; Tsuge, K.; Itaya, M.; Doi, N.; Yanagawa, H. Metabolic Engineering of Carotenoid Biosynthesis in *Escherichia coli* by Ordered Gene Assembly in *Bacillus subtilis*. *Appl. Environ. Microbiol.* **2007**, *73*, 1355−1361.
(39) Storch, M.; Casini, A.; Mackrow, B.; Fleming, T.; Trewhitt, H.; Ellis, T.; Baldwin, G. S. BASIC: A New Biopart Assembly Standard for Idempotent Cloning Provides Accurate, Single-Tier DNA Assembly for Synthetic Biology. *ACS Synth. Biol.* **2015**, *4*, 781−787.
(40) Sikkema, J.; de Bont, J. A.; Poolman, B. Mechanisms of Membrane Toxicity of Hydrocarbons. *Microbiol. Rev.* **1995**, *59*, 201−222.
(41) Espah Borujeni, A.; Channarasappa, A. S.; Salis, H. M. Translation Rate Is Controlled by Coupled Trade-Offs between Site Accessibility, Selective RNA Unfolding and Sliding at Upstream Standby Sites. *Nucleic Acids Res.* **2014**, *42*, 2646−2659.
(42) Tietze, L.; Lale, R. Importance of the 5′ Regulatory Region to Bacterial Synthetic Biology Applications. *Microb. Biotechnol.* **2021**, 2291.
(43) Woodruff, L. B. A.; Gorochowski, T. E.; Roehner, N.; Mikkelsen, T. S.; Densmore, D.; Gordon, D. B.; Nicol, R.; Voigt, C. A. Registry in a Tube: Multiplexed Pools of Retrievable Parts for Genetic Design Space Exploration. *Nucleic Acids Res.* **2017**, *45*, 1553−1565.
(44) McGuffie, M. J.; Barrick, J. E. PLannotate: Engineered Plasmid Annotation. *Nucleic Acids Res.* **2021**, *49*, W516.
(45) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403−410.
(46) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(47) Smith, T. F.; Waterman, M. S. Identification of Common Molecular Subsequences. *J. Mol. Biol.* **1981**, *147*, 195−197.
(48) Wieds, G. *Bioinformatics Explained: BLAST versus Smith-Waterman*; CLC Bio, 2007.
(49) Armstrong, G. A. Genetics of Eubacterial Carotenoid Biosynthesis: A Colorful Tale. *Annu. Rev. Microbiol.* **1997**, *51*, 629−659.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291

(50) Takehara, M.; Nishimura, M.; Kuwa, T.; Inoue, Y.; Kitamura, C.; Kumagai, T.; Honda, M. Characterization and Thermal Isomerization of (All-E)-Lycopene. *J. Agric. Food Chem.* **2014**, *62*, 264−269.
(51) Espah Borujeni, A.; Zhang, J.; Doosthosseini, H.; Nielsen, A. A. K.; Voigt, C. A. Genetic Circuit Characterization by Inferring RNA Polymerase Movement and Ribosome Usage. *Nat. Commun.* **2020**, *11*, No. 5001.
(52) Cardinale, S.; Arkin, A. P. Contextualizing Context for Synthetic Biology − Identifying Causes of Failure of Synthetic Biological Systems. *Biotechnol. J.* **2012**, *7*, 856−866.
(53) Taylor, G. M.; Mordaka, P. M.; Heap, J. T. Start-Stop Assembly: A Functionally Scarless DNA Assembly System Optimized for Metabolic Engineering. *Nucleic Acids Res.* **2019**, *47*, e17.
(54) RBS Calculator. https://www.denovodna.com/software/predict_rbs_calculator.
(55) Jack, B. R.; Leonard, S. P.; Mishler, D. M.; Renda, B. A.; Leon, D.; Suárez, G. A.; Barrick, J. E. Predicting the Genetic Stability of Engineered DNA Sequences with the EFM Calculator. *ACS Synth. Biol.* **2015**, *4*, 939−943.
(56) EFM Calculator. https://barricklab.org/django/efm/.
(57) Zhang, F.; Carothers, J. M.; Keasling, J. D. Design of a Dynamic Sensor-Regulator System for Production of Chemicals and Fuels Derived from Fatty Acids. *Nat. Biotechnol.* **2012**, *30*, 354−359.
(58) Park, W.-J.; You, S.-H.; Choi, H.-A.; Chu, Y.-J.; Kim, G.-J. Over-Expression of Recombinant Proteins with N-Terminal His-Tag via Subcellular Uneven Distribution in *Escherichia coli*. *Acta Biochim. Biophys. Sin.* **2015**, *47*, 488−495.
(59) Borràs, E.; Sabidó, E. What Is Targeted Proteomics? A Concise Revision of Targeted Acquisition and Targeted Data Analysis in Mass Spectrometry. *PROTEOMICS* **2017**, *17*, No. 1700180.
(60) Sinha, A.; Mann, M. A Beginner's Guide to Mass Spectrometry−Based Proteomics. *Biochemist* **2020**, *42*, 64−69.

https://doi.org/10.1021/acssynbio.1c00562
ACS Synth. Biol. 2022, 11, 1272−1291