# BioScan: Parameter-Space Exploration of Synthetic Biocircuits Using MEDA Biochips*

Mohamed Ibrahim[§], Bhargab B. Bhattacharya[‡], and Krishnendu Chakrabarty[§]
[§]Department of ECE, Duke University, Durham, NC, USA [‡]Indian Statistical Institute, Kolkata, India

***Abstract*—Recent advances in microfluidic technology offer efficient platforms to emulate complex molecular networks of biological pathways (biocircuits) on a lab-on-chip. The behavior of biocircuits is governed by a number of gene-regulatory parameters. A fundamental challenge in synthesizing and verifying biocircuits is the lack of design tools that implement biocircuit-regulatory scanning (BRS) assays to explore the large parameter-space efficiently, while optimizing synthesis time and reagent cost. In this paper, we introduce an optimization flow named BioScan for systematic exploration of the parameter-space of a biocircuit. BioScan includes: (1) a statistical approach to determine a subset of mixing ratios of reagents that span the entire parameter space as densely as possible under cost constraints; (2) an ILP-based synthesis method that implements a BRS-assay on a micro-electrode dot-array biochip. Simulation results show that BioScan reduces reagent cost and enhances space-filling properties.**

## I. Introduction

Synthetic biology has emerged recently with the goal of engineering biological parts that can perform new and useful functions. Applications of synthetic biology include environmental monitoring, therapeutics, and creation of new materials [1]. Such applications can be assembled via pathways that function like integrated circuits, enabling *synthetic biocircuits* [2].

A synthetic biocircuit consists of a cell-free regulatory network of genetic parameters that interact through gene-expression activities. For example, consider a simple biocircuit that has the following parameters (Fig. 1(a)) [2]: (1) $X_1$, a gene that expresses "AraC" (a transcription activator); (2) $X_2$, a gene that expresses "TetR" (a transcription repressor); (3) $X_3$, a gene that expresses GFP (the fluorescent protein output of the circuit). As shown in Fig. 1(a), the activity of "AraC" regulates (through activation) both $X_2$ and $X_3$, but the activity of "TetR" in $X_2$ regulates (through repression) $X_3$. Such interactions are demonstrated in the regulatory network diagram in Fig. 1(b). Clearly, there are two incoherent regulations of GFP expression at $X_3$. As a result, careful modulation of $X_1$, $X_2$, and $X_3$ is necessary to allow rapid testing of the biocircuit *in vitro* and to achieve the desired function [1].

An essential step toward reliable performance of synthetic biocircuits is to maintain the expression of circuit parameters in an optimal range. Such an optimization can be realized using a technique known as *parameter-space exploration* (PSE) [2], which scans all possible combinations of circuit parameters. In practice, PSE can be implemented by generating numerous droplets (a droplet is referred to as a *mixture* $M_i$) of equal volume with varying volumetric ratios of biochemical reagents. A volumetric ratio of a reagent $j$ in $M_i$ is referred to as a *concentration factor* (CF) and is denoted by $c_{(i,j)}$. This ratio can also be used to stimulate gene transcription and translation of biocircuit parameters. Such an assay is referred to as biocircuit-regulatory scanning (BRS). For example, to study the parameter space of the circuit in Fig. 1(a), we can implement a BRS assay that generates three mixtures $M_1$, $M_2$, and $M_3$, as shown in Fig. 1(c). Each mixture constitutes four types of reagents: (1) $Y_1$, which modulates parameter $X_1$; (2) $Y_2$, which modulates parameter $X_2$; (3) $Y_3$, which modulates parameter $X_3$; (4) distilled water $Y_4$, which is used to ensure that the droplet volume remains constant. Also, each mixture contains a unique combination of CFs, referred to as the *CF profile*, as shown in Fig. 1(d). For simplicity, we allow $Y_1$ to have a fixed CF value ($c_{(i,1)}$ = 0.2) among all mixtures, whereas the CF values of $Y_2$, $Y_3$, and $Y_4$ are changed. The CF profiles of the three mixtures are selected such that their representation in the *CF space* can fill as much space as possible; see Fig. 1(c).

Ideally, by generating numerous droplets, the range of any constituent CF can span the full concentration range between 0% and 100% of a droplet volume. This process, however, is cost-prohibitive, especially with the exponential growth in the number of parameters. Hence, a barrier facing PSE is the need

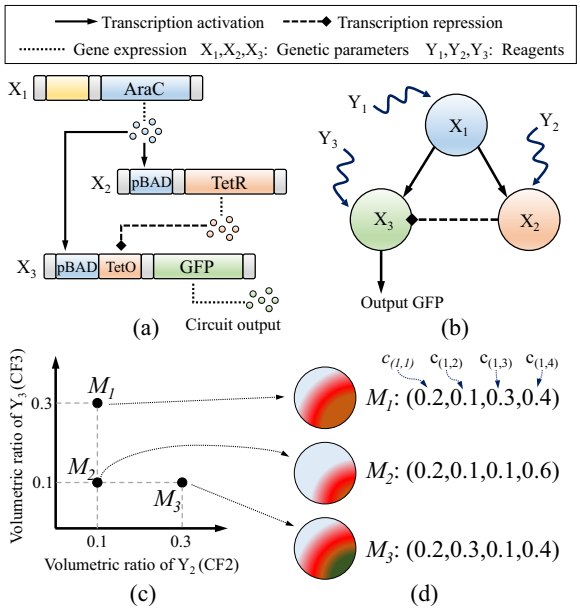



Fig. 1: Study of a 3-parameter biocircuit [2]: (a) schematic of the biocircuit; (b) a gene-expression regulatory network representation of the biocircuit; (c) CF space associated with the biocircuit (reduced to a 2-D space); (d) CF profiles of three mixtures used for PSE.

*This work was supported in part by the National Science Foundation under grant CCF-1702596.

for a systematic methodology that enables dense scanning of the CF space. Recently, a framework based on a flow-based biochip supported with dynamic control of reagent-flow rates has enabled PSE for the circuit of Fig. 1(a) [2]. Despite the novelty of this design, it suffers from the following drawbacks:
**(1) Scalability Limitation:** The flow-based solution performs passive mixing, which relies only on molecular diffusion of fluids without any external energy. Therefore, passive mixing in flow-based systems is prohibitively slow.
**(2) Manual Configuration:** The above design utilizes sinusoidal flow rates of reagents with pre-computed function periods to scan a large volume of the CF space. As the number of reagents is varied, a new set of sinusoidal functions must be computed; such a process can be a significant challenge when a large number of reagents is involved.
**(3) Abandoned CF Subspace:** A fixed droplet volume cannot be maintained unless an input flow of distilled water (DW) is introduced. Note that the flow-rate of DW is computed based on the other sinusoidal functions. This technique allows all mixtures to have equal volume; however, they contain large amounts of DW and therefore this approach causes a large CF subspace to be abandoned. For example, by applying the above technique to a 3-CF setting, the value of each CF can vary only between 0% and 33% (i.e., 100%/3); see Fig. 1(c).

To overcome the above limitations, we need a PSE microfluidic framework that offers reconfigurable mixing and droplet-actuation capabilities, thus enabling flexible composition of biochemical reagents. This requirement can be met using a digital microfluidic biochip (DMFB), which allows the manipulation of nanoliter droplets using an array of electrodes [3]. We consider a specific DMFB architecture referred to as the micro-electrode-dot-array (MEDA), which can support mixing of reagents with considerably higher resolution [4].

In this paper, we introduce the first optimization flow, called BioScan, for PSE in synthetic biology based on MEDA biochips. The key contributions of this paper are as follows:

- BioScan uses a statistical sampling method that scans the CF space and generates a representative collection of CF profiles. The proposed method is scalable in terms of the number of mixtures and space dimensionality, and it ensures that the selected CF profiles fill the entire space.
- We present a synthesis method based on integer linear programming (ILP) that maps the generated CF profiles to a BRS assay on MEDA.

## II. Preliminaries

### A. MEDA Microfluidic Biochips

Digital microfluidics has shown exceptional promise for synthetic biology experiments, including DNA assembly, transformation, culturing, and biocircuits [5]. Advantages of DMFBs are further extended by MEDA, which consists of an array of identical basic microfluidic units called *microelectrode cells* (MCs); see Fig. 2(a-b) [4]. Each MC consists of a microelectrode and a control/sensing circuit. Using this configuration, MEDA biochips can employ the concept of a sea-of-microelectrodes, where microelectrodes can be dynamically grouped

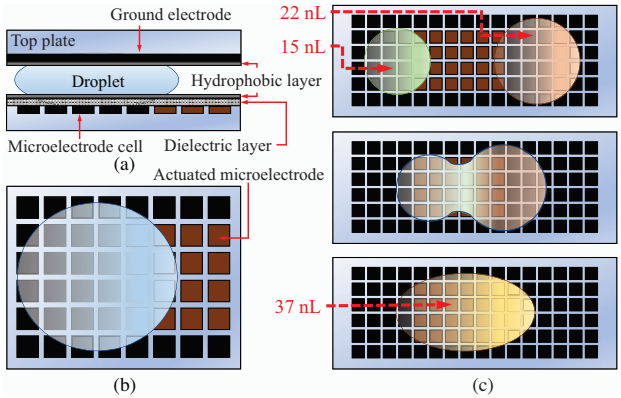


Fig. 2: Droplet actuation in MEDA biochip: (a) side view; (c) top view; (c) MEDA-enabled $(m, n)$ mixing model.

to form a single micro-component (e.g., mixer or diluter) that can perform on-chip biochemical operations.

By using a conventional DMFB, a mixture that comprises several reagents can be systematically prepared using a sequence of $(1:1)$ *mix-and-split* operations only. In this $(1:1)$ model, an intermediate droplet is generated by mixing two reagent droplets of equal volume, and a large droplet can be split only into two small droplets of equal volume. Clearly, this model has limitations, especially if the constituent CFs need to have different values. MEDA, on the other hand, utilizes the dynamic grouping of microelectrodes to extend the above model to a general $(m:n)$ mixing-splitting model [6]; thus enabling droplets with different volumes to be mixed or split with high resolution. As shown in Fig. 2(c), MEDA enables two reagent droplets of volume 15 and 22 *nL* to be mixed to generate a 37 *nL* mixture with CFs $40\%$ and $60\%$, respectively. We exploit such a powerful mixing-splitting model to generate mixtures with a variety of CF profiles used for PSE in biocircuits.

### B. Prior Synthesis Techniques

Advances in genomic analysis and sample preparation on DMFBs have motivated a number of design automation tools in this domain. Although these methods help in bridging the gap between microfluidics and genomics, they are not adequate for handling PSE in synthetic biology [5].

Early research on sample preparation focused on optimizing the dilution process for a single sample with the goal of minimizing reactant-cost, the number of mixing steps, or waste droplets [7]. In [6], the process of sample dilution has been optimized using the $(m:n)$ mixing model offered by MEDA. However, these methods are not capable of handling mixtures that contain three or more reagents. To support dilution gradients in quantitative analysis, multi-target sample-preparation techniques have been introduced [8]. These techniques generate droplets that contain mixtures of only two fluids: sample and buffer. However, these methods are limited to $(1:1)$ mixing and they cannot support the preparation of multiple mixtures that constitute a large number of reagents.

For producing a desired multi-reagent mixture, synthesis methods have been developed to generate a bottom-up mixing tree that encodes the successive composition of reagent volumetric ratios [9]. These methods can be easily adapted to

our space-exploration problem by running multiple iterations of the algorithm; every iteration specifies the mixing of an individual mixture. This approach, however, may lead to a significant increase in the amount of waste droplets and in the protocol completion time.

To make DMFBs useful for dense PSE in synthetic biology, we need a new top-down optimization flow that allows concurrent production of several mixtures with maximum precision, especially in the presence of reagent-usage constraints. Next, we present two enabling components of our proposed flow (BioScan): sampling of the CF space (Section III), and ILP-based synthesis (Section IV). Studies of other components are left for future work.

## III. Sampling of Concentration Factor Space

### A. Stratified Sampling: Latin Hypercubes

Consider a continuous space $\mathcal{X} \subset [0,1)^r$ constructed using input variables $R_x = \{R_{(x,1)}, R_{(x,2)}, ..., R_{(x,r)}\} \in \mathcal{X}$. We seek a sampling technique that selects an ensemble of $n$ samples $x = \{x_1, x_2, ..., x_n\} \subset \mathcal{X}$, where $x_i = \{x_{(i,1)}, x_{(i,2)}, ..., x_{(i,r)}\} | x_{(i,j)} \in R_{(x,j)}$, such that it provides enhanced space-filling properties, i.e., it must ensure that all the space regions are roughly evenly sampled. Theoretically, space-filling criterion can be assessed using the Euclidean maximin ($E_m$) distance, which is defined to be the smallest distance between a pair of points in the sampling space [10]. Let $\boldsymbol{d} = [d_{(1,2)}\ d_{(1,3)}\ ...\ d_{(n-1,n)}]$, where $d_{(i,k)}$ is computed as follows:

$$d_{(i,k)} = \sqrt{(x_{(i,1)} - x_{(k,1)})^2 + ... + (x_{(i,r)} - x_{(k,r)})^2}; \quad i < k$$

The $E_m$ distance is the smallest $d_{(i,k)} \in \boldsymbol{d}$. A larger value of $E_m$ indicates that the distance between the closest points is big, indicating a better space-filling property.

It is known that pseudo-random sampling methods such as Monte Carlo sampling may result in poor space filling [11]. We overcome this limitation by using a classical *stratified* sampling technique known as *Latin Hypercube sampling* (LHS) [10], which divides the range of each input variable $R_{(x,j)} \in R_x$ into $n$ equally probable strata and samples once from each stratum. For each $R_{(x,j)} \in R_x$, the $n$ sampled input values are assigned at random to the $n$ strata, with all $n!$ possible permutations being equally likely. This process is applied independently to each variable $R_{(x,j)}$. LHS can be further enhanced by dividing the sampling space into $l$ equally probable subdivisions, and each subdivision is sampled with the same density; such an enhancement is known as *Orthogonal Array-based Latin Hypercube sampling* (OA-LHS) [10]. Fig. 3 compares these methods using $n = 9$ samples in a 2-D ($r = 2$) space.

According to the above definition, the range of $R_{(x,j)}$ is defined as $R_{(x,j)} \in [0,1)$, and no other constraints are applied to the coordinates $x_{(i,j)} \in R_{(x,j)}$ of the $n$ samples. However, to apply OA-LHS to the CF space to generate a set of $n$ CF profiles $\{M_1, M_2, ..., M_n\}$, we must ensure that the sum of CFs in any CF profile (mixture) must equal 1. In other words, for any mixture $M_i$, the following condition must be satisfied: $\sum_{j=1}^{r} x_{(i,j)} = 1$. Such a requirement cannot be fulfilled through direct application of OA-LHS to the CF space. We explain how to overcome this limitation in Section III-B.

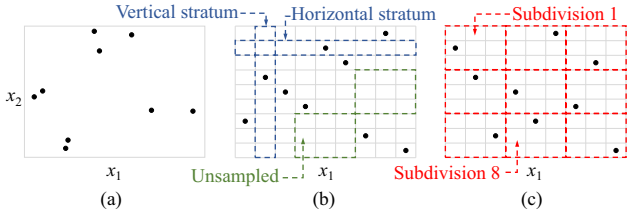


Fig. 3: Sampling 9 points in a 2-D space using: (a) random sampling, (b) LHS, (c) OA-LHS ($l = 9$).

### B. Mapping to CF Space

Consider a continuous space $\mathcal{Y} \subset [0,1)^r$ constructed using input variables $R_y = \{R_{(y,1)}, R_{(y,2)}, ..., R_{(y,r)}\} \in \mathcal{Y}$. A point $y$ in the space $\mathcal{Y}$ is defined as $y = \{y_1, y_2, ..., y_n\} \subset \mathcal{Y}$, where $y_i = \{y_{(i,1)}, y_{(i,2)}, ..., y_{(i,r)}\} | y_{(i,j)} \in R_{(y,j)}$, and it must satisfy the following condition $\sum_{j=1}^{r} y_{(i,j)} = 1; \forall i \in \{1, ..., n\}$. The space $\mathcal{Y}$ can be graphically represented using a simplex that is formed using a barycentric coordinate system [12]. Fig. 4(a) depicts the shapes of 1-simplex (2-D space) and 2-simplex (3-D space).

To adapt OA-LHS to the "simplex" CF space, we seek a mapping function $f_s$ that is defined as $f_s : \mathcal{X} \rightarrow \mathcal{Y}$. A trivial implementation of the function $f_s$ is to uniformly sample points in the space $\mathcal{X}$ using OA-LHS then re-scale the points using the relation $y_{(i,j)} = \frac{x_{(i,j)}}{\sum_{j=1}^{r} x_{(i,j)}}$; this method is referred to as *scaling-based mapping* . However, this approach severely degrades space filling since it tampers with the stratification property; see Fig. 4(b). We develop an alternative implementation of $f_s$ that does not change the uniformity of the sampling by using the *Dirichlet distribution*, which is an exponential family distribution over a simplex, i.e., positive vectors that sum to one. Formally, if $x_{(i,j)} \in [0,1)$ is sampled using OA-LHS based on a uniform distribution, then $f_s : x_{(i,j)} \rightarrow y_{(i,j)}$ can be defined as

$$z_{(i,j)} = \frac{x_{(i,j)}^{\alpha-1} e^{-x_{(i,j)}}}{\Gamma(\alpha)} = \frac{x_{(i,j)}^{\alpha-1} e^{-x_{(i,j)}}}{(\alpha-1)!}; \quad y_{(i,j)} = \frac{z_{(i,j)}}{\sum_{j=1}^{r} z_{(i,j)}}$$

Note that $\sum_{j=1}^{r} y_{(i,j)} = 1$. A key property of this mapping, referred to as *Dirichlet-based mapping*, is that it applies an affine transformation that does not alter the space-filling properties in the simplex space [12]; see Fig. 4(c).

## IV. Synthesis Methodology

BioScan is designed to be an iterative PSE flow that enables composition of new mixtures on a MEDA biochip; new iterations are executed if the accuracy of the constructed model needs to be enhanced. A cost-effective design can generate the new mixtures not only using reagents stored on chip, but also by exploiting mixtures generated during previous iterations.

**Target Mixtures:** At any iteration of the flow, BioScan aims to compute the synthesis solution associated with $n$ new *target mixtures* $\{M_1^t, M_2^t, ..., M_n^t\}$. The CF profile of a mixture $M_i^t$ is defined as the set $\{c_{(i,1)}^t, ..., c_{(i,r)}^t\} = \{y_{(i,1)}, ..., y_{(i,r)}\}$; the values $y_{(i,j)}$ are obtained from the sampling step (Section III).

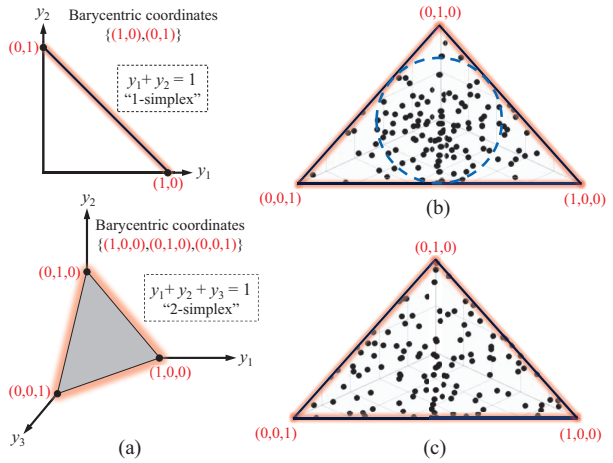



Fig. 4: Mapping OA-LHS-sampled data to simplex CF space: (a) graphical representation of 1-simplex and 2-simplex; (b) scaling-based mapping leads to poor space filling; (c) Dirichlet-based mapping ($\alpha = 3$) preserves enhanced space filling.

Also, each $M_i^t$ is characterized with a "reference" volume $V_i^t$ that the synthesis framework aims to produce. Note that a potential solution for reducing reagent usage during mixture production is to reduce $V_i^t$.

For simplicity (but without loss of generality), we consider the same reference volume for all target mixtures, i.e., $\forall i : V_i^t = V^t$, and we also assume that a droplet volume is represented as a natural number that is multiple of the droplet volume on a single microelectrode. We refer to this measure of volume as a microelectrode (MC) unit.

**Source Mixtures and Reagents:** To generate target mixtures, a number of $m$ *source mixtures* $\{M_1^s, M_2^s, ..., M_m^s\}$, generated in a previous iteration, are used along with $r$ reagent fluids $\{G_1, G_2, ..., G_r\}$—recall that $r$ is also equal to the number of CF-space dimensions. The CF profile of a source mixture $M_k^s$ is defined as the set $\{c_{(k,1)}^s, ..., c_{(k,r)}^s\}$. Also, the CF profile of a reagent fluid $G_j$ is defined as the set $\{c_{(j,1)}^g, ..., c_{(j,r)}^g\}$; where $c_{(j,a)}^g = 1$ only if $j = a$, and $c_{(j,a)}^g = 0$ otherwise.

**Aliquots:** MEDA biochips enable the aliquoting of a source mixture $M_k^s$ of volume $V_k^s$ or a reagent fluid $G_j$ of volume $V_j^g$ into smaller droplets that can vary in volume. Hence, an aliquot of volume $\tau_{(i,k)} \leq V_k^s$ from a source mixture $M_k^s$ contributes to the generation of a target mixture $M_i^t$. Similarly, an aliquot of volume $\theta_{(i,j)} \leq V_j^g$ from a reagent $G_j$ contributes to the generation of the same mixture $M_i^t$. Note that the lower bounds on $\tau_{(i,k)}$ and $\theta_{(i,j)}$, denoted by $V_{k.min}^s$ and $V_{j.min}^g$, respectively, are controlled by the aliquoting constraints imposed by MEDA [13].

**Degree of Concentration Accuracy:** MEDA biochips discretize the CF space. We define parameter $\delta$ as the degree of concentration accuracy, whereby any concentration $c_{(i,j)}^t$ can be expressed as $C_{(i,j)}^t = \lceil c_{(i,j)}^t / \frac{1}{\delta} \rceil$, where $C_{(i,j)}^t$ is an integer. Similarly, $C_{(i,j)}^s = \lceil c_{(i,j)}^s / \frac{1}{\delta} \rceil$. For example, if $\delta = 128$, a concentration of 64% is expressed as $\lceil 0.64 / \frac{1}{128} \rceil = 82$. We use $C_{(i,j)}^t$ or $C_{(k,j)}^s$ to represent a CF in our synthesis flow.

Note that the largest number of target mixtures $\hat{n}$ is impacted by $\delta$—by using stars-and-bars combinatorics, $\hat{n}$ can be computed as follows: $\hat{n} = \binom{\delta-1}{r-1} = \frac{(\delta-1)!}{(r-1)!(\delta-r)!}$. For example, with $\delta = 8$ and $r = 4$, we find that $\hat{n} = 35$.

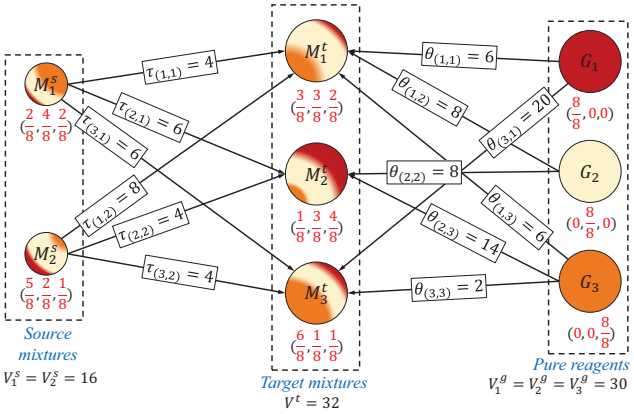



Fig. 5: Preparation of three target mixtures using aliquots from two source mixtures and three reagent fluids ($\delta = 8$).

**BioScan Imprecision:** The first objective of BioScan is to compute the values of $\tau_{(i,k)}$ and $\theta_{(i,j)}$ that enable the production of target mixtures, considering the reference volume $V^t$. However, if $V^t$ is small, the search space of all feasible values of $\tau_{(i,k)}$ and $\theta_{(i,j)}$ becomes limited. As a result, the "actual" volume, denoted by $\hat{V}_i^t$, that is computed by BioScan may not be the same as $V^t$. The value of $\hat{V}_i^t$ can be computed as follows: $\hat{V}_i^t = \sum_k \tau_{(i,k)} + \sum_j \theta_{(i,j)}$. Intuitively, the variation between $\hat{V}_i^t$ and $V^t$ also leads to a variation between the target CF $C_{(i,j)}^t$ and the "actual" CF, denoted by $\hat{C}_{(i,j)}^t$. The actual CF is computed as follows: $\hat{C}_{(i,j)}^t = \lceil \frac{\sum_k (C_{(k,j)}^s \cdot \tau_{(i,k)}) + \delta \cdot \theta_{(i,j)}}{\hat{V}_i^t} \rceil$.

To assess the impact of the variation between $V^t$ and $\hat{V}_i^t$, we introduce a metric named *synthesis imprecision* $\lambda$, where

$$\lambda = \sum_{i,j} \lambda_{(i,j)} = \frac{1}{\delta} \sum_{i,j} \left| \hat{C}_{(i,j)}^t \cdot \hat{V}_i^t - C_{(i,j)}^t \cdot V^t \right|$$
$$= \frac{1}{\delta} \sum_{i,j} \left| \sum_k \left( (C_{(k,j)}^s \cdot \tau_{(i,k)}) + \delta \cdot \theta_{(i,j)} - C_{(i,j)}^t \cdot V^t \right) \right|$$

Hence, our goal is to minimize $\lambda$ while also minimizing $\sum_j \theta_{(i,j)}$ (reagent usage).

Fig. 5 shows an example that illustrates the working principle of the synthesis method. In this example, three target mixtures of volume 32 MC units are generated using aliquots from two source mixtures of volume 16 MC units and three reagent fluids of volume 30 MC units. Using MEDA, our synthesis method first applies split or aliquot operations (to generate the aliquots) followed by mixing operations (to produce the target mixtures). Therefore, we present new types of operation models, namely *split-and-mix* and *aliquot-and-mix*, that replace the classical *mix-and-split* model in previous sample-preparation algorithms.

### A. Problem Formulation

We describe the optimization problem as follows:

**Inputs:** **(i)** The number of target mixtures $n$ and source mixtures $m$. **(ii)** The number of reagents $r$. **(iii)** The degree of accuracy $\delta$. **(iv)** The source mixtures, each with concentration factors $(C_{(k,1)}^s, C_{(k,2)}^s, ..., C_{(k,r)}^s)$ and volume $V_k^s$. **(v)** The target mixtures, each with reference concentration factors $(C_{(i,1)}^t, C_{(i,2)}^t, ..., C_{(i,r)}^t)$ and volume $V^t$. **(vi)** Total liquid

volume $V_j^g$ of reagent $G_j$.

**Output:** **(i)** Volume of aliquots $\tau_{(i,k)}$ and $\theta_{(i,j)}$. **(ii)** Actual CF profile $\{\hat{C}^t_{(i,1)}, ..., \hat{C}^t_{(i,r)}\}$ and actual volume $\hat{V}_i^t$ of every target mixture $M_i^t$. **(iii)** BRS assay.

**Objectives:** **(i)** Minimize reagent usage. **(ii)** Minimize synthesis imprecision. **(iii)** Minimize the completion time of BRS.

To solve the above problem, we develop a two-stage synthesis framework. The objectives of the first stage are to minimize reagent usage and synthesis imprecision; the objective of the second stage is to minimize the protocol completion time.

### B. ILP-Based High-Level Synthesis

High-level synthesis can be optimally solved by mapping the problem to an ILP model. The proposed model is developed to co-optimize synthesis imprecision and reagent usage. The model is described below.

$$\forall i \in \{1, ..., n\}; j \in \{1, ..., r\}; k \in \{1, ..., m\}$$

$$\text{minimize} \quad \frac{\beta}{f^*} \cdot \sum_{i,j} \theta_{(i,j)} + \frac{1-\beta}{g^*} \cdot \sum_{i,j} \lambda_{(i,j)} \tag{1}$$

subject to:

$$V^s_{k.min} < \tau_{(i,k)} \leq V^s_k; \ \sum_i \tau_{(i,k)} \leq V^s_k \tag{2}$$

$$V^g_{j.min} < \theta_{(i,j)} \leq V^g_j; \ \sum_i \theta_{(i,j)} \leq V^g_j \tag{3}$$

$$\sum_k \tau_{(i,k)} + \sum_j \theta_{(i,j)} - V^t \geq 0 \tag{4}$$

$$\lambda_{(i,j)} = \sum_k (C^s_{(k,j)} \cdot \tau_{(i,k)}) + \delta \cdot \theta_{(i,j)} - C^t_{(i,j)} \cdot V^t \geq 0 \tag{5}$$

In (1), we describe the multi-objective function, which aims to minimize reagent usage (by minimizing $\sum_{i,j} \theta_{(i,j)}$) and synthesis imprecision (by minimizing $\sum_{i,j} \lambda_{(i,j)}$) simultaneously. The two objectives are contradicting, thus proper modeling of the problem requires the two objectives to be normalized. For this purpose, we first optimize each objective individually, then divide the related objective term by the corresponding optimum value [14]. The parameters $f^*$ and $g^*$ represent the optimal values of the first and second objectives, respectively.

The parameter $\beta \in [0, 1]$ is introduced to represent the weights for the problem objectives. If $\beta = 1$, then the objective is to minimize reagent usage regardless of the synthesis imprecision. On the other hand, if $\beta = 0$, then achieving the lowest imprecision becomes the only objective. Inequalities (2)-(3) specify volume constraints related to the aliquots. Inequality (2) (Inequality (3)) ensures that the volume of any source aliquot $\tau_{(i,k)}$ (reagent aliquot $\theta_{(i,j)}$) is bounded. We also capture the variation between $V^t$ and $\hat{V}_i^t$ using inequality (4) and the associated impact on the synthesis imprecision using inequality (5).

It is obvious that the value of $V^t$ has a significant impact on the reagent usage and synthesis imprecision; therefore, it needs to be carefully selected. A small value of $V^t$ may lead to reduction in the reagent usage, but it may also increase synthesis imprecision. On the other hand, a large value of $V^t$ may lower synthesis imprecision, but it may also increase reagent usage, thus raising platform cost. In Algorithm 1, we provide a solution to this challenge; We investigate the impact of $V^t$ on the synthesis performance in Section V.

**Algorithm 1** High-Level Synthesis

```
β ← Initialization(); Sol ← ∅;
{V^s_{k.min}, V^g_{j.min}} ← ComputeLowerBoundVolumes(V^s_k, V^g_j);
for V^t ∈ V do                                   ▷ V[1] < V[2]
    {τ_(i,k), θ_(i,j), λ_(i,j)} ← SolveILP(V^t, β, V^s_{k.min}, V^g_{j.min});
    if Σ_{i,j} θ_(i,j) > T_θ OR Σ_{i,j} λ_(i,j) > T_λ then break;
    else Sol ← Sol ∪ {τ_(i,k), θ_(i,j), λ_(i,j)};
BestSol ← SelectLowestImprecision(Sol); return BestSol;
```

### C. Physical-Level Synthesis

The objective of physical-level synthesis is to map the high-level synthesis solution to a sequence of microfluidic operations that implements a BRS assay with the lowest completion time. The droplet-aliquot operation is a key enabler of fine-grained mixture production [13].

Hence, a BRS assay consists of a sequence of MEDA-enabled operations (mixing, splitting, and droplet aliquoting), which can be modeled as a directed acyclic graph $B = (A, U)$, named a *composition graph*. A vertex $a_i \in A$ represents a MEDA-enabled operation, which can be one of four types: (1) mixing; (2) splitting; (3) aliquoting. Each operation type $\Delta$ is associated with a cost value $\rho(\Delta)$, where $\rho(\text{aliquoting}) > \rho(\text{mixing}) > \rho(\text{splitting})$—mixing and splitting are easier and quicker in execution compared to aliquoting. Also, an edge $u_k = \{(a_i, a_j), w_k\} \in U$ models the interdependency between a pair of operations $a_i$ and $a_j$, and $u_k$ is associated with a parameter $w_k$ that captures the droplet volume resulting from $a_i$ and used by $a_j$. The efficient generation of a composition graph is left for future work.

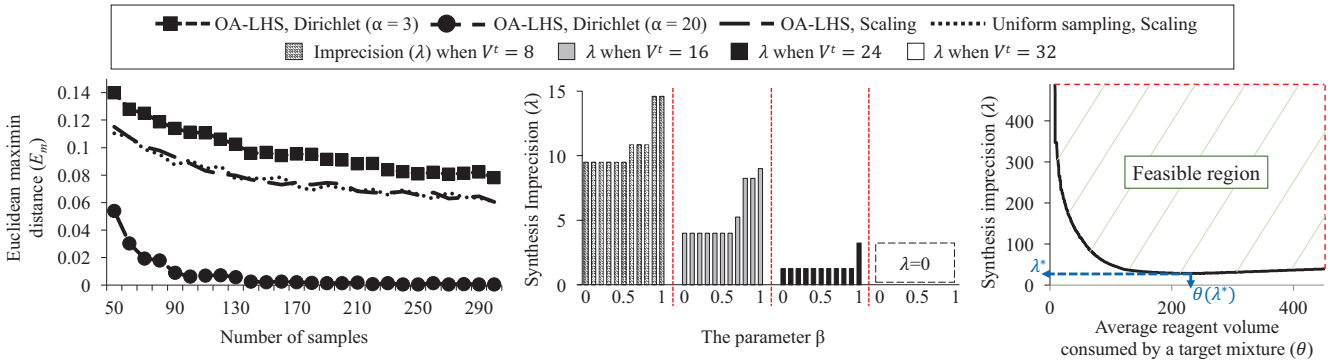
## V. Simulation Results

We implemented BioScan using C++. We solved the ILP model using *lpsolve*, which was integrated in our C++ environment. We assess two aspects of the proposed flow: (1) space filling of the CF profiles, evaluated based on the $E_m$ metric; (2) performance of the synthesis method, evaluated in terms of the synthesis imprecision $\lambda$ and the average reagent volume consumed by a target mixture ($\theta = \frac{\sum_{i,j} \theta_{(i,j)}}{n}$).

### A. Analysis of CF Sampling

Recall that the sampling of the CF space is accomplished in two steps: regular sampling within the interval $[0, 1]$ followed by mapping of samples to the simplex CF space. Therefore, we evaluate the space filling of the CF profiles based on these two steps; we compare four sampling approaches: (1) OA-LHS (stratified sampling) followed by Dirichlet-based mapping where $\alpha = 3$; (2) OA-LHS followed by Dirichlet-based mapping where $\alpha = 20$; (3) OA-LHS followed by scaling-based mapping; (4) uniform sampling followed by scaling-based mapping. The number of CFs, i.e., reagents $r$, is set to 8, and the number of sampling trials in each case is 1000. Results based on other values of $r$ also lead to the same conclusion, showing that our methodology is scalable with $r$—results were not reported here due to lack of space.

Fig. 6(a) compares the above sampling approaches using $E_m$ as a metric while varying the number of targeted samples $n$. We observe that scaling-based mapping degrades the space-filling property, i.e., reduces $E_m$, regardless of which




(a) (b) (c)

Fig. 6: Assessment of BioScan: (a) Impact of sampling and mapping techniques on the space filling. (b) Impact of $\beta$ and $V^t$ on the synthesis performance. (c) The trade-off between $\lambda$ and $\theta$.

sampling method is utilized. This is expected since scaling-based mapping tamper with the stratification property. We also observe that space filling can be severely degraded if the Dirichlet-based mapping is not properly tuned. As shown in Fig. 6(a), Dirichlet-based mapping with $\alpha = 20$ leads to the lowest $E_m$. Large values of $\alpha$ cause the generated samples to be highly concentrated instead of being uniformly distributed. In the following evaluations, we use OA-LHS with Dirichlet-based mapping where $\alpha = 3$.

### B. Impact of $\beta$ and $V^t$

We next evaluate the impact of $\beta$ and $V^t$ on the performance of the high-level synthesis using the example in Fig. 5. For this purpose, we compare four sets of synthesis solutions associated with $V^t = \{8, 16, 24, 32\}$; comparison is performed in terms of the synthesis imprecision $\lambda$. In each set, we compute 11 synthesis solutions by varying the value of $\beta$ from 0 to 1 using 0.1 as the increment between two subsequent values.

Based on Fig. 6(b), we observe that increasing the value of $V^t$ from 8 to 32 gradually lowers the synthesis imprecision. This result corroborates our argument that using a small value of $V^t$ may limit the search space of all feasible aliquots, thus increasing synthesis imprecision (Section IV). We also note that the impact of $\beta$ becomes less significant when $V^t$ is increased.

### C. Synthesis Imprecision vs. Reagent Usage

We next investigate the trade-off between synthesis imprecision and reagent usage. For this study, we consider a more complex setting where the number of reagents $r$ is 6 and the degree of accuracy $\delta$ is 128. The number of source mixtures $m$ is 10 and the number of target mixtures $n$ is 20. We compute the optimal values of $\lambda$ and $\theta$ while varying the target volume $V^t = \{8, 9, 10, ..., 1000\}$ and $\beta = \{0, 0.1, ..., 1\}$ in the second iteration. We then sort these results based on $\theta$ and plot the values of $\lambda$ against the associated values of $\theta$ to perform the trade-off analysis.

Fig. 6(c) shows the trade-off curve and the associated feasible region obtained using the above steps. A first observation is that the lowest values of $\lambda$, denoted by $\lambda^*$, is 28.3; this indicates that the problem setting described above is complex enough that the optimal synthesis solution cannot reach zero synthesis imprecision. Reducing the problem complexity (by minimizing $n$, $m$, and $r$) can result in $\lambda = 0$. We also observe that increasing the amount of reagents beyond $\theta(\lambda^*)$ causes increase in the synthesis imprecision $\lambda$. This counter-intuitive result is observed when $V^t$ becomes significantly large, and even though $\hat{C}^t_{(i,j)} \approx C^t_{(i,j)}; \forall i, j$. The reason for such a finding is that when $V^t$ is increased, a large volume difference $\sum_i(\hat{V}^t_i) - nV^t$ outweighs the impact of $\hat{C}^t_{(i,j)} \approx C^t_{(i,j)}$. This observation is confirmed by computing $\frac{\lambda}{\sum_i(\hat{V}^t_i) - nV^t}$ instead of $\lambda$. We use $\lambda$ to analyze the performance because this allows us to determine the global minimum $\lambda^*$ and the associated $\theta(\lambda^*)$.

## VI. Conclusion

We have introduced an optimization framework for PSE in synthetic biology. We formulated the PSE problem in terms of a BRS assay that is implemented on a MEDA biochip. The proposed framework uses statistical sampling to select reagent mixtures and ILP-based synthesis to generate the reagent mixtures using a BRS assay. Simulation results have shown the effectiveness of BioScan in implementing BRS while minimizing reagent usage and synthesis imprecision.

## References


[1] W. Weber *et al.*, "Emerging biomedical applications of synthetic biology," *Nat. Rev. Genet .*, vol. 13, no. 1, pp. 21–35, 2012.

[2] Y. Hori *et al.*, "Cell-free extract based optimization of biomolecular circuits with droplet microfluidics," *Lab Chip*, vol. 17, 2017.

[3] R. B. Fair, "Digital microfluidics: is a true lab-on-a-chip possible?" *Microfluid. Nanofluid.*, vol. 3, no. 3, pp. 245–281, 2007.

[4] G. Wang *et al.*, "Digital microfluidic operations on micro-electrode dot array architecture," *IET Nanobiotech.*, vol. 5, no. 4, pp. 152–160, 2011.

[5] P. Gach *et al.*, "Droplet microfluidics for synthetic biology," *Lab Chip*, vol. 17, no. 20, pp. 3388–3400, 2017.

[6] Z. Li *et al.*, "Droplet size-aware and error-correcting sample preparation using micro-electrode-dot-array digital microfluidic biochips," *IEEE TBioCAS*, vol. 11, no. 6, pp. 1380–1391, 2017.

[7] T. A. Dinh *et al.*, "A network-flow-based optimal sample preparation algorithm for digital microfluidic biochips," in *Proc. IEEE/ACM ASP-DAC*, 2014, pp. 225–230.

[8] S. Poddar *et al.*, "Optimization of multi-target sample preparation on-demand with digital microfluidic biochips," *IEEE TCAD*, 2018.

[9] S. Roy *et al.*, "Layout-aware mixture preparation of biochemical fluids on application-specific digital microfluidic biochips," *ACM TODAES*, vol. 20, no. 3, p. 45, 2015.

[10] T. Cioppa *et al.*, "Efficient nearly orthogonal and space-filling latin hypercubes," *Technometrics*, vol. 49, no. 1, pp. 45–55, 2007.

[11] H. Niederreiter, *Random number generation and quasi-Monte Carlo methods*. SIAM, 1992, vol. 63.

[12] P. Schneider *et al.*, *Geometric Tools for Computer Graphics*. Elsevier, 2002.

[13] Z. Zhong, Z. Li, and K. Chakrabarty, "Adaptive and roll-forward error recovery in MEDA biochips based on droplet-aliquot operations and predictive analysis," *IEEE TMSCS*, 2018.

[14] R. T. Marler *et al.*, "The weighted sum method for multi-objective optimization: New insights," *Struct. Mult. Opt.*, vol. 41, 2010.