Models of Throughput for Multi-Cell, Multi-Type Droplet Microfluidics

William Krinsman

Graduate Group in Biostatistics; University of California, Berkeley
Division of Computing, Data Science, and Society; University of California, Berkeley
Center for Computational Biology; University of California, Berkeley
BioSciences Area; Lawrence Berkeley National Laboratory

Corresponding author:
William Krinsman
Email address: krinsman@berkeley.edu

ABSTRACT

New experimental platforms encapsulate multiple cells per microfluidic droplet, with each cell belonging to one of multiple possible types. The motivating example comes from microbial ecology, where we want to observe the interactions of microbial strains. Because droplets are formed randomly, we want to accurately predict the data throughput, the numbers of droplets containing desired combinations of cell types.

Herein I identify the default statistical model for predicting the data throughput of multi-cell, multi-type droplet microfluidics experiments, which fits to cell type count data. I explain the assumptions behind this model and issues that in practice may cause these assumptions to fail. One such issue, “compositional heterogeneity”, is unique to multi-type experiments. I show how to modify the default statistical model to describe the consequences of these issues, without needing to mechanistically model their causes.

In practice, only two of these issues may substantially change the data throughput predictions. The changes depend on both (1) which combination of these issues are present, and (2) the precise definition of data throughput. Finally, I show that for a given experimental platform one can estimate the severity of these two issues, enabling more accurate data throughput predictions that account for these two issues.

INTRODUCTION

This paper highlights the work presented in Part II of the PhD Thesis (Krinsman, 2022).

Droplet microfluidics has proven useful for single-cell, single-type experiments. See, for example, (Lagus and Edd, 2013) or (Matula et al., 2020), for a review. New experimental platforms, such as kChip (Kehe et al., 2019) or MINI-Drop (Hsu et al., 2019), use droplet microfluidics to probe the interactions of multiple types of cells. For kChip and MINI-Drop specifically the motivation is to enable more systematic explorations of microbial interactions. Microbial ecology would benefit from experimental, rather than merely observational, data characterizing the interactions of large numbers of microbial strains. The current state of the art for such experiments, by manually culturing and plating combinations of cell types, probes the interactions of fewer than 20 microbial strains simultaneously (Venturelli et al., 2018). Because droplet microfluidics chips produce large numbers of microfluidic droplets, multi-cell, multi-type droplet microfluidics experimental platforms potentially could drastically increase the number of microbial strains whose interactions can be probed simultaneously.

A major tradeoff associated with this approach, compared to manually culturing and plating combinations of cell types, is that droplet microfluidics chips form microfluidic droplets randomly. Cf. Supplementary Figure S4. Therefore, even starting with known microbial communities, it is a priori unclear approximately how many droplets will be produced containing any given combination of cell types. This paper aims to introduce a framework that fills this gap. Doing so will increase the relative advantage of multi-cell, multi-type droplet microfluidics experimental platforms over manually culturing and plating combinations of cell types. Hopefully this can abet future dramatic increases in the state of
the art for the number of microbial strains whose interactions can be probed simultaneously.

**Notation**

Let $S$ denote the total number of cell types in the sampling pool. In the motivating microbial ecology example, “$S$” for “strains” is the total number of microbial strains in the microbial community sample.

Given a positive integer $N \in \mathbb{N}$, define $[N] := \{1, \ldots, n, \ldots, N\}$. Unless specified otherwise, given a (non-negative) vector $\mathbf{v} \in \mathbb{R}^{S}_{\geq 0}$, its entries are denoted $v^{(s)}$, i.e.

$$
\mathbf{v} \overset{\text{def}}{=} (v^{(1)}, \ldots, v^{(s)}, \ldots, v^{(S)}).
$$

(1)

The shorthand $v := \sum_{i=1}^{S} v^{(s)}$ is used for the sum of the entries of $\mathbf{v}$.

For a given droplet, the random variable $N(0)$ is the number of cells from the population encapsulated in the droplet during its formation (“$N$” for “number”). Equivalently, $N(0)$ is the number of cells at “time 0”. This is the sum of the entries of the random vector:

$$
\mathbf{N}(0) := \left(N^{(1)}(0), \ldots, N^{(s)}(0), \ldots, N^{(S)}(0)\right),
$$

(2)

where for each type $s$ (“$s$” for “strain”) of the $S$ total types, $N^{(s)}(0)$ is the random variable equalling the number of cells from type $s$ at time 0.

For each cell type $s$, let $f^{(s)}$ (“$f$” for “frequency”) denote the relative abundance of cell type $s$ in the sampling pool. In particular, we necessarily have that $\sum_{i=1}^{S} f^{(s)} = 1$.

**Precise Definition of Throughput**

The below sections give two different possible definitions of throughput. Cf. supplementary section S4.2.1 for mathematically rigorous versions of these definitions.

**Gluttonous Definition of Data Throughput**

The treatment group $\mathcal{T}^{(s_1,s_2)}_g$ for inferring the effect of type $s_1$ on the growth of type $s_2$ corresponds to droplets where $s_1$ and $s_2$ co-occur. This is called the “gluttonous” treatment group because all droplets where $s_1$ and $s_2$ co-occur are included. The control group $\mathcal{C}^{(s_1,s_2)}_g$ for inferring the effect of type $s_1$ on the growth of type $s_2$ corresponds to droplets where $s_2$ occurs but $s_1$ does not. This is called the “gluttonous” control group because all droplets where $s_2$ occurs but $s_1$ does not are included. Cf. figure 1.

These definitions place no constraints on the presence or absence of types $s \in [S]$ besides $s_1$ and $s_2$. This can be favorable for example when studying very rare types, for which there may be no or very few droplets containing $s_1$ and $s_2$ only, but some containing $s_1$ and $s_2$ along with other types. These definitions give us the best possible chance of avoiding “data starvation”.

**Picky Definition of Data Throughput**

The above “gluttonous” definitions correspond to more possible combinations of cell types than we probably would have included when studying the effect of type $s_1$ on the growth of type $s_2$ by manually plating combinations of cell types. This motivates the following definitions.

The set of droplets where $s_1$ and $s_2$ co-occur, and all other types are absent, is called the “picky” treatment group $\mathcal{T}^{(s_1,s_2)}_p$. The set of droplets where only $s_2$ occurs, and all other types are absent, is called the “picky” control group $\mathcal{C}^{(s_1,s_2)}_p$. Cf. figure 1.

Unlike the gluttonous groups, the picky groups do place constraints on the presence or absence of types $s \in [S]$ besides $s_1$ and $s_2$. For both the treatments and the controls, for “picky” groups all types besides $s_1$ or $s_2$ must have zero counts. Every picky group is a subset of its gluttonous counterpart.

The picky definitions can be favorable when we have plenty of droplets from which to make estimates. They reduce the possibility of confounding effects on the growth of type $s_2$ that could be caused by types that are not type $s_1$. On the other hand, in instances where there are only very few or no droplets without “extra” types, the picky definitions could lead to “data starvation”.

**Statistical Motivation**

The treatment group $\mathcal{T}^{(s_1,s_2)}$ for the effect of type $s_1$ on type $s_2$ consists of droplets containing both types. Likewise, the control group $\mathcal{C}^{(s_1,s_2)}$ for the effect of type $s_1$ on type $s_2$ consists of droplets containing type...
Figure 1. Gluttonous vs. Picky Definitions of Data Throughput. Cf. text.
s_2 but not type s_1. With these groups, we can quantify the effect of type s_1 on type s_2 by targeting an average treatment effect (ATE) or other similar causal estimand.

Even when we can implement estimators for such causal estimands, sometimes the results may be unreliable. Consider e.g. a situation where our treatment group T^{(s_1,s_2)} consists of only T^{(s_1,s_2)} = 5 droplets and our control group C^{(s_1,s_2)} consists of C^{(s_1,s_2)} \approx 50,000 droplets. This example may seem unrealistic, but it is actually optimistic compared to some of the disparities between the treatment and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)} that one may expect to see in practice. Cf. the discussion in (Krinsman, 2022, Appendix 2.C). Regardless of how we choose to quantify interaction effects, we need to estimate the reliability of our estimates in order for them to be scientifically useful. We need to estimate their statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}).

Of course (even for a fixed causal estimand) the statistical power \mathcal{P} function will depend on the particular estimator, so at this level of generality we cannot complete such a calculation. However, a priori we do also know that the statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) will always depend on the sizes T^{(s_1,s_2)}, C^{(s_1,s_2)} of the treatment group and control group, respectively. Thus, to estimate statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}), we always need to know the treatment group and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)}, regardless of the particular estimator.

Unfortunately, due to the random nature of droplet formation in droplet-based microfluidics experiments, the scientist can never directly control the treatment group and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)}. The treatment group and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)} are random variables. Therefore, for any estimator of any causal estimand, the statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) will itself be a random variable.

This is not inherently a “deal-breaker”, but we do still need to constrain the possibilities for the distribution of the statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}). At the very least, we would like to be able to compute the expected statistical power E[\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})].

Specifically, for a given estimator of a given causal estimand, the statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) will be a deterministic function \mathcal{P} of the treatment and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)} (and possibly other factors which for simplicity we will assume to be known and fixed). If we know specific values T^{(s_1,s_2)} = T^{(s_1,s_2)}, C^{(s_1,s_2)} = C^{(s_1,s_2)} for the treatment and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)}, then we can evaluate this statistical power function \mathcal{P} and get a deterministic value \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}). If instead the treatment and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)} are random variables, then when evaluating this statistical power function we get a new random variable \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) whose distribution is determined by the pushforward measure.

Obviously, without specifying the particular causal estimand and its particular estimator, we still cannot complete the calculation of the expectation E[\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})] of a random statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}). Nevertheless, for any given estimator of any given causal estimand, the distribution of the statistical power \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) will be determined by the distribution of the treatment group and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)}. Thus, for any estimator of any causal estimand, as a prerequisite for being able to estimate expected statistical power E[\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})], we must first understand the distribution of the treatment group and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)}.

Note that the treatment group and control group sizes T^{(s_1,s_2)}, C^{(s_1,s_2)} can be understood as sums of indicator random variables:

\begin{equation}
T^{(s_1,s_2)} = \sum_{d \in [D]} \mathbb{I}\left\{ d^{\text{'th droplet in } T^{(s_1,s_2)} \right\}, \quad C^{(s_1,s_2)} = \sum_{d \in [D]} \mathbb{I}\left\{ d^{\text{'th droplet in } C^{(s_1,s_2)} \right\}.
\end{equation}

To completely characterize the distribution of indicator random variables it suffices to know their expectations, i.e. the probabilities:

\begin{equation}
\mathbb{P}\left( d^{\text{'th droplet in } T^{(s_1,s_2)} \right), \quad \mathbb{P}\left( d^{\text{'th droplet in } C^{(s_1,s_2)} \right).
\end{equation}

If we assume the droplets are identically distributed, then linearity of expectation allows us to compute the expected treatment and control group sizes:

\begin{equation}
E\left[ T^{(s_1,s_2)} \right] = D \cdot \mathbb{P}\left( d^{\text{'th droplet in } T^{(s_1,s_2)} \right), \quad E\left[ C^{(s_1,s_2)} \right] = D \cdot \mathbb{P}\left( d^{\text{'th droplet in } C^{(s_1,s_2)} \right).
\end{equation}

This is already enough to compute \mathcal{P}\left( E\left[ T^{(s_1,s_2)} \right], E\left[ C^{(s_1,s_2)} \right) \right), a quantity that we can use as a (potentially suboptimal) stand-in for the expected statistical power E[\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})] in situations where
we don’t want to make additional assumptions about the distributions of the droplets.

However, if we additionally assume that the droplets are mutually independent, then we can compute the distributions of the treatment and control group sizes as marginals of a Multinomial distribution. Given the statistical power function $\mathcal{P}$, i.e. choice of a particular causal estimand and estimator thereof, this would then in principle be enough information to compute the distribution of the statistical power $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})$ or any summary statistic thereof.

One might object to treating the statistical power $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})$ as a random variable. Certainly, after the experiment has been run, the treatment group and control group sizes $T^{(s_1,s_2)}, C^{(s_1,s_2)}$ will be known, allowing us to condition on their observed values $T^{(s_1,s_2)}, C^{(s_1,s_2)}$ and “restore determinism” by evaluating $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) = \mathbb{E} \left[ \mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)}) \mid T^{(s_1,s_2)} = T^{(s_1,s_2)}, C^{(s_1,s_2)} = C^{(s_1,s_2)} \right]$. Yet this is insufficient to help the scientist before running the experiment. The scientist has no direct control in advance over what exactly the treatment group and control group sizes $T^{(s_1,s_2)}, C^{(s_1,s_2)}$ will turn out to be.

However, the distribution of the treatment group and control group sizes $T^{(s_1,s_2)}, C^{(s_1,s_2)}$ should depend on deterministic factors that the scientist can directly control, like the total number of droplets $D$, the expected number of cells per droplet $\lambda$, and the relative abundances $f(t)$ of the various types $s$ in the sampling population. The scientist would like to plan the experiment (before running it) in a way that allows them to (most likely) achieve certain thresholds of statistical power $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})$ for certain interactions of types. Therefore the best way to help the scientist is to predict which values of statistical power $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})$ they will most likely be able to achieve as a function of the deterministic factors that the scientist can directly control. From that perspective it is inescapable that

1. the statistical power $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})$ must be treated as random,

2. the most basic prerequisite for constraining the distribution of the statistical power $\mathcal{P}(T^{(s_1,s_2)}, C^{(s_1,s_2)})$ is to characterize the distributions of the treatment group and control group sizes $T^{(s_1,s_2)}, C^{(s_1,s_2)}$, and

3. (if possible) the distributions of the treatment group and control group sizes $T^{(s_1,s_2)}, C^{(s_1,s_2)}$ should be described as functions of the deterministic factors that the scientist can directly control.

There is no need to incubate the droplets in order to estimate the probabilities (4). Indeed, it should be possible to estimate the probabilities using a much smaller number of total droplets $D$ than what might be required in the final experiment to achieve certain (expected) statistical power thresholds. Therefore I propose that scientists, to be able to design the experiment in a way that allows them achieve certain (expected) statistical power thresholds, first run a “$r = 0$” version of the experiment that skips incubating the droplets and uses a smaller number of total droplets $D$. This both (i) maximizes the usefulness of the data for estimating the distributions of the treatment group and control group sizes (because it should mitigate artifacts resulting from microbial growth and competition, e.g. censoring) and (ii) decreases the cost (in money and time) of this “pre-experiment”. 

A Default Working Model: hPoMu (hierarchical Poisson Multinomial)

The distributions of the “hierarchical Poisson Multinomial” (hPoMu) working model are hierarchical distributions. The total number of cells in the droplet has a Poisson distribution, while ($\mathcal{P}$)

$$\mathcal{P}(\mathcal{N}(0) = \mathbf{n}) = \mathcal{P}(N(0) = n) \cdot \mathcal{P}(\mathcal{N}(0) = \mathbf{n} \mid N(0) = n) = \frac{e^{-\lambda} \lambda^n}{n!} \cdot \left( \frac{n}{n^{(1)} \ldots n^{(S)}} \right) \prod_{s=1}^{S} \left( f^{(s)} \right)^{n^{(s)}}.$$  

1 Cf. sections S3.2.2 and S4.2.2.

2 In practice, given realistic problem sizes, evaluating this exact formula would most likely be intractable, such that one would probably also need to use additional heuristics and approximations. The choice of those heuristics and approximations would depend on the particular causal estimand and estimator thereof, and thus is outside the scope of this discussion.
An alternative form of expressing the likelihood for hPoMu, which reveals that its marginal distributions correspond to mutually independent Poisson-distributed random variables, is

\[ P(\hat{N}(0) = \vec{n}) = \prod_{s=1}^{S} \frac{e^{-f(s)\lambda} (f(s)\lambda)^n}{n!}. \]  

Heuristically speaking, equation (7) says that during the formation of each droplet the sampling rate \( \lambda \) is evenly spread out among the multiple types according to their frequencies in the population.

**Assumptions Justifying Default Working Model**

Both of the two derivations justifying hPoMu as the default working model, cf. supplementary section S2.3.2, make the following implicit assumptions (among others):

- For every type \( s \), the number of cells in the population which belong to type \( s \) is very large.

- Each cell (regardless of its type) has the same small probability of ending up in the droplet as any other cell.

- Whether a given cell (regardless of its type) ends up in the droplet is completely independent of what happens to any other cell.

The first can be thought of as stating that the population size for each type is effectively infinite. The second and third can together be thought of as stating that the sampling pool is perfectly homogeneous.

**Failures of Default Working Model Assumptions**

While hPoMu is definitely a sensible working model to start with, its implicit assumptions failing to be realistic could undermine its usefulness in practice.

**Populations are Finite**

We assumed that for each individual type \( s \) the number of cells is “effectively infinite”. For “rare” types it is a priori unclear whether this assumption is reasonable. However in practice, even for types with relative abundance as low as 0.01\%, taking the finiteness of the populations into account does not lead to substantially different predictions than the default hPoMu model. (Cf. sections S3.4.1 and S4.4.1)

**Density Heterogeneity**

In practice different sections of the sampling pool could conceivably have average numbers of cells per unit volume that differ from the average number of cells per unit volume for the entire sampling pool (i.e. the total number of cells divided by the volume of the sampling pool). Herein I call this phenomenon “density heterogeneity”. Higher density heterogeneity implies more variance in the numbers of cells per unit volume throughout the sampling pool. Cf. figures 2 and 3.
The default Poisson model for the distribution of the total number of cells in each droplet does not account for additional variance which might be introduced in practice by density heterogeneity. Because this describes extra variance in the total number of cells, regardless of their types, density heterogeneity is relevant even when there is only one type. Thus density heterogeneity is relevant even for the single-cell\(^3\), single type experiments which have already been intensively developed.

**Compositional Heterogeneity**

Although the relative abundance of any given cell type \(s\) across the entire sampling pool is a fixed value \(f^{(s)}\), conceivably different sections of the sampling pool could have relative abundances for the cell types that differ from the relative abundances for the entire sampling pool. Herein I call this phenomenon

\(^3\)The goal for these experiments is to have one cell per droplet, whence the name “single-cell”. However, because the droplet formation process is random, in practice some droplets in “single-cell” experiments may still end up with more than one cell.
“compositional heterogeneity”. Higher compositional heterogeneity implies more variance in the relative abundances for the cell types throughout the sampling pool. Cf. figures 4 and 5.

Figure 4. Lower Compositional Heterogeneity. The average relative abundances of each cell type for the entire sampling pool are $\frac{1}{2}$ blue, $\frac{1}{4}$ pink, $\frac{1}{4}$ yellow.

Figure 5. Higher Compositional Heterogeneity. The average relative abundances of each cell type for the entire sampling pool are still $\frac{1}{2}$ blue, $\frac{1}{4}$ pink, $\frac{1}{4}$ yellow.

The default multinomial model for the relative abundances of the cell types in the droplet (conditional upon the total number of cells) does not account for additional variance that might be introduced in practice by compositional heterogeneity. Because compositional heterogeneity describes extra variance in the relative abundances of the cell
types, compositional heterogeneity is relevant only for multi-cell, multi-type experiments. Because such experiments are newer and still less common than single-cell, single type experiments, it seems that compositional heterogeneity may have never before been identified as a potential issue affecting the data throughput of droplet microfluidics experiments.

**Working Model Accounting for Assumption Failures: ghNBDM**

This working model describes how excess variances relative to the completely homogeneous situation (modelled by hPoMu) could affect data throughput. This working model is not a mechanistic model and it does not predict how physical processes create those excess variances. (Cf. supplementary section S2.4.4.) The density concentration $\zeta_D$ and compositional concentration $\zeta_C$ parameters specify this description of the excess variances. These parameters are separate from and not functions of the relative abundances of the cell types $f^{(i)}$ nor the mean number of cells per droplet $\lambda$. This working model can be fit to real data of cell type counts using the procedures from supplementary section S5 for estimating the density concentration $\zeta_D$ and compositional concentration $\zeta_C$ parameters. That alone is sufficient to access the working model’s predictions for how excess variances relative to the completely homogeneous situation could affect data throughput.

The “generalized hierarchical Negative Binomial Dirichlet-Multinomial” (ghNBDM) family is a family of hierarchical distributions. The number of cells in the droplet is Negative Binomial distributed with “density concentration” parameter $\zeta_D S$, while (conditional upon the total number of cells) the numbers of cells belonging to the multiple types have a joint Dirichlet-Multinomial distribution with “compositional concentration” parameter $\zeta_C S$. Thus the likelihood equals

$$\begin{align*}
P\left(N(0) = \mathbf{n}\right) &= P\left(N(0) = \mathbf{n} \mid N_j(0) = n\right) \\
&= \frac{\Gamma(\zeta_D S + n)}{\Gamma(\zeta_D S)} \frac{(\zeta_D S)^{\zeta_D S}}{(\zeta_D S + \lambda)^{\zeta_D S + n}} \frac{\lambda^n}{n!} \frac{\Gamma(\zeta_C S)}{\Gamma(\zeta_C S + n)} \left[ \prod_{i=1}^S \frac{\Gamma(\zeta_C S f^{(i)} + n^{(i)})}{\Gamma(\zeta_C S f^{(i)})} \right] \left(\frac{n}{n^{(1)} \ldots n^{(S)}}\right).
\end{align*}$$

In general the marginal distributions of members of the ghNBDM family have non-zero cross-covariance. Thus the entries $N^{(i)}(0)$ of the random vector $\mathbf{N}(0)$ will in general not be mutually independent. This is unlike the default working model (hPoMu), cf. again equation (7).

Notice how this working model has two additional parameters compared to the hPoMu working model, namely the density concentration $\zeta_D$ parameter and the compositional concentration $\zeta_C$ parameter. Higher values of the density concentration $\zeta_D$ parameter correspond to lower density heterogeneity (cf. supplementary section S2.7). As $\zeta_D \to \infty$ the total number of cells in the droplet approaches a Poisson distribution. Cf. section S2.7.2. Similarly, higher values of the compositional concentration $\zeta_C$ parameter correspond to lower compositional heterogeneity (cf. supplementary section S2.6). As $\zeta_C \to \infty$ the joint distribution of the numbers of cells belonging to the multiple types approaches a Multinomial distribution. Cf. supplementary section S2.6.2. The hPoMu model is approached in the limit as both $\zeta_D \to \infty$ and $\zeta_C \to \infty$ jointly. Cf. supplementary figure S11 and supplementary section S2.8.1.

For details about the calculations demonstrating how the somewhat complicated expressions involving the $\Gamma$ function do indeed generalize the likelihoods of the Poisson and Multinomial distributions, cf. sections S2.7.2, S2.6.2, and S2.10. At a high level, the idea is to use Stirling’s approximation, in particular the formulations in terms of explicit bounds given by (Robbins, 1955) and (Gordon, 1994).

**METHODS**

I computed and stored results using NumPy (Harris et al., 2020) version 1.20.2. I plotted results using Matplotlib (Hunter, 2007) version 3.4.1 and Seaborn (Waskom, 2021) version 0.11.1. Kernel density estimates were made using the default Seaborn (Waskom, 2021) settings. The bar plot (Figure 6) was made using the microbiome (Lahti and Shetty, 2019), phyloseq (McMurdie and Holmes, 2013), and ggplot2 (Wickham, 2016) packages for the R programming language. Complete implementation details can be found in the code at the relevant GitLab repository. See https://gitlab.com/krinsman/droplets.

---

1 Via SciPy (Virtanen et al., 2020), which chooses bandwidths according to Scott’s Rule (Scott, 1992).
For all simulations and distributions, the same “simulated community” of 91 types was used. These correspond to 90 types distributed across 9 distinct relative abundances: 10 types for each of .01%, .02%, .05%, .1%, .2%, .5%, 1%, 2%, and 5%. The 91st type was a “remainder” type with abundance ≈ 12%. All distributions used the same value of the rate parameter, λ = 2, the expected total number of cells for each droplet. Each distribution was statistically independently simulated 500 times, with each simulation having 15 million statistically independently simulated droplets. See section S3.3 for details.

For details about how the expected values under hPoMu were computed, see section S4.3.1. For each simulated distribution, I stratified and then counted the droplets from each of the 500 simulations according to how many types were present. I then computed the average values for each distribution as the arithmetic mean over the 500 simulations. I then computed the average percent change for each distribution relative to hPoMu using the computed expected values under hPoMu.

For details about how the expected values under hPoMu were computed, see sections S4.3.2 and S4.2.1. For each simulated distribution, for each of the 500 simulations I computed via brute force the number of occurrences of each of the gluttonous groups, and then reported the arithmetic means over all 500 simulations. I again computed the average percent change as the observed value minus the value expected under hPoMu divided by the value expected under hPoMu.

Using the formulae (S64) and (S65), I computed the expected count for each picky group under hPoMu. The procedure was then otherwise exactly the same as for the gluttonous groups.

For each simulated distribution, for each of the 500 simulations, for each of the three batch sizes (small = 10,000 droplets, medium = 500,000 droplets, large = 15,000,000 droplets), I partitioned the simulation results into equal parts of the given batch size. Each simulation contains 15,000,000 droplets, so this corresponds to 1,500 small batches per simulation, 30 medium batches per simulation, and 1 large batch per simulation. Then for each of the batches, I computed the maximum likelihood density concentration estimate and the maximum likelihood compositional concentration estimate and stored the results. For details about how maximum likelihood estimators were computed, see sections S5.3.2 and S5.2.

Compositional heterogeneity, which is unique to multi-type experiments, differently affects gluttonous and picky data throughput, depending on the presence or absence of density heterogeneity. These differences can be explained by breaking down the changes according to the number of types in each droplet.

From figures 6 and 7 we see that compositional heterogeneity decreases the number of droplets with three or more types in both absolute and relative terms. Since droplets with three or more types are a component of the gluttonous groups, their decrease with increased compositional heterogeneity explains the decreases in gluttonous throughput visible in figures 8, and 9. We can also see from figures 7, 8, and 9 that the decrease compositional heterogeneity causes in the number of droplets with three or more types is more pronounced in the absence of density heterogeneity.

How compositional heterogeneity affects the numbers of droplets with two or fewer types depends on whether density heterogeneity is also present.
Figure 6. "Comp. Het. Only" = Compositional Heterogeneity Only.
"Comp. + Dens. Het." = Compositional and Density Heterogeneities.
Cf. Supplementary Figure S36.
Figure 7. "Comp. Het. Only" = Compositional Heterogeneity Only. "Comp. + Dens. Het." = Compositional and Density Heterogeneities. Cf. Supplementary Figure S37.

Compositional Heterogeneity Only
In the absence of density heterogeneity, compositional heterogeneity increases the number of droplets with one type. Compositional heterogeneity also causes a smaller increase in the number of droplets with two types. This is most visible in figure 7. This slight increase in the number of droplets with two types appears to cause the slight increase in the number of picky treatments visible in figure 10. As is clear also from figure 6, there is little change in the number of empty droplets.

Compositional and Density Heterogeneity
In the presence of density heterogeneity, compositional heterogeneity still increases the number of droplets with a single type, as visible in figure 7. However, in this case compositional heterogeneity also decreases the number of droplets with two types. This is visible in figure 7, but more clearly evident from the decrease in the number of picky treatments visible in figure 11.

Nevertheless, it is also clear from figure 7, and from a comparison of figures 11 and 9, that the decrease in the number of droplets with two types is not as pronounced (at least in relative terms) as the decrease in the number of droplets with three or more types. Also in contrast to what happens in the absence of density heterogeneity, in the presence of density heterogeneity compositional heterogeneity markedly increases the number of empty droplets, as shown in figures 6 and 7.

Estimators Behave Well
The estimators satisfy the two minimal requirements for the behavior of a “decent” estimator. Both their accuracy and precision increase with the size of the batches. The distribution of the estimated concentrations is centered around the true value. Moreover, the variance of the distribution of estimated concentrations decreases as the batch size increases. Cf. figures 12 and 13.

DISCUSSION
The Effect of Heterogeneity on Data Throughput is Equivocal
The general effect that heterogeneities will have on the data throughput of multi-cell, multi-type droplet microfluidics experiments is not straightforward. The answer depends on the particular combination of
Figure 8. Cf. Supplementary Figure S45.
Figure 9. Cf. Supplementary Figure S47.
Figure 10. Cf. Supplementary Figure S49.
Figure 11. Cf. Supplementary Figure S51.
Figure 12. Cf. Supplementary Figure S89.
Figure 13. Cf. Supplementary Figure S87.
density heterogeneity and compositional heterogeneity present.

Given both density and compositional heterogeneity, it is clear that the effect on data throughput is negative. Both the numbers of picky treatments and gluttonous treatments decrease. Regardless of whether one uses picky or gluttonous controls, the number of treatments is still the limiting “bottleneck”.

Given only compositional heterogeneity but no density heterogeneity, whether the effect on data throughput is positive depends primarily on our opinion of what is useful data. If our primary concern is avoiding “data starvation” and being able to say anything about the interactions of the rarest types, then the strong decrease in the number of gluttonous treatment groups clearly means that the effect of heterogeneity is negative. On the other hand, if we consider only the data from the picky treatments useful, then in the absence of density heterogeneity the effect of compositional heterogeneity is slightly positive.

The Estimators are Good Enough for Use in Practice
The above results show that the proposed estimators both (1) converge towards correct answers as the data size increases and (2) have decreasing variance as the data size increases. Therefore they satisfy the minimal requirements one generally expects of “decent” or “adequate” statistical estimators.

The estimators were already guaranteed to be asymptotically consistent when the statistical model was correctly specified. Nevertheless, that theoretical guarantee does not necessarily translate in practice into estimates usefully approximating the truth for realistic data sizes (represented by the medium or large batches). Consistency of an estimator need not imply efficiency of an estimator. The above results show that the proposed estimators are in fact useful in practice. Not only that, but they often manage to give useful results even for unrealistically small data sizes (represented by the small batches).

CONCLUSIONS
We need to be able to estimate both density and compositional heterogeneity from empirical data produced by multi-cell, multi-type droplet microfluidics experiments in order to predict how these heterogeneities might affect the throughput of a given experimental platform. We have viable statistical estimators for both density heterogeneity and compositional heterogeneity. Therefore, neither quantity needs to be treated as an “unfathomable unknown”. We are able to actively account for the effects of both. Given empirical data, even of only moderate size, we can get estimates for concentration parameters describing levels of heterogeneity that are realistic for the experimental platform.

Having some sense for the amount of either type of heterogeneity one is likely to encounter in practice for any given experimental platform is important. It enables more accurately predicting the throughput from multi-cell, multi-type droplet microfluidics experiments. More accurate predictions of throughput will make droplet microfluidics experiments more viable alternatives to manually co-culturing multiple types of cells. More accurate predictions of throughput are also necessary to accurately simulate multi-cell, multi-type droplet microfluidics experiments. Accurate simulations in turn will help us to identify the statistical analyses that perform best for the novel kinds of datasets produced by multi-cell, multi-type droplet microfluidics experiments. Identifying best performing statistical analyses for these experiments will further make them more viable alternatives to manually co-culturing multiple types of cells.

ACKNOWLEDGMENTS
Professor Adam P. Arkin for critical suggestions, essential scientific and technical advice, important introductions, and access to funding and computational resources without which this project would have been impossible. Dr. Fangchao Song for suggesting the topic of the dissertation (Krisman, 2022) and generous consultations, feedback, and advice throughout. Professor Mark van der Laan for critical suggestions, essential scientific and technical advice, and invaluable recommendations of strategies and resources for scientific writing. Dr. Lauren M. Lui for invaluable recommendations of strategies and resources for scientific writing. Professor Perry de Valpine for the useful suggestion of investigating the Dirichlet-Multinomial and related models for weakening the ‘well-mixed’ droplets assumption.

REFERENCES
Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M.,
Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585:357–362.

Hsu, R. H., Clark, R. L., Tan, J. W., Ahn, J. C., Gupta, S., Romero, P. A., and Venturelli, O. S. (2019). Microbial interaction network inference in microfluidic droplets. *Cell Systems*, 9(3):229–242.e4.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Kehe, J., Kulesa, A., Ortiz, A., Ackerman, C. M., Thakku, S. G., Sellers, D., Kuehn, S., Gore, J., Friedman, J., and Blaine, P. C. (2019). Massively parallel screening of synthetic microbial communities. *Proceedings of the National Academy of Sciences*, 116(26):12804–12809.

Krinsman, W. (2022). *Statistics of High-Throughput Characterization of Microbial Interactions*. PhD thesis, University of California, Berkeley.

Lagus, T. P. and Edd, J. F. (2013). A review of the theory, methods and recent applications of high-throughput single-cell droplet microfluidics. *Journal of Physics D: Applied Physics*, 46(11):114005.

Lahti, L. and Shetty, S. (2012-2019). microbiome r package.

Matula, K., Rivello, F., and Huck, W. T. S. (2020). Single-cell analysis using droplet microfluidics. *Advanced Biosystems*, 4(1):1900188.

McMurdie, P. J. and Holmes, S. (2013). phyloseq: An r package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, 8(4):1–11.

Scott, D. (1992). *Multivariate Density Estimation : Theory, Practice, and Visualization*. John Wiley & Sons, New York, Chicester.

Venturelli, O. S., Carr, A. V., Fisher, G., Hsu, R. H., Lau, R., Bowen, B. P., Hromada, S., Northen, T., and Arkin, A. P. (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biology*, 14(6):1–19.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayoror, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, I., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.