Toward Predictive Inference of Cell Population Heterogeneity through Birth-Death Approximation and Maximum Entropy

Karthik Desingu
Yale University

May 13, 2024

1 Introduction

The response of genetically identical cells to the same extracellular signal is largely heterogeneous in both response time and strength, even in a homogeneous environment [1]. This heterogeneity ultimately leads to diverse, often unpredictable clinical complications, such as drug response variability [2] and phenotypic plasticity [3] among others. Understanding, quantifying, and predicting this heterogeneity is an important research question that can help dissect cell signaling dynamics in diverse contexts. One of the major sources of cell signaling heterogeneity arises from heterogeneity in the underlying signal cascades, such as the number of signaling receptors [4]. These signaling cascades can be described using two constituent elements: (1) species abundances of participating signaling molecules, and (2) parameters that outline the signaling dynamics by relating the abundances of participating species. Species abundances can be measured experimentally. Defining the cascade parameters as probability distributions to capture the underlying heterogeneity, and constraining these parameter distributions to agree with experimentally measured species abundance values is an effective approach to infer parameters [5]. For such constrained inference problems, the maximum entropy (MaxEnt) approach is effective in providing low bias estimates [6].

Species abundances, however, can be tracked experimentally for only a few steps of signaling cascades in a single cell [7]. Hence, a simplified representation of the signaling network is inevitable to make parameter inference tractable. However, lumping several signaling interactions into a single step, as MERIDIAN [5] does, is an over-simplistic representation that can mask stochasticity and signaling heterogeneity crucial for accurate parameter inference. Moreover, existing MaxEnt-based inference approaches can only constrain one species' abundance at a time, and are characterized by very high inference times even just for constraining one species. Hence, the broader future scope of this project is to apply a more biologically realistic motif to simplify the signaling network representation and employ a faster

parameter inference strategy, both while leveraging the strength of the MaxEnt approach in low-bias parameter inference. Figure 1 depicts this workflow where the signaling network (left-most) is represented as a birth-death process (second from left): the birth of X_0 modeled as a Poisson process, death of X_n modeled as another Poisson process, and the time between the two modeled as a gamma distribution. The parameters of these distributions can then be inferred using MaxEnt estimators or flow networks.

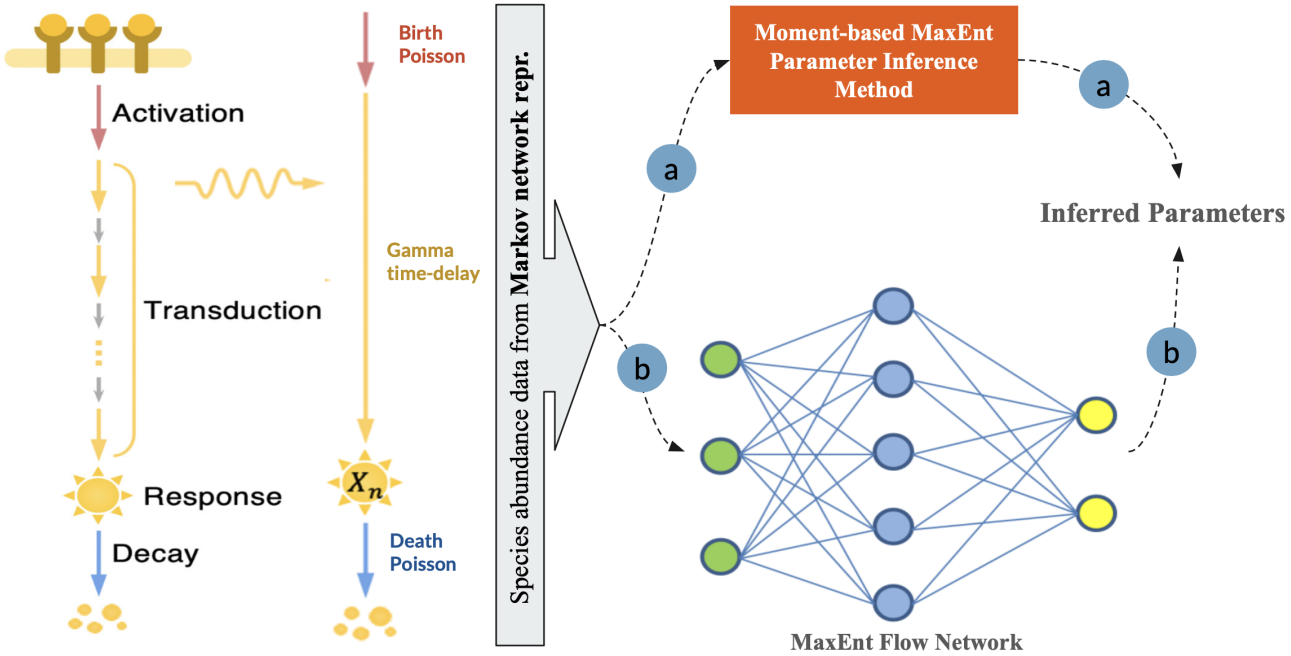
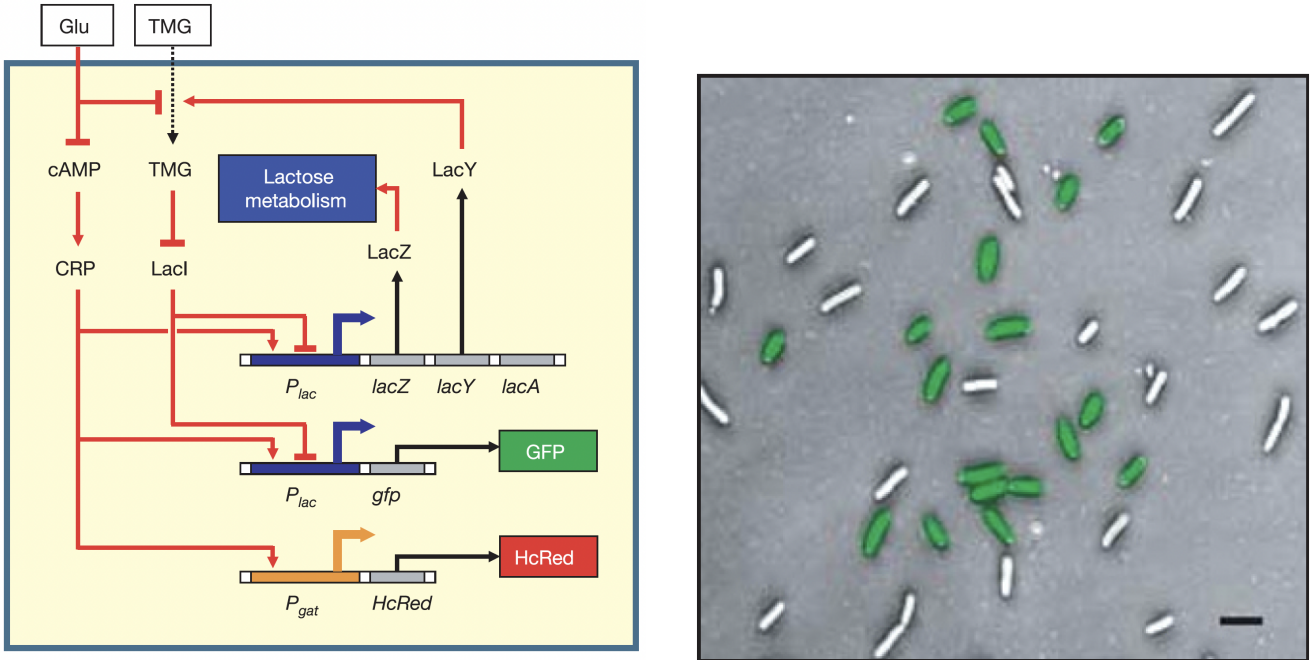


Figure 1: Broader scope of structured network simplification and MaxEnt-based parameter inference.

This project presents a first-level proof of concept. Herein, I first analyzed a sample network – the *lac operon* network – for its dynamics using previously described [8] ordinary differential equations (ODE). I then generated data for two species' concentrations by simulating the network and finally used this data to perform MaxEnt-based parameter inference.

2 Methods

To analyze and simulate the *lac operon* network, I used the lactose utilization network in *Escherichia coli (E. coli)* depicted in Figure 2a. This network is known to exhibit bistable behavior at specific parameter settings and concentrations. Figure 2b shows a population of *E. Coli* showing a bimodal distribution; the green individuals are in a state of higher stable *lacY* state, fluoresced by the green fluorescent protein (GFP) while the white-contrasted ones are at the lower stable state. Hence, the network formed a good candidate for testing parameter inference in the context of response heterogeneity.



(a) (b)

Figure 2: Schematic of the lactose utilization network in *E. Coli* and visual observation of bistability in the dynamics of the network across an *E. coli* population. *Figures adapted from [8].*

2.1 Lactose utilization network as a dynamical system

The network represented in Figure 2a can be represented by a simplified dynamical model describing the levels of *lacY*, intracellular TMG (TMGi), and *LacI* [8]. The active fraction of *LacI*, R, is a decreasing sigmoidal function of the TMG concentration x, with half-saturation concentration x_0, and Hill coefficient n as shown in Equation 1. $n \approx 2$ has been shown as a good approximation [4].

$$\frac{R}{R_T} = \frac{1}{1 + (x/x_0)^n} \tag{1}$$

where x is the intracellular TMG concentration, R_T is the total concentration of *LacI* tetramers, and R is the concentration of active *LacI*. The interaction of a single active *LacI* tetramer with multiple operator sites on the lac promoter generates a DNA loop that blocks transcription. The rate of generation of *LacY* is therefore a decreasing hyperbolic function of R, with maximal value α, half-saturation concentration R_0, and minimal value α/ρ achieved at $R = R_T$. The repression factor $\rho = 1 + \frac{R_T}{R_0}$ describes how tightly *LacI* can regulate lac expression. *LacY* is depleted in a first-order reaction with a time constant τ_y. This is summarized in Equation 2, where y is the concentration of *LacY*.

$$\tau_y \frac{dy}{dt} = \alpha \frac{1}{1 + R/R_0} - y \tag{2}$$

Finally, TMG enters the cell at a rate proportional to y and is similarly depleted in a first-order reaction with time constant τ_x. This is shown in Equation 3. The

parameter β measures the TMG uptake rate per *LacY* molecule. Once inside the cell, TMG can inactivate *LacI*, completing the positive feedback loop.

$$\tau_x \frac{dx}{dt} = \beta y - x \tag{3}$$

2.2 Simulating the network to generate data

This *LacY*-mediated positive feedback loop is responsible for the characteristic bistability pictured in Figure 2b. I used the experimentally obtained [8] parameters as initial values. To generate data for trajectory inference, I simulated the ODE system numerically over time starting from different initial concentrations to collect species concentration trajectories of TMG and *LacY* until they reached the steady state. This typically occurred within 50 time steps. The data generated serves as the ideal trajectory that we want to infer. To generate a synthetic dataset that mimics experimental measures of observable species abundances, I sampled parameters, simulated the ODEs, and added random normal noise to these observations to incorporate errors in experimental measurement. I generated these datasets for a population of $n = 1000$ *E. coli* for each parameter setting.

2.3 Inferring species abundance trajectories using MaxEnt

To infer species trajectories from simulated abundance data, I used a MaxEnt flow network approach. Given the experimental abundance measures as constraints, the flow neural network learns the maximum-likelihood parametric family within which the optimal trajectory parameters lie [9]; this massively compresses the exploration space. The flow network then samples distributions from this compressed space to find optimal parameters. Furthermore, the network can also perform fast simultaneous sampling of MaxEnt parameter- and Lagrange multiplier-distributions from this space [10], offering a significant computation speed-up over traditional MaxEnt inference approaches.

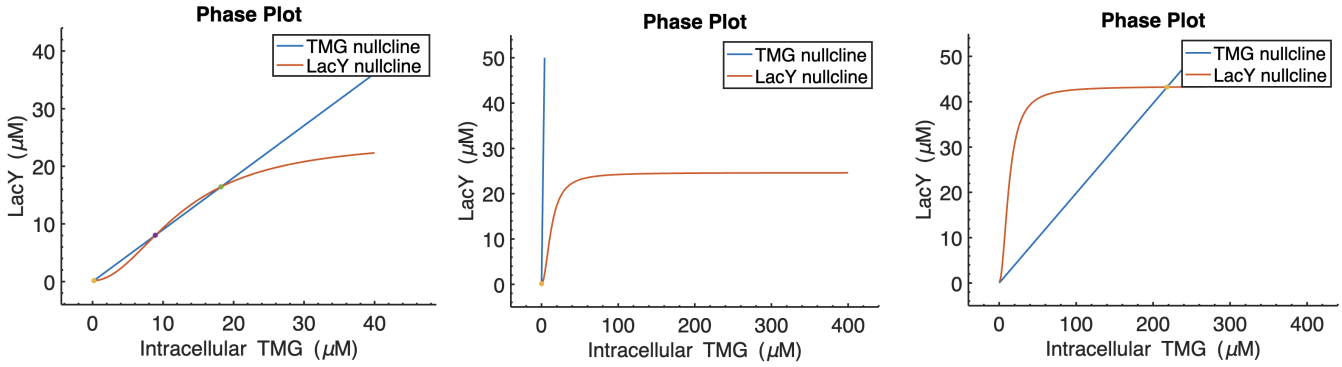
3 Results

This section first presents the findings from analyzing the dynamical model of the *lactose utilization* network and follows it up with the results from trajectory inference using MaxEnt.

3.1 Analytical solutions of steady states show three regions of stability

By setting rate equations to zero and solving the ODE system analytically, I obtained the nullclines of the system and plotted them on the *LacY*-TMG plane to obtain a phase plot. These analytical solutions are presented in Figure 3. With the

initial parameter setting, the system shows three stable points; two stable ones at $(0.18, 0.14)$ and $(21.60, 17.04)$, and one unstable one at $(7.49, 5.91)$. Such a state would explain the observed bistability in the *E. coli* population (see Figure 2b); the *E. coli* bacteria resting at the unstable state could switch to either the upper or the lower fixed point upon perturbation. Consequently, the population becomes bimodal with about one-half exhibiting GFP ("on" state) and the rest not exhibiting fluorescence ("off" state).



(a) Bistable region (b) Uninduced stable region (c) Induced stable region

Figure 3: Analytically obtained phase plots of the lactose utilization network.

To perturb the network and find regions where the system is not bistable, I modified extracellular concentrations of glucose and TMG following previously established results [8, 11]. One such region occurs when TMG concentration is very low, regardless of the glucose concentration (see Figure 3b). The system has only one stable fixed point that occurs at $(0.01, 0.15)$. On the other extreme, the system has exactly one stable system with a fixed point that occurs at $(218.03, 43.22)$ when the TMG concentration is high. For this region, however, the minimum TMG concentration threshold for monostability increases with glucose concentration. These monostable regions are termed uninduced and induced monostable regions, respectively (see below). Hence, when the system's conditions and parameters are perturbed, the system shifts between three different regions: uninduced monostable, bistable, and induced bistable. Further, as the system moves through the bistable region with changing extracellular concentrations, the system exhibits hysteresis; the final stable state of a bacterium in this region when stimulated by a change in extracellular concentration depends on its initial state, inducing a memory-like effect. These regions are analyzed further in a subsequent section.

3.2 Extracellular concentrations can determine the stability region

Even without perturbing the model parameters, the system can be placed in one of the three stability regions by simply controlling extracellular concentrations of glucose and TMG. I conducted a grid-search experiment to identify specific extracellular concentrations and their corresponding stability region. For each concentration set-

ting, I solved the ODE system analytically and registered the number of real points at which the *LacY* and TMG nullclines intersected. The region map so obtained is shown in Figure 4 where yellow denotes regions of bistability (three real solutions to the nullclines) and blue marks regions of monostability.

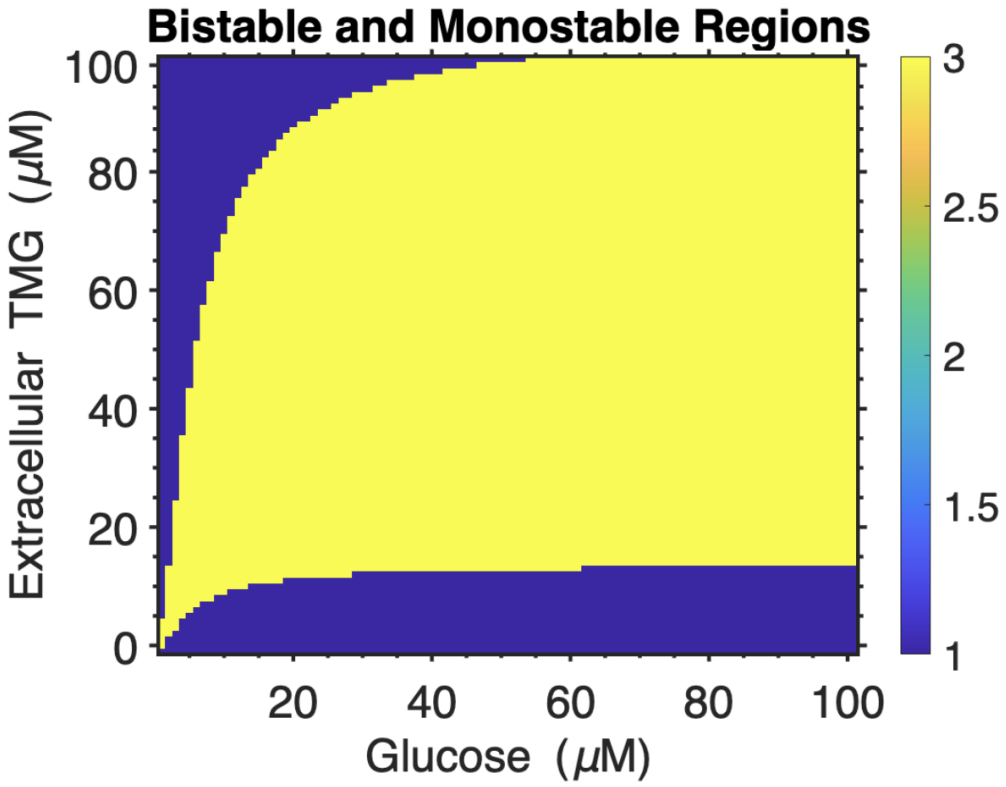


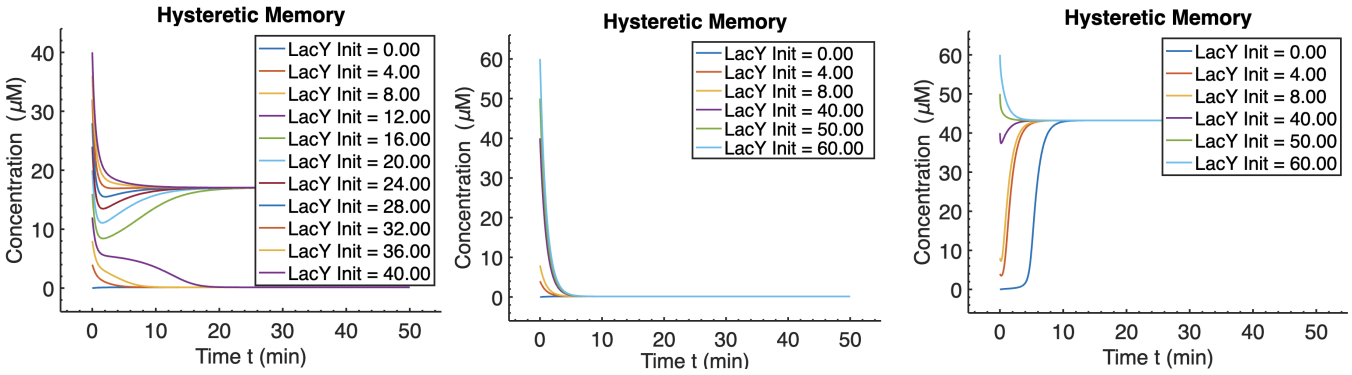
Figure 4: Stability regions as a phase plot on the extracellular glucose-TMG plane. Colorbar indicates the number of real solutions of the system of *LacY* and TMG nullclines.

The bistable region separates the lower uninduced monostable region from the higher induced one. Interestingly, the maximum TMG concentration for uninduced monostability does not depend on the glucose concentration, while they are positively correlated for induced monostability. These results are comparable to Figure 2c of Ozbudak et al. [8], although the exploration shown here is for a smaller range of extracellular concentrations since grid search does not scale well computationally.

3.3 Hysteretic memory affects the final cell state

Since the system is hysteretic in the bistable region, the history of the bacterium influences the side of the unstable fixed point on which it ends up in response to stimulation. Essentially, a bacterium that is in the "off" state moves along the region lower *LacY* nullcline in the hysteretic region; hence, it would require a much higher stimulation to switch "on". Similarly, a previously induced cell approaches along the higher *LacY* nullcline and requires very low concentrations to switch "off".

To validate these observations on the dynamical model, I perturbed extracellular concentrations to place the system in each stability region (see Figure 4) and then



(a) Bistable region (b) Uninduced stable region (c) Induced stable region

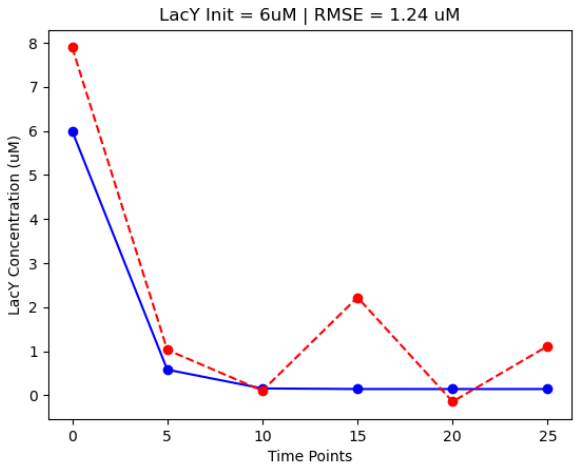
Figure 5: Analytically obtained phase plots of the lactose utilization network.

simulated it with a range of initial $LacY$ concentrations. A large initial concentration indicates that the bacterium starts from the "on" state and *vice-versa*. The results presented in Figure 5 agree with experimental observations [8, 11]. In the bistable region, systems starting at a higher $LacY$ concentration settle at the higher stable steady state ($LacY \approx 17\mu M$), while the lower concentrations end up at the lower state ($LacY \approx 0.14\mu M$). This bifurcation occurs somewhere between $LacY = 16\mu M$ and $LacY = 12\mu M$. In contrast, when the system is in one of the monostable regions, the concentrations equilibrate to the same state regardless of the initial concentration; uninduced to a low concentration and induced to a higher one (the exact values depend on the extracellular concentrations).

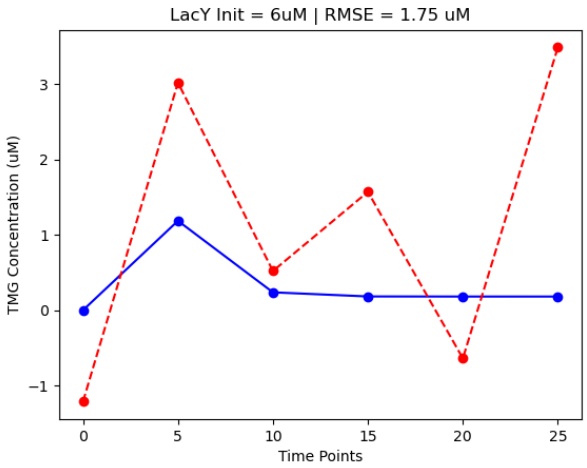
3.4 MaxEnt can estimate species abundance trajectories

I generated two synthetic datasets (see Methods) with the system placed in the bistable state by controlling extracellular concentrations; one dataset was drawn from a species that was initially in the "off" state ($LacY = 6\mu M$) and the other with the species in the "on" state ($LacY = 36\mu M$). Each dataset recorded the species concentrations over time of $LacY$ and intracellular TMG. I then retained the abundance measures at six evenly-spaced time points, picking the simulation time point closest to each desired time point, to both keep the complexity of the network inference low and because the ODE simulation does not yield regularly-spaced trajectory points making some form of binning necessary.

The flow network weights were adjusted by iterating through the entire dataset of $n = 1000$ *E. coli* samples over 10 epochs, each time constraining corresponding time points from the reference to match with the distribution-sampled. Over iterations, the flow network maximizes the underlying parameter distribution from where the trajectory points are drawn such that the constraints are satisfied and the distribution has the maximum possible entropy given the constraints. Ensuring maximum entropy helps eliminate biased inference. The results from the "off" and "on" species are presented in Figures 6 and7, respectively.

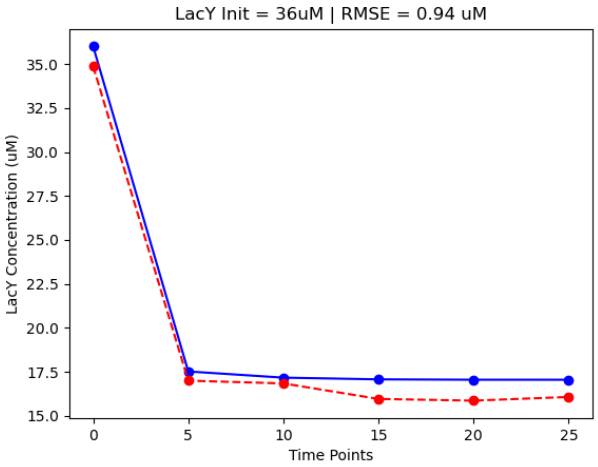


(a) $LacY$ trajectory

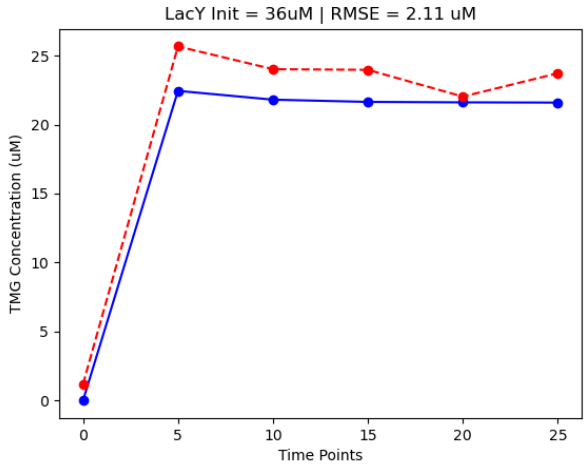


(b) TMG trajectory

Figure 6: MaxEnt inferred (red) vs. actual (blue) trajectory of species at initial $LacY = 6\mu M$.



(a) $LacY$ trajectory



(b) TMG trajectory

Figure 7: MaxEnt inferred (red) vs. actual (blue) trajectory of species at initial $LacY = 36\mu M$.

The predicted and reference values match quite well; the MSE errors are under 25% and are much lower for the initially "on" case. Increasing the number of time points could be a straightforward way to better estimate the trajectory.

4 Conclusions and Next Steps

This project explored the *lactose utilization* network of *E. coli* and analyzed it as a dynamical system. The network's bistability was characterized in terms of its participating extracellular and intracellular molecular species toward explaining heterogeneous bacterial population states. It was used as a model network to experiment with a basic MaxEnt-based species abundance trajectory estimation in a network exhibiting a simple, binary heterogeneous response.

Using data from rapidly evolving high-throughput single-cell technologies to

quantitatively represent biological networks and predict their response heterogeneity has far-reaching applications in precision medicine and therapeutics. Toward that end, these reported analyses and results are a first step toward a broader goal of simplifying biological networks more realistically as birth-death processes rather than merely clumping multiple reaction steps together while leveraging maximum entropy for low-bias parameter estimation. Together, these steps will improve parameter inference in the context of explaining and predicting heterogeneity. Further, the use of flow networks can significantly speed up MaxEnt inference by compressing the parameter space to be explored and in addition, they can perhaps provide valuable insight into parameter distribution through interpretation of their latent space patterns that are not easy to discover analytically.

Code availability

The code for all the experiments performed in the project can be accessed from `https://github.com/karthik-d/maxent-parameter-inference`. This project is primarily based on ideas presented in Dixit et al. [5] and Ozbudak et al. [8].

References

[1] Dae Wook Kim, Hyukpyo Hong, and Jae Kyoung Kim. Systematic inference identifies a major source of heterogeneity in cell signaling dynamics: The rate-limiting step number. *Science advances*, 8(11):eabl4598, 2022.

[2] Bruce R Levin and Daniel E Rozen. Non-inherited antibiotic resistance. *Nature Reviews Microbiology*, 4(7):556–562, 2006.

[3] Piyush B Gupta, Ievgenia Pastushenko, Adam Skibinski, Cedric Blanpain, and Charlotte Kuperwasser. Phenotypic plasticity: driver of cancer initiation, progression, and therapy resistance. *Cell stem cell*, 24(1):65–78, 2019.

[4] Gad Yagil and Ezra Yagil. On the relation between effector concentration and the rate of induced enzyme synthesis. *Biophysical journal*, 11(1):11–27, 1971.

[5] Purushottam D Dixit, Eugenia Lyashenko, Mario Niepel, and Dennis Vitkup. Maximum entropy framework for predictive inference of cell population heterogeneity and responses in signaling networks. *Cell systems*, 10(2):204–212, 2020.

[6] Wouter Boomsma, Jesper Ferkinghoff-Borg, and Kresten Lindorff-Larsen. Combining experiments and simulations using the maximum entropy principle. *PLoS computational biology*, 10(2):e1003406, 2014.

[7] Alejandro Sarrion-Perdigones, Lyra Chang, Yezabel Gonzalez, Tatiana Gallego-Flores, Damian W Young, and Koen JT Venken. Examining multiple cellular pathways at once using multiplex hextuple luciferase assaying. *Nature communications*, 10(1):5710, 2019.

[8] Ertugrul M Ozbudak, Mukund Thattai, Han N Lim, Boris I Shraiman, and Alexander Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427(6976):737–740, 2004.

[9] Gabriel Loaiza-Ganem, Yuanjun Gao, and John P Cunningham. Maximum entropy flow networks. *arXiv preprint arXiv:1701.03504*, 2017.

[10] Sean R Bittner, Agostina Palmigiano, Alex T Piet, Chunyu A Duan, Carlos D Brody, Kenneth D Miller, and John P Cunningham. Interrogating theoretical models of neural computation with deep inference. biorxiv. 2019.

[11] Martin Ackermann. A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews Microbiology*, 13(8):497–508, 2015.
