Maximum Entropy Framework For Inference Of Cell Population Heterogeneity In Signaling Networks

Purushottam D. Dixit*[1], Eugenia Lyashenko*[1], Mario Niepel[2], and Dennis Vitkup[1,3,4]

Correspondence should be addressed to P. D. (dixitpd@gmail.com) or D.V. (dv2121@columbia.edu).

[1]Department of Systems Biology, Columbia University
[2]Department of Systems Biology, Harvard Medical School
[3]Department of Biomedical Informatics, Columbia University
[4]Center for Computational Biology and Bioinformatics, Columbia University

Predictive models of signaling networks are essential tools for understanding cell population heterogeneity and designing rational interventions in disease. However, using network models to predict signaling dynamics heterogeneity is often challenging due to the extensive variability of network parameters across cell populations. Here, we describe a **M**aximum **E**ntropy-based f**R**amework for **I**nference of heterogeneity in **D**ynamics of s**I**g**A**ling **N**etworks (MERIDIAN). MERIDIAN allows us to estimate the joint probability distribution over network parameters that is consistent with experimentally observed cell-to-cell variability in abundances of network species. We apply the developed approach to investigate the heterogeneity in the signaling network activated by the epidermal growth factor (EGF) and leading to phosphorylation of protein kinase B (Akt). Using the inferred parameter distribution, we also predict heterogeneity of phosphorylated Akt levels and the distribution of EGF receptor abundance hours after EGF stimulation. We discuss how MERIDIAN can be generalized and applied to problems beyond modeling of heterogeneous signaling dynamics.

Introduction

Signaling cascades in genetically identical cells often respond in a heterogeneous manner to extracellular stimuli [1]. This heterogeneity arises largely due to cell-to-cell variability in network parameters such as reaction rates and species abundances [2-5]. The variability in parameters can have important functional consequences. For example, in fractional killing of cancer cells treated with chemotherapeutic drugs [3, 5] and in differential infection rates of viruses in mammalian cells [6]. Therefore, the knowledge of the distribution over network parameters is essential to understanding phenotypic heterogeneity in cell populations.

Several experimental techniques such as flow cytometry [7], immunofluorescence [7], and live cell assays [8] have been developed to investigate the variability of network species abundances. However, it is often difficult to estimate the distributions over network parameters from these experimental measurements. The reasons for this challenge are primarily twofold. First, parameters such as protein abundances and biochemical rates vary substantially across cells in a population [1]. For example, previous studies have reported the coefficients of variation of protein abundances in the range 0.1-0.6 [9]. Similarly, reaction rate constants also vary between cells by several orders of magnitude [6]. As a result of the extensive parameter variability, multiple cell types characterized by distinct high probability regions in the parameter space may coexist in a population [10]. Second, the multivariate parameter distribution can potentially have a complex shape. For example, as is often the case in single cell data, some parameters may exhibit multimodality while others may be unimodally distributed [11, 12].

Over the last decade, several computational methods have been developed to estimate the joint distribution of network parameters consistent with experimentally measured cell-to-cell variability in network species [13-18]. However, these methods rely on assuming a specific shape of the parameter distribution. For example, Hasenauer et al. [13, 14] (see also [16] and [15])

approximate the parameter distribution as a linear combination of pre-defined distribution functions. Similarly, Zechner et al. [17, 18] assume that the parameters are distributed according to a log-normal or a gamma distribution. Consequently, it may be difficult for these previously developed approaches to infer a complex multivariate parameter distribution of an unknown shape.

Building on our previous work [19, 20], we developed **MERIDIAN**: a **M**aximum **E**ntropy-based f**R**amework for **I**nference of heterogeneity in **D**ynamics of s**I**gn**A**ling **N**etworks. Notably, MERIDIAN infers the functional form of the parameter distribution that is consistent with data-derived constraints. The maximum entropy principle was first introduced more than a century ago in statistical physics [21]. Among all candidate distributions that agree with imposed constraints, the maximum entropy approach selects the one with the least amount of over-fitting. Maximum entropy-based approaches have been successfully applied previously to a variety of biological problems, including protein structure prediction [22], protein sequence evolution [23], neuron firing dynamics [24], molecular simulations [25-27], and dynamics of biochemical reaction networks [28].

Following a description of the key ideas behind MERIDIAN, we illustrate its performance using synthetic data on a simplified model of growth factor signaling. Next, we use the framework to study the heterogeneity in the signaling network leading to phosphorylation of protein kinase B (Akt). Epidermal growth factor (EGF)-induced Akt phosphorylation governs key intracellular processes [29] including, metabolism, apoptosis, and cell cycle entry. Due to its central role in mammalian signaling, aberrations in the Akt pathway are implicated in multiple diseases [29, 30]. We apply MERIDIAN to infer the distribution over network parameters using experimentally measured phosphorylated Akt (pAkt) and cell surface EGFR (sEGFR) levels in MCF10A cells following EGF stimulation [31]. We then demonstrate that the obtained parameter distribution allows us to accurately predict the heterogeneity in single cell pAkt levels at late time points, as well as heterogeneity in cell surface EGFRs in response to EGF stimulation.

Finally, we discuss generalizations of the framework to study problems beyond modeling heterogeneity in signaling networks.

Results

Outline of MERIDIAN

We consider a signaling network comprising *N* chemical or biological species whose intracellular abundances we denote by $\bar{X} = \{X_1, X_2, \dots, X_N\}$. We assume that the molecular interactions among the species are described by a system of ordinary differential equations

$$\frac{d}{dt}\bar{X}(t,\bar{\theta}) = f(\bar{X},\bar{\theta}) \quad (1)$$

where $f(\bar{X},\bar{\theta})$ is a function of species abundances, $\bar{X}$. Here $\bar{\theta} = \{\theta_1, \theta_2, \dots\}$ is a vector of parameters that describe the dynamics of the signaling network. We denote by $x_a(t,\bar{\theta})$ the solution of equations 1 for species "*a*" with parameters $\bar{\theta}$.

Our computational approach is illustrated in Figure 1. We use experimentally measured the cell-to-cell variability of protein species "a" at multiple time points (illustrated by histograms in Figure 1) to constrain the parameter distribution $P(\bar{\theta})$. Specifically, we quantify the measured variability by estimating fraction ϕ_{ik} of cells that belong to a particular 'bin' in the abundance distribution histogram at each time point; in our notation, *i* indicates time and *k* indicates the bin number. Every distinct dynamical trajectory $x_a(t,\bar{\theta})$ (illustrated by red and blue curves in Figure 1) generated by specific parameter values $\bar{\theta}$ passes through a unique set of bins (red curve through red bins and blue curve through blue bins in Figure 1) at multiple time points. Using MERIDIAN, we find a corresponding probability distribution for network parameters $P(\bar{\theta})$ (see inset, Figure 1) such that the distribution over trajectories $P[x_a(t,\bar{\theta})]$ is consistent with all experimentally measured bin fractions. Notably, because we simultaneously identify all temporal bins crossed by a specific trajectory $x_a(t,\bar{\theta})$, our approach

naturally accounts for statistical correlations between data at different time points and thus avoids over-constraining the probability distribution. Below we present our development to derive the functional form of $P(\bar{\theta})$ consistent with experimental data.

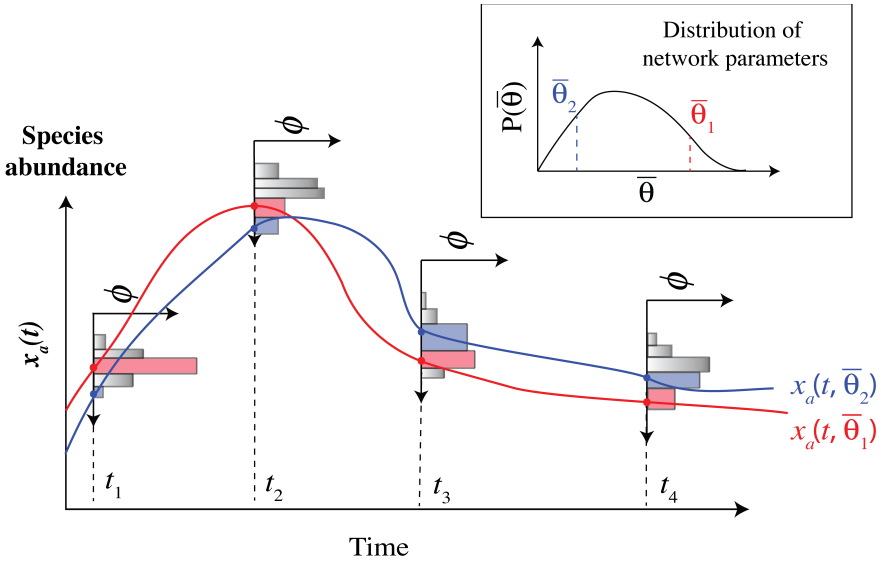

Figure 1. **Illustration of the MERIDIAN inference approach**. Cell-to-cell variability in a protein *"a"* is measured at 4 time points t_1, t_2, t_3, and t_4. From the data, we determine the fraction ϕ_{ik} of cells that populate the k^{th} abundance bin at the i^{th} time point by binning the cell-to-cell variability data in B_i bins. The histograms show the bin fractions ϕ_{ik} at multiple time points. We find $P(\bar{\theta})$ with the maximum entropy while requiring that the distribution $P[x_a(t,\bar{\theta})]$ of simulated trajectories of $x_a(t,\bar{\theta})$ simultaneously reproduce all bin fractions ϕ_{ik}**s**.

Derivation of $P(\bar{\theta})$ using MERIDIAN

For simplicity, we first consider the case when the distribution of cell-to-cell variability in one species x_a is available only at one time point t (for example, $t = t_1$ in Figure 1). We denote by $\bar{\phi} = \{\phi_1, \phi_2, \ldots, \phi_B\}$ the fraction of cells whose experimental measurement of x_a lies in individual bins. Given a particular parameter distribution $P(\bar{\theta})$, the *predicted* fractions $\bar{\psi} = \{\psi_1, \psi_2, \ldots, \psi_B\}$ can be obtained as follows. Using Markov chain Monte Carlo (MCMC), we generate multiple parameter sets from $P(\bar{\theta})$. For each generated parameter set $\bar{\theta}$, we numerically solve equations 1 and find $x_a(t_1, \bar{\theta})$, i.e. the predicted value of the abundance at time t_1. Using the samples from the ensemble of trajectories, we

estimate ψ_k as the fraction of sampled trajectories where $x_a(t_1, \bar{\theta})$ belonged to the k^{th} bin. Mathematically,

$$\psi_k = \int I_k \left(x_a(t_1, \bar{\theta}) \right) P(\bar{\theta})\, d\bar{\theta} \qquad [24]$$

where $I_k(x)$ is an indicator function; i.e. $I_k(x)$ is equal to one if x lies in the k^{th} bin and zero otherwise.

The central idea behind MERIDIAN is to find the maximum entropy distribution $P(\bar{\theta})$ over parameters such that all predicted fractions ψ_k agree with those estimated from experiments ϕ_k. Formally, we seek $P(\bar{\theta})$ with the maximum entropy,

$$S = -\int P(\bar{\theta}) \log \frac{P(\bar{\theta})}{q(\bar{\theta})}\, d\bar{\theta} \qquad (3)$$

subject to constraints $\phi_k = \psi_k$ and normalization, $\int P(\bar{\theta})\, d\bar{\theta} = 1$ [32]. Here, $q(\bar{\theta})$ plays a similar role to the prior distribution in Bayesian approaches [33]. In this work, we choose $q(\bar{\theta})$ to be a uniform distribution within literature-derived ranges of parameters, but other choices can be implemented as well.

To impose aforementioned constraints and perform the entropy maximization, we use Lagrange multipliers. To that end, we use an unconstrained optimization function

$$L = S + \beta\left(\int P(\bar{\theta})\, d\bar{\theta} - 1\right) - \sum_{k=1}^{B} \lambda_k \left(\int I_k \left(x_a(t_1, \bar{\theta}) \right) P(\bar{\theta})\, d\bar{\theta} - \phi_k \right) \qquad (4)$$

where β is the Lagrange multiplier associated with normalization and λ_k are the Lagrange multipliers associated with fixing the predicted fractions ψ_k to their experimentally measured values ϕ_k in all bins. Differentiating equation 4 with respect to $P(\bar{\theta})$ and setting the derivative to zero, we obtain

$$P(\bar{\theta}) = \frac{1}{\Omega} q(\bar{\theta}) \exp\left(-\sum_{k=1}^{B} \lambda_k I_k \left(x_a(t_1, \bar{\theta}) \right) \right) \qquad (5)$$

where $\Omega = \int q(\bar{\theta}) \exp\left(-\sum_{k=1}^{B} \lambda_k I_k \left(x_a(t_1, \bar{\theta}) \right) \right) d\bar{\theta}$ is the partition function that normalizes the probability distribution.

The generalization of equation 5 when abundances of multiple species are measured at several time points is straightforward. Specifically, if we constrain distributions of cell-to-cell abundance variability of n species ($x_1, x_2, x_3, \ldots x_n$) measured at times t_{ij} (where i denotes species and j denotes time points) the maximum entropy probability distribution over parameters $P(\bar{\theta})$ can be written as

$$P(\bar{\theta}) = \frac{1}{\Omega} q(\bar{\theta}) \exp\left(-\sum_{i=1}^{n}\sum_{j=1}^{T_i}\sum_{k=1}^{B}\lambda_{k_{ij}} I_{k_{ij}}\left(x_i\left(t_{ij},\bar{\theta}\right)\right)\right) \quad (6)$$

where $x_i\left(t_{ij},\bar{\theta}\right)$ is the solution of equation 1 for the i^{th} species measured at the j^{th} time point. In equation 6, the experimentally measured distributions of abundances for species i at time j are represented by B_{ij} bins and the Lagrange multipliers corresponding to the k_{ij}^{th} bin are $\lambda_{k_{ij}}$ respectively.

The Lagrange multipliers λs in equations 6 need to be numerically optimized such that the predicted bin fractions are consistent with the experimentally estimated ones. Notably, the search for the Lagrange multipliers is a convex optimization problem and we solve it using an iterative algorithm proposed in [34] (see Figure 2). Briefly, we start from a randomly chosen point in the space of Lagrange multipliers. In the n^{th} iteration of the optimization algorithm, using the current vector of the Lagrange multipliers $\bar{\lambda}_n$, we estimate the predicted bin fractions $\bar{\psi}_n$ using Markov chain Monte Carlo (MCMC). Next, we estimated the error vector $\bar{\Delta}_{n,i} = \left(\bar{\psi}_{n,i} - \bar{\phi}_i\right)/\bar{\phi}_i$ for the n^{th} iteration. We then update the multipliers for the n+1st iteration as $\bar{\lambda}_{n+1} = \bar{\lambda}_n + \alpha_n \bar{\Delta}_n$ (see Figure 2) where the positive "learning rate" α_n is chosen to minimize the error $|\bar{\Delta}_{n+1}|$ (see SI section III for details).

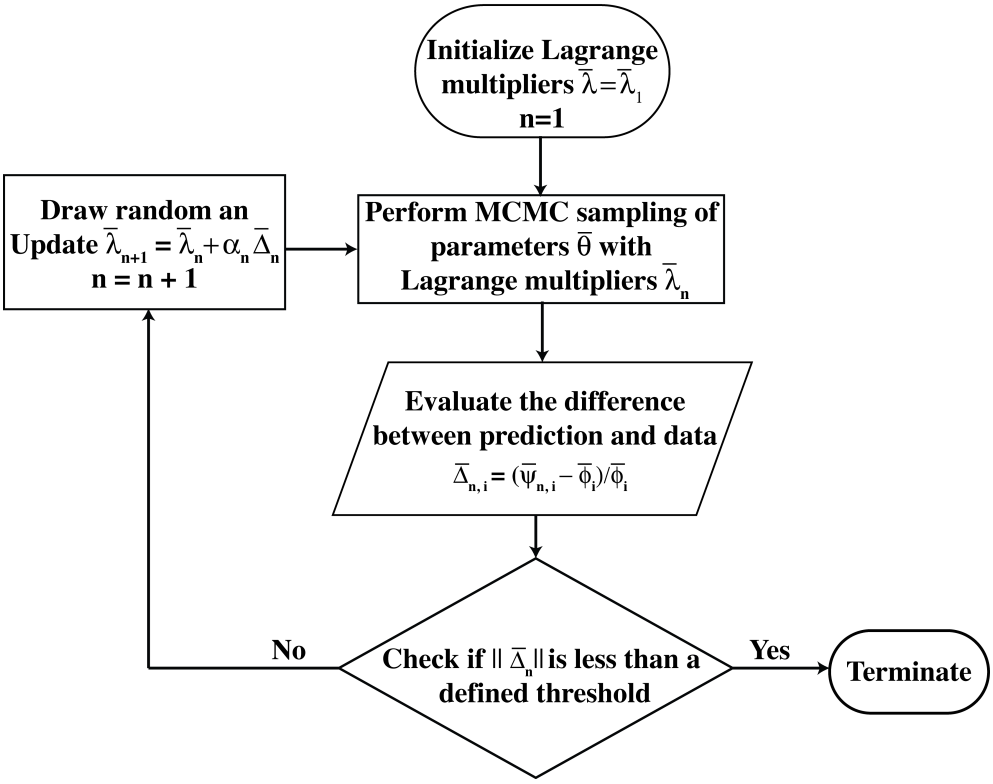

Figure 2. The workflow to numerically determine the values of Lagrange multipliers. In each iteration we evaluated the error vector $\bar{\Delta}_n$ between predicted bin fractions $\bar{\psi}_n$ and the experimentally measured bin fractions $\bar{\phi}$ using MCMC. We proposed a new set of Lagrange multipliers based on the error vector. We repeated until the error reached below a predefined accuracy cutoff.

MERIDIAN performance on synthetic data

First, we used synthetic data to illustrate the utility of MERIDIAN and compared its performance with a previously developed discretized Bayesian (DB) approach by Hasenauer *et al.* [14]. In DB, a crucial step is to approximate the joint parameter distribution as a linear combination

$$P(\bar{\theta}) = \sum \gamma_k \Phi_k(\bar{\theta}) \tag{7}$$

where $\Phi_k(\bar{\theta})$ are a set of predefined distribution functions and $\gamma_k \geq 0$ are the corresponding weights. DB then discretizes the multidimensional parameter space using a Cartesian grid and assumes that $\Phi_k(\bar{\theta})$ are multivariate Gaussian

distributions centered at grid points [14]. The weights γ_k are then determined by maximizing the likelihood of the data given the model. Notably, several other methods (see [13, 15, 16]) also rely on approximating the parameter distribution as a linear combination over known functions. Thus, a comparison with DB is sufficient to underscore the advantages of MERIDIAN over previous approaches.

To compare MERIDIAN with DB we used a simplified growth factor network model (SI Figure 1). Specifically, the considered network included three chemical species: the ligand *L*, ligand-free inactive receptors *R*, and ligand-bound active receptors *P*. Ligand binding to receptors leads to their activation. Inactive receptors are constantly delivered to the cell surface, and active and inactive receptors are removed from the cell surface at different rates (see SI section I). In the model, the dynamics of the species was described by five parameters.

Using the aforementioned model, we generated synthetic single cell data by varying two key parameters: (1) rate of degradation of active receptors k^*_{deg} and (2) the steady state number of receptors on the cell surface R_T. Both parameters were sampled from gamma distributions (see SI section I). Next, using the joint parameter distribution, we generated distributions of activated receptor levels corresponding to four different experimental conditions (ligand L = 2 ng/ml and L = 10 ng/ml, and time t = 10 minutes and steady state, see Figure 3, dashed red lines). We then used the synthetic data to infer the joint distribution $P\left(k^*_{deg}, R_T\right)$ using MERIDIAN and DB. Similar to [14], we used in the DB approach a 10x10 Cartesian grid and placed multivariate Gaussian distributions of specified variance on grid points. The weight coefficients γ_k of the 100 Gaussian distributions were determined by minimizing the L_2 error between the simulated experimental distribution and the predicted distribution of receptor abundance (see SI section I for details).

Notably, both MERIDIAN (red diamonds in SI Figure 2) and DB (blue circles in SI Figure 2) were able to accurately fit the synthetic data (dashed black lines in SI Figure 2). However, the two approaches predicted substantially different parameter distributions. Specifically, while MERIDIAN approach was

able to accurately capture both the distributions $P\left(k_{deg}^{*}\right)$ and $P(R_T)$ (red lines in Figure 3a and b), the distributions inferred using DB were substantially different from the underlying distributions (blue lines in Figure 3a and b). Finally, we note that, the accuracy of the DB approach can potentially be increased with implementation of denser grids in the inference procedure. However, the finer discretization will lead to unfeasibly large number of required grid evaluations in a multidimensional parameter space. This will likely prevent accurate inferences for even a moderately sized signaling network with more than a dozen parameters (see SI section I).

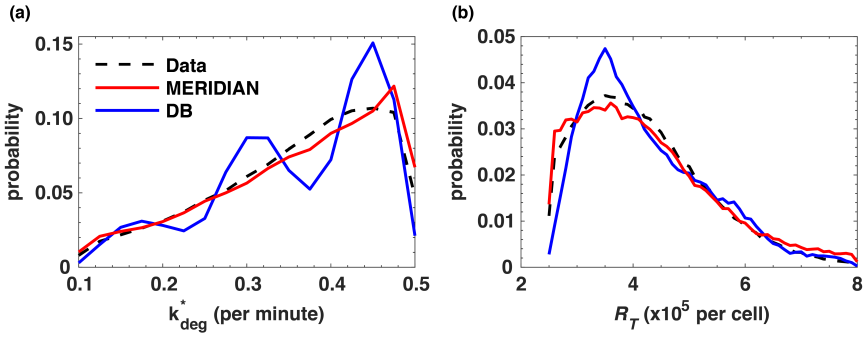

Figure 3. A comparison between the inferred parameter distributions and the data. We show the comparison between the true parameter distributions ($P\left(k_{deg}^{*}\right)$ in panel a and $P(R_T)$ in panel b) (dashed black lines) and the corresponding MERIDIAN-inferred distribution (red lines) and the DB inferred distribution (blue lines).

Using MERIDIAN to study EGFR/Akt signaling

Computational model of the EGFR/Akt signaling network

Signal transduction in the EGFR/Akt network is illustrated in Figure 4. Following stimulation of cells with EGF, the ligand binds to cell surface EGFRs. Ligand-bound receptors dimerize with other ligand-bound as well as ligand-free receptors. EGFR dimers then phosphorylate each other and phosphorylated receptors (active receptors, pEGFRs) on the cell surface lead to downstream phosphorylation of Akt (pAkt). Both active and unphosphorylated (inactive) receptors are internalized with different rates from the cell surface through receptor endocytosis. After addition of EGF in the extracellular medium, pAkt

levels increase transiently within minutes and then, as a results of receptor endocytosis and action of phosphatases, both pAkt and surface EGFR (sEGFR) levels decrease within hours after EGF stimulation [35].

To explore the cell-to-cell variability in this pathway, we used a dynamical model of EGF/EGFR dependent Akt phosphorylation based on Chen *et al*. [35]. The model includes reactions describing EGF binding to EGFR and subsequent dimerization, phosphorylation, dephosphorylation, internalization, and degradation of the receptors. To keep the model relatively small, we simplified pEGFR-dependent phosphorylation of Akt by assuming a single step activation of pAkt by pEGFR with an effective rate constant (see SI section II for the model details).

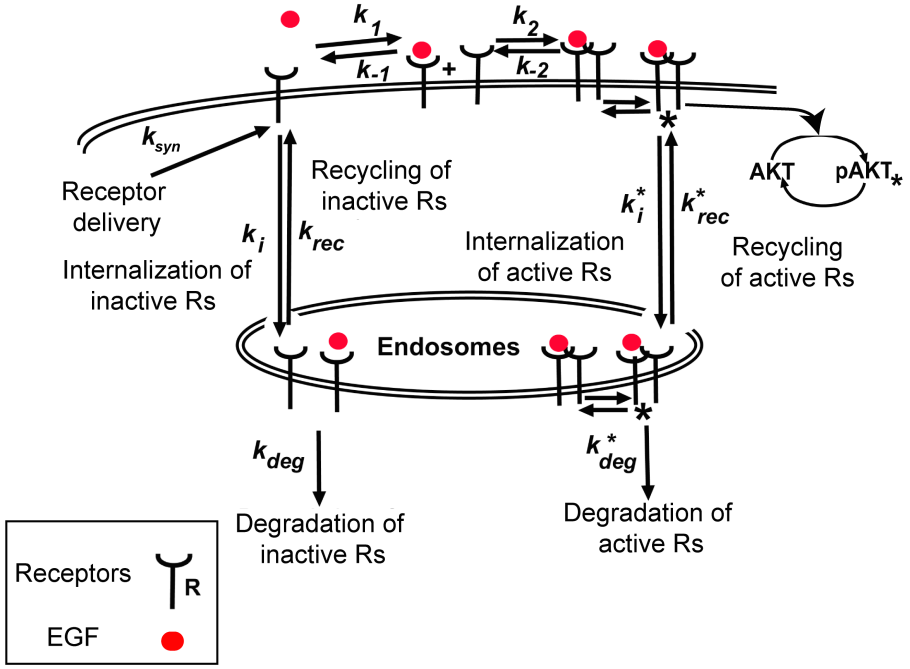

Figure 4. A schematic of the EGF/EGFR pathway leading to phosphorylation of Akt. Extracellular EGF binds to cell surface EGFRs leading to their dimerization. Dimerized EGFRs are autophosphorylated and in turn lead to phosphorylation of Akt. Receptors are also removed from cell surface through internalization into endosomes. See SI section II for details of the model.

Numerical inference of the parameter distribution

To derive the network parameter distribution consistent with experimental data using MERIDIAN, we used experimentally measured cell-to-cell variability in pAkt and sEGFR levels at early times after EGF stimulation. Specifically, we used measured pAkt levels after stimulation with five different EGF doses (0.1, 0.316, 3.16, 10, and 100 ng/ml) at 4 early time points (5, 15, 30, and 45 minutes) [36]. Additionally, we used sEGFR levels without EGF stimulation and after 3 hours of EGF stimulation at 0 ng/ml and 1 ng/ml (see SI section II and III for details). We used 11 bins to represent each experimentally measured distribution; the bin sizes and locations were chosen to cover the entire range of observed variability. We numerically determined the Lagrange multipliers corresponding to the bind fractions using the procedure described above (see equation 6 and Figure 2).

Notably, the best-fit Lagrange multipliers accurately reproduced the experimentally measured bin fractions (Pearson's r^2 = 0.84, $p < 10^{-10}$, median relative error ~ 12%). Furthermore, fitted bin fractions obtained in two independent calculations showed excellent agreement with each other as expected for a convex optimization problem (Pearson's r^2 = 0.98, $p < 10^{-10}$, see SI Figure 3). In Figure 5, we show the temporal profile of measured cell-to-cell variability in pAkt levels (colored circles) at EGF stimulation of 10 ng/ml and the corresponding fits (dashed black lines) based on the inferred parameter distribution.

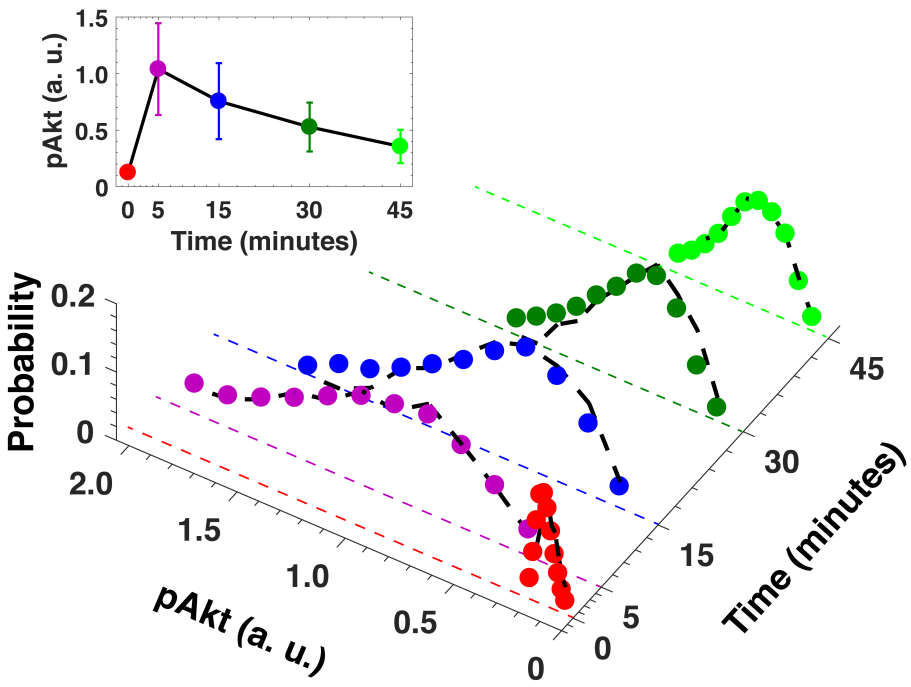

Figure 5. Experimental estimated cell-to-cell variability in pAkt levels used to infer the parameter distribution. We show distribution of pAkt levels at 0, 5, 15, 30, and 45 minutes after exposure to 10 ng/ml EGF. The colored circles represent the experimentally measured pAkt distributions used in the inference of the parameter distribution. The black dashed lines represent fitted distributions. The inset shows the experimentally measured population average pAkt levels at multiple time points. Error bars in the inset represent population standard deviations.

Prediction of single cell dynamics

Akt is a key hub of mammalian cell signaling [29]. Naturally, sustained activity of phosphorylated Akt (pAkt) is implicated in diverse human diseases, such as psychiatric disorders [37] and cancer [38, 39]. Using the developed approach, we investigated whether we could predict pAkt levels hours after EGF stimulation using the parameter distribution inferred from pAkt variability at early times after EGF stimulation. To that end, we numerically sampled multiple parameter sets using the inferred parameter distribution and predicted pAkt levels at late time across a range of EGF stimulation levels corresponding to each parameter set. We compared predicted and experimentally observed distribution of pAkt levels across cells at late times (180 minutes) after sustained EGF stimulation (Figure 6a, b). Our simulations correctly predicted that a

significant fraction of cells have high pAkt levels hours after stimulation; the predicted and observed coefficient of variation (CV) of the pAkt distributions in cells stimulated with 10 ng/ml EGF for 180 minutes were in good agreement, 0.41 and 0.37 respectively. Notably, the inferred parameter distribution accurately captures the population mean and variability (Figure 6c) in pAkt levels at late times across four orders of magnitude of EGF concentrations used to stimulate cells.

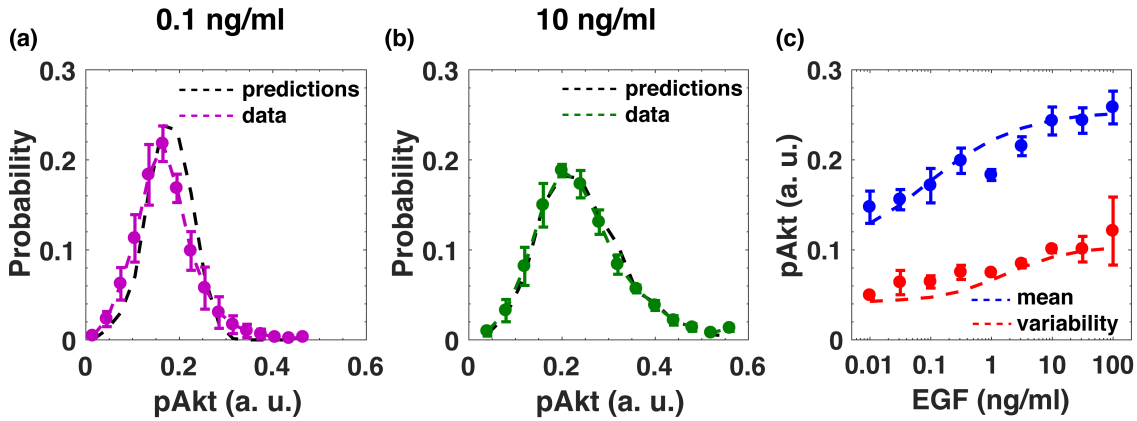

Figure 6. Prediction pAkt levels at late times. (a and b) Measured distributions (colored circles and lines) and the corresponding predictions (dashed black lines) of cell-to-cell variability in pAkt levels at 180 minutes after stimulation with 0.1 ng/ml and 10 ng/ml EGF respectively. (c) Measured mean pAkt levels (blue circles) and measured standard deviation in pAkt levels (red circles) at 180 minutes after sustained stimulation with EGF (x-axis) and the corresponding predictions (dashed blue line and dashed red line respectively). The error bars represent standard deviation estimated from four replicates.

MERIDIAN allowed us to investigate the biochemical parameters that significantly correlate with high pAkt levels at steady state. Interestingly, across all simulated trajectories, the levels of cell surface EGFR showed the highest correlation with pAkt levels among all receptor-related parameters (Pearson r = 0.44, $p < 10^{-10}$, EGF stimulation 10 ng/ml). This suggests that cells with high EGFR levels likely predominantly contribute to the sub-population of cells with high steady state pAkt activity. This demonstrates how MERIDIAN can be used to gain mechanistic insight into heterogeneity in signaling dynamics based on single cell data.

We next investigated whether MERIDIAN could predict the heterogeneity in EGFR levels after prolonged stimulation with EGF. To that end, we compared the predicted and the experimentally measured the steady state EGFR levels across EGF stimulation doses. Similar to pAkt, the simulations accurately captured both the population mean and variability of the EGFR receptor levels across multiple doses of EGF stimulations (Figure 7c). The simulations and experiments demonstrate that in agreement with model prediction that even hours after the growth factor stimulation there is a significant fraction of cells with relatively high levels of EGFR (Figure 7a, b).

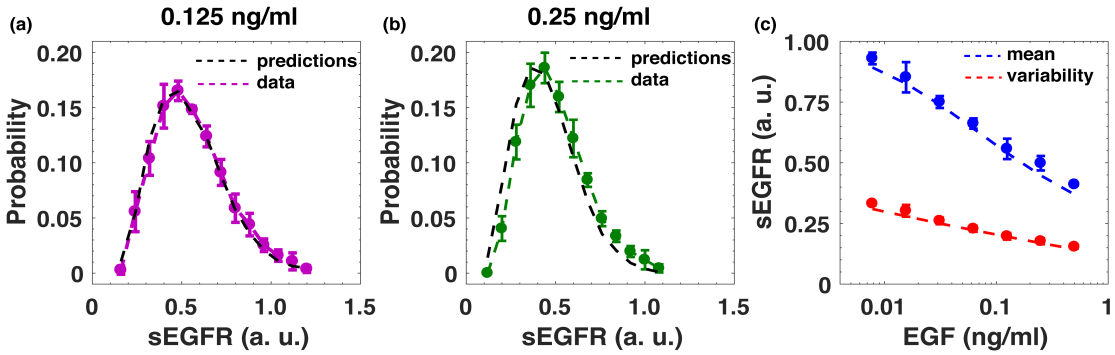

Figure 7. Prediction of sEGFR levels at late times. (a) and (b) Measured distributions (colored circles lines) and the corresponding predictions (dashed black lines) of cell-to-cell variability in sEGFR levels at 180 minutes after stimulation with 0.125 ng/ml and 0.25 ng/ml EGF respectively. (c) Measured mean sEGFR levels (blue circles) and measured standard deviation in sEGFR levels (red circles) at 180 minutes after sustained stimulation with EGF (x-axis) and the corresponding predictions (dashed blue line and dashed red line respectively). The error bars represent standard deviation estimated from three replicates.

Possible extensions of the MERIDIAN framework

Importantly, a straightforward extension makes it possible to use the MERIDIAN framework for networks when the time evolution of species abundances is intrinsically stochastic. For example, transcriptional networks and prokaryotic signaling networks with relatively small species abundances [1]. To that end, we can modify the definition of the predicted fraction $\psi_k = \int P\big(x(t,\bar{\theta}) =$

$x|\bar{\theta})I_k(x)dx$ of a chemical species *x*, where $P(x(t,\bar{\theta}) = x|\bar{\theta})$ is the distribution of *x* values at time *t* with parameters $\bar{\theta}$. The distributions can be obtained numerically using Gillespie's stochastic algorithm [40] and its fast approximations [41] or approximated using moment closure techniques [42].

MERIDIAN can also be used to infer parameter distributions when, instead of the entire abundance distributions only a few moments of the distribution are available, such as average protein abundances measured using quantitative western blots or mass spectrometry [43]. For example, we consider the case where the population mean *m* and the variance *v* of one species *x* are measured at a fixed time point *t*. Instead of constraining fractions ψ_k that represent cell-to-cell variability in different bins of the relevant abundance distribution, we can constrain the population mean $\mu_1 = \int x(t,\bar{\theta})P(\bar{\theta})\,d\bar{\theta}$ and the second moment $\mu_2 = \int x(t,\bar{\theta})^2 P(\bar{\theta})\,d\bar{\theta}$ to their experimentally measured values *m* and $v+m^2$ respectively. Entropy maximization can then be carried out with these constraints. In this case, we have

$$P(\bar{\theta}) = \frac{1}{\Omega} q(\bar{\theta}) \exp(-\lambda_1 x(t,\bar{\theta}) - \lambda_2 x(t,\bar{\theta})^2) \tag{8}$$

Lastly, we can use MERIDIAN to infer parameters from experiments where dynamics of species abundances within single cells are measured using live cell imaging [8]. For example, consider that the time evolution of a species $x(t)$ is measured in n_c cells from time *t=0* to *t* = *T*. We can discretize the n_c continuous time observations into *K* discrete times $\{t_1, t_2, \ldots, t_K\}$. At each time point t_i, we can then divide the range of observed abundances in B_i bins. Then, each individual dynamical trajectory $x(t)$ can be characterized by a vector of discrete indices $x(t) \sim \{B_{1a_1}, B_{1a_1}, \ldots, B_{Ka_K}\}$ where B_{ia_i} is the index of the abundance distribution bin through which the trajectory $x(t)$ passed at time point t_i. Given a sufficiently large number of trajectories, we then can constrain the fraction of trajectories that populate a given sequence of bins to infer the parameter distribution.

Discussion

In this work we presented MERIDIAN: a maximum entropy framework to infer the joint distribution $P(\bar{\theta})$ over signaling network parameters based on experimentally measured cell-to-cell variability in species abundances. We demonstrated that the obtained parameter distribution allow accurate prediction of the time evolution of cell-to-cell variability in species abundances. The inferred distribution takes into account both parameter non-identifiability and their cell-to-cell variability. The contribution due to parameter non-identifiability can be further minimized by explicitly incorporating known constraints in parameter values using prior [33]. Notably, the maximum entropy framework naturally avoids over-constraining the parameter distribution as redundant constraints lead to redundant Lagrange multipliers [44].

Recent developments in cytometry [45] and single cell RNA sequencing [46] make it possible to simultaneously quantify multiple species abundances in single cells. Elegant statistical approaches have been developed to reconstruct trajectories of intracellular species dynamics consistent with time-stamped single cell abundance data [33, 47, 48]. Complementary to these statistical methods, our approach (1) allows us to infer the distribution over parameters that describe mechanistic interactions in the signaling network and moreover (2) the inferred parameter distribution can be used to predict the ensemble of single cell trajectories for time intervals and experimental conditions beyond the measured abundance distributions.

In this work, we applied the developed framework to signaling network data. However, it can also be used in other diverse research contexts. For example, the framework can be applied to computationally reconstruct the distribution of longitudinal behaviors from cross-sectional time-snapshot data in fields such public health, economics, and ecology or to estimate parameter distributions from lower dimensional statistics [49].

Acknowledgment and Funding

We would like to thank Jan Hasenauer and Caroline Loos for suggestions about the manuscript.

References

1. Raj, A. and A. van Oudenaarden, *Nature, nurture, or chance: stochastic gene expression and its consequences.* Cell, 2008. **135**(2): p. 216-26.
2. Llamosi, A., et al., *What Population Reveals about Individual Cell Identity: Single-Cell Parameter Estimation of Models of Gene Expression in Yeast.* PLoS Comput Biol, 2016. **12**(2): p. e1004706.
3. Spencer, S.L., et al., *Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis.* Nature, 2009. **459**(7245): p. 428-32.
4. Meyer, R., et al., *Heterogeneous kinetics of AKT signaling in individual cells are accounted for by variable protein concentration.* Front Physiol, 2012. **3**: p. 451.
5. Albeck, J.G., et al., *Quantitative analysis of pathways controlling extrinsic apoptosis in single cells.* Mol Cell, 2008. **30**(1): p. 11-25.
6. Snijder, B., et al., *Population context determines cell-to-cell variability in endocytosis and virus infection.* Nature, 2009. **461**(7263): p. 520-3.
7. Wu, M. and A.K. Singh, *Single-cell protein analysis.* Current Opinion in Biotechnology, 2012. **23**(1): p. 83-88.
8. Meyer, R., et al., *Heterogeneous kinetics of AKT signaling in individual cells are accounted for by variable protein concentration.* Frontiers in Physiology, 2012. **3**: p. 451.
9. Niepel, M., S.L. Spencer, and P.K. Sorger, *Non-genetic cell-to-cell variability and the consequences for pharmacology.* Current Opinion in Chemical Biology, 2009. **13**(5-6): p. 556-61.
10. Arendt, D., et al., *The origin and evolution of cell types.* Nat Rev Genet, 2016. **17**(12): p. 744-757.
11. Kaern, M., et al., *Stochasticity in gene expression: from theories to phenotypes.* Nat Rev Genet, 2005. **6**(6): p. 451-64.
12. Shalek, A.K., et al., *Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells.* Nature, 2013. **498**(7453): p. 236-40.
13. Hasenauer, J., et al., *ODE constrained mixture modelling: a method for unraveling subpopulation structures and dynamics.* PLoS Comput Biol, 2014. **10**(7): p. e1003686.
14. Hasenauer, J., et al., *Identification of models of heterogeneous cell populations from population snapshot data.* BMC Bioinformatics, 2011. **12**: p. 125.
15. Loos, C., et al., *A Hierarchical, Data-Driven Approach to Modeling Single-Cell Populations Predicts Latent Causes of Cell-To-Cell Variability.* Cell Syst, 2018. **6**(5): p. 593-603 e13.

16. Waldherr S., H.J., Allgöwer F., *Estimation of biochemical network parameter distributions in cell populations.* IFAC Proceedings Volumes, 2009. **42**(10): p. 1265-1270.
17. Zechner, C., et al., *Moment-based inference predicts bimodality in transient gene expression.* Proc Natl Acad Sci U S A, 2012. **109**(21): p. 8340-5.
18. Zechner, C., et al., *Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings.* Nat Methods, 2014. **11**(2): p. 197-202.
19. Dixit, P.D., *Quantifying extrinsic noise in gene expression using the maximum entropy framework.* Biophysical Journal, 2013. **104**(12): p. 2743-50.
20. Eydgahi, H., et al., *Properties of cell death models calibrated and compared using Bayesian approaches.* Mol Syst Biol, 2013. **9**: p. 644.
21. Dixit, P.D., et al., *Perspective: Maximum caliber is a general variational principle for dynamical systems.* J Chem Phys, 2018. **148**(1): p. 010901.
22. Weigt, M., et al., *Identification of direct residue contacts in protein-protein interaction by message passing.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(1): p. 67-72.
23. Mora, T., et al., *Maximum entropy models for antibody diversity.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(12): p. 5405-10.
24. Schneidman, E., et al., *Weak pairwise correlations imply strongly correlated network states in a neural population.* Nature, 2006. **440**(7087): p. 1007-12.
25. Dixit, P.D., et al., *Inferring Transition Rates of Networks from Populations in Continuous-Time Markov Processes.* Journal of Chemical Theory and Computation, 2015. **11**(11): p. 5464-72.
26. Dixit, P.D. and K.A. Dill, *Inferring Microscopic Kinetic Rates from Stationary State Distributions.* Journal of Chemical Theory and Computation, 2014. **10**(8): p. 3002-3005.
27. Tiwary, P. and B.J. Berne, *Spectral gap optimization of order parameters for sampling complex molecular systems.* Proceedings of the National Academy of Sciences of the United States of America, 2016. **113**(11): p. 2839-44.
28. Dixit, P.D., *Communication: Introducing prescribed biases in out-of-equilibrium Markov models.* J Chem Phys, 2018. **148**(9): p. 091101.
29. Manning, B.D. and A. Toker, *AKT/PKB Signaling: Navigating the Network.* Cell, 2017. **169**(3): p. 381-405.
30. Herbst, R.S., *Review of epidermal growth factor receptor biology.* Int J Radiat Oncol Biol Phys, 2004. **59**(2): p. 21-6.
31. Soule, H.D., et al., *Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10.* Cancer Res, 1990. **50**(18): p. 6075-86.
32. Presse, S., et al., *Principles of maximum entropy and maximum caliber in statistical physics.* Reviews of Modern Physics, 2013. **85**(3): p. 1115-1141.
33. Mukherjee, S., et al., *Connecting the dots across time: reconstruction of single-cell signalling trajectories using time-stamped data.* Royal Society Open Science, 2017. **4**: p. 170811.
34. Tkacik, G., et al., *Ising models for networks of real neurons.* arXiv, 2006. **q-bio/0611072**.

35. Chen, W.W., et al., *Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data.* Mol Syst Biol, 2009. **5**: p. 239.
36. Lyashenko, E., et al., *Receptor-based mechanism of relative sensing in mammalian signaling networks.* bioRxiv, 2017: p. 10.1101/158774.
37. Gilman, S.R., et al., *Diverse types of genetic variation converge on functional gene networks involved in schizophrenia.* Nat Neurosci, 2012. **15**(12): p. 1723-8.
38. Vivanco, I. and C.L. Sawyers, *The phosphatidylinositol 3-Kinase AKT pathway in human cancer.* Nat Rev Cancer, 2002. **2**(7): p. 489-501.
39. Nicholson, K.M. and N.G. Anderson, *The protein kinase B/Akt signalling pathway in human malignancy.* Cell Signal, 2002. **14**(5): p. 381-95.
40. Gillespie, D.T., *Stochastic simulation of chemical kinetics.* Annu Rev Phys Chem, 2007. **58**: p. 35-55.
41. Cao, Z. and R. Grima, *Linear mapping approximation of gene regulatory networks with stochastic dynamics.* Nat Commun, 2018. **9**(1): p. 3305.
42. Gillespie, C.S., *Moment-closure approximations for mass-action models.* IET Syst Biol, 2009. **3**(1): p. 52-8.
43. Shi, T., et al., *Conservation of protein abundance patterns reveals the regulatory architecture of the EGFR-MAPK pathway.* Sci Signal, 2016. **9**(436): p. rs6.
44. Dixit, P.D. and K.A. Dill, *Caliber Corrected Markov Modeling (C2M2): Correcting Equilibrium Markov Models.* J Chem Theory Comput, 2018. **14**(2): p. 1111-1119.
45. Chattopadhyay, P.K., et al., *Single-cell technologies for monitoring immune systems.* Nat Immunol, 2014. **15**(2): p. 128-35.
46. Saliba, A.E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic Acids Res, 2014. **42**(14): p. 8845-60.
47. Gut, G., et al., *Trajectories of cell-cycle progression from fixed cell populations.* Nat Methods, 2015. **12**(10): p. 951-4.
48. Mukherjee, S., et al., *In silico modeling identifies CD45 as a regulator of IL-2 synergy in the NKG2D-mediated activation of immature human NK cells.* Sci Signal, 2017. **10**(485).
49. Das, J., S. Mukherjee, and S.E. Hodge, *Maximum Entropy Estimation of Probability Distribution of Variables in Higher Dimensions from Lower Dimensional Data.* Entropy (Basel), 2015. **17**(7): p. 4986-4999.