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]

* Contributed equally
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

The dynamics of intracellular signaling networks can vary substantially among cells in a population. Predictive models of signaling networks are key to our understanding of cellular function and in design of rational interventions in disease. However, using network models to predict heterogeneity in signaling network dynamics is challenging due to cell to cell variability of network parameters, such as reaction rates and species abundances, and parameter *non-identifiability*. In this work, we present an inference framework based on the principle of maximum entropy (ME) to estimate the joint probability distribution over network parameters that is consistent with experimentally measured cell to cell variability in abundances of network species. We apply the framework to study the heterogeneity in the signaling network activated by the epidermal growth factor (EGF) resulting in phosphorylation of protein kinase B (Akt); a central signaling hub in mammalian cells. Notably, the inferred parameter distribution allows us to accurately predict population heterogeneity in phosphorylated Akt (pAkt) levels at early and late times after EGF stimulation as well as the heterogeneity in the levels of cell surface EGF receptors (sEGFRs) after prolonged stimulation with EGF. We discuss how the developed framework can be generalized and applied to problems beyond signaling networks.

Introduction

Signaling networks within individual cells in a cell population often respond to extracellular stimuli in a heterogeneous manner even if the cells are isogenic[1]. Cell to cell variability in signaling network parameters is directly responsible for this observed heterogeneity. Notably, network heterogeneity has important functional consequences, for example, in aiding stochastic transitions in development[2] or in fractional killing of cancer cells treated with chemotherapeutic drugs[3].

Several experimental techniques such as flow cytometry[4], immunofluorescence[4], and live cell assays[5] have been developed to probe cell to cell variability in abundances of species participating in signaling networks. However, it is often challenging to computationally estimate the distribution over network parameters that is consistent with these experimental measurements. The reasons are twofold. First, parameters such as protein abundances and various biochemical rates, can themselves vary substantially from cell to cell in a population[1]. For example, the coefficient of variation of protein abundances ranges between 0.1 to 0.6[6]. Similarly, reaction rate constants may vary between cells by a couple of orders of magnitude[7]. Second, many network parameters are *non-identifiable* given limited experimental measurements of a few species at a few time points[8-10] --- network dynamics are likely to be insensitive to coupled variations in related parameters, such as association rates and the corresponding dissociation rates.

Over the last decade, computational methods have been developed to estimate the joint distribution of network parameters consistent with experimentally measured variability in network species[11-13]. However, these methods rely on several simplifying assumptions. For example, the parameter distributions are restricted to a pre-defined functional family, such as the multivariate log-normal distribution[13]. Or, data collected at different time points and experimental conditions are assumed to be statistically independent of each other thus simplifying the likelihood of observing multiple experimental conditions as a product of the likelihoods of observing individual experiments[12]. However, while cell to cell variability estimated at different time points for different species is measured in independent experiments, the data are statistically correlated through the parameters of the underlying signaling network[13]. Consequently,

the assumption of probabilistic independence of individual experimental measurements is likely to substantially over-constrain the parameter distribution.

Here, building on our previous work[9, 14], we present a maximum entropy (ME) based framework to infer the joint distribution over network parameters. Notably, our approach circumvents the aforementioned simplifying assumptions. ME is a tool first introduced more than a century ago in statistical physics[15]. Among all candidate distributions that agree with imposed constraints, ME chooses the one with the least amount of bias. ME has been successfully used in a variety of biological problems including protein structure prediction[16], protein sequence evolution[17], collective firing of neurons[18], molecular dynamics simulations[19-21], and simulation of bimolecular reaction networks[14, 22, 23].

We apply the developed framework to study heterogeneity in the signaling network leading to phosphorylation of protein kinase B (Akt) induced by the epidermal growth factor (EGF)-dependent activation of its receptor (EGFR); a central mammalian signaling cascade implicated in diseases. Notably, EGF induced Akt phosphorylation governs key intracellular processes[24] including metabolism, apoptosis, and cell cycle entry. Concomitantly, aberrations in the pathway are implicated in many diseases[24, 25].

We infer the distribution over network parameters in a model of the EGF/EGFR/Akt signaling network using experimentally measured cell to cell variability in phosphorylated Akt (pAkt) levels and cell surface EGFR (sEGFR) levels in MCF 10A cells[26]. The parameter distribution allows us to accurately predict the cell to cell variability in pAkt levels at early and late time points as well as cell to cell variability in cell surface EGFRs at steady state in response to a constant stimulation with EGF. We also discuss how to generalize the framework to other types of biological systems and experimental data.

Results

Consider a signaling network comprising *N* chemical species whose intracellular abundances are denoted by $\bar{X} = \{X_1, X_2, \ldots, X_N\}$. We assume that the molecular interactions among the species within the network 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 abundances $\bar{X}$. Here, $\bar{\theta} = \{\theta_1, \theta_2, \ldots\}$ is a vector of parameters that describe the dynamics of the signaling network. $x_a(t,\bar{\theta})$ denotes the solution of Equation 1 for species *a* at time *t* when parameters are fixed at $\bar{\theta}$.

Our approach is schematically shown in Figure 1. Consider that we have experimentally measured the cell to cell variability in a protein species x_a in the signaling network at multiple time points. First, we quantify the measured cell to cell variability by estimating fraction ϕ_{ik} of cells that belong to a particular 'bin' in the histogram of abundances for each measurement (*i* for time, *k* for bin number) by dividing the observed abundances at time every t_i in B_i bins. As shown in Figure 1, every dynamical trajectory $x_a(t,\bar{\theta})$ (generated by parameters $\bar{\theta}$) passes through a unique set of bins corresponding to the different time points. Using the ME framework, we then find $P(\bar{\theta})$ such that the corresponding distribution over trajectories $P[x_a(t,\bar{\theta})]$ is consistent with all measured bin fractions. Notably, because we simultaneously identify all temporal bins crossed by any trajectory $x_a(t,\bar{\theta})$ the approach naturally accounts for statistical correlations in the data and avoids over-constraining the probability distribution.

Below we derive $P(\bar{\theta})$ for networks whose dynamics can be effectively modeled using ordinary differential equations. Later, we discuss how to treat networks that are inherently stochastic.

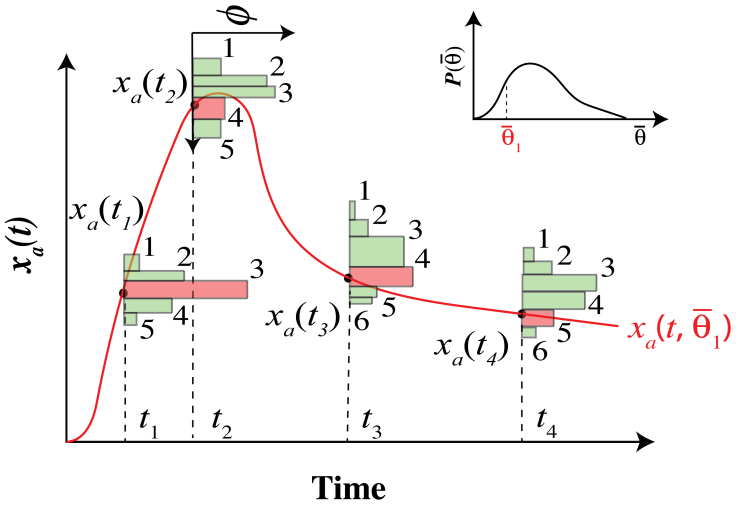

Figure 1. In an illustration of our approach, cell to cell variability in protein x is measured at 4 time points t_1, t_2, t_3, and t_4. From the experimental data, we determine the fraction ϕ_{ik} of cells that populated k^{th} abundance bin at the i^{th} time point by binning the cell to cell variability data in B_i bins. The horizontal 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 reproduces all ϕ_{ik}s.

Derivation of $P(\bar{\theta})$ consistent with experimental data

For simplicity, consider the case when the distribution of cell to cell variability in one species x is measured at a single fixed time t (for example, $t = t_1$ in Figure 1). We first estimate the experimentally determined fractions $\bar{\phi} = \{\phi_1, \phi_2, \ldots, \phi_B\}$ by dividing the range of observed abundances in B bins. Here, ϕ_k is the fraction of cells whose experimental measurement of x lies in the k^{th} bin.

Next, given a parameter distribution $P(\bar{\theta})$, the *predicted* fractions ψ_k can be obtained as follows. Using Markov chain Monte Carlo (MCMC), we generate multiple parameter sets $\bar{\theta}_1, \bar{\theta}_2, \bar{\theta}_3, \cdots$ from $P(\bar{\theta})$. For each parameter set $\bar{\theta}$, we solve Eq. (1) and find the value of $x_a(t,\bar{\theta})$. ψ_k is the fraction of parameter sample points where $x_a(t,\bar{\theta})$ belonged to the k^{th} bin. Mathematically,

$$\psi_k = \int I_k\left(x_a(t,\bar{\theta})\right) \cdot P(\bar{\theta}) d\bar{\theta}\,. \qquad (2)$$

In Equation 2, $I_k(x)$ is an indicator function; $I_k(x)$ is equal to one if x lies in the k^{th} bin and zero otherwise.

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

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

subject to constraints $\psi_k = \phi_k$ and normalization, $\int P(\bar{\theta}) d\bar{\theta} = 1$ [15]. Here, $q(\bar{\theta})$ is equivalent to the prior distribution in Bayesian approaches[27]. In this work, we choose $q(\bar{\theta})$ as a uniform distribution over a literature-derived range of parameters, but other choices are possible as well.

Using Lagrange multipliers to impose aforementioned constraints, we carry out the entropy maximization. To that end, we write an unconstrained optimization function

$$L = S + \beta\left(\int P(\bar{\theta}) d\bar{\theta} - 1\right) - \sum_{b=1}^{B} \lambda_b \left(\int I_b\left(x_a(t,\bar{\theta})\right) \cdot P(\bar{\theta}) d\bar{\theta} - \phi_b\right) \tag{4}$$

Here, β is the Lagrange multiplier associated with normalization and λ_b are the Lagrange multipliers associated with fixing the fraction ψ_b to their experimentally measured value ϕ_b in bins $b \in \{1, 2, \ldots, B\}$. Differentiating Equation 4 with respect to $P(\bar{\theta})$ and setting the derivative to zero, we have

$$P(\bar{\theta}) = \frac{1}{\Omega} q(\bar{\theta}) \exp\left(-\sum_{b=1}^{B} \lambda_b I_b\left(x_a(t,\bar{\theta})\right)\right) \tag{5}$$

Here, $\Omega = \int q(\bar{\theta}) \exp\left(-\sum_{b=1}^{B} \lambda_b I_b\left(x_a(t,\bar{\theta})\right)\right) d\bar{\theta}$ is the partition function that normalizes the probabilities.

The generalization of Equation 5 when abundances of multiple species are measured at multiple time points is straightforward. If we constrain distributions of cell to

cell variability of abundances of n species (x_1, x_2, x_3, … x_n) measured at times t_{ij} (i for species, j for time point) the probability distribution over parameters $P(\bar{\theta})$ is

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

Here, $x_i\left(t_{ij}, \bar{\theta}\right)$ is the solution of Equation 1 for the i^{th} species measured at the j^{th} time point. The experimentally measured distributions of cell to cell variability for species i at time j are split into B_{ij} bins. $\lambda_{b_{ij}}$ are the Lagrange multipliers corresponding to the b_{ij}^{th} bin where $b_{ij} = \{1, 2, 3, \ldots, B_{ij}\}$ and $I_{b_{ij}}(x)$ are the corresponding indicator functions.

The ME method derives the functional form of the maximum entropy distribution. The Lagrange multipliers λs need to be optimized numerically so that the predicted bin counts are equal to the experimentally estimated ones.

Below, we first discuss the computational model of the EGF/EGFR pathway and the experimental data used in this work. Next, we discuss the specific numerical procedure to estimate the Lagrange multipliers that were implemented in this work.

Computational model of the EGF/EGFR signaling network and experimental data

Briefly, the EGF/EGFR signaling network operates as follows. Upon introduction of EGF in the extracellular environment, EGF binds to cell surface EGF receptors (sEGFRs). Ligand-bound receptors dimerize with other ligand-free as well as ligand-bound receptors. Dimerized EGFRs phosphorylate each other. Phosphorylated receptors (pEGFRs) on the cell surface lead to downstream phosphorylation of Akt (pAkt). Both monomeric and dimeric receptors are internalized from the cell surface through receptor endocytosis. Upon addition of EGF in the extracellular medium, pAkt levels increase transiently within minutes, and as a result of receptor endocytosis, both pAkt and surface EGFR (sEGFR) levels reach steady state within hours of constant EGF stimulation[28]. See Figure 2 for a simplified schematic of the network.

We constructed a dynamical model of EGF/EGFR dependent Akt phosphorylation based on Chen et al. [28]. The model includes detailed description of EGF

binding to EGFR and subsequent dimerization, phosphorylation, dephosphorylation, internalization, and degradation of the receptors. 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 I for details of the model.

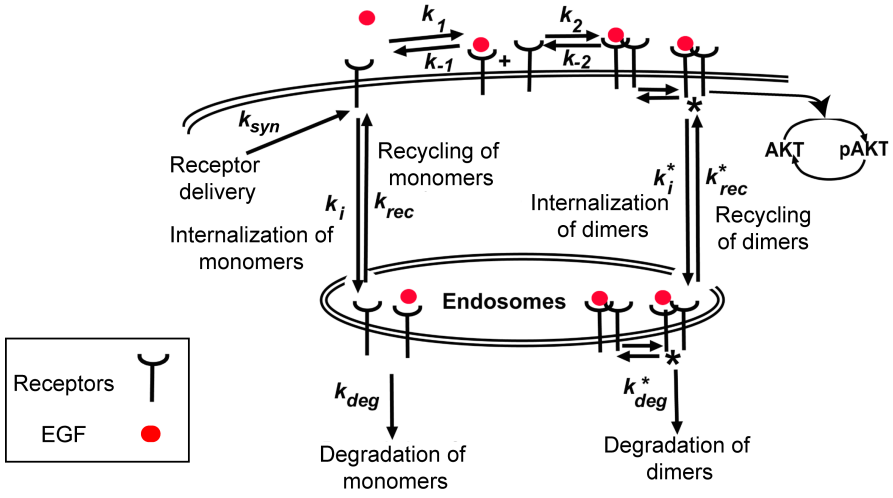

Figure 2. A schematic of the EGF/EGFR pathway leading to phosphorylation of Akt. Extracellular EGF binds to EGFR leading to its dimerization. Dimerized EGFRs are phosphorylated and in turn lead to phosphorylation of Akt. Receptors are also removed from cell surface through internalization into endosomes. See supplementary materials for details of the model.

We applied the developed framework to investigate the dynamical changes in cell to cell variability in pAkt levels in MCF10A cells[26] after stimulation with a constant dose of extracellular EGF. We measured the distribution of cell to cell variability in pAkt levels at 7 time points ranging between 5 to 180 minutes after EGF stimulation at 8 different doses of EGF (between 0.01 ng/ml to 100 ng/ml) covering the entire range of pAkt responses in 9 x 7 = 63 independent experiments using immunofluorescence. We also measured the cell to cell variability in pAkt levels in the absence of EGF stimulation. Finally, we measured the cell to cell variability in the abundance of cell surface EGFRs in the absence of EGF exposure and at 180 minutes after exposure to 9 different EGF doses (from 0.0078 ng/ml to 1 ng/ml along with a saturating dose of 100 ng/ml). We note that the distribution of cell to cell variability at each time point and at each EGF dose was measured in independent sets of cells. See SI section II for details of the experimental procedure.

Numerical inference of the joint parameter distribution $P(\bar{\theta})$ using experimental data

We inferred $P(\bar{\theta})$ given by Equation 6 by fitting 20 out of the 63 distributions of experimentally measured cell to cell variability in pAkt levels, specifically, the distribution of pAkt levels at 5, 15, 30, and 45 minutes after stimulation with 0.1, 0.31, 3.16, 10, and 100 ng/ml of EGF. In addition, we also used the measured distribution of pAkt and sEGFR levels without EGF stimulation and 2 distributions of sEGFR variability measured after 180 minutes of constant EGF exposure at 1 ng/ml and 100 ng/ml. The 20 pAkt distributions captured essential features of Akt phosphorylation dynamics such as the rapid increase in pAkt response within 5-10 minutes of EGF stimulation and the subsequent down-regulation due to receptor endocytosis.

In summary, we used 24 out of the 74 measured distributions of pAkt and sEGFR levels to infer the distribution of model parameters. We used 11 bins to represent each distribution. The bin sizes and locations were chosen to cover the entire range of observed variability while allowing reliable numerical estimates of $\bar{\phi}$. There were a total of 24 x 11 = 264 Lagrange multipliers that constrained experimentally estimated bin fractions $\bar{\phi}$.

Determination of values of the Lagrange multipliers in Equation 6 is a convex optimization problem[29] and we solved it by closely following an iterative algorithm proposed by Broderick et al[29]. We started from a randomly chosen point in the space of Lagrange multipliers. In the n^{th} iteration, using the Lagrange multipliers $\bar{\lambda}_n$, we estimated the predicted bin fractions $\bar{\psi}_n$ using Markov chain Monte Carlo (MCMC). Next, we estimated the error vector $\bar{\Delta}_n = \bar{\psi}_n - \bar{\phi}_n$ for the n^{th} iteration. For the n+1st iteration, we update the multipliers as $\bar{\lambda}_{n+1} = \bar{\lambda}_n + \alpha_n \bar{\Delta}_n$ (see Figure 3) where α_n is a real number chosen randomly.

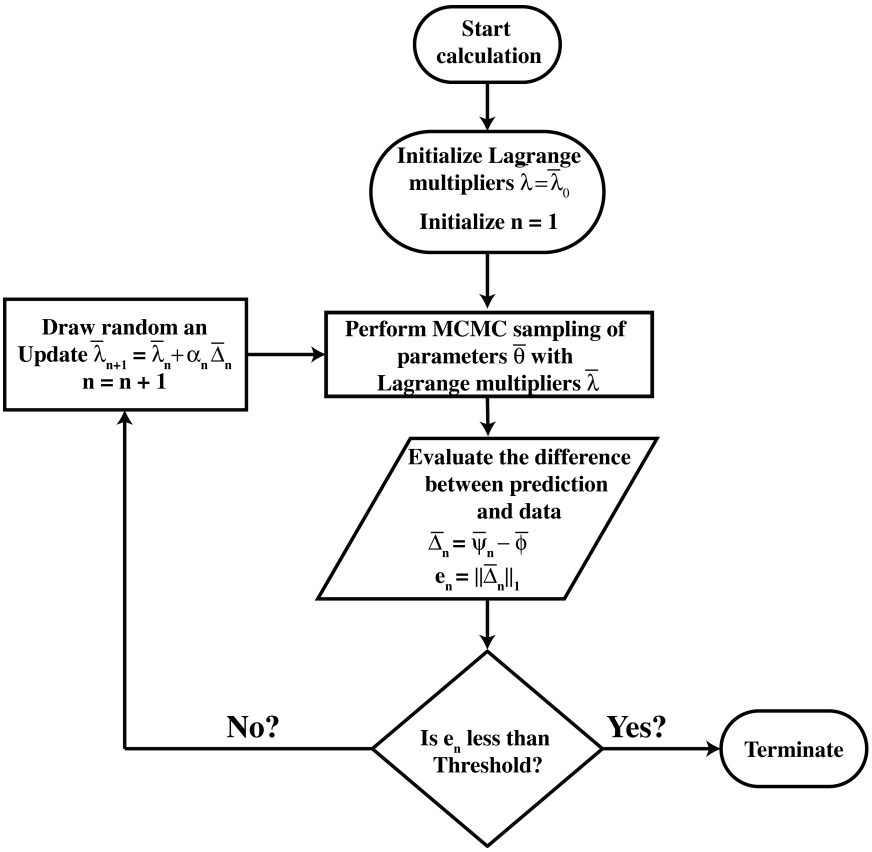

Figure 3. The workflow to numerically determine the values of the Lagrange multipliers. In every iteration we evaluated the error between predicted bin fractions and the experimentally measured bin fractions. We proposed a new set of Lagrange multipliers based on the error. We repeated until the error reached below a predefined accuracy cutoff.

Once the Lagrange multipliers were determined with sufficient accuracy, we sampled multiple parameter points from the inferred distribution $P(\bar{\theta})$ for further analysis. The inferred distribution captured with high accuracy the bin fractions ϕ_{ij}s of the distributions that were used to constrain it ($r^2 \sim 0.91, p < 10^{-10}$). In Figure 4, we show the time evolution of cell to cell variability in pAkt levels at 5, 15, 30, and 45 minutes after exposure to 10 ng/ml EGF. The dashed black lines represent the fitted bin fractions calculated using $P(\bar{\theta})$ and filled circles represent the corresponding experimental data. Notably, fitted bin fractions obtained in two independent calculations agreed with each other with a high degree of accuracy as expected for a convex optimization problem ($r^2 \sim 0.94, p < 10^{-10}$, see SI Figure 2). See SI section III for details of the numerical procedure.

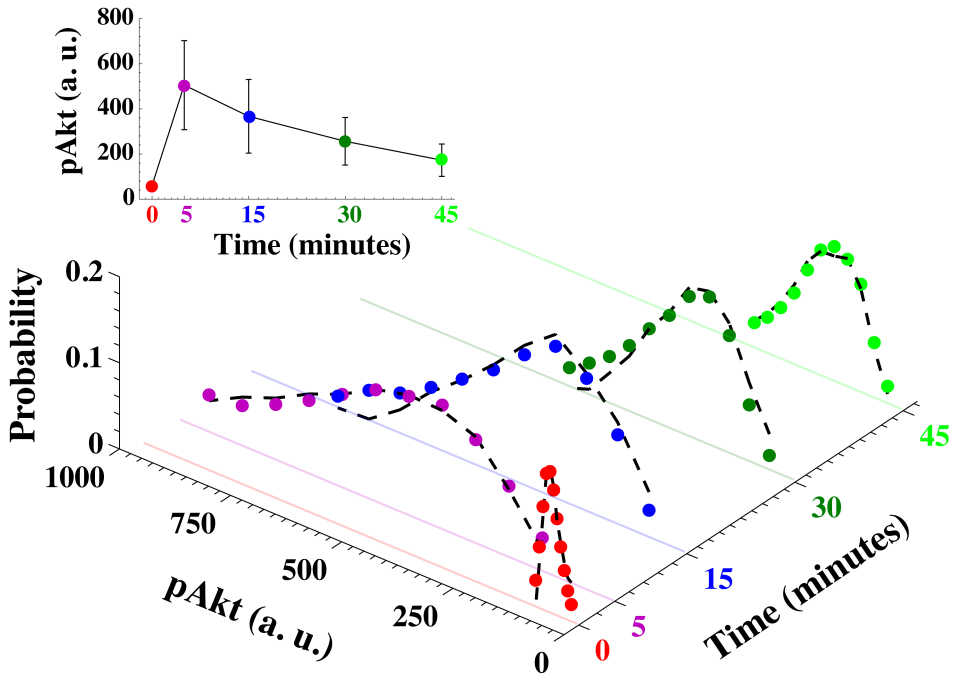

Figure 4. Distribution of pAkt levels at 0, 5, 15, 30, and 45 minutes after exposure to 10 ng/ml EGF. The colored circles represent the experimental data used in the inference of the parameter distribution. The dashed black lines represent the fitted pAkt distributions using the inferred $P(\bar{\theta})$. The inset shows population average pAkt levels at multiple time points. In the inset, filled circles are experimentally measured population averages. Error bars represent the experimentally measured standard deviation of the distributions of pAkt levels.

Predictions of network dynamics

Akt phosphorylation results in upregulation of metabolic activities such as protein synthesis and glycolysis and in cell proliferation[24]. Indeed, high pAkt levels are implicated in many types of cancers[30]. Consequently, the fraction of cells with high pAkt levels near the maximum of the phosphorylation as well as at steady state could serve as predictive markers for abnormal behavior. Using the estimated parameter distribution, we predicted the cell to cell variability in pAkt at 10 minutes and three hours after EGF stimulation. Importantly, these two time points were not used to constrain the parameter distribution $P(\bar{\theta})$. In Figure 5 a, b, c, and d we show the agreement between predicted (dashed black lines) and measured (filled circles) distributions of cell to cell variability in pAkt levels after stimulation with EGF. In fact, the ME framework predicted with great accuracy the cell to cell variability in pAkt levels across all time points and EGF

exposures that were not used in constraining $P(\bar{\theta})$; there was a significant correlation between experimentally estimated population average (Pearson $r^2 \sim 0.97$, $p < 10^{-10}$) as well as heterogeneity (quantified as the standard deviation in cell to cell variability) (Pearson $r^2 \sim 0.98$, $p < 10^{-10}$) in pAkt levels and the corresponding predictions (see SI Figure 3 in SI section IV).

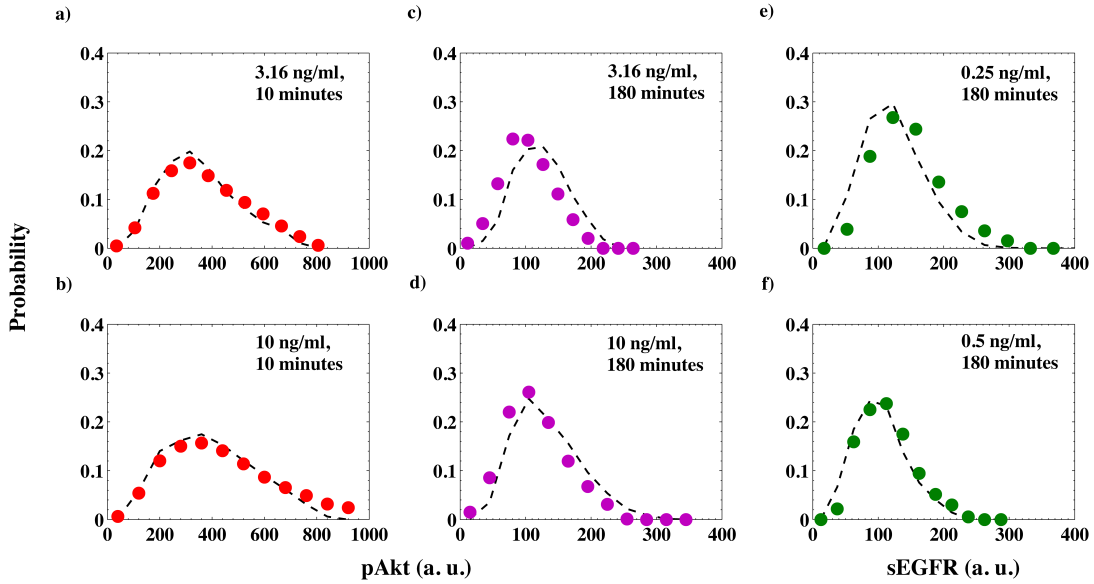

Figure 5. Experimentally measured distributions (filled circles) and maximum entropy predictions (dashed black lines) of cell to cell variability in pAkt levels at 10 minutes (a and b) and 180 minutes (c and d) after exposure to extracellular EGF 3.16 ng/ml and 10 ng/ml respectively. e) and f) Cell to cell variability in steady state sEGFR levels (180 minutes after EGF exposure) at EGF doses 0.25 ng/ml and 0.5 ng/ml respectively.

The abundance of surface EGF receptors determines cells' ability to phosphorylate signaling components downstream of EGFRs in response to EGF stimulation[25]. Steady state sEGFR abundances after prolonged stimulation with EGF are thus crucial in quantifying the sensitivity of EGF/EGFR signaling cascade within individual cells[31]. Using the inferred parameter distribution, in addition to cell to cell variability in pAkt levels, we can also predict the distribution of cell surface receptor levels at steady state (180 minutes after EGF exposure) across different doses of EGF. In Figure 5 e and f, we show the agreement between experimentally measured cell to cell variability in sEGFR levels (black lines) at different doses of EGF along with the predictions (red lines). Notably, only two steady state sEGFR levels (1 ng/ml and 100

ng/ml) were used to constrain the parameter distribution $P(\bar{\theta})$. Similar to pAkt, the ME framework predicted with great accuracy the experimentally estimated population average (Pearson $r^2 \sim 0.98$, $p \sim 10^{-5}$) and cell to cell variability (Pearson $r^2 \sim 0.96$, $p \sim 8 \times 10^{-5}$) in steady state sEGFR levels (see SI Figure 4 in SI section IV).

Extensions of the framework

A straightforward modification allows us to use the developed framework for cases when the time evolution of species abundances $\bar{X}$ is intrinsically stochastic, for example, transcriptional networks and prokaryotic signaling networks[1]. To that end, we can modify the definition of the predicted fraction $\psi_b = \int P\left(x(t,\bar{\theta}) = x \middle| \bar{\theta}\right) \bullet I_b(x)dx$ of a chemical species x where $P\left(x(t,\bar{\theta}) = x \middle| \bar{\theta}\right)$ is the distribution of x values at time t when parameters are fixed at $\bar{\theta}$. The distributions can be obtained numerically using Gillespie's stochastic algorithm[32] or approximated using moment closure techniques[33].

The presented framework can also be used to infer parameter distributions when instead of the entire distribution of cell to cell variability only a few moments of that distribution are available, for example, when average protein abundances are measured using techniques such as quantitative western blots or mass spectrometry. For simplicity we elucidate the case where population mean m and the variance v of one species are measured at a fixed time point t. Instead of constraining fractions ψ_b that represent cell to cell variability in different bins, we constrain the population mean $\mu_1 = \int x(t,\bar{\theta})P(\bar{\theta})d\bar{\theta}$ and the squared mean $\mu_2 = \int x(t,\bar{\theta})^2 P(\bar{\theta})d\bar{\theta}$ to their experimentally measured values m and $v+m^2$ respectively. Entropy maximization can then be carried out with these constraints. We have

$$P(\bar{\theta}) = \frac{1}{\Omega} q(\bar{\theta}) \exp\left(-\lambda_1 x(t,\bar{\theta}) - \lambda_2 x(t,\bar{\theta})^2\right) \quad (7)$$

Lastly, we can use the ME framework to infer parameters from experiments where dynamical changes in abundances of chemical species within single cells are measured using live cell tracking[5]. Consider that the time evolution of a species $x(t)$ is

measured in n_c cells from time $t=0$ to $t = T$. For individual cells, we can estimate features of the trajectory $x(t)$, for example, the maximum response, the time taken to reach the maximum, the rate of signal decay, or the steady state value. We can then use the distributions of these features obtained from the data to constrain the parameters.

Alternatively, we can discretize the n_c continuous time observations into K discrete times $t = \{t_1, t_2, \ldots, t_K\}$. 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 $x(t) \sim \{B_{1a_1}, B_{2a_2}, B_{3a_3}, \ldots, B_{Ka_K}\}$ where B_{ia_i} is the index of the bin through which the trajectory $x(t)$ passed at time point t_i. If we have sufficiently large number of trajectories, similar to constraining trajectories that populated individual bins, we then can constrain the fraction of trajectories that populate a given sequence of bins and infer $P(\bar{\theta})$.

Discussion

In this work we presented a maximum entropy based framework to estimate the joint distribution $P(\bar{\theta})$ over parameters of a signaling network based on measured cell to cell variability in signaling network dynamics. Notably, the parameter distribution allowed us to accurately predict the time evolution of cell to cell variability in species abundances.

The inferred parameter distribution comprises both the *non-identifiability* and the true cell to cell variability in parameters. The effect of parameter *non-identifiability* can be minimized by explicitly incorporating constraints related to population averages as well as cell to cell variability of rate parameters in the inference[27]. Notably, the maximum entropy framework naturally avoids over-constraining the parameter distribution; Lagrange multipliers corresponding to redundant constraints automatically evaluate to zero[15].

In this work, we employed the developed inference framework to signaling network data. However, the theoretical development can also be used in a more general

setting. For example, the framework can be applied to computationally reconstruct the distribution of longitudinal behaviors in a population from cross-sectional time-snapshot data in other fields such public health, economics, and ecology or to estimate parameter distributions from lower dimensional statistics[34].

References

1. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216-226 (2008).
2. Chastanet, A. et al. Broadly heterogeneous activation of the master regulator for sporulation in Bacillus subtilis. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8486-8491 (2010).
3. Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M. & Sorger, P.K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428-432 (2009).
4. Wu, M. & Singh, A.K. Single-cell protein analysis. *Curr Opin Biotechnol* **23**, 83-88 (2012).
5. Meyer, R. et al. Heterogeneous kinetics of AKT signaling in individual cells are accounted for by variable protein concentration. *Front Physiol* **3**, 451 (2012).
6. Niepel, M., Spencer, S.L. & Sorger, P.K. Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr Opin Chem Biol* **13**, 556-561 (2009).
7. Snijder, B. et al. Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461**, 520-523 (2009).
8. Raue, A. et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923-1929 (2009).
9. Eydgahi, H. et al. Properties of cell death models calibrated and compared using Bayesian approaches. *Mol Syst Biol* **9**, 644 (2013).
10. Transtrum, M.K. et al. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys* **143**, 010901 (2015).
11. Zechner, C. & Koeppl, H. Uncoupled analysis of stochastic reaction networks in fluctuating environments. *PLoS Comput Biol* **10**, e1003942 (2014).
12. Zechner, C. et al. Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci U S A* **109**, 8340-8345 (2012).
13. Hasenauer, J. et al. Identification of models of heterogeneous cell populations from population snapshot data. *BMC Bioinformatics* **12**, 125 (2011).
14. Dixit, P.D. Quantifying extrinsic noise in gene expression using the maximum entropy framework. *Biophys J* **104**, 2743-2750 (2013).
15. Presse, S., Ghosh, K., Lee, J. & Dill, K.A. Principles of maximum entropy and maximum caliber in statistical physics. *Rev Mod Phys* **85**, 1115-1141 (2013).

16. Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* **106**, 67-72 (2009).
17. Mora, T., Walczak, A.M., Bialek, W. & Callan, C.G., Jr. Maximum entropy models for antibody diversity. *Proc Natl Acad Sci U S A* **107**, 5405-5410 (2010).
18. Schneidman, E., Berry, M.J., 2nd, Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440**, 1007-1012 (2006).
19. Dixit, P.D., Jain, A., Stock, G. & Dill, K.A. Inferring Transition Rates of Networks from Populations in Continuous-Time Markov Processes. *J Chem Theory Comput* **11**, 5464-5472 (2015).
20. Dixit, P.D. & Dill, K.A. Inferring Microscopic Kinetic Rates from Stationary State Distributions. *J Chem Theory Comput* **10**, 3002-3005 (2014).
21. Tiwary, P. & Berne, B.J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc Natl Acad Sci U S A* **113**, 2839-2844 (2016).
22. Presse, S., Ghosh, K. & Dill, K.A. Modeling stochastic dynamics in biochemical systems with feedback using maximum caliber. *J Phys Chem B* **115**, 6202-6212 (2011).
23. Presse, S., Ghosh, K., Phillips, R. & Dill, K.A. Dynamical fluctuations in biochemical reactions and cycles. *Phys Rev E Stat Nonlin Soft Matter Phys* **82**, 031905 (2010).
24. Manning, B.D. & Toker, A. AKT/PKB Signaling: Navigating the Network. *Cell* **169**, 381-405 (2017).
25. Herbst, R.S. Review of epidermal growth factor receptor biology. *Int J Radiat Oncol Biol Phys* **59**, 21-26 (2004).
26. Soule, H.D. et al. Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer research* **50**, 6075-6086 (1990).
27. Caticha, A. & Preuss, R. Maximum entropy and Bayesian data analysis: Entropic prior distributions. *Phys Rev E Stat Nonlin Soft Matter Phys* **70**, 046127 (2004).
28. 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* **5**, 239 (2009).
29. Broderick T., D.M., Tkacik G., Schapire R. E., Bialek W. Faster solutions of the inverse pairwise Ising problem. *arXiv* **0712.2437** (2007).
30. Vivanco, I. & Sawyers, C.L. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. *Nat Rev Cancer* **2**, 489-501 (2002).
31. Sorkin, A. & Goh, L.K. Endocytosis and intracellular trafficking of ErbBs. *Exp Cell Res* **315**, 683-696 (2009).
32. Gillespie, D.T. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem* **58**, 35-55 (2007).
33. Gillespie, C.S. Moment-closure approximations for mass-action models. *IET Syst Biol* **3**, 52-58 (2009).

34. Das, J., Mukherjee, S. & Hodge, S.E. Maximum Entropy Estimation of Probability Distribution of Variables in Higher Dimensions from Lower Dimensional Data. *Entropy (Basel)* **17**, 4986-4999 (2015).