SCIENTIFIC REPORTS

OPEN

Received: 2 November 2017
Accepted: 18 May 2018
Published online: 03 August 2018

Fitting mathematical models of biochemical pathways to steady state perturbation response data without simulating perturbation experiments

Tapesh Santra

Fitting Ordinary Differential Equation (ODE) models of signal transduction networks (STNs) to experimental data is a challenging problem. Computational parameter fitting algorithms simulate a model many times with different sets of parameter values until the simulated STN behaviour match closely with experimental data. This process can be slow when the model is fitted to measurements of STN responses to numerous perturbations, since this requires simulating the model as many times as the number of perturbations for each set of parameter values. Here, I propose an approach that avoids simulating perturbation experiments when fitting ODE models to steady state perturbation response (SSPR) data. Instead of fitting the model directly to SSPR data, it finds model parameters which provides a close match between the scaled Jacobian matrices (SJM) of the model, which are numerically calculated using the model's rate equations and estimated from SSPR data using modular response analysis (MRA). The numerical estimation of SJM of an ODE model does not require simulating perturbation experiments, saving significant computation time. The effectiveness of this approach is demonstrated by fitting ODE models of the Mitogen Activated Protein Kinase (MAPK) pathway using simulated and real SSPR data.

Computational modelling of STNs is about formulating the biochemical reactions of these networks using systems of differential equations. These models help us understand how environmental stimuli, growth factors, stress signals etc. induce various cellular phenotypes via sequences of biochemical reactions[1]. ODE models can also be used to make quantitative predictions about the behaviour of SNTs, when experimental measurements are unavailable. These models have many parameters which represent physicochemical quantities such as rates of biochemical reactions, synthesis and degradation rates of macromolecules, delays incurred in transcription and translation of genes and proteins etc. The values of these parameters cannot always be experimentally measured and are often inferred using computational algorithms. The basic strategy of these algorithms is to simulate the model repeatedly with different sets of parameter values, and then compare the simulated activities of the STN with experimental data, until a close match is found. Inferring parameter values using computational algorithms can be slow, because there are infinitely many possible parameter values to explore. Additionally, numerical simulation of ODE models can also be computation intensive. To speed up the process, existing methods[2–11] focus on developing (a) clever search algorithms which quickly narrow down the potential values of parameters from infinitesimally large number of possibilities to a relatively manageable set of likely values[2,4–9], (b) fast numerical simulators to simulate the ODE models or solve its rate equations. Despite significant progresses in both avenues, fitting even moderately large ODE models involving more than ten biochemical species to multi-perturbation datasets can be computationally challenging. A particularly popular type of multi-perturbation data which are quantified by perturbing the STNs using chemical inhibitors, siRNAs, viral vectors or plasmids; letting all components of the STN to relax into a steady state following each perturbation; and subsequently measuring the phosphorylation levels of each component[2,12–15]. SSPRs are relatively easy to generate using multiplexed antibody arrays such as Luminex, Reverse Phase Protein arrays etc. and

Systems Biology Ireland, School of Medicine, University College Dublin, Belfield, Dublin, 4, Ireland. Correspondence and requests for materials should be addressed to T.S. (email: tapesh.santra@ucd.ie)

highly useful in reconstructing the wiring diagrams of the STN[2,12–18]. However, using this data to fit ODE model parameters can be challenging. This is because, existing algorithms work by matching simulated SSPRs with the experimental data, i.e. these methods need to simulate all perturbation experiments using the ODE model for each set of parameter value. For instance, if a dataset contains the SSPR responses of an STN to twenty drugs or inhibitors, a parameter calibration algorithm will need to simulate the ODE model twenty times for each potential set of parameter values. This can be computationally challenging. Additionally, in-order to simulate these perturbations using ODE models, one needs to know the exact targets of the perturbing reagents. This information is often unavailable, since most chemical inhibitors are known to influence proteins other than their designated targets. This makes simulating perturbation experiments infeasible.

Here, I propose a method which allows calibrating ODE model parameters using SSPR data without simulating perturbation experiments. Instead of fitting the model to the SSPR data itself, the proposed method first estimates the SJM of the model from SSPR data using MRA[12]. For a given set of parameter values the SJM of an ODE model is calculated by analytically or numerically differentiating its rate equations, without simulating perturbation experiments. Any existing parameter search algorithm[4–9,19] can then be used to explore different sets of parameter values until a reasonable match between the SJMs which are calculated from SSPR data and by differentiating model equations is found. For the purpose of demonstration, I used the Adaptive Weight Approximate Bayesian Computation based Sequential Monte Carlo (AW-ABC-SMC)[19] algorithm for exploring the parameter space, mainly due to its relative simplicity of implementation. The AW-ABC-SMC algorithm, combine with the SJM based parameter fitting method proposed in this study was used to calibrate two separate models of the MAPK pathway to simulated and real SSPR data respectively. In the following sections I describe the details of this algorithm and demonstrate its applicability using simulated and real SSPRs of the MAPK STN.

Method

Linking Jacobian matrix of ODE model with SSPR data using MRA.

Let us assume that an STN contains N nodes which regulate each other's concentrations. A mathematical model ($M_{\boldsymbol{x}}$) that formulates how the interactions between the different nodes influence their concentrations consists of a set of ordinary differential equations (ODE) of the form $\dot{x}_i(t) = f_i(\boldsymbol{x}_{ri}(t), \boldsymbol{\Theta}_i)$, $i = 1, \ldots, N$; where $\dot{x}_i(t)$ represent the rate at which the concentration ($\dot{x}_i(t)$) of the i^{th} node changes with time (t), f_i is a continuous function, $\boldsymbol{x}_{ri}(t)$ are the concentrations of the regulators of node i including itself, $\boldsymbol{\Theta}_i$ are the parameters of the function f_i. The values of the parameters ($\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_i, i = 1, \ldots, N\}$) are unknown and needs to be estimated from experimentally observed data. The experimental data is generated by perturbing the network many times using different biochemical reagents. Following a perturbation (p_i) to each node (i), the STN is allowed to relax into a steady-state and the changes in the concentrations of all nodes ($\mathbf{x} = \{x_i, i = 1, \ldots, N\}$) in response to each perturbation (p_i) are measured. Our objective is to use this data to fit the parameters ($\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_i, i = 1, \ldots, N\}$) of the mathematical model ($M_{\boldsymbol{x}}$) of the STN without simulating the perturbation experiments during the fitting process. To do so, we exploit a relationship between the Jacobian matrix ($\boldsymbol{J}(t)$) of the ODE model and the experimentally observed SSPRs of the STN.

Note that at steady state ($t = t_{ss}$), $\dot{x}_i(t) = f_i(\boldsymbol{x}_{ri}(t), \boldsymbol{\Theta}_i) = 0$ and therefore $df_i(\boldsymbol{x}_{ri}(t), \boldsymbol{\Theta}_i) = 0$. Using chain rule of derivative

$$df_i(\boldsymbol{x}_{ri}(t),\ \boldsymbol{\Theta}_i)\big|_{t=t_{ss}} = \sum_{x_j \in \boldsymbol{X}_{ri}(t)} \frac{\partial f_i(\boldsymbol{x}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_j(t)} .\ dx_j(t)\Bigg|_{t=t_{ss}} = 0 \tag{1}$$

If we assume a hypothetical scenario where all but the ith and j^{jh} nodes of the STN are kept fixed then the above equation reduces to:

$$\frac{\partial f_i(\boldsymbol{x}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_j(t)} dx_j(t) + \frac{\partial f_i(\boldsymbol{x}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_i(t)} dx_i(t)\Bigg|_{t=t_{ss}} = 0$$

$$\text{i.e.}\quad \frac{dx_i(t)}{dx_j(t)} = -\frac{\frac{\partial f_i(\boldsymbol{x}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_j}}{\frac{\partial f_i(\boldsymbol{x}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_i}}\Bigg|_{t=t_{ss}} \tag{2}$$

$\frac{dx_i(t)}{dx_j(t)}$ quantifies the change in the concentration (x_i) of node i, due to an infinitesimally small perturbation to node j, when the concentrations of all other nodes are kept fixed. In other words, $\frac{dx_i}{dx_j}$ represents the influence of an infinitesimally small perturbation to node j on the concentration of node i when all interactions but the regulation of node i by j are disconnected. The numerator $\left(\frac{\partial f_i(\ \boldsymbol{X}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_i(t)}\right)$ and denominator $\left(\frac{\partial f_i(\ \boldsymbol{X}_{ri}(t),\ \boldsymbol{\Theta}_i)}{\partial x_j(t)}\right)$ of the right hand side of Eq. 2 are in fact (i, i)th and (i, j)th elements ($\boldsymbol{J}_{\boldsymbol{ii}}$, $\boldsymbol{J}_{ij}$) of the STN's Jacobian Matrix ($\boldsymbol{J}(t)$) which is defined as $\boldsymbol{J}(t) = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$, where $\boldsymbol{f} = \{f_i(\boldsymbol{x}_{ri}(t), \boldsymbol{\Theta}_i), i = 1 \ldots N\}$ and $\boldsymbol{x} = \{x_j\,(t), j = 1 \ldots N\}$. Scaling both sides of Eq. 3 by the ratio ($x_j(t_{ss})/x_i(t_{ss})$) of the steady state concentrations of nodes j & i gives us

$$r_{ij} = \frac{\frac{dx_i(t)}{x_i(t)}}{\frac{dx_{j(t)}}{x_j(t)}}\Bigg|_{t_{ss}} = \frac{dln(x_i\ (t))}{dln(x_j(t))}\Bigg|_{t_{ss}} = -\frac{J_{ij}(t)\ x_j(t)}{J_{ii}(t)\ x_i(t)}\Bigg|_{t=t_{ss}} \tag{3}$$

r_{ij} is the $(I, j)^{th}$ element of the Jacobian matrix, scaled by the diagonal element in the same row and the ratio of the steady state concentrations of the ith and jth nodes. This quantity is formally known as the local response coefficient (LRC) of the regulation of node *i* by *j* and represents the change in the logarithmic steady state concentration of node *i* due to a small perturbation to node *j*, when all other nodes are disconnected. To be consistent with existing literature we shall refer to this quantity as local response coefficients or LRCs in short, instead of SJMs. It was shown by Kholodenko *et al.*[12] that the local response matrix ($\boldsymbol{r} = \{r_{ij}, i, j = 1, \ldots, N\}$ for notational convenience) can be calculated from the experimentally observed SSPRs by solving the following linear equations

$$\boldsymbol{rR} = (dg((\boldsymbol{R})^{-1}))^{-1} \tag{4}$$

where $\boldsymbol{R} = \{R_{ij}, i, j = 1, \ldots, N\}$ is the global response matrix whose elements $R_{ij} = (\Delta ln(x_i)/\Delta ln(x_j))$ are known as global response coefficients which represent the 'global change' (i.e. when the perturbation propagates through the network) in the logarithmic concentration of node *i* due to an infinitesimally small perturbation to node *j*. Experimental perturbations are never infinitesimally small, therefore R_{ij} is calculated in an approximate sense from experimental SSPR data using the following formula[12]:

$$R_{ij} = 2\frac{x_i^j - x_i}{x_i + x_i^j} \tag{5}$$

here x_i and x_i^j is the experimentally measured steady state concentrations of node *i* prior to and following a perturbation to node *j* respectively.

Eq. 3 allows us to calculate the local response matrix (**r**) of and STN using its ODE models without explicitly simulating the perturbation experiments. Therefore, estimating the model parameters ($\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_i, i = 1, \ldots, N\}$) boils down to the following steps:

- **Step1:** Calculate the local response matrix ($\boldsymbol{r}$) from experimentally observed SSPRs using Eqs 5 and 4
- **Step2:** Calculate the local response matrix ($\boldsymbol{r}_s$) using Eq. 3 for different values of model parameters and chose the sets of values which provide close match between $\boldsymbol{r}$ and $\boldsymbol{r}_s$.

Search for parameter values that minimizes the difference between $\boldsymbol{r}$ and $\boldsymbol{r}_s$ out of infinitesimally many possibilities can be performed using any existing parameter inference algorithm. For the purpose of demonstration we used AW- ABC-SMC[19], an improved version of the original ABC-SMC[7]. The details of this algorithm is discussed in a later subsection.

Computational efficiency gained by the proposed approach.

Let us consider a dataset containing SSPRs of an STN to N_p perturbations and measurements from a control experiment where the STN was unperturbed[13,14], altogether the dataset contains $N_p + 1$ sets of STN responses. Fitting an ODE model of the STN to this dataset in the traditional way requires simulating the model $N_p + 1$ times for each set of parameter. However, to fit the model parameters using the approach proposed above, one needs to calculate the LRCs ($\boldsymbol{r}_s$, Eq. 3) of the ODE model for each set of parameter values. These calculated by differentiating the rate equations of the unperturbed model at steady-state, meaning that the steady-state of the model needs to be calculate once by solving the rate equations. No other model simulation is required. Therefore, when using the proposed approach the model equations need to be solved once for each set of parameter values as opposed to traditional methods which require $N_p + 1$ ODE simulations; i.e. a typical parameter fitting algorithm will require only $\approx \frac{1}{1 + N_p} \times 100\%$ of the execution time by fitting parameters to local response matrices instead of the SSPR data, assuming that most of the execution time is spent by simulating/solving the ODE models. For a typical SSPR dataset containing 8–15 SSPRs[13,14], fitting model parameter to the local response matrix will take only 7–11% of the execution time of the alternative approach which requires simulating the perturbation experiments. The time saving is even more significant when N_p is larger.

Experimental requirements for the proposed approach.

Since the above approach relies on the local response matrix of the STN, there need to be enough experimental data to calculate this matrix. How much data is required to calculate local response matrix depend on the method being used for this calculation. The classical MRA[12] requires exactly as many perturbations as the number of STN components to calculate this matrix. The total-least-square regression based MRA formulations requires at least as many perturbations as the classical MRA[20]. More recent Bayesian and Maximum Likelihood method based MRA realizations can calculate local response matrices using data from less number of perturbation experiments than required by the classical MRA[2,13]. There is no rule of thumb for estimating the minimum number of perturbations that are required for calculating local response matrices with reasonable accuracy. However, the Bayesian formulations of MRA was successfully[2] used to calculate the local response matrices using SSPRs from half as many perturbations as the number of components in the STN.

Exploring the parameter space using Adaptive Weight ABC-SMC.

ABC is inspired by Bayesian Statistics and relies on Bayes principle which provides a framework for updating prior knowledge about an unknown variable or quantity using observed data. The prior and updated knowledge are represented in the form of probability distributions, known as prior and posterior distributions respectively, which formulates our initial guesses and updated estimates about the potential values of the unknown variables. Adaptive Weight ABC-SMC algorithm starts by assigning prior distributions ($P(\boldsymbol{\Theta})$) to the model parameters ($\boldsymbol{\Theta}$), and initializing a monotonically decreasing set of error thresholds (ε_t, $t = 1, \ldots, T$, $\varepsilon_1 > \varepsilon_2 > \ldots > \varepsilon_T$) which will be used to refine the posterior distribution ($\boldsymbol{P(\Theta|D_{obs})}$, ($\boldsymbol{D_{obs}}$ is the observed data, which, in our case, is the local response matrix, i.e. $\mathbf{D}_{\mathrm{obs}} = \boldsymbol{r}$) of the model parameters ($\boldsymbol{\Theta}$) in a stepwise manner as described below (for details see[7,19]).

Step 1: In the first step ($t = 1$), a number of potential values ($\Theta^{\boldsymbol{1k}}$, $k = 1, \ldots.$) of the model parameters ($\boldsymbol{\Theta}$) are sampled from their prior distributions ($P(\boldsymbol{\Theta})$). For each set of values ($\Theta^{\mathbf{1k}}$) a local response matrix ($\boldsymbol{r}_s{}^{1k}$) is simulated using the ODE model (M_x), and the error between the simulated ($\boldsymbol{r}_s{}^{1k}$) and the SSPR derived ($\boldsymbol{r}$) local response matrices are calculated using a distance measure ($d(\boldsymbol{r}, \boldsymbol{r}_s{}^{1k})$). If the error ($d(\boldsymbol{r}, \boldsymbol{r}_s{}^{1k})$) is less than the error threshold ε_1, i.e. $d(\boldsymbol{r}, \boldsymbol{r}_s{}^{1k}) < \varepsilon_1$, then corresponding parameter values ($\Theta^{\boldsymbol{1k}}$) are kept for next iteration, otherwise discarded. This process is repeated until a desired (N_{ABC}) number of parameters are kept ($\boldsymbol{\Theta^1} = \{\Theta^{\boldsymbol{1n}}, n = 1, .., N_{ABC}\}$). Each of the selected values ($\Theta^{\boldsymbol{1n}}$) is assigned a weight ($\omega^{1n} = 1/N_{abc}$).

Step 2: In the next step ($t = 2$), one ($\Theta^{\boldsymbol{1n}}$) of the parameter values ($\boldsymbol{\Theta^1}$) which were not discarded in the previous step is selected with probability $p^{1n} \propto \omega^{1n}\, K_c(\boldsymbol{r}_s{}^{1n}, \boldsymbol{r})$ where $K_c(\boldsymbol{r}_s{}^{1n}, \boldsymbol{r})$ measures the closeness between $\boldsymbol{r}_s{}^{1n}$ and $\boldsymbol{r}$. A new parameter value ($\Theta^{\boldsymbol{2k}}$) is then proposed by sampling a proposal distributions ($P(\Theta^2|\Theta^{1n})$) which is conditioned on the selected value ($\Theta^{\boldsymbol{1n}}$). A local response matrix ($\boldsymbol{r}_s{}^{2k}$) is then simulated using this parameter value ($\Theta^{\boldsymbol{2k}}$) and the error ($d(\boldsymbol{r}, \boldsymbol{r}_s{}^{2k})$) between the simulated ($\boldsymbol{r}_s{}^{2k}$) and the SSPR derived ($\boldsymbol{r}$) local response matrices is calculated. If the error ($d(\boldsymbol{r}, \boldsymbol{r}_s{}^{2k})$) is less than the error threshold ε_2, i.e. $d(\boldsymbol{r}, \boldsymbol{r}_s{}^{2k}) < \varepsilon_2$, then newly sampled value ($\Theta^{\boldsymbol{2k}}$) is kept, otherwise discarded. This process is repeated until a desired (N_{ABC}) number of parameters are kept ($\boldsymbol{\Theta^2} = \{\Theta^{\boldsymbol{2k}}, k = 1, .., N_{ABC}\}$). The weights of the selected values are updated as follows $\omega^{2k} \propto P(\Theta^{\boldsymbol{2k}})/\sum_{\mathrm{k}} \omega^{1k}\, P(\Theta^{2k}|\Theta^{1k})$, where the proportionality constant is the sum over all weights ($\sum_{\mathrm{k}} \omega^{tk}$).

Step 3: Step 2 is repeated T times ($t = 3, 4, \ldots, T$) when the algorithm terminates. The last set of parameter values ($\boldsymbol{\Theta^T}$) kept by the algorithm represent samples from the approximate posterior distribution of the model parameters ($\boldsymbol{\Theta}$).

Weighted Euclidean distance function was used for calculating errors ($d(\boldsymbol{r}, \boldsymbol{r}_s{}^{1k})$); Gaussian function was used for calculating both the closeness measures ($K_c(\boldsymbol{r}_s{}^{tk}, \boldsymbol{r})$) and proposal distributions ($P(\Theta^t|\Theta^{(t-1)k})$). The above algorithm was parallelized in the following manner. In each step t, instead of sampling one set of parameter values at a time and checking whether it passes the error threshold, N_B numbers of parameters values $\{\Theta^{\boldsymbol{tk}}, k = 1, \ldots, N_B\}$ were sampled at a time. Simulating and evaluating the local response matrices using each of these sampled parameter values ($\{\Theta^{\boldsymbol{tk}}, k = 1, \ldots, N_B\}$) were performed in parallel using multiple processors. Since simulating each local response matrix requires solving an ODE model, which is computation intensive, performing several such simulations in parallel saves significant computation time.

Parameter identifiability issues and potential remedies.

An STN consisting of N proteins can have up to ($N^2 - N$) possible interactions excluding self-regulation. The local response matrix ($\boldsymbol{r}$) of the STN provides a quantitative representation of each of these interactions. However a typical STN has far less interactions (N_c) than theoretically possible, i.e. $N_c \ll (N^2 - N)$. The LRCs corresponding to the non-self-regulatory interactions that are theoretically possible but do not occur in reality are close to zero[12] and do not contribute in the parameter inference process. The remaining N_c LRCs are useful for fitting parameters. However, in a typical scenario, a mathematical model requires more than N_c parameters to formulate N_c interactions, i.e. the number of parameters (N_p) in the model is typically larger than the number of interactions (N_c) it formulates ($N_p > N_c$). Generally speaking, fitting a model with less data points than the number of model parameters causes parameter identifiability and model overfitting problems. There are several ways of avoiding this problem as described below.

- One way of resolving the parameter identifiability problem is to generate SSPR data in different experimental conditions. For instance, the STN can be stimulated with different ligands or different doses of the same ligand, and following each stimulation the full set of perturbation experiments (including unperturbed measurements) needs to be performed. This will allow one to calculate multiple local response matrices ($\boldsymbol{r}^l$, $l = 1, 2, \ldots$) for the same STN. The model can then be fitted to all local response matrices simultaneously using the same method described in the previous section. In this case, when using the proposed method of parameter fitting, the model equations need to be solved for each type or dose of ligand for each set of parameter values, but the individual perturbations at each type or dose of ligand need not be simulated. Therefore the gain in computational efficiency remain the same as discussed before.
- Network features other than the local response matrix can also be obtained from multi-conditional SSPR data and used for parameter fitting. For instance, the changes in the steady state concentrations ($\boldsymbol{x}^l = \{x^l_i, i = 1, \ldots, N\}$) of the STN components due to changes in the dose or type of ligand (l) can also be useful for parameter calibration. To elaborate, let the steady state concentrations of the STN components in response to ligand stimulation l (but no other perturbation) be denoted by $\boldsymbol{x}^l = \{x^l_i, i = 1, \ldots, N\}$. The ratio ($\boldsymbol{\rho}^{jk}{}_i$) of the concentration of node i at two different types or doses of ligands ($l = j, k$), i.e. $\boldsymbol{\rho}^{jk}{}_i = \{x^j{}_i/x^k{}_i\}$, $i = 1, \ldots, N$, represents the change in concentration of node i when the ligand or ligand concentration is changed from j to k. Therefore, these ratios ($\boldsymbol{\rho}^{jk}{}_i$) quantify how different ligands influence the STN components, as opposed to local response matrices which contain information about how different nodes influence each other. Here, these ratios ($\boldsymbol{\rho}^{jk}{}_i$, $i = 1, \ldots., N$) contain information which is complementary to the local response matrices and can be augmented with these matrices ($\boldsymbol{D}_{obs} = \{\boldsymbol{r}^l, \boldsymbol{\rho}^{jk}{}_i; l = 1, 2, \ldots; j, k = 1, 2 \ldots, j \neq k, i = 1, 2, \ldots, N\}$) to further improve parameter identifiability. Incorporating the ligand response rations in the parameter fitting process does not require additional model simulation and therefore does not make noticeable difference in the computational complexity of the parameter fitting process.
- Any additional experimental data can also be incorporated in the parameter inference process, especially if it does not incur additional computational cost. For instance, time course measurements ($x_i(t)$) of the concentrations of any node (i) of the STN can also be incorporated. Incorporating time course measurements do not incur any significant additional computational cost since the model needs to be simulated once per

ligand or ligand concentration, with or without such data. The only difference is, in case of time course data, the models needs to be simulated using ODE solvers which can be slower than the numerical solvers used to solve model equations at steady state. Nevertheless, one model simulation is not likely to incur computational cost comparable to $N_p \gg 1$ perturbation experiments.

Availability. All source codes and data needed to replicate the results in this manuscript are available from https://github.com/SBIUCD/MRA_SMC_ABC1.

Results

Evaluating the proposed method using simulated data.

Simulating perturbation response of the MAPK pathway. To test our algorithm we simulated SSPR data using a mathematical model of the ERK pathway (Fig. 1A), which is a three tiered MAPK cascade that controls cell fate[21–23]. It comprises of three kinases, RAF, MEK and ERK. RAF is at the top of the cascade which is activated by RAS-GTP when ligands such as Epidermal Growth Factor (EGF) binds to EGF receptor on the cell surface. Activated RAF (aRAF) then activates MEK by phosphorylating it on two sites. Active MEK (aMEK) in turn activates ERK by doubly phosphorylating it. Activated RAF, MEK and ERK (aERK) are subsequently inactivated by phosphatases which de-phosphorylate them. Activated ERK (aERK) can inhibit the activation and assist in the inactivation of RAF and MEK respectively, thereby forming two negative feedback loops[12,24,25]. A few simplifying assumptions were made to develop a mathematical model of this pathway. For instance, while in relality EGF activates RAF via a network of adaptor proteins and RAS-GTPs, for the purpose of modelling it was assumed that RAF is directly activated by EGF. Additionally, the activations of RAF, MEK, ERK are two stage processes involving phosphorylations of two distinct sites on these kinases. For simplicity, we combined the two stage activation process of these kinases into one stage in which the inactive form of the kinase (iRAF, iMEK, iERK) are converted into their active forms (aERK, aMEK and aERK)[26]. Finally, the negative feedbacks from aERK to aMEK and aRAF operates via two different mechanisms. Finally, in reality the two ERK mediated negative feedback loops are mediated by two different mechanisms of inhibition of upstream kinase activities (aRAF and aMEK). However, for modelling, it was assumed that both feedback are caused by aERK mediated inactivation of aMEK and aRAF. Activation and inhibition of each kinase were formulated using Michaelis Menten functions as shown below

$$\begin{aligned}\frac{d(aRAF)}{dt} &= M_a(k_{f1}, K_{mf1}, iRAF, EGF) - M_a(k_{r1}, K_{mr1}, aRAF, aERK)\\ &\quad - M_0(V_{m1}, K_{m1}, aRAF)\\ \frac{d(aMEK)}{dt} &= M_a(k_{f2}, K_{mf2}, iMEK, aRAF) - M_a(k_{r2}, K_{mr2}, aMEK, aERK)\\ &\quad - M_o(V_{m2}, K_{m2}, aMEK)\\ \frac{d(aERK)}{dt} &= M_a(k_{f3}, K_{mf3}, iERK, aMEK) - M_0(V_{m3}, K_{m3}, aERK)\end{aligned} \tag{6}$$

where, $M_a(k, K, S, M) = kS\frac{M}{K+S}$, $M_0(V, K, S) = V\frac{S}{K+S}$, $iRAF = (RAF_{TOT} - aRAF)$, $iMEK = (MEK_{TOT} - aMEK)$, $iERK = (ERK_{TOT} - aERK)$, $k_{f1} = 1$, $K_{mf1} = 10$, $V_{m1} = 2$, $K_{m1} = 10$, $k_{mf2} = 1$, $K_{mf2} = 10$, $V_{m2} = 1$, $K_{m2} = 10$, $k_{mf3} = 0.1$, $K_{mf3} = 10$, $V_{m3} = 10$, $K_{m3} = 10$, $k_{r1} = 1$, $K_{r1} = 10$, $k_{r2} = 1$, $K_{r2} = 10$.

The following experimental scenario was simulated using the above mathematical model (Eq. 6). Following common practice[13–15,26], it was assumed that the cells are starved prior to stimulation by EGF. Since the phosphorylation levels of kinases are negligible in starved cells, the initial concentrations ($aRAF_{t=0}$, $aMEK_{t=0}$, $aERK_{t=0}$) of aRAF, aMEK and aERK were set to zero. The starved cells are stimulated by adding EGF to the growth medium. This was simulated by setting the EGF level of the model to a positive constant. The cells are then allowed to relax until they reach steady state. This was simulated by running the model until steady state. Once the cells attained steady state, the concentrations of active RAF, MEK and ERK are measured using antibodies or fluorescent reporters which amplify the changes in concentrations by several orders of magnitude. The amplifying effects of antibodies and reporters were simulated by multiplying the simulated concentrations by a large constant ($k_f \gg 1$)[27]. Biological measurements are typically noisy, which was simulated by adding random Gaussian noise to the amplified concentrations. Since biological data are typically generated in replicates, we generated six replicates for each measurements, each of which is a noisy realization of the amplified concentrations.

The above data represents active RAF, MEK and ERK in EGF stimulated, but otherwise unperturbed cells. To simulate perturbation experiments, it was assumed that the ERK pathway was perturbed by transfecting the cells with siRNAs targeting RAF, MEK, and ERK. Since siRNAs reduce the total concentration of their target proteins, the perturbations were simulated by reducing the total amount of RAF, MEK and ERK (aRAF + iRAF, aMEK + iMEK, aERK + iERK respectively) in the ODE model. Following each perturbation, the model was simulated until steady state, the steady state concentrations were amplified and measurement noise were added as described in the previous paragraph. Six replicate measurements were generated following each perturbation. These simulated concentrations of the aRAF, aMEK and aERK can be used to calculate the local response matrix of the ERK STN using Eqs 4 and 5. However, since the STN has three active components, the local response matrix is a 3 × 3 matrix whose diagonal elements are by definition −1 (see. Eqs 2, 3) regardless of the parameter values, leaving us with six LRCs, only four of which represent true interactions, to fit the model (Eq. 6) which has sixteen parameters. Since the model has significantly more parameters than the number of LRCs, it is evident the parameters of the model are not identifiable from a single local response matrix. To solve the model identifiability issue, we generated data for four different levels of EGF (*0.1 ng/ml*, *1 ng/ml*, *2 ng/ml*, *5 ng/ml*; see Fig. 1A).

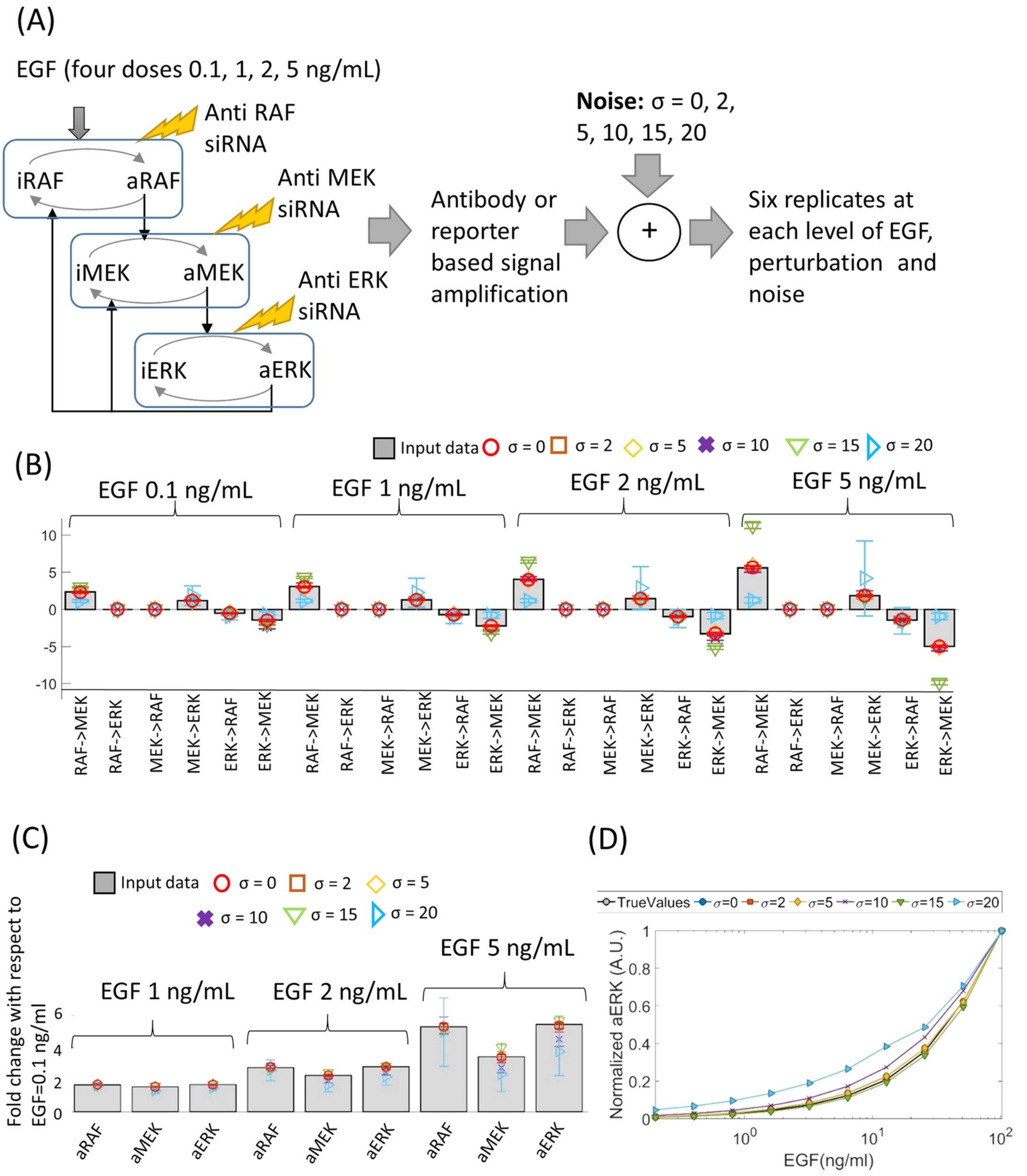

Figure 1. Parameter calibration using local response coefficients calculated from simulated data. (**A**) Schematic diagram of the MAPK model that was used to simulate perturbation response data, along with an outline of the data generation process. (**B**,**C**) Local response coefficients and steady state levels of aRAF, aMEK and aERK, simulated with the original (grey bars) and inferred parameters (coloured markers). Parameters were inferred from data contaminated with different levels ($\sigma = 0, 2, 5, 10, 15, 20$) of noise. The steady state levels of aRAF, aMEK and aERK at EGF levels 1,2,5 ng/mL are shown in terms of fold-change with respect to the same at EGF = 0.1 ng/mL. (**D**) aERK levels following stimulation by different doses of EGFs.

To evaluate the robustness of our algorithm against experimental noise, we generated data for six different levels (standard deviation $\sigma = 0, 2, 5, 10, 15, 20$) of noise (Fig. 1A). Six replicate datasets were generated at each levels of EGF and noise (Fig. 1A).

Calibrating model parameters using simulated SSPR data. The ODE (Eq. 6) was separately fitted to data containing different levels of noise. At each level of noise ($\sigma > 0$) and EGF stimulation, the means of the steady state concentrations of aRAF, aMEK and aERK were first estimated by calculating sample mean of the replicate measurements. The mean concentrations of the perturbed and the unperturbed STNs were then used to calculate four global response matrices ($\boldsymbol{R}^E$, $E = 0.1, 1, 2, 5\,ng/ml$), one for each EGF level, using Eq. 5. The global response matrices were then converted into local response matrices ($\boldsymbol{r}^E$, $E = 0.1, 1, 2, 5\,ng/ml$) using Eq. 4. The ratios ($\boldsymbol{\rho}^{E,0.1}$, $E = 1, 2, 5\,ng/ml$) between the concentrations of aRAF, aMEK and aERK at EGF levels $1, 2, 5\,ng/ml$ to those at the lowest EGF level ($0.1\,ng/ml$) were also calculated and were used for model fitting.

Adaptive weight ABC-SMC algorithm[19] was then used to calibrate the model parameters. Total concentrations of RAF, MEK and ERK were assumed to be known and therefore set to the same values that were used for data simulation. The initial concentrations of the model were set to zero reflecting characteristics of starved cells. The prior distributions of remaining sixteen model parameters was set to log-normal distributions with mean and standard deviations *2.3* and 2 respectively. For each set of parameter values, four local response matrices ($\boldsymbol{r_M}^{\boldsymbol{E}}$, $E = 0.1, 1, 2, 5\,ng/ml$), one for each EGF level, were calculated using Eqs 1–3. Three sets of ligand response ratios ($\boldsymbol{\rho_M}^{E,0.1}$, $E = 1, 2, 5\,ng/ml$) were also calculated. These were then compared with those ($\boldsymbol{r}^E$, $E = 0.1, 1, 2, 5\,ng/ml$; $\boldsymbol{\rho}^{E,0.1}$, $E = 1, 2, 5\,ng/ml$) calculated from simulated data using weighted Euclidean distance. The overall distance (d_o) between four pairs of local response matrices ($\boldsymbol{r}^E, \boldsymbol{r}_M^E$, $E = 0.1, 1, 2, 5\,ng/ml$) and three pairs of ligand response coefficients ($\boldsymbol{\rho}^{E,0.1}$, $\boldsymbol{\rho_M}^{E,0.1}$, $E = 1, 2, 5\,ng/ml$) was calculated as

$$d_o = \sum_E \left(\frac{1}{n_r^E} \right) d(\boldsymbol{r}^{\boldsymbol{E}}, \boldsymbol{r}_{\boldsymbol{M}}^{\boldsymbol{E}}) + \sum_{E=\{1,2,5\}} \left(\frac{1}{n_\rho^E} \right) d(\boldsymbol{\rho}^{E,0.1}, \boldsymbol{\rho_M}^{E,0.1}) \tag{7}$$

where $n_r^E = \sqrt{(\sum_{I,j,i \neq j} (r_{ij}^E)^2)}$ and $n_\rho^E = \sqrt{(\sum_i (\rho_i^{E,0.1})^2)}$ are weights, $d(\boldsymbol{x}, \boldsymbol{y})$ represents Euclidian distances between $\mathbf{x}$ and $\mathbf{y}$. At each stage (t) of the weighted ABC-SMC algorithm, $N_{ABC} = 1000$ sets of parameters were selected, for which the distance d_o is less than the error threshold ε_t. The set of parameters ($\boldsymbol{\Theta}^{\boldsymbol{T}} = \{\boldsymbol{\Theta}^{\boldsymbol{T}}_{\boldsymbol{i}}, i = 1, \ldots, N_{abc}\}$) that were selected at the final stage ($t = T$) of the algorithm were then used as samples from the posterior distributions of the parameters. Parameters were inferred separately from SSPR data sets containing different levels of noise. For national convenience, parameters sampled from data containing different levels of noise will be denoted by $\boldsymbol{\Theta}^{\boldsymbol{T}}_{\boldsymbol{\sigma}}$ hereafter. To see whether the sampled parameters provide a good fit to the data the LRCs and the ligand response ratios of the pathway were simulated from each set of sampled parameters ($\boldsymbol{\Theta}^{\boldsymbol{T}}_{\boldsymbol{\sigma}}$, $\sigma = 0, 2, 5, 10, 15, 20$). The mean and standard errors of these quantities are shown using markers and error bars in Fig. 1B,C. Those calculated from noise-free SSPR data are also shown in these figures using bar charts. These figures suggest that the LRCs and ligand response ratios calculated from the noise free SSPR data and simulated using the sampled parameters match closely when the noise in the data is less than $\sigma = 10$. The model fit worsens at higher noise level.

Predicting active ERK levels in response to different doses of EGF using the calibrated models. Depending on several extrinsic and intrinsic factors such as types and concentrations of ligands, reaction rates etc. a biochemical pathways may take very different temporal trajectories to arrive at the same or very similar steady states[28–30]. Therefore, pathway models that are fitted to steady state data can only be expected to predict the steady state behaviour of the pathway but not its kinetic behaviour. To see if the calibrated models can predict the steady state behaviour of the in-silico MAPK pathway, we simulated steady state dose response of aERK using the parameters sampled at each noise level. We chose EGF doses (EGFs = 0.2 0.4 0.8 1.6 3.2 6.4 12.8 25.6 51.2 102.4]; which were not used to simulate the training data. For each noise level ($\sigma = 0, 2, 5, 10, 15, 20$), aERK levels in response to different doses of EGFs were simulated using the sampled parameter values. The means and standard errors of aERK at different EGF doses were computed and plotted for each level of noise. The gold-standard in-silico aERK dose response was also plotted in the same diagram for comparison. The simulations by calibrated models (models fitted with sampled paramters) closely matched the gold-standard data when the parameters were inferred from less noisy data ($\sigma < 20$). The predictions were worse at the highest level of noise ($\sigma = 20$).

Influence of the prior parameters on model fitting. We further investigated how the choice of the prior distribution influence parameter inference. In the simulation study discussed above we chose log normal prior distributions with mean and standard deviations *2.3* and *2* respectively for all parameters. We varied the means of the prior distributions between 1 and 1000 (mean = 1, 5, 10, 20, 30, 50, 100, 1000) and for each prior mean we inferred parameters from SSPR data containing four levels of noises ($\sigma = 0, 2, 5, 10$). The inferred parameters were then used to estimate the mean local and legand response coefficients of the pathway, and the sum of squared (SSQ) distances between the estimated coefficients and the original data were calculated. The SSQs represent the model fitting errors for different prior means. The SSQs for different values of prior means at different noise levels are shown in Fig. 2. When the noise is low ($\sigma = 0, 2, 5$), the SSQs vary between 0.2–0.4 independently of the value of the prior mean. At higher noise ($\sigma = 10$), the SSQ vary between 4–6 independently of the prior mean. These results suggests the model fitting error is negligible when noise is small, it depends only on the level of noise in data and not the choice of prior mean. Therefore, the proposed algorithm is robust against choices of prior parameters.

Computation time. Fitting the ODE model of the MAPK pathway as described above took an average of ~13 minutes on a laptop computer with Intel Core i7 processor and 20 GB of RAM. When the model was fitted to the SSPR data directly it took ~47 minutes on the same computer. This roughly agrees with the general estimate of computational gain described in the methods section. It should be noted that the time complexity of the overall parameter fitting process depend on several factors related to the AW-ABC-SMC algorithm, e.g. the granularity of its error schedule, prior distribution and parallelization parameters. Therefore, the time complexity can be very different depending on the values of these factors. However, the relative gain between the proposed approach and those which require simulating perturbation experiments should be more or less the same.

Fitting an ODE model of the MAPK pathway using experimentally measured SSPR data.

Calibrating an ODE model of the MAPK pathway using experimental data. We further implemented the algorithm on a real SSPR dataset that was generated to study an interesting biological phenomena involving PC12

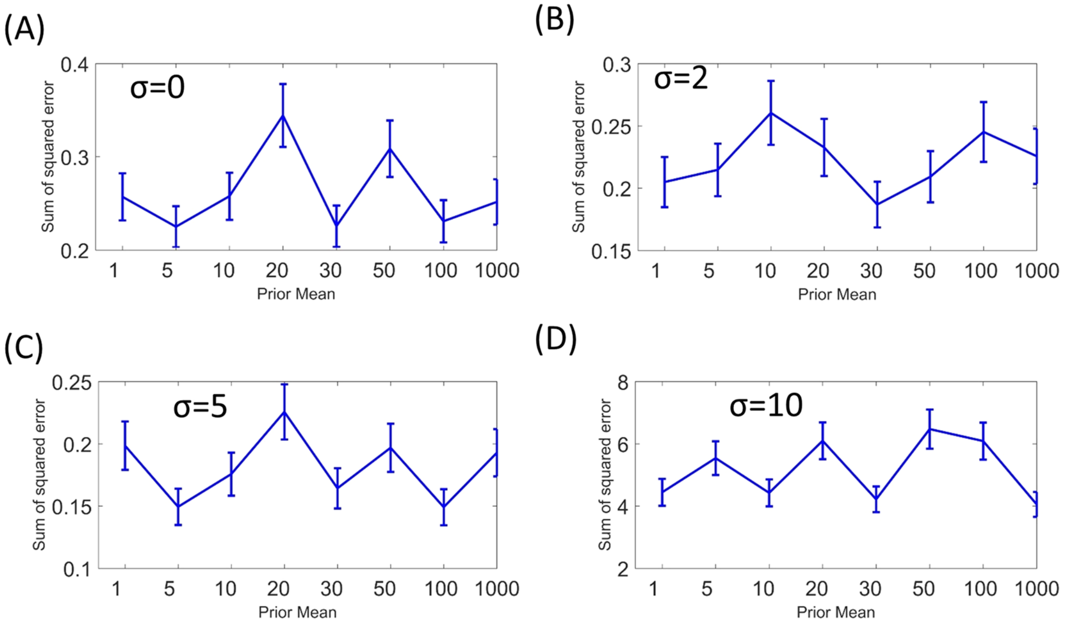

Figure 2. Effects of hyper-parameters on model fitting error. X-axis represents hyper-parameter values, Y-axis represent sum of square error between original and predicted LRCs and SSFCs. Error bars represent standanrd-deviations. Panels (A–D) show the effect of hyper-parameter choice on the ABPIPRD algorithm at different levels of measurement noise ($\sigma = 0, 2, 5, 10$ respectively).

cells which are derived from pheochromocytoma of the rat adrenal medulla. These cells proliferate and differentiate when stimulated by EGF and Nerve Growth Factor (NGF) respectively despite the fact that both of these ligands activate the ERK pathway via the same receptor (the EGFR receptor). The molecular mechanism by which EGF and NGF induce two different phenotypes is a matter of continued research. Santos et. al. studied the ERK pathway in EGF and NGF stimulated cells to understand how these ligands induce different phenotypes via this pathway[15]. They treated PC12 cells by EGF or NGF after perturbing the components (RAF, MEK, and ERK) of the ERK pathway using siRNAs and also without any perturbation, and subsequently measured the phosphorylation levels of RAF, MEK and ERK. NGF induced sustained phosphorylation of the ERK pathway, but EGF mediated phosphorylation was transient, which peaked (reached maximum level) at 5 minutes and then completely diminished at around 15 minutes[15]. Therefore, the SSPRs (~15 minutes) of ERK pathway were quantifiable in NGF stimulated cells but not in EGF stimulated cells[15]. However, 5 minutes after EGF stimulation, the phosphorylation levels of the ERK pathway reached maximum level, where the rate of change in phosphorylation levels is temporarily zero, attaining a pseudo-steady state. Therefore the perturbation responses of the ERK pathway following 5 minutes of EGF stimulation represent pseudo-SSPRs of this pathway. Santos *et al.* used the SSPRs and pseudo-SSPRs to calculated the LRCs for NGF and EGF stimulated ERK pathway (Fig. 3A)[15] respectively. The LRCs indicated different topologies of ERK pathway in response to different ligand stimulation (Fig. 3A). When EGF was used, the interaction from ERK to RAF had a negative LRC indicating the presence of a negative feedback in this condition (Fig. 3A). But when NGF was used, the LRC of the same interaction was positive, indicating the presence of a positive feedback loop. Additionally, the interaction from RAF to ERK had a relatively high positive LRC when the cells were treated with NGF, but this LRC was negligibly small in the presence of EGF. Therefore, it was concluded by Santos et. al. that there was a feedforward loop from RAF to ERK in presence of NGF, but this loop was not operational in presence of EGF (Fig. 3A). To test our method on Santos *et al.*'s dataset, an ODE model which accounted for the topological variations of the ERK pathway in response to EGF and NGF stimulations was developed. The model is shown below.

$$\begin{aligned}
d(aRAF)/dt \;&=\; b_e(M_a(k_{f11}, K_{mf11}, iRAF, RAS\_EGF(t)) \\
&\quad - M_a(k_{r12}, K_{mr12}, aRAF, aERK)) \\
&\quad + b_n(M_a(kf_{13}, Kmf_{13}, iRAF, RAS\_NGF(t)) \\
&\quad + M_a(kf_{14}, Kmf_{14}, iRAF, aERK)) - \; M_0(V_{m1}, K_{m1}, aRAF) \\
d(aMEK)/dt \;&=\; M_a(k_{f21}, K_{mf21}, iMEK, aRAF) \\
&\quad - M_a(k_{f22}, K_{mf22}, aMEK, aERK) - M_0(V_{m2}, K_{m2}, aMEK) \\
d(aERK)/dt \;&=\; M_a(k_{f31}, K_{mf31}, iERK, aMEK) \\
&\quad + b_nM_a(kf_{32}, Kmf_{32}, iERK, aRAF) - M_0(V_{m3}, K_{m3}, aERK)
\end{aligned}$$

where, $M_a(k, K, S, M) = k.S.M/(K + S)$, $M_0(V, K, S) = V.S/(K + S)$,

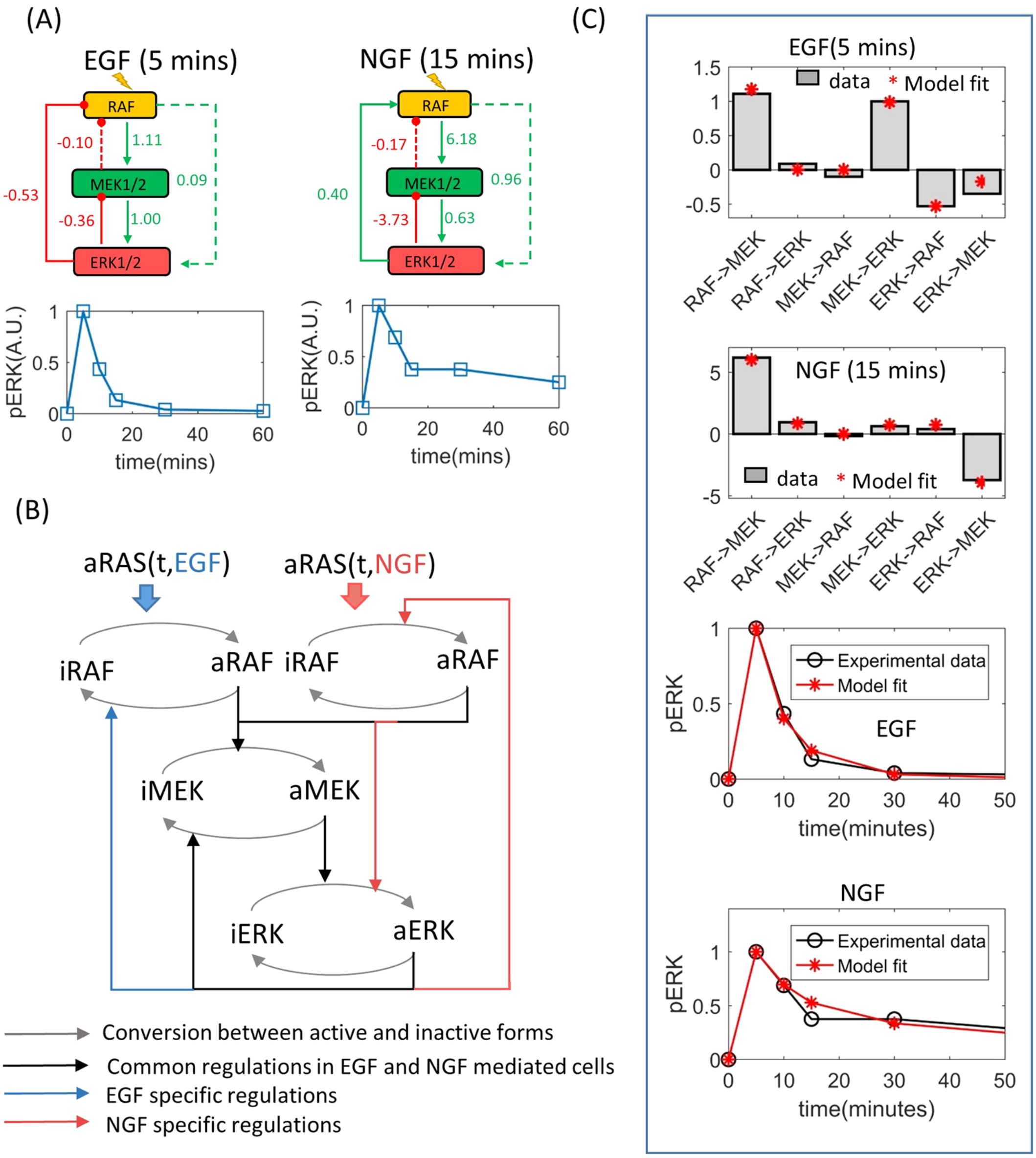

Figure 3. Fitting an ODE model of the MAPK pathway to experimental data. (**A**) LRCs of the ERK pathway and time-dependent relative pERK concentrations in EGF and NGF stimulated PC12 cells. (**B**) Schematic diagram of the ODE model that was fitted to the data presented in (A). (**C**) LRCs and time dependent pERK concentrations calculated using the fitted models. LRCs calculated from experimental data and experimentally observed pERK kinetics are also shown in this panel for comparison. Model fits represent average of an ensemble of one thousand models fitted to 1000 sets of parameters sampled by the variable weight ABC-SMC algorithm. Error bars represent standard error. Error bars are not visible due to having negligible standard error.

$$
\begin{aligned}
iRAF &= (RAF_{TOT} - aRAF),\ iMEK\\
&= (MEK_{TOT} - aMEK),\ iERK = (ERK_{TOT} - aERK),\\
RAS\_EGF(t) &= (EGF/(1 + EGF))(t^5 exp(-kd_{egf}t));\\
RAS\_NGF(t) &= (NGF/(1 + NGF))_{*}(t^5 exp(-kd_{ngf}t));
\end{aligned}
\tag{8}
$$

The above model (Eq. 8) of the ERK pathway is in some ways different from the one (Eq. 6) used for the simulation study.

- Firstly, unlike in the previous case (Eq. 6), it is no longer assumed that EGF or NGF directly activates RAF. This simplification step is avoided to reflect the biological reality that RAF is activated by the RAS proteins which are activated by EGF and NGF via a series of biochemical interactions involving the receptor and adaptor proteins. The SSPR dataset does not encompass receptor, adaptor and RAS proteins, therefore these are not exclusively incorporated in the above model (Eq. 8). But, it is known that RAS, which directly activates RAF, experiences rapid activation and successive deactivation following ligand (EGF, NGF) stimulation[26]. This transient nature of the input signal to RAF was formulated using gamma functions ($RAS\_EGF(t)$, $RAS\_NGF(t)$) with unknown parameters (kd_{egf}, kd_{ngf}) which were inferred from data.

- Secondly, the model in Eq. 8 incorporates the influence of NGF on the kinetics and the topology of the ERK pathway. Since EGF and NGF activate RAF via RAS at different rates, the influence of these ligands on RAF were formulated using two separate Michelis Menten functions (Eq. 8, Fig. 3B). Binary variables b_e and b_n were used to characterize interactions which occur selectively in response to EGF and NGF respectively (Fig. 3B).

The resulting model has twenty four unknown parameters. Additionally, the total concentrations (RAF_{TOT}, MEK_{TOT}, ERK_{TOT}) of RAF, MEK and ERK are also unknown and therefore need to be estimated. However, there are only two sets of LRCs, a total of 12 data points (excluding LRCs for self interactions which are by definition -1) available to fit the model. Fitting such a parameter rich model using such a small number of data points will almost certainly run into model identification problems. Additional data was incorporated in the inference process. Santos et. al. measured the SSPR data at 5 and 15 minutes after EGF and NGF stimulation respectively since the PC12 cells were seen to reach pseudo steady statse at these time points. This implies that the rates of changes ($d(aRAF)/dt$, $d(aMEK)/dt$, $d(aERK)/dt$) in the phosphorylation levels temporarily became zero ($d(aRAF)/dt=0$, $d(aMEK)/dt=0$, $d(aERK)/dt=0$) at these time points. This provides us six additional data points, i.e. the rates of changes in aRAF, aMEK, and aERK at 5 and 15 minutes after EGF and NGF stimulation respectively, totaling 18 data points which is still too little to calibrate a model with 27 parameters. Therefore, phosphorylation levels of ERK measured at 0, 5, 10, 15, 30 and 60 minutes following EGF and NGF stimulation[15] were also incorporated in our inference algorithm to supplement the LRCs and pseudo steady state data.

For parameter inference it was assumed that all model parameters and the total concentrations $\mathrm{RAF_{TOT}}$, $\mathrm{MEK_{TOT}}$ and $\mathrm{ERK_{TOT}}$ have log-normal prior distributions. The means of the prior distributions of all model parameters except those of the gamma functions (kd_{egf}, kd_{ngf}) were set to 2, those of the gamma function parameters (kd_{egf}, kd_{ngf}) were set to 0.2, and those of the total concentrations $\mathrm{RAF_{TOT}}$, $\mathrm{MEK_{TOT}}$ and $\mathrm{ERK_{TOT}}$ were set to 25, 100 and 400 respectively. The standard deviations of all priors were set to 2. The initial concentrations of aRAF, aMEK and aERK were all set to 0 since in Santos et. al's experiments cell were starved prior to stimulations. The Weighted ABC-SMC based algorithm was run using the above settings. LRCs of the ERK pathway model (Eq. 7) and temporal activities of aERK in response to EGF and NGF were simulated using the inferred parameters and then plotted against those derived from experimental data (Fig. 3C), showing a close match between the two. The inferred parameters were then used to predict different kinetic and steady state pathway behaviours which were not used for model calibration.

Predicting active RAF and MEK levels in EGF and NGF stimulated PC12 cells using the calibrated model. Firstly, the relative changes in the concentrations of aRAF and aMEK within a 60 minutes period after EGF and NGF stimulations were simulated. Simulations suggested that aRAF and aMEK levels peak at 5 minutes after both EGF and NGF stimulations, but diminish much quicker after EGF stimulation than NGF stimulation. The simulated aRAF and aMEK activities (Fig. 4A) qualitatively reflected the experimental data[15].

Predicting the effect of growth factor neutralizing antibodies on aERK level in PC12 cells. The response of the ERK pathway to growth factor neutralizing antibodies applied at 10 minutes after EGF and NGF stimulation were simulated using the sampled parameters. The effect of the neutralizers were formulated by setting the ligand concentrations to zero after 10 minutes. The simulation results partially agreed with the experimentally observed behaviour (Fig. 4B). In simulation, aERK level diminished completely at 60 minutes after EGF stimulation; whereas following NGF stimulation aERK level diminished at ~22% of its peak value at the same time point (Fig. 4B). While these general trends were also observed in experimental data (obtained from[15] and also shown in Fig. 4 for convenience), there were also some differences between the simulation and experimental data. The first noticeable difference is that aERK level diminished significantly faster between its peak at 5 minutes and 10 minutes (when the growth factor neutralizing antibody was applied) after EGF stimulation in the experimental observations, compared to the simulation (Fig. 4B). The second obvious difference is that the aERK level peaked at 10 minutes after NGF stimulation in the biochemical experiments, but in simulation the peak occurred at 5 minutes (Fig. 4B). In both cases, the differences between the experimental data and model simulation occur before the application of growth factor neutralizing antibodies. Therefore, it is unlikely that the difference is caused by error in simulating the effect of the neutralizing factors. A closer look at the two sets of experimentally measured phospho- ERK levels, one without the neutralizers and was used for model calibration (Fig. 3C) and the other with the neutralizers (Fig. 4B), reveals that these two sets of measurements are at odds with other. This is most likely due to biological variability between the samples used in these two experiments and/or batch effects. Therefore, in this case, the apparent differences in experimental data and model simulation can be attributed to these factors.

Predicting active ERK concentrations in response to different doses of EGF and NGF. The response of aERK at five minutes following different doses of EGFs and NGFs were simulated (Fig. 5A) using the sampled parameters. The average simulated aERK levels in response to different doses of EGF and NGF are shown in Fig. 5A. Two different sigmoidal curves were fitted to the EGF and NGF dose responses of aERK, mainly to show that (at 5 minutes after stimulation) the concentration of active ERK has a sigmoidal relationship with those of these ligands (5 A). Similar sigmoidal relationship between ligand concentrations and active-ERK levels (at five minutes

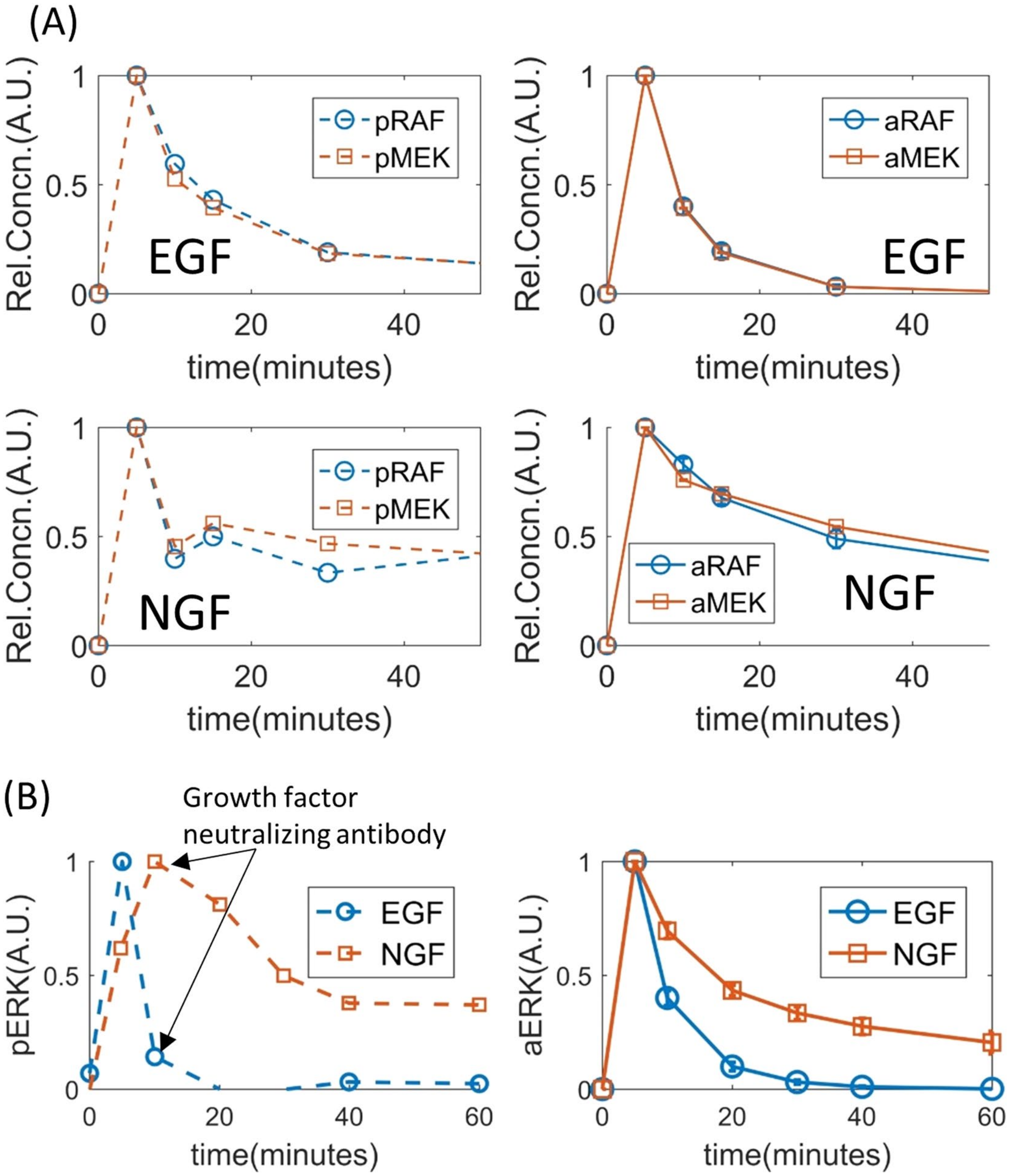

Figure 4. Simulating temporal concentrations of pRAF and pMEK using the fitted models. (**A**) Time dependent relative concentrations of pRAF and pMEK in response to EGF (top two sub-panels) and NGF (bottom two subpanels). Experimental data are shown in the left sub-panels and the model simulations are shown in the right sub-panels. (**B**) Temporal response of pERK to the application of growth factor neutralizing antibody at 10 minutes. The left and right sub-panels show experimental data and model simulation respectively. Model simulations represent average of an ensemble of one thousand models fitted to different sets of parameters sampled by the variable weight ABC-SMC algorithm. Error bars represent standard error.

after stimulation) were experimentally observed in PC12 cells (data obtained from[31], also shown in Fig. 5A for convenience).

Predicting active ERK concentrations in response to different doses of NGF at single cell resolution. The inferred parameters were used to simulate steady state response of aERK to different doses of NGF at a single cell level. To do so, each sampled set of parameter values were assumed to represent a single cell, thereby, the ensemble of all sampled parameter sets represents a cell population. Steady state aERK levels (at 60 minutes) were simulated for each set of parameter values at each level of NGF (0.01, 0.1, 1, 3, 5, 10, 30, 50, 100 ng/ml). The distribution of steady state aERK levels in a cell population in response to different doses of NGF were estimated using a kernel density estimator (https://uk.mathworks.com/help/stats/ksdensity.html). It was previously shown that at steady state, in response to NGF > 1 ng/ml phosphorylated ERK levels have bimodal distributions in populations of PC12 cells[15]. To see if the same is true for the simulated aERK levels, we fitted one or two Gaussian probability

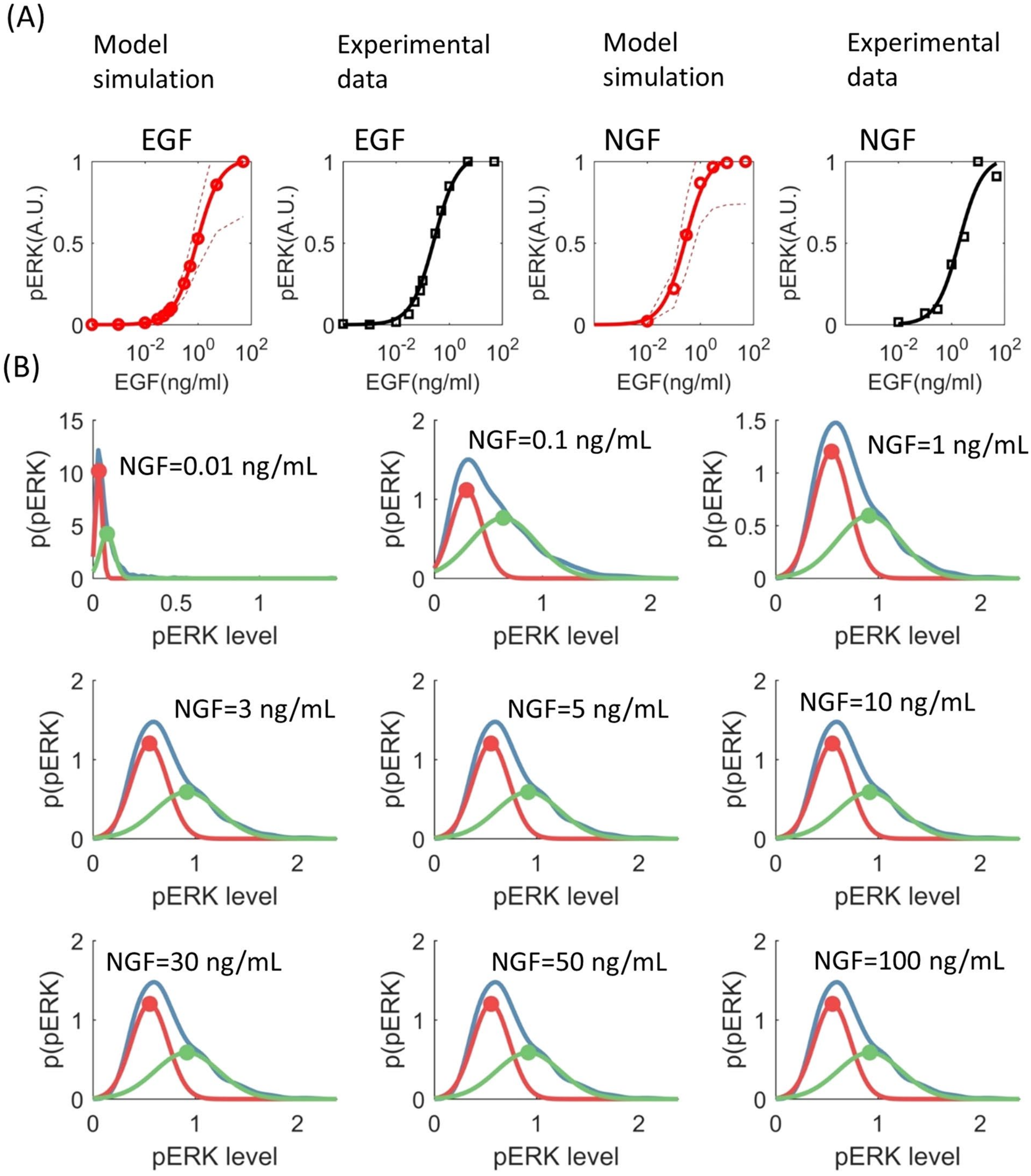

Figure 5. pERK concentrations at different doses of growth factors. (**A**) Simulated (shown in red) and experimentally measured (shown in black) relative pERK concentrations following five minutes of EGF and NGF treatments. A.U. means arbitrary units. The dashed lines represent 67% confidence interval. An ensemble of one thousand models fitted to different sets of parameters sampled by the VW-ABC-SMC algorithm were used to calculate mean response (solid red lines in panel A) and confidence intervals (dashed red lines in panel A). Bimodal distribution of steady-state (60 minutes after NGF stimulation) pERK levels following treatment by different doses of NGF. For each level of NGF, pERK levels were simulated using an ensemble of a thousand models. The empirical distributions (the blue lines in panel (B) of the simulated pERK levels are shown in blow. Individual Gaussian components that make up the empirical distributions are shown in red and green. The peak of the individual components are marked using dots of the respective colour.

density functions to each of the aERK distributions depending on whichever produced the minimum fitting (SSQ) error. In all cases, two Gaussian Distributions provided better fits than a single Gaussian distribution, suggesting that, in our simulations, aERK has bimodal distribution at all levels (0.01, 0.1, 1, 3, 5, 10, 30, 50, 100 ng/ml) of NGF stimulation (Fig. 5B). However the locations of the different modes are nearly inseparable at NGF = 0.01 ng/ml and have the highest separations at or more than 1 ng/ml NGF. Therefore, the overall simulation results largely reflects the experimental observation with one exception which occur at NGF = 0.1 ng/ml. At this concentration of NGF the simulated aERK levels were seen to have bimodal distribution whereas experimentally observed phospho-ERK levels had a single mode. One possible reason behind this difference is that, in experiments, the phospho-ERK levels were measured at 16 hours after EGF stimulation. By that time, the ERK pathway is known to be influence by transcriptional events which are not accounted for in our model[32–34]. This might cause some differences between the dose responses of the model and the real ERK pathway.

Discussion

I proposed a method that can be used to calibrate ODE models to SSPR data without exclusively simulating the perturbation experiments during the calibration process. This has several benefits beyond reducing computational cost. In many scenarios exact mechanism or 'direct' effect of the biochemical perturbations are not known, making it impossible to simulate these experiments in the first place. For instance, the mechanism of action or the exact targets of biochemical inhibitors are often either not known or not straightforward to incorporate in a model without significantly increasing the model complexity. Therefore, the data produced by the perturbation experiments where such inhibitors are used are not useful for fitting ODE models in the traditional way. The proposed approach does not require detailed knowledge of the perturbation experiments, thereby expanding the periphery of usable data for fitting ODE models. It can also be used in any existing parameter fitting algorithm to speed up the overall calibration process when using SSPR data. The models fitted using this method were shown to be able to largely reproduce STN behavior both at population and single cell level.

However, this approach of model fitting is not without its caveats. It relies on fitting parameters of a model to the LRCs of the STN. In any condition, STNs only have as many LRCs as the number of their interactions. Since each of these interactions are formulated using kinetic equations that usually have more than one parameters, in almost all cases there are more parameters to fit than the number available LRCs. This becomes even more of an issue for large networks which have many interactions, each of which is formulated using kinetic equations that may have several parameters. In such cases the difference between the number of parameters to fit and the number of available LRCs become even more apparent. Model complexity also plays a role in parameter identifiability. Mathematical models containing detailed equations for various intermediate stages of biochemical interactions are parameter rich and therefore are not easy to calibrate using LRCs. There are various ways of determining which of the model parameters are identifiable, sensitivity analysis[2] and Fisher Information Matrix[35] are some of the popular options. A common way[2] of circumventing the parameter identifiability issue is to first determine which parameters are not identifiable, assign these parameters reasonable fixed values, and then infer the values of the rest of the parameters from data. Further information about parameter identifiability issues and potential remedies are described in detail by Raue *et al.*[36]. A more straightforward way of improving parameter identifiability is to following various ligand stimulation. The behavior of biochemical networks varies depending on the dose and type of ligand stimulations, and so do the LRCs of the systems. Therefore, it is possible to estimate LRCs of the STN in response to different doses or types of ligand stimulations and use these LRCs to calibrate model. The upside of performing perturbation experiments in multiple conditions is that the resulting data is more informative than data from only one condition, but downside is the increased experimental burden.

Another potential weakness of the proposed method also stems from its inherent reliance on the LRCs. For the method to be effective, it is crucial that the LRCs are accurately estimated from SSPR data. The accuracy of the estimated LRCs depend on many factors ranging from noise, numbers and types of perturbation experiments, number of replicate experiments in the SSPR data, to the nature of the MRA based algorithms used to estimate LRCs. There are currently no rule of thumb for either designing optimal perturbation experiments to produce the most informative SSPR data, or identifying an algorithm which will produce the most accurate estimates of LRCs from an SSPR dataset. Designing optimal experimental protocols and computational algorithms to obtain the most accurate estimate of LRCs is a matter of ongoing research.

References

1. Aldridge, B. B., Burke, J. M., Lauffenburger, D. A. & Sorger, P. K. Physicochemical modelling of cell signalling pathways. *Nat Cell Biol* **8**, 1195–1203 (2006).
2. Halasz, M., Kholodenko, B. N., Kolch, W. & Santra, T. Integrating network reconstruction with mechanistic modeling to predict cancer therapies. *Sci. Signal.* **9**, ra114–ra114 (2016).
3. Degasperi, A., Fey, D. & Kholodenko, B. N. Performance of objective functions and optimisation procedures for parameter estimation in system biology models. *npj Systems Biology and Applications* **3**, 20, https://doi.org/10.1038/s41540-017-0023-2 (2017).
4. Girolami, M., Calderhead, B., Girolami, M. & Calderhead, B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **73**, 123–214, https://doi.org/10.1111/j.1467-9868.2010.00765.x (2011).
5. Jensch, A., Thomaseth, C. & Radde, N. E. Sampling-based Bayesian approaches reveal the importance of quasi-bistable behavior in cellular decision processes on the example of the MAPK signaling pathway in PC-12 cell lines. *BMC Systems Biology* **11**, 11, https://doi.org/10.1186/s12918-017-0392-6 (2017).
6. Kramer, A. *et al.* Hamiltonian Monte Carlo methods for efficient parameter estimation in steady state dynamical systems. *BMC Bioinformatics* **15**, 253–253, https://doi.org/10.1186/1471-2105-15-253 (2014).
7. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface* **6**, 187–202, https://doi.org/10.1098/rsif.2008.0172 (2009).
8. Fiedler, A., Raeth, S., Theis, F. J., Hausser, A. & Hasenauer, J. Tailored parameter optimization methods for ordinary differential equation models with steady-state constraints. *BMC Systems Biology* **10**, 80, https://doi.org/10.1186/s12918-016-0319-7 (2016).
9. Hasenauer, J., Waldherr, S., Wagner, K. & Allgöwer, F. Parameter identification, experimental design and model falsification for biological network models using semidefinite programming. *IET systems biology* **4**, 119–130 (2010).

10. Rosenblatt, M., Timmer, J. & Kaschek, D. Customized Steady-State Constraints for Parameter Estimation in Non-Linear Ordinary Differential Equation Models. *Frontiers in Cell and Developmental Biology* **4**, 41, https://doi.org/10.3389/fcell.2016.00041 (2016).
11. Rumschinski, P., Borchers, S., Bosio, S., Weismantel, R. & Findeisen, R. Set-base dynamical parameter estimation and model invalidation for biochemical reaction networks. *BMC Systems Biology* **4**, 69–69, https://doi.org/10.1186/1752-0509-4-69 (2010).
12. Kholodenko, B. N. *et al*. Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. *Proceedings of the National Academy of Sciences* **99**, 12841–12846 (2002).
13. Klinger, B. *et al*. Network quantification of EGFR signaling unveils potential for targeted combination therapy. *Molecular Systems Biology* **9**, 673–673, https://doi.org/10.1038/msb.2013.29 (2013).
14. Sahin, Ö. *et al*. Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. *BMC systems biology* **3**, 1 (2009).
15. Santos, S. D. M., Verveer, P. J. & Bastiaens, P. I. H. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat Cell Biol* **9**, 324–330 (2007).
16. Bastiaens, P. *et al*. Silence on the relevant literature and errors in implementation. *Nature biotechnology* **33**, 336–339 (2015).
17. Santra, T. A bayesian framework that integrates heterogeneous data for inferring gene regulatory networks. *Frontiers in bioengineering and biotechnology* **2**, 13 (2014).
18. Santra, T., Kolch, W. & Kholodenko, B. N. Integrating Bayesian variable selection with Modular Response Analysis to infer biochemical network topology. *BMC systems biology* **7**, 57 (2013).
19. Bonassi, F. V. & West, M. Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis* **10**, 171–187 (2015).
20. Andrec, M., Kholodenko, B. N., Levy, R. M. & Sontag, E. Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy. *Journal of theoretical biology* **232**, 427–441 (2005).
21. Kholodenko, B. N. Cell signalling dynamics in time and space. *Nature reviews. Molecular cell biology* **7**, 165 (2006).
22. Kholodenko, B. N. & Birtwistle, M. R. Four-dimensional dynamics of MAPK information-processing systems. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* **1**, 28–44 (2009).
23. Rauch, N., Rukhlenko, O. S., Kolch, W. & Kholodenko, B. N. MAPK kinase signalling dynamics regulate cell fate decisions and drug resistance. *Current opinion in structural biology* **41**, 151–158 (2016).
24. Hu, Y. & Bowtell, D. Sos1 rapidly associates with Grb2 and is hypophosphorylated when complexed with the EGF receptor after EGF stimulation. *Oncogene* **12**, 1865–1872 (1996).
25. Kolch, W. Meaningful relationships: the regulation of the Ras/Raf/MEK/ERK pathway by protein interactions. *Biochemical Journal* **351**, 289–305 (2000).
26. Borisov, N. *et al*. Systems-level interactions between insulin–EGF networks amplify mitogenic signaling. *Molecular Systems Biology* **5**, 256–256, https://doi.org/10.1038/msb.2009.19 (2009).
27. Degasperi, A. *et al*. Evaluating strategies to normalise biological replicates of Western blot data. *PloS one* **9**, e87293 (2014).
28. Ben Messaoud, N., Katzarova, I. & López, J. M. Basic Properties of the p38 Signaling Pathway in Response to Hyperosmotic Shock. *PLOS ONE* **10**, e0135249, https://doi.org/10.1371/journal.pone.0135249 (2015).
29. Fritsche-Guenther, R. *et al*. Strong negative feedback from Erk to Raf confers robustness to MAPK signalling. *Molecular Systems Biology* **7**, 489–489, https://doi.org/10.1038/msb.2011.27 (2011).
30. Nguyen, L. K. & Kholodenko, B. N. Feedback regulation in cell signalling: Lessons for cancer therapeutics. *Seminars in Cell & Developmental Biology* **50**, 85–94, https://doi.org/10.1016/j.semcdb.2015.09.024 (2016).
31. Shindo, Y. *et al*. Conversion of graded phosphorylation into switch-like nuclear translocation via autoregulatory mechanisms in ERK signalling. *Nature communications* **7** (2016).
32. Blüthgen, N. *et al*. A systems biological approach suggests that transcriptional feedback regulation by dual-specificity phosphatase 6 shapes extracellular signal-related kinase activity in RAS-transformed fibroblasts. *The FEBS journal* **276**, 1024–1035 (2009).
33. Lake, D., Corrêa, S. A. L. & Müller, J. Negative feedback regulation of the ERK1/2 MAPK pathway. *Cellular and Molecular Life Sciences* **73**, 4397–4413, https://doi.org/10.1007/s00018-016-2297-8 (2016).
34. Nagashima, T. *et al*. Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation. *Journal of biological chemistry* **282**, 4045–4056 (2007).
35. Komorowski, M., Costa, M. J., Rand, D. A. & Stumpf, M. P. H. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences* **108**, 8645 (2011).
36. Raue, A. *et al*. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics* **25**, 1923–1929, https://doi.org/10.1093/bioinformatics/btp358 (2009).

Acknowledgements

This project was funded by the Irish Cancer Society CCRC BREAST-PREDICT [grant number CCRC13GAL].

Author Contributions

T.S. designed the study, performed the analysis and wrote the manuscript.

Additional Information

Competing Interests: The author declares no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.