A hierarchical, data-driven approach to modeling single-cell populations predicts latent causes of cell-to-cell variability

Carolin Loos[a,b,1], Katharina Moeller[c,1], Fabian Fröhlich[a,b], Tim Hucho[c], and Jan Hasenauer[a,b,2]*

[a]Helmholtz Zentrum München-German Research Center for Environmental Health, Institute of Computational Biology, 85764 Neuherberg, Germany;

[b]Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, 85748 Garching, Germany;

[c]Division of Experimental Anesthesiology and Pain Research, Department of Anesthesiology and Intensive Care Medicine, University Hospital Cologne, 50937 Cologne, Germany

[1]C.L. and K.M. contributed equally to this work

*corresponding author and lead contact (jan.hasenauer@helmholtz-muenchen.de)

SUMMARY

All biological systems exhibit cell-to-cell variability. Frameworks exist for understanding how stochastic fluctuations and transient differences in cell state contribute to experimentally observable variations in cellular responses. However, current methods do not allow to identify the sources of variability between and within stable subpopulations of cells. We present a data-driven modeling framework for the analysis of populations comprising heterogeneous subpopulations. Our approach combines mixture modeling with frameworks for distribution approximation, facilitating the integration of multiple single-cell datasets and the detection of causal differences between and within subpopulations. The computational efficiency of our framework allows hundreds of competing hypotheses to be compared. We initially validate our method using simulated data with an understood ground truth, then we analyze data collected using quantitative single-cell microscopy of cultured sensory neurons involved in pain initiation. This approach allows us to quantify the relative contribution of neuronal subpopulations, culture conditions, and expression levels of signaling proteins to the observed cell-to-cell variability in NGF/TrkA-initiated Erk1/2 signaling.

INTRODUCTION

Cellular heterogeneity is a common phenomenon in biological processes (Elsasser, 1984; De Vargas Roditi and Claassen, 2015). Even isogenic cells of the same cell-type may respond differently to identical stimuli (Tay et al., 2010). This cellular heterogeneity is critical for cellular decision making and the formation of complex organisms (Balázsi et al., 2011). It is also a cause of failure in treatments of cancer, pain, and a wide range of common diseases (Willyard, 2016). Many studies have attempted to gain a deeper understanding of cell-to-cell variability (Rubin, 1990), and recently even a large-scale initiative was found to investigate this heterogeneity (Regev et al., 2017).

Experimentally, most common approaches use methods giving single-cell resolution, such as microscopy (Schroeder, 2011), flow and mass cytometry (Pyne et al., 2009), and single-cell RNA sequencing (Islam et al., 2014). These techniques yield increasing amounts of data, which are commonly analyzed using statistical techniques. Accordingly, a large number of powerful statistical methods have been developed for the analysis of single-cell data (see, e.g., Qiu et al. (2011); Kharchenko et al. (2014); Lun et al. (2017)). Unfortunately, these are unable to identify causalities and latent causes, or to reconstruct the governing equations of the process. Improved methods of data analysis are therefore required. We propose a model-based analysis framework for systems exhibiting cell-to-cell variability at different levels:

- differences between cell-types or cellular subpopulations, for example, caused by the cellular micro-environment (Ebinger et al., 2016) or stable epigenetic markers established during cell differentiation (Reik, 2007), and

- differences between cells of the same cell population that arise, for example, from differences in the cell state (Buettner et al., 2015) or from intrinsic stochastic fluctuations (Elowitz et al., 2002).

The differences on both levels can be caused by extrinsic or intrinsic noise (see definition by Elowitz et al., 2002).

In the case of homogeneous cell populations, the reaction rate equations (RREs) provide a description of the population behavior in the form of ordinary differential equations (ODEs) (Figure 1A). Stochastic fluctuations or latent differences between cells result in cell-to-cell variability and a distribution of cell states (Hasenauer et al., 2011; Zechner et al., 2012; Yao et al., 2016; Filippi et al., 2016) (Figure 1B). The statistical moments of this distribution are described by moment-closure approximation equations (Engblom, 2006) and system size expansions (van Kampen, 2007; Fröhlich et al., 2016). These methods provide scalable approximations for a range of processes in which variability arises from different sources. The approximation might be wrong, e.g. even negative variances might be predicted (Schnoerr et al., 2014). Additionally, they fail to provide an accurate description of the population heterogeneity when subpopulations are present and cannot be used to study the causal differences between cells and subpopulations.

To address parameter differences between cell population, we recently (Hasenauer et al., 2014) introduced a method that combines mixture modeling and mechanistic RRE modeling of the subpopulation means (Figure 1C). Cell-to-cell variability within a subpopulation is treated naively as an additional parameter that is to be estimated. Thus, the method assumes that the subpopulations are homogeneous and no mechanistic description of cell-to-cell variability within a subpopulation is possible. Moreover, the extant method can only be applied to one-dimensional measurements. When multivariate measurements are used, only marginal distributions can be analyzed and correlations between measurements are neglected, which may result in a substantial loss of information (Altschuler and Wu, 2010; Buchholz et al., 2013).

In this study, we introduced a non-trivial combination of mixture models that is able to capture subpopulation structures and models for individual subpopulations that account for differences between individual cells (Figure 1D). The approach therefore covers several levels of heterogeneity simultaneously (Figure 1A-D). This was not possible using the afore-mentioned approaches, which are all special cases of our model. The means and covariances of the observed species in each subpopulation are linked to a mixture distribution, allowing the entire cell population to be described and providing a mechanistic description of inter- and intra-subpopulation variability. We used the sigma-point approximation (van der Merwe, 2004), a scalable approach allowing for the analysis of large models, to capture the distribution of cell properties within a subpopulation. Similarly, our framework is able to exploit moment equations and system size expansion for the description of individual subpopulations. In contrast with previous work in (Hasenauer et al., 2014), the proposed framework can fully leverage the correlation information in multivariate data, rendering a better conditioned problem and improving identifiability.

We applied this framework to study signal transduction in the extracellular-signal regulated kinase (Erk) pathway, a signaling cascade that is involved in a range of biological processes. Our specific focus was on the pain sensitization signaling in highly heterogeneous primary sensory neurons in response to nerve growth factor (NGF) stimulation (Hucho and Levine, 2007; Ji et al., 2009; Andres et al., 2012). Our findings suggest that extracellular scaffolds, which provide important structural and biochemical cues to cells, play a crucial modulatory role in pain sensitization signaling and that several changes such as relative TrkA expression, Erk1/2 expression, but not subgroup composition is involved therein.

RESULTS

Mechanistic hierarchical population model for single-cell data

We considered populations comprising heterogeneous subpopulations. To allow coverage of multiple levels of heterogeneity, we linked a mixture distribution ϕ to a mechanistic model of the means and covariances of individual subpopulations. The distribution of the parameters, e.g., initial conditions or kinetic rates, produces a distribution of cell states and observables (Figure 2A-B). This distribution can be simulated using Monte Carlo methods by drawing parameters from the parameter distribution and simulating the single-cell model. Since this approach is computationally demanding, we approximated the distribution of parameters, states, and observables using finite mixture distributions. The components of the mixture describe the individual subpopulations.

Each cell j has cellular properties encoded in the parameter vector $\boldsymbol{\psi}^j$. In the hierarchical framework (Figure 2C), these parameters are considered to be drawn from a mixture distribution, as follows:

$$\boldsymbol{\psi}^j \sim \sum_s w_s N(\boldsymbol{\beta}_s, \mathbf{D}_s)\,,$$

with subpopulation weight w_s, mean $\boldsymbol{\beta}_s$ and covariance $\mathbf{D}_s$ for subpopulation $s = 1, \ldots N$. The subpopulation parameters $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \mathbf{D}_s)$ classify the variability of a property $\boldsymbol{\psi}^j$ as follows:

$$\beta_{s,i} = \begin{cases} \beta_i & \text{homogeneous} \\ \beta_i & \text{cell-to-cell variable} \\ \beta_{s,i} & \text{subpopulation variable} \\ \beta_{s,i} & \text{inter- and intra-subpopulation variable} \end{cases}$$

$$D_{s,ii} = \begin{cases} 0 & \text{homogeneous} \\ D_{ii} & \text{cell-to-cell variable} \\ 0 & \text{subpopulation variable} \\ D_{s,ii} & \text{inter- and intra-subpopulation variable} \end{cases}$$

allowing correlated parameters, $D_{s,ij} \neq 0$. The temporal evolution of the statistical properties of the cells of a subpopulation, including the mean and covariance, are computed using scalable methods. System size expansions and moment equations (van Kampen, 2007; Engblom, 2006) are used to describe stochastic single-cell dynamics, whereas sigma-points (van der Merwe, 2004) are used otherwise. These approaches yield an ODE model of the statistical moments, comprising the means and covariances $\mathbf{z}_s = (\mathbf{m}_s, \mathbf{C}_s)$ of species $\mathbf{x}$. The model is simulated for each of the N subpopulations

$$\dot{\mathbf{z}}_s = g(\mathbf{z}_s, \boldsymbol{\xi}_s, u),\ \mathbf{z}_s(0) \ = \ \mathbf{z}_0(\boldsymbol{\xi}_s, u) \tag{1}$$

with initial conditions $\mathbf{z}_0$ and experimental condition u. The moments of the species in a subpopulation are then mapped to the distribution parameters $\phi_s = h(\mathbf{z}_s, \boldsymbol{\xi}_s, u)$ of the distribution ϕ, including measurement noise $\boldsymbol{\Gamma}$ which is assumed to be the same for all subpopulations. The observables, the quantities of the biological system that can be measured experimentally, are assumed to have the distribution

$$\bar{\mathbf{y}} \sim \sum_s w_s \phi(\boldsymbol{\varphi}_s) \tag{2}$$

at the population level. In this study, we used mixtures of multivariate log-normal distributions, yielding $\boldsymbol{\varphi}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. The sigma-point approximation (detailed in STAR Methods) provides time-dependent moments of the system defined in (1) and accounts for cell-to-cell variability. When combined with subpopulation variability, this yields both the inter- and intra-subpopulation variability. For a comparison of our approach to existing methods, we refer to STAR Methods**.**

Parameter estimation and model selection

The parameters of biochemical processes, the sources of cell-to-cell and subpopulation variability, and the precise network structure are in general unknown. We therefore calibrated the hierarchical population model using single-cell snapshot data $\bar{\mathbf{y}}^{e,k,j}$ with cell j measured at time point t_k under experimental condition u_e, for example, representing a drug dosage. The parameters $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ usually comprise characteristics of a subpopulation (e.g., the means and covariances of the parameter distributions), subpopulation sizes and measurement noise. Maximum likelihood estimation was used to derive these parameters from the data. The maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ was obtained by solving the following optimization problem:

$$\begin{aligned}
&\max_{\boldsymbol{\theta}\in\Theta}\left\{\mathcal{L}(\boldsymbol{\theta}) = \prod_{e,k,j}\sum_{s} w_s^e(t_k,\boldsymbol{\theta})\,\phi\left(\bar{\mathbf{y}}^{e,k,j}|\boldsymbol{\varphi}_s^e((t_k,\boldsymbol{\theta},u_e)\right)\right\}\\
&\text{subject to } \dot{\mathbf{z}}_s^e = g(\mathbf{z}_s^e, \boldsymbol{\xi}_s, u_e), \qquad \mathbf{z}_s^e(0) \ = \ \mathbf{z}_0^e(\boldsymbol{\xi}_s(\boldsymbol{\theta}), u_e)\\
&\qquad\qquad\quad \boldsymbol{\varphi}_s^e \ = \ h(\mathbf{z}_s^e, \boldsymbol{\xi}_s(\boldsymbol{\theta}), u_e).
\end{aligned}$$

The likelihood function $\mathcal{L}$ incorporates all cells, time points, and experimental conditions. For efficient parameter estimation, we performed multi-start local optimization with a robust evaluation scheme for the log-likelihood function and its gradient. The gradient of the log-likelihood function with respect to the parameters was computed using forward sensitivity analysis (see (Loos et al., 2016) and STAR Methods). The practical identifiability and uncertainty of the parameter estimates was evaluated using profile likelihoods as well as sampling methods. For parameter sampling we employed an adaptive parallel tempering method.

To infer the subpopulation structure, the difference between subpopulations, the variability within subpopulations, and the influence of the experimental condition, a collection of hierarchical models is formulated. We compare these models and the corresponding hypotheses using the Bayesian Information Criterion (BIC) (Raftery, 1999). The BIC provides a computationally relatively inexpensive approximation to the Bayes factors, which gives the favor of a model over another. To justify the use of the BIC, we compared the results with those obtained by using (i) Bayes factors computed using thermodynamic integration (Hug et al., 2016) and (ii) log pointwise posterior predictive densities (Gelman et al., 2014).

The subpopulation structures and parameters inferred with the hierarchical population models were subsequently used as prior information for the calibration of the single-cell models. The regularization provided by the prior allows the prediction of single-cell trajectories although the dataset for each individual cell is scarce. These can then be used to predict individual single-cell trajectories (see STAR Methods for more details).

The hierarchical models were implemented in the MATLAB toolbox, incorporating efficient simulations for the individual subpopulations. While any simulation that provides means and covariances of the subpopulations can be employed, in this study, we used the sigma-point approximation. This approach accounts for cell-to-cell variability, which is manifested in the parameters (see STAR Methods for more details).

Unraveling sources of heterogeneity

To demonstrate the advantages of the hierarchical population model, which incorporates a mechanistic description of the means and variances, over the method proposed by Hasenauer et al. (2014), we applied our approach to simulated data on a simple conversion process. Such conversions are common in biological systems, for example, in phosphorylation. The conversion process comprised two species A and B, with cell-to-cell variable conversions from B to A (Figure 3A), corresponding to different levels of phosphatase in the cells. Two subpopulations were assumed with different responses to stimulus u. This produced subpopulations with different rates of stimulus-dependent conversion from A to B. Artificial measurement noise was added to allow the capability of the framework to distinguish measurement noise from biological variability to be assessed. We assumed the underlying subpopulation structure, i.e., the subpopulation variability of k_1, to be known (detailed in STAR Methods).

The simulated data were analyzed using (i) the approach introduced in Hasenauer et al. (2014) which describes the subpopulations using RREs and (ii) the proposed approach using hierarchical single-cell analysis. The first approach does not model the temporal evolution of the variance, requiring different parameterizations to be compared, i.e., constant, time-dependent, and time/subpopulation-dependent variability. Model selection with the BIC indicates that different parameters for each subpopulation at every time point are required to be used to describe the data (Figure 3B). A full Bayesian analysis using the model evidence as well as the log pointwise predictive density justified the use of the computationally less expensive BIC (see Figure S1 and STAR Methods). This demonstrates that the observed cell-to-cell variability changes over time but provides no information about the sources of the observed cell-to-cell variability.

The mechanistic modeling of multiple levels of heterogeneity facilitates the prediction of its causal source via model selection. We considered a range of hypotheses and performed model selection using BIC (Figure 3B). Given the subpopulation structure, the additional source of heterogeneity, namely, the conversion from B to A, was correctly predicted using the BIC and the corresponding model provided a good fit to the data (Figure 3C). The BICs for most of the hierarchical models were substantially lower than that of the best model that incorporates only the mean. This confirms that a mechanistic description of the variability is more appropriate.

We analyzed the ability of the hierarchical model to predict the different contributions of cell-to-cell variability and measurement noise, as both are normally present in single-cell experiments. The uncertainty analysis suggested that the hierarchical modeling approach was able to distinguish between the two (Figure 3D).

To evaluate the predictive power of the method for single-cell trajectories, we inferred the parameter of individual cells from the single data point available for each cell in combination with the calibrated hierarchical population model as a prior. We found that the information about the behavior of a single- cell encoded in the measurement of the first time point was limited (Figure 3E), e.g., the prediction is off. However, using data from late time points, we obtained an good estimate of the (latent) single-cell trajectory (Figure 3F). The prediction of the trajectories for 100 single-cells from measurements at time point $t = 120$ min (Figure 3G) reveals a correlation between true and predicted values > 0.9 for all but early time points.

This example shows how the hierarchical population model outperforms the variants of models presented in Hasenauer et al. (2014). We confirmed the power of the proposed approach by studying a model of stochastic gene expression (Figure S6) and comparing the approach to the method by Zechner et al. (2012) (see STAR Methods). Our model employs a mechanistic description of the variability, thereby enabling a more detailed insight into the heterogeneity of the population and reducing the number of parameters that need to be estimated from the data.

Identification of differential protein expression

Many single-cell technologies provide multivariate measurements and therefore convey information about the correlations between the observables. To incorporate this, we extended our hierarchical modeling framework to multivariate data and demonstrated its capability to reconstruct the differential protein expression of cellular subpopulations (Sauvageau et al., 1994; Kharchenko et al., 2014) using simulated data. We considered a model describing the abundance of two proteins, the expression of which is regulated by stimulus u (Figure 4A). The influence of u varies between cell populations and is therefore able to capture, e.g., different levels of membrane receptors. We generated multivariate data by simulating a single-cell model (see STAR Methods for more details).

An analysis using our hierarchical approach confirmed the ability of the proposed model to reproduce the data (Figure 4B) and to provide reliable parameter estimates (Figure 4C). Such multivariate data cannot be exploited by the existing model-based approaches. When the temporal evolution of proteins is measured individually, the correlation information is missing and a symmetry arises in the system (Figure 4D). This is reflected in the multimodal profiles of the parameters $\lambda_{\mathrm{B},1}$ and $\lambda_{\mathrm{B},2}$, indicating a lack of practical identifiability.

Our framework exploits the correlation structures of multivariate data, which in this simulation example allowed us to conclude that each subpopulation had a high expression of only a single protein. This only becomes possible when the correlations are analyzed.

Modeling signal transduction in sensory neurons

We applied the hierarchical modeling approach to investigate the sources of variability of NGF-induced Erk1/2 activation in cultures of adult sensory neurons (Figure 5A). This was done by monitoring the rates of NGF-mediated Erk1/2 phosphorylation in dissociated cultures of the primary sensory neurons of rat dorsal root ganglia. Primary sensory neurons form a heterogeneous population, from which, upon NGF stimulation, a subpopulation reacts with a graded Erk1/2 phosphorylation response. Previous models have attempted to approximate this by assuming the existence of responders and non-responders with differing levels of the NGF receptor TrkA (Hasenauer et al., 2014). In the current study, we refined this substantially by modeling the overall population using two heterogeneous subpopulations that differed in their average response. To calibrate this refined model, we collected quantitative single-cell microscopy data on NGF-induced Erk1/2 phosphorylation kinetics and dose response curves using immunofluorescence labeling of pErk1/2 alone, co-labeled with Erk1/2 and TrkA antibodies (see STAR Methods for more details). Our analysis used the ODE model introduced in (Hasenauer et al., 2014). This has six structurally identifiable parameters $k_1, k_2, k_4, k_5, k_3[\mathrm{TrkA}]_0$ and $c[\mathrm{Erk}]_0$.

Causal differences between subpopulations of cultured sensory neurons

In this test case, the ultimate goal of our modeling is to provide a mechanistic explanation for why a subpopulation of cultured neuron reacts to NGF stimulation with a graded Erk1/2 phosphorylation response (phosporylated Erk is active). Erk1/2 is activated by TrkA and differences between the responses of responders and non-responders are likely caused by variation in TrkA levels. We first validated our modeling approach by predicting causal differences between subpopulations and its accordance with described differences in TrkA expression. We used experimental kinetic and dose response data from sensory neurons cultured on the adherence substrate poly-D-lysine (PDL). We fitted 64 models with up to 33 parameters, accounting for all combinations of the six potential differences between subpopulations, which was only feasible due to the computational efficiency of our approach. Our assessment of the importance of individual differences between the subpopulations using a BIC-based ranking scheme suggested that cellular TrkA activity ($k_3[\mathrm{TrkA}]_0$) made the greatest contribution (Figure 5B). This was indicated by a high BIC weight, which captures differences by Bayesian model averaging (see STAR Methods for more details), and the substantially better mean rank of the models using differences in cellular TrkA activity compared with those using other differences. The additional subpopulation variability of TrkA expression levels was also confirmed experimentally in the cultures (Figure 5D) and use of this difference alone produced an excellent fit to the experimental data (Figures 5C and S4). The following potential differences are the relative Erk1/2 expression levels ($c[\mathrm{Erk}]_0$) and the dephosphorylation rate (k_5). However, our experimental data showed no statistically significant difference in total Erk1/2 levels between responders and non-responders (Figure 5E). To assess the relevance of the dephosphorylation rate and thus the corresponding phosphatase activity we performed experiments in which we monitored the pErk1/2 decline dynamics after inhibiting the mitogen-activated protein kinase (Mek) that phosphorylates Erk1/2. If the phosphatase activity does vary, we would expect to observe different equilibration dynamics. However, this could not be confirmed (Figures 5F and S3).

This demonstrates that the hierarchical approach using experimental data provided an appropriate ranking of differences which could be demonstrated experimentally and is in line with literature (reviewed in. e.g., Mantyh et al. (2011)).

Influence of extracellular scaffolds on sensitization signaling

As second test of our approach, we systematically varied the extracellular environment and asked whether our modeling approach could generate mechanistic hypothesis to explain the altered cellular responses we observed. Specifically, we characterized NGF-stimulated signaling when neurons were either grown on collagen type I (Col I), a classical extracellular matrix protein that forms receptor-matrix interactions, or on poly-D-lysine (PDL), an organic molecule that promotes cell adherence by electrostatic interaction. We determined the kinetics and dose response curves of NGF-induced Erk1/2 phosphorylation in sensory neurons cultured overnight on Col I or PDL (see STAR Methods for more details). We found that the mean Erk1/2 activation was approximately 17% higher in Col I compared to PDL after NGF treatment (Figure 6A for pErk1/2 dose responses and Figure S5A for the other datasets). In addition to showing increased NGF-induced Erk1/2 activation, the number of cells was observed to be 1.5 times lower in the collagen cultures than in the poly-D-lysine cultures. These observations raised questions about the source of the measured increase in mean NGF-mediated Erk1/2 activation. We considered two hypotheses: (i) the increase results from a biological action of the different scaffolds onto the neurons and (ii) the increase reflects a shift of the subpopulation sizes arising from a nonrandom loss of parts of the high-responder subpopulation due to reduced cell adherence in the collagen cultures. To unravel the causal differences between the primary sensory neurons cultured on PDL and on Col I, we applied 128 hierarchical models with up to 36 parameters, using the previously derived subpopulation structure. These models considered all combinations of differences between the cell population on different scaffolds, including the size of subpopulations. The model for each adherence substrate accounted for the cell-to-cell variability of Erk1/2 and the inter- and intra-subpopulation variability of cellular TrkA activity. The model ranked first by the BIC (Figures 6B) gave a good fit to the data and suggested differences not only in cellular TrkA activity ($k_3[\mathrm{TrkA}]_0$) but also in Erk1/2 expression ($c[\mathrm{Erk}]_0$), and Erk1/2 dephoshorylation (k_5)) (Figures 6C-D and S5B-D). These differences were assumed to explain the higher response on Col I, and therefore supported hypothesis (i). The model that assumed no difference between the extracellular scaffolds (rank 128) or changes only in the relative size of the subpopulations (rank 127) performed worst, indicating that hypothesis (ii) failed to explain the data. Indeed, the differences in relative TrkA and Erk1/2 expression levels predicted by the models with the highest ranked could be confirmed (Figures 6D-E). These results confirmed the model-based analysis and suggested an impact of the classical extracellular matrix protein collagen I on protein expression.

DISCUSSION

Elucidating the causes of cellular heterogeneity is a challenging task in systems biology and requires appropriate mechanistic models for use with single-cell data. In this study, we introduced a hierarchical modeling framework that allowed different levels of heterogeneity to be investigated, including subpopulation structures and cell-to-cell variability within subpopulations. It also provides mechanistic insights. Beyond cell-to-cell variability, the method accounts for measurement noise and is able to deconvolute these sources.

This modeling approach unifies available mechanistic modeling and inference frameworks (Zechner et al., 2012; Hasenauer et al., 2014), complements available statistical methods, and exploits efficient simulation methods for cellular subpopulations. We focused on the cell-to-cell variability encoded in parameter values (Koeppl et al., 2012) and used sigma-point approximation to determine the subpopulation means and variances. To address variability arising from stochastic fluctuations, moment equations (Figure S6) and other methods, including the system size expansion (Fröhlich et al., 2016), can be used. The proposed method facilitates the integration and simultaneous analysis of multiple datasets, without requiring complex pre-processing of the data (Lee et al., 2011). The modeling approach is implemented in the open-source MATLAB Toolbox ODE-MM which is available on GitHub and ready to be reused by the community.

Procedures such as a forward-backward algorithm (e.g., Hastie et al. (2009)) or reversible jump Markov Chain Monte Carlo (Green, 1995) could be implemented to perform parameter estimation and model selection simultaneously. An alternative approach to obtain the model evidence would be the use of sequential Monte Carlo methods, as, e.g., done by Filippi et al., 2016. In this study, mixtures of log-normal distributions were used to model the cell population. However, other distributions, including the Laplace distribution, could be integrated with the computational framework to improve robustness against outliers (Maier et al., 2017).

The inference of mechanistic models from single-cell data relies on statistical models for the measurement and sampling process. In many modeling studies using single-cell data, no distinction is made between cells from different batches, obscuring cell-to-cell variability and differences between experimental batches (Hicks et al., 2015). In this study, we observed that the derived likelihood function can be overly sensitive and that model selection is biased towards complex models. To circumvent this issue, we used a ranking of potential differences rather than a precise measure of statistical significance. However, this problem will need to be addressed, as the use of single-cell data is increasingly common.

In summary, we proposed the use of hierarchical population models as a novel tool to study heterogeneity in multivariate single-cell data and evaluated their performance. Our framework is the first to account for multiple levels of heterogeneity simultaneously. Our results on simulation and application examples suggest that this method can be used to obtain a more holistic understanding of heterogeneity.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT
    - Antibodies
    - Reagents
    - Animals
    - Coating
    - Primary sensory neuron culture
    - Stimulation and fixation of neuronal culturs
    - Immunocytochemistry
    - Quantitative microscopy
- METHOD DETAILS
    - Models for individual subpopulations
    - Mechanistic hierarchical population model
    - Comparison with existing models
    - Parameter estimation
    - Calibration of single-cell model
    - Conversion process
    - Differential protein expression
    - NGF-induced Erk1/2 signaling
    - Accounting for intrinsic noise
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

ACKNOWLEDGEMENTS

C.L., F.F., and J.H. acknowledge financial support from the German Research Foundation (DFG) and the Postdoctoral Fellowship Program of the Helmholtz Zentrum Mu ̈nchen. K.M. was supported by the Evangelisches Studienwerk Villigst and the graduate program in Pharmacology and Experimental Thera- peutics at the University of Cologne which is financially and scientifically supported by Bayer. This work was supported by the German Research Foundation (DFG; http://www.dfg.de) through the Graduate School of Quantitative Biosciences Munich (QBM; FF).

AUTHOR CONTRIBUTIONS

C.L., F.F., and J.H. developed the method. K.M. and T.H. designed the experiments. K.M. performed the experiments. C.L. analyzed the data. C.L., K.M., F.F., T.H., and J.H. wrote the paper.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Altschuler, S.J., and Wu, L.F. (2010). Cellular heterogeneity: when do differences make a difference? Cell, 141 (4):559–563.

Andres, C., Meyer, S., Dina, O.A., Levine, J.D., and Hucho, T. (2010). Quantitative automated microscopy (QuAM) elucidates growth factor specific signalling in pain sensitization. Molecular Pain, 6(98):1–16.

Andres, C., Hasenauer, J., Allgöwer, F., and Hucho, T. (2012). Threshold-free population analysis identifies larger DRG neurons to respond stronger to NGF stimulation. PLoS ONE, 7(3):e34257

Balázsi, G., van Oudenaarden, A., and Collins, J.J. (2011). Cellular decision making and biological noise: from microbes to mammals. Cell, 144(6):910–925.

Buchholz, V.R., Flossdorf, M., Hensel, I., Kretschmer, L., Weissbrich, B. , Gräf, P., Verschoor, A., Schiemann, M., Höfer, T., and Busch, D.H. (2013). Disparate individual fates compose robust CD8+ T cell immunity. Science, 340(6132):630–635.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA- sequencing data reveals hidden subpopulations of cells. Nat. Biotechnol., 33(2):155–160.

Chen, Z.L., Yu, W.M., and Strickland, S. (2007). Peripheral regeneration. Annu. Rev. Neurosci., 30:209–233.

De Vargas Roditi, L., and Claassen, M. (2015). Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. Curr. Opin. Biotechnol., 34:9–15.

Ebinger, S., Özdemir, E.Z., Ziegenhain, C., Tiedt, S., Alves, C.C., Grunert, M., Dworzak, M., Lutz, C., Turati, V.A., Enver, T. et al. (2016). Characterization of rare, dormant, and therapy-resistant cells in acute lymphoblastic leukemia. Cancer Cell, 30(6):849–862.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science, 297(5584):1183–1186.

Elsasser, W.M. (1984). Outline of a theory of cellular heterogeneity. Proc. Natl. Acad. Sci. U S A, 81(16): 5126–5129.

Engblom, S. (2006). Computing the moments of high dimensional solutions of the master equation. Appl. Math. Comp., 180:498–515.

Filippi, S., Barnes, C.P., Kirk, P.D., Kudo, T., Kunida, K., McMahon, S.S., Tsuchiya, T., Wada, T., Kuroda, S., and Stumpf, M.P. (2016). Robustness of MEK-ERK dynamics and origins of cell-to-cell variability in MAPK signaling. Cell Reports, 15(11):2524–2535.

Fröhlich, F., Thomas, P., Kazeroonian, A., Theis, F.J., Grima, R., and Hasenauer, J. (2016). Inference for stochastic chemical kinetics using moment equations and system size expansion. PLoS Comput. Biol., 12(7): e1005030.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. Statistics and Computing, 24(6):997–1016.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In Bayesian Statistics, Bernardo, J.M., Smith, A.F.M., Dawid, A.P., and Berger, J.O., ed. (University Press, Oxford, UK), pp. 169–193.

Gillespie, D.T. (1977). Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem., 81(25): 2340–2361.

Green, P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 82(4):711–732.

Hasenauer, J., Waldherr, S., Doszczak, M., Radde, N., Scheurich, P., and Allgöwer, F. (2011). Identification of models of heterogeneous cell populations from population snapshot data. BMC Bioinformatics, 12(125).

Hasenauer, J., Hasenauer, C., Hucho, T., and Theis, F.J. (2014). ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. PLoS Comput. Biol., 10(7):e1003686.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning, volume 2. Springer.

Hicks, S.C., Teng, M., and Irizarry, R.A. (2015). On the widespread and critical impact of systematic bias and batch effects in single-cell RNA-seq data. bioRxiv 025528.

Hross S., and Hasenauer, J. (2016). Analysis of CFSE time-series data using division-, age- and label-structured population models. Bioinformatics, 32(15):2321–2329.

Hucho T., and Levine, J.D. (2007). Signaling pathways in sensitization: toward a nociceptor cell biology. Neuron, 55(3):365–376.

Hug, S., Schwarzfischer, M., Hasenauer, J., Marr, C., and Theis, F.J. (2016). An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule. Stat. Comput., 26(3): 663–677.

Isensee, J., Diskar, M., Waldherr, S., Buschow, R., Hasenauer, J., Prinz, A., Allgöwer, F., Herberg, F.W., and Hucho, T. (2014). Pain modulators regulate the dynamics of PKA-RII phosphorylation in subgroups of sensory neurons. J. Cell Sci., 127:216–229.

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lonnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nat. Methods, 11(2):163–166.

Ji, R.R., Gereau, R.W., Malcangio, M., and Strichartz, G.R. (2009). MAP kinase and pain. Brain Res Rev., 60 (1):135–148.

Kass, R.E., and Raftery, A.E. (1995). Bayes factors. J. Am. Stat. Assoc., 90(430):773–795.

Kazeroonian, A., Fröhlich, F., Raue, A., Theis, F.J., and Hasenauer, J. (2016). CERENA: ChEmical REaction Network Analyzer–A toolbox for the simulation and analysis of stochastic chemical kinetics. PLoS ONE, 11(1):e0146732.

Kharchenko, P.V., Silberstein, L., and Scadden, D.T. (2014). Bayesian approach to single-cell differential expression analysis. Nat. Methods, 11(7):740–742.

Koeppl, H., Zechner, C., Ganguly, A., Pelet, S., and Peter, M. (2012). Accounting for extrinsic variability in the estimation of stochastic rate constants. Int. J. Robust Nonlinear Control, 22(10):1–21.

Lee, C.H., Kim, K.H., and Kim, P. (2009). A moment closure method for stochastic reaction networks. J. Chem. Phys., 130(13):134107.

Lee, G., Finn, W., and Scott, C. (2011). Statistical file matching of flow cytometry data. J. Biomed. Inform., 44 (4):663–676.

Loos, C., Fiedler, A., and Hasenauer, J. (2016). Parameter estimation for reaction rate equation constrained mixture models. In Proc. 13th Int. Conf. Comp. Meth. Syst. Biol., Lecture Notes in Bioinformatics, Bartocci, E., Lio, P., and Paoletti, N., ed. (Springer International Publishing) pp. 186–200.

Lun, X.K., Zanotelli, V.R., Wade, J.D., Schapiro, D., Tognetti, M., Dobberstein, N., and Bodenmiller, B. (2017). Influence of node abundance on signaling network state and dynamics analyzed by mass cytometry. Nature Biotechnology, 35(2):164–172.

Maier, C., Loos, C., and Hasenauer, J. (2017). Robust parameter estimation for dynamical systems from outlier-corrupted data. Bioinformatics, 33(5):718–725.

Mantyh, P.W., Koltzenburg, M., Mendell, L.M., Tive, L., and Shelton, D.L. (2011). Antagonism of nerve growth factor-TrkA signaling and the relief of pain. The Journal of the American Society of Anesthesiologists, 115(1):189–204.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T., Maier, L., Baecher-Allan, C., McLachlan, G., Tamayo, P., Hafler, D., De Jager, P., and Mesirov, J. (2009). Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. U S A, 106(21):8519–8124.

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs Jr, K.D., Bruggner, R.V., Linderman, M.D. Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with spade. Nature Biotechnology, 29(10):886–891.

Raftery, A.E. (1999). Bayes factors and BIC. Socio. Meth. Res., 27(3):411–417.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics, 25(25):1923–1929.

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B.D., Theis, F.J., Klingmüller, U., and Timmer J. (2013). Lessons learned from quantitative dynamical modeling in systems biology. PLoS One, 8(9):e74335, Sept. 2013.

Regev, A., Teichmann, S., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M. et al. (2017). The Human Cell Atlas. bioRxiv 121202.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature, 447 (7143):425–432, 2007

Roederer, M. (2002). Compensation in flow cytometry. Curr. Protoc. Cytom., pp. 1–14.

Rubin, H. (1990). The significance of biological heterogeneity. Cancer Metastasis Rev., 9(1):1–20.

Sauvageau, G., Lansdorp, P.M., Eaves, C.J., Hogge, D.E., Dragowska, W.H., Reid, D.S. Largman, C., Lawrence, H.J., and Humphries, R.K. (1994). Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells. Proc. Natl. Acad. Sci. U S A, 91(25):12223–12227.

Schnoerr, D., Sanguinetti, G., and Grima, R. (2014). Validity conditions for moment closure approximations in stochastic chemical kinetics. J. Chem. Phys., 141(8):084103.

Schroeder, T. (2011). Long-term single-cell imaging of mammalian stem cells. Nat. Methods, 8(4):30–35.

Stapor, P., Weindl, D., Ballnus, B., Hug, S., Loos, C., Fiedler, A., Krause, S., Hross, S., Fröhlich, F., and J. Hasenauer, J. (2018). PESTO: Parameter EStimation TOolbox. Bioinformatics, 34(4):705–707.

Tay, S., Hughey, J.J., Lee, T.K., Lipniacki, T., Quake, S.R., and Covert, M.W. (2010). Single-cell NF-κB dynamics reveal digital activation and analogue information processing. Nature, 466:267–271.

van der Merwe, R. (2004). Sigma-point Kalman filters for probabilistic inference in dynamic state-space models. Ph.D. thesis, Oregon Health & Science University.

van Kampen, N.G. (2007). Stochastic processes in physics and chemistry. 3rd edition. (North-Holland, Amsterdam).

Williams, P.M. (1999). Matrix logarithm parametrizations for neural network covariance models. Neural Netw., 12(2):299–308.

Willyard, C. (2016). Cancer therapy: an evolved approach. Nature, 532:166–168, 2016 J. Yao, A. Pilko, and R. Wollman. Distinct cellular states determine calcium signaling response. Mol. Syst.

Biol., 12(12):894.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., and Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. Proc. Natl. Acad. Sci. U S A, 109(21):8340–8345.

STAR METHODS

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for software and algorithms should be directed to the Lead Contact Jan Hasenauer (jan.hasenauer@helmholtz-muenchen.de).

EXPERIMENTAL MODEL AND SUBJECT

Antibodies

The following antibodies were used in this study: chicken polyclonal antibody against UCHL1 (1:4000; Novus, #NB110-58872), mouse monoclonal antibody against UCHL1 (1:1000, MorphoSys, #7863-2004), rabbit monoclonal antibody against phospho-Erk1/2 (1:250, Cell Signaling, #4370L), mouse monoclonal antibody against ERK1/2 (1:500, Cell Signaling, cat#4696 S), goat polyclonal antibody against TrkA (1:500, R&D Systems, #AF1056), and highly cross adsorbed Alexa Fluor 488, Alexa Fluor 568, Alexa Fluor 594-, and Alexa Fluor 488-conjugated secondary antibodies (Invitrogen).

Reagents

NGF (50 μg/ml in 0.1% BSA), GDNF (20 μg/ml in 0.1% BSA), U0126 (50 mM in DMSO) were purchased from Alomone labs (#N-240), PeproTech (cat#450-51), and Calbiochem (#662005), respectively, and were prepared as indicated. The concentrations used are indicated in the text or figure legends. Collagen type I (Cell Systems, #5056-A) and poly-D-lysine (Sigma, #P6407-5MG) were diluted in 1xPBS to final concentrations of 3.4 μg/ml and 10 μg/ml.

Animals

Male Sprague Dawley rats (200250 g, 8-10 weeks old) were obtained from Harlan Laboratories.

All experiments were performed in accordance with the German animal welfare law with permission of the District Government for Nature and Environment, NRW (LANUV NRW, license 84-02.05.20.13.045). Rats were sacrificed by CO_2 intoxication for tissue isolation.

Coating

96-well imaging plates (Greiner) were coated with 50 μl volume of matrix protein dilutions per well for 3 h at 37 C Wells were washed one time with 1xPBS for 10 min PDL coatings were dried and washing solution of Col I treated wells was removed immediately before cell seeding.

Primary sensory neuron culture

L1-L6 dorsal root ganglia (DRG) were isolated, desheathed, pooled and incubated in Neurobasal-A (NB) medium supplemented with collagenase P for for 1 h in 5% CO_2 atmosphere at 37°C. Neurons were dissociated by trituration with fire-polished siliconated Pasteur pipettes and axonal debris and disrupted cells were removed by a 14% BSA gradient centrifugation (120 g, 8 min). Cells were resuspended in NB medium supplemented with B27 medium, L-Glutamine, L-Glutamate and Penicillin-Streptomycin. Subsequently, they were plated on pre-coated 96-well imaging plates and incubated overnight in a 5% CO_2 atmosphere at 37°C.

Stimulation and fication of neuronal cultures

Neuronal cultures were stimulated 15 h after isolation by removal of 50 μl culture medium, mixing with the compound and returning to the corresponding culture well. Solvent controls were treated alike. Stimulation was performed with automated eight-channel pipettes (Eppendorf) on pre-warmed heating blocks (37°C), and stimulated cells were placed back into the incubator. Neurons were fixed by adding 8% PFA (final concentration 4% PFA) for 10 min at RT and subsequently washed three times with 1xPBS for 10 min. Kinetic experiments involved time courses of 0, 1, 5, 15, 30, 60 and 120 min NGF stimulation (20 ng/ml), whereas dose response curves were obtained by NGF stimulations with the following NGF concentrations for 1 h: 0.16, 0.8, 4, 20, 100, 500 ng/ml.

Immunocytochemistry

Cells were blocked and permeabilized with 2% normal goat serum or 2% normal donkey serum supplemented with 1% BSA, 0.1% Triton X-100, 0.05% Tween 20 for 1 h at RT. Primary antibodies were added in 1% BSA in 1xPBS and cells were incubated overnight at 4°C. After three washes with 1xPBS for 10 min at RT, cells were incubated with secondary antibodies diluted in 1xPBS for 1 h at RT. Plates were stored at 4°C after three additional washes with

1xPBS (10 min, RT) until scanning.

Quantitative microscopy

Immunofluorescently labelled neurons were imaged via the Cellomics ArrayScan microscope using a 10x objective as described previously (Isensee et al., 2014). Images of 512 x 512 pixels were analyzed using the Cellomics software package. Briefly, images of all channels were background corrected (low pass filter), objects were identified using fixed thresholding (intensity 900) and segmentation by shape (parameter 15). Neurons were validated by the following object selection parameters: size: 1657500 μm^2; circularity (perimeter^2/4π area): 12; length-to-width ratio: 12.67; average intensity: 90012.000; total intensity: 2×10^5 to 5×10^7. The image masks were then used to quantify signals in other channels. Raw values of three to four independent experiments were further processed via the R software. Raw fluorescence data was compensated and normalized. In brief, three controls were prepared for a triple staining: 1. UCHL1 alone, 2. UCHL1 + antibody 1, and 3. UCHL1 + antibody 2. Raw fluorescence data of the controls were used to calculate the bleed-through between fluorescence channels. The slope of best fit straight lines were determined by linear regression and used to compensate bleed through as described previously (Roederer, 2002). Compensated data were scaled to a mean value of 1000 for the unstimulated cells of the poly-D-lysine control to adjust for variability between experimental days.

METHOD DETAILS

Models for individual subpopulations

The hierarchical modeling approach introduced in this manuscript describes the population dynamics based on the dynamics of individual subpopulations. In this section, we introduce the modeling approaches at the subpopulation level that are used in our study.

First, we considered the simple case that only the mean of a subpopulation is modeled mechanistically, whereas the variance and higher order moments are not linked to the underlying biochemical reaction network. For this, the reaction rate equation (RRE) was used

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= f_x(\mathbf{x}, \boldsymbol{\psi}, u), \qquad \mathbf{x}(0) = \mathbf{x}_0(\boldsymbol{\psi}), \\ \mathbf{y} &= f_y(\mathbf{x}, \boldsymbol{\psi}, u). \end{aligned} \tag{3}$$

Here, $\mathbf{x} = (x_1, \ldots, x_n)^{\boldsymbol{T}}$ denotes the biochemical species, $\mathbf{y} = (y_1, \ldots, y_d)^{\boldsymbol{T}}$ the observables of the system, and $\boldsymbol{\psi}$ the parameters, such as reaction rates, protein abundances, or initial conditions. This follows the method introduced in Hasenauer et al., 2014.

The RRE is based on the assumption that the subpopulations are homogeneous. However, many cellular processes exhibit substantial intrinsic or extrinsic cell-to-cell variability. To account for this variability, we considered models accounting for random parameters and stochastic reaction kinetics.

Sigma-point approximation In this study, we modeled extrinsic variability by heterogeneity in L parameters of the parameter vector $\boldsymbol{\psi} \in \mathbb{R}^{n_\psi}$ of individual cells. The parameters $\boldsymbol{\psi}$ were assumed to follow a probability distribution $p_\psi(\boldsymbol{\psi})$. This distribution in the parameters $p_\psi(\boldsymbol{\psi})$ is mapped to a distribution of cell states and observables of the subpopulation, which need to be computed for the parameter estimation. A detailed analysis of this image requires sampling from $p_\psi(\boldsymbol{\psi})$ and subsequent evaluation of the state and observable vectors by simulation. This procedure is, however, computationally demanding. We employed the sigma-point approximation (van der Merwe, 2004) to obtain an approximation of the statistical moments of the image, mean and covariance and their dynamics in time, using a small number of simulations. The sigma-point approximation uses only the image of $2L + 1$ deterministically chosen parameter vectors. These parameter vectors, the so called sigma-points, are chosen to represent the mean $\boldsymbol{\beta}$ and the covariance $\mathbf{D}$ of p_ψ. For the parameters that were considered to be homogeneous, i.e., not variable across the cells, it was assumed that $\beta_i = \psi_i$ and $D_{ii} = D_{ij} = 0, \forall j$.

Following van der Merwe (2004), the sigma-points $\{\nu_l, \boldsymbol{\mathcal{S}}_l\}$ are defined as

$$\begin{aligned}
&\boldsymbol{\mathcal{S}}_0 = \boldsymbol{\beta}, && \nu_0^{(m)} = \frac{\eta_3}{L+\eta_3}, && \text{for } l = 0\\
&\boldsymbol{\mathcal{S}}_l = \boldsymbol{\beta} + \left(\sqrt{(L+\eta_4)\mathbf{D}}\right)_{\boldsymbol{l}}, && \nu_l^{(c)} = \frac{\eta_3}{L+\eta_3} + 1 - \eta_1^2 + \eta_2, && \text{for } l = 1,\dots,L\\
&\boldsymbol{\mathcal{S}}_l = \boldsymbol{\beta} - \left(\sqrt{(L+\eta_4)\mathbf{D}}\right)_{\boldsymbol{l}}, && \nu_l^{(m)} = \nu_l^{(c)} = \frac{1}{2(L+\eta_3)}, && \text{for } l = L+1,\dots,2L.
\end{aligned}$$

We used $\eta_2 = 2$ and $\eta_3 = \eta_1^2(L + \eta_4) - L$, with $\eta_1 = 0.7$ and $\eta_4 = 0$ as proposed by van der Merwe (2004). The superscripts for ν_l indicate whether it is used for the calculation of the mean $^{(m)}$ or the covariance $^{(c)}$.

For the examples and applications presented in the manuscript, we assumed that the variability between cells is completely explained by differences in the model parameters. For a set of given parameters, the dynamics of individual cells were described by the RRE (3). Accordingly, the images of the sigma-points in the state and the observation space, $\boldsymbol{\mathcal{X}}_l$ and $\boldsymbol{\mathcal{Y}}_l$, were computed as

$$\begin{aligned}
\frac{d\boldsymbol{\mathcal{X}}_l}{dt} &= f_x(\boldsymbol{\mathcal{X}}_l, \boldsymbol{\mathcal{S}}_l, u), \qquad l = 0,\dots,2L\\
\boldsymbol{\mathcal{Y}}_l &= f_y(\boldsymbol{\mathcal{X}}_l, \boldsymbol{\mathcal{S}}_l, u).
\end{aligned} \tag{4}$$

The mean and covariances of the species were computed as

$$\mathbf{m}_x \approx \sum_{l=0}^{2L} v_l^{(m)} \boldsymbol{\mathcal{X}}_l,$$
$$\mathbf{C}_x \approx \sum_{l=0}^{2L} v_l^{(c)} (\boldsymbol{\mathcal{X}}_l - \mathbf{m}_x)(\boldsymbol{\mathcal{X}}_l - \mathbf{m}_x)^T.$$

The mean and covariances of the observables read

$$\begin{aligned} \mathbf{m}_y &\approx \sum_{l=0}^{2L} v_l^{(m)} \boldsymbol{\mathcal{Y}}_l, \\ \mathbf{C}_{xy} &\approx \sum^{2L} v_l^{(c)} (\boldsymbol{\mathcal{Y}}_l - \mathbf{m}_y)(\boldsymbol{\mathcal{Y}}_l - \mathbf{m}_y)^T. \end{aligned} \tag{5}$$

In our MATLAB SPToolbox, the parametrization of $\mathbf{D}$ was implemented by either a diagonal matrix logarithm or a matrix logarithm (Williams, 1999), in case of correlations between parameters. For our study, we assumed a log-normal distribution of the parameters, i.e., $\boldsymbol{\beta}$ and $\mathbf{D}$ described the median and scale matrix of the corresponding log-normal distribution and the exponent of $\boldsymbol{\mathcal{S}}_l$ was used in (4).

Moment-closure approximation In this study, we also considered intrinsic variability of biochemical reactions as introduced by discreteness and stochasticity of biochemical reactions. The single-cell dynamics are described by continuous time discrete state Markov chains (CTMCs). We approximated the time-dependent moments of this process using the moment-closure approximation (Engblom, 2006; Lee et al., 2009). This method provided equations for the temporal evolution of moments of the species, i.e., the mean

$$m_{x,i}(t) = \sum_{\mathbf{x}\in\Omega} x_i p(\mathbf{x}, t), \qquad i = 1, \dots, n$$

of species x_i, and higher order moments such as the covariance

$$C_{x,ij}(t) = \sum_{\mathbf{x}\in\Omega} (x_i - m_{x,i}(t))((x_j - m_{x,j}(t)), \qquad i, j = 1, \dots, n$$

between species x_i and x_j. Here, $p(\mathbf{x}, t)$ denotes the chemical master equation, Ω the set of possible states, and n the number of species. Given the moments of the species, we calculated the moments of the observables by

$$\begin{aligned} m_{y,i}(t) &= \sum_{\mathbf{x}\in\Omega} f_{y,i}(\mathbf{x}) p(\mathbf{x}, t) \\ C_{y,ij}(t) &= \sum_{\mathbf{x}\in\Omega} f_{y,i}(\mathbf{x}) - m_{y,i}(t))(f_{y,j}(\mathbf{x}) - m_{y,j}(t)) p(\mathbf{x}, t). \end{aligned} \tag{6}$$

For the automatic generation of the moment-closure approximation and the corresponding simulation files, we employed the MATLAB toolbox CERENA (Kazeroonian et al., 2016). In addition, this toolbox provided the equations for the system size expansion, which can also be incorporated into our modeling framework as an alternative to the moment equations.

Mechanistic hierarchical population model

For the hierarchical population model, the mechanistic description of individual subpopulations, as introduced in the previous section, is combined with mixture models to describe the entire cell population.

Hierarchical model and its approximations We considered heterogeneous cell populations consisting of multiple subpopulations, $s = 1, \dots, N$. Assuming independence, the distribution of the states and observables in the overall population is the weighted sum of the distribution of the states and observables in the subpopulations, $p_s(\boldsymbol{x}|t)$ and $p_s(\boldsymbol{y}|t)$. The weights $w_s(t)$ are the relative populations sizes, with $\forall t\colon \sum_s w_s(t) = 1$. This yields the hierarchical population model

$$p(\boldsymbol{x}|t) = \sum\nolimits_s w_s(t) p_s(\boldsymbol{x}|t),$$
$$p(\boldsymbol{y}|t) = \sum\nolimits_s w_s(t) p_s(\boldsymbol{y}|t).$$

The distribution of states and observables in the subpopulations originate according to the single cell properties. As the measurements $\bar{\boldsymbol{y}}$ are in general noise corrupted, $\boldsymbol{y} \sim p(\bar{\boldsymbol{y}}|\boldsymbol{y})$ we also considered the distribution

$$\begin{aligned} p(\bar{\boldsymbol{y}}|t) &= \int p(\bar{\boldsymbol{y}}|\boldsymbol{y})\, p(\boldsymbol{y}|t) d\boldsymbol{y} \\ &= \sum\nolimits_s w_s(t) \underbrace{\int p(\bar{\boldsymbol{y}}|\boldsymbol{y}) p_s(\boldsymbol{y}|t) d\boldsymbol{y}}_{=: p_s(\bar{\boldsymbol{y}}|t)}. \end{aligned}$$

To ensure computational efficiency, the probability distributions $p_s(\boldsymbol{x}|t)$, $p_s(\boldsymbol{y}|t)$ and $p_s(\bar{\boldsymbol{y}}|t)$ were approximated using the statistical moments. For the measured observables, the computed statistical moments were encoded in $\boldsymbol{\varphi}_s$, yielding

$$p(\bar{\boldsymbol{y}}|t) = \sum\nolimits_s w_s(t) \phi(\bar{\boldsymbol{y}}|\boldsymbol{\varphi}_s(t))$$

with parametric probability distribution ϕ. In this study, we employed the multivariate normal distribution

$$N(\bar{\boldsymbol{y}}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma}_s)^{\frac{1}{2}}} e^{-\frac{1}{2}(\bar{\boldsymbol{y}}-\boldsymbol{\mu}_s)^T (\boldsymbol{\Sigma}_s)^{-1} (\bar{\boldsymbol{y}}-\boldsymbol{\mu}_s)}, \qquad (7)$$

and multivariate log-normal distribution

$$\log N(\bar{\boldsymbol{y}}|\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \frac{1}{(2\pi)^{\frac{d}{2}}\det(\boldsymbol{\Sigma}_s)^{\frac{1}{2}}(\prod_{i=1}^{d} \bar{y}_{yi})} e^{-\frac{1}{2}(\log(\bar{\boldsymbol{y}})-\boldsymbol{\mu}_s)^T(\boldsymbol{\Sigma}_s)^{-1}(\log(\bar{\boldsymbol{y}})-\boldsymbol{\mu}_s)}, \tag{8}$$

with distribution parameters $\boldsymbol{\varphi}_s = (\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. For example, for the multivariate normal distribution and no measurement noise, the distributions parameters were obtained by $\boldsymbol{\mu}_s = \mathbf{m}_{s,y}$ and $\boldsymbol{\Sigma}_s = \mathbf{C}_{s,y}$.

Likelihood function The parameters of the hierarchical population model $\boldsymbol{\theta}$ comprise the means/medians of the cell parameters $\boldsymbol{\beta}, \boldsymbol{\beta}_s$ as well as the entries of the scale matrices $\mathbf{D}, \mathbf{D}_s$, the mixture weights w_s, and measurement noise $\boldsymbol{\Gamma}$**.** These parameters were estimated using maximum likelihood estimation. The likelihood function for multivariate measurement data $\bar{\mathbf{y}}^{e,k,j} \in \mathbb{R}^{\boldsymbol{d}}$ is given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \prod_{e,k,j} \sum_{s} w_s^e(t_k, \boldsymbol{\theta})\, \phi\left(\bar{\mathbf{y}}^{e,k,j} | \boldsymbol{\varphi}_s^e((t_k, \boldsymbol{\theta}, u_e)\right) \\ \text{with } \dot{\mathbf{z}}_s^e &= g(\mathbf{z}_s^e, \boldsymbol{\xi}_s, u_e), \qquad \mathbf{z}_s^e(0) = \mathbf{z}_0^e(\boldsymbol{\xi}_s(\boldsymbol{\theta}), u_e) \\ \boldsymbol{\varphi}_s^e &= h(\mathbf{z}_s^e, \boldsymbol{\xi}_s(\boldsymbol{\theta}), u_e) \end{aligned} \tag{9}$$

with means and covariances $\mathbf{z}_s^e = (\mathbf{m}_s^e, \mathbf{C}_s^e)^T$ of species $\mathbf{x}$. The means and covariances are provided by some map g, e.g., the sigma-point approximation or the moment-closure approximation. The subpopulation parameters $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \mathbf{D}_s)$ are given by

$$\beta_{s,i} = \begin{cases} \beta_i & \text{homogeneous} \\ \beta_i & \text{cell-to-cell variable} \\ \beta_{s,i} & \text{subpopulation variable} \\ \beta_{s,i} & \text{inter- and intra-subpopulation variable} \end{cases}$$

$$D_{s,ii} = \begin{cases} 0 & \text{homogeneous} \\ D_{ii} & \text{cell-to-cell variable} \\ 0 & \text{subpopulation variable} \\ D_{s,ii} & \text{inter- and intra-subpopulation variable} \end{cases}$$

The mapping h links the computed moments to the moments of the measurand including measurement noise, which are denoted by $\mathbf{m}_y = (m_{y,1}, \ldots, m_{y,d})^T$ and $\mathbf{C}_y$ and can be calculated as described, e.g., in (5) and (6). For a mixture of normal distributions (7), the means and covariances were linked to the parameters of the normal distribution

$$\boldsymbol{\mu}_s^e = \mathbf{m}_{s,y}^e,\ \boldsymbol{\Sigma}_s^e = \mathbf{C}_{s,y}^e + \boldsymbol{\Gamma},$$

including additive normally distributed measurement noise parametrized by

$$\mathbf{\Gamma} = \left(\Gamma_{i,j}\right)_{i,j=1,\dots,n} = \begin{pmatrix} \sigma_{1,\mathrm{noise}}^2 \; 0 \; 0 \\ 0 \ddots 0 \\ 0 \; 0 \; \sigma_{d,\mathrm{noise}}^2 \end{pmatrix}.$$

For the log-normal distribution (8), the distribution parameters were directly simulated with the sigma-point approximation for the logarithm of the observable, yielding the relation

$$\boldsymbol{\mu}_s^e = \mathbf{m}_{s,\log(y)}^e, \; \mathbf{\Sigma}_s^e = \mathbf{C}_{s,\log(y)}^e + \mathbf{\Gamma},$$

accounting for multiplicative log-normally distributed measurement noise. Alternatively, the mean of the simulation was linked to the mean of the log-normal distribution by

$$\mu_{s,i}^e = \log(m_{s,y,i}^e) - \frac{1}{2}\Sigma_{s,ii}^e,$$

$$\Sigma_{s,ii}^e = \log\left(\frac{C_{s,y,i}^e}{m_{s,y,i}^e m_{s,y,j}^e} + 1\right) + \Gamma_{ij}.$$

In principle also other distributions can be incorporated in the presented modeling framework. Due to numerical reasons, we used the log-likelihood function (Loos et al., 2016).

Gradient of likelihood function To promote efficiency of the numerical optimization and robust convergence, we derived the gradient of the log-likelihood function. For this, the gradient of the corre- sponding mixture ditribution ϕ with respect to θ was calculated using

$$\frac{\partial}{\partial\theta} N\left(\bar{\boldsymbol{y}}^{e,k,j} | \boldsymbol{\mu}_s^e(t_k), \mathbf{\Sigma}_s^e(t_k)\right) = -\frac{1}{2} N\left(\bar{\boldsymbol{y}}^{e,k,j} | \boldsymbol{\mu}_s^e(t_k), \mathbf{\Sigma}_s^e(t_k)\right)\Bigg(\mathrm{Tr}\left(\left(\mathbf{\Sigma}_s^e(t_k)\right)^{-1} \frac{\partial \mathbf{\Sigma}_s^e(t_k)}{\partial\theta}\right) + (\boldsymbol{\mu}_s^e(t_k) - \bar{\boldsymbol{y}}^{e,k,j})^T\left(\mathbf{\Sigma}_s^e(t_k)\right)^{-1}\left(\frac{\partial \boldsymbol{\mu}_s^e(t_k)}{\partial\theta}\right)^{\boldsymbol{T}} + \left(\frac{\partial \boldsymbol{\mu}_s^e(t_k)}{\partial\theta}\right)^T \left(\mathbf{\Sigma}_s^e(t_k)\right)^{-1}(\boldsymbol{\mu}_s^e(t_k) - \bar{\boldsymbol{y}}^{e,k,j}) + (\boldsymbol{\mu}_s^e(t_k) - \bar{\boldsymbol{y}}^{e,k,j})^T\left(\frac{\partial\left(\mathbf{\Sigma}_s^e(t_k)\right)^{-1}}{\partial\theta}\right)(\boldsymbol{\mu}_s^e(t_k) - \bar{\boldsymbol{y}}^{e,k,j})\Bigg),$$

and the relation

$$\log N(\bar{\boldsymbol{y}}^{e,k,j} | \boldsymbol{\mu}_s^e(t_k), \mathbf{\Sigma}_s^e(t_k)) = N\left(\log(\bar{\boldsymbol{y}}^{e,k,j}) | \boldsymbol{\mu}_s^e(t_k), \mathbf{\Sigma}_s^e(t_k)\right)\left(\prod_{i=1}^{e} y_i^{e,k,j}\right)^{-1}.$$

Additionally, the sensitivities of the distribution parameters $\frac{\partial \boldsymbol{\mu}_s^e}{\partial\theta}$ and $\frac{\partial \mathbf{\Sigma}_s^e}{\partial\theta}$ were required, which were obtained by simulating the sensitivity equation for the sigma-point or the moment-closure approximation and mapping it to the distribution parameters using h.

Comparison with existing models

A comparison of the hierarchical population model with existing methods is given in the following:

method	mechanistic description of		subpopulatons	multivariate data	reference
	dynamics	variabliity			
mixture model			✓	✓	e.g., Hastie et al. (2009)
moment-closure approximation	✓	✓		✓	e.g., Zechner et al. (2012)
ODE-constrained mixture model	✓		✓		Hasenauer et al. (2014)
hierarchical population model	✓	✓	✓	✓	this manuscript

Parameter estimation

For parameter estimation, we used the MATLAB toolbox PESTO (Stapor et al., 2018), which employs the function fmincon.m for local optimization. We used the interior-point algorithm and provided the analytic gradient of the log-likelihood function. Due to numerical better properties, we estimated the $\log_{10}$-transformed parameters. To explore the full parameter space, we performed multi-start optimization which has shown to outperform global optimization methods (Raue et al., 2013; Hross and Hasenauer, 2016). For this, randomly drawn initial parameter values were used for the optimization. For the uncertainty analysis, we calculated profile likelihoods (Raue et al., 2009) and the confidence intervals using the corresponding PESTO functions. We used the maximum likelihood estimates as initial values for the sampling of the posterior distribution with parallel tempering.

Calibration of single-cell model

The calibrated hierarchical population model provides estimates for $\beta_{s,i}$, and $D_{s,ii}$ which can then be used as prior information for the single-cell parameters $\boldsymbol{\xi}^j$ of cell j:

$$p(\xi_i^j) = \begin{cases} \delta(\xi_i^j - \beta_i) & \text{homogeneous} \\ N(\beta_i, D_{ii}) & \text{cell-to-cell variable} \\ \sum_s w_s \delta(\xi_i^j - \beta_{s,i}) & \text{subpopulation variable} \\ \sum_s w_s N(\beta_{s,i}, D_{s,ii}) & \text{inter- and intra-subpopulation variable} \end{cases}$$

in which δ denotes the Dirac delta distribution. The posterior distribution for the parameters of cell

j, $\boldsymbol{\xi}^j$, is given by

$$p(\boldsymbol{\xi}^j|\bar{\boldsymbol{y}}^j,\boldsymbol{\Gamma}) \propto p(\bar{\boldsymbol{y}}^j|\boldsymbol{\xi}^j,\boldsymbol{\Gamma})p(\boldsymbol{\xi}^j)$$

in which $p(\bar{\boldsymbol{y}}^j|\boldsymbol{\xi}^j,\boldsymbol{\Gamma})$ denotes the likelihood of the single-cell measurement $\bar{\boldsymbol{y}}^j$ for single-cell parameters $\boldsymbol{\xi}^j$ and noise parameters $\boldsymbol{\Gamma}$. The likelihood is $p(\bar{\boldsymbol{y}}^j|\boldsymbol{\xi}^j,\boldsymbol{\Gamma}) = \mathrm{N}(\bar{\boldsymbol{y}}^j|\boldsymbol{y}^j,\boldsymbol{\Gamma})$ for additive normally distributed measurement noise and is $p(\bar{\boldsymbol{y}}^j|\boldsymbol{\xi}^j,\boldsymbol{\Gamma}) = \log\mathrm{N}(\bar{\boldsymbol{y}}^j|\boldsymbol{y}^j,\boldsymbol{\Gamma})$ for multiplicative log-normally distributed measurement noise.

Conversion process

In the manuscript, we considered a model of a conversion process. In the following, we provide a detailed description of the data generation and data analysis. We first introduce the single-cell model of the conversion process. Afterwards, we present the results for the model accounting for the mean, and the hierarchical model accounting for the mean and covariances.

Single-cell model The conversion process is described by the following reactions

$$\begin{aligned} &R_1\colon \mathrm{A} \to \mathrm{B}, && \mathrm{rate} = k_1 u[\mathrm{A}],\\ &R_2\colon \mathrm{A} \to \mathrm{B}, && \mathrm{rate} = k_2[\mathrm{A}],\\ &R_3\colon \mathrm{B} \to \mathrm{A}, && \mathrm{rate} = k_3[\mathrm{B}]. \end{aligned}$$

Reaction R_1 describes the stimulus-dependend conversion, whereas reaction R_2 models the basal conversion from A to B. The conversion from B to A, reaction R_3, does not depend on stimulus u (Hasenauer et al., 2014). The concentrations of the species A and B are denoted by $[\mathrm{A}]$ and $[\mathrm{B}]$. The RRE for $(x_1, x_2) = ([\mathrm{A}],[\mathrm{B}])$ is given by

$$\begin{aligned} \frac{dx_1}{dt} &= k_3 x_2 - (k_1 u + k_2)x_1,\\ \frac{dx_2}{dt} &= (k_1 u + k_2)x_1 - k_3 x_2, \end{aligned}$$

with initial conditions

$$x_1(0) = \frac{k_3}{k_2}, \qquad x_2(0) = 1 - \frac{k_3}{k_2},$$

accounting for mass conservation $[\mathrm{A}] + [\mathrm{B}] = 1$ and the assumption that the system was in steady state before the stimulus was added at 0 min. We assumed the conversion from B to A to be cell-to-cell variable,

$$k_3 \backsim \log\mathrm{N}(\beta_{k_3}, \sigma_{k_3}^2), \tag{10}$$

yielding cell-to-cell variable initial conditions. The parameter k_1 was considered to differ between subpopulations and therefore was parametrized by $k_{1,1}$ and $k_{1,2}$. The weight w_1 indicated the proportion of the low responsive subpopulation. We generated artificial data for the parameters

$$\begin{aligned}\boldsymbol{\theta}^{\text{true}} &= \left(k_{1,1}, k_{1,2}, k_2, \beta_{k_3}, \sigma_{k_3}, \sigma_{\text{noise}}, w_1\right)^T \\ &= (10^{-0.1}, 10^{0.1}, 10^{-0.45}, 10^{-0.2}, 10^{-1}, 10^{-1.8}, 0.7)^T.\end{aligned}$$

We observed the concentration of B, i.e., $\mathrm{y} = x_2$. The data was created including 1000 cells at 5 time points for $u = 1$ by sampling from the distribution for k_3 (10) and simulating the corresponding RREs. Of the 1000 cells, 700 cells belonged to subpopulation 1 with low response to stimulation and 300 cells to the high responsive subpopulation 2. Additionally, the measurements of both subpopulations were assumed to be subject to logarithmic multiplicative measurement noise parameterized by σ_{noise}. We assumed the parameters $\boldsymbol{\theta}$ to be unknown and estimated them from the data with

(i) the approach introduced by Hasenauer et al. (2014) using the means (obtained by the RRE) and

(ii) hierarchical population model describing the means and covariances (obtained by the sigma-point approximation).

For both approaches, the underlying subpopulation structure was given, i.e., subpopulation variability of k_1.

Hierarchical model using RREs We considered a hierarchical model with subpopulation means that were described by the RRE. The distribution of the observables was assumed to be log-normal and the scale parameters were estimated from the data. We distinguished the following scenarios:

- one scale parameter that is shared across time points and subpopulations,
- one scale parameter for every subpopulations, which is shared between time points,
- 10 scale parameters that differ for each subpopulation and time-point.

These scale parameters were estimated along with $k_{1,1}, k_{1,2}, k_2, \beta_{k_3}$, and w_1 for this setting, which corresponds to the ODE constrained mixture modeling described by Hasenauer et al. (2014). For optimization, the kinetic parameters k_i were assumed to be in the interval $[10^{-3}, 10^3]$, the weight w_1 in $[0,1]$, and the scale parameters for the log-normal distribution were restricted to the interval $[10^{-2}, 10^2]$. For each model we performed 50 multi-starts at randomly drawn initial points. The fits corresponding to the optimal parameter values are shown in Figure S1A.

Hierarchical model using sigma-point approximations For the hierarchical population model, the parameter vector for subpopulation s was given by $\boldsymbol{\xi}_s = (\boldsymbol{\beta}_s, \mathbf{D}_s)$ with

$$\boldsymbol{\beta}_s = \begin{pmatrix} k_{1,s} \\ k_2 \\ \beta_{k_3} \\ \sigma_{\text{noise}} \end{pmatrix} \begin{matrix} \text{subpopulation variable} \\ \text{homogeneous} \\ \text{cell-to-cell variable} \\ \text{homogeneous} \end{matrix}$$

and

$$D_{s,ij} = \begin{cases} \sigma_{k_3}^2 & \text{for } i = j = 3 \\ 0 & \text{otherwise.} \end{cases}$$

To describe the introduced cell-to-cell variability in k_3 (10) we used the sigma-point approximation for the log-parameters.

To assess whether the true source of heterogeneity can be detected, we tested all possible combinations of additional cell-to-cell variability in $k_{1,s}$, k_2, or k_3. For this, the sigma-point approximation was applied to the logarithm of the observable, to link the mean and variance of the simulation directly to the distribution parameters of the log-normal distribution. The case of no additional cell-to-cell variability corresponds to the RRE models and is therefore not covered here.

For optimization, the kinetic parameters or their means (in case of cell-to-cell variability) were assumed to be in the interval $[10^{-3}, 10^{3}]$, the scale parameters σ_{k_i} and measurement noise σ_{noise} in $[10^{-3}, 10^{2}]$ and the weight w_1 in $[0,1]$. As for the RRE model, we performed 50 multi-starts. The fits corresponding to the optimal parameter values for each model are shown in Figure S1.

To evaluate how the method scales with the number of measured cells, we generated datasets with $n_j = \{10^1, 10^2, 10^3, 10^4, 10^5\}$ measured cells per time point. The average computation time for three replicates for 10 optimization starts for the varying number of data points is shown in Figure S2. The contribution of the evaluation of the density $\phi(\bar{\boldsymbol{y}}|\boldsymbol{\varphi}_s)$ increased linearly with the number of data points. However, the simulation time was almost constant for increasing number of data points, since the simulation did not depend on the number of measured cells. The slight increase can be explained by the increased number of iterations needed for optimization, which might have occurred due to different effective optimizer tolerances that were not comparable for varying number of data points.

Bayesian parameter estimation and model selection In the main manuscript, we used profile likelihoods for parameter uncertainty analysis and the BIC for model selection. We compared these approaches with their fully Bayesian counterparts. To facilitate this comparison, we considered uniform prior distributions.

In a first step, we evaluated the confidence intervals obtained using profile likelihoods. Therefore, we sampled the posterior distribution of the ground truth model using the parallel tempering algorithm implemented in the parameter estimation toolbox PESTO. The chains were initialized at the maximum likelihood estimates and their convergence was assessed using the Geweke test (Geweke, 1992). The comparison of the marginal posterior distributions and the profile likelihoods revealed an excellent agreement (Figure S1B). We note that the initialization of the parallel tempering algorithm using a sample from the prior instead of using the pre-computed maximum likelihood estimates, yielded substantially longer computation times and often did not result in a converged chain for 2×10^5 iterations (corresponding to roughly 4 CPU hours). This indicates that for this problem, optimization is an important step. In a second step, we evaluated the ranking obtained with the BIC to the ranking obtained by fully Bayesian approaches. Therefore, we computed the log marginal likelihood as well as the log pointwise predictive density (Gelman et al., 2014) for each model. The log marginal likelihood was determined using thermodynamic integration with the Simpsons rule (Hug et al., 2016) (Figure S1C). The log pointwise predictive density was determined by sampling the posterior distribution for a subset of the data, for the measurements for all but one time point, and computing the average log-likelihood on the remaining data. The comparison of BIC values, log marginal likelihoods, and log pointwise predictive densities revealed a good agreement. The Spearman's rank correlation coefficient between BICs and log marginal likelihoods is $r = 0.98$, and $r = 0.83$ between BICs and log pointwise predictive densities. Furthermore, all criteria suggest the rejection of the models which include only the mechanistic description of the mean but not the variance. For the remaining models the methods provide a sightly different ordering, but all of them indicate the importance of the variability of k_3. Interestingly, model complexity seems to be more penalized by the BIC. As the model selection did not reject all models but the ground truth model, we evaluated the contribution of the variability of individual parameters to the variability of the observable. Therefore, we evaluated the reduction of the variability of the observable achieved by removing the variability in the parameter of interest. This analysis was performed for samples from the posterior distribution (Figure S1D). We performed this analysis for the models which can not be rejected based on a Bayes factor cutoff of 100 (Kass and Raftery, 1995) and found that clearly the main contribution to the variability comes from variability in k_3. This means that even for plausible models which account for variability in k_1 or k_2, the main source of variability is k_3. To confirm this further, we computed the BIC weights, also known as Schwarz weights, for a certain variability by summing the BIC weights

$$\frac{\exp(-0.5\mathrm{BIC}_m)}{\sum_{\widetilde{m}}\exp(-0.5\mathrm{BIC}_{\widetilde{m}})} \tag{11}$$

for all models accounting for this variability. To detect the source of variability, we took the models for all possible combinations into account. Similarly, we calculated the evidence of a variability based on the computed marginal likelihoods. Both approaches agree in the presence of variability in k_3, confirming the agreement of the results. The BIC weights for the parameters k_1 and k_2 are higher than the evidences computed from the log marginal likelihoods, which, however, does not have a big contribution to the overall variability (Figure S1D&E).

Differential protein expression

In the main manuscript, we investigated multivariate measurements of differential protein expression. Here, we provide the detailed description of the data generation and the data analysis using the hierarchical model for the full and the marginal distributions.

Single-cell model The simple model of differential protein expression considers six reactions

$$\begin{aligned}
&R_1 : \emptyset \to \mathrm{A}, && \mathrm{rate} = \lambda_0,\\
&R_2 : \emptyset \to \mathrm{B}, && \mathrm{rate} = \lambda_0,\\
&R_3 : \emptyset \to \mathrm{A}, && \mathrm{rate} = \lambda_\mathrm{A} u,\\
&R_4 : \emptyset \to \mathrm{B}, && \mathrm{rate} = \lambda_B u,\\
&R_5 : \mathrm{A} \to \emptyset, && \mathrm{rate} = \gamma[\mathrm{A}],\\
&R_6 : \mathrm{B} \to \emptyset, && \mathrm{rate} = \gamma[\mathrm{B}],
\end{aligned}$$

comprising the basal expression with rate λ_0, degradation with rate γ and stimulus-induced expression, depending on u, with rate λ_A and λ_B for protein A and B, respectively. The corresponding ODE system for the temporal evolution of $(x_1, x_2) = ([\mathrm{A}], [\mathrm{B}])$ is

$$\begin{aligned}
\frac{dx_1}{dt} &= \lambda_0 + \lambda_\mathrm{A} u - \gamma x_1,\\
\frac{dx_2}{dt} &= \lambda_0 + \lambda_\mathrm{B} u - \gamma x_2,
\end{aligned}$$

with initial conditions

$$x_1(0) = x_2(0) = \frac{\lambda_0}{\gamma},$$

obtained by assuming that the system was in steady state before the stimulus was added at 0 min. Two subpopulations were assumed, one showing high expression of A while the other shows high expression of B after stimulation with u. The degradation rate γ was considered to be cell-to-cell variable,

$$\gamma \backsim \log\mathrm{N}(\beta_\gamma, \sigma_\gamma^2), \qquad (12)$$

with median β_γ and scale σ_γ which were equal between the subpopulations. The measurements were exposed to log-normally distributed multiplicative measurement noise parametrized by σ_{noise}.

Hierarchical model The hierarchical model accounted for the subpopulation variability of λ_{A} and λ_{B} and the cell-to-cell variability of γ. This yielded the subpopulation parameters

$$\boldsymbol{\beta}_s = \begin{pmatrix} \lambda_0 \\ \lambda_{\text{A},s} \\ \lambda_{\text{B},s} \\ \beta_\gamma \\ \sigma_{\text{noise}} \end{pmatrix} \begin{array}{l} \text{homogeneous} \\ \text{cell-to-cell variable} \\ \text{subpopulation variable} \\ \text{inter- and intra-subpopulation variable} \\ \text{homogeneous} \end{array}$$

$$D_{s,ij} = \begin{cases} \sigma_\gamma^2 & \text{for } i = j = 4 \\ 0 & \text{otherwise.} \end{cases}$$

As before, the sigma-point approximation was applied to the log-transformed parameters accounting for the log-normal distribution of γ. We performed 100 starts using as data either the full or the marginal distribution of A and B. The parameters and corresponding boundaries are:

symbol	description	θ^{lb}	θ^{ub}	θ^{true}
λ_0	basal protein expression	10^{-3}	10^{3}	$10^{1.7}$
$\lambda_{\text{B},1}$	induced protein expression of A in subpop. 1	10^{-3}	10^{3}	$10^{2.7}$
$\lambda_{\text{A},2}$	induced protein expression of A in subpop. 2	10^{-3}	10^{3}	10^{2}
$\lambda_{\text{B},1}$	induced protein expression of B in subpop. 1	10^{-3}	10^{3}	10^{2}
$\lambda_{\text{B},2}$	induced protein expression of B in subpop. 2	10^{-3}	10^{3}	$10^{2.7}$
β_γ	median of protein degradation	10^{-3}	10^{3}	10^{-1}
σ_γ	variability of protein degradation	10^{-3}	10^{1}	10^{-1}
σ_{noise}	measurement noise	10^{-3}	10^{2}	10^{-1}
w_1	weight of subpop.1	0	1	0.5

Using a statistical approach to obtain the number of converged starts (Hross and Hasenauer, 2016), we found that 84/100 starts converged for the full distribution and 91/100 for the marginal distributions.

NGF-induced Erk1/2 signaling

Here, we provide details for the analysis of NGF-induced Erk1/2 signaling. We employed the model proposed by Hasenauer et al. (2014), which comprises the reactions

$$\begin{aligned} &R_1\text{: TrkA} + \text{NGF} \rightarrow \text{ TrkA: NGF}, && \text{rate} = k_1[\text{TrkA: NGF}], \\ &R_2\text{: TrkA: NGF} \rightarrow \text{ TrkA} + \text{NGF}, && \text{rate} = k_2[\text{TrkA}][\text{NGF}], \end{aligned}$$

$$R_3\text{: Erk} \rightarrow \text{ pErk}, \qquad \text{rate} = k_3[\text{TrkA: NGF}][\text{Erk}],$$
$$R_4\text{: pErk} \rightarrow \text{ Erk}, \qquad \text{rate} = k_4[\text{Erk}],$$
$$R_5\text{: pErk} \rightarrow \text{ Erk}, \qquad \text{rate} = k_5[\text{pErk}].$$

Conservation of mass yields

$$\begin{aligned} [\text{TrkA}] + [\text{TrkA: NGF}] &= [\text{TrkA}]_0\,, \\ [\text{NGF}] + [\text{TrkA: NGF}] &= [\text{NGF}]_0, \\ [\text{Erk}] + [\text{pErk}] &= [\text{Erk}]_0. \end{aligned}$$

To eliminate structurally non-identifiable parameters, the model was reparametrized to

$$\begin{aligned} \frac{dx_1}{dt} &= k_1[\text{NGF}]_0(k_3[\text{TrkA}]_0 - x_1) - k_2x_1, \qquad x_1(0) = 0 \\ \frac{dx_2}{dt} &= (x_1 + k_4)(c_P[\text{Erk}]_0 - x_2) - k_5x_2, \qquad x_2(0) = \frac{k_4c_P[\text{Erk}]_0}{(k_4 + k_5)}, \end{aligned} \tag{13}$$

with $x_1 \;=\; k_3[\text{TrkA: NGF}]$ and $x_2 = \; c_P[\text{pErk}]_0$. The observables for the considered experimental conditions are

$$\boldsymbol{y}^e = \begin{cases} c_P^e[\text{pErk}]_0 + o_P^e, e = 1,2, & \text{(pErk1/2 kinetics and dose responses)}, \\ (c_P^e[\text{pErk}]_0 + o_P^e, c_T[\text{TrkA}]_0 + o_T)^{\boldsymbol{T}}, e = 3, & \text{(pErk1/2 and TrkA dose responses)}, \\ (c_P^e[\text{pErk}]_0 + o_P^e, c_E[\text{Erk}]_0 + o_E)^{\boldsymbol{T}}, e = 4, & \text{(pErk1/2 and Erk1/2 dose responses)}, \end{cases}$$

to compare the subpopulations on poly-D-lysine (PDL) and

$$\boldsymbol{y}^e = \begin{cases} c_P^e[\text{pErk}]_0 + o_P^e, e = 1,\ldots,4, & \text{(pErk1/2 kinetics and dose responses)}, \\ (c_P^e[\text{pErk}]_0 + o_P^e, c_T[\text{TrkA}]_0 + o_T)^{\boldsymbol{T}}, e = 5,6, & \text{(pErk1/2 and TrkA dose responses)}, \\ (c_P^e[\text{pErk}]_0 + o_P^e, c_E[\text{Erk}]_0 + o_E)^{\boldsymbol{T}}, e = 7,8, & \text{(pErk1/2 and Erk1/2 dose responses)}, \end{cases}$$

to study the effects of the extracellular scaffolds PDL and collagen type I (Col I) on the neurons (PDL: $e \;= 1,3,5,7$, Col I: $e \;= 2,4,6,8$).

The pErk1/2, TrkA and Erk1/2 levels could only be measured up to some scaling constants denoted by c_P, c_T, and c_E , respectively, and with some offsets denoted by o_P , o_T, and o_E . Each observable was assumed to be subject to multiplicative log-normally distributed measurement noise parameterized by $\sigma_{P,\text{noise}}^e$, $\sigma_{T,\text{noise}}$, and $\sigma_{E,\text{noise}}$. For the comparison of the extracellular scaffold, the same scaling, offset, and measurement noise parameters were used for PDL and Col I. For each subpopulation, we used the sigma- point approximation accounting for cell-to-cell variability in cellular TrkA activity and Erk1/2 levels. The covariance between TrkA activity and relative Erk1/2 expression was parametrized, accounting for correlations, with the matrix logarithm parametrization $M(\sigma_T, \sigma_E, \sigma_{TE}) \in \mathbb{R}^{2\times 2}$. All other entries of $\mathbf{D}_s$ were assumed to be 0.

Data pre-processing For our analysis, we scaled each replicate such that the quadratic difference of the log-transformed fluorescence mean intensities across replicates is minimal (see getScalingFactors.m). The scaled intensities of the cells of each replicate were then pooled and analyzed together.

Subpopulation differences We accounted for all possible combinations of subpopulation variability of $k_1, k_2, k_4, k_5, k_3[\mathrm{TrkA}]_0$, and $c_P^e[\mathrm{Erk}]_0$. This yielded in total $2^6 = 64$ models that were tested, ranging from $n_\theta = 26$ parameters, for the model assuming no subpopulations at all, to $n_\theta = 33$ parameters, assuming that the subpopulations differ in all parameters. To take into account all hierarchical models, we considered the BIC weights for individual differences as in (11).

We compared the results of model selection by BIC and log pointwise posterior density. This was done for the models accounting for no or one difference between the subpopulations (Figure S4B). We considered this reduced set of models for the comparison, as the sampling for the calculation of the log pointwise predictive density and the calculation of the Bayes factors took (on average 780 CPU hours per model for the Bayes factors). The BIC values, the log pointwise posterior density, and the Bayes factors strongly prefer the model accounting for differing TrkA levels over all other models ($\Delta\mathrm{BIC} > 7{\times}10^3$). We found that the log pointwise posterior density highly depends on the splitting of the data set, with smaller test and training data sets preferring less complex models. The results in Figure S4B are shown for splitting the data set in two parts, which gave a rank correlation of $r = 0.61$. The Bayes factors even yielded a rank correlation of $r = 1$, indicating that the Bayes factors are indeed well approximated by the BIC for these models.

Dephosphorylation rates To validate, whether the two subpopulations differ in their dephosphorylation/phosphotase activity (parameterized by k_5), we inhibited cells with the Mek-inhibitor U0126 (10 μM). NGF binds to the TrkA+ subpopulation and activates pErk1/2 signaling, whereas GDNF binds to the Ret receptor on the opposing subpopulation (TrkA-) and yields pErk1/2 signaling in this neuronal subgroup. Cells were pre-stimulated for 1 h with the combined stimuli NGF (20 ng/ml) and GDNF (100 ng/ml) to obtain responses in both subpopulations. We measured pErk1/2 levels to obtain the dynamics of the dephosphorylation as well as TrkA levels to distinguish the two subpopulations. Cells were considered to belong to the TrkA+ subpopulation if their intensity was above 670 and to the TrkA- subpopulation if their intensity was below 630. The measurements were taken at 0, 1, 4, 7, 10, 13, 16, 18, 22, 25, 28, 31, 34, and 37 min and collected for four replicates.

To obtain the de-phosphorylation rate k_5, we normalized the values of pErk1/2 to 1 at $t = 0$ min and 0 at $t_{\max} = 37$ min. We fitted an exponential decay

$$E(t) = E_c \exp(-k_5 t) + E_0,$$

to the scaled data of the four replicates. The scaling E_c and offset E_0 could be determined from the boundary conditions

$$E_0 = 1 \text{ and } E(t_{\max}) = 0.$$

This yielded the four values for the de-phosphorylation in the TrkA+ subpopulation and in the TrkA-subpopulation shown in Figure S3. A two-sample t-test with Welch's correction gave a p-value of 0.6163, indicating that the dephosphorylation rates in the two subpopulations were not significantly different.

Final model The final model accounted for subpopulation differences in cellular TrkA activity (Figure S4A) and also took into account differences in the variance of TrkA activity between the subpopulations. The fits for the data, which are not shown in the main manuscript are visualized in Figure S4C for the multivariate measurements of pErk1/2 and TrkA, and in Figure S4D for the measurements of pErk1/2 and Erk1/2. Using the final calibrated model, we predicted the relation between pErk1/2 levels at 0 and 120 min by drawing parameters from the inferred single-cell parameter distribution and simulating the ODE model (Figure S4E).

Differences mediated by extracellular scaffolds For the mechanistic comparison of the influence of the extracellular scaffolds, we used the model which assumes subpopulation differences in TrkA levels. The differences between the extracellular scaffolds were parameterized as

$$\kappa_{k_1}, \kappa_{k_2}, \kappa_{k_4}, \kappa_{k_5}, \kappa_{\beta_{k_3[\mathrm{TrkA}]_0}}, \kappa_{\beta_{c[\mathrm{Erk}]_0}}, \kappa_w$$

and the parameters were related by

$$k_{1,\mathrm{ColI}} = k_{1,\mathrm{PDL}} 10^{\kappa_{k_1}}.$$

Accounting for these 7 potential differences, we defined 128 hierarchical models. Each model was fitted to the data with multi-start local optimization using at least 20 starts. We sorted the models with respect to their BIC value, for which a low value indicates a good trade-off between model complexity and goodness of fit. The BIC weights for the differences were computed by summing over the BIC weights (11) of the models accounting for the corresponding differences. We found that the best model comprised differences in Erk1/2 expression, Erk1/2 dephosphorylation and cellular TrkA activity. The least suitable model was the model which did not allow differences between the extracellular scaffolds at all. This model was directly followed by the model only accounting for differences in the subpopulation weighting. The fit for the model accounting for differences in Erk1/2 expression, Erk1/2 dephosphorylation and cellular TrkA activity is shown in Figure 6 and Figure S5. The estimated parameters, their boundaries and the 95% confidence interval based on the profiles are

symbol	description	θ^{lb}	θ^{ub}	$\hat{\theta}$	95% confidence interval
k_1	binding affinity	10^{-6}	10^{6}	1.514	$[1.114, 2.179]$
k_2	release of NGF	10^{-6}	10^{6}	0.091	$[0.067, 0.131]$
k_4	basal Erk1/2 phosphorylation	10^{-6}	10^{6}	0.014	$[0.012, 0.017]$
k_5	Erk1/2 de-phosphorylation	10^{-6}	10^{6}	0.177	$[0.141, 0.222]$
$\beta_{k_3[\mathrm{TrkA}]_{0,1}}$	median of cellular TrkA activity in subpop. 1	10^{-6}	10^{6}	1.9×10^{4}	$[1.51\times10^{-4}, 2.41\times10^{-4}]$
$\beta_{k_3[\mathrm{TrkA}]_{0,2}}$	median of cellular TrkA activity in subpop. 2	10^{-6}	10^{6}	0.141	$[0.107, 0.185]$
$\beta_{c_P^{1,2}[\mathrm{Erk}]_0}$	median of relative Erk1/2 expression	10^{-6}	10^{6}	1.04×10^{4}	$[8.989\times10^{3}, 1.256\times10^{4}]$
$\sigma_{T,1}$	TrkA variability in subpop. 1	10^{-4}	10^{4}	5.366	$[4.742, 6.121]$
$\sigma_{T,2}$	TrkA variability in subpop. 2	10^{-4}	10^{4}	0.303	$[0.277, 0.330]$
σ_E	Erk1/2 variability	10^{-4}	10^{4}	0.250	$[0.230, 0.271]$
σ_{TE}	correlation of TrkA and Erk1/2	10^{-4}	10^{4}	2.263	$[2.162, 2.373]$
$o_P^{1,2}$	offset pErk1/2 ($e\ =\ 1, 2$)	10^{-6}	10^{6}	8.064×10^{-6}	$[0, 4.994]$
c_T/k_3	scaling TrkA	10^{-6}	10^{6}	1.735×10^{4}	$[1.318\times10^{4}, 2.296\times10^{4}]$
o_T	offset TrkA	10^{-6}	10^{6}	239.7	$[234.3, 245.2]$
$c_E/c_P^{5,6}$	scaling Erk1/2	10^{-6}	10^{6}	0.040	$[0.032, 0.049]$
o_E	offset Erk1/2	10^{-6}	10^{6}	592.2	$[549.8, 631.9]$
$c_P^{3,4}$	scaling pErk1/2 ($e\ =\ 3,4$)	10^{-6}	10^{6}	0.495	$[0.478, 0.511]$
$o_P^{3,4}$	offset pErk1/2 ($e\ =\ 3,4$)	10^{-6}	10^{6}	175.4	$[159.8, 190.5]$
$c_P^{5,6}$	scaling pErk1/2 ($e\ =\ 5,6$)	10^{-6}	10^{6}	0.810	$[0.783, 0.837]$
$o_P^{5,6}$	offset pErk1/2 ($e\ =\ 5,6$)	10^{-6}	10^{6}	292.1	$[266.2, 317.3]$
$c_P^{7,8}$	scaling pErk1/2 ($e\ =\ 7, 8$)	10^{-6}	10^{6}	1.029	$[0.997, 1.061]$
$o_P^{7,8}$	offset pErk1/2 ($e\ =\ 7, 8$)	10^{-6}	10^{6}	49.91	$[21.91, 76.90]$
$\sigma_{P,\mathrm{noise}}^{1,2}$	measurement noise pErk1/2 ($e\ =\ 1, 2$)	10^{-3}	10^{1}	0.335	$[0.306, 0.361]$
$\sigma_{P,\mathrm{noise}}^{3,4}$	measurement noise pErk1/2 ($e\ =\ 3,4$)	10^{-3}	10^{1}	0.370	$[0.354, 0.385]$
$\sigma_{T,\mathrm{noise}}$	measurement noise TrkA	10^{-3}	10^{1}	0.433	$[0.418, 0.448]$
$\sigma_{P,\mathrm{noise}}^{5,6}$	measurement noise pErk1/2 ($e\ =\ 5,6$)	10^{-3}	10^{1}	0.462	$[0.450, 0.473]$

$\sigma_{E,\text{noise}}$	measurement noise Erk1/2	10^{-3}	10^{1}	0.257	$[0.251, 0.263]$
$\sigma_{P,\text{noise}}^{7,8}$	measurement noise pErk1/2 ($e\ =\ 7{,}8$)	10^{-3}	10^{1}	0.267	$[0.241, 0.299]$
w_1	weight of subpopulation 1	10^{-4}	1	0.294	$[0.289, 0.298]$
κ_{k_5}	diff. between extracellular scaffold in k_5	10^{-3}	10^{3}	1.257	$[1.221, 1.297]$
$\kappa_{\beta_{k_3[\text{TrkA}]_0}}$	diff. between extracellular scaffold in TrkA	10^{-3}	10^{3}	1.043	$[1.019, 1.068]$
$\kappa_{\beta_{c_P^{1,2}[\text{Erk}]_0}}$	diff. between extracellular scaffold in Erk1/2	10^{-3}	10^{3}	1.393	$[1.357, 1.433]$

Accounting for intrinsic noise

To study the possibility of accounting for intrinsic noise in the hierarchical population model, we generated artificial data of a two stage gene expression (Figure S6A) using Gillespie's stochastic simulation algorithm (Gillespie, 1977) incorporated in the MATLAB Toolbox CERENA (Kazeroonian et al., 2016). The system comprises the following reactions

$$\begin{aligned}
&R_1: \emptyset \to \text{mA}, && \text{rate} = k_1,\\
&R_2: \emptyset \to \text{mA}, && \text{rate} = uk_3,\\
&R_3: \text{mA} \to \emptyset, && \text{rate} = k_3[\text{mA}],\\
&R_4: \text{mA} \to \text{A}, && \text{rate} = k_4[\text{mA}],\\
&R_5: \text{A} \to \emptyset, && \text{rate} = k_5[\text{A}].
\end{aligned}$$

Here, mA denotes the mRNA and A the protein and we assumed that only A could be observed. The two subpopulations differed in their response to stimulus u yielding different rates $k_{2,1}$ and $k_{2,2}$. For this setting, we only accounted for homogeneous and subpopulation variable parameters. However, the intrinsic variability of the births and deaths of individual molecules gave cell-to-cell variability in the cellular states. Cell-to-cell variability in parameters can also be incorporated using the moment-closure approximation.

The ODEs for the temporal evolution of the means and covariances were provided by the toolbox CERENA. In particular, the means m_1 and m_2 and the variances C_{11} and C_{22} of mRNA mA and protein A, respectively, were described as well as the correlation C_{12} of mA and A. The ODE system reads

$$\begin{aligned}
\frac{dm_1}{dt} &= \frac{k_1}{\Omega} + \frac{uk_2}{\Omega} - k_3 m_1,\\
\frac{dm_2}{dt} &= k_4 m_1 - k_5 m_2,\\
\frac{dC_{11}}{dt} &= \frac{k_1}{\Omega} + \frac{uk_2}{\Omega^2} - 2C_{11}k_3 + \frac{k_3 m_1}{\Omega},
\end{aligned}$$

$$\frac{dC_{12}}{dt} = C_{11}k_4 - C_{12}(k_3 + k_5),$$
$$\frac{dC_{22}}{dt} = 2C_{12}k_4 - 2C_{22}k_5 + \frac{k_4 m_1}{\Omega} + \frac{k_5 m_2}{\Omega},$$

with system size $\Omega = 1000$. Under the assumption that the system was in steady state before stimulation with u the initial conditions are

$$m_1(0) = \frac{k_1}{\Omega k_3},$$
$$m_2(0) = \frac{k_1 k_4}{\Omega k_3 k_5},$$
$$C_{11}(0) = \frac{k_1}{\Omega^2 k_3},$$
$$C_{12}(0) = \frac{k_{1k_4}}{\Omega^2 k_3(k_3 + k_5)},$$
$$m_1(0) = \frac{1}{\Omega^2}\left(\frac{k_1 k_4}{k_3 + k_5} + \frac{k_1 {k_4}^2}{k_3 k_5 (k_3 + k_5)}\right),$$

The true parameters used for the generation of the data were

$$\begin{aligned}\boldsymbol{\theta}^{\text{true}} &= \left(k_1, k_{2,1,}, k_{2,2}, k_3, k_4, k_5, w_1\right)^T \\ &= (10, 10, 20, 1, 5, 0.1, 0.5)^T.\end{aligned}$$

In this example, we employed mixtures of normal distributions, for which the mean and variance were linked to the distribution parameters by $\boldsymbol{\mu}_s = \mathbf{m}_s$ and $\boldsymbol{\Sigma}_s = \mathbf{C}_s$. First, we compared a model accounting for the mean, which was obtained by the RRE (Hasenauer et al., 2014), and a hierarchical model accounting for the mean and covariances, which were obtained by the moment-closure approximation (MA), both accounting for two subpopulations. For the RRE model 10 parameters for the parametrization of the variances were introduced, yielding in total $n_\theta = 17$. The model using the MA only comprised $n_\theta = 7$, since a mechanistic description of the variances was incorporated. For parameter estimation, the kinetic parameters were restricted to the interval $[10^{-3}, 10^3]$ and the $\log_{10}$-transformed parameters were fitted, whereas the weight w_1 was restricted to $[0,1]$ and fitted linearly. For the RRE model, the parameters for the variance were assumed to lie within $[10^{-4}, 10^2]$ and also fitted in $\log_{10}$-space. We also studied two models that incorporate the mechanistic description of the variance by the MA, but did not consider the presence of two subpopulations (MA, no subpop.). One of these models, however, accounts for cell-to-cell variability of each parameter (MA, cell-to-cell variability, no subpop.), which corresponds to the description by Zechner et al. (2012).

The models not accounting for subpopulation structures did not fit the data at all (Figure S6B). Even the included variability in parameters did not improve the fit substantially. In contrast, both subpopula- tion models provided a good fit to the data. However, the BIC for the MA model was substantially better than for the RRE model (BIC_{RRE}-BIC_{MA}=79.09). We found that the MA model gave the optimal value for 40% of the starts and the optimization for the RRE model ended in the optimum for 36% of the starts (Figure S6C). In terms of computation time there was a clear benefit using the mechanistic description of the variance (Figure S6D). The time required for one optimization start was about two-fold faster when using the MA (median=6.43 sec) instead of RREs (median=13.13 sec).

Furthermore, we studied the uncertainty of the parameter estimates using profile likelihoods (Figure S6E). Using the MA with subpopulations, all parameters were identifiable, indicated by a narrow profile. This was not the case for RREs, for which some parameters could not be identified from the the data and showed a flat profile. For the case of no subpopulations, most of the true parameters do not lie within the estimated intervals (Figure S6F-G). This emphasizes the importance of taking into account subpopulation structures.

QUANTIFICATION AND STATISTICAL ANALYSIS

For the analysis of the differences in pErk1/2 activity in the kinetic and dose responses for PDL and Col I, we employed a two-way ANOVA and Sidak's post-hoc test using GraphPad Prism. For assessing the statistical significance of the predicted differences, we applied the two-sample Welch's t-test employed by the MATLAB function ttest2. Significances are indicated as * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$). Model selection was performed using the BIC. We computed confidence intervals based on the profile likelihoods.

DATA AND SOFTWARE AVAILABILITY

The toolbox ODE-MM was used to implement the proposed hierarchical modeling framework as well as previous versions. This toolbox also provided the likelihood function and analytical gradient required for parameter estimation. The simulation of the means and covariance using sigma-points was implemented in the SPToolbox. Simulation of the RREs and corresponding sensitivity equations was conducted using the toolbox AMICI (Fröhlich et al., 2016). For the parameter estimation, we employed the toolbox PESTO (Stapor et al., 2018). All toolboxes and the experimental data are available at https://github.com/ICB-DCM. The versions of the toolboxes to reproduce the results of this manuscript are available at http://doi.org/10.5281/zenodo.1211553.

Figure 1

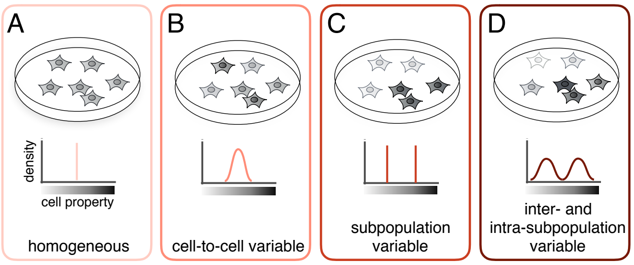

Figure 1: Cell populations exhibiting different levels of heterogeneity. Properties of cells, e.g., receptor levels or reaction rates, indicated by different gray shades for individual cells, can be (A) homogeneous: the property is the same for the entire cell population; (B) cell-to-cell variable: the property has a unimodal distribution across the cells; (C) subpopulation variable: the population can be separated into subpopulations, but within each subpopulation, the property does not vary; (D) inter- and intra-subpopulation variable: the property splits the population into subpopulations and also varies between cells within a subpopulation.

Figure 2

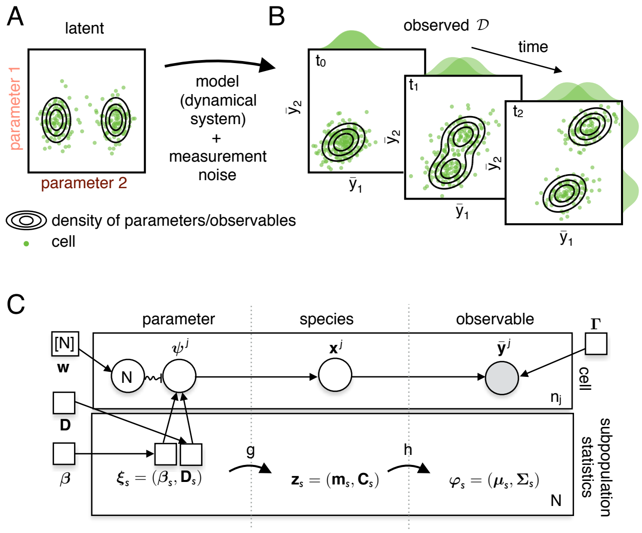

Figure 2: Illustration of the dynamics of a heterogeneous cell population and the mechanistic hierarchical population model. (A) Parameter distribution of a cell population consisting of two subpopulations. The contour lines illustrate the (approximated) parameter density of the cell-to-cell variable parameter 1 and the inter-and intra-subpopulation variable parameters 2. The heterogeneity of parameters is propagated from the latent parameter space to the observed measurement space. (B) Heterogeneity in parameters yields heterogeneous observables $\mathbf{y} = (y_1, y_2)^T$ that separate into two subpopulations after stimulation at time point t_0. (C) Structure of the single-cell system and approximation by the hierarchical population model using plate notation. Squares indicate fixed parameters, whereas circles indicate random variables. Gray shading of the circles/squares indicates a known value, whereas the other values are latent. The upper plate illustrates the variables associated with a cell j. Each of the n_j cells has parameters $\boldsymbol{\psi}^j$ drawn from a distribution defined by $\boldsymbol{\xi}_s$ and $\mathbf{w}$. The states of the species $\mathbf{x}^j$, resulting from the single-cell dynamics, yield the observables $\bar{\mathbf{y}}^j$, additionally influenced by measurement noise $\boldsymbol{\Gamma}$. The bottom plate visualizes the statistics of the corresponding cells of a subpopulation. For each subpopulation, the subpopulation parameters $\boldsymbol{\xi}_s$ are mapped to the means and covariances of the species of a subpopulation $\mathbf{z}_s$, which then are mapped to the distribution parameters $\boldsymbol{\varphi}_s$. The observables at the population level are considered to be distributed according to

Figure 3

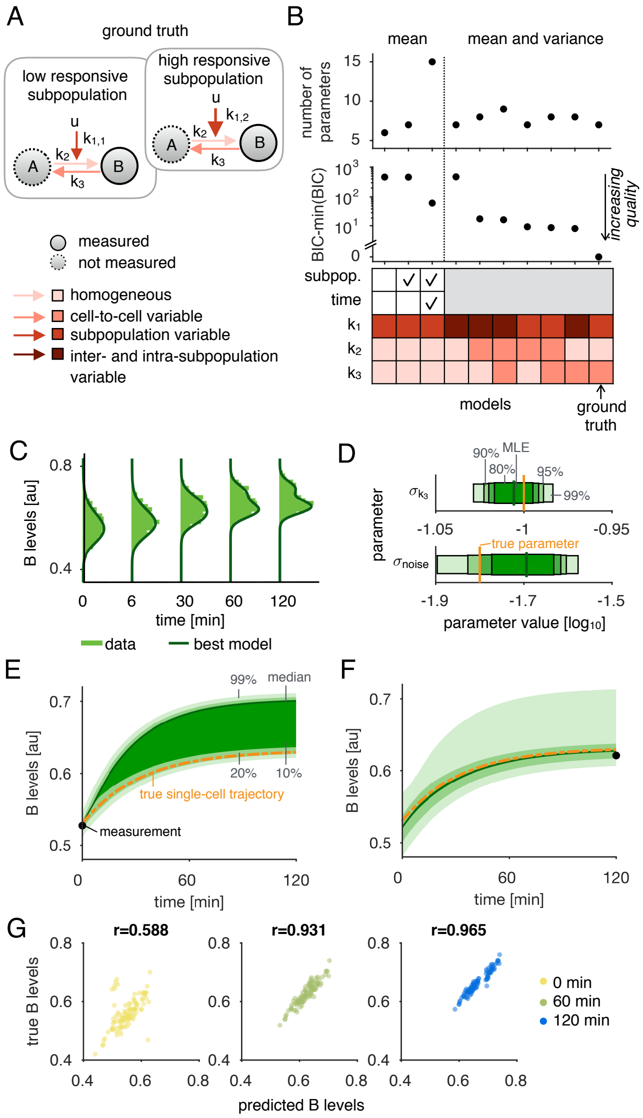

Figure 3: Inference of cell-to-cell variability using mechanistic models. . (A) Model of a conversion between two species A and B comprising two subpopulations differing in their response to stimulus u. Different colors indicate the variability of the reaction rates. (B) Model selection with the Bayesian Information Criterion (BIC). The first three models use RREs according to (Hasenauer et al., 2014) and vary in the number of additional parameters (1, 2, and 10) for the variances of the mixture distribution. The last models use the mean and variance obtained by sigma-points and differ in their sources of heterogeneity. (C) Data on the conversion process (1000 cells per time point) and fit corresponding to the best and true underlying model. (D) Confidence intervals for the variability of k_3 and the measurement noise (σ_{noise}). Horizontal bars show the confidence intervals corresponding to the 80%, 90%, 95%, and 99% confidence levels, and the vertical lines the maximum likelihood estimates (MLE). (E) Single- cell trajectories inferred using a single measurement at (E) t=0 min and (F) t=120 min. The inference is regularized using the hierarchical population model as prior. Shaded areas indicate the confidence intervals which were evaluated for samples of the posterior distribution and the dotted line indicates the single-cell trajectory from which the measurement point was generated. (G) Correlation of predicted and true level of B at 0, 60 and 120 min. True values were extracted from the (noise-free) simulation. Predictions are obtained using the single-cell data at time t=120 min.

Figure 4

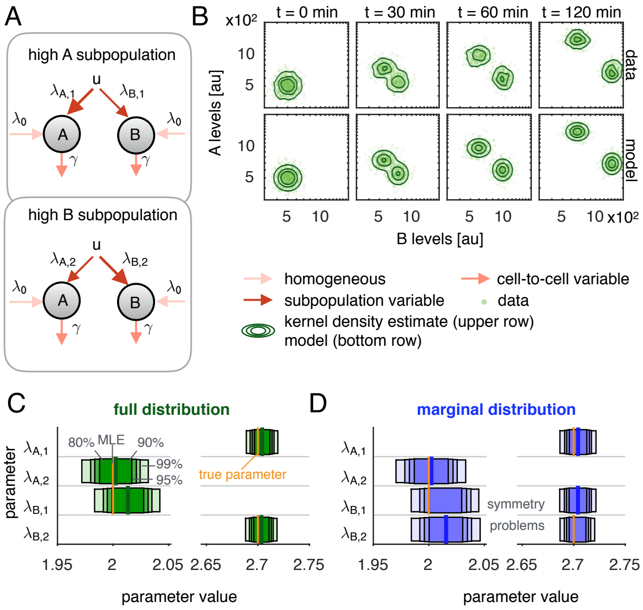

Figure 4: Reconstruction of differential protein expression in heterogeneous populations using multivariate data. (A) Model of differentially expressed proteins A and B. (B) Upper row: data points (1000 cells per time point) and kernel density estimation. Lower row: data points and model for the full distribution. (C,D) Confidence intervals for the parameters of the model using (C) the full distribution and (D) the marginal distributions. Horizontal bars show the confidence intervals corresponding to the 80%, 90%, 95%, and 99% confidence levels. The vertical lines show the MLE.

Figure 5

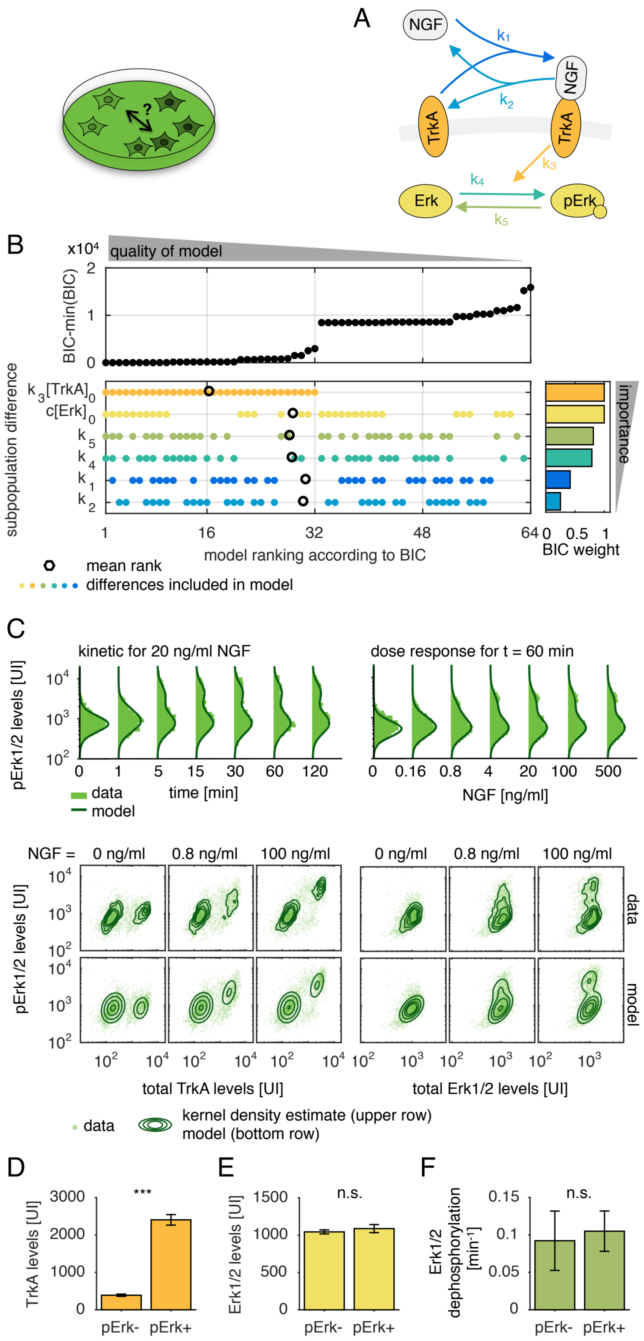

Figure 5: Sources of heterogeneity between subpopulations in primary sensory neurons. (A) Pathway model of NGF-induced Erk signaling. (B) Ranking according to the BIC values for the 64 hierarchical models, in which the colored dots indicate those parameters that are assumed to differ between the subpopulations. The importance of the differences is ranked according to the BIC weights, also known as Schwarz weights. The black circles indicate the mean rank of the models including the corresponding difference. (C) Data and fit for measurements of pErk1/2 levels (approximately 1400 cells per time point and 4300 cells per dosage) and multivariate measurements of pErk/TrkA and pErk/Erk levels (approximately 3000 cells per dosage) measured for 60 min under NGF stimulation with indicated concentrations. The measured values are in arbitrary units of intensity. For the multivariate data, the contour lines of the kernel density estimation of the data and the level sets of the density of the hierarchical model are shown. Mean and standard deviation of (D) TrkA levels ($n_r = 4$ replicates) (E) Erk1/2 levels ($n_r = 4$) and (F) Erk1/2 dephosphorylation ($n_r = 4$) of non-responsive (pErk-) and responsive (pErk+) sensory neurons after NGF stimulation with varying concentrations (as indicated in (C) for 60 min).

Figure 6

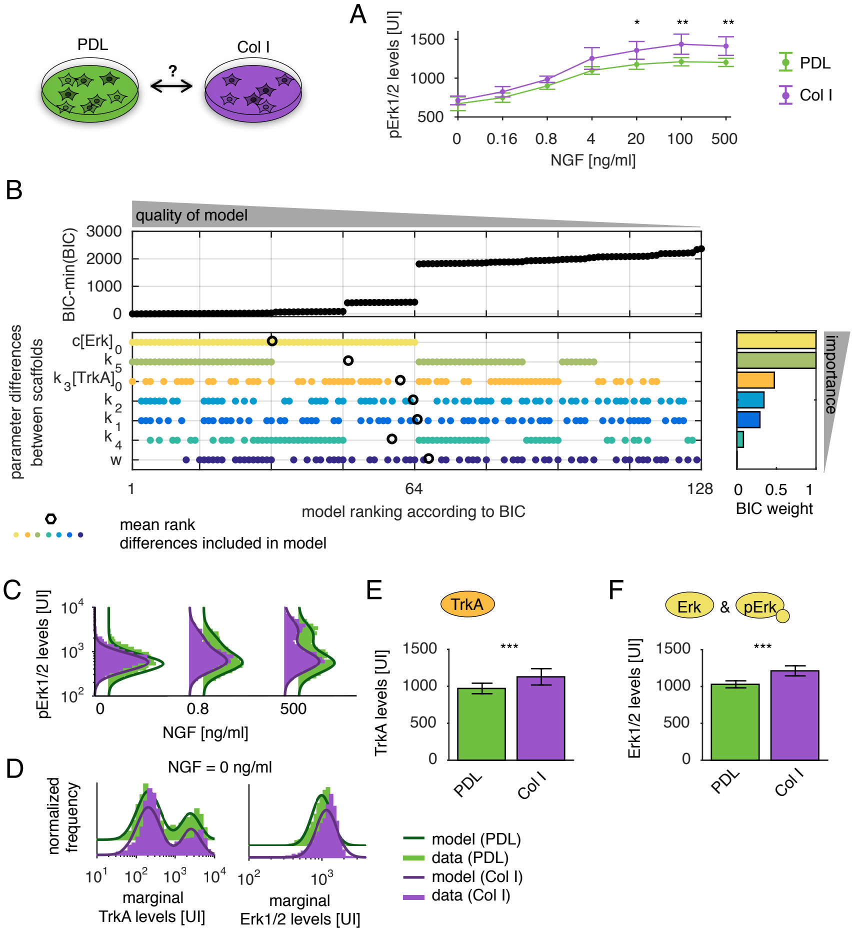

Figure 6: Differences in NGF-induced Erk1/2 phosphorylation mediated by different extracellular scaffolds. Primary sensory neurons were provided with the two different scaffolds poly-D-lysine (PDL) and collagen type I (Col I) in an overnight culture. (A) Sensory neurons grown on the Col I substrate showed a significantly higher mean phospho-Erk1/2 response to indicated doses of NGF after 1 h of stimulation. Means and standard deviations of four replicates are shown. (B) BIC-based ranking for the potential differences between culture conditions. The colored dots indicate which parameters are assumed to differ between the extracellular scaffolds. (C) Experimental data and fit for measurements of pErk1/2 distributions from Col I (approximately 2300 cells per dosage) and PDL (approximately 4300 cells per dosage) cultured neurons after treatment with indicated NGF concentrations for 1 h. (D) Marginal levels for TrkA and Erk1/2, which were assumed to be constant over varying doses and time (approximately 2000 cells in Col I and 2900 in PDL). Mean and standard deviation of (E) TrkA and (F) Erk1/2 levels of NGF dose response curve data, which showed significant elevations in Col I treated neurons. For this calculation, 24 samples were used (4 replicates for 6 doses).

Figure S1

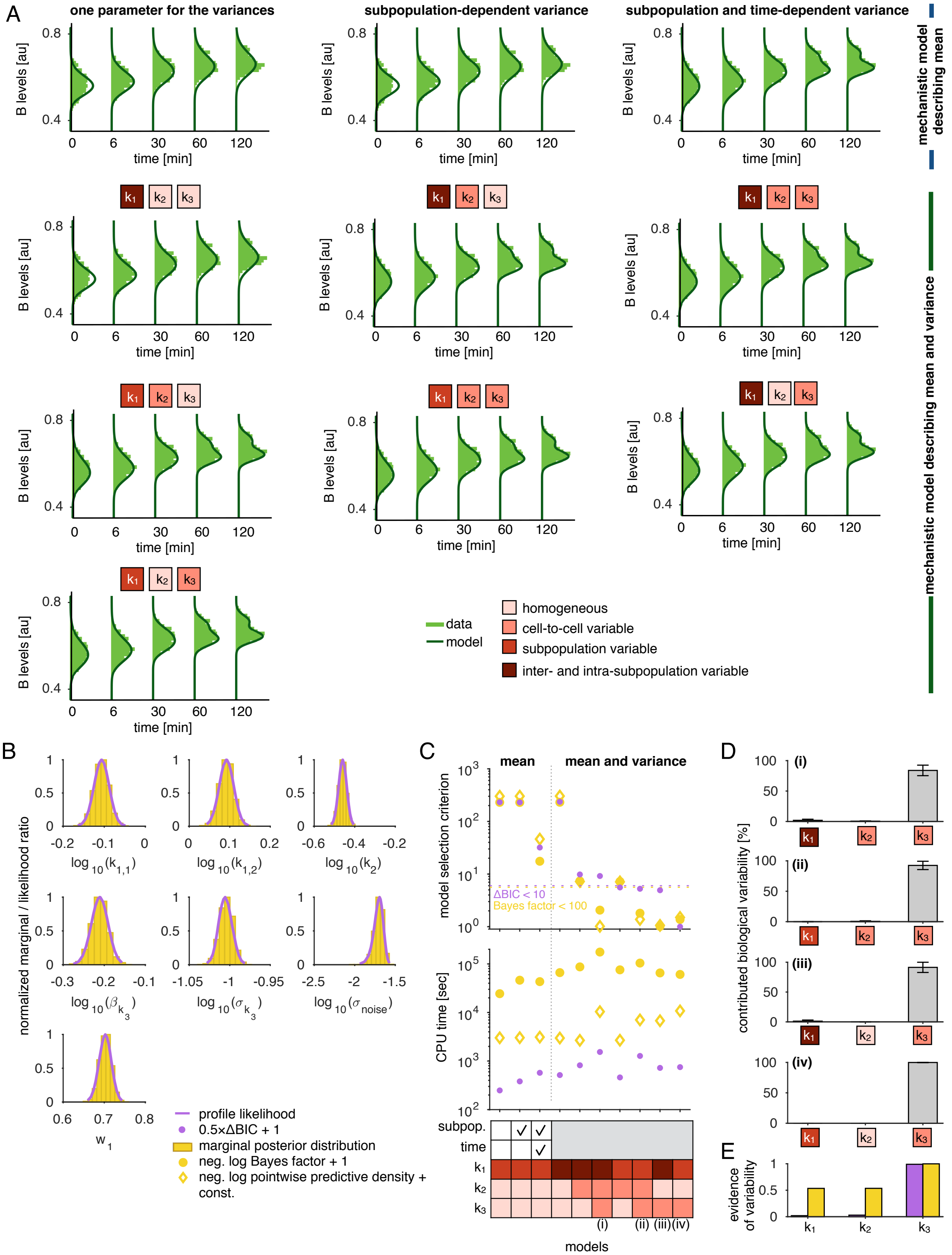

Figure S1, related to Figure 3. Analysis of the models for the conversion process.

(A) Fitted models for the conversion process. Upper part: Models accounting only for the means using RREs according to Hasenauer et al. (2014). Lower part: Models accounting for the means and variances using sigma-point approximations.

(B) Normalized marginal posterior distribution computed from samples of the posterior distribution and likelihood ratio obtained by profile likelihoods for all parameters.

(C) Model selection criteria and required computation times for all models. Lower values indicate a higher evidence for the corresponding model. The horizontal dotted lines indicate the cutoff corresponding to a BIC difference of 10 and a Bayes factor of 100.

(D) Contribution to overall cell-to-cell variability of the observable for the models with Bayes factor $<$ 100. The errorbars indicate deviation over time points.

(E) Evidence for variabilities in parameters computed based on BIC weights (left, purple) and marginal likelihoods (right, yellow).

(F) Predicted single-cell trajectories based on indicated measurements using estimated population parameters as priors. Shaded areas indicate the standard deviation which was evaluated for samples of the posterior distribution.

Figure S2

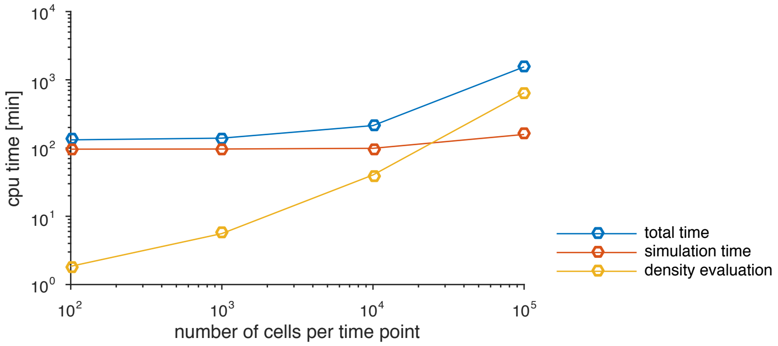

Figure S2, related to Figure 3. Computation time of the method for the conversion process. The overall computation time for 10 starts is depicted for varying number of measured cells per time point. The circles indicate the mean for three replicates. Different contributions to the overall computation time needed are shown: The time needed for the evaluation of the values for the cells under the density of the mixture distribution (yellow) and the time needed for simulation (orange).

Figure S3

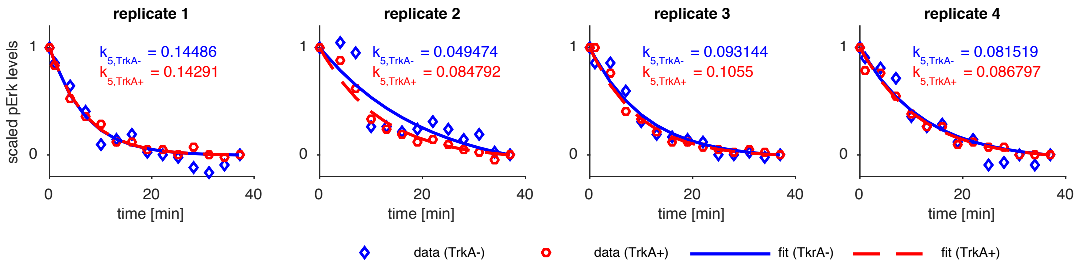

Figure S3, related to Figure 5. Erk1/2 dephosphorylation in TrkA- and TrkA+ subpopulations. Erk1/2 phosphorylation was induced in both neuronal subgroups TrkA- (blue) and TrkA+ (red) by an 1 h treatment with the combined stimuli NGF (acting on TrkA+ neurons) and GDNF (acting on TrkA- neurons expressing the GDNF receptor Ret). Subsequent inhibition of Mek by U0126 induced a pErk1/2 decline. Data of four individual experiments are shown with estimated exponential decay fit. The corresponding values k_5 for the dephosphorylation are noted for both subpopulations.

Figure S4

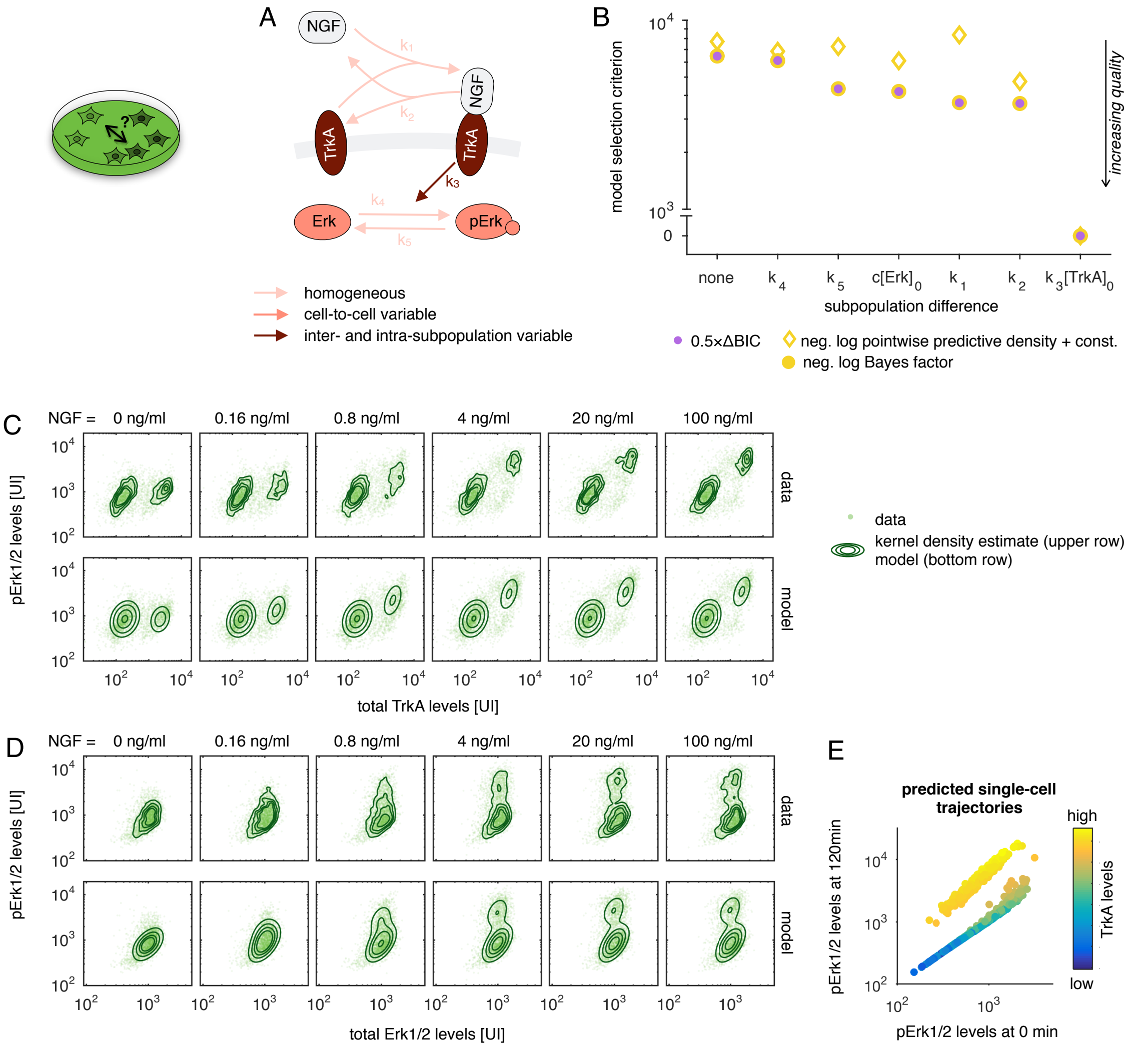

Figure S4, related to Figure 5. Data and fit for NGF-induced Erk1/2 signaling on PDL.

(A) Pathway model with color-coding of the variability of cellular properties.

(B) Comparison of models, which account for one or no difference between subpopulations, using BIC, log pointwise predictive densities, and Bayes factors.

(C,D) Data and fit for combined measurements of (C) TrkA and pErk1/2 levels and (D) Erk1/2 and pErk1/2 levels. The upper rows illustrate the data together with a kernel density estimate. The bottom rows visualize the data together with the contour lines of the hierarchical model.

(E) Predicted single-cell trajectories for the optimal parameter values, showing the relation between pErk1/2 levels in steady state (0 min) and after stimulation with NGF (120 min). The color of the cells indicates the TrkA level, which is assumed to be constant over time.

Figure S5

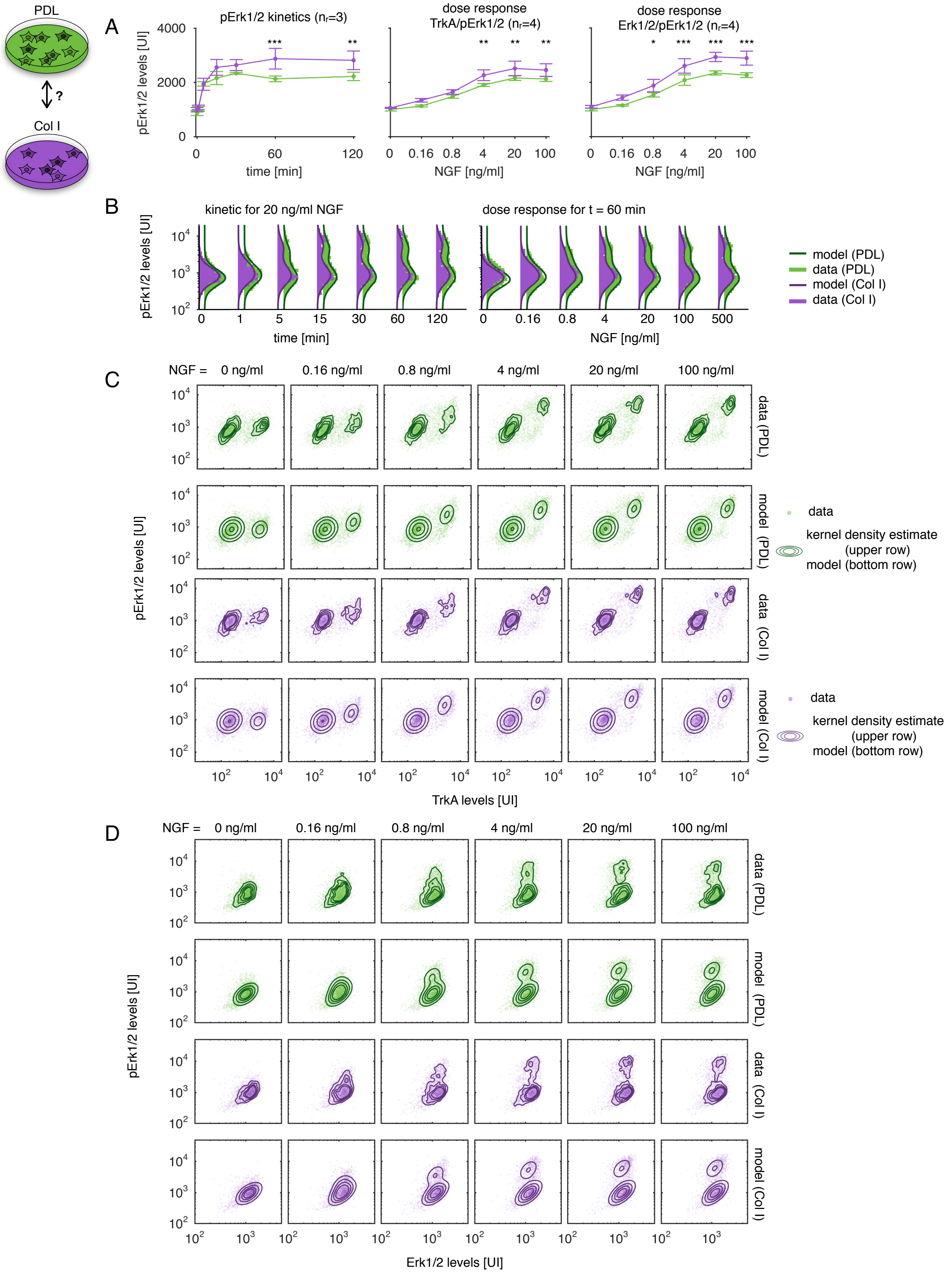

Figure S5, related to Figure 6. NGF-induced Erk1/2 signaling on different extracellular scaffolds.

(A) Mean response to NGF stimulation on Col I compared to PDL. A two-way ANOVA showed significant differences ($p < 0.01$) between the extracellular scaffolds for each experiment. Significances for individual time points/doses obtained by Sidak's multiple comparisons test are indicated by * ($p < 0.05$), ** ($p < 0.01$), and *** ($p < 0.001$).

(B-D) Data and fit for NGF-induced Erk1/2 signaling on different extracellular scaffolds.

(B) pErk1/2 kinetics and dose responses for PDL (green) and Col I (purple).

(C,D) Multivariate measurements of (C) pErk/TrkA levels and (D) pErk/Erk levels. The upper rows illustrate the data together with a kernel density estimate. The bottom rows visualize the data together with the contour lines of the hierarchical model, accounting for differences in Erk1/2 levels, Erk1/2 dephosphorylation, and cellular TrkA activity.

Figure S6

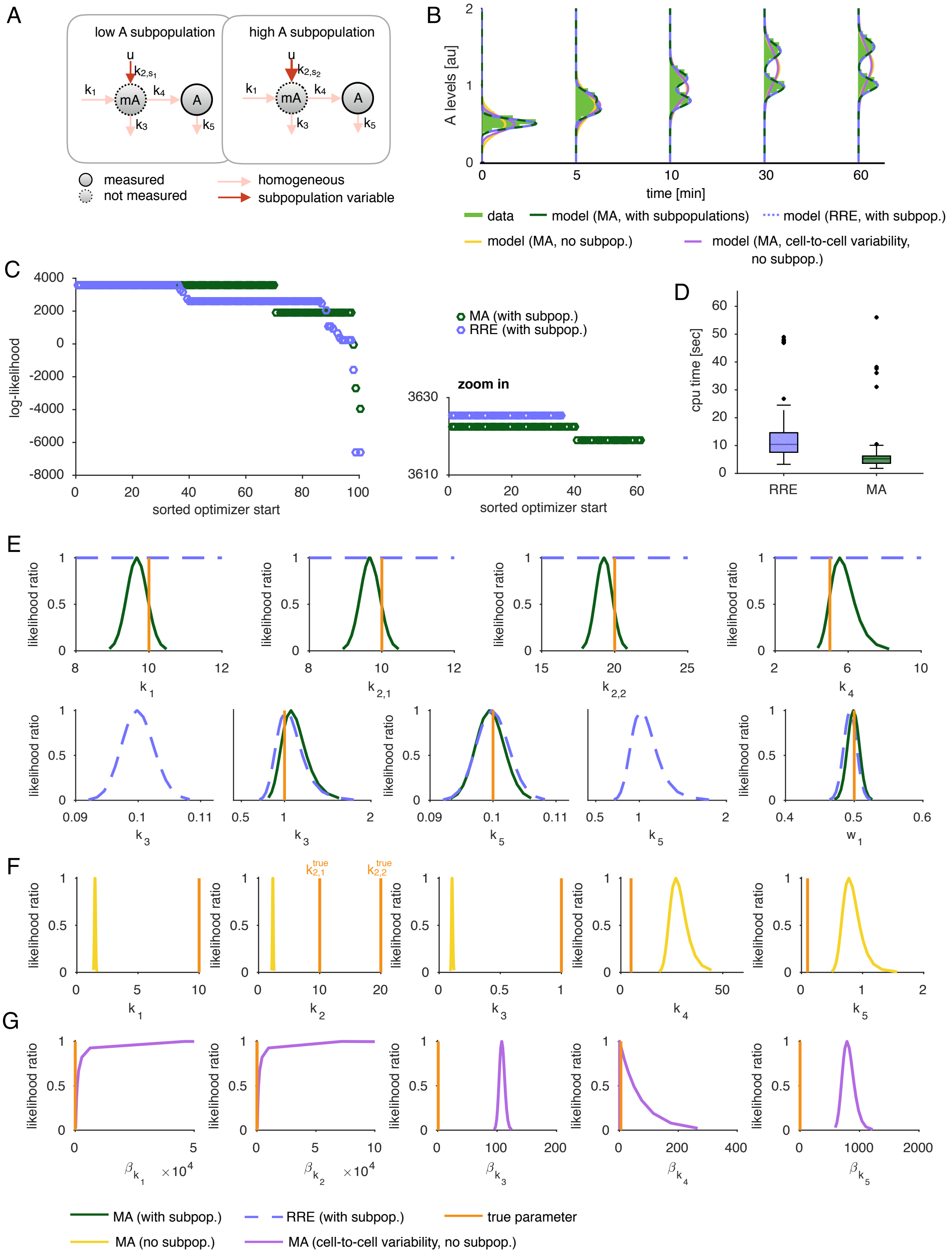

Figure S6, related to STAR methods, Accounting for intrinsic noise. Analysis of a stochastic two stage gene expression model.

(A) Illustration of the system.

(B) Data and fitted models for the moment-closure approximation (MA), for the case of accounting for subpopulation structures and disregarding subpopulation structures, and reaction rate eqations (RRE).

(C) Log-likelihood values for 100 optimization starts sorted decreasingly for MA (green) and RRE (blue). The zoom in shows the 60 best optimization runs.

(D) Boxplot for the CPU time needed for one optimizer start.

(E) Profile likelihoods of the parameters for the models capturing the subpopulation structure.

(F) Profile likelihoods of the parameters for the model using the MA without accounting for subpopulations.

(G) Profile likelihoods of the means rates for the model using the MA, accounting for cell-to-cell variability of all parameters but not for subpopulations. This corresponds to the method proposed by Zechner et al. (2012). Note that the range in x-direction differs for subplots (E)-(G).