HELMHOLTZ ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE UNIVERSITÄT MÜNCHEN

Scalable Simulation and Optimization Methods for Differential Equation Models describing Biochemical Reaction Networks

Fabian Fröhlich

April 2018

TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Mathematik — Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

Scalable Simulation and Optimization Methods for Differential Equation Models describing Biochemical Reaction Networks

Fabian Valentin Karl-Hanno Gero Maria Yoshio Fröhlich

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende:

Prof. Dr. Christina Kuttler

Prüfer der Dissertation:

1. TUM Junior Fellow Dr. Jan Hasenauer
2. Prof. Dr. Oliver Junge
3. Prof. Dr. Nicole Radde

Die Dissertation wurde am 10.01.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 19.03.2018 angenommen.

Acknowledgments

I am very grateful to my supervisor Jan Hasenauer for his outstanding support and scientific guidance. I thank Fabian Theis for his guidance and for providing an invaluably rich scientific environment at the Institute of Computational Biology.

I thank all former and current members of the Institute of Computational Biology for an unforgettable time. In particular, I thank all the former and current members of the Data-Driven Computational Biology group, namely Anna Fiedler, Atefeh Kazeroonian, Benjamin Ballnus, Carolin Loos, Daniel Weindl, Elba Raimundez, Leonard Schmiester, Nick Jagiella, Paul Stapor, Sabine Hug, Sabrina Krause, Sabrina Hross and Yannik Schälte for inspiring discussions and memorable gatherings. I thank Dennis Rickert for his patience with my ever increasing demand for computation time on our cluster and Marianne Antunes, Elisabeth Noheimer and Sabine Kunz for their exemplary administrative support.

I thank Joachim Rädler for excellent guidance during my thesis comittee meetings and project related discussions. I thank Julio Banga and Alejandro Villaverde and the other member os the Bioprocess Engineering group for interesting discussions and for hosting me in Vigo. I thank Thomas Kessler, Christoph Wierling, Bodo Lange and Hans Lehrach and everybody else at ALACRIS Theranostics GmbH for the unique opportunity to work on large scale models of cancer related signalling. I thank Phillip Thomas and Ramon Grima for the inspiring discussions on stochastic modeling. I thank Barbara Kaltenbacher for discussions and insights regarding the use of adjoint methods. Moreover, I thank all my other collaboration partners, especially Anita Reiser, Laura Fink, Daniel Woschee, Andreas Raue, Thomas Ligon and Matthias Heinig for their valuable contributions.

I thank Ulrike Gaul and Erwin Frey for founding and coordinating the Quantitative Biosciences Munich graduate school and all my fellow Qubies for four extraordinary years. I thank Michael Mende, Filiz Civril, Markus Hohle, Julia Schlehe and Mara Kieke for doing their best making these four years extraordinary. I thank the Deutsche Forschungsgemeinschaft for funding my fellowship and for funding the QBM graduate school.

Last, I want to thank my family, in particular my parents Gero Decher and Gabi Fröhlich for their unending support and constant encouragement. I thank Nora Fröhlich for being a great little sister. Finally, I thank Judith Pramsohler for supporting me to the fullest through all the ups and downs of my thesis.

Abstract

The field of systems biology strives for a holistic understanding of biological systems. A common approach that contributes to this holistic understanding is the construction of comprehensive, accurate differential equation models of the underlying biochemical processes. The differential equations usually describe the temporal evolution of the abundance of molecular species and depend on kinetic parameters such as reaction rates. These parameters have to be estimated from experimental data, which can be achieved by model-constrained optimization. Comprehensive models often need to describe a large number of molecular species and parameters to capture the complexity of the underlying biochemical processes. Moreover, for an unbiased description, the effect of stochastic fluctuations needs to be taken into account. These two requirements result in a high, often intractable, computational cost for simulation and model-constrained optimization.

In this thesis, I developed scalable methods for simulation and model-constrained optimization. Firstly, I formulated and developed methods for model-constrained optimization problem, i.e., parameter estimation, for ordinary differential equation approximations to the statistical moments of the solution of stochastic models. These approximations are defined by the system size expansion and the moment-closure approximation. I applied the developed methods to simulation examples and biological applications and was able to obtain accurate parameter estimates – even if stochastic fluctuations are non-negligible. Secondly, I developed and applied methods for parameter estimation that improve the scalability with respect to the number of parameters by using adjoint sensitivity analysis. Fourthly, I applied sparse solvers to reduce the computational cost of numerical simulations by multiple orders of magnitude for models that describe a large number of molecular species. Lastly, I formulated the parameter estimation problem for models with discrete events that approximate fast timescales and developed methods that enable efficient parameter estimation by using forward sensitivity analysis.

I provide an efficient, easy-to-use implementation of all developed methods in the open-source toolbox `AMICI`, which is employed by a growing number of researchers in a variety of different projects. Moreover, I apply the methods to a series of simulation and real biological applications, including a model with over 1200 molecular species and over 4000 parameters. The parameter estimation for this large-scale model was previously intractable, which underlines the importance and practical relevance of the developed methods. Taking all contributions together, they improve the efficiency and scalability of simulation and model-constrained optimization for stochastic and deterministic differential equation models. This enables the consideration of more comprehensive models, which is an important step towards a more holistic understanding of biological systems.

List of contributed articles

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

(ii) F. Fröhlich, F.J. Theis, J.O. Rädler, J. Hasenauer. **Parameter estimation for dynamical systems with discrete events and logical operations.** Bioinformatics 33(7):1049-1056 (2017).

(iii) F. Fröhlich, B. Kaltenbacher, F.J. Theis, J. Hasenauer. **Scalable parameter estimation for genome-scale biochemical reaction networks.** PLoS Computational Biology 13(1):e1005331 (2017).

(iv) F. Fröhlich, T. Kessler, D. Weindl, A. Shadrin, L. Schmiester, H. Hache, A. Muradyan, M. Schütte, J. Lim, M. Heinig, F.J. Theis, H. Lehrach, C. Wierling, B. Lange, and J. Hasenauer. **Efficient parameterization of large-scale mechanistic models enables drug response prediction for cancer cell lines.** bioRxiv:174094.

I am the sole first author and author in charge of all articles listed above.

Contents

Chapter 1

Introduction

In systems biology, differential equation models have become a standard tool for the analysis of biochemical reaction networks (Klipp et al., 2005). The model equations can be derived from biological knowledge of the underlying biochemical processes (Kitano, 2002a,b) and thus allow the systematic integration of prior knowledge. Such models are particularly valuable as they can be used to predict the temporal evolution of latent variables as well as derived quantities (Adlung et al., 2017; Buchholz et al., 2013). Moreover, they provide executable formulations of biological hypothesis and therefore allow the direct verification and falsification of these hypotheses (Hross et al., 2016; Hug et al., 2016; Intosalmi et al., 2016; Molinelli et al., 2013; Schilling et al., 2009; Toni et al., 2012), thus deepening the biological understanding. Furthermore, differential equation models have been applied to derive model-based biomarkers (Eduati et al., 2017; Fey et al., 2015), that enable the personalized design of targeted therapies in precision medicine.

In order to construct predictive models, model parameters have to be estimated from experimental data. This estimation task typically requires the repeated numerical solution to the model equations. Consequently, parameter estimation is computationally demanding if the required computation time for the numerical solution is high. For most applications, a strongly simplified, deterministic description of the underlying biological processes, which is computationally undemanding to simulate, is sufficient to construct predictive models (Kitano, 2002a; Maiwald et al., 2016; Snowden et al., 2017; Transtrum and Qiu, 2016). Such simplified models can be obtained by lumping multi-step reactions together to one-step reactions (Dano et al., 2006), by ignoring the influence of certain biological processes and by ignoring the influence of stochastic fluctuations. Such simplifications yield small- to medium-scale ordinary differential equation (ODE) models for which standard algorithms implemented in established toolboxes often work out-of-the-box (Hoops et al., 2006; Raue et al., 2015; Somogyi et al., 2015).

For systems that are studied in an isolated context, e.g., in a single cell line, the lumping and ignoring of processes might be appropriate. Yet, these simplifications might not be appropriate if a large number of different contexts is considered. For example, when multiple cell lines or patients are considered, certain lumping or neglecting of reactions

might not be appropriate for all cell lines or patients (Adlung et al., 2017; Santos et al., 2007; Yao et al., 2016). Consequently, comprehensive descriptions are required when building predictive models that generalize to a large number of different contexts. Such predictive, generalizing models are particularly important for applications such as precision medicine, where the model must accurately predict treatment outcomes for many different patients. These comprehensive models can easily describe thousands of molecular species involved in thousands of reactions with thousands of parameters. Most algorithms cannot be applied at such a scale as they are prohibitively computationally expensive (Babtie and Stumpf, 2017; Schillings et al., 2015). Furthermore, the stochasticity of reactions can have substantial impact on the behavior of a systems (Ramaswamy et al., 2012). When this is the case, using an ODE model that ignores stochasticity, is not an appropriate simplification and might lead to a bias in the parameter estimates and thus possibly also in the predictions. Yet, the numerical simulation for stochastic models is typically computationally much more demanding than the simulation of ODE models. Therefore, the application of stochastic models is currently limited to small-scale systems.

In my doctoral research, I developed scalable methods that render parameter estimation for ODE models describing many molecular species and many parameters tractable. In particular, I identified bottlenecks in the established numerical methods and developed better scalable alternatives. Moreover, I evaluated how more accurate ODE approximations to stochastic models can be used to render scalable algorithms accessible – even if the effect of stochastic fluctuations is non-negligible. In the next section, I first provide an introductory overview of stochastic and ODEs models for biological systems and describe bottlenecks in parameter estimation for these model classes. In Section 1.2, I give a brief overview of how I addressed these critical points in the course of my thesis work. In Section 1.3, I outline the structure of the following chapters.

1.1 Parameter Estimation for Differential Equation Models

To build predictive models, the model parameters have to be estimated from experimental data. The parameter estimation problem is usually formulated as a differential equation constrained optimization problem. In this optimization problem, an objective function, describing the distance between measurements and simulation, is minimized. The differential equation constraint is required, as the evaluation of the objective function requires the solution to the model equations. In general, the objective function for ODE models (Raue et al., 2013b) are different to those for stochastic models (Andreychenko et al., 2011; Hasenauer et al., 2011). In both cases, the corresponding optimization problem is typi-

cally non-convex and can possess multiple local minima. Consequently, the optimization problem can only be reliably solved when global optimization methods, hybrid global-local optimization methods or repeated local optimization methods are applied (Balsa-Canto and Banga, 2011; Banga, 2008; Egea et al., 2007, 2014; Karr et al., 2015; Moles et al., 2003; Raue et al., 2013b; Reinker et al., 2006; Rodriguez-Fernandez et al., 2006). The efficiency of most of these methods can be improved by providing the gradient of the objective function to the optimization method (Griewank and Walther, 2008; Nocedal and Wright, 2006; Raue et al., 2013b). Yet, the objective function and the objective function gradient are typically not available in closed form, but have to be computed numerically. For large-scale models the computational cost of computing the objective function and it's gradient is high, which makes parameter estimation computationally demanding. Depending on the class of the employed model and simulation algorithm, the computation time will depend on different features of the underlying model (see Figure 1.1), which I will discuss in detail in the following.

1.1.1 Stochastic and Deterministic Modeling Approaches

Due to the discreteness of molecular species and the stochasticity of the reactions they are involved in, biochemical processes are usually described using continuous time Markov chains (CTMCs) (Norris, 1998). This descriptions requires the assumption of a thermally equilibrated and well-mixed environment and the assumption that biological processes are memoryless. The CTMC describes the transition of the system between different states as stochastic firing of different reactions. For every reaction, the propensity defines the rate at which the reaction occurs. The stochastic simulation algorithm (SSA) (Gillespie, 1977) provides exact simulations by randomly sampling the next occurring reaction as well as the reaction time according to the reaction propensities. Consequently, the computational cost of the algorithm is determined by the evaluation time of every propensity as well as the sum of reaction propensities, which determines the time-stepping. Typically both quantities increase with the number of modeled reactions. Moreover, due to the sampling of the next reactions, the evaluation of the algorithm and thus also of the objective function is stochastic. As most local optimization schemes require deterministic evaluation of objective function and gradient, it can be necessary to average over multiple realizations of stochastic simulations. The averaging will reduce the stochasticity, but the repeating of evaluations will further increase the computational cost.

The temporal evolution of the probability to be in a specific state at a specific time is described by the Chemical Master Equation (CME) (Gillespie, 1992). The CME is a set of coupled ODEs for every attainable state of the system. As the number of attainable

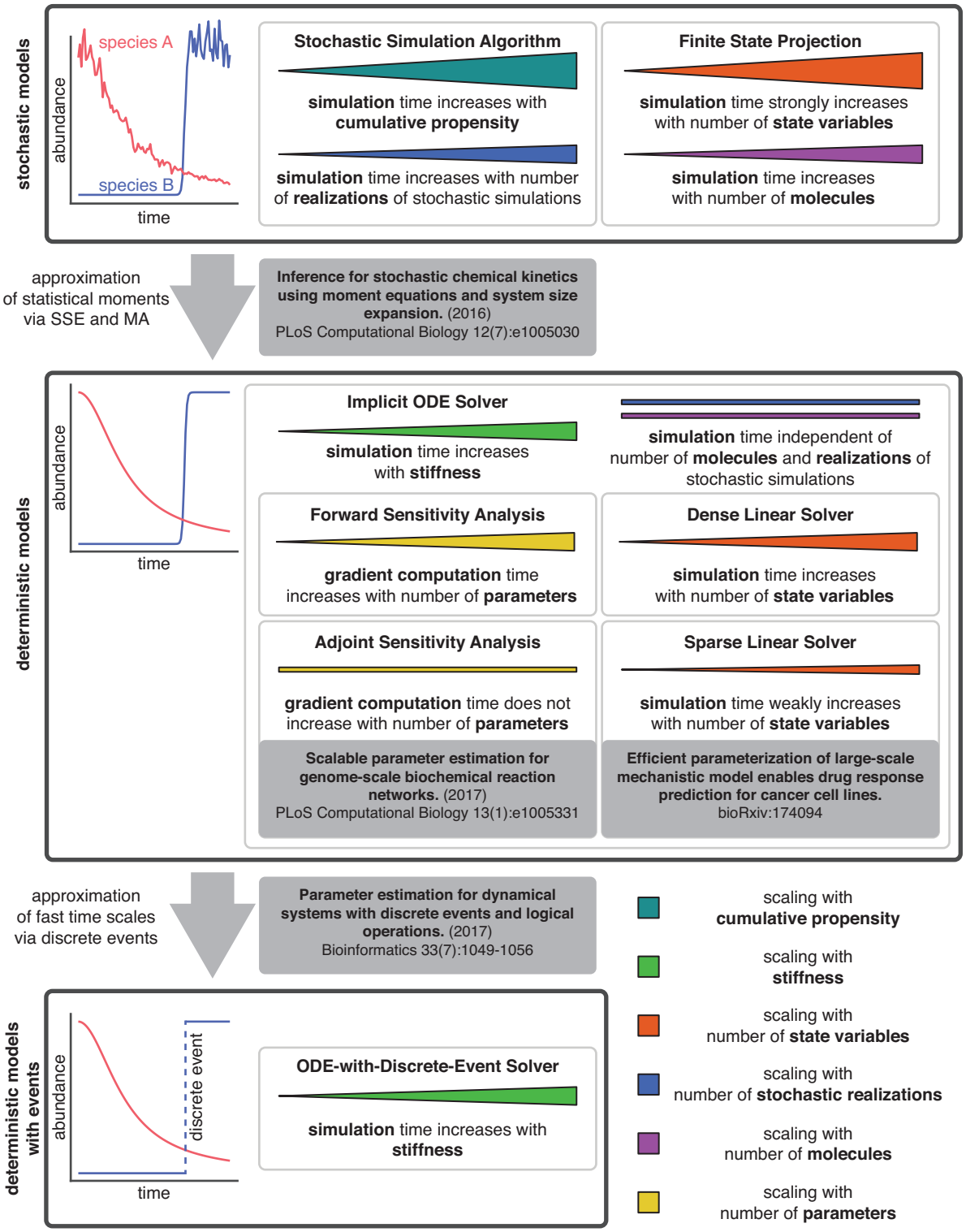


Figure 1.1: Schematic overview of the scalability of various algorithms for simulation and sensitivity computation. Width and height of triangles are not indicative of scaling order in a strongly mathematical sense, but only used to visually represent the statements in the text. Main (first author) contributions are indicated by grey boxes.

states is often large and the propensities are often non-linear, closed form solutions are rarely available. Furthermore, due to large or infinite number of states, a numerical solution is often intractable. To circumvent these problems, the finite state projection (FSP) (Munsky and Khammash, 2006), which approximates the solution by truncating the equations to a reasonable number of states, can be employed. As the CME can be deterministically evaluated, the approach is usually more efficient than the SSA as no repeated evaluations are necessary (Neuert et al., 2013). However, the computational cost increases exponentially with the number of modeled species.

Despite several improvements to the SSA (Anderson and Higham, 2012; Gillespie, 2001; Menz et al., 2012; Ramaswamy et al., 2009; Rathinam et al., 2003; Sanft et al., 2011) and the FSP algorithm (Kazeev et al., 2014; Mateescu et al., 2010; Sunkara and Hegland, 2010), the methods remain computationally demanding for models with many state variables. For larger models, approximative methods, which describe the temporal evolution of statistical moments using ODEs, can be used. These include the system size expansion (SSE) (Grima, 2010, 2011; van Kampen, 2007) and moment-closure approximation (MA) (Ale et al., 2013; Engblom, 2006a; Gillespie, 2009; Lee et al., 2009). The accuracy of the approximations can theoretically be improved by including higher order terms. Although the resulting number of differential equations increases exponentially with the employed order, it only scales polynomially with the number of molecular species, where the leading order is equal to the highest simulated moment order (Engblom, 2006b). As the approximation schemes are deterministic, no repeated simulations are necessary. This allows the numerical simulation of small- to medium-scale models when first or second order approximations are employed.

The first order of the system size expansion of the mean is the Reaction Rate Equation (RRE), which is commonly used throughout the systems biology community (Klipp et al., 2005). The RREs are a coupled system of ODEs and can thus be solved deterministically and the number of equations depend linearly on the number of described molecular species. This enables the numerical simulation of small- to large-scale models. Still, the specific features of biological models require tailored algorithms, which I will discuss in the following.

1.1.2 Numerics of Ordinary Differential Equations

The timescales of biochemical processes span multiple orders of magnitude (Shamir et al., 2016). As comprehensive models often cover a large variety of different biological processes, they are particularly prone to possess multiple timescales (Smallbone and Mendes, 2013).

This results in the stiffness of corresponding ODEs. As the stiffness of the equations typically depend on the choice of parameters, it is rarely possible to assess the stiffness *a priori*. Consequently, it is always advisable to use implicit solvers, which can adequately handle stiffness, for parameter inference (Gonnet et al., 2012). For most implicit solvers, the computational complexity is determined by the cost of solving a linear problem defined by the Jacobian of the right hand side of the ODE (see Section 2.3.1). The computation time of standard direct solvers, such as LU decomposition, scales cubically with respect to the number of state variables. For certain ODE classes, tailored solvers, that exploit the structure of the Jacobian, have been used to reduce the computation time. For example, for finite element discretizations of partial differential equations, banded solvers have been developed (Thorson, 1979). Yet, it is unclear whether the Jacobians of ODE models of biological systems generally possess such a banded structure.

When the separation between timescales is too strong, it can become numerically intractable to simulate the system for reasonable time intervals. For such models, the approximation of fast timescales by discrete events has been proposed (Le˜Novère, 2015; McAdams and Shapiro, 1995). Besides fast biological processes, such processes can also be used to model input profiles that are assumed to have immediate effect, such as bolus injections. Discrete events give rise to hybrid or switched systems which are frequently used in engineering disciplines, such as circuit simulation and process control applications (Antsaklis et al., 1993; Lennartson et al., 1996; Liberzon and Morse, 1999). Sill, optimization in the context of hybrid systems is generally considered to be challenging (Barton et al., 1998).

1.1.3 Sensitivity Equations

Providing an accurate gradient to the objective function can improve the efficiency of most optimization algorithms (Griewank and Walther, 2008; Nocedal and Wright, 2006; Raue et al., 2013b). For differential equation constrained optimization problems, the gradient of the objective function can be computed based on the parametric derivative of the solution to the differential equation. These derivatives are often called the sensitivities of the model. Several approaches to compute sensitivities for ODE models exist, including the direct approach via forward sensitivity analysis (Dickinson and Gelinas, 1976; Kokotovic and Heller, 1967) as well as finite differences (Milne-Thompson, 1933). For models with parameter dependent discrete events, the forward sensitivity analysis requires updates at all discontinuity points (Barton et al., 1998; Rozenvasser, 1967). The computational cost of forward sensitivity analysis and finite differences scales linearly with the number of parameters, which can be prohibitive for large-scale models as they can possess thousands

of parameters. The alternative, adjoint approach, that computes the objective function gradient via adjoint sensitivity analysis, has long been deemed to be computationally more efficient for systems with many parameters (Kokotovic and Heller, 1967). The adjoint approach is supposed to be more efficient when the number of parameters is high, as the dependence of computation time on the number of parameters is theoretically negligible. In other research fields, e.g., for partial differential equation constrained optimization problems, adjoint sensitivity analysis (Hindmarsh et al., 2005) has been adopted in the past decades. In contrast, in the systems biology community there are only isolated applications of adjoint sensitivity analysis (Fujarewicz et al., 2005; Lu et al., 2008, 2012).

1.2 Overview and Contribution of this Thesis

The application of parameter estimation methods to comprehensive, large-scale models requires efficient algorithms. In particular, the scaling with the number of modeled species and model parameters is crucial. For ODE models as well as stochastic models key issues are:

(i) The consideration of large-scale stochastic models is generally limited by the high computational cost of exact numerical simulations. However, the statistical moments of the solution can be approximated by ODEs. The approximation renders the simulation time independent of the cumulative propensity, number of realizations of stochastic simulations and number of molecules. Moreover, it also renders all approaches for ODE models, which are outlined in the following, applicable. Yet, the approximative solution can be biased. It is unclear if and how this bias affects parameter estimation.

(ii) For ODE models the scalability with respect to parameters depends primarily on how the gradient of the objective function is computed. Adjoint sensitivity analysis has been proposed, but a general implementation is currently lacking. Moreover, a rigorous evaluation of the concrete scaling of with respect to the number of parameters for biological models has not been carried out. Thus, the superiority of the approach for this problem class remains unclear.

(iii) For ODE models the scalability with respect to the number of state variables depends primarily on the choice of linear solvers. Tailored solvers that exploit the structure of the Jacobian have been proposed for certain model classes, but no appropriate solver has been established for models of biochemical reaction networks.

(iv) Comprehensive models are prone to stiffness. This can be mitigated by the introduction of discrete events. Yet, the support for sensitivity computation for models with discrete events is poor in most state-of-the-art toolboxes.

I addressed these issues in the main contributions of my thesis. In the following, I provide a brief descriptions of these contributions, while the corresponding articles are provided in the Appendices A-D. These contributions resulted in scalable algorithms that enable reliable parameter estimation for models with hundreds of state variables and thousands of parameters.

- **Assessment of the effect of approximation bias on parameter estimates.** To address issue (i) I assessed how the approximation order influences the error of parameter estimates. I demonstrated that certain parameters can only be reliably estimated when higher order approximations are used. Moreover, I showed that the benefit from higher order approximation is most pronounced for an intermediate strength of stochastic fluctuations. Furthermore, I showed that the identifiability of parameters improves if higher order moment approximations are employed, even when only population average data is available.

- **Facilitation of scalable parameter estimation for models with many parameters.** To address issue (ii) I developed and implemented an easy-to-use, efficient algorithm for adjoint sensitivity analysis in the open-source toolbox `AMICI`. I used this implementation to demonstrate that the computation of the gradient using adjoint sensitivity analysis for models of biochemical reaction networks practically does not depend on the number of parameters. Moreover, I showed that this results in a reduction of computation time compared to forward sensitivity analysis already for medium-scale models. Furthermore, I demonstrated that the achieved speedup increases with respect to both the number of parameter and the number of state variables.

- **Facilitation of scalable parameter estimation for models with many state variables.** To address issue (iii) I demonstrated that the sparsity in the Jacobian of models of biochemical reaction networks can be exploited by using solvers that were designed for circuit simulation (Davis and Palamadai Natarajan, 2010) and implemented an interface to these solvers in the open source toolbox `AMICI`. For a model with over 1200 state variables, I was able to achieve a reduction in gradient computation time of over 100 fold.

- **Facilitation of sensitivity analysis for models with discrete events.** To address issue (iv) I implemented the previously derived update formulas for forward sensitivities in the open source toolbox `AMICI`. I demonstrated that parameter estimation using forward sensitivity analysis is more efficient compared to finite differences. Moreover, I formulated and applied the parameter estimation problem for events that were experimentally observed. By this, I enabled the consideration of new data-types such as time-to-event data with ODE models, which is not possible in any other software toolbox.

- **Application of parameter estimation for models with many state variables and many parameters.** To underline the practical relevance of issues (ii) and (iii) and the applicability of the corresponding developed methods, I applied the methods to a model with over 1200 state variables and over 4000 parameters. I demonstrated that the parameterized model outperforms all investigated statistical models in the prediction of the drug response of different cancer cell lines. I validated the predictions for response to combination treatment on additional experimental data. This application is highly relevant for precision medicine and could at some point improve personalized cancer treatments.

A detailed summary of all main contributions is provided in Chapter 3.

Other Contributions

In addition to the main contributions of this thesis, I had five side-projects which were published in the following articles:

- F. Fröhlich, S. Hross, F.J. Theis, J. Hasenauer. **Radial basis function approximations of Bayesian parameter posterior densities for uncertainty analysis.** Lecture Notes in Computer Science 8859:73-85 (2014).

  In (Fröhlich et al., 2014a), we presented novel methods for analysis of Bayesian posterior probability densities of parameter estimates using weighted sums of radial basis functions. We employed lattice generation algorithms, adaptive interacting particle sampling schemes as well as Markov chain based sampling schemes for the generation of approximation nodes. We used several weighting scheme and compare the computational efficiency of combinations of weighting and node generation schemes on different application examples. Our analysis demonstrated that the novel method can yield an expected l_2 approximation error in marginals that is several

orders of magnitude lower compared to Markov chain Monte Carlo based approximations. Consequently, the method may allow for a drastic reduction of the number of model evaluations without a loss in accuracy, which facilitates the analysis of uncertainty for problems with high computational complexity.

- F. Fröhlich, F.J. Theis, J. Hasenauer. **Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more.** Lecture Notes in Computer Science 8859:61-72 (2014).

  In (Fröhlich et al., 2014b), we analyzed and compared bootstrapping, profile likelihood, Fisher information matrix, and multi-start based approaches for frequentist uncertainty analysis. The analysis was carried out on two models with structurally non-identifiable parameters. We demonstrated that bootstrapping, multi-start optimization, and Fisher information matrix based approaches produce deceptive results for parameters which are structurally non-identifiable. We used geometric arguments to provide a simple and intuitive explanation for the deceptive results. These results are highly relevant as many research groups frequently use bootstrapping, Fisher information matrix and multi-start based approaches without prior assessment of the structural identifiability of parameters.

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

  In (Kazeroonian et al., 2016), we presented the `CERENA` toolbox which automatically generates model equations for the moment-closure approximation and the system size expansion from a specification of the underlying biochemical reaction network. The automated generation of model equations is crucial to address issue (i) as the methods have to be evaluated on several application examples. The `CERENA` toolbox generates native C code for every model an uses simulation routines that were also used in the `AMICI` toolbox. Thus `CERENA` employs the highly optimized simulation routines as well as respective forward and adjoint sensitivity analysis routines developed in the scope of `AMICI`. Besides the routines for moment-closure approximation and system size expansion, `CERENA` also implements algorithms to solve the Finite State Projection as well as the stochastic simulation algorithm. The implemented algorithms support multi-compartment models as well as time dependent propensities, which is not possible in most other state-of-the-art toolboxes. To complement the simulation routines, `CERENA` also implement various visualization routines to display simulation results. In the manuscript, we evaluated and analyzed the approximation error as well as the efficiency of the various implemented methods on a

set of application examples. We found that for the moment-closure approximation, the approximation on average decreases with respect to the order of the considered statistical moments. Moreover, we demonstrated that the computation time for the moment-closure approximation and the system size expansion in `CERENA` is over one order of magnitude lower than in other state-of-the-art toolboxes. `CERENA` was essential for addressing issue (i) as all model equations were generated using the toolbox.

- T. Ligon, F. Fröhlich, O. T. Chis, J. R. Banga, E. Balsa-Canto, J. Hasenauer. **GenSSI 2.0: Multi-experiment structural identifiability analyisis of SBML models.** Bioinformatics, btx735

  In (**?**), we presented the second version of the `GenSSI` toolbox which can perform identifiability analysis for ODE model parameters. `GenSSI` implements a generating series approach to assess the structural parameter identifiability. In the updated version of GenSSI we implemented support for newer versions of MATLAB, for the Systems Biology Markup Language (SBML), for parameter and state transformations as well as for multiple experiments. Moreover, we improved the computational efficency of the implemented algorithms and thus enabled the consideration of larger models. Structural non-identifiability results in a singularity of parameter estimation problem and should thus be assessed beforehand. Consequently, the algorithms implemented in `GenSSI` are an important preprocessing step when addressing issues (i)-(iv).

- P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hross, F. Fröhlich, J. Hasenauer. **PESTO: Parameter EStimation TOolbox.** Bioinformatics, btx676

  In (Stapor et al., 2017), we presented the `PESTO` toolbox which can perform parameter estimation, uncertainty analysis and provides visualization routines for the respective results. `PESTO` implements local, global, and hybrid local-global optimization schemes for parameter estimation and frequentist and Bayesian methods for uncertainty analysis. I used `PESTO` for parameter estimation and uncertainty analysis in all contributions in this thesis. Consequently, the algorithms implemented in `PESTO` were essential to address issues (i)-(iv).

1.3 Outline

This is a cumulative dissertation based on the research that is published in my first-author articles. In this thesis I summarize the respective mathematical background in Chapter 2. In Section 2.1 and 2.2, I introduce modeling formalisms to derive ODE models for biochemical processes. In Section 2.3, I discuss the optimization problem which allows the estimation of parameters from experimental data for such models. Moreover, I discuss how the gradient of the corresponding objective function to the optimization problem can be computed efficiently. In Chapter 3, I provide a summary of all contributed articles. In Chapter 4, I discuss possible extensions to the presented work. The full text of my first-author articles is attached in the Appendices A-D.

Chapter 2

Methods

This thesis focuses on ODE models for biochemical reaction networks. In the following, I will introduce the modeling formalism that gives rise to the respective ODE models. Subsequently, I will introduce all methods for parameter estimation and uncertainty analysis for ODE models that were used in my thesis work.

2.1 Modeling of Biochemical Reaction Networks

This thesis considers a set of d_R reactions, involving M chemical species confined in a reaction volume of size Ω. Denoting the set of reactants by $(X_1, ..., X_M)$, the r^{th} reaction can be written as

$$\sum_{i=1}^{M} \nu_{ir}^{-} X_i \xrightarrow{k_r(\theta)} \sum_{i=1}^{M} \nu_{ir}^{+} X_i.$$

Here $k_r(\theta) : \mathbb{R}^{d_\theta} \to \mathbb{R}$ is the reaction rate constant, which depends on the vector of parameters $\theta = (\theta_1, \ldots, \theta_{d_\theta})$. $\nu_{ir}^{\pm}$ are the integer stoichiometric coefficients and $\nu_{ir} = \nu_{ir}^{+} - \nu_{ir}^{-}$ denotes the net change in molecules of the i^{th} species due to firing of the r^{th} reaction.

In the following I will demonstrate how such a reaction formalism can be translated into an ODE model

$$\dot{x} = f(t, x, \theta), \quad x(t_0) = x_0(\theta);$$

on a finite time interval $\mathbb{T} = [t_0, t_f]$ with initial condition $x_0(\theta)$. For reaction rate equations f will describe the temporal evolution of the average concentrations $x_i = c_i = \mathbb{E}[\frac{n_i}{\Omega}]$ of the i^{th} species, where $n = (n_1, \ldots, n_M)$ is the vector of molecule numbers. For the moment-closure approximation and the system size expansion the vector $x \in \mathbb{R}^{d_x}$ will also contain higher order statistical moments of n or $\frac{n}{\Omega}$ and f will also describe the respective temporal evolution of these higher order moments. Accordingly, the length d_x of the vector x will depend on the considered orders of statistical moments.

2.1.1 Chemical Master Equation

Under well-mixed conditions the state of a biochemical reaction network can be characterized by the corresponding vector of molecule numbers n (van Kampen, 2007). The time-evolution of the probability $P(t, n, \theta) : \mathbb{T} \times \mathbb{R}^M \times \mathbb{R}^{d_\theta}$ of observing the system in state n at time t, then obeys the Chemical Master Equation (CME) (Gillespie, 1992)

$$\dot{P}(t, n, \theta) = \Omega \sum_{r=1}^{d_r} \Big[\hat{v}_r(n - \nu_r, \theta) P(t, n - \nu_r, \theta) - \hat{v}_r(n, \theta) P(t, n, \theta) \Big]. \tag{2.1}$$

Here, $\Omega \hat{v}_r(n, \theta)$ denotes the microscopic propensity function, i.e., the probability per unit time for reaction r to occur somewhere in the volume Ω.

The derivation of an analytical solution to the CME is usually intractable. For the numerical simulation of individual realizations, the Stochastic Simulation Algorithm (SSA) (Gillespie, 1977) can be used. The SSA computes individual realizations of the underlying stochastic process. This is achieved by initializing the simulation at some specified initial state n_0. The probability distribution $P_\xi(\xi)$ of the time ξ until the next reaction is then computed according to the cumulative propensity $\hat{v}_0(n, \theta) = \sum_r \hat{v}_r(n, \theta)$:

$$P_\xi(\xi) = \exp(-\hat{v}_0(n, \theta)\xi)\hat{v}_0(n, \theta).$$

The reaction that happens is also determined by chance, where the probability P_r of the r^{th} reaction to happen is given by the fraction

$$P_r(r) = \frac{\hat{v}_r(n, \theta)}{\hat{v}_0(n, \theta)}.$$

Evidently, the length of the time-step ξ will decrease with an increasing cumulative propensity $\hat{v}_0(n, \theta)$. This renders this approach problematic when models with a large number of reactions or individual reactions with high propensities are considered (Rathinam et al., 2003). Moreover, as previously discussed, multiple realizations of stochastic simulations can be necessary, which further increases the computational cost. This often renders the SSA computationally intractable.

Several more efficient approximation methods to solutions of the CME have been developed, including the Finite State Projection (Munsky and Khammash, 2006), the chemical Langevin Equation (Gillespie, 2000) and the Fokker-Planck Equation (Risken, 1996). For the purpose of this thesis, I focus on approximation methods for the statistical moments of the concentrations $\mu_i = \mathbb{E}[n_i/\Omega]$, and the corresponding covariances of the concentration

fluctuations about them, $\Sigma_{ij} = \mathbb{E}[(n_i/\Omega - \mu_i)\,(n_j/\Omega - \mu_j)]$.

2.1.2 Reaction Rate Equation

The most commonly used differential equation model to describe the mean concentration $c = \mu$ in biochemical reaction networks are Reaction Rate Equations (RREs). Under the assumption of the law of mass action, the RRE is defined by the reaction flux $v_r(c, \theta) : \mathbb{R}^M \times \mathbb{R}^{d_\theta} \to \mathbb{R}$ of the r^{th} reaction

$$v_r(c, \theta) = k_r(\theta) \prod_{i=1}^{d_x} c_i^{\nu_{ir}},$$

which denotes the rate law at which a reaction takes place, and the stoichiometric matrix

$$S = (\nu_{ir})_{i=1,\ldots,d_x, r=1,\ldots,d_R}\,.$$

The RRE ODE is then obtained by multyplying $v(c, \theta)$ with S:

$$\dot{c} = S \cdot v(c, \theta), \quad c(0) = c_0(\theta), \tag{2.2}$$

where $c_0(\theta)$ is the vector of initial conditions. The RRE is a deterministic description of biochemical reaction networks which is a good approximation of the average concentration for large volumes (Engblom, 2006a). In the limit of $\lim_{\Omega\to\infty}$, the RRE yields the (exact) mean of the solution to the CME (2.1), under assumption of monostability (Grima, 2012; van Kampen, 2007). Due to this relationship, the RRE can be called a **macroscopic** description, while the CME is called a **microscopic** description. Therefore, the reaction flux $v_r(c, \theta)$, which may not always agree with the microscopic propensity function $\hat{v}_r(n, \theta)$, is sometimes called the macroscopic rate function. In the next section, I will establish the connection of the macroscopic RRE to the microscopic CME, by discussing two different **mesoscopic** descriptions.

2.1.3 Moment-Closure Approximation

Evolution equations for the moments of the solution of the CME are obtained by differentiating the definition of the moments with respect to time and substituting in the CME.

For the mean and covariances, this yields:

$$\begin{aligned}\dot{\mu}_i &= \sum_{n\geq 0} n\dot{P}(t,n,\theta)\\ \dot{\Sigma}_{ij} &= \sum_{n\geq 0}(n-\mu)(n-\mu)^T\dot{P}(t,n,\theta)\end{aligned}$$

According to Engblom (2006a), these evolution equations can also be written as

$$\begin{aligned}\dot{\mu}_i &= \sum_r \mathbb{E}[-\nu_{ir}\hat{v}_r(\Omega\mu,\theta)]\\ \dot{\Sigma}_{ij} &= \sum_r \mathbb{E}[(\nu_{ir}\nu_{jr} - n_j\nu_{ir} + \mu_j\nu_{ir} - n_i\nu_{jr} + \mu_i\nu_{jr})\hat{v}_r(\Omega\mu,\theta)].\end{aligned}$$

The expectations that involve the propensities can be computed by applying a Taylor expansion. Yet, for systems involving non-linear propensities the equations for lower order moments are typically coupled to higher-order moments resulting in an infinite system of equations. A common procedure to break this hierarchy of moment equations is to set higher than second order cumulants to zero (Engblom, 2006a), which is equivalent to assuming that the third order cumulant is consistent with a Gaussian distribution. Assuming at most bimolecular reactions, the resulting set of non-linear ODEs is called the 2MA and is given by

$$\begin{aligned}\dot{\mu}_i &= \sum_{r=1}^{d_r}\nu_{ir}\left(\hat{v}_r(\Omega\mu,\theta) + \frac{1}{2}\sum_{s,l=1}^{M}\frac{\partial^2\hat{v}_r(\Omega\mu,\theta)}{\partial\mu_s\partial\mu_l}\Sigma_{sl}\right),\\ \dot{\Sigma}_{ij} &= \sum_{r=1}^{d_r}\sum_{s=1}^{M}\left(\nu_{ir}\frac{\partial\hat{v}_r(\Omega\mu,\theta)}{\partial\mu_s}\Sigma_{sj} + \nu_{jr}\frac{\partial\hat{v}_r(\Omega\mu,\theta)}{\partial\mu_s}\Sigma_{si}\right)\\ &\quad + \frac{1}{\Omega}\sum_{r=1}^{d_r}\nu_{ir}\nu_{jr}\left(\hat{v}_r(\Omega\mu,\theta) + \sum_{s,l=1}^{M}\frac{1}{2}\frac{\partial^2\hat{v}_r(\Omega\mu,\theta)}{\partial\mu_s\partial\mu_l}\Sigma_{sl}\right).\end{aligned}$$

The 2MA is precise for unimolecular reactions and fairly accurate if the third order moment is negligible (Engblom, 2006a). The latter is mostly the case for large reaction volume and molecule numbers (Engblom, 2006a). For small volumes, higher-order moment-closure approximations are more appropriate. Setting higher than third order cumulants zero, yields the third order moment-closure approximation (3MA) (Engblom, 2006a; Lee et al., 2009).

Several other choices for closures, such as the low dispersion closure (Hespanha, 2008), the derivative-matching or log-normal closure (Singh and Hespanha, 2007) or the mean-

field closure (Gandhi et al., 2000), exist. These closure schemes often make distribution assumptions of P with respect to n, for which the validity is typically not clear *a priori*. This shortcoming of the moment-closure approximation is sometimes criticized to be "*ad hoc*" (Grima, 2012), while the alternative systems size expansion is claimed to be "systematic" (Grima, 2012; van Kampen, 2007). In the following, I will discuss the system size expansion and it's properties.

2.1.4 System Size Expansion

A different technique to approximate the moments of the CME is given by the system size expansion (SSE) (Grima, 2010; van Kampen, 2007). The procedure allows the expansion of the CME around the solution of the RRE (2.2). The RRE is given by

$$\dot{c}_i = \sum_{r=1}^{d_r} \nu_{ir} v_r(c, \theta).$$

Here $v_r(c,\theta) = \lim_{\Omega\to\infty} \hat{v}_r(\Omega c, \theta)$ is also called the macroscopic rate function. While the RRE represents the leading order term of the SSE and yields the average concentrations for large volumes Ω. The next term, the Linear Noise Approximation (LNA), describes the fluctuations about these concentrations. The covariance of these fluctuations obeys (Elf and Ehrenberg, 2003; van Kampen, 2007):

$$\dot{\Sigma}_{ij} = \sum_{r=1}^{d_r}\sum_{s=1}^{M}\left(\nu_{ir}\frac{\partial v_r(c,\theta)}{\partial c_s}\Sigma_{sj} + \nu_{jr}\frac{\partial v_r(c,\theta)}{\partial c_s}\Sigma_{si}\right) + \frac{1}{\Omega}\sum_{r=1}^{d_r}\nu_{ir}\nu_{jr}v_r(c,\theta).$$

These results are exact for reaction networks comprising up to unimolecular reactions and for a small subset of networks with bimolecular reactions (Grima, 2015). For most networks involving bimolecular reactions, the SSE enables us to systematically correct the mean concentrations of the RRE and the variance predictions of the LNA, by considering higher order terms in the expansion. A more accurate estimate for the mean concentrations than the RRE, is given by the Effective Mesoscopic Rate Equation (EMRE) (Grima, 2010), and follows

$$\dot{\mu}_i = \dot{c}_i + \sum_{r=1}^{d_r}\nu_{ir}\left(\sum_{s=1}^{M}\frac{\partial v_r(c,\theta)}{\partial c_s}(\mu_s - c_s) + \frac{1}{2}\sum_{s,l=1}^{M}\frac{\partial^2 v_r(c,\theta)}{\partial c_s \partial c_l}\Sigma_{sl} - \frac{1}{2}\sum_{s=1}^{M}\frac{c_s}{\Omega}\frac{\partial^2 v_r(c,\theta)}{\partial c_s^2}\right).$$

Note that these equations yield a correction term of order Ω^{-1} to the RREs. Correspondingly, expressions for the covariances about these more accurate concentrations can be derived using the Inverse Omega Square (IOS) approximation, which corrects the LNA

estimate to order Ω^{-2} (Grima, 2011). In contrast to RRE and LNA, EMRE and IOS do not assume large volumes and hence these estimates are expected to be closer to the true moments predicted by the CME.

The MA and the SSE only yield the exact statistical moments of the solution to the CME in special cases (Grima, 2015). In all other cases, they yield approximations to the statistical moments. The approximation error can usually not be computed exactly and only rough error-bounds exist (Grima, 2012).

2.1.5 Models with Multiple Timescales

Biological systems can act on a variety of different timescales (Hasenauer et al., 2015; Shamir et al., 2016). For example, the response time of signaling and metabolic processes typically is in the order of seconds to minutes, while the response of gene regulation can take hours to days. Coupling such models can be problematic, as the time stepping of the numerical solver has to be adapted to the fastest timescales. Alternatively, changes that occur on fastest timescales can be assumed to happen instantaneously, introducing quasi-steady-states (Bowen et al., 1963) and discrete events. This allows time stepping appropriate to the slower timescales, but requires the introduction of algebraic constraints describing quasi-steady-states as well as discrete events. While it is often possible to integrate the algebraic constraints into model equations by substitution, the discrete events require special attention.

In this thesis we consider models with d_e different event types. The m^{th} event type is defined by a trigger function $g^{(m)}(t, x, \theta) : \mathbb{T} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \mapsto \mathbb{R}$ and update function $\Delta^{(m)}(t, x, \theta) : \mathbb{T} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \mapsto \mathbb{R}^{d_x}$ (Barton et al., 1998). The time point of the ℓ^{th} occurrence $\tau_\ell^{(m)}$ of the j^{th} event type is defined by the ℓ^{th} root of the trigger function $g^{(m)}$:

$$\forall\, \ell, j: \quad g^{(m)}(\tau_\ell^{(m)}, x(\tau_\ell^{(m)}, \theta), \theta) = 0 \text{ with } t_0 < \tau_1^{(m)} < \tau_2^{(m)} < \dots.$$

For Boolean trigger functions $g^{(m)}$, output values `true` can be mapped to positive values and output values `false` can be mapped to negative values such that $g^{(m)}$ has a root at every change of the Boolean value. The changes that are induced at every event occurrence are defined by the update function $\Delta^{(m)}$:

$$x(\tau_{\ell,+}^{(m)}, \theta) - x(\tau_\ell^{(m)}, \theta) = \Delta^{(m)}(\tau_\ell^{(m)}, x(\tau_\ell^{(m)}, \theta), \theta), \tag{2.3}$$

where $\tau_{\ell,+}^{(m)} = \lim_{\epsilon \to 0} \tau_\ell^{(m)} + \epsilon, \epsilon > 0$. It is possible that different events occur simultaneously. In this thesis, I assume that the cumulative change in the states is equal to the sum

over all triggered updates. The special case $\Delta^{(m)} \equiv 0$ can be used to treat discontinuities in the vector field f or define events that only produce model outputs.

2.2 Parameter Estimation

For most biological processes the vector of parameters θ is unknown. For the model to be predictive, θ has to be inferred from experimental data. This can be achieved by minimizing an objective function $J(\theta)$, which defines the distance between simulation results and experimental measurements. In the following, I will outline how the likelihood function can be used to define an objective function according to a specific noise model.

2.2.1 Likelihood Function

Experimental techniques usually do not provide direct measurements of modeled quantities x, but only of derived quantities y, which may depend linearly or non-linearly on x as well as θ. This relationship can be described by an output map $h(x, \theta) : \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \mapsto \mathbb{R}^{d_y}$. This output map defines the outputs (or observables) $y(t, \theta) : \mathbb{T} \times \mathbb{R}^{d_\theta} \to \mathbb{R}^{d_y}$ at time point t:

$$y(t, \theta) = h(x(t, \theta), \theta). \tag{2.4}$$

For non-linear output maps, the output map of lower order moments can depend on higher order moments, that might not be modeled. In that case, closure schemes need to be employed. For applications where RREs are employed, a zero cumulants closure, replacing all higher order contributions, is generally used.

In this thesis I assume independent, normally distributed, additive measurement noise. The time-resolved, noise corrupted measurements

$$\bar{y}_{ij} = y_i(t_j, \theta) + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{ij}^2(\theta))$$

and measurement time points t_j make up the experimental data $\mathcal{D} = \{((\bar{y}_{ij})_{i=1}^{d_y}, t_j)\}_{j=1}^{d_t}$. The number of time points, at which measurements have been collected, is denoted by d_t. For RREs, $\bar{y}_{ij}$ should only contain the mean measurements over multiple stochastic realizations, i.e., multiple single cells. The mean can be derived from bulk as well as single cell experiments. For higher-order approximations, $\bar{y}_{ij}$ may also describe higher order statistical moments, which can only be derived from single cell experiments. The standard deviation $\sigma_{ij}(\theta)$ can be modeled to be parameter dependent. The corresponding

error

$$\epsilon_{ij} = \epsilon_{ij,tech} + \epsilon_{ij,stat}$$

decomposes into $\epsilon_{ij,tech} \sim \mathcal{N}\left(0, \sigma^2_{ij,tech}(\theta)\right)$, which describes the noise in the measurement process itself, and $\epsilon_{ij,stat} \sim \mathcal{N}\left(0, \sigma^2_{ij,stat}\right)$, which describes the uncertainty of the empirical moment computed from the measured sample of single cells. Thus, $\sigma^2_{ij}(\theta)$ is given by

$$\sigma^2_{ij}(\theta) = \sigma^2_{ij,tech}(\theta) + \sigma^2_{ij,stat}.$$

Typically only the uncertainty of the measurement process is assumed to be parameter dependent, as $\sigma_{ij,stat}$ can be explicitly computed. The magnitude of $\sigma_{ij,stat}$ decreases with increasing sample size. Consequently, it can typically be safely ignored for bulk measurements, but it is non-neglible for single cell measurements with low sample sizes.

This noise assumption yields the likelihood function

$$\mathcal{L}(\mathcal{D}|\theta) = \prod_{j=1}^{d_t}\prod_{i=1}^{d_y} \frac{1}{\sqrt{2\pi\sigma^2_{ij}(\theta)}} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}_{ij} - y_i(t_j,\theta)}{\sigma_{ij}(\theta)}\right)^2\right)$$

Due to the asymptotic normality of the empirical moment estimators (Hansen, 1982), the distribution assumption for $\epsilon_{ij,stat}$ is usually appropriate for sufficiently large sample sizes. However, for covariances and other higher order moments, the independence assumption might be violated and more appropriate noise models should be employed (Ruess and Lygeros, 2015). Similarly, several assumptions on the distribution of $\epsilon_{ij,tech}(\theta)$ are reasonable, and the choice should depend on the employed measurement process (Arriaga, 2008; Reiter et al., 2011).

Besides time-resolved outputs y, it is also possible to consider event-resolved outputs z. Based on the results in (Barton et al., 1998), I introduce the event-resolved outputs z which describe measurements which are triggered by events and ordered accordingly. To allow for arbitrary event-resolved outputs, I introduce the output function $\lambda(t,x,\theta)^{(m)} : \mathbb{T} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_\theta} \mapsto \mathbb{R}^{d_z{}^{(m)}}$, which defines the event-resolved output for the ℓ^{th} occurrence of the m^{th} event type

$$z^{(m)}_\ell(\theta) = \lambda^{(m)}(\tau^{(m)}_\ell, x(\tau^{(m)}_\ell, \theta), \theta). \tag{2.5}$$

To consider event-resolved data $\bar{z}$ in addition to the time-resolved data $\bar{y}$, the measurement data

$$\mathcal{D} = \left\{ \left\{ t_j, \{\bar{y}_{ij}\}_{i=1}^{d_y} \right\}_{j=1}^{d_t}, \left\{ \bar{z}^{(m)}_{\ell q} \right\}_{q=1,l=1,m=1}^{d_z{}^{(m)}, d_\tau{}^{(m)}, d_e} \right\},$$

can be extended by ${d_z}^{(m)}$ different outputs $\bar{z}_{\ell q}^{(m)}$ for the ${d_\tau}^{(m)}$ occurrences of the m^{th} out of d_e different event types. For this thesis I also assume that the event-resolved experimental data is corrupted by independent, identically normally distributed, additive noise:

$$\bar{z}_{\ell q}^{(m)} = z_{\ell q}^{(m)}(\theta) + \epsilon_{\ell q}^{(m)}, \quad \epsilon_{\ell q}^{(m)} \overset{iid}{\sim} \mathcal{N}(0, {\omega_{\ell q}^{(m)}}^2(\theta)), \tag{2.6}$$

with parameter dependent standard deviation ${\omega_{\ell q}^{(m)}}^2(\theta)$. This assumption yields the likelihood function

$$\begin{aligned}\mathcal{L}(\mathcal{D}|\theta) = &\prod_{j=1}^{d_t}\prod_{i=1}^{d_y} \frac{1}{\sqrt{2\pi\sigma_{ij}^2(\theta)}} \exp\left(-\frac{1}{2}\left(\frac{\bar{y}_{ij} - y_i(t_j,\theta)}{\sigma_{ij}(\theta)}\right)^2\right) \\ &\cdot \prod_{m=1}^{d_e}\prod_{q=1}^{{d_z}^{(m)}}\prod_{\ell=1}^{{d_\tau}^{(m)}} \frac{1}{\sqrt{2\pi {\omega_{\ell q}^{(m)}}^2(\theta)}} \exp\left(-\frac{1}{2}\left(\frac{\bar{z}_{\ell q}^{(m)} - z_{\ell q}^{(m)}(\theta)}{\omega_{\ell q}^{(m)}(\theta)}\right)^2\right).\end{aligned}$$

As for the time-resolved data, other noise distributions may be applicable and a decomposition into different noise sources is possible.

The negative likelihood defines a distance between model simulation and experimental measurements. To obtain parameter values with the best agreement between model simulation and experimental measurements, the negative likelihood needs to be minimized. I will discuss this optimization problem in the following.

2.2.2 Parameter Estimation

The evaluation of the likelihood function $\mathcal{L}(\mathcal{D}|\theta)$ involves the computation of several products, which can be numerically unstable. Therefore, the negative log-likelihood

$$\begin{aligned}J(\theta) = -\log(\mathcal{L}(\mathcal{D}|\theta)) = &\frac{1}{2}\sum_{i=1}^{d_y}\sum_{l=1}^{d_t}\left(\frac{\bar{y}_{ij} - y_i(t_j,\theta)}{\sigma_{ij}(\theta)}\right)^2 + \log\left(2\pi\sigma_{ij}^2(\theta)\right) \\ &+\frac{1}{2}\sum_{m=1}^{d_e}\sum_{q=1}^{{d_z}^{(m)}}\sum_{\ell=1}^{{d_\tau}^{(m)}}\left(\frac{\bar{z}_{\ell q}^{(m)} - z_{\ell q}^{(m)}(\theta)}{\omega_{\ell q}(\theta)}\right)^2 + \log\left(2\pi {\omega_{\ell q}^{(m)}}^2(\theta)\right)\end{aligned} \tag{2.7}$$

is often used as objective function for minimization. As the log is a strictly monotoneously increasing function, the minimization of $J(\theta) = -\log(\mathcal{L}(\mathcal{D}|\theta))$ is equivalent to the minimization of $-\mathcal{L}(\mathcal{D}|\theta)$. The corresponding minimization problem

$$\theta^* = \arg\min_{\theta\in\Theta} J(\theta), \tag{2.8}$$

over a suitable parameter domain Θ, yields the Maximum Likelihood estimate of the parameters. Although this optimization problem is convex in $y_i(t_j, \theta)$, it is usually non-convex in θ. Thus, the objective function $J(\theta)$ can possess multiple local minima and appropriate optimization strategies must be used.

2.2.3 Optimization Methods

According to the infamous No Free Lunch Theorems for optimization (Wolpert and Macready, 1997), there exists no single optimization method that performs best on all classes of optimization problems. For the class of biologically motivated ODE constrained optimization problems such as (2.8), repeated, i.e., multi-start, local optimization has been shown to perform well (Raue et al., 2013b). In multi-start local optimization, independent local optimization runs are initialized at randomly sampled initial points in parameter space. For this purpose, the interior-point (Byrd et al., 2000; Waltz et al., 2006) and trust-region (Byrd et al., 1987; Sorensen, 1982) algorithms are suitable local, Newton-type optimization schemes.

For box-constraints, i.e., when Θ is a hypercube, the interior-point algorithm augments the objective function $J(\theta)$ by addition of a barrier function

$$-\Gamma \sum_{r=1}^{2d_\theta} \log(\rho_r),$$

which regularizes the problem and implements parameter bounds defined by boundary $\partial\Theta$ of the parameter domain Θ. This introduces additional slack variables ρ_r, which describe the distance to the parameter boundary $\partial\Theta$ and the weighting factor Γ, which is successively reduced to 0 over the course of the optimization (Byrd et al., 2000; Waltz et al., 2006).

Many interior-point implementations, such as the implementation in MATLAB routine `fmincon`, internally also rely on a trust-region algorithm to compute the optimization step. However, the trust-region algorithm can also be employed without the interior-point regularization, when a reflective method (Coleman and Li, 1994) is used to enforce parameter bounds. In the reflective method, steps that would result in values $\theta \notin \Theta$ are reflected at the boundary $\partial\Theta$.

In every step of the trust-region algorithm, the minimization problem is approximated by a quadratic problem

$$\min_{s \in B_{D,\varepsilon}(\theta)} \frac{1}{2} s^T \nabla^2 J(\theta) s + s^T \nabla J(\theta), \tag{2.9}$$

where $B_{D,\varepsilon}(\theta) = \{s : \|D(s-\theta)\|_2 \leq \varepsilon\}$ is the trust-region. The trust-region is an ellipsoid with radius ε and diagonal scaling matrix D around the current parameter θ. If the proposed step does not yield a decrease in the objective function, i.e., if $J(\theta + s) \geq J(\theta)$, the trust-region radius ε is reduced and the quadratic problem (2.9) is solved again, until a decrease is achieved.

The quadratic approximation (2.9) can also be used to compute a prediction J_{pred} for the expected objective function value $J(\theta + s)$ after the update step. In every iteration, the trust-region radius ε, can also be increased based on the ratio $\frac{J_{\text{pred}}}{J(\theta+s)}$.

As the trust-region algorithm uses the Hessian of the objective function $\nabla^2 J(\theta)$, it belongs to the class of Newton-type algorithms. When the Hessian is not available, approximate methods such as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm can be used, which yields a quasi-Newton method. The BFGS algorithm iteratively computes approximations to the Hessian based on the dyadic product of the gradient $\nabla J(\theta)$ of the objective function. For independent, normally distributed measurement noise, as assumed in (2.7), and known noise parameters σ and ω, the optimization problem (2.8) is of least squares type. This structure can be exploited in algorithms such as `NL2SOL` (Dennis et al., 1981) using a Gauss-Newton (Björck, 1996) or a Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963; More, 1978) method, which both ignore second-order partial derivatives in the Hessian. The respective approximations of the Hessian, coincide with the (damped) Fisher information matrix (FIM) (Fisher, 1922) of the respective parameter estimate.

In general, the Hessian $\nabla^2 J(\theta)$ is not guaranteed to be positive definite as problem (2.7) can be – and often is – non-convex. In constrast, both the FIM and the BFGS approximation are always positivie semi-definite and can easily be regularized to be positive definite. This leads to better convergence as Newton-type algorithms are known to be unstable if positive definiteness is not guaranteed (Schraudolph, 2002). Therefore, the use of these approximate methods might be advantageous – even if $\nabla^2 J(\theta)$ is available.

Before I describe the different methods to compute the gradient $\nabla J(\theta)$, I will shortly discuss uncertainty analysis, where problems similar to (2.8) have to be solved. For these problems, the same optimization methods can be applied.

2.2.4 Profile Likelihood based Uncertainty Analysis

Experimental data of biochemical processes is often scarce and noise corrupted, resulting in non-identifiabilities and parameter uncertainties. Parameter identifiability is typically assessed using structural and practical identifiability analysis (see (Chis et al., 2011; Raue

et al., 2009) and references therein). Structural identifiability analysis provides information for the considered model topology and measured output, independent of a specific dataset. In contrast, practical identifiability and uncertainty analysis provide information about the uncertainty of parameter estimates for a given dataset. In general, parameter uncertainties can be assessed using frequentist methods such as profile likelihoods (Murphy and van der Vaart, 2000; Raue et al., 2013a) and Bayesian methods, which often rely on sampling (Ballnus et al., 2017; Girolami and Calderhead, 2011; Kramer et al., 2010; Wilkinson, 2007).

The profile likelihood of a parameter θ_i, denoted by $\mathrm{PL}_i(\theta_i)$, is given by the likelihood maximized over the remaining parameters,

$$\mathrm{PL}_i(\theta_i) = \max_{\theta_{j \neq i}, \theta \in \Theta} \mathcal{L}(\theta).$$

Accordingly, profile likelihoods can be computed by solving a set of parameter constrained optimization problems. This requires repeated local optimization. For this purpose the optimization methods introduced in Subsection 2.2.3 are suitable.

Frequentist confidence intervals can be computed by comparing the profile likelihood $\mathrm{PL}(\theta_i)$ to the likelihood $\mathcal{L}(\mathcal{D}|\theta^*)$ at the Maximum Likelihood estimate θ^* (Venzon and Moolgavkar, 1988). For the models considered in this thesis, that contain structural non-identifiabilities, profile likelihoods are the only viable frequentist technique for global uncertainty analysis (Fröhlich et al., 2014b).

In contrast to this frequentist approach, Bayesian uncertainty analysis methods rely on Bayes' theorem,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})},$$

in which $p(\theta)$, $p(\mathcal{D}|\theta)(= \mathcal{L}(\mathcal{D}|\theta))$, $p(\mathcal{D})$ and $p(\theta|\mathcal{D})$ denote prior probability, likelihood, evidence and posterior distribution respectively (Wilkinson, 2007). Effectively, the choice of a particular prior $p(\theta)$ corresponds to the regularization of the negative log-posterior $\tilde{J}(\theta) = -\log(p(\theta|\mathcal{D}))$. For example, a Gaussian prior corresponds to a Tikhonov (l_2) regularization, while a Laplacian prior corresponds to a l_1 regularization.

To determine Bayesian credibility intervals of the parameters, methods such as Markov chain Monte Carlo (MCMC) based sampling schemes (Hastings, 1970) can be applied. For many applications, sampling is computationally more demanding than optimization and profile computation, as a high number of function evaluations can be necessary. Similar to optimization schemes, some MCMC schemes, such as the Metropolis adjusted Langevin

algorithm (Roberts and Rosenthal, 1998), the Hamilton Monte Carlo method (Girolami and Calderhead, 2011) or Riemann manifold Langevin method (Girolami and Calderhead, 2011) can also exploit the gradient of the posterior.

The evaluation of $\mathcal{L}(\mathcal{D}|\theta)$ and accordingly $J(\theta)$ requires the solution to the underlying ODE model. Thus, the corresponding gradient depends on the derivates of the solution to the differential equation with respect to the parameters θ. In the next section, I will outline how these gradients can be computed efficiently.

2.3 Simulation and Sensitivity Analysis

For ODE constrained optimization problems, the computational cost of evaluating the objective function is dominated by the numerical simulation. The objective function evaluation as well as the computation of the gradient of the objective function require the numerical solution to an ODE. Consequently, efficient numerical methods to solve ODEs are desirable. In the following, I will outline how these ODEs can be solved efficiently and how the gradient can be computed as solution to one or multiple ODEs.

2.3.1 Numerical Simulation

As previously discussed, biochemical processes can act on multiple timescales. Consequently, the corresponding model equations are often stiff (Resat et al., 2009). However, the time-scale of processes is often not known *a priori* and thus encoded in parameter values. Consequently, the model equations might be stiff in certain parameter regimes – even if the real biological system does not involve multiple timescales (Jia et al., 2011). Therefore, it is always advisable to use an implicit differential equation solver, which can adequately handle stiff systems (Gonnet et al., 2012).

Implicit differential equations solvers include the fully implicit Runge-Kutta solver family (Butcher, 1964), the Singly Diagonally Implicit Runge-Kutta solver family (Alexander, 1977) as well as the Rosenbrock solver family (Rosenbrock, 1963). Implicit methods generally compute the state variables at the next time step ξ_{i+1} based on the state variables at previous iterations $x(\xi_i), x(\xi_{i-1}), x(\xi_{i-2}), \ldots$ by solving an implicit equation

$$G(x(\xi_{i+1}), x(\xi_i), x(\xi_{i-1}), x(\xi_{i-2}), \ldots) = 0,$$

where the function G depends on the choice of the method and on right hand side of the differential equation f. For single step methods, such as Runge-Kutta type solvers, the

function G will only depend on $x(\xi_i)$, and not on previous values. For multi-step methods the function G will depend on previous values of x. This thesis will consider the implicit linear multi-step Backwards Differentiation Formula (BDF), implemented in the CVODES solver (Serban and Hindmarsh, 2005). In every iteration i, the BDF solves an equation of the form

$$h_i \beta_{i,0} \dot{x}(\xi_i) + \sum_{j=0}^{q} \alpha_{i,j} x(\xi_{i-j}) = 0,$$

where q is the (variable) order of the method, α and β are the coefficients that depend on the method type, the order q, and the history of recent step sizes h_i. The order q and the step size h_i will determine the local error of the numerical solution (Serban and Hindmarsh, 2005).

This implicit equation is typically solved using a Newton's method (Hindmarsh et al., 2005; Zhang and Sandu, 2014). For the BDF, the function G depends on $\dot{x}$ and thus on f. Consequently, the Newton solver computes multiple solutions to linear systems defined by the Jacobian $\nabla_x f(t, x, \theta)$ of the right hand side of the differential equation at every integration step.

The computation time of the BDF method primarily depends on two factors: (i) the evaluation time of the function f and the Jacobian $\nabla_x f(t, x, \theta)$, which usually scales linearly with d_x and (ii) the time to solve the linear systems defined by $\nabla_x f(t, x, \theta)$. The matrix $\nabla_x f(t, x, \theta)$ is typically not symmetric and neither positive nor negative definite. For such unstructured problems, LU decomposition, which factorizes $\nabla_x f(t, x, \theta)$ into a lower-triangular matrix L and an upper-triangular matrix U, is the method of choice to solve the linear system. After performing the decomposition, the solution to the linear systems can be computed by matrix multiplication. When no additional structure of the matrix is exploited, the computational complexity of matrix multiplication with state-of-the-art algorithms increases at least with exponent 2.376 with respect to d_x (Coppersmith and Winograd, 1990) and thus dominates the computation time for sufficiently large d_x.

For certain ODE model classes, specialized solvers exist. For finite element discretizations of partial differential equations, the Jacobian can be brought into banded form, for which specialized solvers that scale with the number of off-diagonals of the Jacobian have been developed (Thorson, 1979). Unfortunately, ODE models of biochemical reaction networks cannot generally be brought into a banded structure. For example, in polymerization reactions that include dissociation of monomers, the monomer species will always be influenced by all other species and the number of off-diagonals in the Jacobian will be equal to d_x. Other frequently occurring motifs, such as feedback loops and single highly in-

teractive species (Barabasi and Oltvai, 2004), will also increase the number of necessary off-diagonals.

As alternative to banded solvers, sparse solvers have been introduced in the context of circuit simulations (Davis and Palamadai Natarajan, 2010). For sparse solvers, the computation time depends on the number of non-zero entries. The method relies on an approximate minimum degree (AMD) ordering (Amestoy et al., 1996) which is a graph theoretical approach that can be used to minimize the fill-in of the L and U decomposition. The algorithm relies on the fact that matrix $M \in \mathbb{R}^{d_x \times d_x}$ can be represented as directed graph (V, E) with d_x vertices V and number of non-zero entries $nnz(M)$ edges E. For every non-zero entry M_{ij}, the corresponding graph includes a directed edge $(i \to j)$. The LU-decomposition can be represented as operations on this graph, where the cumulative degree of the vertices in the final graph corresponds to the fill in the L and U matrices (George and Liu, 1981). The AMD algorithm uses an approximation of the resulting vertice degree in combination with a greedy algorithm to compute a reordering of M to minimize the fill-in in L and U. By reducing the fill-in in L and U, the computational cost of subsequently necessary matrix multiplications can be reduced (Davis and Palamadai Natarajan, 2010). As computing the AMD reordering is computationally inexpensive, the overall computational cost usually decreases. Currently, no formulas for the expected speedup or the general scaling with respect to non-zero entries exist. For biochemical reaction networks, the application of such a sparse solver seems reasonable (Gonnet et al., 2012). Typically, most molecular species only interact with a handful of other species (Barabasi and Oltvai, 2004), resulting in few non-zero entries in the Jacobian.

Sparse numerical solvers can be used to compute the numerical solution of the model ODE that are required for objective function evaluation. They can also be used to compute objective function gradients as solution to one or more ODEs. Several of different gradient computation approaches exist and in the following, I will discuss the three most common approaches.

2.3.2 Finite Differences

A näive approximation to the gradient of the objective function (2.7) with respect to θ_k can be obtained by finite differences:

$$\frac{\mathrm{d}J}{\mathrm{d}\theta_k} \approx \frac{J(\theta + a\, e_k) - J(\theta - b\, e_k)}{a + b},$$

with $a, b \geq 0$ and the k^{th} unit vector e_k. In practice, forward differences ($a = \zeta$, $b = 0$), backward differences ($a = 0$, $b = \zeta$) and central differences ($a = \zeta$, $b = \zeta$), with $\zeta \ll 1$, are widely used. For the computation of forward finite differences, this yields a procedure with three steps:

Step 1: The state trajectory $x(t, \theta)$ and output trajectory $y(t, \theta)$ are computed.

Step 2: The state trajectories $x(t, \theta^{(k)})$ and the output trajectories $y(t, \theta^{(k)})$ are computed for the perturbed parameters $\theta^{(k)} = \theta + \zeta e_k$ for $k = 1, \ldots, d_\theta$.

Step 3: The objective function gradient elements $\frac{\mathrm{d}J}{\mathrm{d}\theta_k}$, are computed from the output trajectory $y(t, \theta)$ and the perturbed output trajectory $y(t, \theta^{(k)})$ for $k = 1, \ldots, d_\theta$.

In theory, forward and backward differences provide approximations of order ζ while central differences provide more accurate approximations of order ζ^2, provided that J is sufficiently smooth. In practice, the optimal choice of a and b depends on the accuracy of the numerical integration (Raue et al., 2013b). If the integration accuracy is high, an accurate approximation of the gradient can be achieved using $a, b \ll 1$. For lower integration accuracies, larger values of a and b usually yield better approximations. A good choice of a and b is typically not clear *a priori* (see (Hanke and Scherzer, 2001) and the references therein). Alternative methods based on complex numbers that circumvent the specification of a and b have been proposed (Lyness and Moler, 1967; Squire and Trapp, 1998), but no general-purpose implementations of such methods for ODEs have been established.

The computational complexity of evaluating gradients using finite differences is affine linear in the number of parameters. Forward and backward differences require in total $d_\theta + 1$ function evaluations. Central differences require in total $2d_\theta$ function evaluations. Asa single simulation of a large-scale model is already time-consuming, the gradient calculation using finite differences can be limiting.

2.3.3 Forward Sensitivity Analysis

State-of-the-art systems biology toolboxes, such as Data2Dynamics (Raue et al., 2015), use forward sensitivity analysis for gradient evaluation. The gradient of the objective

function (2.7) is given by

$$\begin{aligned}\frac{\mathrm{d}J}{\mathrm{d}\theta_k} =& \sum_{i=1}^{d_y}\sum_{j=1}^{N}\left(\frac{\bar{y}_{ij}-y_i(t_j,\theta)}{\sigma_{ij}^2}\right)s_{i,k}^{y}(t_j,\theta)\\ &+\sum_{m=1}^{d_e}\sum_{q=1}^{d_z{}^{(m)}}\sum_{\ell=1}^{d_\tau{}^{(m)}}\left(\frac{\bar{z}_{\ell q}^{(m)}-z_{\ell q}^{(m)}(\theta)}{\omega_{\ell q}(\theta)}\right)s_k^{z_{\ell q}^{(m)}}(\theta)\\ &+\frac{\partial J}{\partial\theta_k},\end{aligned}$$

with $s_{i,k}^{y}(t,\theta):[t_0,t_N]\times\mathbb{R}^{d_\theta}\mapsto\mathbb{R}$ denoting the sensitivity of time-resolved output y_i at time point t with respect to parameter θ_k and $s_k^{z_{\ell,q}^{(m)}}(\theta):\mathbb{R}^{d_\theta}\mapsto\mathbb{R}$: denoting the sensitivity of event-resolved output $z_{\ell,q}^{(m)}$. These two quantities can be computed by applying the total derivative to the functions h and λ:

$$\begin{aligned}s_{i,k}^{y}(t_j,\theta) &= \left.\frac{\partial h_i}{\partial x}\right|_{x(t,\theta),\theta}s_k^{x}(t_j,\theta)+\left.\frac{\partial h_i}{\partial\theta_k}\right|_{x(t,\theta),\theta}\\ s_k^{z_{\ell,q}^{(m)}}(\theta) &= \left.\frac{\partial\lambda_q^{(m)}}{\partial t}\right|_{\tau_\ell^{(m)},x(\tau_\ell^{(m)}\theta),\theta}s_k^{\tau_\ell^{(m)}}(\theta)+\left.\frac{\partial\lambda_q^{(m)}}{\partial x}\right|_{\tau_\ell^{(m)},x(\tau_\ell^{(m)}\theta),\theta}s_k^{x}(x(\tau_\ell^{(m)}\theta),\theta)\\ &\quad+\left.\frac{\partial\lambda_q^{(m)}}{\partial\theta_k}\right|_{\tau_\ell^{(m)},x(\tau_\ell^{(m)}\theta),\theta}.\end{aligned}$$

with $s_k^{x}(t,\theta):[t_0,t_N]\times\mathbb{R}^{d_\theta}\mapsto\mathbb{R}^{d_x}$ denoting the sensitivity of the state x with respect to θ_k. The state sensitivity to a specific parameter can be computed as solution to an ODE system (Kokotovic and Heller, 1967):

$$\dot{s}_k^{x}(t,\theta)=\left.\frac{\partial f}{\partial x}\right|_{x(t,\theta),\theta}s_k^{x}(t,\theta)+\left.\frac{\partial f}{\partial\theta_k}\right|_{x(t,\theta),\theta},\qquad s_k^{x}(t_0,\theta)=\left.\frac{\partial x_0}{\partial\theta_k}\right|_{\theta}.$$

Moreover, $s_k^{\tau_\ell^{(m)}}(\theta):\mathbb{R}^{d_\theta}\mapsto\mathbb{R}^{d_x}$ denotes the sensitivity of the event-time of the ℓ^{th} occurrence of the m^{th} event type with respect to θ_k. The sensitivity $s_k^{\tau_\ell^{(m)}}(\theta)$ can be computed according to the implicit function theorem

$$s_k^{\tau_\ell^{(m)}}(\theta)=\frac{\partial\tau_\ell^{(m)}}{\partial\theta_k}=-\left(\frac{\partial g^{(m)}}{\partial t}\right)^{-1}\left.\frac{\partial g^{(m)}}{\partial\theta_k}\right|_{\tau_\ell^{(m)}},$$

assuming that $\left(\frac{\partial g^{(m)}}{\partial t}\right)$ is non-zero. At the occurrence of every event the state sensitivities s_k^x need to be updated according to the formula

$$s_k^x(\tau_{\ell,+}^{(m)}) - s_k^x(\tau_\ell^{(m)}) = -(\dot{x}(\tau_{\ell,+}^{(m)}) - \dot{x}(\tau_\ell^{(m)}))s_k^{\tau_\ell^{(m)}} + \frac{\partial \Delta^{(m)}}{\partial x}\left(s_k^x + \dot{x}s_k^{\tau_\ell^{(m)}}\right)\Bigg|_{\tau_\ell^{(m)}} + \frac{\partial \Delta^{(m)}}{\partial t}s_k^{\tau_\ell^{(m)}}\Bigg|_{\tau_\ell^{(m)}} + \frac{\partial \Delta^{(m)}}{\partial \theta}\Bigg|_{\tau_\ell^{(m)}}.$$

Forward sensitivity analysis consists of three steps:

Step 1: The state trajectory $x(t,\theta)$ and output trajectory $y(t,\theta)$ are computed.

Step 2: The state sensitivities $s_k^x(t,\theta)$ and the output sensitivities $s_k^y(t,\theta)$ and $s_k^{z_{\ell,q}^{(m)}}(\theta)$ are computed using the state trajectory $x(t,\theta)$ for $k = 1,\ldots,d_\theta$.

Step 3: The objective function gradient elements $\frac{\mathrm{d}J}{\mathrm{d}\theta_k}$, are computed from the output sensitivities $s_k^y(t,\theta)$ and $s_k^{z_{\ell,q}^{(m)}}(\theta)$ and the outputs $y(t,\theta)$ and $z_\ell^{(m)}$ for $k = 1,\ldots,d_\theta$.

Step 1 and 2 are often combined, which enables simultaneous error control and the reuse of the Jacobian (Hindmarsh et al., 2005). The simultaneous error control allows for the calculation of accurate and reliable gradients. The reuse of the Jacobian improves the computational efficiency.

The number of state and output sensitivities increases linearly with the number of parameters. While this is unproblematic for small- and medium-sized models, solving forward sensitivity equations for systems with several thousand state variable bears technical challenges. Code compilation can take multiple hours and requires more memory than what is available on standard machines. Furthermore, while forward sensitivity analysis is usually faster than finite differences, the complexity, in practice, still increases roughly linearly with the number of parameters.

2.3.4 Adjoint Sensitivity Analysis

In the mathematics and engineering community, adjoint sensitivity analysis is frequently used to compute the gradients of a functional with respect to the parameters if the functional depends on the solution of a differential equation (Plessix, 2006). In these applications, measurements are continuous in time and $J(\theta)$ is assumed to be a functional of the solution $x(t)$ of a differential equation. However, this approach can also be applied to

discrete-time measurements and in contrast to forward sensitivity analysis, adjoint sensitivity analysis does not rely on the state sensitivities $s_k^x(t)$, but on the adjoint state $p(t)$.

For discrete-time measurements – the usual case in systems and computational biology – the adjoint state is piece-wise continuous in time and defined by a sequence of backward differential equations. For $t > t_N$, the adjoint state is zero, $p(t) = 0$. Starting from this end value the trajectory of the adjoint state is calculated backwards in time, from the last measurement $t = t_N$ to the initial time $t = t_0$. At the measurement time points $t_N, \ldots, t_1$, the adjoint state is reinitialized as

$$p(t_j) = \lim_{t \to t_j^+} p(t) + \frac{\partial J}{\partial x}, \tag{2.10}$$

which usually results in a discontinuity of $p(t)$ at t_j. Starting from the end value $p(t_j)$ as defined in (2.10) the adjoint state evolves backwards in time until the next measurement point t_{j-1} or the initial time t_0 is reached. This evolution is governed by the time dependent linear ODE

$$\dot{p} = -\left(\frac{\partial f}{\partial x}\right)^T p. \tag{2.11}$$

The repeated evaluation of (2.10) and (2.11) until $t = t_0$ yields the trajectory of the adjoint state. Given this trajectory, the gradient of the objective function with respect to the individual parameters is

$$\frac{\mathrm{d}J}{\mathrm{d}\theta_k} = -\int_{t_0}^{t_N} p^T \frac{\partial f}{\partial \theta_k} dt - p(t_0)^T \frac{\partial x_0}{\partial \theta_k} + \frac{\partial J}{\partial \theta_k}. \tag{2.12}$$

The calculation of the objective function gradient using adjoint sensitivity analysis consists of three steps:

Step 1: The state trajectory $x(t, \theta)$ and output trajectory $y(t, \theta)$ are computed.

Step 2: The trajectory of the adjoint state $p(t)$ is computed.

Step 3: The objective function gradient elements $\frac{\mathrm{d}J}{\mathrm{d}\theta_k}$, $k = 1, \ldots, n_\theta$, are computed from the state trajectory $x(t, \theta)$, the adjoint state trajectory $p(t)$ and the output trajectory $y(t, \theta)$.

Step 1 and 2, which are usually the computationally intensive steps, are independent of the parameter dimension. The complexity of Step 3 increases linearly with the number

of parameters, yet the computation time required for this step is typically negligible. Unfortunately, adjoint sensitivities are currently not applicable for systems with parameter dependent discrete events as respective update formulas for the adjoint and quadrature states are missing.

Chapter 3

Summary of Contributed Articles

In this chapter, I provide detailed summaries of all of four articles that constitute this publication-based dissertation. I am the sole first author of all of these articles and was in charge of their preparation. A detailed description of my contributions to each article is provided below. All articles are peer-reviewed and published in international, well-established journals and not used in any other publication-based dissertation. The articles are sorted according to the issues (i) to (iv), listed in Section 1.2. The full text of all of my main contributions follows in Appendices A-D.

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

Quantitative mechanistic models are valuable tools for disentangling biochemical pathways and for achieving a comprehensive understanding of biological systems. However, to be quantitative, the parameters of these models have to be estimated from experimental data. In the presence of significant stochastic fluctuations, this is a challenging task as stochastic simulations are usually too time-consuming (see Section 2.1.1) and a macroscopic description using RREs may not be accurate (see Section 2.1.2). Mesoscopic moment-closure approximation and the system size expansion have been proposed for simulation as they can be evaluated more efficiently than stochastic simulations but are more accurate than macroscopic rate equations (see Section 2.1.3 and Section 2.1.4). However, these approximations were previously not used for parameter estimation and the effect of the approximation error on parameter estimates was unknown.

In this article, I derived and implemented novel gradient-based parameter optimization methods and uncertainty analysis methods for these mesoscopic descriptions, as conceived by Philipp Thomas, Fabian Theis, Ramon Grima, Jan Hasenauer and me. I considered a model of protein synthesis followed by enzymatic degradation, a model of a trimerization reaction as well as a model Epo-induced JAK/STAT signaling. For all models, I generated the model equations for the macroscopic and mesoscopic descriptions, with support from Atefeh Kazeroonian and Philipp Thomas, using the

MATLAB toolbox `CERENA` (Kazeroonian et al., 2016). To address issue (i), I performed a first comprehensive assessment of the dependence of the simulation time on the system volume, a proxy for the magnitude of stochastic fluctuations and cumulative reaction propensity. This revealed that the simulation time for MA and SSE does not depend on the system volume. Furthermore, I assessed how the approximation order influences the error of parameter estimates obtained using the parameter estimation toolbox `PESTO` (Stapor et al., 2017). I evaluated the accuracy of parameter estimates on simulated data for the model of protein synthesis followed by enzymatic degradation and the model of a trimerization reactions as well as experimental data for the model of Epo-induced JAK/STAT signaling with assistance of Jan Hasenauer, Philipp Thomas and Ramon Grima. For the model of Epo-induced JAK/STAT signaling I employed the profile likelihood approach to demonstrate that moment-closure approximation and system size expansion can improve parameter identifiability in comparison to RRE – even if merely population-average data are available. I validated the estimated parameter value with additional experimental data from previous studies. This result was a novel finding and underlined the importance of considering stochastic formulation of the underlying biochemical reaction network. Furthermore, the simulation examples revealed that the estimates obtained using mesoscopic descriptions are more reliable than estimates obtained using the RRE, for an intermediate volume regime, where the effect of stochastic fluctuations is weak but non-negligible. I was able to relate the observed error in parameter estimates to the approximation order of the system size expansion and the moment-closure expansion as well as the variance of the employed empirical moment estimator and could thus explain the emergence of such an intermediate volume regime. Moreover, I proposed methods to determine the regime boundaries based on model selection and model rejection methods. These methods can also be used to determine whether a stochastic description is necessary for a specific dataset. These results illustrated that inference using moment-closure approximation and system size expansion is feasible and possesses a high sensitivity.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I wrote the first complete draft of the manuscript and consolidated the draft with the other authors. The final manuscript includes contributions from Jan Hasenauer, Philipp Thomas and Ramon Grima.

(2) F. Fröhlich, B. Kaltenbacher, F.J. Theis, J. Hasenauer. **Scalable parameter estimation for genome-scale biochemical reaction networks.** PLoS Computational Biology 13(1):e1005331 (2017).

Mechanistic mathematical modeling of biochemical reaction networks using ODE models has improved our understanding of small- and medium-scale biological processes. While the same should in principle hold for large- and genome-scale processes, the computational methods for the analysis of ODE models which describe hundreds to thousands of biochemical species and reactions were missing. While individual simulations were feasible, the estimation of the model parameters from experimental data was computationally too intensive. Adjoint sensivitities have been proposed as means of efficiently evaluating gradients for functionals where measurements are continuous, but were never evaluated for ODE models of biochemical reaction networks, which are usually trained on time-discrete measurements (see Section 2.3.4).

In this article, I provided the first thorough evaluation of adjoint sensitivity analysis for parameter estimation in large-scale biochemical reaction networks, as conceived by Jan Hasenauer and me. I implemented the algorithms in the open-source toolbox `AMICI`, which is used by a growing number of other researchers in various research projects (Ballnus et al., 2017; Boiger et al., 2016; Loos et al., 2016, 2017; Maier et al., 2017; Stapor et al., 2017). `AMICI` performs all necessary symbolic processing of the differential equations to generate native C++ code that allows for highly efficient simulation and sensitivity analysis. To make the toolbox accessible to other researchers, I implemented an easy-to-use mex-interface that renders the simulation routines accessible from MATLAB. To compare the adjoint sensitivity analysis approach for time-discrete measurement to state-of-the-art forward sensitivity and finite difference methods used in systems and computational biology, I implemented both in `AMICI`. To address issue (ii), I evaluated the performance of all methods on seven published models of biochemical reaction networks with 116 to 1801 parameters and 18 to 500 state variables with the assistance of Barbara Kaltenbacher and Jan Hasenauer. The comparison revealed a reduction of the computational cost by up to 334 fold and a superior scalability with the number of parameters for adjoint sensitivity analysis for all but one of the considered models. These results suggest that adjoint sensitivity analysis is generally favorable for most models with more than 10 to 100 parameters. However, for specific models forward sensitivity analysis may still be more efficient. The evaluation of the accuracy of the computed sensitivities revealed that the numerical error in adjoint sensitivity analysis was as low as for forward sensitivity analysis. For finite differences I observed a much higher numerical error. The observed computational cost for adjoint sensitivity analysis is effectively independent of the number of parameters, enabling the analysis of large- and genome-scale models. The study of a comprehensive kinetic model of ErbB signaling revealed that parameter estimation using adjoint sensitivity analysis re-

quired a fraction of the computation time of established methods, which underlines the practical relevance of the developed method. The study provides the first solid evidence for the benefits of adjoint sensitivity analysis for biochemical reaction networks. The implementation I provided in AMICI is currently used in several research projects by multiple research groups.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I wrote the first complete draft of the manuscript and consolidated the first draft with the other authors. The final manuscript includes contributions from Barbara Kaltenbacher, Fabian Theis and Jan Hasenauer.

(3) F. Fröhlich, T. Kessler, D. Weindl, A. Shadrin, L. Schmiester, H. Hache, A. Muradyan, M. Schütte, J. Lim, M. Heinig, F.J. Theis, H. Lehrach, C. Wierling, B. Lange, and J. Hasenauer. **Efficient parameterization of large-scale mechanistic models enables drug response prediction for cancer cell lines.** bioRxiv: 174094.

The effect of cancer drugs often affects multiple different pathways and substantially varies across individuals. Thus large-scale mechanistic models are likely to be required for accurate predictions of drug efficacy. Several such large scale models have been derived (Barrette et al., 2017; Chen et al., 2009), but parameter estimation using experimental data was never performed due to the high computational cost. Besides the large number of parameters, the large number of state variables also poses a challenge for parameter estimation. Sparse solvers for the efficient simulation of ODE models with many state variables have been developed in the context of circuit simulation, but require a sparsity structure in the respective ODE model (see Section 2.3.1). The applicability and efficiency of such sparse solvers for ODE models of biochemical reaction networks was unclear.

In this article, I developed a novel framework of computational methods for the parameterization of large-scale quantitative mechanistic models and its application to the prediction of drug response of cancer cell lines based on exome and transcriptome sequencing data. The considered model describes several cancer-associated signaling pathways and has in total over 1200 state variables and over 4000 parameters. I implemented the model, which was constructed by our collaboration partners Thomas Kessler, Alexey Shadrin, Hendrik Hache, Artur Muradyan, Moritz Schütte, Ji-Hyun Lim, Hans Lehrach, Christoph Wierling and Bodo Lange at ALACRIS Theranostics GmBH, in `AMICI`. To address issue (iii), I proposed the use of a sparse solver and implemented an interface to the sparse solver KLU (Davis and Palamadai Natarajan, 2010) in `AMICI`. To achieve scalability with respect to the number of experimental

conditions, I implemented a parallelization scheme in the MATAB toolbox `PESTO`. I performed parameter estimation using the sparse solver and adjoint sensitivity analysis implemented in `AMICI` and parallelization implemented in `PESTO` with assistance of Daniel Weindl and Jan Hasenauer, as conceived by Christoph Wierling, Fabian Theis, Jan Hasenauer, Thomas Kessler and me. Overall the developed framework achieved an computational speedup of over 240.000 fold over state-of-the-art methods. Only the combination of all three approaches enabled parameterization of the model from experimental data, as the computational cost with state-of-the-art approaches was in the order of hundreds of thousands of years.

The developed mechanistic model can be individualized to specific cell lines and experiments by using values from exome and transcriptome sequencing as well as drug concentrations in the right hand side of the ODE. This allowed the parameterization of the model from several thousand drug assays for cancer cell lines from various tissues in the Cancer Cell Line Encyclopedia (Barretina et al., 2012). I compared the prediction accuracy of the parameterized mechanistic model to several statistical methods, including logistic regression approaches as well as a random forest classifier, which were implemented and trained by Matthias Heinig. I found that the mechanistic model outperforms all statistical approaches in terms of classification accuracy for in-tissue as well as out-of-tissue predictions. To analyse the uncertainty of estimated parameters I performed uncertainty analysis based on the Fisher Information Matrix, which suggested substantial uncertainty in the estimated parameters. However, the comparison with proteomic data, which was carried out by Leonard Schmiester, revealed that reliable prediction of molecular phenotypes could still be achieved. Furthermore, I demonstrated that the mechanistic model has the remarkable capacity of predicting outcome of combination treatments, even when only trained on individual treatments, which is not possible with any of the established statistical approaches. These results show that the scalable methods, that were developed in this thesis, enable parameter estimation for ODE at unprecedented scale. The application to a massive characterization of cancer cell lines underlines the relevance of the developed methods to practical applications.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I wrote the first complete draft of the manuscript and consolidated the draft with the other authors. The final manuscript includes contributions from Christoph Wierling, Daniel Weindl, Jan Hasenauer, Matthias Heinig and Thomas Kessler.

(4) F. Fröhlich, F.J. Theis, J.O. Rädler, J. Hasenauer. **Parameter estimation for dy-**

namical systems with discrete events and logical operations. Bioinformatics 33(7):1049-1056 (2017).

ODE models are frequently used to describe the dynamic behaviour of biochemical processes. Such ODE models can be extended by events to describe the effect of fast latent processes on the process dynamics (see Section 2.1.5). For some biochemical processes, such as cell-death, the instant of time at which an event takes places can be experimentally measured as event-resolved datapoint. However, no formulation of the likelihood and of respective gradients existed. Thus, the consideration of such event-resolved data was only possible with statistical methods, but not with mechanistic ODE models. This prohibited a deeper understanding of underlying biochemical mechanisms. Moreover, the support for ODE models with discrete events was generally poor in state-of-the-art toolboxes, in particular with respect to gradient calculation.

In this article, I described the sensitivity equations for differential equation models with discrete events. I derived and applied a novel mathematical framework for parameter estimation using event-resolved experimental data by modeling event-triggered observations. To address issue (iv), I implemented forward sensitivity analysis for ODE models with events in the `AMICI` toolbox. I developed a model for GFP expression after transfection together with Joachim Rädler, who also provided experimental data for the model. Additionally I considered a model for spiking neurons, for which I simulated data. Both models include discrete events. I implemented both models in `AMICI` and performed parameter estimation using `PESTO`. For the model for GFP expression, I demonstrated that the computational cost is 2.2 fold lower when using forward sensitivity analysis compared to finite differences. For the model of spiking neurons, I showed that the computational cost is 3.9 fold lower for forward sensitivity analysis compared to finite differences. Moreover, the reproducibility of parameter estimates over multiple local optimizations runs was higher or forward sensitivity analysis for the model of spiking neurons compared to finite differences. For the model of spiking neurons, the parameters were only estimated from event-resolved outputs, a novel approach, which I introduced in this thesis. In the manuscript, I also introduced a novel regularization scheme which further improves the reproducibility of parameter estimates across multiple optimization runs. The regularization scheme is based on the final value of the event trigger function for events that did occur not in the simulation during the experimentally observed time-frame. The consideration of event-resolved data in the context of ODE models was previously not possible as the respective mathematical framework was not established and the methods were not implemented in any other systems biology

toolbox. Thus, the proposed approach enables the consideration of a new data-type of event-resolved datapoints, which is relevant for many biological applications that deal with time-to-event data such as recorded cell-death events.

In addition to the scientific contributions, I was the author in charge of the preparation of this publication. I had the idea for the mathematical framework to consider event-resolved experimental data and the respective regularization scheme. I wrote the first complete draft of the manuscript and consolidated the draft with the other authors. The final manuscript includes contributions from Jan Hasenauer and Fabian Theis.

Chapter 4

Discussion and Outlook

4.1 Discussion

Small- to medium-scale ordinary differential equation models are widely used to describe the temporal evolution of biological systems and established toolboxes exist for parameter estimation and uncertainty analysis. Yet, these ordinary differential equation models ignore the influence of stochastic fluctuations, which can lead to a bias in parameter estimates. Moreover, these small- to medium-scale models often ignore the complexity of biological systems, which limits the predictive power of respective models. For many applications, such as precision medicine for complex diseases, predictive models with many parameters and many state-variables are necessary. Yet, many of the established algorithms for parameter estimation and uncertainty analysis do not scale well with the number of parameter and state variables and are not applicable to large-scale models, as they are prohibitively computationally expensive. Furthermore, comprehensive models are prone to pronounce stiffness which results in high computational complexity.

The computational cost of parameter estimation and uncertainty analysis primarily depend on the computation time for numerical simulation of the model and corresponding sensitivity equations. For stochastic models, computationally much cheaper mesoscopic approximations such as the moment-closure approximation and the systems size expansion have been introduced, but the effect of the approximation error on parameter estimation was never evaluated. For deterministic models several approaches to reduce this computation time, such as sparse solvers and adjoint sensitivities, have been proposed, but were never rigorously implemented or evaluated for biological systems. Moreover, discrete events have been proposed to mitigate stiffness, but sensitivity based parameter estimation was never properly implemented or evaluated. To address these shortcomings, I studied the respective approaches and provided an efficient implementation in the software toolbox `AMICI`. Using this implementation, I studied the properties and performance of respective approaches on several application examples.

Firstly, I demonstrated that mesoscopic MA and SSE descriptions can provide computa-

tionally cheap, but accurate parameter estimates. Secondly, I demonstrated that adjoint sensitivities can be used to compute objective function gradients almost independent of the number of parameters. Thirdly, I demonstrated that adjoint sensitivity analysis can be used together with a sparse linear solver and parallelization to enable the consideration of large-scale ODE models. Lastly, I demonstrated how sensitivity equations for models with discrete events can be used to efficiently compute parameter estimates.

My thesis work resulted in a careful selection of scalable algorithms for simulation of model equations and corresponding sensitivity equations which enable the consideration of large-scale models for parameter estimation and uncertainty analysis. All of these methods are implemented in the software toolbox `AMICI`, which has already been used in many other research projects (Ballnus et al., 2017; Boiger et al., 2016; Loos et al., 2016; Maier et al., 2017). However several additional improvements to further reduce the simulation time for certain problem classes are possible. Moreover, improvements in optimization algorithms could further reduce the computation time. These possible extensions are discussed in the following.

4.2 Adjoint Sensitivity Analysis for Models with Events

The timescales of biological processes typically do not depend on environmental settings and usually generalize well. Consequently, the approximation of fast timescales via discrete events is relevant for detailed large-scale models and could reduce the stiffness of equations. This would be particularly useful, when coupling signaling models to gene regulator networks.

However, adjoint sensitivity analysis currently cannot be applied to models with parameter dependent events as respective update formulas for the adjoint state and quadrature equations are missing. Consequently, large-scale models that approximate fast timescales with discrete events currently remain intractable. For the discrete adjoint approach (Giles et al., 2003) respective update formulas have already been derived (Zhang et al., 2017), but do not directly extend to the continuous adjoint approach described in this thesis.

4.3 Stochastic Gradient Descent Optimization

In this thesis, I only considered parameter estimation using standard local optimizers namely interior-point and trust-region methods. As previously discussed, these approaches

require exact gradient information to perform well. The computation time of the exact gradient scales linearly with the number of considered experimental conditions, as individual numerical simulation of model and sensitivity equations are required for every experimental condition. This renders parameter estimation and uncertainty analysis intractable, when considering massive datasets such as the Cancer Cell Line Encyclopedia (Barretina et al., 2012), which include thousands to millions of experimental conditions.

In such settings, stochastic gradient descent could be a viable alternative, which have already been successfully applied for optimization for deep learning and other machine learning problems (Bottou, 2010; Sutskever et al., 2013). Stochastic gradient descent algorithms such as Root Mean Square Propagation, Adaptive Moment Estimation (Kingma and Ba, 2014) or the momentum method (Polyak, 1964) assume that the gradient information in some of the datasets is to some extent redundant and exploits this by subsampling the experimental conditions in every iteration. The complete dataset is considered over the span of multiple iterations, which together form an epoch. Stochastic gradient descent can easily be combined with adjoint sensitivity analysis, when applied to individual datasets, which is possible at no additional cost. Consequently, this approach could improve the scaling with respect to the number of considered experimental conditions. Furthermore, the Stochastic Gradient Descent approach could be combined with trust-region or variational methods (Wibisono et al., 2016) for the step size computation to further accelerate convergence speed.

4.4 Automatic Reconstruction of Large-Scale Models

The creation of comprehensive, genome-scale models remains challenging. For metabolic models, community driven endeavors have yielded genome-scale models (Herrgård et al., 2008; Thiele et al., 2013). For signaling dynamics, comprehensive reconstructions endeavors such as the Atlas of Cancer Signaling Network (Kuperstein et al., 2015) have been initiated, but currently lack rate laws, so no ODE model can be constructed. Moreover, several online resources that aggregate vast amounts of biological knowledge about interactions such as openBEL (Fluck et al., 2014), Pathway Commons (Cerami et al., 2011) or OmniPathDB (Turei et al., 2016) have been established. Complimentary, tools for the automatic construction comprehensive ODE models from these resources are currently in development (Gyori et al., 2017). However, the necessary level of detail at which such large-scale comprehensive models should describe the biological model is not clear.

Standard, data-driven model reduction methods (Maiwald et al., 2016; Transtrum and Qiu,

2016) could be used to assess the appropriate level of detail. However, these methods were developed for small- to medium-scale models and are computationally highly demanding. Even when combined with methods developed in this thesis, they are likely to be computationally intractable. As alternative to model reduction, model selection (Akaike, 1974; Arlot and Celisse, 2010) on a set of candidate models could be performed. With the availability of massive datasets in public databases (Barretina et al., 2012; Li et al., 2017; Yang et al., 2012), cross-validation methods can be employed. These cross-validation methods are data-intensive, but only require parameter estimation for which scalable methods have been developed in this thesis. Complimentary, automated model construction methods allow the generation of large sets of candidate models at varying level of detail. This would allow the data-driven identification of an optimal level of detail, which provides a minimally detailed but predictive model. Comprehensive models constructed according to this level of detail could be employed in a variety of systems biology applications and would bring us a large step closer to a truly holistic understanding of biological systems.

Bibliography

L. Adlung, S. Kar, M.-C. Wagner, B. She, S. Chakraborty, J. Bao, S. Lattermann, M. Boerries, H. Busch, P. Wuchter, A. D. Ho, J. Timmer, M. Schilling, T. Höfer, and U. Klingmüller. Protein abundance of AKT and ERK pathway components governs cell type-specific regulation of proliferation. *Mol. Syst. Biol.*, 13(1):904, 2017.

H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723, 1974.

A. Ale, P. Kirk, and M. P. H. Stumpf. A general moment expansion method for stochastic kinetic models. *J. Chem. Phys.*, 138(17):174101, 2013.

R. Alexander. Diagonally implicit Runge–Kutta methods for stiff O.D.E.'s. *SIAM J. Numer. Anal.*, 14(6):1006–1021, 1977.

P. R. Amestoy, T. A. Davis, and I. S. Duff. An approximate minimum degree ordering algorithm. *SIAM J. Matrix Anal. A.*, 17(4):886–905, 1996.

D. F. Anderson and D. J. Higham. Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics. *Multiscale Model. Simul.*, 10(1):146–179, 2012.

A. Andreychenko, L. Mikeev, D. Spieler, and V. Wolf. Parameter identification for Markov models of biochemical reactions. In *Computer Aided Verification*, pages 83–98. Springer, 2011.

P. J. Antsaklis, J. A. Stiver, and M. Lemmon. *Hybrid system modeling and autonomous control systems*. Springer, 1993.

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

E. A. Arriaga. Determining biological noise via single cell analysis. *Anal. Bioanal. Chem.*, 393(1):73–80, 2008.

A. C. Babtie and M. P. H. Stumpf. How to deal with parameters for whole-cell modelling. *J. R. Soc. Interface*, 14(133), 2017.

B. Ballnus, S. Hug, K. Hatz, L. Görlitz, J. Hasenauer, and F. J. Theis. Comprehensive benchmarking of Markov chain Monte Carlo methods for dynamical systems. *BMC Syst. Biol.*, 11(63), 2017.

E. Balsa-Canto and J. R. Banga. AMIGO, a toolbox for advanced model identification in systems biology using global optimization. *Bioinformatics*, 27(16):2311–2313, 2011.

J. R. Banga. Optimization in computational systems biology. *BMC Syst. Biol.*, 2(47), 2008.

A.-L. Barabasi and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, 2004.

J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, Jr, M. de˜Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

A. M. Barrette, M. Bouhaddou, and M. R. Birtwistle. Integrating transcriptomic data with mechanistic systems pharmacology models for virtual drug combination trials. *ACS Chemical Neuroscience*, 2017.

P. I. Barton, R. J. Allgor, W. F. Feehery, and S. Galán. Dynamic optimization in a discontinuous world. *Ind. Eng. Chem. Res.*, 37(3):966–981, 1998.

Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, 1996.

R. Boiger, J. Hasenauer, S. Hross, and B. Kaltenbacher. Integration based profile likelihood calculation for PDE constrained parameter estimation problems. *Inverse Prob.*, 32(12): 125009, 2016.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT*, pages 177–186. Springer, 2010.

J. R. Bowen, A. Acrivos, and A. K. Oppenheim. Singular perturbation refinement to quasi-steady state approximation in chemical kinetics. *Chem. Eng. Sci.*, 18(3):177–188, 1963.

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

J. C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18(85):50–64, 1964.

R. H. Byrd, R. B. Schnabel, and G. A. Shultz. A trust region algorithm for nonlinearly constrained optimization. *SIAM J. Numer. Anal.*, 24(5):1152–1170, 1987.

R. H. Byrd, J. C. Gilbert, and J. Nocedal. A trust region method based on interior point techniques for nonlinear programming. *Math. Program.*, 89(1):149–185, 2000.

E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. D. Bader, and C. Sander. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, 39(Database issue):D685–D690, 2011.

W. W. Chen, B. Schoeberl, P. J. Jasper, M. Niepel, U. B. Nielsen, D. A. Lauffenburger, and P. K. Sorger. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.*, 5(1):239, 2009.

O.-T. Chis, J. R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6(11):e27755, 2011.

T. F. Coleman and Y. Li. On the convergence of interior-reflective newton methods for nonlinear minimization subject to bounds. *Math. Program.*, 67(1):189–224, 1994.

D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. *J. Symb. Comp.*, 9(3):251–280, 1990.

S. Dano, M. F. Madsen, H. Schmidt, and G. Cedersund. Reduction of a biochemical model with preservation of its basic dynamic properties. *FEBS Journal*, 273(21):4862–4877, 2006.

T. A. Davis and E. Palamadai Natarajan. Algorithm 907: KLU, a direct sparse solver for circuit simulation problems. *ACM T. Math. Software.*, 37(3):36, 2010.

J. E. Dennis, Jr., D. M. Gay, and R. E. Welsch. Algorithm 573: Nl2sol—an adaptive nonlinear least-squares algorithm [E4]. *ACM T. Math. Software.*, 7(3):369–383, 1981.

R. P. Dickinson and R. J. Gelinas. Sensitivity analysis of ordinary differential equation systems—A direct method. *J. Comput. Phys.*, 21(2):123–143, 1976.

F. Eduati, V. Doldàn-Martelli, B. Klinger, T. Cokelaer, A. Sieber, F. Kogera, M. Dorel, M. J. Garnett, N. Blüthgen, and J. Saez-Rodriguez. Drug resistance mechanisms in colorectal cancer dissected with cell Type–Specific dynamic logic models. *Cancer Res.*, 77(12):3364–3375, 2017.

J. A. Egea, M. Rodriguez-Fernandez, J. R. Banga, and R. Marti. Scatter search for chemical and bio-process optimization. *J. Global Optim.*, 37(3):481–503, 2007.

J. A. Egea, D. Henriques, T. Cokelaer, A. F. Villaverde, A. MacNamara, D. P. Danciu, J. R. Banga, and J. Saez-Rodriguez. MEIGO: An open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. *BMC Bioinf.*, 15(136), 2014.

J. Elf and M. Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res.*, 13:2475–2484, 2003.

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

S. Engblom. *Numerical methods for the chemical master equation.* Ph.D. thesis, Uppsala University, Uppsala, Sweden, 2006b.

D. Fey, M. Halasz, D. Dreidax, S. P. Kennedy, J. F. Hastings, N. Rauch, A. G. Munoz, R. Pilkington, M. Fischer, F. Westermann, W. Kolch, B. N. Kholodenko, and D. R.

Croucher. Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.*, 8(408), 2015.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London, Ser. A*, 222:309–368, 1922.

J. Fluck, S. Madan, S. Ansari, et al. BELIEF – a semiautomatic workflow for BEL network creation. In *Proc. 6th Int. Symp. Semant. Min. Biomed*, pages 109–113, 2014.

F. Fröhlich, S. Hross, F. J. Theis, and J. Hasenauer. Radial basis function approximation of Bayesian parameter posterior densities for uncertainty analysis. In P. Mendes, J. O. Dada, and K. O. Smallbone, editors, *Proc. 12th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture Notes in Bioinformatics, pages 73–85. Springer International Publishing Switzerland, 2014a.

F. Fröhlich, F. J. Theis, and J. Hasenauer. Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more. In P. Mendes, J. O. Dada, and K. O. Smallbone, editors, *Proc. 12th Int. Conf. Comp. Meth. Syst. Biol.*, Lecture Notes in Bioinformatics, pages 61–72. Springer International Publishing Switzerland, 2014b.

K. Fujarewicz, M. Kimmel, and A. Swierniak. On fitting of mathematical models of cell signaling pathways using adjoint systems. *Math Bio Eng*, 2(3):527–534, 2005.

A. Gandhi, S. Levin, and S. Orszag. Moment expansions in spatial ecological models and moment closure through gaussian approximation. *Bull. Math. Biol.*, 62(4):595–632, 2000.

A. George and J. W. Liu. *Computer Solution of Large Sparse Positive Definite Systems.* Prentice Hall Professional Technical Reference, 1981.

M. B. Giles, M. C. Duta, J.-D. Muller, and N. A. Pierce. Algorithm developments for discrete adjoint methods. *AIAA J.*, 41(2):198–205, 2003.

C. S. Gillespie. Moment-closure approximations for mass-action models. *IET Syst. Biol.*, 3(1):52–58, 2009.

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

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188 (1):404–425, 1992.

D. T. Gillespie. The chemical Langevin equation. *J. Chem. Phys.*, 113(1):297–306, 2000.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reaction systems. *J. Chem. Phys.*, 115:1716–1733, 2001.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Statist. Soc. B*, 73(2):123–214, 2011.

P. Gonnet, S. Dimopoulos, L. Widmer, and J. Stelling. A specialized ODE integrator for the efficient computation of parameter sensitivities. *BMC Syst. Biol.*, 6(46), 2012.

A. Griewank and A. Walther. *Evaluating Derivatives*. Society for Industrial and Applied Mathematics, 2008.

R. Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *J. Chem. Phys.*, 133(035101), 2010.

R. Grima. Construction and accuracy of partial differential equation approximations to the chemical master equation. *Phys. Rev. E*, 84(5 Pt 2):056109, 2011.

R. Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.*, 136(15):154105, 2012.

R. Grima. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. *Phys. Rev. E*, 92(4):42–124, 2015.

B. M. Gyori, J. A. Bachman, K. Subramanian, J. L. Muhlich, L. Galescu, and P. K. Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11), 2017.

M. Hanke and O. Scherzer. Inverse problems light: Numerical differentiation. *Am. Math. Mon.*, 108(6):512–521, 2001.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

J. Hasenauer, N. Radde, M. Doszczak, P. Scheurich, and F. Allgöwer. Parameter estimation for the CME from noisy binned snapshot data: Formulation as maximum likelihood problem. Extended abstract at *Conf. of Stoch. Syst. Biol.*, Monte Verita, Switzerland, 2011.

J. Hasenauer, N. Jagiella, S. Hross, and F. J. Theis. Data-driven modelling of biological multi-scale processes. *Journal of Coupled Systems and Multiscale Dynamics*, 3(2):101–121, 2015.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 51(1):97–109, 1970.

M. J. Herrgård, N. Swainston, P. Dobson, W. B. Dunn, K. Y. Arga, M. Arvas, N. Blüthgen, S. Borger, R. Costenoble, M. Heinemann, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat. Biotechnol.*, 26(10):1155, 2008.

J. Hespanha. Moment closure for biochemical networks. In *Proc. Int. Symp. on Communications, Control and Signal Processing*, pages 42–147, 2008.

A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM T. Math. Software.*, 31(3):363–396, 2005.

S. Hoops, S. Sahle, R. Gauges, C. Lee, J. Pahle, N. Simus, M. Singhal, L. Xu, P. Mendes, and U. Kummer. COPASI – a COmplex PAthway SImulator. *Bioinformatics*, 22(24): 3067–3074, 2006.

S. Hross, A. Fiedler, F. J. Theis, and J. Hasenauer. Quantitative comparison of competing PDE models for Pom1p dynamics in fission yeast. In R. Findeisen, E. Bullinger, E. Balsa-Canto, and K. Bernaerts, editors, *Proc. 6th IFAC Conf. Found. Syst. Biol. Eng.*, volume 49, pages 264–269. IFAC-PapersOnLine, 2016.

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

J. Intosalmi, K. Nousiainen, H. Ahlfors, and H. Lähdesmäki. Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks. *Bioinformatics*, 32(12):i288–i296, 2016.

G. Jia, G. N. Stephanopoulos, and R. Gunawan. Parameter estimation of kinetic models from metabolic profiles: Two-phase dynamic decoupling method. *Bioinformatics*, 27 (4):196–1970, 2011.

J. R. Karr, A. H. Williams, J. D. Zucker, A. Raue, B. Steiert, J. Timmer, C. Kreutz, DREAM8 Parameter Estimation Challenge Consortium, S. Wilkinson, B. A. Allgood, B. M. Bot, B. R. Hoff, M. R. Kellen, M. W. Covert, G. A. Stolovitzky, and P. Meyer. Summary of the DREAM8 parameter estimation challenge: Toward parameter identification for whole-cell models. *PLoS Comput. Biol.*, 11(5):e1004096, 2015.

V. Kazeev, M. Khammash, M. Nip, and C. Schwab. Direct solution of the Chemical Master Equation using quantized tensor trains. *PLoS Comput. Biol.*, 10(3):e1003359, 2014.

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

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

H. Kitano. Systems biology: A brief overview. *Science*, 295(5560):1662–1664, 2002a.

H. Kitano. Computational systems biology. *Nature*, 420(6912):206–210, 2002b.

E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach. *Systems biology in practice.* Wiley-VCH, Weinheim, 2005.

P. Kokotovic and J. Heller. Direct and adjoint sensitivity equations for parameter optimization. *IEEE T. Autom. Contr.*, 12(5):609–610, 1967.

A. Kramer, J. Hasenauer, F. Allgöwer, and N. Radde. Computation of the posterior entropy in a Bayesian framework for parameter estimation in biological networks. In *Proc. IEEE Multi-Conf. Syst. Contr.*, pages 493–498, Yokohama, Japan, 2010.

I. Kuperstein, E. Bonnet, H.-A. Nguyen, D. Cohen, E. Viara, L. Grieco, S. Fourquet, L. Calzone, C. Russo, M. Kondratova, M. Dutreix, E. Barillot, and A. Zinovyev. Atlas of cancer signalling network: A systems biology resource for integrative analysis of cancer data with Google maps. *Oncogenesis*, 4:e160, 2015.

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

B. Lennartson, M. Tittus, B. Ehardt, and S. Pettersson. Hybrid systems in process control. *IEEE Control Syst. Mag.*, 16(5):45–56, 1996.

N. Le˜Novère. Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.*, 16(3):146–58, 2015.

K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Q Appl Math*, 2(2):164–168, 1944.

J. Li, W. Zhao, R. Akbani, W. Liu, Z. Ju, S. Ling, C. P. Vellano, P. Roebuck, Q. Yu, A. K. Eterovic, L. A. Byers, M. A. Davies, W. Deng, Y. N. V. Gopal, G. Chen, E. M. von Euw, D. Slamon, D. Conklin, J. V. Heymach, A. F. Gazdar, J. D. Minna, J. N. Myers, Y. Lu, G. B. Mills, and H. Liang. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell*, 31(2):225–239, 2017.

D. Liberzon and A. S. Morse. Basic problems in stability and design of switched systems. *IEEE Control Systems*, 19(5):59–70, 1999.

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

C. Loos, K. Moeller, F. Fröhlich, T. Hucho, and J. Hasenauer. Mechanistic hierarchical population model identifies latent causes of cell-to-cell variability. *bioRxiv*, 2017.

J. Lu, S. Muller, R. Machné, and C. Flamm. SBML ODE solver library: Extensions for inverse analysis. In *Proc. 5th Int. W. Comp. Syst. Biol.*, 2008.

J. Lu, E. August, and H. Koeppl. Inverse problems from biomedicine: Inference of putative disease mechanisms and robust therapeutic strategies. *J. Math. Biol.*, 67(1):143–168, 2012.

J. N. Lyness and C. B. Moler. Numerical differentiation of analytic functions. *SIAM J.*

Numer. Anal., 4(2):202–210, 1967.

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

T. Maiwald, H. Hass, B. Steiert, J. Vanlier, R. Engesser, A. Raue, F. Kipkeew, H. H. Bock, D. Kaschek, C. Kreutz, and J. Timmer. Driving the model to its limit: Profile likelihood based model reduction. *PLoS ONE*, 11(9), 2016.

D. W. Marquardt. An algorithm for least-squares estimation of non-linear parameters. *SIAM J Appl Math*, 11(22):431–441, 1963.

M. Mateescu, V. Wolf, F. Didier, and T. Henzinger. Fast adaptive uniformisation of the Chemical Master Equation. *IET Syst. Biol.*, 4(6):441–452, 2010.

H. McAdams and L. Shapiro. Circuit simulation of genetic networks. *Science*, 269(5224): 650–656, 1995.

S. Menz, J. C. Latorre, C. Schütte, and W. Huisinga. Hybrid stochastic deterministic solution of the Chemical Master Equation. *SIAM J. Multiscale Model. Simul.*, 10(4): 1232–1262, 2012.

L. Milne-Thompson. *The calculus of finite differences*. Macmillan, 1933.

C. G. Moles, P. Mendes, and J. R. Banga. Parameter estimation in biochemical pathways: A comparison of global optimization methods. *Genome Res.*, 13:2467–2474, 2003.

E. J. Molinelli, A. Korkut, W. Wang, M. L. Miller, N. P. Gauthier, X. Jing, P. Kaushik, Q. He, G. Mills, D. B. Solit, C. A. Pratilas, M. Weigt, A. Braunstein, A. Pagnani, R. Zecchina, and C. Sander. Perturbation biology: Inferring signaling networks in cellular systems. *PLoS Comput. Biol.*, 9(12):e1003290, 2013.

J. J. More. The Levenberg-Marquardt algorithm: Implementation and theory. *Lecture Notes in Mathematics*, 630:105–116, 1978.

B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.*, 124(4):044104, 2006.

S. A. Murphy and A. W. van der Vaart. On profile likelihood. *J. Am. Stat. Assoc.*, 95 (450):449–485, 2000.

G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden. Systematic identification of signal-activated stochastic gene regulation. *Science*, 339 (6119):584–587, 2013.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer Science & Business Media, 2006.

J. R. Norris. *Markov Chains*. Cambridge university press, 1998.

R.-E. Plessix. A review of the adjoint-state method for computing the gradient of a

functional with geophysical applications. *Geophys. J. Int.*, 167(2):495–503, 2006.

B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Comp. Math. Math. Phys.*, 4(5):1–17, 1964.

R. Ramaswamy, N. González-Segredo, and I. F. Sbalzarini. A new class of highly efficient exact Stochastic Simulation Algorithms for chemical reaction networks. *J. Chem. Phys.*, 130(2009):244104, 2009.

R. Ramaswamy, N. González-Segredo, I. Sbalzarini, and R. Grima. Discreteness-induced concentration inversion in mesoscopic chemical systems. *Nat. Comm.*, 3(779), 2012.

M. Rathinam, L. Petzold, Y. Cao, and D. T. Gillespie. Stiffness in stochastic chemically reaction systems: The implicit tau-leaping method. *J. Chem. Phys.*, 119(24):12784–12794, 2003.

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

A. Raue, C. Kreutz, F. J. Theis, and J. Timmer. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. *Philos T Roy Soc A*, 371(1984), 2013a.

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

A. Raue, B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tönsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Höfer, E. Klipp, F. J. Theis, U. Klingmüller, B. Schöberl, and J.Timmer. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, 2015.

S. Reinker, R. M. Altman, and J. Timmer. Parameter estimation in stochastic biochemical reactions. *IEE Proc. Syst. Biol.*, 153(4):168–178, 2006.

M. Reiter, B. Kirchner, H. Müller, C. Holzhauer, W. Mann, and M. W. Pfaffl. Quantification noise in single cell experiments. *Nucleic Acids Res.*, 39(18):e124–e124, 2011.

H. Resat, L. Petzold, and M. F. Pettigrew. Kinetic modeling of biological systems. *Methods Mol. Biol.*, 541:311–335, 2009.

H. Risken. *The Fokker-Planck equation: Methods of solution and applications*. Springer, 1996.

G. O. Roberts and J. S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *J. R. Soc. B*, 60(1):255–268, 1998.

M. Rodriguez-Fernandez, J. A. Egea, and J. R. Banga. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinf.*, 7(483), 2006.

H. H. Rosenbrock. Some general implicit processes for the numerical solution of differential equations. *Comput. J.*, 5(4):329–330, 1963.

E. N. Rozenvasser. General sensitivity equations of discontinuous systems. *Avtomat i Telemekh*, 3:52–56, 1967.

J. Ruess and J. Lygeros. Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. *ACM T Math Softwares. Model. Comput. S.*, 25(2):8, 2015.

K. R. Sanft, S. Wu, M. Roh, J. Fu, R. K. Lim, and L. R. Petzold. StochKit2: Software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458, 2011.

S. D. M. Santos, P. J. Verveer, and P. I. H. Bastiaens. Growth factor-induced MAPK network topology shapes Erk response determining PC-12 cell fate. *Nat. Cell Biol.*, 9 (3):324–330, 2007.

M. Schilling, T. Maiwald, S. Hengl, D. Winter, C. Kreutz, W. Kolch, W. D. Lehmann, J. Timmer, and U. Klingmüller. Theoretical and experimental analysis links isoform-specific ERK signalling to cell fate decisions. *Mol. Syst. Biol.*, 5(334), 2009.

C. Schillings, M. Sunnåker, J. Stelling, and C. Schwab. Efficient characterization of parametric uncertainty of complex (bio)chemical networks. *PLoS Comput. Biol.*, 11(8): e1004457, 2015.

N. N. Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.

R. Serban and A. C. Hindmarsh. CVODES: An ODE solver with sensitivity analysis capabilities. *ACM Math. Software*, 31(3):363–396, 2005.

M. Shamir, Y. Bar-On, R. Phillips, and R. Milo. SnapShot: Timescales in cell biology. *Cell*, 164(6):1302–1302, 2016.

A. Singh and J. Hespanha. A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.*, 69(6):1909–1925, 2007.

K. Smallbone and P. Mendes. Large-scale metabolic models: From reconstruction to differential equations. *Ind. Biotechnol.*, 9(4):179–184, 2013.

T. J. Snowden, P. H. van der Graaf, and M. J. Tindall. Methods of model reduction for large-scale biological systems: A survey of current methods and trends. *B. Math. Biol.*, 79(7):1449–1486, 2017.

E. T. Somogyi, J.-M. Bouteiller, J. A. Glazier, M. König, J. K. Medley, M. H. Swat, and H. M. Sauro. libRoadRunner: A high performance SBML simulation and analysis

library. *Bioinformatics*, 31(20):3315–3321, 2015.

D. C. Sorensen. Newton's method with a model trust region modification. *SIAM J. Numer. Anal.*, 19(2):409–426, 1982.

W. Squire and G. Trapp. Using complex variables to estimate derivatives of real functions. *SIAM Rev.*, 40(1):110–112, 1998.

P. Stapor, D. Weindl, B. Ballnus, S. Hug, C. Loos, A. Fiedler, S. Krause, S. Hross, F. Fröhlich, and J. Hasenauer. PESTO: Parameter EStimation TOolbox. *Bioinformatics*, btx676, 2017.

V. Sunkara and M. Hegland. An optimal Finite State Projection method. *Procedia Comput. Sci.*, 1(1):1579–1586, 2010.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *Proc. Int. Conf. Machine Learning*, pages 1139–1147, 2013.

I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, S. K. Chavali, P. Dobson, W. B. Dunn, L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le˜Novère, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, Sr, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van˜Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. O. Palsson. A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, 31(5): 419–425, 2013.

J. Thorson. Gaussian elimination on a banded matrix, 1979.

T. Toni, Y.-i. Ozaki, P. Kirk, S. Kuroda, and M. P. H. Stumpf. Elucidating the in vivo phosphorylation dynamics of the ERK MAP kinase using quantitative proteomics data and Bayesian model selection. *Mol. Biosyst.*, 8:1921–1929, 2012.

M. K. Transtrum and P. Qiu. Bridging mechanistic and phenomenological models of complex biological systems. *PLoS Comput. Biol.*, 12(5):1–34, 2016.

D. Turei, T. Korcsmaros, and J. Saez-Rodriguez. OmniPath: Guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods*, 13(12):966–967, 2016.

N. G. van Kampen. *Stochastic processes in physics and chemistry*. North-Holland, 2007.

D. Venzon and S. Moolgavkar. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37(1):87–94, 1988.

R. Waltz, J. Morales, J. Nocedal, and D. Orban. An interior algorithm for nonlinear optimization that combines line search and trust region steps. *Math. Program.*, 107(3): 391–408, 2006.

A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proc. Natl. Acad. Sci. U˜S˜A*, 113(47):E7351–E7358, 2016.

D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinf.*, 8(2):109–116, 2007.

D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, 1(1):67–82, 1997.

W. Yang, J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J. A. Smith, I. R. Thompson, et al. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.*, 41(D1):D955–D961, 2012.

J. Yao, A. Pilko, and R. Wollman. Distinct cellular states determine calcium signaling response. *Mol. Syst. Biol.*, 12(12):894, 2016.

H. Zhang and A. Sandu. FATODE: A library for forward, adjoint, and tangent linear integration of ODEs. *SIAM J. Sci. Comput.*, 36(5):C504–C523, 2014.

H. Zhang, S. Abhyankar, E. Constantinescu, and M. Anitescu. Discrete adjoint sensitivity analysis of hybrid dynamical systems with switching. *IEEE T. Circ. Syst.*, 64(5):1247–1259, 2017.

Appendix A

Inference for stochastic chemical kinetics using moment equations and system size expansion.

This is the copyedited PDF of an article accepted for publication in PLoS Computational Biology: F. Fröhlich, P. Thomas, A. Kazeroonian, F.J. Theis, R. Grima, J. Hasenauer. **Inference for stochastic chemical kinetics using moment equations and system size expansion.** PLoS Computational Biology 12(7):e1005030 (2016).

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion

Fabian Fröhlich[1,2], Philipp Thomas[3], Atefeh Kazeroonian[1,2], Fabian J. Theis[1,2], Ramon Grima[4]*, Jan Hasenauer[1,2]*

1 Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany, **2** Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Garching, Germany, **3** Department of Mathematics, Imperial College London, London, United Kingdom, **4** School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

* ramon.grima@ed.ac.uk (RG); jan.hasenauer@helmholtz-muenchen.de (JH)

Citation: Fröhlich F, Thomas P, Kazeroonian A, Theis FJ, Grima R, Hasenauer J (2016) Inference for Stochastic Chemical Kinetics Using Moment Equations and System Size Expansion. PLoS Comput Biol 12(7): e1005030. doi:10.1371/journal.pcbi.1005030

Editor: Daniel A Beard, University of Michigan, UNITED STATES

Received: September 28, 2015

Accepted: June 23, 2016

Published: July 22, 2016

Data Availability Statement: All relevant code and data are within the paper and its Supporting Information files.

Funding: This work was supported by the German Research Foundation (DFG; http://dfg.de) through the Graduate School of Quantitative Biosciences Munich (QBM, FF), the Federal Ministry of Education and Research (BMBF; http://www.bmbf.de) within the SYS-Stomach project (Grant No. 01ZX1310B; JH), the Postdoctoral Fellowship Program (PFP) of the Helmholtz Zentrum München (https://www.helmholtz-muenchen.de/en/fellows/index.html; JH), the

Abstract

Quantitative mechanistic models are valuable tools for disentangling biochemical pathways and for achieving a comprehensive understanding of biological systems. However, to be quantitative the parameters of these models have to be estimated from experimental data. In the presence of significant stochastic fluctuations this is a challenging task as stochastic simulations are usually too time-consuming and a macroscopic description using reaction rate equations (RREs) is no longer accurate. In this manuscript, we therefore consider moment-closure approximation (MA) and the system size expansion (SSE), which approximate the statistical moments of stochastic processes and tend to be more precise than macroscopic descriptions. We introduce gradient-based parameter optimization methods and uncertainty analysis methods for MA and SSE. Efficiency and reliability of the methods are assessed using simulation examples as well as by an application to data for Epo-induced JAK/STAT signaling. The application revealed that even if merely population-average data are available, MA and SSE improve parameter identifiability in comparison to RRE. Furthermore, the simulation examples revealed that the resulting estimates are more reliable for an intermediate volume regime. In this regime the estimation error is reduced and we propose methods to determine the regime boundaries. These results illustrate that inference using MA and SSE is feasible and possesses a high sensitivity.

Author Summary

In this manuscript, we introduce efficient methods for parameter estimation for stochastic processes. The stochasticity of chemical reactions can influence the average behavior of the considered system. For some biological systems, a microscopic, stochastic description is computationally intractable but a macroscopic, deterministic description too inaccurate. This inaccuracy manifests itself in an error in parameter estimates, which impede the

European Union within the ERC grant 'Latent Causes' (http://erc.europa.eu; FJT), the Royal Commission for the Exhibition of 1851 (https://www.royalcommission1851.org) in form of a fellowship (PT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

predictive power of the proposed model. Until now, no rigorous analysis on the magnitude of the estimation error exists. We show by means of two simulation examples that using mesoscopic descriptions based on the system size expansions and moment-closure approximations can reduce this estimation error compared to inference using a macroscopic description. This reduction is most pronounced in an intermediate volume regime where the influence of stochasticity on the average behavior is moderately strong. For the JAK/STAT pathway where experimental data is available, we show that one parameter that was not structurally identifiable when using a macroscopic description becomes structurally identifiable when using a mesoscopic description for parameter estimation.

Introduction

On the single-cell level many biological processes are influenced by stochastic fluctuations [1–3]. This stochasticity must be accounted for when constructing quantitative mechanistic models for the behavior of cells. Traditionally, dynamics of stochastic biochemical processes are modeled using the Chemical Master Equation (CME) [4]. The CME provides an accurate microscopic description of stochastic chemical kinetics [5] and enables the prediction of the behavior of biochemical reaction networks. To achieve high prediction accuracy, however, the parameters of the CME have to be inferred from experimental data. This inference is challenging and the development of new methods to perform efficient inference is the subject of current research.

In the literature, methods to perform statistical inference for single-cell time-lapse data [6–13] and populations snapshot data [14–20] have been proposed. These methods use the Stochastic Simulation Algorithm (SSA) [21], as well as various approximations of the CME such as the Finite State Projection (FSP) [22], moment closure approximations (MA) [23] and the linear-noise approximation (LNA) [24]. We next provide a brief discussion of these methods, in particular their use to infer the parameters from experimental single-cell data—a visual summary of these methods and their properties is provided in Fig 1. In this manuscript we will only consider population snapshot data and thus focus on the respective methods.

The parameters of stochastic processes are frequently inferred using Approximate Bayesian Computing approaches [25]. These methods rely on exhaustive stochastic simulations and accept parameter values if the differences between simulation and experimental data is sufficiently small [7, 13, 19]. While many methods which exploit stochastic simulations are asymptotically exact, their computational efficiency suffers from the required number of simulations. While SSA-based methods are asymptotically exact, appropriate stopping criteria and distance measures are difficult to obtain [26]. Furthermore, the computational efficiency of Approximate Bayesian Computing methods suffers from the tremendous number of required SSA runs.

Inference using FSP methods is usually more efficient than using the SSA [20]. The parameter dependent probability distribution of the process is simulated and the likelihood of the data under this distribution is evaluated (Fig 1b). This likelihood function is a multinomial probability distribution [15, 16] and efficient gradient-based optimization methods can be used [18]. The ODE systems might however be large and hence their simulation is intractable even when using state-of-the-art sliding window [27] and tensor train approaches [12]. Even with tailored methods [12, 27, 28], the simulation of many reaction networks remains computationally intractable and hence FSP-based inference is still very limited.

To circumvent the computational complexity of evaluating the full probability distribution, MA [29–32] and the SSE methods [24] have been introduced. Both classes of methods approximate the statistical moments of the stochastic process which is described by the CME:

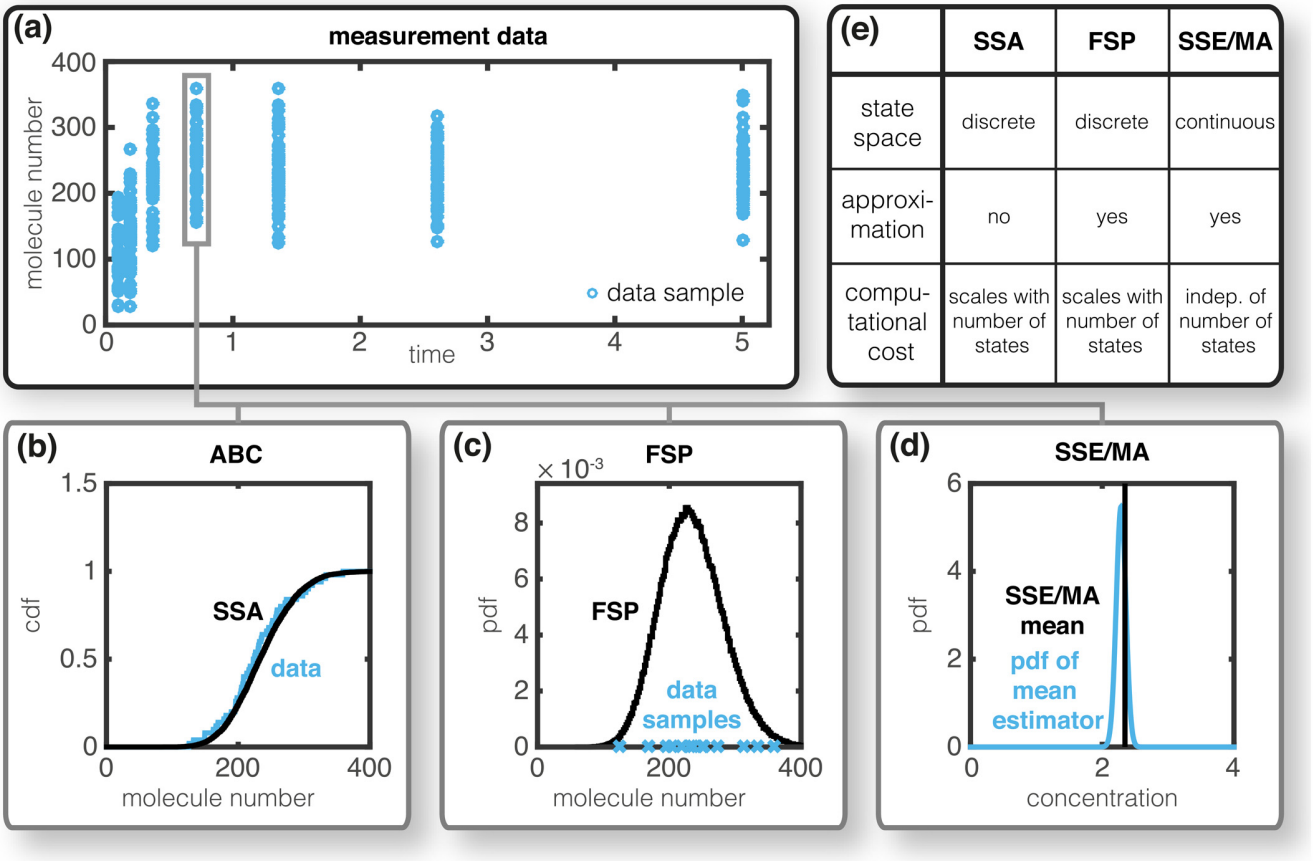


Fig 1. Inference methods for stochastic processes. (a) Single-cell snapshot data collected using a high-throughput technique, such as flow cytometry. (b) Empirical density functions for SSA runs (black —) and experimental data (blue —), the difference is used as distance measure in Approximate Bayesian Computing. (c) Instantaneous probability distribution computed using FSP (black —) to evaluate the likelihood of the observing the individual cells (blue ×). (d) Mean computed using MA/SSE (black —) as well as measured mean and its uncertainty (blue —). (e) Summary of the properties of the displayed methods.

doi:10.1371/journal.pcbi.1005030.g001

- The MA is based on the hierarchy of evolution equations for the statistical moments of the CME solution. This hierarchy is truncated at an order N and the $(N+1)$th order moments usually contained in the remaining system are approximated by functions of the lower-order moments. This approximation is based on an assumption of the distribution solution of the CME [33–35]. The Nth order MA is in the following denoted by NMA.
- The SSE of the CME is a series expansion in the inverse volume of the compartment in which the system is confined [24]. The leading order in the mean gives the reaction rate equations (RRE) while the leading order in the variance gives the LNA. The consideration of additional terms in the expansion gives the expected mesoscopic rate equation (EMRE) [36] (the first-order correction to the RRE) and the inverse omega square (IOS) method [37] (the first-order correction to the LNA).

Both MA and SSE approaches generate a system of coupled ordinary differential equations (ODEs) for the approximate moments. It has been shown that the difference between MA and SSE methods decreases with increasing volume and approaches the solution of the CME [23]. The accuracy of MA equations and the conditions under which they provide physically meaningful results have recently been studied for several distribution choices [23, 33, 38].

For statistical inference of parameters the LNA and 2MA have recently been used [10, 17, 18, 39, 40]. The comparison of the measured and simulated moments often provides good parameter estimates [17, 18] and the corresponding estimation problems are tractable. Besides reducing the computational complexity, MA and SSE approaches also enable the application of techniques which were already established for deterministic models, e.g., structural identifiability analysis [41]. Accordingly, the literature for the application of MA and SSE methods for inference is promising, there is however plenty of room for improvement: (i) in none of the studies have gradient-based optimization methods with sensitivity equations been employed, even though they have been shown to be superior for a wide range of dynamical systems [42]; (ii) the estimation error of inferred parameter values is influenced by the fact that the MA and LNA typically provide an approximation of the moments for chemical systems with at least one bimolecular reaction (see [43] for more details on when the LNA is exact). Hence a systematic evaluation of estimation errors in the inferred parameter values, say as a function of the compartment volume is direly needed so that one can decide which modeling approach is best suited for a given compartment volume. (iii) it has been shown that EMRE and IOS yield more accurate approximations to the CME than possible using the LNA and RRE [36, 43–45] (although there are exceptions such as when the LNA is exact up to second-order moments with the CME [43]). Similarly in the limit of large volumes, it has been shown that higher-order MA equations are more accurate than lower-order ones [23]; for example the 3MA is more accurate than the commonly used 2MA. However to-date the equations derived by considering the terms in the SSE beyond the LNA and the equations obtained using the 3MA have not been used for inference.

In this manuscript, we will introduce an efficient gradient- and sensitivity-based method for parameter estimation for population snapshot data using MA and SSE-based approaches. This method is evaluated on experimental data available for the JAK/STAT signaling pathway model, which is traditionally modeled using the RRE. For this model, we demonstrate that our approach yields additional insight. Subsequent to this application part, we systematically evaluate the estimation error for two biochemical networks, each with at least one bimolecular reaction. We will provide a first quantification of the improvement achieved using the 3MA and the SSE truncated beyond the next to leading-order term over the RREs, 2MA and LNA. Using this evaluation, two simple approaches for the selection of the correct inference approach will be proposed.

Methods

In the following we outline the considered modeling approaches, parameter estimation, uncertainty analysis, model selection. The workflow is shown in Fig 2.

Chemical master equation

Consider a set of R reactions, involving M chemical species confined in a reaction volume of size Ω. Denoting the set of reactants by $(X_1, \ldots, X_M)$, the r^{th} reaction can be written as

$$\sum_{i=1}^{M} v_{ir}^{-} X_i \xrightarrow{k_r} \sum_{i=1}^{M} v_{ir}^{+} X_i.$$

Here k_r is the reaction rate constant, $v_{ir}^{\pm}$ are the integer stoichiometric coefficients, and we denote by $v_{ir} = v_{ir}^{+} - v_{ir}^{-}$ the change in molecules of the i^{th} species in the r^{th} reaction. Under well-mixed conditions the state of this biochemical system is characterized by the corresponding vector of molecule numbers $n = (n_1, \ldots, n_M)$. The time-evolution of the probability of

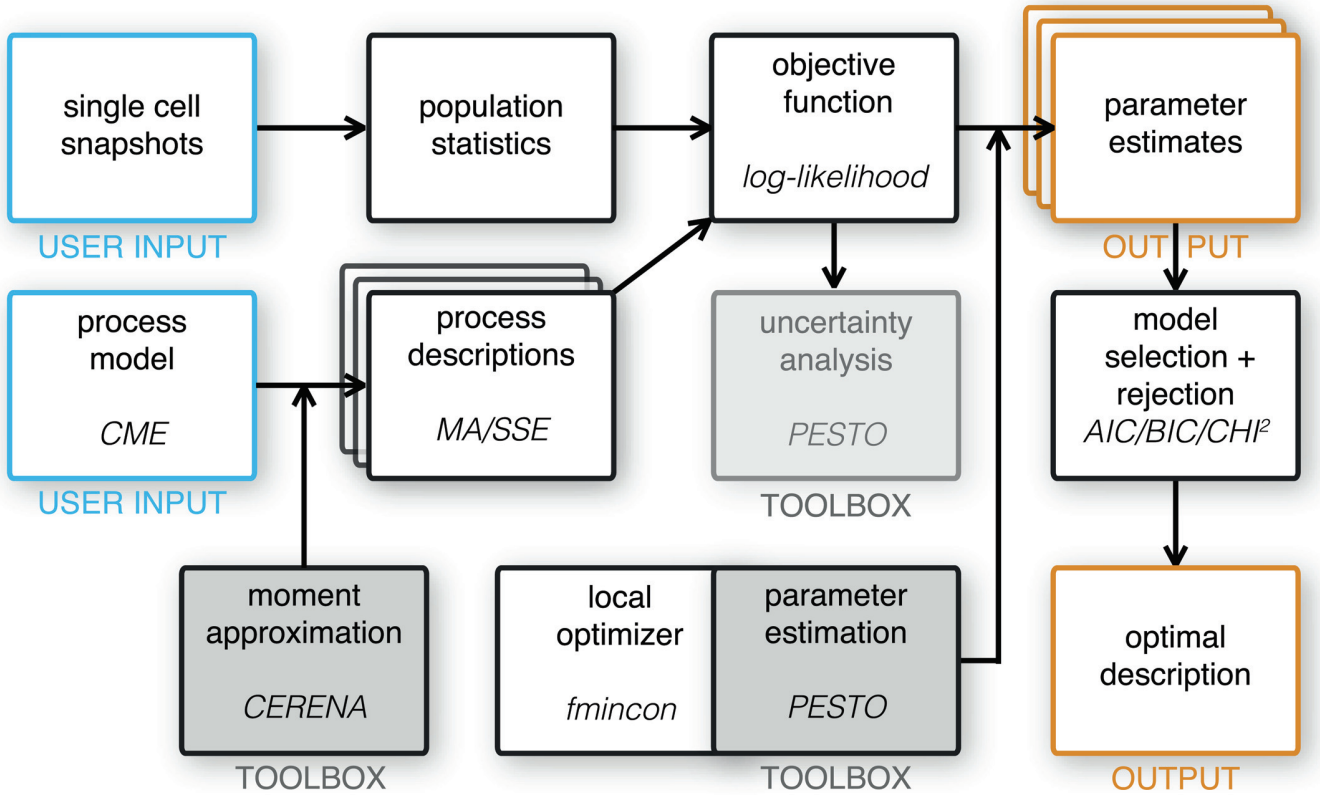


Fig 2. Workflow for modeling, parameter estimation and model selection. User inputs are colored in blue, workflow outputs are colored in orange. MATLAB toolboxes are indicated by gray boxes. The employed method/function/toolbox is indicated as oblique text in every box where applicable.

doi:10.1371/journal.pcbi.1005030.g002

observing the system in state n, then obeys the CME

$$\frac{\mathrm{d}P(n,t)}{\mathrm{d}t} = \Omega \sum_{r=1}^{R} \left[\hat{f}_r(n - v_r)P(n - v_r, t) - \hat{f}_r(n)P(n,t) \right]. \quad (1)$$

Here, v_r denotes the stoichiometry $(v_{1r}, \ldots, v_{Mr})$ of the r^{th} reaction and $\Omega\hat{f}_r(n)$ is the propensity function, i.e., the probability per unit time for reaction r to occur somewhere in the volume Ω. Since the CME is often intractable for analytical solution, we here focus on approximation methods for the mean concentrations $\mu_i = \langle n_i/\Omega \rangle$, and the corresponding covariances of the concentration fluctuations about them, $\Sigma_{ij} = \langle (n_i/\Omega - \mu_i)(n_j/\Omega - \mu_j) \rangle$, which is outlined in the following.

Moment-closure approximation

Equations for the moments are straightforwardly derived from the CME Eq (1). For systems involving non-linear propensities, however, these equations are intractable because the equation for a certain moment is typically coupled to higher-order moments resulting in an infinite system of equations. A common procedure to break this hierarchy of moment equations is to neglect higher than second order cumulants [29]; this indeed is the same as assuming that the third order cumulant is consistent with a Gaussian distribution. Assuming at most bimolecular reactions, the result is a set of non-linear ODEs coupling mean and variance called the 2MA

and is given by

$$\begin{aligned}\frac{\partial \mu_i}{\partial t} &= \sum_{r=1}^{R} v_{ir}\left(\hat{f}_r(\Omega\mu) + \frac{1}{2}\sum_{s,l=1}^{M}\frac{\partial^2 \hat{f}_r(\Omega\mu)}{\partial\mu_s\partial\mu_l}\Sigma_{sl}\right),\\ \frac{\partial \Sigma_{ij}}{\partial t} &= \sum_{r=1}^{R}\sum_{s=1}^{M}\left(v_{ir}\frac{\partial \hat{f}_r(\Omega\mu)}{\partial\mu_s}\Sigma_{sj} + v_{jr}\frac{\partial \hat{f}_r(\Omega\mu)}{\partial\mu_s}\Sigma_{si}\right)\\ &\quad + \frac{1}{\Omega}\sum_{r=1}^{R} v_{ir}v_{jr}\left(\hat{f}_r(\Omega\mu) + \sum_{s,l=1}^{M}\frac{\partial^2 \hat{f}_r(\Omega\mu)}{\partial\mu_s\partial\mu_l}\Sigma_{sl}\right).\end{aligned}$$

The 2MA is precise for unimolecular reactions and fairly accurate if the third order moment is negligible [29]. The latter is mostly the case for large reaction volume and molecule numbers [29]. For small volumes higher-order moment equation must be used. Neglecting higher than third order cumulants yields the 3rd order moment-closure approximation (3MA) that are outlined in Ref. [29, 30]. The simulation routines were generated using the CERENA toolbox [46].

System size expansion

A different technique to approximate the moments of the CME is given by the SSE. The procedure allows us to expand the CME about the solution of the RREs which are valid for large reaction volumes Ω and are given by

$$\frac{\partial \phi_i}{\partial t} = \sum_{r=1}^{R} v_{ir} f_r(\phi)\,.$$

Here $f_j(\phi) = \lim_{\Omega\to\infty}\hat{f}_j(\Omega\phi)$ denotes the macroscopic rate function. While the RREs represent the leading order term of the SSE and yield the average concentrations for large volumes Ω, the next term, the LNA, describes the fluctuations about these concentrations. The covariance of these fluctuations obeys [24, 47]:

$$\frac{\partial \Sigma_{ij}}{\partial t} = \sum_{r=1}^{R}\sum_{s=1}^{M}\left(v_{ir}\frac{\partial f_r(\phi)}{\partial\phi_s}\Sigma_{sj} + v_{jr}\frac{\partial f_r(\phi)}{\partial\phi_s}\Sigma_{si}\right) + \frac{1}{\Omega}\sum_{r=1}^{R} v_{ir}v_{jr}f_r(\phi).$$

These results are exact for reaction networks comprising up to unimolecular reactions and for a small subset of networks with bimolecular reactions [43]. For most networks involving bimolecular reactions, the SSE enables us to systematically correct the mean concentrations of the RREs and the variance predictions of the LNA, by considering higher order terms in the expansion. A more accurate estimate for the mean concentrations than the RREs is given by the EMRE [36], and follows

$$\frac{\partial \mu_i}{\partial t} = \frac{\partial \phi_i}{\partial t} + \sum_{r=1}^{R} v_{ir}\left(\sum_{s=1}^{M}\frac{\partial f_r(\phi)}{\partial\phi_s}(\mu_s - \phi_s) + \frac{1}{2}\sum_{s,l=1}^{M}\frac{\partial^2 f_r(\phi)}{\partial\phi_s\phi_l}\Sigma_{sl} - \frac{1}{2}\sum_{s=1}^{M}\frac{\phi_s}{\Omega}\frac{\partial^2 f_r(\phi)}{\partial\phi_s^2}\right).$$

Note that these equations yield a correction term of order Ω^{-1} to the RREs. Correspondingly, expressions for the covariances about these more accurate concentrations can be derived using the IOS approximation, which corrects the LNA estimate to order Ω^{-2} [37]. In contrast to RRE and LNA, EMRE and IOS do not assume large volumes and hence these estimates are expected to be closer to the true moments predicted by the CME.

In what follows we shall collectively refer to the EMRE and IOS as higher-order SSEs, meaning they are obtained using the SSE truncated to a higher-order than that giving the LNA. The simulation routines were generated using the CERENA toolbox [46].

Statistical model of experimental data

In this study we consider population average data as well as single-cell snapshot data. Population average data could, among others, be obtained by Western blot and (bulk) mRNA sequencing. Single-cell snapshot data could be obtained by flow and mass cytometry. Some statistical properties of these data types are introduced in the following.

Population average data. These data provide information about the mean $\mu_i(t_k, \theta)$ of measured quantities $\hat{\mu}_{i,k}$ at times t_k,

$$\hat{\mu}_{i,k} = \mu_i(t_k, \theta) + \epsilon_{i,k,\mathrm{T}}\,.$$

These measurements are noise corrupted. The measurement noise $\epsilon_{i,\,k,\,\mathrm{T}}$ is in the following assumed to be independently and distributed with mean zero and variance $\sigma^2_{\hat{\mu}_{i,k},\mathrm{T}}$, $\epsilon_{i,k,\mathrm{T}} \sim \mathcal{N}(0, \sigma^2_{\hat{\mu}_{i,k},\mathrm{T}})$ and true population mean $\mu_i(t_k, \theta)$.

Single-cell snapshot data. These data provide information about the measured quantities y_i at times t_k for individual cells. The single cell measurements are given by

$$\hat{y}^{(j)}_{i,k} = y^{(j)}_{i,k} + \epsilon_{i,k,\mathrm{T}}, \qquad j = 1, \ldots, N\,,$$

with $y^{(j)}_{i,k}$ denoting a sample from the cell population, $y^{(j)}_{i,k} \sim p(y_i|t_k, \theta)$ with mean $\mu_i(t_k, \theta)$, variance $\Sigma_{ii}(t_k, \theta)$ and fourth order central moment $\Sigma_{iiii}(t_k, \theta)$. The technical noise is assumed to depend on the replicate and therefore independent of j. From these samples mean and variances,

$$\hat{\mu}_{i,k} = \frac{1}{N}\sum_{j=1}^{N} \hat{y}^{(j)}_{i,k} \quad \text{and} \quad \hat{\Sigma}_{ii,k} = \frac{1}{N}\sum_{j}^{N} \left(\hat{y}^{(j)}_{i,k} - \hat{\mu}_{i,k}\right)^2,$$

as well as higher-order moments can be estimated. According to the central limit theorem, these estimators are approximately normally distributed for $N \gg 1$. The estimator of the mean, $\hat{\mu}_{i,k}$, possesses the variance

$$\sigma^2_{\hat{\mu}_{i,k}} = \mathrm{E}\left[\left(\hat{\mu}_{i,k} - \mu_i(t_k, \theta)\right)^2\right] = \underbrace{\mathrm{E}\left[\left(\frac{1}{N}\sum_{j=1}^{N} y^{(j)}_{i,k} - \mu_i(t_k, \theta)\right)^2\right]}_{=\ \text{statistical uncertainty}} + \underbrace{\mathbb{E}\left[\epsilon^2_{i,k,\mathrm{T}}\right]}_{=\ \text{technical noise}},$$

where the last reformulation exploits independence of $y^{(j)}_{i,k}$ and $\epsilon^2_{i,k,\mathrm{T}}$. The first summand has the value $\frac{1}{N}\Sigma_{ii,k}$ (see [17, 48]) and describe the statistical noise resulting from the finite number of measured cells. As the sample size N grows, this variance contribution goes to zeros. In contrast, the second summand is the variance of the technical noise, $\sigma^2_{\hat{\mu}_{i,k},\mathrm{T}}$, which is independent of the sample size. This yields the overall variance

$$\sigma^2_{\hat{\mu}_{i,k}} = \frac{1}{N}\Sigma_{ii}(t_k, \theta) + \sigma^2_{\hat{\mu}_{i,k},\mathrm{T}},$$

The estimator of the variance, $\hat{\Sigma}_{ii,k}$, possesses the variance

$$\sigma^2_{\hat{\Sigma}_{ii,k}} = \mathrm{E}\left[\left(\hat{\Sigma}_{ii,k} - \Sigma_{ii}(t_k, \theta)\right)^2\right] = \frac{1}{N}\left(\Sigma_{iiii}(t_k, \theta) - \frac{N-3}{N-1}\Sigma^2_{ii}(t_k, \theta)\right),$$

which is independent of the technical noise. For a detailed derivation we refer the reader to the supplement. Note that the estimates of mean and variance are potentially correlated if both are computed from the same sample [48].

The statistical description of population snapshot data also provides a framework for population average data. Experimental techniques providing population average typically analyze millions of single cells simultaneously. Accordingly, N is rather large, yielding the variance $\sigma^2_{\hat{\mu}_{i,k}} \approx \sigma^2_{\hat{\mu}_{i,k},\mathrm{T}}$.

Modeling of noise variance. The variance of mean and variance estimators, $\sigma^2_{\hat{\mu}_{i,k}}$ and $\sigma^2_{\hat{\Sigma}_{ii,k}}$ depends on the statistical moments of the process and the variance of the technical noise. The moments $\Sigma_{ii}(t_k, \theta)$ and $\Sigma_{iiii}(t_k, \theta)$ could be computed using higher-order MA and SSE. However, this can be computationally intensive and subject to approximation errors. Instead, we used the sample-based estimates of these statistical moments, $\hat{\Sigma}_{ii,k} = \frac{1}{N}\sum_j^N \left(\hat{y}^{(j)}_{i,k} - \hat{\mu}_{i,k}\right)^2$ and $\hat{\Sigma}_{iiii,k} = \frac{1}{N}\sum_j^N \left(\hat{y}^{(j)}_{i,k} - \hat{\mu}_{i,k}\right)^4$. These estimates are rather reliable (for $N \gg 1$) and are not influenced by technical noise. Accordingly, the variance of the technical noise, $\sigma^2_{\hat{\mu}_{i,k},\mathrm{T}}$, can either be obtained by computing the statistics over multiple experimental replicates with large sample sizes ($N \gg 1$), or by modeling them as a possibly parameter dependent function. For generality, we assume in the following that the variances of the estimators are parameter dependent, $\sigma^2_{\hat{\mu}_{i,k}}(\theta)$ and $\sigma^2_{\hat{\Sigma}_{ii,k}}(\theta)$.

Parameter estimation

To infer the parameters of biochemical reaction networks we employ maximum likelihood and Bayesian parameter estimation. Based upon the statistical model introduced above, the likelihood function becomes

$$\mathcal{L}(\theta) = \underbrace{\prod_{i,k} \frac{1}{\sqrt{2\pi\sigma^2_{\hat{\mu}_{i,k}}(\theta)}} \exp\left(-\frac{1}{2}\left(\frac{\mu_i(t_k,\theta) - \hat{\mu}_{i,k}}{\sigma_{\hat{\mu}_{i,k}}(\theta)}\right)^2\right)}_{(1)\ \text{likelihood of measured means}} \times \underbrace{\prod_{i,k} \frac{1}{\sqrt{2\pi\sigma^2_{\hat{\Sigma}_{ii,k}}(\theta)}} \exp\left(-\frac{1}{2}\left(\frac{\Sigma_{ii}(t_k,\theta) - \hat{\Sigma}_{ii,k}}{\sigma_{\hat{\Sigma}_{ii,k}}(\theta)}\right)^2\right)}_{(2)\ \text{likelihood of measured variances}}.$$

The two contributions, (1) and (2), provide the likelihood of measured mean and measured variance of the data, respectively. In the absence of information about the variance, part (2) is set to one. To improve the numerical robustness and the convergence properties of optimizers, instead of maximizing the likelihood, the negative log-likelihood

$$J(\theta) = \frac{1}{2}\sum_{k,i}\left(\log 2\pi\sigma^2_{\hat{\mu}_{i,k}}(\theta) + \left(\frac{\mu_i(t_k,\theta) - \hat{\mu}_{i,k}}{\sigma_{\hat{\mu}_{i,k}}(\theta)}\right)^2\right) + \frac{1}{2}\sum_{k,i}\left(\log 2\pi\sigma^2_{\hat{\Sigma}_{ii,k}}(\theta) + \left(\frac{\Sigma_{ii}(t_k,\theta) - \hat{\Sigma}_{ii,k}}{\sigma_{\hat{\Sigma}_{ii,k}}(\theta)}\right)^2\right)$$

is minimized [42]. The corresponding minimization problem is

$$\hat{\theta} = \arg\min_{\theta\in\Theta} J(\theta)\,,$$

with plausible parameter domain Θ. The minimizer $\hat{\theta}$ of $J(\theta)$ is the maximum likelihood estimate. In practice, a further improvement is often achieved by optimizing the log-transformed parameter $\xi = \log\theta$ instead of θ [42].

The optimization of the objective function has been implemented in MATLAB using our in-house Parameter Estimation Toolbox (PESTO). PESTO uses a multi-start local optimization scheme, an approach which has been shown to perform well for similar problems [42]. To ensure a good coverage of the domain Θ [42], the starting points for the local solvers were generated using a latin hypercube sampling between the lower and upper bounds for the parameters defined by Θ. In order to exploit gradient and curvature information in the local optimization we made use of the trust-region-reflective algorithm [49, 50] implemented in the MATLAB routine `fmincon.m`.

The gradient of the objective function with respect to parameter θ_l is given by

$$\frac{\partial J}{\partial \theta_l} = \sum_{i,k} \frac{1}{\sigma^2_{\hat{\mu}_{i,k}}(\theta)} \left(1 - \left(\frac{\mu_i(t_k,\theta) - \hat{\mu}_{i,k}}{\sigma^2_{\hat{\mu}_{i,k}}(\theta)}\right)^2\right) \frac{\partial \sigma^2_{\hat{\mu}_{i,k}}}{\partial \theta_l}\bigg|_\theta + \frac{\mu_i(t_k,\theta) - \hat{\mu}_{i,k}}{\sigma^2_{\hat{\mu}_{i,k}}(\theta)} \frac{\partial \mu_i}{\partial \theta_l}\bigg|_{t_k,\theta}$$
$$+ \sum_{i,k} \frac{1}{\sigma^2_{\hat{\Sigma}_{ii,k}}(\theta)} \left(1 - \left(\frac{\Sigma_i(t_k,\theta) - \hat{\Sigma}_{ii,k}}{\sigma^2_{\hat{\mu}_{i,k}}(\theta)}\right)^2\right) \frac{\partial \sigma^2_{\hat{\Sigma}_{i,k}}}{\partial \theta_l}\bigg|_\theta + \frac{\Sigma_{ii}(t_k,\theta) - \hat{\Sigma}_{i,k}}{\Sigma^2_{\hat{\mu}_{i,k}}(\theta)} \frac{\partial \Sigma_{ii}}{\partial \theta_l}\bigg|_{t_k,\theta},$$

in which $\partial\mu_i/\partial\theta_l := (s_l^\mu)_i$ and $\partial\Sigma_{ii}/\partial\theta_l := (s_l^\Sigma)_i$ denote the sensitivity of mean and variance with respect to the parameters. The governing equations for the sensitivities s_l^μ and s_l^Σ are derived by differentiation of the evolution equations and subsequent reordering. For the 2MA the sensitivities are governed by:

$$\frac{\partial s_l^\mu}{\partial t} = \left(\frac{\partial}{\partial \mu}\left(\frac{\partial \mu}{\partial t}\right)\right) s_l^\mu + \left(\frac{\partial}{\partial \Sigma}\left(\frac{\partial \mu}{\partial t}\right)\right) s_l^\Sigma + \frac{\partial}{\partial \theta_l}\left(\frac{\partial \mu}{\partial t}\right),$$
$$\frac{\partial s_l^\Sigma}{\partial t} = \left(\frac{\partial}{\partial \mu}\left(\frac{\partial \Sigma}{\partial t}\right)\right) s_l^\mu + \left(\frac{\partial}{\partial \Sigma}\left(\frac{\partial \Sigma}{\partial t}\right)\right) s_l^\Sigma + \frac{\partial}{\partial \theta_l}\left(\frac{\partial \Sigma}{\partial t}\right),$$

in which $\partial\mu/\partial t$ and $\partial\Sigma/\partial t$ denote the right-hand side of the evolution equations for the 2MA. For the EMRE the sensitivities are governed by:

$$\frac{\partial s_l^\Phi}{\partial t} = \left(\frac{\partial}{\partial \Phi}\left(\frac{\partial \Phi}{\partial t}\right)\right) s_l^\Phi + \left(\frac{\partial}{\partial \mu}\left(\frac{\partial \Phi}{\partial t}\right)\right) s_l^\mu + \left(\frac{\partial}{\partial \Sigma}\left(\frac{\partial \Phi}{\partial t}\right)\right) s_l^\Sigma + \frac{\partial}{\partial \theta_l}\left(\frac{\partial \Phi}{\partial t}\right),$$
$$\frac{\partial s_l^\mu}{\partial t} = \left(\frac{\partial}{\partial \Phi}\left(\frac{\partial \mu}{\partial t}\right)\right) s_l^\Phi + \left(\frac{\partial}{\partial \mu}\left(\frac{\partial \mu}{\partial t}\right)\right) s_l^\mu + \left(\frac{\partial}{\partial \Sigma}\left(\frac{\partial \mu}{\partial t}\right)\right) s_l^\Sigma + \frac{\partial}{\partial \theta_l}\left(\frac{\partial \mu}{\partial t}\right),$$
$$\frac{\partial s_l^\Sigma}{\partial t} = \left(\frac{\partial}{\partial \Phi}\left(\frac{\partial \Sigma}{\partial t}\right)\right) s_l^\Phi + \left(\frac{\partial}{\partial \mu}\left(\frac{\partial \Sigma}{\partial t}\right)\right) s_l^\mu + \left(\frac{\partial}{\partial \Sigma}\left(\frac{\partial \Sigma}{\partial t}\right)\right) s_l^\Sigma + \frac{\partial}{\partial \theta_l}\left(\frac{\partial \Sigma}{\partial t}\right),$$

in which $\partial s_l^\Phi = \partial\Phi/\partial\theta_l$ is the sensitivity of the solution of the reaction rate equation and $\partial\Phi/\partial t$, $\partial\mu/\partial t$ and $\partial\Sigma/\partial t$ denote the right-hand side of the evolution equations for the EMRE. The sensitivity equations for RRE, 3MA and IOS possess a similar structure as those for 2MA and EMRE. In principle all the sensitivity equations can be obtained by rewriting the respective systems into systems of ODEs and using generic methods (see, e.g., [51]).

The gradient of the objective function was computed using forward sensitivity equations to ensure robust and efficient evaluation [42]. In addition to gradient information, we supplied `fmincon.m` with the Fisher-Information Matrix as approximation to the Hessian of the objective function to accelerate the optimization. This approximation of the Hessian is equivalent to the formulation in Levenberg-Marquardt [52] type optimization schemes. Parameter and objective function tolerances were both set to 10^{-6}. For every dataset, the multi-start scheme was initialized at 50 initial values using a latin hypercube sampling. Convergence to a local and supposedly global optimum was checked by ensuring that a minimum of 5 of the 50

starts yielded the same minimal objective function value. If convergence was not observed, we doubled the number of multi-starts until this criterion was met.

Uncertainty analysis

Experimental data of biochemical processes is often scarce and noise corrupted, resulting in non-identifiabilities and parameter uncertainties. Parameter identifiability is typically assessed using structural and practical identifiability analysis (see [41, 53] and references therein). Structural identifiability analysis provides information for the considered model topology and measured output, independent of a specific dataset. In contrast, practical identifiability and uncertainty analysis provide information about the reliability of parameter estimates for a given dataset. In this study we use profile likelihoods [54, 55] and Bayesian methods [56, 57] to study practical identifiability and parameter uncertainties.

The profile likelihood of a parameter θ_i, denoted by $\mathrm{PL}(\theta_i)$, is given by the likelihood maximized over the remaining parameters,

$$\mathrm{PL}(\theta_i) = \max_{\theta_j \neq i, \theta \in \Theta} \mathcal{L}(\theta).$$

Accordingly, profile likelihoods can be computed by solving a set of constrained optimization problems requiring repeated local optimization. In this study this task was carried out using the toolbox PESTO. Frequentist confidence intervals can be computed by comparing the profile likelihood $\mathrm{PL}(\theta_i)$ to the likelihood $\mathcal{L}(\hat{\theta})$ at the globally optimal parameter point $\hat{\theta}$ [58]. As the models considered here can contain structurally non-identifiable parameters, profile likelihoods are the only viable frequentist technique for global uncertainty analysis [59].

Bayesian uncertainty analysis methods rely on Bayes' theorem,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})},$$

in which $p(\theta)$, $p(\mathcal{D}|\theta)(=\mathcal{L}(\theta))$, $p(\mathcal{D})$ and $p(\theta|\mathcal{D})$ denote prior probability, likelihood, evidence and posterior distribution, respectively [56]. For determining Bayesian credibility intervals of the parameters, we sampled from the posterior distribution using the efficient adaptive Markov Chain Monte Carlo (MCMC) method delayed rejection adaptive metropolis [60]. From the multivariate samples the respective univariate Bayesian confidence intervals were computed. We collected a total of 10^5 samples after a burn-in period of 10^4. In accordance with the log-transformed parameters used for optimization, a log-uniform prior over the parameter domain Θ has been employed.

Model selection

For comparing competing model alternatives, we used Akaike's Information Criterion (AIC),

$$\mathrm{AIC}_l = -2\log\mathcal{L}(\hat{\theta}_l) + 2n_{\theta,l}\,.$$

The AIC of the l-th model depends on the maximum of the likelihood, $\hat{\theta}_l$, and the number of estimated parameters $n_{\theta,l}$. Therefore, the AIC accounts for the match of model and data as well as for model complexity. The model with the lowest AIC value and index l^* is selected. In order to simplify the interpretation of individual AIC values, we employ Akaike weights [61] defined by

$$w_{\mathrm{AIC},l} = \frac{\exp\left(-\frac{1}{2}(\mathrm{AIC}_l - \mathrm{AIC}_{l^*})\right)}{\sum_{l'}\exp\left(-\frac{1}{2}(\mathrm{AIC}_{l'} - \mathrm{AIC}_{l^*})\right)}.$$

The AIC weight $w_{\mathrm{AIC},\,l}$ of the l-th model is related to its posterior probability [61].

Reliability of our results has been ensured by comparing these values to the Bayesian information criterion (BIC) [62] and their corresponding BIC weights. As the number of parameters of the different models (e.g., RRE, EMRE and 2MA) is very similar, the results of these model selection criteria were comparable.

Model falsification

Model selection criteria provide information about the relative quality of competing models, but not about their respective goodness-of-fit. The best model l^* may still fail to adequately describe the measured data. To assess whether a model fits the data appropriately, we considered the sum of squared residuals at the optimal parameter value $\hat{\theta}$ [63],

$$\chi^2(\hat{\theta}) = \sum_{i,k} \underbrace{\left(\frac{\mu_i(t_k,\hat{\theta}) - \hat{\mu}_{i,k}}{\sigma_{\hat{\mu}_{i,k}}(\hat{\theta})} \right)^2}_{:=r^2_{\mu_{i,k}}} + \sum_{i,k} \underbrace{\left(\frac{\Sigma_{ii}(t_k,\hat{\theta}) - \hat{\Sigma}_{ii,k}}{\sigma_{\hat{\Sigma}_{ii,k}}(\hat{\theta})} \right)^2}_{:=r^2_{\Sigma_{ii,k}}}.$$

The sum of squared residuals is a standard goodness-of-fit statistic and is equal to $-2\log\mathcal{L}(\hat{\theta})$ put to a negative constant. As for adequate models the residuals $r_{\mu_{i,\,k}}$ and $r_{\Sigma_{ii,\,k}}$ should be normally distributed with unit variance, the sum of squared residuals should be drawn from a χ^2-distribution [64]. The number of degrees of freedom of the χ^2 distribution is the number of data points minus number of parameters. Accordingly, the χ^2-test can be used for model rejection [65, 66].

Results

In the following, we will illustrate how MA and SSE can be used to infer the parameters of stochastic biochemical processes. We will outline how the results can be interpreted and tested, and which novel insights can be gained even when only population-average data is available. For this purpose, we study an example for which experimental data is available and two examples for which artificial data was generated using stochastic simulations. The application to experimental data should substantiate the relevance of the developed methods in real-world application whereas the application to simulation examples allows for a more detailed analysis of the method properties.

Application to experimental data: The JAK/STAT signaling pathway

To evaluate MA and SSE based inference in a real-world application, we study the dynamics of the Janus family of kinases (JAK)-signal transducer and activator of transcription (STAT) signaling pathway [67]. Constitutive activation of STATs is related to the malignancy of many tumors [68]. Moreover, Erythropoietin (Epo), the upstream activation factor of the JAK/STAT signaling pathway, is administered as therapeutic agent for treatment of cancer related anaemia [69]. This is the case although several adverse effects such as increased tumour progression and thromboembolic events have been attributed to Epo [69, 70].

The core module of the JAK/STAT signaling pathway is composed of the Erythropoietin receptor (EpoR) and the transcription factor STAT5. Upon phosphorylation, the Epo receptor induces phosphorylation of STAT5 via the JAK2 kinase. Phosphorylated STAT5 (pSTAT) can dimerize and the pSTAT dimer can translocate to the nucleus to activate the transcription of target genes. The dimer dissociates and is exported to the cytoplasm after some delay, which is described by a sequence of intermediate states. The biochemical reaction network is depicted

in Fig 3(a). A more detailed description of the employed mathematical model is provided in S1 Supporting Information Section 1.2.

The JAK/STAT signaling pathway is a well studied system [53, 67]. For inference, we use Western Blot data for the phosphorylated Epo receptor (pEpoR), the cytoplasmic phosphorylated STAT (pSTAT), and the cytoplasmic STAT (tSTAT). These Western Blots average concentrations in thousands of cells, thereby provide information about the population mean but not about cell-to-cell variability. Due to the large cell numbers, statistic uncertainty can be ignored ($\frac{1}{N}\Sigma_{ii}(t_k,\theta)=0$). The technical noise of each measured species was estimated as additional log-scaled parameters ($\sigma^2_{\bar{\mu}_{\mathrm{pSTAT},k},\mathrm{T}}=10^{\theta_{15}}$,$\sigma^2_{\bar{\mu}_{\mathrm{tSTAT},k},\mathrm{T}}=10^{\theta_{16}}$,$\sigma^2_{\bar{\mu}_{\mathrm{pEpoR},k},\mathrm{T}}=10^{\theta_{17}}$). The data have been recorded by Swameye et al. [67] and are depicted by the black stars in Fig 3(b).

Mesoscopic description of the JAK/STAT signaling. A RRE model for the JAK/STAT signaling pathway has been introduced by Swameye et al. [67] and analyzed/extended in subsequent publications [53, 71]. Microscopic and mesoscopic descriptions of the process have not been studied yet. Thus, it remains unclear which role stochasticity plays in this process and how valid the RRE description is. To address this, we derived 2MA and EMRE models for the process (S1 Code). As the JAK/STAT pathway involves two compartments, the cytoplasm and the nucleus, we applied a simple extension of the MA and SSE to multiple compartments (see SI for details). The extension essentially leads to a rescaling of propensities for reactions that transport chemical species between compartments and ensures the correctness of parameter estimates of the associated kinetic rates.

The 2MA and EMRE models are studied along with the well-known RRE model by Raue et al. [53]. All three descriptions possess 5 mechanistic parameters: 4 kinetic parameters ($p_1,\ldots,p_4$); and the initial concentration of STAT5 in the cytoplasm ($[\mathrm{STAT}]_0$). For all descriptions, the pEpoR concentration is modeled as a time-dependent cubic spline function with 5 parameters. Furthermore, 7 nuisance parameters are used, i.e. scaling factors, noise variances. The number of state variables for RRE, EMRE and 2MA are 8, 52 and 44, respectively. As the dimerization reaction possesses a nonlinear propensity, the predictions for the mean of the underlying stochastic process differ between the models. Moreover, the phosphorylation of STAT5 depends on the pEpoR concentration which, as the concentration is modeled as a spline function, gives rise to a time-dependent propensity.

Efficient multi-start local optimization makes parameter inference feasible. As parameter estimation for RRE was reported to be challenging [53, 55], we evaluated multi-start local optimization for 2MA, EMRE and RRE using a large number of multi-starts (1000). Similar to previous studies we used the trust-region-reflective method in the MATLAB routine `fmincon.m`. To demonstrate the importance of accurate gradient calculation we compared results obtained using forward sensitivity equations and finite difference approximations. Forward sensitivities were computed using CVODES while finite differences were evaluated with a step-size of 10^{-4}. The results are illustrated in Fig 3(b)–3(f).

Optimization using finite differences does not work reliably for the three considered descriptions. This can be attributed to poor accuracy in gradient computations. In regions where the objective function gradient entries are small, for instance close to the optimum, approximation errors caused by numerical integration of the ODE models can dominate over actual entries and thus lead to poor search directions. This can lead to premature termination of the optimization, if the objective function is locally ascending in the chosen search direction. The lowest objective function value achieved for finite differences is greater than the value obtained using forward sensitivities (Fig 3(a)). Moreover, no plateaus are observed [42]. This is the case for the RRE as well as 2MA and EMRE model. Using forward sensitivity equations we observed reproducible optimization results, substantiating that the global optimum is found. In

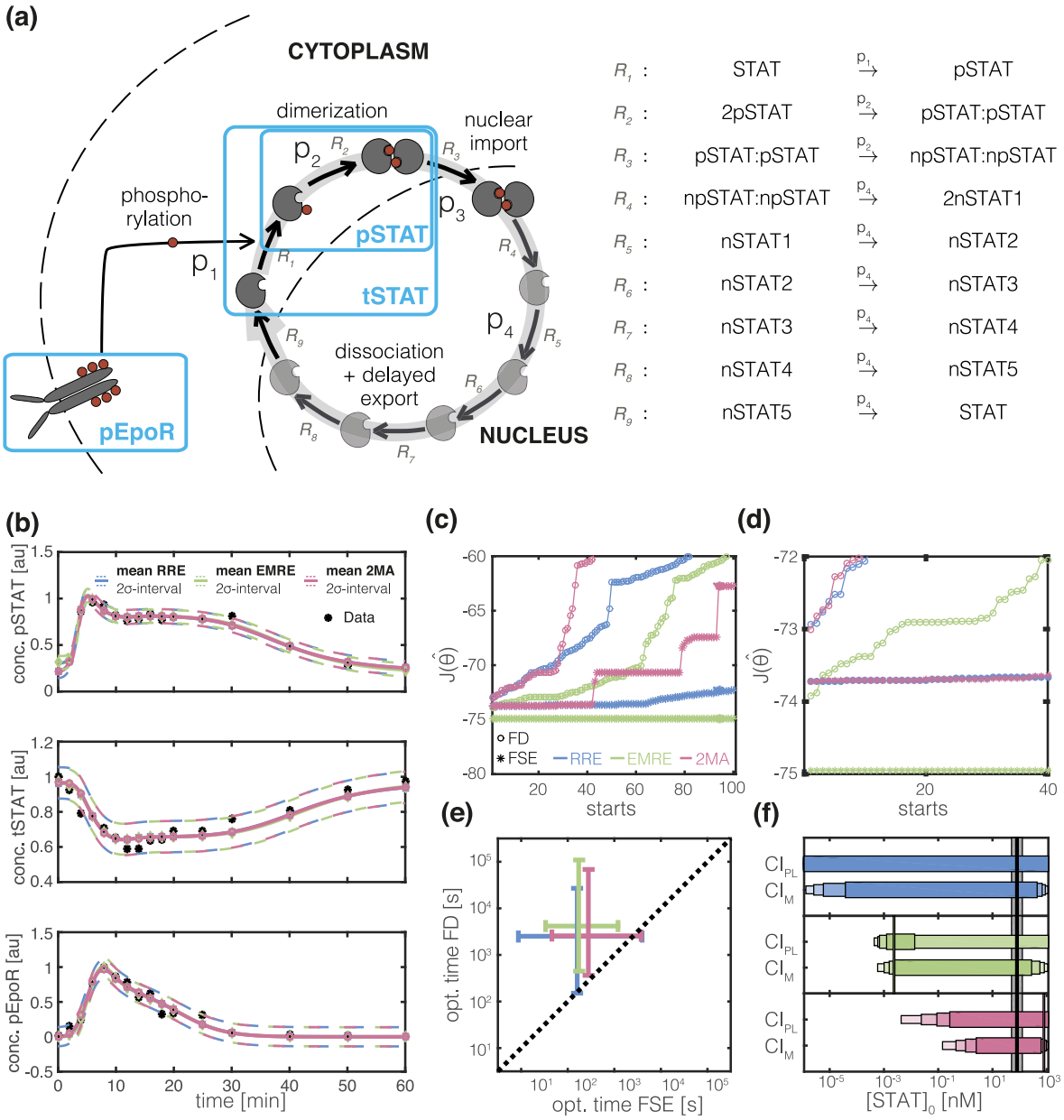


Fig 3. Parameter inference using EMRE and 2MA for JAK/STAT signaling pathway. (a) Schematic of JAK/STAT signaling pathway including biochemical reactions (→), biochemical species (gray elements) and observed outputs (blue boxes). Elements introduced to capture the delayed export of pSTAT from the nucleus are indicates as light gray. For subplots (b)-(e): RRE (blue), EMRE (green) and 2MA (red). (b) Experimental data (∗), fitted mean (—) and estimated 2σ interval of the measurement noise (- -). (c) Objective function values for the best 100 (out of 1000) multi-starts obtained using forward sensitivity analysis (FSE, *) and finite differences (FD, °) for gradient calculation. Local optimization for RRE, EMRE and 2MA used the same initial parameter values. (d) Zoom-in of the 40 best multi-starts. (e) Median (+) and 80% percentile interval of computation time per local optimizer run. (f) Estimate of initial STAT concentration. Vertical lines mark the maximum likelihood estimates and the horizontal bars represent the confidence(CI_{PL})/credibility(CI_M) intervals corresponding to different significance levels (80%, 90%, 95% and 99%) computed using profile likelihoods/MCMC samples. The reference value with 95% confidence intervals [71] is depicted by a black line and gray bar respectively.

doi:10.1371/journal.pcbi.1005030.g003

addition to the superior convergence rate, the median computation times for one local optimization were consistently more than 10-fold faster using forward sensitivity analysis compared to finite differences (Fig 3(e)). This finding supports previous findings for ODE models [42] and underlines the importance of employing forward sensitivities as an efficient and robust gradient computation scheme.

A comparison across models revealed that the fitting results for RRE and 2MA are visually indistinguishable, while the EMRE differ slightly from both (Fig 3(b)). Furthermore, optimization of the RRE was indeed computationally most efficient (Fig 3(e)). The computation times for EMRE and 2MA were however only slightly higher. Interestingly, the minimal objective function value was more frequently reached for the EMRE and MA compared to RRE (Fig 3 (d)). This indicates a larger region of attraction, reducing the number of required multi-starts and the convergence of alternative global optimization methods. Our results verify the practical feasibility of parameter inference using mesoscopic descriptions and potentially simpler objective function landscape.

Mesoscopic descriptions improve data exploitation. Optimization yielded the maximum likelihood estimates for the parameters of the biochemical process. Due to limited and noise corrupted data, these maximum likelihood estimates are often unreliable. We evaluated the uncertainty of the parameters obtained using RRE, EMRE and 2MA via profile likelihood calculation and Markov chain Monte-Carlo sampling. Profile likelihoods and marginal densities are provided in Figure B in S1 Supporting Information.

Profiles and marginals indicate identifiability of the four kinetic parameters p_1-p_4. Confidence intervals for these parameters are finite and agree for RRE, EMRE and 2MA. The initial STAT concentration, $[STAT]_0$, has been shown to be structurally non-identifiable when using RRE [53]. This implied that independent of the amount of measurement data, the initial STAT concentration cannot be inferred using the RRE. Accordingly, the RRE yielded flat profiles for the initial STAT concentration. This was different for EMRE and 2MA. For EMRE, the lower bound of the 99% confidence and credibility intervals computed using profiles and marginals is $8 \cdot 10^{-3}$ nM for the initial STAT concentration. For 2MA, we found lower bounds of $2 \cdot 10^{-2}$ nM and $1 \cdot 10^{-1}$ nM using profiles and marginals, respectively. This lower bound could only be derived as the reaction propensities are nonlinear and the reaction volumes as well as molecule numbers are finite. In this case the dynamics of the population mean are affected by fluctuations, which are controlled by initial concentrations. This dependency established structural identifiability and enabled us to exploit features of the data that could not be used by the RRE.

This finding is in line with results reported in the literature, which suggested that stochasticity can be exploited to improve the identifiability of parameters [18, 40, 72, 73]. Yet, previous analysis relied on using the process mean and variance for inference. The latter is only available for single-cell measurements. We demonstrated that stochasticity can be exploited even if only the process mean is available for inference. This renders stochastic inference attractive even if single-cell data is not available.

Literature validates lower bound for previously structurally non-identifiable parameter. To verify the lower bound for the initial STAT concentration derived using EMRE and 2MA, we screened additional literature. We found that Bachmann et al. [71] determined a STAT concentration of 80 nM under similar experimental conditions. This value is within the confidence/credibility bounds for both, EMRE and 2MA. While Bachman et al. [71] considered a different cell types, their results provide a partial confirmation of our finding.

In summary, the study of the JAK/STAT signaling pathway using EMRE and 2MA demonstrated the applicability of mesoscopic descriptions to real-world data. Using multi-start local optimization with accurate gradients, model parameters can be inferred from experimental

data. Frequentist and Bayesian uncertainty analysis revealed that MA and SSE can provide additional insights, even if merely population-average data are available.

Application to artificial data: Trimerization and enzymatic degradation

To assess the properties and potential of inference using mesoscopic descriptions (MA and SSE) in more detail, we study two processes: trimerization and enzymatic degradation. The use of artificial data enabled us to: (i) assess the estimation error introduced by macroscopic and mesoscopic descriptions; (ii) deduce a rule-of-thumb for the *a priori* selection of modeling approaches; and (iii) develop methods for the *a posteriori* selection and verification of modeling approaches.

Model description and artificial data generation. In the remainder, we study the trimerization process and the enzymatic degradation process depicted in Fig 4(a) and 4(b). The icons for the models introduced in Fig 4(a) and 4(b) will be used in the following figures to indicate the model in the respective study.

The trimerization process describes the bursty synthesis of monomers and their subsequent dimerization and trimerization [44]. Relevant biological applications of this model include

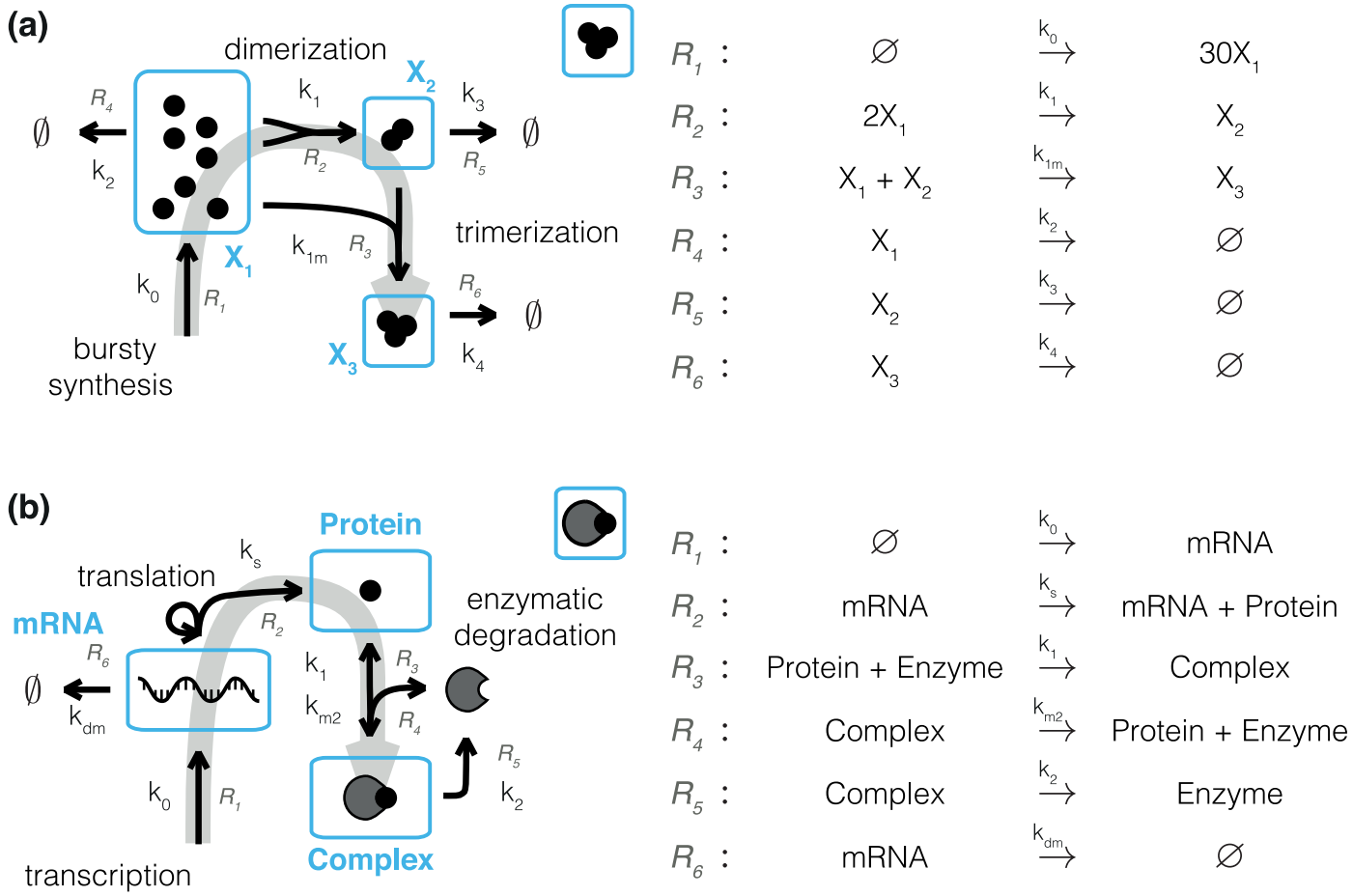


Fig 4. Reaction networks for comprehensive *in silico* evaluation of mesoscopic and macroscopic approaches. (a) Schematic of the trimerization process. (b) Schematic of the enzymatic degradation process. Arrows indicate reactions with the corresponding rate and reaction index next to them. Observed states are outlined and labeled in blue. A gray arrow represents the direction of information flow.

doi:10.1371/journal.pcbi.1005030.g004

receptor clustering and heat-shock factor trimerization [74, 75]. The trimerization process consists of 6 reactions and possesses 7 parameters (6 kinetic parameters and the reaction volume). Two reactions are bimolecular and hence have nonlinear propensities. Monomer, dimer and trimer concentrations are assumed to be measurable.

The enzymatic degradation process is an extension of the well-known two-stage model of gene expression [39, 76] and it has previously been studied in [77]. The enzymatic degradation process describes transcription and translation as well as enzymatic degradation of the gene product. It comprises several models of gene expression as special cases, e.g. [78–80]. The process consists of 6 reactions and possesses 8 parameters (6 kinetic parameters, the initial concentration of the enzyme and the reaction volume). The reaction resulting in the formation of the protein-enzyme complex is bimolecular and hence its propensity is nonlinear. The measured outputs are the mRNA, protein and complex concentrations.

A detailed mathematical description of trimerization and enzymatic degradation process is provided in S1 Supporting Information, Section 1.2.

For trimerization and enzymatic degradation process artificial data are generated using the SSA [21] with the parameter values in (Table B,D in S1 Supporting Information). A range of volumes Ω is considered to facilitate a comprehensive analysis of stochastic effects on estimation accuracy and to assess the regimes of validity for the different approximations. We considered realistic sample sizes in the range of $N_k = 10^1 - 10^4$, which are accessible by recent single-cell technologies [81]. The results of the parameter inference of the trimerization and the enzymatic degradation process, which are depicted schematically in Fig 4, are presented in the following.

Approximate descriptions result in estimation errors. Macroscopic and mesoscopic descriptions provide only approximate estimates of the statistical moments of microscopic processes. These approximation errors may result in erroneous parameter estimates. This happens, for instance, when the approximation error can be partially or completely compensated by changing the parameter values, as we have illustrated in Fig 5 for the trimerization process. For small volumes, we find pronounced differences between the mean of the stochastic process determined using SSA and the means predicted by the RRE, EMRE and 2MA (Fig 5(a)). We regarded the mean of the SSA runs as artificial data and optimized parameters of RRE, EMRE and 2MA using the aforementioned multi-start local optimization with accurate gradients. The optimized trajectories for RRE, EMRE and 2MA agree well with the mean of the SSA runs as shown in Fig 5(b). This agreement is achieved for parameter values deviating from the true parameter values used for the stochastic simulation. The objective function landscapes of the individual models shown in Fig 5(c) indicate that the optimum of the objective function does generally not coincide with the true parameters. This pattern is reproducible and is caused by the error of the approximation methods resulting in erroneous, biased parameter estimates.

Mesoscopic descriptions improve the estimation accuracy at intermediate volumes. As the estimation error is caused by the approximation error of the statistical moments on which the inference is based, a relation between the magnitude of the approximation error and the estimation error is to be expected. Since mesoscopic descriptions (EMRE,2MA) tend to have smaller approximation errors than macroscopic descriptions (RRE) [23, 29, 36], the former are expected to lead to smaller estimation errors as we have have demonstrated in Fig 5(c). We will now give a verification of these arguments. To assess the estimation error we generated 100 artificial datasets, each containing 10^5 cells for different volumes, and evaluated the estimation accuracy of the parameter estimation. The workflow is illustrated in Figure C in S1 Supporting Information. For the inference we used MA and the SSE truncated to various orders:

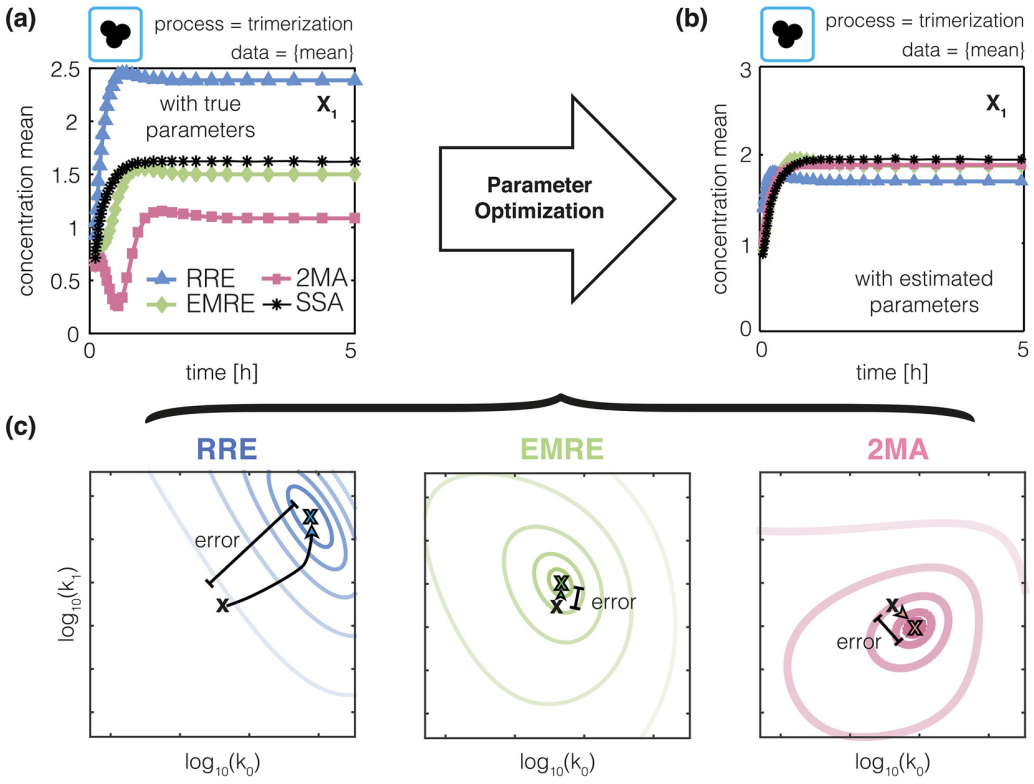


Fig 5. Approximation error introduces estimation error. (a) Mean monomer concentration in the trimerization process for $\Omega = 6\mu m^3$ computed from 10^5 SSA trajectory realizations (black line). Approximate mean monomer concentrations obtained using RRE, EMRE and 2MA (colored lines). (b) Mean monomer concentration for RRE, EMRE and 2MA obtained after parameter estimation using the SSA mean as artificial dataset. (c) True (black ×) and optimized parameter values (colored ×) for RRE, EMRE and 2MA. Contour lines of objective function are colored. The opacity increases with increasing likelihood values.

doi:10.1371/journal.pcbi.1005030.g005

- `data = {mean}` → inference using RRE, EMRE and 2MA.
- `data = {mean, variance}` → inference using LNA, IOS and 3MA.

Medians and 80% symmetric percentile intervals of the squared estimation error,

$$\text{error}^2 = \|\theta^{\text{true}} - \hat{\theta}\|_2^2,$$

were calculated and the results are shown in Fig 6 for both processes.

In accordance with our hypothesis, we found that mesoscopic descriptions using higher-order SSEs and MAs tend to yield a lower estimation error compared to macroscopic descriptions, here RRE and LNA. The difference between meso- and macroscopic descriptions is most pronounced for intermediate volumes ($10^1\ \mu m^3 - 10^3\ \mu m^3$). As expected, for large volumes—where micro-, meso- and macroscopic descriptions agree—all descriptions resulted in small estimation errors. For small volumes, meso- and macroscopic descriptions depart from the underlying process resulting in large estimation errors which render results meaningless. For the enzymatic degradation process, higher-order MAs and SSEs sometimes yield higher estimation errors than low-order MAs and SSEs. This might come surprising, but the approximation order is only informative about the approximation error in the large volume limit and

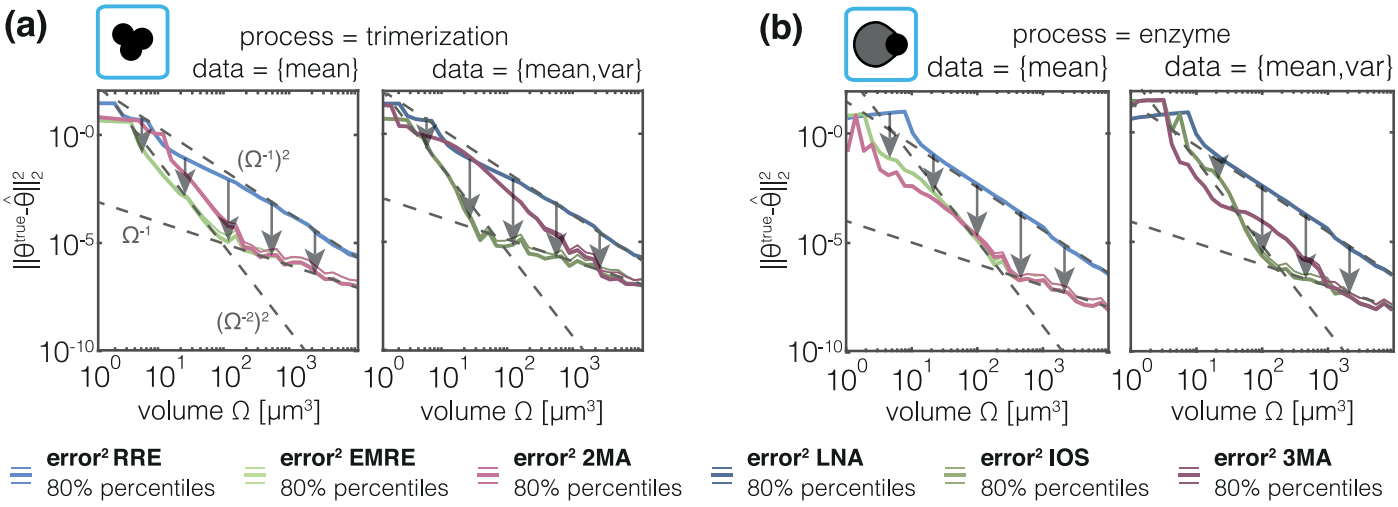


Fig 6. Quantification of volume dependence of estimation error. Medians (thick line) and symmetric 80% percentile based confidence intervals (thin lines) of the errors for two representative parameters of (a) the trimerization process and (b) enzymatic degradation process. Results for different meso- and macroscopic models are color-coded and panels show datasets computed from 10^5 single-cell measurements: (left) `data = {mean}` ; and (right) `data = {mean,variance}` . The estimated convergence order for the intermediate and high-volume regimes is indicated as gray dotted lines.

doi:10.1371/journal.pcbi.1005030.g006

does not allow conclusions for low volume regimes. Accordingly, the superiority of higher-order approximations cannot be expected in low volume regimes.

In the medium- to high-volume regime we would expect an approximation order of Ω^{-1} for RNA/LNA and Ω^{-2} for EMRE/IOS. Accordingly, in the absence of measurement noise, the convergence order of the mean squared error should be $(\Omega^{-1})^2$ and $(\Omega^{-2})^2$ respectively. In fact, the observed convergence rates agree with these theoretical rates which are indicated by dashed gray lines in Fig 6. In the medium- to high-volume regime, the convergence rates are dominated by the bias fraction of the mean squared error. However, for high volume regimes we observe a convergence rate of approximately Ω^{-1} for EMRE/IOS. In this regime, the convergence rate of the mean squared error is dominated by the variance of the parameter estimator. Thus the convergence rate can be expected to be proportional to the variances of sample means and variances $\sigma^2_{\hat{\mu}_{i,k}}$ and $\sigma^2_{\hat{\Sigma}_{ii,k}}$. For the considered setting the convergence rate seems to be dominated by $\sigma^2_{\hat{\mu}_{i,k}} = \frac{1}{N}\Sigma_{ii}$, which scales, according to the LNA, as $\frac{1}{N}\Omega^{-1}$. We expect that for higher volumes, the convergence rate of RRE/LNA will also be limited by the estimation variance and thus attune to $\frac{1}{N}\Omega^{-1}$. The decomposition of the mean squared error for the two models is provided in Figure E in S1 Supporting Information. Furthermore, this theoretical limit suggest that an increase in the number of measured cells N should result in a shift of this variance limit to lower values.

For the simulation examples, including variance information did not yield any consistent reduction of the estimation error. This might come as a surprise as a previous study suggested that the variance carries considerable amounts of information which when included can even render previously non-identifiable parameters identifiable [39]. However, for the simulation examples we considered a data-rich setting where all parameters are well identifiable and the estimation error is mainly due to the approximation error of the description. In less data-rich situations and in the presence of technical noise, we expect that including variance information could also reduce the estimation error.

SSE and MA methods achieved similar estimation accuracies for the trimerization and enzymatic degradation processes. However, optimization using SSE turned out to be

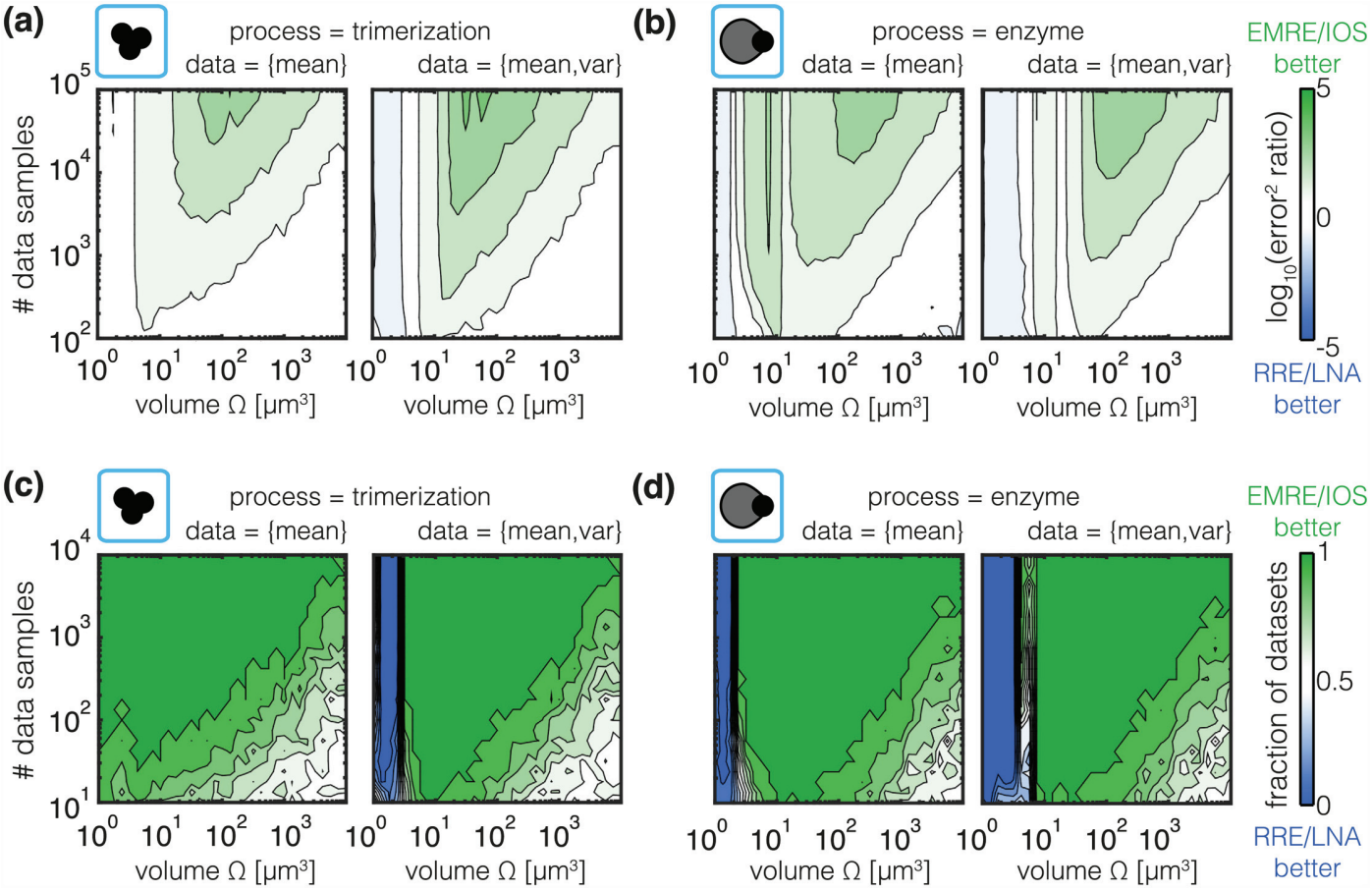


Fig 7. Quantification of sample size dependence of estimation error. (a,b) Ratio of the absolute estimation errors. Green indicates a lower estimation error for EMRE and IOS while blue indicates a lower estimation error for RRE and LNA. (c,d) Frequency for lower estimation error for EMRE and IOS compared to RRE and LNA. The color indicates the fraction of datasets for which EMRE and IOS yields a lower estimation error than RRE and LNA.

doi:10.1371/journal.pcbi.1005030.g007

computationally more efficient than MA, as robust numerical integration of the respective differential equations was less problematic, see Figure D in S1 Supporting Information. In the following we present the results for RRE, LNA, EMRE and IOS while the results for 2MA and 3MA are reported in Figure F-I in S1 Supporting Information.

Mesoscopic descriptions are beneficial for the analysis of high-throughput single-cell data. As we have seen in the previous section, the sample size influences the estimation of mean and variance, we studied its impact on the accuracy of inference with different models. We determined the estimation errors for RRE, LNA, EMRE and IOS using 100 artificial data sets of different sample sizes and volumes. This detailed analysis confirmed that RRE and LNA generally yield larger estimation errors than EMRE and IOS. Interestingly, the regime of volumes for which this is consistently observed increases with sample size as we show in Fig 7(a) and 7(b, green area). In Fig 7(c) and 7(d) we verify that this relation holds not only on average but also for individual datasets resulting in lower estimation errors. As expected, this is the case for intermediate to large volumes. Only for small volumes, the approximation was unsatisfactory and RRE and LNA were occasionally favored over EMRE and IOS methods.

Depending on the experimental devices, the number of single-cell recordings ranges from tens to hundreds of thousands measured cells. High-content single-cell methods, such as

single-cell RNAseq and single-cell time-lapse microscopy, are mostly used to study tens to hundreds of cells [82, 83]. High-throughput single-cell methods, flow and mass cytometry, enable the assessment of thousands of cells but provide merely a smaller number of features [84]. Intuitively, the high-throughput single-cell methods reduce the sampling error as many cells are recorded and can therefore be well characterized by moments. Hence higher-order SSEs are particularly valuable for the analysis of high-throughput single-cell data. This simple rule-of-thumb for the *a priori* selection of the modeling approach is also corroborated by our findings for MA Figure F,G in S1 Supporting Information.

Model selection pinpoints regimes in which mesoscopic descriptions yield improved approximation accuracy. Our results suggest that meso- and macroscopic descriptions are only appropriate for inference in certain volume and sample size regimes. In practice, the boundaries of these regimes remain unknown. To identify the most appropriate description in a certain regime *a posteriori*, we propose the use of model selection methods.

We employed AIC to select the most appropriate among a set of candidate models given by the macro- and mesoscopic descriptions of the processes. Fig 8 depicts the AIC weights—interpretable as posterior probabilities—of EMRE and IOS for different volumes and sample sizes. We find that EMRE and IOS are favored over RRE and LNA everywhere except in two regimes that provide additional insights:

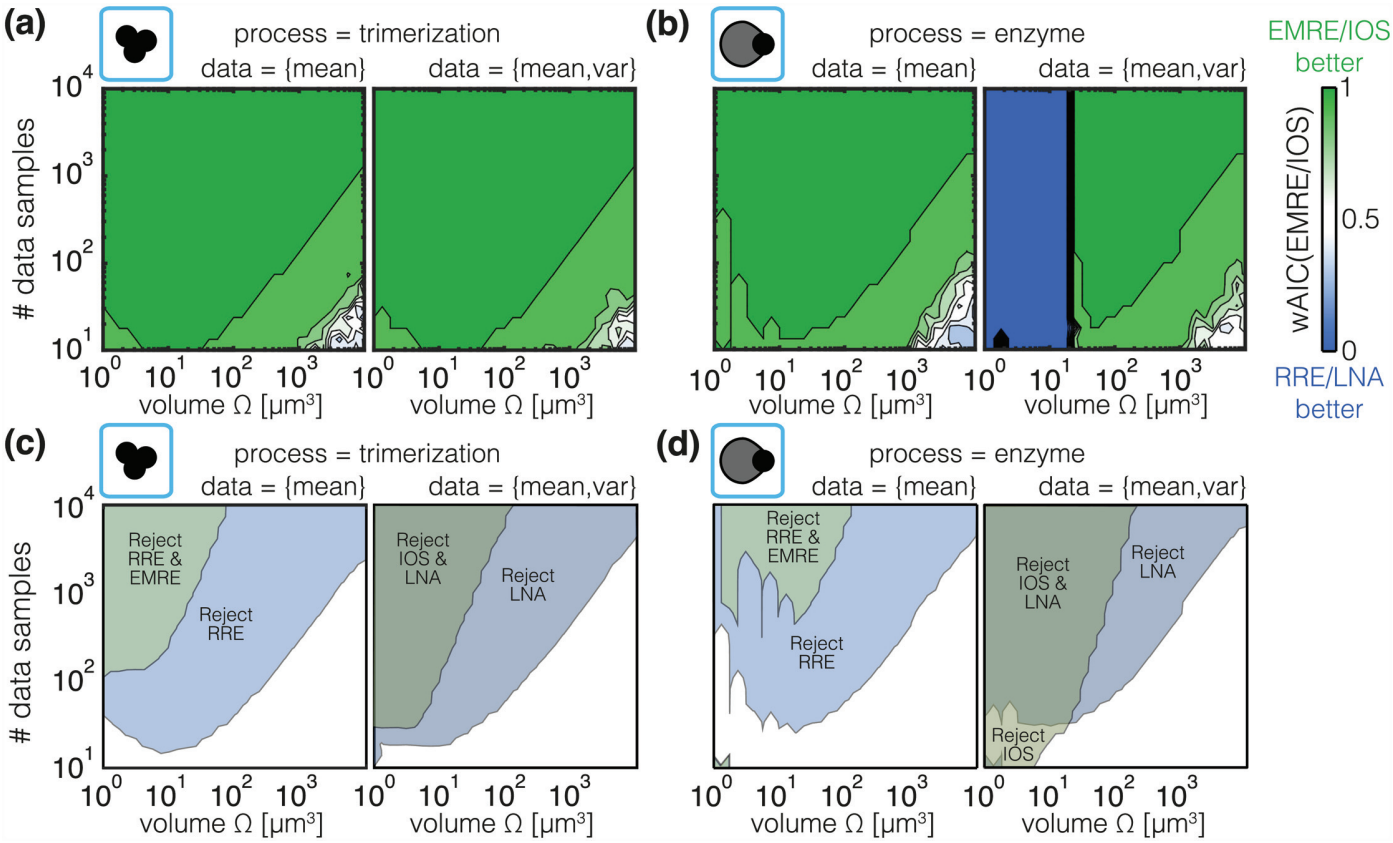


Fig 8. Analysis of model selection and rejection criteria. (a) and (b) Median AIC weight for EMRE and IOS at respective estimated parameters. A green color indicates that the EMRE and IOS description is more probable and a blue color indicates the RRE and LNA description is more probable. (c) and (d) area in which the models can on average be rejected based on a chi-square test to confidence level 0.01. The coloring indicates the method to which the area corresponds.

doi:10.1371/journal.pcbi.1005030.g008

- **Regime I** is classified by large volumes and low sample sizes. The AICs of RRE and EMRE as well as LNA and IOS are comparable—AIC weights close to 0.5—as the models fit the limited data fairly well. If the statistical power of the data is however increased by increasing the sample size, EMRE and IOS are favored as descriptions provided by RRE and LNA are no longer sufficiently accurate. This indicates that the statistical power is simply not sufficient to reveal the small differences in between EMRE/IOS and RRE/LNA.
- **Regime II** appears only for the inference of the enzymatic degradation model using data for mean and variance. For volumes below 10 μm^3, LNA is favored over IOS. The reason for this is that the LNA leads to a physically meaningful description, i.e. positive variances for all volumes, whereas the IOS leads to positive variances (which correct the LNA) for large enough volumes but can potentially give rise to negative variances for small enough volumes. The latter is possible since terms in the SSE beyond the LNA, i.e., those involving third- and higher-order derivatives, do not lead to a Fokker-Planck description which can imply negative values of the approximated probability density function [37, 85]). Hence the LNA becomes favorable over the IOS for small volumes.

Accordingly, model selection favors the macroscopic description over the mesoscopic one either when the statistical power is too limited to reject them (Regime I) or when they are indeed more accurate (Regime II). Otherwise, mesoscopic descriptions based on higher-order SSEs or MAs (Figure H,I in S1 Supporting Information) are selected.

The selection also resembles the results for the estimation error in Fig 7(a) and 7(b). The critical volume for which the AIC weights depart from unity coincides with the upper bound of the intermediate regime in which mesoscopic description provide lower estimation errors. Furthermore, in Regime II IOS yields large estimation errors. In summary, this suggests that model selection can be used (i) to decide whether a mesoscopic or a macroscopic description is appropriate and (ii) to improve the quality of parameter estimates.

Model rejection criteria can reveal the necessity of a microscopic description. The superiority of a model according to model selection criteria does not imply that the favored model accurately represents the data. Specific applications may indeed require microscopic descriptions to model experimental data. To check this, simulation and parameter estimation using microscopic descriptions could be performed. While efficient algorithms have been developed, such procedure is often time-consuming. We therefore considered model rejection to assess the necessity of microscopic modeling without performing the microscopic analysis.

We computed the goodness-of-fit and employed a χ^2-test with a confidence level of 0.01 for model rejection. Fig 8(c) and 8(d) illustrates the regimes in which the meso-/macroscopic descriptions have been rejected for at least 50% of the artificial datasets. We find that regions in which higher-order SSEs are rejected are mostly contained in regions for which lower-order SSEs are rejected.

As sample size increases higher-order SSEs and MAs are rejected for increasingly larger volumes. This is plausible as the improved statistical power allows us to resolve smaller differences between microscopic and the corresponding meso-/macroscopic descriptions. The statistical power is determined by the number of samples and the statistical moments of the samples. If the difference between approximative descriptions and the process is large, a small sample size is sufficient to rule out a model, while a large number of samples is required to detect smaller differences. As the difference between approximative descriptions and the process is volume-dependent and process-specific, the regions in which the approximative description can be rejected might possess a complex shape. For the enzymatic degradation model we find for instance that for low sample sizes the RRE can merely be rejected for an intermediate volume

regime but not for small or large volumes. The dependence on the number of samples supports also the finding that for the analysis of high-throughput data accurate models need to be employed.

The proposed approach based on model rejection reveals the need for a more accurate description without performing the corresponding analysis. Accordingly, macroscopic models such as RRE and LNA can be used to perform the initial analysis. Only if these models are rejected using the χ^2-test, mesoscopic descriptions need to be employed. In the same way also the necessity of microscopic descriptions can be assessed without actually performing the corresponding time-consuming analysis.

In summary, the study of trimerization and enzymatic degradation model clearly revealed that higher-order SSEs and MAs are generally more reliable. The increased computational complexity is tractable and the investment becomes worthwhile for high-throughput data in particular. Further improvement could be achieved by combining model selection criteria and rejection criteria.

Discussion

Many biological processes exhibit stochastic fluctuations which are relevant for cells and organisms [1–3]. Quantitative mechanistic models facilitate an understanding of the relevance of these fluctuations to dynamics over various length scales. Despite significant progress, the parameterization of such quantitative mechanistic models remains challenging. In this work, we implemented sophisticated parameter estimation and uncertainty analysis methods relying on mesoscopic descriptions of stochastic processes, namely higher-order SSEs and MAs.

We verified the developed methods using simulation examples. We found that for intermediate and large volumes, for which inference using microscopic descriptions is computationally already demanding, our approximate methods provides reliable estimation results. The computation time required for optimization was a fraction of the computation time required for the stochastic simulation of the stochastic process (c.f. Figure A in S1 Supporting Information). Compared to estimation methods using macroscopic descriptions, such as RRE or LNA [40], a significantly decreased estimation error is observed for intermediate volumes. This intermediate regime increases with the number of single-cell measurements. Our parameter estimation methods using higher-order SSEs and MAs are therefore especially suited for the data-driven modeling of high-throughout data, such as flow and mass cytometry data.

As the unnecessary study of meso-/microscopic descriptions can be time-consuming, we also considered model rejection approaches. We found that the application of such methods can guide model refinement. The methods cannot however distinguish between inappropriateness arising from meso- and microscopic descriptions due to an inaccurate knowledge of the biochemical reaction network as both result in a disagreement of model and data. In addition, for applications with multiple candidate models it is not guaranteed that model selection results obtained for macroscopic descriptions will be reproduced for the corresponding mesoscopic or microscopic descriptions. Thus model selection and model rejection methods should always be combined. If the microscopic description of all candidate models were rejected using the χ^2-test, the set of candidates would not contain a model which accurately represents the data and should be extended.

Beyond the study of artificial data, we employed the proposed methods to study experimental data for the JAK/STAT signaling pathway. This revealed that mesoscopic modeling can also provide additional insights if merely population-average data are available. For processes with non-linear reaction propensities, the mean encodes information about the volume and the molecular numbers, respectively [29, 36]. This enabled the estimation of a lower bound for the

initial STAT concentration, a parameter, which is structurally non-identifiable when macroscopic descriptions are employed. To assess the lower bound we implemented profile likelihood calculations and MCMC methods for higher-order SSEs and MAs. MCMC methods for MAs had already been proposed [10], the combination of Bayesian and frequentist methodology is however known to provide more robust results [55, 86]. The derived lower bound for the STAT concentration could be confirmed with literature data. The insight could be obtained for a well-studied system and pinpoints the great potential of mesoscopic descriptions for data-driven modeling.

However, the use of mesoscopic descriptions also has certain drawbacks. It is for instance not completely clear how practical and structural identifiability of the stochastic process (described by the CME) and the approximative descriptions are related [18]. Furthermore, as higher-order SSEs and MAs are merely approximations to data generating processes, the resulting estimators are inconsistent. Hence, parameter estimates and confidence intervals can be erroneous. In principle, this problem can be addressed using ideas developed in the fields of model reduction [87, 88] or probabilistic numerical simulations [89, 90]. These methods require upper bounds for the approximation error or the error distribution of vector field approximation, respectively. Approximations for both might be obtained by using a sequence of higher-order expansions. A rigorous treatment would yet require exact bounds, as available for the FSP [22].

In this study we employed higher-order SSEs and MAs to approximate the moments of the stochastic process for inference. A further improvement could be achieved by using hybrid approaches, such as the method of conditional moments [91] or the conditional system size expansion [92]. These approaches exploit a microscopic description of low-copy number species and a mesoscopic description for medium- to high-copy number species. Complementarily, higher-order SSEs and MAs could be used to enhance the accuracy of ODE constrained mixture modeling [93]. This modeling and analysis method accounts for the subpopulation structure but relies on simple macroscopic descriptions for the subpopulation dynamics. The use of macroscopic descriptions could result in a reduction of the number of parameters and an improved data exploitation.

Until now, the stochasticity of biological systems is often disregarded as its analysis is computationally demanding. The emergence of measurement techniques such as single-cell fluorescent microscopy [83, 94], flow and mass cytometry [84], single-cell qPCR [95] and single-cell RNA-seq [82] renders the consideration of stochastic effects a necessity [96, 97]. The presented methods are computationally efficient and scalable. This will facilitate the quantitative mechanistic modeling of complex cellular process es and the exploitation of cell-to-cell variability for biological discovery.

Supporting Information

S1 Supporting Information. Supplementary notes regarding modeling and computational analysis. This document provides a detailed description of the biochemical reaction networks and their parameters, system size expansion and moment approximation, as well as the parameter estimation and the uncertainty analysis results.
(PDF)

S1 Code. MATLAB code used for inference using SSE and MA. This zip-file contains the MATLAB code for the simulation and application example presented in the paper. We provide implementations of all models, parameter estimation and uncertainty analysis to allow everybody to reproduce the results.
(ZIP)

Author Contributions

Conceived and designed the experiments: FF PT FJT RG JH. Performed the experiments: FF. Analyzed the data: FF PT RG JH. Contributed reagents/materials/analysis tools: FF PT AK JH. Wrote the paper: FF PT RG JH.

References

1. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science; 2002; 297(5584):1183–1186. doi: 10.1126/science.1070919 PMID: 12183631
2. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB. Gene regulation at the single-cell level. Science; 2005; 307(5717):1962–1965. doi: 10.1126/science.1106914 PMID: 15790856
3. Raj A, van Oudenaarden A. Nature, nurture, or chance: Stochastic gene expression and its consequences. Cell; 2008; 135(2):216–226. doi: 10.1016/j.cell.2008.09.050 PMID: 18957198
4. Maheshri N, O'Shea EK. Living with noisy genes: how cells function reliably with inherent variability in gene expression. Annu Rev Biophys Biomol Struct; 2007; 36:413–434. doi: 10.1146/annurev.biophys.36.040306.132705 PMID: 17477840
5. Gillespie DT. A rigorous derivation of the chemical master equation. Physica A; 1992; 188(1):404–425. doi: 10.1016/0378-4371(92)90283-V
6. Wilkinson DJ. Parameter inference for stochastic kinetic models of bacterial gene regulation: A Bayesian approach to systems biology. In: Bernardo JM, Bayarri MJ, amd A P Dawid JOB, Heckerman D, Smith AFM, West M, editors. Proc. of 9th Valencia Int. Meet. (Bayesian Statistics 9), Valencia, Spain. Oxford University Press; 2010. p. 679–705.
7. Toni T, Stumpf MPH. Simulation-based model selection for dynamical systems in systems and population biology. Bioinf; 2010; 26(1):104–110. doi: 10.1093/bioinformatics/btp619
8. Fuchs C. Inference for diffusion processes with applications in life sciences. 1st ed. Berlin / Heidelberg: Springer; 2013.
9. Mikeev L, Wolf V. Parameter estimation for stochastic hybrid models of biochemical reaction networks. In: Proc. of the 15th ACM International Conference on Hybrid Systems: Computation and Control. New York, NY, USA: ACM; 2012. p. 155–166.
10. Milner P, Gillespie CS, Wilkinson DJ. Moment closure based parameter inference of stochastic kinetic models. Stat Comp; 2013; 23(2):287–295. doi: 10.1007/s11222-011-9310-8
11. Zechner C, Unger M, Pelet S, Peter M, Koeppl H. Scalable inference of heterogeneous reaction kinetics from pooled single-cell recordings. Nat Meth; 2014; 11:197–202. doi: 10.1038/nmeth.2794
12. Kazeev V, Khammash M, Nip M, Schwab C. Direct solution of the Chemical Master Equation using quantized tensor trains. PLoS Comput Biol; 2014; 10(3):e1003359. doi: 10.1371/journal.pcbi.1003359 PMID: 24626049
13. Loos C, Marr C, Theis FJ, Hasenauer J. Approximate Bayesian Computation for stochastic single-cell time-lapse data using multivariate test statistics. In: Roux O, Bourdon J, editors. Computational Methods in Systems Biology. vol. 9308 of Lecture Notes in Computer Science. Springer International Publishing; 2015. p. 52–63.
14. Munsky B, Khammash M. Identification from stochastic cell-to-cell variation: a genetic switch case study. IET Syst Biol; 2010; 4(6):356–366. doi: 10.1049/iet-syb.2010.0013 PMID: 21073235
15. Nüesch T. Finite state projection-based parameter estimation algorithms for stochastic chemical kinetics [Master Thesis]. Swiss Federal Institute of Technology, Zürich; 2010.
16. Hasenauer J, Radde N, Doszczak M, Scheurich P, Allgöwer F. Parameter estimation for the CME from noisy binned snapshot data: Formulation as maximum likelihood problem; 2011. Extended abstract at *Conf. of Stoch. Syst. Biol.*, Monte Verita, Switzerland.
17. Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J, et al. Moment-based inference predicts bimodality in transient gene expression. Proc Natl Acad Sci U S A; 2012; 109(21):8340–8345. doi: 10.1073/pnas.1200161109 PMID: 22566653
18. Kazeroonian A, Hasenauer J, Theis FJ. Parameter estimation for stochastic biochemical processes: A comparison of moment equation and finite state projection. In: Autio R, Shmulevich I, Strimmer K, Wiuf C, Sarbu S, Yli-Harja O, editors. Proceedings of 10th International Workshop on Computational Systems Biology. Tampere, Finland: Tampere International Center for Signal Processing; 2013. p. 66–73.
19. Lillacci G, Khammash M. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. Bioinf; 2013; 29(18):2311–2319. doi: 10.1093/bioinformatics/btt380

20. Neuert G, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. Systematic identification of signal-activated stochastic gene regulation. Science; 2013; 339(6119):584–587. doi: 10.1126/science.1231456 PMID: 23372015

21. Gillespie DT. Exact stochastic simulation of coupled chemical reactions. J Phys Chem; 1977; 81 (25):2340–2361. doi: 10.1021/j100540a008

22. Munsky B, Khammash M. The finite state projection algorithm for the solution of the chemical master equation. J Chem Phys; 2006; 124(4):044104. doi: 10.1063/1.2145882 PMID: 16460146

23. Grima R. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. J Chem Phys; 2012; 136(15):154105. doi: 10.1063/1.3702848 PMID: 22519313

24. van Kampen NG. Stochastic processes in physics and chemistry. 3rd ed. Amsterdam: North-Holland; 2007. doi: 10.1016/B978-044452965-7/50010-6

25. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in population genetics. Genetics; 2002; 162(4):2025–2035. PMID: 12524368

26. Robert CP, Cornuet JM, Marin JM, Pillai NS. Lack of confidence in Approximate Bayesian Computation model choice. Proc Natl Acad Sci U S A; 2011; 108(37):15112–15117. doi: 10.1073/pnas.1102900108 PMID: 21876135

27. Mateescu M, Wolf V, Didier F, Henzinger TA. Fast adaptive uniformisation of the chemical master equation. IET Syst Biol; 2010; 4(6):441–452. doi: 10.1049/iet-syb.2010.0005 PMID: 21073242

28. Sunkara V, Hegland M. An optimal Finite State Projection method. Procedia Computer Science; 2010; 1(1):1579–1586. doi: 10.1016/j.procs.2010.04.177

29. Engblom S. Computing the moments of high dimensional solutions of the master equation. Appl Math Comp; 2006; 180:498–515. doi: 10.1016/j.amc.2005.12.032

30. Lee CH, Kim KH, Kim P. A moment closure method for stochastic reaction networks. J Chem Phys; 2009; 130(13):134107. doi: 10.1063/1.3103264 PMID: 19355717

31. Gillespie CS. Moment-closure approximations for mass-action models. IET Syst Biol; 2009; 3(1):52–58. doi: 10.1049/iet-syb:20070031 PMID: 19154084

32. Ale A, Kirk P, Stumpf MPH. A general moment expansion method for stochastic kinetic models. J Chem Phys; 2013; 138(17):174101. doi: 10.1063/1.4802475 PMID: 23656108

33. Schnoerr D, Sanguinetti G, Grima R. Comparison of different moment-closure approximations for stochastic chemical kinetics. The Journal of Chemical Physics; 2015; 143(18). doi: 10.1063/1.4934990 PMID: 26567686

34. Whittle P. On the use of the normal approximation in the treatment of stochastic processes. J R Stat Soc B; 1957; 19(2):268–281.

35. Singh A, Hespanha JP. Lognormal moment closures for biochemical reactions. In: Proc. IEEE Conf. on Dec. and Contr. (CDC); 2006. p. 2063–2068.

36. Grima R. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. J Chem Phys; 2010; 133 (035101).

37. Grima R. Construction and accuracy of partial differential equation approximations to the chemical master equation. Physical Review E; 2011; 84(5 Pt 2):056109. doi: 10.1103/PhysRevE.84.056109

38. Schnoerr D, Sanguinetti G, Grima R. Validity conditions for moment closure approximations in stochastic chemical kinetics. J Chem Phys; 2014; 141(8):084103. doi: 10.1063/1.4892838 PMID: 25173001

39. Munsky B, Trinh B, Khammash M. Listening to the noise: random fluctuations reveal gene network parameters. Mol Syst Biol; 2009; 5(318). doi: 10.1038/msb.2009.75 PMID: 19888213

40. Komorowski M, Costa MJ, Rand DA, Stumpf MPH. Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. Proc Natl Acad Sci U S A; 2011; 108(21):8645–8650. doi: 10.1073/pnas.1015814108 PMID: 21551095

41. Chis OT, Banga JR, Balsa-Canto E. Structural identifiability of systems biology models: A critical comparison of methods. PLoS ONE; 2011; 6(11):e27755. doi: 10.1371/journal.pone.0027755 PMID: 22132135

42. Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. PLoS ONE; 2013; 8(9):e74335. doi: 10.1371/journal.pone.0074335 PMID: 24098642

43. Grima R. Linear-noise approximation and the chemical master equation agree up to second-order moments for a class of chemical systems. Phys Rev E; 2015; 92:042124. doi: 10.1103/PhysRevE.92.042124

44. Ramaswamy R, González-Segredo N, Sbalzarini I, Grima R. Discreteness-induced concentration inversion in mesoscopic chemical systems. Nat Comm; 2012; 3(779).

45. Thomas P, Matuschek H, Grima R. How reliable is the linear noise approximation of gene regulatory networks? BMC Genomics; 2013; 14(Suppl 4)(S5). doi: 10.1186/1471-2164-14-S4-S5 PMID: 24266939

46. Kazeroonian A, Fröhlich F, Raue A, Theis FJ, Hasenauer J. CERENA: ChEmical REaction Network Analyzer—A Toolbox for the Simulation and Analysis of Stochastic Chemical Kinetics. PLoS ONE; 2016; 11(1):e0146732. doi: 10.1371/journal.pone.0146732 PMID: 26807911

47. Elf J, Ehrenberg M. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. Genome Res; 2003; 13:2475–2484. doi: 10.1101/gr.1196503 PMID: 14597656

48. Ruess J, Lygeros J. Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. ACM Transactions on Modeling and Computer Simulation (TOMACS); 2015; 25(2):8. doi: 10.1145/2688906

49. Coleman TF, Li Y. On the convergence of reflective Newton Methods for large-scale nonlinear minimization subject to bounds. Math Prog; 1992;p. 1–36.

50. Coleman TF, Li Y. An interior trust region approach for nonlinear minimization subject to bounds. SIAM J Optim; 1996; 6:418–445. doi: 10.1137/0806023

51. Hindmarsh AC, Brown PN, Grant KE, Lee SL, Serban R, Shumaker DE, et al. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. ACM T Math Software; 2005; 31(3):363–396. doi: 10.1145/1089014.1089020

52. More JJ. The Levenberg-Marquardt algorithm: Implementation and theory. Lecture Notes in Mathematics; 1978; 630:105–116. doi: 10.1007/BFb0067700

53. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinf; 2009; 25(25):1923–1929. doi: 10.1093/bioinformatics/btp358

54. Murphy SA, van der Vaart AW. On profile likelihood. J Am Stat Assoc; 2000; 95(450):449–485. doi: 10.1080/01621459.2000.10474219

55. Raue A, Kreutz C, Theis FJ, Timmer J. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. Phil Trans Royal Soc A; 2013; 371 (1984).

56. Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. Briefings in Bioinf; 2007; 8(2):109–116. doi: 10.1093/bib/bbm007

57. Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J R Statist Soc B; 2011; 73(2):123–214. doi: 10.1111/j.1467-9868.2010.00765.x

58. Venzon D, Moolgavkar S. A Method for Computing Profile-Likelihood Based Confidence Intervals. Applied Statistics; 1988; 37(1):87–94. doi: 10.2307/2347496

59. Fröhlich F, Theis FJ, Hasenauer J. Uncertainty analysis for non-identifiable dynamical systems: Profile likelihoods, bootstrapping and more. In: Mendes P, Dada JO, Smallbone KO, editors. Proceedings of the 12th International Conference on Computational Methods in Systems Biology (CMSB 2014), Manchester, UK. Lecture Notes in Bioinformatics. Springer International Publishing Switzerland; 2014. p. 61–72.

60. Haario H, Laine M, Mira A, Saksman E. DRAM: Efficient adaptive MCMC. Stat Comp; 2006; 16(4):339–354. doi: 10.1007/s11222-006-9438-0

61. Akaike H. On the likelihood of a time series model. The Statistician; 1978; 27(3/4):217–235. doi: 10.2307/2988185

62. Schwarz G. Estimating the dimension of a model. Ann Statist; 1978; 6(2):461–464. doi: 10.1214/aos/1176344136

63. Huber-Carol C, Balakrishnan N, Nikulin M, Mesbah M. Goodness-of-fit tests and model validity. Statistics for Industry and Technology. Birkhäuser Basel; 2002.

64. Jaqaman K, Danuser G. Linking data to models: Data regression. Nature Reviews Molecular Cell Biology; 2006; 7:813–819. doi: 10.1038/nrm2030 PMID: 17006434

65. Papoulis A. Probability, Random Variables, and Stochastic Processes. Tata McGraw-Hill; 2002.

66. Koch KR. Parameter estimation and hypothesis testing in linear models. Springer-Verlag Berlin Heidelberg; 1999. doi: 10.1007/978-3-662-03976-2

67. Swameye I, Müller TG, Timmer J, Sandra O, Klingmüller U. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. Proc Natl Acad Sci U S A; 2003; 100 (3):1028–1033. doi: 10.1073/pnas.0237333100 PMID: 12552139

68. Rawlings JS, Rosler KM, Harrison DA. The JAK/STAT signaling pathway. J Cell Sci; 2004; 117(Pt 8):1281–1283. doi: 10.1242/jcs.00963 PMID: 15020666

69. Spivak JL, Gascón P, Ludwig H. Anemia management in oncology and hematology. Oncologist; 2009; 14 (Suppl 1):43–56. doi: 10.1634/theoncologist.2009-S1-43 PMID: 19762516

70. Hedley BD, Chu JE, Ormond DG, Beausoleil MS, Boasie A, Allan AL, et al. Recombinant human erythropoietin in combination with chemotherapy increases breast cancer metastasis in preclinical mouse models. Clin Cancer Res; 2011; 17(19):6151–6162. doi: 10.1158/1078-0432.CCR-10-3298 PMID: 21856770

71. Bachmann J, Raue A, Schilling M, Böhm ME, Kreutz C, Kaschek D, et al. Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. Mol Syst Biol; 2011; 516(7).

72. Liao S, Vejchodský T, Erban R. Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. Journal of The Royal Society Interface; 2015; 12(108). doi: 10.1098/rsif.2015.0233

73. Kügler P. Moment fitting for parameter inference in repeatedly and partially observed stochastic biological models. PloS one; 2012; 7(8):e43001. doi: 10.1371/journal.pone.0043001 PMID: 22900079

74. Rabindran S, Haroun R, Clos J, Wisniewski J, Wu C. Regulation of heat shock factor trimer formation: role of a conserved leucine zipper. Science; 1993; 259(5092):230–234. doi: 10.1126/science.8421783 PMID: 8421783

75. Kim E, Cho KO, Rothschild A, Sheng M. Heteromultimerization and NMDA Receptor-Clustering Activity of Chapsyn-110, a Member of the PSD-95 Family of Proteins. Neuron; 1996; 17(1):103–113. doi: 10.1016/S0896-6273(00)80284-6 PMID: 8755482

76. Golding I, Paulsson J, Zawilski SM, Cox EC. Real-time kinetics of gene activity in individual bacteria. Cell; 2005; 123(6):1025–1036. doi: 10.1016/j.cell.2005.09.031 PMID: 16360033

77. Thomas P, Matuschek H, Grima R. Computation of biochemical pathway fluctuations beyond the linear noise approximation using iNA. IEEE International Conference on Bioinformatics and Biomedicine; 2012;p. 192–196.

78. Peccoud J, Ycart B. Markovian modelling of gene product synthesis. Theor Popul Biol; 1995; 48 (2):222–234. doi: 10.1006/tpbi.1995.1027

79. Friedman N, Cai L, Xie XS. Linking stochastic dynamics to population distribution: an analytical framework of gene expression. Phys Rev Lett; 2006; 97(16):168302. doi: 10.1103/PhysRevLett.97.168302 PMID: 17155441

80. Shahrezaei V, Swain PS. Analytical distributions for stochastic gene expression. Proc Natl Acad Sci U S A; 2008; 105(45):17256–17261. doi: 10.1073/pnas.0803850105 PMID: 18988743

81. De Vargas Roditi L, Claassen M. Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. Curr Opin Biotechnol; 2015; 34:9–15. doi: 10.1016/j.copbio.2014.10.010 PMID: 25461506

82. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Gen; 2009; 10(1):57–63. doi: 10.1038/nrg2484

83. Rimon N, Schuldiner M. Getting the whole picture: combining throughput with content in microscopy. J Cell Sci; 2011; 124(22):3743–3751. doi: 10.1242/jcs.087486 PMID: 22124141

84. Bodenmiller B, Zunder ER, Finck R, Chen TJ, Savig ES, Bruggner RV, et al. Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. Nat Biotechnol; 2012; 30(9):858–867. doi: 10.1038/nbt.2317 PMID: 22902532

85. Thomas P, Grima R. Approximate distributions of the Master equation. Physical Review E; 2015; 92:012120.

86. Hug S, Raue A, Hasenauer J, Bachmann J, Klingmüller U, Timmer J, et al. High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. Math Biosci; 2013; 246 (2):293–304. doi: 10.1016/j.mbs.2013.04.002 PMID: 23602931

87. Hasenauer J, Löhning M, Khammash M, Allgöwer F. Dynamical optimization using reduced order models: A method to guarantee performance. Journal of Process Control; 2012; 22(8):1490–1501. doi: 10.1016/j.jprocont.2012.01.017

88. Dihlmann M, Haasdonk B. Certified nonlinear parameter optimization with reduced basis surrogate models. In: Cvetković L, Atanacković T, Kostić V, editors. Proceedings of Applied Mathematics and Mechanics (PAMM). vol. 13; 2013. p. 3–6.

89. Chkrebtii OA. Probabilistic solution of differential equations for Bayesian uncertainty quantification and inference. Simon Fraser University, Canada; 2013.

90. Chkrebtii OA, Campbell DA, Girolami MA, Calderhead B. Bayesian uncertainty quantification for differential equations. arXiv:13062365 [statME]; 2014;.

91. Hasenauer J, Wolf V, Kazeroonian A, Theis FJ. Method of conditional moments (MCM) for the chemical master equation. Journal of Mathematical Biology; 2014; 69(3):687–735.

92. Thomas P, Popovic N, Grima R. Phenotypic switching in gene regulatory networks. Proc Natl Acad Sci U S A; 2014; 111(19):6994–6999. doi: 10.1073/pnas.1400049111 PMID: 24782538

93. Hasenauer J, Hasenauer C, Hucho T, Theis FJ. ODE constrained mixture modelling: A method for unraveling subpopulation structures and dynamics. PLoS Comput Biol; 2014; 10(7):e1003686. doi: 10.1371/journal.pcbi.1003686 PMID: 24992156

94. Schroeder T. Long-term single-cell imaging of mammalian stem cells. Nat Methods; 2011; 8(4):30–35. doi: 10.1038/nmeth.1577

95. Taniguchi K, Kajiyama T, Kambara H. Quantitative analysis of gene expression in a single cell by qPCR. Nature Methods; 2009; 6(7):503–506. doi: 10.1038/nmeth.1338 PMID: 19525960

96. Pelkmans L. Using cell-to-cell variability–A new era in molecular biology. Science; 2012; 336 (6080):425–426.

97. Munsky B, Neuert G, von Oudenaarden A. Using gene expression noise to understand gene regulation. Science; 2012; 336(6078):183–187. doi: 10.1126/science.1216379 PMID: 22499939

Appendix B

Scalable parameter estimation for genome-scale biochemical reaction networks.

This is the copyedited PDF of an article accepted for publication in PLoS Computational Biology: F. Fröhlich, B. Kaltenbacher, F.J. Theis, J. Hasenauer. **Scalable parameter estimation for genome-scale biochemical reaction networks.** PLoS Computational Biology 13(1):e1005331 (2017).

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks

Fabian Fröhlich[1,2], Barbara Kaltenbacher[3], Fabian J. Theis[1,2], Jan Hasenauer[1,2]*

1 Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany, **2** Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Garching, Germany, **3** University of Klagenfurt, Institute of Mathematics, Klagenfurt, Austria

* jan.hasenauer@helmholtz-muenchen.de

Citation: Fröhlich F, Kaltenbacher B, Theis FJ, Hasenauer J (2017) Scalable Parameter Estimation for Genome-Scale Biochemical Reaction Networks. PLoS Comput Biol 13(1): e1005331. doi:10.1371/journal.pcbi.1005331

Editor: Jorg Stelling, ETH Zurich, SWITZERLAND

Received: May 4, 2016

Accepted: December 20, 2016

Published: January 23, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the German Research Foundation (DFG; http://www.dfg.de) through the Graduate School of Quantitative Biosciences Munich (QBM; FF), the German Federal Ministry of Education and Research (BMBF; www.bmbf.de) within the SYS-Stomach project (Grant No. 01ZX1310B; JH and FJT) and the Postdoctoral Fellowship Program (https://www.helmholtz-muenchen.de/en/fellows/index.html) of the Helmholtz Zentrum München (JH). The funders

Abstract

Mechanistic mathematical modeling of biochemical reaction networks using ordinary differential equation (ODE) models has improved our understanding of small- and medium-scale biological processes. While the same should in principle hold for large- and genome-scale processes, the computational methods for the analysis of ODE models which describe hundreds or thousands of biochemical species and reactions are missing so far. While individual simulations are feasible, the inference of the model parameters from experimental data is computationally too intensive. In this manuscript, we evaluate adjoint sensitivity analysis for parameter estimation in large scale biochemical reaction networks. We present the approach for time-discrete measurement and compare it to state-of-the-art methods used in systems and computational biology. Our comparison reveals a significantly improved computational efficiency and a superior scalability of adjoint sensitivity analysis. The computational complexity is effectively independent of the number of parameters, enabling the analysis of large- and genome-scale models. Our study of a comprehensive kinetic model of ErbB signaling shows that parameter estimation using adjoint sensitivity analysis requires a fraction of the computation time of established methods. The proposed method will facilitate mechanistic modeling of genome-scale cellular processes, as required in the age of omics.

Author Summary

In this manuscript, we introduce a scalable method for parameter estimation for genome-scale biochemical reaction networks. Mechanistic models for genome-scale biochemical reaction networks describe the behavior of thousands of chemical species using thousands of parameters. Standard methods for parameter estimation are usually computationally intractable at these scales. Adjoint sensitivity based approaches have been suggested to have superior scalability but any rigorous evaluation is lacking. We implement a toolbox for adjoint sensitivity analysis for biochemical reaction network which also supports the import of SBML models. We show by means of a set of benchmark models that adjoint sensitivity based approaches unequivocally outperform standard approaches for large-scale models and that the achieved speedup increases with respect to both the

had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

number of parameters and the number of chemical species in the model. This demonstrates the applicability of adjoint sensitivity based approaches to parameter estimation for genome-scale mechanistic model. The MATLAB toolbox implementing the developed methods is available from http://ICB-DCM.github.io/AMICI/.

Introduction

In the life sciences, the abundance of experimental data is rapidly increasing due to the advent of novel measurement devices. Genome and transcriptome sequencing, proteomics and metabolomics provide large datasets [1] at a steadily decreasing cost. While these genome-scale datasets allow for a variety of novel insights [2, 3], a mechanistic understanding on the genome scale is limited by the scalability of currently available computational methods.

For small- and medium-scale biochemical reaction networks mechanistic modeling contributed greatly to the comprehension of biological systems [4]. Ordinary differential equation (ODE) models are nowadays widely used and a variety of software tools are available for model development, simulation and statistical inference [5–7]. Despite great advances during the last decade, mechanistic modeling of biological systems using ODEs is still limited to processes with a few dozens biochemical species and a few hundred parameters. For larger models rigorous parameter inference is intractable. Hence, new algorithms are required for massive and complex genomic datasets and the corresponding genome-scale models.

Mechanistic modeling of a genome-scale biochemical reaction network requires the formulation of a mathematical model and the inference of its parameters, e.g. reaction rates, from experimental data. The construction of genome-scale models is mostly based on prior knowledge collected in databases such as KEGG [8], REACTOME [9] and STRING [10]. Based on these databases a series of semi-automatic methods have been developed for the assembly of the reaction graph [11–13] and the derivation of rate laws [14, 15]. As model construction is challenging and as the information available in databases is limited, in general, a collection of candidate models can be constructed to compensate flaws in individual models [16]. For all these model candidates the parameters have to be estimated from experimental data, a challenging and usually ill-posed problem [17].

To determine maximum likelihood (ML) and maximum a posteriori (MAP) estimates for model parameters, high-dimensional nonlinear and non-convex optimization problems have to be solved. The non-convexity of the optimization problem poses challenges, such as local minima, which have to be addressed by the selection of optimization methods. Commonly used global optimization methods are multi-start local optimization [18], evolutionary and genetic algorithms [19], particle swarm optimizers [20], simulated annealing [21] and hybrid optimizers [22, 23] (see [18, 24–26] for a comprehensive survey). For ODE models with a few hundred parameters and state variables multi-start local optimization methods [18] and related hybrid methods [27] have proven to be successful. These optimization methods use the gradient of the objective function to establish fast local convergence. While the convergence of gradient based optimizers can be significantly improved by providing exact gradients (see e.g. [18, 28, 29]), the gradient calculation is often the computationally most demanding step.

The gradient of the objective function is usually approximated by finite differences. As this method is neither numerically robust nor computationally efficient, several parameter estimation toolboxes employ forward sensitivity analysis. This decreases the numerical error and computation time [18]. However, the dimension of the forward sensitivity equations increases linearly with both the number of state variables and parameters, rendering its application for

genome-scale models problematic. In other research fields such as mathematics and engineering, adjoint sensitivity analysis is used for parameter estimation in ordinary and partial differential equation models. Adjoint sensitivity analysis is known to be superior to the forward sensitivity analysis when the number of parameters is large [30]. Adjoint sensitivity analysis has been used for inference of biochemical reaction networks [31–33]. However, the methods were never picked up by the systems and computational biology community, supposedly due to the theoretical complexity of adjoint methods, a missing evaluation on a set of benchmark models, and an absence of an easy-to-use toolbox.

In this manuscript, we provide an intuitive description of adjoint sensitivity analysis for parameter estimation in genome-scale biochemical reaction networks. We describe the end value problem for the adjoint state in the case of discrete-time measurement and provide an user-friendly implementation to compute it numerically. The method is evaluated on seven medium- to large-scale models. By using adjoint sensitivity analysis, the computation time for calculating the objective function gradient becomes effectively independent of the number of parameters with respect to which the gradient is evaluated. Furthermore, for large-scale models adjoint sensitivity analysis can be multiple orders of magnitude faster than other gradient calculation methods used in systems biology. The reduction of the time for gradient evaluation is reflected in the computation time of the optimization. This renders parameter estimation for large-scale models feasible on standard computers, as we illustrate for a comprehensive kinetic model of ErbB signaling.

Methods

In this section we introduce the model class and the corresponding estimation problem. Subsequently, gradient calculation using finite differences, forward sensitivity analysis and adjoint sensitivity analysis is described and the theoretical complexity as well as some aspects of the numerical implementation are discussed.

Mathematical model and experimental data

We consider ODE models for biochemical reaction networks,

$$\dot{x} = f(x, \theta), \quad x(t_0) = x_0(\theta), \tag{1}$$

in which $x(t, \theta) \in \mathbb{R}^{n_x}$ is the concentration vector at time t and $\theta \in \mathbb{R}^{n_\theta}$ denotes the parameter vector. Parameters are usually kinetic constants, such as binding affinities as well as synthesis, degradation and dimerization rates. The vector field $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n_x}$ describes the temporal evolution of the concentration of the biochemical species. The mapping $x_0 : \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n_x}$ provides the parameter dependent initial condition at time t_0.

As available experimental techniques usually do not provide measurements of the concentration of all biochemical species, we consider the output map $h : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \mapsto \mathbb{R}^{n_y}$. This map models the measurement process, i.e. the dependence of the output (or observables) $y(t, \theta) \in \mathbb{R}^{n_y}$ at time point t on the state variables and the parameters,

$$y(t, \theta) = h(x(t, \theta), \theta). \tag{2}$$

The i-th observable y_i can be the concentration of a particular biochemical species (e.g. $y_i = x_l$) as well as a function of several concentrations and parameters (e.g. $y_i = \theta_m(x_{l_1} + x_{l_2})$).

We consider discrete-time, noise corrupted measurements

$$\bar{y}_{ij} = y_i(t_j, \theta) + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2), \tag{3}$$

yielding the experimental data $\mathcal{D} = \{((\bar{y}_{ij})_{i=1}^{n_y}, t_j)\}_{j=1}^{N}$. The number of time points at which measurements have been collected is denoted by N.

Remark: For simplicity of notation we assume throughout the manuscript that the noise variances, σ_{ij}^2, are known and that there are no missing values. However, the methods we will present in the following as well as the respective implementations also work when this is not the case. For details we refer to the S1 Supporting Information.

Maximum likelihood (ML) estimation

We estimate the unknown parameter θ from the experimental data $\mathcal{D}$ using ML estimation. Parameters are estimated by minimizing the negative log-likelihood, an objective function indicating the difference between experiment and simulation. In the case of independent, normally distributed measurement noise with known variances the objective function is given by

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n_y}\sum_{j=1}^{N}\left(\frac{\bar{y}_{ij} - y_i(t_j,\theta)}{\sigma_{ij}}\right)^2, \tag{4}$$

where $y_i(t_j, \theta)$ is the value of the output computed from Eqs (1) and (2) for parameter value θ. The minimization,

$$\theta^* = \arg\min_{\theta\in\Theta} J(\theta), \tag{5}$$

of this weighted least squares J yields the ML estimate of the parameters.

The optimization problem Eq (5) is in general nonlinear and non-convex. Thus, the objective function can possess multiple local minima and global optimization strategies need to be used. For ODE models multi-start local optimization has been shown to perform well [18]. In multi-start local optimization, independent local optimization runs are initialized at randomly sampled initial points in parameter space. The individual local optimizations are run until the stopping criteria are met and the results are collected. The collected results are visualized by sorting them according to the final objective function value. This visualization reveals local optima and the size of their basin of attraction. For details we refer to the survey by Raue *et al.* [18]. In this study, initial points are generated using latin hypercube sampling and local optimization is performed using the interior point and the trust-region-reflective algorithm implemented in the MATLAB function `fmincon.m`. Gradients are computed using finite differences, forward sensitivity analysis or adjoint sensitivity analysis.

Finite differences

A näive approximation to the gradient of the objective function with respect to θ_k is obtained by finite differences,

$$\frac{\partial J}{\partial \theta_k} \approx \frac{J(\theta + a\,e_k) - J(\theta - b\,e_k)}{a+b}, \tag{6}$$

with $a, b \geq 0$ and the kth unit vector e_k. In practice forward differences ($a = \epsilon$, $b = 0$), backward differences ($a = 0$, $b = \epsilon$) and central differences ($a = \epsilon$, $b = \epsilon$) are widely used. For the computation of forward finite differences, this yields a procedure with three steps:

Step 1 The state trajectory $x(t, \theta)$ and output trajectory $y(t, \theta)$ are computed.

Step 2 The state trajectories $x(t, \theta^{(k)})$ and the output trajectories $y(t, \theta^{(k)})$ are computed for the perturbed parameters $\theta^{(k)} = \theta + \epsilon e_k$ for $k = 1, \ldots, n_\theta$.

Step 3 The objective function gradient elements $\frac{\partial J}{\partial \theta_k}$, are computed from the output trajectory $y(t, \theta)$ and the perturbed output trajectory $y(t, \theta^{(k)})$ for $k = 1, \ldots, n_\theta$.

In theory, forward and backward differences provide approximations of order ϵ while central differences provide more accurate approximations of order ϵ^2, provided that J is sufficiently smooth. In practice the optimal choice of a and b depends on the accuracy of the numerical integration [18]. If the integration accuracy is high, an accurate approximation of the gradient can be achieved using $a, b \ll 1$. For lower integration accuracies, larger values of a and b usually yield better approximations. A good choice of a and b is typically not clear *a priori* (cf. [34] and the references therein).

The computational complexity of evaluating gradients using finite differences is affine linear in the number of parameters. Forward and backward differences require in total $n_\theta + 1$ function evaluations. Central differences require in total $2n_\theta$ function evaluations. As already a single simulation of a large-scale model is time-consuming, the gradient calculation using finite differences can be limiting.

Forward sensitivity analysis

State-of-the-art systems biology toolboxes, such as the MATLAB toolbox Data2Dynamics [7], use forward sensitivity analysis for gradient evaluation. The gradient of the objective function is

$$\frac{\partial J}{\partial \theta_k} = \sum_{i=1}^{n_y} \sum_{j=1}^{N} \left(\frac{\bar{y}_{ij} - y_i(t_j, \theta)}{\sigma_{ij}^2} \right) s_{i,k}^y(t_j), \tag{7}$$

with $s_{i,k}^y(t) : [t_0, t_N] \mapsto \mathbb{R}$ denoting the sensitivity of output y_i at time point t with respect to parameter θ_k. Governing equations for the sensitivities are obtained by differentiating Eqs (1) and (2) with respect to θ_k and reordering the derivatives. This yields

$$\begin{aligned} \dot{s}_k^x &= \frac{\partial f}{\partial x} s_k^x + \frac{\partial f}{\partial \theta_k}, \qquad s_k^x(t_0) = \frac{\partial x_0}{\partial \theta_k} \\ s_{i,k}^y &= \frac{\partial h_i}{\partial x} s_k^x + \frac{\partial h_i}{\partial \theta_k} \end{aligned} \tag{8}$$

with $s_k^x(t) : [t_0, t_N] \mapsto \mathbb{R}^{n_x}$ denoting the sensitivity of the state x with respect to θ_k. Note that here and in the following, the dependencies of f, h, x_0 and their (partial) derivatives on t, x and θ are not stated explicitly but have the to be assumed. For a more detailed presentation we refer to the S1 Supporting Information *Section 1*.

Forward sensitivity analysis consists of three steps:

Step 1 The state trajectory $x(t, \theta)$ and output trajectory $y(t, \theta)$ are computed.

Step 2 The state sensitivities $s_k^x(t)$ and the output sensitivities $s_k^y(t)$ are computed using the state trajectory $x(t, \theta)$ for $k = 1, \ldots, n_\theta$.

Step 3 The objective function gradient elements $\frac{\partial J}{\partial \theta_k}$, are computed from the output sensitivities $s_k^y(t)$ and the output trajectory $y(t, \theta)$ for $k = 1, \ldots, n_\theta$.

Step 1 and 2 are often combined, which enables simultaneous error control and the reuse of the Jacobian [30]. The simultaneous error control allows for the calculation of accurate and reliable gradients. The reuse of the Jacobian improves the computational efficiency.

The number of state and output sensitivities increases linearly with the number of parameters. While this is unproblematic for small- and medium-sized models, solving forward

sensitivity equations for systems with several thousand state variable bears technical challenges. Code compilation can take multiple hours and require more memory than what is available on standard machines. Furthermore, while forward sensitivity analysis is usually faster than finite differences, in practice the complexity still increases roughly linearly with the number of parameters.

Adjoint sensitivity analysis

In the numerics community, adjoint sensitivity analysis is frequently used to compute the gradients of a functional with respect to the parameters if the function depends on the solution of a differential equation [35]. In contrast to forward sensitivity analysis, adjoint sensitivity analysis does not rely on the state sensitivities $s_k^x(t)$ but on the adjoint state $p(t)$.

The calculation of the objective function gradient using adjoint sensitivity analysis consists of three steps:

Step 1 The state trajectory $x(t, \theta)$ and output trajectory $y(t, \theta)$ are computed.

Step 2 The trajectory of the adjoint state $p(t)$ is computed.

Step 3 The objective function gradient elements $\frac{\partial J}{\partial \theta_k}$, $k = 1, \ldots, n_\theta$, are computed from the state trajectory $x(t, \theta)$, the adjoint state trajectory $p(t)$ and the output trajectory $y(t, \theta)$.

Step 1 and 2, which are usually the computationally intensive steps, are independent of the parameter dimension. The complexity of Step 3 increases linearly with the number of parameters, yet the computation time required for this step is typically negligible.

The calculation of state and output trajectories (Step 1) is standard and does not require special methods. The non-trivial element in adjoint sensitivity analysis is the calculation of the adjoint state $p(t) \in \mathbb{R}^{n_x}$ (Step 2). For discrete-time measurements—the usual case in systems and computational biology—the adjoint state is piece-wise continuous in time and defined by a sequence of backward differential equations. For $t > t_N$, the adjoint state is zero, $p(t) = 0$. Starting from this end value the trajectory of the adjoint state is calculated backwards in time, from the last measurement $t = t_N$ to the initial time $t = t_0$. At the time points at which measurements have been collected, $t_N, \ldots, t_1$, the adjoint state is reinitialised as

$$p(t_j) = \lim_{t \to t_j^+} p(t) + \sum_{i=1}^{n_y} \left(\frac{\partial h_i}{\partial x} \right)^T \frac{\bar{y}_{ij} - y_i(t_j)}{\sigma_{ij}^2}, \tag{9}$$

which usually results in a discontinuity of $p(t)$ at t_j. Starting from the end value $p(t_j)$ as defined in Eq (9) the adjoint state evolves backwards in time until the next measurement point t_{j-1} or the initial time t_0 is reached. This evolution is governed by the time-dependent linear ODE

$$\dot{p} = -\left(\frac{\partial f}{\partial x} \right)^T p. \tag{10}$$

The repeated evaluation of Eqs (9) and (10) until $t = t_0$ yields the trajectory of the adjoint state. Given this trajectory, the gradient of the objective function with respect to the individual parameters is

$$\frac{\partial J}{\partial \theta_k} = -\int_{t_0}^{t_N} p^T \frac{\partial f}{\partial \theta_k} dt - \sum_{i,j} \frac{\partial h_i}{\partial \theta_k} \frac{\bar{y}_{ij} - y_i(t_j)}{\sigma_{ij}^2} - p(t_0)^T \frac{\partial x_0}{\partial \theta_k}. \tag{11}$$

Accordingly, the availability of the adjoint state simplifies the calculation of the objective

function to n_θ one-dimensional integration problems over short time intervals whose union is the total time interval $[t_0, t_N]$.

Algorithm 1: Gradient evaluation using adjoint sensitivity analysis

```
% State and output
Step 1 Compute state and output trajectories using Eqs (1) and (2).

% Adjoint state
Step 2.1 Set end value for adjoint state, ∀t > t_N: p(t) = 0.
for j = N to 1 do
    Step 2.2 Compute end value for adjoint state according to the jth measure-
             ment using Eq (9).
    Step 2.3 Compute trajectory of adjoint state on time interval t = (t_{j-1}, t_j]
             by solving Eq (10).
end

% Objective function gradient
for k = 1 to n_θ do
    Step 3 Evaluation of the sensitivity ∂J/∂θ_k using Eq (11).
end
```

Pseudo-code for the calculation of the adjoint state and the objective function gradient is provided in Algorithm 1. We note that in order to use standard ODE solvers the end value problem Eq (10) can be transformed in an initial value problem by applying the time transformation $\tau = t_N - t$. The derivation of the adjoint sensitivities for discrete-time measurements is provided in the S1 Supporting Information *Section 1*.

The key difference of the adjoint compared to the forward sensitivity analysis is that the derivatives of the state and the output trajectory with respect to the parameters are not explicitly calculated. Instead, the sensitivity of the objective function is directly computed. This results in practice in a computation time of the gradient which is almost independent of the number of parameters. A visual summary of the different sensitivity analysis methods is provided in Fig 1. Besides the procedures also the computational complexity is indicated.

Implementation

The implementation of adjoint sensitivity analysis is non-trivial and error-prone. To render this method available to the systems and computational biology community, we implemented the Advanced Matlab Interface for CVODES and IDAS (AMICI). This toolbox allows for a simple symbolic definition of ODE models (1) and (2) as well as the automatic generation of native C code for efficient numerical simulation. The compiled binaries can be executed from MATLAB for the numerical evaluation of the model and the objective function gradient. Internally, the SUNDIALS solvers suite is employed [30], which offers a broad spectrum of state-of-the-art numerical integration of differential equations. In addition to the standard functionality of SUNDIALS, our implementation allows for parameter and state dependent discontinuities. The toolbox and a detailed documentation can be downloaded from http://ICB-DCM.github.io/AMICI/.

Results

In the following, we will illustrate the properties of adjoint sensitivity analysis for biochemical reaction networks. For this purpose, we study several models provided in the BioPreDyn benchmark suite [27] and from the curated branch of the Biomodels Database [37]. We compare adjoint sensitivity analysis with forward sensitivity analysis and finite differences

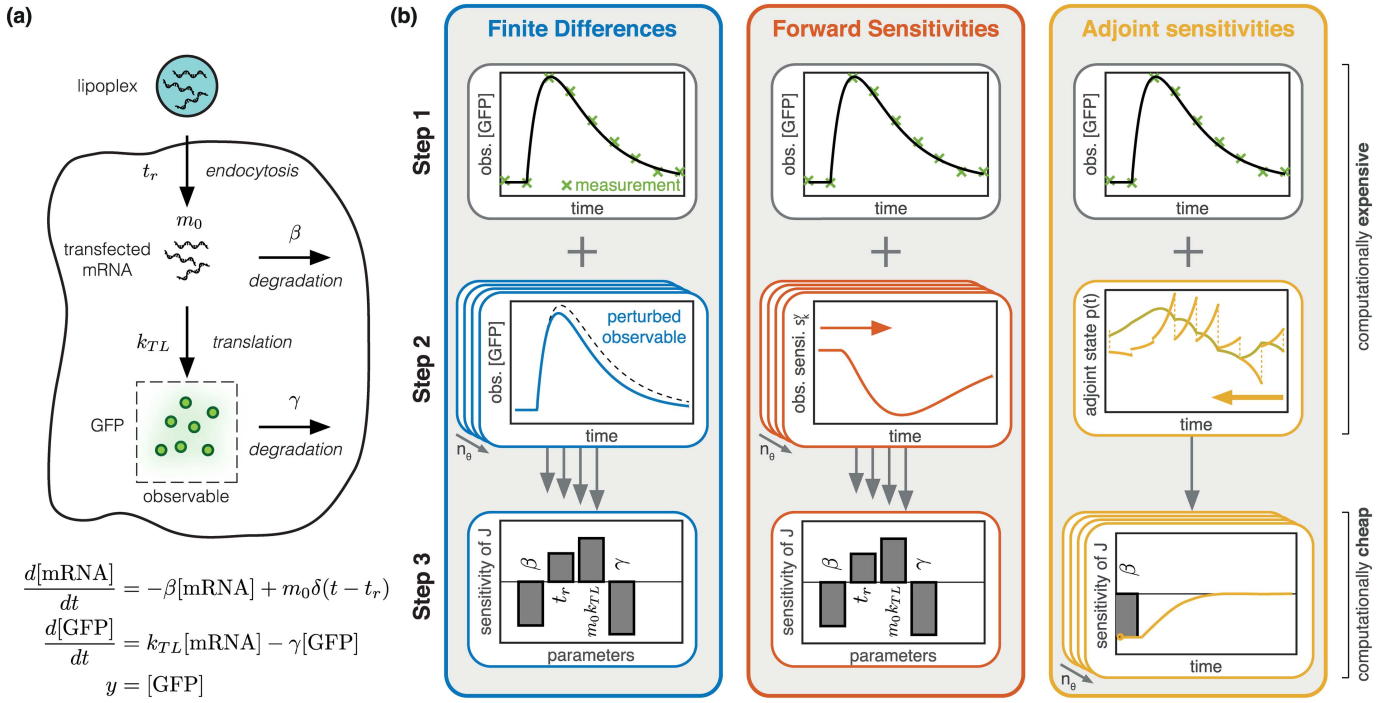


Fig 1. Illustration of gradient calculation using finite differences, forward sensitivity analysis and adjoint sensitivity equations for a model of mRNA transfection. (a) Sketch and mathematical formulation of the mathematical model of mRNA transfection presented by [36]. The intracellular release of mRNA at time point t_r is modeled using the Dirac delta distribution δ. (b) Illustration of finite differences, forward sensitivity analysis and adjoint sensitivity analysis for the model of mRNA transfection: (top) Step 1: simulation of model; (middle) Step 2: intermediate step for gradient calculation; and (bottom) Step 3: calculation of gradient from intermediate results. For all methods, Step 1 and 2 involve numerical simulation (the direction indicated by the arrow) and are computationally demanding, while Step 3 is computationally negligible.

doi:10.1371/journal.pcbi.1005331.g001

regarding accuracy, computational efficiency and scalability for a set of medium- to large-scale models.

Investigated models

For the comparison of different gradient calculation methods, we consider a set of standard models from the Biomodels Database [37] and the BioPreDyn benchmark suite [27]. From the biomodels database we considered models for the regulation of insulin signaling by oxidative stress (BM1) [38], the sea urchin endomesoderm network (BM2) [39], and the ErbB sigaling pathway (BM3) [40]. From BioPreDyn benchmark suite we considered models for central carbon metabolism in *E. coli* (B2) [41], enzymatic and transcriptional regulation of carbon metabolism in *E. coli* (B3) [42], metabolism of CHO cells (B4) [43], and signaling downstream of EGF and TNF (B5) [44]. Genome-wide kinetic metabolic models of *S. cerevisiae* and *E.coli* (B1) [45] contained in the BioPreDyn benchmark suite and the Biomodels Database [15, 45] were disregarded due to previously reported numerical problems [27, 45]. The considered models possess 18-500 state variable and 86-1801 parameters. A comprehensive summary regarding the investigated models is provided in Table 1.

To obtain realistic simulation times for adjoint sensitivities realistic experimental data is necessary (see S1 Supporting Information Section 3). For the BioPreDyn models we used the data provided in the suite, for the ErbB signaling pathway we used the experimental data

Table 1. List of investigated models and their properties.

ID	Parameters	State Variables	Model Type	Time Points	Data Points	Ref
B2	116	18	Metabolic	51	110	[41]
B3	178	47	Metabolic/Gene Reg.	161	7567	[42]
B4	117	34	Metabolic	12	156	[43]
B5	86	26	Signaling	16	960	[44]
BM1	383	104	Signaling	10	120	[38]
BM2	1801	431	Gene Reg.	21	3780	[39]
BM3	219	500	Signaling	21	105	[40]

doi:10.1371/journal.pcbi.1005331.t001

provided in the original publication and for the remaining models we generated synthetic data using the nominal parameter provided in the SBML definition.

In the following, we will compare the performance of forward and adjoint sensitivities for these models. As the model of ErbB signaling has the largest number of state variables and is of high practical interest in the context of cancer research, we will analyze the scalability of finite differences and forward and adjoint sensitivity analysis for this model in greater detail. Moreover, we will compare the computational efficiency of forward and adjoint sensitivity analysis for parameter estimation for the model of ErbB signaling.

Scalability of gradient evaluation using adjoint sensitivity analysis

The evaluation of the objective function gradient is the computationally demanding step in deterministic local optimization. For this reason, we compared the computation time for finite differences, forward sensitivity analysis and adjoint sensitivity analysis and studied the scalability of these approaches at the nominal parameter θ_0 which was provided in the SBML definitions of the investigated models.

For the comprehensive model of ErbB signaling we found that the computation times for finite differences and forward sensitivity analysis behave similarly (Fig 2a). As predicted by the theory, for both methods the computation time increased linearly with the number of parameters. Still, forward sensitivities are computationally more efficient than finite differences, as reported in previous studies [18].

Adjoint sensitivity analysis requires the solution to the adjoint problem, independent of the number of parameters. For the considered model, solving the adjoint problem a single time takes roughly 2-3-times longer than solving the forward problem. Accordingly, adjoint sensitivity analysis with respect to a small number of parameter is disadvantageous. However, adjoint sensitivity analysis scales better than forward sensitivity analysis and finite differences. Indeed, the computation time for adjoint sensitivity analysis is almost independent of the number of parameters. While computing the sensitivity with respect to a single parameter takes on average 10.09 seconds, computing the sensitivity with respect to all 219 parameters takes merely 14.32 seconds. We observe an average increase of $1.9 \cdot 10^{-2}$ seconds per additional parameter for adjoint sensitivity analysis which is significantly lower than the expected 3.24 seconds for forward sensitivity analysis and 4.72 seconds for finite differences. If the sensitivities with respect to more than 4 parameters are required, adjoint sensitivity analysis outperforms both forward sensitivity analysis and finite differences. For 219 parameters, adjoint sensitivity analysis is 48-times faster than forward sensitivities and 72-times faster than finite differences.

To ensure that the observed speedup is not unique to the model of ErbB signaling (BM3) we also evaluated the speedup of adjoint sensitivity analysis over forward sensitivity analysis

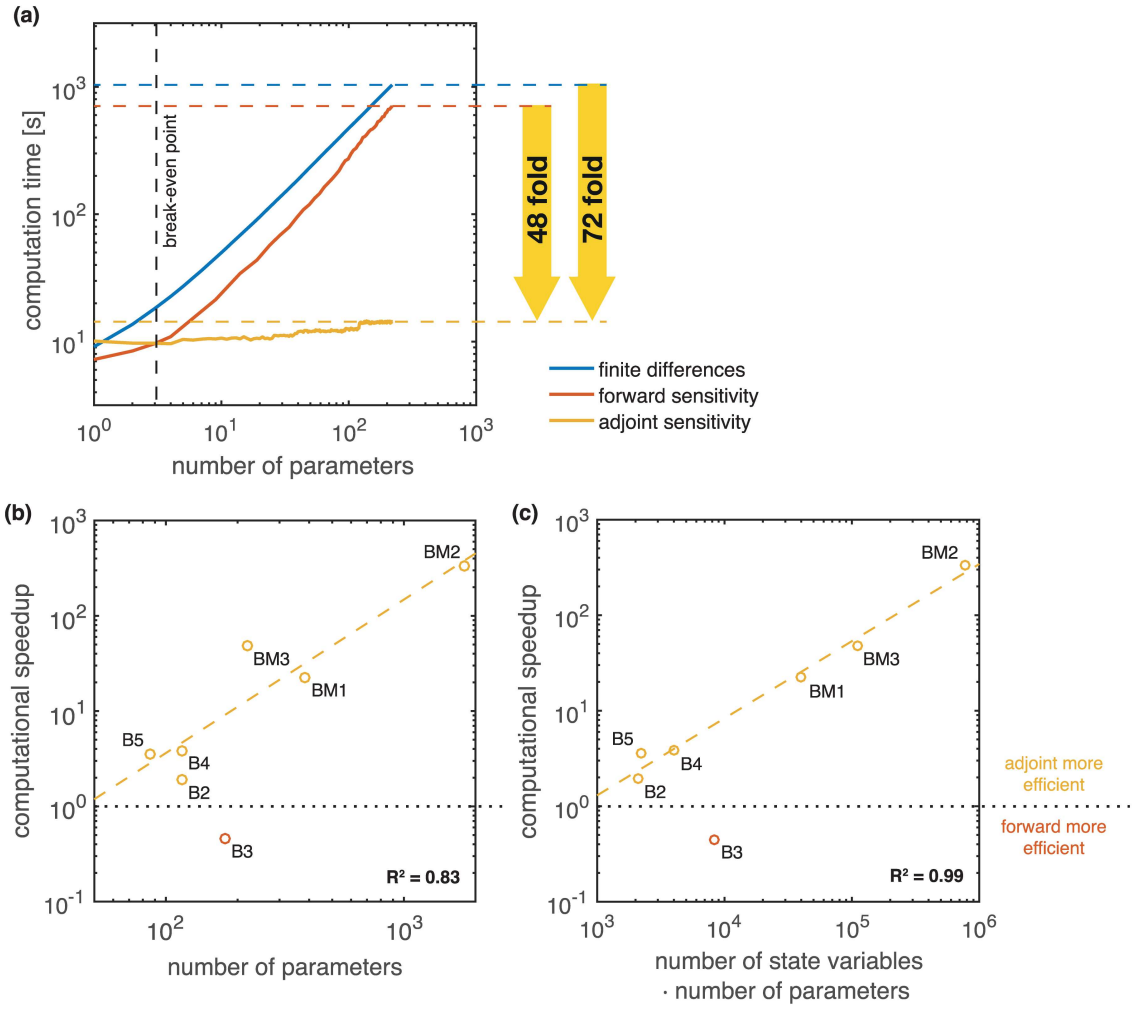


Fig 2. Comparison of gradient computation times for finite differences and forward and adjoint sensitivity analysis. (a) Scaling of computation time with respect to the number of parameters for the model of ErbB signaling (BM3). Computation time for finite differences and forward sensitivity equations increases roughly linearly. Computation time for adjoint sensitivity analysis is almost independent of the number of parameters but possesses a higher initial cost. Adjoint sensitivity analysis is 48 times faster than forward sensitivity analysis when considering all parameters. (b,c) Speedup when using adjoint sensitivity analysis over forward sensitivity analysis for gradient computation evaluated for all investigated models compared against n_θ and $n_x \cdot n_\theta$. Regression curves (dashed lines) have been fitted to the results of all models excluding B3, which seems to be an outlier. All computations were performed on a MacBook Pro with an 2.9 GHz Intel Core i7 processor.

doi:10.1371/journal.pcbi.1005331.g002

on models B2-5 and BM1-2. The results are presented in Fig 2b and 2c. We find that for all models, but model B3, gradient calculation using adjoint sensitivity is computationally more efficient than gradient calculation using forward sensitivities (speedup > 1). For model B3 the backwards integration required a much higher number of integration steps ($4 \cdot 10^6$) than the forward integration ($6 \cdot 10^3$), which results to a poor performance of the adjoint method. One reason for this poor performance could be that, in contrast to other models, the right hand side of the differential equation of model B3 consists almost exclusively of non-linear, non-mass-action terms.

Excluding model B3 we find an polynomial increase in the speedup with respect to the number of parameters n_θ (Fig 2b), as predicted by theory. Moreover, we find that the product

$n_\theta \cdot n_x$, which corresponds to the size of the system of forward sensitivity equations, is an even better predictor ($R^2 = 0.99$) than n_θ alone ($R^2 = 0.83$). This suggest that adjoint sensitivity analysis is not only beneficial for systems with a large number of parameters, but can also be beneficial for systems with a large number of state variables. As we are not aware of any similar observations in the mathematics or engineering community, this could be due to the structure of biological reaction networks.

Our results suggest that adjoint sensitivity analysis is an excellent candidate for parameter estimation in large-scale models as it provides good scaling with respect to both, the number of parameters and the number of state variables.

Accuracy and robustness of gradients computing adjoint sensitivity analysis

Efficient local optimization requires accurate and robust gradient evaluation [18]. To assess the accuracy of the gradient computed using adjoint sensitivity analysis, we compared this gradient to the gradients computed via finite differences and forward sensitivity analysis. Fig 3 visualizes the results for the model of ErbB signaling (BM3) at the nominal parameter θ_0 which was provided in the SBML definition. The results are similar for other starting points.

The comparison of the gradients obtained using finite differences and adjoint sensitivity analysis revealed small discrepancies (Fig 3a). The median relative difference (as defined in S1 Supporting Information Section 2) between finite differences and adjoint sensitivity analysis is $1.5 \cdot 10^{-3}$. For parameters θ_k to which the objective function J was relatively insensitive, $\partial J/\partial \theta_k < 10^{-2}$, there are much higher discrepancies, up to a relative error of $2.9 \cdot 10^{3}$.

Forward and adjoint sensitivity analysis yielded almost identical gradient elements over several orders of magnitude (Fig 3b). This was expected as both forward and adjoint sensitivity analysis exploit error-controlled numerical integration for the sensitivities. To assess numerical robustness of adjoint sensitivity analysis, we also compared the results obtained for high and low integration accuracies (Fig 3c). For both comparisons we found the similar median relative and maximum relative error, namely $2.6 \cdot 10^{-6}$ and $9.3 \cdot 10^{-4}$. This underlines the

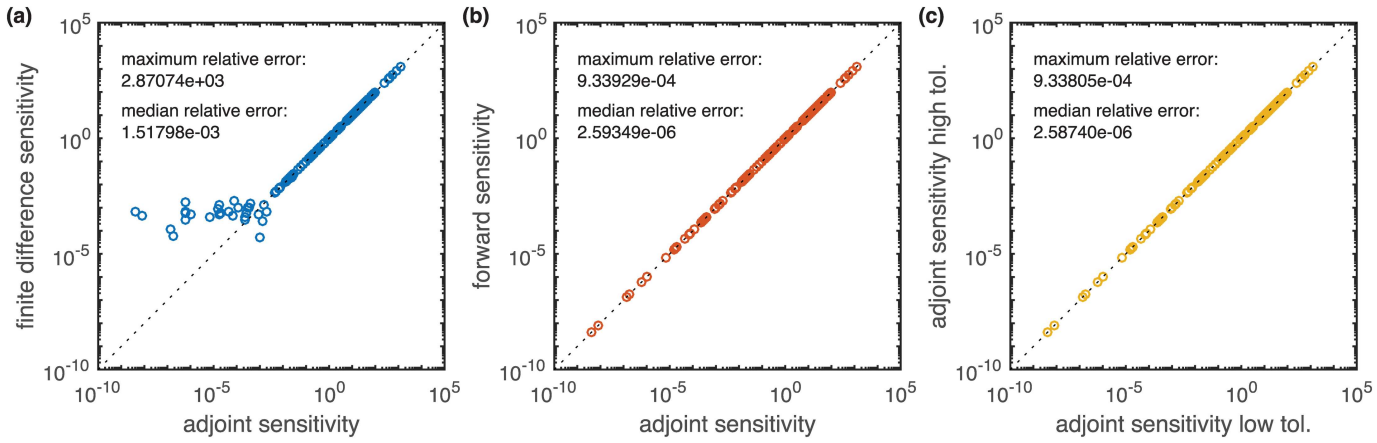


Fig 3. Comparison of the gradients computed using adjoint sensitivity equations with gradients computed using finite differences and forward sensitivity equations with default accuracies (absolute error < 10^{-16}, relative error < 10^{-8}). Each point represents the absolute value of one gradient element. Points on the diagonal indicate a good agreement. (a) Forward finite differences with $\epsilon = 10^{-3}$ vs. adjoint sensitivities. (b) Forward sensitivities vs. adjoint sensitivities. (c) Adjoint sensitivities with high accuracies (absolute error < 10^{-32}, relative error < 10^{-16}) and default accuracies (absolute error < 10^{-16}, relative error < 10^{-8}).

doi:10.1371/journal.pcbi.1005331.g003

robustness of the sensitivitity based methods and ensures that differences observed in Fig 3a indeed originate from the inaccuracy of finite differences.

Our results demonstrate that adjoint sensitivity analysis provides objective function gradients which are as accurate and robust as those obtained using forward sensitivity analysis.

Optimization of large-scale models using adjoint sensitivity analysis

As adjoint sensitivity analysis provides accurate gradients for a significantly reduced computational cost, this can boost the performance of a variety of optimization methods. Yet, in contrast to forward sensitivity analysis, adjoint sensitivities do not yield sensitivities of observables and it is thus not possible to approximate the Hessian of the objective function via the Fisher Information Matrix [46]. This prohibits the use of possibly more efficient Newton-type algorithms which exploit second order information. Therefore, adjoint sensitivities are limited to quasi-Newton type optimization algorithms, e.g. the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [47, 48], for which the Hessian is iteratively approximated from the gradient during optimization. In principle, the exact calculation of the Hessian and Hessian-Vector products is possible via second order forward and adjoint sensitivity analysis [49, 50], which possess similar scaling properties as the first order methods. However, both forward and adjoint approaches come at an additional cost and are thus not considered in this study.

To assess whether the use of adjoint sensitivities for optimization is still viable, we compared the performance of the interior point algorithm using adjoint sensitivity analysis with the BFGS approximation of the Hessian to the performance of the trust-region reflective algorithm using forward sensitivity analysis with Fisher Information Matrix as approximation of the Hessian. For both algorithms we used the MATLAB implementation in `fmincon.m`. The employed setup of the trust-region algorithm is equivalent to the use of `lsqnonlin.m` which is the default optimization algorithm in the MATLAB toolbox Data2Dynamics [7], which was employed to win several DREAM challenges. For the considered model the computation time of forward sensitivities is comparable in Data2Dynamics and AMICI. Therefore, we expect that Data2Dynamics would perform similar to the trust-region reflective algorithm coupled to forward sensitivity analysis.

We evaluated the performance for the model of ErbB signaling based on 100 multi-starts which were initialized at the same initial points for both optimization methods. For 41 out of 100 initial points the gradient could not be evaluated due numerical problems. These optimization runs are omitted in all further analysis. To limit the expected computation to a bearable amount we allowed a maximum of 10 iterations for the forward sensitivity approach and 500 iterations for the adjoint sensitivity approach. As the previously observed speedup in gradient computation was roughly 48 fold, we expected this setup should yield similar computation times for both approaches.

We found that for the considered number of iterations, both approaches perform similar in terms of objective function value compared across iterations (Fig 4a and 4b). However, the computational cost of one iteration was much cheaper for the optimizer using adjoint sensitivity analysis. Accordingly, given a fixed computation time the interior-point method using adjoint sensitivities outperforms the trust-region method employing forward sensitivities and the FIM (Fig 4c and 4d). In the allowed computation time, the interior point algorithm using adjoint sensitivities could reduce the objective function by up to two orders of magnitude (Fig 4c). This was possible although many model parameters seem to be non-identifiable (see S1 Supporting Information Section 4), which can cause problems.

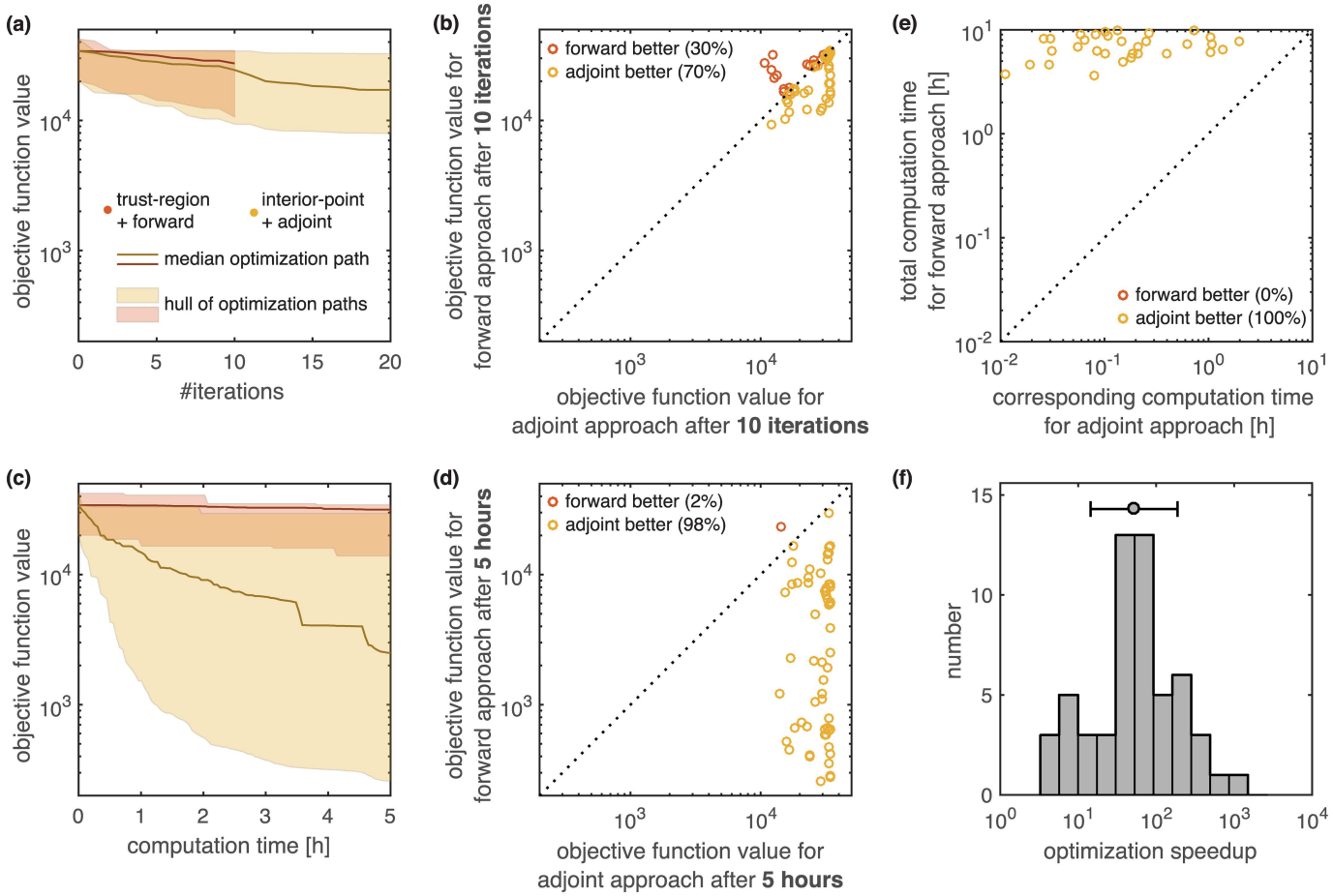


Fig 4. Comparison of optimization speed using forward and adjoint sensitivities for the model of ErbB signaling. For local optimization using forward sensitivity analysis (trust-region method) and local optimization using adjoint sensitivity analysis (interior-point method) we quantified the computation time across 100 local optimization runs with different initial conditions. For 41 out of 100 initial points the gradient could not be evaluated due to numerical problems. These optimization runs are omitted in all further analysis. (a,c) Comparison of objective function value with respect to iteration number and computation time. The hulls and medians computed for both methods are depicted as shaded areas and solid lines. (b,d) Pairwise comparison of objective function value after 10 iterations and 5 hours for both methods. Each dot corresponds to one initial point for the optimization. The coloring indicates which method performed better. (e) Pairwise comparison of the time required to reach the final objective function value achieved in the forward approach. For the adjoint approach the equivalent time is the minimal time to reach the same objective function value. Each dot corresponds to one initial point for the optimization. (f) Histogram of speedup by using adjoint sensitivity analysis over forward sensitivity analysis for individual initial points, computed from (e). All computations were performed on a linux cluster. Runs with same initial conditions were carried out on the same computation node.

doi:10.1371/journal.pcbi.1005331.g004

To quantify the speedup of the optimization using adjoint sensitivity analysis over the optimization using forward sensitivity analysis, we performed a pairwise comparison of the minimal time required by the adjoint sensitivity approach to reach the final objective function value of the forward sensitivity approach for the individual points (Fig 4e). The median speedup achieved across all multi-starts was 54 (Fig 4f), which was similar to the 48 fold speedup achieved in the gradient computation. The availability of the Fisher Information Matrix for forward sensitivities did not compensate for the significantly reduced computation time achieved using adjoint sensitivity analysis. This could be due to the fact that adjoint sensitivity based approach, being able to carry out many iterations in a short timeframe, can build a reasonable approximation of the Hessian approximation relatively fast.

In summary, this application demonstrates the applicability of adjoint sensitivity analysis for parameter estimation in large-scale biochemical reaction networks. Possessing similar accuracy as forward sensitivities, the scalability is improved which results in an increased optimizer efficiency. For the model of ErbB signaling, optimization using adjoint sensitivity analysis outperformed optimization using forward sensitivity analysis.

Discussion

Mechanistic mathematical modeling at the genome scale is an important step towards a holistic understanding of biological processes. To enable modeling at this scale, scalable computational methods are required which are applicable to networks with thousands of compounds. In this manuscript, we present a gradient computation method which meets this requirement and which renders parameter estimation for large-scale models significantly more efficient. Adjoint sensitivity analysis, which is extensively used in other research fields, is a powerful tool for estimating parameters of large-scale ODE models of biochemical reaction networks.

Our study of several benchmark models with up to 500 state variables and up to 1801 parameters demonstrated that adjoint sensitivity analysis provides accurate gradients in a computation time which is much lower than for established methods and effectively independent of the number of parameters. To achieve this, the adjoint state is computed using a piecewise continuous backward differential equation. This backward differential equation has the same dimension as the original model, yet the computation time required to solve it usually is slightly larger. As a result, finite differences and forward sensitivity analysis might be more efficient if the sensitivities with respect to a few parameters are required. The same holds for alternatives like complex-step derivative approximation techniques [51] and forward-mode automatic differentiation [28, 52]. For systems with many parameters, adjoint sensitivity analysis is advantageous. A scalable alternative might be reverse-mode automatic differentiation [28, 53], which remains to be evaluated for the considered class of problems.

For the model of ErbB signaling we could show that adjoint sensitivity based optimization outperforms forward sensitivity based optimization, which is the standard in most systems biology toolboxes. With the availability of the MATLAB toolbox AMICI the adjoint sensitivity based approach becomes accessible for other researchers. AMICI allows for the fully automated generation of executables for adjoint or forward sensitivity analysis from symbolic model definitions. This way, the toolbox is easy-to-use and can easily be integrated with existing toolboxes. Also other MATLAB toolboxes for computational modeling, e.g. AMIGO [6], Data2Dynamics [7], MEIGO [54] and SBtoolbox2 [55] could be extended to exploit adjoint sensitivity analysis. In addition to adjoint sensitivity analysis, these MATLAB toolboxes could exploit forward sensitivity analysis available via AMICI, as AMICI yields computation times comparable to those of tailored numerical methods such as odeSD [56] (S1 Supporting Information *Section 5*) or Data2Dynamics [7]. Moreover AMICI comes with detailed documentation and is already now used by several research labs.

Our study of the model of ErbB signaling suggests that for the available data, a large number of parameters remains non-identifiable. While novel technologies provide rich dataset, we expect that non-identifiability will remain a problem. In particular if merely relative measurements are available, as the case for many measurement techniques, additional unknown scaling factors need to be introduced. These scaling factors are, in combination with initial conditions and total abundances, often the source of practical and structural non-identifiabilites [18]. Fortunately, for a broad range of biological questions, these information are not necessary and also state-of-the-art methods optimization seem to work reasonably well in the

presence of non-identifiabilities. For the considered model of EreB signaling, we were able to achieve a significant decrease in the objective function value, despite the non-identifiability of parameters. This demonstrates that gradient based optimization is still feasible for large-scale problems. Yet, we believe that convergence of the optimizer could be improved by regularizing the objective function by integrating prior knowledge, possibly in a Bayesian framework [57], from databases such as SABIO-RK [58] or BRENDA [59].

Beyond the use in optimization, gradients computed using adjoint sensitivity analysis will also facilitate the development of more efficient uncertainty analysis methods. Riemann manifold Langevin and Hamiltonian Monte Carlo methods [60, 61] exploit the first and second order local structure of the posterior distribution and profit from more efficient gradient evaluation. The same holds for novel emulator-based sampling procedures [62] and approaches for posterior approximation [63]. By exploiting the proposed approach, rigorous Bayesian parameter estimation for models with hundreds of parameters could become a standard tool instead of an exception [64, 65].

In conclusion, adjoint sensitivity analysis will facilitate the development of large- and genome-scale mechanistic models for cellular processes as well as other (multi-scale) biological processes [66]. This will complement available statistical analysis methods for omics data [67] by providing mechanistic insights and render a holistic understanding feasible.

Supporting Information

S1 Supporting Information. Supplementary notes regarding sensitivity analysis and additional numerical examples. This document provides a detailed derivation of forward and adjoint sensitivity analysis and one additional numerical example for the comparision to the MATLAB toolbox odeSD.
(PDF)

S1 Code. MATLAB code. This zip-file contains the MATLAB code for the simulation and application examples presented in the paper. We provide implementations of all models, parameter estimation to allow everybody to reproduce the results.
(ZIP)

Author Contributions

Conceived and designed the experiments: FF JH.

Performed the experiments: FF.

Analyzed the data: FF JH BK.

Wrote the paper: FF BK FJT JH.

References

1. Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. Mol Syst Biol; 2013; 9(640). doi: 10.1038/msb.2012.61 PMID: 23340846
2. Chen R et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell; 2012; 148(6):1293–1307. doi: 10.1016/j.cell.2012.02.009 PMID: 22424236
3. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Gen; 2013; 45(10):1127–1133. doi: 10.1038/ng.2762
4. Kitano H. Computational Systems Biology. Nature; 2002; 420(6912):206–210. doi: 10.1038/nature01254 PMID: 12432404

5. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, et al. COPASI—a COmplex PAthway SImulator. Bioinf; 2006; 22:3067–3074. doi: 10.1093/bioinformatics/btl485
6. Balsa-Canto E, Banga JR. AMIGO, a toolbox for advanced model identification in systems biology using global optimization. Bioinf; 2011; 27(16):2311–2313. doi: 10.1093/bioinformatics/btr370
7. Raue A, Steiert B, Schelker M, Kreutz C, Maiwald T, Hass H, et al. Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems. Bioinf; 2015; 31(21):3558–3560. doi: 10.1093/bioinformatics/btv405
8. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res; 2000; 28 (1):27–30. doi: 10.1093/nar/28.1.27 PMID: 10592173
9. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, et al. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res; 2011; 39(Database issue):D691–7. doi: 10.1093/nar/gkq1018 PMID: 21067998
10. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res; 2013; 41 (Database issue):D808–15. doi: 10.1093/nar/gks1094 PMID: 23203871
11. Büchel F, Rodriguez N, Swainston N, Wrzodek C, Czauderna T, Keller R, et al. Path2Models: large-scale generation of computational models from biochemical pathway maps. BMC Syst Biol; 2013; 7 (116).
12. Thiele I, Swainston N, Fleming RMT, Hoppe A, Sahoo S, Aurich MK, et al. A community-driven global reconstruction of human metabolism. Nat Biotechnol; 2013; 31(5):419–25. doi: 10.1038/nbt.2488 PMID: 23455439
13. Ganter M, Kaltenbach HM, Stelling J. Predicting network functions with nested patterns. Nat Commun; 2014; 5:3006. doi: 10.1038/ncomms4006 PMID: 24398547
14. Dräger A, Zielinski DC, Keller R, Rall M, Eichner J, Palsson BO, et al. SBMLsqueezer 2: context-sensitive creation of kinetic equations in biochemical networks. BMC Syst Biol; 2015; 9:68. doi: 10.1186/s12918-015-0212-9 PMID: 26452770
15. Stanford NJ, Lubitz T, Smallbone K, Klipp E, Mendes P, Liebermeister W. Systematic construction of kinetic models from genome-scale metabolic networks. PLoS ONE; 2013; 8(11):e79195. doi: 10.1371/journal.pone.0079195 PMID: 24324546
16. Villaverde AF, Bongard S, Mauch K, Müller D, Balsa-Canto E, Schmid J, et al. A consensus approach for estimating the predictive accuracy of dynamic models in biology. Computer Methods and Programs in Biomedicine; 2015; 119(1):17–28. doi: 10.1016/j.cmpb.2015.02.001 PMID: 25716416
17. Hadamard J. Sur les problémes aux dérivées partielles et leur signification physique. In: Princeton University Bulletin; 1902. p. 49–52.
18. Raue A, Schilling M, Bachmann J, Matteson A, Schelke M, Kaschek D, et al. Lessons learned from quantitative dynamical modeling in systems biology. PLoS ONE; 2013; 8(9):e74335. doi: 10.1371/journal.pone.0074335 PMID: 24098642
19. Bäck T. Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. New York and Oxford: Oxford University Press; 1996.
20. Yang XS. Nature-inspired metaheuristic algorithms. 2nd ed. Bristol, UK: Luniver Press; 2010.
21. Kirkpatrick S, Gelatt CD Jr, M P Vecchi MP. Optimization by simulated annealing. Science; 1983; 220 (4598):671–680. doi: 10.1126/science.220.4598.671 PMID: 17813860
22. Balsa-Canto E, Peifer M, Banga JR, Timmer J, Fleck C. Hybrid optimization method with general switching strategy for parameter estimation. BMC Syst Biol; 2008; 2(26). doi: 10.1186/1752-0509-2-26 PMID: 18366722
23. Vaz A, Vicente L. A particle swarm pattern search method for bound constrained global optimization. J Global Optim; 2007; 39(2):197–219. doi: 10.1007/s10898-007-9133-5
24. Moles CG, Mendes P, Banga JR. Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Res; 2003; 13:2467–2474. doi: 10.1101/gr.1262503 PMID: 14559783
25. Banga JR. Optimization in computational systems biology. BMC Syst Biol; 2008; 2(47).
26. Weise T. Global Optimization Algorithms: Theory and Application. Nature Inspired Computation and Applications Laboratory (NICAL), University of Science and Technology, China; 2009.
27. Villaverde AF, Henriques D, Smallbone K, Bongard S, Schmid J, Cicin-Sain D, et al. BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology. BMC Syst Biol; 2015; 9(8). doi: 10.1186/s12918-015-0144-4 PMID: 25880925
28. Griewank A, Walther A. Evaluating Derivatives. 2nd ed. Society for Industrial and Applied Mathematics; 2008.

29. Nocedal J, Wright S. Numerical optimization. Springer Science & Business Media; 2006.
30. Hindmarsh AC, Brown PN, Grant KE, Lee SL, Serban R, Shumaker DE, et al. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. ACM T Math Software; 2005; 31(3):363–396. doi: 10.1145/1089014.1089020
31. Lu J, Muller S, Machné R, Flamm C. SBML ODE Solver library: Extensions for inverse analysis. In: Proceedings of the Fifth International Workshop on Computational Systems Biology, WCSB; 2008.
32. Fujarewicz K, Kimmel M, Swierniak A. On fitting of mathematical models of cell signaling pathways using adjoint systems. MBE; 2005; 2(3):527–534. doi: 10.3934/mbe.2005.2.527 PMID: 20369938
33. Lu J, August E, Koeppl H. Inverse problems from biomedicine: Inference of putative disease mechanisms and robust therapeutic strategies. J Math Biol; 2012; 67(1):143–168. doi: 10.1007/s00285-012-0523-z PMID: 22526835
34. Hanke M, Scherzer O. Inverse problems light: Numerical differentiation. The American Mathematical Monthly; 2001; 108(6):512–521. doi: 10.2307/2695705
35. Plessix RE. A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. Geophys J Int; 2006; 167(2):495–503. doi: 10.1111/j.1365-246X.2006.02978.x
36. Leonhardt C, Schwake G, Stögbauer TR, Rappl S, Kuhr JT, Ligon TS, et al. Single-cell mRNA transfection studies: Delivery, kinetics and statistics by numbers. Nanomedicine: Nanotechnology, Biology, and Medicine; 2014; 10(4):679–688.
37. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol; 2010; 4:92. doi: 10.1186/1752-0509-4-92 PMID: 20587024
38. Smith GR, Shanley DP. Computational modelling of the regulation of Insulin signalling by oxidative stress. BMC Syst Biol; 2013; 7(1):1. doi: 10.1186/1752-0509-7-41
39. Kühn C, Wierling C, Kühn A, Klipp E, Panopoulou G, Lehrach H, et al. Monte carlo analysis of an ode model of the sea urchin endomesoderm network. BMC Syst Biol; 2009; 3(1):1.
40. Chen WW, Schoeberl B, Jasper PJ, Niepel M, Nielsen UB, Lauffenburger DA, et al. Input–output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. Mol Syst Biol; 2009; 5(239).
41. Chassagnole C, Noisommit-Rizzi N, Schmid JW, Mauch K, Reuss M. Dynamic modeling of the central carbon metabolism of Escherichia coli. Biotechnol Bioeng; 2002; 79(1):53–73. doi: 10.1002/bit.10288 PMID: 17590932
42. Kotte O, Zaugg JB, Heinemann M. Bacterial adaptation through distributed sensing of metabolic fluxes. Mol Syst Biol; 2010; 6(1). doi: 10.1038/msb.2010.10 PMID: 20212527
43. Villaverde AF, Bongard S, Mauch K, Müller D, Balsa-Canto E, Schmid J, et al. High-Confidence Predictions in Systems Biology Dynamic Models. In: 8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014). Cham: Springer International Publishing; 2014. p. 161–171.
44. MacNamara A, Terfve C, Henriques D, Bernabé BP, Saez-Rodriguez J. State–time spectrum of signal transduction logic models. Phys Biol; 2012; 9(4):045003. doi: 10.1088/1478-3975/9/4/045003 PMID: 22871648
45. Smallbone K, Mendes P. Large-scale metabolic models: From reconstruction to differential equations. Ind Biotechnol; 2013; 9(4):179–184. doi: 10.1089/ind.2013.0003
46. Longford NT. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Biometrika; 1987; 74(4):817–827. doi: 10.1093/biomet/74.4.817
47. Fletcher R, Powell MJ. A rapidly convergent descent method for minimization. Comp J; 1963; 6(2):163–168. doi: 10.1093/comjnl/6.2.163
48. Goldfarb D. A Family of Variable-Metric Methods Derived by Variational Means. Math Comp; 1970; 24 (109):23–26. doi: 10.1090/S0025-5718-1970-0258249-6
49. Guay M, McLean DD. Optimization and sensitivity analysis for multiresponse parameter estimation in systems of ordinary differential equations. Comp and Chem Eng; 1995; 19(12):1271–1285. An International Journal of Computer Application in Chemical Engineering. doi: 10.1016/0098-1354(94)00120-0
50. Özyurt DB, Barton PI. Cheap Second Order Directional Derivatives of Stiff ODE Embedded Functionals. SIAM Journal on Scientific Computing; 2005; 26(5):1725–1743. doi: 10.1137/030601582
51. Martins JRRA, Sturdza P, Alonso JJ. The complex-step derivative approximation. ACM Trans Math Softw; 2003; 29:245–262. doi: 10.1145/838250.838251
52. Chang YF, Corliss GF. ATOMFT: Solving ODEs and DAEs using Taylor series. Comp Math App; 1994; 28(10–12):209–233. doi: 10.1016/0898-1221(94)00193-6

53. Linnainmaa S. Taylor expansion of the accumulated rounding error. BIT; 1976; 16(2):146–160. doi: 10.1007/BF01931367

54. Egea JA, Henriques D, Cokelaer T, Villaverde AF, MacNamara A, Danciu DP, et al. MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. BMC Bioinf; 2014; 15(136).

55. Schmidt H, Jirstrand M. Systems biology toolbox for MATLAB: a computational platform for research in systems biology. Bioinf; 2006; 22(4):514–515. doi: 10.1093/bioinformatics/bti799

56. Gonnet P, Dimopoulos S, Widmer L, Stelling J. A specialized ODE integrator for the efficient computation of parameter sensitivities. BMC Syst Biol; 2012; 6(46). doi: 10.1186/1752-0509-6-46 PMID: 22607742

57. Gábor A, Banga JR. Robust and efficient parameter estimation in dynamic models of biological systems. BMC Syst Biol; 2015; 9:74. doi: 10.1186/s12918-015-0219-2 PMID: 26515482

58. Wittig U, Kania R, Golebiewski M, Rey M, Shi L, Jong L, et al. SABIO-RK—database for biochemical reaction kinetics. Nucl Acids Res; 2012; 40(D1):D790–D796. doi: 10.1093/nar/gkr1046 PMID: 22102587

59. Chang A, Scheer M, Grote A, Schomburg I, Schomburg D. BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. Nucleic Acids Res; 2009; 37(Database issue):D588–92. doi: 10.1093/nar/gkn820 PMID: 18984617

60. Neal RM. MCMC using Hamiltonian dynamics. In: Brooks S, Gelman A, Jones G, Meng XL, editors. Handbook of Markov Chain Monte Carlo. London, United Kingdom: Chapman & Hall / CRC Press; 2011.

61. Girolami M, Calderhead B. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J R Statist Soc B; 2011; 73(2):123–214. doi: 10.1111/j.1467-9868.2010.00765.x

62. Lan S, Bui-Thanh T, Christie M, Girolami M. Emulation of higher-order tensors in manifold Monte Carlo methods for Bayesian inverse problems. arXiv; 2015. arXiv:1507.06244v2 [stat.CO].

63. Fröhlich F, Hross S, Theis FJ, Hasenauer J. Radial basis function approximation of Bayesian parameter posterior densities for uncertainty analysis. In: Mendes P, Dada JO, Smallbone KO, editors. Proceedings of the 12th International Conference on Computational Methods in Systems Biology (CMSB 2014), Manchester, UK. Lecture Notes in Bioinformatics. Springer International Publishing Switzerland; 2014. p. 73–85.

64. Xu TR, Vyshemirsky V, Gormand A, von Kriegsheim A, Girolami M, Baillie GS, et al. Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. Sci Signal; 2010; 3(113):ra20. doi: 10.1126/scisignal.2000517

65. Hug S, Raue A, Hasenauer J, Bachmann J, Klingmüller U, Timmer J, et al. High-dimensional Bayesian parameter estimation: Case study for a model of JAK2/STAT5 signaling. Math Biosci; 2013; 246 (2):293–304. doi: 10.1016/j.mbs.2013.04.002 PMID: 23602931

66. Hasenauer J, Jagiella N, Hross S, Theis FJ. Data-driven modelling of biological multi-scale processes. Journal of Coupled Systems and Multiscale Dynamics; 2015; 3(2):101–121. doi: 10.1166/jcsmd.2015.1069

67. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet; 2013; 14(5):333–346. doi: 10.1038/nrg3433 PMID: 23594911

Appendix C

Efficient parameterization of large-scale mechanistic models enables drug response prediction for cancer cell lines.

This is an author-produced PDF of a bioRxiv preprint.: F. Fröhlich, T. Kessler, D. Weindl, A. Shadrin, L. Schmiester, H. Hache, A. Muradyan, M. Schütte, J. Lim, M. Heinig, F.J. Theis, H. Lehrach, C. Wierling, B. Lange, and J. Hasenauer. **Efficient parameterization of large-scale mechanistic models enables drug response prediction for cancer cell lines.** bioRxiv:174094.

Efficient parameterization of large-scale mechanistic models enables drug response prediction for cancer cell lines

Fabian Fröhlich[1,2,*], Thomas Kessler[3,4,*], Daniel Weindl[1], Alexey Shadrin[6,7], Leonard Schmiester[1,2], Hendrik Hache[3], Artur Muradyan[3], Moritz Schütte[3], Ji-Hyun Lim[3], Matthias Heinig[1,2], Fabian J. Theis[1,2], Hans Lehrach[4,5], Christoph Wierling[3,4], Bodo Lange[3,4], and Jan Hasenauer[1,2]

[1]Institute of Computational Biology, Helmholtz Zentrum München, 85764 Neuherberg, Germany
[2]Chair of Mathematical Modeling of Biological Systems, Center for Mathematics, Technische Universität München, 85748 Garching, Germany
[3]Alacris Theranostics GmbH, 12489 Berlin, Germany
[4]Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany
[5]Dahlem Centre for Genome Research and Medical Systems Biology, 12489 Berlin, Germany
[6]NORMENT, KG Jebsen Centre for Psychosis Research, Institute of Clinical Medicine, University of Oslo, 0450 Oslo, Norway
[7]Division of Mental Health and Addiction, Oslo University Hospital, 0450 Oslo, Norway

*These authors contributed equally to this work.
Correspondence should be addressed to B.L. (b.lange@alacris.de) or J.H. (jan.hasenauer@helmholtz-muenchen.de).

The response of cancer cells to drugs is determined by various factors, including the cells' mutations and gene expression levels. These factors can be assessed using next-generation sequencing. Their integration with vast prior knowledge on signaling pathways is, however, limited by the availability of mathematical models and scalable computational methods. Here, we present a computational framework for the parameterization of large-scale mechanistic models and its application to the prediction of drug response of cancer cell lines from exome and transcriptome sequencing data. With this framework, we parameterized a mechanistic model describing major cancer-associated signaling pathways (>1200 species and >2600 reactions) using drug response data. For the parameterized mechanistic model, we found a prediction accuracy, which exceeds that of the considered statistical approaches. Our results demonstrate for the first time the massive integration of heterogeneous datasets using large-scale mechanistic models, and how these models facilitate individualized predictions of drug response. We anticipate our parameterized model to be a starting point for the development of more comprehensive, curated models of signaling pathways, accounting for additional pathways and drugs.

Introduction

Personalized tumor therapy relies on our ability to predict the drug response of cancer cells from genomic data[1]. This requires the integration of genomic data with available prior knowledge, and its interpretation in the context of cancer-associated processes[2]. At the heart of this endeavor are statistical and mechanistic mathematical models[3]. In patient

stratification, statistical models are used to derive prognostic and predictive signatures of tumor subtypes[4,5]. Linear and nonlinear regression, machine learning methods and related approaches have been used to obtain such signatures[6]. Yet, purely statistical models do not provide mechanistic insights or information about actionable targets. High-quality mechanistic models of cancer signaling are thus of interest to researchers and clinicians in systems biology and systems medicine.

Mechanistic models aim to quantitatively describe biological processes. Consequently, they facilitate the systematic integration of prior knowledge on signaling pathways, as well as the effect of somatic mutations and gene expression. These models have been used for the identification of drug targets[7] as well as the development of prognostic signatures[8,9]. Furthermore, mechanistic modeling has facilitated the study of oncogene addiction[10], synthetic-lethal phenotypes[11] and many other relevant phenomena[12].

Various pathways have been modeled mechanistically using Ordinary Differential Equations (ODEs) of varying detail[13]. ERBB, MAPK and PI3K signaling attracted special attention as they are altered in many cancer types[14] and targeted by many drugs[15]. Tailoring the models to individual pathways ensures manageability of the development effort, but neglects crosstalk. The Atlas of Cancer Signaling Network (ACSN) addresses this issue by covering a majority of molecular processes implicated in cancer[16]. However, like other pathway maps[17,18], the ACSN lacks kinetic rate laws and rate constants, preventing numerical simulation and quantitative prediction. This might be addressed in the future by using comprehensive databases[13,19,20] in combination with semi-automatic[21–23] or automatic reconstruction methods[8,24,25].

After the construction of a mechanistic model, parameterization from experimental data is necessary to render the model predictive. Optimization methods achieve this by iteratively minimizing the objective function, i.e. the distance between model simulation and experimental data[26,27]. This requires repeated numerical simulations. As even for medium-scale models millions of simulations are necessary, the computational burden is often immense[28]. Accordingly, parameterizing large-scale pathway models is often deemed intractable and has not been done in practice. A scalable method for parameterization of large-scale mechanistic models is therefore essential for the community as it enables the comprehensive integration of prior knowledge and experimental data.

Here, we describe a large-scale mechanistic model of cancer signaling which is individualized using information about somatic mutations and gene expression levels. We introduce a computational framework for the parameterization of large-scale ODE models that reduces the computation time by multiple orders of magnitude compared to state-of-the-art methods. We demonstrate the parameterization of the model from thousands of drug assays from over 100 human cancer cell lines and validate the predictive power of the model. Moreover, we show that the mechanistic model outperforms all investigated statistical models in terms of classification accuracy and generalizes to cancer cell lines originating from tissues not used for training.

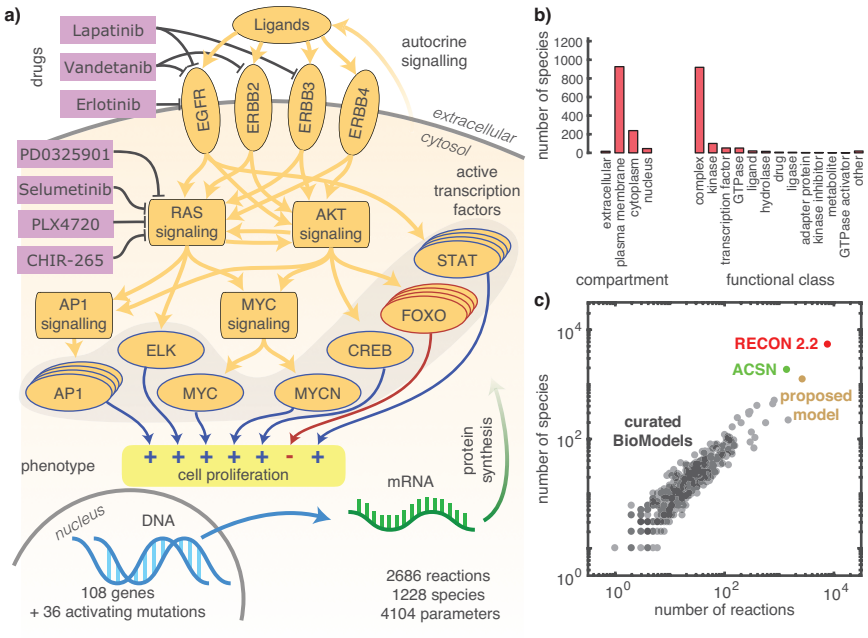


Figure 1. Model structure and properties. (a) Sketch of modeled signaling pathways. The developed model describes the synthesis and protein-protein interactions for protein products of 108 genes and 36 activating mutations. The visualization depicts drugs (purple), selected molecular species (orange) and cell proliferation as phenotypic readout (yellow). (b) Distribution of modeled molecular species on compartments and functional classes (c) Comparison of complexity of the proposed model with curated models from the BioModels database[13], Recon 2.2[36] and the ACSN[16].

Results

Large-scale mechanistic model integrates knowledge of cancer signaling pathways

To predict the drug response of cancer cell lines, we developed a mechanistic model integrating signaling modules reflecting the human ERBB, RAS and PI3K/AKT signaling pathways, as well as regulation of the transcription factors MYC and AP1[29] (Fig. 1a). This model describes synthesis, degradation, translocation, complex formation, phosphorylation and various other types of reactions for proteins and their functional variants (Supplementary Fig. 1). We assembled it manually using the web-based platform PyBioS[30,31] and provide it as annotated Systems Biology Markup Language (SBML) file (Supplementary File 1). The model is based on curated information from ConsensusPathDB[32], a meta-database integrating more than 20 public databases (e.g., DrugBank[33], KEGG[34] and Reactome[35]), and additional publications.

The model accounts for 108 genes and 36 activating mutations yielding a total of 1228 molecular species in 4 cellular compartments (Fig. 1b) involved in 2686 reactions. The modeled mutations cover 7 of the 10 most frequent driver mutations reported by Rubio-Perez et al.[36] and account for 22.1% of driver mutations observed in patient samples[36]. The model describes the action of 7 different small molecule kinase inhibitors, of which 4 are FDA-approved. For 17 additional FDA-approved kinase inhibitors, one or more main targets are included in the model, but the action is currently not described. Thus, the model covers main targets for 27.3% of FDA-approved targeted cancer therapies.

To quantify the scale of our model, we compared it to curated models available in the BioModels database[13]. The proposed model describes more biochemical species and

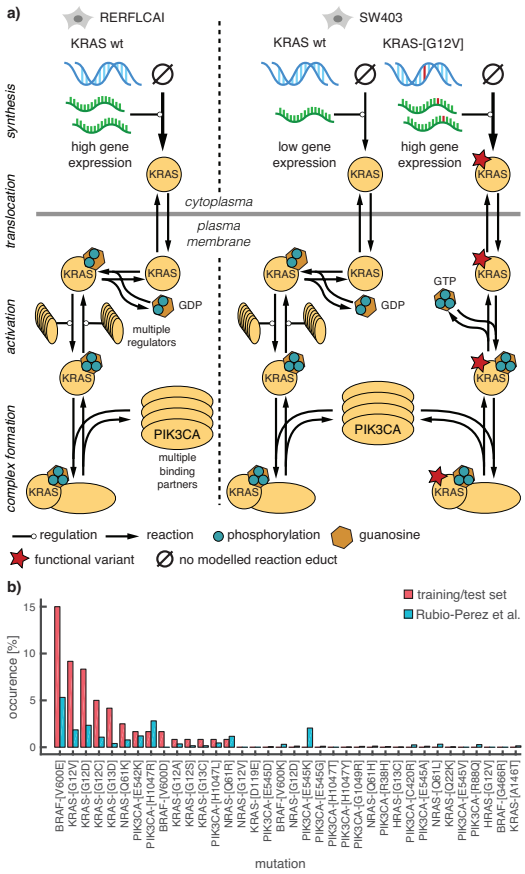


Figure 2. Individualization of the model with genomic and transcriptomic data. (a) Individualization of the generic mechanistic model for the two different cell lines: RERFLCAI (wild-type KRAS); and SW403 (wild-type and mutated KRAS). KRAS signaling model is illustrated from synthesis until complex formation. Degradation reactions are omitted. (b) Comparison of the occurrence frequency of mutations included in the model between the training/test set extracted from the Cancer Cell Line Encyclopedia and the InTOGen database by Rubio-Perez *et al.*[36], which provides an extensive characterization of somatic mutations in human tumors.

reactions than any other of the curated models (Fig. 1c). Two pathway maps, Recon 2.2[18] and the ACSN, possess a similar size than our model. Yet these pathway maps do not provide kinetic rate laws and Recon 2.2 does not focus on cancer signaling. Hence, we conclude that the proposed model is one of the most extensive mechanistic models of cancer signaling.

Most drug response assays provide information about the cell proliferation rate relative to the untreated condition. Cell proliferation of cancer cells is governed by a complex interplay of cellular signaling processes regulating e.g. the balance between pro-growth and (anti-) apoptotic signals in response to extracellular stimuli or presence of activating mutations within respective signal transduction cascades. A major function in regulation of cell proliferation has been attributed to transcription factor (TF) activation, e.g. of the MYC, AP1 and FOXO transcription factors, and regulation of target gene expression in response to extracellular or oncogenic stimuli. In the current model we used the weighted sums of the simulated molecularly activated state of these TFs as a surrogate for proliferation (see Online Methods, Section Model Development). This semi-mechanistic description provides a simple model of down-stream regulatory processes.

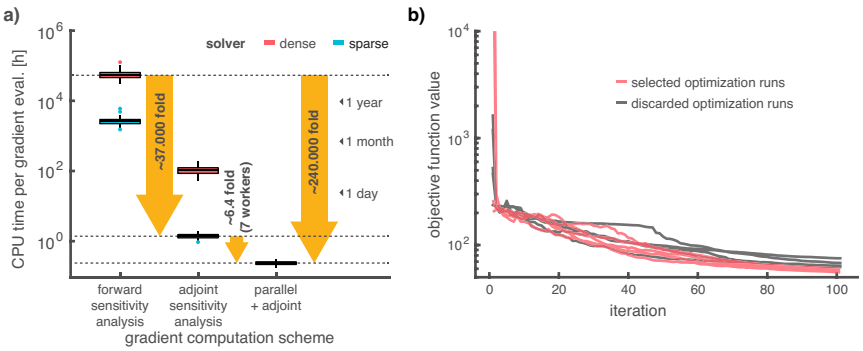


Figure 3. Parameterization of the mechanistic model. (a) Computation time for one evaluation of the objective function gradient, which determines the per-iteration time for a single local optimization step. For the non-parallelized evaluation, the time was computed based on representative samples (See Online Methods, Numerical Benchmark). The gradient evaluation time was dramatically reduced by using adjoint sensitivities, exploiting sparsity and parallelization. (b) Objective function traces for ten different local optimization runs for the first cross-validation set. Initial conditions for the local optimization runs are sampled from a latin hypercube. Although higher initial objective function values were observed, the corresponding axis was cropped at 10^4. The 5 best optimization runs are colored in red and used for subsequent analysis.

Genomic data provides basis for individualization of the mechanistic model

The mechanistic model provides a generic template for a subset of signaling processes in human cells. To obtain a model for a particular cancer cell line, we individualized the mechanistic model by incorporating gene expression levels as synthesis rates for proteins and their mutated functional variants (Fig. 2a). We assumed that all other kinetic parameters such as transport, binding and phosphorylation rates, only depend on the chemical properties of the involved biochemical species. Accordingly, these parameters differ between proteins and their functional variants, while they are assumed to be identical across cell lines. This enables the simultaneous consideration of multiple cell lines and drugs for the model parameterization, increasing the available training sample size. Furthermore, the assumption allows us to predict the drug response of new cell lines from information about gene expression levels and functional variants.

In this study, we considered data for 120 human cancer cell lines from 5 tissues (breast, large-intestine, lung, pancreas and skin) provided in the Cancer Cell Line Encyclopedia (CCLE)[37]. We processed the included genetic characterization of cell lines in the untreated condition using a standardized bioinformatics pipeline (Online Methods, Section Data Processing). Of the modeled driver-mutations, 14 are present in more than one cell line (Fig. 2b).

Scalable, parallel optimization method enables model parameterization

The mechanistic model includes more than 4,100 unknown parameters, i.e. kinetic constants and weighting factors. To describe the available data and to predict future experiments, we parameterized the model using measured proliferation data from 120 cell lines treated with 7 different drugs at up to 9 concentrations provided in the CCLE. In total, this dataset provides more than 6,900 experimental conditions. To assess the prediction uncertainty, we performed 5-fold cross-validation with 5 pairs of training (80%; 96 cell lines) and test datasets (20%; 24 cell lines).

To parameterize the model from the training data, we minimized the sum of squared residuals of measured and simulated relative proliferation. This non-linear and non-convex

ODE-constrained optimization problem was solved using multi-start local optimization, an efficient and reliable approach that outperformed global optimization methods in several studies[26,38]. As the optimization problem is high dimensional, we first assessed the applicability of state-of-the-art methods, such as forward sensitivity analysis[26]. Therefore, we determined the computation time per gradient evaluation. This revealed that due to (i) the large-scale ODE model, (ii) the large number of parameters and (iii) the large number of experimental conditions, a single evaluation of the objective function gradient would require more than $5 \cdot 10^4$ CPU hours (> 6 CPU years) (Fig. 3a). As the gradient has to be evaluated hundreds of times for a single optimization, available toolboxes were not applicable.

To render parameterization tractable, we addressed challenges (i)-(iii). Firstly, we reduced the CPU time per model evaluation by using a sparse linear solver[39] for ODE integration (0.5% non-zero entries in the Jacobian). Secondly, we implemented adjoint sensitivity analysis[40] which improves scaling with the number of parameters. These two methodological advancements reduced the computation time over 37,000-fold (Fig. 3a).

Thirdly, we established scalability with respect to the number of experimental conditions by parallelization on the level of cell lines (Fig. 3a). Using 8 cores (7 workers), we observed a 6.4-fold acceleration. In total, our flexible and easily extendable parameterization framework reduced the expected wall time by over 240,000-fold.

Using 400 cores and a trivial parallelization over local optimizations, our computational framework enabled the parameterization for all cross-validations in less than one week. In comparison, state-of-the-art approaches would have required hundreds of thousands of years. The local optimization achieved a substantial reduction of the sum of squared residuals within a few iterations and then the curve flattened out (Fig. 3b). The optimization was stopped early at 100 iterations to improve the prediction accuracy by avoiding overfitting[41]. To filter insufficient optimization runs and improve robustness, we used ensemble averaging (see Online Methods, Section Ensemble Averaging) over the 5 optimization runs that achieved lowest objective function value in each cross-validation for all following analysis and prediction.

Mechanistic model yields quantitative description of experimental data and generalizes to test data

The parameterized model describes the drug dose-dependent proliferation of cell lines. To assess the combined quality of the model and the parameterization, we quantified the model-data mismatch. Our study of the parameterized model revealed a good agreement of measured data and model simulation and little variation in the prediction (Fig. 4a), despite large parameter uncertainties (Supplementary Fig. 2). For the highest drug concentration (8μM) of each dose response curve, the correlation of measured and simulated proliferation was $r=0.82$ ($p<10^{-150}$) (Fig. 4b, for more concentrations see Supplementary Fig. 3).

The agreement of measured and simulated proliferation is similar for all tissues but varies between drugs (Fig. 4c). Similarly, no dependence of the model-data mismatch on mutation status could be identified (Supplementary Fig. 4). Particularly good correlations were

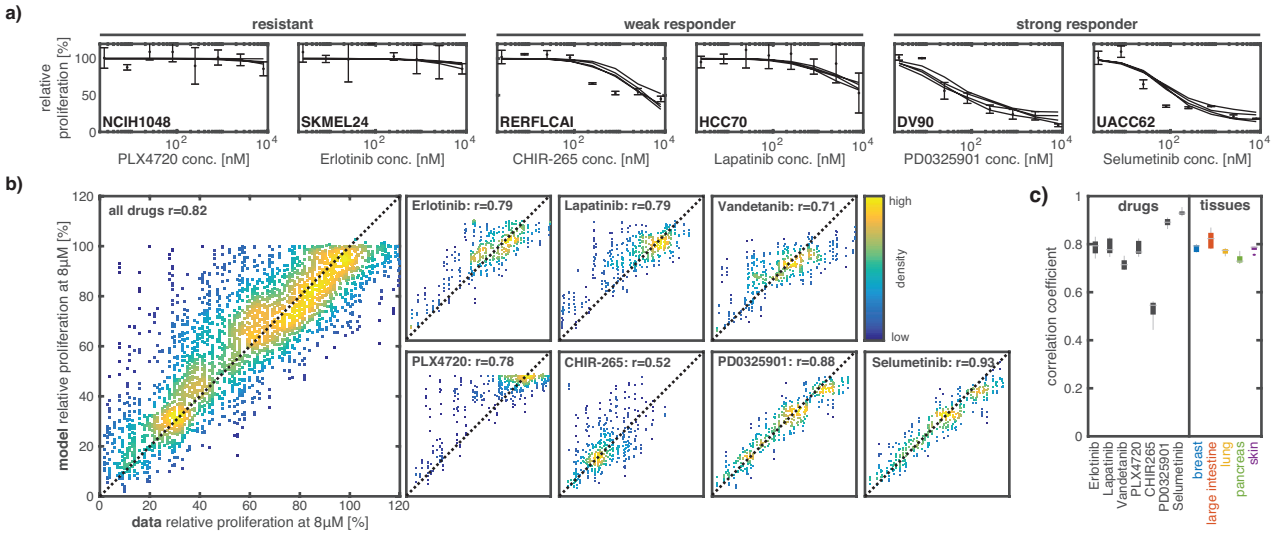


Figure 4. Analysis of fitting properties of the model. (a) Representative examples of model simulations for six combinations of drugs (x-label) and cell lines (bold text on bottom, left). The five plotted lines are the median fit for the five best optimization runs for every cross-validation set. (b) Correlation of simulation and measured proliferation for the response at maximal concentration. For the simulation the ensemble prediction based on the median is shown. Smaller subplots show the correlation filtered for individual drugs. (c) Correlation statistics over cross-validations for individual drugs and tissues.

achieved for selumetinib (r=0.93) and PD0325901 (r=0.88). Interestingly, CHIR-265 (r=0.52) and PLX4720 (r=0.78) have distinct correlation coefficients although they share B-RAF^{V600E} as main target with similar affinity. Still, many cell lines respond to CHIR-265, but appear to be resistant against PLX4720. This suggests that the molecular understanding of the drug/target effects or the implementation in the model may be incomplete. For example, inhibition of VEGFR2 activation by CHIR-265 has been described[42], but is not captured by the current version of the model.

To evaluate the predictive power of the parameterized mechanistic model, we turned to the test sets of the cross-validation (see Fig. 5a left). We quantified (i) the correlation of measured and predicted proliferation as well as (ii) the classification accuracy for responder cell lines in terms of the area under the receiver-operating-characteristic (ROC) and the precision-recall (PR) curve. A cell line was considered a responder to a drug when the proliferation at the highest drug concentration was below 50% compared to the untreated control. Our analysis of the correlation revealed a good quantitative agreement of measured and predicted relative proliferation for the test set ($r=0.55$, $p<10^{-8}$) (Fig. 5b), a lower correlation comparing to the training set data ($r=0.82$, $p<10^{-150}$) (Fig. 5b). For the qualitative predictions, we found an average classification accuracy of 76.7±1.8% (area under ROC=0.767±0.018, area under PR=0.73±0.022) (Fig. 5c and Supplementary Figure 5).

Mechanistic model outperforms established statistical models

To provide a reference for the performance of the parameterized mechanistic model, we trained several well-established statistical models on the training set. The statistical models include a random forest[43], sparse linear and nonlinear regression models[44], as well as a network-constrained sparse regression model (with network derived from the mechanistic model)[45]. The training of all statistical models was performed using state-of-the-art toolboxes (see Online Methods, Section Statistical Analysis). The best statistical model achieved a classification accuracy of 67.8±2.9% on the test set (Fig. 5d), which is 8.9

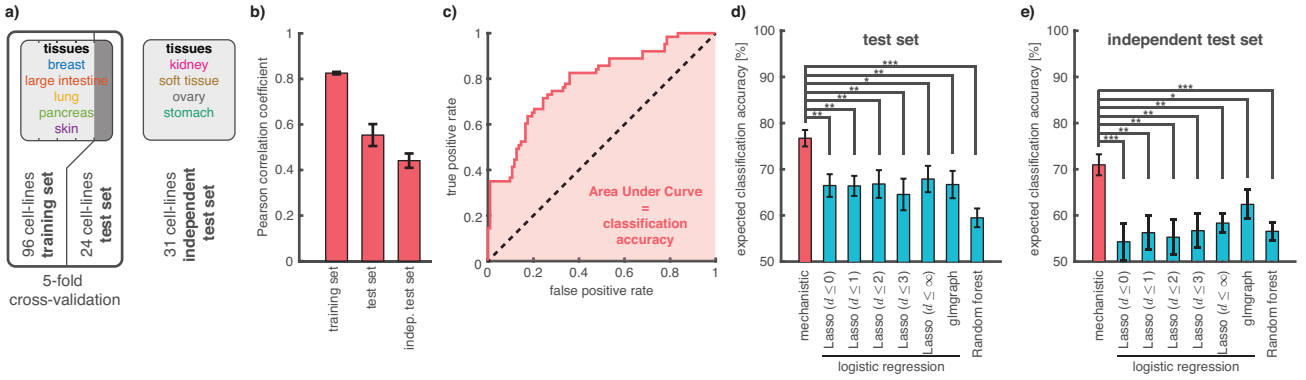


Figure 5. Validation of model prediction for single-drug treatment. (a) Overview over CCLE datasets used to evaluate the classification accuracy: (i) test set for cell lines originating from the same tissues on which the model was trained; and (ii) independent test set using cell lines from different tissues. (b) Correlation coefficients between measured and predicted proliferation for 8μM drug concentration on all considered datasets. Error bars show the standard error (c) Representative receiver-operating-characteristic curve for the mechanistic model on the test set (d-e) Comparison of classification accuracy across all tested methods for test set and independent test set. For the Lasso approach the d defines up to which network distance interactions are considered. Error bars show the standard error. Stars indicate statistical significance (*: p<0.05; **: p<0.01;***: p<0.001).

percentage points lower than the classification accuracy of the mechanistic model (76.7±1.8%). In conclusion, the parameterized mechanistic model provides significantly ($p<3.6\cdot10^{-2}$ according to Welch's *t*-test) more accurate classification than all considered statistical models.

Following these positive results, we assessed the generalization error of the mechanistic model. We processed experimental data for 31 additional cell lines from 4 additional tissues (kidney, ovary, soft tissue and stomach) available in the CCLE database that were not part of the initial training set (see Fig. 5a right). For this independent dataset, the parameterized mechanistic model achieved a classification accuracy of 70.7±2.3% (Fig. 5e), which is 6 percentage points lower than on the test set. Interestingly, the tested statistical models achieve a maximal classification accuracy of 62.1±3.1%, suggesting that our proposed parameterization framework for mechanistic models may achieve better generalization properties.

Combination treatment outcomes predicted from single treatment measurements

For an additional assessment, we predicted the outcome of combination treatments. We considered the dataset published by Friedman *et al.*[46] reporting the response of cancer cell lines to individual drugs as well as to combinations of two drugs. The dataset includes 7 cell lines and 5 drugs (selumetinib, CHIR-265, erlotinib, lapatinib, PLX4720) contained in our training set. To establish a reference for the accuracy of the prediction, we assessed the agreement of the proliferation measurements by Friedman *et al.*[46] and the measurements from the CCLE database. The comparison yielded a correlation of r=0.33. This weak correlation is likely due to differences in the experimental procedure and systematic bias in the proliferation measurements. These are known problems for similar pharmacogenetic studies[47,48] and apparently limit the achievable correlation between the predictions of the parameterized model and the Friedman data. The predicted proliferation from the mechanistic model achieves comparable agreement with the Friedman data for both individual drug treatments (r=0.35) and combinatorial drug treatments (r=0.26) (Fig. 6). Accordingly, for the combinatorial drug treatments, the model achieved a correlation with the Friedman data only 20% lower than the agreement between the datasets

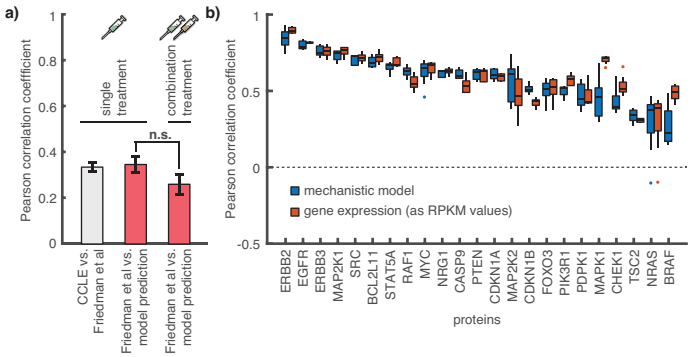


Figure 6. Assessment of model prediction for combination treatment dataset. (a) Correlation coefficients of the mechanistic model and data interpolation on the single and drug combination measurements by Friedman et al.[46] evaluated for cell lines contained in the training sets of the cross-validation. Error bars show the standard error. (b) Pearson correlation coefficients of the mechanistic model and gene expression levels (as reads per kilobase of transcript per Million mapped reads (RPKM)) values with proteomic data from the MD Anderson Cell Line Project[49] (For details see Online Methods, Section Validation). For the mechanistic model the correlation was evaluated for cell lines contained in the training sets of the cross-validation.

(0.26/0.33=0.79 compared to 1). The correlation difference between single and combination treatment was not statistically significant. We conclude that the prediction accuracy for combinatorial treatments of the mechanistic model is reasonable, given the reproducibility of the proliferation data.

Protein abundances predicted using mechanistic model

For the estimation of the model parameters only proliferation measurements were employed. As expected, our assessment of model uncertainties revealed that this limits the prediction accuracy (Online Methods, Section Uncertainty Analysis). The analysis suggested a low reliability of parameter estimates (Supplementary Fig. 2a), a higher reliability of prediction for protein abundances (Supplementary Fig. 2b) and the highest reliability for the proliferation readout (Supplementary Fig. 2c). To determine the accuracy of the predicted protein abundances in the untreated condition, we compared them to the measurements contained in the MD Anderson Cell Line Project (MCLP)[49]. The MCLP provides normalized Reverse Phase Protein Array (RPPA) data for 33 proteins described by the mechanistic model in 68 of the cell lines considered for the parameterization. We implemented the same normalization for the simulation and considered all proteins that were measured in at least 10 of the considered cell lines. The mechanistic model achieved an average correlation of r=0.57±0.03 (Figure 6b). This is similar to the correlation between gene expression and RPPA data (Figure 6b). We conclude that the individualization of the model with cell-specific gene expression levels allows for a reasonable prediction of the protein levels, despite the dependence on a large number of kinetic parameters, such as degradation rates, that were not constrained by any molecular data.

Discussion

We generated a large-scale mechanistic model that integrates large amounts of prior knowledge and expands upon previous large-scale models[16,18] by implementing kinetic rate laws, mutation variants of key regulators and possibilities for individualization. As the model can be individualized to particular cell lines and covers many relevant driver mutations, the

model provides a valuable resource for analysis of various cancer types and drug treatments.

To parameterize this model, we established a computational framework that provides scalability with respect to the number of parameters and number of state variables, and employs parallelization to handle the large number of experimental conditions. The final wall time requirement for all optimization runs (~$4 \cdot 10^3$ hours) was more than one order of magnitude lower than the wall time required for a single gradient evaluation using established methods (~$6 \cdot 10^4$ hours). This allowed, to the best of our knowledge for the first time, the parameterization of a large-scale mechanistic model from experimental data from over 100 cell lines, each under dozens of experimental conditions. The computational efficiency of the approach renders iterative rounds of optimization, hypothesis generation and model refinement of large-scale mechanistic models and multiple (heterogeneous datasets) feasible in a reasonable time frame. Our implementation of the methods is available as Supplementary File 1 and can be freely reused by other research groups.

The assessment of the parameterized model revealed that the prediction of cell proliferation – a key readout to cancer therapy – is accurate for single drug treatments. Hence, the large-scale mechanistic model we derived and parameterized can predict the drug response of cancer cell lines from sequencing data. This is in our opinion a result of combining extensive integration of prior knowledge on network structure and reaction kinetics parameterized with our scalable methods. We illustrated the broad capabilities of the mechanistic model by predicting protein abundances and the outcome of combination treatments from single treatment proliferation measurements, neither of which is possible with statistical models.

Our analysis, however, also revealed limitations of the available datasets and the parameter estimates. Firstly, for combination treatments, the weak correlation of the available datasets[37,46] limited the validation of our model and highlighted the need for accurate, reproducible phenotypic characterizations. Secondly, the parameterization of the model using only proliferation data resulted in large parameter uncertainty, which suggests that inclusion of proteomic and phosphoproteomic data in the training process will be necessary to render reliable predictions on the molecular level feasible. Thirdly, even more cell lines are necessary to capture the effect of driver mutations with low recurrence. However, as the model can be individualized to arbitrary cell lines and other experimental systems, it is particularly suited for the study of rare mutation patterns.

The developed model is currently limited to cell lines but can be extended in several ways. The modeling of additional intracellular processes, cell-cell communication, cancer heterogeneity or pharmacokinetics might improve the prediction of patients' response. To obtain a refined description, our model could be integrated with agent-based models for tumor growth[28,50] or physiology-based pharmacokinetic models[51]. The resulting models could provide valuable tools for the identification of novel drug targets[3], virtual clinical trials[52] and personalized medicine. The extensive mechanistic modeling of biological processes will therefore be an important area of future research.

Acknowledgements

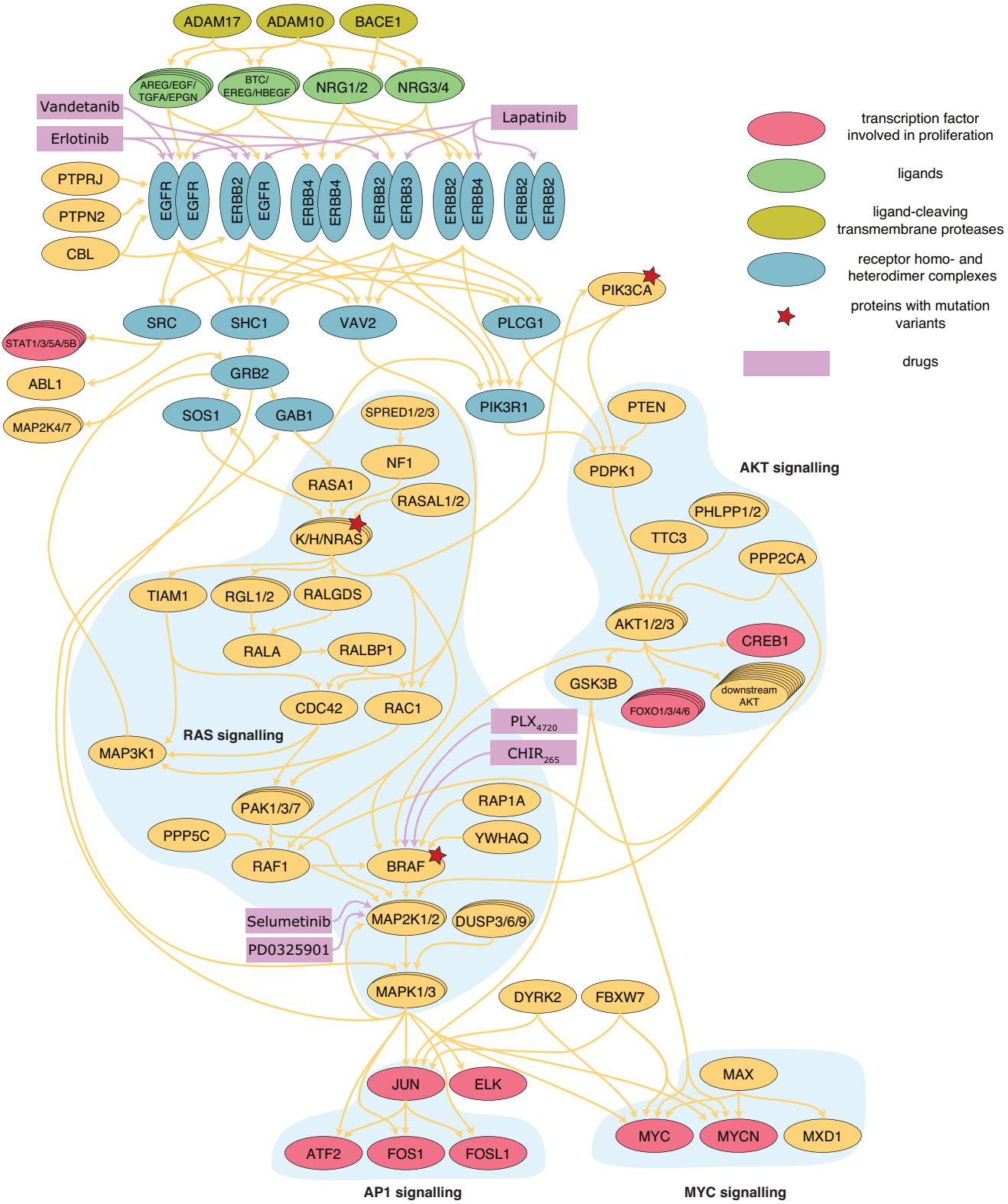
This work was supported by the German Research Foundation (DFG) through the Graduate School of Quantitative Biosciences Munich (QBM; F.F.), the European Union's Horizon 2020 research and innovation program (CanPathPro; Grant No. 686282; A.M., B.L., C.W., D.W., J.H., L.S. and M.S.), the German Federal Ministry of Education and Research (BMBF) within the SYS-Stomach project (Grant No. 01ZX1310B; J.H.) and the Postdoctoral Fellowship Program of the Helmholtz Zentrum München (J.H.).

Author Contributions

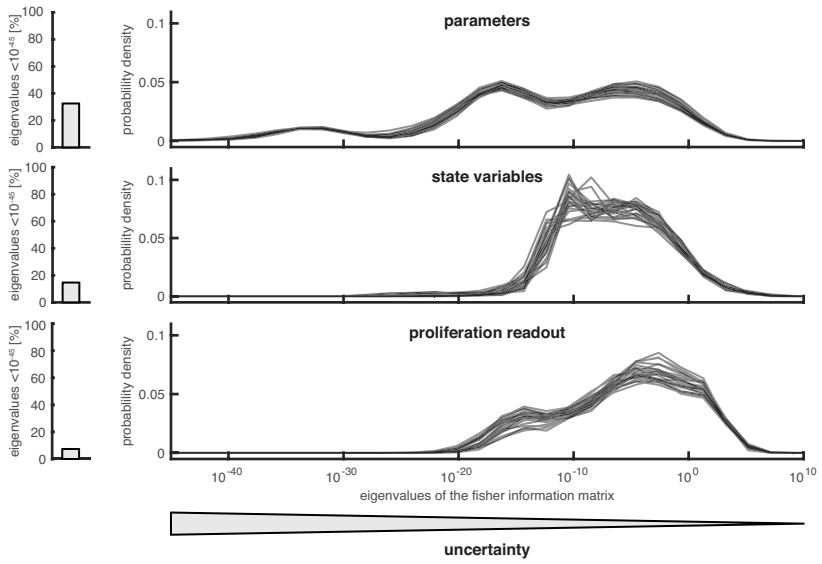
A.M. B.L., C.W., H.L. and T.K. designed the model. F.F. and J.H. designed the methods for numerical simulation, parameter optimization und uncertainty analysis. C.W., F.F., F.J.T., J.H., M.H. and T.K. designed the experiments. A.S., J.L., H.H., H.L., M.S. and T.K wrote and ran the code for model/data mapping and integrated and assembled model input data. D.W., F.F., J.H. and L.S. wrote and ran the code for the parameterization and assessment of the mechanistic model. M.H. wrote and ran the code for the parameterization and assessment of the statistical models. D.W., F.F., J.H., L.S. and M.H. analyzed output data. C.W., D.W., F.F., J.H., M.H. and T.K. wrote the manuscript. All authors discussed the results and implications and commented on the manuscript at all stages.

Competing Interests

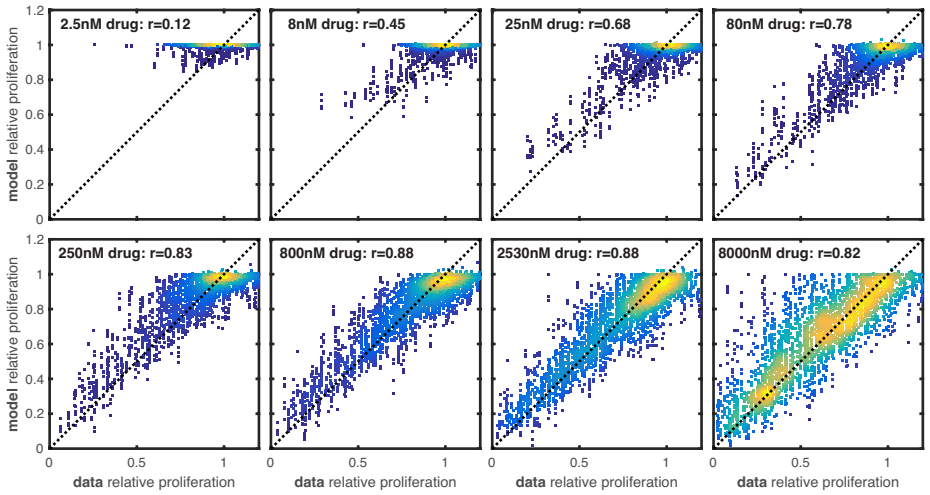
Several of the authors are employees (A.M., B.L., C.W., J.L., M.S. and T.K.), former employees (A.S. H.H) and founders (H.L.) of Alacris Theranostics GmbH. This company did however not influence the interpretation of the data, or the data reported, or financially profit by the publication of the results.



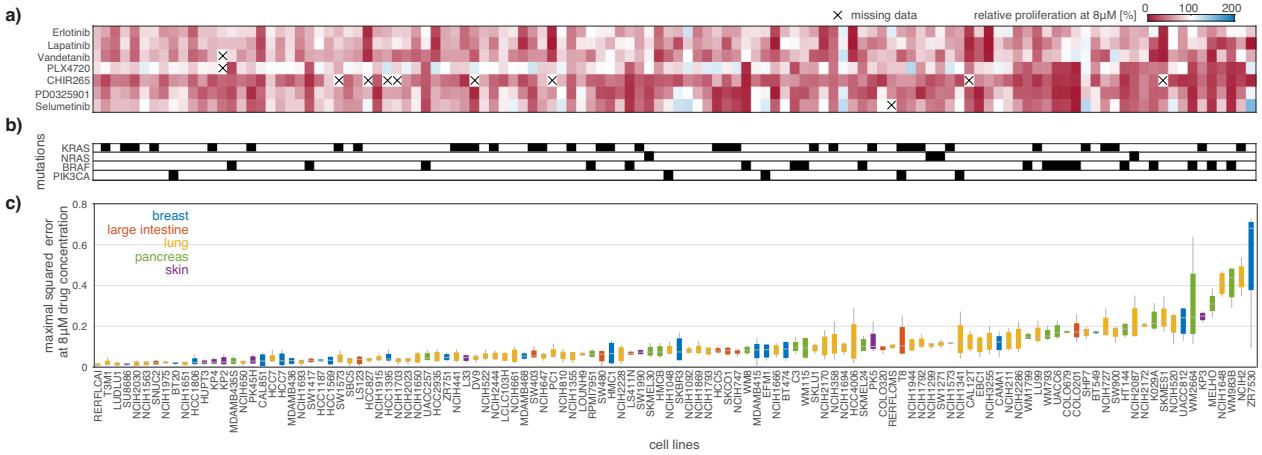
Supplementary Figure 1. Simplified overview of the model. The figure illustrates modeled interactions. Complex formation and phosphorylation as well as activation and repression are not discriminated here. Synthesis, translocation and degradation are omitted. All species are colored according to their function.



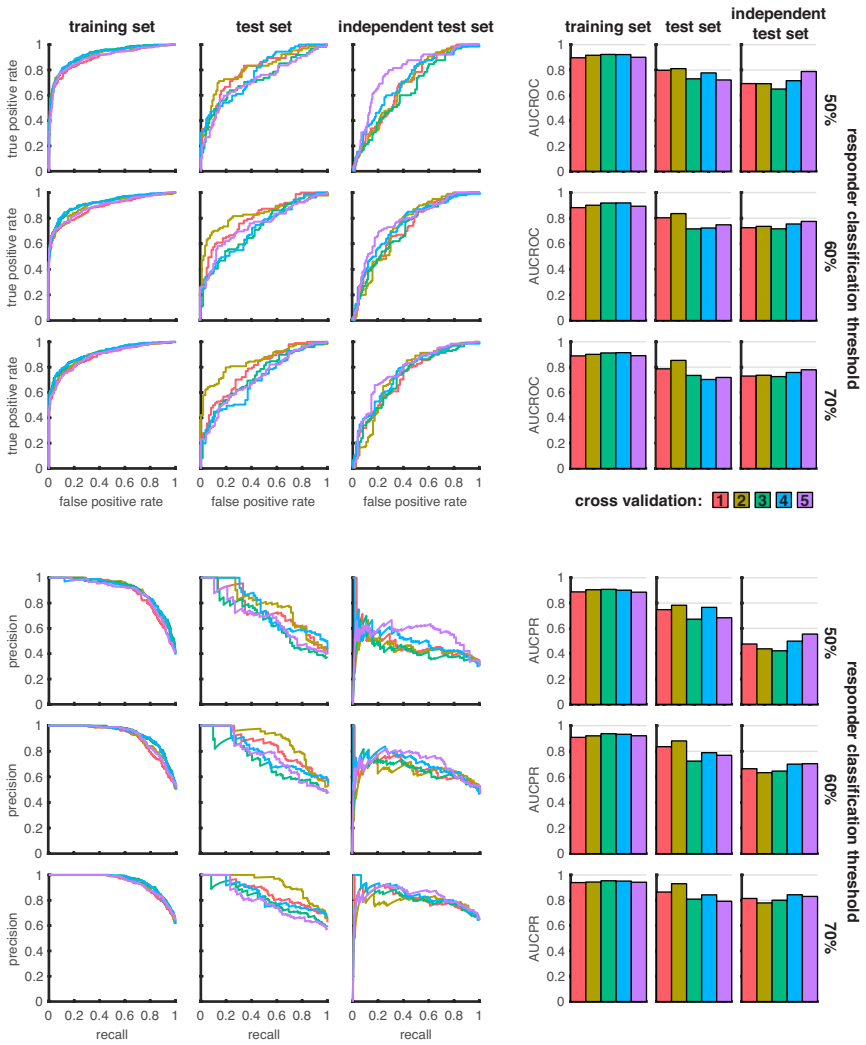
Supplementary Figure 2. Eigenvalues densities of the Fisher Information Matrix for parameters, state variables and proliferation readouts. Small eigenvalues correspond to large uncertainties of readout combinations defined by respective eigenvectors. One line for each of the best 5 optimization runs for each of the 5 cross validation is shown. All eigenvalues below 10^{-45} are not shown in the density plot but the corresponding fraction of eigenvalues is indicated in the barplot on the left.



Supplementary Figure 3. Correlation of model simulation and experimental data at all measured drug concentrations. The correlation for the drug concentrations from 250nM to 2530nM is higher than at 8000nM, which is likely due to an inflation of cell lines not responding to drugs (relative proliferation=1). For lower drug concentrations the correlation is lower than at 8000nM, which is due to the lower dynamic range of simulated and experimentally observed relative proliferation values.



Supplementary Figure 4. Overview of parameterization results. One column corresponds to an individual cell line. The cell lines are sorted according to the median over cross-validations of the maximal squared error at 8µM drug concentration shown in c). (a) Measured relative proliferation in response to the treatment with 8µM of the different drugs. (b) Gain-of-function mutations in the individual cell lines. Mutation status is summarized per gene and does not distinguish individual variants. (c) Boxplots of the maximal squared error at 8µM drug concentration over the 5 cross-validations. The squared error is evaluated for the median of the simulation from the 5 best optimization runs. The maximum is taken over all drugs. The boxplots are colored according to the tissue of origin of the cell lines.



Supplementary Figure 5. Receiver-operating-characteristic and precision-recall analysis for different datasets and classification thresholds. Every column corresponds to a dataset. Every row corresponds to a different value for the classification threshold. Colors indicate the 5 different cross-validations.

Online Methods

Model Development

The mechanistic model was developed using PyBioS[30], a web-based platform for modeling of complex molecular systems. We exploited several features of PyBioS, including the modular formulation of large-scale models based on individual pathways and their interactions. For model development we employed information from ConsensusPathDB[32]. The information was manually curated and implemented in the model using a standard operating procedure (SOP). The SOP ensured the model quality and the compatibility of different pathway models. The model structure was refined several times by different experts to ensure highest standards. To validate the model structure a plethora of logical test was used, e.g. to certify that the known effects of growth factor stimulations are correctly captured.

The developed model is made publicly available as supplement to this manuscript. The model features an exhaustive annotation, including UniProt and Ensembl IDs. Phosphorylations are indicated in the name of the species by a preceding "P[$X;$Y;…]-" where $X and $Y specify the phosphorylation site using a one letter amino acid code, followed by the amino acid number. Mutations are indicated in the name of the species by a preceding "MutAA[$Z]-" where $Z specifies the mutation site using standard sequence variant nomenclature. Homodimers are indicated by a trailing "[2x]".

The SBML file encodes an ordinary differential equation (ODE) model of the form

$$\frac{dx}{dt} = S \cdot v(x, \theta, d, c), \qquad x(0) = x_0,$$

with concentration vector x and its initial condition x_0, stoichiometric matrix S and flux vector v. The parameter vector θ provides the reaction rates, e.g. binding affinities. The vector d provides the drug concentrations used for simulation and the vector c provides the expression levels for the gene products and respective variants for a particular cell line. To consider different drug treatments and cell lines, only d and c need to be changed. The parameter vector is generic and transferable. The SBML model provides a representative parameter estimate.

In the SBML model the proliferation output variable is specified as an assignment rule. The proliferation output is computed as fraction of weighted sums of concentrations of active forms of transcription factors

$$y^{c,d} = \frac{\sum_i^N \omega_i^{pos} x_{i,c,d}^{pos}}{1+\sum_j^M \omega_j^{neg} x_{j,c,d}^{neg}},$$

in which $x_{i,c,d}^{pos}$ and $x_{i,c,d}^{neg}$ denote the concentrations of transcription factors, for a particular cell line and drug treatment combination, with a positive and negative influence on proliferation, respectively. The corresponding weights are denoted by ω_i^{pos} and ω_i^{neg} and were estimated during model parameterization. The model captures the effect of $\mathrm{N} = 12$ transcription factors with positive influence: P[S63;S73]-JUN[2x], P[S252;S265]-

FOSL1:P[S63;S73]-JUN, P[T69;T71]-ATF2:P[S63;S73]-JUN, P[S374;T325;T331]-FOS:P[S63;S73]-JUN P[Y701]-STAT1[2x], P[Y705]-STAT3[2x], P[Y694]-STAT5A[2x], P[Y699]-STAT5B[2x], MAX-001:P[S62]-MYCN, MAX:P[S62]-MYC, P[S324;S383]-ELK1, and P[S133]-CREB1. Furthermore, $M = 4$ transcription factors with negative influence were included: FOXO1, FOXO3, FOXO4 and FOXO6. In all cases only the species with nuclear localization were considered to be active.

The model employs experimentally derived drug-target binding affinity (Kd) values for the drugs CHIR-265, erlotinib, lapatinib, PLX4720, selumetinib, sorafenib and vandetanib, which were obtained from Davis *et al.*[53]. For PD0325901 the model employs the inhibitory concentrations (IC50), which was measured in a cell-free assay by Barrett *et al.*[54].

We note that the model includes several components that were not used in the presented analysis. This includes the option to specify gene specific scaling constants to individually adjust synthesis rates. Furthermore, the small molecular kinase inhibitor sorafenib was modeled. However, as none of the considered cell lines responded to sorafenib and as sorafenib targets several components that are not captured by the model, the corresponding response data was not considered in this study.

CCLE Data Processing

We downloaded RNAseq BAM-files for 780 CCLE cell lines from the Cancer Genomics Hub (https://cghub.ucsc.edu/) in April 2014. The same data, including additional cell lines is now available for download in the Cancer Genomics Cloud (https://cgc.sbgenomics.com/). The gene expression values were normalized as Reads Per Kilobase of transcript per Million (RPKM) using gene models from Ensembl Release 73. Mutation data was downloaded from the CCLE data portal (https://portals.broadinstitute.org/ccle/data/, file CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.05.07.maf). RNA allele frequencies for the mutations were determined from the downloaded RNAseq BAM-files using SAMtools mpileup (http://www.htslib.org). Drug response data were downloaded from https://portals.broadinstitute.org/ccle/downloadFile/DefaultSystemRoot/exp_10/ds_27/CCLE_NP24.2009_Drug_data_2015.02.24.csv?downloadff=true&fileId=20777.

Of the 780 cell lines, for which we processed RPKM values, 123 originated from the tissues breast, large-intestine, lung, pancreas and skin which were considered in the training/test data-set and 31 originated from the tissues kidney, soft tissue, ovary and stomach which were considered for the independent test set. For the training/test data we considered 120 of the 123 available cell lines to ensure equally sized training and test sets in all cross-validations.

To generate test and training datasets from the processed CCLE data, we performed 20-80% splits on the cell-line level, which yielded 5 training sets with 96 cell lines and test sets with 24 cell lines. The split was performed such that the tissue distribution in the individual training sets is maximally similar. The number of experimental condition in the training sets varies from 5390 to 5403 due to incomplete data for individual cell lines.

Numerical Simulation and Gradient Evaluation

The compilation and numerical simulation of the model was performed using the MATLAB toolbox AMICI[40] (http://dx.doi.org/10.5281/zenodo.579891). AMICI employs the backward differentiation method implemented in the SUNDIALS solver package[55]. We used the KLU linear solver with AMD reordering and relative and absolute error tolerance 10^{-8}. As the proliferation measurement was taken after 72 to 84 hours, we assumed that the state of the model reached a steady state. To find the steady state for the untreated condition of a cell line, the forward simulation was initialized with zero and run until the regularized maximal absolute relative derivative was smaller than 10^{-6},

$$\delta = \max_i \left| \frac{S \cdot v(x, \theta, d, c)_i}{x_i + 10^{-6}} \right| 10^{-6}.$$

For all treated conditions of a cell line, the forward simulation was initialized with the steady state of the corresponding untreated condition.

The objective function gradient was computed using adjoint sensitivity analysis[40]. As the model only possesses a single model output, the proliferation y, we computed the sensitivity of this output. From this output sensitivity, we computed the objective function gradient and the Fisher Information Matrix (FIM). The FIM is not accessible when adjoint sensitivity analysis is used to directly compute the objective function gradient.

The forward and backward simulation of experimental conditions was parallelized using the MATLAB command *parfor*, which implements OpenMP parallelization. As our cluster infrastructure features 8 core nodes, we parallelized each gradient computation over 8 cores (1 master, 7 workers), thereby avoiding inter-node communication overhead. The different local optimizations were performed on different nodes.

Numerical Benchmark

To compare different methods for gradient evaluation, we assessed the computation time for a single gradient evaluation on the full training set. For sequential and parallel gradient evaluation using adjoint sensitivity analysis, we measured the computation time. As this would have been too time consuming for forward sensitivities, we first assured that the computation time for individual experimental conditions is comparable and then extrapolated to all experimental conditions. The computation time was evaluated on the training set of the first cross-validation for 10 randomly sampled parameter vectors. For the difference between forward and adjoint sensitivity analysis and sparse and dense solvers, we only evaluated the simulation time for the untreated condition of a single cell-line. The performance was evaluated based on 100 samples with a randomly drawn parameter vector and a randomly drawn cell-line. The computation time was then normalized such that the median for the sparse adjoint approach matched the computation time for the full training set.

Parameterization

To estimate the model parameters θ, we used the measurement data for the proliferation in the treated condition relative to the untreated condition, $y_m^{c,d}/y_m^c$, provided in the CCLE dataset. These data were fitted using a sum-of-squared-residuals objective function

$$J(\theta) = \sum_{c \in C} \sum_{d \in D_c} \left(\frac{y_m^{c,d}}{y_m^{c,0}} - \frac{y^{c,d}(\theta)}{y^{c,0}(\theta)} \right)^2,$$

in which $c \in C$ is cell-line specific and $d \in D_c$ denotes the drug treatment. This objective function is equivalent to the negative log-likelihood function under the assumption of additive independent and identically distributed standard normally distributed measurement noise. To minimize the objective function we used multi-start local optimization[26] implemented in the MATLAB toolbox PESTO (http://dx.doi.org/10.5281/zenodo.579890). Parameters were constrained to a $[10^{-2},10^2]$ hypercube. For each local optimization run, parameters were drawn from this hypercube, followed by 100 optimization iterations of the MATLAB fmincon interior-point algorithm. The gradient of the objective function was computed from adjoint sensitivities and supplied to the interior-point algorithm. A high-performance-computing-ready standalone executable was generated from the parameterization pipeline implemented in MATLAB using the MATLAB Compiler toolbox. For every cross-validation we performed 10 local optimization runs. As no communication was necessary between optimization runs, each could be submitted as a separate job to the cluster. In total we submitted 50 jobs using 8 cores each, resulting in a total parallelization over 400 cores.

Ensemble Averaging

We used ensemble averaging to reduce the effect of overfitting and the variance of predictors. For the mechanistic model we used an ensemble model based on five optimization runs that achieved the lowest objective function value. The parameter values from these optimization runs were then used to simulate the model individualized to cell lines from the test and independent test set. For quantitative predictions the median of the five simulations was used and for classification a majority vote was used. The ensemble averaging was solely based on results from the training set. The test and independent test set were only used for validation.

Uncertainty Analysis

We assessed the uncertainty of parameters using the eigenvalue spectrum of the Fisher Information Matrix (FIM). Small eigenvalues indicate large uncertainties in the direction of the respective eigenvector while large eigenvalues indicate small uncertainties. The eigenvalue spectrum was evaluated for the best 5 optimization runs for every cross-validation.

The FIM was computed by summing the dyadic product of adjoint sensitivities over all experimental conditions

$$FIM_\theta = \sum_{c\in C}\sum_{d\in D_c}\frac{1}{y^c(\theta)^4}\left(\frac{\partial y^{c,d}(\theta)}{\partial\theta}y^c(\theta) - \frac{\partial y^c(\theta)}{\partial\theta}y^{c,d}(\theta)\right)^T\left(\frac{\partial y^{c,d}(\theta)}{\partial\theta}y^c(\theta) - \frac{\partial y^c(\theta)}{\partial\theta}y^{c,d}(\theta)\right).$$

As the number of experimental conditions (~5.400) exceeds the number of parameters (~4,100) the FIM could theoretically have full rank.

For parameter derived readouts z, such as proliferation readouts as well as state variables, a similar quantification of the uncertainty is possible by considering a transformation FIM_z of the FIM. The transformation is obtained by multiplication with the respective parameter derivatives

$$FIM_z = \frac{\partial z}{\partial\theta}FIM_\theta\frac{\partial z}{\partial\theta}^T.$$

For state variables the formula for steady state parameter derivatives were computed according to the implicit function theorem

$$\frac{\partial x}{\partial\theta} = -\left(S\cdot\frac{\partial v}{\partial x}\right)^{-1}S\cdot\frac{\partial v}{\partial\theta},$$

assuming that the system is in steady state,

$$S\cdot v = 0.$$

For the FIM_z for state variables, we only considered state variables with non-zero simulated steady state. The state variables with steady state equal to zero correspond to molecular species that are not expressed.

Statistical Methods

For the comparison of model performances, we trained a series of statistical models for the prediction of response to treatment, based on the exact same training data sets, cross-validation setup and test data sets that were used for the mechanistic model. Responder and non-responder cell lines were defined for each drug by applying the threshold 0.5 on the proliferation at the highest dose used for the drug (in addition results for thresholds 0.7 and 0.9 were also obtained). The classification of cell lines into responders and non-responders was evaluated for three sets of input variables: 1) mutation genotype data; 2) gene expression data; and 3) genotype and gene expression data. In addition we also provided the network topology as input to some of the classifiers. Model training was performed by nested cross-validation, where the outer 5-fold cross validation loop split the data into training set (80% of the data) and test set (20% of the data). For each classifier and each training set we estimated the model parameters by optimizing the classification performance in the inner cross validation loop splitting the training data again into training and validation sets with balanced class labels or by bootstrapping of the training data.

We used the R implementations of the following classifiers: 1) logistic regression with LASSO penalty (glmnet package); 2) Random Forest (randomForest package); 3) graph regularized logistic regression (glmgraph package); and 4) logistic regression with LASSO

penalty on augmented data, including additional interaction terms. The interaction terms were defined based on the network topology used in the mechanistic model. The adjacency matrix was extracted from the Jacobian of the right hand side of the differential equation. For a pair of genes the Jacobian was reduced to rows or columns corresponding species that include the corresponding proteins in any form (phosphorylated, cleaved or bound). Two genes were defined to be adjacent when the corresponding submatrix of the Jacobian has at least one non-zero entry. We augmented the data set either with all pairwise interactions (product) between variables of the same type (genotype or gene expression) or interactions between genes that are connected by paths in the network not longer than 1, 2 or 3 steps. The optimal parameter λ for LASSO models, λ_1 and λ_2 for the graph regularized LASSO model were selected as the largest λ that produces an AUC ROC within one standard error of the maximum AUC ROC[56] in a 8-fold inner cross validation. Random Forest classifiers were trained by selecting parameters that minimize the out of bag error[43]. We optimized over the number of variables randomly sampled as candidates for each split, the number of trees in the forest ranging from 50 to 500 and the maximum leaf node size criterion ranging from 1 to 1/3 of the data set. All classifiers were then applied to the test set and performance was assessed as the area under the ROC curve.

Finally, we used the classifiers trained in each round of the cross validation and applied them to an independent test set. Performance was assessed as the area under the ROC curve and averaged over the five cross validation rounds.

Validation

Area Under the Curve: For the quantification of the classification accuracy for the mechanistic model we computed receiver-operating-characteristic (ROC) curves and precision-recall (PR) curves. For the data and model, a cell line was classified as responder to a drug when the measured/simulated relative proliferation at 8μM was smaller or equal to the specified threshold. For the data, this threshold was fixed to 0.5 and for the model the threshold was continuously varied between 0 and 1 and the precision, sensitivity and specificity was evaluated for every threshold value. From these evaluations the ROC and PR curves were constructed and the area under the curve was evaluated using a trapezoidal rule.

Exactly the same split in training and test set was applied for all statistical models and the mechanistic model. For both approaches the test and independent test sets were never used for parameterization. For the statistical model the AUC was averaged over drugs and cross-validation as individual models were constructed for every drug. For the mechanistic model the AUC was averaged only over cross-validations as a single model for all drugs could be constructed.

Combination Therapy: The data by Friedman *et al.*[46] includes measurements for single and paired treatments at low and high concentrations for the drugs selumetinib, CHIR-265, erlotinib, lapatinib and PLX4720. For the paired treatment, the experiment was repeated twice and the average value was used for our analysis. As the treatment concentrations employed in the two studies did not agree, we interpolated measurements from the CCLE data to concentrations employed in the Friedman dataset. The interpolation was performed

in logarithmic concentration space but results were comparable for linear concentration space. For every cross validation we only considered the measurements from cell lines which were also contained in the training set.

Proteomics Measurements: For the comparison of model prediction and RPPA data from the MD Anderson Cell Lines Project, we computed the median total protein concentrations for the 5 best optimization runs. To compute total protein concentrations, we computed the sum of concentrations of all protein and complex species that contain a specific protein, taking into account the individual stoichiometry. As the RPPA data in the MD Anderson Cell Lines Project are normalized, we also normalized the simulation data using the same procedure: 1) The simulated concentrations were $\log_2$ transformed; 2) The median concentration over all cell lines was subtracted for every protein and then the median of all concentrations was subtracted for every cell-line. While this procedure resembles the procedure described in the manuscript providing the data, we note that the model only captures a fraction of the proteins and the normalization might therefore be suboptimal. For the RPKM data, only a $\log_2$ transformation was applied. The correlation coefficients of normalized data and predictions were computed for all proteins that were measured in at least 10 cell lines. For every cross validation we only considered the measurements from cell lines which were also contained in the training set.

The considered dataset only provides measurements for the untreated condition. However, the parameterization was performed based on relative proliferation values, which is computed based on the ratio of sums of active transcription factor concentrations in the model. Accordingly, the training data provides some information about the ratio of concentrations in treated and untreated condition, but little information about absolute concentrations in the treated and untreated condition.

Bibliography

1. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32,** 1202–1212 (2014).
2. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
3. Kim, J. & Schöberl, B. Beyond static biomarkers – The dynamic response potential of signaling networks as an alternate biomarker? *Sci. Signal.* **8,** fs21 (2015).
4. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45,** 1127–1133 (2013).
5. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439,** 353–357 (2006).
6. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10,** 1108–1115 (2013).
7. Schöberl, B. *et al.* Therapeutically targeting ErbB3: A key node in ligand-induced activation of the ErbB receptor-PI3K axis. *Sci. Signal.* **2,** ra31 (2009).

8. Fey, D. *et al.* Signaling pathway models as biomarkers: Patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci. Signal.* **8,** ra130 (2015).
9. Eduati, F. *et al.* Drug Resistance Mechanisms in Colorectal Cancer Dissected with Cell Type–Specific Dynamic Logic Models. *Cancer Res.* **77,** 3364 (2017).
10. Weinstein, I. B. & Joe, A. K. Mechanisms of Disease: oncogene addiction – a rationale for molecular targeting in cancer therapy. *Nat. Clin. Pract. Oncol.* **3,** 448–457 (2006).
11. Kaelin, William G. The Concept of Synthetic Lethality in the Context of Anticancer Therapy. *Nat. Rev. Cancer* **5,** 689–698 (2005).
12. Zhang, J., Yang, P. L. & Gray, N. S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **9,** 28–39 (2009).
13. Li, C. *et al.* BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst. Biol.* **4,** 92 (2010).
14. Network, C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
15. Wu, P., Nielsen, T. E. & Clausen, M. H. Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. *Drug Discov. Today* **21,** 5–10 (2016).
16. Kuperstein, I. *et al.* Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* **4,** e160 (2015).
17. Mizuno, S. *et al.* AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. *BMC Syst. Biol.* **6,** 52 (2012).
18. Swainston, N. *et al.* Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics* **12,** 109 (2016).
19. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39,** D685–D690 (2011).
20. Turei, D., Korcsmaros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat. Methods* **13,** 966–967 (2016).
21. Stanford, N. J. *et al.* Systematic construction of kinetic models from genome-scale metabolic networks. *PLoS ONE* **8,** e79195 (2013).
22. Smallbone, K. & Mendes, P. Large-scale metabolic models: From reconstruction to differential equations. *Ind. Biotechnol* **9,** 179–184 (2013).
23. Büchel, F. *et al.* Path2Models: large-scale generation of computational models from biochemical pathway maps. *BMC Syst. Biol.* **7,** (2013).
24. Molinelli, E. J. *et al.* Perturbation biology: Inferring signaling networks in cellular systems. *PLoS Comput. Biol.* **9,** e1003290 (2013).
25. Halasz, M., Kholodenko, B. N., Kolch, W. & Santra, T. Integrating network reconstruction with mechanistic modeling to predict cancer therapies. *Sci. Signal.* **9,** (2016).
26. Raue, A. *et al.* Lessons learned from quantitative dynamical modeling in systems biology. *PLoS ONE* **8,** e74335 (2013).
27. Villaverde, A. F. *et al.* BioPreDyn-bench: a suite of benchmark problems for dynamic modelling in systems biology. *BMC Syst Biol* **9,** (2015).
28. Jagiella, N., Rickert, D., Theis, F. J. & Hasenauer, J. Parallelization and high-performance computing enables automated statistical Inference of multiscale models. *Cell Syst.* **4,** 194–206 (2017).
29. Wierling, C. *et al.* Network and systems biology: essential steps in virtualising drug discovery and development. *SI Netw.-Based Discov. Syst. Biol.* **15,** 33–40 (2015).

30. Wierling, C., Herwig, R. & Lehrach, H. Resources, standards and tools for systems biology. *Brief Funct Genomic Proteomic* **6,** 240–51 (2007).
31. Klipp, E., Herwig, R., Kowald, A., Wierling, C. & Lehrach, H. *Systems biology in practice*. (Wiley-VCH, Weinheim, 2005).
32. Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat. Protoc.* **11,** 1889–1907 (2016).
33. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34,** D668–D672 (2006).
34. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38,** D355–D360 (2010).
35. Croft, D. *et al.* Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* **39,** D691–697 (2011).
36. Rubio-Perez, C. *et al.* In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27,** 382–396 (2015).
37. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483,** 603–607 (2012).
38. Hross, S. & Hasenauer, J. Analysis of CFSE time-series data using division-, age- and label-structured population models. *Bioinformatics* **32,** 2321–2329 (2016).
39. Davis, T. A. & Palamadai Natarajan, E. Algorithm 907: KLU, a direct sparse solver for circuit simulation problems. *ACM Trans. Math. Softw. TOMS* **37,** 36 (2010).
40. Fröhlich, F., Kaltenbacher, B., Theis, F. J. & Hasenauer, J. Scalable parameter estimation for genome-scale biochemical reaction networks. *PLoS Comput. Biol.* **13,** 1–18 (2017).
41. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12,** (2016).
42. Venetsanakos, E. *et al.* CHIR-265, a novel inhibitor that targets B-Raf and VEGFR, shows efficacy in a broad range of preclinical models. *Cancer Res.* **66,** 1140 (2006).
43. Breiman, L. Random Forests. *Mach. Learn.* **45,** 5–32 (2001).
44. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Sociecte B* **58,** 267–288 (1996).
45. Chen, L., Liu, H., Kocher, J.-P. A., Li, H. & Chen, J. glmgraph: an R package for variable selection and predictive modeling of structured genomic data. *Bioinformatics* **31,** 3991–3993 (2015).
46. Friedman, A. A. *et al.* Landscape of Targeted Anti-Cancer Drug Synergies in Melanoma Identifies a Novel BRAF-VEGFR/PDGFR Combination Treatment. *PLoS ONE* **10,** e0140310 (2015).
47. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504,** 389–393 (2013).
48. Hafner, M., Niepel, M., Chung, M. & Sorger, P. K. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat. Methods* **13,** 521–527 (2016).
49. Li, J. *et al.* Characterization of Human Cancer Cell Lines by Reverse-phase Protein Arrays. *Cancer Cell* **31,** 225–239
50. Waclaw, B. *et al.* A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* **525,** 261–264 (2015).

51. Edginton, A. N., Theil, F. P., Schmitt, W. & Willmann, S. Whole body physiologically-based pharmacokinetic models: their use in clinical drug development. *Expert Opin. Drug Metab. Toxicol.* **4,** 1143–1152 (2008).
52. Lehrach, H. Virtual clinical trials, an essential step in increasing the effectiveness of the drug development process. *Public Health Genomics* **18,** 366–371 (2015).
53. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29,** 1046–1051 (2011).
54. Barrett, S. D. *et al.* The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901. *Bioorg. Med. Chem. Lett.* **18,** 6501–6504 (2008).
55. Hindmarsh, A. C. *et al.* SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM Trans. Math. Softw.* **31,** 363–396 (2005).
56. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of statistical learning*. **2,** (Springer, 2009).

Appendix D

Parameter estimation for dynamical systems with discrete events and logical operations.