Computergestützte Untersuchung stochastischer biochemischer Reaktionssysteme

Dissertation
zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing)

von Dipl.-Phys. Dennis Pischel
geb. am 06.03.1989 in Magdeburg
genehmigt durch die Fakultät für Verfahrens-und Systemtechnik
der Otto-von-Guericke-Universität Magdeburg

Promotionskommission:	Prof. Dr. rer. nat. habil. Dieter Schinzer	(Vorsitz)
	Prof. Dr.-Ing. habil. Kai Sundmacher	(Gutachter)
	Prof. Dr.-Ing. Robert J. Flassig	(Gutachter)
	Prof. Dr. rer. nat. Wilhelm Huisinga	(Gutachter)

eingereicht am: 06.05.2019
Promotionskolloquium am: 20.12.2019

Abstract

Biochemical reaction systems can be understood as complex reaction networks, which facilitate various life essential processes, *e.g.* metabolism, signal transduction and cell communication. Furthermore, they enable the synthesis of organic products *via* biochemically active substances in bioreactors. Generally, biochemical reaction systems are characterized by strong nonlinearities. These can lead to significant changes regarding the dynamical behaviour in presence of perturbations of system parameters. The thereby originated variability and heterogeneity represent fundamental biological properties, which can be observed in numerous experiments. Especially on the cellular level the individual character turns out to be exceptionally dominant. Thus, single cells within a population differ in all kinds of attributes including cell cycle state, morphology and cellular compounds. Striking examples of this phenomenon are *i.a.* the stochastic occurence of cell death, the asymmetric devision of a mother cell into unequal daughter cells and the formation of inhomogeneous cancer tissue from tumor stem cells. For real-world applications biological variability constitutes a major challenge, since it goes along with the optimization of bioprocesses concerning a diverse spectrum of cells. In this context the term bioprocess is not limited to industrial applications, but also comprises medicamentous treatments or natural biochemical occurrences in organisms and their surroundings. Concerning the optimal control and design of bioprocesses stochastic effects must be considered in order to increase their quality, yield and outcome. This applies to experimental measurements, data analysis and mathematical modeling.

During the last decades immense progress regarding the measurements of single cells was achieved, which enabled the detailed observation of biological variability. In contrast to bulk population measurements, single cell measurements allow to track a cell population or single cells through time instead monitoring solely the population mean. The advances in data curation go along with increasing data size and complexity leading to so called Big Data. The analysis and interpretation of Big Data unlocks great potential to gain deep insight into biological processes. Furthermore, mathematical modeling of single cells or cell populations enables the comparison of theoretical and experimental studies, which is only partly possible by simulating the population

mean.

This work focuses on the theoretical aspects of integrating biological variability into mathematical modeling and data analysis. First, a novel algorithm to efficiently simulate the impact of different sources of noise is presented. The algorithm captures stochasticity from intrinsic probabilistic chemical reactions with uncertain model parameters. By combining the Sigma Point method with the Gillespie algorithm an approximate solution of the chemical master equation is derived. This method is benchmarked on several example systems and successfully used to calibrate a complex apoptosis model utilizing western blot and imaging flow cytometry data. The apoptosis model gives new insight regarding the ambivalent cellular decision between life and death, which leads to a novel mechanism determining cell fate. Furthermore, an automated machine learning workflow is established to distinguish cells with different phenotypes based on label-free imaging flow cytometry data. In contrast, traditional approaches rely on two-dimensional gating methods based on fluorescence staining. These approaches are characterized by manual adjustment of borders to separate subpopulations belonging to different phenotypes. Thereby valuable information that is hidden in the enormous amount of data is ignored. The proposed method is benchmarked on an apoptosis scenario and turns out to yield a similar accuracy while showing a number of advantages compared to the traditional gating approach.

Zusammenfassung

Biochemische Reaktionssysteme können als komplexe Reaktionsnetzwerke verstanden werden, die überlebenswichtige Prozesse biologischer Organismen, wie den Metabolismus, die Signaltransduktion und die Zellkommunikation, ermöglichen. Darüber hinaus erlauben sie die Herstellung organischer Produkte aus biochemisch aktiven Substanzen in Bioreaktoren. In der Regel sind biochemische Reaktionssysteme durch starke Nichtlinearitäten gekennzeichnet, weshalb Störungen der Systemparameter oft signifikante Veränderungen der Dynamik hervorrufen können. Die dadurch verursachte Variabilität und Heterogenität stellen fundamentale Eigenschaften biologischer Systeme dar, die in einer Vielzahl von Experimenten belegt worden sind. Von besonderer Dominanz ist dieses Phänomen auf der Ebene einzelner Zellen. Diese unterscheiden sich in der Gesamtheit ihrer Merkmale und demonstrieren damit, dass der individuelle Charakter zellulärer Systeme sehr ausgeprägt ist. Prominente Beispiele sind unter anderem das zufällige Eintreten des Zelltodes, die asymmetrische Teilung einer Mutterzelle in ungleiche Tochterzellen sowie die Bildung von inhomogenem Krebsgewebe aus Tumorstammzellen. Für praktische Anwendungen stellt die biologische Variabilität eine große Hürde dar, da sie im Allgemeinen mit der Optimierung von Bioprozessen hinsichtlich eines Spektrums verschiedenartiger Zellen einhergeht. In diesem Zusammenhang soll der Begriff des Bioprozesses nicht ausschließlich im Kontext industrieller Anwendungen betrachtet werden, sondern er soll auch medikamentöse Behandlungen oder natürliche biochemische Abläufe in Organismen sowie deren Umgebung einschließen. Hinsichtlich der optimalen Steuerung und Auslegung realer Bioprozesse müssen demnach stochastische Effekte berücksichtigt werden, um beispielsweise die Produktqualität und -ausbeute zu erhöhen oder die Wirksamkeit medizinischer Therapien zu steigern. Dies gilt für experimentelle Messungen, die Analyse experimenteller Daten und deren mathematische Modellierung.

Ein großer technischer Fortschritt gelang mit der Entwicklung von Verfahren zur Einzelzellmessung, die es ermöglichen, biologische Variabilität zu beobachten. Einzelzellmessungen erlauben es, die Dynamik von Zellpopulationen oder einzelner Zellen zu verfolgen. Im Gegensatz dazu ist es mit Bulk-Populationsmessung lediglich möglich, den Populationsmittelwert zu beobachten. Der Fortschritt der Datenerhebung geht

einher mit einem Anstieg des Datenumfangs und der Datenkomplexität, was letztendlich zu „Big Data“ führt. Die Analyse und Interpretation von Big Data ist ein vielversprechender Ansatz, der das Potential besitzt, einen tieferen Einblick in biologische Prozesse zu gewinnen, um bisher unbekannte Muster und Strukturen zu erkennen. Darüber hinaus erlaubt die mathematische Modellierung einzelner Zellen oder Zellpopulationen die Gegenüberstellung theoretischer und experimenteller Ergebnisse, was mittels der Simulation des Populationsmittelwertes nur teilweise realisierbar ist.

Diese Arbeit konzentriert sich auf theoretische Aspekte, die biologische Variabilität in die wissenschaftliche Forschung integrieren, nämlich mathematische Modellierung und Datenanalyse. Dazu wird zunächst ein neuartiger Algorithmus vorgestellt, um die Auswirkungen verschiedener Störungen effizient zu simulieren. Der Algorithmus erfasst Stochastizität von intrinsisch probabilistischen chemischen Reaktionen mit unsicheren Modellparametern. Durch die Kombination der Sigma-Punkt-Methode mit dem Gillespie-Algorithmus wird eine approximative Lösung der chemischen Mastergleichung abgeleitet. Diese Methode wird an mehreren Beispielsystemen getestet und erfolgreich zur Kalibrierung eines komplexen Apoptosemodells verwendet. Das Apoptosemodell gibt neue Einblicke in die ambivalente zelluläre Entscheidung zwischen Leben und Tod, woraus ein neuartiger Mechanismus abgeleitet wird, der das Schicksal der Zelle bestimmt. Darüber hinaus wird ein automatisiertes Machine-Learning-Verfahren vorgeschlagen, um Zellen unterschiedlicher Phänotypen zu unterscheiden. Das Verfahren beruht auf markierungsfreien Eigenschaften, gemessen mit bildgebender Flusszytometrie. Im Gegensatz dazu basieren traditionelle Methoden vorrangig auf zweidimensionalen Gating-Verfahren, die mittels Fluoreszenzfarbstoffen gemessen werden. Diese sind stark durch die händische Anpassung von Grenzen geprägt, die Zellen unterschiedlicher Phänotypen separieren. Darüber hinaus ignorieren sie wertvolle Informationen, die in den komplexen, hochdimensionalen Daten verborgen sind. Die in dieser Arbeit vorgestellte Methode wird an einem Apoptoseszenario getestet. Dabei stellt sich heraus, das eine mit der traditionellen Methode vergleichbare Genauigkeit erreicht wird. Zudem ergeben sich eine Reihe von Vorteilen gegenüber dem traditionellen, zweidimensionalen Gating-Ansatz.

Inhaltsverzeichnis

1 Einleitung

1.1 Ziel der Arbeit

Die Systembiologie ist ein junger Zweig der Naturwissenschaften, der in den letzten Jahrzehnten mehr und mehr an Bedeutung dazugewann [Chuang et al., 2010]. Mit der Absicht biologische Systeme als Gesamtheit zu begreifen, beschäftigt sie sich mit der Synergie von Prozessen verschiedenster Skalen: vom Genom und dem Proteom über den Metabolismus bis hin zum Organismus. Daraus ergibt sich ein integriertes Bild, das Aufschluss über fundamentale biologische Gesetzmäßigkeiten sowie deren Wechselwirkung und Kopplung auf der Systemebene liefert [Kitano, 2004]. Biologische Systeme sind in der Regel hochgradig komplex und aus einer Vielzahl von multifunktionalen Elementen aufgebaut, die in einer komplizierten Weise nichtlinear und selektiv miteinander interagieren [Kitano, 2002; Wilkinson, 2009]. Aus diesem Grund ist es allein mit experimentellen Mitteln nicht möglich, die Funktionsweise der einzelnen Elemente und deren Zusammenspiel zu verstehen. Erst durch die Kombination experimenteller Methoden mit Computersimulationen können biologische Modelle quantitativ an experimentelle Daten angepasst werden [Jaqaman et al., 2006], neue Experimente geplant werden, um konkurrierende Modelle abzuwägen [Fedorov, 2010] oder Vorhersagen in Situationen getroffen werden, die experimentell nicht durchführbar sind [Klipp et al., 2016].

Die Modellierung biologischer Systeme ist vielfältig und kann mittels verschiedener Ansätze erfolgen [Chuang et al., 2010], die in Abhängigkeit davon gewählt werden, welche Systemeigenschaften vorliegen und welche Fragestellung zu beantworten ist:

- gewöhnliche Differentialgleichung → deterministische Systeme
- Boolesche Netzwerke → logische Systeme
- stochastische Differentialgleichung → probabilistische Systeme

Damit wird die Biologie zu einer quantitativen Wissenschaft, die stark durch Konzepte der Physik, Chemie und Ingenieurwissenschaften geprägt ist [Klipp et al., 2016] und es somit erlaubt, biologische Systeme auf einer fundamentalen Ebene zu erfassen, um

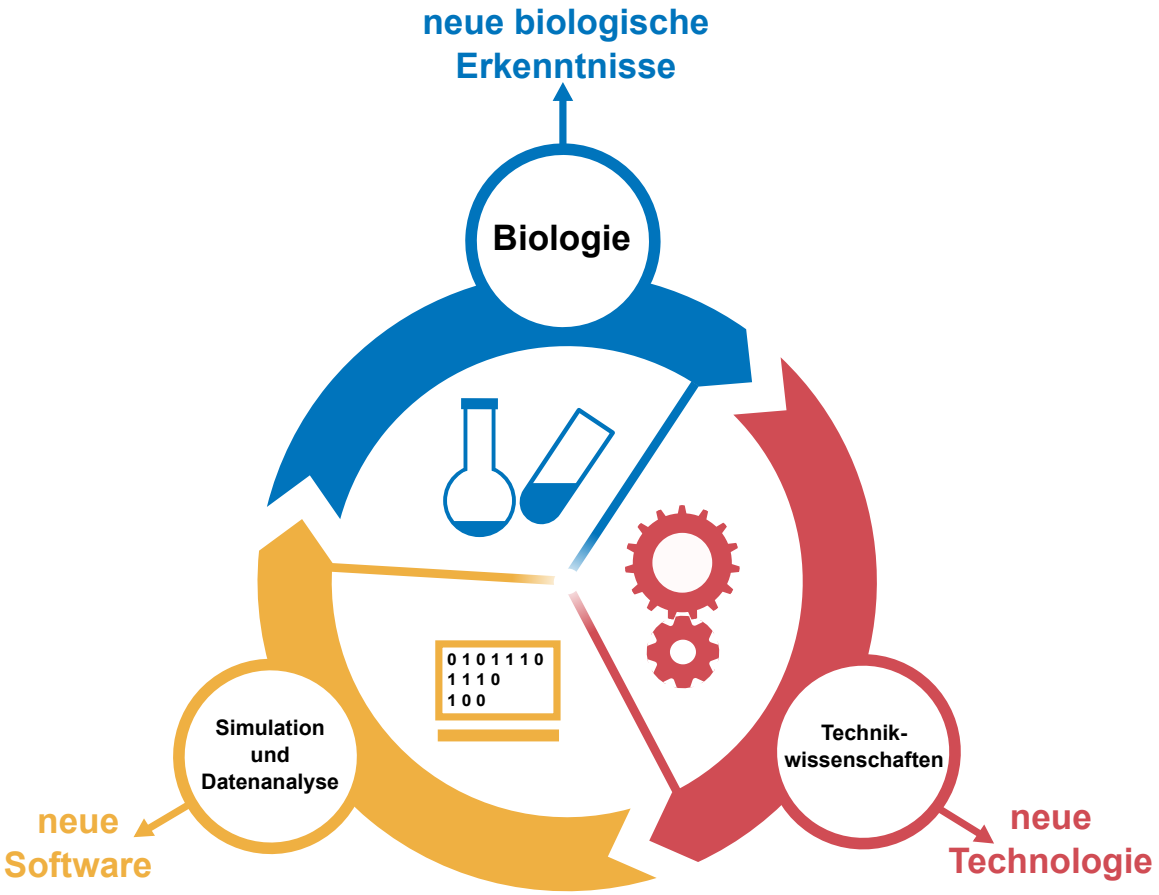

Abbildung 1.1: Die Systembiologie als interdisziplinäre Wissenschaft. Die Beantwortung biologischer Fragestellungen treibt die Entwicklung der Messtechnik voran, die wiederum neue, effiziente Algorithmen zur Simulation und Datenauswertung benötigt. Mittels Simulationen können dann neue biologische Hypothesen aufgestellt werden, die anschließend validiert werden müssen.

tieferes Verständnis zu erzeugen [Schrödinger, 1944; Lazebnik, 2002]. Der interdisziplinäre Exkurs führt dazu, dass nicht nur die Biologie voranschreitet, sondern auch technische Disziplinen, siehe Abb. 1.1. Neue biologische Erkenntnisse gehen deshalb oft mit der Weiterentwicklung von Messmethodiken und experimenteller Apparaturen einher. Diese generieren im Gegenzug größere Mengen an Daten oder geben Aufschluss über bisher nicht detektierbare Signale. Zur Auswertung dieser Messungen werden wiederum neue Algorithmen benötigt, die größeren Datenmengen effektiv verarbeiten oder neue beobachtbare Phänomene detailliert beschreiben. Ein eindrucksvolles Beispiel dafür stellt die rasche Entwicklung der Technologie zur Hochdurchsatzmessung von Einzelzellen dar. Dieser technische Fortschritt ermöglichte nicht nur die Untersuchung der Variabilität und Heterogenität von Zellpopulationen, sondern verhalf der stochastischen Modellierung und automatisierten Datenanalyse zu großer Popularität.

Die Zukunftsvision besteht darin mittels der Systembeschreibung die Biologie noch weiter voranzutreiben, um beispielsweise robustere und ertragreichere Nutzpflanzen zu entwerfen [Jogaiah et al., 2013], medizinische Behandlungen individuell auf jeden Patienten abzustimmen [van der Greef et al., 2006] oder mikrobielle Stämme für industrielle Anwendungen zu optimieren [Wang et al., 2016]. Diese Vorhaben sind teilweise schon umgesetzt und erzielen vielversprechende Ergebnisse. Im Gegensatz dazu steckt die biotechnologische Nutzung synthetischer Organismen noch immer in den Kinderschuhen [Rollié et al., 2012]. Die Ursache dafür liegt in der schieren Komplexität und Variabilität biologischer Systeme, die Experimente, deren Auswertung sowie Computersimulationen erschweren [Macklin et al., 2014]. Fokussiert auf die effiziente Simulation stochastischer biologischer Systeme und die automatisierte Charakterisierung heterogener Zellpopulationen, zielt diese Arbeit darauf ab, einige der genannten rechentechnischen Hürden zu überwinden. Um die Effektivität der im Rahmen dieser Arbeit entwickelten Methoden zu demonstrieren, werden diese zur Analyse und Charakterisierung von Apoptose in HeLa-CD95 Zellen genutzt. Aufbauend auf den neuen Methoden können biologisch relevante Fragestellungen beantwortet werden, die ein tieferes Verständnis fundamentaler biologischer Prozesse ermöglichen. Damit belegt diese Arbeit den facettenreichen Charakter der Systembiologie und verdeutlicht anschaulich das Potenzial interdisziplinärer Zusammenarbeit.

1.2 Inhalt und Gliederung der Arbeit

Diese Arbeit beschäftigt sich mit der Charakterisierung stochastischer biochemischer Reaktionssysteme. Biologische Systeme sind von hoher Variabilität geprägt, die durch verschiedene Störungseinflüsse verursacht wird. Besonders eindrucksvoll ist dies in heterogenen Differenzierungsprozessen zu erkennen, wie zum Beispiel der Apoptose, siehe Abb. 1.2. Der Fokus wird in dieser Arbeit vor allem auf die Simulation stochastischer Systeme sowie die Auswertung von Hochdurchsatzmessung einzelner Zellen gelegt. Durch deren Kombination ist es möglich, hinsichtlich des dynamischen Prozesses der Apoptose neue Hypothesen aufzustellen und biologisch relevante Fragestellungen zu beantworten. Im Folgenden ist eine kurze Gliederung der Arbeit gegeben, die diese Punkte aufgreift:

Kapitel 2 - Mathematische Modellierung biochemischer Systeme Zunächst wird ein Überblick hinsichtlich der Modellierung störungsfreier biochemischer Reaktionssysteme ohne räumliche Konzentrationsgradienten gegeben. Dabei wird darauf eingegangen, wann diese Modellierung adäquat ist und welche Stärken bzw. Schwächen sie besitzt. Anschließend wird demonstriert, dass biochemische Systeme stark von

Variabilität und Heterogenität geprägt sind. Zur Erweiterung auf stochastische biochemische Reaktionssysteme werden extrinsische, intrinsische und externe Störungen eingeführt. Die Simulation stochastischer Reaktionssysteme ist sehr aufwendig, weshalb für Optimierungszwecke oft approximative Methoden verwendet werden. Verschiedene approximative Methoden zur Simulation dieser Störungen werden vorgestellt. Es zeigt sich dabei, dass die simultane Simulation verschiedener Störungseinflüsse in der Literatur nur selten behandelt wird. Aus diesem Grund ist ein neuer Algorithmus entwickelt worden, der effizient verschiedene Störungseinflüsse simuliert. Der Algorithmus ist an verschiedenen Modellsystemen getestet worden und liefert exzellente Ergebnisse bei Parameteroptimierungsproblemen. Der Abschnitt über den neuen Algorithmus basiert auf zwei eigenen Publikationen [Pischel et al., 2016, 2017].

Kapitel 3 - Der stochastische Prozess der Apoptose Aufbauend auf dem vorigen Kapitel, wird der entwickelte Algorithmus genutzt, um ein komplexes Apoptosemodell anhand von experimentellen Daten, gemessen mit Western Blots und bildgebender Flusszytometrie, zu kalibrieren[1]. Das kalibrierte Modell wird anschließend zur Simulation der Dynamik von Schlüsselmolekülen des pro- und antiapoptotischen Pfadwegs genutzt. Basierend auf den Simulationsergebnissen wird ein neuartiger Mechanismus postuliert, der die ambivalente Entscheidung einzelner Zellen zwischen Leben und Tod bestimmt. Zur Überprüfung des Mechanismus werden Experimente mit Inhibitionen des pro- und antiapoptotischen Pfadweges getätigt. Dabei zeigt sich eine qualitative Übereinstimmung von Theorie und Experiment. Dieses Kapitel basiert auf einer weiteren eigenen Publikation [Buchbinder et al., 2018].

Kapitel 4 - Zelldiskriminierung mittels Machine-Learning Die bildgebende Flusszytometrie stellt eine Messmethodik dar, die den Hochdurchsatzcharakter des Flusszytometers mit dem Auflösungsvermögen eines Mikroskops kombiniert. Aus den Hochdurchsatzdaten (Big Data) lässt sich eine große Menge an Information extrahieren. Manuelle Methoden stoßen bei der Interpretation dieser Daten schnell an ihre Grenzen, weshalb automatisierte Machine-Learning-Algorithmen zur Datenverarbeitung stark an Popularität zugenommen haben. In diesem Kapitel wird eine einfache, modulare Vorgehensweise entwickelt, die in der Lage ist, die großen Datenmengen effizient zu verarbeiten. Dazu wird die Selektion der Merkmale mittels Filterung in die Modellwahl integriert. Als Anwendungsbeispiel wird die Vorgehensweise zur Diskriminierung lebendiger und apoptotischer Zellen anhand markierungsfreier morphologischer Eigenschaften getestet. Es stellt sich dabei heraus, dass die entwickelte Vorgehensweise eine Reihe von Vorteilen gegenüber den traditionellen Methoden basierend

[1] Kapitel 3 kann somit als Anwendung der theoretischen Grundlagen verstanden werden, die in Kapitel 2 erarbeitet worden sind.

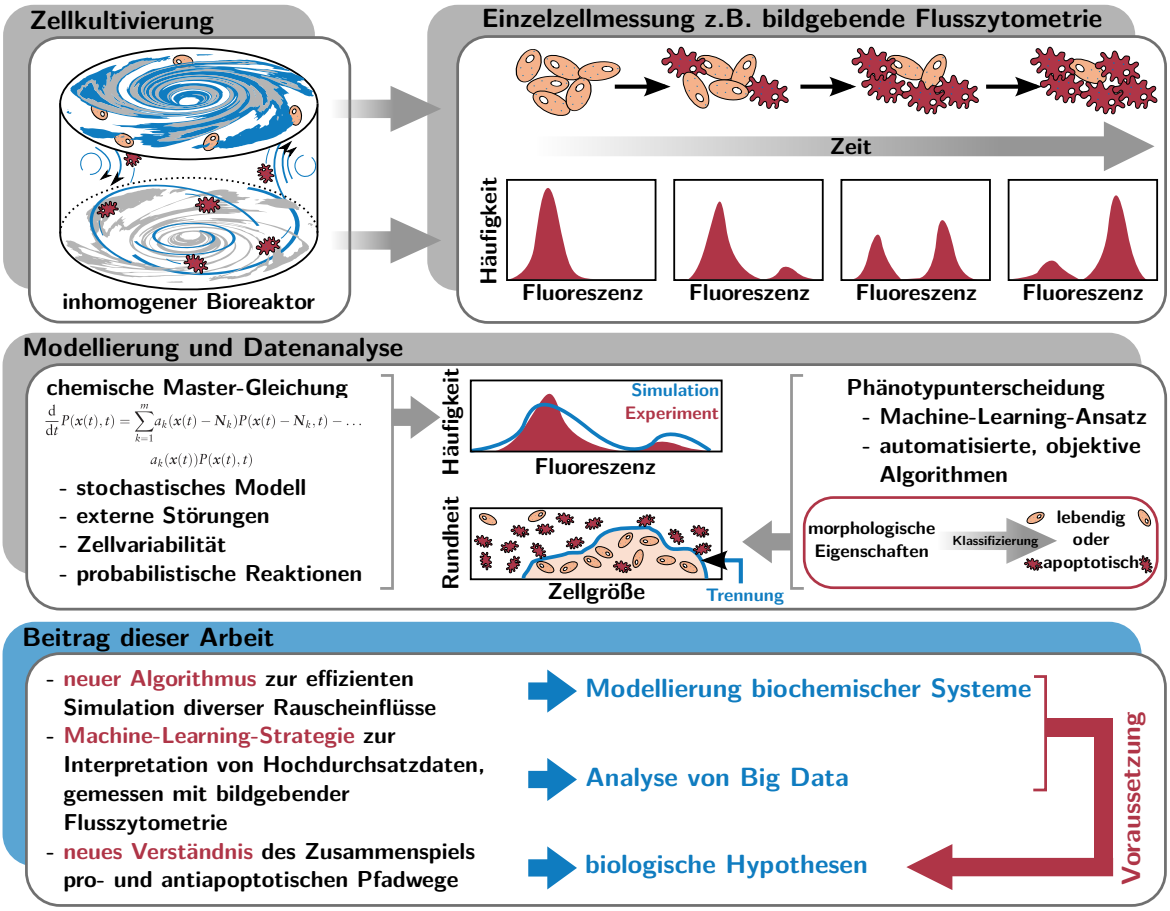

Abbildung 1.2: Überblick der Dissertation. Mittels systematischer Auswertung von Einzelzellexperimenten und adäquater Modellierung stochastischer zellulärer Prozesse werden in dieser Arbeit neue biologische Hypothesen bezüglich der Apoptose von HeLa-Zellen aufgestellt und überprüft. Dazu wird ein Algorithmus zur effizienten Simulation dynamischer System unter verschiedenen Störeinflüssen und ein Machine-Learning-Ansatz zur automatisierten Detektion apoptotischer Zellen entwickelt.

auf Fluoreszenzfarbstoffen aufweist, während vergleichbare Ergebnisse erzielt werden. Darüber hinaus wird aufgezeigt, welche Fehler und Schwierigkeiten häufig bei der automatisierten Analyse mittels Machine-Learning auftreten. Dieses Kapitel basiert auf einer vierten eigenen Publikation [Pischel et al., 2018].

Kapitel 5 - Zusammenfassung und Ausblick Abschließend werden die wichtigsten Ergebnisse dieser Arbeit zusammengefasst. Dabei wird darauf eingegangen, welcher wissenschaftliche Mehrwert generiert worden ist. Es werden offene Probleme und potenzielle Anwendungen der entwickelten Methoden und Hypothesen diskutiert.

2 Mathematische Modellierung biochemischer Systeme

Biologische Systeme, wie Organismen, Zellen oder Zellorganellen, stellen hoch strukturierte Einheiten dar, die in der Lage sind, verschiedenste Funktionen zu erfüllen. Bei der Beobachtung dieser Systeme werden häufig komplexe Prozesse identifiziert, die nicht durch Grundprinzipien erklärt werden können oder deren Verlauf nicht intuitiv vorhersagbar ist [Klipp et al., 2016]. Zum Verständnis dieser Prozesse werden in der Regel vereinfachte Modelle betrachtet, die eine adäquate Beschreibung der Realität erlauben. Dabei werden Modelle so gewählt, dass sie dem Prinzip der Sparsamkeit (Ockhams Rasiermesser) folgen [Guyon et al., 2010]. Das bedeutet, dass einfache Modelle mit wenigen Variablen oder Annahmen zu bevorzugen sind, wenn sie in der Lage sind, die beobachteten Sachverhalte zu erklären. Demnach richtet sich die Komplexität des Modells (*i*) nach der Detailliertheit der Beobachtung und (*ii*) nach dem Sachverhalt, der beschrieben werden soll. Die Detailliertheit der Beobachtung wird maßgeblich durch die Messmethodik bestimmt. Beispielsweise liefern Bulk-Populationsmessungen, wie Western Blots, lediglich den Mittelwert des Proteingehalts einer Zellpopulation. Sie geben keinen Aufschluss über die Streuung des biologischen Prozesses und die Existenz von Subpopulationen. Da Bulk-Populationsmessungen nur das Verhalten einer durchschnittlichen Zelle widerspiegeln, können aus diesen Messungen keine Einzelzellmodelle abgeleitet werden [Altschuler et al., 2010]. Sie dienen ausschließlich als Grundlage deterministischer Modelle, die den Mittelwert einer Zellpopulation beschreiben. Im Gegensatz dazu ist es mittels Einzellzellmessungen, wie zum Beispiel Konfokalmikroskopie, Fluss- oder Massenzytometrie, möglich Aufnahmen einzelner Zellen durchzuführen. Mithilfe dieser Messmethoden können stochastische Modelle aufgestellt werden, die es durch ihre detaillierte Betrachtungsweise erlauben, tiefere Einblicke in biochemische Prozesse vorzunehmen und deren probabilistische Natur zu verstehen [Spiller et al., 2010]. Dabei wird zwischen Methoden unterschieden, die in der Lage sind, der zeitlichen Dynamik einzelner Zellen (Zeitrafferaufnahmen) bzw. mittels Momentaufnahmen einzelner Zellen die zeitliche Dynamik der gesamten Population zu verfolgen (Einzelzellaufnahmen). Einzelzellaufnahmen zeichnen sich

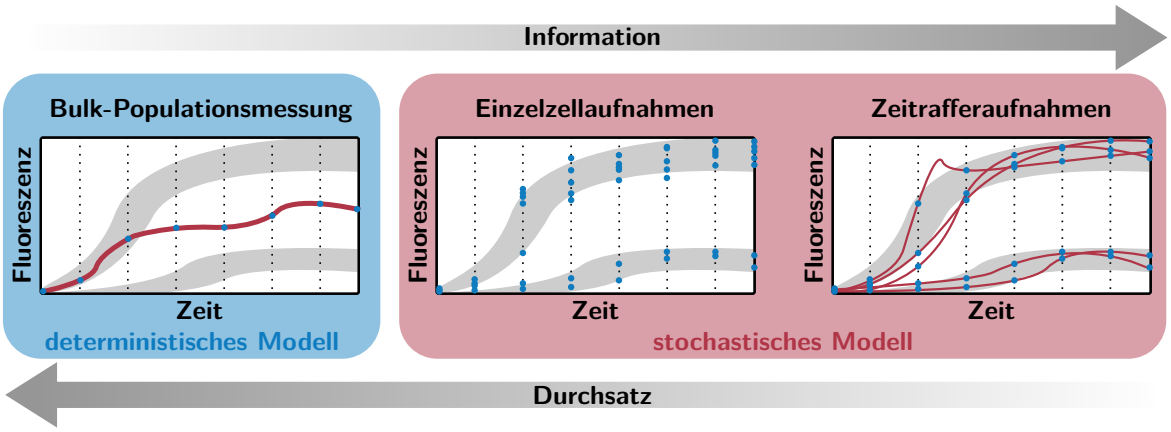

Abbildung 2.1: Beobachtung der Bildung von Subpopulationen mittels verschiedener Messmethodiken. Obwohl Bulk-Populationsmessungen im Gegensatz zu Einzelzellmessungen eine immense Anzahl von Zellen verwenden, liefern sie einen geringeren Informationsgewinn. Zur Erstellung stochastischer Modelle sind sie nicht geeignet, da sie lediglich den Mittelwert der Population charakterisieren und keinen Aufschluss über die Verteilung der Population zulassen.

im Kontrast zu Zeitrafferaufnahmen durch einen hohen Durchsatz aus. Im Gegenzug liefern Einzelzellaufnahmen jedoch keine Information über den zeitlichen Verlauf einzelner Zellen, sondern beschreiben den zeitlichen Verlauf einer Zellpopulation. Dies kann als Konflikt zwischen zeitlicher und statistischer Auflösung betrachten werden. In Abb. 2.1 sind Bulk-Populationsmessung, Einzelzell- und Zeitrafferaufnahmen für einen binären Entscheidungsprozess illustriert. Gut zu erkennen ist, dass der Verlauf der durchschnittlichen Zelle stark von der eigentlichen Verteilung der Population abweicht. Im Gegensatz dazu sind die Einzelzellmessmethoden in der Lage die Verteilung der Population akkurat wiederzugeben. Komplexe stochastische Modelle sind demnach nur anwendbar, wenn die Messungen hinreichend Information über die gesamte Population beinhalten, wie dieses Beispiel eindeutig zeigt. In den folgenden Abschnitten werden die Grundlagen deterministischer und stochastischer Modellierung besprochen und in den Kontext biologischer Beobachtungen gesetzt. Dabei wird erläutert, welche Annahmen den jeweiligen Modellierungsweisen zugrunde liegen und durch welche Stärken bzw. Schwächen sie gekennzeichnet sind.

2.1 Deterministische Beschreibung - Modellierung der durchschnittlichen Zelle

Die quantitative Beschreibung biologischer Prozesse beruht in der Regel auf ihrer Interpretation als biochemische Reaktionen $\boldsymbol{R}$, die die Umwandlung chemischer Spezies

S beschreiben [Klipp et al., 2016]

$$R_j: \quad \sum_{i=1}^{n_S} \nu_{ij}^e \mathrm{S}_i \rightarrow \sum_{i=1}^{n_S} \nu_{ij}^p \mathrm{S}_i. \tag{2.1}$$

In diesem Fall bezeichnen ν_{ij}^e und ν_{ij}^p die stöchiometrischen Koeffizienten der chemischen Spezies i und Reaktion j der Edukte e und Produkte p. Sie geben an, wie viele Moleküle einer chemischen Spezies bei einer Reaktion verbraucht bzw. erzeugt werden. Im Folgenden wird angenommen, dass die biologischen Prozesse in einem ideal gemischten System stattfinden, in dem keine räumlichen Konzentrationsgradienten existieren. Zudem muss das System hinreichend groß sein, damit chemische Spezies als Kontinuum betrachtet und Störungen bzw. Fluktuationen vernachlässigt werden können [Wilkinson, 2009; Grima, 2010]. In diesem Fall wird die Änderung des Systemzustandes $\boldsymbol{x}(t)$, der durch die Abundanzen der chemischen Spezies charakterisiert wird, durch die Ratengleichung bestimmt

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{N}\boldsymbol{r}(\boldsymbol{x}). \tag{2.2}$$

Dabei beschreibt die stöchiometrische Matrix $\boldsymbol{N}$ die Änderung des Systemzustandes durch die Reaktionen

$$\boldsymbol{N} = \begin{pmatrix} \nu_{11}^p - \nu_{11}^e & \dots & \nu_{1n_r}^p - \nu_{1n_r}^e \\ \vdots & \ddots & \vdots \\ \nu_{n_S1}^p - \nu_{n_S1}^e & \dots & \nu_{n_Sn_r}^p - \nu_{n_Sn_r}^e \end{pmatrix} \tag{2.3}$$

und die Reaktionsraten $\boldsymbol{r}(\boldsymbol{x})$ geben die Geschwindigkeiten der Reaktionen an. Die Abundanzen der chemischen Spezies werden dabei in Teilchenzahlen ausgedrückt. Mittels Division durch das Systemvolumen Ω ergeben sich die makroskopischen Konzentrationen $\boldsymbol{\phi}(t)$

$$\boldsymbol{\phi}(t) = \frac{\boldsymbol{x}(t)}{\Omega}. \tag{2.4}$$

Abhängig davon, welcher Reaktionsmechanismus zugrunde liegt, nehmen die effektiven Reaktionsraten unterschiedliche Formen an. Ein einfacher kinetischer Ansatz ist die Massenwirkungskinetik, die durch das Massenwirkungsprinzip der chemischen Thermodynamik inspiriert ist [Klipp et al., 2016]. Für eine Reaktion der Form

$$R_j: \quad \sum_{i=1}^{n_S} \nu_{ij}^e \mathrm{S}_i \rightarrow \dots \tag{2.5}$$

ergeben sich damit die Raten zu

$$r_j = k_j \prod_{i=1}^{n_S} x_i^{\nu_{ij}^e} \tag{2.6}$$

mit der Ratenkonstante k_j als Proportionalitätsfaktor. Basierend auf der Massenwirkungskinetik ist es möglich weitere Kinetiken herzuleiten, wie beispielsweise die Michaelis-Menten- [Michaelis et al., 1913] oder Hill-Kinetik [Hill, 1913].

In der Regel stellt Gl. 2.2 eine nichtlineare Differenzialgleichung dar, deren Dynamik komplexes Verhalten aufzeigen kann [Argyris et al., 2015]. Eine kleine Auswahl ist in Abb. 2.2 exemplarisch illustriert. Durch einfache Reaktionsnetzwerke, wie beispielsweise eine Aktivierung bzw. Inhibition, ist es möglich, unbeschränkte bzw. beschränkte Systeme zu schaffen. Systeme dieser Art werden unter anderem bei der Modellierung des Wachstums von Bakterien verwendet [Novick, 1955]. Mittels geeigneter Kombination von Aktivierungen (doppeltes positives Feedback) und Inhibitionen (doppeltes negatives Feedback) können komplexere Strukturen, wie bistabile Systeme, erzeugt werden [Ferrell Jr, 2002]. Diese können durch einen passenden Stimulus von einem Zustand in den anderen überführt werden und in diesem Zustand verharren, sogar nach dem Verschwinden des Stimulus [Ferrell Jr, 2002; Strasser et al., 2012]. Ist der Systemzustand periodischen Änderungen durch die nichtlineare Dynamik des Systems unterworfen, liegt oszillatorisches Verhalten vor. Prominente Beispiele von Oszillationen in chemischen Reaktionssystemen sind unter anderem die Briggs-Rauscher- [Briggs et al., 1973], Belousov-Zhabotinsky- [Field et al., 1974] oder Bray-Liebhafsky-Reaktion [Bray, 1921]. Die oben genannten Effekte stellen nur einen kleinen Ausschnitt der in biochemischen Systemen auftretenden nichtlinearen Phänomene dar [Wolf et al., 2003]. Obwohl bis jetzt ausschließlich eine vereinfachte deterministische Betrachtungsweise gewählt worden ist, wird deutlich, dass sich durch die Synergie nichtlinearer Prozesse komplexe Verhaltensmodule aufbauen lassen, deren Kompliziertheit mit der Anzahl der Reaktionen und chemischen Spezies ansteigt. Dies gilt nicht ausschließlich für die hier deterministisch betrachteten Systeme, sondern lässt sich auch auf die in den folgenden Abschnitten behandelten, weitaus komplizierteren stochastischen Systeme anwenden.

2.2 Unsicherheiten in biochemischen Systemen

Die deterministische Beschreibung biochemischer Systeme ist sehr erfolgreich gewesen und hat schon früh große Erfolge gefeiert, wie die Erklärung der Ausbreitung des Aktionspotentials in Nervenzellen [Hodgkin et al., 1952], die 1963 mit dem Nobelpreis ausgezeichnet wurde. Mit der raschen technischen Entwicklung von Einzelzellmessmethodiken, wie Mikroskopie [Amos, 2000] und Flusszytometrie [Herzenberg et al., 2002], stellt sich jedoch heraus, dass eine rein deterministische Beschreibung oft nicht in der Lage ist zufriedenstellende Resultate zu erzielen [Ori et al., 2018]. In Experi-

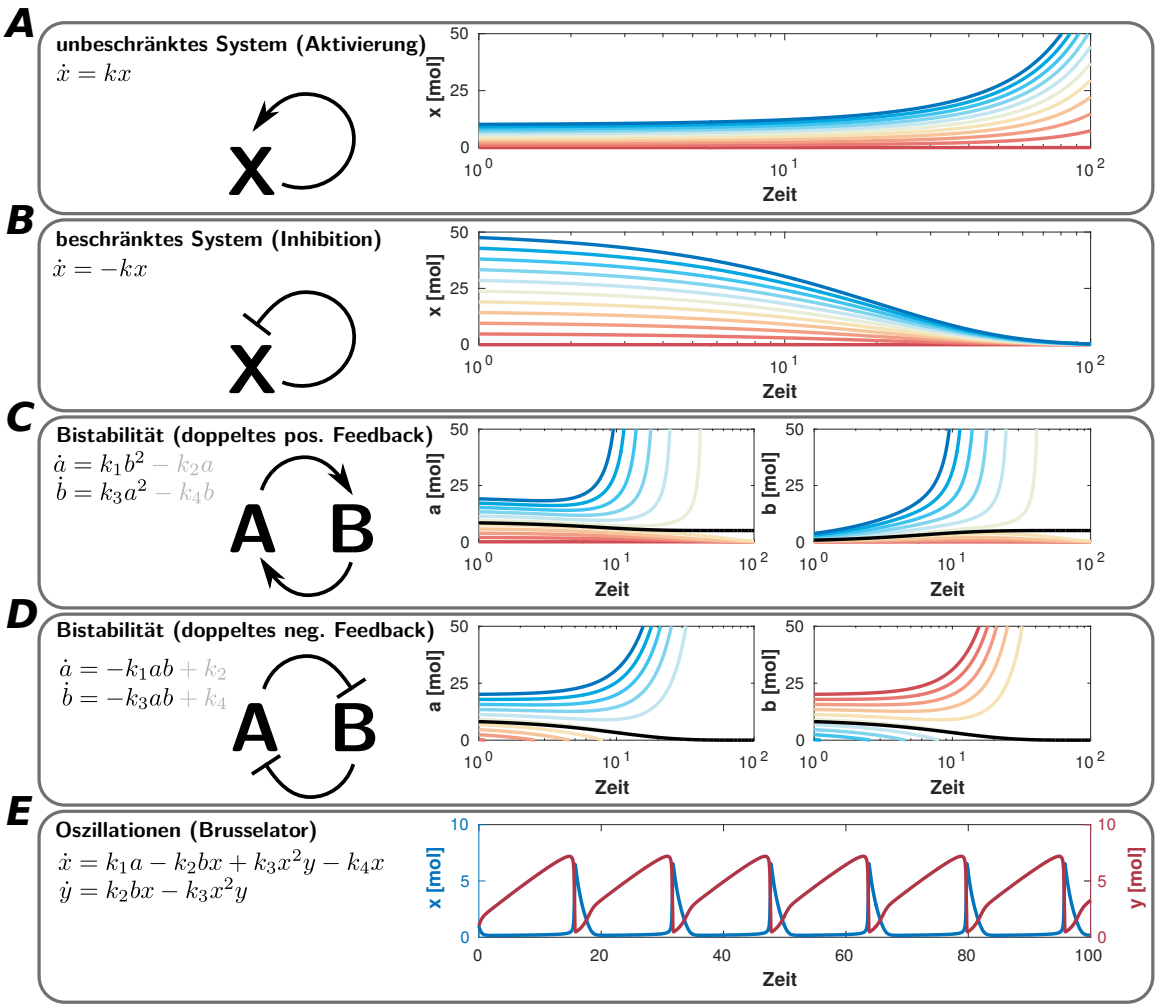

Abbildung 2.2: Dynamik biochemischer Reaktionssysteme. Aktivierungen **(A)** und Inhibitionen **(B)** sind in der Lage, unbeschränkte bzw. beschränkte Systeme zu erzeugen. Durch ihre Kombination können bistabile Systeme konzipiert werden. In diesem Beispiel ist doppeltes positives Feedback **(C)** für die gleichzeitige Aktivierung bzw. Inaktivierung von A und B verantwortlich. Im Gegensatz dazu bewirkt doppeltes negatives Feedback **(D)** die Aktivierung von A und Inaktivierung von B bzw. die Umkehrung davon [Ferrell Jr, 2002]. Die schwarze Linie trennt beide Zustände voneinander. Die grauen Terme in den Ratengleichungen sind nicht in den Pfeilschemata eingezeichnet. Sie sind lediglich zur besseren Darstellung des zeitlichen Verlaufs der Zustände eingefügt. Das letzte Beispiel **(E)** zeigt das oszillatorische Verhalten des Brusselators [Prigogine et al., 1968].

menten kann gezeigt werden, dass der individuelle Charakter biochemischer Systeme sehr dominant ist, weshalb Variabilität und Heterogenität als fundamentale Eigenschaften dieser Systeme angesehen werden. Aus diesem Grund kann die Dynamik einer Zellpopulation in der Regel nicht durch die Dynamik der durchschnittlichen Zelle abgebildet werden [Spiller et al., 2010]. Beispiele dieser Erscheinungen sind unter anderem die Bildung von inhomogenem Krebsgewebe aus Tumorstammzellen [Meacham

et al., 2013], das zufällige Eintreten des Zelltodes durch Apoptose [Xia et al., 2014] und die Zusammensetzung mikrobieller Populationen bestehend aus Hefezellen [Blake et al., 2003], Bakterien [Lidstrom et al., 2010; Ackermann, 2015] oder Mikroalgen [Fachet et al., 2016]. Die Ursache der Heterogenität biochemischer Systeme und der damit verbundenen phänotypischen Vielfalt stellen Störungen der Dynamik durch stochastische Einflüsse dar [Kærn et al., 2005; Wilkinson, 2009; Delvigne et al., 2017]. Unterschiedliche Phänotypen werden dabei durch den Wechsel von einem stabilen Zustand aufgrund von Störungen in einen anderen hervorgerufen [Ferrell Jr, 2002]. Es ist möglich, dass der neue Zustand im deterministischen Fall nicht existiert und lediglich durch Störungen induziert wird [Qian et al., 2009; Strasser et al., 2012; Biancalani et al., 2014].

Der Einfluss von Störungen auf biochemische Systeme wirft nicht nur biologisch interessante Fragestellungen auf, wie zum Beispiel:

- Wie gelingt es Zellen, trotz interner und externer Fluktuationen ihre Funktion aufrechtzuerhalten (Robustheit) [Stelling et al., 2004]?
- Wie wägen Zellen zwischen Robustheit und Effizienz ab [Kitano, 2004]?

sondern er ist auch für die Bioprozesstechnik von großer Bedeutung:

- Wie setzt sich die Gesamtvariabilität eines Bioprozesses zusammen [Delvigne et al., 2014a]?
- Ist es möglich die Gesamtvariabilität eines Bioprozesses extern zu steuern [Gernaey et al., 2010]?
- Wie optimiert man einen Bioprozess mit Unsicherheiten hinsichtlich Ausbeute, Ertrag oder Produktqualität [Gernaey et al., 2010]?

Zur Illustration sind in Abb. 2.3 zwei Beispiele gezeigt, die für die Optimierung von Bioprozessen typische störungsinduzierte Probleme darstellen. Zum einen sind viele Prozesse durch eine hohe Variabilität gekennzeichnet, die die Qualität des Produktes mindert. Außerdem wird die Ausbeute oft durch die Bildung von unproduktiven Subpopulationen verringert [Delvigne et al., 2014b]. Die Hoffnung ist, dass ein tieferes Verständnis der Variabilität biochemischer Systeme genutzt werden kann, um Bioprozesse zu optimieren und durch externe Steuerung die störungsinduzierten Probleme zu minimieren.

Zur detaillierteren Untersuchung des Einflusses von Störungen in biochemischen Systemen wird im Folgenden zwischen intrinsischen Störungen, gekennzeichnet durch probabilistische Reaktionen, extrinsischen Störungen, gekennzeichnet durch Zellvariabilität, und externen Störungen, gekennzeichnet durch fluktuierende Umwelteinflüsse,

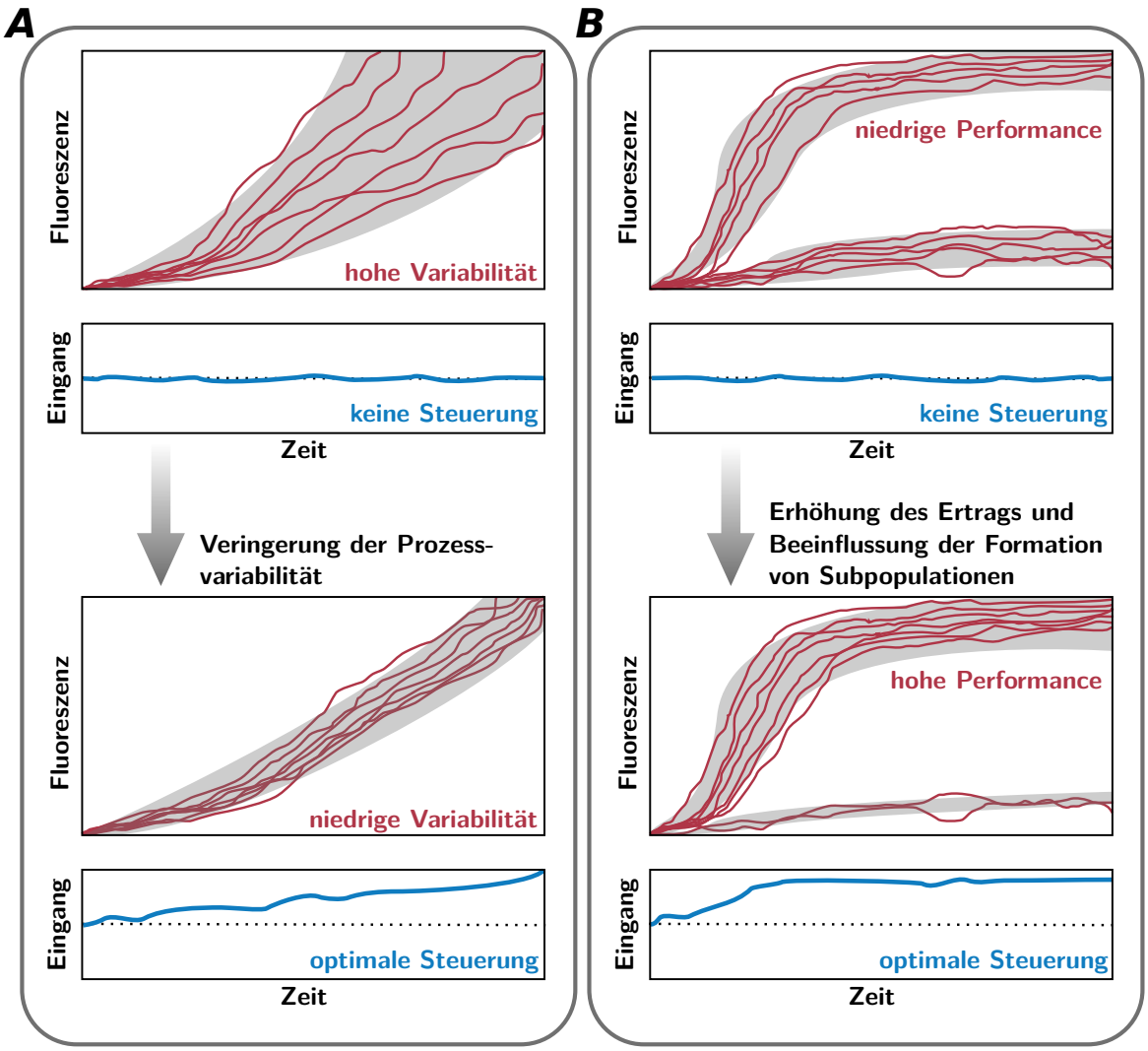

Abbildung 2.3: Steuerung der Prozessvariabilität. Prozessstreuung **(A)** und Ausbildung von Subpopulationen **(B)** stellen große Hürden in der Bioprozesstechnik dar. Für einige Anwendungen ist es nötig die Streuung eines Prozesses zu minimieren, um beispielsweise Produkte mit möglichst gleicher Qualität zu erzielen, wohingegen andere Anwendungen die Eliminierung gewisser Subpopulationen zur Ertragssteigerung erfordern.

unterschieden [Patnaik, 2006; Pischel et al., 2017]. Intrinsische Störungen sind besonders dominant in kleinen Systemen mit geringen Abundanzen der chemischen Spezies [Kærn et al., 2005], wohingegen extrinsische und externe Störungen mit der Systemgröße ansteigen [Bar-Even et al., 2006; Wilkinson, 2009; Delvigne et al., 2017]. Die Überlagerung dieser Störungen und deren Auswirkung auf die Systemdynamik ist wegen experimenteller und rechentechnischer Hürden jedoch kaum erforscht [Spiller et al., 2010; Lencastre Fernandes et al., 2011; Delvigne et al., 2014b]. Zum einen gestaltet sich das Design von Experimenten mit definierten Umwelteinflüssen schwierig.

Herkömmliche Bioreaktoren sind durch räumliche Inhomogenitäten und unzureichende Durchmischung gekennzeichnet [Epstein, 1995; Hartman et al., 2011], weshalb der Einfluss externer Störungen nur ungenügend abgeschätzt werden kann. Erst mit der Entwicklung von mikrofluidischen Chips und deren Charakterisierung mittels Computersimulationen war es möglich, Einzelzellexperimente unter definierten Bedingungen durchzuführen [Dusny et al., 2012; Westerwalbesloh et al., 2015]. Des Weiteren ist die Simulation stochastischer Systeme sehr zeitaufwändig und rechenintensiv, da der Systemzustand $\boldsymbol{x}$ eine kontinuierliche Zufallsvariable darstellt. Konkrete Realisierungen des Systemzustandes im Experiment sind deshalb nicht vorhersagbar. Lediglich Wahrscheinlichkeitsaussagen bezüglich des Eintretens von Realisierungen können getroffen werden. Die Zentrale Größe zur Charakterisierung stochastischer Systeme ist die Wahrscheinlichkeitsdichte ρ, die die Berechnung statistisch relevanter Größen erlaubt. Dazu zählen die Momente m-ter Ordnung

$$\boldsymbol{\mu}_m = \int \boldsymbol{x}^m \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{2.7}$$

die zentralen Momente m-ter Ordnung

$$\boldsymbol{\mu}'_m = \int (\boldsymbol{x} - \boldsymbol{\mu}_1)^m \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{2.8}$$

aber auch die Wahrscheinlichkeit des Auffindens des Systems innerhalb eines bestimmten Intervalls $[\boldsymbol{x}_-, \boldsymbol{x}_+]$

$$P(\boldsymbol{x}_- \leq \boldsymbol{x} \leq \boldsymbol{x}_+) = \int_{\boldsymbol{x}_-}^{\boldsymbol{x}_+} \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}. \tag{2.9}$$

Unglücklicher Weise lässt sich die Wahrscheinlichkeitsdichte nur unter großem rechnerischen Aufwand ermitteln. Besonders in Optimierungsproblemen, wie der Bestimmung unbekannter Parameter [Jaqaman et al., 2006], müssen zeitaufwändige Modellsimulationen oft wiederholt werden. Aus diesem Grund werden in der Regel ausschließlich approximative Rechenansätze verwendet, die den Systemzustand durch statistische Momente niedriger Ordnung charakterisieren [Wu et al., 2006; Gillespie et al., 2013]. Man beschränkt sich dabei meist auf den Mittelwert $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}_1$ und die Varianz $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\mu}'_2$ bzw. die Standardabweichung $\mathrm{std}(\boldsymbol{x}) = \sqrt{\mathrm{Var}(\boldsymbol{x})}$.

In den folgenden Abschnitten werden die wichtigsten Modellierungsansätze der oben aufgeführten Störungen besprochen. Anschließend wird auf die Schwierigkeiten der simultanen Simulation verschiedener Störungen eingegangen und wie diese effektiv approximiert werden können. Eine dafür in dieser Arbeit entwickelte Methode wird dann an verschiedenen Modellsystemen getestet und zur Schätzung unbekannter Parameter verwendet.

2.3 Simulation extrinsischer Störungen

Die Variabilität isogener Zellpopulationen wird im Allgemeinen als extrinsische Störung bezeichnet. Hinter diesem Begriff verbergen sich nicht nur Unterschiede hinsichtlich des Phänotyps, sondern auch Abweichungen in den Proteinabundanzen oder der Zellmorphologie. Da sich jede Zelle von den anderen Zellen einer Population unterscheidet, werden die Abundanzen der chemischen Spezies in der Regel nicht als einfache reelle Zahlen betrachtet, wie beispielsweise in Gl. 2.2. Stattdessen werden die Abundanzen durch Zufallsvariablen charakterisiert, deren Wahrscheinlichkeit für eine bestimmte Realisierung durch die Wahrscheinlichkeitsverteilung gegeben ist. Am häufigsten anzutreffen sind unimodale Verteilungen, wie die Normalverteilung [Furusawa et al., 2005] oder logarithmischen Normalverteilung [Limpert et al., 2001]. In Kombination mit nichtlinearen Prozessen biochemischer Reaktionssysteme können jedoch Multistabilitäten eine Aufspaltung von unimodalen in bimodale Verteilungen bewirken [Ferrell Jr, 2002].

Um diesen Sachverhalt genauer zu verstehen, wird im Folgenden ein einfaches Eingangs-Ausgangs-Modell betrachtet. Dieses stellt eine nichtlineare Transformation dar, die sich durch Integration der Ratengleichung ergibt. Sie kann als funktionale Abhängigkeit des Systemzustandes zu einem bestimmten Zeitpunkt von gewissen Parametern interpretiert werden. Zur Beschreibung einer isogenen Zellpopulation mit solch einem Modell wird angenommen, dass sich die Topologie des biochemischen Reaktionsnetzwerkes zwischen den einzelnen Zellen nicht unterscheidet. Lediglich Parameter, die das Reaktionsnetzwerk charakterisieren, variieren zwischen den Zellen und stellen somit Zufallsvariablen dar. In dem in Abb. 2.4 betrachteten Fall ist die funktionale Abhängigkeit durch einen sigmoidalen Verlauf gekennzeichnet. Kleine Eingangssignale induzieren stets ein kleines Ausgangssignal, wohingegen große Eingänge ein großes Ausgangssignal bewirken. Zudem gibt es eine schmale Übergangszone zwischen beiden Bereichen, die durch einen steilen Anstieg gekennzeichnet ist. Zur Untersuchung des Einflusses externer Störungen werden normalverteilte Zufallsvariablen mit unterschiedlichen Mittelwerten und Standardabweichungen als Systemeingang verwendet. Diese werden abhängig davon, welche Form die Transformation im interessierenden Bereich aufzeigt, unterschiedlich abgebildet. In Abb. 2.4A wird dazu der Einfluss konkaver bzw. konvexer Transformationen eines normalverteilten Systemeingangs untersucht. Konkave Funktionen zeichnen sich dadurch aus, dass sie stets unter der Verbindungslinie zweier Punkte der Funktion liegen. Im Gegensatz dazu liegen konvexe Funktionen über der Verbindungslinie. Es stellt sich heraus, dass in konvexen Bereichen die symmetrische Verteilung in eine rechtsschiefe Verteilung transformiert wird, wohingegen

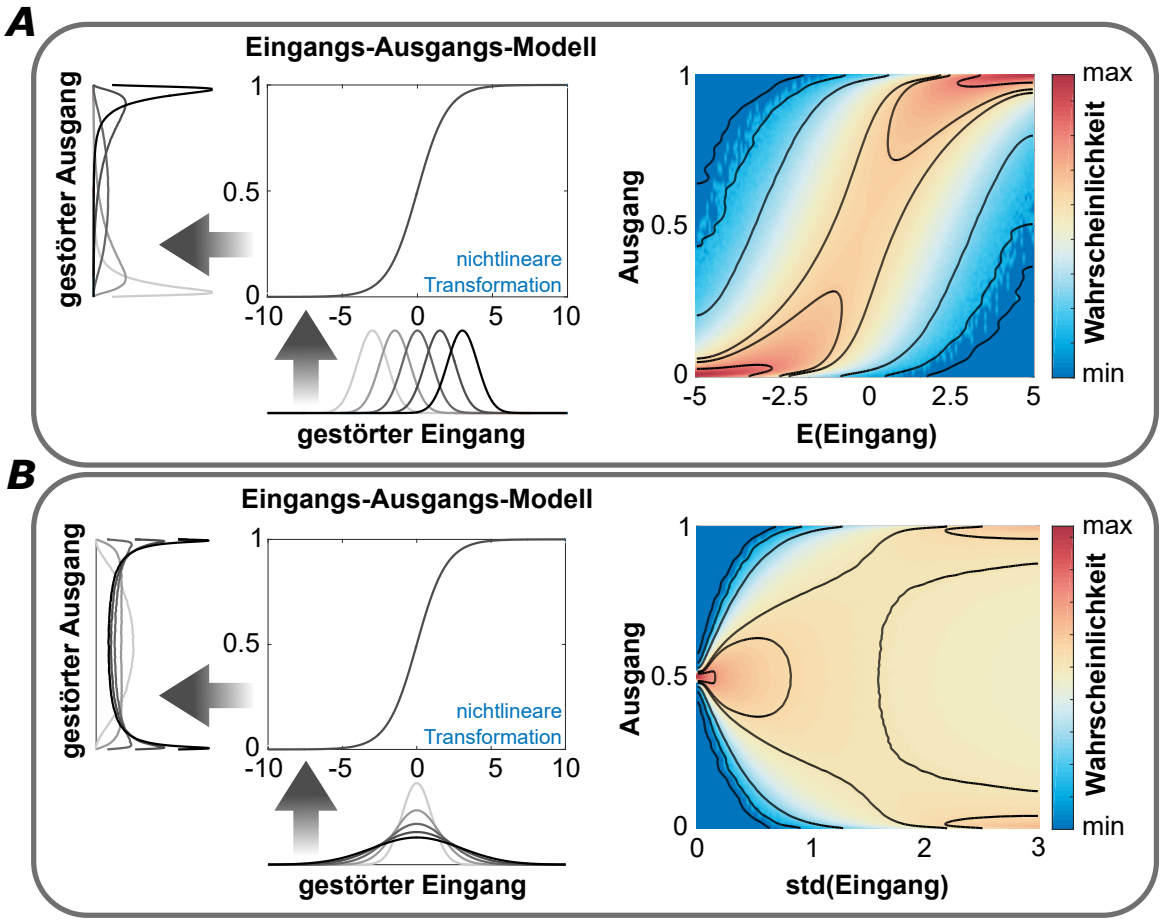

Abbildung 2.4: Einfluss externer Störungen auf biochemische Reaktionssysteme. **(A)** Das sigmoidale Eingangs-Ausgangs-Modell kann in einen konkaven und einen konvexen Bereich unterteilt werden. Eine konkave Funktion transformiert eine Normalverteilung in eine linksschiefe Verteilung, wohingegen eine konvexe Funktion eine Normalverteilung in eine rechtsschiefe Verteilung transformiert. Dies wird für Normalverteilungen mit unterschiedlichen Mittelwerten E(·) und gleicher Standardabweichung std(·) illustriert. **(B)** Nichtlineare Funktionen können unimodale Verteilungen in bimodale Verteilungen transformieren. Schmale Verteilungen, die lediglich im linearen Übergangsbereich existieren erzeugen unimodale Ausgangsverteilungen. Im Gegensatz dazu erzeugen breite Eingangsverteilungen, die über die Übergangszone herausragen bimodale Ausgangsverteilungen. Dies wird für Normalverteilungen mit gleichen Mittelwerten und unterschiedlicher Standardabweichung illustriert.

konkave Bereiche linksschiefe Verteilungen erzeugen [Kim et al., 2013]. Ausschließlich im linearen Bereich ändert sich die Form durch die Transformation nicht, da die Ausgangsverteilung lediglich eine Umskalierung der Eingangsverteilung darstellt. Außergewöhnliche Ausgangsverteilungen ergeben sich, wenn die Transformation im interessierenden Bereich konvexes und konkaves Verhalten gleichzeitig aufzeigt. Dazu werden in Abb. 2.4B normalverteilte Zufallsvariablen mit konstantem Mittelwert und unterschiedlicher Standardabweichung in das Zentrum des linearen Übergangsberei-

ches platziert. Es zeigt sich, dass breite Eingangsverteilungen nicht nur den linearen Bereich abdecken, sondern auch konvexe und konkave Bereiche. Die resultierende Ausgangsverteilung ergibt sich demnach als Superposition von Verteilungen verschiedener Formen. Je breiter die Eingangsverteilung ist, desto stärker werden die konvexen und konkaven Anteile gewichtet, was eine bimodale Ausgangsverteilung induziert. Im Gegensatz dazu wichten schmale Verteilungen den normalverteilten Anteil stärker.

Obwohl das hier betrachtete Beispiel sehr einfach ist, zeigt es eindrucksvoll, wie nichtlineare Transformationen symmetrische, unimodale Verteilungen verzerren und sogar in bimodale Verteilungen aufspalten. In der Regel sind biochemische Reaktionssysteme sehr komplex und bestehen aus einer Vielzahl von Reaktionen und chemischen Spezies. Die Transformationen dieser Systeme sind weitaus komplizierter, weshalb noch zusätzliche, hier nichtgezeigte, nichtlineare Phänomene auftreten können. Um diese zu verstehen, muss von einer qualitativen Diskussion zu einer quantitativen Beschreibung übergegangen werden. Analog wie zuvor wird dazu der Systemausgang η als nichtlineare Transformation $g(\xi)$ des Systemeingangs ξ betrachtet

$$\eta = g(\xi). \tag{2.10}$$

Da der Systemeingang eine Zufallsvariable darstellt, ist auch der Systemausgang von stochastischer Natur. Die Transformation $g(\xi)$ ist jedoch ein deterministischer Prozess, der keine zusätzliche Unsicherheit mit einbringt. Für monotone Transformationen kann die Verteilung des Systemausgangs ρ_η einfach aus der Eingangsverteilung ρ_ξ analytisch berechnet werden [Hines et al., 1990]

$$\rho_\eta = \rho_\xi(\xi)\left|\frac{\mathrm{d}\xi}{\mathrm{d}\eta}\right| = \rho_\xi(g^{-1}(\eta))\left|\frac{\mathrm{d}g^{-1}(\eta)}{\mathrm{d}\eta}\right|. \tag{2.11}$$

Nicht monotone Funktionen werden in monotone Stücke aufgeteilt, die separat transformiert werden müssen. Für reale Systeme ist es unmöglich, eine analytische Form der Transformation $g(\xi)$ und damit der Ausgangsverteilung ρ_η zu ermitteln. Zudem bereiten numerische Berechnungen Schwierigkeiten aufgrund ihres immensen Zeitaufwandes und ihrer hohen Rechenintensität. Zur Berechnung der Ausgangsverteilung oder ihrer statistischen Momente werden deshalb oft approximative Ansätze verwendet, die Rechenaufwand gegen Präzision abwägen. Im Folgenden wird ein Überblick bezüglich der wichtigsten Vertreter dieser Methoden sowie deren Stärken und Schwächen gegeben.

2.3.1 Approximation mittels Taylor-Entwicklung

Die Tatsache, dass die nichtlineare Funktion $g(\xi)$ in den meisten Fällen nicht berechenbar oder unbekannt ist, stellt Wissenschaftler vor große Probleme. Eine beliebte

Strategie, um dennoch die Ausgangsverteilung oder ihre statistischen Momente abzuschätzen, besteht darin $g(\xi)$ durch eine Ersatzfunktion $\hat{g}(\xi)$ zu approximieren. Die Taylor-Entwicklung stellt eine geeignete Ersatzfunktion dar, die mit beliebiger Güte an die unbekannte Funktion $g(\xi)$ angenähert werden kann

$$\eta = g(\xi) \approx \hat{g}(\xi) = \sum_{i=0}^{N} \left. \frac{\partial^i g(\xi)}{\partial \xi^i} \right|_{\xi=\xi_0} \frac{(\xi - \xi_0)^i}{i!}. \tag{2.12}$$

Sie ist einfach zu berechnen, da lediglich die Ableitungen von $g(\xi)$ an der Stelle ξ_0 ermittelt werden müssen. In der Regel wird um den Mittelwert des Eingangssignals $\xi_0 = \mathrm{E}(\xi)$ entwickelt. Obwohl die Güte der Approximation mit zunehmender Anzahl der aufsummierten Terme N ansteigt, wird in der Praxis meist $N = 1$ gewählt. Dadurch wird die Funktion $g(\xi)$ mittels einer linearen Funktion angenähert

$$\eta \approx g(\mathrm{E}(\xi)) + \left. \frac{\partial g(\xi)}{\partial \xi} \right|_{\xi=\mathrm{E}(\xi)} (\xi - \mathrm{E}(\xi)), \tag{2.13}$$

was eine schnelle Berechnung der Ausgangsverteilung mithilfe von Gl. 2.11 ermöglicht. Des Weiteren lassen sich die statistischen Momente, wie Mittelwert und Varianz, der Ausgangsverteilung einfach ermitteln [Ku, 1966; Hines et al., 1990]

$$\mathrm{E}(\eta) \approx \mathrm{E}(\hat{g}(\xi)) = g(\mathrm{E}(\xi)), \tag{2.14}$$

$$\mathrm{Var}(\eta) \approx \mathrm{Var}(\hat{g}(\xi)) = \left(\left. \frac{\partial g(\xi)}{\partial \xi} \right|_{\xi=\mathrm{E}(\xi)} \right)^2 \mathrm{Var}(\xi). \tag{2.15}$$

Wie bereits in Abb. 2.4 gezeigt, stellen lineare Transformationen von Wahrscheinlichkeitsverteilungen lediglich Skalierungen dar. Dies ist zu erkennen in Gl. 2.11, wenn die funktionale Abhängigkeit durch einen linearen Zusammenhang ersetzt wird, aber auch in Gl. 2.15, da die Varianz des Ausgangssignals proportional zur Varianz des Eingangssignals ist. Es stellt sich heraus, dass dies nicht nur für normalverteilte Zufallsvariablen, sondern auch für beliebige Verteilungen der Fall ist [Hines et al., 1990], wie in Abb. 2.5A dargestellt. Trotz der Eleganz und Einfachheit der Linearisierung lässt sich rasch erkennen, dass sie nur für wenige Spezialfälle präzise Vorhersagen liefert, nämlich wenn $g(\xi)$ in der unmittelbaren Umgebung der Eingangsverteilung akkurat durch eine lineare Funktion angenähert werden kann. Das bedeutet, dass nichtlineare Phänomene, wie die Entstehung von schiefen Verteilungen aus symmetrischen oder die Aufspaltung von unimodalen Verteilungen in bimodale, nicht mittels der Linearisierung beschrieben werden können. Um zu demonstrieren, dass lineare Transformationen in biochemischen Reaktionssystemen nur selten vorkommen, sind in Abb. 2.5B-D die zeitlichen Verläufe und die zugehörigen Transformationen bezüglich der Reaktionskonstanten für Reaktionen nullter, erster und zweiter Ordnung dargestellt. Ausschließlich

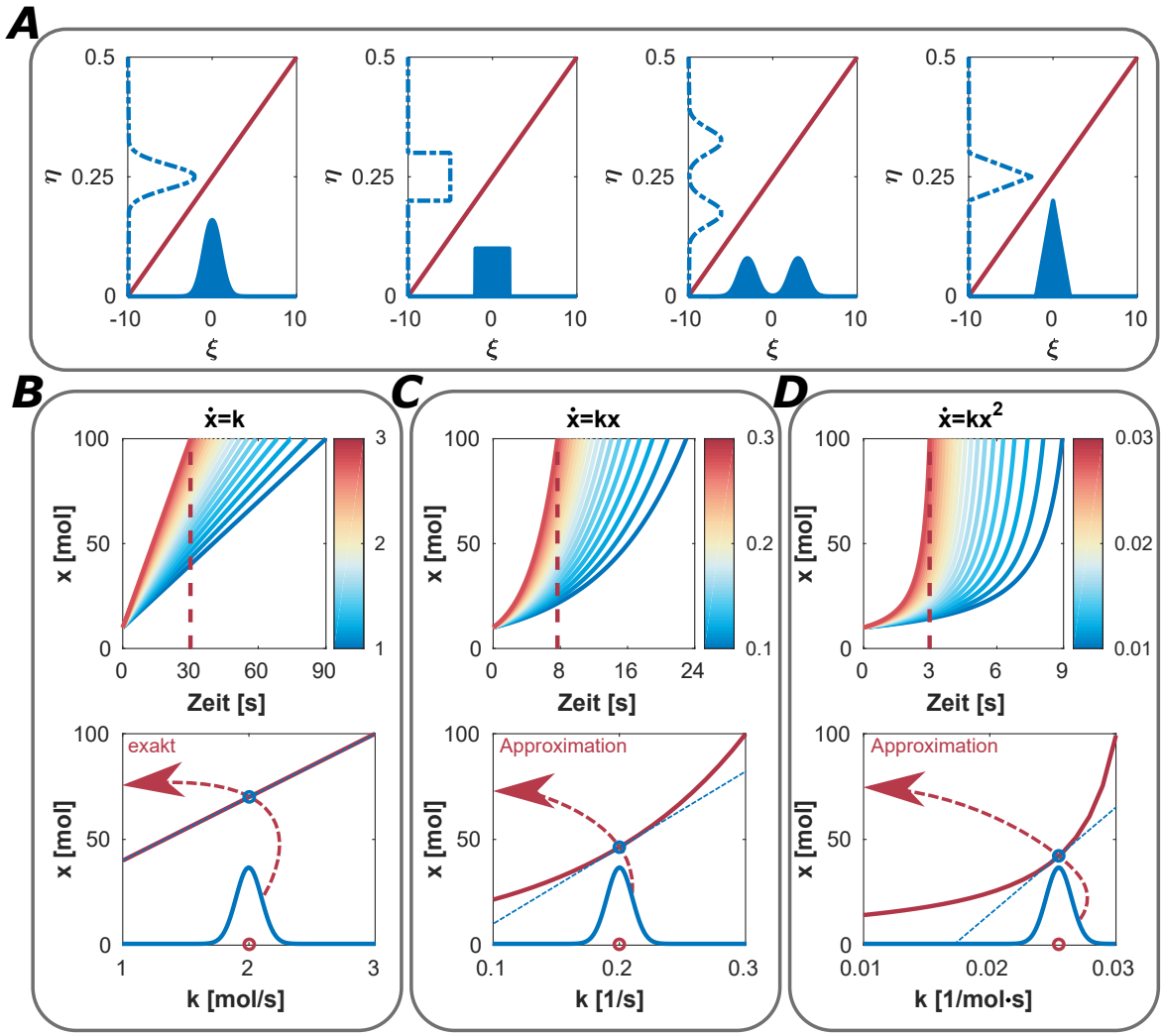

Abbildung 2.5: Approximation mittels Linearisierung. **(A)** Lineare Transformationen bewirken lediglich eine Skalierung des Eingangssignals. Die Form des Signals ändert sich dabei nicht. Für Reaktionen nullter **(B)**, erster **(C)** und zweiter Ordnung **(D)** sind die Zeitverläufe für verschiedene Reaktionskonstanten skizziert. Dabei bezeichnen rote Kurven große und blaue Kurven kleine Reaktionskonstanten. Für die mit einer roten, vertikal gestrichelten Linie markierten Zeitpunkte ist zusätzlich die Abhängigkeit des Systemzustandes $\boldsymbol{x}$ von den Reaktionskonstanten gezeigt. Für die Reaktion nullter Ordnung ergibt sich eine lineare Transformation (rot, durchgezogen), die exakt durch eine Linearisierung (blau, gestrichelt) abgebildet werden kann. Im Gegensatz dazu stellen die Transformationen der anderen Reaktionen nichtlineare Funktionen dar, die nicht exakt durch die Linearisierung angenähert werden können.

die Reaktion nullter Ordnung lässt sich durch einen linearen Zusammenhang beschreiben und somit akkurat durch die Linearisierung approximieren. Die übrigen Reaktionen zeigen nichtlineare Abhängigkeiten auf, deren Nichtlinearität mit zunehmender Kompliziertheit des Reaktionsmechanismus ansteigt. Für diese Reaktionen stellt die

Linearisierung nur eine Näherung dar, deren Fehler abhängig davon ist, wie stark nichtlinear die Transformation in unmittelbarer Umgebung des Eingangssignals ist.

Obwohl in verschiedenen Anwendungen gezeigt worden ist, dass die Berücksichtigung von Termen höherer Ordnung der Taylor-Entwicklung die Güte der Approximation steigert und somit zuverlässigere Ergebnisse erzielt werden können, wird in der Praxis selten eine Approximation zweiter oder höherer Ordnung verwendet [Xue et al., 2012]. Grund dafür ist der steigende Rechenaufwand, da Ableitungen höherer Ordnung von $g(\xi)$ berechnet werden müssen [Julier et al., 2004; Xue et al., 2012]. Darüber hinaus ist $g(\xi)$ oft keine analytische Funktion und kann verschiedene Unstetigkeitsstellen aufweisen, weshalb womöglich keine Taylor-Entwicklung vorgenommen werden kann. Diese Limitationen sorgten dafür, dass Linearisierungen durch zuverlässigere Techniken zur Propagation von Unsicherheiten in biochemischen Reaktionssystemen ersetzt worden sind.

2.3.2 Approximation mittels Gauß-Quadratur

Besteht kein Interesse daran, die Wahrscheinlichkeitsdichte des Systemausgangs zu rekonstruieren, sondern lediglich dessen statistische Momente zu ermitteln, ist es möglich diese durch Integration zu berechnen. Für den Mittelwert und die Varianz ergeben sich mit Gl. 2.7-2.8 folgende Ausdrücke

$$\mathrm{E}(\eta) = \int \eta \rho_\eta \, \mathrm{d}\eta, \tag{2.16}$$

$$\mathrm{Var}(\eta) = \int (\eta - \mathrm{E}(\eta))^2 \rho_\eta \, \mathrm{d}\eta = \int \eta^2 \rho_\eta \, \mathrm{d}\eta - \mathrm{E}(\eta)^2. \tag{2.17}$$

Da der Systemausgang η abhängig von Eingang ξ ist, können die auftauchenden Integrale der Momente erster und zweiter Ordnung umgeschrieben werden zu

$$\mu_1 = \int \eta \rho_\eta \, \mathrm{d}\eta \; = \int g(\xi) \rho_\xi \, \mathrm{d}\xi \tag{2.18}$$

$$\mu_2 = \int \eta^2 \rho_\eta \, \mathrm{d}\eta = \int g(\xi)^2 \rho_\xi \, \mathrm{d}\xi. \tag{2.19}$$

Wie schon in den vorigen Abschnitten erläutert, ist der funktionale Zusammenhang $g(\xi)$ in der Regel unbekannt, weshalb die Integrale nicht direkt ausgewertet werden können. Oft werden aus diesem Grund approximative Verfahren verwendet, die eine numerische Integration ermöglichen. Die Gauß-Quadratur stellt eines dieser Verfahren dar. Sie beruht auf der Ersetzung des Integrals einer Funktion $f(\xi)$ durch eine gewichtete Summe von Funktionswerten

$$\int_{-1}^{1} f(\xi) \, \mathrm{d}\xi \approx \sum_{i}^{n_L} w_i f(\xi_i). \tag{2.20}$$

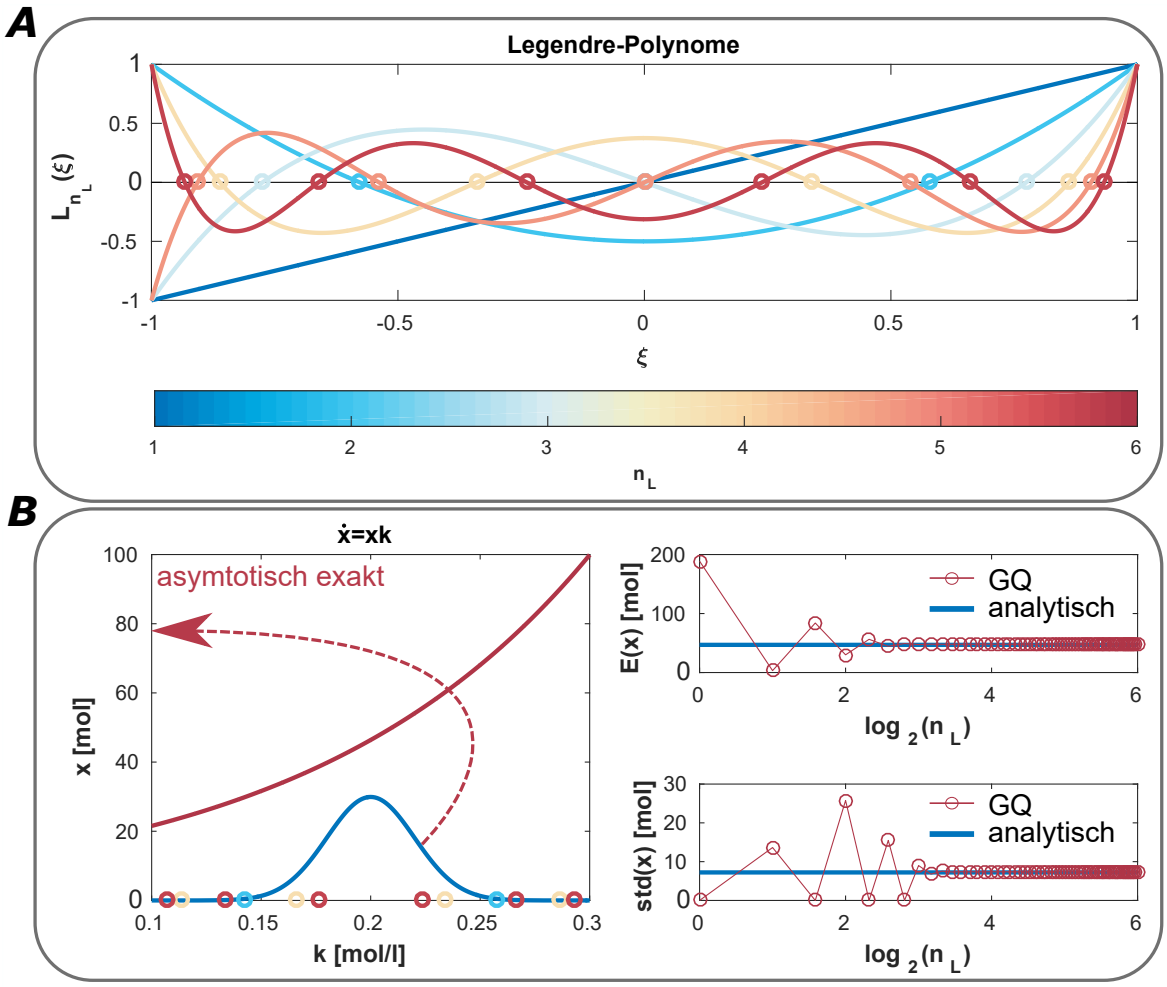

Abbildung 2.6: Approximation mittels Gauß-Quadratur. **(A)** Die Nullstellen der Legendre-Polynome n_L-ter Ordnung werden zur approximativen Berechnung von Integralen benutzt. Dabei bezeichnen rote Kurven große und blaue Kurven kleine Ordnungen. **(B)** Zur Berechnung des Mittelwertes und der Standardabweichung der Reaktion erster Ordnung werden die Nullstellen der Legendre-Polynome benutzt. Mit steigender Ordnung nähern sich die Approximationen mittels Gauß-Quadratur an die analytisch berechneten Werte des Mittelwertes und der Standardabweichung an.

Die Stellen ξ_i, an denen die Funktion ausgewertet wird, bezeichnen Nullstellen der Legendre-Polynome $L_{n_L}(\xi)$ der Ordnung n_L, siehe Abb. 2.6A. Zudem lassen sich die Gewichte w_i der Funktionswerte $f(\xi_i)$ mittels

$$w_i = \frac{2}{(1-\xi_i)^2 \left.\frac{\mathrm{d}L_{n_L}}{\mathrm{d}\xi}\right|_{\xi=\xi_i}} \tag{2.21}$$

berechnen [Press, 2002]. Der Vorteil bei der Berechnung der statistischen Momente mittels numerischer Integration besteht darin, dass keine Ableitungen der nichtlinearen Funktion $f(\xi)$ benötigt werden. Die Funktion wird lediglich an diskreten Punkten

ausgewertet, weshalb auch statistische Momente nicht differenzierbarer Funktionen berechnet werden können. Die Gauß-Quadratur liefert exakte Ergebnisse für Polynome bis zum Grad $2n_L - 1$ [Wu et al., 2006]. Für Funktionen, die keine Polynome sind, oder Polynome höheren Grades stellt diese Methode demnach nur eine Näherung dar, die umso präziser ist, je besser sich $f(\xi)$ durch ein Polynom annähern lässt. Zur Illustration der Gauß-Quadratur wird das bereits in Abb. 2.5C betrachtete Beispiel der Reaktion erster Ordnung untersucht. Dabei werden Mittelwert und Varianz unter Zuhilfenahme von Gl. 2.18-2.19 berechnet. Obwohl die Legendre-Polynome lediglich im Intervall $[-1, 1]$ definiert sind, lässt sich das in Gl. 2.20 gezeigte Integral für beliebige Intervalle $[\xi_-, \xi_+]$ erweitern

$$\int_{\xi_-}^{\xi_+} f(\xi)\,\mathrm{d}\xi \approx \frac{\xi_+ - \xi_-}{2} \sum_i^{n_L} w_i f\left(\frac{\xi_+ - \xi_-}{2}\xi_i + \frac{\xi_- + \xi_+}{2}\right). \tag{2.22}$$

Zur Veranschaulichung dieser Methode wird die Reaktion erster Ordnung mit einer normalverteilten Reaktionsgeschwindigkeitskonstante untersucht, siehe Abb. 2.6B. In diesem Fall wird das Intervall zu $[\mathrm{E}(k) - 5\mathrm{std}(k), \mathrm{E}(k) + 5\mathrm{std}(k)]$ gesetzt. Es stellt sich heraus, dass sich Mittelwert und Standardabweichung mit zunehmender Ordnung n_L genauer an die analytisch berechneten Werte annähern. Die Standardabweichung konvergiert dabei langsamer als der Mittelwert, da für ihre Berechnung das Moment zweiter Ordnung ermittelt werden muss. Dieses ist stark nichtlinear und kann darum nur durch ein Polynom hohen Grades hinreichend genau approximiert werden.

Selbst für dieses einfache System müssen die Legendre-Polynome bis zur etwa zehnten Ordnung ausgewertet werden, um die Momente präzise anzunähern. Mit zunehmender Komplexität des Modells und damit der Funktion $f(\xi)$ steigt die benötigte Ordnung der Legende-Polynome weiter. Für bestimmte Funktionen des Typs $f(\xi) = g(\xi)\rho(\xi)$ ist es möglich, eine geeignetere Wahl der orthogonalen Polynome und deren Wichtung zu treffen, die eine schnellere Konvergenz ermöglicht [Wu et al., 2006]. Zudem skaliert die Gauß-Quadratur exponentiell mit der Dimension des Systems, weshalb selten Ordnungen höher als fünf verwendet werden [Wu et al., 2006]. Es gibt zwar Ansätze, wie Sparse Sampling, die dem Fluch der Dimension entgegenwirken [Bungartz et al., 2004]. Dennoch ist der Rechenaufwand für die meisten realen Probleme zu hoch, um diese Methode erfolgreich anzuwenden.

2.3.3 Approximation mittels Monte Carlo Simulationen

Monte Carlo (MC) Simulationen sind approximative Verfahren, die zur numerischen Berechnung von Integralen benutzt werden. Sie sind nicht nur in der Lage statistische Momente zu berechnen, sondern können auch zur Bestimmung von Wahrschein-

lichkeitsdichten genutzt werden. Dazu wird, ähnlich wie bei der Gauß-Quadratur, eine diskrete Anzahl von Realisierungen (Samples) der Zufallsvariable ξ transformiert $\eta_i = g(\xi_i)$. Die Samples werden jedoch nicht deterministisch gewählt, sondern stellen zufällige Realisierungen der Wahrscheinlichkeitsdichte ρ_ξ dar [James, 1980]. Der Mittelwert und die Varianz dieser Stichprobe ergeben sich zu

$$\mathrm{E}(\eta) = \frac{1}{n_{MC}} \sum_i^{n_{MC}} \eta_i, \tag{2.23}$$

$$\mathrm{Var}(\eta) = \frac{1}{n_{MC} - 1} \sum_i^{n_{MC}} (\eta_i - \mathrm{E}(\eta))^2. \tag{2.24}$$

Dabei bezeichnet n_{MC} die Anzahl der verwendeten Samples. Zusätzlich lässt sich die Wahrscheinlichkeitsdichte der transformierten Samples mittels Kerndichteschätzung bestimmen. Diese wird dazu als Superposition einzelner Kerne K (Kerndichteschätzung) dargestellt

$$\rho_\eta \approx \frac{1}{n_{MC}} \sum_i^{n_{MC}} K(\eta - \eta_i), \tag{2.25}$$

die ebenfalls als Wahrscheinlichkeitsdichten interpretiert werden können [Pérez et al., 2009]. Grundsätzlich lässt sich sagen, dass breite Kerne besonders glatte Verteilungen ergeben, die lokale Eigenschaften der eigentlichen Dichte maskieren oder verschmieren können. Mit steigender Anzahl der generierten Samples können jedoch immer schmalere Kerne gewählt werden, die eine genauere Schätzung der Dichte erlauben. Sowohl die statistischen Momente, als auch die Wahrscheinlichkeitsdichte werden mittels MC Simulationen asymptotisch exakt bestimmt.

Zur Veranschaulichung dieser Methode wird erneut die bereits gezeigte Reaktion erster Ordnung mit einer normalverteilten Reaktionsgeschwindigkeitskonstante untersucht. Es werden zufällige Realisierungen aus dieser Wahrscheinlichkeitsdichte generiert, die anschließend in die Ratengleichung eingesetzt werden, um den zeitlichen Verlauf des Systemzustandes zu berechnen, siehe Abb. 2.7A. Aus den einzelnen Trajektorien können Mittelwert, Standardabweichung und die Wahrscheinlichkeitsdichte zu jedem Zeitpunkt berechnet werden. Da MC Simulationen auf der Generierung von Zufallszahlen beruhen, liefern wiederholte Simulationen stets unterschiedliche Ergebnisse. Die Unsicherheit der ermittelten statistischen Größen skaliert mit $\frac{1}{\sqrt{n_{MC}}}$ und nimmt demnach mit zunehmender Anzahl der verwendeten Samples ab. Dies ist für wiederholte MC Simulationen zur Bestimmung des Mittelwertes in Abb. 2.7A skizziert. Mit zunehmender Anzahl der Samples gleichen sich die Ergebnisse der einzelnen MC Simulationen an, weshalb die Standardabweichung und damit die Unsicherheit des Mittelwertes abnimmt. Zur Beschleunigung der Konvergenz werden oft Quasi Monte Carlo Verfahren, wie zum Beispiel Latin Hyper Cube Sampling, verwendet. Im Gegensatz zum MC

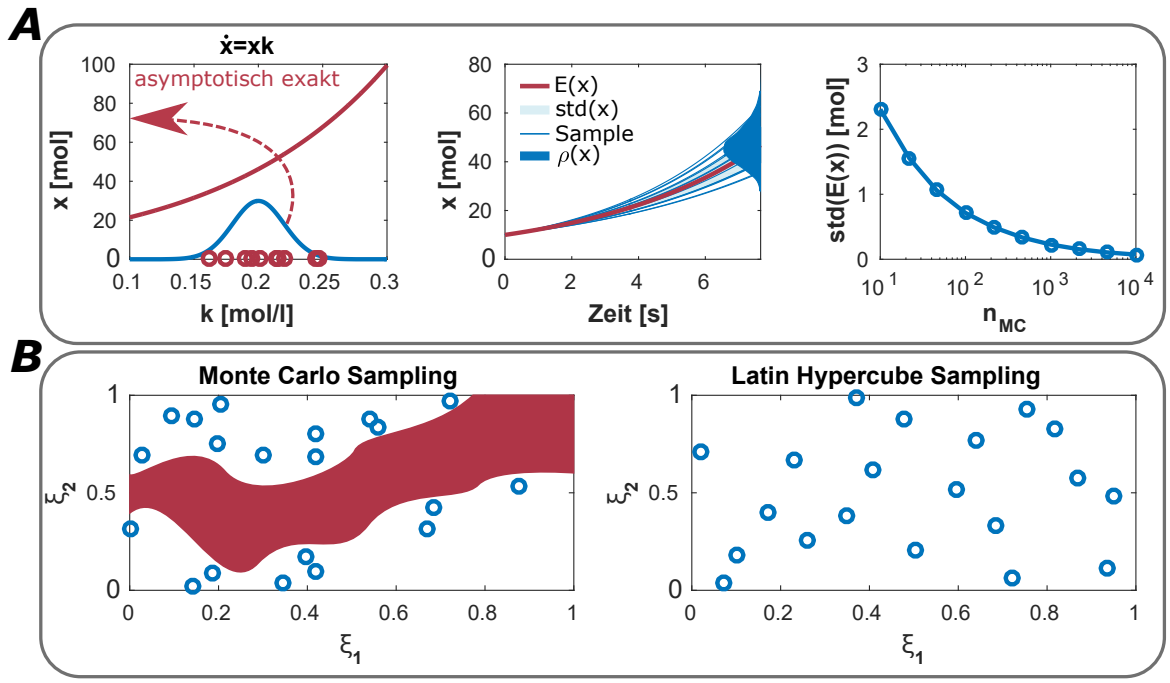

Abbildung 2.7: Berechnung statistischer Größen mittels Monte Carlo Simulation. **(A)** Zur statistischen Beschreibung der Reaktion erster Ordnung werden Zufallszahlen (Samples) der Wahrscheinlichkeitsdichte gezogen. Für jedes Sample wird eine Trajektorie berechnet, aus denen der Mittelwert E($\boldsymbol{x}$), die Standardabweichung std($\boldsymbol{x}$) und die Wahrscheinlichkeitsdichte $\rho(\boldsymbol{x})$ zum Endzeitpunkt bestimmt wird. Je größer die Anzahl der verwendeten Sample ist, desto kleiner wird die Unsicherheit der statistischen Größen. Dies ist exemplarisch für den Mittelwert gezeigt worden (asymptotisch exakte Konvergenz). **(B)** Eine Alternative zum Monte Carlo Sampling stellt Latin Hyper Cube Sampling dar, welches sich durch eine schnellere Konvergenz auszeichnet. Für eine zweidimensionale gleichverteilte Wahrscheinlichkeitsdichte sind beide Verfahren dargestellt. Wegen der geringen Anzahl von Samples sind beim Monte Carlo Sampling große Bereiche nicht exploriert (rot), wohingegen andere Bereiche sehr dicht besetzt sind. Im Gegensatz dazu stellt Latin Hyper Cube Sampling auch für eine geringe Anzahl von Samples eine gute Näherung der Gleichverteilung dar.

Sampling werden hier keine zufälligen Samples generiert, sondern Sequenzen benutzt, die eine gleichmäßige Verteilung der Samples bewirken [James, 1980]. Dies ist am Beispiel einer zweidimensionalen Gleichverteilung in Abb. 2.7B verdeutlicht.

2.3.4 Approximation mittels der Sigma-Punkt-Methode

Die bisher präsentierten Methoden dienen ausschließlich dazu, einen Überblick bezüglich der Modellierung extrinsischer Störungen zu vermitteln. Dabei sind lediglich die für diese Arbeit relevanten Verfahren vorgestellt worden. Auf andere populäre Methoden, wie das Importance Sampling oder die Polynomial Chaos Expansion, wird in

dieser Arbeit nicht eingegangen. Bislang sind der Einfachheit halber eindimensionale Probleme betrachtet worden, da die verwendeten Methoden daran anschaulich diskutiert werden können und die mathematischen Ausdrücke eine besonders simple Form annehmen. Bei der Sigma-Punkt-Methode, die bei dieser Arbeit im Fokus steht, bietet es sich jedoch an, die Herleitung mittels einer n_ξ-dimensionalen Zufallsvariablen $\boldsymbol{\xi}$ zu beginnen. Ähnlich wie die Gauß-Quadratur und Monte Carlo Sampling transformiert die Sigma-Punkt-Methode eine diskrete Anzahl von Samples zur approximativen Berechnung von Integralen [Julier et al., 2000, 2004]. Dazu wird ein Satz von $2n_\xi + 1$ Sigma-Punkten $\boldsymbol{\xi}_i$ gewählt, dessen Mittelwert $\mathrm{E}(\boldsymbol{\xi})$ und Kovarianz $\mathrm{Cov}(\boldsymbol{\xi})$ mit denen der zu transformierenden Verteilung übereinstimmen

$$\boldsymbol{\xi}_0 = \mathrm{E}(\boldsymbol{\xi}) \tag{2.26}$$

$$\boldsymbol{\xi}_i = \mathrm{E}(\boldsymbol{\xi}) + \left(\sqrt{(n_\xi + \kappa)\mathrm{Cov}(\boldsymbol{\xi})}\right)_i \tag{2.27}$$

$$\boldsymbol{\xi}_{i+n_\xi} = \mathrm{E}(\boldsymbol{\xi}) - \left(\sqrt{(n_\xi + \kappa)\mathrm{Cov}(\boldsymbol{\xi})}\right)_i. \tag{2.28}$$

Dabei bezeichnet κ eine reelle Konstante und $\left(\sqrt{(n_\xi + \kappa)\mathrm{Cov}(\boldsymbol{\xi})}\right)_i$ die i-te Spalte der Quadratwurzel der skalierten Kovarianzmatrix. Anschließend werden die Sigma-Punkte mittels der nichtlinearen Funktion $\boldsymbol{\eta}_i = \boldsymbol{g}(\boldsymbol{\xi}_i)$ transformiert. Der Mittelwert und die Kovarianz ergeben sich dann zu

$$\mathrm{E}(\boldsymbol{\eta}) = \sum_0^{2n_\xi} w_i \boldsymbol{\eta}_i \tag{2.29}$$

$$\mathrm{Cov}(\boldsymbol{\eta}) = \sum_0^{2n_\xi} w_i (\boldsymbol{\eta}_i - \mathrm{E}(\boldsymbol{\eta})) \cdot (\boldsymbol{\eta}_i - \mathrm{E}(\boldsymbol{\eta}))^\top \tag{2.30}$$

mit den Wichtungen

$$w_0 = \frac{\kappa}{n_\xi + \kappa} \tag{2.31}$$

$$w_i = \frac{1}{2(n_\xi + \kappa)} \tag{2.32}$$

$$w_{i+n_\xi} = \frac{1}{2(n_\xi + \kappa)}. \tag{2.33}$$

Dies ist für die Reaktion erster Ordnung, aber auch für ein zweidimensionales Beispiel in Abb. 2.8A illustriert. Der freie Parameter κ regelt den Abstand der Sigma-Punkte vom Mittelwert der Verteilung und kann beliebig gewählt werden. Es zeigt sich, dass für $\kappa = 3 - n_\xi$ der Fehler vierter Ordnung der Kovarianz für eine normalverteilte Zufallsvariable minimiert wird [Julier et al., 2000]. Dieses Verfahren ist jedoch nicht ausschließlich für normalverteilte Zufallsvariablen geeignet, sondern findet in leicht

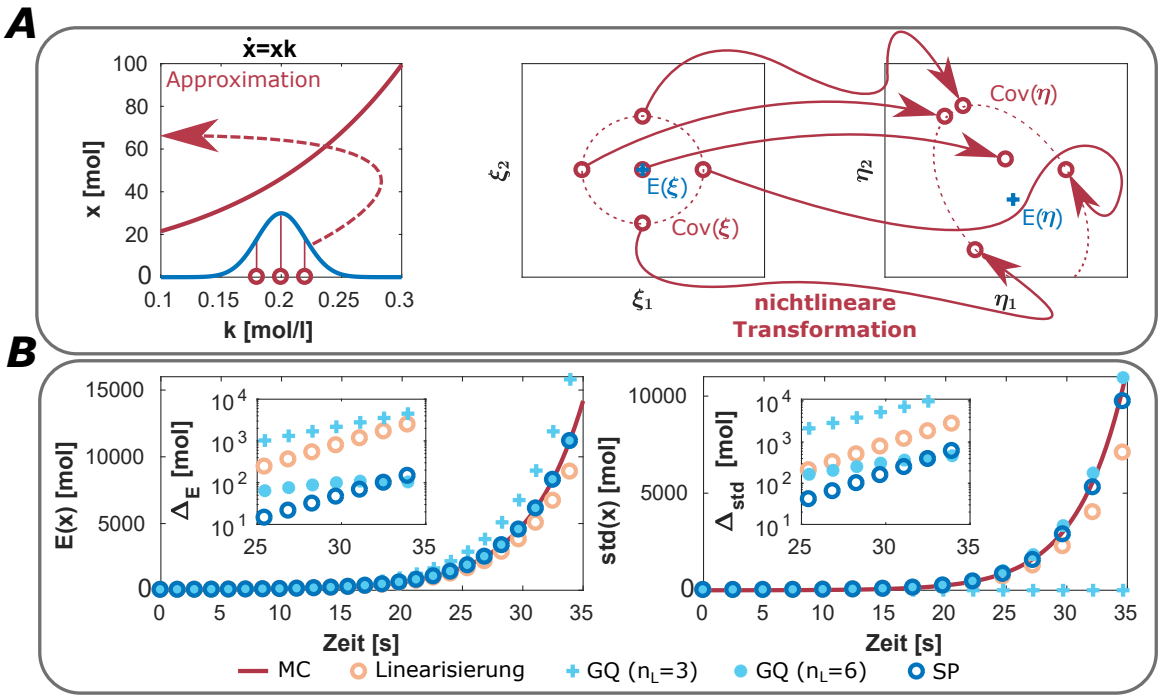

Abbildung 2.8: Approximation mittels Sigma-Punkt-Methode. **(A)** Die Sigma-Punkt-Methode wählt lediglich $2n_\xi + 1$ Samples, die mittels der nichtlinearen Abbildung transformiert werden. Dies ist für die bereits gezeigte Reaktion erster Ordnung demonstriert, aber auch für ein zweidimensionales Beispiel. Aus den transformierten Samples können der Mittelwert und die Varianz geschätzt werden. Für nichtlineare Transformationen stellen diese jedoch nur Näherungen dar. **(B)** Zur Bestimmung des Mittelwertes und der Standardabweichung der Reaktion erster Ordnung werden Linearisierung, Gauß-Quadratur, Monte Carlo Simulation und Sigma-Punkt-Methode verglichen. Der Fehler des Mittelwertes Δ_E und der Standardabweichung Δ_{std} bezogen auf die asymptotisch exakte Monte Carlo Simulation zeigen in diesem Fall, dass die Sigma-Punkt-Methode die beste Näherung darstellt.

abgewandelter Form auch für asymmetrische Verteilungen Anwendung [Tenne et al., 2003; Flassig et al., 2012].

Die Sigma-Punkt-Methode grenzt sich klar von den bislang präsentierten Verfahren ab [Wu et al., 2006]. Zum einen beruht sie nicht, wie die Linearisierung, auf der Berechnung von Ableitungen, um die Transformation durch ein Ersatzmodell anzunähern. Aus diesem Grund bereiten unstetige Transformationen keine Probleme und können einfach gehandhabt werden. Dennoch ist die Sigma-Punkt-Methode eine Näherung, die ausschließlich für Polynome dritten Grades oder niedriger exakte Ergebnisse liefert [Adurthi et al., 2017]. Im Gegensatz dazu können mittels Gauß-Quadratur und Monte Carlo Simulationen die statistischen Momente asymptotisch exakt bestimmt werden. Der rechnerische Aufwand ist jedoch für beide Verfahren immens. Die Gauß-Quadratur skaliert exponentiell mit der Dimension des Systems $\mathcal{O}(n_L^{n_\xi})$, weshalb diese Methode nur für kleine Systeme anwendbar ist. Monte Carlo Simulationen sind zwar nicht

vom Fluch der Dimension betroffen, dennoch muss eine enorme Anzahl von Samples transformiert werden, um den Approximationsfehler zu minimieren. Die Sigma-Punkt-Methode stellt im Kontrast dazu einen exzellenten Kompromiss aus Aufwand und Präzision dar, da sie lediglich mit $\mathcal{O}(2n_\xi + 1)$ skaliert. Um dies zu verdeutlichen, wird in Abb. 2.8B die zeitliche Entwicklung des Mittelwertes und der Standardabweichung der bereits betrachteten Reaktion erster Ordnung untersucht. Die Monte Carlo Simulation mit 10^6 Samples stellt eine zuverlässige Referenzlösung dar, die mit der Linearisierung, der Gauß-Quadratur ($n_L = 3$ und $n_L = 6$) und der Sigma-Punkt-Methode verglichen wird. Aus diesem Grund wird die Abweichung des Mittelwertes Δ_E und der Standardabweichung Δ_{std} von den Ergebnissen der Monte Carlo Simulation als Maß für die Güte der Näherung angesehen. Hinsichtlich beider Größen ist zu erkennen, dass die Linearisierung und Gauß-Quadratur der Ordnung $n_L = 3$ stark von eigentlichen Kurvenverlauf abweichen und deshalb große Fehler aufweisen. Die Sigma-Punkt-Methode und Gauß-Quadratur der Ordnung $n_L = 6$ stellen jedoch akzeptable Näherungen dar. Dabei ist anzumerken, dass die Sigma-Punkt-Methode für fast alle Zeitpunkte genauere Ergebnisse liefert, obwohl die gewöhnliche Differenzialgleichung lediglich dreimal ausgewertet wird. Im Gegensatz dazu benötigt die Gauß-Quadratur der Ordnung $n_L = 6$ sechs Auswertungen zur Berechnung der statistischen Momente. Damit sind an einem eindimensionalen, simplen Beispiel die Vorzüge der Sigma-Punkt-Methode im Vergleich zu konventionellen Verfahren zur Propagation von Unsicherheiten gezeigt worden.

Aufgrund der Einfachheit der Implementierung, ihrer numerischen Stabilität und der Genauigkeit bei geringem Rechenaufwand findet die Sigma-Punkt-Methode vielfältige Anwendung in der Biologie [Schenkendorf et al., 2009; Flassig et al., 2012], der Verfahrenstechnik [Kaiser et al., 2016] und weiteren Disziplinen [Wu et al., 2006]. Zudem gibt es zahlreiche Erweiterungen und verwandte Arten der hier besprochenen Form der Sigma-Punkt-Methode [Wu et al., 2006; Menegaz et al., 2015; Afshari et al., 2017]. Im weiteren Verlauf dieser Arbeit wird gezeigt, wie zusätzliche Modifikationen hinsichtlich der Simulation von Systemen, die intrinsischen, extrinsischen und externen Störungen unterworfen sind, vorgenommen werden können [Pischel et al., 2016, 2017]. Diese Modifikationen stellen einen Kernpunkt der Arbeit dar, die in Abschnitt 2.6 genauer untersucht werden.

2.4 Simulation intrinsischer Störungen

Auf molekularer Ebene können chemische Reaktionen als stochastische, ungeordnete Kollisionen einzelner Teilchen betrachtet werden [Gillespie, 1977]. Damit eine Reaktion

stattfindet, müssen die Zusammenstöße entlang der Verbindungslinie der Atomkerne erfolgen und dabei eine gewisse Aktivierungsenergie überschreiten. In Systemen mit einer großen Anzahl von Teilchen stoßen diese ständig zusammen und Reaktionen finden ununterbrochen statt. Einzelne Stöße, die den Systemzustand ändern, können deshalb nicht beobachtet werden, weshalb die Abundanzen der chemischen Spezies als Kontinuum betrachtet werden können. Wird die Systemgröße bei gleichbleibender Konzentration der chemischen Spezies verringert, sollte laut der Ratengleichung keine Veränderung der Systemdynamik verzeichnet werden. Es zeigt sich jedoch, dass sehr kleine Systeme mit geringen Abundanzen aufgrund intrinsischer Störungen stark vom Verlauf der Ratengleichung abweichen können [Wilkinson, 2009], siehe Abb. 2.9A. Diese Systeme verzeichnen weniger Stöße, die zu Reaktionen führen, weshalb die stochastische Komponente der ungeordneten Teilchenbewegung sehr dominant ist. Chemische Reaktionen können deshalb als diskrete, zufällige Ereignisse identifiziert werden, die einen Sprung des Systemzustandes in einen anderen Zustand induzieren. Die Änderung des Systemzustandes wird dabei durch die Stöchiometrie der stattfindenden Reaktion bestimmt. Da bei einer Reaktion ausschließlich ganzzahlige Mengen an Molekülen gebildet bzw. verbraucht werden, stellen die Abundanzen der chemischen Spezies ebenfalls diskrete Größen dar $\boldsymbol{x}(t) \in \mathbb{N}_0^{ns}$ [Gillespie, 1977]. Intrinsische Störungen sind besonders ausgeprägt bei der Genexpression, da nur wenige Kopien eines Gens innerhalb des Zellkerns existieren [Paulsson, 2005; Kaufmann et al., 2007; Eldar et al., 2010]. Durch ihren binären Wechsel vom aktiven in den inaktiven Zustand und umgekehrt sorgen sie für rasche Änderungen und Schwankungen der Abundanzen nachgeschalteter Produkte, siehe Abb. 2.9B. Die stochastische Genexpression stellt die Grundlage fundamentaler Prozesse dar, die in biologischer Variabilität und Vielfalt resultieren. Neben der Heterogenität mikrobieller Populationen [Ackermann, 2015] lassen sich auch Prozesse, wie Zelldifferenzierung [Eldar et al., 2010; Chubb, 2017] oder die Bildung von Krebszellen [Capp, 2017], darauf zurückführen. Im Laufe dieser Arbeit wird gezeigt werden, dass auch Apoptose von der intrinsischen Störung der Genexpression betroffen ist [Buchbinder et al., 2018].

Mathematisch interpretiert stellen intrinsische Störungen lediglich zufällige Sprünge des Systemzustandes in einem diskreten Zustandsraum dar. Die Änderung des Zustandes wird ausschließlich vom momentanen Zustand bestimmt und ist unabhängig von dessen Vorgeschichte, weshalb Systeme dieser Art adäquat mittels eines Markov-Prozesses modelliert werden können [Gillespie, 1977]. Der Systemzustand stellt dabei, analog zur deterministischen Betrachtungsweise der Ratengleichung, einen Vektor mit den Abundanzen der chemischen Spezies dar. Die Abundanzen sind jedoch nicht eindeutig bestimmt, sondern werden durch Zufallsvariablen beschrieben. Änderungen des Zustandes geschehen nur durch das Eintreten von Reaktionen. Tritt die Reaktion J

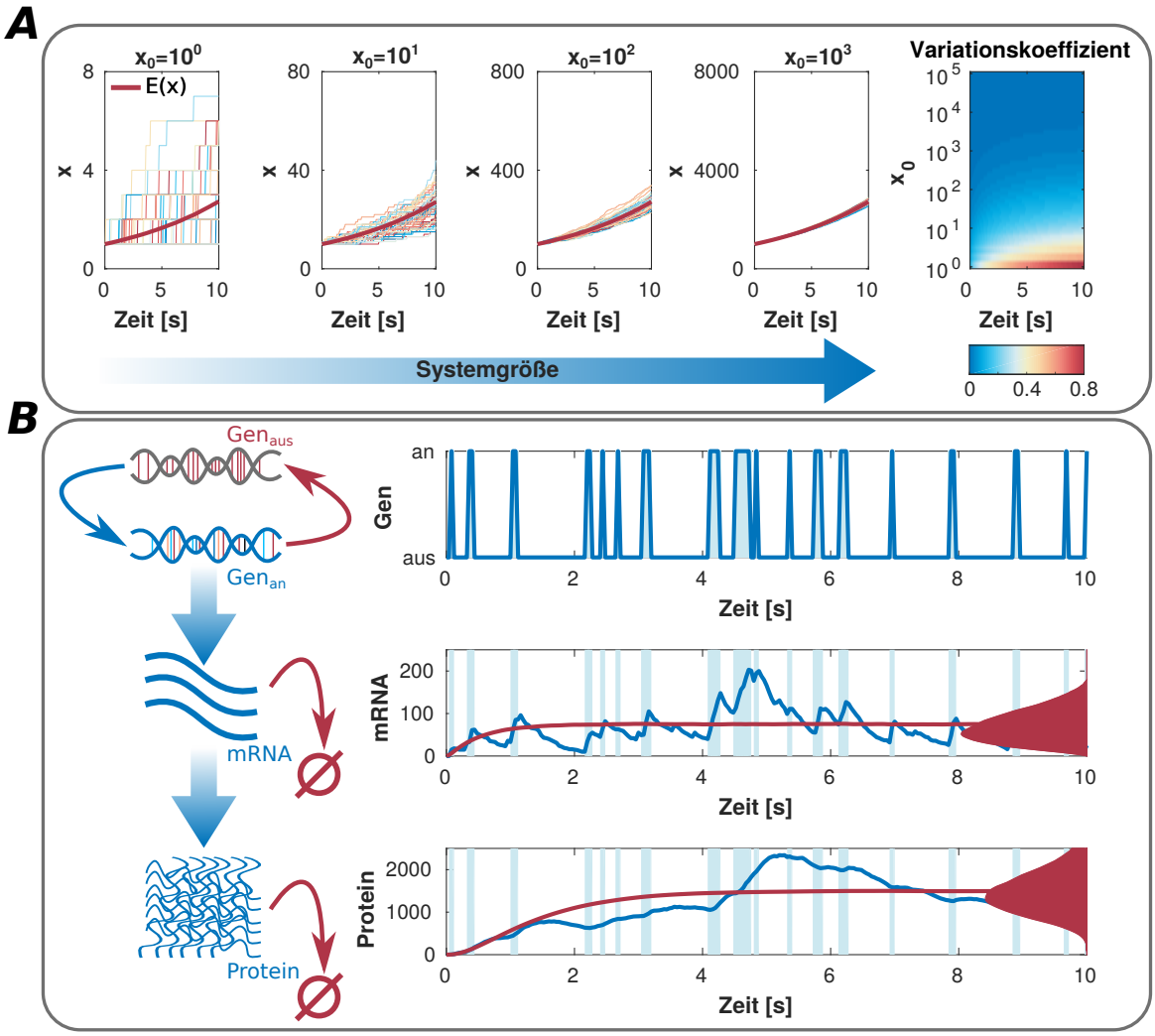

Abbildung 2.9: Intrinsische Störungen in biochemischen Reaktionssystemen. **(A)** Am Beispiel der Reaktion erster Ordnung wird gezeigt, dass mit steigender Systemgröße $\boldsymbol{x}_0$ (initialer Systemzustand) intrinsische Störungen abnehmen und sich asymptotisch der Ratengleichung annähern. Dies ist gut an der Streuung der einzelnen Trajektorien zuerkennen, aber auch am Variationskoeffizienten. Die Simulationen sind mittels des Gillespie-Algorithmus durchgeführt worden. **(B)** Gene können in einem aktiven (an) oder inaktiven (aus) Zustand vorkommen. Aufgrund der geringen Anzahl der Kopien einzelner Gene sind sie besonders von intrinsischen Störungen betroffen. Der stochastische Prozess der Aktivierung bzw. Inaktivierung verursacht große Schwankungen nachgeschalteter Produkte, wie mRNA und Proteine. Die roten Kurven bezeichnen den Mittelwert, wohingegen in blau eine einzelne Realisierung gekennzeichnet ist. Hellblaue Schatten markieren die Zeitpunkte, bei denen das Gen aktiv ist und die rote Verteilung zeigt die Dichte der chemischen Spezies zum Endzeitpunkt.

ein, springt der Systemzustand in einen benachbarten gemäß der Reaktionsstöchiometrie

$$\boldsymbol{x}(t) \leftarrow \boldsymbol{x}(t) + \boldsymbol{N}_J, \tag{2.34}$$

Tabelle 2.1: Propensitäten verschiedener Reaktionen [Gillespie, 1976].

Reaktion	$h_j(\boldsymbol{x})$	c_j
$\emptyset \rightarrow \ldots$	1	$k_j\Omega$
$\mathrm{S}_i \rightarrow \ldots$	S_i	k_j
$\mathrm{S}_i + \mathrm{S}_l \rightarrow \ldots$	$\mathrm{S}_i\mathrm{S}_l$	$\frac{k_j}{\Omega}$
$2\mathrm{S}_i \rightarrow \ldots$	$\frac{1}{2}\mathrm{S}_i(\mathrm{S}_i - 1)$	$\frac{k_j}{\Omega}$
$\mathrm{S}_i + 2\mathrm{S}_l \rightarrow \ldots$	$\frac{1}{2}\mathrm{S}_i\mathrm{S}_l(\mathrm{S}_l - 1)$	$\frac{2k_j}{\Omega^2}$
$3\mathrm{S}_i \rightarrow \ldots$	$\frac{1}{6}\mathrm{S}_i(\mathrm{S}_i - 1)(\mathrm{S}_i - 2)$	$\frac{6k_j}{\Omega^2}$

wobei $\boldsymbol{N}_J$ die J-te Spalte der stöchiometrischen Matrix bezeichnet. Der Index der nächsten Reaktion J sowie der Zeitpunkt ihres Eintretens stellen Zufallsvariablen dar, die von den stochastischen Übergangsraten (Propensitäten) $\boldsymbol{a}(\boldsymbol{x})$ abhängen. Die Propensitäten können analog zu den deterministischen Reaktionsraten aus den Reaktionsgleichungen abgeleitet werden. Üblicher Weise werden sie in Faktoren zerlegt, die vom Reaktionsmechanismus $h_j(\boldsymbol{x})$ bzw. den Systemeigenschaften c_j abhängen

$$a_j(\boldsymbol{x}) = h_j(\boldsymbol{x})c_j. \tag{2.35}$$

Eine Zusammenfassung der wichtigsten Propensitäten sowie deren Bezug zu den deterministischen Reaktionskonstanten k_j und dem Systemvolumen Ω ist in Tab. 2.1 gegeben [Gillespie, 1992].

Da chemische Reaktion intrinsisch stochastisch sind, können lediglich Wahrscheinlichkeitsaussagen bezüglich des Systemzustandes getroffen werden. Die zeitliche Entwicklung der Wahrscheinlichkeit $P(\boldsymbol{x})$, das System in einem bestimmten Zustand anzutreffen, wird durch die chemische Master-Gleichung (CME, chemical master equation) bestimmt

$$\frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}t} = \sum_{j=1}^{n_R} a_j(\boldsymbol{x} - \boldsymbol{N}_j)P(\boldsymbol{x} - \boldsymbol{N}_j) - a_j(\boldsymbol{x})P(\boldsymbol{x}). \tag{2.36}$$

Diese Gleichung stellt eine Differenzial-Differenzengleichung dar (kontinuierliche Zeit, aber diskrete Zustände), die zu einem System von gewöhnlichen Differenzialgleichungen umgeformt werden kann [Munsky et al., 2006]. Der zeitliche Verlauf der Wahrscheinlichkeit jedes möglichen Zustandes wird dabei durch eine Differenzialgleichung beschrieben. Aufgrund der Tatsache, dass die meisten biochemischen Systeme eine sehr große Anzahl von möglichen Zuständen besitzen, erweist es sich als besonders schwierig die CME zu lösen (Fluch der Dimension). Lediglich einfache Spezialfälle, wie unimolekulare Reaktionen, können analytisch gelöst werden [Gillespie, 2007; Jahnke et al.,

2007]. Infolgedessen steht die Entwicklung approximativer Verfahren zur Lösung der CME im Fokus der aktuellen Forschung [Gillespie, 2007; Gillespie et al., 2013]. Im Weiteren werden die wichtigsten Vertreter dieser Methoden sowie deren Stärken und Schwächen vorgestellt.

2.4.1 Gillespie-Algorithmus

Die Tatsache, dass die CME selbst für sehr einfache biochemische Reaktionsnetzwerke nicht lösbar ist, hat Wissenschaftler rasch auf die Idee gebracht, einen komplementären Weg einzuschlagen und stattdessen zufällige Trajektorien oder Realisierungen des Systems zu berechnen, die konsistent mit der CME sind [Wilkinson, 2009]. Aus den Realisierungen können anschließend die statistischen Momente sowie die Wahrscheinlichkeit, das System in einem gewissen Zustand zu finden, bestimmt werden. Die frühesten Ansätze dieser Art betrachteten lediglich sehr spezielle Systeme mit heuristischen Näherungen [Nakanishi, 1972; Bunker et al., 1974]. Erst mit der Entwicklung des Gillespie-Algorithmus [Gillespie, 1977] wurde eine exakte Methode etabliert, die allgemein anwendbar ist und auf soliden physikalischen Annahmen beruht [Gillespie et al., 2013]. Zur Berechnung zufälliger Realisierungen der CME müssen zwei Fragen beantwortet werden:

- Wie groß ist das Zeitintervall τ^* bis zur nächsten Reaktion?
- Welche Reaktion J tritt ein?

Es kann gezeigt werden, dass die Wahrscheinlichkeit der gesuchten Größen in Abhängigkeit des momentanen Systemzustands gegeben ist durch

$$P(\tau^*, j|\boldsymbol{x}) = \mathrm{e}^{-a_0(\boldsymbol{x})\tau^*} a_j(\boldsymbol{x}) \tag{2.37}$$

mit $a_0(\boldsymbol{x}) = \sum_{j=1}^{n_R} a_j(\boldsymbol{x})$ [Gillespie, 1976]. Mittels der Erzeugung zweier gleichverteilter Zufallszahlen r_1 und r_2 aus dem Intervall $[0, 1]$ können daraus das Zeitintervall bis zur nächsten Reaktion

$$\tau^* = \frac{1}{a_0(\boldsymbol{x})} \ln\left(\frac{1}{r_1}\right) \tag{2.38}$$

und dessen Index J

$$\sum_{j=1}^{J-1} a_j(\boldsymbol{x}) < r_2 a_0(\boldsymbol{x}) \leq \sum_{j=1}^{J} a_j(\boldsymbol{x}) \tag{2.39}$$

bestimmt werden. Durch systematisches Aktualisieren des Systemzustandes und der Zeit sowie sukzessive Wiederholung der gesamten Prozedur können Realisierungen bis zu einem gewünschten Endzeitpunkt t_{end} berechnet werden. Ähnlich wie bei der Monte

Carlo Simulation liefert die Berechnung unabhängiger Realisierungen der Systemtrajektorien ein umfassendes Bild der Dynamik des Reaktionsnetzwerkes. Mit zunehmender Anzahl der berechneten Realisierungen konvergiert die Lösung asymptotisch gegen die exakte Lösung der CME. Eine knappe Veranschaulichung des Gillespie-Algorithmus in Pseudocode ist in Alg. 2.1 dargestellt. Zudem sind die in Abb. 2.9 gezeigten Kurven mit dem Gillespie-Algorithmus berechnet worden.

Algorithmus 2.1: Gillespie-Algorithmus [Gillespie, 1977]

Ergebnis: Berechnung zufälliger Systemtrajektorien

Initialisierung: $t \leftarrow t_0$, $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$;

while $t < t_{end}$ **do**

  Berechnung der Propensitäten $\boldsymbol{a}(\boldsymbol{x})$ und deren Summe $a_0(\boldsymbol{x})$;

  Berechnung zweier gleichverteilter Zufallszahlen $r_{1,2}$ aus dem Intervall $[0, 1]$;

  Berechnung des Zeitintervalls bis zur nächsten Reaktion τ^*;

  if $t + \tau^* > t_{end}$ **then**

    Aktualisierung der Zeit $t \leftarrow t_{end}$;

  else

    Berechnung des Index der nächsten Reaktion J;

    Aktualisierung der Zeit $t \leftarrow t + \tau^*$ und des Zustandes

    $\boldsymbol{x}(t + \tau^*) \leftarrow \boldsymbol{x}(t) + \boldsymbol{N}_J$;

  end

end

Der Gillespie-Algorithmus ist eine sehr etablierte Methode, die einfach implementiert werden kann. Es zeigt sich jedoch, dass mit zunehmenden Abundanzen der chemischen Spezies die Anzahl der stattfindenden Reaktionen und damit auch der Rechenaufwand stark ansteigen. Die mittlere Zeit zwischen zwei Reaktionen $\frac{1}{a_0(\boldsymbol{x})}$ ist irgendwann so kurz, dass die Berechnung technisch unmöglich ist [Gillespie et al., 2013]. Aus diesem Grund sind verschiedenste Approximationen dieses Algorithmus entwickelt worden, die auf unterschiedlichen Näherungen beruhen [Gillespie, 2007; Gillespie et al., 2013]. Ausgehend von der Tatsache, dass biochemische Reaktionen eines Netzwerkes oft in langsame und schnelle Reaktionen eingeteilt werden können, bietet es sich an, eine Zeitskalenseparation vorzunehmen [Haseltine et al., 2002, 2005; Cao et al., 2005]. Eine sehr effiziente Formulierung ist durch eine hybride deterministisch-stochastische Beschreibung gegeben, die langsame Reaktionen als diskrete Ereignisse und schnelle Reaktionen als kontinuierliche Prozesse simuliert [Haseltine et al., 2002, 2005]. Dazu werden die schnellen Reaktionen mittels gewöhnlicher Differenzialgleichungen so lange integriert, bis eine langsame Reaktion stattfindet. Der Systemzustand und die Zeit werden entsprechend aktualisiert und die ganze Prozedur wiederholt, bis der Endzeit-

punkt erreicht ist. Ein kurzer Abriss des Algorithmus, wie in [Haseltine et al., 2002] vorgeschlagen, ist in Alg. A.1 zu finden.

Komplementäre Ansätze beruhen auf der Beobachtung, dass der Gillespie-Algorithmus viel Zeit für die Berechnung der Propensitäten nach der Aktualisierung des Systemzustandes benötigt. Es stellt sich jedoch heraus, dass sich die Propensitäten bei hohen Abundanzen der chemischen Spezies nach dem Eintreten einer einzelnen Reaktion nicht signifikant ändern [Gillespie, 2001], weshalb sie für ein gewisses Zeitintervall τ als konstant angenommen werden können. Ansätze dieser Art werden deshalb als τ-Leaping-Algorithmen bezeichnet. Das Zeitintervall τ muss groß genug sein, damit möglichst viele Reaktionen eintreten können, aber gleichzeitig klein genug, um zu gewährleisten, dass die Propensitäten keinen großen Änderungen unterworfen sind [Gillespie, 2001]. Diese Einschränkung wird als Leaping-Bedingung bezeichnet. Zur Bestimmung eines angemessenen Zeitintervalls und der darin erfolgenden Zustandsänderung wird ausgenutzt, dass die Häufigkeit des Eintretens einer Reaktion eine Poisson-verteilte Zufallsvariable mit Mittelwert und Varianz $a_j(\boldsymbol{x}\tau)$ darstellt [Cao et al., 2006]. Abhängig vom Mittelwert und der Varianz der Anzahl der zu erwartenden Reaktionen wird das Zeitintervall τ so gewählt, dass die Leaping-Bedingung stets gewährleistet ist. Die Änderung des Systemzustandes kann anschließend durch die Berechnung Poisson-verteilter Zufallszahlen für jede Reaktion bestimmt werden. Stellt sich heraus, dass die Abundanzen einiger Spezies negativ werden oder dies durch kritische Reaktionen möglicherweise verursacht werden kann, wird für eine gewisse Anzahl von Reaktionen der Gillespie-Algorithmus verwendet oder das Zeitintervall τ verkürzt. Die gesamte Prozedur wird wiederholt, bis der Endzeitpunkt erreicht ist. In Alg. A.2 ist eine Zusammenfassung des Algorithmus in Pseudocode, wie in [Cao et al., 2006] vorgeschlagen, dargestellt.

Neben den hervorgehobenen speziellen Formen der Zeitskalenseparation [Haseltine et al., 2002] und des τ-Leaping-Algorithmus [Cao et al., 2006] gibt es noch weitere Methoden, wie diese umgesetzt werden können [Gillespie, 2007; Gillespie et al., 2013]. Beide der hier vorgestellten Formen des approximativen Gillespie-Algorithmus werden jedoch im späteren Verlauf dieser Arbeit erneut auftauchen. Obwohl die Ursprünge der stochastischen Modellierung biochemischer Reaktionssysteme weit in die Vergangenheit zurückreichen [Gillespie, 1977], zeigt sich, dass dies auch heute noch ein aktueller Forschungszweig ist. Besonders die Implementierung schneller und robuster Algorithmen steht dabei im Vordergrund, um diese Methoden einer breiten Anwenderschaft zugänglich zu machen [Sanft et al., 2011; Somogyi et al., 2015; Kazeroonian et al., 2016; Drawert et al., 2016].

2.4.2 Finite-State-Projection-Algorithmus

Wie bereits diskutiert, ist es möglich, die CME als System von gekoppelten Differenzialgleichungen zu interpretieren. Dabei wird die zeitliche Entwicklung der Wahrscheinlichkeit jedes Zustandes durch eine Differenzialgleichung beschrieben. Mittels Kombination aller möglichen Reaktionen, die in einem gewissen Zustand beginnen oder enden, ist es möglich, dessen zeitliche Entwicklung mithilfe von

$$\frac{\mathrm{d}P(\boldsymbol{x})}{\mathrm{d}t} = \begin{bmatrix} -\sum_{j=1}^{n_R} a_j(\boldsymbol{x}) \\ a_1(\boldsymbol{x}-\boldsymbol{N}_1) \\ \vdots \\ a_{n_R}(\boldsymbol{x}-\boldsymbol{N}_{n_R}) \end{bmatrix}^{\top} \cdot \begin{bmatrix} P(\boldsymbol{x}) \\ P(\boldsymbol{x}-\boldsymbol{N}_1) \\ \vdots \\ P(\boldsymbol{x}-\boldsymbol{N}_{n_R}) \end{bmatrix} \tag{2.40}$$

zu berechnen [Munsky et al., 2006]. Wird eine geeignete Auflistung der Menge aller möglichen Zustände $\boldsymbol{X}$ vorgenommen, lässt sich dessen zeitliche Entwicklung anhand einer linearen Gleichung bestimmen

$$\frac{\mathrm{d}\boldsymbol{P}(\boldsymbol{X})}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{P}(\boldsymbol{X}). \tag{2.41}$$

In diesem Fall bezeichnet $\boldsymbol{P}$ einen Vektor, der jedem möglichen Zustand eine Wahrscheinlichkeit zuordnet. Die Koeffizienten der zeitunabhängigen Matrix $\boldsymbol{A}$ ergeben sich zu

$$A_{ij} = \left\{ \begin{array}{cl} -\sum_{k=1}^{n_R} a_k(\boldsymbol{x}_i) & \text{, wenn } i = j \\ a_k(\boldsymbol{x}_i) & \text{, } \forall\ j \text{ mit } \boldsymbol{x}_j = \boldsymbol{x}_i + \boldsymbol{N}_k \\ 0 & \text{, sonst} \end{array} \right\}, \tag{2.42}$$

wobei $\boldsymbol{X}_i$, $\boldsymbol{X}_j$ und $\boldsymbol{X}_k$ die i-te, j-te und k-te Realisierung bezüglich der Auflistung der Menge aller möglichen Zustände darstellen [Munsky et al., 2006]. Demnach kann die Wahrscheinlichkeit, das System in einem bestimmten Zustand zu finden, durch Integration von Gl. 2.41 bestimmt werden. Da die Matrix $\boldsymbol{A}$ und die Menge aller möglichen Zustände $\boldsymbol{X}$ in der Regel sehr groß sind, wird eine Reduktion des Zustandsraumes zu $\boldsymbol{X}^{red}$ vorgenommen, siehe Abb. 2.10. Reaktionen, die aus dem reduzierten Zustandsraum hinausführen, werden dabei in einem einzigen Zustand gespeichert, der eine Senke darstellt. Je größer der reduzierte Zustandsraum und je kleiner die Senke ist, desto genauer ist der Finite-State-Projection-Algorithmus. Die Genauigkeit des Algorithmus kann durch Summation der Wahrscheinlichkeit zu einem bestimmten Zeitpunkt t_{end} aller Zustände abgeschätzt werden, da diese im Idealfall eins ergibt

$$\epsilon = 1 - \sum_{\boldsymbol{X}_i \in \boldsymbol{X}^{red}} P(\boldsymbol{X}_i, t_{end}). \tag{2.43}$$

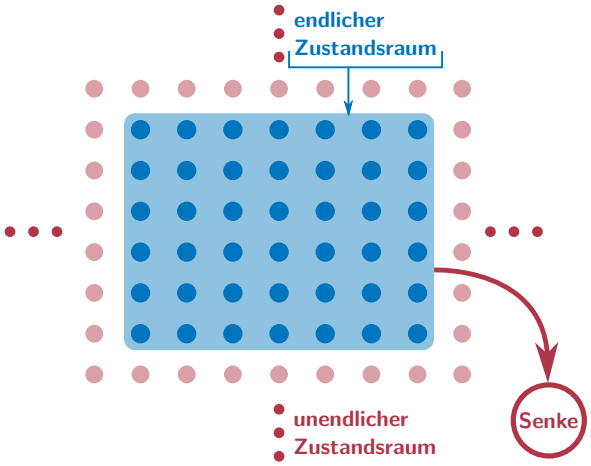

Abbildung 2.10: Reduktion des Zustandsraumes beim Finite-State-Projection-Algorithmus. Ausschließlich relevante Zustände werden betrachtet, wodurch der in der Regel unendlich dimensionale Zustandsraum (rot) in einen endlich dimensionalen (blau) überführt wird. Reaktionen, die aus dem reduzierten Zustandsraum hinausführen (roter Pfeil), münden in einem einzigen Zustand, der eine Senke darstellt.

Die durch die Näherung verursachte Abweichung ϵ stellt die Wahrscheinlichkeit dar, die durch die Senke absorbiert wird und kann als Maß der Güte verstanden werden. Durch die Vergrößerung des reduzierten Zustandsraumes kann die Güte der Näherung erhöht werden, bis ϵ kleiner als eine definierte Schranke ist [Munsky et al., 2006].

Der Finite-State-Projection-Algorithmus ist eine sehr elegante Methode zur Lösung der CME, die sich durch Integration eines Gleichungssystems ergibt. Es stellt sich jedoch heraus, dass diese Methode nur für kleine Systeme geeignet ist, da reale Systeme nicht hinreichend präzise durch den reduzierten Zustandsraum angenähert werden können. Weiterentwicklungen des Algorithmus zeigen eine breitere Anwendbarkeit [Peleš et al., 2006; Munsky et al., 2008; Waldherr et al., 2012], doch diese reichen bei weitem nicht aus, um reale Modelle akkurat zu simulieren. Zudem fehlen geeignete Implementierungen dieser Ansätze, weshalb sie selten verwendet werden.

2.4.3 Ω-Entwicklung

Die Ω-Entwicklung stellt eine Taylor-Entwicklung der CME dar [van Kampen, 2007], die es erlaubt, die statistischen Momente des Systemzustandes näherungsweise zu bestimmen. Dazu wird eine Variablentransformation des Systemzustandes vorgenommen

$$\frac{\boldsymbol{x}(t)}{\Omega} = \boldsymbol{\phi}(t) + \frac{\boldsymbol{\epsilon}}{\sqrt{\Omega}}. \tag{2.44}$$

In diesem Fall bezeichnet Ω das Systemvolumen und $\boldsymbol{\phi}(t)$ die Konzentrationen der chemischen Spezies, die durch die Ratengleichung gegeben sind. Die statistischen Eigenschaften der kontinuierlichen Zufallsvariable $\boldsymbol{\epsilon}$ bestimmen die Abweichung der Lösung der CME von der Lösung der Ratengleichung. Die CME kann dann durch Einsetzen von Gl. 2.44 und anschließende Entwicklung bezüglich $\Omega^{-\frac{1}{2}}$ approximiert werden. In der niedrigsten Ordnung ergibt sich für den Mittelwert des Systemzustandes die bereits bekannte Ratengleichung Gl. 2.2. Die niedrigste Ordnung der Kovarianz führt auf die Linear-Noise-Approximation [Elf et al., 2003; Paulsson, 2005; van Kampen, 2007]

$$\frac{\mathrm{dCov}(\boldsymbol{x})}{\mathrm{d}t} = \boldsymbol{J}\mathrm{Cov}(\boldsymbol{x}) + \mathrm{Cov}(\boldsymbol{x})\boldsymbol{J} + \boldsymbol{D} \tag{2.45}$$

Dabei bezeichnet $\boldsymbol{J}$ die Jacobi-Matrix

$$\boldsymbol{J} = \boldsymbol{N}\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{x}} \tag{2.46}$$

und $\boldsymbol{D}$ die Diffusionsmatrix mit den Einträgen

$$D_{ij} = \sum_{k=1}^{n_R} N_{ik}N_{jk}r_k(\boldsymbol{x}) \tag{2.47}$$

mit $i, j = 1 \ldots n_S$. Unter Einbeziehung höherer Ordnungen ergeben sich zusätzliche Korrekturterme für den Mittelwert und die Kovarianz [Grima, 2010, 2011], wodurch die Genauigkeit noch weiter gesteigert werden kann. Im Gegensatz zum Gillespie- und Finite-State-Projection-Algorithmus ist es mit dieser Methode nicht möglich, die Wahrscheinlichkeitsverteilung des Systemzustandes zu rekonstruieren. Es lassen sich lediglich statistische Momente bestimmen, durch die die Lösung der CME charakterisiert wird. Zudem ist eine absolute Fehlerabschätzung nicht möglich, weshalb zur Prüfung der Resultate der Ω-Entwicklung asymptotisch exakte Verfahren, wie der Gillespie-Algorithmus, verwendet werden müssen. Dennoch findet die hier beschriebene Methode eine breite Anwendung, da sie es ermöglicht, die statistischen Momente mittels einfacher Integration gewöhnlicher Differenzialgleichungen zu berechnen. Zudem ermöglichen schnelle Implementierungen, diese Methoden benutzerfreundlich anzuwenden [Kazeroonian et al., 2016]. Ähnlich wie andere Methoden, die die Lösung der CME mittels der statistischen Momente des Systemzustandes charakterisieren, ist auch die Ω-Entwicklung von numerischen Instabilitäten gekennzeichnet [Lee et al., 2009; Azunre et al., 2011]. Dies kann zu unphysikalischem Verhalten führen, wie beispielsweise negative Abundanzen, negative Varianzen oder falsche Konvergenzeigenschaften der chemischen Spezies. Je mehr Korrekturterme der Ω-Entwicklung berücksichtigt werden, desto präziser werden ihre Vorhersagen. Da *a priori* nicht klar ist, wie viele Terme zu berücksichtigen sind, um ein zuverlässiges Ergebnis zu erhalten, ist die Genauigkeit der Vorhersagen ungewiss.

2.5 Simulation externer Störungen

In den bisherigen Abschnitten ist die mathematische Formulierung und Handhabung stochastischer Effekte in biochemischen Systemen, wie Zellvariabilität und probabilistische Reaktionen, detailliert erläutert worden. Zellvariabilität kann als Unsicherheit hinsichtlich gewisser Parameter in der Modellstruktur aufgefasst werden, wohingegen probabilistische Reaktionen als inhärent stochastischer Prozess interpretiert werden können. Um die Dynamik einzelner Zellen zur Beschreibung einer Population präzise vorherzusagen, müssen jedoch neben intrinsischen und extrinsischen auch externe Störungen betrachtet werden, die als Fluktuationen der Umgebung angesehen werden können [Pischel et al., 2017]. Diese sind von besonderer Wichtigkeit für die Bioprozesstechnik, da die Steuerung und Regelung industrieller Prozesse ausschließlich über externe Einflussnahme geschieht. Neben natürlichen Umwelteinflüssen, wie Sonneneinstrahlung oder Niederschlag, sind auch experimentell einstellbare Größen, wie Reaktortemperatur, pH-Wert oder Zelldichte, von Schwankungen nicht ausgeschlossen. Einzelne Zellen sind deshalb unterschiedlichen externen Einflüssen ausgesetzt und weisen eine individuelle Entwicklung auf. Je größer das betrachtete System ist, desto ausgeprägter sind externe Störungen [Bar-Even et al., 2006; Wilkinson, 2009; Delvigne et al., 2017]. Aus diesem Grund stellt die Skalierung verfahrenstechnischer Prozesse vom Labormaßstab auf die industrielle Skala eine große Herausforderung dar.

Hinsichtlich der Modellierung lassen sich externe Störungen in Unsicherheiten der Modellparameter und stochastische Prozesse einteilen. Beispiele für unsichere Parameter, die sich signifikant auf die zelluläre Dynamik auswirken können, sind unter anderem Inhomogenitäten des Agarbodens in Petrischalenexperimenten [Croze et al., 2011] oder lokale Unterschiede der Zelldichte [Weber et al., 2012]. Diese Effekte können analog zu extrinsischen Störungen modelliert werden, was die mathematische Beschreibung externer Störungen stark vereinfacht. Stellen externe Störungen jedoch stochastische Prozesse dar, werden diese in der Regel mittels aufwendiger kinetischer Monte Carlo Simulationen, ähnlich wie der Gillespie-Algorithmus zur Beschreibung intrinsischer Störungen, berechnet. So lassen sich beispielsweise stochastische Trajektorien einzelner Zellen in inhomogenen Bioreaktoren [Lapin et al., 2004; Delafosse et al., 2015] simulieren. Im Folgenden wird vereinfacht angenommen, dass externe Störungen stets durch Unsicherheiten der Modellparameter ausdrücken werden können. Somit lassen sich extrinsische und externe Störungen nicht hinsichtlich ihrer mathematischen Handhabung unterscheiden, weshalb zukünftig der Einfachheit halber die Bezeichnung extrinsische Störung synonym für extrinsische und externe Störungen verwendet wird.

2.6 Simultane Simulation extrinsischer und intrinsischer Störungen

Biologische Zellen stellen stochastische Reaktionssysteme dar, die in einer inhomogenen Umgebung agieren, weshalb intrinsische und extrinsische Störungen nie einzeln auftreten, siehe Abb. 2.11A. Experimentell ist gezeigt worden, dass diese Störungen in einer komplexen Weise miteinander wechselwirken und nicht unabhängig voneinander sind [Raser et al., 2004]. So werden intrinsische Störungen durch extrinsische Störungen beeinflusst [Swain et al., 2002], aber auch umgekehrt [Elowitz et al., 2002]. Um die Wechselwirkung beider Arten von Störungen zu illustrieren, ist in Abb. 2.11B-C ihre Auswirkung auf einen Zerfallsprozess $X \rightarrow \emptyset$ dargestellt. Intrinsische Störungen beschreiben probabilistische Reaktionen, die mittels des Gillespie-Algorithmus berechnet werden. Im Gegensatz dazu resultieren extrinsische Störungen in unterschiedlichen Anfangsbedingungen einzelner Zellen, die mithilfe von Monte Carlo Simulationen in Kombination mit der Integration der Ratengleichung ermittelt werden. Beide Arten von Störungen führen zu Unsicherheiten des Systemzustandes, die mittels der Wahrscheinlichkeitsdichte $\boldsymbol{\rho}$, dem Mittelwert $\mathrm{E}(\boldsymbol{x})$ und der Standardabweichung $\mathrm{std}(\boldsymbol{x})$ charakterisiert werden können. Die Kombination beider Störungen bewirkt eine zusätzliche Streuung, die sich durch eine Verbreiterung der Verteilung bemerkbar macht, siehe Abb. 2.11D.

Die simultane Modellierung unterschiedlicher Arten von Störungen wird selten behandelt. In der Regel wird vereinfacht angenommen, dass intrinsische oder extrinsische Störungen zu vernachlässigen sind. Dies lässt sich dann rechtfertigen, wenn alle chemischen Spezies zahlreich vorhanden sind oder externe Störungen nur sehr langsam

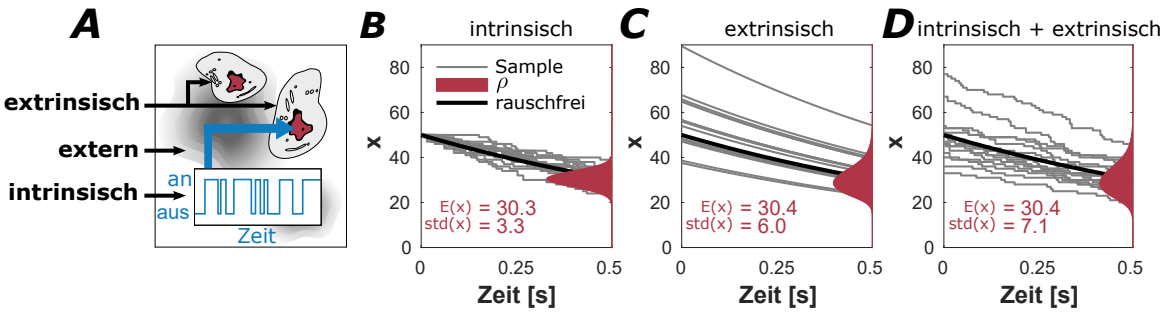

Abbildung 2.11: Störungen in biochemischen Reaktionssystemen [Pischel et al., 2017]. **(A)** Unterschiedliche Arten von Störungen nehmen Einfluss auf die zelluläre Dynamik. Die Auswirkung intrinsischer **(B)**, extrinsischer **(C)** sowie die Kombination intrinsischer und extrinsischer Störungen auf einen Zerfallsprozess resultieren in Unterschieden hinsichtlich der Wahrscheinlichkeitsdichte ρ, des Mittelwertes $\mathrm{E}(\boldsymbol{x})$ und der Standardabweichung $\mathrm{std}(\boldsymbol{x})$.

im Vergleich zur Systemdynamik fluktuieren [Voliotis et al., 2016]. Mit der schnellen technischen Entwicklung der Computerchips [Moore, 1998] folgte ein verstärktes Interesse an detaillierten Simulationen, um der Realität ein Stück näher zu rücken. Erste Ansätze die CME zu erweitern, stellen die Berücksichtigung von Diffusionsprozessen dar [Gardiner, 1976; Andrews et al., 2004; Bernstein, 2005], die es erlauben stochastische Transportprozesse und Reaktionen zu koppeln. Andere Methoden sind in der Lage Fluktuationen der dynamischen Umgebung als zeitabhängige Störungen einzubeziehen [Shahrezaei et al., 2008; Hilfinger et al., 2011; Zechner et al., 2014; Thanh et al., 2015; Voliotis et al., 2016]. Dies kann beispielsweise mittels zeitabhängiger Ratenkonstaten erfolgen. In dieser Arbeit werden extrinsische Störungen ausschließlich als Unsicherheiten bezüglich der Modellparameter angesehen. Diese unterliegen einer gewissen Verteilung, die jeder Zelle zufällige Realisierungen dieser Parameter zuordnet. Die einfachste Möglichkeit, Systeme dieser Art zu simulieren besteht darin, die Anfangsbedingungen für jede Zelle mittels Monte Carlo Simulationen zu ermitteln und anschließend die zeitliche Entwicklung mithilfe des Gillespie-Algorithmus zu berechnen [Wilkinson, 2009]. Die Kombination beider Methoden ist zwar asymptotisch exakt, jedoch geht sie mit einem immensen Rechenaufwand einher. Approximative Ansätze zur Reduzierung des Rechenaufwandes gibt es nur wenige. Probleme bei denen unsichere Parameter lediglich die Anfangsbedingungen der chemischen Spezies betreffen, können mittels des Finite-State-Projection-Algorithmus oder der Ω-Entwicklung angegangen werden. Beide Methoden weisen jedoch verschiedene Nachteile auf, die in den vorigen Abschnitten bereits benannt worden sind. Ein vielversprechendes Verfahren, das nicht nur für Unsicherheiten bezüglich der Anfangsbedingungen chemischer Spezies, sondern auch für andere unsichere Parameter anwendbar ist, stellt die Erweiterung der Ω-Entwicklung dar. Dies kann durch Kombination mit der Sigma-Punkt-Methode [Toni et al., 2013] oder der Gauß-Quadratur erfolgen [Bayati, 2017]. Dabei wird angenommen, dass sich die Gesamtvariabilität durch Superposition intrinsischer und extrinsischer Komponenten zusammensetzt. Ähnlich wie die Ω-Entwicklung sind auch ihre Erweiterungen hinsichtlich extrinsischer Störungen von numerischen Instabilitäten nicht befreit. Es ist deshalb nicht klar, wie zuverlässig die erzielten Ergebnisse sind. Zudem wird der Systemzustand nur durch seine statistischen Momente charakterisiert und nicht durch die zugrunde liegende Wahrscheinlichkeitsdichte. Um die Schwächen der bestehenden Methoden zu überwinden, wird im Folgenden eine Methode entwickelt, die in der Lage ist, simultan extrinsische und intrinsische Störungen in effizienter Weise zu simulieren.

2.6.1 Kombination der Sigma-Punkt-Methode und des Gillespie-Algorithmus

Wie bereits erläutert stellt die Kombination von Monte Carlo Simulationen und des Gillespie-Algorithmus zur Beschreibung extrinsischer und intrinsischer Störungen eine adäquate Methode dar, um die CME mit unsicheren Parametern zu lösen [Wilkinson, 2009]. Mit steigender Anzahl der berechneten Trajektorien nähert sich die Lösung der Monte Carlo Methode beliebig genau an die Lösung der CME an. Der große Nachteil dieser Methode besteht in der schlechten Konvergenz, die zu einem immensen Rechenaufwand und großen Simulationszeiten führt. Besonders in Optimierungsproblemen, die Modellsimulationen für diverse Parameterkonfigurationen erfordern, stellt der große Rechenaufwand eine unüberwindbare Hürde dar. Systeme mit Störungen auf unterschiedlichen Skalen werden deshalb äußerst selten in theoretischen Studien behandelt.

Die Kombination des Gillespie-Algorithmus mit Monte Carlo Simulationen kann als Berechnung von Trajektorien einzelner Zellen verstanden werden, die sich in der konkreten Realisierung ihrer unsicheren Parameter unterscheiden und deren zeitliche Entwicklung durch stochastische Prozesse bestimmt wird. In Abb. 2.12A wird zur Illustration das in Abb. 2.5 bereits gezeigte Beispiel der Reaktion erster Ordnung erneut aufgegriffen. Zusätzlich zu den intrinsischen Störungen wird angenommen, dass die Ratenkonstante ein unsicherer Parameter ist, der durch eine Normalverteilung beschrieben wird. Aus den einzelnen Trajektorien können statistische Momente und die Wahrscheinlichkeitsdichte berechnet werden, die jedoch nur langsam konvergieren. Der Ansatz, der in dieser Arbeit verfolgt wird, besteht darin, nicht die Zeitverläufe einzelner Zellen, sondern die stochastische Dynamik von Zellpopulationen zu berechnen, deren Superposition als Näherung der Lösung der CME betrachtet wird, siehe Abb. 2.12B. Für das hier betrachtete Beispiel zeigt sich, dass der Mittelwert und die Standardabweichung bestimmt mit beiden Methoden sehr gut übereinstimmen, was darauf hindeutet, dass der hier vorgeschlagene Ansatz eine akzeptable Näherung darstellt. Darüber hinaus werden weniger Funktionsauswertungen des zeitaufwändigen Gillespie-Algorithmus benötigt, wodurch die Simulation erheblich beschleunigt wird.

Bisher wurde lediglich qualitativ das Konzept des neuen Ansatzes diskutiert, aber keine Angaben darüber gemacht, wie die stochastische Dynamik der Populationen berechnet und deren Superposition erfolgen soll. In den vorigen Abschnitten sind verschiedene Methoden zur Simulation extrinsischer und intrinsischer Störungen, sowie deren Vor- und Nachteile, diskutiert worden. Hinsichtlich der Simulation extrinsischer Störungen hat sich die Sigma-Punkt-Methode, die einen guten Kompromiss aus Rechenaufwand

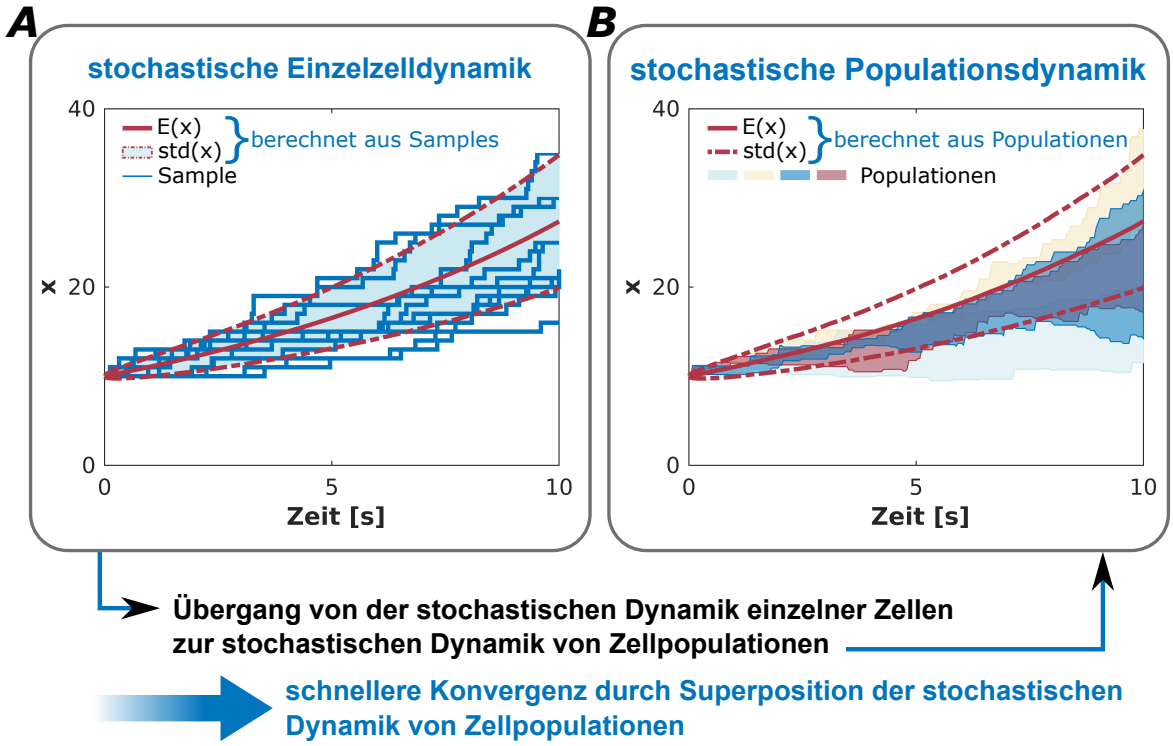

Abbildung 2.12: Stochastische Einzelzelldynamik *vs.* stochastische Populationsdynamik am Beispiel der Reaktion erster Ordnung. **(A)** Die Simulation von stochastischen Trajektorien einzelner, unterschiedlicher Zellen ergibt ein umfassendes Bild der Populationsdynamik. Der Nachteil dieser Methode ist die sehr langsame Konvergenz. **(B)** Im Gegensatz dazu konvergiert die Superposition der stochastischen Dynamik von Zellpopulationen schneller zu einer approximativen Lösung. Die approximative Lösung ist eine exzellente Näherung der exakten Lösung, was am fast identischen Verlauf des Mittelwertes $\mathrm{E}(\boldsymbol{x})$ und der Standardabweichung $\mathrm{std}(\boldsymbol{x})$ zu erkennen ist.

und Präzision darstellt, als vielversprechender Ansatz herausgestellt. Diese Methode ermöglicht es, mithilfe einer geringen Anzahl von Samples (Sigma-Punkte) die statistischen Momente einer nichtlinearen Funktion, die von unsicheren Parametern abhängt, zu berechnen. Zur Erweiterung auf Systeme mit intrinsischen Störungen bietet sich die Kombination mit dem Gillespie-Algorithmus oder einer approximativen Version dieses Algorithmus an [Pischel et al., 2016, 2017]. Die nichtlineare Funktion zur Transformation der Sigma-Punkte ist in diesem Fall durch einen stochastischen Prozess gegeben. Auf diese Weise lässt sich eine intrinsisch stochastische Komponente zur Populationsdynamik, die mittels der Sigma-Punkt-Methode berechnet wird, hinzufügen. Die zugrunde liegende Wahrscheinlichkeitsverteilung, die den Systemzustand charakterisiert, kann aus den statistischen Momenten unter Einbeziehung von Annahmen rekonstruiert werden. Wiederholte Simulationen mittels der Kombination beider Methoden liefern aufgrund des stochastischen Prozesses, der für die nichtlineare Transformation benutzt wird, stets unterschiedliche Populationsdynamiken. Dies drückt sich in Abweichungen

Algorithmus 2.2: Kombination der Sigma-Punkt-Methode und des Gillespie-Algorithmus [Pischel et al., 2016, 2017]

Ergebnis: Berechnung der approximativen Lösung der CME
Berechnung der Sigma-Punkte $\boldsymbol{\xi}_i$;
for $j = 1 : n$ **do**
  Berechnung der transformierten Sigma-Punkte $\boldsymbol{\eta}_i$ mittels Gillespie-Algorithmus;
  Berechnung des Mittelwertes $\mathrm{E}(\boldsymbol{\eta})$ und der Kovarianz $\mathrm{Cov}(\boldsymbol{\eta})$;
  Berechnung der Wahrscheinlichkeitsdichte $\hat{\boldsymbol{\rho}}_j = \mathcal{N}(\mathrm{E}(\boldsymbol{\eta}), \mathrm{Cov}(\boldsymbol{\eta}))$;
end
Superposition der Dichten $\tilde{\boldsymbol{\rho}}_j = \sum_{j=1}^{n} w_j \hat{\boldsymbol{\rho}}_j$ mit $w_j = \frac{1}{n}$;

hinsichtlich der ermittelten statistischen Momente und der daraus berechneten Wahrscheinlichkeitsdichten aus. Durch Superposition einer hinreichend großen Anzahl von Dichten n ergibt sich dann die gesuchte Näherung der Lösung der CME. Um dieses hier im Wortlaut beschriebene Verfahren näher zu erläutern, wird in Alg. 2.2 eine Darstellung in Pseudocode gegeben. Zudem wird im folgenden Abschnitt eine detaillierte Analyse bezüglich der Konvergenz und Präzision der hier vorgestellten approximativen Methode zur Lösung der CME präsentiert.

Vergleich mit konventionellen Methoden

Für das Beispiel der Reaktion erster Ordnung ist bereits gezeigt worden, dass die in dieser Arbeit vorgestellte Methode in der Lage ist, biochemische Reaktionssysteme mit extrinsischen und intrinsischen Störungen akkurat zu simulieren. Dies ist an der guten Übereinstimmung des Mittelwertes und der Standardabweichung mit der asymptotisch exakten Monte Carlo Methode zu erkennen. In vielen Anwendungen stellt sich jedoch heraus, dass es nicht ausreichend ist, eine Verteilung mithilfe der ersten beiden statistischen Momente zu charakterisieren. Besonders multimodale Verteilungen, die stark von der Normalverteilung abweichen, werden am besten durch ihre Wahrscheinlichkeitsdichte beschrieben. Aus diesem Grund wird im Folgenden die Güte und Konvergenz der Wahrscheinlichkeitsdichte, ermittelt mit der in dieser Arbeit vorgestellten Methode, untersucht und mit konventionellen Methoden verglichen. Konventionelle Methoden beruhen auf Monte Carlo Simulationen kombiniert mit dem Gillespie-Algorithmus, die mittels Kerndichteschätzung oder Histogrammeinteilung in der Lage sind, die Wahrscheinlichkeitsdichte zu ermitteln. Zur Veranschaulichung sind

diese in Abb. 2.13A am Schlögl-Modell illustriert

$$2\mathrm{X} + \mathrm{A} \underset{k_2}{\overset{k_1}{\rightleftarrows}} 3\mathrm{X} \tag{2.48}$$

$$\mathrm{B} \underset{k_4}{\overset{k_3}{\rightleftarrows}} \mathrm{X}, \tag{2.49}$$

das trotz seiner Einfachheit interessante Effekte, wie bimodales Verhalten, aufzeigen kann. Das betrachtete System ist neben intrinsischen auch von extrinsischen Störungen betroffen, die durch eine log-normalverteilte Initialmenge der chemischen Spezies X gekennzeichnet sind. Im Gegensatz zu den Monte Carlo Ansätzen, die auf der Berechnung zufälliger Realisierungen der Initialmenge von X basieren, zeichnet sich der Sigma-Punkt-Ansatz dadurch aus, dass lediglich eine geringe Anzahl von Samples deterministisch berechnet wird, siehe Abb. 2.13A. Diese werden wiederholt als Initialzustand für den stochastischen Prozess benutzt. Die sich ergebende Approximation der Wahrscheinlichkeitsdichte $\tilde{\boldsymbol{\rho}}$ aller Methoden wird mit einer Referenzlösung $\boldsymbol{\rho}$ verglichen. Als Ähnlichkeitsmaß wird die euklidische Distanz [Cha, 2007]

$$\Delta = \sqrt{(\tilde{\boldsymbol{\rho}} - \boldsymbol{\rho})^2} \tag{2.50}$$

verwendet. Je kleiner die euklidische Distanz ist, desto ähnlicher sind die Dichten. Da alle hier verwendeten Methoden zur Berechnung der zeitlichen Entwicklung des Systems auf stochastischen Prozesses beruhen, liefern wiederholte Berechnungen der Wahrscheinlichkeitsdichte $\tilde{\boldsymbol{\rho}}$ stets unterschiedliche Ergebnisse. Um eine statistische Aussage treffen zu können, werden deshalb wiederholte Berechnungen vorgenommen. Der Mittelwert der euklidischen Distanz $\mathrm{E}(\Delta)$ kann dabei als Maß für die Güte der Näherung und deren Standardabweichung $\mathrm{std}(\Delta)$ als Maß für die Konvergenz betrachtet werden. In Abb. 2.13B sind beide Größen in Abhängigkeit von den Funktionsauswertungen der jeweils benutzten Form des Gillespie-Algorithmus dargestellt. Die Monte Carlo Ansätze verwenden den Gillespie-Algorithmus, der konsistent mit der CME ist, wohingegen der Sigma-Punkt-Ansatz den approximativen τ-Leaping-Algorithmus verwendet, um die Berechnung der Trajektorien zusätzlich zu beschleunigen. Dabei ergibt sich die Anzahl der Funktionsauswertungen der Monte Carlo Ansätze aus der Anzahl der verwendeten Realisierungen n_{MC}. Im Gegensatz dazu ist die Anzahl der Funktionsauswertungen des Sigma-Punkt-Ansatzes durch die Anzahl der Sigma-Punkte $(2n_\xi + 1)$ multipliziert mit der Anzahl der aufsummierten Verteilungen n gegeben. Deutlich zu erkennen ist, dass für bis zu etwa 300 Funktionsauswertungen der Sigma-Punkt-Ansatz genauere Ergebnisse liefert als die Monte Carlo Ansätze. Zudem konvergiert der Sigma-Punkt-Ansatz weitaus schneller, weshalb die resultierenden Verteilungen dieses Ansatzes robuster und stabiler sind.

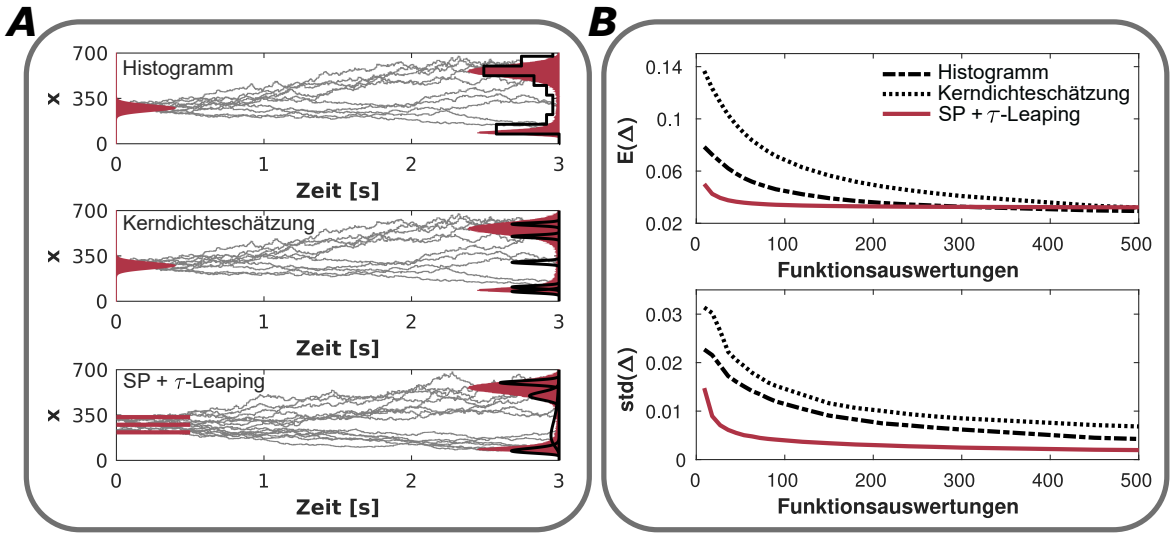

Abbildung 2.13: Approximation der Monte Carlo Methoden am Beispiel des Schlögl-Modells [Pischel et al., 2017]. **(A)** Die Wahrscheinlichkeitsdichte der mittels Monte Carlo Simulation kombiniert mit dem Gillespie-Algorithmus berechneten Samples kann mithilfe eines Histogrammansatzes oder einer Kerndichteschätzung ermittelt werden. Im Gegensatz dazu verwendet der Sigma-Punkt-Ansatz lediglich $2n_\xi + 1$ Samples, die mehrmals mittels des τ-Leaping-Algorithmus transformiert werden. Auf diese Weise werden extrinsische und intrinsische Störungen effizient approximiert. Anschließend kann aus der Superposition der resultierenden Dichten die approximative Lösung der CME ähnlich wie bei der Kerndichteschätzung berechnet werden. Für eine bessere visuelle Darstellung sind die gezeigten Verteilungen skaliert. **(B)** Die Güte E($\boldsymbol{\Delta}$) und Konvergenz std($\boldsymbol{\Delta}$) in Abhängigkeit der Funktionsauswertungen des Gillespie-Algorithmus oder seiner approximativen Form zeigen die rechentechnischen Vorteile des Sigma-Punkt-Ansatzes.

Um die rechentechnischen Vorteile weiter zu untermauern, wird eine umfassende Analyse an weiteren Modellen vorgenommen. Dazu wird als erstes ein einfaches Gen-Modell betrachtet [Pischel et al., 2016, 2017]

$$\mathrm{Gen_{OFF}} \underset{k_2}{\overset{k_1}{\rightleftarrows}} \mathrm{Gen_{ON}} \tag{2.51}$$

$$\mathrm{Gen_{ON}} \xrightarrow{k_3} \mathrm{Gen_{ON}} + \mathrm{A} \tag{2.52}$$

$$\mathrm{A} + \mathrm{B} \xrightarrow{k_4} \varnothing. \tag{2.53}$$

Das Modell besteht aus einem einzelnen Gen, das zwischen einem aktiven Zustand $\mathrm{Gen_{ON}}$ und einem inaktiven Zustand $\mathrm{Gen_{OFF}}$ wechseln kann. Ausschließlich im aktiven Zustand wird Protein A gebildet, das von Protein B zerstört wird. Dieses System ist von starken intrinsischen Störungen geprägt, die durch die geringe Abundanz des Gens verursacht werden. Zusätzlich werden extrinsische Störungen durch eine lognormalverteilte Initialmenge des Proteins B hervorgerufen. Die Synergie beider Störungen verursacht bimodales Verhalten, das für Protein A in Abb. 2.14A dargestellt

ist. Zusätzlich ist die mit dem approximativen Sigma-Punkt-Ansatz ermittelte Dichte, die sich durch Superposition einzelner Verteilungen zusammensetzt, illustriert. Es stellt sich dabei heraus, dass der approximative Ansatz qualitativ in der Lage ist, das Verhalten der scharfen bimodalen Verteilung widerzuspiegeln. Im Gegensatz zur vorigen Untersuchung wird diesmal der Fokus ausschließlich auf den Vergleich des Sigma-Punkt-Ansatzes und der Kerndichteschätzung gelegt. Die Kerndichteschätzung liefert im Kontrast zum Histogramm glatte Verteilungen, die keine Unstetigkeiten beinhalten. Da *a priori* nicht klar ist, welche Bandbreite der Kerne zu wählen ist, werden verschiedene Bandbreiten systematisch getestet. In diesem Fall werden Normalverteilungen als Kerne verwendet, deren Bandbreite durch ihre Standardabweichung gegeben ist. Ähnlich wie auch im vorigen Beispiel werden der Mittelwert und die Standardabweichung der euklidischen Distanz als Maß für die Güte und Konvergenz der Verteilung genutzt. Zum Vergleich der Güte wird die Differenz des Mittelwerts der euklidischen Distanz beider Verfahren $\mathrm{E}(\Delta_{Kern}) - \mathrm{E}(\Delta_{SP})$ gebildet. Analog dazu wird für den Vergleich der Konvergenz die Differenz der Standardabweichung $\mathrm{std}(\Delta_{Kern}) - \mathrm{std}(\Delta_{SP})$ ermittelt. Nehmen diese Größen positive Werte an, erzielt die in dieser Arbeit vorgestellte Kombination aus Sigma-Punkt-Methode und einer approximativen Version des Gillespie-Algorithmus genauere bzw. robustere Ergebnisse als die konventionelle Kerndichteschätzung. In Abb. 2.14B-C ist der Vergleich der Güte und Konvergenz für die Proteine A und B, sowie deren zeitliche Entwicklung dargestellt. Gut zu erkennen ist für Protein A, dass für bis zu etwa 1500 Funktionsauswertungen der Sigma-Punkt-Ansatz genauere Ergebnisse liefert. Für eine größere Anzahl von Funktionsauswertungen ist die Kerndichteschätzung in einem gewissen Bereich der Bandbreite in der Lage bessere Ergebnisse zu erzielen. Hinsichtlich der Konvergenz ist auffällig, dass die Kerndichteschätzung lediglich für sehr große Bandbreiten robustere Verteilungen ermittelt. Diese Verteilungen stellen jedoch aufgrund ihrer Breite eine schlechte Näherung der Lösung der CME dar. Im Gegensatz zu Protein A wird Protein B durch eine unimodale Verteilung beschrieben, die einfacher anzunähern ist. Aus diesem Grund übertrifft der Sigma-Punkt-Ansatz für alle betrachteten Kombinationen der Bandbreite und Anzahl der Funktionsauswertungen die Kerndichteschätzung hinsichtlich Güte und Konvergenz.

Im Appendix dieser Arbeit sind die Analysen zusätzlicher Modelle dokumentiert, siehe Abb. A.1-A.3. Es handelt sich dabei um das Schlögl-Modell [Schlögl, 1972], ein Virus-Modell [Gupta et al., 2014] und die Michaelis-Menten-Kinetik [Michaelis et al., 1913], deren konkrete Modellstruktur in Tab. 2.2 angegeben ist. Analog zum Gen-Modell wird dabei die Güte und Konvergenz des Sigma-Punkt-Ansatzes mit der Kerndichteschätzung verglichen, siehe Abschnitt A.2. Es stellt sich heraus, dass für eine geringe Anzahl von Funktionsauswertungen der Sigma-Punkt-Ansatz stets genauere

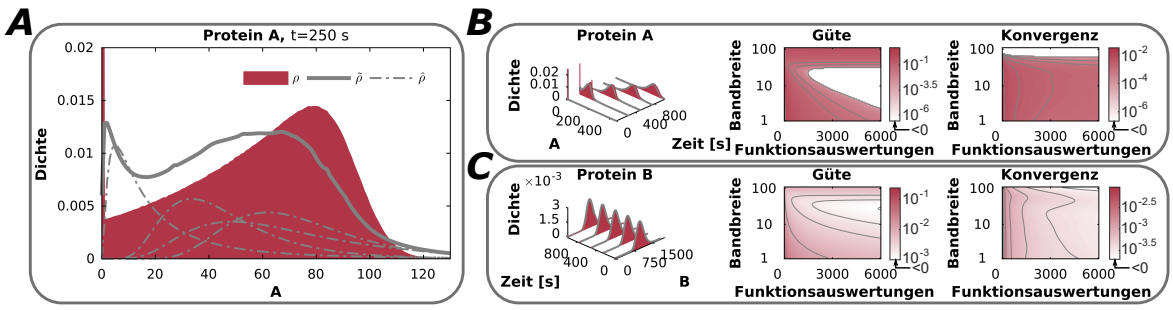

Abbildung 2.14: Benchmark der Sigma-Punkt-Methode kombiniert mit dem τ-Leaping-Algorithmus am Beispiel eines Gen-Modells [Pischel et al., 2017]. **(A)** Die Superposition der $\hat{\boldsymbol{\rho}}$ (grau, gestrichelt) ergibt die approximative Lösung der CME $\tilde{\boldsymbol{\rho}}$ (grau, durchgezogen). Die exakte Lösung, berechnet mit Monte Carlo Simulationen kombiniert mit dem Gillespie-Algorithmus, ist in Rot dargestellt. Für eine bessere visuelle Darstellung sind die Verteilungen $\hat{\boldsymbol{\rho}}$ skaliert. Zusätzlich sind für Protein A **(B)** und Protein B **(C)** die Zeitverläufe der Wahrscheinlichkeitsdichten sowie der Vergleich hinsichtlich Güte und Konvergenz des Sigma-Punkt-Ansatzes und der Kerndichteschätzung dargestellt. Dabei bezeichnen rote Schattierungen Bereiche, in denen die Sigma-Punkt-Methode bessere Ergebnisse erzielt, wohingegen in weißen Bereichen die Kerndichteschätzung bessere Ergebnisse liefert.

Ergebnisse erzielt. Für einige Modelle konnten für eine große Anzahl von Funktionsauswertungen jedoch auch Bereiche gefunden werden, in denen die Kerndichteschätzung präzisere Ergebnisse liefert. Hinsichtlich der Konvergenz zeigt sich, dass mit dem Sigma-Punkt-Ansatz für alle Modelle robustere Verteilungen berechnet werden. Lediglich für sehr breite Kerne, die keine akkurate Approximation der Lösung der CME zulassen, konvergiert die Kerndichteschätzung schneller. Damit ist gezeigt worden, dass die Kombination aus Sigma-Punkt-Methode und τ-Leaping-Algorithmus ein effizientes und präzises Verfahren darstellt, um biochemische Reaktionssysteme, gestört durch extrinsische und intrinsische Faktoren, zu simulieren.

Anwendung der Methode zur Parameterschätzung

In biochemischen Systemen ist es in der Regel nicht möglich alle Parameter, die für die mathematische Modellierung nötig sind, zu messen. Aus diesem Grund werden die experimentell nicht bestimmbaren Parameter mittels Optimierungsmethoden geschätzt. Die mathematischen Modelle müssen dafür wiederholt mit unterschiedlichen Parameterkonfigurationen simuliert werden. Um den immensen Rechenaufwand der asymptotisch exakten Monte Carlo Methoden zu umgehen, ist es möglich, approximative Methoden zu verwenden, die eine schnellere Konvergenz aufweisen. In diesem Abschnitt wird überprüft, ob die in dieser Arbeit vorgestellte Methode zur Parameterschätzung geeignet ist. Dazu werden die in Tab. 2.2 beschriebenen Modellsysteme

Tabelle 2.2: Modellbeschreibung für die Parameterschätzung [Pischel et al., 2017].

Modell	Anfangsbedingung		k		k_{opti}
Gen-Modell [Pischel et al., 2016, 2017]:					
$\mathrm{Gen_{OFF}} \underset{k_2}{\overset{k_1}{\rightleftarrows}} \mathrm{Gen_{ON}}$	N_{ON}	0	k_1	10^{-2}	$8.8 \cdot 10^{-3}$
$\mathrm{Gen_{ON}} \xrightarrow{k_3} \mathrm{Gen_{ON}} + \mathrm{A}$	N_{OFF}	1	k_2	10^{-3}	$6.6 \cdot 10^{-4}$
$\mathrm{A} + \mathrm{B} \xrightarrow{k_4} \varnothing$	N_A	0	k_3	$5 \cdot 10^{-1}$	$5.2 \cdot 10^{-1}$
	N_B	$e^{\mathcal{N}}(\mu = 10^3, \sigma = 1.5 \cdot 10^2)$	k_4	$5 \cdot 10^{-7}$	$5.2 \cdot 10^{-7}$
Schlögl-Modell [Schlögl, 1972]:					
$2\mathrm{X} + \mathrm{A} \underset{k_2}{\overset{k_1}{\rightleftarrows}} 3\mathrm{X}$	N_X	$e^{\mathcal{N}}(\mu = 2.75 \cdot 10^2, \sigma = 5)$	k_1	$3 \cdot 10^{-7}$	$3.1 \cdot 10^{-7}$
$\mathrm{B} \underset{k_4}{\overset{k_3}{\rightleftarrows}} \mathrm{X}$	N_A	10^5 (const.)	k_2	10^{-4}	10^{-4}
	N_B	$2 \cdot 10^5$ (const.)	k_3	10^{-3}	10^{-3}
			k_4	3.5	3.5
Michaelis-Menten-Kinetik [Michaelis et al., 1913]:					
$\mathrm{E} + \mathrm{S} \underset{k_2}{\overset{k_1}{\rightleftarrows}} \mathrm{ES}$	N_E	$2.5 \cdot 10^2$	k_1	10^{-4}	10^{-4}
$\mathrm{ES} \xrightarrow{k_3} \mathrm{P}$	N_S	$\mathcal{N}(\mu = 10^4, \sigma = 10^3)$	k_2	$5 \cdot 10^{-3}$	$2.4 \cdot 10^{-3}$
	N_{ES}	0	k_3	10^{-1}	10^{-1}
	N_P	0			
Virus-Modell [Gupta et al., 2014]:					
$\mathrm{V} \xrightarrow{k_1} \mathrm{G}$	N_V	$e^{\mathcal{N}}(\mu = 50, \sigma = 5)$	k_1	$1.5 \cdot 10^{-1}$	$1.4 \cdot 10^{-1}$
$\mathrm{G} \xrightarrow{k_2} \mathrm{G} + \mathrm{M}$	N_G	0	k_2	$2 \cdot 10^{-2}$	$2.1 \cdot 10^{-2}$
$\mathrm{G} \xrightarrow{k_3} 2\mathrm{G}$	N_M	0	k_3	$5 \cdot 10^{-2}$	$5.1 \cdot 10^{-2}$
$\mathrm{M} \xrightarrow{k_4} \mathrm{M} + \mathrm{P}$	N_P	0	k_4	1	$9.9 \cdot 10^{-1}$

benutzt. Zuerst werden mit der asymptotisch exakten Monte Carlo Methode und den in Tab. 2.2 angegebenen Ratenkonstanten k Referenzdichten für sechs Zeitpunkte mit konstantem zeitlichen Abstand berechnet. Dies wird für alle chemischen Spezies außer Gen_{ON} und Gen_{OFF} durchgeführt. Anschließend wird versucht, mittels eines genetischen Algorithmus [McCall, 2005] die Ratenkonstanten k_{opti} zu bestimmen, die die bestmögliche Annäherung der approximativen Dichten $\tilde{\boldsymbol{\rho}}$ zur Referenzlösung $\boldsymbol{\rho}$ ermöglichen. Die Zielfunktion, die dazu minimiert wird, setzt sich zusammen aus der Summe der euklidischen Distanzen aller chemischen Spezies und Zeitpunkte. Zum Vergleich der optimierten Parameter und der Referenzparameter sind diese in Tab. 2.2 aufgeführt. Für alle Modelle kann eine sehr gute Übereinstimmung festgestellt werden, wodurch demonstriert wird, dass der in dieser Arbeit vorgestellte Sigma-Punkt-Ansatz für die Parameterschätzung stochastischer biochemischer Reaktionsnetzwerke besonders gut geeignetist.

2.7 Zusammenfassung

In diesem Kapitel ist die mathematische Modellierung der Dynamik biochemischer Reaktionssysteme behandelt worden. Beginnend mit der Beschreibung störungsfreier Reaktionssysteme ohne räumliche Konzentrationsgradienten wird verdeutlicht, wie nichtlineare Effekte eine komplexe Systemdynamik verursachen können. Von besonderem biologischen Interesse stellt sich dabei multistationäres und oszillatorisches Verhalten heraus. Im weiteren Verlauf wird gezeigt, dass die deterministische Modellierung biologischer Systeme nicht in der Lage ist, deren Variabilität und Heterogenität zu beschreiben. Aus diesem Grund wird die deterministische Beschreibung auf stochastische Systeme beeinflusst durch intrinsische, extrinsische und externe Störungen erweitert. Intrinsische Störungen beziehen sich auf stochastische chemische Reaktionen, extrinsische Störungen auf Zellvariabilität und externe Störungen auf fluktuierende Umwelteinflüsse. Analog zu den störungsfreien Reaktionssystemen werden auch in stochastischen Systeme komplexe Dynamiken durch nichtlineare Effekte erzeugt. Zudem gibt es Erscheinungen, wie störungsinduzierte Bistabilität, die ausschließlich in stochastischen Systemen auftreten.

Die Simulation stochastischer Reaktionssysteme stellt sich als äußerst zeit- und rechenaufwendig heraus. Für die Optimierung, die auf wiederholten Simulationen des Systems mit unterschiedlichen Parameterkonfigurationen beruhen, stellt dies eine große Hürde dar. Aus diesem Grund werden für Optimierungszwecke in der Regel approximative Verfahren verwendet, die Präzision gegen Rechenaufwand abwägen. In diesem Kapitel wird ein Überblick verschiedener Verfahren hinsichtlich der Anwendbarkeit

sowie Vor- und Nachteilen gegeben. Die simultane Simulation intrinsischer und extrinsischer Störungen wird in der Literatur nur selten behandelt, da häufig vereinfachend angenommen wird, dass nur eine Störungsart einen signifikanten Beitrag leistet. Das in diesem Kapitel vorgestellte Verfahren zur simultanen Simulation intrinsischer und extrinsischer Störungen stellt damit einen wichtigen Beitrag zur Systembiologie dar. Das Verfahren kombiniert die Sigma-Punkt-Methode zur Approximation extrinsischer Störungen mit dem τ-Leaping-Algorithmus zur Simulation intrinsischer Störungen, um effizient die CME mit unsicheren Parametern zu lösen. Zur Untersuchung der Konvergenzeigenschaften wird ein Vergleich mit asymptotisch exakten Verfahren an einfachen Modellsystemen vorgenommen. Dabei stellt sich heraus, dass das in dieser Arbeit vorgestellte Verfahren schnell zu einer approximativen Lösung konvergiert, die eine gute Näherung der Lösung der CME darstellt. Es zeigt sich, dass die Präzision des Verfahrens nicht von der Modellgröße bestimmt wird, sondern abhängig von der Nichtlinearität des Systems ist. Dies ist daran erkennbar, dass einfache unimodale Verteilungen besser approximiert werden können als bimodale Verteilungen. Zusätzlich ist das Verfahren auf Parameteroptimierungsprobleme angewendet worden. Es stellt sich heraus, dass die unbekannten Parameter mit hoher Genauigkeit identifiziert werden können. Demnach eignet sich das vorgestellte Verfahren besonders, um biochemische Reaktionssysteme mit extrinsischen und intrinsischen Störungen effizient zu optimieren.

3 Der stochastische Prozess der Apoptose

Der programmierte Zelltod, Apoptose, ist ein essenzieller Mechanismus multizellulärer Organismen, der eine wichtige Rolle in biologischen Prozessen, wie der Gewebebildung oder der Zellhomöostase, spielt [Krammer et al., 2007]. Wird das Gleichgewicht zwischen neu gebildeten und sterbenden Zellen gestört, kann es zur Ausbildung von Krankheiten kommen. Dazu zählen AIDS, Krebs sowie verschiedene neurodegenerative und Autoimmunkrankheiten [Thompson, 1995]. Obwohl in den letzten Jahren ein immenser Fortschritt hinsichtlich der Regulation und Steuerung apoptotischer Prozesse erzielt worden ist, stellt sich der heutige Wissensstand noch weitgehend deskriptiv dar. Das bedeutet, dass die komplexen Prozesse pro- und antiapoptotischer Signalkaskaden zum Großteil unverstanden sind. Aus diesem Grund ist die erfolgreiche Implementierung quantitativer biologischer Modelle gestützt durch Experimente unabdingbar, um das Zusammenspiel dieser Prozesse aufzuklären [Spencer et al., 2011]. Zukünftig soll das tiefere Verständnis genutzt werden, um Krankheiten effektiver zu kurieren. Eine besondere Herausforderung der modernen Medizin stellt die medikamentöse Behandlung von Krebs dar, der als unkontrollierbare Gewebeneubildung verstanden werden kann [Cotter, 2009]. Wird eine Zelle derart beschädigt, dass sie nicht in der Lage ist, ihre Funktion zu erfüllen oder sich korrekt zu reproduzieren, dann stirbt sie durch Apoptose. Dies gewährleistet, dass lediglich funktionstüchtige Zellen existieren und die Arbeitsweise einzelner Organe bzw. des gesamten Organismus nicht beeinträchtigt wird. Im Gegensatz dazu sind Krebszellen dadurch gekennzeichnet, dass sie aufgrund der sehr dominanten antiapoptotischen Prozesse sehr resistent hinsichtlich Apoptose sind. Krebszellen können darum leicht der Abtötung durch medikamentöse Behandlungen widerstehen [Cotter, 2009]. Zudem zeigt sich, dass Krebszellen in der Regel sehr heterogen sind und sich stark zwischen verschiedenen Individuen unterscheiden können [Meacham et al., 2013]. Darüber hinaus sind antiapoptotische Prozesse eng mit der Genexpression verknüpft, die einen inhärent stochastischen Prozess darstellen [Tay et al., 2010]. Damit ergibt sich die Frage, wie Krebszellen in Gegenwart verschiedener Störungen besonders effizient und gezielt abzutöten sind.

Um dieser Fragestellung auf den Grund zu gehen, wird in den folgenden Abschnitten zunächst ein kurzer Überblick hinsichtlich der komplexen Biologie pro- und antiapoptotischer Prozesse gegeben. Anschließend wird auf das verwendete mathematische Modell eingegangen und wie extrinsische und intrinsische Störungen darin integriert sind. Das Modell wird mit der im vorigen Kapitel vorgestellten Methode optimiert und anschließend genutzt, um neue biologische Hypothesen aufzustellen. Dabei wird ein einfacher Mechanismus abgeleitet, der Aufschluss darüber gibt, wie graduelle molekulare Prozesse zu einem binären Entscheidungsprozess zwischen Leben und Tod führen.

3.1 Signaltransduktion der Apoptose

Apoptose kann über den extrinsischen und intrinsischen Pfadweg aktiviert werden. Der extrinsische Pfadweg wird durch zellexterne Stimulation eines Todesrezeptors mittels eines spezifischen Todesliganden ausgelöst [Krammer, 2000]. In dieser Arbeit liegt der Fokus auf CD95, der neben TNF-R1, TRAIL-R1, TRAIL-R2, DR3 und DR6 zu den bedeutendsten Todesrezeptoren gehört. CD95 befindet sich in der Zellmembran und besitzen eine zytosolische Region, die Todesdomäne genannt wird [Lavrik et al., 2009]. Nach der Stimulation von CD95 mittels des Todesliganden CD95L kommt es zur Komplexbildung der Todesdomäne mit FADD. Anschließend können die Proteine c-FLIP$_{\mathrm{L}}$, c-FLIP$_{\mathrm{RS}}$ und Caspase-8 an diesen Komplex binden. Dabei kommt es zur Ausbildung kettenartiger Strukturen, deren Zusammensetzung von der Stimulationsstärke des Todesliganden abhängt [Schleich et al., 2012]. Aus den Ketten können sich Homodimere, gebildet aus Caspase-8, sowie verschiedene Heterodimere gebildet aus Caspase-8 und c-FLIP$_{\mathrm{L}}$ bzw. c-FLIP$_{\mathrm{RS}}$ lösen. Einige der Homo- und Heterodimere sind in der Lage Caspase-3 zu aktivieren. Aktive Caspase-3 stellt ein besonders reaktives Molekül dar, das an der Spaltung zelleigener Proteine, wie Aktin und Lamin, beteidigt ist. Die damit verbundene Zerstörung der Zellmembran und verschiedener Organellen führt dann zum Tod der Zelle durch Apoptose [Lavrik, 2010].

Neben dem proapoptotischen Pfadweg wird durch die Stimulation des Todesrezeptors auch ein antiapoptotischer Pfadweg aktiviert. Das Protein c-FLIP$_{\mathrm{L}}$ ist dafür bekannt, dass seine Heterodimere nicht nur Caspase-3, sondern auch NF-κB aktivieren [Lavrik et al., 2009]. NF-κB ist ein zytosolischer Transkriptionsfaktor, der an einen Inhibitor IκBα gebunden ist. Durch Aktivierung der IκB-Kinase (IKK) mittels des c-FLIP$_{\mathrm{L}}$-Heterodimers p43-FLIP wird NF-κB von seinem Inhibitor getrennt und anschließend in den Zellkern transportiert [Golks et al., 2006; Neumann et al., 2010]. Dort beeinflusst es die Genexpression und beschleunigt die Bildung verschiedener Proteine. Dadurch

werden zum einen Proteine ersetzt, die durch Caspase-3 zerstört werden, aber auch Regulatoren, die die Aktivierung von NF-κB steuern. Eine schematische Darstellung dieser Prozesse ist in Abb. 3.1 gegeben.

Im Gegensatz zum extrinsischen Pfadweg wird der intrinsische Pfadweg durch die Permeabilisierung der äußeren Mitochondrienmembran ausgelöst, was künstlich mittels UV- bzw. γ-Strahlung oder genotoxischen Stress hervorgerufen werden kann. Durch die permeabilisierte Membran gelangt Cytocrom-C in das Zytoplasma, wo es die Bildung von Apoptosomen verursacht [Lavrik, 2010]. Apoptosome führen über verschiedener Reaktionskaskaden zur Aktivierung von Caspasen und spielen damit eine zentrale Rolle in der Initiierung der Apoptose. In dieser Arbeit werden ausschließlich Experimente mit extrinsischer Aktivierung des Todesrezeptors CD95 durch den Todesliganden CD95L analysiert. Aus diesem Grund wird der intrinsische Pfadweg der Apoptose nicht weiter thematisiert.

3.2 Modellierung der Apoptose

In der Systembiologie werden mathematische Modelle mit experimentellen Daten kombiniert, um das komplexe Verhalten biologischer Systeme zu verstehen. Dies ermöglicht es sowohl pro- [Fussenegger et al., 2000; Fricker et al., 2010] als auch antiapoptotische Pfadwege [Spencer et al., 2009; Tay et al., 2010; Pękalski et al., 2013] systematisch zu untersuchen und neue modellgestützte Hypothesen aufzustellen. Dabei zeigt sich, dass für den dynamischen Prozess der Apoptose das Zusammenspiel beider Pfadwege von besonderer Bedeutung ist [Neumann et al., 2010; Buchbinder et al., 2018]. Um zu untersuchen, wie pro- und antiapoptotische Pfadwege miteinander wechselwirken und welchen Einfluss unterschiedliche Störungen auf die Entscheidung zwischen Leben und Tod haben, wird in dieser Arbeit das in Abb. 3.1 dargestellte Modell untersucht. Dieses Modell unterscheidet sich von seinen Vorgängern dahingehend, dass es essenzielle Aspekte methodisch und biologisch relevanter, vorangegangener Arbeiten kombiniert, die zuvor lediglich einzeln oder noch nicht betrachtet worden sind:

- extrinsische Störungen der kettenartigen Strukturen
- intrinsische Störungen der Genexpression [Tay et al., 2010; Pękalski et al., 2013]
- Kalibrierung stochastischer biologischer Einzelzellmodelle [Poovathingal et al., 2010; Lillacci et al., 2013; Pischel et al., 2017]
- Abhängigkeit der Zusammensetzung der kettenartigen Strukturen von der Stimulationsstärke [Schleich et al., 2012]

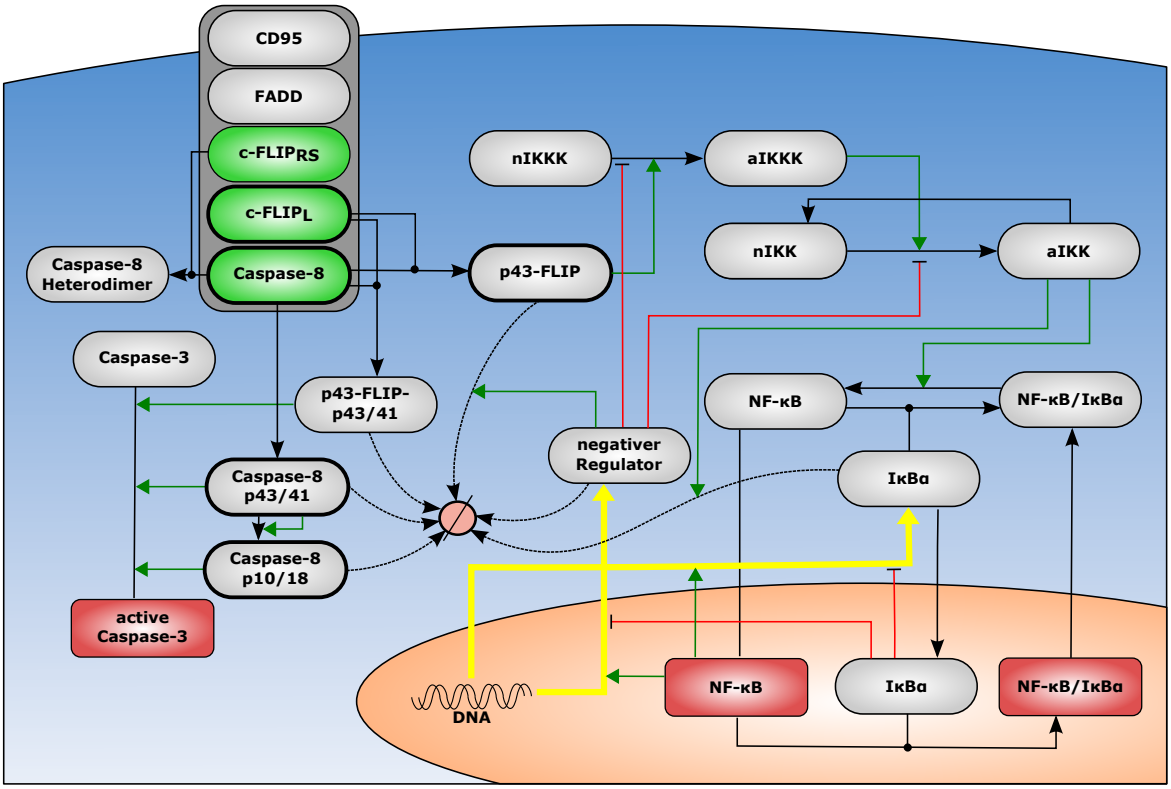

Abbildung 3.1: Modelltopologie des Apoptosenetzwerks [Buchbinder et al., 2018]. Das CD95-Netzwerk ist ein Multikompartmentmodell, das das Zytosol (blau), die Zellmembran und den Zellkern (orange) beinhaltet. An den Komplex aus dem aktivierten Todesrezeptor und FADD binden kettenartige Strukturen, die aus Caspase-8, c-FLIP$_L$ und c-FLIP$_{RS}$ bestehen. Sie stellen die Quellen der extrinsischen Störung dar (grün). Intrinsische Störungen sind durch die stochastische Genexpression gegeben (gelbe Pfeile). Die mit Western Blot gemessenen Proteinabundanzen sind durch einen dicken Rahmen markiert (c-FLIP$_L$, Caspase-8 und deren Dimere). Messungen mit bildgebender Flusszytometrie sind rot markiert (Caspase-3, NF-κB). Inhibitionen sind durch rote Linien und Aktivierungen durch grüne Pfeile gekennzeichnet.

- Integration CD95-induzierter pro- und antiapoptotischer Pfadwege [Neumann et al., 2010]

Das Modell integriert somit die Stärken vorangegangener Arbeiten und ermöglicht eine sehr detaillierte Beschreibung des dynamischen Prozesses der Apoptose[1]. Eine Übersicht der chemischen Spezies sowie der chemischen Reaktionen ist in Tab. A.1-A.2 zu finden. In den folgenden Abschnitten wird darauf eingegangen, wie diese Aspekte in die mathematische Modellierung integriert worden sind.

[1]Das Modell ist strukturell besonders durch [Fricker et al., 2010] und [Pękalski et al., 2013] inspiriert.

3.2.1 Störungen apoptotischer und antiapoptotischer Pfadwege

Werden Zellen einer Population mittels eines Todesliganden stimuliert, ergibt sich in der Regel eine sehr heterogenes Antwortverhalten [Albeck et al., 2008; Spencer et al., 2009; Roux et al., 2015]. Apoptotische und antiapoptotische Pfadwege werden unterschiedlich stark aktiviert, wodurch einige Zellen durch Apoptose sterben und andere nicht. Die Quellen der Variabilität beinhalten genetische Variationen, Zellzykluseffekte und stochastische Effekte der Genexpression [Spencer et al., 2009; Xia et al., 2014]. Diese können, wie im vorigen Kapitel erläutert, in extrinsische und intrinsische Störungen eingeteilt und akkurat durch die CME beschrieben werden.

Hinsichtlich der Abundanzen chemischer Spezies des Zelltyps HeLa-CD95 sind lediglich die Unsicherheiten der Proteine Caspase-8, c-FLIP$_{\mathrm{L}}$ und c-FLIP$_{\mathrm{RS}}$ aus vorangegangenen Arbeiten bekannt [Fricker et al., 2010]. Im Folgenden wird angenommen, dass die Abundanzen dieser chemischen Spezies zum initialen Zeitpunkt log-normal verteilt sind [Limpert et al., 2001]. Zudem wird davon ausgegangen, dass Mittelwert und Standardabweichung zueinander proportional sind. Der dosisunabhängige Proportionalitätsfaktor ist aus [Fricker et al., 2010] entnommen worden. Aufgrund des Mangels an Wissen über die Unsicherheit der übrigen chemischen Spezies wird angenommen, dass diese nicht von extrinsischen Störungen betroffen sind.

Neben extrinsischen spielen auch intrinsische Störungen im betrachteten Modellsystem eine wichtige Rolle. Besonders dominant ist dieser Effekt bei Reaktionen, die chemische Spezies mit einer geringen Abundanz involvieren. In diesem Fall betrifft das Reaktionen der Genexpression [Eldar et al., 2010], da angenommen wird, dass lediglich zwei Kopien jedes Gens vorhanden sind. Die übrigen chemischen Spezies kommen relativ dazu sehr häufig vor, weshalb eine hybride deterministisch-stochastische Approximation des Gillespie-Algorithmus verwendet wird [Haseltine et al., 2002]. Dabei werden die Wechsel der Gene zwischen aktiven und inaktiven Zuständen durch stochastische Prozesse beschrieben. Für die übrigen Reaktionen werden im Gegensatz dazu gewöhnliche Differenzialgleichungen zur Simulation benutzt.

Zur Modellkalibrierung wird für die simultane Simulation beider Störungen die im vorigen Kapitel vorgestellte Methode verwendet [Pischel et al., 2016, 2017]. Durch die Kombination der Sigma-Punkt-Methode mit dem approximativen Gillespie-Algorithmus ist es möglich, die CME mit unsicheren Parametern numerisch effizient zu lösen. Die folgenden Simulationen des Modells mit den bereits optimierten Parametern werden dann mittels Monte Carlo Simulationen in Kombination mit dem hybriden Gillespie-Algorithmus durchgeführt.

3.2.2 Modellannahmen und -kalibrierung

Zur Erstellung eines zuverlässigen Modells muss dieses qualitativ mit biologischen Beobachtungen übereinstimmen und anhand experimenteller Daten kalibriert werden. Aus vorangegangenen Arbeiten ist bekannt, dass sich die Proteine c-FLIP$_{\mathrm{L}}$, c-FLIP$_{\mathrm{RS}}$ und Caspase-8 zytosolisch an den Komplex aus dem aktivierten Todesrezeptor und FADD anlagern. Dabei kommt es zur Bildung kettenartiger Strukturen, deren Zusammensetzung von der Stimulationsdosis abhängt [Schleich et al., 2012]. Obwohl dies für Zellen des Typs SKW6.4 gezeigt worden ist, wird angenommen, dass die gleiche Zusammensetzung gegeben als Verhältnis zu FADD auch bei den hier verwendeten Zellen des Typs HeLa-CD95 vorherrscht. Die Daten werden deshalb extrapoliert und für die Simulationen dieser Arbeit verwendet, siehe Abb. 3.2A. Zudem wird davon ausgegangen, dass das am Todesrezeptor gebundene FADD proportional zur Stimulationsdosis ist. Der Proportionalitätsfaktor wird mit den übrigen freien Parametern während der Modellkalibrierung optimiert. Für das Protein c-FLIP$_{\mathrm{RS}}$ ist das Verhältnis zu FADD bezüglich verschiedener Stimulationsdosen unbekannt. Aus diesem Grund wird angenommen, dass c-FLIP$_{\mathrm{L}}$ und c-FLIP$_{\mathrm{RS}}$ proportional zueinander sind, wobei der Proportionalitätsfaktor aus [Fricker et al., 2010] entnommen worden ist.

Um das Modell zu kalibrieren, wird es an Bulk-Populations- (Western Blot) und Einzelzellmessungen (bildgebende Flusszytometrie) angepasst. Für die Western Blot Daten wird ein multiplikatives Störungsmodell zur Proteinquantifizierung angenommen [Kreutz et al., 2007]

$$M_{WB} = M^0_{WB} + \alpha_{WB}\bar{x}\eta_{WB}. \tag{3.1}$$

Dabei bezeichnet M_{WB} den gemessenen Grauwert, M^0_{WB} einen konstanten Offset, α_{WB} einen Proportionalitätsfaktor, $\bar{x}$ den simulierten Mittelwert und η_{WB} eine lognormalverteilte Störung. Der Offset M^0_{WB} wird aus Gl. 3.1 zum initialen Zeitpunkt entnommen, wohingegen α_{WB} mit den übrigen freien Parametern während der Modellkalibrierung optimiert wird. In dieser Arbeit sind Western Blots für die in Abb. 3.1 fett umrahmten Proteine zur Modellkalibrierung verwendet worden.

Für die Daten, aufgenommen mittels bildgebender Flusszytometrie, wird folgendes Modell zur Proteinquantifizierung angenommen

$$M_{FZ} = M^0_{FZ} + \alpha_{FZ}\hat{x}(1 + \eta_{FZ}). \tag{3.2}$$

In diesem Fall bezeichnet M_{FZ} die gemessene Intensitätsverteilung, M^0_{FZ} eine konstante Offset-Verteilung, α_{FZ} einen Proportionalitätsfaktor, $\hat{x}$ die simulierte Verteilung und η_{FZ} eine normlverteilte Störung. Analog zu den Western Blot Daten wird die Offset-Verteilung aus Gl. 3.2 zum initialen Zeitpunkt entnommen und α_{FZ} mit den übrigen

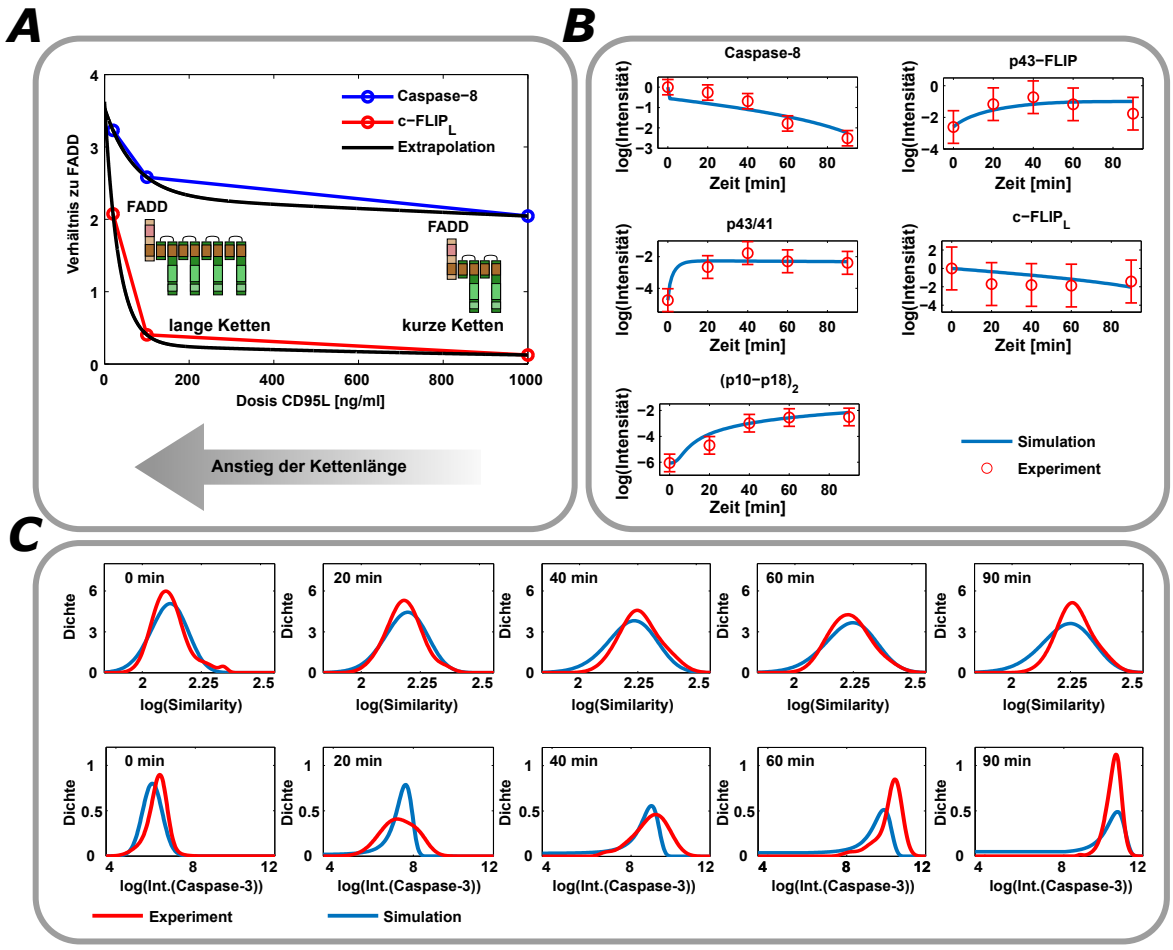

Abbildung 3.2: Modellkalibrierung [Buchbinder et al., 2018]. **(A)** Die Zusammensetzung der Ketten ist abhängig von der Stimulationsdosis. **(B)** Vergleich des simulierten Populationsmittelwertes mit Western Blot Experimenten. Die Balken zeigen die Standardabweichung ermittelt aus drei Replikationssexperimenten. **(C)** Vergleich der simulierten Populationsverteilung mit Experimenten, gemessen mit bildgebender Flusszytometrie.

freien Parametern während der Modellkalibrierung optimiert. In dieser Arbeit werden Einzellzellmessungen, gemessen mit bildgebender Flusszytometrie, aktiver Caspase-3 und nuklearem NF-κB verwendet. Diese Proteine sind in Abb. 3.1 rot dargestellt. Für Caspase-3 wird die Fluoreszenzintensität als Messsignal verwendet, wohingegen für NF-κB der Similarity-Wert genutzt wird, der es erlaubt zu quantifizieren, welche Menge an NF-κB sich im Zellkern befindet [Buchbinder et al., 2018].

Die Bestimmung der unbekannten Parameter wird mittels der Methode der kleinsten Fehlerquadrate vorgenommen. Dazu wird die Zielfunktion

$$L = \sum_{i,j,k} \frac{(M^{exp}_{i,j,k} - M^{sim}_{i,j,k})^2}{\sigma^2_{i,j,k}} \tag{3.3}$$

mittels eines genetischen Algorithmus minimiert. Die Indizes exp und sim kennzeichnen dabei experimentelle bzw. simulierte Daten. Die Summe läuft über alle Proteine i, Zeitpunkte j und Stimulationsdosen k. σ bezeichnet in diesem Fall die Standardabweichung von Replikationsexperimenten. Für Western Blot Daten stellen M^{exp} und M^{sim} die gemessenen bzw. simulierten Grauwerte auf logarithmischer Skala dar. Für Daten gemessen mit bildgebender Flusszytometrie wird die Differenz $M^{exp} - M^{sim}$ als euklidische Distanz der gemessenen und simulierten Verteilungen interpretiert [Cha, 2007]. Die Vorhersagen des kalibrierten Modells mit den zugehörigen experimentellen Daten sind in Abb. 3.2B,C und A.4-A.5 dargestellt.

3.3 Modellvorhersagen

Nachdem das Modell kalibriert worden ist, wird es genutzt, um eine Analyse der zentralen Knotenpunkte durchzuführen. Die Knotenpunkte beinhalten aktive Caspase-3, die durch die Spaltung verschiedener Proteine zur Apoptose führt, p43-FLIP, das den pro- und antiapoptotischen Pfadweg miteinander verknüpft und nukleares NF-κB, das die Genexpression aktiviert. Dabei zeigt sich, dass die Aktivierung von Caspase-3 stark von der Stimulationsdosis abhängig ist. Mit steigender Stimulationsdosis wird ein exponentieller Anstieg des Mittelwertes und der Standardabweichung von Caspase-3 verzeichnet, siehe Abb. 3.3A. Dies stimmt mit Beobachtungen vorangegangener Arbeiten überein [Neumann et al., 2010; Spencer et al., 2011]. Die Variabilität wird dabei ausschließlich durch extrinsische Störungen der Proteine c-FLIP$_{\mathrm{L}}$, c-FLIP$_{\mathrm{RS}}$ und Caspase-8 hervorgerufen. Intrinsische Störungen haben keinen Einfluss auf die Aktivierung von Caspase-3, da es kein Feedback des antiapoptotischen Pfades auf den apoptotischen gibt, siehe Abb. 3.1.

Hinsichtlich p34-FLIP, dem Bindeglied zwischen dem apoptotischen und antiapoptotischen Pfadweg, stellt sich heraus, dass Mittelwert und Standardabweichung ebenfalls mit steigender Stimulationsdosis zunehmen. Darüber hinaus zeigt sich, dass durch Überexpression des Proteins c-FLIP$_{\mathrm{L}}$ ein Anstieg von p43-FLIP zu verzeichnen ist, siehe Abb. 3.3B und A.6A. c-FLIP$_{\mathrm{L}}$ ist dafür bekannt, dass es neben dem apoptotischen auch antiapoptotische Pfadwege aktiviert [Lavrik et al., 2009; Neumann et al., 2010]. Der Anstieg von p43-FLIP induziert durch eine Überexpression von c-FLIP$_{\mathrm{L}}$ ist somit plausibel und lässt sich biologisch belegen. Im Gegensatz zu Caspase-3 wird p43-FLIP durch extrinsische und intrinsische Störungen beeinflusst, siehe Abb. 3.1. In Abb. 3.3B und Abb. A.6A werden jedoch nur die Auswirkungen der intrinsischen Störungen dargestellt. Das Zusammenspiel intrinsischer und extrinsischer Störungen ist in Abb. A.6B zu finden. Die zusätzlichen Störungen bewirken lediglich eine Verbreiterung

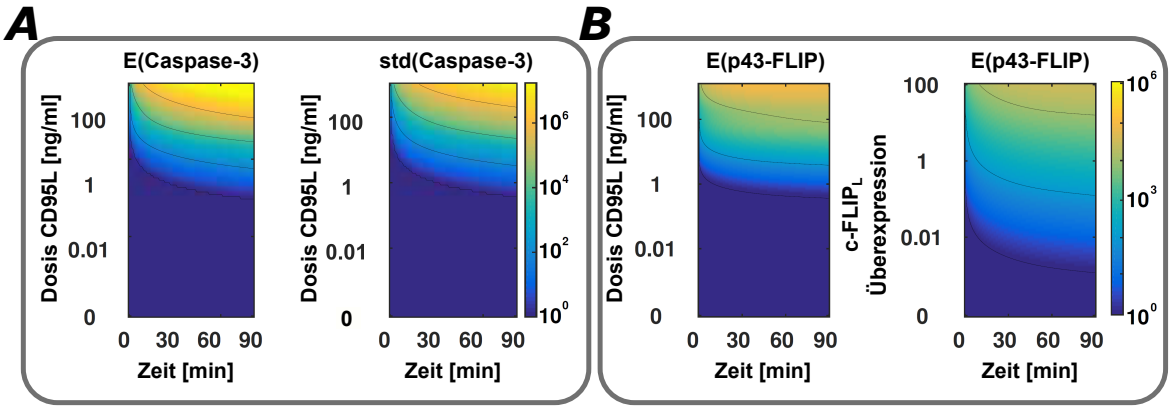

Abbildung 3.3: Modellvorhersagen bezüglich der Aktivierung von **(A)** Caspase-3 und **(B)** p43-FLIP in Abhängigkeit von der Stimulationsstärke und der c-FLIP$_L$-Überexpression [Buchbinder et al., 2018]. Die Abundanzen der chemischen Spezies werden in Teilchenanzahlen angegeben.

der Verteilung, jedoch keine signifikante Änderung der Dynamik.

Die Aktivierung von p43-FLIP bewirkt die Translokation von zytosolischem NF-κB in den Zellkern. Es zeigt sich, dass für eine breite Menge an Stimulationsdosen sehr ähnliche Verläufe von nuklearem NF-κB beobachtet werden können, siehe Abb. 3.4A. Eine systematische Analyse der Abhängigkeit der Dynamik von nuklearem NF-κB lässt verschiedene Bereiche gekennzeichnet durch qualitativ unterschiedliches Verhalten erkennen. Zwischen 0.5 und 5 ng/ml CD95L findet eine schwache Aktivierung von nuklearem NF-κB im betrachteten Zeitintervall statt, die von großer Variabilität geprägt ist. Für Stimulationsstärken über 5 ng/ml zeigt sich eine starke Aktivierung, die robust gegen intrinsische Störungen ist. Dies ist an stabilen zeitlichen Verläufen des Mittelwertes und der Standardabweichung zu erkennen, siehe Abb. 3.4B. Auch für Variationen des initialen Verhältnisses R_i von nuklearem und zytosolischem NF-κB kann das gleiche Verhalten beobachtet werden, siehe Abb. 3.4C und A.6C. Experimentell registrierbare Variabilitäten der initialen Menge an nuklearem NF-κB [Lee et al., 2014] haben somit keine Auswirkungen auf die Modellvorhersagen. Bezüglich der Überexpression von c-FLIP$_L$ zeigt sich, dass auch unter der endogenen Menge an c-FLIP$_L$ eine Aktivierung von NF-κB stattfindet. Diese ist jedoch stark durch intrinsische Störungen geprägt. Für höhere Expressionen sind stabile Verläufe des Mittelwertes und der Standardabweichung erkennbar, siehe Abb. 3.4B. In Abb. 3.4 ist lediglich der Einfluss intrinsischer Störungen dargestellt. Es zeigt sich jedoch, dass zusätzliche extrinsische Störungen keinen Einfluss auf die Dynamik haben, siehe Abb. A.6D.

Damit stellt sich heraus, dass pro- (Caspase-3) und antiapoptotische Pfadwege (NF-κB) fundamental unterschiedliche Dynamiken in Abhängigkeit der Stimulationsdosis aufzeigen. Caspase-3 und dessen Variabilität steigen mit zunehmender Stimulations-

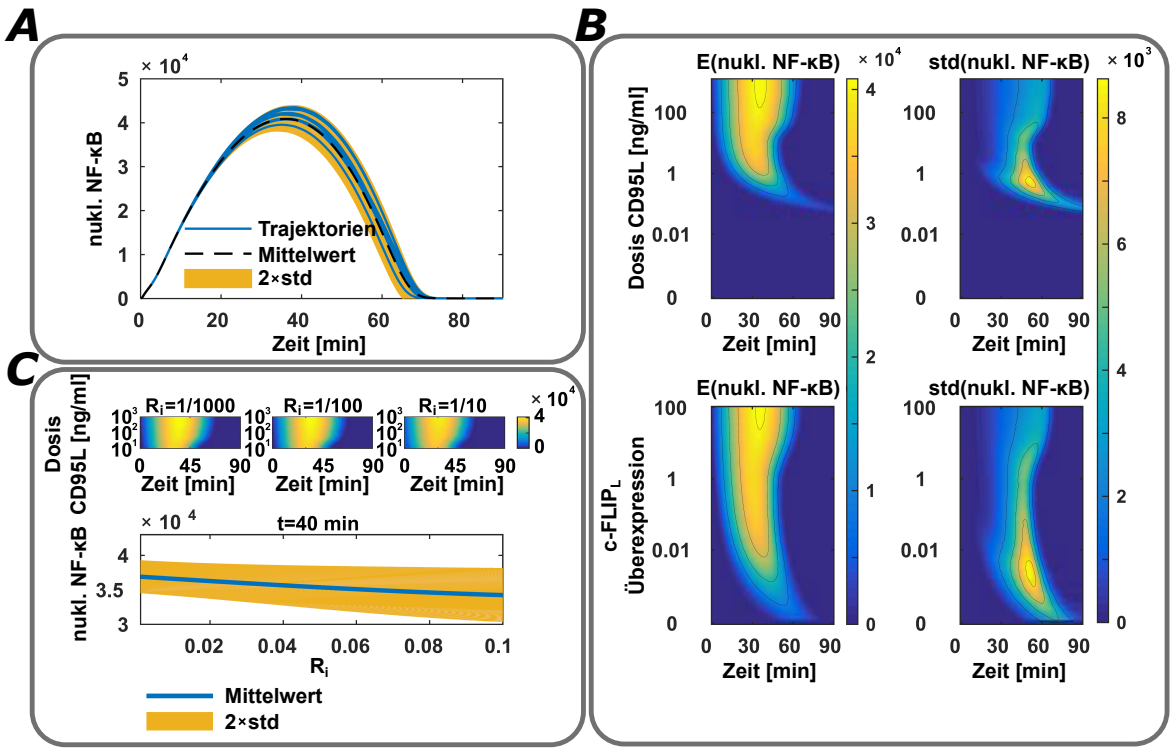

Abbildung 3.4: Modellvorhersagen bezüglich der Aktivierung von nuklearem NF-κB [Buchbinder et al., 2018]. **(A)** Der typische Verlauf einzelner nuklearer NF-κB-Trajektorien zeigt im betrachteten Zeitraum ein Maximum, dessen Höhe und Zeitpunkt nur leicht variiert. **(B)** Die Dynamik von nuklearem NF-κB ist abhängig von der Stimulationsstärke und der c-FLIP$_\mathrm{L}$-Überexpression. **(C)** Die Variation des Verhältnisses von nuklearem und zytosolischem NF-κB R_i bewirkt keine Änderung der Systemdynamik. Die Abundanzen der chemischen Spezies werden in Teilchenanzahlen angegeben.

dosis exponentiell an. NF-κB und dessen Variabilität nehmen jedoch nur bis etwa 5 ng/ml CD95L mit steigender Stimulationsdosis zu. Darüber hinaus wird ein robustes Verhalten beobachtet, das nur gering durch intrinsische Störungen beeinflusst wird. Demnach werden Heterogenitäten des dynamischen Prozesses der Apoptose über 5 ng/ml CD95L durch extrinsische Störungen dominiert. Im Folgenden wird diese Hypothese weiter untersucht und dabei ein Parameter definiert, der die Entscheidung zwischen Leben und Tod bestimmt.

3.4 Entscheidung zwischen Leben und Tod

In Abwesenheit der Aktivierung des antiapoptotischen Pfadweges stirbt eine Zelle durch Apoptose, kurz nachdem sie eine kritische Konzentration an Caspase-3 überschritten hat, die als „Point of no Return“ bezeichnet wird [Spencer et al., 2011;

Roux et al., 2015]. Zur Bestimmung der kritischen Konzentration an Caspase-3 wird die Fluoreszenz aktiver Caspase-3 lebendiger und apoptotischer Zellen mittels einer quadratischen Diskriminanzanalyse untersucht[2]. Dabei wird eine kritische Intensität identifiziert, die lebendige und apoptotische Zellen separiert. Mittels Gl. 3.2 kann die Umrechnung der kritischen Intensität in die kritische Menge an Caspase- vorgenommen werden, um den Point of no Return zu bestimmen. Der Zeitpunkt, wann der Point of no Return erreicht wird, ist stark abhängig von der Stimulationsdosis, siehe Abb. 3.5A. Wird der antiapoptotische Pfadweg aktiviert, ist es möglich, dass der Zelltod über den Point of no Return hinaus verzögert oder sogar aufgehalten wird. Nukleares NF-κB regt die Genexpression an und sorgt dafür, dass antiapoptotische Proteine gebildet werden. Zur Charakterisierung der Synergie beider Pfadwege werden zwei Parameter eingeführt, die deren zeitlichen Zusammenhang beschreiben. TOD (Time of Decision) stellt die Zeitspanne von der Stimulation bis zum Erreichen des Point of no Return dar, wohingegen TOS (Time of Survival) als zeitlicher Abstand der maximalen Aktivierung nuklearem NF-κBs zum Point of no Return definiert ist. Tritt die maximale Aktivierung von nuklearem NF-κB vor dem Point of no Return ein ist TOS positiv, andernfalls negativ. Sowohl TOD als auch TOS sind abhängig von der Stimulationsdosis, da die Dynamik von Caspase-3 dosisabhängig ist. Es wird die Hypothese aufgestellt, dass Zellen mit einem hohen Verhältnis von TOS/TOD dazu tendieren, die apoptotische Stimulation zu überleben, da sie mehr Zeit haben, um der Aktivierung des apoptotischen Pfadweges entgegenzuwirken. Im Gegensatz dazu scheitern Zellen mit einem niedrigen Verhältnis von TOS/TOD die Apoptose abzuwehren.

Um das Verhältnis TOS/TOD des Modells zu experimentellen Daten in Verbindung zu setzen, wird ein kritische Verhältnis r_{crit} ermittelt, welches die Entscheidung zwischen Leben und Tod bestimmt. Zellen mit einem Verhältnis TOS/TOD $> r_{crit}$ überleben die apoptotische Stimulation, wohingegen Zellen mit TOS/TOD $< r_{crit}$ durch Apoptose sterben. Mittels des Prinzips der kleinsten Fehlerquadrate wird durch Anpassung von r_{crit} die simulierte Dosis-Response-Kurve an den experimentellen Verlauf angepasst, siehe Abb. 3.5B. Es sind dazu die Daten aus Abb. 3.5C ohne Inhibition genutzt worden. Die Ergebnisse der Simulation stimmen qualitativ mit den experimentellen Daten überein. Dabei stellt sich heraus, dass der sigmoidale Verlauf der Simulationen viel steiler als der experimentell beobachtbare ist. Es wird vermutet, dass die Diskrepanz durch extrinsische Störungen verursacht wird, die bislang nicht im Modell integriert sind. Lediglich die Variabilitäten der Proteine c-FLIP$_{\mathrm{L}}$, c-FLIP$_{\mathrm{RS}}$ und

[2] Eine detaillierte Diskussion der quadratischen Diskriminanzanalyse ist im folgenden Kapitel zu finden.

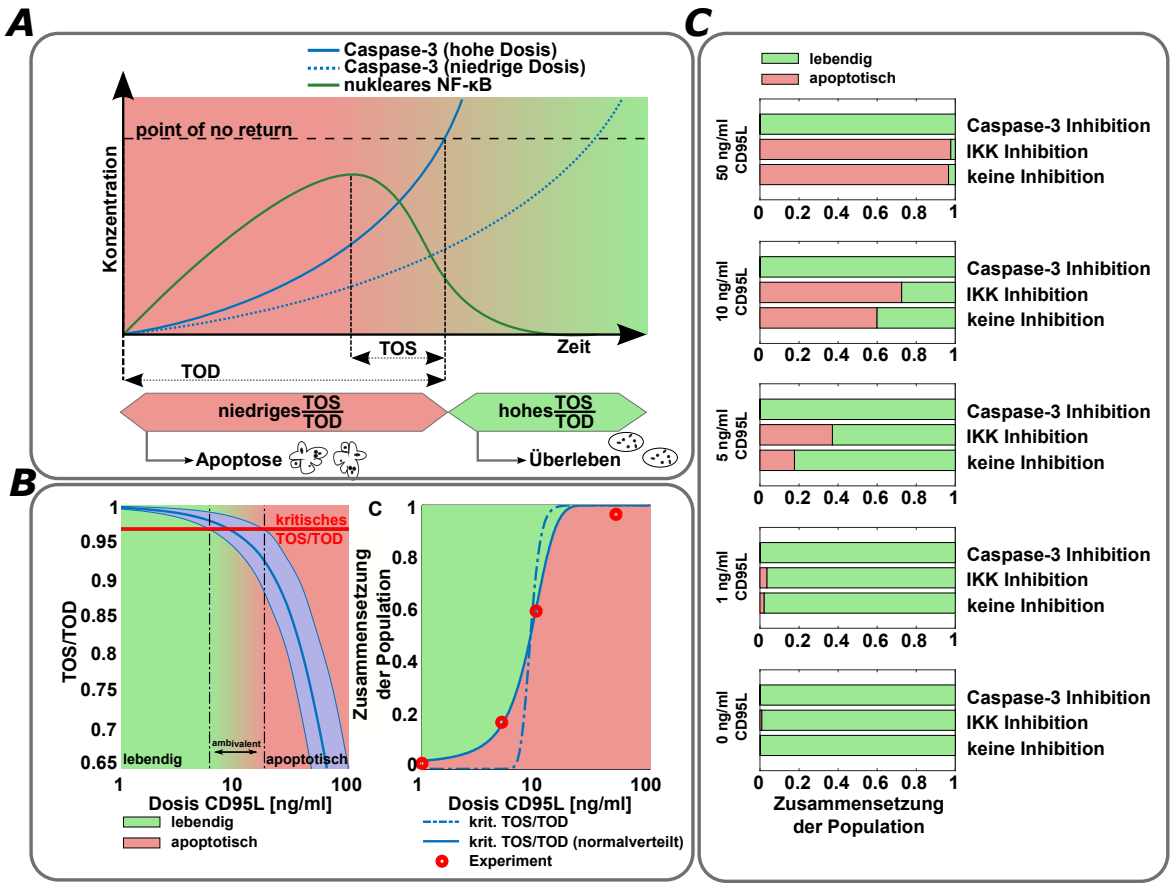

Abbildung 3.5: TOS/TOD-Mechanismus und dessen Validierung [Buchbinder et al., 2018]. **(A)** Dynamik von nuklearem NF-κB und Caspase-3 für verschiedene Stimulationsdosen. Der Point of no Return markiert die kritische Caspase-3 Konzentration, die eine Zelle in Abwesenheit der Aktivierung von NF-κB durch Apoptose tötet. Der Zeitpunkt, an dem der Point of no Return erreicht wird (TOD, Time of Decision) sowie dessen Abstand zur maximalen Aktivierung von NF-κB (TOS, Time of Survival), sind wichtige Parameter, die Aufschluss über die Entscheidung zwischen Leben und Tod geben. Zellen mit einem kleinen Verhältnis TOS/TOD tendieren dazu durch Apoptose zu sterben, wohingegen Zellen mit einem größeren Verhältnis überleben. **(B)** Das Verhältnis TOS/TOD kann in Abhängigkeit von der Stimulatiosstärke berechnet werden. Das kritische Verhältnis wird mittels Anpassung der simulierten Daten an eine experimentell ermittelte Dosis-Response-Kurve aufgenommen. Wird angenommen, dass das kritische Verhältnis ein verteilter Parameter ist, lässt sich die Qualität der Anpassung erheblich verbessern. **(C)** Zur qualitativen Überprüfung der vorgestellten Theorie sind Experimente mit unterschiedlichen Inhibitionen durchgeführt worden.

Caspase-8 sind berücksichtigt worden, da für die restlichen Proteine keine Angaben bezüglich der extrinsischen Störung vorgelegen haben. Wird nun angenommen, dass diese nicht beachteten Störungen eine Variabilität des kritischen Verhältnisses verursachen, können die Simulationen sehr genau an die experimentellen Daten angepasst werden. Das kritische Verhältnis entspricht einer Normalverteilung mit $\mathrm{E}(r_{crit}) = 0.97$ und einer Standardabweichung von 1.5% des Mittelwertes. Somit lassen sich die schwa-

che apoptotische Response für 1 ng/ml CD95L, die ambivalente Response für 5 bis 10 ng/ml CD95L und die starke apoptotische Response für 50 ng/ml CD95L erklären. Die Heterogenität der ambivalenten Response für moderate Stimulationsdosen wird vor allem durch extrinsische Störungen verursacht, die die Aktivität von Caspase-3, damit auch TOD und TOS, signifikant beeinflussen. Die Dynamik von NF-κB ist jedoch nicht sensitiv bezüglich extrinsischer Störungen und zeigt nur eine geringe Variabilität, weshalb intrinsisch stochastische Effekte der Genexpression nur eine untergeordnete Rolle spielen.

Zur Validierung des TOS/TOD-Mechanismus wird untersucht, ob durch experimentelle Manipulation der pro- und antiapoptotischen Pfadwege das Verhältnis lebendiger und apoptotischer Zellen beeinflusst werden kann. Dazu werden Experimente mit Stimulationen von bis zu 50 ng/ml CD95L für eine Zeitspanne von 15 h durchgeführt. Für jede Stimulationsdosis sind drei Messungen getätigt worden. Davon ist eine Messung durch die Inhibition des apoptotischen Pfadwegs (Caspase-3) und eine Messung durch die Inhibition des antiapoptotischen Pfadwegs (IKK) gekennzeichnet. Die übrige Messung ist ohne Inhibition durchgeführt worden, siehe Abb. 3.5C. Das Modell sagt vorher, dass durch Inhibition von Caspase-3 die Zeitspanne zwischen Stimulation und dem Erreichen des Point of no Return stark vergrößert wird. Das Maximum der Aktivierung von NF-κB bleibt unverändert, weshalb TOD und TOS annähernd gleich groß sind. Das Verhältnis von TOS/TOD ist damit etwa eins und größer als r_{krit}. Fast alle Zellen sind deshalb lebendig, wie im Experiment beobachtet. Die Inhibition von IKK geht einher mit der Inhibition von NF-κB. Caspase-3 wird durch den IKK-Inhibitor nicht beeinflusst. Die Zeitspanne zwischen Stimulation und dem Erreichen des Point of no Return bleibt deshalb unverändert. Das Maximum der Aktivierung von nuklearem NF-κB wird jedoch verzögert, wodurch TOS verkleinert wird. Das Verhältnis TOS/TOD wird damit verringert, was eine Zunahme apoptotischer Zellen bewirkt, wie im Experiment beobachtet.

3.5 Zusammenfassung

Verschiedene Modelle zur Beschreibung von Apoptose auf der Ebene einzelner Zellen sind in der Vergangenheit vorgestellt worden [Spencer et al., 2009; Xia et al., 2014; Roux et al., 2015]. Die Analyse des Zusammenspiels pro- und antiapoptotischer CD95-Pfadwege, die in dieser Arbeit unternommen worden ist, war bisher jedoch nicht existent. Nur durch das Verständnis der Synergie beider Pfadwege hinsichtlich des stochastischen Prozesses der Apoptose ist es möglich, die Entscheidung zwischen Leben und Tod einzelner Zellen vorherzusagen. Dies spielt eine wichtige Rolle im Design

und der Anwendung der medikamentösen Behandlung von Krebserkrankungen, die auf der effizienten und zuverlässigen Abtötung von Krebszellen durch Apoptose beruhen. Dabei stellt die Heterogenität und Variabilität der apoptotischen Response auf die Stimulation mittels eines Todesliganden eine große Hürde dar.

Zur Beschreibung dieser stochastischen Effekte ist in dieser Arbeit ein Modell genutzt worden, das sowohl extrinsische Störungen der Abundanzen verschiedener Proteine als auch intrinsische Störungen der Genexpression beinhaltet. Stochastische Simulationen sind äußerst rechenintensiv und erschweren damit die Anpassung des Modells an die experimentellen Daten. Mittels der in dieser Arbeit vorgestellten Kombination der Sigma Punkt Methode mit einem approximativen Gillespie-Algorithmus ist es gelungen, die unbekannten Parameter dieses komplexen Modells effizient zu identifizieren. Zur Kalibrierung des Modells sind Bulk-Populationsmessungen (Western Blot) und Einzelzellmessungen (bildgebende Flusszytometrie) genutzt worden. Das kalibrierte Modell wird dann verwendet, um Schlüsselmoleküle des pro- und antiapoptotischen Pfadwegs zu analysieren. Dabei zeigt sich, dass pro- und antiapoptotische Pfadwege durch qualitativ unterschiedliche Dynamiken charakterisiert werden. Die Aktivierung und Variabilität von Caspase-3 nimmt stetig mit steigender Stimulationsdosis zu, wohingegen NF-κB für Stimulationen von über 5 ng/ml CD95L eine robuste, dosisunabhängige Dynamik aufweist und lediglich für geringere Dosen von hoher Variabilität geprägt ist. Um das mathematische Modell mit experimentellen Daten zu verknüpfen, sind die Parameter TOD und TOS sowie deren Verhältnis eingeführt worden. Sie erlauben es, den apoptotischen Response-Mechanismus und die dosisabhängige Variabilität zu erklären. Darüber hinaus ermöglichen sie die Verknüpfung der kontinuierlichen Dynamik chemischer Spezies mit kategorischen Phänotypen (lebendig *vs.* apoptotisch). Zellen mit einem TOS/TOD Verhältnis, das einen bestimmten Schwellwert r_{krit} übersteigt, überleben die apoptotische Stimulation, wohingegen Zellen mit einem TOS/TOD Verhältnis kleiner als r_{krit} durch Apoptose sterben. Damit ist es möglich die schwache apoptotische Response für niedrige Stimulationsdosen, die ambivalente Response für moderate Stimulationsdosen und die starke apoptotische Response für hohe Stimulationsdosen zu erklären. Zudem stellt sich heraus, dass extrinsische Störungen die Hauptursache der Heterogenität bezüglich der Entscheidung zwischen Leben und Tod sind.

Die Bestimmung der optimalen Stimulationsstärke ist ein wichtiger Schritt der Krebstherapie zur Abtötung krankhafter Zellen. Die Dynamik des stochastischen Prozesses der Apoptose wird dabei oft außer acht gelassen, obwohl sie essenziell für die Entscheidung zwischen Leben und Tod ist. In dieser Arbeit sind für alle Experimente krebsähnliche Zellen des Typs HeLa-CD95L verwendet und analysiert worden. Zukünftige

Studien müssen sich damit auseinandersetzen, wie sich die Dynamik gesunder Zellen davon unterscheidet. Nur so kann sichergestellt werden, dass man Krebszellen so effizient wie möglich abtötet, ohne dabei gesundes Gewebe zu zerstören. Dies sollte, wie in dieser Arbeit demonstriert, mittels Einzelzellmessungen in Kombination mit mathematischer Modellierung geschehen.

4 Zelldiskriminierung mittels Machine-Learning

Wie bereits in den vorigen Kapiteln ausgiebig diskutiert, stellen Zellpopulationen heterogene Gruppierungen einzelner Zellen dar, die stark durch Variabilität gekennzeichnet sind. Oft bilden sich innerhalb einer Population Subpopulationen aus, die hinsichtlich ihrer Phänotypen mittels Einzelzellmessgeräten, wie dem Flusszytometer, unterschieden werden können. Bei der Flusszytometrie fließen einzelne Zellen in kurzen Zeitabständen an einem Laser vorbei und erzeugen abhängig von ihrer Form, Textur oder Färbung verschiedene Effekte, die mittels eines Detektors aufgezeichnet werden. Gemessen werden dabei zwei Eigenschaften bzw. Merkmale pro Farbkanal, die Intensität des in Laserrichtung und des senkrecht dazu gestreuten Lichtes, woraus anschließend auf den Phänotyp einzelner Zellen geschlossen wird. Das in Laserrichtung gestreute Licht korreliert mit der Zellgröße, wohingegen das senkrecht dazu gestreute Licht Aufschluss über die Textur gibt [Pedreira et al., 2008]. Die Flusszytometrie stellt einen Meilenstein in der biomedizinischen Forschung dar [Herzenberg et al., 2002] und erwies sich als äußerst nützlich hinsichtlich der Erforschung verschiedener biologischer Phänomene [Jaye et al., 2012]. Sie zeichnet sich durch ihren hohen Zelldurchsatz aus, liefert jedoch nur eine geringe Anzahl gemessener Merkmale. Aus diesem Grund werden die Phänotypen der Zellen traditionell mittels zweidimensionaler Gating-Methoden, oft unter Verwendung von Fluoreszenzfarbstoffen, bestimmt. Dazu werden manuell Grenzen in einem zweidimensionalen Merkmalsraum gezogen, die Subpopulationen mit unterschiedlichen Phänotypen separieren. Die Festlegung der Grenzen sowie die Auswahl der verwendeten Merkmale beruhen auf dem subjektiven Urteil des Wissenschaftlers und sind deshalb nicht eindeutig reproduzierbar [O'Neill et al., 2013; Saeys et al., 2016].

Mit der fortschreitenden technologischen Entwicklung sind Flusszytometer weiter optimiert worden, was sich beispielsweise in der steigenden Anzahl der zur Verfügung stehender Farbkanäle zeigt [Perfetto et al., 2004]. Messungen mittels Flusszytometrie lieferten immer größere und komplexere Datensätze, wodurch deren Analyse und Interpretation erschwert worden ist. Dabei stellt sich heraus, dass manuelle Gating-

Methoden sehr ineffizient sind, da sie nur eine eingeschränkte Sicht mittels der zweidimensionalen Betrachtungsweise erlauben. Zudem sind manuelle Methoden sehr zeitaufwendig, da für Messungen an unterschiedlichen Tagen oder mit unterschiedlichen Patienten bzw. Proben stets kleine Anpassungen der Grenzen händisch vorgenommen werden müssen [O'Neill et al., 2013; Saeys et al., 2016]. Ein gewaltiger Sprung ist mit der Entwicklung der bildgebenden Flusszytometrie verzeichnet worden. Diese stellt eine technische Weiterentwicklung der gewöhnlichen Flusszytometrie dar, die Mikroskopie und Flusszytometrie miteinander kombiniert [Basiji et al., 2007]. Sie liefert Hochdurchsatzdaten in Form einzelner Bilder mit mikroskopischer Auflösung jeder Zelle, aus denen Hunderte von Merkmalen extrahiert werden können [Saeys et al., 2016]. Traditionelle Gating-Methoden stoßen bei der Analyse derartiger Daten schnell an ihre Grenzen, weshalb neue, speziell auf Big Data Probleme zugeschnittene Ansätze etabliert werden müssen [Fan et al., 2014].

Machine-Learning stellt einen vielversprechenden Ansatz dar, der Objekte oder Samples basierend auf ihren Merkmalen kategorischen Klassen zuordnet. Im Kontext dieser Arbeit wird darunter die Klassifizierung einzelner Zellen hinsichtlich ihres Phänotyps verstanden. Die Zuordnung beruht auf algorithmischen Regeln, die ein objektives Vorgehen erlauben und dabei subjektive Einflüsse durch das Vorwissen und die Erfahrung des Wissenschaftlers minimieren. Das Ziel ist es, die Klasse bzw. den Phänotyp einer Zelle anhand ihrer mit bildgebender Flusszytometrie gemessenen Merkmalen korrekt zu bestimmen. Da die bildgebende Flusszytometrie ein recht junges Verfahren zur Einzelzellmessung ist, hat sich noch kein einheitliches Vorgehen zur Analyse und Interpretation der dabei anfallenden Daten herausgebildet. Aus diesem Grund wird in dieser Arbeit eine einfache Methode präsentiert, die es erlaubt, effizient die signifikante Information aus dem hochdimensionalen Merkmalsraum zu extrahieren und diese für eine präzise Klassifizierung zu nutzen. Dazu wird in den folgenden Abschnitten ein Überblick hinsichtlich etablierter Machine-Learning-Methoden gegeben und demonstriert, wie diese in der in dieser Arbeit vorgestellten Methode integriert werden können. Die Nützlichkeit der Methode wird anschließend am Beispiel der Detektion des Zelltodes demonstriert. Dabei wird versucht, apoptotische und lebendige Zellen einer heterogenen Population zu diskriminieren, siehe Abb. 4.1.

4.1 Von der Messung zur Klassifizierung

Die bildgebende Flusszytometrie ermöglicht es, Hochdruchsatzmessungen mittels der Aufnahme von Bildern einzelner Zellen in verschiedenen Farbkanälen durchzuführen, siehe Abb. 4.2A. Die Einzellzellbilder beinhalten Störungen und Artefakte, wie bei-

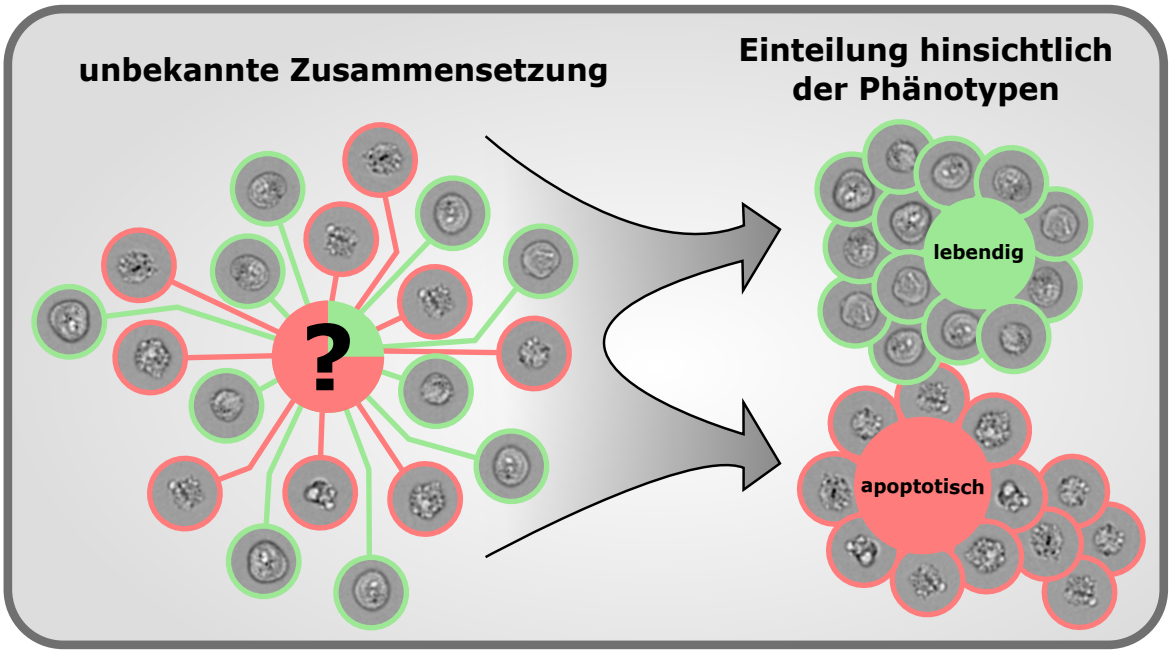

Abbildung 4.1: Zelldiskriminierung mittels bildgebender Flusszytometrie [Pischel et al., 2018].

spielsweise überlappende Zellen oder Luftblasen, die durch angemessene Vorverarbeitung entfernt werden können. Aus den artefaktfreien Bildern kann anschließend Information extrahiert werden, die jede Zelle durch numerische Merkmale charakterisiert. Die Merkmale stellen physikalisch messbare Größen dar und können in vier Gruppen eingeteilt werden [Pischel et al., 2018]:

- morphologische Merkmale
- spektrale Merkmale
- morpho-spektrale Merkmale
- abstrakte Merkmale

Morphologische Merkmale geben Aufschluss über die Form, die Textur sowie die Beschaffenheit der Zellen und werden vor allem aus dem Hellfeldkanal abgeleitet. Im Gegensatz dazu basieren spektrale Merkmale auf der Fluoreszenz von Farbstoffen mit deren Hilfe die Abundanz gewisser Proteine, das Membranpotenzial oder die Membranpermeabilität abgeschätzt werden können. Durch Kombination morphologischer und spektraler Merkmale (morpho-spektral) ist es möglich, die Translokation von Farbstoffen in bestimmte Kompartimente zu beobachten und damit gezielt die Dynamik von Zellorganellen zu analysieren. Die bisher aufgezählten Merkmale beruhen auf Transmissionsmessungen und lassen durch den Aufbau des Messgerätes keine Kontamination durch Streulicht zu. Komplementär dazu ist es mittels des Dunkelfeldkanals möglich,

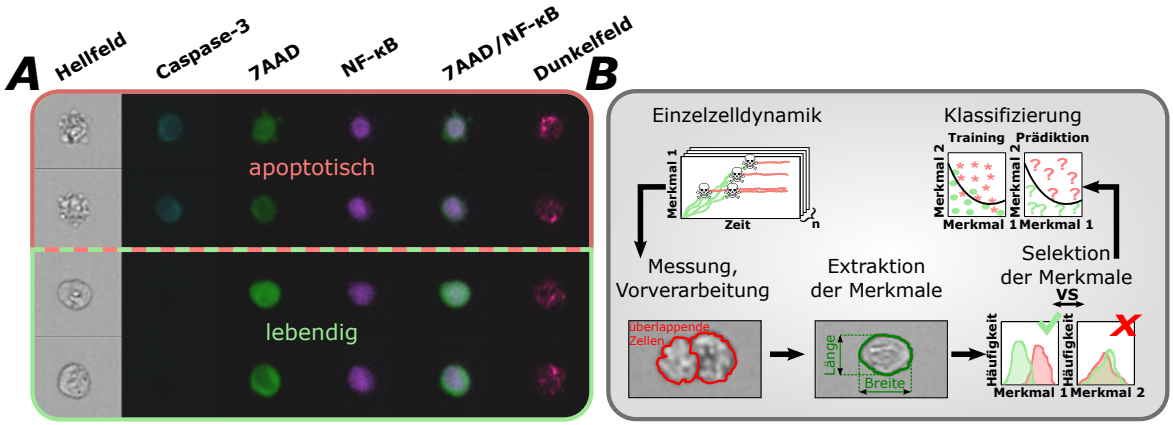

Abbildung 4.2: Bildgebende Flusszytometrie: Messergebnisse und schematisches Vorgehen zur Datenanalyse [Pischel et al., 2018]. **(A)** Zwei apoptotische und lebendige Zellen wurden mit bildgebender Flusszytometrie in verschiedenen Farbkanälen vermessen. Aus diesen Bildern können Hunderte von Merkmalen extrahiert werden. **(B)** Vorgehen zur Analyse von experimentellen Daten gemessen mit bildgebender Flusszytometrie.

das Streulicht unter Ausschluss des transmittierten Lichtes zu messen. Merkmale aus dem Dunkelfeldkanal sind sehr abstrakt und ihre Interpretation ist oft nicht trivial. Dennoch enthalten sie wertvolle Information, die zur Klassifizierung genutzt werden kann und nicht verworfen werden sollte [Blasi et al., 2016; Hennig et al., 2017].

Aufgrund der hohen Dimension des Merkmalsraumes sind viele Merkmale korreliert, redundant und unterscheiden sich stark im Informationsgehalt. Folglich ergibt sich die Frage, welche Merkmale für die Klassifizierung zu verwenden sind und wie diese ausgewählt werden sollen, um eine optimale Diskriminierung zu gewährleisten. Dies ist keine triviale Aufgabe, da die Anzahl aller möglichen Kombinationen von Merkmalen exponentiell mit der Anzahl der Merkmale wächst. Zur Auswahl oder Selektion der Merkmale haben sich verschiedene Ansätze etabliert, die grob in drei Klassen gegliedert werden können [Saeys et al., 2007]:

- Filter
- Wrapper
- eingebettete Methoden

Filter sind unabhängig vom Machine-Learning-Algorithmus, der für die Klassifizierung genutzt wird und wählen Merkmale ausschließlich anhand ihrer intrinsischen Charakteristiken aus. Im Gegensatz dazu verwenden Wrapper die Genauigkeit des Machine-Learning-Algorithmus, um Merkmale zu wählen. Eingebettete Methoden stellen wiederum einen Machine-Learning-Algorithmus mit integrieter Selektion der Merkmale dar. Da Wrapper und eingebettete Methoden sehr rechenintensiv sind und zur Über-

anpassung der Daten neigen, fokussiert diese Arbeit auf Filter, die sich durch Effizienz und Einfachheit auszeichnen [Bolón-Canedo et al., 2013]. Besonders der geringe Rechenaufwand und die Skalierbarkeit zu großen Datensätzen mit einer enormen Anzahl von gemessenen Merkmalen und Samples machen Filter zu idealen Kandidaten hinsichtlich der Analyse von Experimenten, durchgeführt mit bildgebender Flusszytometrie.

Um die vermessenen Zellen hinsichtlich ihrer Phänotypen zu unterscheiden, wird ein algorithmisches Verfahren zur Klassifizierung verwendet. Zunächst wird der Klassifizierungsalgorithmus auf Daten mit bekanntem Phänotyp trainiert. Der Algorithmus erkennt dabei Muster und Strukturen innerhalb der Daten anhand derer neue, bisher ungesehene Zellen mit unbekannten Phänotypen klassifiziert werden. In der Regel sind Machine-Learning-Algorithmen abhängig von Hyperparametern, die die Genauigkeit der Klassifizierung signifikant beeinflussen. Demnach steht eine Schar von Modellen zur Verfügung, aus der das beste Modell mittels Optimierung identifiziert werden kann. Um das Modell, das die verschiedenen Klassen bestmöglich unterscheidet, zu wählen, werden Genauigkeit, Robustheit gegen Störungen der Daten und Generalisierung zu unbekannten Daten als Kriterien in Betracht gezogen.

Das hier beschriebene modulare Vorgehen von der Messung bis hin zur erfolgreichen Klassifizierung heterogener Zellpopulationen ist in Abb. 4.2B schematisch dargestellt. Das experimentelle Vorgehen beschränkt sich dabei vor allem auf die Präparation des zu beobachtenden biologischen Systems, der Messung und der Vorverarbeitung der gewonnenen Daten zur Bereinigung von Artefakten. Die Extraktion der Merkmale kann direkt mittels der Software des Messgerätes [Basiji et al., 2007] oder nachträglich unter Zuhilfenahme externer Software [Eliceiri et al., 2012] erfolgen. Im Gegensatz dazu stellen die Auswahl der Merkmale sowie das anschließende Training des Machine-Learning-Algorithmus und dessen Anwendung auf neue Daten hauptsächlich rechentechnische Verfahren dar, die im folgenden näher beschrieben werden.

4.2 Selektion der Merkmale mittels Filterung

Die bildgebende Flusszytometrie ermöglicht es, jede Zelle mittels einer enormen Anzahl von Merkmalen zu charakterisieren. Dies erlaubt es im hochdimensionalen Merkmalsraum nach Mustern und Strukturen zu suchen, die zur Klassifizierung der Zellen genutzt werden können (Segen der Dimension). Parallel nimmt jedoch das Volumen des Merkmalsraumes mit zunehmender Anzahl der Merkmale zu, wodurch es vorkommen kann, dass die Daten dünn besetzt sind (Fluch der Dimension). Da Machine-

Learning-Methoden auf statistischer Signifikanz beruhen, stellt sich die Analyse von Datensätzen mit einer großen Anzahl von Merkmalen und einer relativ dazu geringen Anzahl von Samples als problematisch dar [Saeys et al., 2007]. Aus diesem Grund wird versucht, die Dimension des Problems zu reduzieren und in einem Unterraum des Merkmalsraumes vorzunehmen. Neben Projektionsmethoden, wie der Hauptkomponentenanalyse, der multidimensionalen Skalierung oder der Unabhängigkeitsanalyse, hat sich die Selektion der Merkmale als effiziente Methode zur Dimensionsreduktion etabliert [Sommer et al., 2013]. Das Ziel der Selektion besteht darin, eine Menge von Merkmalen auszuwählen, die eine optimale Klassifizierung erlaubt. Die Selektion der Merkmale kann als Abbildung h vom n-dimensionalen Merkmalsraum in einen n_{FS}-dimensionalen Unterraum verstanden werden

$$h: \ \mathbb{R}^n \rightarrow \mathbb{R}^{n_{FS}} \ \text{ mit } \ \mathbb{R}^{n_{FS}} \subseteq \mathbb{R}^n. \tag{4.1}$$

Dabei bezeichnet n_{FS} die Anzahl der ausgewählten Merkmale. Um die Merkmale entsprechend ihres Informationsgehaltes anzuordnen, wird eine Sortierung mittels eines Filters vorgenommen. Nur die n_{FS} Merkmale mit dem größten Informationsgehalt werden für die Klassifizierung verwendet.

Filter können hinsichtlich des verwendeten Informationsmaßes in uni- und multivariate Verfahren unterteilt werden [Saeys et al., 2007; Bolón-Canedo et al., 2013]. Univariate Verfahren ordnen die Merkmale ausschließlich anhand ihrer Korrelation den verschiedenen Klassen zu. Die Interaktion zwischen verschiedenen Merkmalen wird dabei außer Acht gelassen. Dies erlaubt es, die Relevanz eines Merkmals zu beurteilen, jedoch nicht die Redundanz zu anderen Merkmalen. Im Gegensatz dazu beziehen multivariate Verfahren Relevanz und Redundanz in den Informationsgehalt der Merkmale ein, was in einem höheren Rechenaufwand resultiert.

Im Folgenden wird angenommen, dass die Daten in Matrixform $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ gegeben sind, wobei m die Anzahl der beobachteten Samples und n die Anzahl der Merkmale darstellen. Des Weiteren werden die Zufallsvariblen $\boldsymbol{x}$ und χ_i eingeführt. Dabei bezeichnet $\boldsymbol{x}$ einen Vektor, der die Realisierungen aller Merkmale eines bestimmte Samples beinhaltet, wohingegen χ_i das i-te Merkmal darstellt. Zusätzlich ordnet der Vektor $\boldsymbol{y}$ bestimmten Samples eine der n_c verschiedenen Klassen $\boldsymbol{c} = \{c_1, \ldots, c_{n_c}\}$ zu.

4.2.1 Transinformation

Die Transinformation oder auch gegenseitige Information $\boldsymbol{I}$ ist ein Maß aus der Informationstheorie, das angibt, wie groß der Informationszuwachs hinsichtlich einer Zufallsvariablen ist, wenn eine andere Zufallsvariable bekannt ist. Sie kann dazu genutzt

werden, um die Korrelation zwischen Merkmalen und Klassen zu bestimmen

$$\boldsymbol{I}(\chi_i, \boldsymbol{c}) = \sum_j \sum_k P(\chi_{i,j}, c_k) \log \left(\frac{P(\chi_{i,j}, c_k)}{P(\chi_{i,j}) P(c_k)} \right). \tag{4.2}$$

Diese Darstellung ist gültig für diskretisierte Merkmale $\chi_{i,j} \subseteq \chi_i$, wobei P die Wahrscheinlichkeit, j den Diskretisierungsindex und k den Klassenindex bezeichnet. Die Merkmale werden anschließend anhand ihrer Transinformation geordnet [Lewis, 1992; Brown et al., 2012]. Dabei signalisieren Merkmale mit einer hohen Transinformation bezüglich der Klassen ein großes Diskriminierungspotenzial und folglich einen hohen Informationsgehalt. Da ausschließlich intrinsische Eigenschaften der Merkmale und keine Interaktionen untereinander in Betracht gezogen werden, stellt dieses Verfahren einen univariaten Ansatz zur Selektion dar. Zur Veranschaulichung sind in Abb. 4.3A drei eindimensionale Zweiklassenprobleme dargestellt. Mittels der Transinformation kann das hohe Diskriminierungspotenzial der nicht überlappenden, separablen Verteilungen und der geringe Informationsgehalt der stark überlappenden Verteilungen erkannt werden. Ein weiterer Vorteil der Transinformation im Vergleich zu anderen Korrelationsmaßen, wie dem Korrelationskoeffizienten, ist, dass mittels Transinformation auch nichtlineare Zusammenhänge erkannt werden, siehe Abb. 4.3B.

4.2.2 Minimale Redundanz und maximale Relevanz

Das Prinzip der minimalen Redundanz und maximalen Relevanz (MRMR) [Peng et al., 2005; Brown et al., 2012] ist ein multivariates, informationstheoretisches Verfahren zur Selektion von Merkmalen, das sowohl den Informationsgehalt als auch Abhängigkeiten zu anderen Merkmalen einbezieht. Der Selektionsprozess ist ein iteratives Verfahren, das ein Merkmal nach dem anderen zu der Menge der ausgewählten Merkmale $\boldsymbol{S}$ hinzufügt. Das erste Merkmal wird anhand der Transinformation gewählt. Anschließend wird das Merkmal zu $\boldsymbol{S}$ hinzugefügt, das den größten MRMR-Wert Θ aufweist

$$\Theta(\chi_i, \boldsymbol{c}) = I(\chi_i, \boldsymbol{c}) - \frac{1}{|\boldsymbol{S}|} \sum_{s \subseteq \boldsymbol{S}} I(\chi_s, \chi_i). \tag{4.3}$$

Dabei bezeichnet der erste Term die Korrelation zwischen einem Merkmal und den Klassen (Relevanz), wohingegen der zweite Term die Korrelation zwischen einem Merkmal und den bereits gewählten Merkmale misst (Redundanz). Nachdem $\boldsymbol{S}$ aktualisiert worden ist, wird das Schema wiederholt, bis die gewünschte Anzahl an Merkmalen erreicht ist.

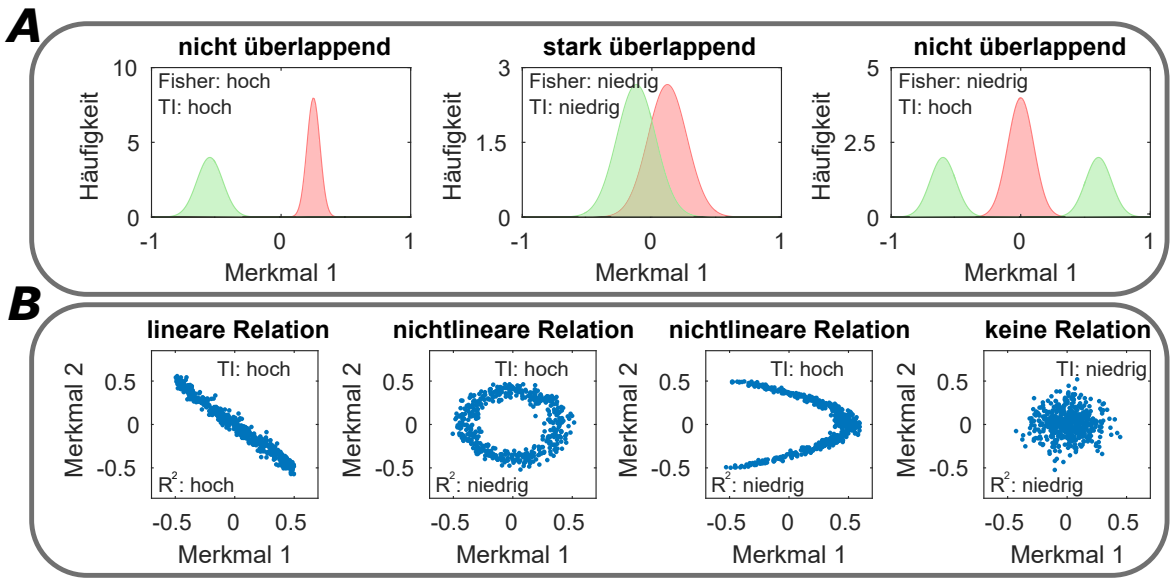

Abbildung 4.3: Selektion der Merkmale [Pischel et al., 2018]. **(A)** Die Transinformation (TI) und der Fisher-Wert sind univariate Verfahren zur Schätzung des Informationsgehaltes eines Merkmals, die auf zwei separable und ein nicht separables Zweiklassenproblem (rot und grün) angewendet werden. **(B)** Im Vergleich zu anderen Verfahren, wie dem Korrelationskoeffizienten, ist die Transinformation in der Lage, nichtlineare Relationen zu erkennen.

4.2.3 Fisher-Wert

Der Fisher-Wert beruht auf der Annahme, dass Merkmale mit einem großen Informationsgehalt Samples einer Klasse ähnliche Werte zuordnen, die sich stark von denen anderer Klassen unterscheiden. Solche Merkmale besitzen Klassenmittelwerte, die relativ zu ihrer Standardabweichung weit auseinander liegen. Mathematisch formuliert lässt sich dies folgendermaßen ausdrücken

$$\psi(\chi_i, \boldsymbol{c}) = \frac{\sum_k m_k(\mu_{i,k} - \mu_i)^2}{\sum_k m_k \sigma_{i,k}^2}. \tag{4.4}$$

Dabei bezeichnen μ_i und σ_i den Mittelwert und die Standardabweichung des i-ten Merkmals. Wie zuvor gibt m die Anzahl der Samples an. Ist zusätzlich der Klassenindex k angegeben, beziehen sich die Mittelwerte $\mu_{i,k}$, Standardabweichungen $\sigma_{i,k}$ und Anzahl der Samples $m_{i,k}$ ausschließlich auf die k-te Klasse. Die Selektion der Merkmale mittels des Fisher-Wertes ist ein univariates Verfahren, das keine Abhängigkeiten der Merkmale erfasst. In Abb. 4.3A ist der Fisher-Wert für drei eindimensionale Zweiklassenprobleme illustriert. Er ist in der Lage, den hohen Informationsgehalt der nicht überlappenden, linear separablen Verteilungen und den niedrigen Informationsgehalt der überlappenden Verteilungen zu erkennen. Er ist jedoch nicht in der Lage, das hohe Diskriminierungspotenzial der linear nicht separablen, nicht überlappenden Verteilun-

gen auszumachen.

4.3 Klassifizierung mittels Machine-Learning

Machine-Lerning bietet die Möglichkeit, heterogene Zellpopulationen hinsichtlich ihrer Phänotypen auf eine effiziente Art und Weise zu klassifizieren. Die Klassifizierung stellt ein algorithmisches Verfahren dar, das die Klassenzugehörigkeit einzelner Samples ausschließlich anhand ihrer Merkmale bestimmt. Dies kann als Abbildung f verstanden werden, die aus dem n-dimensionalen Merkmalsraum auf eine der möglichen n_c Klassen abbildet

$$f(\boldsymbol{x}) : \mathbb{R}^n \rightarrow \{c_1, \ldots, c_{n_c}\}. \tag{4.5}$$

Der zufriedenstellendste Ansatz der Klassifizierung besteht darin, Samples der Klasse zuzuordnen, die bezüglich der Realisierung der Merkmale $\boldsymbol{x}$ am wahrscheinlichsten ist. Unglücklicherweise ist es in der Regel nicht möglich, die zugrunde liegende Wahrscheinlichkeitsverteilung zu rekonstruieren [Jain et al., 2000], weshalb Annahmen bezüglich der Verteilung oder heuristische Algorithmen zur Klassifikation angewendet werden. Im Folgenden werden vier populäre Machine-Learning-Algorithmen vorgestellt, die im weiteren Verlauf dieser Arbeit genutzt werden.

4.3.1 Diskriminanzanalyse

Die lineare und quadratische Diskriminanzanalyse (LDA und QDA) ordnen unbekannte Samples derjenigen Klasse zu, die die *a posteori*-Wahrscheinlichkeit $P(c_k|\boldsymbol{x})$ maximiert. Diese kann mittels des Satz von Bayes wie folgt ausgedrückt werden

$$P(\boldsymbol{c}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{c})P(\boldsymbol{c})}{P(\boldsymbol{x})}. \tag{4.6}$$

Wird angenommen, dass die Wahrscheinlichkeit des Auftretens der verschiedenen Klassen gleichverteilt ist, zum Beispiel weil kein Vorwissen bezüglich der Verteilung der Klassen vorhanden ist, lässt sich die Klassifizierung als Optimierungsproblem formulieren

$$f(\boldsymbol{x}) = \arg \max_{c_k \in \boldsymbol{c}} P(\boldsymbol{x}|c_k). \tag{4.7}$$

Die bedingte Wahrscheinlichkeit $P(\boldsymbol{x}|c_k)$ wird als multivariate Normalverteilung angenommen mit verschiedenen Mittelwerten für jede Klasse. Wird zusätzlich von einer gemeinsamen Kovarianzmatrix aller Klassen ausgegangen, ergibt sich eine lineare Separationsgrenze (LDA), die zur Trennung der Klassen genutzt wird. Im Gegensatz dazu

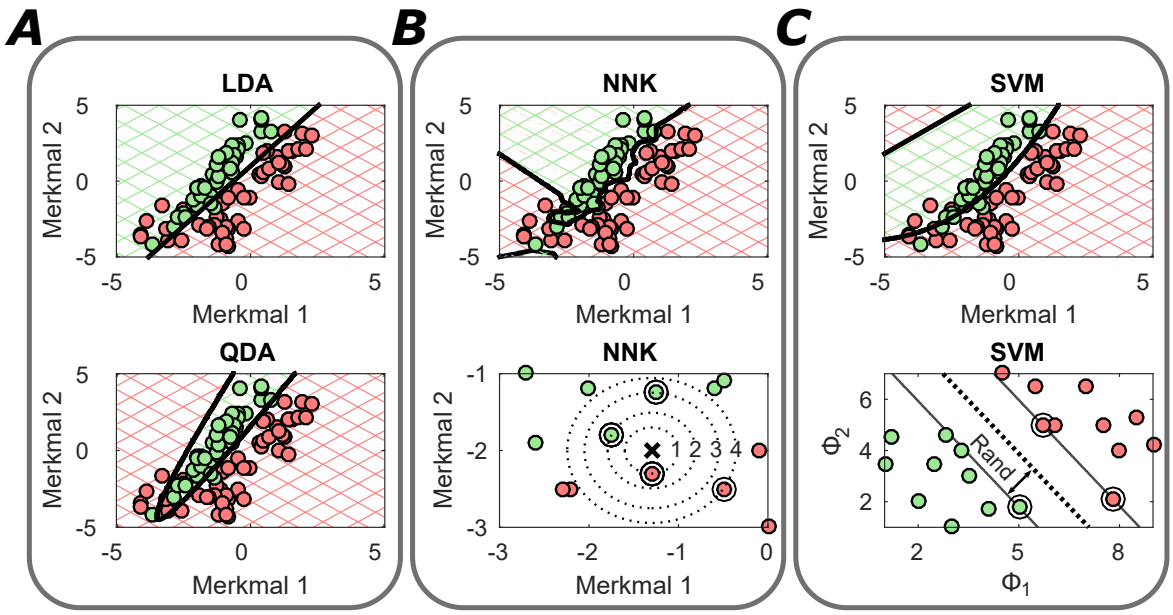

Abbildung 4.4: Klassizizierung mittels Machine-Learning [Pischel et al., 2018]. Die Klassifizierung eines Zweiklassenproblems (rot und grün) mittels Machine-Learning-Algorithmen unterschiedlicher Komplexität zeigt Variationen hinsichtlich der Separationsgrenze. Dies ist für **(A)** die lineare (LDA) bzw. quadratische Diskriminanznalyse (QDA), **(B)** die Nächste-Nachbarn-Klassifizierung (NNK) und **(C)** den Support-Vector-Machine-Algorithmus (SVM) dargestellt.

führt die Annahme von klassenabhängigen Kovarianzen zu einer nichtlinearen Separationsgrenze (QDA). Beide Machine-Learning-Algorithmen sind äußerst robust gegen Störungen und einfach zu implementieren, da sie unabhängig von algorithmenspezifischen Hyperparametern sind [Jain et al., 2000]. In Abb. 4.4A sind die Anwendungen beider Algorithmen auf ein nicht separables Zweiklassenproblem illustriert. Es stellt sich heraus, dass die QDA mittels der nichtlinearen Separationsgrenze in der Lage ist, die Klassen besser zu trennen als die LDA.

4.3.2 Nächste-Nachbarn-Klassifizierung

Die Nächste-Nachbarn-Klassifizierung (NNK) ist ein Machine-Learning-Alorithmus, dessen Klassenzuordnung auf der Mehrheitsentscheidung der k nächsten Nachbarn beruht. Dabei wird einem unbekannten Sample die am häufigsten auftretende Klasse der nächsten Nachbarn zugeordnet [Hastie et al., 2003]. Zur Berechnung der Distanzen zwischen den einzelnen Samples können verschiedene Metriken verwendet werden [France et al., 2012]. In dieser Arbeit wird jedoch ausschließlich die euklidische Metrik benutzt. In Fig. 4.4B ist die NNK am Beispiel eines nicht separablen Zweiklassenproblems dargestellt. Die NNK ist in der Lage, die Nichtlinearität mittels einer rauen Separations-

grenze abzubilden. Je mehr nächste Nachbarn für die Klassifizierung eines unbekannten Samples (schwarzes Kreuz) genutzt werden, desto größer wird der Einfluss entfernter Punkte und desto glatter erscheint die Separationsgrenze, siehe Fig. 4.4B. Die NNK ist ein einfacher Klassifizierungsalgorithmus, der nicht auf Annahmen bezüglich der zugrunde liegenden Wahrscheinlichkeitsverteilung der Merkmale beruht. Aus diesem Grund ist die NNK einfach zu implementieren und anzuwenden. Der Nachteil des Algorithmus besteht im immensen Rechenaufwand, besonders für große Datenmengen, da die Distanzen zu allen Trainingsdaten berechnet werden müssen.

4.3.3 Support-Vector-Machine

Um Samples verschiedener Klassen zu unterscheiden, bildet eine Support-Vector-Machine (SVM) den Merkmalsraum mittels einer nichtlinearen Transformation $\boldsymbol{\Phi}(\boldsymbol{x})$ in einen höherdimensionalen Raum ab. Dort werden unbekannte Samples mittels eines Modells folgender Form klassifiziert [Tarca et al., 2007; Ben-Hur et al., 2008]

$$f(\boldsymbol{x}) = \operatorname{sgn}(\boldsymbol{w}^\top \cdot \boldsymbol{\Phi}(\boldsymbol{x}) + b). \tag{4.8}$$

Dabei bezeichnet $\boldsymbol{w}$ einen Wichtungsvektor und b eine additive Konstante. Die nichtlineare Transformation kann effizient mittels des Kernel-Tricks, mit beispielsweise gaußschen oder polinomialen Kerneln κ, berechnet werden. Das Modell lässt sich damit wie folgt formulieren [Cortes et al., 1995]

$$f(\boldsymbol{x}) = \sum_{i=2}^{m} y_i \alpha_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{4.9}$$

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{\Phi}(\boldsymbol{x}_i)^\top \boldsymbol{\Phi}(\boldsymbol{x}). \tag{4.10}$$

Die Faktoren α_i bezeichnen in diesem Fall Lagrange-Multiplikatoren. Zur Veranschaulichung ist in Fig. 4.4C die Anwendung der SVM mit einem gaußschen Kernel auf das linear nicht separable Zweiklassenproblem gezeigt. Die Separationsgrenze ist so festgelegt, dass sie den Rand zwischen beiden Klassen maximiert. Der Rand ist dabei als kürzeste Distanz zwischen der Separationsgrenze und den Samples der Trainingsdaten definiert. Zudem werden die Samples, die sich auf dem Rand befinden (eingekreiste Punkte), als Support Vectors bezeichnet, siehe Fig. 4.4C. Sollten die zugrunde liegenden Verteilungen der Merkmale verschiedener Klassen stark überlappen, führt eine exakte Separation in der Regel zu einer schlechten Generalisierung hinsichtlich neuer Daten. Deshalb wird die Komplexität des Machine-Learning-Algorithmus durch Regularisierung gemindert. Damit werden Missklassifikationen erlaubt und entsprechend bestraft, wodurch sich eine glatte Separationsgrenze ergibt. Die SVM ist ein sehr kom-

plexer Machine-Learning-Algorithmus, der wegen seines immensen Speicherbedarfs besonders für große Datensätze schnell an seine Grenzen stößt. Aus diesem Grund wird in dieser Arbeit die Low Rank Linearization SVM verwendet [Zhang et al., 2012; Djuric et al., 2013], die eine approximative Version der SVM darstellt.

4.4 Modellwahl

Machine-Learning stellt ein effizientes Verfahren zur automatisierten Klassifizierung großer Datenmengen dar. Um eine präzise Klassifizierung zu gewährleisten, müssen verschiedene Größen an den jeweiligen Datensatz angepasst werden. Dazu zählen die Hyperparameter der Machine-Learning-Algorithmen, wie die Anzahl der nächsten Nachbarn k, aber auch die Zusammensetzung und Anzahl der verwendeten Merkmale. Jeder Machine-Learning-Algorithmus mit einer bestimmten Einstellung dieser Größen repräsentiert ein Modell, das trainiert werden kann, um es anschließend zur Klassifizierung unbekannter Objekte zu verwenden. Das Ziel der Modellwahl ist es aus der Schar aller möglichen Modelle das optimale Modell zu identifizieren, das die verschiedenen Klassen bestmöglich unterscheidet. Um die Genauigkeit eines Modells zu ermitteln, wird häufig die 0-1-Gewinnfunktion Γ verwendet

$$\Gamma(y_i, f(\boldsymbol{x}_i)) = \begin{cases} 1 & \text{, wenn } y_i = f(\boldsymbol{x}_i) \\ 0 & \text{, sonst} \end{cases}, \tag{4.11}$$

die eine eins zuordnet, wenn ein Sample korrekt klassifiziert wurde und eine 0, wenn nicht. Durch Mittelung über alle Samples ergibt sich die Genauigkeit Δ

$$\Delta = \frac{1}{m} \sum_{i=1}^{m} \Gamma(y_i, f(\boldsymbol{x}_i)), \tag{4.12}$$

die als Maß für die Güte der Klassifizierung verstanden werden kann.

Zur Identifizierung des optimalen Modells wird zunächst der Trainingsdatensatz $\boldsymbol{D}$ in zwei Teile $\boldsymbol{D}_{MW} \subset \boldsymbol{D}$ und $\boldsymbol{D}_{Test} = \boldsymbol{D} \backslash \boldsymbol{D}_{MW}$ geteilt, siehe Abb. 4.5A. Dabei wird die Strategie verfolgt verschiedene, konkurrierende Modelle auf $\boldsymbol{D}_{MW}$ zu trainieren und deren Genauigkeit grob abzuschätzen (Modellwahl). Im Gegensatz dazu wird der zweite Teil $\boldsymbol{D}_{Test}$ zur zuverlässigen Ermittlung der Modellgenauigkeit verwendet. Mittels Kreuzvalidierung (KV) wird $\boldsymbol{D}_{MW}$ weiter in k_{KV} Teilmengen $\boldsymbol{D}_{MW,1}, \ldots, \boldsymbol{D}_{MW,k_{KV}}$ gleicher Größe und Klassenkomposition gegliedert. Jedes Modell wird k_{KV}-mal auf $k_{KV} - 1$ Teilmengen trainiert und anschließend auf der ausgelassenen Teilmenge validiert. Das optimale Modell maximiert die über alle Iterationen der KV gemittelte Genauigkeit und zeichnet sich damit durch seine Robustheit gegen Störungen der

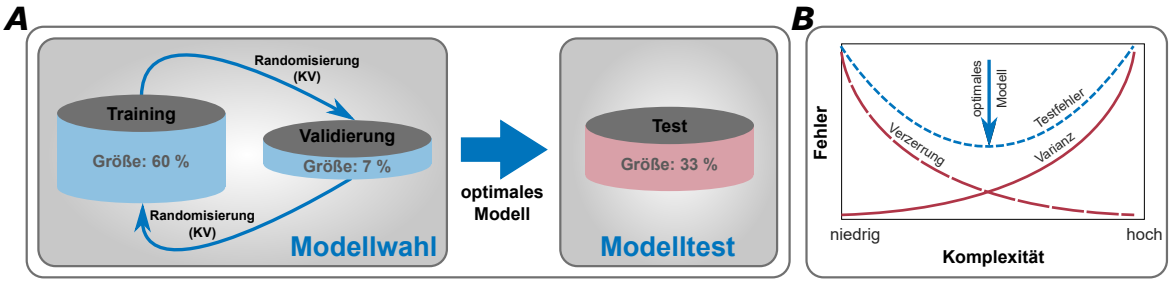

Abbildung 4.5: Modellwahl [Pischel et al., 2018]. **(A)** Der Trainingsdatensatz wird in zwei Teile gegliedert. Ein Teil (67% der Daten) wird zur Modellwahl genutzt, wohingegen der zweite Teil (33% der Daten) zum Testen der Genauigkeit verwendet wird. Um das Modell hinsichtlich Robustheit gegen Störungen der Trainingsdaten zu untersuchen, wird eine Kreuzvalidierung in der Modellwahl integriert. **(B)** Zur Identifizierung des optimalen Modells (blauer Pfeil) wird die Summe der Verzerrung (rot, gestrichelt) und der Varianz (rot, durchgezogen) minimiert.

Trainingsdaten, Generalisierung zu unbekannten Daten und Präzision aus. Da die ermittelte Genauigkeit des optimalen Modells basierend auf Daten ermittelt worden ist, die bereits für die Kalibrierung genutzt worden sind, ist die Genauigkeit nicht zuverlässig und möglicherweise überoptimistisch [Lemm et al., 2011]. Aus diesem Grund wird das Modell auf $\boldsymbol{D}_{MW}$ erneut trainiert und auf den bisher unangerührten Teil des Trainingsdatensatzes $\boldsymbol{D}_{Test}$ getestet. Anschließend wird das Modell auf dem gesamten Datensatz trainiert und kann für die Klassifizierung neuer, bisher unbekannter Daten verwendet werden. Auf diese Weise ist es möglich das optimale Modell mit einer zuverlässigen Fehlerabschätzung zu erstellen, das auf den vollständigen Trainingsdaten kalibriert wird.

Die Genauigkeit eines Modells wird stark von den Eigenschaften der verwendeten Attribute beeinflusst. Besonders ausgeprägt ist die Einflussnahme von Merkmalen, die sich um mehrere Größenordnungen unterscheiden, da Merkmale auf großen Skalen leicht die Klassifizierung dominieren können. Um dies zu verhindern und eine gerechte Wichtung zu erzwingen, sollten Merkmale vor der Klassifizierung transformiert werden. Die am häufigsten verwendeten Transformationen stellen die Standardisierung und die Normalisierung dar. Bei der Standardisierung werden die Daten so transformiert, dass ihr Mittelwert und ihre Varianz festgelegte Werte annehmen. Für den Fall, dass der Mittelwert auf null und die Standardabweichung auf eins gesetzt werden, ergeben sich die standardisierten Merkmale $\chi_{i,stand}$ mittels

$$\chi_{i,stand} = \frac{\chi_i - \mathrm{E}(\chi_i)}{\mathrm{std}(\chi_i)}. \tag{4.13}$$

Als Alternative dazu lassen sich die Merkmale mittels Normalisierung auf ein bestimmtes Intervalls beschränken. Wird das Intervall zu $[0, 1]$ gewählt, können die normali-

sierten Merkmale $\chi_{i,norm}$ mittels

$$\chi_{i,norm} = \frac{\chi_i - \min(\chi_i)}{\max(\chi_i) - \min(\chi_i)} \tag{4.14}$$

berechnet werden. Die Skalierung der Merkmale ist kein Bestandteil der Vorverarbeitung, sondern ein integraler Teil der Modellwahl [Lemm et al., 2011]. Sowohl Trainings- als auch Validierungsdaten sollten durch die gleiche Transformation skaliert werden. Idealer Weise basiert die Skalierung ausschließlich auf den Eigenschaften der Trainingsdaten. Das bedeutet, dass keine Information der zu klassifizierenden Daten für die Berechnung der Größen $\mathrm{E}(\chi_i)$, $\mathrm{std}(\chi_i)$, $\min(\chi_i)$ oder $\max(\chi_i)$ verwendet wird. Analog sollte auch bei der Selektion der Merkmale vorgegangen werden, sodass diese ausschließlich auf den Trainingsdaten beruht. Dies würde jedoch darin resultieren, dass sich in jeder Iteration der KV eine andere Reihenfolge der hinsichtlich ihres Informationsgehaltes geordneten Merkmale ergibt. Aus diesem Grund wird die Selektion der Merkmale auf dem gesamten Datensatz $\boldsymbol{D}_{MW}$ durchgeführt. Während der KV wird deswegen zwar Information der zu validierenden Daten für das Training genutzt, was zu einer Überanpassung führen kann. Jedoch ist $\boldsymbol{D}_{Test}$ immer noch unabhängig, weshalb die damit ermittelte Genauigkeit der Klassifizierung stets zuverlässig ist.

Die Suche nach dem geeignetsten Modell stellt ein Optimierungsproblem im Raum der Hyperparameter dar, bei dem die Klassifizierungsgenauigkeit der KV als Zielfunktion maximiert wird [Bergstra et al., 2011]. Die Hyperparameter variieren abhängig davon, welcher Machine-Learning-Algorithmus genutzt wird. Die LDA und QDA sind einfache Algorithmen, bei denen keine Optimierung der Hyperparameter nötig ist. Im Gegensatz dazu erfordern die NNK und die SVM eine angemessene Wahl der Anzahl der nächsten Nachbarn k bzw. der Regularisierungskonstante C und der Bandbreite des Kernels γ. Zudem müssen die Merkmale festgelegt werden, die für die Klassifizierung genutzt werden sollen. Die Menge aller möglichen Kombinationen von Merkmalen beträgt 2^m und ist damit immens, weshalb es rechentechnisch nicht möglich ist, alle Kombinationen zu testen. Aus diesem Grund bietet es sich an die Anzahl der ausgewählten Merkmale n_{FS} als zusätzlichen Hyperparameter zu nutzen. Da alle Merkmale bereits hinsichtlich ihres Informationsgehaltes geordnet sind, kann die Selektion der Merkmale effizient mittels eines einzigen Hyperparameters, der lediglich $m - 1$ Werte annehmen kann, durchgeführt werden. Hinsichtlich der Optimierung der Hyperparameter haben sich verschiedene Strategien etabliert. Probleme mit wenigen Hyperparametern werden oft mittels Grid Search optimiert [Bergstra et al., 2012], wohingegen für hochdimensionale Probleme häufig Partikelschwarmalgorithmen [Escalante et al., 2009], genetische Algorithmen [Coroiu, 2016] oder gradientenbasierte Verfahren [Bengio, 2000] genutzt werden. In dieser Arbeit wird ausschließlich die Optimierung mittels

Grid Search durchgeführt. Zum einen werden lediglich Probleme mit drei oder weniger Hyperparameter betrachtet. Darüber hinaus erlaubt Grid Search eine einfache Illustration der Zielfunktion in Abhängigkeit der Hyperparameter.

Die Maximierung der Klassifizierungsgenauigkeit geht einher mit der Minimierung der Summe zweier Fehlerquellen (Verzerrung-Varianz-Dilemma) [Dietterich et al., 1995], siehe Abb. 4.5B. Die Verzerrung stellt einen systematischen Fehler dar, der durch falsche Annahmen des Machine-Learning-Algorithmus hinsichtlich des funktionalen Zusammenhangs zwischen Merkmalen und Klassen verursacht wird. Zusätzlich wird die Genauigkeit durch Störungen der Trainingdaten, Messrauschen und stochastischen Eigenschaften der Machine-Learning-Algorithmen beeinträchtigt, was als Varianz bezeichnet wird. Um das optimale Modell zu identifizieren, muss ein Kompromiss aus Verzerrung und Varianz gefunden werden, siehe Abb. 4.5B. Zu komplexe Modell können leicht an Überanpassung leiden, da sie lediglich Fluktuationen der Trainingsdaten modellieren. Im Gegensatz dazu können zu einfache Modelle durch Unteranpassung beeinträchtigt werden und die Zusammenhänge zwischen Merkmalen und Klassen falsch abbilden.

4.5 Detektion apoptotischer Zellen

Wie bereits erwähnt, bezeichnet Apoptose den programmierten Zelltod, der durch extrinsische oder intrinsische Stimulation hervorgerufen werden kann [Lavrik, 2014]. Apoptose ist gekennzeichnet durch die Aktivierung von Caspasen, Zellschrumpfung, Blasenbildung der Membran sowie der Verkleinerung und Verdichtung des Zellkerns [Henry et al., 2013], siehe Abb. 4.6. Aus diesem Grund wird Apoptose in Experimenten durchgeführt mit bildgebender Flusszytometrie üblicher Weise mittels diverser Fluoreszenzfarbstoffe gemessen, die Veränderungen in der Zellmorphologie verdeutlichen [Pietkiewicz et al., 2015]. Alternativ dazu zeigen Studien, dass morphologische Merkmale einen hohen Informationsgehalt hinsichtlich der Diskriminierung aoptotischer und lebendiger Zellen aufweisen [Schmidt et al., 2015]. Der Vorteil der Klassifizierung basierend auf morphologischen Merkmalen besteht darin, dass keine Fluoreszenzfarbstoffe verwendet werden müssen (markierungsfreie Messung). Dadurch können die nicht genutzten Farbkanäle für andere Zwecke verwendet werden, wie beispielsweise die Messung von Proteinen. Außerdem benötigen Fluoreszenzfarbstoffe zur Detektion lebendiger und toter Zellen eine spezielle experimentelle Handhabung, die oft inkompatibel mit anderen Farbstoffen ist [Specht et al., 2017]. Um diese Nachteile zu umgehen, wird in dieser Arbeit eine systematische Untersuchung zweier Fallstudien unter Verwendung von bildgebender Flusszytometrie durchgeführt. In der ersten Studie werden manuel-

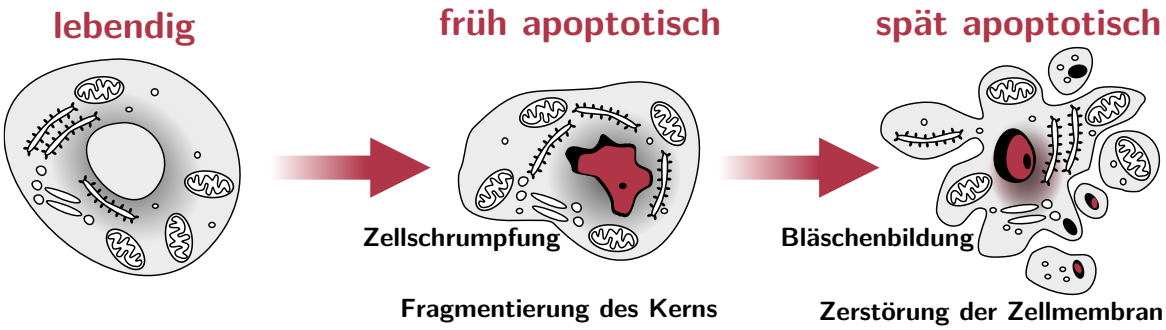

Abbildung 4.6: Veränderungen der Zellmorphologie hervorgerufen durch Apoptose.

le Gating-Methoden, basierend auf der Fluoreszenz zelltodspezifischer Farbstoffe Annexin V und Propidiumiodid (PI), mit automatisierten Machine-Learning-Verfahren verglichen, die ausschließlich markierungsfreie Merkmale extrahiert aus dem Hell- und Dunkelfeldkanal in Betracht ziehen. Dabei stellt sich heraus, dass Machine-Learning-Ansätze und traditionelle Gating-Methoden eine vergleichbare Genauigkeit erzielen. Im Gegensatz dazu werden in der zweiten Studie typische Fehler aufgedeckt, die häufig bei Machine-Learning-Verfahren auftreten und wie diese behoben werden können.

4.5.1 Fallstudie 1

Die traditionelle, zweidimensionale Gating-Strategie basierend auf den Fluoreszenzfarbstoffen Annexin V und PI ist eine etablierte Methode, die präzise und zuverlässige Ergebnisse liefert [Pietkiewicz et al., 2015], jedoch einige Nachteile in der experimentellen Handhabung aufweist. Alternativ dazu beruht der hier vorgestellte Machine-Learning-Ansatz auf markierungsfreien Merkmalen extrahiert aus dem Hell- und Dunkelfeldkanal, die ähnlich wie Farbstoffe morphologische Eigenschaften charakterisieren. Durch Demonstration der Konformität beider Methoden wird gezeigt, dass die Färbung mittels zelltodspezifischer Farbstoffe entbehrlich ist und bedenkenlos durch markierungsfreie Verfahren ersetzt werden kann.

Es wird dazu ein Datensatz betrachtet, der aus zwei Messungen besteht. Die erste Messung beinhaltet unstimulierte, fast ausschließlich lebendige Zellen, wohingegen die zweite Messung 180 min nach der Stimulation mit 250 ng/ml CD95L fast ausschließlich apoptotische Zellen beinhaltet. Der traditionelle Gating-Ansatz anhand der Intensität der Fluoreszenzfarbstoffe Annexin V und PI ist in Abb. 4.7A dargestellt. Dabei werden die Zellen im unteren Linken Gate als lebendig und die restlichen Zellen als apoptotisch klassifiziert. Das Gating-Verfahren kann nun auf unbekannte Daten angewendet werden, wie beispielsweise dem in Abb. 4.7A dargestellten Kinetikexperiment, das den

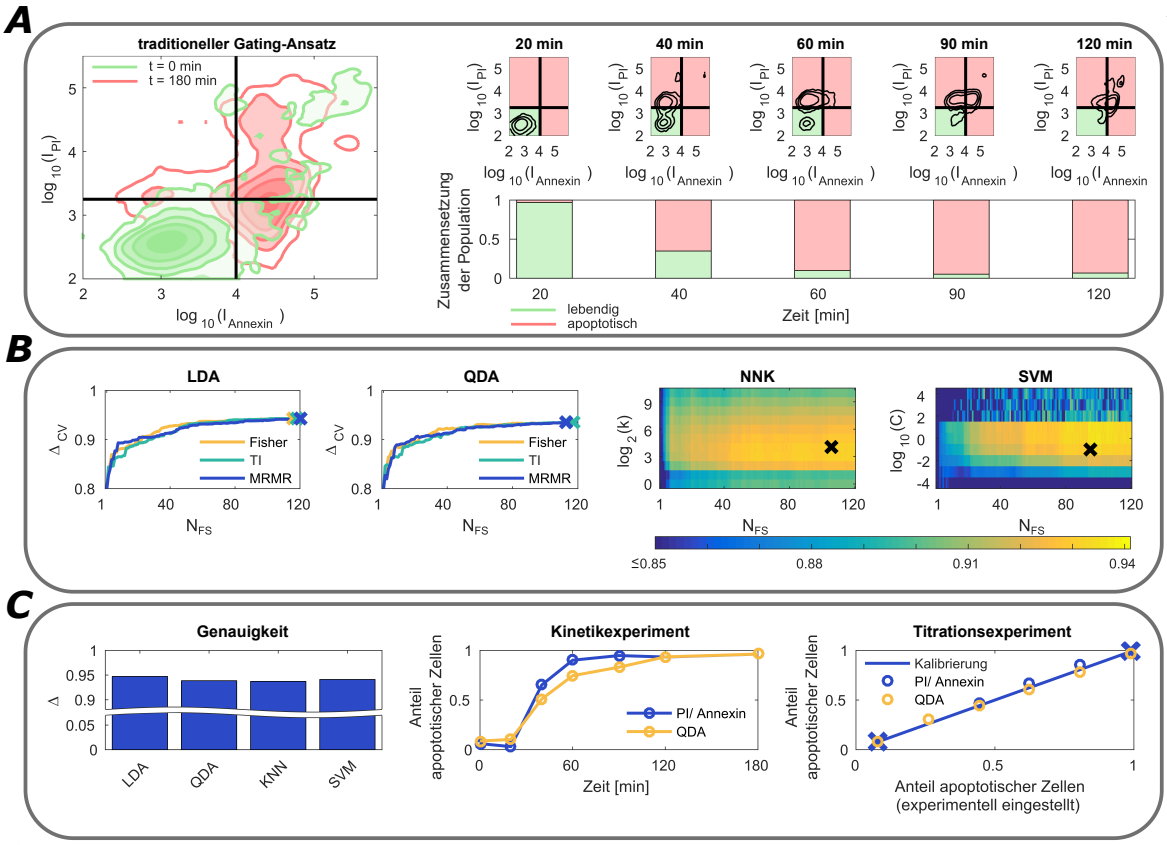

Abbildung 4.7: Vergleich des traditioneller Gating-Ansatzes mit dem markierungsfreien Machine-Learning-Verfahren [Pischel et al., 2018]. **(A)** Traditionelles Gating mittels der Farbstoffe Annexin V und Propidiumiodid (PI) erlaubt es lebendige und apoptotische Zellen zu unterscheiden. **(B)** Die Modellwahl stellt ein Optimierungsproblem im Raum der Hyperparameter dar. Die Zielfunktion ist für verschiedene Machine-Learning-Algorithmen dargestellt. **(C)** Die verwendeten Machine-Learning-Algorithmen zeigen eine hohe Klassifizierungsgenauigkeit gemessen an einem unabhängigen Testdatensatz. Zudem ist ein Vergleich beider Methoden an einem Kinetik- bzw Titrationsexperiment vorgenommen worden. Stellvertretend für alle Machine-Learning-Algorithmen wird lediglich die quadratische Diskriminanzanalyse (QDA) gezeigt.

zeitlichen Verlauf einer sterbenden Zellpopulation zeigt.

Um die Machine-Learning-Algorithmen zu kalibrieren, werden die nach Abb. 4.7A mittels des traditionellen Gating-Ansatzes klassifizierten unstimulierten und 180 min nach der Stimulation gemessenen Zellen als Trainingsdatensatz verwendet. Im Gegensatz zum Gating-Ansatz werden ausschließlich die bisher nicht in Betracht gezogenen markierungsfreien Merkmale für die Klassifizierung mittels Machine-Lerning genutzt. Wie in den vorigen Abschnitten erläutert, wird der Trainingsdatensatz in einen Teil für die Modellwahl $\boldsymbol{D}_{MS}$ (67 % der Daten) und einen Teil zum Testen der Genauigkeit $\boldsymbol{D}_{Test}$ (33 % der Daten) gegliedert. Die Algorithmen zur Selektion der Merk-

male werden auf ganz $\boldsymbol{D}_{MS}$ angewendet. Anschließend werden die optimalen Modelle anhand ihrer über alle Iterationen der KV gemittelten Genauigkeit identifiziert. Als Optimierungsmethode wird Grid Search verwendet, da auf diese Weise die Zielfunktion in Abhängigkeit der Hyperparameter anschaulich dargestellt werden kann, siehe Abb. 4.7B. Die optimalen Modelle werden dabei durch ein Kreuz markiert. Da die LDA und QDA unabhängig von algorithmenspezifischen Hyperparametern sind, wird lediglich die Anzahl der selektierten Merkmale optimiert. Dies ist für alle Methoden zur Selektion der Merkmale in Abb. 4.7B dargestellt. Um die NNK und SVM zu optimieren, müssen zusätzliche Hyperparameter angepasst werden. Diese beinhalten für die NNK die Anzahl der nächsten Nachbarn k und für die SVM mit einem gaußschen Kernel die Regularisierungskonstante C sowie die Kernelbandbreite γ. Für die NNK und SVM sind in Abb. 4.7B die Zielfunktionen lediglich für eine Methode der Selektion der Merkmale dargestellt. Zudem ist für die SVM die Zielfunktion nur abhängig von zwei Hyperparametern illustriert. Die fehlenden Darstellungen zur Modellwahl sind in Abb. A.7-A.8 zu finden. Um die Genauigkeit der optimalen Modelle zu testen, werden diese erneut auf $\boldsymbol{D}_{MS}$ und anschließend auf $\boldsymbol{D}_{Test}$ angewendet. Dabei stellt sich heraus, dass alle Machine-Learning-Algorithmen eine sehr hohe Genauigkeit von etwa 0.95 erzielen, obwohl sie sich stark in ihrer Komplexität unterscheiden, siehe Abb. 4.7C und Tab. A.3.

Analog zum traditionellen Gating-Ansatz können automatisierten Machine-Learning-Algorithmen auf neue, bisher unbekannte Daten angewendet werden. Zur Demonstration werden die optimalen Modelle auf dem kompletten Trainingsdatensatz $\boldsymbol{D}$ trainiert und anschließend genutzt, um ein Kinetik- bzw. Titrationsexperiment zu klassifizieren, siehe Abb. 4.7C. Das Kinetikexperiment stellt den bereits in Abb. 4.7A illustrierten zeitlichen Verlauf einer sterbenden Zellpopulation dar. Im Gegensatz dazu zeigt das Titrationsexperiment experimentell eingestellte Verhältnisse lebendiger und apoptotischer Zellen. Es stellt sich heraus, dass die QDA, stellvertretend für alle Machine-Learning-Algorithmen, vergleichbare Ergebnisse liefert wie der traditionelle Gating-Ansatz. Beim Kinetikexperiment ergibt sich der typische sigmoidale Verlauf und beim Titrationsexperiment liegen die Berechnungen sehr nah an der Kalibrierungskurve, siehe Abb. 4.7C. Die übrigen Machine-Learning-Algorithmen erzielen ebenfalls präzise Ergebnisse, die in Abb. A.9 dokumentiert sind. Damit wird eindeutig demonstriert, dass automatisierte Machine-Learning-Verfahren basierend auf markierungsfreien Merkmalen in der Lage sind, lebendige und apoptotische Zellen akkurat zu unterscheiden.

4.5.2 Fallstudie 2

In der zweiten Fallstudie wird ebenfalls ein Datensatz betrachtet, der aus zwei Messungen besteht. Die erste Messung beinhaltet unstimulierte Zellen, wohingegen die zweite Messung Zellen 180 min nach der Stimulation mit 250 ng/ml CD95L beinhaltet. Zusätzlich zum Hell- und Dunkelfeldkanal sind drei weitere Fluoreszenzfarbstoffe verwendet worden. Dazu zählen der zelltodspezifische Fluoreszenzfarbstoff Zombie Aqua, der DNA-Farbstoff 7AAD und ein Farbstoff sensitiv bezüglich aktiver Caspase-3. Das traditionelle Gating-Verfahren mittels Annexin V und PI kann nicht angewendet werden, da es aufgrund der experimentellen Handhabung nicht kompatibel mit den anderen Fluoreszenzfarbstoffen ist. Deshalb wird vereinfachend angenommen, dass die Messung unstimulierter Zellen ausschließlich lebendige und die Messung stimulierter Zellen ausschließlich apoptotische Zellen beinhalten. Aus Fallstudie 1 ist bekannt, dass dies eine plausible Annahme darstellt, siehe Abb. 4.7B.

Anhand dieses Datensatzes wird demonstriert, welche Fehler und Schwierigkeiten oft bei der automatisierten Klassifizierung mittels Machine-Learning auftreten und wie diese behoben werden können. Dabei wird der Fokus besonders auf den Einfluss des Klassenungleichgewichts, der Identifizierung irreführender Merkmale sowie der Integration biologischen Fachwissens gelegt.

Klassenungleichgewicht

Unter dem Klassenungleichgewicht werden große Differenzen in der Anzahl der Samples der verschiedenen Klassen verstanden. Der hier betrachtete Datensatz ist davon betroffen, da das Verhältnis von apoptotischen zu lebendigen Zellen etwa 1/10 beträgt. Modelle, die auf solchen Daten trainiert werden, begünstigen oft die Klassifizierung der stark vertretenen Klasse, wohingegen die Klasse mit einer geringen Anzahl von Samples unzureichend erkannt wird [He et al., 2009; Lin et al., 2013], wie in Abb. 4.8A dargestellt. In diesem Fall wird ein zweidimensionales Problem mit stark überlappenden Verteilungen lebendiger und apoptotischer Zellen illustriert, die mittels LDA und NNK klassifiziert werden. Für die NNK wird die Anzahl der nächsten Nachbarn k mittels der über alle Iterationen der Kreuzvalidierung gemittelten Genauigkeit optimiert. Es stellt sich heraus, dass die nichtlineare Klassifizierung der NNK im Vergleich zur LDA eine höhere Genauigkeit erzielt, die nicht sehr sensitiv bezüglich des Hyperparameters k ist. Die LDA ist im Gegensatz dazu unabhängig von Hyperparametern und wird deshalb als konstante Linie dargestellt. Werden die Separationsgrenzen im Detail untersucht, stellt sich jedoch heraus, dass die LDA, obwohl sie eine geringere Genauigkeit erzielt, die Klassen qualitativ besser trennt. Die NNK ordnet fast alle

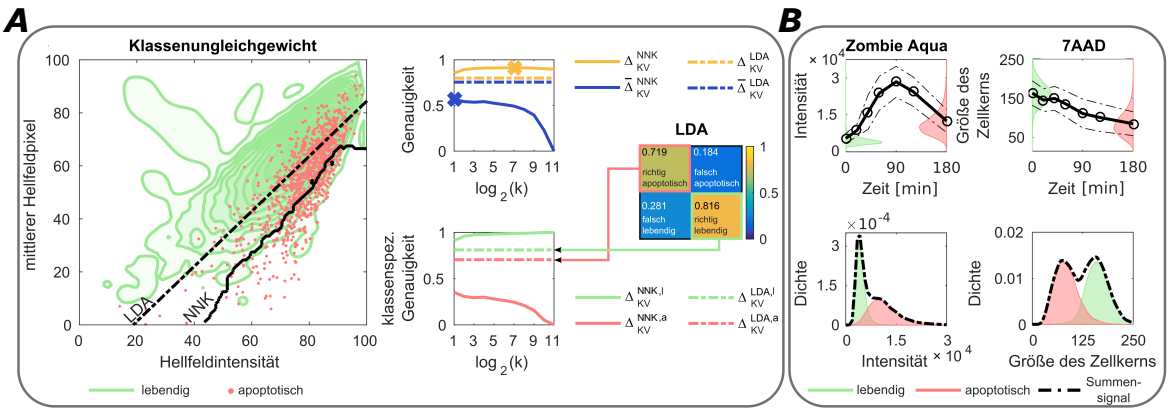

Abbildung 4.8: Auswirkung von Klassenungleichgewicht und irreführender Merkmale auf die Klassifizierung [Pischel et al., 2018]. **(A)** Die Diskriminierung lebendiger (grün) und apoptotischer (rot) Zellen anhand zweier Merkmale ist für die NNK und LDA dargestellt. Obwohl die NNK eine höhere Genauigkeit erzielt, stellt sich heraus, dass die LDA die Klassen qualitativ besser unterscheidet. Dies kann an der klassenspezifischen Genauigkeit und der Wahrheitsmatrix abgelesen werden. Es zeigt sich, dass für Probleme mit Klassenungleichgewicht der geometrische Mittelwert der klassenspezifischen Genauigkeit ein adäquates Maß der Klassifizierungsgüte darstellt. **(B)** Die zeitliche Entwicklung der Intensität des zelltodspezifischen Fluoreszenzfarbstoffes Zombie Aqua zeigt nicht monotones Verhalten. Zudem lässt die Summenverteilung lebendiger und apoptotischer Zellen keine Abschätzung der Zusammensetzung der Zellpopulation zu. Im Gegensatz dazu erlaubt die Größe des Zellkerns gemessen mit 7AAD die Unterscheidung lebendiger und apoptotischer Zellen.

Samples der stark vertretenen Klasse der lebendigen Zellen zu, wohingegen apoptotische Zellen nur unzureichend korrekt identifiziert werden. Dies kann eindeutig anhand der klassenspezifischen Genauigkeit erkannt werden, siehe Abb. 4.8A. Der Anstieg der korrekt erkannten lebendigen Zellen für große k geht einher mit einem Abfall der korrekt erkannten apoptotischen Zellen. Da das Verhältnis beider Phänotypen 1/10 beträgt, ist dieser Effekt jedoch nicht anhand der Genauigkeit erkennbar. Im Gegensatz dazu zeigt die klassenspezifische Genauigkeit der LDA keine großen Unterschiede hinsichtlich lebendiger und apoptotischer Zellen. Beiden Klassen werden annähernd gleich gut erkannt, was auch anhand der Wahrheitsmatrix bestätigt werden kann, siehe Abb. 4.8A.

Ein alternatives Maß zur Messung der Güte der Klassifizierung stellt der geometrische Mittelwert der klassenspezifischen Genauigkeit $\bar{\Delta}$ dar

$$\bar{\Delta} = \sqrt{\Delta_l \Delta_a}. \tag{4.15}$$

Dabei bezeichnen Δ_l und Δ_a die klassenspezifischen Genauigkeiten lebendiger bzw. apoptotischer Zellen. Im Kontrast zur Genauigkeit Δ ist der geometrische Mittelwert der klassenspezifischen Genauigkeit $\bar{\Delta}$ in der Lage die geringe Güte der Klassifizierung mittels NNK für große k zu identifizieren, siehe Abb. 4.8A. Bezüglich der LDA

zeigt das Klassenungleichgewicht keine signifikanten Auswirkungen, da lediglich Mittelwerte und Kovarianzen der Daten bestimmt werden. Δ und $\bar{\Delta}$ liegen deshalb dicht beieinander und liefern vergleichbare Ergebnisse. Damit zeigt sich, dass die Genauigkeit als Maß zur Messung der Klassifizierungsgüte nur geeignet ist, wenn die Samples beider Klassen annähernd gleich stark auftreten. Für Probleme mit Klassenungleichgewicht sollte demnach ein Maß der Güte verwendet werden, das die klassenspezifischen Genauigkeiten adäquat wichtet.

Qualität der Merkmale

Neben der Färbung mittels Annexin V und PI haben sich noch weitere farbstoffbasierte Strategien zur Detektion apoptotischer Zellen etabliert. Dazu zählen die Messung der Größe des Zellkerns [Wen et al., 2017] sowie die Messung der Integrität der Zellmembran [Higuchi et al., 2015]. Zur Prüfung der Eignung beider Methoden bezüglich der Detektion apoptotischer Zellen werden diese auf den zeitlichen Verlauf einer sterbenden Zellpopulation angewendet. Der Fluoreszenzfarbstoff Zombie Aqua kann ausschließlich in Zellen eindringen, wenn deren Zellmembran zerstört wird. Er wird deshalb häufig verwendet, um die Integrität der Zellmembran zu messen und damit lebendige und tote Zellen zu unterscheiden. Hinsichtlich der Dynamik der sterbenden Zellpopulation ist zu erkennen, dass die Intensität zunächst, wie erwartet, ansteigt, siehe Abb. 4.8B. Mit zunehmender Zeit sterben mehr Zellen durch Apoptose, deren zerstörte Zellmembran den Fluoreszenzfarbstoff eindringen lässt. Nach etwa 90 min ist ein starker Abfall der Intensität zu verzeichnen, der jedoch nicht als Wiederauferstehung der apoptotischen Zellen missinterpretiert werden sollte. In diesem Fall ist der Abfall der Intensität höchstwahrscheinlich durch den Ausfluss des Farbstoffes aufgrund der Zerstörung der Zelle verursacht worden. Der Vergleich der Verteilungen der initialen Population (lebendige Zellen) und der Population 180 min nach der Stimuation (apoptotische Zellen) zeigt einen starken Überlapp. In Experimenten wird Zombie Aqua üblicher Weise genutzt, um lebendige und apoptotische Zellen einer Population unbekannter Zusammensetzung zu bestimmen (gestrichelte Kurve). Diese unimodale Verteilung kann nicht in ihre Komponenten (lebendig und apoptotisch) zerlegt werden und beinhaltet deshalb kaum diskriminative Information. Im Gegensatz dazu stellt sich heraus, dass der DNA-Farbstoff 7AAD sehr nützlich hinsichtlich der Messung der Größe des Zellkerns ist, siehe Abb. 4.8B. Es ist bekannt, dass Apoptose einhergeht mit der Verkleinerung und Verdichtung des Zellkerns [Henry et al., 2013; Higuchi et al., 2015], was an einem abfallenden Verlauf der Größe des Zellkerns beobachtet werden kann. Zudem zeigen die Verteilungen lebendiger und apoptotischer Zellen nur einen kleinen Überlapp. Dies resultiert in einer bimodalen Summenverteilung, anhand der

lebendige und apoptotische Zellen unterschieden werden können.

Fluoreszenzfarbstoffe werden in der Regel genutzt, um die Interpretation biologischer Experimente zu vereinfachen. Obwohl belegt ist, dass gewisse Fluoreszenzprodukte zur spektralen Klassifizierung gute Ergebnisse erzielen, können sie in einer speziellen Anwendung versagen. Um diesen Fällen vorzubeugen, wird empfohlen die Eignung farbstoffbasierter Merkmale stets zu prüfen. In diesem Fall stellt sich heraus, dass der zelltodspezifische Farbstoff Zombie Aqua biologisch unplausible Resultate hinsichtlich der Dynamik der sterbenden Zellpopulation aufzeigt. Um eine Störung der automatisierten Machine-Learning-Methoden zu vermeiden, wird empfohlen diese irreführenden Merkmale auszuschließen.

Integration biologischen Wissens

Machine-Learning-Algorithmen können als Blackbox-Modelle verstanden werden, die abhängig vom empfangenen Eingang (Realisierung bestimmter Merkmale) ein Ausgangssignal (zugeordnete Klasse) ausgeben. Die Zuordnung basiert dabei ausschließlich auf den Zusammenhängen zwischen Merkmalen und Klassen, die anhand des Trainingsdatensatzes gelernt worden sind. Neben den Informationen integriert in den Trainingsdaten ist oft zusätzliche Expertise bzw. Fachwissen vorhanden, das für eine akkurate Klassifizierung berücksichtigt werden sollte [Guyon et al., 2003].

Es ist bekannt, dass der in dieser Arbeit betrachtete Prozess der Apoptose durch eine starke Aktivierung von Caspase-3 verursacht wird [Lavrik, 2014]. Aktive Caspase-3 verursacht die Spaltung verschiedener downstream gelegener Proteine, die über unterschiedliche Pfadwege zur Apoptose führen. Wie erwartet, zeigt die Selektion der Merkmale eine starke Korrelation von Caspase-3 und Apoptose. Dies ist erkennbar an der in Abb. 4.9A dargestellten Ordnung der Merkmale mittels Transinformation hinsichtlich ihres Informationsgehaltes[1]. Etwa die vierzig informativsten Merkmale basieren auf Caspase-3, was den immensen Informationsgehalt verdeutlicht. Für die übrigen Merkmale basierend auf 7AAD sowie dem Hell- und Dunkelfeld ist ein derartiges Blockverhalten nicht erkennbar. Zudem lassen sich die Verteilungen der Fluoreszenzintensität von Caspase-3 lebendiger und apoptotischer Zellen gut separieren, siehe Abb. 4.9A.

Um den Einfluss der Merkmale basierend auf Caspase-3 auf die Klassifizierung zu untersuchen, wird die Modellwahl auf allen Merkmalen inklusive Caspase-3 und exklusive Caspase-3 vorgenommen. Dies ist anhand der QDA stellvertretend für alle Machine-Learning-Algorithmen in Abb. 4.9B dargestellt. Dabei stellt sich heraus,

[1] Der Vollständigkeit halber sind die Ergebnisse der übrigen Methoden zur Selektion der Merkmale in Abb. A.10 aufgeführt.

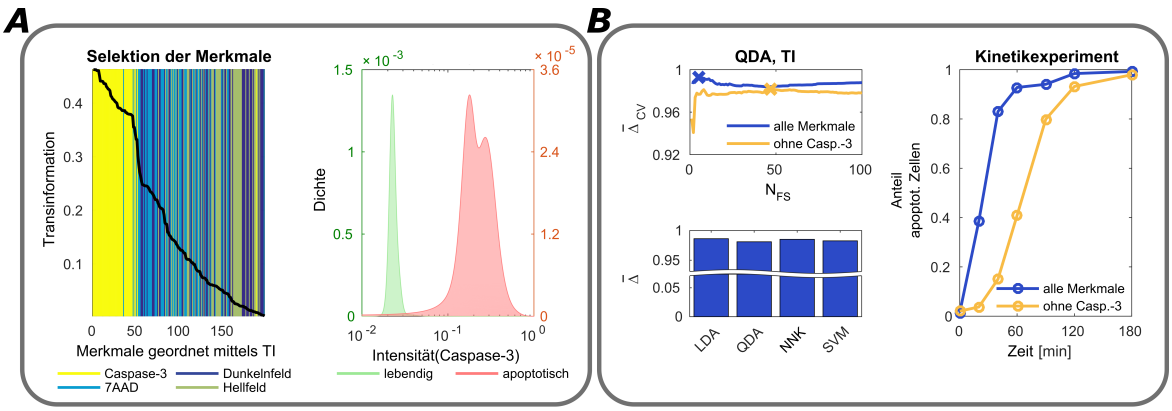

Abbildung 4.9: Identifizierung von Caspase-3 als ungeeignetes Merkmal zur Charakterisierung des dynamischen Prozesses der Apoptose [Pischel et al., 2018]. **(A)** Merkmale basierend auf Caspase-3 zeigen ein hohes Potenzial zur Unterscheidung lebendiger und apoptotischer Zellen. Dies zeigt sich anhand der Anordnung der Merkmale hinsichtlich ihrer Transinformation und der Intensitätsverteilung lebendiger und apoptotischer Zellen. **(B)** Die Modellwahl basierend auf allen Merkmalen inklusive Caspase-3 (gelb) bzw. exklusive Caspase-3 (blau) zeigt keine signifikanten Unterschiede hinsichtlich der Güte der Klassifizierung. Dies ist exemplarisch anhand der quadratischen Diskriminanzanalyse (QDA) stellvertretend für alle verwendeten Machine-Learning-Algorithmen gezeigt. Es stellt sich jedoch heraus, dass beide Ansätze qualitativ unterschiedliche Dynamiken eines Kinetikexperimentes ergeben. Ausschließlich der sigmoidale Verlauf aller Merkmale exklusive Capase-3 ist biologisch plausibel.

dass keine signifikanten Unterschiede hinsichtlich der Güte der Klassifizierung festzustellen sind. Bezüglich der Anzahl der verwendeten Merkmale zeigt sich jedoch, dass der Klassifizierungsansatz exklusive Caspase-3 deutlich mehr Merkmale benötigt. Gemessen am unabhängigen Testdatensatz lässt sich erkennen, dass alle Machine-Learning-Algorithmen eine präzise Klassifizierung basierend auf allen Merkmalen exklusive Caspase-3 ermöglichen, siehe Abb. 4.9B und Tab. A.4.

Zur Verifizierung der Eignung der Merkmale basierend auf Caspase-3 werden die beiden optimalen Modelle der QDA auf ein Kinetikexperiment angewendet, siehe Abb. 4.9B. Aus Fallstudie 1 ist bereits bekannt, dass die Dynamik durch einen sigmoidalen Verlauf gekennzeichnet ist, der korrekt durch das Modell exklusive Caspase-3 wiedergegeben wird. Im Gegensatz dazu beschreibt das Modell inklusive Caspase-3 die Dynamik mittels eines hyperbolischen Verlaufs. Besonders im Übergangsbereich von der lebendigen zur apoptotischen Population (20-90 min) zeigen sich große Abweichungen. Ausschließlich der initiale und der Endzeitpunkt anhand derer das Modell trainiert worden ist, stimmen mit dem erwarteten Verlauf überein. Damit wird demonstriert, dass Merkmale basierend auf Caspase-3 lediglich geeignet sind, um stationäre Populationen hinsichtlich lebendiger und apoptotischer Zellen zu diskriminieren. Der dynamische Prozess der Apoptose kann jedoch nicht korrekt beschrieben werden, weshalb diese Merkmale

mit Vorsicht zu verwenden sind.

Es ist bekannt, dass morphologische Veränderungen apoptotischer Zellen zeitlich verzögert zur Aktivierung von Caspase-3 auftreten [Schmidt et al., 2015], da es einen kausalen Zusammenhang zwischen ihnen gibt. Erst wird Caspase-3 aktiviert und anschließend werden dadurch nachgeschaltete Prozesse ausgelöst, die zur Apoptose führen [Lavrik, 2014]. Aus diesem Grund hätten Merkmale basierend auf Caspase-3 gleich zu Beginn der Analyse als ungeeignet identifiziert werden können. Dies wäre allein durch die rechnergestützte Analyse mittels Machine-Learning nicht möglich gewesen.

4.5.3 Zusammenfassung

Dieses Kapitel beschäftigt sich mit dem Problem der automatisierten Klassifizierung und Analyse hochdimensionaler Daten, gemessen mit bildgebender Flusszytometrie. Dazu wird ein kurzer Überblick hinsichtlich etablierter Machine-Learning-Ansätze gegeben. Im Besonderen wird dabei auf die Selektion der Merkmale, die Modellwahl und Machine-Learning-Algorithmen eingegangen. Als Anwendungsbeispiel dient die Analyse des dynamischen Prozesses der Apoptose, der anhand zweier Fallstudien untersucht worden ist. In der ersten Fallstudie ist die automatisierte Klassifizierung basierend auf markierungsfreien Merkmalen mit manuellen Gating-Methoden basierend auf Fluoreszenzfarbstoffen vergleichen worden. Es stellt sich heraus, dass beide Methoden qualitativ übereinstimmende Ergebnisse liefern. Dies erlaubt es in Zukunft auf zelltodspezifische Farbstoffe zu verzichten, wodurch sich eine Reihe experimenteller Vorteile ergeben. In der zweiten Fallstudie wird der Fokus auf die korrekte Anwendung der Machine-Learning-Ansätze gelegt. Die Automatisierung komplexer Prozesse, wie die Detektion apoptotischer Zellen, ist keine triviale Aufgabe und potenzielle Hürden werden oft nicht erkannt bzw. beachtet. Um dies zu adressieren, werden die Identifizierung irreführender Merkmale und die Integration biologischen Wissens betrachtet. Dabei zeigt sich, dass der zelltodspezifische Farbstoff Zombie Aqua und ein Fluoreszenzfarbstoff sensitiv für Caspase-3 ungeeignet hinsichtlich der Charakterisierung des dynamischen Prozesses der Apoptose sind. Zudem ist der Einfluss des Klassenungleichgewichts auf die Güte der Klassifizierung unterucht worden. Es stellt sich heraus, dass Metriken, die die klassenspezifische Güte nicht adäquat wichten, oft die stark vertretende Klasse bevorzugen.

In dieser Arbeit sind verschiedene Verfahren zur Selektion der Merkmale und zur Klassifizierung verwendet worden, die stark in ihrem Rechenaufwand variieren. Der Vergleich einfacher und komplexer Methoden lässt keine signifikanten Unterschiede hinsichtlich der Klassifizierungsgüte erkennen. Unabhängig davon, welche Methoden

zur Selektion der Merkmale bzw. welcher Machine-Learning-Algorithmen verwendet werden, wird stets eine hohe Klassifizierungsgüte erzielt. Hinsichtlich der Modellwahl zeigt sich, dass komplexe Verfahren deutlich zeitaufwendiger sind, da neben der Anzahl der verwendeten Merkmale zusätzlich algorithmenspezifische Hyperparameter optimiert werden müssen. Aus diesem Grund wird empfohlen, besonders im Kontext von Big Data die Analyse stets mit einfachen Methoden zu beginnen. Sollten diese versagen, kann mittels komplexerer und zeitaufwendigerer Verfahren versucht werden, bisher nicht entdeckte Muster in den Daten zu identifizieren.

Einzellzellmessungen, wie die bildgebende Flusszytometrie, erlauben es riesige Mengen an Daten aufzunehmen, wobei jede Zelle durch diverse Eigenschaften charakterisiert wird. Zur Analyse und Modellierung dieser Daten sind in jüngster Vergangenheit verschiedene Machine-Learning-Ansätze vorgeschlagen worden, die sich hinsichtlich der Dimensionsreduktion, Modellwahl und Klassifizierung unterscheiden [Blasi et al., 2016; Chen et al., 2016; Hennig et al., 2017]. Die hier vorgestellte Vorgehensweise grenzt sich von diesen Ansätzen durch eine effiziente Selektion der Merkmale ab, bei der die Anzahl der verwendeten Merkmale einen zu optimierenden Hyperparameter darstellt. Damit zeichnet sich dieser Ansatz durch seine einfache Handhabung und breite Anwendbarkeit aus. Zudem besteht keine Einschränkung des Ansatzes auf Apoptose. Demmnach kann er auch auf andere Prozesse angewendet werden, die mit morphologischen Veränderungen einhergehen.

5 Zusammenfassung und Ausblick

Die Biologie, die Wissenschaft lebendiger Systeme, zeichnet sich durch ihre Komplexität, oft nur qualitatives Verständnis sowie ausgeprägte Variabilität und Heterogenität aus. Um das Verhalten von Organismen als Ganzes zu begreifen, ist es unabdingbar experimentelle Methoden mit Computersimulationen zu kombinieren. In dieser Arbeit werden dazu methodische Lösungsansätze zur Analyse stochastischer biochemischer Reaktionssysteme im Anwendungsgebiet der Systembiologie untersucht. Die Lösungsansätze beinhalten einen neuen **Algorithmus zur simultanen Simulation extrinsischer und intrinsischer Störungen** sowie eine effiziente **Machine-Learning-Strategie zur automatisierten Analyse von Hochdurchsatzdaten**, gemessen mit bildgebender Flusszytometrie. Beide Verfahren werden genutzt, um ein komplexes Apoptosenetzwerk zu analysieren. Dabei ist ein möglicher **zellulärer Mechanismus** identifiziert worden, der die zelluläre Entscheidung zwischen Leben und Tod bestimmt. Ähnliche Ansätze, die Einzellzellexperimente, Machine-Learning und stochastische Modellierung kombinieren, sind bereits in renommierten Zeitschriften veröffentlicht worden [Roux et al., 2015] und zeigen damit, dass dieses Vorgehen momentan State-of-the-Art ist.

Die Simulation biochemischer Reakionssysteme, die durch unterschiedliche Störungen beeinflusst werden, ist in der Literatur bisher nur selten behandelt worden. In der Regel wird vereinfacht angenommen, dass gewisse Anteile der Störungen zu vernachlässigen sind, wodurch die Komplexität des Systems stark gemindert wird. In dieser Arbeit wird diese Vereinfachung nicht getroffen, sondern stattdessen eine Kombination der Sigma-Punkt-Methode und eines approximativen Gillespie-Algorithmus zur Simulation extrinsischer und intrinsischer Störungen verwendet. Damit wird ein wichtiger wissenschaftlicher Beitrag geleistet, der die realitätsnahe Modellierung biologischer Variabilität erlaubt. Die Vorteile der Sigma-Punkt-Methode gegenüber anderen Verfahren zu approximativen Berechnung extrinsischer Störungen werden in dieser Arbeit an einfachen Modellsystemen dargestellt. Dies deutet zwar auf einen sehr vorteilhaften Kompromiss von Genauigkeit und Rechenaufwand hin, jedoch müsste eine weitaus größere Menge an Modellen untersucht werden, um diese Aussage abzusichern. Auch in der Literatur sind ähnliche Gegenüberstellungen getätigt worden, deren ver-

gleichende Analysen auf einer unzureichenden Anzahl verschiedener Modelle basieren [Wu et al., 2006; Nimmegeers et al., 2016; Maußner et al., 2018; Akkermans et al., 2018]. Zudem hängen die vorgestellten Methoden von verschiedenen Hyperparametern ab, deren Wahl oft nicht eindeutig ist. Analog verhält es sich beim Vergleich approximativer Methoden zur Simulation intrinsischer Störungen [Marchetti et al., 2017]. Momentan fehlen vergleichende Studien, die verschiedene Methoden zur approximativen Simulation extrinsischer bzw. intrinsischer Störungen unter Verwendung einer ausreichenden Anzahl von Modellen gegenüberstellen. In den Ingenieurwissenschaften wird diese Arbeit wohl nie durchzuführen sein, da selten Code als Zusatzmaterial in wissenschaftlichen Zeitschriftenartikeln bereitgestellt wird. Man müsste deshalb jede Publikation einzeln durchsuchen und die Modellgleichungen übernehmen, was sehr zeitaufwändig und fehleranfällig ist. In der Systembiologie hat sich im Gegensatz dazu die Systems Biology Markup Language (SBML) etabliert, die es erlaubt, Modelle schnell und unkompliziert in Computerprogramme zu laden und zu simulieren [Hucka et al., 2003]. Zudem ist es in der Systembiologie üblich, mathematische Modelle in SBML-Format den Publikationen anzuhängen oder sie in Datenbanken wie BioModels [Le Novère et al., 2006], NetPath [Kandasamy et al., 2010] oder Reactome [Joshi-Tope et al., 2005] einzupflegen[1]. Diese Modelle können dann, wie in [Kazeroonian et al., 2017] demonstriert, genutzt werden, um einen aussagekräftigen Vergleich verschiedener Methoden durchzuführen. Das bedeutet für diese Arbeit, dass womöglich nicht die optimalen Methoden hinsichtlich des Kompromisses zwischen Rechenaufwand und Genauigkeit für die simultane Simulation extrinsischer und intrinsischer Störungen kombiniert worden sind. Diese Aussage ändert jedoch nichts an der Nützlichkeit der entwickelten Methode, sondern sie zeigt lediglich auf, dass es Potenzial gibt die Methode noch weiter zu verbessern. Des Weiteren berücksichtigt die entwickelte Methode nur extrinsische bzw. externe Störungen in Form unsicherer Parameter. Die Realisierungen dieser Parameter variieren zwischen einzelnen Zellen, bleiben jedoch zeitlich konstant. Komplementär dazu integrieren verschiedene Methoden zeitabhängige Störungen [Shahrezaei et al., 2008; Hilfinger et al., 2011; Zechner et al., 2014; Thanh et al., 2015; Voliotis et al., 2016]. Diese können beispielsweise als schwankende Wettereinflüsse oder stochastische Zelltrajektorien in einem inhomogenen Bioreaktor interpretiert werden. Durch Kombination der Modellierung extrinsischer bzw. externer Störungen als unsichere Parameter und stochastischen Prozess zusätzlich zu den intrinsischen Störungen könnten noch detailliertere Modelle aufgestellt werden. Dies ist jedoch nicht Gegenstand dieser Arbeit, sondern eine Aufgabe für zukünftige Forschungsvorhaben.

[1] Für die im Rahmen der Promotion verfassten Publikationen [Pischel et al., 2017, 2018; Buchbinder et al., 2018] sind die verwendeten mathematischen Modelle im SBML-Format beigefügt bzw. der Code bereitgestellt, um die Simulationen nachzuvollziehen.

Hinsichtlich der Modellierung des dynamischen Prozesses der Apoptose lässt sich sagen, dass die unbekannten Parameter effizient mit der entwickelten Methode zur Simulation extrinsischer und intrinsischer Störungen identifiziert werden können. Das Modell beinhaltet 40 chemische Reaktionen sowie 29 chemische Spezies und stellt damit eines der komplexesten stochastischen Modelle dar, die bisher kalibriert und analysiert worden sind. Darüber hinaus war die Analyse pro- und antiapoptotischer CD95-Pfadwege auf der Ebene einzelner Zellen bisher nicht existent. Somit ist ein wichtiger Schritt zum Verständnis des Gesamtprozesses der Apoptose vollzogen, der sich als hilfreich beim Design und der Anwendung medikamentöser Krebstherapien erweisen kann. Die aus den Analysen abgeleiteten Größen TOS und TOD sind plausibel eingeführt worden und deren Verhältnis stellt einen informativen Parameter dar, der Aufschluss über die zelluläre Entscheidung zwischen Leben und Tod gibt. Prinzipiell wären auch die Definition anderer Größen möglich gewesen, wie dem zeitlichen Integral der Abundanzen von Caspase-3 und nuklearem NF-κB. Diese Größen sind jedoch nicht analysiert worden, da das einfache TOS/TOD-Modell bereits hervorragende Resultate geliefert hat. Zudem sind Größen wie der Point of no Return, auf denen TOS und TOD aufbauen, bereits etablierte Begriffe in der Systembiologie [Spencer et al., 2011].

Die automatisierte Analyse von Einzelzellexperimenten mittels Machine-Learning ist seit kurzem ein populäres Forschungsgebiet, das unabdingbar zur Auswertung moderner Hochdurchsatzmessungen (Big Data) ist. Von besonderem Interesse ist die Verwendung markierungsfreier Merkmale zur Unterscheidung von Zellen mit unterschiedlichen Phänotypen [Blasi et al., 2016; Hennig et al., 2017]. In dieser Arbeit wird dieses Thema aufgegriffen und am Beispiel der Diskriminierung lebendiger und apoptotischer Zellen veranschaulicht. Dazu werden in Experimenten Bilder mit bildgebender Flusszytometrie aufgenommen, aus denen diverse Merkmale extrahiert werden. Im Gegensatz zu den häufig verwendeten Wrapper-Ansätzen [Blasi et al., 2016] erlaubt die Integration der Anzahl der Merkmale mittels Filterung als zusätzlichen Hyperparameter eine effiziente Modellwahl. Es stellt sich heraus, dass die optimalen Modelle basierend auf markierungsfreien Merkmalen vergleichbare Ergebnisse liefern wie traditionelle, farbstoffbasierte Gating-Verfahren. Zudem sind verschiedene Nachteile farbstoffbasierter Verfahren identifiziert worden. Alternative Machine-Learning-Ansätze, wie neuronale Netze, die in dieser Arbeit nicht thematisiert worden sind, sind ebenfalls weit verbreitet. Der Vorteil dieser Verfahren ist, dass keine Merkmale aus den Bildern einzelner Zellen extrahiert werden müssen. Neuronale Netze verwenden die Bilder als Pixelmatrizen und generieren eigenständig Merkmale daraus, die zur Klassifizierung herangezogen werden [Angermueller et al., 2016]. Für die Kalibrierung neuronaler Netze werden jedoch enorme Mengen an Daten benötigt, die oft nicht zur Verfügung stehen

[Cho et al., 2015; Sun et al., 2017]. Die klassischen Machine-Learning-Algorithmen sind deshalb nicht obsolet, sondern stellen universell anwendbare Verfahren dar.

Zusammenfassend stellt sich heraus, dass die Analyse stochastischer biochemisher Reaktionsnetzwerke keine triviale Aufgabe ist. Zum einen werden die ohnehin schon komplexen mathematischen Modelle durch stochastische Simulationen weiter verkompliziert. Darüber hinaus sind automatisierte Machine-Learning-Algorithmen vonnöten, um die enormen Datenmengen moderner Hochdurchsatzmessungen zu verarbeiten. Innerhalb der letzten Jahre haben sich verschiedene Tools entwickelt, um diese Hürden in Angriff zu nehmen und eine breite Anwendbarkeit zu ermöglichen. Zudem werden zunehmend Code und experimentelle Daten als Zusatzmaterial von Zeitschriftenartikeln zur Verfügung gestellt [Gewin, 2016]. Wird dieses Vorgehen auch in Zukunft beibehalten, so wird die Analyse stochastischer biochemisher Reaktionsnetzwerke bald zur Routine. Ähnlich wie gewöhnliche Differenzialgleichungen zur Simulation deterministischer Systeme oder Gating-Verfahren zur Auswertung von Hochdurchsatzdaten werden die in dieser Arbeit thematisierten oder alternative Ansätze zu Standardverfahren zur Analyse stochastischer biochemischer Systeme.

A Appendix

A.1 Approximative Algorithmen zur Simulation intrinsischer Störungen

Algorithmus A.1: Hybrider Gillespie-Algorithmus [Haseltine et al., 2002]

Ergebnis: Berechnung zufälliger Systemtrajektorien

Initialisierung: $t \leftarrow t_0$, $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$, Einteilung schneller s und langsamer l Reaktionen;

while $t < t_{end}$ **do**

  Berechnung der schnellen Propensitäten $\boldsymbol{a}^s(\boldsymbol{x})$, der langsamen Propensitäten $\boldsymbol{a}^l(\boldsymbol{x})$ und deren Summe $a_0^l(\boldsymbol{x})$;

  Berechnung zweier gleichverteilter Zufallszahlen $r_{1,2}$ aus dem Intervall $[0, 1]$;

  Berechnung Zeitintervalls bis zur nächsten langsamen Reaktion und der Zustandsänderung durch schnelle Reaktionen $\delta\boldsymbol{x}$ durch Integration ihrer Propensitäten bis $\int_t^{t+\tau^*} a_0^l(\boldsymbol{x})\,\mathrm{d}t + \log(r_1) = 0$ erfüllt ist;

  if $t + \tau^* > t_{end}$ **then**

    Aktualisierung der Zeit $t \leftarrow t_{end}$ und des Zustandes $\boldsymbol{x} \leftarrow \boldsymbol{x} + \delta\boldsymbol{x}(t_{end})$;

  else

    Aktualisierung der Zeit $t \leftarrow t + \tau^*$ und des Zustandes $\boldsymbol{x} \leftarrow \boldsymbol{x} + \delta\boldsymbol{x}$;

    Berechnung des Index der nächsten langsamen Reaktion mittels $\sum_{j=1}^{J-1} a_j(\boldsymbol{x}) < r_2 a_0^l(\boldsymbol{x}) \leq \sum_{j=1}^{J} a_j(\boldsymbol{x})$;

    Aktualisierung des Zustandes $\boldsymbol{x} \leftarrow \boldsymbol{x} + \boldsymbol{N}_J$;

  end

end

Algorithmus A.2: τ-Leaping-Algorithmus [Cao et al., 2006]

Ergebnis: Berechnung zufälliger Systemtrajektorien

Initialisierung: $t \leftarrow t_0$, $\boldsymbol{x} \leftarrow \boldsymbol{x}_0$, Festlegung der Parameter ϵ und n_{krit};

while $t < t_{end}$ **do**

- Berechnung der Propensitäten $\boldsymbol{a}(\boldsymbol{x})$;
- Berechnung der Beschränkungsfunktion $\boldsymbol{g}(\boldsymbol{x})$;
- Identifizierung kritischer Reaktionen $\boldsymbol{I}_{krit}$;
- Berechnung des Mittelwertes der Änderungsrate der Propensitäten:
  $\boldsymbol{\mu}_i(\boldsymbol{x}) \leftarrow \sum_{j \notin \boldsymbol{I}_{krit}} (\nu_{ij}^p - \nu_{ij}^e) a_j(\boldsymbol{x}),\ \forall\, i = 1...n_S$;
- Berechnung der Varianz der Änderungsrate der Propensitäten:
  $\boldsymbol{\sigma}_i^2(\boldsymbol{x}) \leftarrow \sum_{j \notin \boldsymbol{I}_{krit}} (\nu_{ij}^p - \nu_{ij}^e)^2 a_j(\boldsymbol{x}),\ \forall\, i = 1...n_s$;
- Berechnung des Zeitintervalls: $\tau \leftarrow \min\limits_{\substack{j \notin \boldsymbol{I}_{krit} \\ i=1...n_S}} \left\{ \dfrac{\max(\epsilon x_i / g_i, 1)}{|\mu_i(\boldsymbol{x})|}, \dfrac{\max(\epsilon x_i / g_i, 1)^2}{|\sigma_i^2(\boldsymbol{x})|} \right\}$;
- Berechnung des Zeitintervalls τ_{krit} bis zur nächsten Reaktion;
- Aktualisierung des Zeitintervalls $\tau \leftarrow \min(\tau, \tau_{krit}, t_{end} - t)$;
- **if** $\tau < \frac{10}{a_0(\boldsymbol{x})}$ **then**
  - Berechnung der Zustandsänderung $\Delta\boldsymbol{x}$ und des Zeitintervalls τ mittels des Gillespie-Algorithmus für 100 Reaktionen
- **else**
  - Berechnung der Zustandsänderung $\Delta\boldsymbol{x}$ durch kritische und unkritische Reaktionen;
  - **while** *any*$(\boldsymbol{x} + \Delta\boldsymbol{x} < 0)$ **do**
    - Aktualisierung des Zeitintervalls $\tau \leftarrow \tau/2$;
    - Berechnung der Zustandsänderung $\Delta\boldsymbol{x}$ durch kritische und unkritische Reaktionen;
  - **end**
- **end**
- Aktualisierung der Zeit $t \leftarrow t + \tau$ und des Zustandes $\boldsymbol{x} \leftarrow \boldsymbol{x} + \Delta\boldsymbol{x}$;

end

A.2 Benchmarking des Sigma-Punkt-Ansatzes

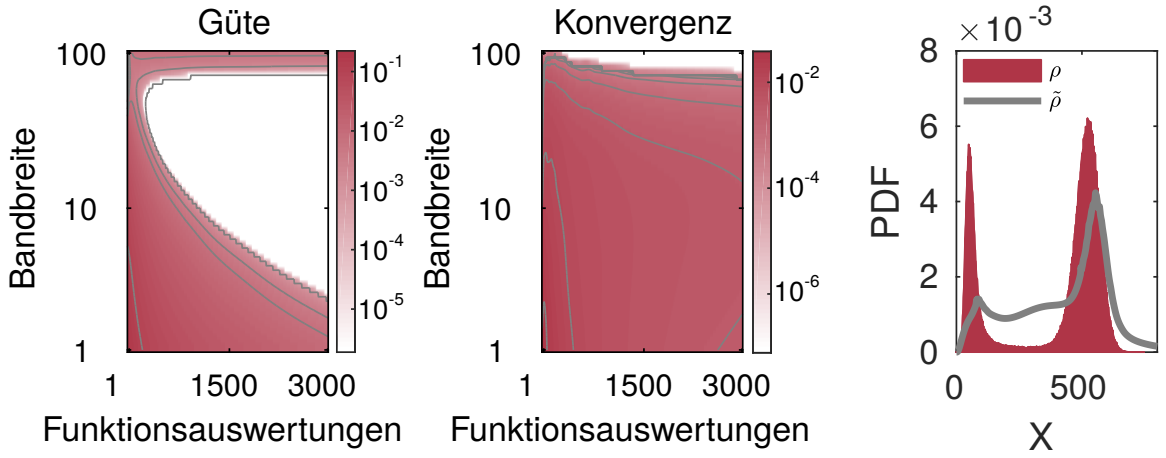

Abbildung A.1: Benchmark der Sigma-Punkt-Methode kombiniert mit dem τ-Leaping-Alorithmus am Beispiel des Schlögl-Modells [Pischel et al., 2017].

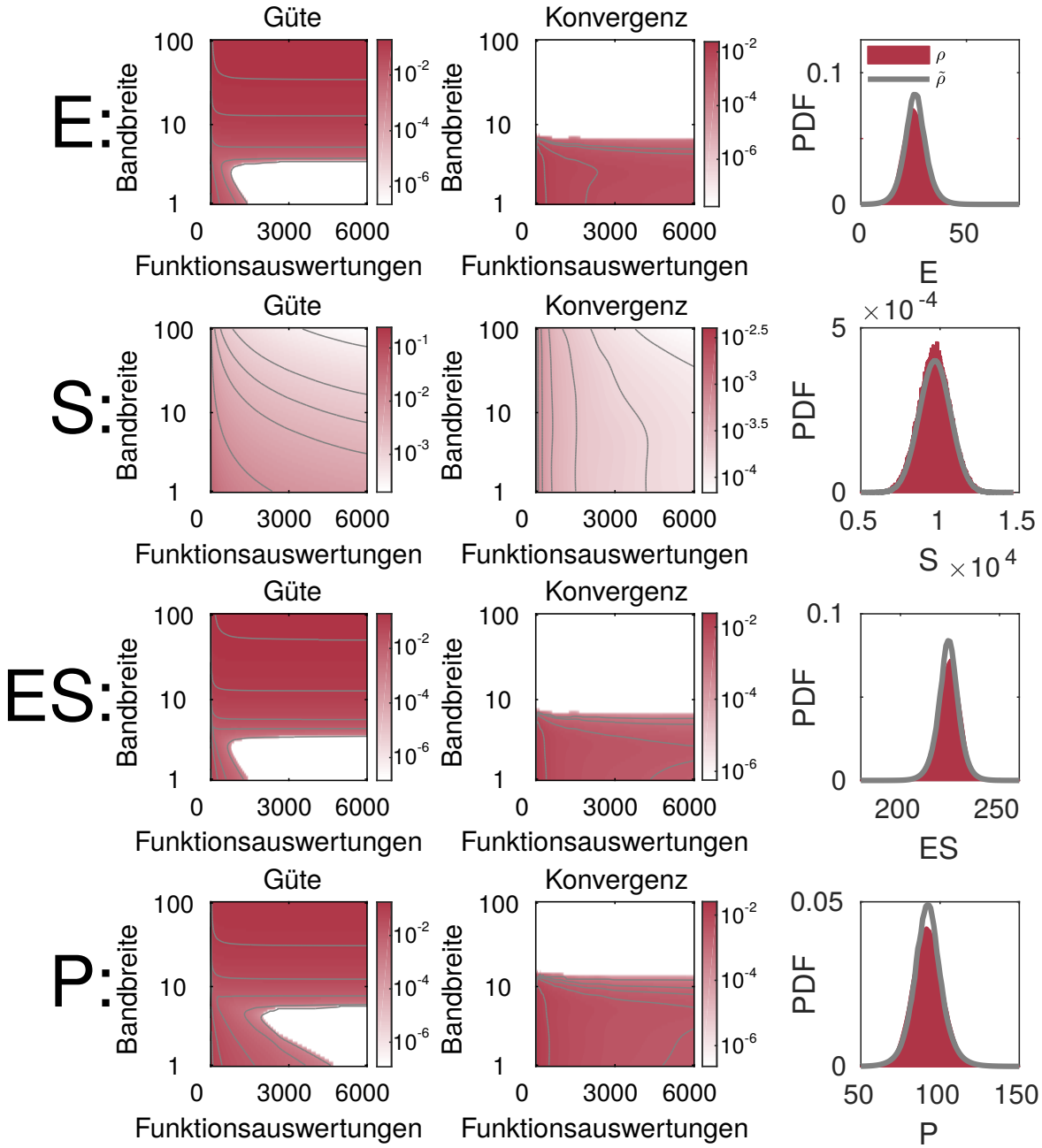

Abbildung A.2: Benchmark der Sigma-Punkt-Methode kombiniert mit dem τ-Leaping-Alorithmus am Beispiel der Michaelis-Menten-Kinetik [Pischel et al., 2017].

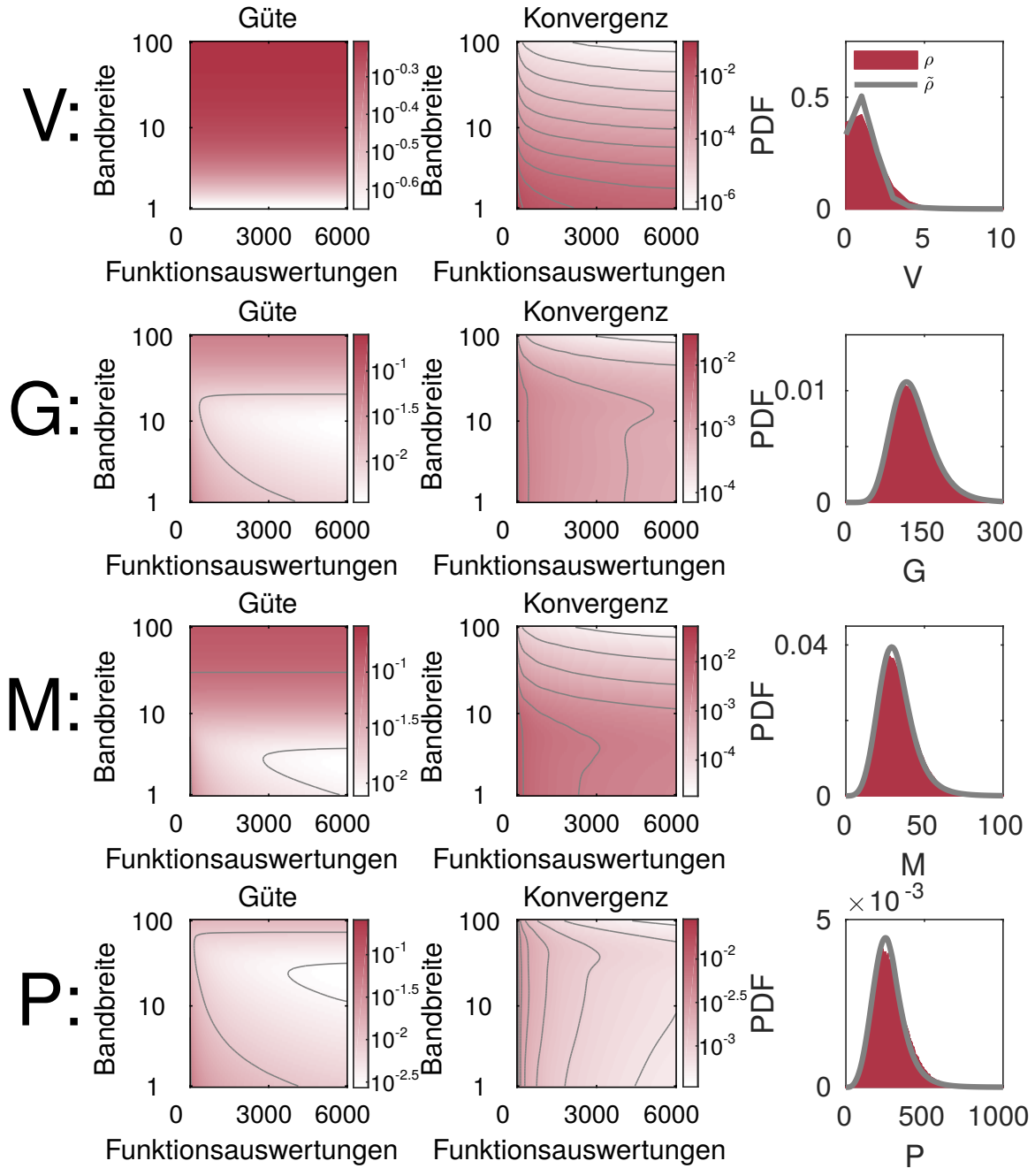

Abbildung A.3: Benchmark der Sigma-Punkt-Methode kombiniert mit dem τ-Leaping-Alorithmus am Beispiel eines Virus-Modells [Pischel et al., 2017].

A.3 Apoptose-Modell

Tabelle A.1: Auflistung chemischer Spezies [Buchbinder et al., 2018].

Spezies	Abkürzung	Anfangsbedingung	Ref.
c-FLIP$_{\mathrm{RS}}$	$cFLIP_L$	$f_{RS}(Dosis)$	—
c-FLIP$_{\mathrm{L}}$	$cFLIP_{RS}$	$f_L(Dosis)$	—
caspase-8	$Casp8$	$f_{Casp8}(Dosis)$	—
caspase-8 p43/p41	$p43hom$	0	[Fricker et al., 2010]
p43-FLIP-p43/p41	$p43het$	0	[Fricker et al., 2010]
caspase-8 p10p18	$p10$	0	[Fricker et al., 2010]
p43-FLIP	$p43FLIP$	0	[Neumann et al., 2010]
caspase-8 hetero dimer	$Casp8het$	0	[Fricker et al., 2010]
caspase-3	A	1 (fixed)	—
active caspase-3	$casp3$	0	[Fricker et al., 2010]
neutral IKKK	$IKKK_n$	10^5	[Pękalski et al., 2013]
active IKKK	$IKKK_a$	0	[Pękalski et al., 2013]
active IKK	IKK_a	0	[Pękalski et al., 2013]
inactive IKK	IKK_i	0	[Pękalski et al., 2013]
neutral IKK	IKK_n	$2 \cdot 10^5$	[Pękalski et al., 2013]
intermediate IKK	IKK_i	0	[Pękalski et al., 2013]
negative regulator	neg	0	[Pękalski et al., 2013]
IκBα	$I\kappa B\alpha$	$1.5 \cdot 10^4$	—
NF-κB	$NF\kappa B$	0	[Pękalski et al., 2013]
IκBα : NF-κB	$NF\kappa Bcompl$	10^5	[Pękalski et al., 2013]
nuclear IκBα	$I\kappa B\alpha_n$	$6 \cdot 10^3$	[Pękalski et al., 2013]
nuclear NF-κB	$NF\kappa B_n$	0	[Pękalski et al., 2013]
nuclear IκBα : NF-κB	$NF\kappa Bcompl_n$	0	[Pękalski et al., 2013]
gene$_{\mathrm{on}}$ IκBα	$I\kappa B\alpha Gen_{on}$	0	[Pękalski et al., 2013]
gene$_{\mathrm{off}}$ IκBα	$I\kappa B\alpha Gen_{off}$	2	[Pękalski et al., 2013]
gene$_{\mathrm{on}}$ negative regulator	$negGen_{on}$	0	[Pękalski et al., 2013]
gene$_{\mathrm{off}}$ negative regulator	$negGen_{off}$	2	[Pękalski et al., 2013]
mRNA IκBα	$I\kappa B\alpha_{mRNA}$	0	—
mRNA negative regulator	neg_{mRNA}	0	—

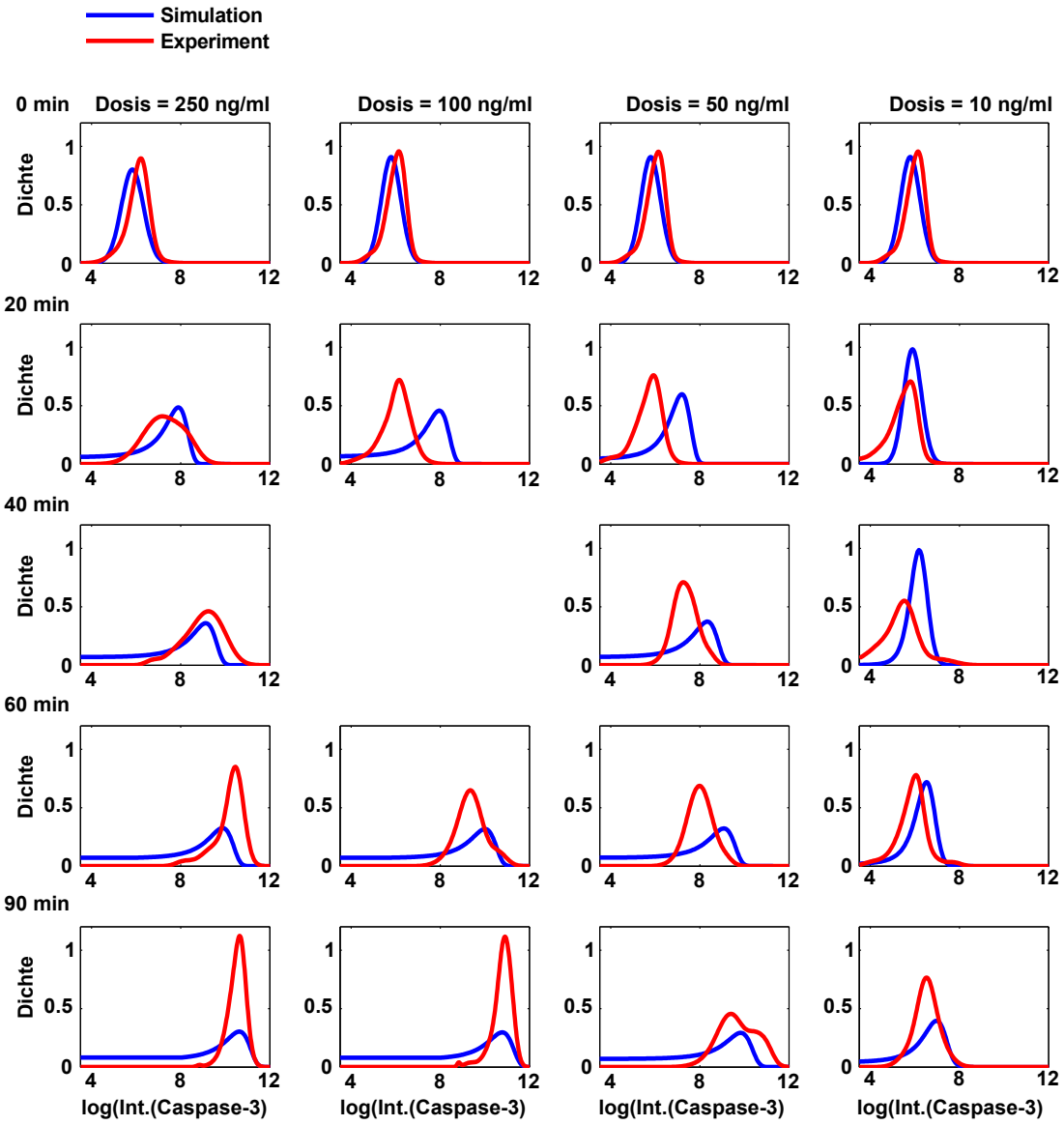

Abbildung A.4: Modellkalibrierung: Caspase-3 [Buchbinder et al., 2018].

Tabelle A.2: Auflistung chemischer Reaktionen [Buchbinder et al., 2018]. Langsame Reaktionen sind blau markiert.

Reaktion	k $[\frac{1}{s}]$	Ref.
$Casp8 + Casp8 \rightarrow p43hom$	$6.1 \cdot 10^{-12}$	–
$cFLIP_L + Casp8 \rightarrow p43het$	$1.3 \cdot 10^{-11}$	–
$cFLIP_L + Casp8 \rightarrow p43FLIP$	$1.4 \cdot 10^{-10}$	–
$p43hom + p43hom \rightarrow p43hom + p10$	$1.9 \cdot 10^{-6}$	–
$A + p43het \rightarrow A + casp3$	$1.5 \cdot 10^{-7}$	–
$A + p10 \rightarrow A + casp3$	$7.5 \cdot 10^{-2}$	–
$A + p43hom \rightarrow A + casp3$	$1.6 \cdot 10^{-2}$	–
$p43FLIP + IKKK_n \rightarrow p43FLIP + IKKK_a$	$3.2 \cdot 10^{-6}$	–
$neg + IKKK_a \rightarrow neg + IKKK_n$	$4.9 \cdot 10^{-5}$	–
$IKKK_a \rightarrow IKKK_n$	10^{-2}	[Pękalski et al., 2013]
$neg + p43FLIP \rightarrow neg$	$8.4 \cdot 10^{-10}$	–
$p43FLIP \rightarrow \oslash$	$3.4 \cdot 10^{-10}$	–
$neg \rightarrow \oslash$	$5 \cdot 10^{-4}$	–
$IKKK_a + IKK_n \rightarrow IKKK_a + IKK_a$	$5 \cdot 10^{-6}$	[Lipniacki et al., 2007]
$IKK_a \rightarrow IKK_i$	$2 \cdot 10^{-3}$	[Tay et al., 2010]
$IKK_a + neg \rightarrow IKK_i + neg$	$2 \cdot 10^{-7}$	–
$IKK_i \rightarrow IKK_{ii}$	10^{-3}	[Pękalski et al., 2013]
$IKK_{ii} \rightarrow IKK_n$	10^{-3}	[Pękalski et al., 2013]
$NF\kappa B \rightarrow NF\kappa B_n$	$3.8 \cdot 10^{-4}$	–
$I\kappa B\alpha \rightarrow I\kappa B\alpha_n$	$2 \cdot 10^{-3}$	[Pękalski et al., 2013]
$I\kappa B\alpha_n \rightarrow I\kappa B\alpha$	$5 \cdot 10^{-3}$	[Pękalski et al., 2013]
$IKK_a + I\kappa B\alpha \rightarrow IKK_a$	10^{-7}	[Pękalski et al., 2013]
$NF\kappa B_n + I\kappa B\alpha Gen_{off} \rightarrow NF\kappa B_n + I\kappa B\alpha Gen_{on}$	$4 \cdot 10^{-7}$	[Pękalski et al., 2013]
$IKK_a + NF\kappa Bcompl \rightarrow IKK_a + NF\kappa B$	$5 \cdot 10^{-7}$	[Pękalski et al., 2013]
$I\kappa B\alpha_n + NF\kappa B_n \rightarrow NF\kappa Bcompl_n$	$2.5 \cdot 10^{-6}$	[Pękalski et al., 2013]
$I\kappa B\alpha_n + I\kappa B\alpha Gen_{on} \rightarrow I\kappa B\alpha_n + I\kappa B\alpha Gen_{off}$	10^{-6}	[Pękalski et al., 2013]
$I\kappa B\alpha_n + negGen_{on} \rightarrow I\kappa B\alpha_n + negGen_{off}$	10^{-6}	–
$NF\kappa B_n + negGen_{off} \rightarrow NF\kappa B_n + negGen_{on}$	$4 \cdot 10^{-7}$	–
$NF\kappa Bcompl_n \rightarrow NF\kappa Bcompl$	$3.9 \cdot 10^{-11}$	–
$I\kappa B\alpha Gen_{on} \rightarrow I\kappa B\alpha Gen_{on} + I\kappa B\alpha_{mRNA}$	10^{-1}	[Pękalski et al., 2013]
$negGen_{on} \rightarrow negGen_{on} + neg_{mRNA}$	10^{-1}	–
$I\kappa B\alpha_{mRNA} \rightarrow I\kappa B\alpha_{mRNA} + I\kappa B\alpha$	$5 \cdot 10^{-1}$	[Pękalski et al., 2013]
$neg_{mRNA} \rightarrow neg_{mRNA} + neg$	$5 \cdot 10^{-1}$	–
$I\kappa B\alpha_{mRNA} \rightarrow \oslash$	$7.5 \cdot 10^{-4}$	[Tay et al., 2010]
$neg_{mRNA} \rightarrow \oslash$	$7.5 \cdot 10^{-4}$	–
$I\kappa B\alpha + NF\kappa B \rightarrow NF\kappa Bcompl$	$5 \cdot 10^{-7}$	[Pękalski et al., 2013]
$cFLIP_{RS} + Casp8 \rightarrow Casp8het$	$1.6 \cdot 10^{-2}$	–
$p43het \rightarrow \oslash$	$1.6 \cdot 10^{-7}$	–
$p10 \rightarrow \oslash$	$4.8 \cdot 10^{-7}$	–
$p43hom \rightarrow \oslash$	$1.1 \cdot 10^{-9}$	–

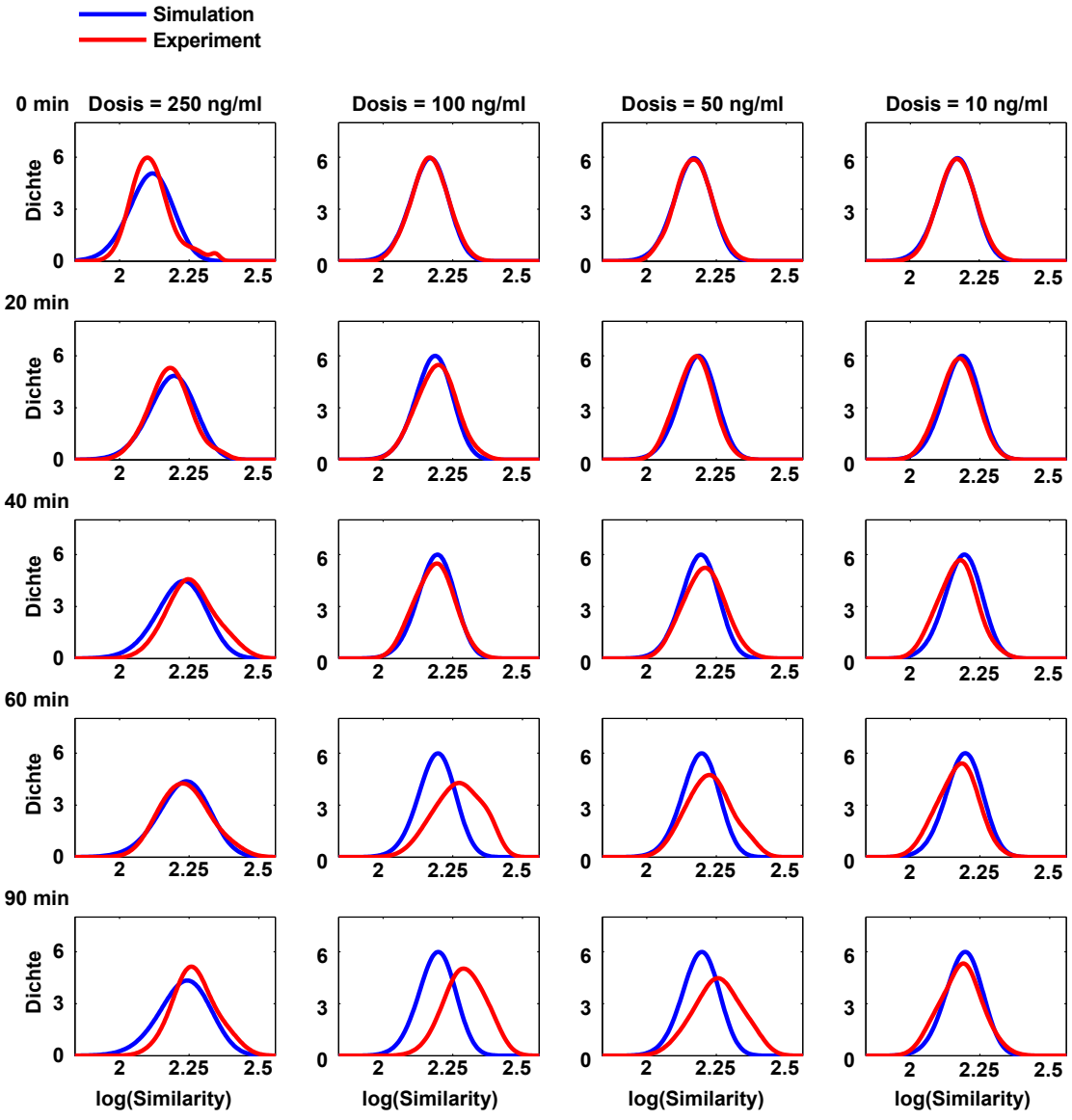

Abbildung A.5: Modellkalibrierung: nukleares NF-κB [Buchbinder et al., 2018].

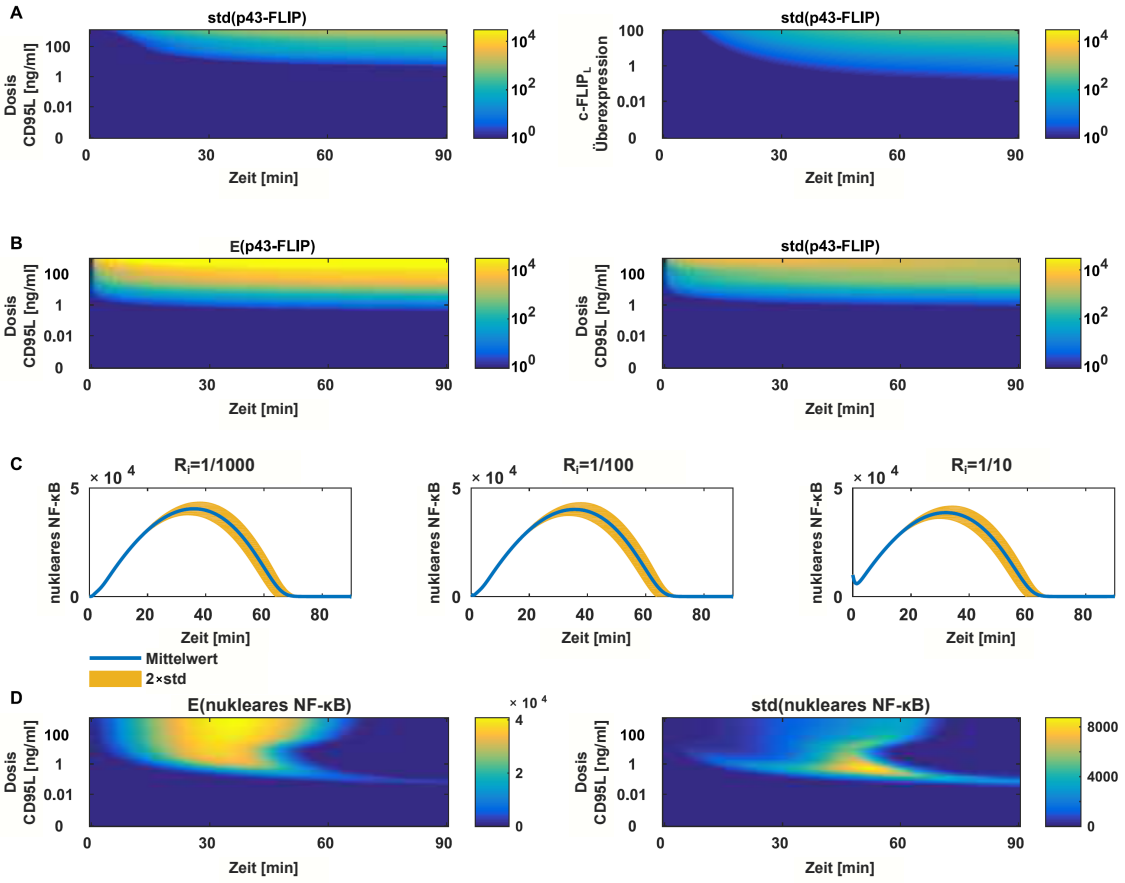

Abbildung A.6: Zusätzliche Modellvorhersagen bezüglich p43-FLIP und nuklearem NF-κB [Buchbinder et al., 2018]. **(A)** Standardabweichung der Dynamik von p34-FLIP in Abhängigkeit von der Stimulationsdosis und c-FLIP$_{\mathrm{L}}$-Überexpression (nur intrinsische Störung). **(B)** Mittelwert und Standardabweichung der Dynamik von p34-FLIP in Abhängigkeit von der Stimulationsdosis (extrinsische und intrinsische Störung). **(C)** Die Variation des Verhältnisses von nuklearem und cytosolischem NF-κB R_i bewirkt keine signifikante Änderung der Systemdynamik. **(D)** Mittelwert und Standardabweichung der Dynamik von nuklearem NF-κB in Abhängigkeit der Stimulationsstärke (extrinsische und intrinsische Störung).

A.4 Machine-Learning: Fallstudie 1

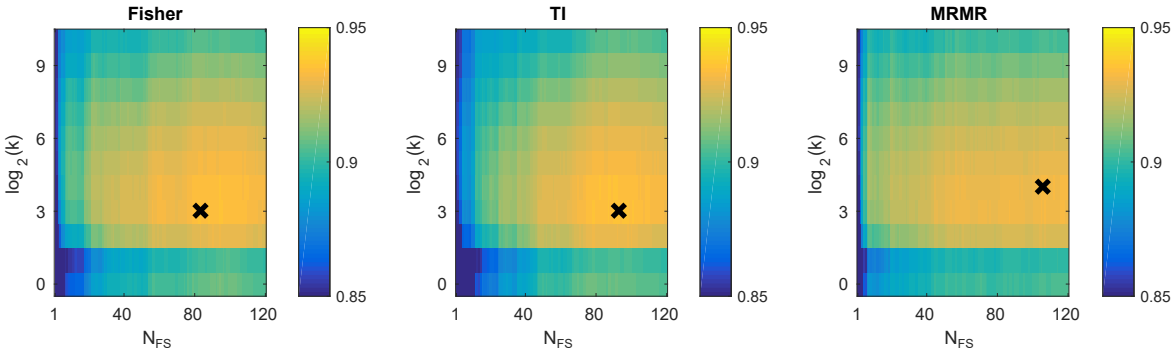

Abbildung A.7: Modellwahl für die NNK [Pischel et al., 2018].

Tabelle A.3: Optimale Modelle [Pischel et al., 2018].

	Fisher		TI		MRMR	
Algo.	Δ	Hyperpara.	Δ	Hyperpara.	Δ	Hyperpara.
LDA	0.947	n : 116	0.948	n : 118	0.947	n : 120
QDA	0.938	n : 117	0.939	n : 117	0.937	n : 113
NNK	0.933	n : 83	0.935	n : 93	0.937	n : 106
		k : 8		k : 8		k : 16
SVM	0.941	n : 95	0.948	n : 99	0.947	n : 102
		C : 10		C : 10		C : 1
		γ : 10^{-2}		γ : 10^{-2}		γ : 10^{-2}

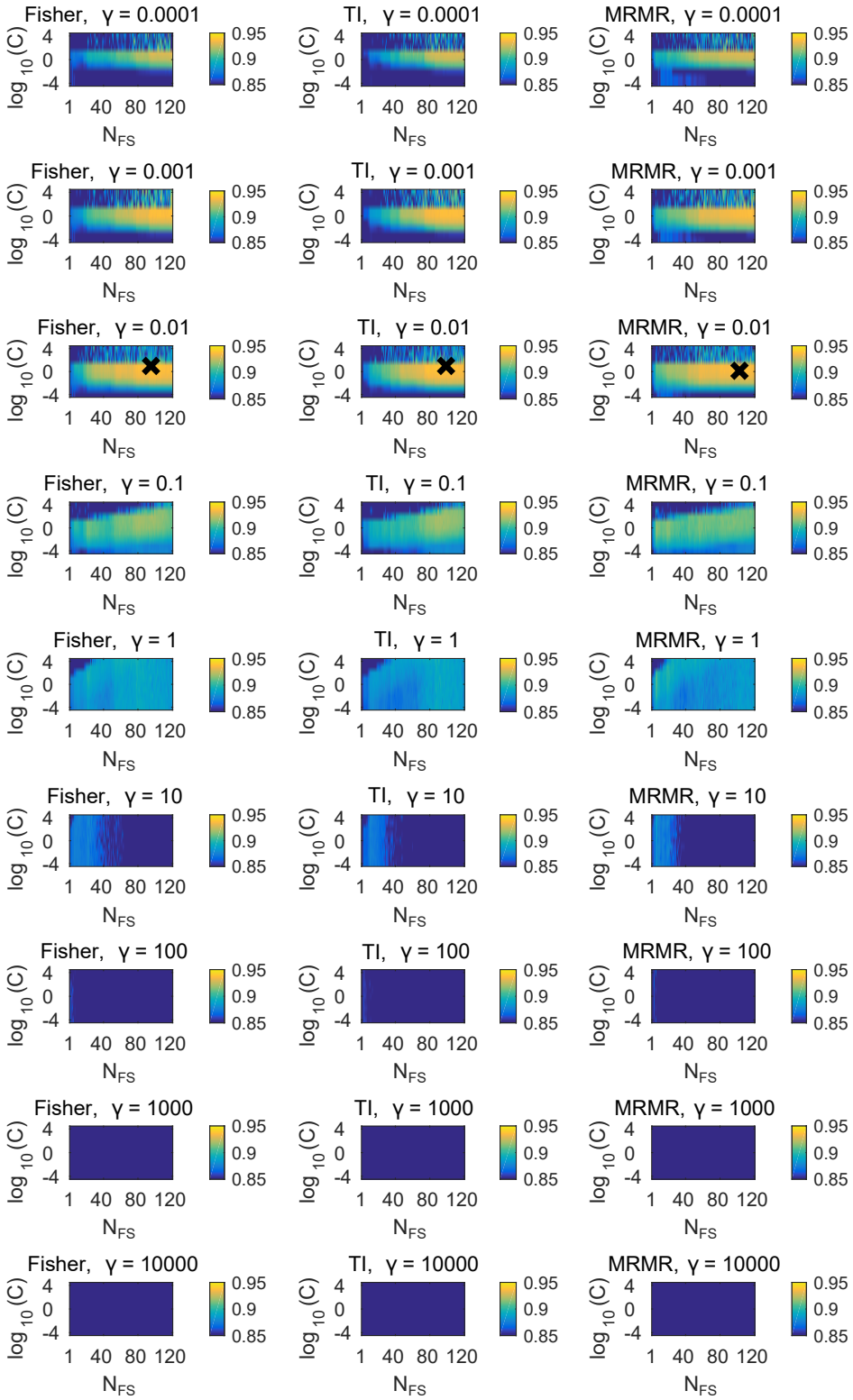

Abbildung A.8: Modellwahl für die SVM [Pischel et al., 2018].

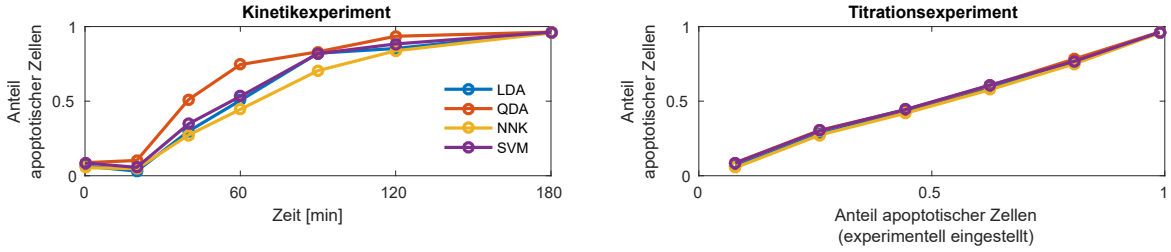

Abbildung A.9: Prädiktion hinsichtlich Kinetik- und Titrationsexperimenten [Pischel et al., 2018].

A.5 Machine-Learning: Fallstudie 2

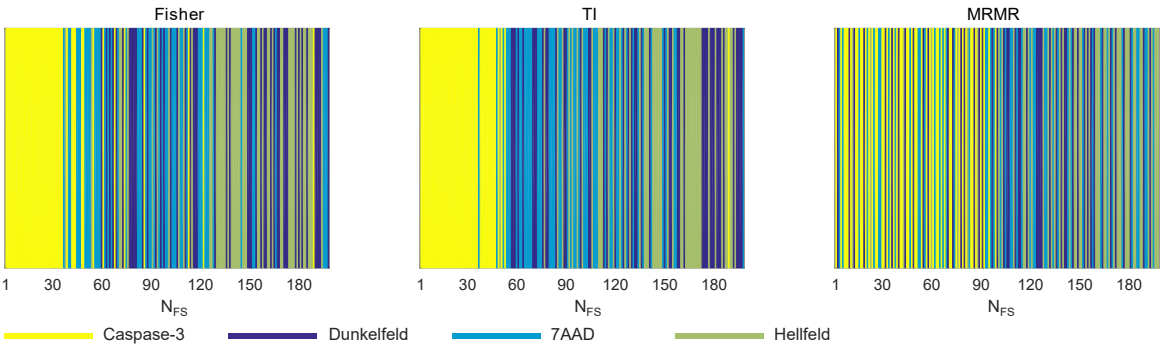

Abbildung A.10: Selektion der Merkmale [Pischel et al., 2018].

Tabelle A.4: Optimale Modelle [Pischel et al., 2018].

	Fisher		TI		MRMR	
Algo.	$\bar{\Delta}$	Hyperpara.	$\bar{\Delta}$	Hyperpara.	$\bar{\Delta}$	Hyperpara.
LDA	0.986	$n:$ 40	0.986	$n:$ 44	0.985	$n:$ 71
QDA	0.981	$n:$ 42	0.979	$n:$ 47	0.981	$n:$ 93
NNK	0.984	$n:$ 26	0.985	$n:$ 47	0.984	$n:$ 29
		$k:$ 16		$k:$ 8		$k:$ 16
SVM	0.983	$n:$ 39	0.980	$n:$ 48	0.979	$n:$ 55
		$C:$ 10		$C:$ 10		$C:$ 1
		$\gamma:$ 10^{-2}		$\gamma:$ 10^{-2}		$\gamma:$ 10^{-3}

Literaturverzeichnis

Ackermann, M. (2015). A functional perspective on phenotypic heterogeneity in microorganisms. *Nature Reviews Microbiology*, 13(8):497–508.

Adurthi, N., Singla, P. und Singh, T. (2017). Conjugate unscented transformation: Applications to estimation and control. *Journal of Dynamic Systems, Measurement, and Control*, 140(3). 030907.

Afshari, H., Gadsden, S. und Habibi, S. (2017). Gaussian filters for parameter and state estimation: A general review of theory and recent trends. *Signal Processing*, 135:218–238.

Akkermans, S., Nimmegeers, P. und Van Impe, J. (2018). A tutorial on uncertainty propagation techniques for predictive microbiology models: A critical analysis of state-of-the-art techniques. *International Journal of Food Microbiology*, 282:1–8.

Albeck, J., Burke, J., Aldridge, B., Zhang, M., Lauffenburger, D. und Sorger, P. (2008). Quantitative analysis of pathways controlling extrinsic apoptosis in single cells. *Molecular Cell*, 30(1):11–25.

Altschuler, S. und Wu, L. (2010). Cellular heterogeneity: Do differences make a difference? *Cell*, 141(4):559–563.

Amos, B. (2000). Lessons from the history of light microscopy. *Nature Cell Biology*, 2(8):E151–E152.

Andrews, S. und Bray, D. (2004). Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Physical Biology*, 1(3):137–151.

Angermueller, C., Pärnamaa, T., Parts, L. und Stegle, O. (2016). Deep learning for computational biology. *Molecular Systems Biology*, 12(7).

Argyris, J., Faust, G., Haase, M. und Friedrich, R. (2015). *An Exploration of Dynamical Systems and Chaos*. Springer, 2. Auflage.

Azunre, P., Gómez-Uribe, C. und Verghese, G. (2011). Mass fluctuation kinetics: Analysis and computation of equilibria and local dynamics. *IET Systems Biology*, 5(6):325–335.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. und Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643.

Basiji, D., Ortyn, W., Liang, L., Venkatachalam, V. und Morrissey, P. (2007). Cellular image analysis and imaging by flow cytometry. *Clinics in Laboratory Medicine*, 27(3):653–670.

Bayati, B. (2017). Quantifying uncertainty in the chemical master equation. *The Journal of Chemical Physics*, 146(24).

Ben-Hur, A., Ong, C., Sonnenburg, S., Schölkopf, B. und Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4(10).

Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural Computation*, 12(8):1889–1900.

Bergstra, J., Bardenet, R., Bengio, Y. und Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011*.

Bergstra, J. und Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13:281–305.

Bernstein, D. (2005). Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 71(4).

Biancalani, T., Dyson, L. und McKane, A. (2014). Noise-induced bistable states and their mean switching time in foraging colonies. *Physical Review Letters*, 112(3).

Blake, W., Kærn, M., Cantor, C. und Collins, J. (2003). Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637.

Blasi, T., Hennig, H., Summers, H., Theis, F., Cerveira, J., Patterson, J., Davies, D., Filby, A., Carpenter, A. und Rees, P. (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nature Communications*, 7.

Bolón-Canedo, V., Sánchez-Maroño, N. und Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3):483–519.

Bray, W. (1921). A periodic reaction in homogeneous solution and its relation to catalysis. *Journal of the American Chemical Society*, 43(6):1262–1267.

Briggs, T. und Rauscher, W. (1973). An oscillating iodine clock. *Journal of Chemical Education*, 50(7):496.

Brown, G., Pocock, A., Zhao, M.-J. und Luján, M. (2012). Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13:27–66.

Buchbinder, J., Pischel, D., Sundmacher, K., Flassig, R. und Lavrik, I. (2018). Quantitative single cell analysis uncovers the life/death decision in CD95 network. *PLoS Computational Biology*, 14(9).

Bungartz, H.-J. und Griebel, M. (2004). Sparse grids. *Acta Numerica*, 13:147–269.

Bunker, D., Garrett, B., Kleindienst, T. und Long III, G. (1974). Discrete simulation methods in combustion kinetics. *Combustion and Flame*, 23(3):373–379.

Cao, Y., Gillespie, D. und Petzold, L. (2005). The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1).

Cao, Y., Gillespie, D. und Petzold, L. (2006). Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4):044109.

Capp, J.-P. (2017). Tissue disruption increases stochastic gene expression thus producing tumors: Cancer initiation without driver mutation. *International Journal of Cancer*, 140(11):2408–2413.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Chen, C., Mahjoubfar, A., Tai, L.-C., Blaby, I., Huang, A., Niazi, K. und Jalali, B. (2016). Deep learning in label-free cell classification. *Scientific Reports*, 6.

Cho, J., Lee, K., Shin, E., G., C. und Do, S. (2015). Medical image deep learning with hospital PACS dataset. *CoRR*, arXiv/1511.06348.

Chuang, H.-Y., Hofree, M. und Ideker, T. (2010). A decade of systems biology. *Annual Review of Cell and Developmental Biology*, 26:721–744.

Chubb, J. (2017). Symmetry breaking in development and stochastic gene expression. *Wiley Interdisciplinary Reviews: Developmental Biology*, 6(6).

Coroiu, A. (2016). Tuning model parameters through a genetic algorithm approach. *Proceedings - 2016 IEEE 12th International Conference on Intelligent Computer Communication and Processing*, pages 135–140.

Cortes, C. und Vapnik, V. (1995). Support-vector networks. *Machine Learning*,

20(3):273–297.

Cotter, T. (2009). Apoptosis and cancer: The genesis of a research field. *Nature Reviews Cancer*, 9(7):501–507.

Croze, O., Ferguson, G., Cates, M. und Poon, W. (2011). Migration of chemotactic bacteria in soft agar: Role of gel concentration. *Biophysical Journal*, 101(3):525–534.

Delafosse, A., Calvo, S., Collignon, M.-L., Delvigne, F., Crine, M. und Toye, D. (2015). Euler-lagrange approach to model heterogeneities in stirred tank bioreactors - comparison to experimental flow characterization and particle tracking. *Chemical Engineering Science*, 134:457–466.

Delvigne, F., Baert, J., Sassi, H., Fickers, P., Grünberger, A. und Dusny, C. (2017). Taking control over microbial populations: Current approaches for exploiting biological noise in bioprocesses. *Biotechnology Journal*, 12(7):1600549.

Delvigne, F. und Goffin, P. (2014a). Microbial heterogeneity affects bioprocess robustness: Dynamic single-cell analysis contributes to understanding of microbial populations. *Biotechnology Journal*, 9(1):61–72.

Delvigne, F., Zune, Q., Lara, A. R., Al-Soud, W. und Sørensen, S. J. (2014b). Metabolic variability in bioprocessing: Implications of microbial phenotypic heterogeneity. *Trends in Biotechnology*, 32(12):608–616.

Dietterich, T. und Kong, E. (1995). Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report.

Djuric, N., Lan, L., Vucetic, S. und Wang, Z. (2013). Budgetedsvm: A toolbox for scalable SVM approximations. *Journal of Machine Learning Research*, 14:3813–3817.

Drawert, B., Hellander, A., Bales, B., Banerjee, D., Bellesia, G., Daigle, B.J., J., Douglas, G., Gu, M., Gupta, A., Hellander, S., Horuk, C., Nath, D., Takkar, A., Wu, S., Lötstedt, P., Krintz, C. und Petzold, L. (2016). Stochastic simulation service: Bridging the gap between the computational expert and the biologist. *PLoS Computational Biology*, 12(12).

Dusny, C., Fritzsch, F., Frick, O. und Schmid, A. (2012). Isolated microbial single cells and resulting micropopulations grow: Faster in controlled environments. *Applied and Environmental Microbiology*, 78(19):7132–7136.

Eldar, A. und Elowitz, M. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173.

Elf, J. und Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical net-

works with the linear noise approximation. *Genome Research*, 13(11):2475–2484.

Eliceiri, K., Berthold, M., Goldberg, I., Ibáñez, L., Manjunath, B., Martone, M., Murphy, R., Peng, H., Plant, A., Roysam, B., Stuurmann, N., Swedlow, J., Tomancak, P. und Carpenter, A. (2012). Biological imaging software tools. *Nature Methods*, 9(7):697–710.

Elowitz, M., Levine, A., Siggia, E. und Swain, P. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186.

Epstein, I. (1995). The consequences of imperfect mixing in autocatalytic chemical and biological systems. *Nature*, 374(6520):321–327.

Escalante, H., Montes, M. und Sucar, L. (2009). Particle swarm model selection. *Journal of Machine Learning Research*, 10:405–440.

Fachet, M., Hermsdorf, D., Rihko-Struckmann, L. und Sundmacher, K. (2016). Flow cytometry enables dynamic tracking of algal stress response: A case study using carotenogenesis in dunaliella salina. *Algal Research*, 13:227–234.

Fan, J., Han, F. und Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1(2):293–314.

Fedorov, V. (2010). Optimal experimental design. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):581–589.

Ferrell Jr, J. (2002). Self-perpetuating states in signal transduction: Positive feedback, double-negative feedback and bistability. *Current Opinion in Cell Biology*, 14(2):140–148.

Field, R. und Noyes, R. (1974). Oscillations in chemical systems. iv. limit cycle behavior in a model of a real chemical reaction. *The Journal of Chemical Physics*, 60(5):1877–1884.

Flassig, R. und Sundmacher, K. (2012). Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks. *Bioinformatics*, 28(23):3089–3096.

France, S., Douglas Carroll, J. und Xiong, H. (2012). Distance metrics for high dimensional nearest neighborhood recovery: Compression and normalization. *Information Sciences*, 184(1):92–110.

Fricker, N., Beaudouin, J., Richter, P., Eils, R., Krammer, P. und Lavrik, I. (2010). Model-based dissection of CD95 signaling dynamics reveals both a pro- and antiapoptotic role of c-FLIP$_{\mathrm{L}}$. *Journal of Cell Biology*, 190(3):377–389.

Furusawa, C., Suzuki, T., Kashiwagi, A., Yomo, T. und Kaneko, K. (2005). Ubiquity of log-normal distributions in intra-cellular reaction dynamics. *BIOPHYSICS*, 1:25–31.

Fussenegger, M., Bailey, J. und Varner, J. (2000). A mathematical model of caspase function in apoptosis. *Nature Biotechnology*, 18(7):768–774.

Gardiner, C. (1976). A comment on chemical langevin equations. *Journal of Statistical Physics*, 15(6):451–454.

Gernaey, K., Lantz, A., Tufvesson, P., Woodley, J. und Sin, G. (2010). Application of mechanistic models to fermentation and biocatalysis for next-generation processes. *Trends in Biotechnology*, 28(7):346 – 354.

Gewin, V. (2016). An open mind on open data. *Nature*, 529(7584):117–119.

Gillespie, D. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434.

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361.

Gillespie, D. (1992). A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1-3):404–425.

Gillespie, D. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733.

Gillespie, D. (2007). Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55.

Gillespie, D., Hellander, A. und Petzold, L. (2013). Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17).

Golks, A., Brenner, D., Krammer, P. und Lavrik, I. (2006). The c-FLIP-NH2 terminus (p22-flip) induces NF-κB activation. *Journal of Experimental Medicine*, 203(5):1295–1305.

Grima, R. (2010). An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *The Journal of Chemical Physics*, 133(3).

Grima, R. (2011). Construction and accuracy of partial differential equation approximations to the chemical master equation. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 84(5).

Gupta, A. und Rawlings, J. (2014). Comparison of parameter estimation methods in

stochastic chemical kinetic models: Examples in systems biology. *AIChE Journal*, 60(4):1253–1268.

Guyon, I. und Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Guyon, I., Saffari, A., Dror, G. und Cawley, G. (2010). Model selection: Beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11:61–87.

Hartman, R., McMullen, J. und Jensen, K. (2011). Deciding whether to go with the flow: Evaluating the merits of flow reactors for synthesis. *Angewandte Chemie - International Edition*, 50(33):7502–7519.

Haseltine, E. und Rawlings, J. (2002). Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of Chemical Physics*, 117(15):6959–6969.

Haseltine, E. und Rawlings, J. (2005). On the origins of approximations for stochastic chemical kinetics. *The Journal of Chemical Physics*, 123(16).

Hastie, T., Tibshirani, R. und Friedman, J. (2003). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

He, H. und Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

Hennig, H., Rees, P., Blasi, T., Kamentsky, L., Hung, J., Dao, D., Carpenter, A. und Filby, A. (2017). An open-source solution for advanced imaging flow cytometry data analysis using machine learning. *Methods*, 112:201–210.

Henry, C., Hollville, E. und Martin, S. (2013). Measuring apoptosis by microscopy and flow cytometry. *Methods*, 61(2):90–97.

Herzenberg, L., Parks, D., Sahaf, B., Perez, O., Roederer, M. und Herzenberg, L. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: A view from stanford. *Clinical Chemistry*, 48(10):1819–1827.

Higuchi, T., Flies, D., Marjon, N., Mantia-Smaldone, G., Ronner, L., Gimotty, P. und Adams, S. (2015). CTLA-4 blockade synergizes therapeutically with parp inhibition in BRCA1-deficient ovarian cancer. *Cancer Immunology Research*, 3(11):1257–1268.

Hilfinger, A. und Paulsson, J. (2011). Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences of the United States of America*, 108(29):12167–12172.

Hill, A. (1913). The combinations of haemoglobin with oxygen and with carbon mon-

oxide. i. *Biochemical Journal*, 7(5):471–480.

Hines, W. und Montgomery, D. (1990). *Probability and statistics in engineering and management science*. Wiley-Blackwell, 3. Auflage.

Hodgkin, A. und Huxley, A. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544.

Hucka, M., Finney, A., Sauro, H., Bolouri, H., Doyle, J., Kitano, H., Arkin, A., Bornstein, B., Bray, D., Cornish-Bowden, A., Cuellar, A., Dronov, S., Gilles, E., Ginkel, M., Gor, V., Goryanin, I., Hedley, W., Hodgman, T., Hofmeyr, J.-H., Hunter, P., Juty, N., Kasberger, J., Kremling, A., Kummer, U., Le Novère, N., Loew, L., Lucio, D., Mendes, P., Minch, E., Mjolsness, E., Nakayama, Y., Nelson, M., Nielsen, P., Sakurada, T., Schaff, J., Shapiro, B., Shimizu, T., Spence, H., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. und Wang, J. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

Jahnke, T. und Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26.

Jain, A., Duin, R. und Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37.

James, F. (1980). Monte carlo theory and practice. *Reports on Progress in Physics*, 43(9):1145–1189.

Jaqaman, K. und Danuser, G. (2006). Linking data to models: Data regression. *Nature Reviews Molecular Cell Biology*, 7(11):813–819.

Jaye, D., Bray, R., Gebel, H., Harris, W. und Waller, E. (2012). Translational applications of flow cytometry in clinical practice. *Journal of Immunology*, 188(10):4715–4719.

Jogaiah, S., Govind, S. und Tran, L.-S. (2013). Systems biology-based approaches toward understanding drought tolerance in food crops. *Critical Reviews in Biotechnology*, 33(1):23–39.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., Lewis, S., Birney, E. und Stein, L. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database Issue):D428–D432.

Julier, S. und Uhlmann, J. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422.

Julier, S., Uhlmann, J. und Durrant-Whyte, H. (2000). A new method for the nonlinear transformation of means and covariances in filters and estimators. *IEEE Transactions on Automatic Control*, 45(3):477–482.

Kærn, M., Elston, T., Blake, W. und Collins, J. (2005). Stochasticity in gene expression: From theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464.

Kaiser, N., Flassig, R. und Sundmacher, K. (2016). Probabilistic reactor design in the framework of elementary process functions. *Computers and Chemical Engineering*, 94:45–59.

Kandasamy, K., Sujatha Mohan, S., Raju, R., Keerthikumar, S., Sameer Kumar, G., Venugopal, A., Telikicherla, D., Navarro, D., Mathivanan, S., Pecquet, C., Gollapudi, S., Tattikota, S., Mohan, S., Padhukasahasram, H., Subbannayya, Y., Goel, R., Jacob, H., Zhong, J., Sekhar, R., Nanjappa, V., Balakrishnan, L., Subbaiah, R., Ramachandra, Y., Abdul Rahiman, B., Keshava Prasad, T., Lin, J.-X., Houtman, J., Desiderio, S., Renauld, J.-C., Constantinescu, S., Ohara, O., Hirano, T., Kubo, M., Singh, S., Khatri, P., Draghici, S., Bader, G., Sander, C., Leonard, W. und Pandey, A. (2010). Netpath: A public resource of curated signal transduction pathways. *Genome Biology*, 11(1).

Kaufmann, B. und van Oudenaarden, A. (2007). Stochastic gene expression: from single molecules to the proteome. *Current Opinion in Genetics and Development*, 17(2):107–112.

Kazeroonian, A., Fröhlich, F., Raue, A., Theis, F. und Hasenauer, J. (2016). Cerena: Chemical reaction network analyzer–a toolbox for the simulation and analysis of stochastic chemical kinetics. *PLoS ONE*, 11(1):e0146732.

Kazeroonian, A., Theis, F. und Hasenauer, J. (2017). A scalable moment-closure approximation for large-scale biochemical reaction networks. *Bioinformatics*, 33(14):i293–i300.

Kim, K., Qian, H. und Sauro, H. (2013). Nonlinear biochemical signal processing via noise propagation. *The Journal of Chemical Physics*, 139(14).

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

Kitano, H. (2004). Biological robustness. *Nature Reviews Genetics*, 5(11):826–837.

Klipp, E., Liebermeister, W., Wierling, C. und Kowald, A. (2016). *Systems Biology: A Textbook*. Wiley-Blackwell, 2. Auflage.

Krammer, P. (2000). CD95's deadly mission in the immune system. *Nature*, 407(6805):789–795.

Krammer, P., Arnold, R. und Lavrik, I. (2007). Life and death in peripheral T cells. *Nature Reviews Immunology*, 7(7):532–542.

Kreutz, C., Rodriguez, M., Maiwald, T., Seidl, M., Blum, H., Mohr, L. und Timmer, J. (2007). An error model for protein quantification. *Bioinformatics*, 23(20):2747–2753.

Ku, H. (1966). Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards*, 70c(4):263–273.

Lapin, A., Müller, D. und Reuss, M. (2004). Dynamic behavior of microbial populations in stirred bioreactors simulated with euler-lagrange methods: Traveling along the lifelines of single cells. *Industrial and Engineering Chemistry Research*, 43(16):4647–4656.

Lavrik, I. (2010). Systems biology of apoptosis signaling networks. *Current Opinion in Biotechnology*, 21(4):551–555.

Lavrik, I. (2014). Systems biology of death receptor networks: Live and let die. *Cell Death and Disease*, 5(5).

Lavrik, I., Eils, R., Fricker, N., Pforr, C. und Krammer, P. (2009). Understanding apoptosis by systems biology approaches. *Molecular BioSystems*, 5(10):1105–1111.

Lazebnik, Y. (2002). Can a biologist fix a radio? - or, what I learned while studying apoptosis. *Cancer Cell*, 2(3):179–182.

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J. und Hucka, M. (2006). Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic acids research.*, 34(Database Issue):D689–691.

Lee, C., Kim, K.-H. und Kim, P. (2009). A moment closure method for stochastic reaction networks. *The Journal of Chemical Physics*, 130(13).

Lee, R., Walker, S., Savery, K., Frank, D. und Gaudet, S. (2014). Fold change of nuclear NF-κB determines TNF-induced transcription in single cells. *Molecular Cell*, 53(6):867–879.

Lemm, S., Blankertz, B., Dickhaus, T. und Müller, K.-R. (2011). Introduction to machine learning for brain imaging. *NeuroImage*, 56(2):387–399.

Lencastre Fernandes, R., Nierychlo, M., Lundin, L., Pedersen, A., Puentes Tellez, P.,

Dutta, A., Carlquist, M., Bolic, A., Schäpper, D., Brunetti, A., Helmark, S., Heins, A.-L., Jensen, A., Nopens, I., Rottwitt, K., Szita, N., van Elsas, J., Nielsen, P., Martinussen, J., Sørensen, S., Lantz, A. und Gernaey, K. (2011). Experimental methods and modeling techniques for description of cell population heterogeneity. *Biotechnology Advances*, 29(6):575–599.

Lewis, D. (1992). Feature selection and feature extraction for text categorization. *Proceedings of the workshop on Speech and Natural Language*, pages 212–217.

Lidstrom, M. und Konopka, M. (2010). The role of physiological heterogeneity in microbial population behavior. *Nature Chemical Biology*, 6(10):705–712.

Lillacci, G. und Khammash, M. (2013). The signal within the noise: Efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–2319.

Limpert, E., Stahel, W. und Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51(5):341–352.

Lin, W. und Chen, J. (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in Bioinformatics*, 14(1):13–26.

Lipniacki, T., Puszynski, K., Paszek, P., Brasier, A. und Kimmel, M. (2007). Single TNFα trimers mediating NF-κB activation: Stochastic robustness of NF-κB signaling. *BMC Bioinformatics*, 8.

Macklin, D., Ruggero, N. und Covert, M. (2014). The future of whole-cell modeling. *Current Opinion in Biotechnology*, 28:111–115.

Marchetti, L., Lombardo, R. und Priami, C. (2017). Hsimulator: Hybrid stochastic/-deterministic simulation of biochemical reaction networks. *Complexity*, 2017.

Maußner, J. und Freund, H. (2018). Optimization under uncertainty in chemical engineering: Comparative evaluation of unscented transformation methods and cubature rules. *Chemical Engineering Science*, 183:329–345.

McCall, J. (2005). Genetic algorithms for modelling and optimisation. *Journal of Computational and Applied Mathematics*, 184(1):205–222.

Meacham, C. und Morrison, S. (2013). Tumour heterogeneity and cancer cell plasticity. *Nature*, 501(7467):328–337.

Menegaz, H., Ishihara, J., Borges, G. und Vargas, A. (2015). A systematization of the unscented kalman filter theory. *IEEE Transactions on Automatic Control*, 60(10):2583–2598.

Michaelis, L. und Menten, M. L. (1913). Die kinetik der invertinwirkung. *Biochemische Zeitschrift*, 49:333–369.

Moore, G. E. (1998). Cramming more components onto integrated circuits. *Proceedings of the IEEE*, 86(1):82–85.

Munsky, B. und Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4).

Munsky, B. und Khammash, M. (2008). The finite state projection approach for the analysis of stochastic noise in gene networks. *IEEE Transactions on Automatic Control*, 53(SPECIAL ISSUE):201–214.

Nakanishi, T. (1972). Stochastic analysis of an oscillating chemical reaction. *Journal of the Physical Society of Japan*, 32(5):1313–1322.

Neumann, L., Pforr, C., Beaudouin, J., Pappa, A., Fricker, N., Krammer, P., Lavrik, I. und Eils, R. (2010). Dynamics within the CD95 death-inducing signaling complex decide life and death of cells. *Molecular Systems Biology*, 6.

Nimmegeers, P., Telen, D., Logist, F. und Van Impe, J. (2016). Dynamic optimization of biological networks under parametric uncertainty. *BMC Systems Biology*, 10(1).

Novick, A. (1955). Growth of bacteria. *Annual Review of Microbiology*, 9(1):97–110.

O'Neill, K., Aghaeepour, N., Špidlen, J. und Brinkman, R. (2013). Flow cytometry bioinformatics. *PLoS Computational Biology*, 9(12).

Ori, H., Marder, E. und Marom, S. (2018). Cellular function given parametric variation in the hodgkin and huxley model of excitability. *Proceedings of the National Academy of Sciences*, 115(35):E8211–E8218.

Patnaik, P. R. (2006). External, extrinsic and intrinsic noise in cellular systems: analogies and implications for protein synthesis. *Biotechnology and Molecular Biology Reviews*, 1(4):121–127.

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175.

Pękalski, J., Zuk, P., Kochańczyk, M., Junkin, M., Kellogg, R., Tay, S. und Lipniacki, T. (2013). Spontaneous NF-κB activation by autocrine TNFα signaling: A computational analysis. *PLoS ONE*, 8(11).

Pedreira, C., Costa, E., Arroyo, M., Almeida, J. und Orfao, A. (2008). A multidimensional classification approach for the automated analysis of flow cytometry data. *IEEE Transactions on Biomedical Engineering*, 55(3):1155–1162.

Peleš, S., Munsky, B. und Khammash, M. (2006). Reduction and solution of the chemical master equation using time scale separation and finite state projection. *The Journal of Chemical Physics*, 125(20).

Peng, H., Long, F. und Ding, C. (2005). Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Pérez, A., Larrañaga, P. und Inza, I. (2009). Bayesian classifiers based on kernel density estimation: Flexible classifiers. *International Journal of Approximate Reasoning*, 50(2):341–362.

Perfetto, S., Chattopadhyay, P. und Roederer, M. (2004). Seventeen-colour flow cytometry: Unravelling the immune system. *Nature Reviews Immunology*, 4(8):648–655.

Pietkiewicz, S., Schmidt, J. und Lavrik, I. (2015). Quantification of apoptosis and necroptosis at the single cell level by a combination of imaging flow cytometry with classical Annexin V/propidium iodide staining. *Journal of Immunological Methods*, 423:99–103.

Pischel, D., Buchbinder, J., Sundmacher, K., Lavrik, I. und Flassig, R. (2018). A guide to automated apoptosis detection: How to make sense of imaging flow cytometry data. *PLoS ONE*, 13(5):e0197208.

Pischel, D., Flassig, R. und Sundmacher, K. (2016). Efficient simulation of heterogeneity and stochasticity in microbial processes. *Computer Aided Chemical Engineering*, 38:1213–1218.

Pischel, D., Sundmacher, K. und Flassig, R. (2017). Efficient simulation of intrinsic, extrinsic and external noise in biochemical systems. *Bioinformatics*, 33(14):i319–i324.

Poovathingal, S. und Gunawan, R. (2010). Global parameter estimation methods for stochastic biochemical systems. *BMC Bioinformatics*, 11.

Press, W. (2002). *Numerical recipes in C: the art of scientific computing*. Cambridge University Press, 2. Auflage.

Prigogine, I. und Lefever, R. (1968). Symmetry breaking instabilities in dissipative systems. ii. *The Journal of Chemical Physics*, 48(4):1695–1700.

Qian, H., Shi, P.-Z. und Xing, J. (2009). Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. *Physical Chemistry Chemical Physics*, 11(24):4861–4870.

Raser, J. und O'Shea, E. (2004). Control of stochasticity in eukaryotic gene expression.

Science, 304(5678):1811–1814.

Rollié, S., Mangold, M. und Sundmacher, K. (2012). Designing biological systems: Systems engineering meets synthetic biology. *Chemical Engineering Science*, 69(1):1–29.

Roux, J., Hafner, M., Bandara, S., Sims, J., Hudson, H., Chai, D. und Sorger, P. (2015). Fractional killing arises from cell-to-cell variability in overcoming a caspase activity threshold. *Molecular Systems Biology*, 11(5):1–17.

Saeys, Y., Inza, I. und Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

Saeys, Y., Van Gassen, S. und Lambrecht, B. (2016). Computational flow cytometry: Helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16(7):449–462.

Sanft, K., Wu, S., Roh, M., Fu, J., Lim, R. und Petzold, L. (2011). Stochkit2: Software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458.

Schenkendorf, R., Kremling, A. und Mangold, M. (2009). Optimal experimental design with the sigma point method. *IET Systems Biology*, 3(1):10–23.

Schleich, K., Warnken, U., Fricker, N., Öztürk, S., Richter, P., Kammerer, K., Schnölzer, M., Krammer, P. und Lavrik, I. (2012). Stoichiometry of the CD95 death-inducing signaling complex: Experimental and modeling evidence for a death effector domain chain model. *Molecular Cell*, 47(2):306–319.

Schlögl, F. (1972). Chemical reaction models for non-equilibrium phase transitions. *Zeitschrift für Physik*, 253(2):147–161.

Schmidt, J., Pietkiewicz, S., Naumann, M. und Lavrik, I. (2015). Quantification of CD95-induced apoptosis and NF-κB activation at the single cell level. *Journal of Immunological Methods*, 423:12–17.

Schrödinger, E. (1944). *What is life? : the physical aspect of the living cell*. Cambridge University Press.

Shahrezaei, V., Ollivier, J. und Swain, P. (2008). Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4.

Sommer, C. und Gerlich, D. (2013). Machine learning in cell biology-teaching computers to recognize phenotypes. *Journal of Cell Science*, 126(24):5529–5539.

Somogyi, E., Bouteiller, J.-M., Glazier, J., König, M., Medley, J., Swat, M. und Sauro, H. (2015). Libroadrunner: A high performance SBML simulation and analysis

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

Specht, E., Braselmann, E. und Palmer, A. (2017). A critical and comparative review of fluorescent tools for live-cell imaging. *Annual Review of Physiology*, 79:93–117.

Spencer, S., Gaudet, S., Albeck, J., Burke, J. und Sorger, P. (2009). Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(7245):428–432.

Spencer, S. und Sorger, P. (2011). Measuring and modeling apoptosis in single cells. *Cell*, 144(6):926–939.

Spiller, D., Wood, C., Rand, D. und White, M. (2010). Measurement of single-cell dynamics. *Nature*, 465(7299):736–745.

Stelling, J., Sauer, U., Szallasi, Z., Doyle III, F. und Doyle, J. (2004). Robustness of cellular functions. *Cell*, 118(6):675–685.

Strasser, M., Theis, F. und Marr, C. (2012). Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophysical Journal*, 102(1):19–29.

Sun, C., Shrivastava, A., Singh, S. und Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *CoRR*, arXiv/1707.02968.

Swain, P., Elowitz, M. und Siggia, E. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800.

Tarca, A., Carey, V., Chen, X., Romero, R. und Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Computational biology*, 3(6).

Tay, S., Hughey, J., Lee, T., Lipniacki, T., Quake, S. und Covert, M. (2010). Single-cell NF-κB dynamics reveal digital activation and analogue information processing. *Nature*, 466(7303):267–271.

Tenne, D. und Singh, T. (2003). The higher order unscented filter. *Proceedings of the American Control Conference*, 3:2441–2446.

Thanh, V. und Priami, C. (2015). Simulation of biochemical reactions with time-dependent rates by the rejection-based algorithm. *The Journal of Chemical Physics*, 143(5).

Thompson, C. (1995). Apoptosis in the pathogenesis and treatment of disease. *Science*, 267(5203):1456–1462.

Toni, T. und Tidor, B. (2013). Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. *PLoS*

Computational Biology, 9(3).

van der Greef, J., Hankemeier, T. und McBurney, R. (2006). Metabolomics-based systems biology and personalized medicine: Moving towards n = 1 clinical trials? *Pharmacogenomics*, 7(7):1087–1094.

van Kampen, N. (2007). *Stochastic processes in physics and chemistry*. North Holland, 3. Auflage.

Voliotis, M., Thomas, P., Grima, R. und Bowsher, C. (2016). Stochastic simulation of biomolecular networks in dynamic environments. *PLoS Computational Biology*, 12(6).

Waldherr, S. und Haasdonk, B. (2012). Efficient parametric analysis of the chemical master equation through model order reduction. *BMC Systems Biology*, 6.

Wang, Y., Ho, S.-H., Cheng, C.-L., Guo, W.-Q., Nagarajan, D., Ren, N.-Q., Lee, D.-J. und Chang, J.-S. (2016). Perspectives on the feasibility of using microalgae for industrial wastewater treatment. *Bioresource Technology*, 222:485–497.

Weber, A., Prokazov, Y., Zuschratter, W. und Hauser, M. (2012). Desynchronisation of glycolytic oscillations in yeast cell populations. *PLoS ONE*, 7(9).

Wen, Y., Chen, Z., Lu, J., Ables, E., Scemama, J.-L., Yang, L., Lu, J. und Hu, X.-H. (2017). Quantitative analysis and comparison of 3D morphology between viable and apoptotic MCF-7 breast cancer cells and characterization of nuclear fragmentation. *PLoS ONE*, 12(9).

Westerwalbesloh, C., Grünberger, A., Stute, B., Weber, S., Wiechert, W., Kohlheyer, D. und Von Lieres, E. (2015). Modeling and CFD simulation of nutrient distribution in picoliter bioreactors for bacterial growth studies on single-cell level. *Lab on a Chip*, 15(21):4177–4186.

Wilkinson, D. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133.

Wolf, D. und Arkin, A. (2003). Motifs, modules and games in bacteria. *Current Opinion in Microbiology*, 6(2):125–134.

Wu, Y., Hu, D., Wu, M. und Hu, X. (2006). A numerical-integration perspective on gaussian filters. *IEEE Transactions on Signal Processing*, 54(8):2910–2921.

Xia, X., Owen, M., Lee, R. und Gaudet, S. (2014). Cell-to-cell variability in cell death: Can systems biology help us make sense of it all? *Cell Death and Disease*, 5.

Xue, J. und Ma, J. (2012). A comparative study of several taylor expansion me-

thods on error propagation. *Proceedings - 2012 20th International Conference on Geoinformatics*.

Zechner, C. und Koeppl, H. (2014). Uncoupled analysis of stochastic reaction networks in fluctuating environments. *PLoS Computational Biology*, 10(12).

Zhang, K., Lan, L., Wang, Z. und Moerchen, F. (2012). Scaling up kernel svm on limited resources: A low-rank linearization approach. *Journal of Machine Learning Research*, 22:1425–1434.

Abkürzungsverzeichnis

CME	Chemical Master Equation ⟨engl., „chemische Master-Gleichung“⟩
KV	Kreuzvalidierung
LDA	lineare Diskriminanzanalyse
MC	Monte Carlo
MRMR	Minimale Redundanz und maximale Relevanz
NNK	Nächste-Nachbarn-Klassifizierung
PI	Propidiumiodid
QDA	quadratische Diskriminanzanalyse
SBML	systembiologische Auszeichnungssprache ⟨engl., „Systems Biology Markup Language“⟩
SVM	Support-Vector-Machine
TI	Transinformation
TOD	Time of Decision
TOS	Time of Survival

Abbildungsverzeichnis

Tabellenverzeichnis

Algorithmenverzeichnis

Veröffentlichungen

Während des Verlaufs der Promotion habe ich verschiedene Veröffentlichungen publiziert, die in dieser Arbeit aufgegriffen und integriert worden sind:

[Pischel et al., 2016] **D. Pischel**, R.J. Flassig, K. Sundmacher. Efficient simulation of heterogeneity and stochasticity in microbial processes. *Computer Aided Chemical Engineering*, 38:12131218, 2016. (**D. Pischel** entwickelte die Methode, führte die theoretischen Analysen durch und schrieb das Manuskript.)

[Pischel et al., 2017] **D. Pischel**, K. Sundmacher, and R.J. Flassig. Efficient simulation of intrinsic, extrinsic and external noise in biochemical systems. *Bioinformatics*, 33(14):i319i324, 2017. (**D. Pischel** entwickelte die Methode, führte die theoretischen Analysen durch und schrieb das Manuskript.)

[Buchbinder et al., 2018] J.H. Buchbinder†, **D. Pischel**†, K. Sundmacher, R.J. Flassig, I.N. Lavrik. Quantitative single cell analysis uncovers the life/death decision in CD95 network. *PLoS Computational Biology*, 14(9): e1006368, 2018. (**D. Pischel** führte die theoretischen Analysen durch und schrieb Teile des Manuskripts.)
† geteilte Erstautorschaft

[Pischel et al., 2018] **D. Pischel**, J.H. Buchbinder, K. Sundmacher, I.N. Lavrik, R.J. Flassig. A guide to automated apoptosis detection: How to make sense of imaging flow cytometry data. *PLoS ONE*, 13(5):e0197208, 2018. (**D. Pischel** entwickelte die Methode, führte die theoretischen Analysen durch und schrieb Teile des Manuskripts.)

Die Publikationen [Pischel et al., 2018; Buchbinder et al., 2018] sind in Zusammenarbeit mit Jörn H. Buchbinder und Inna N. Lavrik (Translationale Entzündungsforschung, Otto-von-Guericke-Universität Magdeburg) erarbeitet worden. Dabei wurden alle experimentellen Daten von Jörn H. Buchbinder aufgenommen. Teile dieser Daten sind in dieser Arbeit dargestellt. Zudem beruhen die von mir vorgenommenen Analysen und Simulationen auf diesen Daten.

Konferenzbeiträge

D. Pischel, R.J. Flassig, K. Sundmacher. Coping with heterogeneity and stochasticity in microbial processes. *International Conference on Mathematics in (bio)Chemical Kinetics and Engineering*, Ghent, Belgien, 08.-09.11.2015 (Vortrag).

D. Pischel, R.J. Flassig, K. Sundmacher. Efficient simulation of heterogeneity and stochasticity in microbial processes. *European Symposium on Computer Aided Process Engineering*, Portorož, Slowenien, 12.-15.06.2016 (Vortrag).

D. Pischel, K. Sundmacher, R.J. Flassig. Efficient simulation of intrinsic, extrinsic and external noise in biochemical systems. *International Conference on Intelligent Systems for Molecular Biology/European Conference on Computational Biology*, Prag, Tschechische Republik, 21.-25.07.2017 (Vortrag).

D. Pischel, J.H. Buchbinder, R.J. Flassig, I.N. Lavrik, K. Sundmacher. Monitoring with Imaging Flow Cytometry - a Machine Learning Approach. *European Congress of Applied Biotechnology*, Barcelona, Spanien, 01.-05.10.2017 (Vortrag).

D. Pischel, K. Sundmacher, R.J. Flassig. Efficient Simulation of Variability and Heterogeneity in Bioprocess Engineering. *International Conference on Mathematical Modelling*, Wien, Österreich, 21.-23.02.2018 (Poster).

Selbstständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Insbesondere habe ich nicht die Hilfe einer kommerziellen Promotionsberatung in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form als Dissertation eingereicht und ist als Ganzes auch noch nicht veröffentlicht.

Magdeburg, den 06. Mai 2019 Dennis Pischel