University of California
Santa Barbara

Data-Driven Bayesian Methods for Analyzing Biochemical Reaction Networks

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Richard Maxwell Jiang

Committee in charge:

Professor Linda R Petzold, Chair
Professor William Wang
Professor Alexander Franks

December 2021

The Dissertation of Richard Maxwell Jiang is approved.

Professor William Wang

Professor Alexander Franks

Professor Linda R Petzold, Committee Chair

September 2021

Data-Driven Bayesian Methods for Analyzing Biochemical Reaction Networks

by

Richard Maxwell Jiang

To my Mom, my Dad, my Brother and Teddy

Acknowledgements

The research in Chapter 2 was supported by the US Army Contracting Command, Aberdeen Proving Ground, Natick Contracting Division under Contract No. W911QY-15-C-0026.

The research in Chapters 3,4, and 5 was supported under NIBIB Award No. 2-R01-EB014877-04A1.

The content of the information does not necessarily reflect the position or the policy of the funding agency, and no official endorsement should be inferred.

Curriculum Vitæ
Richard Maxwell Jiang

Education

2021	Ph.D. in Computer Science (Expected), University of California, Santa Barbara.
2021	M.S. in Computer Science, University of California, Santa Barbara
2013	B.S. in Financial Engineering, Columbia University

Publications

- **Richard Jiang**, Fredrik Wrede, Prashant Singh, and Linda Petzold. Sparse bayesian inference of mass-action biochemical reaction networks using the regularized horseshoe prior. *In Submission to PLOS Computational Biology*
- Fredrik Wrede, Robin Eriksson, **Richard Jiang**, Linda Petzold, Stefan Engblom, Andreas Hellander, and Prashant Singh. Robust and integrative bayesian neural networks for likelihood-free parameter inference. *arXiv preprint arXiv:2102.06521*, 2021
- **Richard M Jiang**, Fredrik Wrede, Prashant Singh, Andreas Hellander, and Linda R Petzold. Accelerated regression-based summary statistics for discrete stochastic systems via approximate simulators. *BMC Bioinformatics*, 22(1):1–17, 2021
- **Richard Jiang**, Arya A Pourzanjani, Mitchell J Cohen, and Linda Petzold. Associations of longitudinal d-dimer and factor ii on early trauma survival risk. *BMC Bioinformatics*, 22(1):1–13, 2021
- **Richard Jiang***, Bruno Jacob*, Matthew Geiger, Sean Matthew, Bryan Rumsey, Prashant Singh, Fredrik Wrede, Tau-Mu Yi, Brian Drawert, Andreas Hellander, and Linda Petzold. Epidemiological modeling in stochss live! *Bioinformatics*, 2021
- Yun Zhao, **Richard Jiang**, Zhenni Xu, Elmer Guzman, Paul K Hansma, and Linda Petzold. Scalable bayesian functional connectivity inference for multi-electrode array recordings. In *BioKDD'20*, 2020
- Arya A Pourzanjani, **Richard Jiang**, Brian Mitchell, Paul J Atzberger, and Linda R Petzold. General bayesian inference over the stiefel manifold via the givens representation. *Bayesian Analysis*, 2020
- **Richard Jiang***, Arya A Pourzanjani*, and Linda Petzold. Improving the identifiability of neural networks for bayesian inference. In *2017 NIPS Workshop on Bayesian Deep Learning*, 2017
- Arya A. Pourzanjani, Tie Bo Wu, **Richard M. Jiang**, Mitchell J. Cohen, and Linda R. Petzold. Understanding coagulopathy using multi-view data in the presence of sub-cohorts: A hierarchical subspace approach. In *Proceedings of the 2nd Machine*

Learning for Healthcare Conference, volume 68 of *Proceedings of Machine Learning Research*, pages 338–351. PMLR, 18–19 Aug 2017

- Yuanyang Zhang, **Richard Jiang**, and Linda Petzold. Survival topic models for predicting outcomes for trauma patients. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1497–1504. IEEE, 2017

Abstract

Data-Driven Bayesian Methods for Analyzing Biochemical Reaction Networks

by

Richard Maxwell Jiang

Significant modern advances in faster and cheaper measurement techniques for biological processes has led to an explosion in the availability of biological data, from the clinical scale down to the molecular scale with the promise to vastly increase our understanding of these complex systems. However, a critical step in accomplishing this is developing flexible data-driven and statistical methods to make sense of these rich datasets. As measurements in this domain are frequently noisy and sparse, Bayesian methods are promising for providing not only accurate estimates that capture prior knowledge, but also uncertainties in drawing conclusions.

In this thesis, we describe our contributions to the analysis and development of biochemical reaction networks using Bayesian methods, both from applied and computational directions. We begin with an application of a Bayesian model for understanding a biochemical process at the clinical level. Then, we follow by describing our contributions to inferring the parameters and the structure of the biochemical reaction networks from experimental data using Bayesian techniques.

Specifically, we first describe our work in applying a hierarchical Bayesian joint longitudinal survival model to analyze the clinical risks of the protein biomarkers D-Dimer and Factor II, which play a crucial role in the coagulation cascade. Next, we transition to the molecular scale and start by describing our innovations in accelerating a critical step component of the Approximate Bayesian Computation algorithm for Bayesian inference of the parameters of stochastic biochemical reaction networks. Then, we discuss

the problem of inferring the structure of biochemical reaction systems from data and describe our contributions to this problem using a Bayesian formulation. We close with our brief work demonstrating the application of stochastic biochemical reaction networks to the field of epidemiology along with some supporting software.

Lastly, we provide summary of our contributions and a few future directions.

Contents

Chapter 1

Introduction

Knowledge of the complex mechanisms behind biological processes is critical to the development of better drugs and therapies for numerous medical conditions and diseases. By identifying the pathways and interactions that govern functional biological systems, such as gene regulation and blood coagulation, biologists can create targeted treatments that can properly regulate systems while also being aware of possible downstream side effects. As an example, we can observe the coagulation cascade, shown in Fig. (**??**), which regulates the complex process of coagulating blood around a wound. Occasionally in trauma patients, this process is subject to critical failure, resulting in the life-threatening condition coagulopathy. However, as significant research efforts have demonstrated, the failure mechanisms of coagulopathy are multi-modal [23, 60], and explainable by both a lack of necessary clotting factors and an abnormally high rate of clot breakdown, which could potentially be impacted by the nature of the injury as well. This key knowledge manifests at the clinical level, as these distinct mechanisms, and the potential causes, can inform what assays to run and, ultimately, what treatments to administer at a moments notice.

Many of these inherently non-linear and dynamic biological processes can be math-

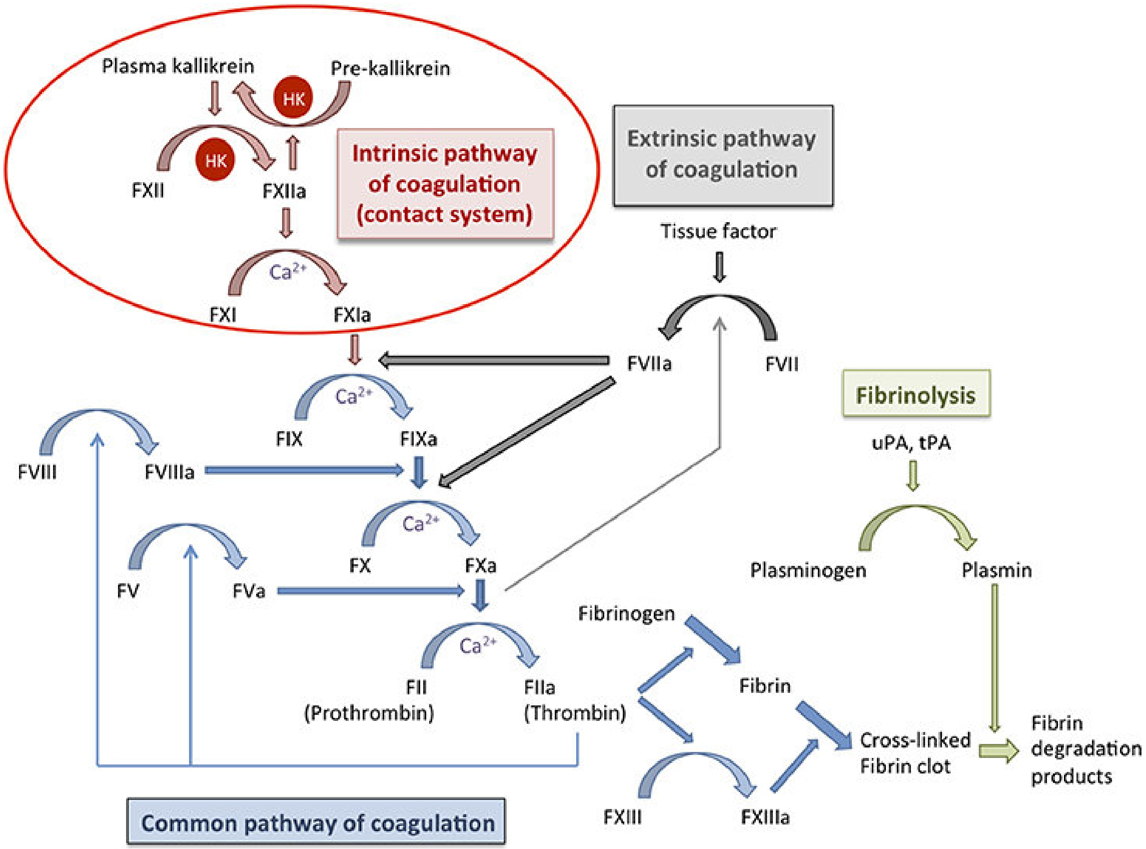


Figure 1.1: **The coagulation cascade and fibrinolysis [62].** The coagulation cascade is responsible for formation of a fibrin clot, while fibrinolysis is responsible for breaking down fibrin clots. Balance in the system is crucial for the regulation of overall coagulation.

ematically described using the framework of biochemical reaction networks [63]. These models allow researchers to specify how any number of biochemical species can interact via any number of reactions to produce interesting and unique effects. Furthermore, the flexibility of the framework allows for the dynamics to take into account both deterministic and stochastic effects, as recent research has demonstrated to be important [29]. However, while these tools allow scientists to describe potential hypotheses, a necessary step in building a complete understanding of a biological system is validating potential models to experimental data. With rapid advancements in better and cheaper measurement techniques, such as flouorescence-activated cell sorting (FACS) and single molecule fluorescence in situ hybridization (smFISH) [119], the ability for biologists to gather quality experimental data for many complex biological systems has greatly accelerated. Consequently, this brings to forefront the main motivation of this work, which is the increased necessity of powerful statistical methods to aid in drawing biological insights from these vast datasets.

In this work, we utilize Bayesian methods for the study and validation of biochemical reaction networks to experimental data. The powerful Bayesian method for statistical inference offers substantial benefits for systems biologists by providing an intuitive way to specify probabilistic models and incorporate prior knowledge while also providing uncertainty estimates around parameter estimates and predictions. Furthermore, parallel to the explosion of biological data, recent advances in computational algorithms [69] and software for statistical inference [12] has enabled the Bayesian framework to be applied to more complex problems. For stochastic biochemical reaction systems, yield intractable likelihood functions, Bayesian methods such as Approximate Bayesian Computation (ABC) have been gaining traction [110] due to the availability of the slow, but exact simulation algorithms.

At the time of this work, adoption of Bayesian methods in this field has been limited, due in part to the added complexity of these techniques as well as standing computational challenges. To this end, we demonstrate a improvements and applications to Bayesian methods for the study of biochemical reaction networks, largely motivated and enabled by the recent availability of high-throughput measurement technologies and the significant developments in statistical computation and Machine Learning. Our hopes are that these developments expand the capabilities and adoption of these powerful techniques to this very important and growing field.

1.1 Organization

The rest of this thesis will proceed as follows. In Chapter 2, we will discuss an application of Bayesian methods to study the effects of the coagulation biochemical reaction networks at a clinical level. In Chapter 3 and Chapter 4, we will present our work on novel Bayesian methods for the improving inference of the parameters and the structure of biochemical reaction networks. In Chapter 5, we will discuss our work demonstrating the application of the biochemical reaction networks to epidemiological models and data. We close with a summary of our work and a few future directions in Chapter 7.

Chapter 2

Bayesian Joint Survival Models for Quantifying Coagulopathic Survival Risk

Coagulopathy is a life-threatening condition experienced by many trauma patients who have lost large amounts of blood. The blood coagulation process can be described by a set of biochemical reactions that ultimately lead to formation of a clot at a specific injury site. Malfunctions in this complex biochemical network can lead to the failure to properly form and maintain a clot, commonly referred to as coagulopathy, especially among trauma patients. In this work, we focus on the clinical side of this reaction system and utilize joint survival models to investigate the survival risks of the time varying biomarkers D-Dimer, a protein fragment formed in the process of clot breakdown, and Factor-II, a protein consumed in clot formation, using assays from the ICU.

Joint survival modeling is an increasingly popular technique in the area of clinical data science, used to study the effect of longitudinally measured factors on outcomes. This Chapter covers the material in our work *Associations of Longitudinal D-Dimer and*

Factor II on Early Trauma Survival Risk [102]. We begin by providing some background on the coagulation system. Then we describe the technique of joint survival models. Lastly, we use the technique to study the early survival risk implied from longitudinal D-Dimer and Factor II levels.

2.1 Background

Coagulopathy (as defined here) is a condition in which blood fails to properly form robust clots. Following injury and shock from a major trauma, patients often become coagulopathic, coinciding with increased bleeding, higher resuscitation requirements and much higher rates of death [17, 18, 97]. However, despite the increased urgency for treatment, the complexity of the underlying coagulation system makes understanding and diagnosis of trauma-induced coagulopathy (TIC) extremely difficult, especially in a clinical setting with so much interpatient and intrapatient variability. The main objective of this study is to quantify the level to which markers of two possible mechanisms of TIC affect survival odds, accounting for patient variability, and to understand what this tells us about possible targets for intervention.

The Coagulation System and Coagulopathy

The standard model for the coagulation system consists of two distinct physical processes: coagulation (clot formation) and fibrinolysis (clot breakdown). Coagulation is the process by which a sequence of protein interactions ultimately leads to the formation of cross-linked fibrin clots, which physically block off a wound site [9]. To balance this process, fibrinolysis breaks down fibrin clots and produces fibrin degradation products, which are then flushed out of the system. Properly regulated, these two systems prevent excessive bleeding. A schematic is shown in Fig. 2.1.

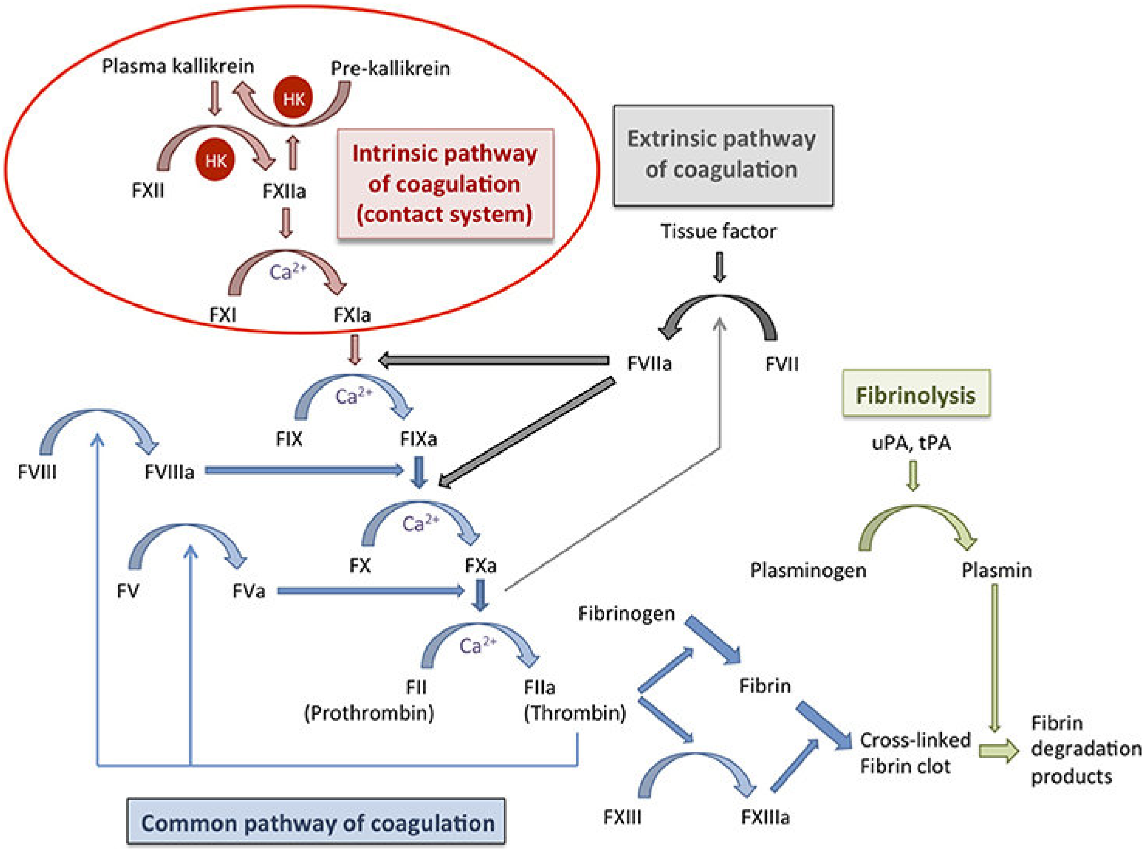


Figure 2.1: **The coagulation cascade and fibrinolysis [62].** The coagulation cascade is responsible for formation of a fibrin clot, while fibrinolysis is responsible for breaking down fibrin clots. Balance in the system is crucial for the regulation of overall coagulation.

Malfunctions in the coagulation system lead to the inability to form clots or to keep clots in place, resulting in excessive bleeding at the wound site. Several hypotheses exist to explain the driving factors of TIC [23, 60]. Two important coagolopathic conditions are consumptive coagulopathy and hyperfibrinolysis. Consumptive coagulopathy focuses on the inability to form fibrin clots, due to a lack of necessary pro-coagulants, while hyperfibrinolysis emphasizes the inability to keep a sufficient number of fibrin clots active due to overactive fibrinolysis. Though the mechanisms are different, both manifest as increased, uncontrollable bleeding at the wound, often through a complex interdependent mechanism.

In this study we used data collected from trauma patients to quantify how these two mechanisms may be realized in patient survival odds. We chose Factor II and D-Dimer as representative biomarkers of consumptive coagulopathy and hyperfibrinolysis respectively. Factor II, or prothrombin, is a protein that is converted into thrombin in the coagulation cascade [125]. Thrombin is the central protein in the coagulation cascade, responsible for forming fibrin clots and activating platelets to essentially seal a wound. On the other hand, D-Dimer is a fibrin degradation product created when plasmin breaks down fibrin clots. We fit a joint survival model to this data and examined the distribution of patient longitudinal curves and the hazards of both longitudinal covariates.

Methods

Dataset

Our dataset consists of severely injured patients admitted to the ICU at the UCSF Level I Trauma Center. Upon admission, age, sex, injury severity score, injury type, and the presence of a traumatic brain injury, in addition to many other measurements, were recorded. Blood draws were attempted for each patient at hours close to 0, 2, 3, 4, 6, 12, and 24 as measured from admission. The time and outcome of each patient was recorded post dispatch. From each blood draw, a variety of coagulation activity levels were measured, of which only the protein Factor II (% activity) and the protein fragment D-Dimer ($\mu g/ml$) were used in this analysis, for the aformentioned reasons. Blood assays were conducted using the Stago Compact Analyzer (Diagnostica Stago, Parsippany, NJ) according to manufacturer instructions. For D-Dimer, the upper limit normal value is ~0.5 $\mu g/ml$ [117] while for Factor II standard operating range falls within $50\% - 200\%$ activity. The hour 0 measurements of most patients fell within these values though with a slight skew due to the nature of the dataset. Patients with no Factor II or D-Dimer measurements were omitted. Post pre-processing, a total of 891 patients remained with 2062 longitudinal observations. In this work we define outcome as survival at hour 25, which is on the order of when deaths from TIC are most prevalent [52]. Past this window, many patients die from other causes such as sepsis. From a survival analysis perspective, patients were considered censored if death was not recorded within the observation window. A summary of the distributions in the data are presented in Table 2.1.

2.1.1 Statistical Model

To uncover the effects of Factor II and D-Dimer on early trauma survival, we employ a joint survival model [86, 111, 54]. Joint survival models relate the effects of time-dependent covariates, such as measured clinical biomarkers, on time-to-event data, such as death, accounting for irregular measurement times and intrinsic measurement variability. Recently, joint survival models have been used to study survival in a variety of other diseases [81, 99]. In particular, they have gained prominence due to their ability to robustly model how the continuous evolution of biomarkers affects survival. In the following, we describe the two subcomponents of the joint model: the longitudinal submodels and the survival submodel. We note that for applying this model, we first apply the natural log to values of D-Dimer and henceforth refer to this quantity as log D-Dimer.

Longitudinal Submodels

The longitudinal submodels describe how each time-dependent covariate evolves over the observation window. By explicitly specifying the form, as opposed to naively imputing values, we can account for measurement variability when associating the covariate to the

Characteristic	Estimate
Total number of individuals	891
Death within 24h, n (%)	61 (6.8%)
Sex, n (%)	
Male	728 (81.7%)
Age, n (%)	
$\geq 15, < 20$	58 (6.5%)
$\geq 20, < 30$	286 (32.1%)
$\geq 30, < 40$	170 (19.1%)
$\geq 40, < 50$	127 (14.3%)
$\geq 50, < 60$	115 (12.9%)
$\geq 60, < 70$	62 (6.9%)
$\geq 70, < 80$	42 (4.7%)
≥ 80	31 (3.5%)
Injury Severity Score, n (%)	
$\geq 0, < 10$	344 (38.6%)
$\geq 10, < 20$	168 (18.9%)
$\geq 20, < 30$	192 (21.5%)
$\geq 30, < 40$	108 (12.1%)
$\geq 40, < 50$	25 (2.8%)
$\geq 50, < 60$	38 (4.3%)
≥ 60	16 (1.8%)
Trauma Type, n (%)	
Penetrating	385 (43.2%)
Blunt	506 (56.8%)
Traumatic Brain Injury, n (%)	343 (38.5%)

Table 2.1: **Characteristics of Cohort**

survival outcome. This has been shown to reduce bias in estimates [111] compared to traditional treatments of time-dependent covariates in survival models.

Let $y_{ij}(t)$ denote the measured activity level of coagulopathic biomarker j for patient i at time t. For our study, the longitudinal biomarkers Factor II and D-Dimer are modeled using generalized linear mixed effects models with grouping at the individual level. Specifically, we set $y_{ij}(t) \sim \mathcal{N}(\eta_{ij}(t), \sigma_F)$, with

$$\eta_{ij}(t) = \beta_{0j} + \beta_{1ij} + \beta_{2j}t + \beta_{3ij}t + \beta_{4l_ij}t + \sum_k \beta_{5jk}x_{ik},$$

the expected value of the respective marker at time t for patient i, and σ_F the estimated standard deviation. In this formulation, β_{0j} and β_{1ij} denote the population and individual level intercepts and β_{2j} and β_{3ij} denote the population level and individual level slopes. β_{4jk} specifies the effect of the $k-th$ fixed covariate on the $j-th$ time-dependent biomarker. The included fixed covariates are age, sex, injury severity score, traumatic brain injury, and injury type, selected due to their relevance in other studies in this area. This is equivalent to fitting a regression line to each of the coagulopathic biomarkers.

Survival Submodel

The survival submodel connects the longitudinal submodel to the observed patient outcomes. We use the standard proportional hazards model. For each patient, we have a tuple (T_i, D_i) indicating the time that the patient died or was censored and the binary outcome of death. Let $h_i(t)$ be the hazard function for the i-th patient at time t,

$$h_i(t) = h_0(t) \exp\left(\sum_k \gamma_k x_{ik} + \sum_j \alpha_j \eta_{ij}(t) \right),$$

with $h_0(t)$ the baseline hazard function, α_j the coefficient indicating the strength of the association between longitudinal covariate j and survival, and γ_k the strength of the association between fixed covariate k and survival. The baseline hazard $h_0(t)$ was selected to be a 6-th order B-Spline, as this choice offers maximum flexibility in fitting the unique survival curves of subgroups while avoiding overparameterization [8]. This hazard function at time t can be interpreted as the instantaneous rate at which the subject accumulates hazard toward the outcome, assuming that they have survived up to time t. Compared to standard time-dependent survival models, the hazard function depends on the expected value of the time dependent covariate, as opposed to the observed or imputed value. The hazard function is linked to the time of the outcome via the survival function

$$S(t) = P(T_i \geq t) = \exp\left(-\int_0^t h_i(x)dx\right).$$

For interpretation, we observe the association strengths, α, which indicate the change in survival odds for every unit change in the covariate.

Estimation

We use the rstanarm package [98, 12] and the joint model function to obtain a Bayesian fit for our model to the data. To estimate the patient-level effects in the longitudinal covariates, we use hierarchical priors to induce shrinkage in the case of few observations [34]. Posterior predictive checks were performed on the longitudinal trajectories to verify that the resulting fit were consistent with the observed data and convergence metrics were checked to validate that the chains were consistent. Analysis was performed using 4000 posterior draws over 4 chains.

2.2 Results

Factor II and D-Dimer Trajectories

In Table 2.2 and Table 2.3 we show the estimated fixed effect coefficients for Factor II and log D-Dimer in the longitudinal submodels. Traumatic brain injury is tied to significantly higher levels of both covariates while penetrating injuries tend to decrease the predicted log D-Dimer levels. Higher injury severity score and age slightly increase the level of log D-Dimer and decrease the level of Factor II. These effects indicate that older and more severely injured patients have higher D-Dimer and lower Factor II, which would be intuitive as they indicate higher levels of fibrinolysis and lower levels of available pro-coagulants. Penetrating injuries provide an uncertain effect on Factor II but are associated with lower levels of log D-Dimer. At the population level, Factor II tends to decrease over time while D-Dimer tends to increase. For healthy patients, these would be the expected patterns as clotting factors are used and fibrin degradation products are produced. Fig. 2.2 and Fig. 2.3 show the estimated mean Factor II and log D-Dimer trajectories for 4 patients. Crucially for diagnosis, the 4 patients show varying individual behavior but also reversion to the population level distribution in the case of patients with few measurements.

		Factor II
	coefficient	95% credible interval
intercept	84.03	80.04, 88.04
slope	-0.23	-0.30, -0.17
age	-0.15	-0.21, -0.10
sex (ref: Male)	-1.58	-4.07, 0.90
injury severity score	-0.37	-0.44, -0.30
traumatic brain injury (ref: Yes)	2.65	0.18, 5.07
penetrating injury (ref: Yes)	-0.28	-2.73, 2.10

Table 2.2: **Coefficients for Longitudinal Factor II**

Factor II and log D-Dimer Associations with Survival

Estimated association strengths, interpreted as the increase in odds for every unit increase in the biomarker, as well as 95% credible intervals are shown in Table 2.4. For exogenous covariates, we find minimal evidence that a higher initial injury severity score and age increases the risks of death. The large uncertainty in the gender hazard ratio is

	log D-Dimer	
	coefficient	95% credible interval
intercept	-0.64	-0.41, -0.18
slope	0.011	0.008, 0.016
age	0.008	0.004, 0.011
sex (ref: Male)	-0.20	-0.34, -0.06
injury severity score	0.043	0.039, 0.047
traumatic brain injury (ref: Yes)	0.44	0.30, 0.59
penetrating injury (ref: Yes)	-0.25	-0.38, -0.11

Table 2.3: **Coefficients for Longitudinal log D-Dimer**

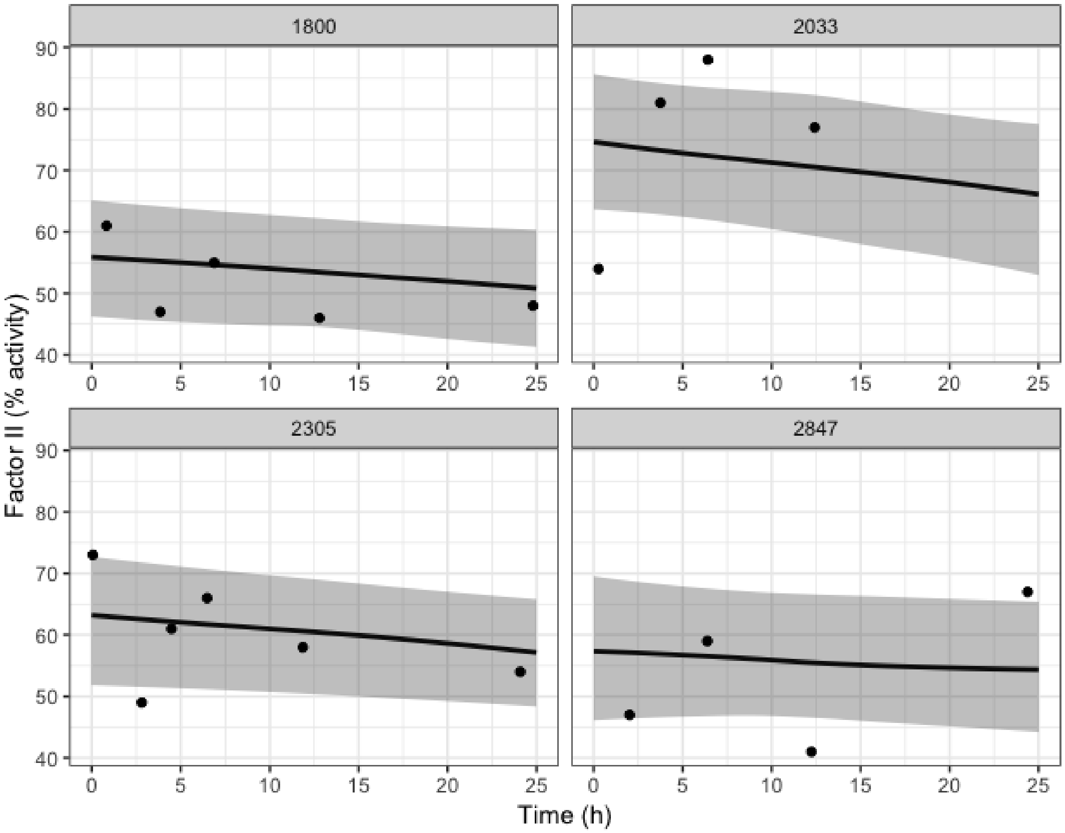


Figure 2.2: **Sample Factor II Patient Trajectories.** Each plot shows a random patient along with their estimated mean trajectory for Factor II and confidence intervals of the mean trajectory. Scattered points are observed data points.

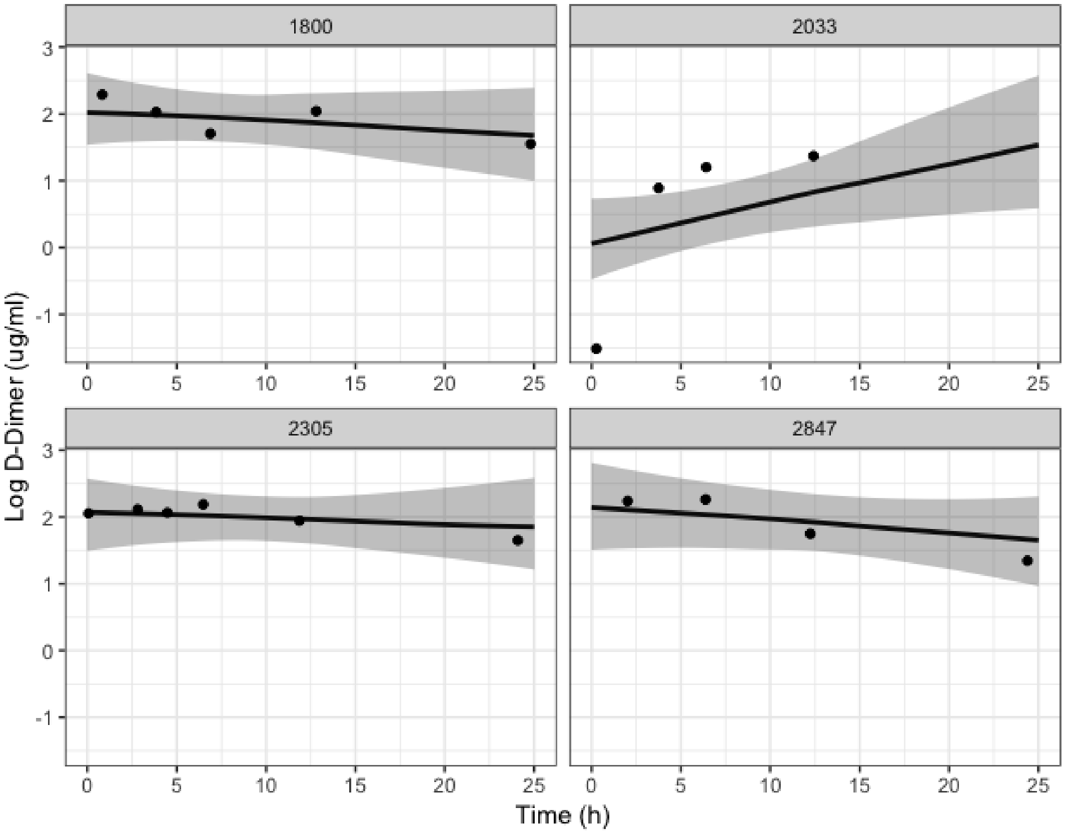


Figure 2.3: **Sample log D-Dimer Patient Trajectories.** Each plot shows a random patient along with their estimated mean trajectory for log D-Dimer and confidence intervals of the mean trajectory. Scattered points are observed data points.

likely due to an insufficient sample size of women in the dataset. As previously known, we find that traumatic brain injury has an extremely large effect on the risk of early death. Interestingly, penetrating injuries seem to significantly increase the risk of early death (hazard ratio [6.08, 3.37 - 11.19]), however, the large credible intervals indicate a relative lack of data for patients who ultimately died.

For the longitudinal coagulopathic covariates, we find that unit increases in log D-Dimer significantly increase the risk of early death (hazard ratio [2.22, 1.57 - 3.28]). At the same time, unit increases in Factor II only marginally decrease the risk of death (hazard ratio [0.94, 0.91 - 0.96]) but with high certainty. This is in good agreement with

	Hazard Ratios	95% credible interval
Factor II	0.94	0.91, 0.96
log D-Dimer	2.22	1.57, 3.28
age	1.02	1.01, 1.03
sex (ref: Male)	0.76	0.45, 1.32
injury severity score	1.03	1.01, 1.04
traumatic brain injury (ref: Yes)	2.71	1.51, 5.04
penetrating injury (ref: Yes)	6.08	3.37, 11.19

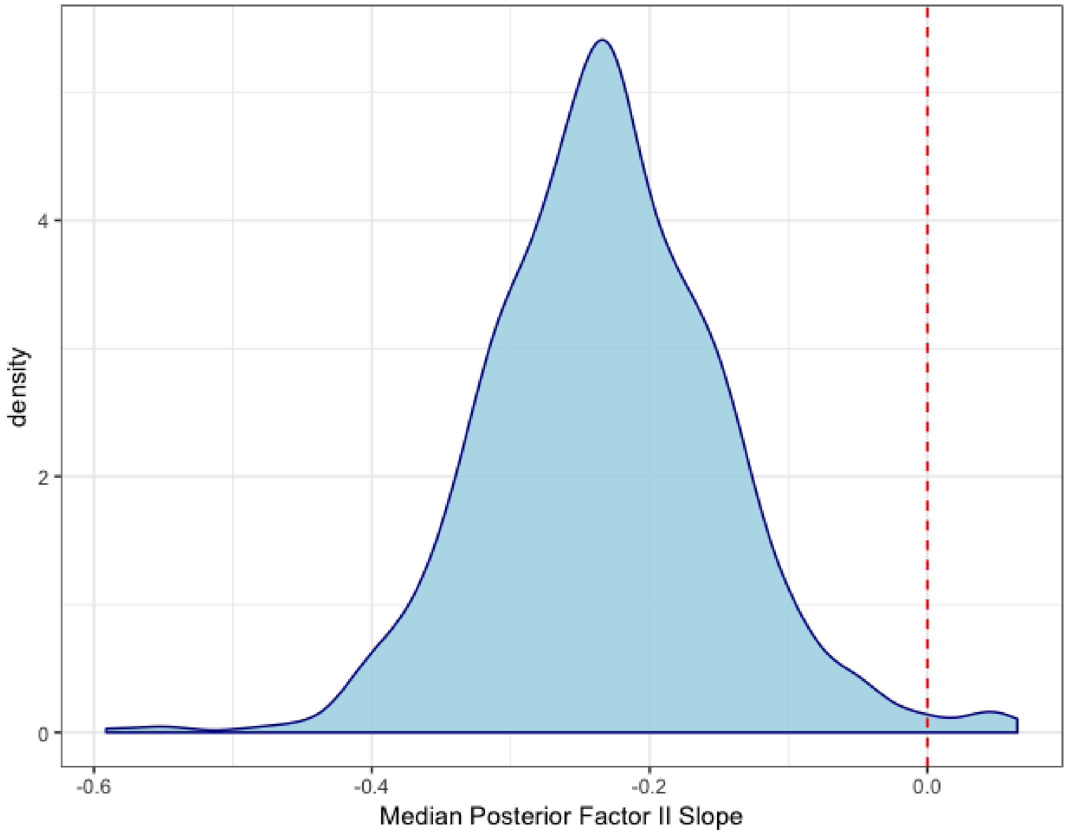
Table 2.4: **Median and 95% credible interval for Hazard Ratios**

[47], which concludes that high log D-Dimer levels are the more definitive predictor of death regardless of fibrinogen levels. The significant effect of log D-Dimer suggests that maintaining or lowering the rate of fibrinolysis and thus D-Dimer generation is a key component in reducing the risk of early death in a hospital setting.

Variation Among Longitudinal Trajectories

In addition to associations, our model estimates individual trajectories for each patient. In this cohort, the vast majority of patients gradually decrease in Factor II levels over the 25 hour window, as shown by the distribution of median slopes in Fig. 2.2. The relatively low rate seems to indicate that, for the majority of patients, Factor II is being held relatively consistent in this 25 hour window. We see no cases where the model indicates that Factor II is being consumed at a significantly large rate. In comparison, for log D-Dimer trajectories, we observe large variation from expected behavior. As shown in Fig. 2.5, patients are centered around 0 but have significant probability mass at both increasing and decreasing D-Dimer levels. However, as D-Dimer is only a product of fibrinolysis, it is difficult to predict what a traditionally healthy trajectory would consist of.

The level of variation among patients, controlling for all of the fixed covariates, in-



dicates that these biomarkers are likely subject to some unaccounted for patient level variability. From a treatment perspective, the varying trends among patients indicate that when making risk assessments for a particular patient, it is important to understand both the estimated hazard ratio as well as the projected trajectory of their biomarkers. As an example, if a patient exhibits high D-Dimer but it is seemingly decreasing, perhaps treatment for fibrinolysis is not necessary.

Figure 2.4: **Distribution of Median log D-Dimer Slopes** Estimated median slopes of log D-Dimer for each patient. Red, dashed line indicates the zero-line

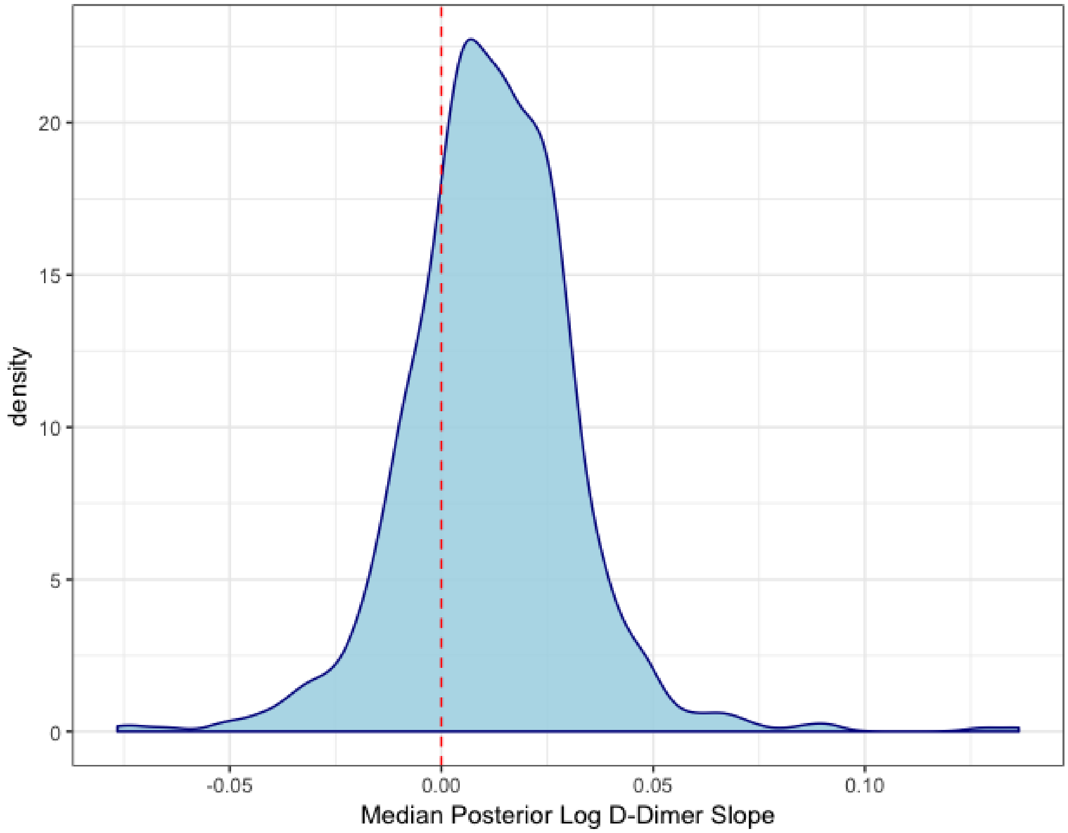


Figure 2.5: **Distribution of Median Factor II Slopes** Estimated median slopes of Factor II for each patient. Red, dashed line indicates the zero-line

2.3 Discussion

Hazard Ratios

The hazard ratios indicate that, in this cohort, unit increases in Factor II levels only marginally increase survival odds, while a doubling of D-Dimer (due to the log transformation) largely affects survival odds. Observing the data, a doubling of D-Dimer is not uncommon. Thus, although both consumptive coagulopathy and hyperfibrinolysis do seem to affect survival in some regard, increased rates of fibrinolysis are much more likely to be damaging to survival. From this perspective, greater benefit would be gained by controlling hyperfibrinolysis rather than further managing or increasing factor availability levels.

High levels of D-Dimer have often been associated with poor patient outcomes as a proxy for hyperfibrinolysis. This is further consistent with the growing literature which indicates the importance of addressing hyperfibrinolysis in TIC. Hyperfibrinolysis is estimated to occur in a large number of trauma cases, often with significantly higher mortality rates [92]. The relatively low, but positive impact of Factor II levels suggests that managing Factor II levels is not a significant problem in this cohort. Indeed, the importance of coagulation consumption has long been studied [58] and linked to poor outcomes. As this is a single hospital, retrospective study, it is possible that monitoring and treatment for factor depletion is better monitored and maintained, leading to better outcomes for patients that exhibit signs of coagulopathy.

Also of note is the significant effect of both traumatic brain injury and penetrating trauma, independent of the levels of both Factor II and D-Dimer. Largely, the increased mortality from injuries of this types are well known in trauma [109]. The scale of the hazard ratios provides a rough perspective on the priority of treatment, with concern based on the the type of injury preceding further monitoring of hyperfibrinolysis and consumptive coagulopathy.

D-Dimer Modulation in Trauma Care

D-Dimer, while often used as a surrogate for measuring fibrinolysis, can also be affected by other factors. Due to the risk associated with high levels of D-Dimer as indicated by our model, it is important to further describe some of these alternative factors.

From a physiological perspective, as D-Dimer is a protein fragment created from the breakdown of a fibrin clot, any processes which effect the rate of clot breakdown could result in measured changes in D-Dimer levels. In ICU patients, while initially elevated levels of D-Dimer are expected due to the nature of the injuries involved, the type of injury can have a significant effect on D-Dimer levels over time, as the condition of the patient evolves. As seen in Table 2.4, patients with non-penetrating or traumatic brain injuries tend to see an increase in D-Dimer levels over time. Typically, for healthy patients recovering from injury, coagulation and fibrinolysis are expected to slow down, resulting in declining D-Dimer levels. Sathe et al. [90] further mentions several non-hyperfibrinolytic pathological and non-pathological conditions which have also been shown to increase D-Dimer. An important possibility that may affect D-Dimer levels without clearly indicating increased fibrinolysis is the decreased ability to clear D-Dimer from the blood, as has been found in patients with liver disease and cirrhosis [45]. In these cases, an underlying liver problem may result in abnormally high levels of D-Dimer as it accumulates over time, even when the patient is not hyperfibrinolytic.

Common interventions may also cause D-Dimer levels to change. A recent standard treatment for hyperfibrinolysis is administration of the anti-fibrinolytic drug Tranexamic acid (TXA). As its mechanism of action is to prevent plasmin formation and thus slow down fibrinolysis, it naturally decreases D-Dimer levels. This has been demonstrated in both laboratory and clinical settings [87, 75] with time-scales as short as 30 minutes after administration [93].

Clinical Considerations

Our findings broadly suggest that, from an early clinical perspective, managing fibrinolysis is typically more of a concern than managing consumptive coagulopathy over a 24 hour window of care. Furthermore, as shown in Fig.2.2 and Fig.2.5, the trends of these factors can vary significantly between different patients and thus treatment and evaluation of patient state can possibly improve by projecting how a patient's state is trending. This follows exactly the thinking of the clinician, where they are constantly evaluating the current physiologic/biologic state of a patient and trying to predict and modify the trajectory. Although high D-Dimer levels are linked to poor outcomes, if the patient is projected to be improving, further treatment may not be necessary. The development of explicit risk metrics which provide individual projected trajectories as such could provide valuable information in acute decision making.

As the factors analyzed in this work are not typically measured in real-time, our work primarily aims to explore the risk factors in TIC and to observe patient-level variations over their ICU stay. State of the art treatment of TIC in the ICU typically includes providing blood products such as crystalloids, fresh frozen plasma, and packed red blood cells through transfusion and by administering drugs such as Tranexamic acid [55, 87] both of which aim to control hyperfibrinolysis as well as consumptive coagulopathy. A significant amount of recent research has focused on implementing better protocols for these interventions using viscoelastic assays, such as TEG and ROTEM [43, 42]. These measurements aim to provide a more holistic picture of blood coagulation, which can lead to significant advantages in accuracy or diagnosis of coagulation malfunctions. Additionally, in the future, we expect that results can be extracted at the point-of-care and used for a truly precision medicine individualized approach to diagnosis and treatment.

The use of viscoelastic measurements in a similar computational study can extend the conclusions of this work to more precisely capture malfunctions of the coagulation system as well as provide for a practical component in a dynamic risk-prediction system that can aid in acute decision making over a patient's stay.

Model Limitations

Importantly, there are a few limitations to full interpretation of this model. Due to the retrospective and single hospital nature of the data, these results can be understood more as an evaluation of early trauma hospital protocol. As interventions such as mass transfusion are not accounted for, from this perspective we find that the trauma protocol mediates the effects of most covariates but does not seem to adequately control for the effects of increasing log D-Dimer levels. To improve interpretation, we would need to utilize data from multiple hospitals. Furthermore, certain studies indicate that elevated log D-Dimer is not necessarily a definitive sign of hyperfibrinolysis [85] and can be rather thought of as a confounded measure of injury severity and the need for an activated coagulation system. Thus, utilization of viscoelastic assays, such as TEG and ROTEM, that offer different measurements may help to better distinguish the effect of the two components of coagulation on survival. Despite this however, our data show that D-dimer, whatever its biologic interpretation (fibrinolysis or enhanced clot breakdown) is an important predictor of future mortality. Similarly, use of other proteins in the coagulation cascade may reveal more informative results with respect to how much of an impact consumptive coagulopathy over time actually has on survival odds. A secondary model for interventions may also help for improving treatment for TIC.

Conclusions

We fit a joint-survival model to trauma to quantify the effect of activity levels of Factor II and log D-Dimer on survival in an early 25 hour window. From this work, we find that increases in Factor II levels have a small, but positive effect on survival, while increases in log D-Dimer levels have a large negative effect on survival. The nature of this study suggests further investigation into methods to prevent excessive fibrinolysis in hospital protocol. Furthermore, this model can also be used to better understand individualized and dynamic risk prediction from a standard patient, due to the large variability in patient longitudinal trajectories.

Chapter 3

ABC Summary Statistics for Stochastic Biochemical Reactions via Approximate Simulators

Moving to the molecular scale, many biochemical processes exhibit intrinsic stochasticity [29], which manifests as significant variation of the same process across different cells. Mathematically, this can be captured and studied using the framework of stochastic biochemical reaction networks, first described in [37]. Developments in modern measurement techniques such as sMFISH and FACS has lead to significant research effort on inference of the parameters of these stochastic models to experimental data. However, due to the complexity of these models and the inability to evaluate the likelihood function, this is an incredibly challenging task.

In this chapter, we present our work *Accelerated Regression-Based Summary Statistics for Discrete Stochastic Systems via Approximate Simulators* [105], focusing on accelerating Approximate Bayesian Computation (ABC) for Bayesian parameter estimation of these stochastic biochemical models. Specifically, we introduce our method of utilizing approximate simulators for the important step of training summary statistics used within ABC. Furthermore, we detail our proposed method to overcome the potential bias introduced when the approximate simulator is poor. The benefits of our approach are illustrated using a few example stochastic biochemical models and some implementation details, and potential practical pitfalls are explained.

3.1 Background

In recent years, stochasticity has been shown to play a crucial role in many molecular biological processes such as genetic toggle switches [29, 66] and robust oscillators [114]. In many cases, systems biologists will model these stochastic biochemical reaction networks using continuous-time, discrete-space Markov Chains [38], which allow one to capture stochasticity in a system caused by the limited availability of certain reactants, such as transcription factors. A critical step in building an accurate mechanistic model of these stochastic systems is calibrating the kinetic rate constants to experimental data. While efficient methods exist for parameter estimation using maximum likelihood or Bayesian inference for similar models, for these discrete stochastic models, the intractability of the likelihood function forces researchers to rely on the growing class of Likelihood-Free Inference (LFI) methods [21, 91, 116], which depend only on the availability of a model simulator. Recently, Approximate Bayesian Computation (ABC) [96, 22] has become one of the most popular LFI methods for discrete stochastic models due to its simplicity and demonstrated effectiveness.

Approximate Bayesian Computation

Given a prior over parameters $p(\theta)$ and a stochastic simulator $p(X|\theta)$, Approximate Bayesian Computation (ABC) approximates the posterior distribution $p(\theta|X) \propto$

$p(X|\theta)p(\theta)$ using only forward simulations and without computing the likelihood [96]. The basic Rejection ABC is presented in Algorithm 1.

Algorithm 1: Rejection ABC

```
Input: simulator model $p(\theta, X)$, distance function $d$, observed dataset $X_o$,
       tolerance $\epsilon$, $N$ posterior samples
Output: $\{\theta_i\}_N \sim p(\theta|X)$
samples = {};
while length(samples) < $N$ do
    Sample from the prior $\theta \sim p(\theta)$;
    Draw a simulation $X \sim p(X|\theta)$;
    if $d(X_o, X) < \epsilon$ then
        samples = samples $\bigcup \theta$
    end
end
```

When X is high dimensional, comparing exact trajectories often results in very low acceptance rates due to the curse of dimensionality. For this reason, it is standard practice to trade bias for efficiency by first reducing the dimensionality of X using a set of *summary statistics*, $S(X)$, and subsequently comparing trajectories using $d(S(X_o), S(X))$, where d is a user selected distance function. This can lead to much higher acceptance rates, however, selection of an appropriate $S(X)$ for any given model can be difficult.

Regression-Based Summary Statistics

The performance of ABC is highly dependent on having an effective set of summary statistics for the experimental data, which becomes increasingly difficult for domain experts to hand-select as the dimensionality of the problem grows [96]. For stochastic biochemical reaction networks, where data is often in the form of sample paths of molecular species over time, this is a common issue due to complexity of trajectories where simple means and correlations may not effectively capture the features. For this reason, significant focus has recently been given to the automatic learning of summary statistics from model simulations, which we will refer to as regression-based summary statistics.

Fearnhead and Prangle [30] formulate the problem of regression-based summary statistics for ABC as a least squares estimation of the posterior mean:

$$S(X) = \mathbf{E}[\theta|X] = f_\Phi(X) \tag{3.1}$$

$$\theta|X \sim \mathcal{N}(f_\Phi(X), 1) \tag{3.2}$$

$$\theta = f_\Phi(X) + \epsilon, \tag{3.3}$$

where f_Φ is an arbitrary expressive function and ϵ is standard normal noise. The parameters of f_Φ are fit using maximum likelihood on a simulated dataset $\mathcal{D} = \{(\theta_0, X_0) \cdots (\theta_N, X_N)\}$ drawn from the model $p(\theta, X)$. While initially proposed as a linear $f_\Phi(X)$ for each parameter, nonlinear Neural Network architectures have shown promise in producing accurate results [57]. For discrete stochastic models, Akesson et al. [3] show that Convolutional Neural Networks (CNNs) tend to outperform other architectures. The general procedure for this is detailed in Algorithm 2.

Algorithm 2: Constructing a Summary Statistic

```
Input: prior $p(\theta)$, SSA simulator $p(X|\theta)$, $N$ samples
Output: calibrated $S(X) = f_\Phi(X)$
samples = {};
for $i = 1 : N$ do
    Sample from the prior $\theta \sim p(\theta)$;
    Draw a trajectory $X \sim p(X|\theta)$;
    samples = samples $\bigcup$ $(\theta, X)$
end
Train $S(X) = \mathbf{E}[\theta|X] = f_\Phi(X)$ using samples.
```

A major bottleneck of regression-based summary statistics is their requirement to first draw a large number of simulations N to train accurate summary statistics. For discrete stochastic models which rely on expensive simulators such as Gillespie's stochastic simulation algorithm (SSA) [37] for generating exact trajectories, this step introduces a significant overhead in ABC. Fortunately, many faster approximate simulators exist for biochemical reaction networks, such as Ordinary Differential Equations (ODEs) in the form of the reaction rate equations (RRE), the Chemical Langevin Equation (CLE)[39], or τ-Leaping [84]. However, training a regression-based summary statistic using an approximation will inevitably lead to bias due to the unknown approximation error as the summary statistics will learn incorrect features.

In this work we propose to use data driven machine learning models to train approximate summary statistics for discrete stochastic models using a mix of samples from an approximate simulator and the SSA. This is done with the aim of significantly lowering the computational cost while also mitigating the potential introduced bias in a black-box way. The key insight used for this is that, although the quality of an approximate simulator can vary significantly as we move around parameter space, in many parts it is sufficiently accurate, but also often unknown. To take advantage of this, we train an approximate ratio estimator to inform when the approximation is significantly different and thus when we need to simulate using the SSA to prevent bias. In the following, we demonstrate the ability for our algorithm to effectively reduce the number of expensive SSA calls made, while maintaining accuracy of the learned summary statistics.

3.2 Results

Approximate Summary Statistics Overview

The goal of our algorithm is to reduce the computational cost of constructing a set of regression-based summary statistics for ABC by leveraging the availability of a single approximate simulator. This is accomplished by our algorithm in two major steps. First, a ratio estimator is trained to distinguish between approximate and SSA trajectories using M samples from both simulators. Next, to train the summary statistic, $N - M$ additional samples from the approximate simulator are drawn and passed through the ratio estimator. If the ratio estimator falls below a certain threshold, indicating that it is significantly different than the true model, we resample it using the full simulator, preventing unnecessary resamples from the costly SSA. For complete details see ”Methods” and Algorithm 3.

Experiments

To assess the computational savings of our method, we evaluate our method on four discrete stochastic models of varying complexity and compare to the baseline Algorithm 2 which uses no approximate simulations. We report the total number of SSA calls used to train a summary statistic as opposed to wall clock time due to the highly parallelizable nature of the problem. The baseline method utilizes N SSA calls but produces the most accurate summary statistic by definition. Accuracy of the resulting summary statistic is evaluated using normalized posterior mean absolute error $E_{\%}$ [3] on a large hold out test set of SSA trajectories. We briefly explain $E_{\%}$ in the following section.

For each experiment, we denote X for trajectories that are simulated from SSA and $\tilde{X}$ for trajectories simulated from the approximation. Each trajectory is also labeled with $Y = \{0, 1\}$ where $Y = 1$ indicates that the trajectory came from the SSA simulator and $Y = 0$ indicates that the trajectory came from the approximate simulator. Errors are reported using 15 replications of training and evaluation. We also plot the predictions of the trained approximate ratio classifier on samples drawn from the approximate simulator for each experiment. The output of this is interpreted as the probability, under the trained ratio estimator, that the approximate trajectory $\tilde{X}$ at θ came from the SSA model, $P(Y = 1|\tilde{X}, \theta)$. Values near 0 or 1 inform the decision to resample using the SSA model, as we know that the true class label for $\tilde{X}$ is $Y = 0$. Probabilities near 0.5 indicate that the ratio estimator cannot distinguish between SSA and approximate samples and the approximate does not need to be resampled. Complete details for each experiment can be found in the Supplementary Materials.

Normalized Posterior Mean Absolute Error $E_{\%}$

We evaluate the performance of our experiments using the normalized posterior mean absolute error $E_{\%}$ [3], which is defined as,

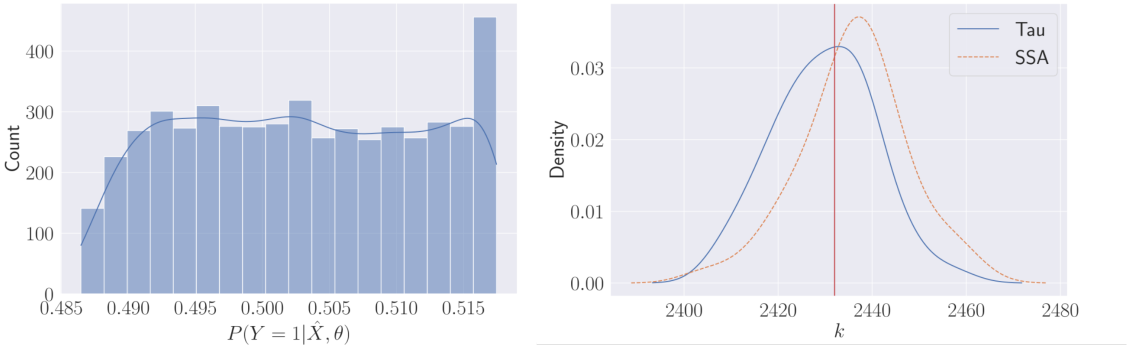
$$E_{\%} = \frac{\mathbb{E}_{\theta \in p(\theta)}|\theta - \hat{\theta}|}{\mathbb{E}_{\theta \in p(\theta)}|\theta - \bar{\theta}|}.$$

In this setup, $\hat{\theta}$ is the posterior mean and $\bar{\theta}$ is the prior mean. This quantity can be approximated for a uniform prior $U(a, b)$ over a set of N test points as

$$E_{\%} \approx \frac{4}{b-a}\frac{1}{N}\sum_{i=1}^{N}|\theta_i - \hat{\theta}_i|,$$

where $\hat{\theta}_i$ is obtained using the regression based summary statistic, which is trained to predict the posterior mean.

$E_{\%}$ aims to quantify the information gained in the posterior distribution. A value of $E_{\%} = 1$ indicates no information gained while values of $E_{\%} < 1$ indicate relative accuracy improvements. The true value of this quantity depends on the informativeness of observations, which is unknown in general for most problems. For this reason, the quality of different summary statistics are compared relative to each other under the assumption that the SSA trained summary statistic is maximally informative and the ground truth.



(a) Ratio Estimates

(b) Sample Posterior

Figure 3.1: **Calibrated Ratio Estimates for the Pure Birth Process (a)** The trained ratio estimator captures that the τ-Leaping approximation is exact, assigning a probability of 0.5 to all samples. **(b)** The posterior from both summary statistics captures the ground truth.

Pure-Birth Process

The Pure-Birth Process, or homogenous Poisson Process, is a trivial example where the likelihood is tractable and the τ-Leaping approximation produces exact trajectories for all parameter values. In a biochemical system, the pure-birth process represents the spontaneous generation of a molecular species at a fixed rate, which, while simplistic by itself, is often a fundamental component in more complex models. The model is described by a single parameterized reaction:

$$\phi \xrightarrow{k} S,$$

with initial condition of $S_0 = 0$. We assign a wide uniform prior $k \sim \mathcal{U}(0, 10000)$ and observe the process at times $t = \{0 : 100 : 1\}$. Though trivial, this example explores the ability to learn the correct approximate ratio-estimator, which should always predict around 0.5 due to the exactness of the approximation.

Figure 3.1a. shows the output of the approximate ratio estimator trained on only

$M = 300$ samples from the parameter space and evaluated on 5000 samples from the approximate model. The concentration around 0.5 indicates that the ratio estimator is able to detect that the two models evaluate the same likelihood. Indeed, in Figure 3.1b, we see that the posterior distributions using summary statistics trained via only τ-Leaping samples or only SSA samples are effectively the same. In this situation, we use a very small amount of samples from the SSA model to build the ratio estimator but otherwise rely entirely on the τ-Leaping approximation for no loss in accuracy.

Lotka-Volterra Stochastic Oscillator

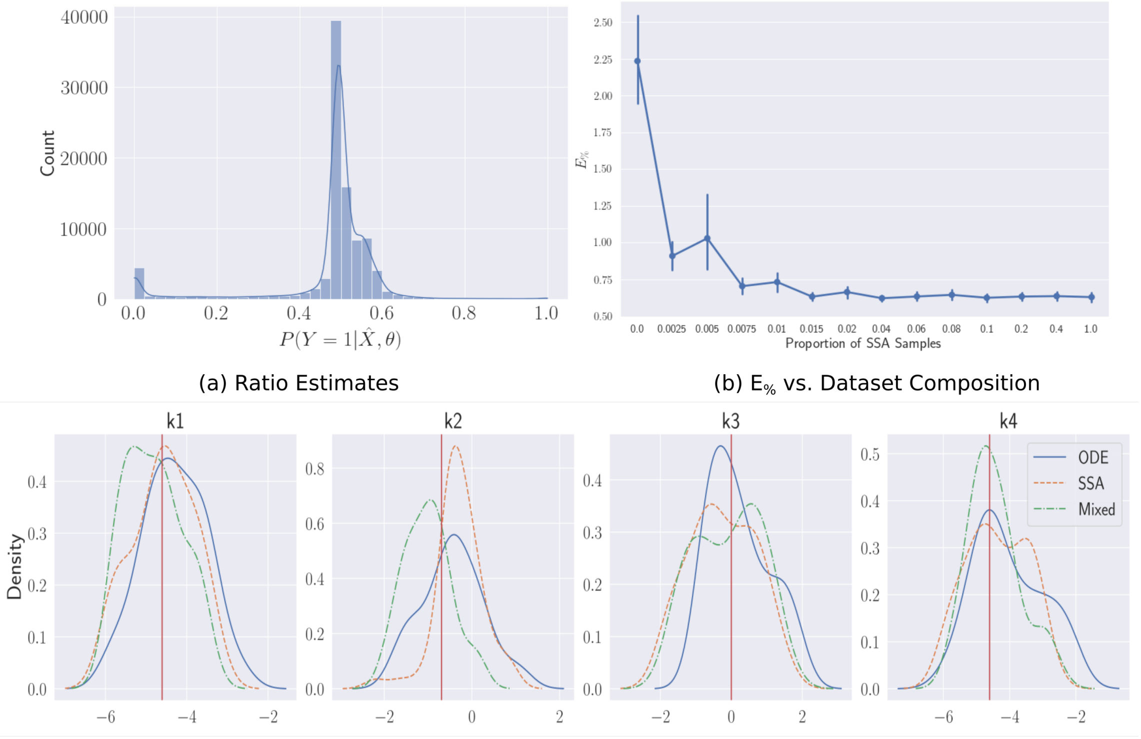
A more challenging and commonly used test problem is the Lotka-Volterra stochastic oscillator. This model describes predator-prey population dynamics and can be modelled as a discrete stochastic system. The system is specified via the following set of reactions:

$$S_1 + S_2 \xrightarrow{k_1} 2S_1 + S_2 \qquad\qquad S_1 \xrightarrow{k_2} \phi$$

$$S_2 \xrightarrow{k_3} 2S_2 \qquad\qquad S_1 + S_2 \xrightarrow{k_2} S_2$$

with initial populations of $S_1(0) = 50, S_2(0) = 100$. We assign the same priors and observation frequency as [73] and select a deterministic ODE as our approximating simulator. A key characteristic of this model is that, over the specified prior, only a small region of parameter space leads to consistent oscillations in both the ODE and the SSA models. In most other regions, population explosions are the typical behavior. We train the ratio estimator using $M = 3000$ samples and train the summary statistic with $N = 10^5$ samples. $E_{\%}$ is evaluated using 300000 hold out SSA test samples.

As shown in Figure 3.2a, the trained ratio estimator assigns significant mass around 0.5 but with heavy tails, suggesting that some proportion of samples should be resampled using SSA for better accuracy. Figure 3.2b shows the sensitivity of $E_{\%}$ as we increase



(a) Ratio Estimates

(b) $E_{\%}$ vs. Dataset Composition

(c) Posterior Marginals

Figure 3.2: **Trained Ratio Estimates for the Lotka-Volterra Stochastic Oscillator (a)** The trained $P(Y = 1|\tilde{X}, \theta)$ for the Lotka-Volterra easily classifies many cases, indicated by the peak at the left tail, but remains uncertain for the majority. **(b)** As the proportion of included SSA calls increase using the ratio estimator, the error quickly falls. Note the nonlinear x-axis, suggesting a very stiff decline in error. **(c)** Posterior marginals for the four parameters shows that all three summary statistics are able to perform roughly equivalently in the oscillating region.

the proportion of SSA samples according to the ratio estimator. In this case, the error rapidly reduces to the level of the full SSA summary statistic by introducing only 1.5% of SSA samples. Assigning an insufficient proportion of SSA samples leads to significantly larger errors.

Figure 3.2c shows the posterior distribution of the trained summary statistics for a set of observations in the oscillatory regime. All three posteriors are able to capture the true parameters, indicating that for certain parts of parameter space, the ODE and the approximate ratio summary statistic can perform just as well as the SSA trained summary statistic. However, the lower $E_{\%}$ indicates that globally the mixed summary statistics may perform better. Of note in this example is that, despite the ODE being deterministic, we still obtain good results, demonstrating the robustness of the method to having a perfectly precise ratio estimator.

Comparison to Random

Proportion of SSA Samples	Random $E_{\%}$	Ratio Estimator $E_{\%}$
0.015	0.79 [0.63, 1.51]	0.65 [0.57, 0.75]
0.025	0.71 [0.56, 0.94]	0.66 [0.56, 0.70]
0.030	0.66 [0.56, 0.87]	0.62 [0.56, 0.71]

Table 3.1: **Approximate Summary Statistic Median $E_{\%}$ for Lotka-Volterra for 100 iterations with 90% intervals**

In Table 3.2, we show the performance of summary statistics for the Lotka-Volterra model trained by randomly including a fixed proportion of SSA samples instead of using the ratio estimator. We see that, while the randomly trained summary statistic can produce results comparable to our ratio estimator approximate summary statistic, it is far less robust, especially when the proportion is small. This makes sense because as we include more random samples, the chance of randomly including the same samples as the ratio estimator becomes much higher. Since the Lotka-Volterra model does not really require many SSA samples as seen from our experiments, this happens relatively quickly.

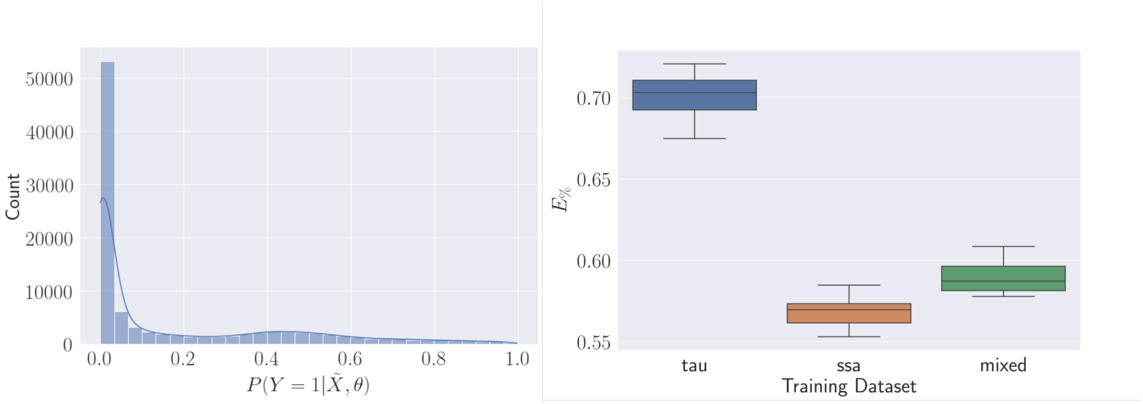
Genetic Toggle-Switch

The Genetic Toggle-Switch is a model for a biological system which exhibits stochastic switching behavior at low-population counts. This system is described by the following set of reactions:

$$
\begin{aligned}
&\phi \overset{\frac{\alpha_1}{1+V^{\beta}}}{\rightarrow} U &\qquad\qquad& \phi \overset{\frac{\alpha_2}{1+V^{\gamma}}}{\rightarrow} V \\
&U \overset{\mu}{\rightarrow} \phi && V \overset{\mu}{\rightarrow} \phi
\end{aligned}
$$

with initial conditions $U = 10, V = 10$. We use an adaptive τ-Leaping solver [11] as our approximate simulator, which in this model produces trajectories with consistently higher population counts than the SSA model. Since these differences correspond to areas that have small population counts, the difference in ensemble results are significant. We train the ratio estimator using $M = 5000$ and train the summary statistic using a budget of $N = 10^5$. $E_{\%}$ is evaluated using 300000 hold out SSA test samples.

Figure 3.3a shows the predicted ratios for all 10^5 low-fidelity samples after training, indicating that the classifier can easily distinguish the correct class of most of the τ-Leaping samples. However, as there is still mass near 0.5, we are able to reduce the number of SSA calls by 50% while only losing 3% in $E_{\%}$. Under this prior, though most of the parameter space leads to small population counts, significant portions lead to growth in the populations of U and V, where the τ-Leaping approximation is more accurate. The trained ratio estimator is able to capture this difference and prevent expensive resampling.



(a) Ratio Estimates

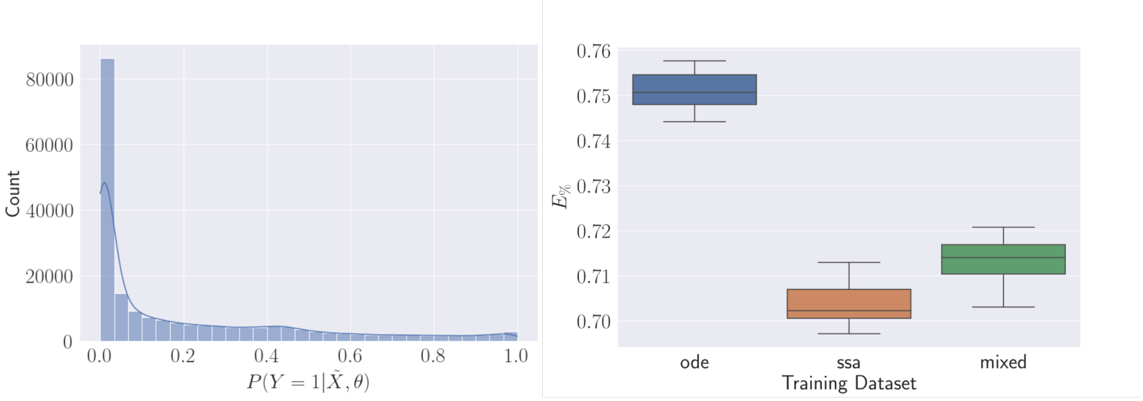
(b) $E_{\%}$ vs. Dataset Composition

Figure 3.3: **Trained Ratio Estimates for the Genetic Toggle-Switch (a)** The trained $P(Y = 1|\tilde{X}, \theta)$ can easily classify most of the cases. **(b)** The $E_{\%}$ error only slightly increases by using our mixed training set but still reduces SSA calls significantly.

Vilar Oscillator

To investigate our method on a larger problem with a questionable approximation, we look at a stable stochastic genetic oscillator [114] modelling a circadian clock. The system is defined with 9 species and 18 reactions controlled by 15 rate constants, and is designed to produce robust oscillations in the presence of intrinsic noise. See the Appendix for further details for the reactions of this model. The Vilar Oscillator is a challenging problem for inference due to oscillations of a certain amplitude being localized to small region of parameter space coupled with the large prior space. We use an ODE model with log-normal noise as our approximation and only observe species C, A, and R of the system. Under the observational settings for this model, the parameters are generally poorly identified [3]. The ratio estimator is trained using $M = 10000$ and the summary statistic is trained using $N = 200000$. $E_{\%}$ is evaluated using 300000 hold out SSA test samples.

Figure 3.4a shows that the trained approximate ratio estimator is able to easily classify most of the ODE solutions with added noise, suggesting that the ODE model is a fairly



(a) Ratio Estimates

(b) $\mathrm{E}_{\%}$ vs. Dataset Composition

Figure 3.4: **Trained Ratio Estimates for the Vilar Oscillator (a)** The trained $P(Y = 1|\tilde{X}, \theta)$ can easily classify most of the cases. **(b)** The $E_{\%}$ error only slightly increases by using our mixed training set but still reduces SSA calls significantly.

poor approximation. While this model is robust to noise and the mean is captured well by the ODE, at the same time, the log-normal noise does not properly capture the variance and the ratio estimator is able to distinguish the two. This is potentially also useful to diagnose whether an approximation is appropriate to study the model. Nevertheless, due to the addition of noise, the ratio estimator remains uncertain about some areas and we are still able to reduce the number of SSA calls by 20% and obtain a similar $E_{\%}$ to that of the full SSA dataset as seen in Figure 3.4b. This shows that, even when the approximation is poor, computational savings can still be accomplished while maintaining accuracy by intelligently selecting resamples according to the ratio estimator.

3.3 Discussion

	Total Simulations	Total SSA Calls	% Reduction in SSA Calls
Pure-Birth	30000	0	**-100%**
Lotka-Volterra	100000	9954	**-90.0%**
Genetic Toggle-Switch	100000	43699	**-56.3%**
Vilar Oscillator	200000	151164	**-24.4%**

Table 3.2: **Ratio-Estimated Approximate Summary Statistic SSA Calls**

	Approximate Only	SSA	Mixed	% Change in $E_{\%}$
Lotka-Volterra	2.24	0.63	0.64	**-1.6%**
Genetic Toggle-Switch	0.70	0.57	0.59	**-3.5%**
Vilar Oscillator	0.75	0.70	0.71	**-1.4%**

Table 3.3: **Approximate Summary Statistic Average** $E_{\%}$

Tables 3.2 and 3.3 summarize the primary results for all of the experiments. The results report the average $E_{\%}$ over 15 replications as aforementioned. Notably, in each case, using our method we are able to train a summary statistic using significantly fewer expensive SSA calls with only a small loss in accuracy. Overall the trained ratio estimator is able to detect when the approximate simulator is good and thus when to lean heavily into the approximate simulator for training.

Practical Implementations

While a precise ratio estimator will inevitably lead to an accurate algorithm, we find that in many cases, the ratio estimator for training a summary statistic does not need to be incredibly accurate. In fact, a very expressive ratio estimator may overfit to noise and lead to perfect classification while less expressive ratio estimators can produce a similar level of accuracy in the summary statistic. This is most apparent when we use an ODE as an approximation, where the ratio estimator can quickly learn to discriminate based on the smoothness of solutions. Nevertheless, this can still be useful for summary statistics, as approximate models can often still represent the high-level features. In our examples, we used a variety of Neural Network architectures to learn the ratio estimator, but we find that often, a simple DNN suffices to obtain similar results. For the Lotka-Volterra ODE model, we use a DNN to prevent overfitting to noise, as mentioned above. We use a CNN architecture similar to [3] for the other models where the approximation is stochastic, and suggest a similar approach based on the approximate simulator used.

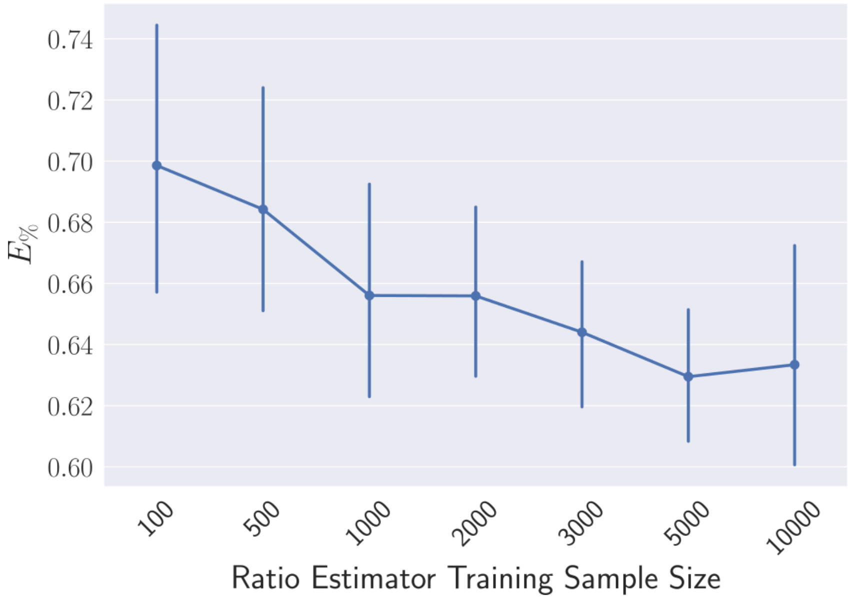


Figure 3.5: $E_{\%}$ **error vs Number of Ratio Estimator Training Samples for Lotka-Volterra** Larger N increases the accuracy and robustness but with diminishing returns. Selecting M is highly model dependent.

Selecting the number of samples M to train the ratio-estimator is important both for the efficiency and accuracy of our method. In general, M depends on how sensitive the output of the model is through parameter space. If the model exhibits heavily varies throughout parameter space, M would naturally need to be larger to capture this. In Figure 3.5, we show the performance of the approximation trained summary statistics as we change the number of initial samples M for the Lotka-Volterra model. While this is highly model dependent, we can see that in this case, the number of samples does not need to be high to obtain good accuracy for the summary statistic. As the approximation is relatively accurate and behavior does not rapidly change through parameter space, we only need to add full simulations from a few locations to obtain an accurate ratio estimator. After which, larger M only marginally changes the accuracy or robustness of the summary statistic.

In selecting ρ, we are trying to maximize accuracy while minimizing the number of

SSA samples. An effective heuristic is to simulate a large batch of cheap, approximate trajectories, pass it through the ratio estimator, and choose ρ to capture the first major mode in the distribution. For the Lotka-Volterra model, Figure 3.2a would suggest to set rho to around 0.01. Empirically, we find that setting the threshold quite low and effectively only correcting for the worst cases can still produce effective summary statistics. Optimal selection of ρ is something to investigate in the future, as it represents a key computational trade-off.

Learning an approximate ratio-estimator via binary classification, while generally an easier task than learning summary statistics, can be expensive if the parameter space is very sensitive or very high dimensional. In these cases, to distinguish between models we may need to set M to a large number to get the precision needed. In our examples, we are able to use a much smaller number of samples than needed to train the summary statistic. As model complexity increases, the number of training samples needed to learn a good ratio estimator will likely increase. One possibility to save some computational cost is to pre-train the first layers of the ratio-estimator to be an encoder, and then fine-tune the encoder layers to learn the summary statistic. This would act as a semi-supervised algorithm [59] that may be useful for learning a good summary statistic.

Related Work

The use of multifidelity simulators for Approximate Bayesian Computation has been explored, but under the assumption of the existence of a set of summary statistics. Prescott and Baker[80] construct a similar decision process for using multifidelity simulators within ABC-MCMC and ABC-SMC algorithms. In their method, they derive optimal continuation probabilities from a set of assumptions, while we take the more black-box approach of using Deep Neural Networks and approximate ratio estimators.

Approximate Likelihood Ratios have been used to perform likelihood free inference within both an MCMC and an ABC framework [48, 106, 20, 7]. These works have mainly focused on estimating the likelihood ratios within a single model at different parameter points, whereas our focus is on estimating the likelihood ratio between approximate and full models.

3.4 Methods

Approximate Summary Statistics

Given access to an approximate simulator $q(\tilde{X}|\theta)$ and the full SSA simulator $p(X|\theta)$ for a given discrete stochastic biochemical system, our goal is to train a summary statistic according to (3.1) that utilizes as many approximate samples as possible, while mitigating the bias in doing so. We assign a computational budget of N total simulations and assume that the approximate simulator is much faster to simulate from than SSA. For discrete stochastic models, this assumption is accurate much more often than not. As the approximation error is often non-trivial, training a summary statistic using only approximate trajectories will likely lead to bias depending on the problem.

Constructing an Approximate Dataset for Training via Likelihood Ratios

Our approach to solving this problem is to treat each sampling step as a decision on whether the approximate simulation is sufficient. Specifically, suppose that for each sample, we draw $\theta \sim p(\theta)$ and then simulate from the approximate simulator $\tilde{X} \sim q(\tilde{X}|\theta)$. The sample $\tilde{X}$ will induce bias in training $S(X)$ if at θ, $q(\tilde{X}|\theta)$ is significantly different from the full SSA simulator $p(\tilde{X}|\theta)$. Intuitively, to avoid this bias, we will need to resample, $X \sim p(X|\theta)$ and discard $\tilde{X}$. Computational savings will be attained if, in a substantial portion of parameter space, the approximate simulator yields a good approximation to that of SSA.

We quantify the difference between the two models using the likelihood ratio between the SSA model and the approximate model evaluated at the approximate sampled trajectory $\tilde{X}$ and θ:

$$r(\tilde{X}, \theta) \triangleq \frac{p(\tilde{X}|\theta)}{q(\tilde{X}|\theta)}. \tag{3.4}$$

This can be seen as conducting a hypothesis test at each step to determine whether there is sufficient evidence to distinguish which simulator the trajectory came from. If at a given θ and $\tilde{X}$, we cannot distinguish whether it came from the approximation or the full model, using the approximate simulation should induce little bias. If the two models produce the exact same likelihood, we would expect a value of 1, expressing indifference between the two. Most importantly, evaluating this ratio often requires simulating only from the approximate simulator, requiring a call to SSA only if we are not confident in the approximate trajectory.

Unfortunately, for discrete stochastic biochemical models, this ratio is unavailable due to the intractability of the likelihood. However, using recent advances in machine learning, we can construct powerful approximations to the likelihood ratio.

Approximate Ratio Estimation

Algorithm 3: Summary Statistic with Approximate Simulators

```
Input: prior $p(\theta)$, SSA simulator $p(X|\theta)$, approximate simulator $q(X|\theta)$, $M$
       ratio-samples, $N$ samples, tolerance range $\rho$
Output: calibrated $S(X) = f_\Phi(X)$
ratio_samples = $\{\}$;
for $i$ = $1$ : $M$ do
    Sample from the prior $\theta \sim p(\theta)$;
    Draw a trajectory $X \sim p(\tilde{X}|\theta)$;
    ratio_samples = ratio_samples $\bigcup$ $(Y = 1, X, \theta)$;
    Draw a trajectory $\tilde{X} \sim q(\tilde{X}|\theta)$;
    ratio_samples = ratio_samples $\bigcup$ $(Y = 0, \tilde{X}, \theta)$;
end
Train $\hat{r}(X, \theta) = p(Y = 1|X, \theta)$ using ratio_samples;
samples = $\{(X_1, \theta_1), \ldots (X_M, \theta_M)\}$;
for $i$ = $M$ : $N$ do
    Sample from the prior $\theta \sim p(\theta)$;
    Draw a trajectory $\tilde{X} \sim q(\tilde{X}|\theta)$;
    if $\hat{r}(\tilde{X}, \theta) < \rho$ or $\hat{r}(\tilde{X}, \theta) > 1 - \rho$ then
        Draw a trajectory $X \sim p(X|\theta)$;
        samples = samples $\bigcup$ $(\theta, X)$;
    else
        samples = samples $\bigcup$ $(\theta, \tilde{X})$;
    end
end
Train $S(X) = \mathbf{E}[\theta|X] = f_\Phi(X)$ using samples.
```

Although the likelihood ratio in (3.4) cannot be directly computed, recent work has shown that it can be well approximated by using a binary classifier to distinguish between samples from the two different models [20, 44, 106]. Specifically, suppose we assign labels $Y = 1$ to trajectories $X \sim p(X|\theta)$ and $Y = 0$ to trajectories from $\tilde{X} \sim q(X|\theta)$. If we have access to a probability $p(Y = 1|X, \theta)$, the likelihood ratio is directly related via:

$$p(Y = 1|X, \theta) = \frac{p(X|\theta)}{p(X|\theta) + q(X|\theta)} \tag{3.5}$$

$$r(X, \theta) \triangleq \frac{p(X|\theta)}{q(X|\theta)} = \frac{p(Y = 1|X, \theta)}{1 - p(Y = 1|X, \theta)}. \tag{3.6}$$

As we do not have access to $p(Y = 1|X, \theta)$, we must approximate it. Recent advances in deep learning have demonstrated how to build powerful approximations to $p(Y = 1|X, \theta)$ despite the dimensionality of the trajectories X:

$$Y \sim \text{Bernoulli}(\phi(X, \theta)) \tag{3.7}$$

$$\hat{p}(Y = 1|X, \theta) = \phi(X, \theta) = \frac{\exp(f_\psi(X, \theta))}{1 + \exp(f_\psi(X, \theta)} \tag{3.8}$$

with dataset

$$\mathcal{D} = \{(\theta_1, X_1, 1), (\theta_1, \tilde{X}_1, 0), \cdots, (\theta_M, X_M, 1), (\theta_M, \tilde{X}_M, 0)\}.$$

Despite the need to train a ratio estimator, the binary classification task is easier than the regression task, allowing us to use fewer training samples than for training the summary statistic. The parameters of f_ψ are estimated via maximum likelihood. With this initial step, we describe the full summary statistic training procedure in Algorithm 3. As $p(Y = 1|X, \theta)$ is directly proportional to $\hat{r}(X, \theta)$, we use the probability as a more interpretable surrogate within the algorithm.

Implementations of this algorithm and replications of the experiments can be found at `https://github.com/rmjiang7/approximate_summary_statistics`.

3.5 Conclusions

We have presented a method to utilize approximate simulators together with exact simulators of discrete stochastic reaction models to train summary statistics for ABC. Using advances in Machine Learning and approximate ratio estimators, we demonstrate that when properly calibrated, we can significantly reduce the number of expensive SSA calls required for learning a summary statistic. Using four examples of reaction systems, we showed that significant computational savings can be achieved while preserving accuracy of approximate summary statistics.

In this work we have focused on utilizing only a single approximation at a time. In practice, there are numerous approximations available for the same model of varying accuracy. Extending this method to choose between different levels of approximations could further reduce the number of full SSA calls needed, even in cases where one of the approximations is sufficiently poor in all regions.

Chapter 4

Bayesian Systems Identification of Mass-Action Biochemical Reaction Networks

In the previous chapter, we discussed the problem of parameter estimation and inference for biochemical reaction networks. However, parameter estimation largely assumes that the biologist is aware of the underlying reaction system, but unaware of the reaction rates. In many realistic situations, a biologist will only have partial knowledge of the reaction system, potentially missing several key components. To this end, a complementary problem to parameter estimation is inference of the structure of the reaction system itself.

In this chapter, we present our work *Identification of Dynamic Mass-Action Biochemical Reaction Networks Using Sparse Bayesian Methods* [104] which focuses on the problem of inferring the structure of a reaction system. Specifically, we describe how we can utilize the assumption of mass-action kinetics to construct a flexible statistical model, and then apply recent techniques in Bayesian sparsity priors to recover interpretable reaction systems. In addition, we show how the latent variable formulation of the problem allows for unique observational models without additional bias while still being amenable to Bayesian inference using modern efficient samplers. We conclude with a discussion of the significant challenge of identifiability of these systems, which poses a problem for any solution.

4.1 Abstract

Identifying the reactions that govern a dynamical biological system is a crucial but challenging task in systems biology. In this work, we present a data-driven method to infer the underlying biochemical reaction system governing a set of observed species concentrations over time. We formulate the problem as a regression over a large, but limited, mass-action constrained reaction space and utilize sparse Bayesian inference via the regularized horseshoe prior to produce robust, interpretable biochemical reaction networks, along with uncertainty estimates of parameters. The resulting systems of chemical reactions and posteriors inform the biologist of potentially several reaction systems that can be further investigated. We demonstrate the method on two examples of recovering the dynamics of an unknown reaction system, to illustrate the benefits of improved accuracy and information obtained.

4.2 Introduction

Reconstructing the correlated reactions that govern a system of biochemical species from observational temporal data is an essential step in understanding many biological systems. To facilitate this process, we propose a robust, data-driven approach based on a sparse Bayesian statistical model. Our approach exploits sparse Bayesian priors and an unbiased observational model to recover a parsimonious, interpretable reaction system from mass-action relations, utilizing very little user input. On a set of simulated test problems, the method demonstrates increased robustness and decreased bias at different levels of measurement variability, while also producing interpretable reaction systems and quantifying uncertainty. As a tool, the approach can be used to flexibly interrogate biological systems while allowing incorporation of potentially uncertain domain knowledge to improve the efficiency and identifiability of the problem.

Developments in high-throughput experimental methodologies in biology have enabled the collection of massive amounts of time varying molecular data at small scales. This has resulted in significant advances in understanding the biochemical networks and mechanisms underlying physiological processes such as gene regulation. Indeed, greater understanding of regulatory processes at the single cell level can aid in the development of targeted therapies for diseases such as cancer [2, 74, 118]. A major challenge in this process is the translation of high-throughput, observational molecular data into analyzable and interpretable reaction networks. Typically, this is accomplished by utilizing significant biological insights to first define a reaction system, and then calibrating the model based on collected data, which while accurate, requires substantial time and effort to iterate. An appealing avenue is to utilize data-driven approaches for systems identification, whereby plausible biochemical reaction networks are generated and estimated directly from data without the need to initially propose a system. While recently, many such methods have been developed to infer networks from a wide variety of different datasets, it remains a challenging statistical and computational task [16]. Most works of estimating networks typically focus on either reconstructing a network without assuming any known dynamics due to destructive time series measurements [65, 61, 53], or producing networks that replicate dynamics, but without focusing on interpretability [64, 121, 68, 72, 35].

In this work, we are primarily interested on identifying interpretable mass-action biochemical reaction networks using only the observed time series of species concentrations. Expanding upon the problem formulation first proposed as Reactive SINDy in [51], we automatically enumerate the allowable mass-action reactions given a set of species and a library of ansatz reactions and utilize advances in sparse Bayesian inference to generate posterior distributions of interpretable biochemical reaction systems. Compared with Reactive SINDy, our method provides uncertainty estimates over potential reaction systems, reduces a major source of bias in the previous method, and produces potentially several interpretable reaction networks. Furthermore, the transparent statistical formulation of the problem allows us to easily incorporate existing, potentially uncertain, domain knowledge via prior distributions to improve the efficiency and identifiability of the problem.

The remainder of this paper proceeds as follows. In Materials and Methods, we describe mass-action biochemical reaction networks and formulate the problem of inferring these networks from observational data. Next, we propose improvements to the existing methodology and describe the specifics of the proposed model applied to inference of reaction networks. In Results, we demonstrate how our methodology can be used in two different examples to retrieve interpretable networks from observational concentration data. We close in Discussion by noting a few details for usage, detailing some future directions, and mentioning the limitations of our method. All implementations and code can be found at `https://github.com/rmjiang7/bayes_reactive_sindy`.

4.3 Materials and methods

4.3.1 Mass-action Biochemical Kinetics

Systems of biochemical species reacting under any number of reaction channels are commonly modeled dynamically using the framework of chemical kinetics. Specifically, denote $\mathbf{X}(t) \in \mathbb{R}^N$ as the vector of concentrations of each of N species at time t. The evolution of the system can be modeled using the following set of coupled ordinary differential equations (ODEs) formally known as the reaction rate equations:

$$\frac{d\mathbf{X}}{dt} = S\mathbf{f}(\mathbf{X}), \tag{4.1}$$

where $S \in \mathbb{Z}^{N \times D}$ is the stoichiometric matrix with D reactions among N species and $\mathbf{f}(\mathbf{X})$ is the vector of all rate functions.

Although theoretically, $\mathbf{f}(\mathbf{X})$ can take the form of any nonlinear function, in this work we assume that the system follows mass-action kinetics [115] and thus, the reaction rates are proportional to the product of the concentrations of each reactant in the case of multiple reactants, and proportional to the concentration of the reactant in single reactant reactions.

4.3.2 Network inference for mass-action reaction systems

Suppose we observe a time series of N species concentrations at T discrete times:

$$\hat{\mathbf{X}}(t_j) \in \mathbb{R}^N, \qquad j = \{0, \ldots, T\}.$$

Given this data and assuming that the system is governed by up to 2nd-order mass-action kinetics and the dynamics of Eq. (4.1), we wish to recover a parsimonious system of expressible reactions that can explain the observed data.

Under these constraints, this problem can be posed as a linear regression, given a library of ansatz reactions. More specifically, suppose we initially specify a large set of D possible reactions among the N species in our system. Each reaction can be expanded into a stoichiometry s and a rate function $f(\mathbf{X})$, where the rate function is known due to the assumption of mass-action rate kinetics. Let $S_c \in Z^{N\times D}$ denote the complete stoichiometric matrix constructed by stacking all D stoichiometries into a matrix. The reaction rate equations then take the form,

$$\frac{d\mathbf{X}}{dt} = S_c^T \begin{pmatrix} k_1 f_1(\mathbf{X}) \\ k_2 f_2(\mathbf{X}) \\ \cdots \\ k_D f_D(\mathbf{X}) \end{pmatrix}, \tag{4.2}$$

where $k_i > 0$ is the unknown rate-constant and f_i is simply a product of the reactants for the i-th reaction. Thus we aim to estimate $\mathbf{k}$ such that, when solved, Eq. (4.2) replicates the observations $\hat{\mathbf{X}}$ at all t_j. Many methods exist to solve these types of problems, such as ridge, LASSO, and Elastic net regression [49, 108, 127, 112].

Although S_c is potentially high dimensional, conditional on the initially specified set of D reactions, in most situations D over-specifies the possible reactions. Hence, to replicate the observations, a safe assumption is that most potential reactions do not exist, which is equivalent to setting $k_i = 0$ when the i-th reaction does not contribute to the dynamics of the system. This assumption can be captured by estimating $\mathbf{k}$ using sparse regression methods. A small reaction system can then be expressed by rewriting the system in terms of only the non-zero reactions.

Sparse regression methods for estimating dynamical systems from data have been widely applied in the last few years. More generally, when Eq. (4.1) is generated from polynomial basis functions rather than ansatz reactions, this becomes Sparse Identification of Nonlinear Dynamics (SINDy) [10], which has been applied to biological systems [64], though without the specific aim to recover interpretable reactions. Reactive SINDy, as described above, expands SINDy by constraining the basis functions to such ansatz mass-action reactions. Both of these methods estimate the coefficients $\mathbf{k}$ using LASSO regularization, resulting in maximum likelihood networks that do not inform about the uncertainty associated with the particular fits, an especially important feature when data is sparse and noisy. Reactive SINDy uses finite difference derivative estimates from observations to transform Eq. (4.2) into a linear regression problem, which can result in significant bias for estimating networks when measurements are sparse and noisy, as is often the case in biological systems.

More specifically, using the assumptions of mass-action kinetics and the law of parsimony, Reactive SINDy solves a mixed LASSO and ridge regression optimization problem. Letting $\frac{d\hat{X}}{dt}(t_i)$ be the derivatives numerically estimated from the observations $\hat{\mathbf{X}}(t_j)$ via second-order finite differences, the optimization problem solved is

$$\Phi(X) = \begin{pmatrix} k_1 f_1(X(t_j)) \\ k_2 f_2(X(t_j)) \\ \cdots \\ k_D f_D(X(t_j)) \end{pmatrix}$$

$$\mathbf{k} = \arg\min_{\mathbf{k}} \left(\frac{1}{2T} \|\hat{\mathbf{X}} - \Phi(X)\|_F^2 + \alpha\lambda\|\mathbf{k}\|_2 + \alpha(1-\lambda)\|\mathbf{k}\|_2^2\right)$$

$$\text{subject to } \mathbf{k} \geq 0.$$

The equivalent statistical model for the LASSO optimization can be summarized as

$$k_i \sim \text{Laplace}(\lambda), \tag{4.3}$$

$$\frac{d\hat{X}}{dt}(t_j) \sim \text{Normal}\left(S_c^T \begin{pmatrix} k_1 f_1(X(t_j)) \\ k_2 f_2(X(t_j)) \\ \dots \\ k_D f_D(X(t_j)) \end{pmatrix}, 1 \right), \quad j = 0, \dots, T,$$

which can also be fit using Bayesian methods to provide uncertainty estimates.

In this work we improve on Reactive SINDy in two key ways. First, we estimate **k** using the sparse Bayesian regularized horseshoe prior to obtain uncertainty estimates as well as to introduce a natural way of incorporate existing domain knowledge via prior distributions. Second, we avoid biased numerical derivative estimates by re-formulating the statistical model in terms of the solution of the ODE. This better captures the observational model and allows us to incorporate alternative models of measurement noise. Using recent advances in automatic differentiation software for sensitivity analysis of ODE systems [4, 13, 15, 27], this can be solved efficiently and provides more accurate solutions, especially in the case of sparsely measured data.

4.3.3 Bayesian Reactive SINDy

In this section we introduce the regularized horseshoe prior [76] used in our Bayesian formulation of the Reactive SINDy model and the modified observational model, which better captures the measurement process and avoids biased, low-order derivative estimates. We construct the complete stoichiometric matrix S_c using a library of possible mass-action ansatz reactions and all reaction rates are specified by **k** as indicated in Eq. (4.2) Details for how we construct a set of ansatz reactions can be found in the

Appendix.

Sparse Bayesian regularized horseshoe priors

A challenge in implementing a Bayesian formulation of this problem is the fact that the LASSO penalization used for sparse parameter estimation, which can be translated as a statistical model to Eq. (4.3), does not result in sparse Bayesian posterior distributions. Instead, we adapt the regularized horseshoe prior, an extension of the standard horseshoe prior [14], which is a drop-in replacement for the LASSO derived Laplace prior.

Letting N be the number of species, T be the number of observations, and D be the number of ansatz reactions, the regularized horseshoe prior placed on the reaction coefficients $\mathbf{k}$ takes the form

$$\begin{aligned} \lambda_i &\sim \text{Cauchy}^+(0,1), \\ \tilde{\lambda_i} &= \sqrt{\frac{c^2\lambda_i^2}{c^2+(\tau\lambda_i)^2}}, \\ k_i &\sim \text{Normal}(0,\tau\tilde{\lambda_i}), \quad i=1,\dots,D. \end{aligned} \tag{4.4}$$

This promotes sparse solutions in the following way: each reaction rate k_i is given a normal prior centered around 0 with a standard deviation of $\tau\lambda_i$, where τ is a global shrinkage parameter shared among all reaction rates and λ_i is a positive parameter specific to each reaction rate. The heavy tailed half-Cauchy priors on the individual λ_i allows for the values to grow extremely large. This has the following effect:

$$\begin{aligned} &\text{if } (\tau\lambda_i)^2 \gg c^2, \qquad \tau\tilde{\lambda_i} \to c \\ &\text{if } (\tau\lambda_i)^2 \ll c^2, \qquad \tau\tilde{\lambda_i} \to \tau\lambda_i. \end{aligned}$$

Thus, if k_i is estimated to be non-zero, λ_i is allowed to become large and k_i breaks away from τ toward a regularized value of c^2, which is an estimate of the scale of the non-zero terms. On the other hand, if k_i is estimated to be zero, λ_i becomes small and k_i is shrunken to 0 with an often very small standard deviation. The horseshoe prior has the effect of placing significant prior mass towards 0 for all parameters, but allowing for any individual parameter to be non-zero if there is sufficient evidence to do so. The regularized horseshoe further shrinks non-zero estimates using a Gaussian slab with variance c^2, to help when parameters are weakly identified and to prevent non-zero values from growing too large.

The pivotal global shrinkage parameter τ specifies the scale of the near-zero reaction rates, which is relevant because, compared to the spike-and-slab prior [56], the regularized horseshoe prior is continuous in all parameters, preventing any parameter from becoming exactly 0. Furthermore, smaller values of τ also result in sparser networks. For our problem, as reaction rates can often be very small, specifying the scale at which a reaction is considered negligible can dramatically affect the interpretation and the simulated dynamics.

Following [76], we place a hyper-prior on the term c with distribution

$$c \sim \text{Inv-Gamma}(a, b).$$

The c parameter regularizes by essentially placing a $\mathcal{N}(0, c^2)$ prior on non-zero rates, preventing them from getting too large. The non-regularized horseshoe is retrieved when $c^2 \to \infty$.

The regularized horseshoe prior offers a few distinct advantages compared to other sparse Bayesian priors. Primarily, the dependency structure formed by introducing the global τ and the local λ_i parameters leads to sparser solutions that can borrow information from other reactions. The regularized horseshoe, as a continuous relaxation of the commonly used sparse Bayesian spike-and-slab prior [56, 67], allows for efficient Bayesian computation using modern gradient based MCMC samplers such as Hamiltonian Monte Carlo (HMC) [69] or Variational Inference [83]. This allows it to be implemented in probabilistic programming languages such as Stan [12], PyMC3 [89], or Pyro [5].

Observational Model

A potentially large source of bias in SINDy and Reactive SINDy as presented in Eq. (4.3) is the need to first estimate $\frac{d\hat{X}}{dt}$ from observations of the system. This presents an issue as standard methods of estimating derivatives, such as finite difference methods, become much less accurate as the time between observations increases, resulting in heavily biased estimates of $\mathbf{k}$. To correct for this, we modify the observational model as follows:

$$\mathbf{Z}(t_j) = \int_0^{t_j} S_c^T \begin{pmatrix} k_1 f_1(Z(t)) \\ k_2 f_2(Z(t)) \\ \dots \\ k_D f_D(Z(t)) \end{pmatrix} dt$$

$$\hat{\mathbf{X}}(t_j) \sim \text{Log-Normal}(\mathbf{Z}(t_j), \sigma), \quad j = 0, \dots, T.$$

Rather than assuming that we observe derivatives of the process, as in the Eq. (4.3), this formulation models that the underlying system follows a latent variable $Z(t_j)$, which is the solution of the ODE. We observe noisy measurements of the underlying system $\hat{\mathbf{X}}(t_j)$ at times t_j. By directly modeling the observations, there is no need to pre-process the data by estimating derivatives.

In this work, we also assume that the measured concentrations of each species are corrupted by log-normal error. This captures both that concentration measurements are strictly positive and that at higher concentrations, measurements are more variable. In addition, this formulation enables us to easily change the measurement error model to better capture the user's beliefs, without modifying the regularized horseshoe prior for inferring the network. As an example, a Poisson error model, such as that explained in [6], can be applied under the assumption that measured values are positive and discrete, and that measurements at some time t_j are distributed with mean and variance of $Z(t_j)$. The use of MCMC for sampling enables the observational model to be configured based on the experimental setup as long as the likelihood remains tractable.

The use of MCMC for sampling enables the observational model to be configured based on the experimental setup as long as the likelihood remains tractable. For biochemical reaction networks, PTLasso [46] apply a similar latent observation model, but with the Laplace prior to the parameters of a biochemical reaction network, further using parallel tempering MCMC to obtain sparse Bayesian estimates on models of up to a dozen different reactions.

The latent variable formulation also allows for the realistic scenario of observing only some of the species in the system. Suppose that in a system consisting of 5 species, we can only observe species $n = \{1, 2\}$. Then the observational model can be easily modified to

$$\begin{aligned} \mathbf{Z}(t_j) &= \int_0^{t_j} S_c^T \begin{pmatrix} k_1 f_1(Z(t)) \\ k_2 f_2(Z(t)) \\ \cdots \\ k_D f_D(Z(t)) \end{pmatrix} dt \\ \hat{X}_n(t_j) &\sim \text{Log-Normal}(Z_n(t_j), \sigma), \quad j = 0, \ldots, T, \quad n = \{1, 2\}. \end{aligned} \tag{4.5}$$

This is possible because the latent trajectory Z does not directly depend on the observed values $\hat{X}$. In comparison, for Eq. (4.3) used in Reactive SINDy and SINDy, the regression directly depends on the observed values, which thus requires complete observations of the system.

Statistical model and estimation

Combining the regularized horseshoe prior and the latent variable observational model, the complete hierarchical statistical model is specified by

$$\begin{aligned}
\lambda_i &\sim \text{Cauchy}^{+}(0,1), \\
\tilde{\lambda_i} &= \sqrt{\frac{c^2\lambda_i^2}{c^2+(\tau\lambda_i)^2}}, \\
k_i &\sim \text{Normal}(0,\tau\tilde{\lambda_i}), \quad i=1,\dots,D \\
\mathbf{Z}(t_j) &= \int_0^{t_j} S_c^T \begin{pmatrix} k_1 f_1(Z(t)) \\ k_2 f_2(Z(t)) \\ \dots \\ k_D f_D(Z(t)) \end{pmatrix} dt \\
\hat{\mathbf{X}}(t_j) &\sim \text{Log-Normal}(\mathbf{Z}(t_j),\sigma).
\end{aligned} \tag{4.6}$$

The algorithm for network identification is then as follows. First, we construct the complete stoichiometric matrix, S_c, and the set of linear and nonlinear reaction rate functions $\mathbf{f}(\mathbf{X})$ implied by mass-action kinetics. In our examples, the library of reactions consists of a large set of zero, first, and second order reactions types between all species modeled, which is automatically defined by our implementation. We note here that our method of generating possible reactions is intended to be general to demonstrate the method. In practice, the set of possible reactions is something the modeler can and should modify according to the constraints of the problem.

Provided with S_c and $\mathbf{f}(\mathbf{X})$, the sparse Bayesian posterior distribution $p(\mathbf{k}|\hat{\mathbf{X}})$ is approximated from the above statistical model using the No-U-Turns [50] sampler implemented in Stan [12].

As the regularized horseshoe is continuous in all parameters, no rate parameter will be set exactly to zero. Thus to decide whether a reaction is to be removed from the system, we employ the pruning technique adopted from [36]. Specifically, we estimate

$$P(\tau\tilde{\lambda_i} < \delta) > p_0, \quad i = 1, \dots, D$$

using the posterior distribution. This can be roughly interpreted as pruning all reactions where the posterior probability that the scale of k_i is less than δ is sufficiently large. This metric is sensible because rates which are shrunken towards 0 in the regularized horseshoe are scaled by $\tau\lambda_i$. This leaves two tuning hyperparameters, δ and p_0. These can be calibrated for a model by choosing the threshold such that, allowing more reactions does not improve the model's fit to the data, while removing reactions degrades the fit. In our examples, we find that $\delta = 1e^{-3}$ and $p_0 = 0.90$ work well for these models.

The complete implementation and all replicating results can be found at `https://github.com/rmjiang7/bayes_reactive_sindy`.

4.4 Results

We demonstrate our method on two synthetic examples where data is first generated from a known system of reactions and our method is used to recover the underlying network from a relatively large set of possible reactions. In each example, results are compared to those of Reactive SINDy, to show the ability of our model to obtain a network, with uncertainty estimates, that replicates the observations in addition to demonstrating the superior performance in the case of sparse observations due to the modified observational model. In the second, larger problem, we demonstrate the ability of our method to discover multiple small reaction systems that can capture the observations and discuss identifiability issues. Further descriptions and more precise model specifications can be found in the supplementary materials.

4.4.1 Lotka-Volterra

The Lotka-Volterra predator-prey system is a simple but informative example of a non-linear system with oscillatory dynamics. Although not strictly a biochemical reaction system, we provide it as an example for evaluating the model formulation and method. Briefly, the Lotka-Volterra system models the interaction dynamics of two species $X := \{P, Y\}$ where P is the predator and Y is the prey. This can be described using the following reactions:

$$\begin{aligned} Y &\overset{k_1}{\to} 2Y \\ P + Y &\overset{k_2}{\to} 2P \\ P &\overset{k_3}{\to} \phi, \end{aligned}$$

which corresponds to the following stoichiometric matrix and rate vectors under mass-action kinetics,

$$S = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad \mathbf{f}(\mathbf{X}) = \begin{pmatrix} k_1[Y] \\ k_2[P][Y] \\ k_3[P] \end{pmatrix}.$$

With $k_1 = 1, k_2 = 0.01, k_3 = 0.3$ and initial conditions $X(t_0) := \{50, 100\}$, this gives rise to stable oscillations.

Data is generated by solving the above system of reactions and perturbing with Log-Normal(0, 0.2) noise at fixed times to simulate a noisy measurement process. The reactions comprising the complete stoichiometric matrix S_c from which we will recover the underlying system is provided in Table (4.1) and adopted from [51]. In total, there are 16 possible reactions in this system, three of which are non-zero in the original system.

Table 4.1: **Library of Ansatz Reactions for the Lotka-Voltera Model**

Reaction Index	Allowed Reactions	True Rate Constant
0	$2X \xrightarrow{k_1} 0$	$k_1 = 0$
1	$2Y \xrightarrow{k_2} 0$	$k_2 = 0$
2	$\mathbf{X \xrightarrow{k_3} 2X}$	$\mathbf{k_3 = 1.0}$
3	$\mathbf{X + Y \xrightarrow{k_4} 2Y}$	$\mathbf{k_4 = 0.01}$
4	$\mathbf{X \xrightarrow{k_5} 0}$	$\mathbf{k_5 = 0.3}$
5	$X + Y \xrightarrow{k_6} 2X$	$k_6 = 0$
6	$X \xrightarrow{k_7} 0$	$k_7 = 0$
7	$2Y \xrightarrow{k_8} Y$	$k_8 = 0$
8	$Y \xrightarrow{k_9} 2Y$	$k_9 = 0$
9	$2X \xrightarrow{k_{10}} X$	$k_{10} = 0$
10	$X + Y \xrightarrow{k_{11}} X$	$k_{11} = 0$
11	$X + Y \xrightarrow{k_{12}} Y$	$k_{12} = 0$
12	$2X \xrightarrow{k_{13}} Y$	$k_{13} = 0$
13	$X \xrightarrow{k_{14}} Y$	$k_{14} = 0$
14	$Y \xrightarrow{k_{15}} X$	$k_{15} = 0$
15	$X \xrightarrow{k_{16}} 2Y$	$k_{16} = 0$

We generate data at three different measurement frequencies dt $= \{0.2, 1, 2\}$ between $t = [0, 15]$ and estimate $\mathbf{k}$ separately for each using the same S_c. Trajectories of the two species are shown in Fig. (4.1). For estimation from the regularized horseshoe model, we set $\tau = 1e^{-8}$ and estimate c along with the other parameters by placing the prior

$c \sim$ Inv-Gamma$(4, 4)$. A total of 4000 samples are drawn using four MCMC chains. We note that while we use MCMC for accuracy and demonstration purposes, variational inference can also be used to obtain fast approximate solutions and is supported in our implementations. In our experiments, we found that the variational approximations were generally reliable, though this largely depends on the problem.

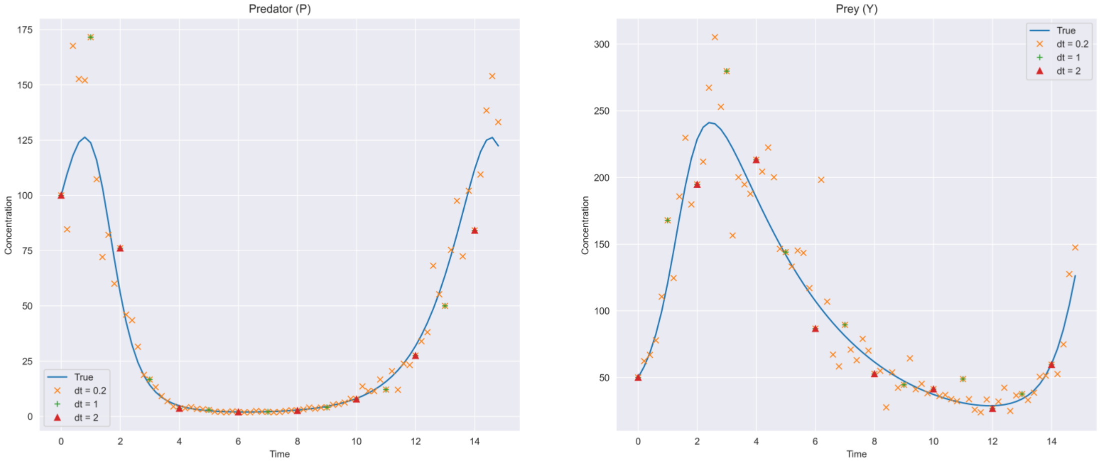
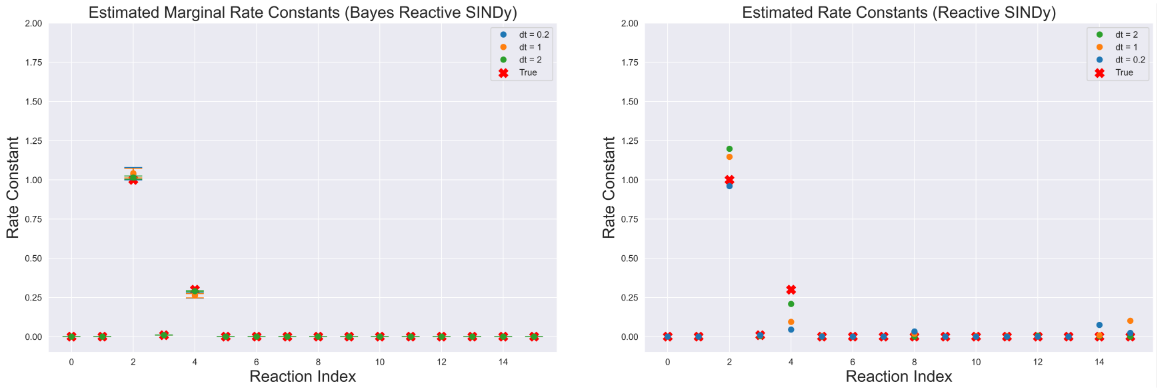


Figure 4.1: **Lotka-Volterra Observation Data** Simulated data used for network identification. Log-Normal noise is added to the true trajectory, and measurement frequency is changed to show the uncertainty in posteriors.

In Fig. (4.2a), we show the posterior credible intervals for the recovered rate constants from each of the three measurement frequencies, which are heavily centered around the true values for all reactions. Fig. (4.2b) shows the point estimates obtained by using Reactive SINDy under equivalent experimental setups. Notably, both methods can recover the reaction system with frequent measurements but Reactive SINDy degrades considerably as measurements become more sparse.

More specifically, the difference in the results demonstrates the bias introduced by estimating derivatives. At observation intervals dt = 1.0 and dt = 2.0, too much information is lost from estimating derivatives coupled with measurement noise to obtain the correct system. Fig. (4.3) shows the differences in inferred dynamics along with predic-

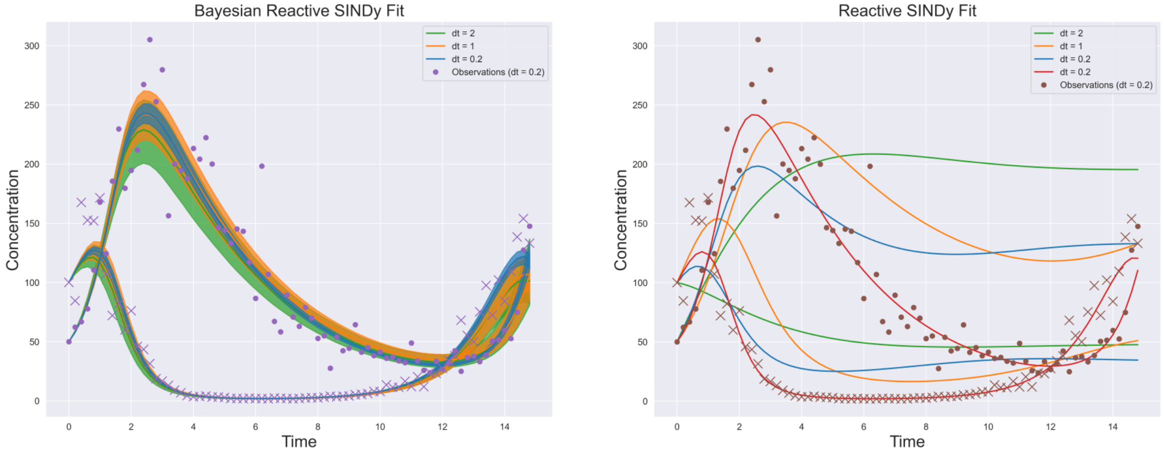


(a) (b)

Figure 4.2: **Lotka-Volterra Estimated Reactions** (a) Estimated parameters using Eq. (4.6). Reactions correspond to the reactions specified in Table. (4.1) (b) Estimated parameters using Reactive SINDy at different measurement frequencies as well as using noise-less measurements.

tive uncertainty intervals from the networks recovered using our observational model, (a), and Reactive SINDy, (b). Our model remains in phase with the observations while the networks derived from using estimated derivatives demonstrate a systematic bias away from the true reaction system, even in the case of dt $= 0.2$ due to measurement noise.

With the Bayesian treatment of the problem, we can also quantify uncertainty in the non-zero reaction rates. This informs us of the plausible range of reaction rates, given the observed data, and can be useful to detect which parameters the model is able to identify with evidence of correlated reactions. In Fig. (4.4), the posterior distributions of the non-zero estimated parameters are shown, demonstrating that as we increase measurement frequency, uncertainty decreases. Furthermore, in this system there is mild correlation between the reaction rates, indicating that they vary together to replicate the oscillating behavior.



(a) (b)

Figure 4.3: **Reconstructed Trajectories** (a) Using posterior samples from Eq. (4.6). Even at smaller observation frequencies, the observed data is accurately captured, though (as expected) with greater uncertainty. (b) As Reactive SINDy estimates derivatives, errors in the numerical methods lead to large deviations in the reconstructed trajectories as sampling frequency and noise increase. Although a single trajectory at dt $= 0.2$ may capture the oscillating behavior, it is clearly biased away from the true observations.

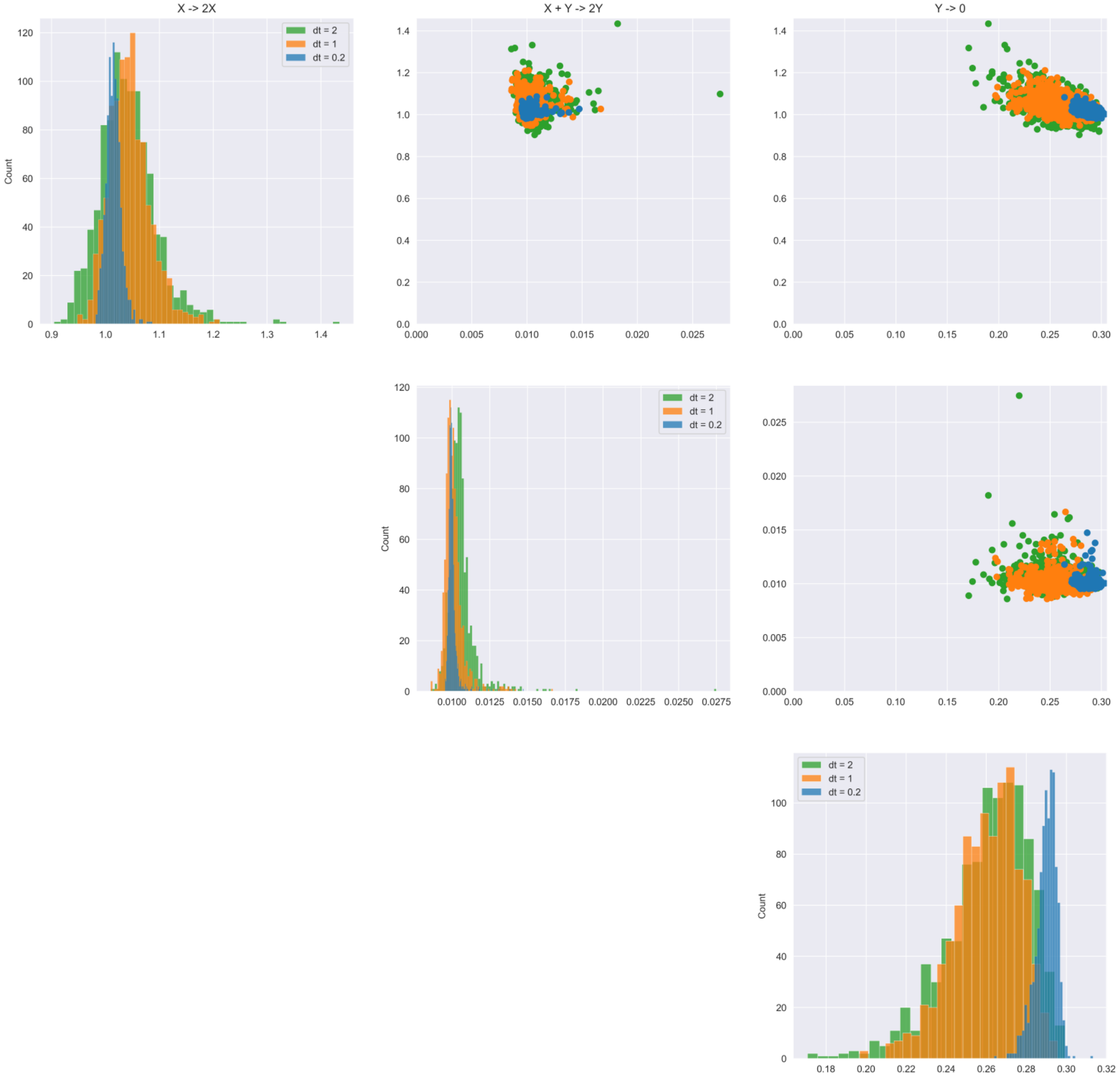


Figure 4.4: **Lotka-Volterra: Posterior Distributions of Non-Zero Reactions using the proposed model** As expected, uncertainty in the parameters decreases as the measurement frequency decreases, but all are concentrated in relatively the same area. Only a single network is consistently identified given this data, indicating that identifiabiltiy is not a problem for this system.

Partially observed species

In the previous example, we assumed that the species were completely observed. However, under the latent variable formulation, this is not strictly required. In this section, we demonstrate inference of the network for the identifiable Lotka-Volterra example, when only the prey species, Y, is observed. In this case, the statistical observational model can be changed to

$$\mathbf{Z}(t_j) = \int_0^{t_j} S_c^T \begin{pmatrix} k_1 f_1(Z(t)) \\ k_2 f_2(Z(t)) \\ \cdots \\ k_D f_D(Z(t)) \end{pmatrix} dt$$

$$\hat{Y}(t_j) \sim \text{Log-Normal}(Z_2(t_j), \sigma), \quad j = 0, \ldots, T, \tag{4.7}$$

where we apply the likelihood only to the observations of Y.

Fig. (4.5) shows the simulated trajectories and posterior distributions obtained by using our model under this scenario. Compared to the situation where both species are observed, the uncertainty is significantly higher for the same reactions because the information gained from observing P is lost. However, the method is still able to retrieve the correct networks, as the oscillating regime for this problem is generally unique.

Sums of observed species

Similar to above, the latent variable formulation we have presented allows for modeling of the situation where a sum of species concentrations is observed, but not any of the individual species. In this case, for the Lokta-Volterra system, letting $W = X + Y$ be the observed sum of X and Y, the statistical model can be stated as

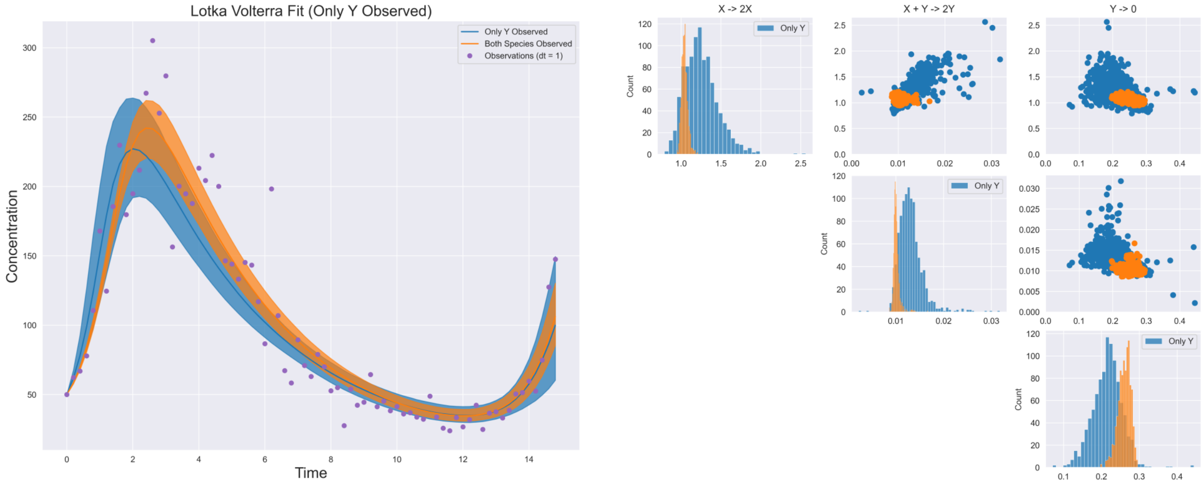


Figure 4.5: **Identified Trajectories and Posterior from Partial Observations** The true network can still be captured using only observations of Y however the credible intervals are significantly higher due to the loss of observations of P.

$$\mathbf{Z}(t_j) = \int_0^{t_j} S_c^T \begin{pmatrix} k_1 f_1(Z(t)) \\ k_2 f_2(Z(t)) \\ \dots \\ k_D f_D(Z(t)) \end{pmatrix} dt$$

$$\hat{W}(t_j) \sim \text{Log-Normal}(Z_1(t_j) + Z_2(t_j), \sigma), \quad j = 0, \dots, T, \tag{4.8}$$

Fig. (4.6) shows that our model under only additive observations can still recover the correct network under this highly identifiable model. Similar to the previous case, uncertainties in the rate constants are, as expected, larger.

Prokaryotic auto-regulation

To evaluate the method on a larger reaction system with more possible reactions, we explore a simple synthetic model of auto-regulation of a protein P by a gene g in a

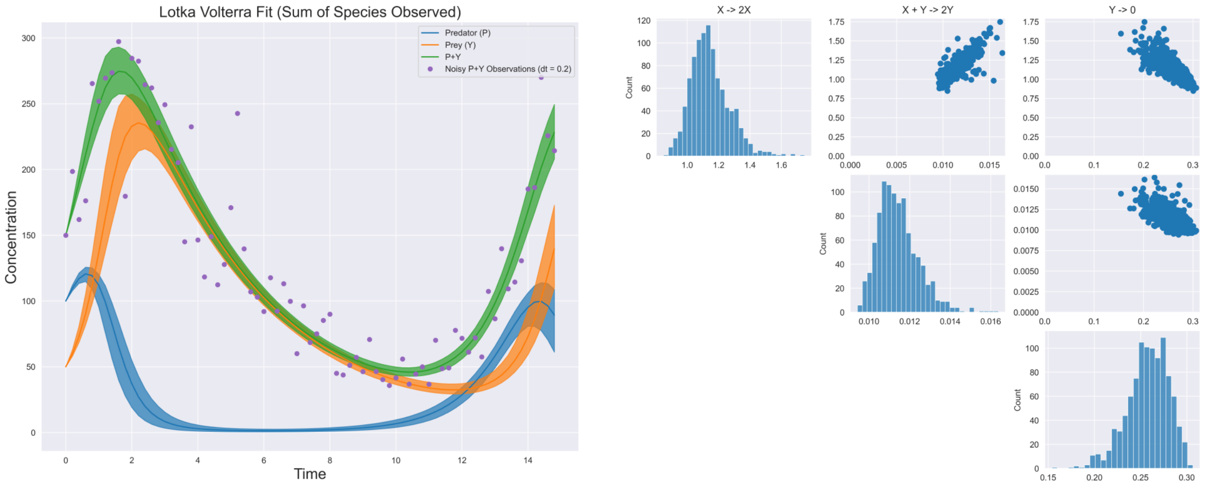


Figure 4.6: **Identified Trajectories and Posterior from Additive** Similar to the case of observing only one Y, the true network can still be recovered in this example however credible intervals are significantly larger.

prokaryotic cell [120]. The model is described by the following reaction system:

$$\begin{aligned} g + P_2 &\xrightarrow{k_1} gP_2 \quad \text{(Repression)} \\ gP_2 &\xrightarrow{k_2} g + P_2 \\ g &\xrightarrow{k_3} g + r \quad \text{(Transcription)} \\ r &\xrightarrow{k_4} r + P \quad \text{(Translation)} \\ 2P &\xrightarrow{k_5} P_2 \quad \text{(Dimerization)} \\ P_2 &\xrightarrow{k_6} 2P \\ r &\xrightarrow{k_7} \phi \quad \text{(mRNA Degradation)} \\ P &\xrightarrow{k_8} \phi \quad \text{(Protein Degradation)}, \end{aligned}$$

where gP_2 is the bound gene and r is the mRNA of protein P. Protein P represses its own transcription by binding to an available gene location. Denoting $X := \{g, P_2, gP_2, r, P\}$, we generate data from the system with parameters $k_1 = 0.5, k_2 = 1, k_3 = 0.15, k_4 =$

$1, k_5 = 0.5, k_6 = 0.5, k_7 = 1.5, k_8 = 0.3$ and initial conditions $X(t_0) := \{20, 20, 20, 20, 20, 20\}$ at dt $= 0.05$ for times in the interval $[0, 0.5]$. Furthermore, Log-Normal(0, 0.07) noise is added to the observations. At these parameter values, g and P decay rapidly, thus a small dt is required to provide sufficient information to the model. True trajectories and observations are shown in Fig. (4.7).

Figure 4.7: **Prokaryotic Auto-Regulation Observation Data** Simulated data used for the prokaryotic auto-regulation model. Log-normal observational noise is added to the true trajectory.

Using our library of ansatz reactions, we construct a complete stoichiometric matrix S_c of 260 possible reactions. The exact reactions included can be explored in the code repository. For estimation from the regularized horseshoe model in this problem, we set $\tau = 1e^{-6}$ and estimate c along with the other parameters by setting $c \sim$ Inv-Gamma$(5, 25)$. We run several MCMC chains to obtain results however only report the best two networks obtained for each experiment.

Including Known Reactions

To replicate the more common situation where the biologist has prior domain knowledge about the system under study, we explored the scenario where the first 4 reactions and rate parameters, k_1, k_2, k_3, and k_4, are known with confidence and the aim is to retrieve a system of reactions which replicates the observations, given these four known reactions. Below, we present the results of this setting to demonstrate a realistic situation where partial knowledge about system. The same experiment when no reactions are known is presented in the Appendix with similar results though converging to different sparse networks.

Table (4.2) lists the two selected networks obtained from MCMC chains with the rate constants set to the posterior median. Notably, each chain converges to different reaction pathways, neither of which are the true generating network. We note that, though we only present two networks here, our method was capable of producing several different reaction pathways with roughly the same number of reactions also capable of capturing the data.

Table 4.2: **Selected Recovered Networks for Prokaryotic Auto-Regulation System** The first 4 reactions are assumed to be known and the remaining reactions are to be inferred by the method.

True Network	Network 1	Network 2
$\mathbf{g + P_2 \xrightarrow{0.5} gP_2}$	$\mathbf{g + P_2 \xrightarrow{0.5} gP_2}$	$\mathbf{g + P_2 \xrightarrow{0.5} gP_2}$
$\mathbf{gP_2 \xrightarrow{1} g + P_2}$	$\mathbf{gP_2 \xrightarrow{1} g + P_2}$	$\mathbf{gP_2 \xrightarrow{1} g + P_2}$
$\mathbf{g \xrightarrow{0.15} g + r}$	$\mathbf{g \xrightarrow{0.15} g + r}$	$\mathbf{g \xrightarrow{0.15} g + r}$
$\mathbf{r \xrightarrow{1} r + P}$	$\mathbf{r \xrightarrow{1} r + P}$	$\mathbf{r \xrightarrow{1} r + P}$
$\mathbf{2P \xrightarrow{0.5} P_2}$	$2r \xrightarrow{0.05} P$	$\mathbf{2P \xrightarrow{0.5} P_2}$
$P_2 \xrightarrow{0.5} 2P$	$2P \xrightarrow{0.26} gP_2$	$2P_2 \xrightarrow{0.06} P$
$r \xrightarrow{1.5} \phi$	$P_2 + gP_2 \xrightarrow{0.04} P$	$gP_2 + r \xrightarrow{0.05} P_2 + gP_2$
$P \xrightarrow{0.3} \phi$	$P_2 + P \xrightarrow{0.4} 2P_2$	

As Fig. (B.1) demonstrates, although the reaction networks are different from the ground truth, the dynamics produced from each inferred reaction system appear plausible, especially given the noise present in data. Fig. (B.2) shows the posterior distributions of the non-zero reactions for both networks provided by our Bayesian approach. The marginals for each reaction rate in both cases are relatively tight, indicating that the reactions are well identified within in each discovered mode.

Reactive SINDy is also capable of inferring a network, however it is considerably less sparse and with larger reaction rates than those from our method. Under a threshold of $1e^{-2}$, selected such that thresholding larger reactions changes the dynamics, the best estimated network was comprised of 24 total reactions. The full network is detailed in the Supplementary Materials. As Fig. (B.1) demonstrates, though, the replicated trajectory is still consistent with the observations. In this example, the scale of the observed

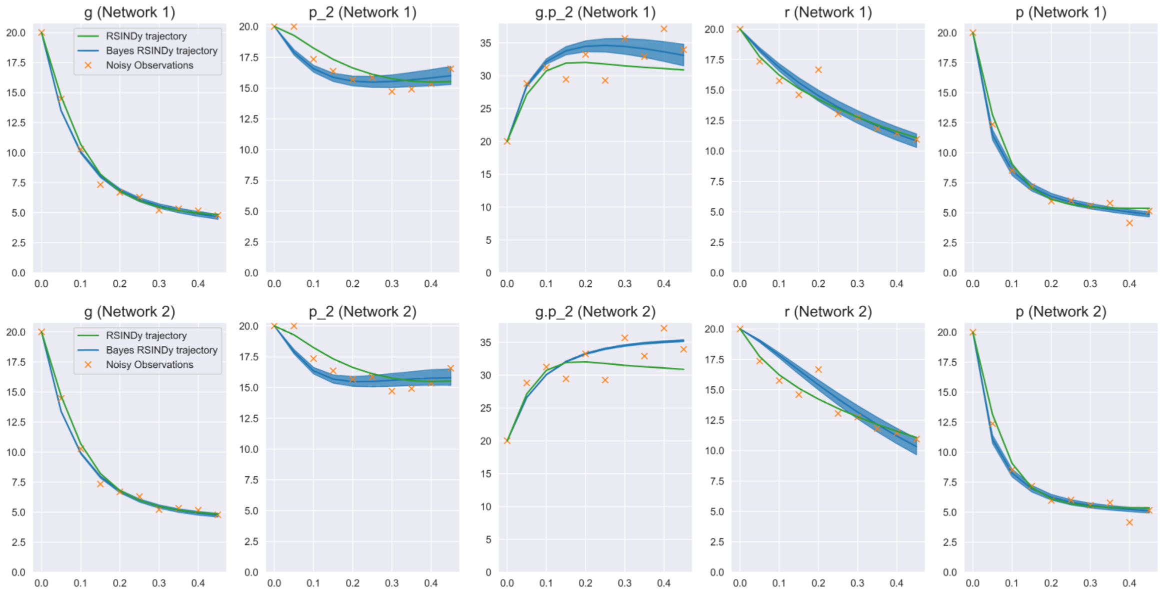


Figure 4.8: **Dynamics from the identified networks.** The dynamics from both recovered networks are different from the truth and each other, but still manage to produce plausible dynamics when compared to the noisy data. This points to an unidentifiability in the system, caused by noise in the data and structural identifiability issues.

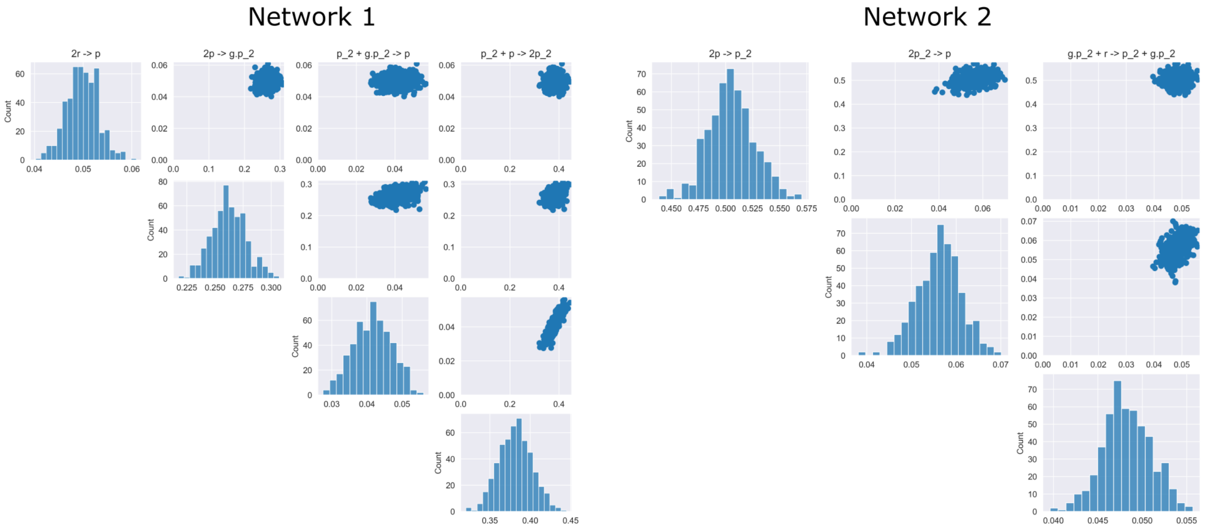


Figure 4.9: **Posterior Distributions over non-zero reaction rates** Pair plots of the two distinct reaction networks inferred by the model. Reaction rates within each network exhibit are relatively well determined. This indicates a distinct multi-modality or unidentifiability in the problem.

concentrations and the small observation frequency provide well estimated derivatives, resulting in minimal bias for the reactive SINDy method. Under these circumstances, it may be preferable to utilize Reactive SINDy as it can be run significantly faster than our method while still providing reasonable results as shown here.

That multiple networks are obtained by different chains in this problem is largely due to the facts that our complete stoichiometric matrix constructed from the above process does not restrict many reactions. In this, an iterative procedure can be applied, where the recovered networks can be examined by the user for plausibility and implausible reactions can be excluded in future runs to converge to a different reaction system. Realistically, we expect that the complete stoichiometric matrix will often be constructed in a more careful manner so as to eliminate many of the implausible reactions before the method is used. We discuss the identifiability issue in the next section. Interestingly, the inferred networks converge largely to 2nd-order reactions to describe the system. While from a combinatorial perspective, this is not surprising considering that the ansatz library contains significantly more 2nd-order reactions than 1st-order, a possibility is to add a bias to the system for 1st-order reactions via a prior weight on certain reactions.

4.5 Discussion

4.5.1 Observational Model

The latent variable formulation for the observational model provides robustness when observations are noisy or observations are not made for all of the species. In these situations, it is valid and desirable to use this model as it takes into account the true measurement process as demonstrated in the Lotka-Volterra example. However, this comes at a substantial computational cost. In some situations, when all of the species are observed and the measurements are not too noisy, simply using the Bayesian Regularized Horseshoe along with estimated derivatives as an extension to Reactive SINDy is sufficient and significantly faster to identify models. The model for this is:

$$\begin{aligned}
&\lambda_i \sim \text{Cauchy}^+(0,1), \qquad (4.9)\\
&\tilde{\lambda}_i = \sqrt{\frac{c^2\lambda_i^2}{c^2+(\tau\lambda_i)^2}},\\
&k_i \sim \text{Normal}(0, \tau\tilde{\lambda}_i), \quad i = 1, \dots, D.\\
\frac{d\hat{X}}{dt}(t_j) \sim \text{Normal}\Bigg(S_c^T &\begin{pmatrix} k_1 f_1(X(t_j)) \\ k_2 f_2(X(t_j)) \\ \dots \\ k_D f_D(X(t_j)) \end{pmatrix}, 1\Bigg), \quad j = 0, \dots, T,
\end{aligned}$$

where $\frac{d\hat{X}}{dt}$ is estimated numerically as previously discussed. This avoids the need to use an ODE solver and can provide Bayesian sparsity estimates similar to PTLasso [46]. We suggest that this method be used initially as it can often result in reasonable networks significantly faster.

4.5.2 Identifiability

A major problem in the identification of reaction systems is the possibility of multiple structural networks which can produce nearly identical results, especially when data is limited and noisy. While the sparsity priors used in this paper aim to resolve this situation by biasing estimates toward systems with fewer reactions, this remains an issue as multiple structural pathways may still exist with a very similar number of reactions. Immediately, this issue can be somewhat relaxed in a few ways.

First, constraining the allowed reactions will naturally bias the solutions away from certain pathways. However, this requires significant domain knowledge of the species or the system under observation. The work of Tuza et al. [112] presents one possible way to restrict the reaction basis to make the problem more identifiable while also using the LASSO with estimated derivatives. Another possibility would be to first pre-process the dictionary of functions to eliminate the indistinguishable graphs aided by the concept of linearly conjugate reaction systems such as demonstrated in acs2016computing. Further exploration in this direction is needed as their algorithm focuses on expanding a known reaction system into it's equivalents while we do not know the reaction graph at all. This can potentially automatically eliminate the structural unidentifiabilities in the problem before inferring the system. An interesting extension in this direction would be to use recent advances in Machine Learning (ML) to search the literature and generate a reasonable set of reactions given the species involved in the system [88, 82].

Alternatively, without introducing any domain knowledge, multiple MCMC chains can be used to explore all of the different networks. As each chain is trivially parallelizable, massive computational power can help explore the space more efficiently. Starting a large number of chains at different initial points will allow the chains to converge to different modalities and present it to the user as is in the case of the Prokaryotic Autoregulation model above. A user can prune implausible reaction networks and re-run the model to converge to better solutions. The use of ML techniques such as cross-validation on a hold-out test set to automatically rank networks based on predictive accuracy [113] could be useful but limitations in the amount of data may pose a problem.

As explored in Reactive SINDy, the incorporation of more data such as trajectories from multiple initial conditions can also aid in improving the identifiabiltiy of the process. Intuitively, this can be relevant in the case where certain dynamics are only present at certain concentration levels. In this case, a straightforward modification to the observational model where L independent trajectories are observed could be stated as follows,

$$\mathbf{Z}(t_j) = \int_0^{t_j} S_c^T \begin{pmatrix} k_1 f_1(Z(t)) \\ k_2 f_2(Z(t)) \\ \dots \\ k_D f_D(Z(t)) \end{pmatrix} dt$$

$$\hat{\mathbf{X}}_{\mathbf{l}}(t_j) \sim \text{Log-Normal}(\mathbf{Z}(t_j), \sigma), \quad j = 0, \dots, T \qquad l = 1, \dots, L,$$

where $\hat{\mathbf{X}}_{\mathbf{l}}(t_j)$ refers to the observed species concentrations at time t_j for the l-th trajectory.

4.5.3 Future Directions & Limitations

Scaling

As the number of species grows, the number of possible reactions grows combinatorially. This poses a significant issue computationally, as it results in a large search space for reactions and possibly further identifiability issues as demonstrated above. The scaling issue limits the applicability of the method to systems with a small number of active species. One possibility is to run the method on smaller subsets of reactions to prune reactions in a sequential procedure. However, this may lead to bias issues as combining the estimates from different subsets is a non-trivial problem, especially if dealing with partial posterior distributions. Practically, biological domain knowledge can substantially help here in limiting the allowed reactions in the system or specifying known reactions as in Example 2.

Computationally, the latent variable approach with Bayesian Inference is significantly more expensive than the approach used by reactive SINDy. A large part of this is the need to compute the sensitivities of the ODE system to obtain efficient sampling. In our experiments, the auto regulatory network with 260 reactions took approximately 4 hours of time on a M1 Apple ARM processor using our approach while roughly 1.5 hours to perform a large grid search using reactive SINDy. We find that this difficulty generally scales as a function of the number of data points in addition to the number of possible reactions. A possibility on this end is to utilize Variational Inference to speed up the inference component as presented in [36]. Furthermore, there are a few different methods for computing the sensitivities of ODE systems as well as a variety of different ODE solvers [15] that may potentially offer speedups for these types of problems. For our experiments we employ the rk45 solver and a forward sensitivity solver as implemented by Stan.

Hyper-parameter selection of τ

Selection of τ determines the level of sparsity of these networks and, in our experience, is a pivotal hyperparameter to tune when using the horseshoe prior. Generally, we find that smaller values of τ will force the near-zero reaction rates to smaller values however, this typically leads to a significant decrease in computational efficiency when estimating the networks. For this reason, through our experiments and set τ to a small enough value such that the above pruning procedure removes a large enough set of reactions while maintaining the dynamics.

Further work exploring how to properly tune and select τ in a more interpretable way for reaction network inference problems is needed. A common strategy employed in other models is to place the prior $\tau \sim \text{Cauchy}^{+}(0, \tau_0)$ to allow the data to adjust τ [77], however this needs to be further explored in the context of the horseshoe for systems of differential equations. For linear regression models, Piironen et al. [76] propose a way to parameterize τ_0 as,

$$\tau_0 = \frac{m_0 \sigma}{(D - m_0)\sqrt{NM}},$$

where m_0 can be derived as a guess for the effective non-zero coefficients and σ is the measurement noise, however our models deviate from linear regression and thus the same interpretations do not hold.

A common concern with Bayesian methods is whether the prior can be overcome with sufficient data. While in our experience, the utilized horseshoe priors are weakly informative, and indeed can be overcome with sufficient data to obtain the true network, however more rigorous study needs to be done for this. The particular case study demonstrated by Golchi et al. [40] offers good insight into the strength and importance of priors in the context of ODEs though further investigation needs to be done with respect our model and for network inference.

Stochastic Models

Many biochemical reaction systems exhibit intrinsic stochasticity. In these situations, Eq. (4.1) no longer sufficiently captures the dynamics of $X(t)$ and the evolution of the system is better described using a stochastic process. While mass-action kinetics can still be applied, they now specify reaction propensities. To accommodate this, Eq. (4.5) can be modified from the observational ODE model,

$$\begin{aligned} \mathbf{Z}(t_j) &\sim P(\mathbf{Z}(t_j)|\mathbf{Z}(t_{j-1}), S_c^T, \mathbf{k}) \\ \hat{\mathbf{X}}(t_j) &\sim P(X|\mathbf{Z}(t_j)), \qquad j = 0, \ldots, T. \end{aligned}$$

where the trajectory $\mathbf{Z}$ comes from the stochastic process as specified by [39] while the regularized horseshoe and S_c remain as previously defined. However, the significant challenge here is that the posterior distribution becomes intractable due to the intractable likelihood term $P(\mathbf{Z}(t_j)|\mathbf{Z}(t_{j-1}, S_c^T, k))$, which corresponds to the solution of the chemical master equation [71]. This prevents the application of standard efficient Bayesian inference methods, which are heavily reliant on tractable likelihoods.

While there is a growing class of likelihood-free Bayesian inference methods [19] that can be applied to stochastic biochemical reaction networks, they are known to scale incredibly poorly to high dimensional parameter spaces. This makes it quite challenging to utilize with our method of network inference, which introduces a new parameter for each ansatz reaction. A possibility is to instead use stochastic approximations to the model, such as the Chemical Langevin Equation or the Linear Noise Approximation, to capture some intrinsic stochasticity, but also provide much more tractable likelihoods [39, 28, 41].

4.6 Conclusion

In this work, we have presented a method to recover a parsimonious system of interpretable mass-action reactions directly from observations of species concentrations over time. Improving on the formulation presented by Reactive SINDy, we have modified the method via the Bayesian regularized horseshoe prior and by adapting the model as to not require derivative estimates. Our experiments show that, when identifiable, our modifications are able to recover the underlying system with uncertainty estimates from the Bayesian formulation even in sparse data scenarios. Alternatively, when unidentifiable, we present multiple sparse reaction networks which can reasonably explain the results and upon which a biologists can iterate.

Chapter 5

StochSS Live! for Epidemiological Modeling

Although most of this thesis focuses on biochemical reaction networks that capture a biological process, the techniques and models used to study these reaction systems are widely applicable to a number of fields, including epidemiology. In this chapter we present our work *Epidemiological Modeling in StochSS Live!* [101], which demonstrates the adaptation of stochastic biochemical reactions networks to epidemiological models. Specifically, we show how one can develop, implement, and calibrate a stochastic epidemiological model using our software package *StochSS Live!*. The techniques are illustrated using an example of specifying and fitting a custom stochastic epidemiological model using Approximate Bayesian Computation to COVID-19 epidemic data in the Santa Barbara and Buncombe counties. The results demonstrate the flexibility of our software and methods for modeling different datasets, in addition to the heterogeneity of the resulting parameter estimates in different regions.

5.1 Introduction

Epidemiological models are essential tools to assist public health authorities in the planning of policy responses to pandemic prevention and control [107]. In general, these models are classified into different categories (deterministic/stochastic), treatment of the populations (continuous/discrete) or spatial dependence, and distribution of the population (homogeneous/heterogeneous) [33].

An example of a recent application of epidemiological modeling is the study of early transmission dynamics and effectiveness of control measures in individuals infected by the novel coronavirus disease (COVID-19). As of September 7, 2020, COVID-19 has been responsible for over 27 million reported cases and 900,000 deaths worldwide [70]. Given the global impact of the virus, several software tools have been developed, mostly focused on either deterministic [31] or stochastic [100] models. These tools typically require some level of technical expertise.

On the mathematical level, most epidemiological models are structurally identical to models of chemical kinetics widely used in systems biology. In the systems biology community, there has been a large focus the last decade on increasingly efficient stochastic simulation algorithms and on tools to improve usability for modelers. We have in previous work developed a wide range of model development and simulation tools for such models in the *StochSS Suite of Software.* We believe that there is great urgency and potential for a software environment that makes epidemiological modeling easily accessible to a wide audience, and that bridges the notation gap needed to effectively re-use simulation tools from systems biology for epidemiological models. To accomplish this we present *StochSS Live!*, a powerful web-based tool that enables users to create models, perform simulations, infer parameters and visualize the results through simple and intuitive workflows, and have developed a stochastic COVID-19 epidemiological model accessible via *StochSS*

Live!.

We present *StochSS Live!*, a web-based service for modeling, simulation, and analysis of a wide range of mathematical, biological and biochemical systems. Using an epidemiological model of COVID-19, we demonstrate the power of *StochSS Live!* to enable researchers to quickly develop a deterministic or a discrete stochastic model, infer its parameters, and analyze the results.

Availability: *StochSS Live!* is freely available at `https://live.stochss.org/`.

Supplementary information: Available at `https://github.com/StochSS/Covid19_Modeling`

StochSS Live! enables easy access to the powerful feature set of the simulation and model analysis toolkits in the *StochSS Suite of Software* [1, 94]. *StochSS Live!* builds on and extends the model development UI from [25] in several ways: Through a set of clear, user friendly interfaces used directly from a web browser (hence requiring no installation), a researcher can explicitly define their model, simulate it using deterministic or stochastic solvers, analyze and explore the parameter space using either traditional parameter sweeps or workflows guided by unsupervised machine learning. Users can also calibrate the model to observed data using highly scalable likelihood-free parameter inference. For analysis needs that goes beyond the capability of the UI, *StochSS Live!* will automatically generate templated Jupyter notebooks that can be shared and extended. This automated dual representation of models and computational workflows via a UI and as code is a defining feature of *StochSS Live!* and greatly simplifies collaboration between domain and computational experts.

The *StochSS Suite of Software* encompasses a hierarchy of open source mathematical toolboxes which allows a researcher maximum flexibility in modeling and analyzing their systems and assumptions. GillesPy2 provides common interfaces and a plethora of advanced solvers for ordinary differential equations (ODE) and discrete stochastic simulations[1]. Spatial extensions to the same models are provided via SpatialPy [1]. For analysis of any implemented model, the Sciope toolbox [95] provides algorithms for efficient model exploration [123] and parameter estimation.

5.2 Epidemiological Model

To demonstrate the use of *StochSS Live!* for epidemiological modeling, we consider the infection dynamics of COVID-19 in two U.S. counties: Santa Barbara, CA, and Buncombe, NC. The data was gathered from Santa Barbara's health department and [24], between March 13 - August 31, 2020. We construct an extended SEIRD model with symptomatic and asymptomatic compartments using the *StochSS Live!* model builder, as shown in Fig. 5.1-A. We divide the population into 7 groups: susceptible, exposed, infected, symptomatic, recovered, deceased, and cleared individuals. Transition events between these groups are shown in Fig. 5.1-A. The user can immediately preview sample trajectories from either deterministic or stochastic versions of the system simply by selecting the respective option (Fig. 5.1-B).

Figures 5.1-C, D show the results of parameter inference for a discrete stochastic version of the model. Inference is performed using Approximate Bayesian Computation, allowing for uncertainty quantification of parameters and predictions. In Fig. 5.1-C, each realization (blue lines) corresponds to a simulation using a parameter sample from the posterior distribution, which are contrasted with the data (black lines). Fig. 5.1-D, shows the posteriors for parameters for both counties. While infectivity rates between the two counties are roughly the same, the estimated lethality rate is a bit higher in Buncombe county, although there is substantial uncertainty. We do note that this particular model does not seem to sufficiently capture the data as evidenced in Fig. 5.1-C and would need

[1] https://github.com/StochSS/SpatialPy

to be further iterated upon before any strong conclusions can be drawn.

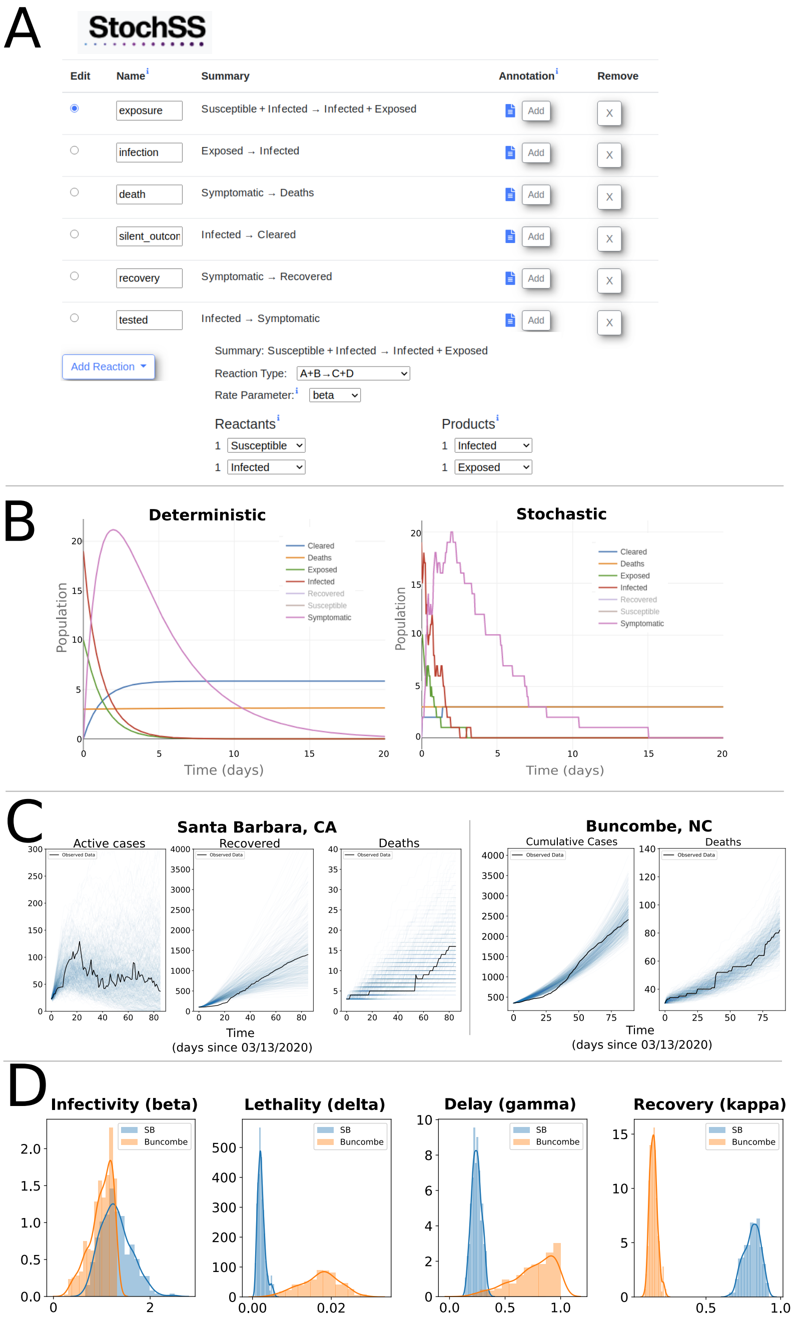
5.2.1 Model Details and Implementation

The epidemiological model we implement is an extended version of the SEIRD model that accounts for symptomatic and asymptomatic cases. The involved compartments (species) are: susceptible (S), exposed (E), infected (I), symptomatic (Y), recovered (R), dead (D), and cleared (C). The compartmental system can be visualized in Fig. 5.2.

The system evolves according to SEIR dynamics but with a chance of becoming symptomatic after being exposed. We fix the rate at which exposed patients become infectious at 0.16, which represents 6.25 day incubation period and estimate the proportion of patients who become infected vs. symptomatic. This is roughly adopted from a similar model [32]. Specifically, we implement the following set of reactions:

$$\begin{aligned}
\text{Susceptible} + \text{Infected} &\overset{\beta/N}{\rightarrow} \text{Infected} + \text{Exposed}\\
\text{Exposed} &\overset{0.16(1-\nu)}{\rightarrow} \text{Infected}\\
\text{Exposed} &\overset{0.16\nu}{\rightarrow} \text{Symptomatic}\\
\text{Symptomatic} &\overset{\kappa}{\rightarrow} \text{Recovered}\\
\text{Symptomatic} &\overset{\delta}{\rightarrow} \text{Dead}\\
\text{Infected} &\overset{\kappa}{\rightarrow} \text{Cleared}
\end{aligned}$$

This model assumes that only asymptomatic transmission is possible, all asymptomatic cases recover, and that all parameters are static. Of importance is that the presented models do not really indicate a sufficient fit to the data to draw any strong



Edit	Name	Summary	Annotation	Remove
●	exposure	Susceptible + Infected → Infected + Exposed	Add	X
○	infection	Exposed → Infected	Add	X
○	death	Symptomatic → Deaths	Add	X
○	silent_outcor	Infected → Cleared	Add	X
○	recovery	Symptomatic → Recovered	Add	X
○	tested	Infected → Symptomatic	Add	X

Summary: Susceptible + Infected → Infected + Exposed

Reaction Type: A+B→C+D

Rate Parameter: beta

Reactants
1 Susceptible
1 Infected

Products
1 Infected
1 Exposed

Figure 5.1: Snapshot of the *StochSS Live!* web interface. (A) The user can explicitly define populations, parameters and reactions. (B) The preview window settings allow the user to preview simulation results for both deterministic and stochastic models. (C) Example of parameter sweep inference in *StochSS Live!*. The blue lines are computed realizations obtained by the stochastic solver, and the black lines correspond to the official data. Notice that, regardless of the fact that data from Buncombe and Santa Barbara counties have different scales and different levels of stochasticity, *StochSS Live!* is capable of modeling both cases. (D) Comparison of posteriors from Santa Barbara and Buncombe counties.

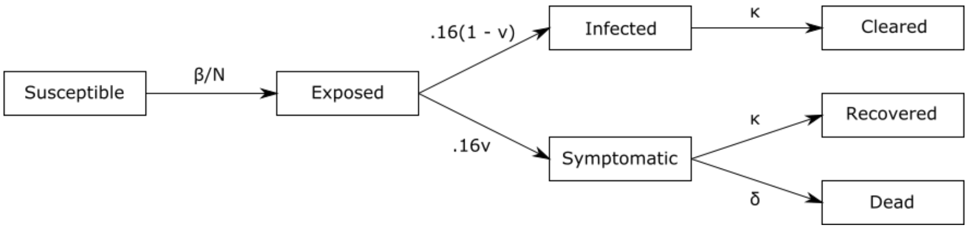


Figure 5.2: SEIYRDC Compartmental Model

conclusions. For a complete analysis, this process needs to be repeated, changing the model to better capture assumptions about the system. For example, we would expect the infectivity to change over time as policies are implemented and we know that there are non-intrinsic measurement error, such as reporting errors in the data.

The parameters of the stochastic models are estimated using the Replenishment Approximate Bayesian Computation algorithm [26] with a separate unpooled fit for each of the two counties. The following common priors are assigned to the unknown parameters:

$$\beta \sim U(0,3), \qquad \kappa \sim U(0,1)$$
$$\delta \sim U(0,0.1), \qquad \nu \sim U(0,1).$$

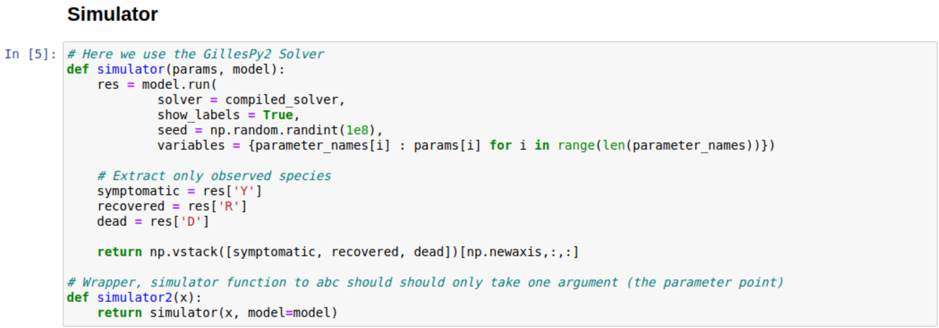
but the observational models differ due to the nature of the data that is recorded. Specifically, for Santa Barbara, which records the symptomatic ($\tilde{Y_t}$), recovered ($\tilde{R_t}$), and dead($\tilde{D_t}$) cases, we use the following observational model where we denote $\tilde{Y_t}$ as the observed species at time t and the probabilistic model corresponds to the biochemical reaction network with stochastic dynamics from above:

$$\tilde{Y_t}, \tilde{R_t}, \tilde{D_t} \sim Y_t, R_t, D_t | \beta, \kappa, \delta, \nu \qquad t = 0, \cdots, T.$$

In comparison, in Buncombe county, where only total active case $\tilde{A}_t$ and deaths $\tilde{D}_t$ are recorded, we instead have the observational model for total active cases $\tilde{A}_t$ as:

$$\hat{Y}_t, \hat{R}_t, \tilde{D}_t \sim Y_t, R_t, D_t | \beta, \kappa, \delta, \nu \qquad t = 0, \cdots, T$$
$$\tilde{A}_t = \hat{Y}_t + \hat{Y}_t.$$

The effect of this is that the observations from Buncombe county are less informative of certain parameters, as demonstrated in the parameters estimated. However, due to the purely simulation based nature of Approximate Bayesian Computation, such an observational model is easy to implement without the need to derive likelihoods. The summary statistics used for inference are the normalized euclidean distances between the observations.

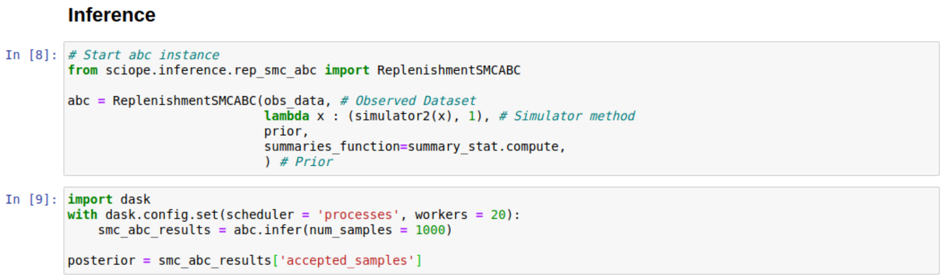


Simulator

```
In [5]: # Here we use the GillesPy2 Solver
        def simulator(params, model):
            res = model.run(
                    solver = compiled_solver,
                    show_labels = True,
                    seed = np.random.randint(1e8),
                    variables = {parameter_names[i] : params[i] for i in range(len(parameter_names))})

            # Extract only observed species
            symptomatic = res['Y']
            recovered = res['R']
            dead = res['D']

            return np.vstack([symptomatic, recovered, dead])[np.newaxis,:,:]

        # Wrapper, simulator function to abc should should only take one argument (the parameter point)
        def simulator2(x):
            return simulator(x, model=model)
```

Figure 5.3: StochSS Live! simulation code

Through *StochSS Live!*, the model can be easily specified as shown in Fig. 5.3. In this, we show the code for Santa Barbara county, which allows for partial observations but no sum type observations as in Buncombe county. Parallelized inference using Replenishment Approximate Bayesian Computation as mentioned above can be accomplished by passing in the simulator, observed data, and the summary statistic as shown in Fig. 5.4.



Inference

```
In [8]: # Start abc instance
        from sciope.inference.rep_smc_abc import ReplenishmentSMCABC

        abc = ReplenishmentSMCABC(obs_data, # Observed Dataset
                                  lambda x : (simulator2(x), 1), # Simulator method
                                  prior,
                                  summaries_function=summary_stat.compute,
                                  ) # Prior
```

```
In [9]: import dask
        with dask.config.set(scheduler = 'processes', workers = 20):
            smc_abc_results = abc.infer(num_samples = 1000)

        posterior = smc_abc_results['accepted_samples']
```

Figure 5.4: StochSS Live! ABC code

In Fig. 5.5 we show the pairs plot of the posterior parameters for Santa Barbara obtain from this process. Within our specified prior, the parameters are relatively well identified.

However, in the case of Buncombe county, as shown in Fig. 5.6, this is much less the case. This is in large part due to the cumulative observations of total active case as opposed to the split observations that we obtained from Santa Barbara county data.

5.3 Conclusion

We have presented the epidemiological modeling capabilities of StochSS, a freely available, user-friendly platform for stochastic and deterministic simulations. *StochSS Live!* and the *StochSS Suite of Software* offer users a unique modeling experience by providing an integrated, web-based, solution that addresses model specification on multiple levels, features state-of-the-art simulation algorithms for efficient simulation, and removes the barrier of scaling computational resources when needed.

Our model of COVID-19 demonstrates epidemiological capabilities of *StochSS Live!*, a freely available, user-friendly web-based service for the development, simulation and analysis of a wide range of models. To make these capabilities as widely accessible as possible, we provide *StochSS Live!*. In addition, the *StochSS Suite of Software* provides

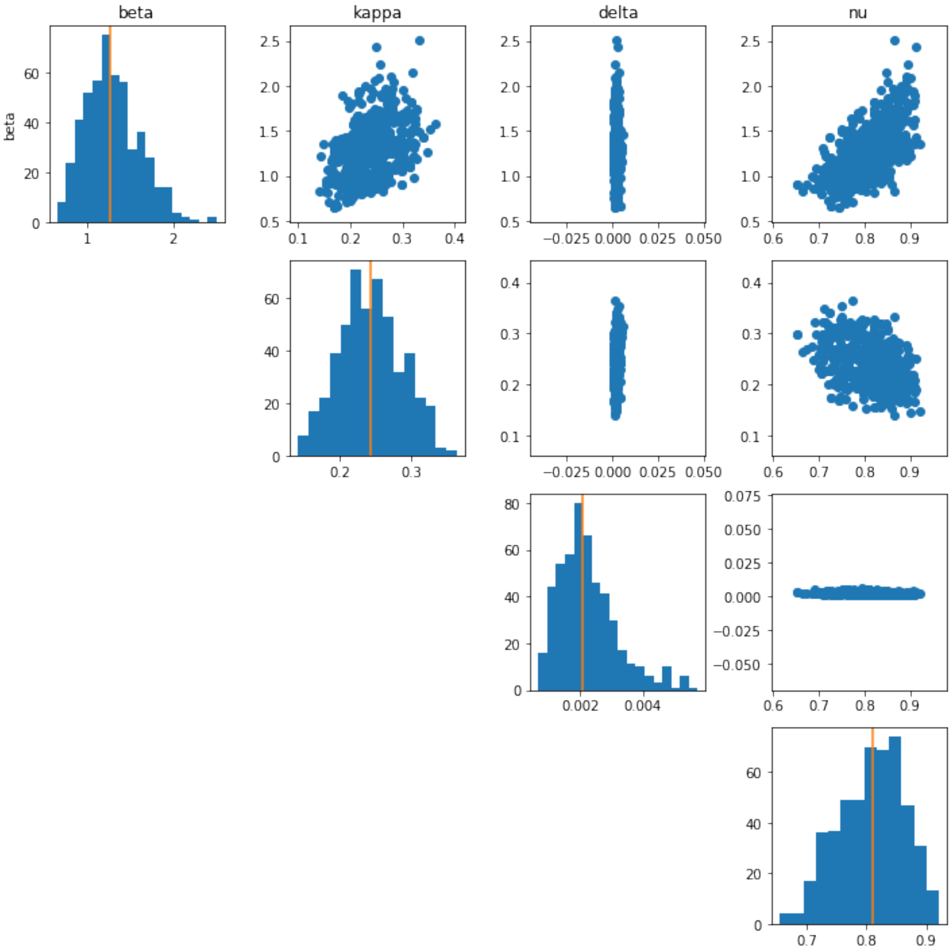


Figure 5.5: Pair plot for parameters calibrated to Santa Barbara data

the individual tools, if you wish to integrate them into your own software.

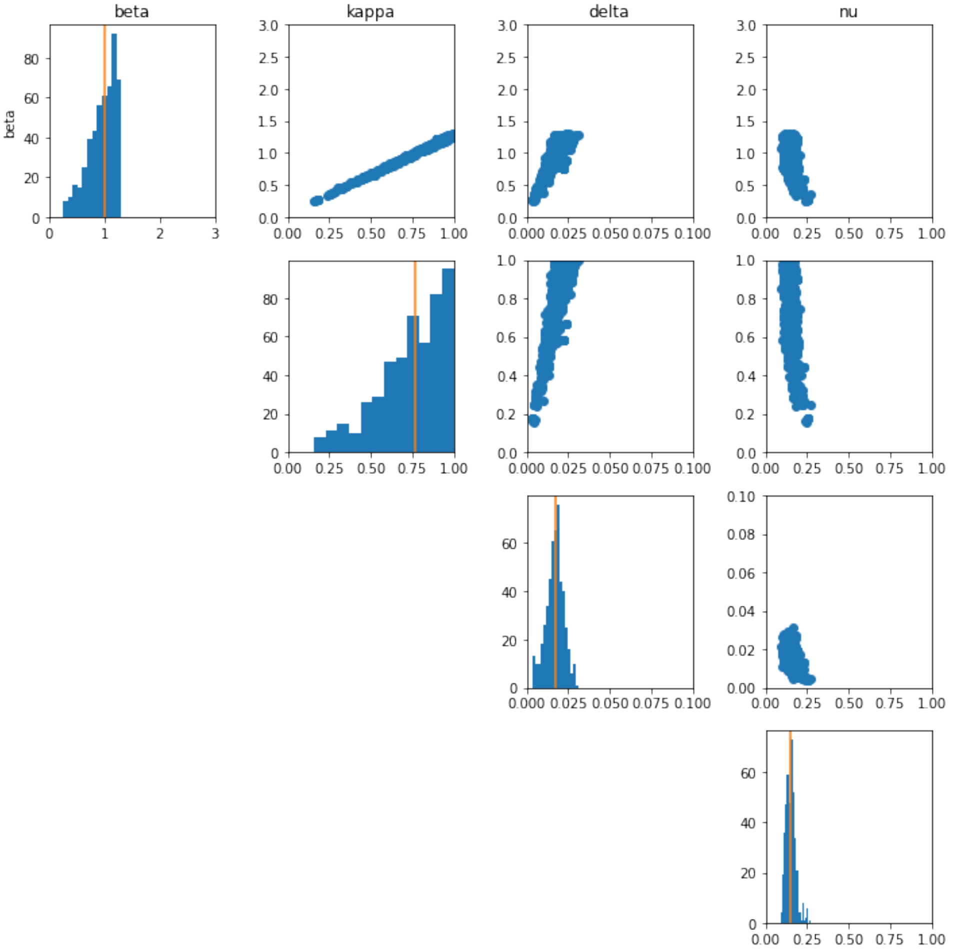


Figure 5.6: Pair plot for parameters calibrated to Buncombe County data

Chapter 6

Conclusion

Throughout this thesis, we have described some of our work applying and improving data-driven Bayesian methods for analyzing biochemical reaction networks, as well as offering some insights into the motivating problems behind these solutions. We have demonstrated improvements and new methods for calibrating the mathematical models describing these complex systems and discussed some of the additional open challenges of inference within this domain. Furthermore, we have provided many open implementations of these methods and algorithms.

We first presented our applied work on quantifying the survival risks based on the levels of several specific blood proteins using Bayesian joint longitudinal models. In this, we demonstrated how data-driven Bayesian techniques could help in understanding biochemical processes via protein assays at the clinical scale. Then, transitioning to the molecular scale, we described our innovations on accelerating Approximate Bayesian Computation (ABC) for inferring the parameters of stochastic biochemical reaction networks. Next, we presented our method on the complimentary problem of inferring the structure of biochemical reaction networks from data through the use of sparse Bayesian method and a latent variable model. Finally, we closed with a short demonstration, along with accompanying open source software, on applying the framework of stochastic biochemical reaction networks and Approximate Bayesian Computation to model and calibrate epidemiological models with intrinsic stochasticity.

There are several remaining challenges in adopting data-driven Bayesian methods for studying biochemical reaction networks. For one, though we propose computational methods, most problems, and particularly larger scale problems, still require significant computational effort. ABC remains a struggle to scale to systems with large numbers of parameters and the use of approximate models needs to be carefully investigated for bias when used for sampling. Identifiability, especially with the prominence of sparsely measured and noisy datasets, can be a significant issue when trying to come to reasonable biological conclusions, though the Bayesian method of introducing prior knowledge is helpful for resolving this. Using and testing these methods under realistic experimental assumptions would substantially help in discovering new practical tools. Finally, the investigation of more recent Deep Learning and Machine Learning tools may help in resolving some of these problems.

Appendix A

Accelerated Regression-Based Summary Statistics for Discrete Stochastic Systems

Pure-Birth Process

The Pure-Birth Process is represented as

$$\phi \xrightarrow{k} S.$$

As this is simply a homogenous Poisson process, we can evaluate the likelihood of an observation at any time t as

$$P(S(t)|S(0),k) = \frac{k^{S(t)-S(0)}e^{-k}}{(S(t)-S(0))!}.$$

We assign prior $k \sim \mathcal{U}(0, 10000)$ and observations are made of S at times $t = \{1 : 100 : 1\}$. We train the ratio estimator using $M = 300$ samples from both the SSA and the Tau-Leaping approximation. The summary statistic is trained using $N = 5000$ samples. The posterior in the main text is obtained using $k = 2432$.

Lotka-Volterra Stochastic Oscillator

The Lotka-Volterra Stochastic Oscillator is described by

$$\begin{aligned} S_1 + S_2 &\xrightarrow{k_1} 2S_1 + S_2 & S_1 &\xrightarrow{k_2} \phi \\ S_2 &\xrightarrow{k_3} 2S_2 & S_1 + S_2 &\xrightarrow{k_2} S_2. \end{aligned}$$

We assign the following priors

$$\begin{aligned} \log(k_1) &\sim \mathcal{U}(-6, 2) & \log(k_2) &\sim \mathcal{U}(-6, 2) \\ \log(k_3) &\sim \mathcal{U}(-6, 2) & \log(k_4) &\sim \mathcal{U}(-6, 2), \end{aligned}$$

and observations are made of both S_1 and S_2 at times $t = 0 : 30 : 0.2$, for a total of 150 time steps. We train the ratio estimator using $M = 3000$ samples from both the SSA and the ODE approximation. The summary statistic is trained using $N = 100000$ samples. The posterior in the main text is obtained from $\mathbf{k} = [0.01, 0.5, 1.0, 0.01]$, giving oscillatory behavior.

Genetic Toggle-Switch

The Genetic Toggle-Switch is described as

$$\phi \overset{\frac{\alpha_1}{1+V^{\beta}}}{\to} U \qquad\qquad \phi \overset{\frac{\alpha_2}{1+V^{\gamma}}}{\to} V$$

$$U \overset{\mu}{\to} \phi \qquad\qquad V \overset{\mu}{\to} \phi.$$

We assign the following priors

$$\alpha_1 \sim \mathcal{U}(0,6) \qquad \alpha_2 \sim \mathcal{U}(0,6)$$

$$\beta \sim \mathcal{U}(0,6) \qquad \gamma \sim \mathcal{U}(0,6) \qquad \mu \sim \mathcal{U}(0,6),$$

and observations are made of both U and V at times $t = 0 : 50 : 0.25$, for a total of 200 time steps. We train the ratio estimator using $M = 5000$ samples from both the SSA and the Tau-Leaping approximation. The summary statistic is trained using $N = 100000$ samples.

Vilar-Oscillator

The Vilar-Oscillator is described as in [114]. We assign the following priors to the parameters:

$$\alpha_A \sim \mathcal{U}(0,80), \qquad \alpha'_A \sim \mathcal{U}(100,600), \qquad \alpha_R \sim \mathcal{U}(0,4), \qquad \alpha'_R \sim \mathcal{U}(20,60),$$

$$\beta_A \sim \mathcal{U}(10,60), \qquad \beta_R \sim \mathcal{U}(1,7), \qquad \delta_{MA} \sim \mathcal{U}(1,12), \qquad \delta_{MR} \sim \mathcal{U}(0,2),$$

$$\delta_A \sim \mathcal{U}(0,3), \qquad \delta_R \sim \mathcal{U}(0,0.7), \qquad \gamma'_A \sim \mathcal{U}(0.5,2.5), \qquad \gamma_R \sim \mathcal{U}(0,4),$$

$$\gamma_C \sim \mathcal{U}(0,3), \qquad \theta_A \sim \mathcal{U}(0,70), \qquad \theta_R \sim \mathcal{U}(0,300),$$

and observations are made of species C, A, and R at times $t = \{0 : 100 : 1\}$. Most parameters are poorly identified under these settings [3]. To simulate more realistic conditions, we also perturb the ODE trajectories with log-normal noise. This prevents the ratio estimator from overfitting to the smooth ODE solutions, as mentioned in the main text.

We train the ratio estimator using $M = 10000$ samples from both the SSA and the noise added ODE approximation. The summary statistic is trained using $N = 200000$ samples to more thoroughly explore the high dimensional parameter space.

Neural Network Architectures

Table A.1: Neural Network Architectures

	Ratio Estimator	Summary Statistics
Pure-Birth	CNN	CNN
Lotka-Volterra	MLP(50,50)	CNN
Genetic Toggle Switch	CNN	CNN
Vilar Oscillator	CNN	CNN

In Table A.1 we list the details to train the approximate ratio estimator and the summary statistic. The referenced CNN follows the construction from [3] while the referenced Multi-Layer Perceptron (MLP) specifies the number of hidden neurons with ReLU activation functions. For all experiments, we implement the model in PyTorch and use the Adam Optimizer with an exponential learning rate scheduler to train the Neural Networks.

Appendix B

Bayesian Systems Identification of Mass-Action Biochemical Reaction Networks

Construction of Ansatz Reactions

To construct the library of ansatz reactions used in our experiments, we use the following naive algorithm, which can be found implemented in the github repository:

1. For each species S_i, generate all reactions of type $S_i \rightarrow 0$

2. For each species S_i, generate all reactions of type, $j, k \neq i$:

    - $S_i \rightarrow S_j$

    - $S_i \rightarrow S_i + S_j$

    - $2S_i \rightarrow S_j$

    - $S_j \rightarrow 2S_i$

- $S_i \to 2S_i$
- $S_i \to S_j + S_k$

3. For each pair of species S_i and S_j, generate all reactions of type $k, l \neq i, j$:

- $S_i + S_j \to 2S_i$
- $S_i + S_j \to 2S_j$
- $S_i + S_j \to S_k$
- $S_i + S_j \to S_i + S_k$
- $S_i + S_j \to S_j + S_k$
- $S_i + S_j \to S_k + S_l$

Experimental Details

Lotka-Volterra Oscillator

The Lotka-Volterra Oscillator is described by

$$
\begin{aligned}
Y &\xrightarrow{k_1} 2Y \\
P + Y &\xrightarrow{k_2} 2P \\
P &\xrightarrow{k_3} \phi,
\end{aligned}
$$

where P represents the predator concentration in an area and Y represents the prey concentration. This is one of the simplest non-linear systems to exhibit oscillatory behavior and is often a building block for such systems. We generate data from this system by solving the corresponding ODE and then adding independent log-normal noise with standard deviation $\sigma = 0.2$.

To test our method under varying sampling frequencies, we first generate data, recording observations every $dt = 0.2$. Then, given this time series, we take every 5th observation to obtain a sampling frequency of $dt = 1$ and every 10th observation to obtain a sampling frequency of $dt = 2$.

Stan Model for Regularized Horseshoe of Lotka-Volterra Model

```
functions {
  vector sys(real t,
             vector y,
             vector theta) {
      vector[2] dydt;
      vector[16] v;
      matrix[2, 16] S = [
        [-2, 0, 1,-1, 0, 1,-1, 0, 0,-1, 0,-1,-2,-1, 1,-1],
        [ 0,-2, 0, 1,-1,-1, 0,-1, 1, 0,-1, 0, 1, 1,-1, 2]
      ];
      v[1] = theta[1] * y[1] * y[1];
      v[2] = theta[2] * y[2] * y[2];
      v[3] = theta[3] * y[1];
      v[4] = theta[4] * y[1]* y[2];
      v[5] = theta[5] * y[2];
      v[6] = theta[6] * y[1]* y[2];
      v[7] = theta[7] * y[1];
      v[8] = theta[8] * y[2] * y[2];
      v[9] = theta[9] * y[2];
      v[10] = theta[10] * y[1] * y[1];
      v[11] = theta[11] * y[1]* y[2];
      v[12] = theta[12] * y[1]* y[2];
      v[13] = theta[13] * y[1] * y[1];
      v[14] = theta[14] * y[1];
      v[15] = theta[15] * y[2];
      v[16] = theta[16] * y[1];

      dydt = S * v;
      return dydt;
  }
}
data {
    int N; // Number of observations
    int M; // Number of species
    int M_obs; // Observed species
    int obs_idx[M_obs]; // Indices of observed speces

    int D; // Number of possible reactions
    int D1; // Number of known rates

    vector[M] y0;
    real y[N, M_obs];
    real ts[N + 1];

    vector[D1] known_rates;

    // horseshoe parameters
    real m0;
    real slab_scale;
    real slab_df;
    real<lower = 0> tau0;


    // noise model parameters
    real<lower = 0> noise_sigma;

}

transformed data {
    real slab_scale2 = square(slab_scale);
    real half_slab_df = 0.5 * slab_df;
}

parameters {
    vector<lower = 0>[D - D1] unknown_rates_tilde;
    vector<lower = 0>[D - D1] lambda;
    real<lower = 0> c2_tilde;
}

transformed parameters {
    vector[D] rates;
    real c2;
    real tau;
    vector[D - D1] lambda_tilde;
    vector[M] y_hat[N];
    {
        tau = tau0;

        c2 = slab_scale2 * c2_tilde;

        lambda_tilde = sqrt((c2 * square(lambda)) ./
        (c2 + square(tau) * square(lambda)));

        if(D1 > 0) {
          rates[:D1] = known_rates;
        }
        rates[D1 + 1:] = tau * lambda_tilde .* unknown_rates_tilde;
    }
    y_hat = ode_rk45(sys,
                     y0,
                     ts[1],
                     ts[2:],
                     rates);
}

model {
    // horseshoe priors
    unknown_rates_tilde ~ normal(0, 1);
    lambda ~ cauchy(0, 1);
    c2_tilde ~ inv_gamma(half_slab_df, half_slab_df);

    // model likelihood
    for(j in 1:M_obs) {
      y[ ,j] ~ lognormal(log(y_hat[ ,obs_idx[j]]), noise_sigma);
    }
}

generated quantities {
  real y_rep[N, M];

  for(i in 1:N) {
    for(j in 1:M) {
      y_rep[i,j] = lognormal_rng(log(y_hat[i,j]), noise_sigma);
    }
  }
}
```

Prokaryotic Auto-Regulatory Network

A simple synthetic model of auto-regulation of a protein P by a gene g in a prokaryotic cell [120] can be described using the following reaction system:

$$\begin{aligned}
g + P_2 &\xrightarrow{k_1} gP_2 \quad \text{(Repression)}\\
gP_2 &\xrightarrow{k_2} g + P_2\\
g &\xrightarrow{k_3} g + r \quad \text{(Transcription)}\\
r &\xrightarrow{k_4} r + P \quad \text{(Translation)}\\
2P &\xrightarrow{k_5} P_2 \quad \text{(Dimerization)}\\
P_2 &\xrightarrow{k_6} 2P\\
r &\xrightarrow{k_7} \phi \quad \text{(mRNA Degradation)}\\
P &\xrightarrow{k_8} \phi \quad \text{(Protein Degradation)},
\end{aligned}$$

In this example, as the steady state is quickly reached, we generate synthetic data from times $t = [0, 1]$ with a sampling frequency o $dt = 0.5$. Our measurement noise model used is a lognormal error model with $\sigma = 0.07$.

Inferring the Prokaryotic Auto-Regulatory Network with no known reactions

Below, we present the results of our method when fitting the prokaryotic auto-regulatory network without assuming the 4 known reactions and using the same data. Interestingly, these networks are incredibly sparse while also successfully reconstructing the network.

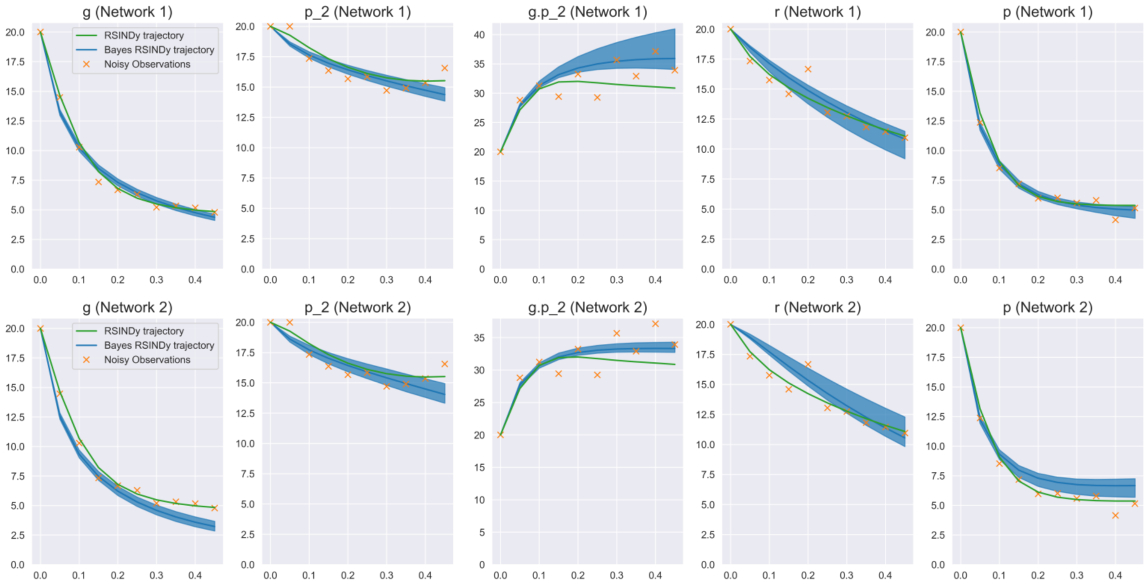


Figure B.1: **Dynamics when inferring all 4 known reactions.** Similar to the case with 4 known reactions, the dynamics from both recovered networks are different from the truth and each other, but still manage to produce plausible dynamics when compared to the noisy data.

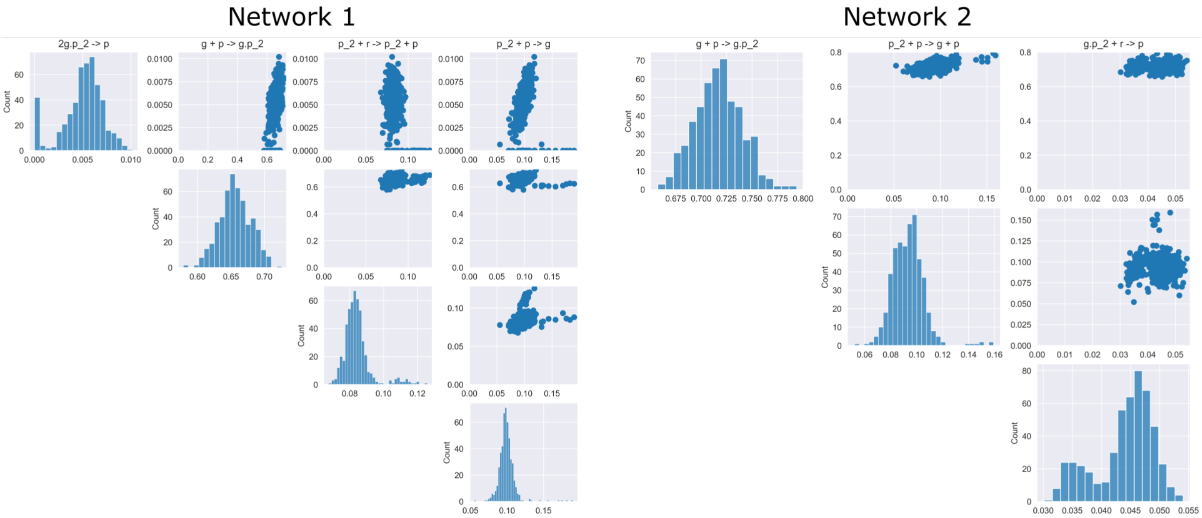


Figure B.2: **Posterior Distributions over non-zero reaction rates** Pair plots of the two distinct reaction networks inferred by the model. Both largely produce similar dynamics despite the differences.

Table B.1: **Selected Recovered Networks for Prokaryotic Auto-Regulation System**

True Network	Network 1	Network 2
$g + P_2 \xrightarrow{0.5} gP_2$	$2gP_2 \xrightarrow{0.005} P$	$g + P \xrightarrow{0.7} gP_2$
$gP_2 \xrightarrow{1} g + P_2$	$g + P \xrightarrow{0.66} gP_2$	$P_2 + P \xrightarrow{0.9} g + P$
$g \xrightarrow{0.15} g + r$	$P_2 + r \xrightarrow{0.8} P_2 + P$	$gP_2 + R \xrightarrow{0.05} P$
$r \xrightarrow{1} r + P$	$P_2 + P \xrightarrow{0.1} g$	
$2P \xrightarrow{0.5} P_2$		
$P_2 \xrightarrow{0.5} 2P$		
$r \xrightarrow{1.5} \phi$		
$P \xrightarrow{0.3} \phi$		

Network inferred by Reactive SINDy for Prokaryotic Auto-Regulation System

Table B.2: **Reactive SINDy inferred Prokaryotic Auto-Regulation Network. Bolded reactions are present in the true network though rates may vary from the true values.**

$$\mathbf{g + P_2 \overset{0.23}{\rightarrow} gP_2}$$

$$\mathbf{r \overset{4.66}{\rightarrow} 0}$$

$$\mathbf{g \overset{7.72}{\rightarrow} 0}$$

$$P_2 \overset{6.27}{\rightarrow} 0$$

$$gP_2 \overset{8.55}{\rightarrow} 2gP_2$$

$$P \overset{1.05}{\rightarrow} P + P_2$$

$$P \overset{17.40}{\rightarrow} 2P$$

$$g + P_2 \overset{0.27}{\rightarrow} 2g$$

$$g + gP_2 \overset{0.47}{\rightarrow} 2gP_2$$

$$g + gP_2 \overset{0.05}{\rightarrow} r$$

$$g + r \overset{0.40}{\rightarrow} 2g$$

$$g + P \overset{0.03}{\rightarrow} P_2$$

$$g + P \overset{0.11}{\rightarrow} 2P$$

$$P_2 + gP_2 \overset{0.11}{\rightarrow} g$$

$$P_2 + gP_2 \overset{0.47}{\rightarrow} 2P_2$$

$$P_2 + gP_2 \overset{0.14}{\rightarrow} P$$

$$P_2 + r \overset{0.18}{\rightarrow} 2r$$

$$P_2 + r \overset{0.05}{\rightarrow} P$$

$$P_2 + P \overset{0.71}{\rightarrow} 2P_2$$

$$gP_2 + r \overset{0.03}{\rightarrow} g$$

$$gP_2 + r \overset{0.09}{\rightarrow} 2r$$

$$gP_2 + P \overset{0.18}{\rightarrow} g$$

$$gP_2 + P \overset{0.39}{\rightarrow} 2gP_2$$

$$r + P \overset{0.25}{\rightarrow} 2r$$

Bibliography

[1] John H Abel, Brian Drawert, Andreas Hellander, and Linda R Petzold. GillesPy: a python package for stochastic model building and simulation. *IEEE life sciences letters*, 2(3):35–38, 2016.

[2] Rinat Abramovitch, Einat Tavor, Jasmine Jacob-Hirsch, Evelyne Zeira, Ninette Amariglio, Orit Pappo, Gideon Rechavi, Eithan Galun, and Alik Honigman. A pivotal role of cyclic amp-responsive element binding protein in tumor progression. *Cancer research*, 64(4):1338–1346, 2004.

[3] Mattias Åkesson, Prashant Singh, Fredrik Wrede, and Andreas Hellander. Convolutional neural networks as summary statistics for approximate bayesian computation. *arXiv preprint arXiv:2001.11760*, 2020.

[4] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18, 2018.

[5] Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

[6] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of the gene expression states of single cells from scrna-seq data. *bioRxiv*, 2019.

[7] Johann Brehmer, Gilles Louppe, Juan Pavez, and Kyle Cranmer. Mining gold from implicit models to improve likelihood-free inference. *Proceedings of the National Academy of Sciences*, 117(10):5242–5249, 2020.

[8] Samuel L Brilleman, Michael J Crowther, Margarita Moreno-Betancur, Jacqueline Buros Novik, James Dunyak, Nidal Al-Huniti, Robert Fox, Jeff Hammerbacher, and Rory Wolfe. Joint longitudinal and time-to-event models for multilevel hierarchical data. *Statistical Methods in Medical Research*, page 0962280218808821, 2018.

[9] Kathleen E Brummel-Ziedins, Thomas Orfeo, Peter W Callas, Matthew Gissel, Kenneth G Mann, and Edwin G Bovill. The prothrombotic phenotypes in familial protein c deficiency are differentiated by computational modeling of thrombin generation. *PloS one*, 7(9):e44378, 2012.

[10] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[11] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics*, 124(4):044109, 2006.

[12] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[13] Bob Carpenter, Matthew D Hoffman, Marcus Brubaker, Daniel Lee, Peter Li, and Michael Betancourt. The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164*, 2015.

[14] Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

[15] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. *arXiv preprint arXiv:1806.07366*, 2018.

[16] Shuonan Chen and Jessica C Mar. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. *BMC Bioinformatics*, 19(1):1–21, 2018.

[17] Mitchell J Cohen and S Ariane Christie. Coagulopathy of trauma. *Critical Care Clinics*, 33(1):101–118, 2017.

[18] Mitchell Jay Cohen, Matt Kutcher, Britt Redick, Mary Nelson, Mariah Call, M Margaret Knudson, Martin A Schreiber, Eileen M Bulger, Peter Muskat, Louis H Alarcon, et al. Clinical and mechanistic drivers of acute traumatic coagulopathy. *The Journal of Trauma and Acute Care Surgery*, 75(1 0 1):S40, 2013.

[19] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[20] Kyle Cranmer, Juan Pavez, and Gilles Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*, 2015.

[21] Bernie J Daigle, Min K Roh, Linda R Petzold, and Jarad Niemi. Accelerated maximum likelihood parameter estimation for stochastic biochemical systems. *BMC bioinformatics*, 13(1):68, 2012.

[22] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. An adaptive sequential monte carlo method for approximate bayesian computation. *Statistics and Computing*, 22(5):1009–1020, 2012.

[23] Geoffrey P Dobson, Hayley L Letson, Rajiv Sharma, Forest R Sheppard, and Andrew P Cap. Mechanisms of early trauma-induced coagulopathy: The clot thickens or not? *Journal of Trauma and Acute Care Surgery*, 79(2):301–309, 2015.

[24] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[25] Brian Drawert, Andreas Hellander, Ben Bales, et al. Stochastic simulation service: bridging the gap between the computational expert and the biologist. *PLoS computational biology*, 12(12):e1005220, 2016.

[26] Christopher C Drovandi and Anthony N Pettitt. Likelihood-free bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis*, 55(9):2541–2556, 2011.

[27] Peter Eberhard and Christian Bischof. Automatic differentiation of numerical integration algorithms. *Mathematics of Computation*, 68(226):717–731, 1999.

[28] Johan Elf and Måns Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome research*, 13(11):2475–2484, 2003.

[29] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[30] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

[31] Seth Flaxman, Swapnil Mishra, Axel Gandy, et al. Estimating the effects of non-pharmaceutical interventions on COVID-19 in europe. *Nature*, pages 1–5, 2020.

[32] Seth Flaxman, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature*, 584(7820):257–261, 2020.

[33] MG Garner, SA Hamilton, et al. Principles of epidemiological modelling. *Revue Scientifique et Technique-OIE*, 30(2):407, 2011.

[34] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, Boca Raton, Florida, 2013.

[35] Pierre Geurts et al. dyngenie3: dynamical genie3 for the inference of gene networks from time series expression data. *Scientific reports*, 8(1):1–12, 2018.

[36] Soumya Ghosh and Finale Doshi-Velez. Model selection in bayesian neural networks via horseshoe priors. *arXiv preprint arXiv:1705.10388*, 2017.

[37] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

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

[39] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[40] Shirin Golchi. Informative priors and bayesian computation. In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, pages 782–789. IEEE, 2016.

[41] Andrew Golightly and Darren J Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005.

[42] Eduardo Gonzalez, Ernest E Moore, and Hunter B Moore. Management of trauma-induced coagulopathy with thrombelastography. *Critical Care Clinics*, 33(1):119–134, 2017.

[43] Eduardo Gonzalez, Ernest E Moore, Hunter B Moore, Michael P Chapman, Theresa L Chin, Arsen Ghasabyan, Max V Wohlauer, Carlton C Barnett, Denis D Bensard, Walter L Biffl, et al. Goal-directed hemostatic resuscitation of trauma-induced coagulopathy: a pragmatic randomized clinical trial comparing a viscoelastic assay to conventional coagulation assays. *Annals of Surgery*, 263(6):1051, 2016.

[44] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[45] J. Gram, H. Duscha, K. H. Zurborn, and H. D. Bruhn. Increased levels of fibrinolysis reaction products (D-dimer) in patients with decompensated alcoholic liver cirrhosis. *Scandinvian Journal of Gastroenterology*, 26(11):1173–1178, Nov 1991.

[46] Sanjana Gupta, Robin EC Lee, and James R Faeder. Parallel tempering with lasso for model reduction in systems biology. *PLoS computational biology*, 16(3):e1007669, 2020.

[47] Mineji Hayakawa, Kunihiko Maekawa, Shigeki Kushimoto, Hiroshi Kato, Junichi Sasaki, Hiroshi Ogura, Tetsuya Matauoka, Toshifumi Uejima, Naoto Morimura, Hiroyasu Ishikura, et al. High d-dimer levels predict a poor outcome in patients with severe trauma, even with high fibrinogen levels on arrival: a multicenter retrospective study. *Shock*, 45(3):308–314, 2016.

[48] Joeri Hermans, Volodimir Begy, and Gilles Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. *arXiv preprint arXiv:1903.04057*, 2019.

[49] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[50] Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[51] Moritz Hoffmann, Christoph Fröhner, and Frank Noé. Reactive sindy: Discovering governing reactions from concentration data. *The Journal of chemical physics*, 150(2):025101, 2019.

[52] John B Holcomb, Deborah J Del Junco, Erin E Fox, Charles E Wade, Mitchell J Cohen, Martin A Schreiber, Louis H Alarcon, Yu Bai, Karen J Brasel, Eileen M Bulger, et al. The prospective, observational, multicenter, major trauma transfusion (prommtt) study: comparative effectiveness of a time-varying treatment with competing risks. *JAMA Surgery*, 148(2):127–136, 2013.

[53] Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5(9):1–10, 2010.

[54] Joseph G Ibrahim, Haitao Chu, and Liddy M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

[55] Hiroyasu Ishikura and Taisuke Kitamura. Trauma-induced coagulopathy and critical bleeding: the role of plasma and platelet transfusion. *Journal of Intensive Care*, 5(1):1–8, 2017.

[56] Hemant Ishwaran, J Sunil Rao, et al. Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, 33(2):730–773, 2005.

[57] Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate bayesian computation via deep neural network. *Statistica Sinica*, pages 1595–1618, 2017.

[58] Jeffry L Kashuk, Ernest E Moore, J Scott Millikan, and John B Moore. Major abdominal vascular trauma–a unified approach. *The Journal of Trauma*, 22(8):672–679, 1982.

[59] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pages 3581–3589, 2014.

[60] Matthew E Kutcher, Adam R Ferguson, and Mitchell J Cohen. A principal component analysis of coagulation after trauma. *The Journal of Trauma and Acute Care Surgery*, 74(5):1223, 2013.

[61] Gwenaël GR Leday, Mathisca CM De Gunst, Gino B Kpogbezan, Aad W Van der Vaart, Wessel N Van Wieringen, and Mark A Van De Wiel. Gene network reconstruction using global-local shrinkage priors. *The annals of applied statistics*, 11(1):41, 2017.

[62] Torsten Loof, Christin Deicke, and Eva Medina. The role of coagulation/fibrinolysis during streptococcus pyogenes infection. *Frontiers in Cellular and Infection Microbiology*, 4:128, 09 2014.

[63] Pavel Loskot, Komlan Atitey, and Lyudmila Mihaylova. Comprehensive review of models and methods for inferences in bio-chemical reaction networks. *Frontiers in genetics*, 10:549, 2019.

[64] Niall M Mangan, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):52–63, 2016.

[65] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. Springer, 2006.

[66] Harley H McAdams and Adam Arkin. It'sa noisy business! genetic regulation at the nanomolar scale. *Trends in genetics*, 15(2):65–69, 1999.

[67] Toby J Mitchell and John J Beauchamp. Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032, 1988.

[68] Edward R Morrissey, Miguel A Juárez, Katherine J Denby, and Nigel John Burroughs. On reverse engineering of gene interaction networks using time course data with repeated measurements. *Bioinformatics*, 26(18):2305–2312, 2010.

[69] Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

[70] World Health Organization et al. Coronavirus disease (COVID-19) weekly epidemiological update, September 7, 2020. 2020.

[71] Jamie Owen, Darren J Wilkinson, and Colin S Gillespie. Likelihood free inference for markov processes: a comparison. *Statistical applications in genetics and molecular biology*, 14(2):189–209, 2015.

[72] Wei Pan, Ye Yuan, Jorge Gonçalves, and Guy-Bart Stan. Reconstruction of arbitrary biochemical reaction networks: A compressive sensing approach. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 2334–2339. IEEE, 2012.

[73] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019.

[74] Omar D Perez, Peter O Krutzik, and Garry P Nolan. Flow cytometric analysis of kinase signaling cascades. In *Flow Cytometry Protocols*, pages 67–94. Springer, 2004.

[75] R. Picetti, H. Shakur-Still, R. L. Medcalf, J. F. Standing, and I. Roberts. What concentration of tranexamic acid is needed to inhibit fibrinolysis? A systematic review of pharmacodynamics studies. *Blood Coagulation & Fibrinolysis*, 30(1):1–10, Jan 2019.

[76] Juho Piironen, Aki Vehtari, et al. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

[77] Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105, 2010.

[78] Arya A Pourzanjani, **Richard Jiang**, Brian Mitchell, Paul J Atzberger, and Linda R Petzold. General bayesian inference over the stiefel manifold via the givens representation. *Bayesian Analysis*, 2020.

[79] Arya A. Pourzanjani, Tie Bo Wu, **Richard M. Jiang**, Mitchell J. Cohen, and Linda R. Petzold. Understanding coagulopathy using multi-view data in the presence of sub-cohorts: A hierarchical subspace approach. In *Proceedings of the 2nd*

Machine Learning for Healthcare Conference, volume 68 of *Proceedings of Machine Learning Research*, pages 338–351. PMLR, 18–19 Aug 2017.

[80] Thomas P. Prescott and Ruth E. Baker. Multifidelity approximate bayesian computation with sequential monte carlo parameter sampling, 2020.

[81] Cécile Proust-Lima, Jean-François Dartigues, and Hélène Jacqmin-Gadda. Joint modeling of repeated multivariate cognitive measures and competing risks of dementia and death: a latent process and latent class approach. *Statistics in Medicine*, 35(3):382–398, 2016.

[82] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[83] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[84] Muruhan Rathinam, Linda R Petzold, Yang Cao, and Daniel T Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics*, 119(24):12784–12794, 2003.

[85] I Raza, R Davenport, C Rourke, S Platton, J Manson, C Spoors, S Khan, HD De'Ath, S Allard, DP Hart, et al. The incidence and magnitude of fibrinolytic activation in trauma patients. *Journal of Thrombosis and Haemostasis*, 11(2):307–314, 2013.

[86] Dimitris Rizopoulos and Pulak Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12):1366–1380, 2011.

[87] I. Roberts, H. Shakur, T. Coats, B. Hunt, E. Balogun, L. Barnetson, L. Cook, T. Kawahara, P. Perel, D. Prieto-Merino, M. Ramos, J. Cairns, and C. Guerriero. The CRASH-2 trial: a randomised controlled trial and economic evaluation of the effects of tranexamic acid on death, vascular occlusive events and transfusion requirement in bleeding trauma patients. *Health Technology Assessment*, 17(10):1–79, Mar 2013.

[88] Rasmus Ros, Elizabeth Bjarnason, and Per Runeson. A machine learning approach for semi-automated search and selection in literature studies. In *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*, pages 118–127, 2017.

[89] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

[90] P. M. Sathe and U. D. Patwa. D Dimer in acute care. *International Journal of Critical Illness & Injury Science*, 4(3):229–232, Jul 2014.

[91] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, 50(9):093001, 2017.

[92] H Schöchl, W Voelckel, M Maegele, and C Solomon. Trauma-associated hyperfibrinolysis. *Hämostaseologie*, 32(01):22–27, 2012.

[93] H. Shakur-Still, I. Roberts, B. Fawole, M. Kuti, O. O. Olayemi, A. Bello, S. Huque, O. Ogunbode, T. Kotila, C. Aimakhu, O. A. Okunade, T. Olutogun, C. O. Adetayo, K. Dallaku, U. Mansmann, B. J. Hunt, T. Pepple, and E. Balogun. Effect of tranexamic acid on coagulation and fibrinolysis in women with postpartum haemorrhage (WOMAN-ETAC): a single-centre, randomised, double-blind, placebo-controlled trial. *Welcome Open Res*, 3:100, 2018.

[94] Prashant Singh, Fredrik Wrede, and Andreas Hellander. Scalable machine learning-assisted model exploration and inference using sciope. *Bioinformatics*, 2020.

[95] Prashant Singh, Fredrik Wrede, and Andreas Hellander. Scalable machine learning-assisted model exploration and inference using Sciope. *Bioinformatics*, 2020.

[96] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.

[97] Donat R Spahn, Bertil Bouillon, Vladimir Cerny, Timothy J Coats, Jacques Duranteau, Enrique Fernández-Mondéjar, Daniela Filipescu, Beverley J Hunt, Radko Komadina, Giuseppe Nardi, et al. Management of bleeding and coagulopathy following major trauma: an updated european guideline. *Critical Care*, 17(2):R76, 2013.

[98] Stan Development Team. Rstanarm: Bayesian applied regression modeling via stan. r package version 2.17.4. 2018.

[99] Michael J Sweeting, Jessica K Barrett, Simon G Thompson, and Angela M Wood. The use of repeated blood pressure measures for cardiovascular risk prediction: a comparison of statistical models in the aric study. *Statistics in Medicine*, 36(28):4514–4528, 2017.

[100] Sanyi Tang, Biao Tang, Nicola Luigi Bragazzi, et al. Stochastic discrete epidemic modeling of COVID-19 transmission in the province of shaanxi incorporating public health intervention and case importation. *medRxiv*, 2020.

[101] **Richard Jiang***, Bruno Jacob*, Matthew Geiger, Sean Matthew, Bryan Rumsey, Prashant Singh, Fredrik Wrede, Tau-Mu Yi, Brian Drawert, Andreas Hellander, and Linda Petzold. Epidemiological modeling in stochss live! *Bioinformatics*, 2021.

[102] **Richard Jiang**, Arya A Pourzanjani, Mitchell J Cohen, and Linda Petzold. Associations of longitudinal d-dimer and factor ii on early trauma survival risk. *BMC Bioinformatics*, 22(1):1–13, 2021.

[103] **Richard Jiang***, Arya A Pourzanjani*, and Linda Petzold. Improving the identifiability of neural networks for bayesian inference. In *2017 NIPS Workshop on Bayesian Deep Learning*, 2017.

[104] **Richard Jiang**, Fredrik Wrede, Prashant Singh, and Linda Petzold. Sparse bayesian inference of mass-action biochemical reaction networks using the regularized horseshoe prior. *In Submission to PLOS Computational Biology.*

[105] **Richard M Jiang**, Fredrik Wrede, Prashant Singh, Andreas Hellander, and Linda R Petzold. Accelerated regression-based summary statistics for discrete stochastic systems via approximate simulators. *BMC Bioinformatics*, 22(1):1–17, 2021.

[106] Owen Thomas, Ritabrata Dutta, Jukka Corander, Samuel Kaski, and Michael U. Gutmann. Likelihood-free inference by ratio estimation. *Bayesian Anal.*, 2020. Advance publication.

[107] Robin N Thompson. Epidemiological models are important tools for guiding COVID-19 interventions. *BMC Medicine*, 18:1–4, 2020.

[108] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[109] Wu-Song Tong, Ping Zheng, Jing-Song Zeng, Yi-Jun Guo, Wen-Jin Yang, Gao-Yi Li, Bin He, Hui Yu, Yong-Sheng Li, Xin-Fen Tang, et al. Prognosis analysis and risk factors related to progressive intracranial haemorrhage in patients with acute traumatic brain injury. *Brain Injury*, 26(9):1136–1142, 2012.

[110] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael PH Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31):187–202, 2009.

[111] Anastasios A Tsiatis and Marie Davidian. Joint modeling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, pages 809–834, 2004.

[112] Zoltan A Tuza and Guy-Bart Stan. An automatic sparse model estimation method guided by constraints that encode system properties. In *2019 18th European Control Conference (ECC)*, pages 2171–2176. IEEE, 2019.

[113] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017.

[114] José MG Vilar, Hao Yuan Kueh, Naama Barkai, and Stanislas Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proceedings of the National Academy of Sciences*, 99(9):5988–5992, 2002.

[115] Eberhard O Voit, Harald A Martens, and Stig W Omholt. 150 years of the mass action law. *PLoS Comput Biol*, 11(1):e1004012, 2015.

[116] David J Warne, Ruth E Baker, and Matthew J Simpson. A practical guide to pseudo-marginal methods for computational inference in systems biology. *Journal of theoretical biology*, page 110255, 2020.

[117] Jeffrey I Weitz, James C Fredenburgh, and John W Eikelboom. A test in context: D-dimer. *Journal of the American College of Cardiology*, 70(19):2411–2420, 2017.

[118] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872–876, 2008.

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

[120] Darren J Wilkinson. *Stochastic modelling for systems biology.* CRC press, 2018.

[121] Mark J Willis and Moritz von Stosch. Inference of chemical reaction networks using mixed integer linear programming. *Computers & Chemical Engineering*, 90:31–43, 2016.

[122] Fredrik Wrede, Robin Eriksson, **Richard Jiang**, Linda Petzold, Stefan Engblom, Andreas Hellander, and Prashant Singh. Robust and integrative bayesian neural networks for likelihood-free parameter inference. *arXiv preprint arXiv:2102.06521*, 2021.

[123] Fredrik Wrede and Andreas Hellander. Smart computational exploration of stochastic gene regulatory network models using human-in-the-loop semi-supervised learning. *Bioinformatics*, 35(24):5199–5206, 05 2019.

[124] Yuanyang Zhang, **Richard Jiang**, and Linda Petzold. Survival topic models for predicting outcomes for trauma patients. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pages 1497–1504. IEEE, 2017.

[125] Yuanyang Zhang, Tie Bo Wu, Bernie J Daigle, Mitchell Cohen, and Linda Petzold. Identification of disease states associated with coagulopathy in trauma. *BMC Medical Informatics and Decision Making*, 16(1):124, 2016.

[126] Yun Zhao, **Richard Jiang**, Zhenni Xu, Elmer Guzman, Paul K Hansma, and Linda Petzold. Scalable bayesian functional connectivity inference for multi-electrode array recordings. In *BioKDD'20*, 2020.

[127] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.