Summary of the thesis

This thesis explores several statistical methodologies developed to tackle key challenges in the analysis of complex biological and medical data. The research spans a broad array of applications and data of various types, necessitating the use of diverse statistical and probabilistic techniques to ensure accurate inference and meaningful biological interpretations. The thesis is organized into four main parts.

The first part introduces *BayVel*, a Bayesian framework for RNA velocity estimation. RNA velocity leverages single-cell RNA sequencing (scRNA-seq) data to model the evolution of cells at different stages of maturity, providing valuable insights into differentiation trajectories. BayVel operates directly on discrete mRNA expression counts and addresses several criticisms of the widely used method *scVelo*. *BayVel* resolves identifiability issues with mathematically and biologically justified solutions, and incorporates a capture efficiency parameter in the likelihood. Extensive simulation studies demonstrate that *BayVel* effectively recovers underlying gene dynamics. Its application to real data raises important questions about the overall reliability of RNA velocity estimates.

The second part focuses on survival analysis, with a particular emphasis on the impact of diagnostic delays on survival estimates. We develop a mathematical framework to quantify the bias introduced by delayed diagnoses. Through simulation studies that vary both the magnitude of the delay and its correlation with covariates, we quantify how these delays distort survival estimates as well as calibration and discrimination metrics. Additionally, we propose a naive correction method to adjust for these biases, demonstrating its effectiveness in improving model calibration and risk prediction. These findings underscore the critical importance of accounting for diagnostic delays.

The third part addresses the challenge of analyzing high-dimensional semicontinuous data, which frequently arise in biomedical and environmental research. Traditional multivariate techniques such as MANOVA struggle in these settings due to the high dimensionality relative to sample size and the presence of zero inflation. We propose a novel regularized MANOVA test that simultaneously compares mean structures and the probability of excess zeros across groups. Ridge-like regularization is incorporated to stabilize covariance estimation, with the penalty parameter chosen by minimizing an information criterion. Our test statistic is derived from a regularized likelihood ratio test, and its null distribution is approximated using permutation methods. Extensive simulation studies confirm that our approach effectively controls type I error and achieves high power, and its utility is demonstrated with applications to real microRNA expression data from human blastocysts and ecological datasets.

The final part adopts a probabilistic framework to model nanoparticle formation using chemical reaction network modeling. Focusing on the final size distribution as the initial number of monomers approaches infinity, we introduce a novel scaling regime in which the growth rate of the first newly formed

nanoparticle is comparable to the nucleation rate, departing from classical scaling assumptions. Simulation studies using Gillespie's algorithm and tau-leaping methods reveal that, even under this alternative scaling, a deterministic size distribution seems to emerge. To mitigate the high computational costs of simulating such complex systems, we propose also an approximation technique that substantially reduces simulation time while preserving the essential statistical properties of the model, rendering it scalable and applicable to realistic experimental scenarios.