Preface

This Ph.D. thesis summarizes my scientific research in collaboration with the Niels Bohr Institute (NBI) and the Globe Institute, University of Copenhagen, and was funded by the Lundbeck Foundation. The research was supervised by Associate Professor Troels C. Petersen (NBI) and Assistant Professor Thorfinn S. Korneliussen (Globe Institute).

Being a cross-disciplinary project, the research presented in this thesis is multi-faceted and covers a wide range of topics with the main scope being the development and integration of novel statistical methods and machine learning models for the analysis of large-scale biological data. The thesis is organized as follows: First I present a brief introduction to the statistical methods and machine learning models used in the thesis and then I present the research in the form of four papers, each of which reflects a different aspect of the research. The introduction is written with my former self in mind, containing the background knowledge I would have liked to know when I started the projects. I hope that it will be useful for anyone interested in the research presented in this thesis.

The first paper presents a novel method I developed for detecting and classifying ancient DNA damage in metagenomic samples taking the full taxonomic information into account. While the first paper focuses on the development of the statistical model in the field of ancient genomics, the second paper focuses on the use of modern machine learning models in medicine and how advanced boosted decision trees can not only improve the accuracy of identifying patients at risk of being readmitted after knee or hip surgery, but doing so in a way that is interpretable as well.

In the beginning of 2020 we all experienced how COVID-19 suddenly changed our lives and impacted our societies in dramatic ways. During this time, I worked for Statens Serum Institut, the Danish Center of Disease Control, on a project to develop an agent based model capable of simulating the spread of COVID-19 in Denmark. This model is presented in the third paper and was used to inform the Danish government on how to best handle the pandemic in the early stages and the effect of contact tracing.

Lastly, in the fourth paper I show how advanced Bayesian methods can be utilized to better estimate the diffusion coefficients in silencing foci in the cell nucleus with single-particle tracking experiments.

Abstract

In recent years, methods such as next generation sequencing in genomics and the use of electronic records in the health care sector has dramatically increased the amount of data in the life sciences. In the field of ancient genomics, newer lab protocols, combined with strict precautions, now allow for the sequencing of ancient environmental DNA millions of years old. In health care, electronic records have allowed for the use of modern machine learning models due to the increased amount of collected data. This has led to a need for new methods and tools to analyze and interpret this vast amount of information that seems to keep increasing in size in the coming years. This thesis focuses on the use cases and potential issues with applying modern statistical and data science related methods on biological data.

The work of this thesis is split into four parts, each with a dedicated paper supporting it. The first paper introduces a novel statistical method that we developed for analysing ancient metagenomic DNA damage. To our knowledge, no prior methods exist which are designed to cover this specific use case in genomics. We show that the work of this project, the `metaDMG` software, is both faster at ancient DNA damage estimation than existing methods and provides more accurate damage estimates – even at taxonomic levels down to 100 reads. As such, `metaDMG` is state-of-the-art for ancient DNA damage estimation for both simple and complex ancient genomic datasets.

The second paper presents a machine learning approach to predict medical complications after surgery, in particular knee and hip operations. The use of machine learning in anaesthesiology is still in its infancy, and this work is a first step towards the use of machine learning in this field. We show that modern machine learning models can be used to predict complications after surgery with higher accuracy than classical statistical methods commonly used in the field. Concretely, we find a 9.7% increase in precision and 1.6 percentage points increase in the area-under-ROC-curve metric when using a boosted decision tree compared to logistic regression. We further show how explainability methods can not only be used to better understand the "black box" of machine learning models, and thus the risk predictions themselves, but also help support the doctors in their decision making process.

The third paper describes how spatial heterogeneities affect the fitted predictions of an epidemic curve in the early phase. In collaboration with Statens Serum Institut, the Danish Center for Disease Control, we developed an agent based

model which extends on the classical SIR models often used in epidemiology. This allowed us to model the spread of disease in the Danish population and introduce complex interaction patterns between the agents in the form of heterogeneities based on geographical density. We found that fitting with classical SEIR models overestimate the peak number of infected and the total number of infected by a factor of two if only fitted on an early-stage epidemic.

All living cells share the same DNA, yet the expression of genes differ wildly between cells. The mechanisms regulating gene expressions and the silencing of specific genes are not yet fully understood, however, it is known that the heterogeneous environment in the cell nucleus is a key factor in this. In particular, the silencing and repair foci play an important role. The fourth paper presents the analysis of these foci by analysing the single molecule dynamics using Bayesian inference based on diffusion models. This allow us to extract and quantify the diffusion coefficients of the foci which describe the physical mechanisms of the formation of the foci.

Dansk Resumé

Metoder som næste-generation sekventering i genetik og brugen af elektroniske journaler i sundhedsvæsenet har i løbet af de seneste år drastisk øget mængden af data. Nye laboratorieprotokoller har inden for arkæogenetik nu muliggjort sekventering af DNA som er millioner år gammelt. Med indførslen af elektroniske patientjournaler blev den tilgængelige mængde data øget kraftigt, hvilket har muliggjort brugen af moderne maskinlæringsmodeller. Tilsammen har disse moderne metoder ført til et øget behov for nye værktøjer til at analysere og fortolke denne enorme mængde information – information som ser ud til at fortsætte med at vokse i størrelse i de kommende år. Denne afhandling fokuserer på udviklingen og brugen af moderne statistiske metoder på forskellig biologisk data.

Indholdet af denne afhandling er delt op i fire dele baseret på hver sin artikel. I den første artikel introducerer vi en ny statistisk metode til analyse af DNA-skade i arkæogenetik. Vi er ikke bekendt med nogen tidligere metoder der er designet til at dække dette specifikke anvendelsesområde. Vi viser i artiklen at produktet af vores forskning, `metaDMG` softwaren, er både hurtigt og præcist til at estimere DNA-skade – selv med kun ganske lidt data (helt ned til kun 100 DNA-sekvenser). Dette viser at `metaDMG` er et førende værktøj indenfor feltet til estimering af DNA-skade for både simple og komplekse arkæogenetiske datasæt

I den anden artikel præsenterer vi en ny tilgang til at forudsige medicinske komplikationer efter en knæ- eller hofteoperation ved brug af moderne maskinlæringsmodeller. Brugen af maskinlæring er stadig forholdsvis ny indenfor anæstesi og dette er et første skridt i at anvende maskinlæring indenfor dette felt. Vi viser i artiklen at moderne maskinlæringsmetoder kan anvendes til at forudsige medicinske komplikationer med højere præcision end de klassiske metoder der ofte er benyttet inden for feltet. Vi finder en 9,7% forbedring i præcision og 1,6 procentpoint forøgelse i arealet-under-ROC-kurven når man sammenligner maskinlæringsmodellen med en logistisk regression. Vi viser yderligere at metoder relateret til model-forklaring ikke blot kan bruges til at forstå modellens inderste dele, og dermed selve risikoforudsigelserne, men også kan hjælpe lægerne i deres beslutningsproceser.

Vi beskriver i den tredje artikel hvordan rumlige uensartetheder påvirker de teoretiske forudsigelser af en epidemikurve, hvis man baserer sine forudsigelser på data fra den tidlige fase af en epidemi. Vi udviklede i samarbejde med Statens Serum Institut en agent-baseret model. Denne model var bygget på de klassiske SIR-modeller som ofte er anvendt i epidemiologien. Brugen af agent-baserede modeller

tillod os at modellere spredningen af sygdom i den danske befolkning og introducere komplekse interaktionsmønstre mellem agenterne i form af uensartetheder baseret på geografisk tæthed. Vi fandt at forudsigelser baseret på SIR-lignende modeller overestimerer det maksimale antal samtidig smittede, og det samlede antal smittede, med en faktor to, hvis man kun kigger på data fra den tidlige fase af en epidemi.

Alle levende celler deler det samme DNA, dog er der stor forskel på hvilke gener som hver enkel celle rent faktisk udtrykker. Mekanismerne bag denne genregulering og dæmpningen af specifikke gener er stadig ikke forklaret fuldstændig, men man ved at den fysiske struktur af cellekernen spiller en stor rolle. Især dæmpnings- og reperationsfokusserne i cellekernen er særdeles vigtige i denne sammenhæng. I den fjerde artikel analyserer vi disse fokusser ved hjælp af Bayesiansk inferens baseret på diffusionsmodeller. Ud fra dette måler vi diffusionskoefficienterne af fokusserne, hvilket kan bruges til at beskrive de fysiske processer som ligger til grund for skabelsen af fokusserne.

Publications

The work presented in this thesis is based on the following publications:

Paper 1: **Christian Michelsen**[†], Mikkel W. Pedersen[†], Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). "metaDMG: A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data". Submitted to Methods in Ecology and Evolution.

Paper 2: **Christian Michelsen**[†], Christoffer C. Jørgensen[†], Mathias Heltberg, Mogens H. Jensen, Alessandra Lucchetti, Pelle B. Petersen, Troels C. Petersen, Henrik Kehlet (2022). "Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty – a machine learning based approach". In review at BMJ Open.

Paper 3: Mathias S. Heltberg[†], **Christian Michelsen**[†], Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". Published in: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

Paper 4: Susmita Sridar[†], Mathias S. Heltberg[†], **Christian Michelsen**[†], Judith M. Hattab, Angela Taddei (2022). "Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci". Unpublished paper draft.

Shared first authorship is indicated with a dagger (†) next to the name.

The appendix contains two papers of which I am a co-author, see Appendix A and Appendix B. The appendix further contains two reports published by Statens Serums Institut that are based on my research during my Ph.D., see Appendix C and Appendix D. These appendices are further explained in Chapter 1.

Acknowledgements

First of all, I want to express my sincere gratitude to my long time supervisor, captain, and friend: Troels. You are truly an inspiration to work with. I want to thank you for opening so many doors for me, both academically, professionally, and nautically. I am looking forward to our future adventures together. I also want to thank my co-supervisor, Thorfinn. I want to thank you for introducing me to the field of bioinformatic and helping me to develop my skills in this area. I also want to thank you for your patience and guiding me through the endless amount of (near) identical biological concepts and helping me to understand the minute differences.

I have been fortunate to work with people from a wide range of backgrounds and disciplines during my Ph.D. The author lists on the papers in this thesis include a particle physicist, bioinformaticians, a clinical professor, epidemiologists, a medical doctor, a bio-physicist, a mathematician, a biologist, and the president of the Royal Danish Academy of Sciences and Letters at the time. Before anything else, I want to thank all of my co-authors for their work and contributions to these papers and for allowing me to be a part of their projects. I have learned a lot from all of you, and I hope I have been able to contribute something to your work as well.

I am thankful for all the people who have helped me with my work and listened to my complaints when I was stuck, when the code did not compile, or when the small bug was almost impossible to find (which was not a small amount of time). In particular I want to thank the people at Globe who I have spent the most time with; Rasa, Alba, and Rasmus. I also want to thank the Korneliussen Group and the people in my office for helpful advice, suggestions and discussions. This also includes Daniel Nielsen and Rasmus Ørsøe from NBI. Finally, I want to thank Mathias Heltberg for many years of fruitful collaboration and for including me in his projects.

This project would not have been possible had it not been for the Lundbeck Foundation which funded my Ph.D. In addition to the funding itself, I am grateful for the inter-disciplinary aspect of project which has allowed me to meet so many inspiring and talented people and for the freedom to pursue my own interests within the project.

I would also like to express my gratitude to Professor Guido Sanguinetti from the International School for Advanced Studies, SISSA, in Trieste, Italy, for hosting me in his group during the Winter of 2021/2022. My gratitude also goes out to

Kosio, Sara, Max, Romina, Noor, Viplove, Anne-Marie, and all the other wonderful people I met during in Trieste; thank you for making my stay in Italy so enjoyable.

I want to thank my friends for always being there for me. A special thanks to my friends from NBI and Borchen who I know I can always count on, whether or not that includes a trip in a party bus (of the Sea), taking Artemis out for a sail, or board games and beer. Thank you for always being there. I also want to thank my family, especially my parents for their unconditional support and encouragement. I am grateful for the opportunities they have given me and for the sacrifices they have made for me.

Lastly, I want to thank my future wife and mother of our child, Anna. I would not have been able to do this without you. Thank you for your patience and support. I am looking forward to our future together. I love you to the moon and back. And Blob, I cannot wait to meet you!

1 *Introduction*

The primary content of my thesis is the four papers included in the thesis in Chapter 2 to Chapter 5. This chapter is meant as a brief introduction to the background needed to understand the basics of the methods used throughout the papers. As such, this chapter is not meant to be a comprehensive guide to all the statistical methods and bioinformatic tools used in the papers. The original research motivation supporting the funding of this Ph.D. was multi-disciplinary and the papers included in my thesis are also highly influenced by this.

In Section 1.1, I will shortly introduce the field of ancient genomics and the statistical methods used to identify ancient DNA will be explained. Paper I, see Chapter 2, utilize modern Bayesian methods to classify which species are ancient, and which ones are not. Bayesian methods are great when possible, however, they also rely on some statistical model being defined. In the case of Paper I, the model is a beta-binomial distribution combined with a modified geometric damage profile (exponential decay).

Sometimes the model is not known and the data generation process has to be inferred by other means. This is the case in Paper II, see Chapter 3, where we utilize machine learning methods to extract this information. This paper deals with estimating the individual risk scores for each patient being re-hospitalized after a knee or hip operation. Section 1.2 introduces the reader to basic classification with machine learning models.

While the former two papers are based on real life data, Paper III, see Chapter 4, concerns the development of a new agent based model for COVID-19. The model is based on the SIR model but by using an agent-based model it allows for more complex and realistic behaviour of the disease and the transmission process. The model is used to simulate the spread of virus in Denmark and to estimate the effect of contact tracing. The model is also used to simulate and predict the spread of the "alpha" variant of COVID-19 in Denmark. Section 1.3 introduces the reader to the basics of agent based models.

Finally, the method of Bayesian model comparison of different diffusion models is introduced in Paper IV, see Chapter 5. In particular, this paper deals with different mixture-models of independent Rayleigh-distributions, and how they can be used to extract important information about the underlying diffusion processes of a polymer bridging model in cell nuclei, see Section 1.4.

1.1 *Ancient DNA and Bayesian Statistics*

The similarity between family members and the degree to which siblings resemble one another has long been a mystery in human history. People have always thought about the balance between nature and nurture, as in the famous fairy tale "The Ugly Duckling" by Hans Christian Andersen from 1843. These questions were addressed two decades later, when Gregor Mendel founded genetics as a modern, scientific discipline with his studies on trait inheritance in pea plants (Mendel, Gregor, 1866).

A century later, a major breakthrough occurred when Watson and Crick discovered the double helix structure of DNA (Watson and Crick, 1953). This lead to other important discoveries within genetics, such as the development of DNA sequencing allowing scientists to identify the genetic makeup for a specific cell. Until the mid 1980s, studies within archaeogenetics were limited to analysis of fossilised samples of plants, animals or other species (Parducci and Petit, 2004). Following the first successful recovery of ancient DNA from 5000 year old ancient Mummies, it was shown that it was indeed possible to extract and sequence DNA (Pääbo, 1985a; Pääbo, 1985b). This discovery, along with a dozen other, pushed the boundary for what is scientifically possible with ancient DNA, and led to Svante Pääbo being awarded with the Nobel Prize in Physiology or Medicine in 2022 for "his discoveries concerning the genomes of extinct hominins and human evolution" (Karolinska Institutet, 2022).

The field of ancient DNA (aDNA) was drastically changed with the invention of the Polymerase Chain Reaction (PCR) method (Mullis et al., 1986) along with the Next Generation Sequencing (NGS) technology which revolutionized the speed and throughput of genomic sequencing, while decimating the cost (Slatko, Gardner and Ausubel, 2018). This technological advance has lead to better understanding of human migration and the genealogical tree of modern humans including the previously unknown human (sub)species; the Denisova hominin (Krause et al., 2010). In 2008, the first human genome was sequenced and since then multiple NGS methods have allowed for cheap, high-quality, in-depth sequencing of genetical samples (Genomics and Mobley, 2021). All of this shows, that the field of genetics has grown exponentially and become a central part of modern biology.

Leaving the homocentric world view, aDNA also allows for the study of archaic animals. The age limitation for when aDNA can be sequenced has in the recent years increased; in 2013 with the early Middle Pleistocene 560–780 kyr BP horse (Orlando et al., 2013) and in 2021 with the million-year-old mammoths (van der Valk et al., 2021). High-throughput sequencing not only allows for the sequencing of single genomes – like single humans, animals, or plants – but also for sequen-

cing of entire communities of organisms, so-called metagenomics. By analysing environmental DNA (eDNA) from a set of samples, one can survey the rich plant and animal assemblages of a given area and at a specific time in the past. Our new paper in Nature shows it is now possible to perform metagenomic sequencing on environmental DNA that is 2 million years old, see Appendix A. This is a direct application of the statistical method developed in Paper I, see Chapter 2, showing that `metaDMG` can help to push the boundary of what is possible with ancient DNA.

Ancient DNA is difficult to work with since it often contains only a limited amount of biological material due to bad preservation, leading to low endogenous content with high duplication rates, making high-depth sequencing difficult[1] (Renaud et al., 2019). Here endogenous content refers to DNA from the species of interest and not e.g. ancient bacteria or modern contamination. In addition to this, ancient DNA is often highly degraded. In particular, the two prominent issues with aDNA is fragmentation and deamination (Dabney, Meyer and Pääbo, 2013; Peyrégne and Prüfer, 2020). Fragmentation refers to the fact that through time the DNA is broken into very short fragments, often with a size of less than 50 bp. A consequence of this, upon alignment, is low mapping quality, multimapping, and reference bias, which can somewhat be mitigated by the use variant graphs (Martiniano et al., 2020).

1 Genotype likelihoods are often used to alleviate the problem of low-coverage data (Nielsen et al., 2011).

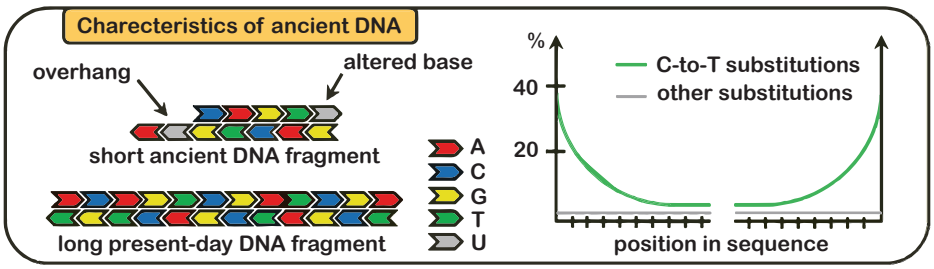


Figure 1. Illustration of DNA damage. Ancient DNA is often highly fragmented with short reads compared to modern, present-day DNA. Due to deamination, aDNA can contain uracils (U), which will be misread as thymines (T) while sequencing, leading to C-to-T nucleotide misincorporations. This is primarily happening at the end of the reads. Modified from Peyrégne and Prüfer, 2020.

Deamination is a process in which cytosine (C) in the single-stranded overhangs in the end of the DNA molecules is often hydrolized to uracil (U) which is read as thymine (T) by the DNA polymerase. This particular type of postmortem damage is known as cytosine deamination, or C-to-T transitions, and is one of the main reasons behind nucleotide misincorporations in ancient DNA (Briggs et al., 2007). Due to the short fragment sizes in ancient DNA, the fragments will often contain overhangs with over-expressed C-to-T frequency. In the case of single-genome analysis, previous solutions have been to either remove all transitions and only keep transversions, apply trimming at the read ends, or enzymatically remove them with USER treatment (Schubert et al., 2012; Rohland et al., 2015). For an illustration of both fragmentation and deamination of ancient DNA, see Figure 1.

Measuring DNA damage is thus a way to prove authentic aDNA. Currently, a handful of different methods for quantifying ancient DNA damage exist. In particular, the mapDamage software has been the standard for how to measure ancient DNA damage in the field (Jónsson et al., 2013). While mapDamage allows for estimating all of the four Briggs parameters, it is often the empirical deamination patterns that mapDamage computes that are used. Newer and faster methods for estimating ancient DNA damage are continuously being developed, including PyDamage (Borry et al., 2021), which tackles some of mapDamage's limitations. However, within metagenomics, which studies the genetic material of all organisms collected from an environmental sample, faster methods suited to analyse this large-scale dataset are still lacking.

Paper I, see Chapter 2, introduces the `metaDMG` software which utilizes the C-to-T deamination pattern[2] to identify ancient DNA damage. One of the key features of this method is the beta-binomial model which allows the uncertainty of the deamination frequency to be fitted independently of the mean of the frequencies leading to improved accuracy of the damage estimation. The deamination frequencies are based on the number of C-to-T transitions, k, out of the total number of C's, N, for a given position within the fragment. The classical likelihood to use for this type of data is a binomial distribution. The mean and variance of the binomial distribution is given by:

2 for the forward strand and the G-to-A deamination pattern for the reverse strand.

$$\begin{aligned} \mathbb{E}[k] &= Np \\ \mathbb{V}[k] &= Np(1-p), \end{aligned} \tag{1}$$

where p is the probability of success (a C-to-T substitution). One of the issues, however, is that the variance of the binomial distribution is proportional to the mean. The binomial distribution is thus not flexible enough to accommodate large amounts of variance in the data, so-called overdispersion (McElreath, 2020). One way to accommodate overdispersion is to instead use a beta-binomial model. The beta-binomial model is a generalization of the binomial distribution where the variance is independent of the mean. Technically, the beta-binomial model assumes that p is a random variable which follows a beta distribution $p \sim \text{Beta}(\mu, \varphi)$ where the beta distribution is parameterized[3] in terms of its mean, μ, and dispersion parameter, φ, (Cepeda-Cuervo and Cifuentes-Amado, 2017). The mean and variance of this beta-binomial model is then given by:

3 This can be reparameterization in term of the classical α, β parameterization by: $\mu = \alpha/(\alpha+\beta)$ and $\varphi = \alpha + \beta$.

$$\begin{aligned} \mathbb{E}[k] &= N\mu \\ \mathbb{V}[k] &= N\mu(1-\mu)\frac{\varphi + N}{\varphi + 1}. \end{aligned} \tag{2}$$

Comparing Equation 1 and Equation 2, it is seen that the variance of the beta-binomial model is no longer (strictly) proportional to the mean, but instead is a function of the dispersion parameter, φ, allowing for higher variance than the binomial-only model. When $\varphi = 0$, the variance of the beta-binomial model is N times larger, and when $\varphi \to \infty$ the variance reduces to the variance of the binomial model, showing that the beta-binomial model is a generalization of the binomial model.

Equation 2 shows how to model the C-to-T damage at a specific base position in the read. We model the position-dependent damage frequency, $f(x) = k(x)/N(x)$, see Figure 1, as a function of the distance from the end of the read, x, with a modified geometric damage profile (exponential decay):

$$y(x; A, q, c) = A(1 - q)^{x-1} + c. \tag{3}$$

Here A is the scale factor, or amplitude, q is the decay rate, and c is a constant offset. The offset can be interpreted as the baseline C-to-T background substitution rate or baseline damage rate. Since x is discrete, this is similar to a (modified) geometric sequence starting from $x = 1$. The combination of equation (2) and (3) is illustrated in Figure 2, which shows the position-dependent decreasing damage frequency. The figure also shows the increase in uncertainty in the beta-binomial model compared to the binomial-only model.

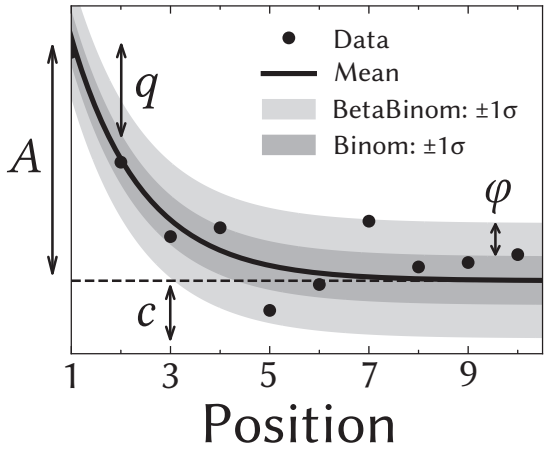


Figure 2.
Illustration of the damage model. The figure shows data points as circles and the fitted damage frequency, $y(x)$, as a solid line. The amplitude of the damage is A, the offset is c, and the relative decrease in damage pr. position is given by q. The damage uncertainty for a binomial model is shown in dark grey and the uncertainty for a beta-binomial model in light grey.

The damage framework described above is based on the nucleotide misincorporations, i.e. the C-to-T transitions. The background for this data can be from either DNA sequence files mapped to a single genome or from metagenomic data consisting of multiple mapped reads. As such, the damage framework is a general tool for estimating damage based on DNA alignment files.

In the metagenomic case, `metaDMG` identifies the lowest common ancestor (LCA) based on the algorithm from ngsLCA (Wang et al., 2022). For each read that maps to multiple reference genomes from separate species, i.e. has multiple alignments, the taxonomic tree is traversed for each alignment until a common ancestor is found. Figure 3 illustrates the LCA for a read that maps to different (sub)species. In this example, the LCA of alignment 1 and 2 is the Subspecies I while the LCA for all four alignments is the Genus X. `metaDMG` works by default with the NCBI taxanomic database but can also be used with custom databases.

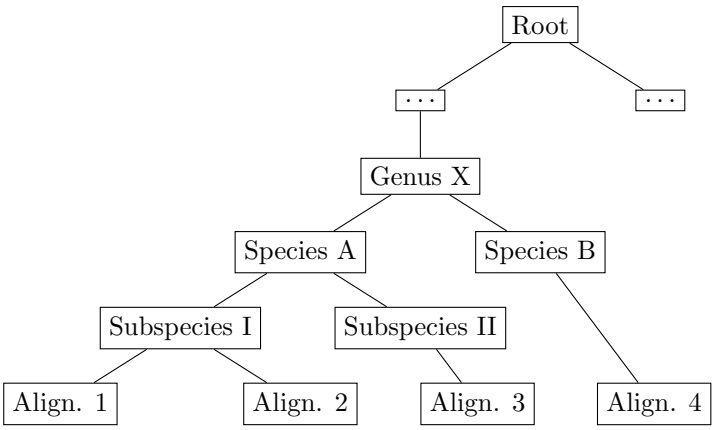


Figure 3.
Illustration of the lowest common ancestor (LCA) for taxonomic trees. Here the LCA of alignment 1 and 2 is Subspecies I, while the LCA of all four reads is Genus X. The dots (...) refers to other taxonomic levels, e.g. family and order.

Given the nucleotide misincorporations, either coming from a single-reference alignment file or after LCA in the metagenomic case, eq. (2) and (3) are fitted with a Bayesian model. This is done to ensure the optimal inference of the parameters, A, q, and c, and to account for the uncertainty in the data. Bayesian inference also allows for the inclusion of domain knowledge in the form of the prior distribution by Bayes theorem. Bayes theorem is based on the law of conditional probability (Barlow, 1993) stating that the probability of two events, A and B, both happening, $P(A \cap B)$, is given by:

$$P(A \cap B) = P(B)P(A|B), \tag{4}$$

where $P(B)$ is the probability of B and $P(A|B)$ is the conditional probability of A given B. Similarly, $P(A \cap B)$ can also be expressed in terms of the probability of A:

$$P(A \cap B) = P(A)P(B|A). \tag{5}$$

Combining Equation 4 and Equation 5 and rearranging terms gives the Bayes theorem:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}, \tag{6}$$

with a change of variables where D refers to the observed data and θ the parameter(s) of the model[4]. The first term in the numerator, $P(\theta)$, is the prior distribution and describes the probability distribution assigned to θ before observing any data. The second term is the likelihood function, $P(D|\theta)$, which is the probability of observing the data given the parameter(s). Together these two terms combine to a compromise between data and prior information.

4 In the case of `metaDMG`, D would be the observed deamination frequencies and θ the four fit parameters.

The numerator, $P(D)$, also known as the evidence, can be treated as a data-related normalization factor. In the case of continuous θ, this can calculated as the marginalization of the likelihood function over θ:

$$P(D) = \int_\theta P(D|\theta)P(\theta)\,\mathrm{d}\theta. \tag{7}$$

This equation, however, is often intractable to compute in the higher-dimensional case. Luckily, it can be shown that Markov Chain Monte Carlo (MCMC) sampling can approximate the posterior distribution, $P(\theta|D)$, and asymptotically converge to the correct distribution (Gelman, Carlin et al., 2015).

Traditionally MCMC methods such as Metropolis Hastings (MH) or Gibbs sampling have been used for Bayesian inference, however, these methods are often slow and require a lot of tuning. In the last decades, a new class of MCMC methods have been developed, namely Hamiltonian Monte Carlo (HMC) methods. While traditional MH uses a Gaussian random walk, HMC is a gradient-based MCMC method that uses Hamiltonian dynamics to guide the sampling. This makes HMC more efficient than traditional MCMC methods and allows for sampling from high-dimensional distributions (Neal, 2011; Betancourt, 2018). A particularly efficient type of HMC is the No-U-Turn Sampler (NUTS). NUTS is a variant of HMC that automatically tunes the step size and number of steps to take in the Hamiltonian dynamics (Homan and Gelman, 2014).

Most statistical domain-specific languages (DSL) such as Stan (Carpenter et al., 2017), Pyro (Bingham et al., 2019), NumPyro (Phan, Pradhan and Jankowiak, 2019) or Turing.jl (Ge, Xu and Ghahramani, 2018), implement HMC and in particular the NUTS algorithm. Since the statistical modelling part of `metaDMG` is implemented in Python, NumPyro is used for the Bayesian inference of the damage model, as it is easy to implement and computationally efficient since it uses JAX (Bradbury

et al., 2018) under the hood for automatic differentiation and just-in-time (JIT) compilation.

Even though NumPyro is fast and `metaDMG` is efficiently implemented, the Bayesian inference of the damage model is still computationally expensive. Thus, it was decided to also include a faster, approximate method of Bayesian inference: the maximum a posteriori (MAP) estimate. The MAP estimate is the point estimate of the posterior distribution that maximizes the posterior probability density function, i.e. the posterior mode:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(\theta)P(D|\theta), \tag{8}$$

where the second equality is due to the evidence being independent of θ. Since this is a point estimate, $\hat{\theta}_{\text{MAP}}$ does not fully explain the full posterior, however, it is often a good approximation[5]. Comparing $\hat{\theta}_{\text{MAP}}$ to the maximum likelihood estimate (MLE):

5 Especially when the posterior is unimodal, which is generally the case for `metaDMG`.

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} P(D|\theta), \tag{9}$$

the MAP estimate can be seen as a regularized version of the MLE estimate (Murphy, 2012). To further optimize the computational efficacy of the MAP estimation in `metaDMG`, the MAP estimation function is JIT compiled using Numba (Lam, Pitrou and Seibert, 2015) and mathematically optimized with iMinuit (Dembinski et al., 2021).

1.2 *Anestesiology – a Machine Learning Approach*

This section explains the technical background behind Paper II, see Chapter 3. This study investigates the potential advantages of using a modern machine-learning model compared to classical logistic regression to predict the risk of patients being re-hospitalized after fast-track hip and knee replacements. In particular, the patients were grouped into two groups. The first group were the so-called "risk-patients" that stayed at least 4 days in the hospital post surgery or were re-hospitalized within 90 days of surgery. The second group were the non-risk-patients. As such, this is a binary classification problem where the patient's risk-score is predicted based on historical data. The machine learning models were trained on 33 variables, of which 7 were continuous, related to the patient's medical record, such as age, gender, the use of walking aid, anaemia, diabetes, etc. A total of 22.017 patients were included in the study, of which 1.476 were risk-patients.

Most classification and regression problems fall under the same machine learning (ML) branch called supervised learning. In supervised learning, the goal is to find the hypothesis h^* in the hypothesis set $\mathcal{H}$ that matches the unknown, "true" data-generating function $f : \mathcal{X} \to \mathcal{Y}$ optimally, where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. Assuming that we have access to realizations of f, the so-called training data $\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, we can use a learning algorithm $\mathcal{A}$ combined with the training data to estimate h^* (Abu-Mostafa, Magdon-Ismail and Lin, 2012). Here N refers to the number of training samples and $\mathbf{x}_i$ is the ith observation with the true label y_i. This process is illustrated in Figure 4.

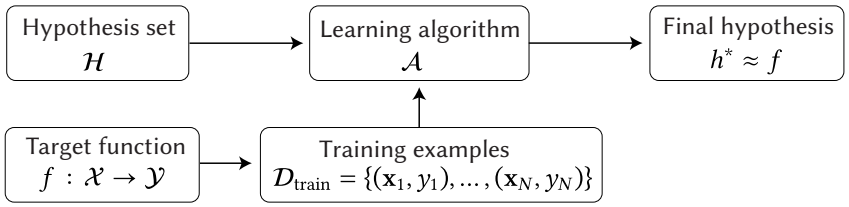


Figure 4. Illustration of how to learn from data in a supervised learning setting. Adapted from (Abu-Mostafa, Magdon-Ismail and Lin, 2012).

Both logistic regression (LR) and ML models can be viewed through the lens of Figure 4, just with $|\mathcal{H}_{\text{LR}}| \ll |\mathcal{H}_{\text{ML}}|$, i.e. the machine learning model is a lot more complex than the logistic regression model and the hypothesis space thus significantly larger. While sufficiently parameterized ML methods can in theory achieve perfect performance on the labelled training set, one is rarely interested in the predictive power of h^* on the training set, as the truth is already known. Instead, one often wish to apply the trained model to new, unseen data where the truth is unknown.

Assessing the performance of h^* on unlabelled data can be difficult. A naive estimate would be to assume that the performance on new, unseen data is the same as on the training data. However, this would likely be a poor estimate due to overfitting and thus bias the predicted performance, especially for high cardinality hypothesis sets. (Abu-Mostafa, Magdon-Ismail and Lin, 2012). The concept of overfitting is illustrated in Figure 5, which shows the training loss as a function of model complexity. The figure shows how more advanced models can achieve lower and lower training losses, however, at some point they start to overfit, leading to higher validation losses. The validation loss is the error on unseen data and is thus the quantity of interest. The goal is to find the sweet spot between underfitting and overfitting.

To avoid overfitting and get accurate estimates of the performance of h^*, we use a technique called cross-validation (CV). In the simplest way, this can be done by splitting the data into two sets, one called the training and one called the validation set, and then only train on the training set. Afterwards the trained

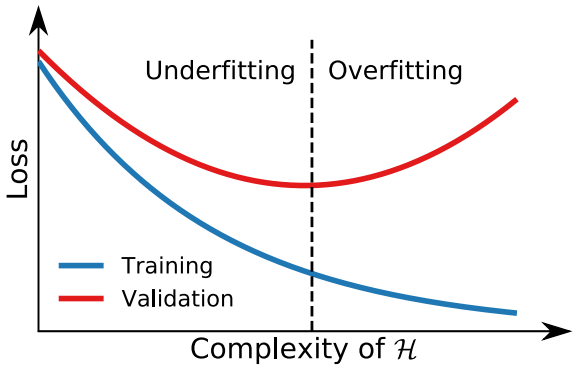
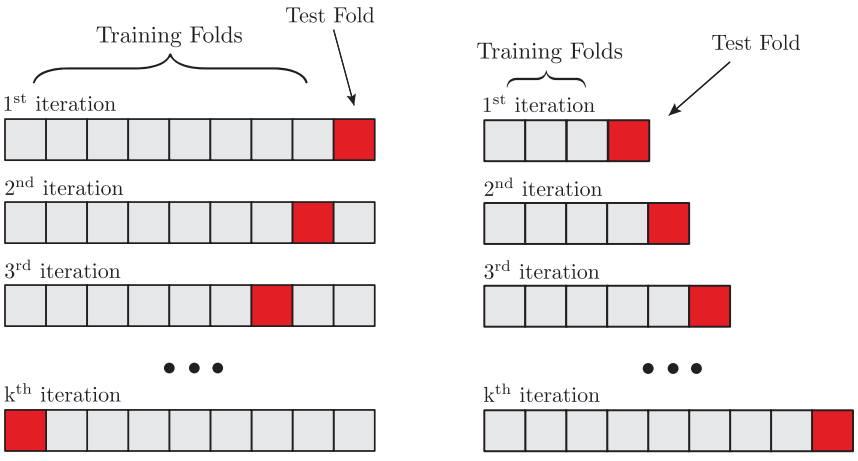


Figure 5. Illustration of the loss as a function of model complexity. The training error is shown in blue and validation error in red. Figure from Michelsen, 2020.

model can be evaluated on the validation set without biasing the performance estimate. This process can further be refined by splitting the data into K folds and then repeating the process K times, where each fold is used as the validation set once. This is called K-fold cross-validation and is illustrated in Figure 6a (Murphy, 2012; Hastie, Tibshirani and Friedman, 2016). K-fold cross validation works well in many cases, yet in the case of temporal data, it also risks introducing bias in the performance estimates, since, in the different folds, it, effectively, is allowed to "look into the future". The most extreme case of this is shown in the bottom of Figure 6a where the model trains on future data and is then evaluated on past data (relative to test fold). In many time dependent datasets, such as the one in Paper II, this is undesirable. Instead, we use a technique called temporal cross validation, see Figure 6b, which circumvents this problem by only allowing the model to train on past data and evaluate on future data (Tashman, 2000). As the patient data is time dependent, this is the technique we use in Paper II.



(a) K-fold cross validation **(b)** Temporal cross validation

Figure 6. Two types of cross validation: K-fold cross validation, and temporal cross validation. Both figures from Michelsen, 2020.

The actual training of the learning model $\mathcal{A}$ is model-dependent and will not be covered in this thesis. The term training refers of the optimization of the internal parameters in the ML model. In most cases, the training depends on the gradient of the loss function with respect to internal parameters to be computed, see Michelsen, 2020 for a more detailed description of the training process.

Training is not the only way to optimize the performance of $\mathcal{A}$, albeit it is the primary one. In addition to the internal parameters of the model, some parameters are external to the model in the sense that they are not optimized by the model itself, but rather by the user. These are called hyperparameters and are often optimized using a technique called hyperparameter optimization (HPO). In the case of logistic regression, the number of variables to include would be an example of a hyperparameter; in the case of a decision tree model, the depth of the tree. Hyperparameter optimization can be performed in many ways, where the common one is through grid search, see Figure 7.

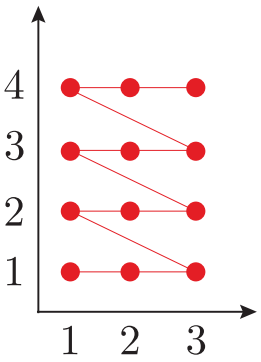


Figure 7.
Illustration of grid search. Figure from Michelsen, 2020.

In grid search, all combinations of the hyperparameters (the cartesian product) are tested and the best combination is chosen. This is a simple and intuitive approach, however, it scales exponentially with the number of hyperparameters. As such, grid search suffers from the curse of dimensionality. In addition to this, it depends on the user-defined grid, which might not be optimal. To circumvent this, a technique called random search (RS) was developed (Bergstra and Bengio, 2012). Random search is a randomized version of grid search, where the hyperparameters are sampled randomly from a distribution. This allows for a more efficient sampling of the hyperparameter space, see Figure 8. Another advantage is that RS lets the user decide on the number of iterations beforehand.

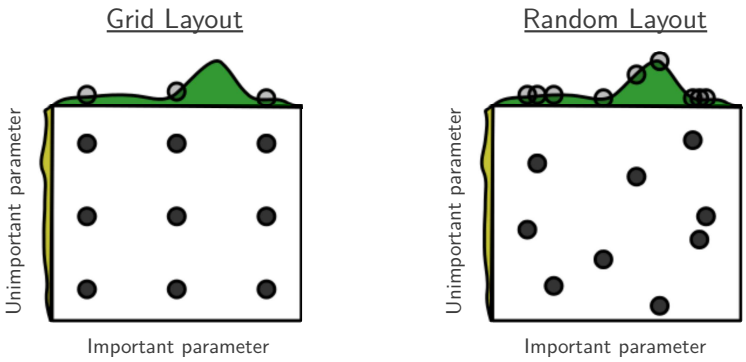


Figure 8.
Illustration comparing grid search to random seach. The height of green surve is the score-function which has to be optimized. Figure from Bergstra and Bengio, 2012.

The disadvantage of random search is that all draws are fully independent. While this allows for easy parallelisation of the algorithm, this also means that each new sample might be infinitesimal close in the hyperparameter space to a previous sample with bad performance, which with high probability will thus also have a high loss. An approach that does take the history of the previous samples'

performance into consideration is Bayesian optimization (Brochu, Cora and de Freitas, 2010). In Bayesian optimization each successive hyperparameter is chosen based on an acquisition function, which optimizes the expected improvement in the performance of the model. This is illustrated in Figure 9. This leaves the user with the task of choosing between "exploitation" and "exploration" of the hyperparameter space in the definition of the acquisition function, yet most implementations of bayesian optimization have decent default settings.

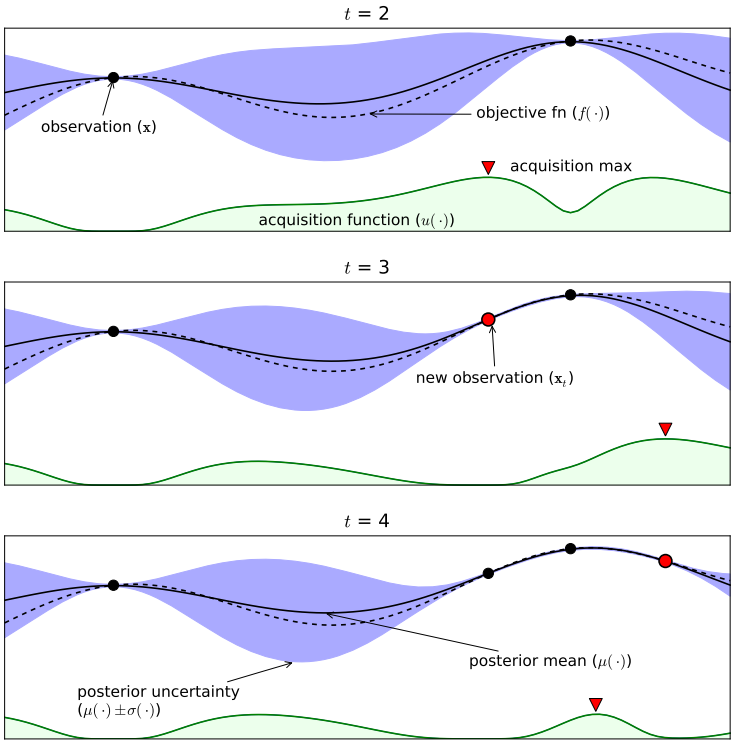


Figure 9. Illustration of the learning process of Bayesian optimization. The previous observations are shown as black dots and the true objective function is shown as a dashed black line. This line is fitted with Gaussian processes which is shown as the solid line with its uncertainty in purple. The acquisition function is shown in green and its maximum decides what the next iteration of the hyperparameter value(s) should be (Michelsen, 2020).

We use the Python package Optuna (Akiba et al., 2019) for HPO in Paper II due to its ease of use and its support for Bayesian optimization. In particular, we use the Tree-structured Parzen Estimator algorithm for the Bayesian optimization and a median stopping rule to minimize optimization time (Bergstra, Bardenet et al., 2011). This allowed for a good comporise between optimization time and performance.

While model performance is often paramount, in some fields – such as medicine – being able to explain the model's predictions is almost as important. This is especially true in the case of medical decision support systems, where the model is used to make decisions about the patient's treatment. Model explainability helps to build trust in the model, for both the patient and the medical staff alike.

In Paper II, we employ the SHapley Additive exPlanations (SHAP) values which provide estimates on which variables contribute most to the risk score predictions (Scott M Lundberg and Lee, 2017; Scott M. Lundberg, Erion et al., 2020). SHAP values allow for not only a global explanation of the model, i.e. which features are most important generally, but also a local explanation, i.e. which features led to a single patient being predicted at risk of being re-hospitalized. It has previously been shown that the interaction between SHAP values and medical doctors can improve the performance of anaesthesiologists (Scott M. Lundberg, Nair et al., 2018).

While the aim of Paper II is to show how modern machine learning techniques can be used to improve the risk prediction process, the usefulness of the SHAP values in a medical context is demonstrated in our paper in Appendix B. The paper uses the SHAP values to compare the preoperative haemoglobin level in the patient with the risk-score, stratified by sex and operation type (knee vs. hip replacement). Currently, the WHO guidelines for the haemoglobin levels are gender specific, however, our study finds no significant gender difference and a haemoglobin threshold close to the WHO suggestions for men (Anaemias and Organization, 1968).

1.3 *COVID-19 and Agent Based Models*

In early 2020, a contagious disease called COVID-19 started to spread in Europe, including Denmark. With new infections showing up faster and faster, governments started to implement different measures to limit the spread of the contagious disease, including lockdowns, travel restrictions, and social distancing, measures not previously seen in peacetime since the Spanish flu in 1918. This was the background for the work that we did in 2020 which became the basis for Paper III, see Chapter 4. This paper deals with the development of a new agent based model for COVID-19 in Denmark in collaboration with Statens Serum Institut (SSI), the Danish Center for Disease Control.

Historically, most mathematical models of infectious diseases were variations of the SIR model, which describe the evolution of a pandemic by approximating all individuals as one population (Kermack, McKendrick and Walker, 1927). As one of the simplest compartmental models, the susceptible-infectious-recovered (SIR) model is based on a system of three non-linear differential equations that describe the transition between each state, or compartment, of the model (Kröger and Schlickeiser, 2020). Initially the entire population is susceptible. At $t = 0$ an outbreak happens where some number of random agents are infected and become infectious, allowing the disease to spread. After having been infectious, the in-

dividuals recover and become immune to the disease and stop being infectious. Several variations of the SIR model exist, including the SIS model, where the recovered individuals become susceptible again (Hethcote, 1989). Another variation is the SEIR model, which includes an exposed state, where individuals are infected but not yet infectious, which is the basis for the model used in Paper III.

SIR-like models suffer from several shortcomings, including the assumptions that the population is homogeneous, and that agents are equally infectious throughout there infectious period. In reality, neither the population nor the transmission rates are homogenous. While multistage SEIR and multicompartment models can help mitigate some of the issues none of these can handle the geographical interactions between agents, which is why we chose to develop an agent based model (ABM) (Tang et al., 2020; Wu et al., 2022). Agent based models simulate individual agents in a population in a way that allows for complex interactions patterns, e.g. based on geographical features such as agent density (Wilensky and Rand, 2015).

In particular, we implemented an event-based, stochastic, spatial ABM using the Gillespie algorithm, a stochastic simulation algorithm (Gillespie, 1977). The model is JIT compiled with Numba (Lam, Pitrou and Seibert, 2015) to speed up the simulation, allowing the simulation of the Danish population of 5.8 million people in a couple of hours instead of days. The model allows for the individual tuning of the three main effects; A) heterogeneities in the infection strength[6], B) heterogeneities in the number of connections[7], C) and the spatial clustering of the agents. In the absence of any of these effects, we find that the ABM's predictions matches the SEIR model's predictions within ±5%. Once we allowed for spatial clustering, we found that the epidemic developed faster and with a higher infection peak compared to the SEIR model, but that the total number of infected in the end of the epidemic was lower.

6 allowing *super-shedders*

7 allowing *super-connecters*

In real-life scenarios, one does not have the opportunity to let the epidemic run loose and afterwards evaluate the strength of the epidemic; the goal is to predict the intensity in the very beginning of the epidemic and implement lockdown-related measured based on this estimate. In the second part of Paper III, we show that once spatial clustering is introduced, fitting standard SEIR-models to infection numbers from the first few days of the epidemic, predictions are overestimated by a factor of two. The result is a significant over-estimation of the impact of the epidemic, in particular the reproduction number $\mathcal{R}_0$ and thus also the number of infected, both the maximal number of simultaneously infected and the endemic steady state number of infected. Since the population is highly susceptible in the beginning of an epidemic, this also highlights the benefits of early lockdowns to reduce the effect of the super+connectors.

The developed ABM was further used by SSI to estimate the effect of contact tracing related to COVID-19 in Denmark, see Appendix C. It was further used to estimate spread of the "alpha" variant of COVID-19 (B.1.1.7) in Denmark, see Appendix D. Based on data available January 2nd 2021, the model predicted that the "alpha" variant would be the dominant variant in Denmark February 10–20, 2021. It became the dominant variant in Week 7: February 15–21, 2021 (Bager et al., 2021).

1.4 *Diffusion Models and Bayesian Model Comparison*

While Section 1.1 discusses the behaviour of ancient DNA, Paper IV focusses on how living cells work and, in particular, how they regulate the transcription of DNA in the cell nucleus. Despite the fact that all cells share the same DNA, the regulation and expression of the genes stored within can vary. The mechanism of the cell-specific expression and silencing of specific genomic regions are one of the most fundamental biological challenges.

Currently, different biological models try to explain the physical principles creating the heterogeneous environment in the cell nucleus of eukaryotic cells. One of these is the polymer-bridging model (PBM) that models the micro compartments called the foci. The cell nucleus contains two different types of loci; the repair foci and the silencing foci. Paper IV studies the physical mechanism of the formation of the silencing foci.

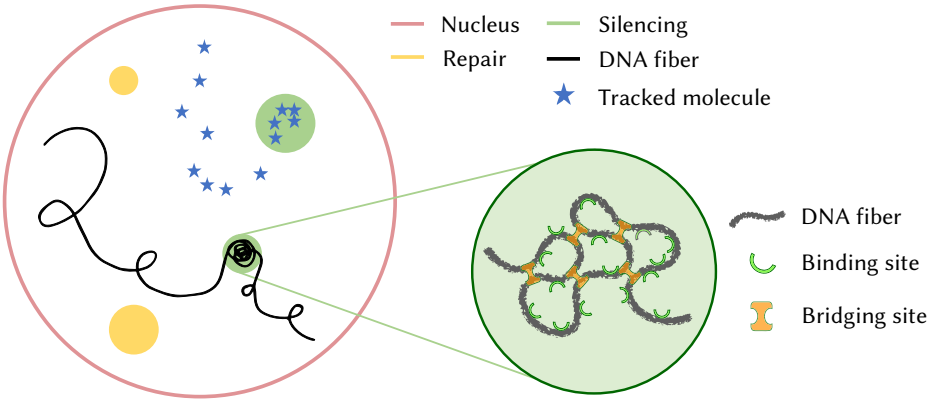


Figure 10. Illustration of the cell nucleus. The nucleus membrane is shown in red and the repair foci in yellow. The black line represents the DNA fiber which is curled up in the silencing foci in green. The right side of the figure shows a zoomed in view of the silencing foci according to the polymer-bridging model with the binding and bridging sites that interact with the SIR proteins. The tracking of the SIR proteins is shown as blue stars. Partly adapted from (Heltberg et al., 2021).

Figure 10 illustrates the parts of the cell nucleus relevant to the polymer-bridging model. Inside the nucleus, DNA fibers are curled up and some parts of the DNA locate inside the silencing foci. Inside the silencing foci, the PBM predicts binding and bridging sites that interact with the DNA fiber through the SIR proteins, which are up-regulated inside the the region of the foci (Heltberg

et al., 2021). The silent Information Regulator (SIR) proteins repress the underlying genes, and, due to the increased concentration inside the focus, the foci are termed silencing foci.

With the use of single particle tracking and photoactivated localization microscopy, it is possible to track the individual SIR protein at high temporal and spatial resolution (Manley et al., 2008; Oswald et al., 2014). As the SIR proteins are assumed to follow a diffusion process, the tracking allows for the determination of the diffusion coefficients of cell nucleus, which help quantify the heterogeneous structure in the nucleus.

Assuming classical Brownian motion in 2D, the displacement lengths, Δr_i, defined as the distances between subsequent observations $\vec{x}$:

$$\Delta r_i = ||\vec{x}_{i+1} - \vec{x}_i||, \tag{10}$$

follows a Rayleigh distribution:

$$\text{Rayleigh}(r;\sigma) = \frac{r}{\sigma^2} e^{-r^2/(2\sigma^2)} \quad r > 0, \tag{11}$$

with scale parameter $\sigma = \sqrt{2d\tau}$, where d is the diffusion coefficient and τ is the time between observations (Anderson et al., 1992) . Using Bayesian mixture models, the switch diffusion process is a simple model describing the system, (Baker, 2021). With $K = 2$ diffusion states, Figure 11 illustrates the model in directed factor graph notation (Dietz, 2022). It shows how the two diffusion coefficients, d_1 and d_2, each define their own Rayleigh distribution, $\mathcal{R}_k$, which are then combined to a mixture distribution, $\mathcal{R}_{1,2}$, with mixing probabilities $\vec{\theta}$. The measured data, Δr, are modelled as N realisations from this mixture distribution.

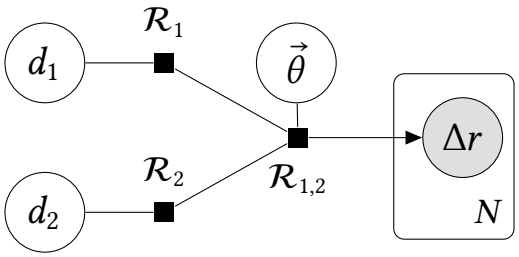


Figure 11.
A graphical representation of the Bayesian model case of two diffusion components using the directed factor graph notation (Dietz, 2022). Here d_1 is the diffusion coefficient, $\mathcal{R}_1$ is the d-parameterized Rayleigh distribution and $\mathcal{R}_{1,2}$ is the mixture model of the Rayleigh distributions with a $\vec{\theta}$ prior.

The diffusion model illustrated in Figure 11 with $K = 2$ diffusion states can be extended to K states, where data shows that both a simpler $K = 1$ model (K_1), the $K = 2$ model (K_2), and a more advanced model with $K = 3$ diffusion states (K_3), all yields appropriate results. Remembering that the formation of the foci depends on the physical properties of the cell nucleus, it is important to be able to evaluate the different models since they provide different diffusion estimates.

The models are compared using the Widely Applicable Information Criterion (WAIC) (Watanabe, 2010) which is a generalized version of the Akaike information criterion (AIC), useful for Bayesian model comparison (Gelman, Hwang and Vehtari, 2014). The WAIC is an approximation of the out-of-sample loss of the model and is defined as:

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}), \tag{12}$$

where the log-pointwise-predictive-density (lppd) is a Bayesian version of the accuracy of the model and p_{WAIC} is a penalty term that penalizes the model for the effective number of parameters (McElreath, 2020). To compare two models, the model with the lowest WAIC is preferred, however, the difference between the WAICs should also be considered. The results for the WT1 dataset from Paper IV is shown in Figure 12. This figure shows the WAIC in black for the K_1, K_2 and K_3 models along with their uncertainties and it is easily seen that the model with only a single diffusion component does not perform well. The difference between the WAIC of the model and the best performing model (K_3) is shown in grey, $\Delta_{A,B}$, where the z-value above the error bars are the number of sigmas the difference is from zero:

$$z = \frac{\Delta_{A,B}}{\sigma_{\Delta_{A,B}}}. \tag{13}$$

Following Occam's razor, the K_2 model is chosen as the optimal model, since the difference between the K_2 model and the K_3 model, the best performing one, is statistically non-significant ($z = 0.57 < 2$).

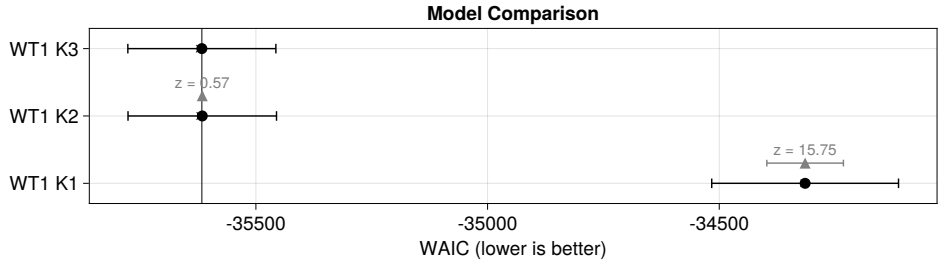


Figure 12. Comparison between diffusion models with $K = 1$, $K = 2$, or $K = 3$ diffusion coefficients for the Wild Type 1 data (WT1). The x-axis shows the WAIC score, where lower values indicate higher-performing models. The WAIC-score for each model is shown in black along with its uncertainty. The difference in WAIC-scores between the model and the best performing model (WT1 K3) is shown in grey with z being the number of standard deviations between them.