Modeling the spatio-temporal organization and segregation of bacterial chromosomes

- Dissertation -
zur Erlangung des Doktorgrades
der Naturwissenschaften
(Dr. rer. nat.)

dem Fachbereich Physik
der Philipps-Universität Marburg
vorgelegt von

David Geisel
aus Bad Homburg v. d. Höhe

Marburg/Lahn, 2021

Erstgutachter und Betreuer: Prof. Dr. P. Lenz
Zweitgutachter: Prof. Dr. K. Drescher

Veröffentlicht in Marburg, 2021

Vom Fachbereich der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am: 10.06.2021

Erstgutachter: Prof. Dr. Peter Lenz
Zweitgutachter: Prof. Dr. K. Drescher

Tag der mündlichen Prüfung: 21.06.2021

Published in Marburg, 2021

Accepted as dissertation by the Department of Physics at the University of Marburg (University ID 1180) on: 06/10/2021

Primary assessor: Prof. Dr. Peter Lenz
Secondary assessor: Prof. Dr. Knut Drescher

Day of thesis defense: 06/21/2021

Abstract

One of the most remarkable findings in biology is that the fundamental processes regulating the inheritance of genetic material and the proliferation of life are conserved over all forms of life on Earth ([4], [144]). This work examined the spatio-temporal organization and segregation of bacterial DNA in order to investigate these fundamental processes. Such analyses are motivated by the multitude of breakthrough discoveries resulting from the study of bacteria that have significantly improved our lives or have the potential to do so in the future ([96], [160], [34]).
For the investigation of the spatio-temporal organization of genetic material in the cell fundamental physical principles were used in this work. The aim was to use concepts of polymer physics to formulate physical models of the complex biological reality. These models were evaluated in computer simulations and compared with experimental data.
In the first project of this thesis, the spatial organization of DNA in multipartite bacteria (= bacteria with multiple replicons) was investigated. Only in recent years, research has recognized that bacterial chromosomes are organized within the cell. This organization is associated with important functional and regulatory processes ([9], [40], [150], [173]). However, there is little evidence for multipartite bacteria. The results of this work reveal high order of spatial organization for multipartite bacteria as well. The organization could be reproduced using a physical model of compacted DNA and geometric constraints on individual genes. Furthermore, it was possible to make accurate predictions for different mutants and to predict interactions between replicons with the developed model. These predictions need to be verified in future experiments.
The second project focused on the study of simultaneous replication and segregation of bacterial DNA. So far, no unified segregation mechanism has been discovered in bacteria ([9], [37], [55]). In the present work, segregation patterns of the origin of replication (ori) were analyzed in the model organism *Bacillus subtilis* (*B. subtilis*). Using Molecular Dynamics (MD) simulations, it was shown that entropic segregation of chromosomes is a plausible mechanism for the segregation of genetic material that would also explain the observed variability in the experimental data.
The model of entropic segregation of bacterial chromosomes was extended in the third project by the implementation of additional segregation mechanisms, so that a large data set of different trajectories of the ori through the cell could be generated. Thus, machine learning (ML) models could be used to classify the different segregation movements. The evaluation of the predictions showed very good results and encourages future classification of experimental data based on the developed models.
This work is intended to provide new perspectives on the organization of DNA in the bacterial cell as well as a better understanding of the physical basis of cellular processes.

Zusammenfassung

Eine der bemerkenswertesten Erkenntnisse der Biologie ist die Tatsache, dass die grundlegenden Prozesse, die die Weitergabe des Erbgutes und die Verbreitung des Lebens regulieren für alle Lebensformen auf der Erde die gleichen sind ([4], [144]). Um diese zu untersuchen, wurden in dieser Arbeit die raum-zeitliche Organisation und Segregation bakterieller DNA untersucht. Motiviert sind solche Untersuchungen durch die Vielzahl von bahnbrechenden Entdeckungen, die das Studium von Bakterien bereits hervorgebracht hat und die unser Leben signifikant verbessert haben oder das Potential haben dies in Zukunft zu tun ([96], [160], [34]).
Ziel dieser Arbeit war es, anhand grundlegender physikalischer Prinzipien die raum-zeitliche Organisation des genetischen Materials in der Zelle zu untersuchen. Dafür wurden Konzepte der Polymerphysik genutzt, um physikalische Modelle der komplexen biologischen Realität zu formulieren. Diese wurden anschließend in Computersimulationen ausgewertet und mit experimentellen Daten verglichen.
Im ersten Projekt dieser Arbeit wurde die räumliche Organisation der DNA multipartiter Bakterien (= Bakterien mit mehreren Replikons) untersucht. Erst seit einigen Jahren hat die Forschung erkannt, dass auch die Chromosome von Bakterien einer Ordnung in der Zelle unterliegen, die mit wichtigen funktionellen und regulatorischen Prozessen verbunden ist ([9], [40], [150], [173]). Allerdings gibt es kaum Erkenntnisse für multipartite Bakterien. Die hier vorgestellten Ergebnisse zeigen auch für diese eine räumliche Organisation der DNA in der Zelle. Diese konnte mit einem physikalischen Modell kompaktifizierter DNA und geometrischen Beschränkungen einzelner Gene reproduziert werden. Außerdem war es anhand des entwickelten Modells möglich, zutreffende Vorhersagen für verschiedene Mutanten zu machen und Wechselwirkungen zwischen Replikons vorherzusagen, die mit zukünftigen Experimenten zu überprüfen sind.
Im Zentrum des zweiten Projekts stand die Untersuchung der gleichzeitigen Replikation und Segregation bakterieller DNA. Bisher konnte in Bakterien noch kein einheitlicher Segregationsmechanismus entdeckt werden ([9], [37], [55]). In der hier vorgestellten Arbeit wurden die Segregationsmuster des Replikationsursprungs, ori, im Modellorganismus *B. subtilis* analysiert. Anhand von MD Simulationen konnte gezeigt werden, dass entropische Segregation der Chromosome ein möglicher Mechanismus für die Separation des genetischen Materials ist, der auch die beobachtete Variabilität in den experimentellen Daten erklären würde.
Das Modell der entropischen Segregation bakterieller Chromosome wurde im dritten Projekt um weitere Segregationsmechanismen erweitert, so dass ein großer Datensatz verschiedener Trajektorien des ori durch die Zelle generiert werden konnte. Dieser ermöglichte es ML Modelle zur Klassifizierung der unterschiedlichen Segregationsbewegungen zu nutzen. Die Auswertung der Vorhersagen zeigte sehr gute Ergebnisse und ermutigt zur zukünftigen Klassifizierung experimenteller Daten auf Basis der hier entwickelten Modelle.
Mit dieser Arbeit sollen neue Perspektiven auf die Organisation der DNA in der bakteriellen Zelle eröffnet und ein besseres Verständnis der physikalischen Grundlagen der zellulären Prozesse vermittelt werden.

Contents

Abbreviations

Alpha anomalous exponent α.

B. subtilis *Bacillus subtilis*.

bp base pair.

C. crescentus *Caulobacter crescentus*.

CID chromosomal interaction domains.

DH Debye-Hueckel.

DNA Deoxyribonucleic acid.

$\triangle$ **pSymA** knock-out mutant with deletion of pSymA.

E efficiency.

E. coli *Escherichia coli*.

EM electron microscopy.

FD fractal dimension.

FJC freely-jointed chain.

G gaussianity.

GB gradient boosting.

kbp kilo base pair.

LJ Lennard-Jones.

LR logistic regression.

Mbp mega base pair.

MC Monte Carlo.

MD Molecular Dynamics.

MSD mean squared displacement.

MSDR mean squared displacement ratio.

ML machine learning.

M. xanthus Myxococcus xanthus.

ori origin of replication.

ParAB parABS system.

dParAB mutant with inactivated ParAB.

PCR polymerase chain reaction.

PDF probability density function.

RF random forest.

RG radius of gyration.

RNA Ribonucleic acid.

S straightness.

SmAB fusion strain of pSymA and pSymB.

SmABC fused strain of the *S. meliloti* replicons.

SMC structural maintenance of chromosomes.

dSMC mutant with inactivated SMC.

dSMCdParAB mutant with inactivated SMC and dParAB.

S. meliloti *Sinorhizobium meliloti*.

SPT single-particle tracking.

SVC support vector classifier.

SVM support vector machine.

V. cholerae *Vibrio cholerae*.

WCA Weeks-Chandler-Anderson.

WLC worm-like chain.

WT wild type.

1. Introduction

1.1. Object of research

The main subject of this work is the examination of two universal features of all cellular life on earth: the organization and segregation of genetic material. Life on earth appears in a great number of different manifestations. It is remarkable that all these different forms of life are very similar in their most basic functions. It is estimated that there are between 10 to 100 million living species on earth today [4]. A central building block of all species is the code in which we store our genetic information: Deoxyribonucleic acid (DNA). We use it to pass our genetic characteristics to our offspring. It does not matter if we look at complex organisms like the human body which is an accumulation of 10^{13} cells or if we investigate the single cell of a bacterium: the basis for the organism is always replication and segregation of the genetic material of one single cell ([4], [144]). This heredity principle stands at the most fundamental definition of life. It specifies the complex system of chemical processes which regulate the maintenance and organization of living cells [4]. In this sense, the cell can be seen as the fundamental unit of life, similar to the atom being the fundamental unit of chemical processes. There is nothing smaller than a cell that is alive [144].
Today, science agrees that all living organisms on earth evolved billions of years ago from a common ancestor [144]. Consequently, it should not matter which organism is studied to understand the basic mechanisms of life, such as metabolism and replication. However, there are many practical reasons for using bacteria as study objects [144]. Some of these advantages include the fact that bacteria are easy to isolate, they grow and replicate well and quickly, and scientists have become very skilled at altering the genetic material of bacteria and creating mutants that can be used to test specific hypotheses. Last but not least, most of the living organisms on earth are single-celled organisms, so that it is reasonable to assign them a corresponding weight in research ([4], [144]).
The study of bacteria has already led to a variety of breakthrough discoveries and applications that improve our lives or have the potential to do so in the future. As early as the seventeenth century, microscopic observations by Hooke and van Leeuwnhoek revealed the cell as the fundamental unit of biological organization [144]. Robert Koch presented another groundbreaking discovery on March 24, 1882, in his paper on the origin of tuberculosis, in which he identified the tubercle bacillus (*Mycobacterium tuberculosis*) as the agent responsible for tuberculosis [96]. In the current COVID-19 pandemic, polymerase chain reaction (PCR) is a key component in containing the spread of the virus by allowing us to test whether a person is infected. PCR makes it possible to create millions of copies of a DNA sequence. This allows to amplify an extremely small sample of DNA to the point where it can be studied in detail ([159], [160]). The foundations of this technique, for which Kary Mullis received the Nobel Prize in Chemistry in 1993, were laid by work on bacteria living in hot springs in Yellowstone National Park [170]. Most recently, the Nobel Prize in Chemistry in 2020 was awarded to Jennifer A. Doudna and Emmanuelle Charpentier for the development of the genome editing method CRISPR Cas-9 ([34], [83]).

CRISPR, or Clustered-Regularly-Interspaced-Short-Palindromic-Repeats and the protein Cas9 together build a system used by bacteria in order to protect themselves from viruses. The Nobel prize was awarded for the conversion of this system into a precise tool for the modification of genomes of living organisms. With CRISPR Cas-9 it is possible to cut the genome of a cell at a specified location in order to remove or add genes *in vivo*. Therefore, there are high hopes that this technology will be able to help people with certain genetic disorders in the future.

Against the background of the great successes achieved in the study of bacteria, the processes within the bacterial cell will also be the object of investigation in the present work. For this purpose, the question of how bacteria manage to organize their DNA in the cell and then pass it on to their offspring in the combined process of replication and segregation will be addressed. Here, the bacterial cell faces major challenges. One of them is the fact that the length of a bacterial chromosome is about three orders of magnitude larger than the cell itself ([9], [197]). As a consequence, cells have to compact their DNA in a manner that is compatible with vital cellular processes. It is therefore not surprising that for a long time it was assumed that bacterial chromosomes fit randomly within cells with no reproducible organization [9]. It is only in recent years that people have begun to understand that bacteria use different physical and biochemical strategies to organize their genomes and establish chromosome architectures on small and large length scales [174]. In doing so, cells depend on being able to accurately duplicate and segregate their DNA to maintain the level of organization ([4], [72]). Another particular difficulty for bacteria is that segregation and replication of DNA occur simultaneously ([9], [173]). Furthermore, bacteria, unlike eukaryotes, do not possess a macromolecular machine for the proper segregation of the duplicated DNA. Instead, bacteria use a variety of segregation mechanism ranging from purely physical forces like entropic segregation of the chromosomes to protein complexes, organizing and separating the chromosomes. Here, too, the molecular mechanisms of chromosome segregation in bacteria are just beginning to emerge. They consist both of specific protein components as well as mechanical-based mechanisms [9].

At this point, the question arises of how to describe and understand the complexity of the bacterial cell and its processes. The cell can be understood as a complex system consisting of a large number of interacting constituents, which is capable of modifying its internal structure and activity patterns due to an exchange of energy or information with the environment [99]. If one wants to describe such a complex system, one has to limit oneself to a certain level of its organization. Thereby the challenge is to neglect the deeper levels of organization while avoiding loss of meaningful information ([99], [127]). In this sense, the goal is to generate an abstraction of the biological problem that is simple enough to be understood by the human mind while at the same time capable of making testable predictions [144]. To achieve this, one needs simple analytical models based on some realistic estimates of the details of the biological system. In the specific case of studying DNA, it will not be useful for us to use an atomic description of the complete molecule. Instead, depending on the question, we need to extract the relevant properties of the DNA to investigate a very specific aspect of its behavior. Therefore, it makes no sense to speak of a sole simple model of DNA. Rather, one uses a variety of models, each as projections of the complex real DNA molecule into a specific conceptual space [144].

An important finding in the search for suitable models to describe DNA was the fact that DNA is a natural polymer composed of monomers called nucleotides. Thus, it was possible to describe DNA with concepts of polymer physics. Today, polymers are a daily part of our

lives. Synthetic polymers are used in a wide range of industrial and medical applications, as well as in the simplest everyday items such as our coffee cup. In addition, silk, DNA or cellulose represent examples of natural polymers. Staudinger laid the foundation for the description of polymers in 1920 with the macromolecular hypothesis. According to this polymers are molecules consisting of covalently bonded monomers ([58], [157]). In the following years, the concepts of polymer physics were further deepened. Important contributions to thermodynamics and conformational statistics were made by P. J. Flory. The study of the macromolecular conformations of DNA was further promoted by the discovery of the double helix structure by Watson and Crick in 1953. The emergence of a harmonic system of models and concepts for describing the fundamental properties of polymers is closely associated with the work of theoretical physicists such as I. M. Lifshitz, S. F. Edwards, and P. G. de Gennes ([58], [157]).
In the present work, the spatial arrangement of genetic material in the bacterial cell, as well as the combined process of replication and segregation of the DNA are analyzed. To this end, physical models are developed to study the complex biological systems. The predictions derived from these models are compared with experimental data. In addition, this work tests known tools from other research areas for their applicability in the context of the bacterial cell. The core of the work consists of three different projects that have dealt with selected of the above-mentioned questions. Taken together, the individual studies provide an impression of how complex biological systems can be transformed into simple models using physical considerations. Thereafter, the obtained models are implemented numerically and finally tested on experimental data.
In the following sections, we begin with a recapitulation of the biological background and the physical approaches to a model-based description of this biological reality.
In chapter 2 the main focus lies on the question of how multipartite bacteria, i.e. bacteria which have a main chromosome and additional plasmids over which their genetic material is distributed, organize their chromosomes in the cell. After several studies have investigated and reproduced the organization of individual chromosomes in typical model organisms, the analysis of multipartite bacteria opens up another field of research ([9], [24], [25], [173], [183], [190], [199], [214]). The project presented here was carried out as part of the Transregional Collaborative Research Center 'Spatiotemporal dynamics of bacterial cells' (TRR174) which is a DFG-funded reseach center comprised of groups from the Marburg and Munich areas. The results presented here were generated in a collaboration with the Becker lab, which provided experimental data on the spatial organization of the model organism *Sinorhizobium meliloti* (*S. meliloti*). On the theoretical side, Monte Carlo (MC) simulations were used to elucidate the differences in the chromosome configuration of monopartite and multipartite bacteria and how these differences affect the interactions of the replicons in the cell. The data and analyses shown here will also be the content of two papers currently in preparation ([134], [193]).
Chapter 3 addresses the simultaneous process of replication and segregation in the model organism *B. subtilis*. In another TRR174 collaboration with the Graumann lab, the segregation of the two origins of replication (oris) was investigated in the model organism *B. subtilis*. Since no uniform mechanism of chromosome segregation in bacteria is known yet, this question is of particular interest ([9], [37], [55]). In this project, a theoretical model of entropic segregation of chromosomes was developed according to a proposal by Arnold and Jun [84] and its predictions were compared with experimental time-lapse data of the segregating oris obtained by the Graumann lab. The results of this collaboration have already been published [37].

Finally, in chapter 4 the model for chromosome segregation is expanded by the implementation of additional segregation mechanisms of the chromosomes. At the same time, however, the focus of the third project is no longer on comparing experimental data with the predictions of a model. Instead, the goal of the third project is to apply the powerful tool of automated classification of trajectories via ML to the classification of the movement of the ori in the cell. In recent years, considerable success has already been achieved in the classification of diffusive motions with ML models ([80], [97], [135], [192]). To try a corresponding application to the study of chromosome segregation in bacteria, synthetic data of segregation trajectories beyond the current possibilities of experiments were produced with the previously developed MD model. For this purpose, different replication and segregation models were combined to simulate distinct cell types which were later classified with ML models. Different ML models as well as different techniques of data preparation were compared and the possibilities of a future application to experimental data, e.g. from single-particle tracking (SPT) experiments, were tested.
A final evaluation of the results of this work, their placement in the current state of research, and possibilities for building future studies can be found in chapter 5.

1.2. Biological context

The biological context of this work is briefly summarized here. We discuss the basic structure of DNA and the challenges bacterial cells face ranging from the compaction and organization of their DNA in the cell to their replication and segregation. Later, we will distill the fundamental parameters for a physical description and computer-based modeling from this broad and complex biological background.

1.2.1. Basic structure of DNA

The information medium that bacteria use to store and organize their genetic information is the same as in any living organism on earth: double-stranded molecules of DNA. The monomers in a DNA strand are called nucleotides. They are made up of a sugar (deoxyribose) that has a phosphate group as well as one of the four bases cytosine (C), guanine (G), adenine (A) or thymine (T) attached to it. The backbone of the DNA is formed by the linkage of the sugars via the phosphate groups. The bases protrude from this backbone and bind to a newly synthesized strand of DNA. Thereby, A always binds to T and C always binds to G via hydrogen bonds. This process is called base-pairing and is responsible for the emergence of a double-stranded structure which consists of two complementary sequences of the bases. The well-known double helix form of the DNA emerges from the twisting of the two strands. In this common helical form the spacing between base pairs is 0.34 nm ([4], [161]). In figure 1.1 a simplified model of a DNA helix is shown.
The order of the nucleotides along the DNA is used to encode information. Thereby, the bases A, C, T and G may be seen as letters of a four-letter alphabet which is used to describe instructions for producing proteins. Proteins are the most important components of the majority of cell functions. They form enzymes which catalyze chemical reactions in the cell, they are used to build the cell structures, they regulate gene expression and enable the cell to move or communicate with other cells [4]. Each segment of DNA containing the coding sequence for the production of a particular protein is called a *gene*. The process in which the genetic information of a gene is converted into a protein is called *gene expression*.

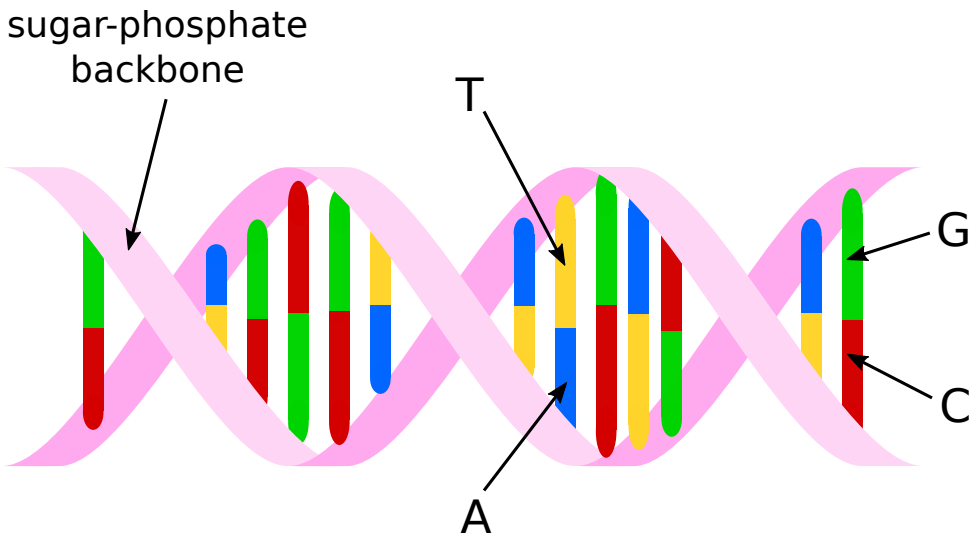

Figure 1.1.: Depiction of a DNA double helix. The sugar-phosphate backbone is shown in pink. The nucleotides are shown in red (cytosine), green (guanine), blue (adenine), and yellow (thymine), depending on the base of the nucleotide.

In the first step the cell converts the nucleotide sequence of the respective gene into another nucleotide sequence coding for an Ribonucleic acid (RNA) molecule. In the next step this is translated into the amino acid sequence of a protein. The number of genes varies dramatically between organisms. The human genome encodes for roughly 24,000 distinct proteins. In contrast, the simplest known cells have under 500 genes [4]. The complete set of genetic information encoded on an organism's DNA is called its *genome*. The genome includes the information for all proteins that the organism will be able to produce. Eukaryotes divide their DNA between a set of different *chromosomes*. However, there is no obvious correlation between the size of the genome and the number of chromosomes or the complexity of the organism. For example the human genome is divided into 46 chromosomes, while somatic cells from a species of small deer contain only 6 chromosomes and a species of carp contain over 100 chromosomes [4]. In contrast to eukaryotes, the majority of bacteria only possess one circular chromosome [133]. However, roughly 10% of all bacteria are multipartite. In order to refer to the different types of DNA molecules that exist within a multipartite genome of bacteria, we use the terminology suggested in [40]. Accordingly, we use *replicon* as a general term referring to any DNA molecule regardless of its specific nature. More specifically, a *secondary replicon* is every replicon that is not the primary chromosome of the cell while obviously the *chromosome* is the primary replicon. The chromosome is always the largest replicon containing the majority of the essential genes. In contrast to a chromosome, *megaplasmids* or *plasmids* are defined by their lack of essential genes. Following [40] we distinguish megaplasmids and plasmids by a lower cutoff of 350 kb for megaplasmid status. A *chromid* is a replicon with an intermediate status between plasmid and chromosome since a chromid carrys at least one essential gene. Finally, a *secondary chromosome* describes a secondary replicon formed as a result of an ancestral chromosome into two replicons [40].
One challenge that eukaryotes and prokaryotes faces alike, whether they have multiple replicons or not, is the compaction of DNA in the cell. If all 46 chromosomes in a human cell would be laid end to end, one would reach a length of approximately 2 m. In contrast, the nucleus to which the DNA is confined, has a diameter of roughly 6 μm [4]. Bacteria face the same problem. Here, the length of a bacterial chromosome is about three orders

of magnitude larger than the cell ([9], [197]). Thus, the task for all cells is to massively compact their DNA in a manner that is compatible with DNA replication, DNA repair and further cellular processes. A brief summary of the main mechanisms that bacteria use for this purpose follows in the next section.

1.2.2. Compaction and organization of DNA in bacteria

The second law of thermodynamics states that closed systems increase their entropy with time. But since living organisms are able to exchange energy and matter with the environment they are no closed systems. This is the reason that we find high degrees of organization in living cells. While it was thought for a long time that bacterial cells do not organize their genome so that it just resides randomly within the cells, nowadays it becomes clearer that there is indeed a very complex organization. This organization is not stochastic but serves functional and regulatory purposes [40]. The high degree of spatial order can already be seen in the variety of cell geometries among different bacterial species. Today the three-dimensional organization of the bacterial genome is expected to play crucial roles in the regulation of gene expression and the establishment of cell fate [150]. Challenged by the above mentioned packing problem, bacteria have come to condense their chromosomes into a spatially ordered structure composed of domains with further subdomains [173]. Thereby, one can differentiate several levels of organization.
At the first level, chromosomal DNA is divided into so-called microdomains. These domains are negatively supercoiled (described below) and form plectonemic loops which are topologically insulated. These independent topological domains are expected to be of size between 10-100kilo base pair (kbp) and are distributed stochastically along the chromosome. Thus, the *Escherichia coli* (*E. coli*) chromosome might consist of approximately 400 domains with an average size of 10kbp. This estimation is in good agreement with the number of loops in chromosomes from lysed cells imaged by electron microscopy (EM) ([9], [149], [172], [197], [206]).
A schematic illustration of a microdomain consisting of plectonemic loops is shown in figure 1.2.

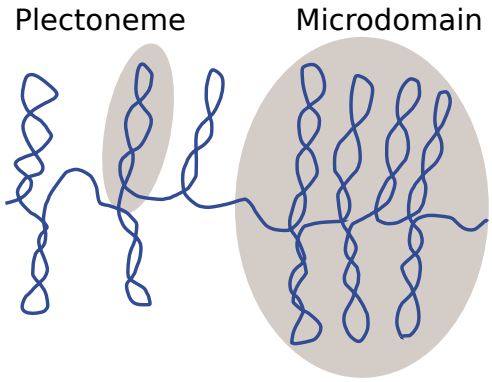

Figure 1.2.: Schematic depiction of plectonemes and microdomain formation. Negatively supercoiled DNA builds plectonemic loops forming microdomains with sizes of $10 - 100$kbp. Adapted from ref. [173]

There are various explanations for how microdomains are formed in the cell. In general, one can differentiate between biochemical and physical mechanisms organizing the bacterial chromosome. On the physical side the DNA can be seen as an oriented helix with a natural

pitch of roughly 10.5 base pair (bp) per twist. Consequently, the DNA exhibits torsional stress if a twist is added or removed. In the case of the circular DNA of bacteria, the strain is partially released by a process called supercoiling. This is the folding of the DNA into plectonemes [173]. Such supercoiling is a first mechanism of condensing the DNA. Another mechanism condensing the DNA in the cell results from the mechanical properties of the chromosome. A typical bacterial chromosome (e.g. *E. coli*) has a contour length of approximately 1.5 mm [86]. At the cellular scale, such a chromosome can be modeled as a very long and flexible polymer with a persistence length of approximately 50 nm. Within the cytoplasm the chromosome is surrounded by a large number of crowding particles. These crowding particles can compress the chromosome into a nucleoid as a result of excluded volume effects. More precisely, whenever the gain in accessible volume for the crowding particles is greater than the loss of conformational entropy of the chromosome, the chromosome will be condensed. It was even shown that this entropic effect produces forces which are sufficient to compact the chromosome to its *in vivo* size ([146], [168]). This phenomenon is called macromolecular crowding. It was suggested that macromolecular crowding provides the basis for the nucleoid compaction which is finalized by biochemical mechanisms as the action of the so-called nucleoid-associated proteins (NAPs) [86]. NAPs influence the structure of DNA locally. Such NAPs bind in large numbers to the DNA and thereby collectively structure the chromosome. One can divide NAPs into the group of DNA benders and DNA bridgers. Some NAPs like FIS in *E. coli* generate local kinks in the chromosome and thereby change the local geometry while others like H-NS in *E. coli* bridge different DNA segments and stabilize topological domains by simultaneously binding to multiple sites ([172], [173], [194], [206]). In figure 1.3 a schematic illustration of NAPs associating with DNA is shown.

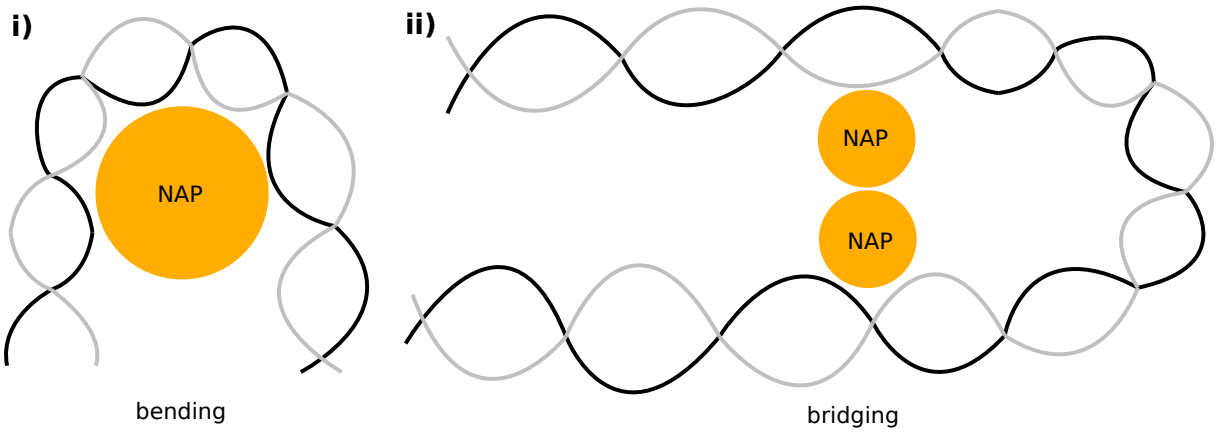

Figure 1.3.: Schematic depiction of DNA bending and bridging by NAPs. (i) Local bending of a DNA segment. (ii) Distant parts of DNA are bridged by NAPs. Adapted from ref. [172]

In addition to the positive effect of the compactification of the DNA through supercoiling, the topological domains also protect the DNA from relaxation, assist in decatenation of chromosomal links and have been proposed to aid in the repair of double strand breaks [197]. A byproduct of the emergence of plectonemes is that distant parts of the DNA are brought into spatial proximity [68]. Another possible structure-function relation associated with the genome packing density is its correlation with genome activity. It is assumed that highly transcribed genes might cluster into transcription factories [21]. Furthermore, transcription contributes to supercoiling as RNA polymerase introduces negative supercoils behind it

and positive supercoils in front [9]. So far it is not known why the microdomains are distributed across the genome as they are, but it is assumed that domain boundaries could help to periodically pause DNA replication in order to promote compaction of recently replicated domains and the decatenation of sister chromosomes [9].
A possible second level of organization of eukaryotic chromosomes are chromosomal interaction domains (CID) identified in contact maps of eukaryotic genomes [21]. They were first described in *Caulobacter crescentus* (*C. crescentus*) where Hi-C analyses revealed that the chromosome is divided into approximately 23 CIDs [101]. The domains received their name from the fact that loci within a domain interact preferentially with each other. In *C. crescentus* the CIDs are created in part by highly expressed genes, thereby supporting the assumption that they are connected to gene activity [101].
On a significantly larger scale than supercoiled domains and CIDs, the bacterial chromosome is further organized into so-called macrodomains [9]. They were first observed in *E. coli* where fluorescence in situ hybridization (FISH) measurements were performed. FISH experiments measure the spatial distance between two DNA segments in single cells using fluorescent probes that bind to specific parts of the chromosome. It was observed that certain loci frequently co-occupy the same restricted space in the cell. In *E. coli* four macrodomains called Ori, Ter, Left, and Right were identified, each with a size of approximately 1 Mb [139]. It is reported that loci within a given macrodomain interact more frequently with each other than with loci in different macrodomains. Furthermore, DNA within macrodomains is more restricted in its movement than DNA in unstructured regions and DNA inversions occuring within a macrodomain are more easily tolerated than those outside of a macrodomain. These findings indicate that macrodomains are an important level of chromosome organization in bacteria ([9], [139], [172]).
At the top level of chromosome organization, the overall arrangement of the chromosome in the cell is investigated. Here, different bacterial species show variations in their chromosome configurations. In *C. crescentus* it was found that the spatial position of the loci within the cell recapitulated the genetic map with the origin of replication at one cell pole and the terminus at the opposite cell pole [190]. This configuration is referred to as the *ori-ter* configuration. In contrast, the origin in slow growing *E. coli* resides near the middle of the cell. The two chromosomal arms arrange to opposite sides of the cell and the terminus is found variably around mid-cell. This configuration is called left-*ori*-right. Interestingly, fast growing *E. coli* cells adopt an *ori-ter* configuration and the chromosome of *B. subtilis* alternates between the two patterns depending on its cell cycle and developemental stage ([9], [173], [190], [197], [199], [206], [216]). Remarkably, an origin proximal centromere-like region *parS* is present in both *C. crescentus* and *B. subtilis*. This region is used for the segregation of the origin and seems to dictate the global orientation of the chromosome since it was found that moving the *parS* region leads to a global rotation of the chromosome such that the relocated *parS* sites are still polar but the origins are not [183]. This discovery suggests that as a consequence of the positioning of parS, the remaining loci are placed indirectly, probably in combination with further processes like compaction. However, the *ori-ter* pattern does not necessarily require pole-anchoring as provided by *parS*. For example Myxococcus xanthus (*M. xanthus*) also adopts an *ori-ter* pattern while having a large cytoplasmic gap between the cell pole and the seemingly not anchored origin [9].
The topic of intracellular organization of the bacterial genome becomes even more complex if we consider the multipartite bacteria. So far, one of the sole studies examining the three-dimensional genome topology in a multipartite genome was done by Val *et al.* [185] with *Vibrio cholerae* (*V. cholerae*). While the main focus of this study was on the

synchronization of replication of the two replicons, the data of the study suggested that the replicons have very different organizations and occupy different locations in the cell. Furthermore, it seems as if the two replicons interact physically. However, so far there is not much known of the organization of multiple replicons within a bacterial cell, with open questions concerning the interreplicon interactions, the possible impact of removal of one or more replicons on the localization of the remaining ones and the examination if there exists a clear spatial division in the cell between the replicons ([40], [185]).

1.2.3. Replication in bacterial cells

Now that we have already addressed how cells can compact and order their genetic material in the cell, we have to turn towards the question of how this order can be maintained while the cell must accurately duplicate its DNA . In doing so, chromosome replication must be highly regulated to ensure a constant number of chromosomes in the cell. Especially chromosomal DNA must be replicated exactly once per cell cycle. The fundamental features of DNA replication have been conserved in all three domains of life (archaea, bacteria, eukaryotes) [4].
The basic mechanism behind the duplication of genetic material is called DNA templating. It makes use of the above mentioned complementary base-pairing within the DNA : A with T and G with C (and vice versa). Thus, a single DNA strand can also be copied into a complementary DNA sequence if each base in the template is recognized by a complementary base. Consequently, the two DNA strands serve as templates for the formation of new strands ([4], [92]). This process is schematically shown in figure 1.4

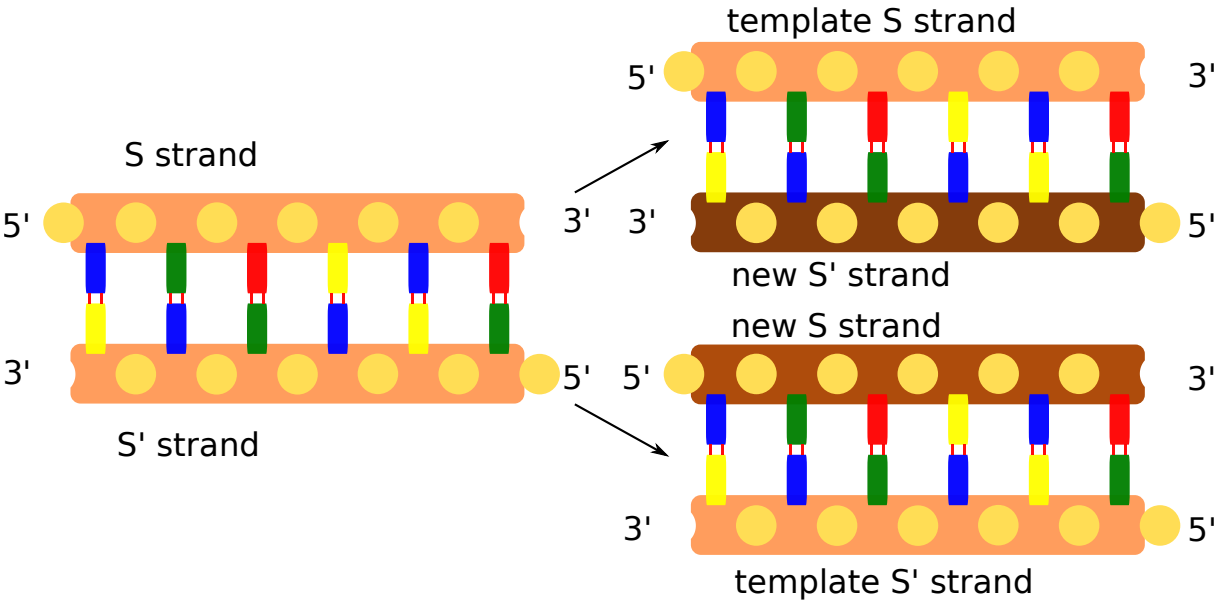

Figure 1.4.: Illustration of the mechanism of DNA templating. The two original strands S and S' each serve as a template for a new strand. As shown in ref. [4]

DNA replication is initiated by a highly organized nucleoprotein complex, the replisome, whose formation is induced by so-called initiator proteins. The localized region of replication that moves along the DNA is called a replication fork [102]. The basic enzymatic functions carried out at the replication fork are well conserved from prokaryotes to eukaryotes. It received its name from its structure in the shape of a Y. It is schematically shown in figure 1.5

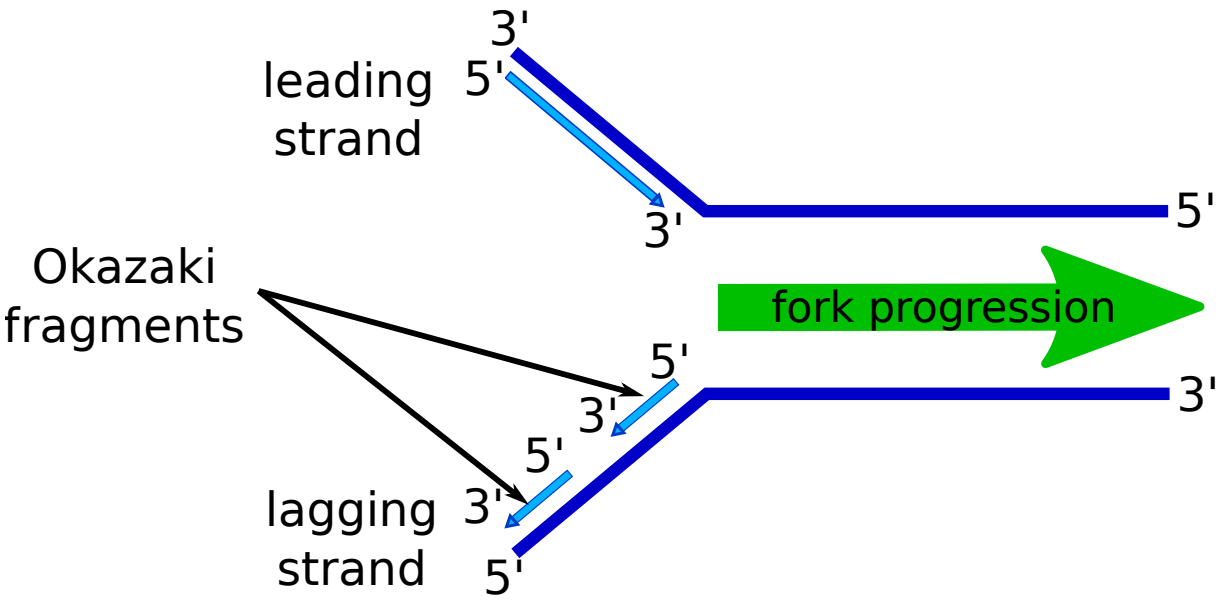

Figure 1.5.: Schematic depiction of the replication fork. Both strands are replicated in the 5' to 3' direction. This happens continuously for the leading strand while the lagging strand is synthesized discontinuously using Okazaki fragments. Adapted from [102].

The two ends of a DNA molecule are named the 5' (five primer) and 3' (three primer) referring to the number of the carbon atom in a deoxyribose sugar molecule to which a phosphate group binds. Synthesis of DNA has a defined 5' to 3' direction. Since the two strands of the DNA are oriented in opposite directions, only the so-called leading strand can be replicated continuously. The remaining lagging strand must be synthesized discontinuously in short, separated segments called Okazaki fragments ([4], [92], [102]).
The replication fork assembles at a defined structure on the DNA called the origin of replication (ori). The origins of replication are specified by signature DNA sequences attracting the initiator proteins. In bacteria, the origin sequences have a length of several hundred base pairs. Here two replication forks assemble which start to replicate the chromosome in opposite directions (bidirectional replication) at a relatively constant speed of approximately 500-1000 nucleotides per second until the replication forks meet in the so-called replication termination region *ter* ([4], [92], [102], [152]).
One question that remains unresolved in this context is whether or not the replisomes are fixed within the cell. In the literature two opposing models are discussed. Within the factory model on the one side, the replisomes are fixed (and possibly linked to each other) at the middle of the cell and the parental chromosome is pushed through this factory-like organization while being duplicated. On the other hand the track model of replication suggests that the relpisomes are individually resolveable and move in opposite directions along the parental chromosome like a train on a track ([81], [103], [104], [117], [152]). A schematic depiction of both models is provided in figure 1.6
The factory model seems to require an additional mechanism to anchor the replisomes to the cell. This might come with the advantage of preventing the replisome from winding along the DNA and interweaving the newly replicated strands. Such an anchor mechanism could be provided by protein complexes organizing the newly synthesized DNA strands behind the replication fork and thereby effectively immobilizing the replisomes [117]. The track model on the other hand does not need such an anchoring mechanism. Instead,

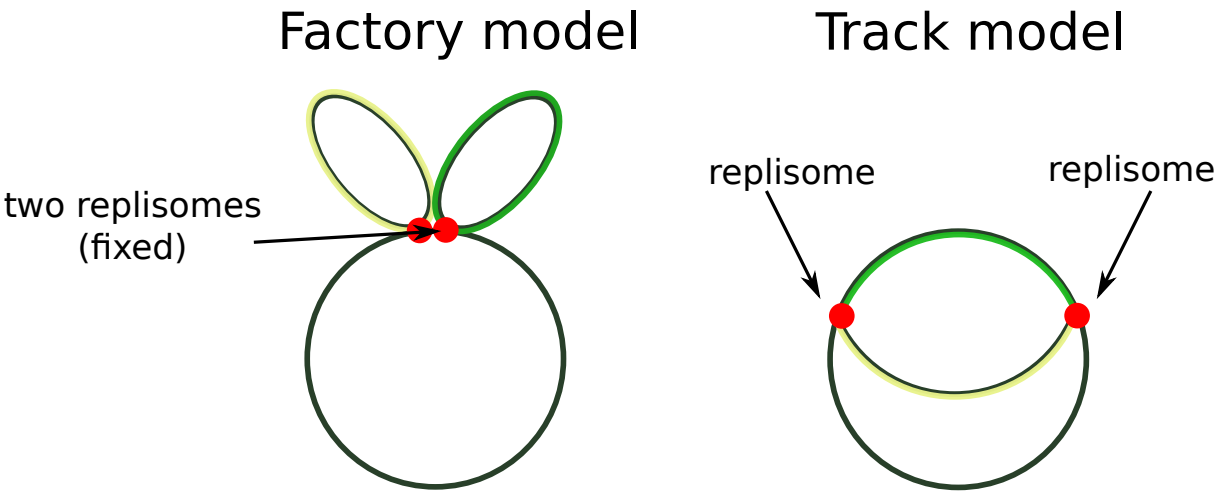

Figure 1.6.: Representation of the factory and track model of replication. The parental chromosome is shown in grey and the newly duplicated strands in yellow and green. In the factory model the two replisomes are fixed at the middle of the cell and the parental chromosome is pulled through this replication factory while being duplicated. In the track model the two replisomes move in opposite directions along the parental chromosome. Adapted from [81].

the localization and movement of the replisomes would be determined by the spatial organization of the chromosome. So far, there are several arguments for both models. Japaridze et al. showed experimental evidence for independently moving replisomes in *E. coli* [81]. This is contrasted by another study of Mangiameli et al. using time-lapse data from *B. subtilis* and *E. coli* reporting that both replisomes reside in close proximity for the most part of the replication period [117]. A further suggestion is that the replication forks are not strictly fixed and connected via a physical link but that they might be confined to a limited region, possibly by the dynamics of the growing nucleoid [152]. So far, it is also unclear whether either of the two replication models provides an advantage in the third big challenge that bacteria are confronted with: the segregation of the newly duplicated DNA into opposite cell halves prior to cell division. It has been suggested that the factory model might facilitate chromosome segregation by pushing the DNA to opposite poles from midcell and thereby preventing mixing of the chromosomes [117]. In the next section, we will discuss further mechanisms by which bacteria segregate their DNA.

1.2.4. Bacterial chromosome segregation

Eukaryotes use a well-understood macromolecular machine, the mitotic spindle, to segregate their genetic material. In contrast, there is no such unique mechanism in bacteria ([9], [37], [55]). Instead, a variety of mechanisms are known to contribute to segregation in bacteria. Again, we can differentiate between physical and biochemical mechanisms, as we did in the section on chromosome compaction. Here, we had already discussed how entropic forces contribute to the compactification of polymers in the cell. The same mechanical properties of a polymer that lead to its compactification in the context of macromolecular crwoding allow us to identify a basic mechanism of spatial separation of two polymers in a confined cell. Again, according to the second law of thermodynamics, the effort of polymers to increase their conformational entropy causes two polymers to repel each other. The reason is that intermingled polymers have less conformational entropy

than completely separated ones. This effect is especially important in confined spaces like the bacterial cell, where polymers behave like loaded entropic springs ([24], [25], [85], [86]). The idea of entropic repulsion as a basic mechanism for chromosome segregation in bacteria was confirmed in both experiments and theoretical simulations. In experiments with *E. coli* cells of increased width it was shown that the probability of successful chromosome segregation decreases with increasing cell width, supporting the prominent role of spatial confinement for chromosome segregation [81]. Also, a number of polymer simulations confirmed the effective segregation of polymers by entropic repulsion resulting from their mechanical properties ([72], [86], [146], [168], [216]). Since entropic repulsion is based on a fundamental physical principle, it can be assumed that it contributes to chromosome segregation in all cells. However, in pure entropic segregation, there is no designated direction of separation. Thus, it may not be sufficient for the the high demands imposed on the organization of the genetic material in a bacterial cell.

One mechanism that provides a directed separation of the chromosomes are partitioning (*par*) proteins. Almost all bacteria use such partitioning systems to segregate their genetic matieral albeit its contribution to segregation varies strongly. The most prominent partitioning complex is the parABS system (ParAB) system. It consists of three components: the DNA sequence *parS*, the DNA-binding protein ParB, and the deviant Walker A-type ATPase ParA. The ParAB system is especially used to segregate the origins. It appears to "pull" the duplicated origin region to the opposite cell pole, where it is anchored while the remaining ori stays at the other cell pole ([37], [42], [79], [108], [199]).

Another group of proteins which play a crucial role in both bacteria and eukaryotes are the structural maintenance of chromosome (structural maintenance of chromosomes (SMC)) proteins. They play a key role in the compaction of DNA and are also believed to facilitate segregation of chromosomes. In *B. subtilis* it was shown that SMC is loaded at ori -proximal *parS* sites where it encircles DNA and thus separates newly replicated origins ([200], [201]). A schematic depiction of how SMC might facilitate the separation of origins is shown in figure 1.7.

It is suggested that SMC and the ParAB work together in the segregation of the genetic material in the cell. While ParAB provides a direction to the segregating origins, SMC ensures compaction of the replicating chromosomes and topologically separates the origins further. Polymer simulations of chromosome dynamics were able to show that loop extrusion by SMC proteins is sufficient to compact and segregate chromosomes ([52], [197], [199], [200]).

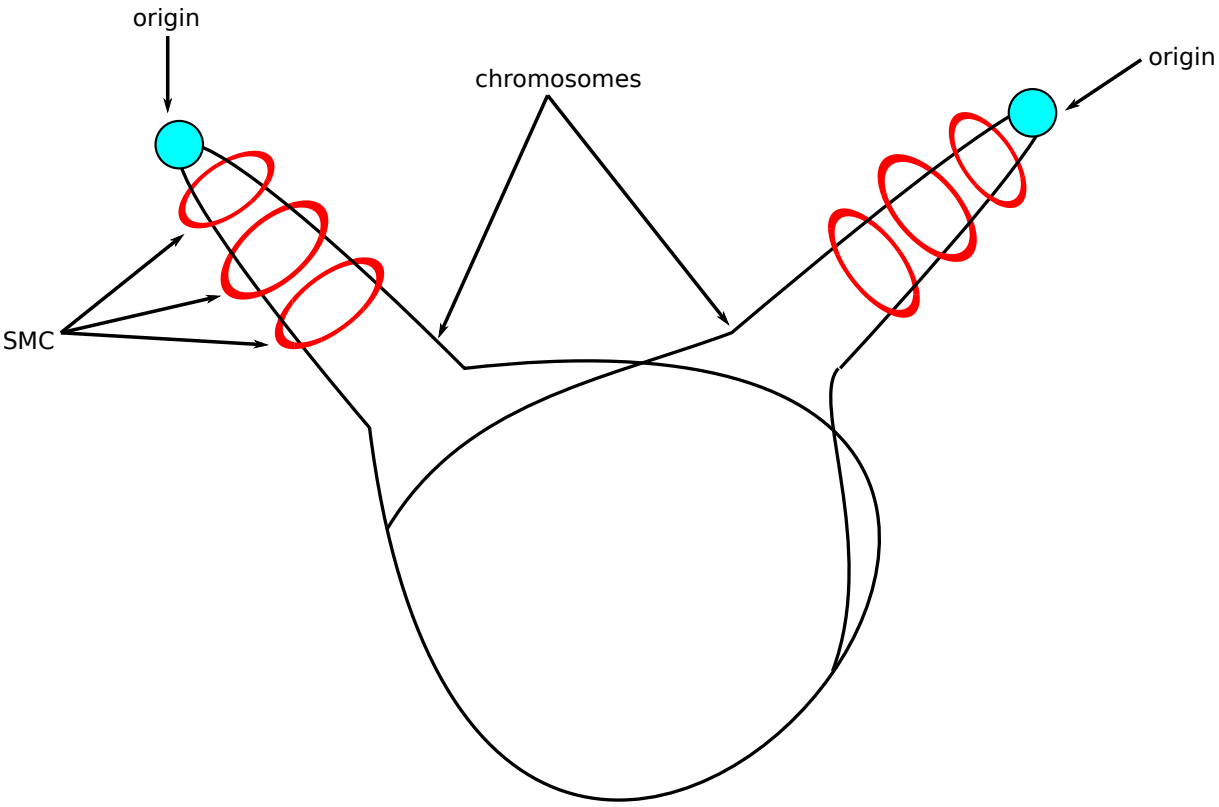

Figure 1.7.: Sister chromosome separation by SMC. SMC is loaded at the origin and encircles the newly replicated DNA strand. Thereby, sister chromosomes are separated. Adapted from [200].

1.3. Modelling approaches

> *"Philosophy is written in this vast book, which continuously lies open before our eyes (I mean the universe). But it cannot be understood unless you have first learned to understand the language and recognise the characters in which it is written. It is written in the language of mathematics, and the characters are triangles, circles, and other geometrical figures. Without such means, it is impossible for us humans to understand a word of it, and to be without them is to wander around in vain through a dark labyrinth."*
> (from Galileo Galilei: Il Saggiatore)

Having gathered the biological knowledge about the bacterial cell in the last section, we now want to evolve the basic concepts for a mathematical-physical description of the phenomena, just as Galileo Galilei requested. The following section describes basic physical models for describing chromosomes with polymer models. In addition, techniques for computer-based modeling are discussed and an overview of current research approaches is given.

1.3.1. Physical models of DNA

In order to create a physical model of DNA in a cell, one has to neglect some chemical details of the monomers and instead develop a coarse-grained model simple enough for theoretical treatment and yet sufficiently detailed to map the most important properties of

DNA on the macromolecular scale. Such properties are, for example, the connectivity, local rigidity and large scale flexibility of the polymer ([30], [127]). Obviously, the more complex the model becomes and the more constraints are considered, the higher the computational cost of analyzing a model. In this sense, there is always a trade-off between the desire to represent as much detail as possible and computational efficiency.
One of the most basic models for a polymer is the freely-jointed chain (FJC). Here, the polymer is simply represented as a succession of N segments (= bonds) of length b. The bonds connect the $N+1$ monomers of which the polymer is composed. A schematic figure of a FJC is given in figure 1.8.

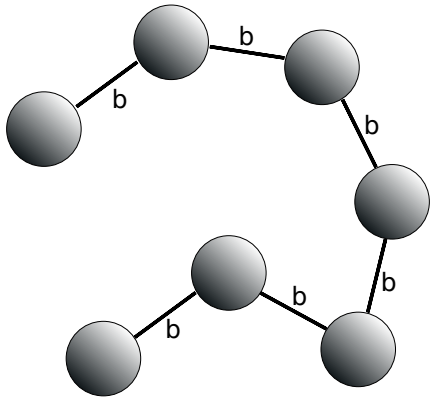

Figure 1.8.: Schematic depiction of a FJC model. The monomers are shown as grey spheres connected by the bonds of length b.

The FJC model only considers the chain connectivity as a property of the polymer but it does not assign a binding energy or torsional stress. Instead, bonds can take arbitrary relative orientations. Furthermore, the model does not account for self-avoidance of the polymer, i.e. the polymer chain is allowed to cross itself and no excluded volume effects are considered. We can represent the configuration of the polymer by noting the positions of each monomer $\vec{r_0}, \vec{r_1}, ..., \vec{r_N}$. The bonds connect consecutive monomers, $\vec{b_i} \equiv \vec{r_i} - \vec{r_{i-1}}$, and are all of length b. A first basic quantitiy characterizing the size of a polymer is the end-to-end vector $\vec{R}$ ([12], [30], [127], [157], [179]). It can be expressed as

$$\vec{R} = \sum_{i=1}^{n} \vec{b_i} \quad . \tag{1.1}$$

We find that the average end-to-end vector $\langle \vec{R} \rangle = 0$ is zero because configurations with end-to-end vector $+\vec{R}$ and $-\vec{R}$ are equally probable. Therefore, the mean square end-to-end distance is used. It can be written as ([12], [30], [127], [157], [179])

$$\langle R^2 \rangle = \sum_{i,j=1}^{N} \langle \vec{b_i} \cdot \vec{b_j} \rangle \quad . \tag{1.2}$$

The bond vectors $\vec{b_i}$ and $\vec{b_j}$ have an angle of ϕ_{ij}. Thus, we can write the scalar product as $\vec{b_i} \cdot \vec{b_j} = b^2 \cdot \cos(\phi_{ij})$. Furthermore, we assumed above that bonds are not correlated with each other and bonds can take arbitrary orientations. Therefore, $\langle \vec{b_i} \cdot \vec{b_j} \rangle = 0$ if $i \neq j$

because all angles have the same probability. This leaves us with contributions only from equal bond vectors $i = j$ and we can write ([12], [30], [127], [157], [179])

$$\langle R^2 \rangle = \sum_{i,j=1}^{N} \langle \vec{b_i} \cdot \vec{b_j} \rangle = Nb^2 \quad . \tag{1.3}$$

With this we can estimate the size of an ideal polymer as

$$R = \sqrt{\langle R^2 \rangle} = \sqrt{N} b \quad . \tag{1.4}$$

We can furthermore define the so called radius of gyration R_g as another measure for the spatial size of the chain. The idea behind the radius of gyration is that the chain roughly occupies a sphere of radius R_g. For its definition we need the center of mass r_G of the chain so that we can write [179]

$$\begin{aligned} R_g^2 &= \langle \frac{1}{N+1} \sum_{i=0}^{N} (\vec{r_i} - \vec{r}_G)^2 \rangle \quad , \\ \text{with } \vec{r}_G &= \frac{1}{N+1} \sum_{i=0}^{N} \vec{r_i} \quad . \end{aligned} \tag{1.5}$$

where we assume that the beads have the same mass and are connected by massless bonds. Instead of using the center of mass of the polymer one can also use the mean square distance between two monomers to obtain the radius of gyration [179]. With this, we can write:

$$R_g^2 = \frac{1}{2} \frac{1}{(N+1)^2} \sum_{i,j=0}^{N} \left\langle |\vec{r_i} - \vec{r_j}|^2 \right\rangle \text{ any conformation.} \tag{1.6}$$

For equation 1.6 no specific chain model is assumed. Therefore, it applies to any chain conformation. It indicates that R_g^2 is half of the average square distance between two monomers on the chain [179].
The advantage of the radius of gyration is that it allows to estimate the size of arbitrary configurations of polymers. We can explicitly calculate the radius of gyration for an ideal chain like the FJC model or a bead-spring model. For this we use that the end-to-end distance of an ideal chain is according to equation 1.4 Nb^2. Since the part of the ideal chain between any i-th and j-th monomer also is an ideal chain, we can write for the end to end distance of this part $\langle |\vec{r_i} - \vec{r_j}|^2 \rangle = b^2 |i - j|$. Thereby, we replaced N with $|i - j|$ [179]. Inserting this into equation 1.6 yields

$$\begin{aligned} 2R_g^2 &= \frac{1}{(N+1)^2} \sum_{i,j=0}^{N} b^2 |i-j| = \frac{2b^2}{(N+1)^2} \sum_{i=0}^{N} \sum_{j=0}^{i} (i-j) \\ &= \frac{2b^2}{(N+1)^2} \sum_{i=0}^{N} \frac{1}{2} i(i+1) = b^2 \frac{N(N+2)}{3(N+1)} \quad . \end{aligned} \tag{1.7}$$

In the limit of large N we can thus write

$$R_g^2 = \frac{b^2 N}{6} \quad . \tag{1.8}$$

Another quantity used to describe a polymer is the probability distribution function $P(\vec{R}, N)$. It describes the probability distribution for the end-to-end vector of the chain consisting of N segments to equal $\vec{R}$. In the case of the FJC the vector $\vec{R}$ equals the sum of N independent, randomly oriented contributions $\vec{b_i}$. Thus, the central limit theorem of probability theory states for $N >> 1$ that the probability distribution becomes a Gaussian

$$P(\vec{R}, N) \approx \left(\frac{3}{2\pi N b^2}\right)^{3/2} \exp\left(-\frac{3}{2}\frac{R^2}{Nb^2}\right) \quad . \tag{1.9}$$

For this reason, such polymers are also called Gaussian polymers or ideal chains ([12], [30], [58], [127], [157]).
We can use the probability distribution of our FJC model to investigate further scaling laws. For example, we can look at the entropy S of an ideal chain. The entropy is defined as

$$S = k_B \ln \Omega \quad , \tag{1.10}$$

with $\Omega(N, \vec{R})$ denoting the number of conformations of a freely jointed chain of N monomers with end-to-end vector $\vec{R}$ and k_B is the Boltzmann constant [157]. One can use the probability distribution function of equation 1.9 to calculate the free energy of an ideal chain from equation 1.10 (see appendix A.1). One finds

$$F(N, \vec{R}) = \frac{3}{2} k_B T \frac{\vec{R}^2}{Nb^2} + F(N, 0) \quad , \tag{1.11}$$

as the free energy of the chain [157]. The result of equation 1.11 indicates that the free energy increases quadratically with $\vec{R}$. This shows similarity to Hooke's law, indicating that an ideal chain has a spring-like entropic elasticity. We can also calculate the force needed to separate the ends of a FJC polymer by a distance $\vec{R}$ as

$$f = \frac{\partial F(N, \vec{R})}{\partial \vec{R}} = \frac{3 k_B T}{Nb^2} \vec{R} \quad . \tag{1.12}$$

With this we find that the force of the spring has an "entropic spring constant" of $3k_BT/Nb^2$ [157]. This is an important result which will be used again below when we estimate the excluded volume effects of self-avoiding polymers with the Flory theory. Furthermore, we will compare the linear dependence of the streching force for an ideal chain with the force required to strech a polymer under confinement below.
Another important polymer model is the worm-like chain (WLC) model (or Kratky-Porod model). Here, DNA conformations are described by a three-dimensional space curve $\vec{r}(s)$ of fixed length L, where s is the arc length of the curve ([122], [123]). Thus, the WLC model represents the continuum limit of the FJC model for the bond length $b \to 0$. The geometric properties of every continuous, differentiable curve in the $\mathbb{R}^3$ are described by the Frenet-Serret formulas (see appendix A.2) that say that each curve is determined by the two parameters of the local Frenet-Serret curvature (= bending of the central axis)

and torsion (= twisting of the curve) [91]. Consequently, we can separate any distortions of a DNA molecule into distortions of the central axis and distortions defining the internal twisting of the double helix. For the description of twisting we use that the double helix repeat states that relaxed DNA makes one turn every $h = 3.5nm$ and thus the spatial angular frequency ω_0 of relaxed DNA can be expressed as $\omega_0 = 2\pi/h = 1.85nm^{-1}$. Deviations in the twisting rate from ω_0 can be described by a scalar field $\Omega(s)$. With this, we can write the elastic energy of a DNA molecule as [123]

$$\frac{E_{el}}{k_B T} = \frac{1}{2}\int_0^L ds \left[\underbrace{A(\partial_s^2 \vec{r})^2}_{\text{bending}} + \underbrace{C\Omega^2}_{twisting} \right] \quad . \tag{1.13}$$

Here, the curvature is given by $|\partial_s^2 \vec{r}|$ and A is the bending persistence length, while C is the twist persistence length. A can be defind by the exponential decay of tangend vecors correlation:

$$\langle \vec{t}(s) \cdot \vec{t}(s') \rangle \sim \exp\left(-|s - s'|/A\right) \quad . \tag{1.14}$$

Thus, we see at this point that the WLC model now also takes into account the energy cost of introducing local bends in the chain, which was not considered in the FJC model. The twisting energy arises from deviations in the double helix twist from the equilibrium state ω_0. In aqueous solution of 0.14M univalent salt the bending persistence length equals $A \approx 50nm$ and the twist persistence length is $C \approx 75nm$ ([121], [122], [123]).

1.3.2. DNA topology

With the WLC model of DNA, we can now make topological considerations to understand the effect of supercoiling, which is elementary for the compaction of DNA in the cell. Therefore, we define a topological invariant, called the linking number L_k. It describes the number of times the two strands of a DNA, described by two curves C and C' paramterized by $\vec{r}(s)$ and $\vec{r'}(s')$ wind around each other. The linking number can be calculated using the Gauss linking integral [191]:

$$L_k = \frac{1}{4\pi} \oint_{C'} ds' \oint_C ds \frac{\vec{r'}(s') - \vec{r}(s)}{|\vec{r'}(s') - \vec{r}(s)|^3} \cdot \left[\frac{d\vec{r'}(s')}{ds'} \times \frac{d\vec{r}(s)}{ds} \right] \quad . \tag{1.15}$$

While the expression of the linking number with the Gauss integral considers two closed curves that do not touch each other, DNA appears as a single filament at long length scales. Thus, one would like to recast the linking number in terms of the single polymer picture of DNA [91]. This is done in the appendix A.3. In this way, an important result obtained by White and Fuller ([49], [208]) is obtained:

$$L_k = T_W + W_r \quad . \tag{1.16}$$

The result of equation 1.16 is also called Calugareanu's theorem [1]. Here, the linking number L_k is written as the sum of the twist T_W and the writhe W_r. Twist and writhe are not topological invariants. The twist describes the number of helical turns of one strand around the other while the writhe states how many times the double helix crosses itself.

Thus, the writhe can adopt positive or negative values depending on the orientation. We can express both quantities analytically (see appendix A.3). For the twist we find

$$T_W = \frac{1}{2\pi} \oint_C ds\ \vec{t}(s) \cdot \left[\vec{n}(s) \times \frac{d\vec{n}(s)}{ds} \right] \quad , \tag{1.17}$$

and the writhe one finds

$$W_r = \frac{1}{4\pi} \oint_{C'} ds' \oint_C ds \frac{\vec{r}(s') - \vec{r}(s)}{|\vec{r}(s') - \vec{r}(s)|^3} \cdot \left(\vec{t}(s') \times \vec{t}(s) \right) \quad . \tag{1.18}$$

Although equation 1.18 has a strong resemblance to the Gauss linking integral of equation A.24, the equations are not identical because equation A.24 considers two different curves and equation 1.18 is for the same curve [91]. An exemplary representation of the formation of twist and writhe at DNA is shown in figure 1.9.

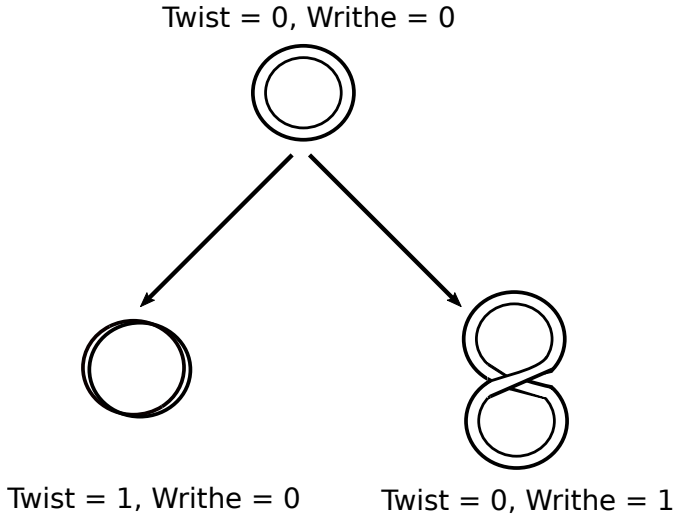

Figure 1.9.: Formation of twist and writhe at DNA.

Important for the understanding of supercoiling is the fact that circular DNA with covalently linked ends is topologically constraint. In this case the linking number cannot be changed without cutting at least one of the DNA strands open. We can further express the linking number of a relaxed DNA molecule in a planar circle. Such a DNA only has twist as a result of the double helix repeat but no writhe. Thus, the linking number can be written as

$$L_{k_0} = \frac{L}{h} = \frac{\omega_0 L}{(2\pi)} \quad . \tag{1.19}$$

The state L_{k_0} is energetically most favorable and DNA is supercoiled when $\triangle L_k = L_k - L_{k_0} \neq 0$ ([22], [23]). We speak of negatively supercoiled DNA when $\triangle L_k < 0$ and of positively supercoiled DNA for $\triangle L_k > 0$. The DNA of most bacteria have linking numbers about 5% less than that of the relaxed double helix. Thus, they are negatively supercoiled. Such changes in the linking number are produced by specialized proteins in the cell called topoisomerases. These proteins are able to introduce breaks in one

DNA strand and pass the other strand through the break before closing it again. More specific, type I topoisomerases are able to break one of the two DNA strands and pass the other strand through the gap, thereby increasing or decreasing the linking number by 1 through increasing or decreasing the twist. On the other hand type II topoisomerases break both DNA strands and pass the entire double helix through the gap. Thereby, they increase or decrease the linking number by 2 through increasing or decreasing the writhe. When the interwinding of the DNA reaches a critical twist density the molecule buckles to form plectonemic structures as a result of competition between entropy and elastic energy. If further turns are introduced, one observes a rapid decrease in extension of the molecule as twist is traded for writhe. In this way, supercoiling contributes to the compaction of DNA. Besides this, supercoiling also has an effect on separation of DNA. If negatively supercoiled DNA is separated, more twists are created in the rest of the DNA causing rewinding of the unwound strands. Thus, the DNA that is still base paired is driven towards the relaxed state which is energetically favoured. Therefore, negatively supercoiled DNA is separated easier than relaxed or positively supercoiled DNA ([22], [23], [121], [122], [123], [175], [176]).

1.3.3. Excluded volume effects

Another important property of real polymers that we have not considered so far are monomer-monomer interactions. Especially important are the so-called excluded-volume effects, taking into account the fact that real polymers are self-avoiding. This additional property leads to important changes in the configurations of polymers. A first simple but successful description of excluded-volume effects is provided by the Flory theory ([12], [41], [157]). The goal of the Flory theory is to describe the balance between the repulsive energy of the self-avoiding monomers and the entropy loss due to the arising chain deformations. Although the Flory theory makes rather rough estimates in order to determine the energetic and entropic terms of the free energy of a self-avoiding polymer, it still yields results which are in supringsingly good agreement with both experiments and more sophisticated theories ([12], [60], [157]).
Again, we start by considering a polymer consisting of N monomers of size b. In order to take self-avoidance into account, we assign an excluded volume ν to every monomer [157]. This extension of our model results in an effective repulsion of the monomers of the polymer on small length scales. We denote with R the size of the swollen polymer and with $R_{id} = b\sqrt{N}$ (see equation 1.4) the size of an ideal chain. Furthermore, we assume that monomers are distributed uniformly within the volume R^3 and that besides the excluded volume interactions no further correlations between monomers exist ([12], [60], [157]). In this case, the volume occupied by the polymer scales like R^3. We can also say that the probability to find a monomer within the excluded volume of another monomer can be written as a product of the excluded volume ν and the monomer number density $N/V \sim N/R^3$. Using this, the Flory theory assumes that the energetic cost of being excluded from this volume is $k_B T \nu N/R^3$ per monomer [157]. Thus, we can write the energetic term of the free energy by multiplying the cost per monomer with the number N of monomers

$$\mathcal{F}_{int} \approx k_B T \nu \frac{N^2}{R^3} \quad . \tag{1.20}$$

What remains is to estimate the entropic term of the free energy. Here, the Flory theory estimates that the entropic contribution to the free energy of a real chain is the energy required to stretch an ideal chain to end-to-end distance which we obtained in equation 1.11 as $\mathcal{F}_{id} \approx k_B T \frac{\vec{R}^2}{Nb^2}$ [157].
Finally, we can write down the Flory estimate of the free energy by summing up the energetic term 1.20 and the entropic term

$$\mathcal{F} = \mathcal{F}_{int} + \mathcal{F}_{id} \approx k_B T \left(\nu \frac{N^2}{R^3} + \frac{R^2}{b^2 N} \right) \quad . \tag{1.21}$$

With this we can now estimate the optimal size of the self-avoiding polymer by minimizing the free energy

$$\begin{aligned} \frac{\partial F}{\partial R} = 0 &= k_B T \left(-3\nu \frac{N^2}{R_F^4} + 2 \frac{R_F}{N b^2} \right) \\ &\rightarrow R_F^5 \approx \nu b^2 N^3 \\ &\rightarrow R_F \approx \nu^{1/5} b^{2/5} N^{3/5} \quad . \end{aligned} \tag{1.22}$$

The result indicates that while the self-avoiding polymer still shows a scaling of its size with the number of monomers, the power law has changed compared to the ideal polymer $R_{id} \approx b N^{1/2}$ from equation 1.4.
One can compare the size of long ideal and long real chains (with same numbers of monomers) with the so-called swelling ratio

$$R_F / R_{id} \approx (\nu N^{1/2} / b^3)^{1/5} \quad . \tag{1.23}$$

The swelling ratio states that due to the excluded volume interactions we find an increase of polymer size for the real polymer compared to an ideal polymer. A further important result of the Flory theory is obtained by computing the free energy of a real polymer for arbitrary dimensions [157]. In this case, one obtains

$$\mathcal{F} \approx k_B T \left(\nu \frac{N^2}{R^d} + \frac{R^2}{b^2 N} \right) \quad , \tag{1.24}$$

where only the energetic term depends on the dimension d while the entropic term still is the one of an ideal polymer, independent of d. Minimization of this generalized form yields a famous universal power law for the scaling of the polymer size with the number of monomers

$$R_F \sim N^{\nu} \quad . \tag{1.25}$$

The exponent ν in equation 1.25 is called the Flory exponent. It depends on the dimension d as $\nu = \frac{3}{d+2}$.
As already mentioned at the beginning of this section, the Flory theory is not perfect in the sense that it makes some estimation errors. The suprisingly good results of the theory are due to some degree to cancellation of the errors. The most important errors are that the Flory theory overestimates the repulsion energy by ignoring correlations between monomers along the backbone of the chain. On the other hand, the entropic energy is also overestimated as the Flory theory simply assumes the conformational entropy of an

ideal chain although real chain conformations are different from the ones of ideal chains. Nevertheless, the Flory theory still is seen as an useful approach to the conformations of self-avoiding polymers and the Flory theory result of $\nu = 3/5$ in three dimensions is very close to the exact value as obtained by renormalization group techniques of $\nu = 0.588$.

1.3.4. Self-avoiding walks

In the following we discuss modelling approaches including the self-avoiding of polymers. While the WLC model is obviously useful to describe effects like supercoiling and polymer behaviour on the scales of the persistence lengths, it is often not necessary to consider molecular details. Instead, one deliberately makes use of coarse-grained models, neglecting structures below the persistence length. Mostly, such coarse-grained models represent the polymer as a random walk on a three dimensional (cubic) lattice ([114], [115], [127], [171], [188]). Typically, self-avoidance is introduced in lattice models by preventing multiple occupancies of grid points, in order to construct a self-avoiding random walk (SAW). Surprisingly, the SAW represents an accurate model for polymers. This is true although a real polymer molecule lives in continuous space, has tetrahedral bond angles, a non-trivial energy surface for the bond rotation angles, and a complicated monomer-monomer interaction potential. On the other hand the SAW lives on a discrete lattice, has non-tetrahedral bond angles, and an energy independent of the bond rotation angles plus a repulsive hard-core monomer-monomer potential. Nevertheless, it was shown that both systems exhibit the same asymptotic behavior independent of their chemical details ([26], [171]). Some basic statistics of random walks can be found in appendix A.4.
If we aim to simulate bacterial chromosomes with a coarse-grained model this implies that we have to construct closed SAWs (or self-avoiding polygons (SAP)) since bacterial chromosomes are circular. It is possible to determine the number of distinct (up to translation) SAPs with N segments embedded in the cubic lattice, p_N, with the formula from Hammersley [66]:

$$\lim_{N\to\infty} \frac{1}{N} \log p_N \equiv \kappa \quad . \tag{1.26}$$

Here, κ is the limiting entropy per step and depends on the lattice. The investigation of such geometrical confined polymers is a challenging task and usually requires the use of numerical techniques such as MC approaches ([100], [114], [115], [127], [171]). Efficient MC simulations of SAWs (as models of polymers) started in 1955 with the invention of the Rosenbluth algorithm [156] which was later generalized by the PERM [53] and GARM [153] algorithms. In addition, progress was made through the development of the BFACF algorithm ([10], [27]) and the pivot algorithm ([100], [114]). Depending on the object of investigation, one can find different algorithms that are adapted to the problem. Thus, SAW can be divided into different ensembles depending on whether the lengths and endpoints are fixed or not. In the simulation of bacterial chromosomes, one is interested in the fixed-length, fixed-endpoint ensemble. Here, the MOS algorithm [115] has proven to be very powerful and ergodic in any dimension for the cubic lattice (see appendix B.1.1 for detailed description).

1.3.5. Confined polymers

In addition to defining the overall length and position of the endpoints of the polymer chains, it is also possible to introduce additional spatial constraints like the confinement of the polymer by the cell. Confined polymers are in general an important field of study as they play an important role in both industrial processes and biological systems. Examples for industrial applications are membrane filtration or oil recovery. For us, the effect of confinement of the DNA by the cell is of particular interest. A number of studies have been able to show that spatial confinement of polymers has a great influence on their spatial arrangement as well as their segregation behavior ([7], [16], [29], [33], [60], [84], [85], [120], [130]).
To find a description for two polymers under confinement, one can start by estimating the free energy cost of two overlapping polymers without confinement. Here, the Flory theory can be used for a first estimation. The relevant considerations of the Flory theory for this case are summarized in the appendix A.5. The Flory theory comes to the estimation that long polymers shuld behave as mutually impenetrable hard spheres. However, as mentioned above, the Flory theory makes some estimation errors. One main mistake of the Flory theory is that it assumes that the self-avoiding monomers are independently distributed in the volume. Thereby, the Flory theory ignores the linkage of the monomers along the backbone of the polymer. This linkage causes correlations between the monomer positions and leads to a different estimate of the free energy of a confined polymer. It was shown by Grosberg *et al.* [57] that in fact

$$\mathcal{F} \sim k_B T \quad . \tag{1.27}$$

This result now indicates that polymers in bulk can rather easy intermingle ([84], [157]). The question now is how this behavior changes if the polymers are exposed to spatial confinement. To consider this case, the so-called blob picture of de Gennes [30] is very helpful. In order to introduce the blob picture it is convenient to start by considering a self-avoiding polymer consisting of N monomers of size b under tension. We already know the end-to-end distance of the polymer in the unperturbed state from equation 1.22 as $R_F \approx \nu^{1/5} b^{2/5} N^{3/5}$. If we assume the excluded volume to scale as $\nu \approx b^3$ we get

$$R_F \approx b N^{3/5} \quad . \tag{1.28}$$

The central idea of the blob picture is to subdivide the polymer chain into segments of size ξ. The sections of size ξ are called blobs and consist of g monomers each. It is assumed that at this small length scales the chain statistics behave like that of an unperturbed chain. Thus, in this concept a 'blob' is defined as the largest unit of a polymer that shows the characteristics of an unperturbed chain. Here, each blob contains an extended piece of the polymer chain. The chain inside each blob does not experience the confining constraints. Thus, blobs can be interpreted as the effective monomers of the polymer. A schematic figure of this 'blob-picture' is shown in figure 1.10
We can estimate the size of a blob using the above formula for the end-to-end distance of an unperturbed chain. Thereby, we receive

$$\xi \approx b g^{3/5} \quad . \tag{1.29}$$

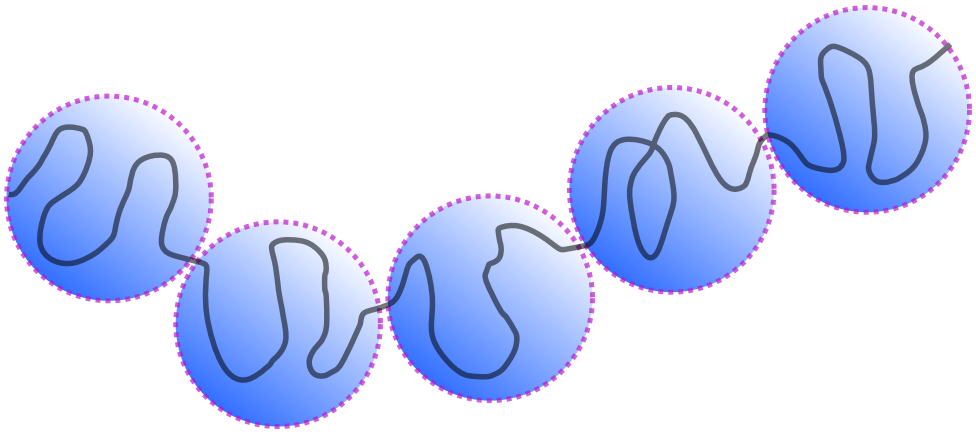

Figure 1.10.: Schematic depiction of a polymer chain in the blob picture. The polymer chain is shown as the grey line and can be divided into the blobs which are shown as blue spheres. The chain segments inside a blob behave as an unconstrained chain. Adapted from [85].

The size of the complete polymer under tension R_f is just the product of the number of blobs $n_{blobs} = \frac{N}{g}$ and the size of a blob

$$\begin{aligned} R_f \approx \xi \frac{N}{g} \underbrace{\approx}_{\text{equ. 1.29}} N \frac{b^{5/3}}{\xi^{2/3}} \approx \frac{R_F^{5/3}}{\xi^{2/3}} \\ \rightarrow \xi \approx \frac{R_F^{5/2}}{R_f^{3/2}} \quad . \end{aligned} \tag{1.30}$$

We can also express the free energy of the streched polymer. For this we make use of the above result stating that the overlap of two blobs comes with a free energy cost of order k_BT. We can use this result because the polymer chain inside a blob behaves as an unperturped chain. Thus, we can write the free energy as

$$\mathcal{F} \approx k_BT \underbrace{\frac{N}{g}}_{n_{blobs}} \approx k_BT \frac{R_f}{\xi} \approx k_BT \left(\frac{R_f}{R_F}\right)^{5/2} \quad . \tag{1.31}$$

From here, we obtain the force required to stretch the polymer as the derivative

$$f = \frac{\partial \mathcal{F}}{\partial R_f} \approx \frac{k_BT}{R_F} \left(\frac{R_f}{R_F}\right)^{3/2} \quad . \tag{1.32}$$

This result contrasts with the behavior of ideal chains where the force to stretch the chain was linearly proportional to R_F (see equation 1.12). Here, we find that for real chains the stretching force scales with $R_f^{3/2}$.
Now we are ready to turn to the case of confined polymers. Assume that a polymer is confined in an infinitely long cylinder of diameter D. In this case, the diameter of the cylinder defines the size of the blobs since on length scales smaller than D the chain

segments are not effected by the confinement and thus behave like unperturbed chains. Thus, we can write analog to equation 1.29

$$D \approx bg^{3/5} \quad . \tag{1.33}$$

For the free energy we can use the result of Grosberg et al. of a free energy cost k_BT per blob and write for the free energy of the polymer

$$\mathcal{F} \approx k_BT \underbrace{\frac{N}{g}}_{n_{blobs}} \approx k_BTN \left(\frac{b}{D}\right)^{5/3} \quad . \tag{1.34}$$

Thus, the repulsion between two confined polymers is very strong and proportional to the chain length N. This is in stark constrast to the weak repulsion found for polymers in bulk. In fact, it was suggested that the repulsive force might be a key element for segregation of bacterial chromosomes ([7], [60], [84], [132], [157]).

1.3.6. Recent trends

At this point, we will briefly review several studies that investigate the organization and dynamics of DNA using computational models. To discuss the different approaches to modeling DNA in bacteria, it makes sense to group them into classes beforehand. A first division of the approaches is the one in mechanism-based modeling strategies on the one hand and data-driven approaches on the other hand [150]. The former are based on a certain mechanistic conception of DNA as a polymer and use this to carry out simulations. The simulation results are subsequently compared with experimental data. In contrast, the latter are fed directly with experimental data. In particular, the advent of the increasingly available Hi-C data sets has led to the emergence of many approaches that attempt to elucidate the structural properties of chromosomes. Hi-C methods are a subclass of chromosome conformation capture (3C) methods which generate genome-wide contact probabilities between loci along the chromosomes. In many studies a relationship between these contact probabilities and the spatial distance of the loci is assumed [214]. The two classes of model approaches can be further subdivided. In the review of [77], a subdivision of data-driven models into consensus structure ensembles and data-driven ensembles is proposed while the mechanistic or *de novo* models might be categorized as structural ensembles and mechanistic ensembles. Data-driven models are distinguished according to whether an attempt is made to reconstruct a single chromosome structure (the consensus structure model) from Hi-C data or if one aims at producing an ensemble of structures which reconstructs the Hi-C data (ensemble methods). Typically these methods use some sort of a polymer description with a set of (flexible) constraints and try to infer interactions between the monomers by fitting the Hi-C contact map. Therafter, they sample the space of possible conformations using MC or MD simulations using the inferred interactions to generate a set of conformations which reproduces the Hi-C data [77].
The *de novo* approaches do not infer chromosome conformations from experimental data. Instead, one typically tests physical hypotheses with the simulations. The structural ensemble methods use typical polymer models like random walks or self-avoiding walks to gain insights into chromosome organization. Thereby, one usually does not consider a specific biological mechanism but rather the statistical properties of the ensembles. An example are lattice MC methods where rather "unphysical" global moves are often applied

to the conformations in order to establish an ergodic sampling of conformation space at the cost of not being able to modulate realistic dynamics of polymers. Mechanistic ensemble methods on the other hand aim to use only biologically plausible interactions. They start with a basic polymer model and test whether the implementation of specific mechanisms is sufficient to explain experimentally observed behavior [77]. While the classification of the different models is useful to get an idea of different approaches, it must also be noted that the transitions between the individual classes are fluid and thus clear classifications are not always possible.

One of the first pioneering data-driven studies was that of Umbarger *et al.* [183] where the three-dimensional structure of the *C. crescentus* genome was modeled with a resolution of 13 kb based on 5C (chromosome conformation capture carbon copy) data. The assumption in this study was an inverse relationship between the contact probabilities measured by the 5C data and the average distance of loci pairs. Furthermore, a calibration curve for the contact probabilites was produced with average distances obtained via FISH (fluorescence in situ hybridization) which was used in many subsequent studies. Also conducted at *C. crescentus* was the study of Le *et al.* [101] which for the first time identified chromosomal interaction domains (CIDs) for a bacterial chromosome from the analysis of Hi-C data. A general problem of many models is that the beforementioned assumption of a direct relationship between contact frequencies of loci and their average distance is somewhat problematic. In fact, a given contact frequency for a pair of loci only reflects in what fraction of cells the loci are close to each other (below a certain threshold) and does not give an average distance of the loci. Consequently, inferring a consensus structure model is rather impossible because of the highly variable ensemble of structures underlying a Hi-C map. Thus, population-based models seem to be an appropriate choice in order to reflect the cell-to-cell variability. Tjong *et al.* [181] presented a model for human lymphoblastoid cells that incorporated the stochastic nature of chromosome conformations. They generated a large population of chromosome structures in a form where the cumulated contacts of all structures recapitulate the Hi-C data using a maximum likelihood estimation. Thus, such an approach does not require a direct functional relation between contact frequencies and spatial distances. In the study by Zhang and Wolynes [218] a maximum entropy approach was developed to infer the least-structured distribution of chromosome conformations matching the Hi-C data which was used by many subsequent approaches like [28] and [126]. Other models have been developed that start with a multi-scale polymer that captures physical properties of the polymers like supercoiling and plectoneme topology. Then, these models are further developed by the incorporation of exprimental data like Hi-C maps or RNA polymerase binding data ([61], [214]).

We can differentiate the *de novo* methods for example by the simulation method. At this point, we restrict ourselves to the simulation methods also used in this work: MC simulations and MD simulations. Typically, both schemes work with coarse-grained polymer models in order to simulate effects on the cell level which comes at the price of a lower structural resolution of the models. A landmark work comes from Arnold and Jun [7] who performed MD simulations modelling the entropic segregation of overlapping polymers in confinement. They modeled the polymer with a bead-spring model in a cylindrical compartment and implemented interactions between the beads and between the beads and the compartment with a Weeks-Chandler-Andersen (WCA) potential to account for excluded-volume effects. Simultaneously, chain connectivity is ensured by harmonic bonds between the beads of a polymer. MD simulations then propagate such a system by solving the equations of motion, for example using the velocity-Verlet algorithm and a Langevin

thermostat to keep the system at constant temperature. Thereby, Arnold and Jun were able to show that the entropic repulsion between polymer chains is sufficiently strong to cause segregation of the chains, promoting the idea that entropy might be the driving force of chromosome segregation in bacteria. This work has inspired a variety of subsequent studies. Jung *et al.* [89] used a similar model to study ring polymers under cylindrical confinement. Their results indicated that the ring topology even further enhances the segregation time of the polymers. Additional works by Minina and Arnold ([130], [131]) on entropic segregation of polymers has illuminated the importance of the initial symmetry of the system for the onset of the separation process. MD simulations also enable the analysis of interactions between DNA and proteins in the cell as demonstrated for example by Pereira *et al.* [147]. In their simulations they showed that the interactions of a polymer with various DNA -binding and non-DNA -binding proteins facilitates the compression and expansion dynamics of the polymer. The effects of macromolecular crowding were also studied by Jeon *et al.* [82] who suggested that the crowding effects are not only important for the organization of DNA in the cell but also for separating the two chromosome arms in *E. coli*. While all the just mentioned MD simulations used a model very similar to the one described above of Jun and Arnold, there also exist MD simulations in which the DNA is modeled as a bottle-brush polymer. Jund and Ha [90] analyzed the helical organization of DNA. They showed that bottle-brush polymers under cylindrical confinement tend to adopt a helical pattern due to the entropic ordering of their side chains. The same result was obtained by Swain *et al.* [177] who also highlighted the importance of the cellular confinement and cytosolic crowders in the cell.

Another very frequently used method for the analysis of polymers and their spatio-temporal organization are MC simulations. An influential study by Vologodskii [191] analyzed the topological properties of the FJC model. The conformational space of such a polymer with excluded volume interactions and intersegment electrostatic interactions was sampled using a Metropolis MC procedure consisting of rotation and reptation moves. Also using the FJC model, Dorier and Stasiak [33] showed that even polymers without excluded volume form chromosomal territories. Self-avoidance was in contrast implemented in a work of Cook and Marenduzzo [29] who studied the effect of entropic forces on self-avoiding polymers under confinement. Instead of a FJC model they implemented the polymer as a string of beads of ~ 30 nm diameter that adopted a random walk. Another prominent model, the elastic filament model, was proposed by Wiggins *et al.* [211]. The idea is that the stochastic organization of a chromosome in the cell can be understood by a fluctuating elasitc filament model with intranucleoid interactions and two mechanisms of external positioning: Confinement of the chromosome by the cell and tethering of specific loci. Such targeting of specific chromosomal loci was also investigated by Junier *et al.* [88]. Their MC simulations suggested that the existence of structured microdomains in combination with the tethering of specific loci in the cell is sufficient to explain the segregation patterns in *E. coli*. Also, a recent paper by Polson and Kerry [148] investigated the segregation behavior of confined polymers by calculating the free energy functions in a MC simulation. In addition, there are a number of other papers investigating the role of topological constraints and macromolecular crowding on polymer organization ([2], [3], [16], [47], [75], [77], [87]).

2. Organization of bacterial DNA

The analyses presented in this chapter result from a project within the Transregional Collaborative Research Center 'Spatiotemporal dynamics of bacterial cells' (TRR174) which is a DFG-funded reseach center comprised of groups from the Marburg and Munich areas. One of the central reseach areas of this collaborative is the chromosome organization and segregation in bacterial cells. In this particular collaboration between the Becker lab and the Lenz lab, the spatial organization of the genetic material of the model organism *S. meliloti* was studied. The special interest here is rooted in the fact that *S. meliloti* is one of the roughly 10% of bacterial species that are multipartite, i.e. bacteria which have a main chromosome and additional plasmids over which their genetic material is distributed. The main focus of our analyses was to investigate the spatial configurations of the replicons in the bacterial cell. Thereby, we aimed to illuminate differences in the three-dimensional organization of chromosomes in bacteria with only one chromosome and those which possess many replicons. Furthermore, possible interactions between the replicons were investigated. Within this chapter first the model organism and its specific features is introduced. In addition, the experimental data basis for the theoretical investigations is explained in section 2.1. Thereafter, in section 2.2 the model of DNA and the computational framework, by which it is analyzed, are discussed. In the results section we start by presenting the results for the wild type (WT) of *S. meliloti* in 2.3.1. Thereafter, in 2.3.2 we discuss the cases of several mutants designed by the Becker lab, where either individual replicons have been eliminated or all replicons were fused into one large chromosome. The data and analyses shown here will also be the content of two papers currently in preparation, which are cited here as "manuscripts in preparation" ([134], [193]).

2.1. Model organism and experimental data

While the majority of bacterial species harbors one chromosome, roughly 10% of the bacteria a multipartite and distribute their genetic material over several replicons [185]. Until now, the origins of multipartite bacteria are not clear. Two general scenarios are discussed in the literature. The first one is the so called *schism hypothesis*. It postulates that the split of an ancestral chromosome is the reason for secondary replicons. In contrast, the *plasmid hypothesis* suggests that a megaplasmid might have been captured and subsequently aquired essential genes from the original chromosome in the course of evolution. The latter hypothesis is strengthened by the fact that one observes a bias for essential genes to be located on one chromosome. Another possibility is the simple duplication of the ancestral chromosome ([112], [133], [184]). While the emergence of a multipartite genome structure may be random, this structure has subsequently been shaped by evolutionary pressures and has led to adaptation to different niches. Several studies on multipartite bacteria reveiled that the majority of multipartite genome-harbouring bacteria are either stress tolerant or pathogens and that the secondary genome elements encode functions associated with adaptation and survival in different niches. In contrast,

the primary chromosome is typically larger and encodes more genes needed for core cellular functions. The primary chromosome also shows a greater conservation of the contents ([4], [40], [133]). The fact that bacteria with multiple replicons have been identified in diverse prokaryotic phyla suggests that they have arisen independently, many times in the course of evolution [184]. The possible advantages of dividing the genome on multiple replicons include a faster replication time and potentially more rapid growth, possible genome expansion (if some sort of maximal size limit should exist for a chromosome), functional division of genes onto separate replicons or an enhanced regulation of genes as the localization of genes on the same replicon facilitates their coordinated regulation. However, the experimental data on these points is ambiguous. Until now, no correlation between genome size and growth rate could be found for bacteria. Furthermore, there exist species with a single chromosome of 9 Mb size while many multipartite bacteria have primary chromosomes of less than 3 Mb even though multipartite genomes are on average larger than genomes of bacteria with a single chromosome ([40], [112]). At the same time, maintaining a multipartite genome structure also entails some fitness costs. Additional replicons require an increased energy for DNA replication and gene expression due to energetically expensive multiprotein transport systems (ABC systems) which are enriched in secondary replicons. There could also emerge negative interactions between pathways encoded by the chromosome and secondary replicons or between the replicons at the transcriptional level ([40], [112]).
At the same time, the question of how to solve the spatial organization of replicons in the cell arises. There are only a few studies on this question especially from Val and coworkers ([184], [185]) who studied the three-dimensional genome topology of *V. cholerae* which carries two replicons called chr1 and chr2. These studies report that both replicons are longitudinally arranged in the cell. However, they both occupy distinct locations as chr1 spreads the entire cell length while chr2 only extends from midcell to the new pole. Although the two studies provide first important results on the topology of multipartite bacteria, the main focus of the studies was on the coordination of replication of the two replicons. Therefore, a lot of questions remain to be answered. It would be interesting to know whether increased interreplicon contacts exist in multipartite bacteria and if these contacts are associated with region of special interest. Furthermore, it is unknown if the spatial organization of multiple replicons in the cell is stable and how it might change if one or more replicons are removed. To address these questions, the model organism *S. meliloti* was studied in this work. It is briefly introduced in the following.
S. meliloti is an α-proteobacterium and as such closely related to bacterial plant and animal pathogens including *Agrobacterium* and *Brucella* ([43], [50]). *S. meliloti* infects roots and induces so-called nodules, specialized organs used by bacterial endosymbionts to fix nitrogen within the plant cytoplasm. Such nitrogen fixation is very important in the environment as many plants rely on this symbiosis with bacteria to obtain nitrogen in poor soils. Thus, the understanding of such processes provides valuable information for agriculture and ecosystem management [50]. *S. meliloti* possesses a tripartite genome composed of one chromosome (3.65 mega base pair (Mbp)) and the megaplasmids pSymA (1.35 Mbp) and pSymB (1.68 Mbp). The megaplasmids belong to the RepABC family that is characterized by the combined replication and partitioning *repABC* locus. It was found that the pSymB megaplasmid in contrast to pSymA does carry essential genes. Both megaplasmids encode the majority of the proteins required for the symbiotic association with plants and thus seem to be important for niche adaptation ([43], [94], [133], [145]). The previous infromation on the spatial arrangement of replicons in the cell is mainly

based on snapshot studies of replication of origins. Here, it was found that all three oris (*oriC, oriA, oriB*) are located near the old cell pole, with *oriA* and *oriB* drifting a little bit more towards midcell. These studies also revealed a highly ordered succession in the partitioning of the free replicons with the chromosome being the first replicon to be segregated. It is followed by pSymA and finally pSymB ([43], [94]).
In the collaboration with the Becker lab, we extended these experimental results ([134], [193]). Therefore, a triple-color fluorescent labeling system was used to image predetermined loci positions in non-perturbed living cells. Thereby, the subcellular localization of loci at an approximate resolution of 60 kbp was obtained for all three replicons. The sites of the markers were designed to not disrupt genes. A schematic representation of the replicons of *S. meliloti* and the distribution of fluorescent markers is provided in figure 2.1. An interesting feature that can already be seen in figure 2.1 is the clearly asymmetric

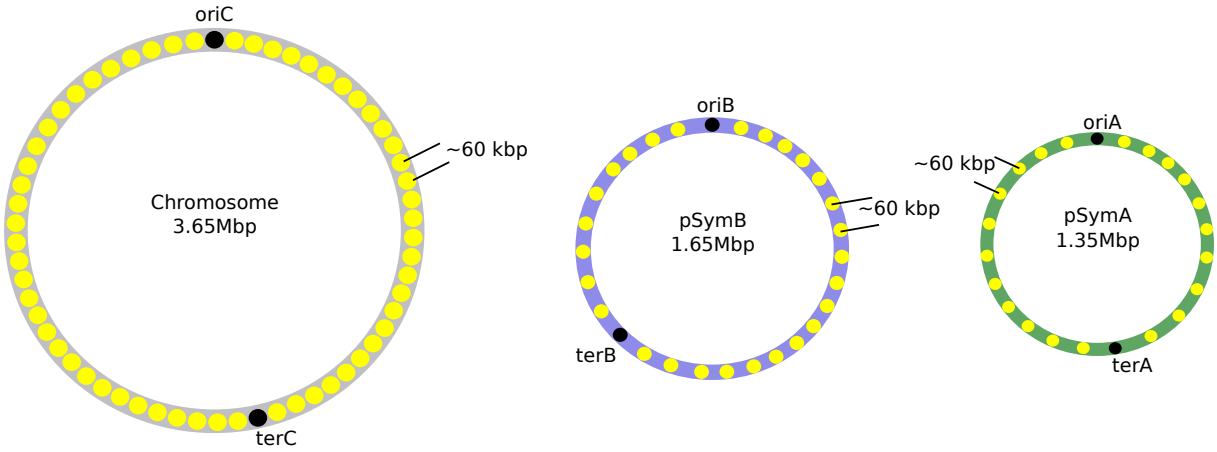

Figure 2.1.: Schematic depiction of the replicons in *S. meliloti*. The chromosome has a size of 3.65 Mbp, the larger megaplasmid (*pSymB*), has a size of 1.65 Mbp and the smaller megaplasmid (*pSymA*) is of size 1.35 Mbp. The experiments performed by the Becker lab use fluorescence markers (yellow dots) to identify the positions of loci from all replicons with a spacing of 60 kbp.

positioning of *terB* on pSymB. This provides an opportunity to examine whether this asymmetry is also reflected in the global organization of the replicon in the cell. To generate an ensemble of cells at similar stage of the cell cycle, the position of the chromosomal origin were controlled and a size limit was implemented discarding cells longer than $2\mu m$ from the data. With this, it was ensured that cells did not recide in the replication or segregation stage. The resulting experimental results were compared with predictions of the physical model presented in the next section.

2.2. DNA model and simulation framework

In the following the theoretical model that was used to make predictions about the global orientation of the three replicons in *S. meliloti* is described. There are various requirements to be met. First, for *S. meliloti* no Hi-C data is available, yet. Therefore, one cannot use a data-driven model but a *de novo* approach was required. For this, we had to use a coarse-grained model that is able to capture the global organization of three replicons in the cell while being fine-grained enough to allow a meaningful comparison with the experimental data in the form described above. Furthermore, the experimental data

provided averaged positions for individual genes in the cell. Therefore, the model should also provide an ensemble of cell configurations that produces similar mean values while reflecting cell-to-cell variability.
To meet the above requirements, a model in which DNA is represented as a SAW on a three-dimensional lattice was used. The model is an extension of the model presented in [24]. In this study, the strong linear correlation between the position of genes on the chromosomal map and their spatial position within the cell of *C. crescentus* was explained successfully. Here, this model had to be extended to represent the organization of multiple replicons in a cell. Bacterial DNA was modeled as a semi-flexible polymer of compacted units which form the effective monomers (called beads). We can imagine that the beads are the result of the action of compaction proteins and supercoiling. From section 1.2.2 we know that supercoiled domains contain roughly $l \approx 10$kbp of DNA (we call l the loopsize of a bead). We estimated the spatial extension of a supercoiled domain using the radius of gyration given in equation 1.8. Thereby, we assumed that the supercoiled domain occupies a sphere with radius $r_b = R_g$. For the calculation we used $b = 0.34$nm for the length of a base pair and N in equation 1.8 is given by l, i.e. by the amount of DNA in a bead (measured in base pairs). Thus, by assuming $l = 10$kbp we obtained

$$R_g \equiv r_b = \frac{\sqrt{lb}}{\sqrt{6}} \approx 14nm \quad . \tag{2.1}$$

In our simulations we used the diameter $d_b = 2r_b$ of a bead as the basic length scale. The grid spacing was set to this value so that we could use it to represent DNA configurations as SAWs on the three-dimensional grid. A schematic depiction of the single steps of the model construction is given in figure 2.2

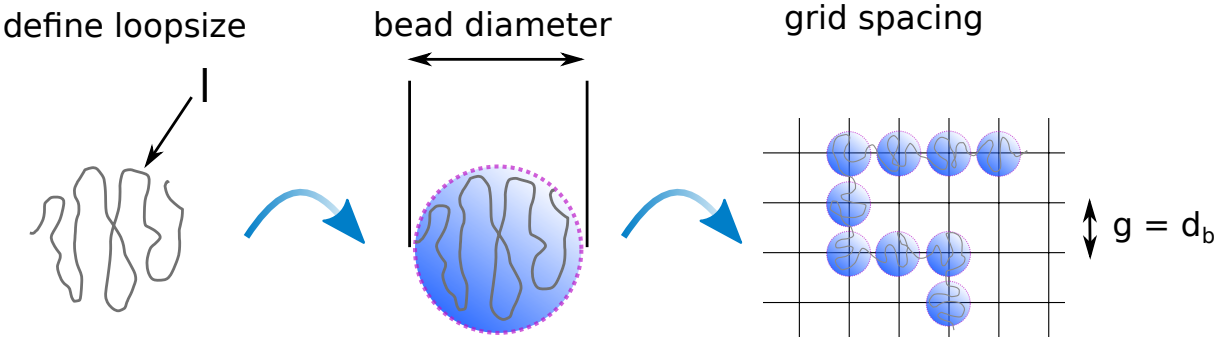

Figure 2.2.: Representation of the individual steps of model construction. First the loopsize of the DNA inside a supercoiled domain is defined. Via the radius of gyration this gives the diameter of a bead. The beads are the effective monomers of the SAW on the lattice which is shown here in two dimensions for the sake of simplicity (three-dimensional in the actual simulations).

We set the grid dimensions to correspond to a cell of the desired length H (e.g., $H = 2\mu m$). The three-dimensional cell was then usually modeled with dimensions $H \times \frac{1}{4}H \times \frac{1}{4}H$.
We already implemented two geometric constraints by the excluded-volume interactions of the beads with each other and the spatial confinement of the replicons by the cell walls. Another important constraint is the fixation of single loci in the cell. To implement this constraint, the algorithm divided each replicon into individual segments (= strands). Each strand is the connection of two loci of the plasmid, which are spatially fixed. The number of beads by which the strand is represented can be calculated using the amount of DNA

per bead l and the genetic distance $\triangle d_{loci}$ (measured in bp) between the two fixed loci of the strand

$$N_{beads}(strand) = \frac{\triangle d_{loci}}{l} \quad . \tag{2.2}$$

Consequently, the following resulted for the length, L_{strand}, of a strand

$$L_{strand} = \frac{\triangle d_{loci}}{l} \cdot d_b \quad . \tag{2.3}$$

Obviously, successive strands within a plasmid share the respective start- and end-beads. After defining the fixed loci for each plasmid on the grid, they need to be connected to form SAWs. As the number of fixpoints and replicons in the cell increases, this task becomes more complicated as the strands already established become obstacles to the strands yet to be constructed due to excluded-colume interactions. Thus, a path finding problem arises. In order to solve this reliably for all desired configurations, the A*-search algorithm was used in the simulations (see appendix B.1.2 for a description). The A* algorithm finds the shortest possible path to connect two fixed loci on the grid. Consequently, the SAWs found between two fixed loci had to be extended to the length of the respective strand in a second step. To accomplish this, a method proposed by Berg and Foester [10] was used. Here, a bead was chosen at random from the inital SAW as found by the A* algorithm and replaced by a randomly oriented hook consisting of three new beads. Thereby, a chain elongation by two was realized. If placing the hook was not possible due to occupied grid positions, a new bead was selected instead to be replaced by a hook. This procedure was repeated until the strand reached the desired length. The process is schematically shown in figure 2.3.

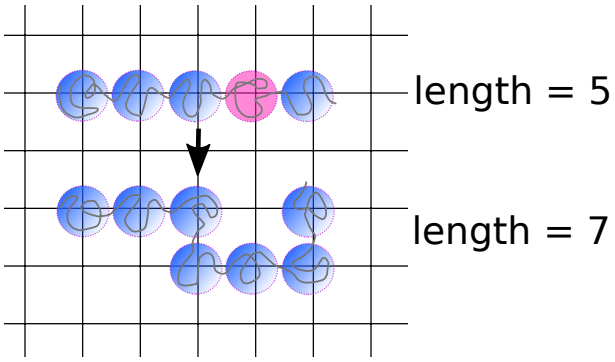

Figure 2.3.: Chain growth by hook expansion in the SAW model. The pink bead on the original SAW was randomly selected and replaced with a hook consisting of three new beads in the following. Method as proposed in [10].

For several replicons in the cell, one can encounter a problem, if the fixed loci of different strands are close to each other. In this case the complete expansion of one strand might "trap" the neighboring strand in the sense that no more grid spaces are available for the latter in order to expand further. To avoid this phenomenon, the algorithm expanded all strands simultaneously, i.e. one hook is added to each strand per iteration.
Finally, we obtained a first configuration for a replicon in the cell, consisting of variable strands between fixed points and satisfying all constraints. However, our goal was to find an ensemble of configurations. To realize this, the MOS algorithm was used in the following. The MOS algorithm is an algorithm for fixed-length, fixed-endpoints ensembles

of SAWs on the cubic lattice [115]. It consists of a set of spatial transformations that allow to sample the phase space of possible configurations of SAWs starting from an initial configuration. The algorithm is applicable for arbitrary dimensions and guarantees ergodicity. A brief description of the algorithm and the three transformations is given in the appendix B.1.1. Also, when transforming the initial configurations using the MOS algorithm, transformations were always performed simultaneously on all available replicons to ensure uniform sampling.
With this, we can summarize our program to generate an ensemble of replicon configurations in the cell:

1. Define size of supercoiled domains in bp. From this, the size and number of beads per replicon as well as the dimensions of the grid are calculated.
2. Define the loci of each replicon that should be spatially fixed. Accordingly, each replicon is divided into strands.
3. Initialization of the strands by connecting fixed loci using the A* algorithm.
4. Simultaneous expansion of initial strands to desired length by hook-expansion mechanism.
5. Sampling of the phase space using the MOS algorithm yields ensemble of replicon configurations.

Example configurations obtained with this procedure are shown in figure 2.4.

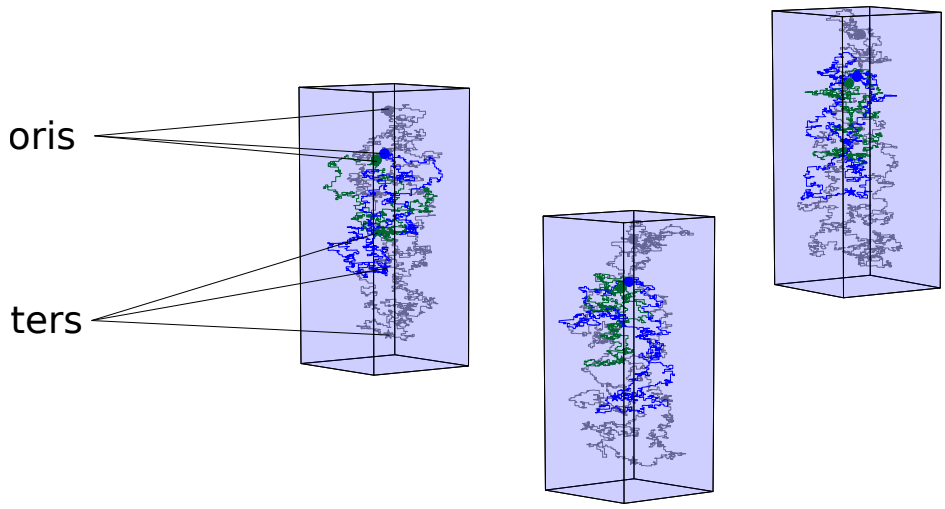

Figure 2.4.: Example configurations of the chromosome (grey) and the megaplasmids pSymA (green) and pSymB (blue) for the WT of *S. meliloti*. For these configurations, the oris (circles) and ters (stars) of the replicons were fixed at the marked positions.

2.3. Results

2.3.1. Wild type

We start the presentation of the results of our analyses on the spatial organization of DNA in *S. meliloti* with the WT results. As part of our model building, we aimed to test the following hypotheses:

1. The average spatial organization of the DNA in the cell is governed by the mechanical properties of the DNA and additional geometric constraints such as excluded-volume interactions, confinement by the cell, and fixation of several loci to specific positions in the cell.

2. In the case of multipartite bacteria interactions between replicons are expected and may affect the organization of DNA in the cell.

The results of previous studies on *C. crescentus* ([25], [190]) suggest that the oris and ters should be considered as potentially fixed loci in particular. Therefore, the experimental data for the three oris and ters were analyzed first and the coordinates were averaged to serve as input for fixated loci in the model. Experimental data were collected using the `ImageJ` plug-in `MicrobeJ` [35] and showed some error rate in assigning the correct cell poles. In addition, non-optimal synchronization of cells could be a source of additional variance. To correct this, an outlier search was performed before further processing. An example for this is shown in the appendix in figure C.1. To get a first impression of the experimental data in figure 2.5 the marker distributions for the entire replichores are shown (in blue) and the distributions for the oris (red) and ters (black) are depicted separately to better recognize them. From the heatmap data of figure 2.5 some first important conclusions could be drawn. Obviously, the chromosome stretches across the complete length of the cell while the megaplasmids are restricted to smaller subvolumes. In agreement with previous studies ([43], [94]), *oriC* is located at the cell pole while *oriA* and *oriB* are slightly subpolar. The average position of *terC* is located at the opposite cell pole while both *terA* and *terB* show significantly greater variance and appear to recide closer to the middle of the cell.

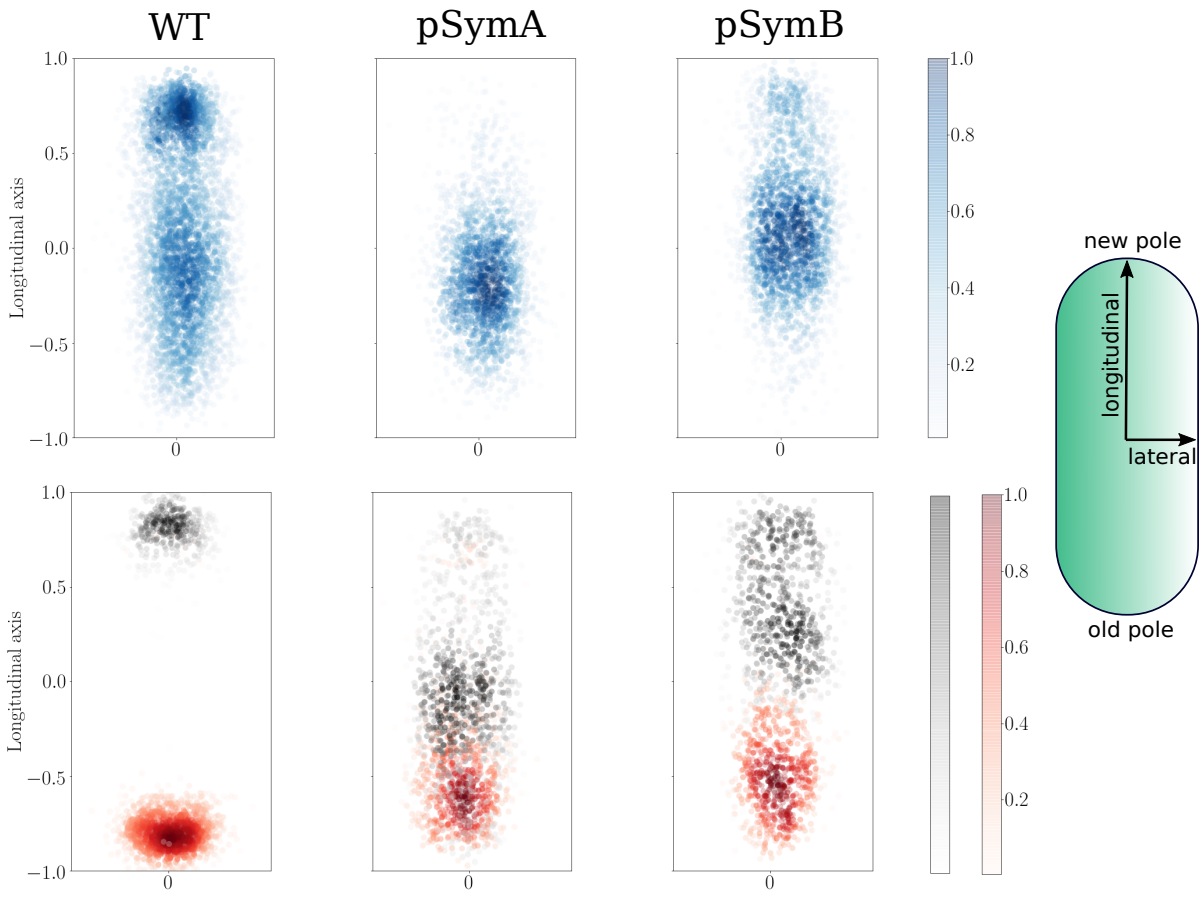

Figure 2.5.: Experimental data for complete replichores and especially for the oris and ters of *S. meliloti* WT. In blue the complete data from all markers of the respective replicons is shown in the upper heatmaps. In the heatmaps below, the distributions of the respective oris (red) and ters (black) are highlighted. The color gradients indicate the respective value of the PDF of a point.

To test the first hypothesis, the mean values of oris and ters were calculated from the heatmap data and initial model simulations were started in which one time only the oris were spatially fixed and once both the oris and ters were spatially fixed. The results can be seen in figure 2.6. In the plots, the position of each marker along the longitudinal axis of the cell (normalized from -1 to 1) was plotted over the relative position of the marker on the genomic map. For the genomic map of a replicon the respective ori is chosen as the reference point and the position ($0 \equiv 1$) is assigned to it. All subsequent loci are ranked according to their relative distance from ori. The experimental data are the same in the two parts whereas in figure 2.6 A only the oris were fixed in the model and in figure 2.6 both the oris and the ters of the replicons were fixed in the cell within the simulations.

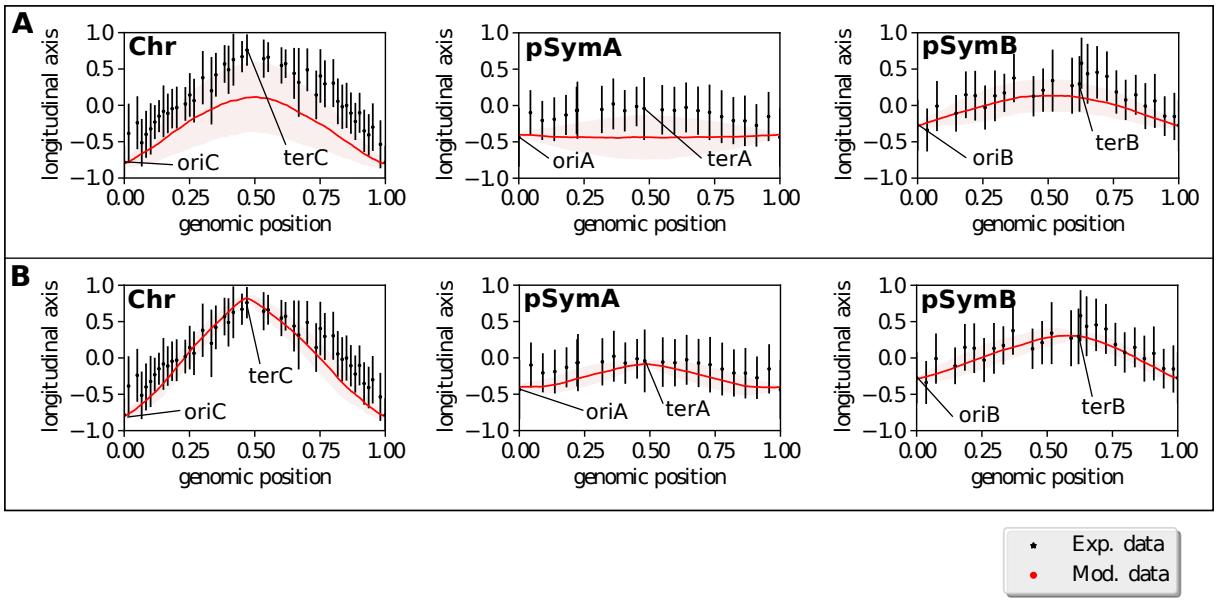

Figure 2.6.: Spatial organization of chromosome and mega plasmids in *S. meliloti* WT. Comparison of experimental data with model predicitions for fixations of oris and ters. Experimental data is shown in black and model predictions are shown in red. The models standard deviations are shown as shaded areas. (A): Model prediction for fixation of oris. Here, the ters are not spatially confined.(B): Model prediction with fixation of all oris and ters. In both, A and B, the three subplots each show the organization of one of the three replicons in the same cell. The chromosome is shown on the far left, followed by pSymA and pSymB.

We can already extract important information from these initial results. From part A we can see that the model predicted symmetric configurations when we constrained only the oris. This is plausible since in this case we are dealing with circular polymers held at only one point. For an average over many configurations, a symmetric arrangement is to be expected. At the same time, we detect an asymmetry for pSymB in the experimental data. From this we concluded that at least one more constraint must affect the average configuration here. Indeed, the model results from (B), in which in addition to the oris the ters (and in particular *terB*) were spatially fixed, show that the asymmetric organization of pSymB is reproduced in the cell. However, we discover a remarkable feature in the experimental data on pSymB. It can be seen that *terB* is not the marker closest to the cell pole. Instead, this is the marker "BR17", which is only 12851.5 bp away from *terB*. At first glance, this could indicate a very sudden jump on the configuration of the plasmid for which no meaningful biological explanation exists so far. However, another possible explanation for the observation is that the BR17 marker in WT is also part of the broader terminus region of pSymB. Since the two markers are very close to each other, this is possible. Thus, BR17 would also be subject to the spatial constraints discussed for *terB*, so we included BR17 in the spatial constraints of the *terB* domain in the model below. Therefore, based on our mechanistic conception of the model, we assume the existence of an extended *terB* domain to which BR17 belongs.
We also find that the intermediate configurations of the chromosome and pSymA were also reproduced quite well by the model. We note that the chromosome has a *C. crescentus* -like organization. Here, too, the loci follow a linear pattern between *oriC* and *terC*. This

pattern was well reproduced by the model in which *oriC* and *terC* were fixed at the poles. This result makes sense and can be understood as a proof of principle in the sense that the model successfully reproduced the linear pattern of the chromosome as found in *C. crescentus*. Obviously, the basic spatial arrangement of the chromosome with fixed *oriC* and *terC* is not fundamentally altered by the existence of additional plasmids.
When looking at pSymA, the first thing that stands out is that the megaplasmid occupies a much smaller volume than the chromosome and is located in the middle of the cell instead of extending from pole to pole. At the same time, the comparison of the two model results shows that only the implementation of an *terA* fixation allowed a comparable arrangement. When *terA* was not fixed, it rather seems that pSymA deviated towards the old pole, probably as a result of the excluded-volume effects of the other replicons. This could be a first indication of inter-plasmid interactions in the cell.
With this, the model provided a clear indication that the global organization of both plasmids pSymA and pSymB is not determined by their oris alone, but that there must be at least one other spatial constraint, presumably in the vicinity of the two terminus regions. Thus, we can already record the first important results at this point. The extension of our geometric model of DNA organization was successful and the new model can reproduce key features of the DNA organization of *S. meliloti*. Moreover, based on our mechanistic understanding of the model, we can conclude that the location of plasmids in the cell must also be regulated by spatial constraints near the terminus regions. If this were not the case, the model would predict a shift of pSymA due to excluded volume effects, and the asymmetric organization of pSymB would also not be reproducible.
In order to clarify the question of how the spatial constraints for the terminus regions of the replicons are constituted, we first considered the results of figure 2.5 again. Even though the experimental data used for this were preliminary and showed quite some dispersion, nevertheless we can see some indications. Obviously, the oris and *terC* are much more spatially restricted than we recognize for *terA* and *terB.* Thus, the heatmap data suggested that it makes sense to consider the oris and *terC* as fixed in the model. In the case of *terA* and *terB*, however, this does not seem to be true. At the same time, our previous results suggested that there is some kind of spatial limitation for the regiones near *terA* and *terB* on the megaplasmids. In the following, we analyzed two possibilities of additional spatial constraints that could influence the location of plasmids.
The first possibility is that *terA* and *terB* are limited to specific "enrichment zones." This could be important to ensure a smooth replication and segregation process. It has already been described that the individual replicon partitioning events follow a strict temporal order in which the chromosome segregates first, followed by pSymA and finally pSymB. Likewise, segregation of the duplicated *terB* starts about 20min after segregation of the duplicated *terA* [43]. Such an orderly temporal sequence of segregation steps makes it seem likely that it is accompanied by a corresponding spatial organization. Therefore, the localization patterns of *terA* and *terB* could be the result of the action of different partitioning proteins or interactions with host cellular structures to ensure the temporal sequence of segregation. This does not necessarily have to happen in the form of strict fixation at a specific location, but can also be gradient guided in the form of an approximate "enrichment zone". However, because little is known about the coordinated control of multireplicon replication and segregation, we could not precisely define this zone. Instead, we made a phenomenological definition based on the experimental variances of the markers. For this, *terA* and *terB* were initially not fixed in the model and a large set of configurations was created. In a post-processing step, we then allowed only those configurations in which

terA and *terB* were within 75 % of the standard deviations of the experimental data by which we defined an enrichment zone. This approach allows a "softer" restriction of the two loci as opposed to strict fixation at one location. The resulting organization is shown in figure 2.7 (A).

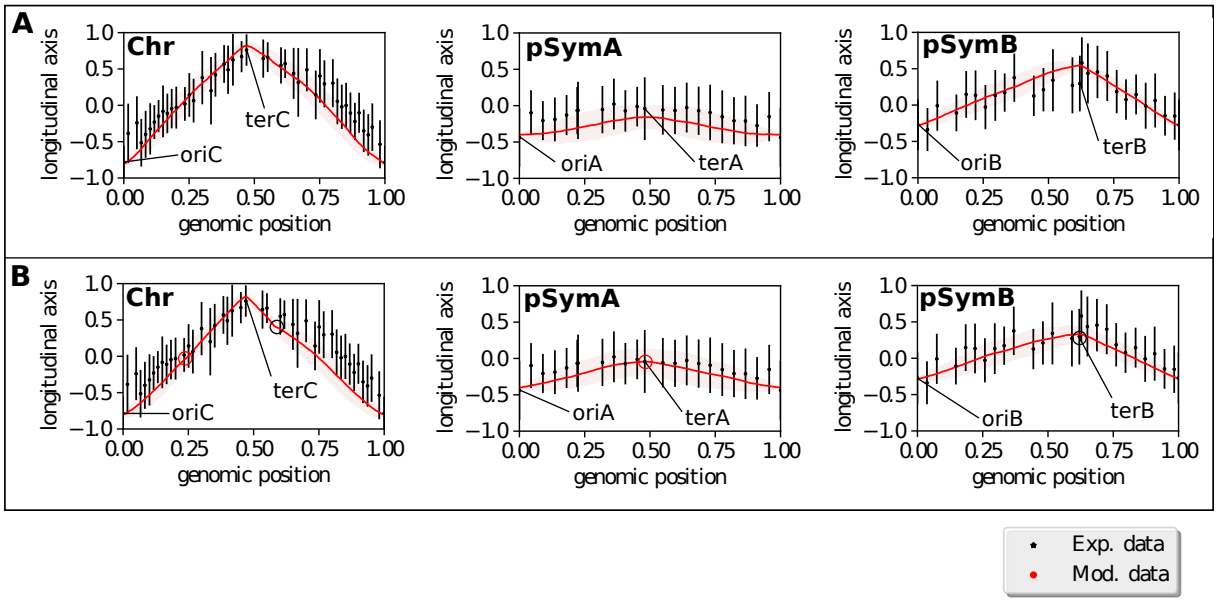

Figure 2.7.: **A:** Spatial organization of the replicons in *S. meliloti* WT due to spatial confinement of *terA* and *terB* to enrichment zones. In black the experimental data is shown while the model results are shown in red (standard deviation as shaded area). *terA* and *terB* are restricted to an enrichment zone defined as 75% of the experimental standard deviation for the markers. **B:** Spatial organization of the replicons in *S. meliloti* WT due to genomic fixation of *terA* and *terB* to the chromosome. The interacting loci are marked with corresponding circles. In both A and B the three subplots each show the organization of one of the three replicons in the same cell.

The result of figure 2.7 (A) shows a very good agreement between model and experimental data. Using the model of enrichment zones for *terA* and *terB* yielded model results which recapitulate the global arrangement of the replicons.
Besides the idea of enrichment zones for *terA* and *terB*, inter-replicon interactions represent an alternative possibility for a mechanism of spatial confinement of specific plasmid regions. Thus, one could imagine that certain regions of a plasmid - in our case, regions near the termini - are spatially coupled to chromosome regions by proteins or other interactions. In such a case, there would not be a "rigid" spatial fixation of the respective plasmid region, but it would vary with the position of the respective chromosomal anchor point. At the same time, such a mechanism could perform important regulatory functions in the cell, e.g. the temporal organization of segregation. Thereby, it could be part of some kind of checkpointing mechanisms as reported in *V. cholerae*, where replication of a locus positioned on chromosome one initiates replication of chromosome two [185]. Such "genomic fixation" of plasmid regions to the chromosome can also be implemented in the model. In a first step, the respective plasmid regions were not fixed and a large number of possible configurations were simulated. In a second post-processing step, the constraint of genomic fixation was implemented in such a way that only those configurations were

allowed in which the plasmid region and the chromosomal anchor point lay within a spatial threshold (e.g. 300 nm) of each other. This drastically restricted the original ensemble of configurations and provided a model prediction for the case of a chromosome-plasmid interaction. This procedure could then be extended iteratively to implement further inter-replicon contacts. Since we already established that for both pSymA and pSymB at least one additional spatial constraint is required to reproduce the experimental data, each of the plasmids was assumed to interact with the chromosome in the vicinity of the respective terminus. The positions on the chromosome that are linked to the plasmids remained to be determined. We assumed that these will be genes that are already close to the positions of *terA* and *terB*. In figure 2.7 (B) *terB* was genomically fixed to a gene on the chromosome with a relative position on the chromosomal map of 0.59 and *terA* was "genomically fixed" to a gene on the chromosome with a relative position on the chromosomal map of 0.25. In figure 2.7 (B) the interconnected loci are marked by corresponding circles.
Evaluation of the resulting spatial arrangements of replicons shows that we already achieved a very good fit to the experimental data. In contrast to the model results for a spatial confinement, it can be seen that the configuration of the chromosome also changes slightly due to the interaction with the plasmids. Especially at the interaction side of the chromosome with pSymB an asymmetry is induced. Such an asymmetry could also be suspected in the experimental data as mentioned above. However, it should be noted at this point that we do not find this observation in the experimental data of the knock-out mutant with deletion of pSymA ($\triangle$ pSymA) (discussed below). Since both the chromosome and pSymB are unchanged there, one would assume the same for their interaction and the resulting asymmetry in the average chromosome configuration. Therefore, we could not clearly determine at this point whether the asymmetry found in WT is truly significant and, if so, whether it results from an interaction with pSymB.
To illustrate the interaction between replicons clearly once again, in figure 2.8 the two-point-correlation matrix from the simulation data of genomic fixation is shown for the whole genome. The matrix shows the two-point correlation coefficients of each genes position along the long axis in the cell with every other gene's position. To display the matrix, a resolution of 10 kbp was chosen. I.e. in the model the positions of loci within a 10 kbp interval were averaged and displayed as a single pixel in the matrix. To illustrate the whole genome in one matrix, the genes located on pSymA and pSymB were just added at the end of the chromosomal genes on the axis. Thus, the chromosomal genes range from 0Mbp-3.65Mbp, pSymA genes follow from 3.65Mbp – 5Mbp and pSymB genes are shown from 5Mbp – 6.65Mbp. In the two-point correlation matrix of figure 2.8 we not only see interactions within genes located on the same plasmid but also between plasmids. Correlation coefficients greater than zero imply positive relationship between the positions of the genes which we can interpret as attraction of genes. On the other hand, correlation coefficients smaller than zero imply a negative interaction, which we interpreted as competition for space in the cell.

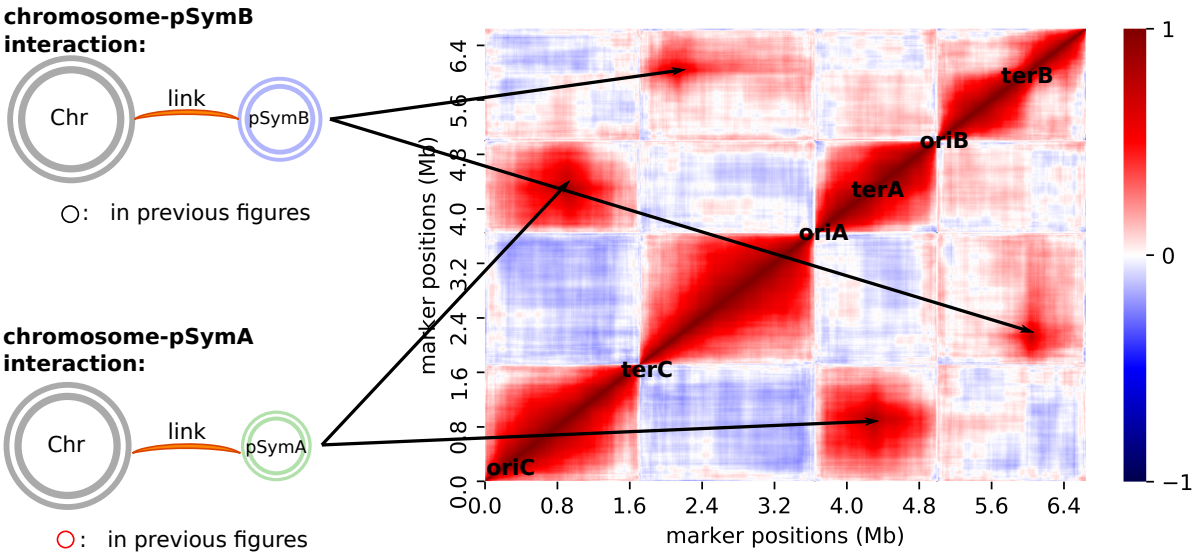

Figure 2.8.: Schematic depiction of the interaction between the plasmids and two-point correlation matrix for the complete genome. The matrix shows the two-point correlation coefficients of each genes position along the long axis in the cell with every other gene's position.

The main diagonal in figure 2.8 represents the expected positive correlations between genes with small distances. We can identify our induced interactions between the plasmids in the correlation matrix as dark red spots next to the main diagonal. We also find that the correlations between the two chromosome arms are negative. These anticorrelations might result from spatial exclusion between the two chromosomal arms. Theoretically, the model could be used to study a wide variety of interactions between individual replicons. The interactions chosen here are those that resulted in a good fit of the averaged model configuration to the experimental data. In the future, Hi-C experiments are planned for *S. meliloti* so that we can use the experimental Hi-C maps to analyze contact frequencies between the replicons and compare them with the model prediction.
In summary, after studying the WT strain, we can say that the model reflects the global organization of replicons in *S. meliloti* very well. Discrepancies between model and experiment therefore concern individual points rather than the general position of the replicons. A striking feature of the data for the chromosome and pSymA is that both oris show a shift in the direction of the cell pole compared to the general pattern of the other loci. To date, no biological interpretation exists for such a jump. However, since the oris are the basic input for the model, their shift causes a systemic deviation of the model from the remaining data. We will address this point again when we discuss the results for the knock-out mutant in the next section.

2.3.2. Mutants

In addition to the studies on the WT strain of *S. meliloti*, we investigated how the deletion of a megaplasmid affects the organization of the remaining replicons in the cell. The experiments on this have shown that deletion of pSymB leads to cell death. However, it is possible to cure cell lines of pSymA and create a △ pSymA strain. One can also introduce fluorescence marker to determine the spatial organization of the chromosome and pSymB in this strain. At the same time, we simulated such a mutant in the model. Here we assumed that spatial constraints or inter-plasmid interactions that we established in WT between the chromosome and pSymB are conserved. In figure 2.9 the experimental data are compared with the model predictions for the assumption of a *terB* enrichment zone (A) and for the assumption of an interaction between pSymB and the chromosome (B).

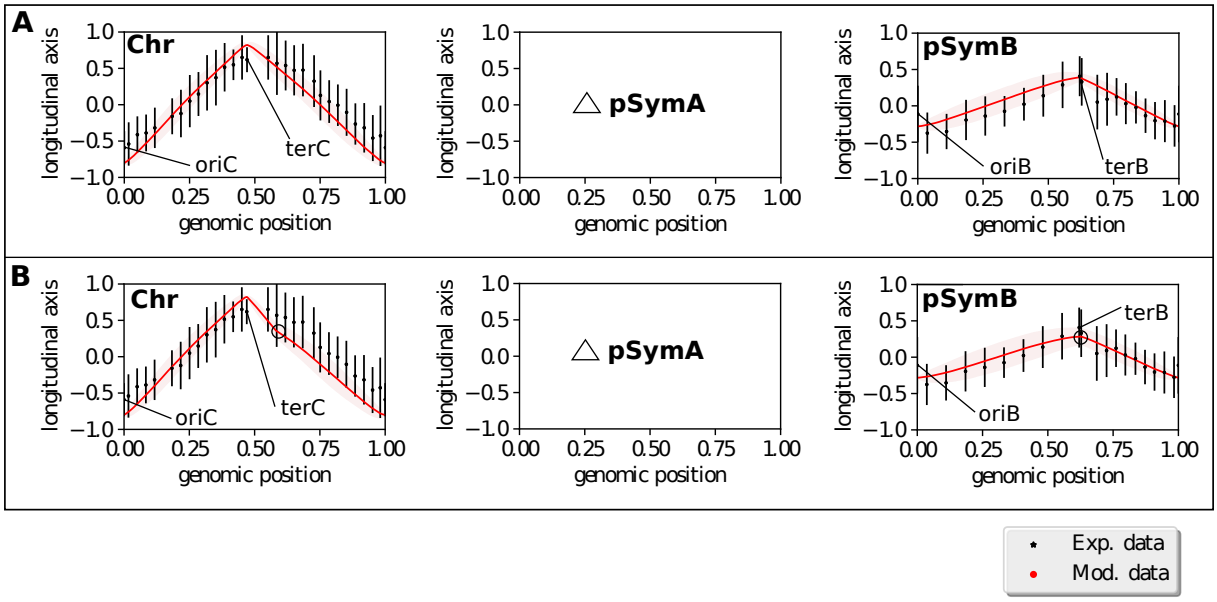

Figure 2.9.: **A:** Model prediction of the spatial organization of the replicons in the knock-out mutant △ pSymA due to spatial confinement of the *terB* region. In black the experimental data is shown while the model results are shown in red (standard deviation as shaded area). **B:** Model prediction due to genomic fixation of the *terB* region to the chromosome. The interaction was implemented between the same loci that were used in the WT strain (interacting loci marked by the circles). The three subplots in A and B each show the organization of one of the three replicons in the same cell.

The results shown in figure 2.9 do not indicate a significant change in the global organization of the chromosome or pSymB. One still finds a linear organization of chromosomal markers between *oriC* and *terC* . The asymmetric configuration of pSymB is also preserved. The results of the model prediction using the WT coordinates provided a reasonable fit to the experimental data. However, to get a more accurate idea of what effect the deletion of pSymA has on the configuration of the other two replicons, it helps to compare the WT data directly with that of △ pSymA. This is done in figure 2.10. There we see in (A) the comparison of the experimental data for WT and △ pSymA. It can be seen that the location of the chromosomal loci is almost unchanged. The only eye-catching difference is the location of *oriC*. Here, the above mentioned circumstance that *oriC* exhibits an unnatural poleward shift in the WT data becomes particularly clear. Since at the same

time all other loci seem to agree in their position, the conjecture that the position of *oriC* measured in △ pSymA is the more realistic one is confirmed. Looking at pSymB, on the other hand, shows a clear difference between WT and △ pSymA. Although the basic asymmetric configuration is preserved, a poleward offset of the entire plasmid can be seen compared to WT. The only loci deviating from this observation is *oriB*, which is even higher than in WT. However, it seems very questionable that only the ori deviates from the complete rest of the plasmid configuration, so it seems more reasonable to assume that *oriB* actually also shows a poleward offset with respect to the WT position like the other loci do. Though, this will have to be further investigated in subsequent measurements and can only be assumed here as a working hypothesis. Figures C.2 and C.3 in the appendix show results that make this assumption.

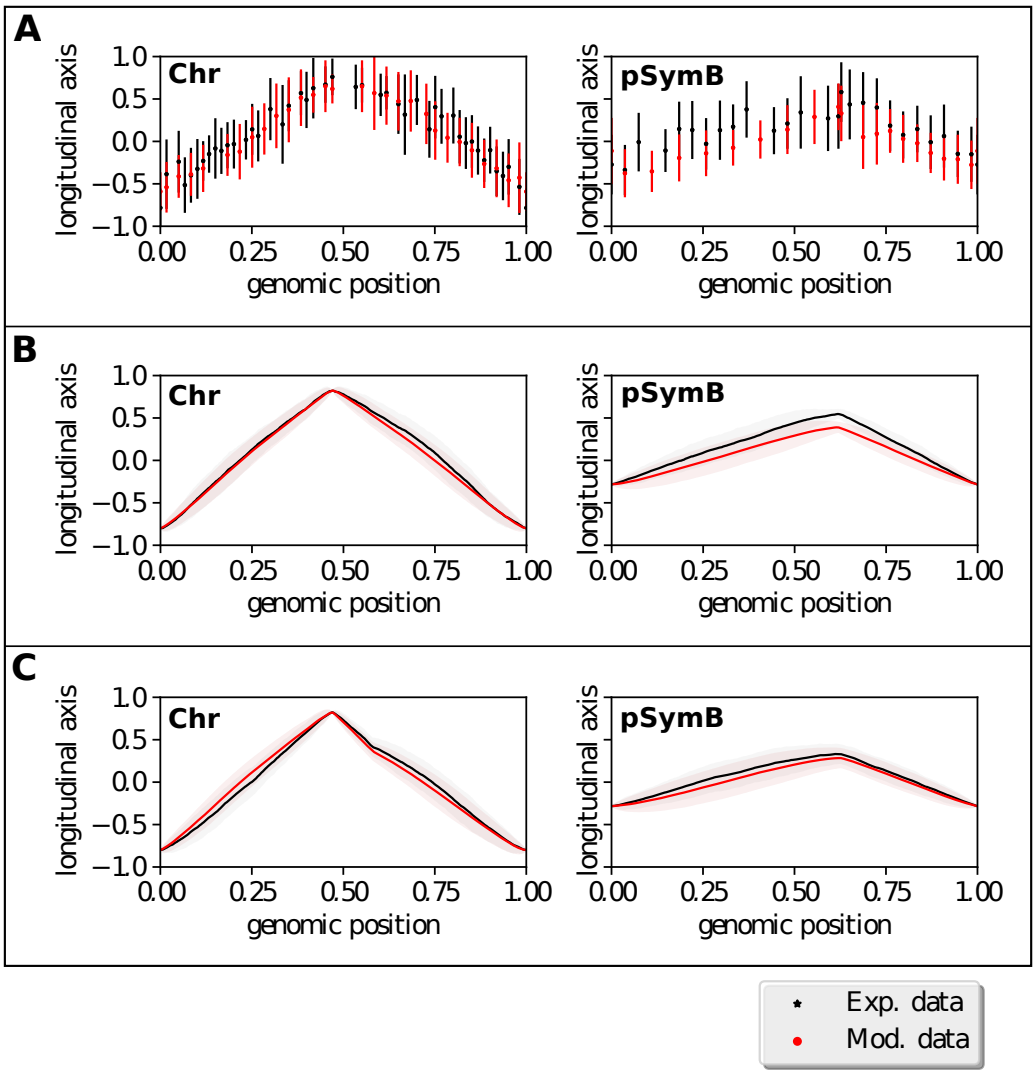

Figure 2.10.: **A:** Comparison of experimental data in WT and in the knock-out mutant △ pSymA In black the experimental data for WT is shown while the △ pSymA results are shown in red. **B:** Model prediction due to spatial confinement of the *terB* region in form of an enrichment zone. **C:** Model prediction due to genomic fixation of the *terB* region to the chromosome. The interaction was implemented between the same loci that were used in the WT strain.

In figure 2.10 (B) the model results assuming a *terB* enrichment zone are shown and compared to the model results assuming an interaction between pSymB and the chromosome as discussed in the WT section shown in (C). It can be seen that in both approaches the poleward shift of pSymB was predicted, which is confirmed in the experimental data of (A). The shift is more pronounced in the prediction based on the *terB* enrichment zone. Thus, at this point, the model succeeds in making an important prediction for a first mutant. The obvious assumption is that the deletion of pSymA frees up spatial volume in the subpolar region of the cell that is partially occupied by pSymB. This leads to a poleward shift in the configuration of pSymB. The fact that this effect is not visible to the chromosome reinforces the assumption that *terC* is fixed at the opposite pole. In the next part, we subjected these hypotheses to a final test in the form of another mutant.

In figure 2.11, the construction of a fused strain for *S. meliloti* is shown schematically. In a first step, the two megaplasmids pSymA and pSymB were fused to form the new fusion strain of pSymA and pSymB (SmAB). Thereafter, this strain was combined with the chromosome. With this, the fused strain of the *S. meliloti* replicons (SmABC) was obtained. The details of the creation of these strains are beyond the scope of this work and can be found in [193]. For our purposes, the most important thing to know is that the resulting SmABC exhibited an entirely merged genome. Furthermore, it should be noted that for design reasons, two separate regions of the original pSymB plasmid are present in the fused strain (see blue sections in figure 2.11). To study the organization of the genome in the cell for this new mutant, fluorescent markers (yellow dots in figure 2.11) were again attached to the chromosome. We tried to place markers especially at transitions between the different replicons. Thereby, the region split off from pSymB (near 10 o'clock) is of particular interest.

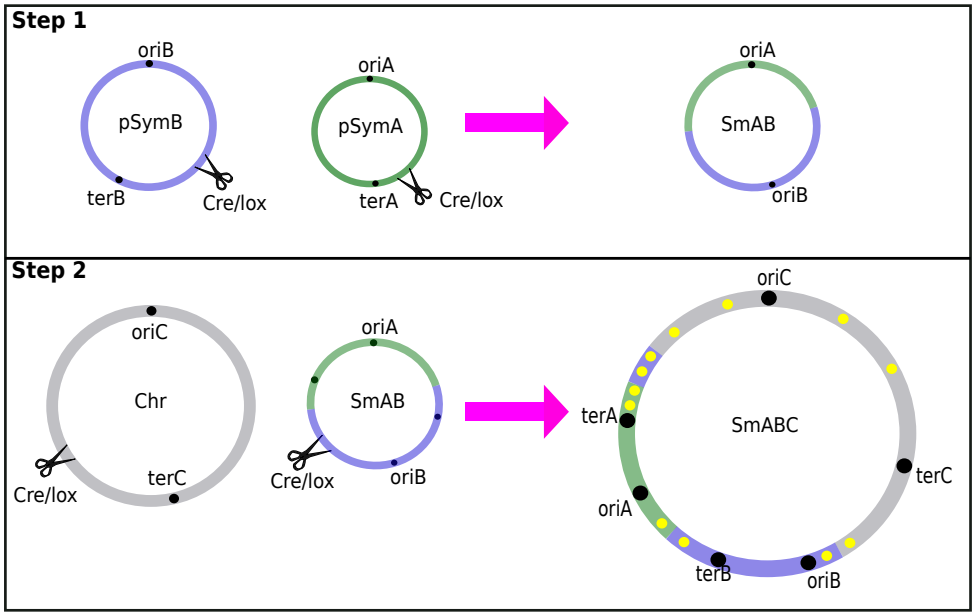

Figure 2.11.: Schematic representation of the constuction of the fused SmABC strain for *S. meliloti*. In a first step (top), the two megaplasmids were fused to form the new SmAB strain using cre/lox recombination technology [193]. Then, in a second step (below), the chromosome was fused with SmAB to form the final SmABC strain. To measure the organization in the cell, fluorescence marker (yellow dots) were placed in the SmABC strain.

The new SmABC strain now provides yet another opportunity to test our understanding of genome organization in *S. meliloti*. In contrast to the WT, we are de facto dealing with a single large "chromosome". However, this new chromosome has three ori - and three *ter*-regions. Thus, we were now able to test whether the configuration of the new chromosome is determined solely by the positions of *oriC* and *terC* , so that we would again expect a *C. crescentus*-like organization. Another possibility would be that not only *oriC* but all oris remain spatially fixed as in WT. In this case, the global organization should deviate significantly from the linear pattern of *C. crescentus*. This seems logical if we assume that the mechanisms leading to the positioning of oris in WT are also active in the new SmABC strain. Consequently, we checked a third possibility, in which we assumed that also the two *ter*-regions of the megaplasmids in SmABC remain close to their WT positions. In this case, we expected to see a serrated pattern in the plot of longitudinal positions over the genomic map, with the chromosome extending in a serrated fashion between the oris and ters. In figure 2.12 the model predictions for all three possibilites are shown. The background colors indicate from which plasmid the respective section of the fused chromosome originates.

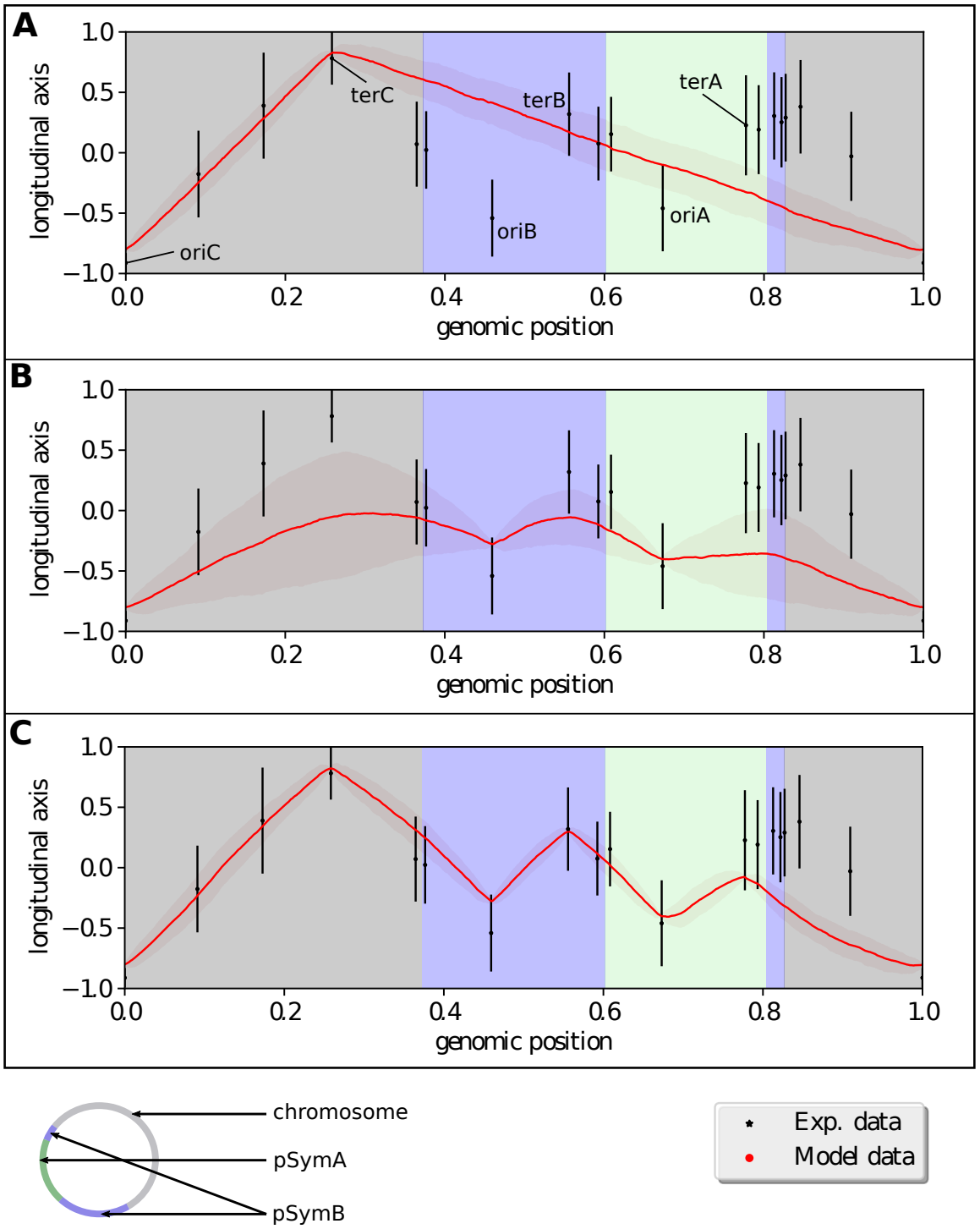

Figure 2.12.: Spatial organization of the fused SmABC chromosome in the cell. Experimental data is compared to model predictions for **A:** fixation of *oriC* and *terC*, **B:** fixation of *oriC*, *oriB*, *oriA* and **C:** fixation of all oris and ters. For the predictions the WT positions of oris and ters were assumed. The background colors indicate from which plasmid the respective part of the fused chromosome originates. The colour code is explained below the plots.

In figure 2.12 (A) the expected genome organization for SmABC for the assumption that only *oriC* and *terC* are fixed according to their WT positions is shown. Although we find that *oriC* and *terC* are quite close to the WT positions, the mean configuration of the rest of the genome deviates significantly from the corresponding prediction. Apparently, the organization of the genome of *S. meliloti* is thus also determined by the fixation of loci of megaplasmids, as already suspected. In figure 2.12 (B) we can see that very likely both oris of the megaplasmids, *oriB* and *oriA*, are spatially fixed as well, although specifically for *oriB* we observe a slight shift compared to the WT position. Assuming a fixation of both the oris and ters of all replicons results in a further improvement of the fit to the experimental data as seen in figure 2.12 (C). However, we make an important observation at this point. The position of *terA* in SmABC is not the WT position. Rather, *terA* is approximately at the level of *terB* . Furthermore, we see that the markers following *terA* are also at this level, deviating from the prediction based on *terB* fixation. From the background colors we can see that these deviations appear exactly at the position of the fused chromosome where the smaller part of pSymB lies as an inlay between the pSymA part and the chromosomal part. In the WT, the genomic distance of the inlay region of *terB* is 31 kbp. Therefore, we can hypothesize that the inlay region belongs to the larger *terB* domain that we postulated earlier. This leads to the conclusion that the inlay region could also be part of the postulated *terB* enrichment zone. In figure 2.13, we assumed this and fixed oris and *terC* in the model (using the now available SmABC coordinates), while *terB* and the inlay region were confined in an enrichment zone.

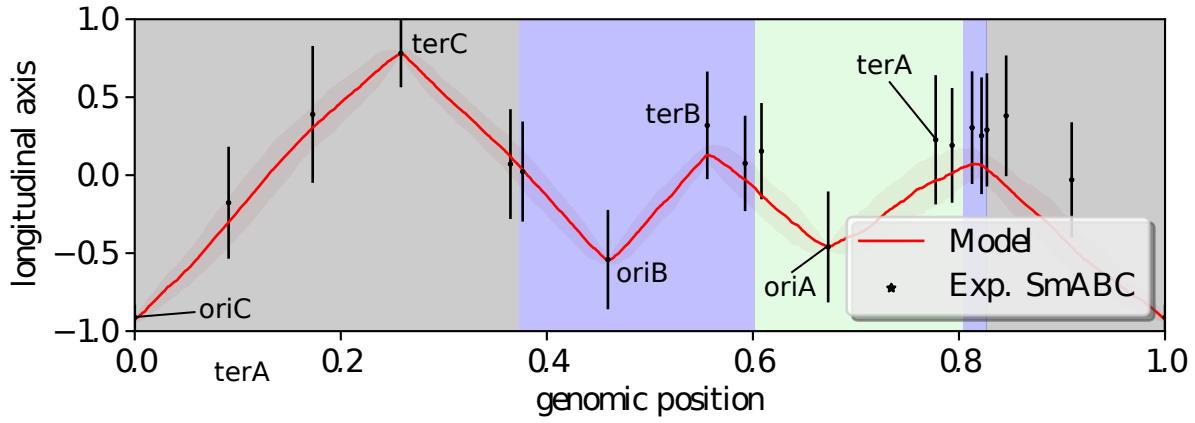

Figure 2.13.: Model prediction of the spatial organization of SmABC under the premise of a terB enrichment zone.

From figure 2.13, it can be seen that the *terB* enrichment zone approach accurately reflects the mean configuration of the SmABC strain. This is particularly compelling since this approach also successfully reproduces the organization in WT as well as for the knock-out version △ pSymA. Furthermore, we observe that the positioning of the *terB* region in the merged strain takes priority over the positioning of *terA*. This could be due to the fact that essential genes are located on pSymB in contrast to pSymA. In the future, additional designs for similar fused strains provide the opportunity to further investigate the genomic organization of *S. meliloti*.

2.4. Project summary and outlook

The goal of this project was to investigate the spatial organization of genetic material in *S. meliloti*. The complex order of genetic material in bacteria has only in recent years been the subject of intensive research. Previously, it was assumed for a long period that DNA is randomly distributed in the cell ([9], [173]). By analyzing the configurations in *S. meliloti* we turned to the special case of multipartite bacteria. Despite the fact that roughly 10% of all bacterial species harbour multipartite genomes, the organization of replicons in multipartite bacteria has been sparsely considered. Here, especially the works of Val *et al.* ([184], [185]) should be mentioned, who reported that the two replicons of *V. cholerae* are longitudinally arranged in the cell.
However, to date it is not known which advantages have led to the formation of multipartite genomes, nor how they are organized in the cell. It is unclear whether a controlled spatial organization of the replicons within multipartite bacteria exists. Furthermore, very little is known about possible interactions between the replicons of *S. meliloti*. If such interactions exist it would be interesting to know if these interactions of the replicons are associated with regions of special interest. In our collaboration with the Becker lab we aimed to analyze the three-dimensional organization of the three replicons in *S. meliloti* by approaching from two sides. On the experimental side fluorescence marker were inserted into the genome to measure the positions of the tagged loci in an ensemble of cells. At the same time, within the framework of this work a theoretical model was developed in order to simulate the spatial organization of multiple confined polymers due to geometric constraints. For this purpose, an existing model for the organization of *C. crescentus* [24] was extended to multiple replicons. In addition, further analysis methods have been developed to predict inter-plasmid interactions or enrichment zones.
The following hypotheses were formulated and tested within the project:

1. The averaged configuration of replicons in the cell is a consequence of the compactification of DNA as well as spatial constraints that affect the DNA's location.

2. The spatial constraints consist of restriction of DNA to the cell interior, self-avoidance of DNA, and spatial restrictions of individual loci of special interest to specific positions in the cell.

3. In the special case of multipartite bacteria, interactions between replicons lead to further mutual spatial constraints that affect the organization of DNA in the cell.

Research approach To simulate the spatial organization of chromosomes in the cell, different approaches are chosen. In recent years, some data-driven studies based mainly on Hi-C data for known model organisms have been performed ([28], [61], [101], [126], [181], [183], [214], [218]). Common to these approaches is the attempt to fit a particular polymer model to produce a set of conformations that reproduces the experimental Hi-C data. Since such data are not available for *S. meliloti*, a *de novo* approach had to be adopted in the present work. For this purpose, a physical model of DNA was designed on which the hypotheses on the organization of replicons in *S. meliloti* could be tested.
In the coarse-grained model used here, DNA is considered a semi-flexible polymer of compacted beads. Thereby, it is assumed that chromosome compaction is achieved by the combined effects of supercoiling, macromolecular crowding, and DNA binding proteins. Corresponding models for describing bacterial chromosomes can be found in many similar

studies ([7], [82], [89], [90], [130], [131], [147], [177]), as the neglection of atomic details enables the study of the global configuration of chromosomes. At the same time, the small-scale effects of, for example, the binding proteins are included in the coarse-grained description of the beads and the used model is fine-grained enough to compare its results with the experimental data providing positioning of genetic loci with a resolution of 60 kbp.
To create ensembles of chromosome configurations with the chosen model, DNA was modeled as a SAW on a three-dimensional grid. Here, the lattice spacing was determined by the diameter of a bead. In *S. meliloti* one has to deal with the particular difficulty of having multiple replicons in the cell. Therefore, a version of the A* algorithm was implemented by which initial configurations for an arbitrary number of replicons could be constructed. These were brought to the desired length afterwards by simultaneous expansion of the walks. In the following, simultaneous applications of the MOS algorithm [115] allowed us to sample the configuration space for these replicons and thus it was possible to produce chromosome configurations reflecting the mean values of the experimentally determined loci positions as well as the cell-to-cell variability observed *in vivo*. At the same time, spatial constraints such as self-avoidance of DNA or fixation of individual loci within the cell could be implemented. Therefore, duplicate grid occupancies were prevented to account for self-avoidance and selected loci could be fixed on respective grid points to account for a spatial constraint. In post-processing procedures configurations were filtered to modulate inter-replicon interactions or spatial enrichment zones of specific loci.

Key findings For the investigation of the above formulated hypotheses, both the WT and two mutants were studied. Thus, assumptions made on the WT could be challenged by testing on the mutants.
In the analyses on the WT strain of *S. meliloti oriC* was found to reside near the cell pole while the *oriA* and *oriB* are found at subpolar regions, in agreement with previous studies ([43], [94], [145]). Furthermore, a linear organization of the chromosome in the cell was found, very similar to the previously described organization of the chromosome of *C. crescentus* ([24], [25], [190]). Here, modeling showed that this organization can indeed be derived from the hypotheses described above, with chromosomal *oriC* and *terC* appearing to be fixed at the two cell poles. Such "strict" fixation of selected loci was thus to be expected for the two mutants as well. Indeed, this was evident in both the knock-out mutant △ pSymA and the fused strain SmABC. The result in SmABC is particularly compelling here, as the relative position of the two loci along the replicon changed significantly from their position in the WT chromosome in the fused strain, so that the constant positioning in the cell is due to an external fixation mechanism and cannot be a consequence of the self-assembly of the replicon. The finding that the main chromosome of *S. meliloti* arranges itself similarly to that of *C. crescentus* in the cell, was expected as both organisms belong to the group of *Alphaproteobacteria*.
Another interesting finding that resulted from the analysis of the WT is that spatial confinement of oris and ters can also be assumed for the two megaplasmids pSymA and pSymB. The smaller megaplasmid, *pSymA*, was found to reside near midcell. Interestingly, the model simulations revealed that without any spatial confinement of *terB* the excluded volume effects between the replicons would cause pSymA to reside closer to the cell pole. This is a first indication of interactions between the replicons. Even more impressive is the asymmetric organization that was found for pSymB. With the help of the model it was possible to demonstrate that this organization arises from the confinement of *terB* and

is found in all mutants. Thus, modeling has provided a first important indication that genome organization in *S. meliloti* is not limited to constraining the chromosomal loci, but it also appears that the configurations of the megaplasmids are regulated.
While the evaluations for all oris of the replicons as well as for *terC* suggest a relatively strict fixation within the cell, for instance by specific anchor proteins, larger variances were found in the experimental data for the two terminus regions of the megaplasmids. Here it was possible to set up two different hypotheses with the model in order to take this into account. First, a spatial enrichment zone would be imaginable into which terminus regions within the cell are drawn, e.g., in preparation for segregation steps. This would result in a less strict spatial confinement of the corresponding regions in the cell, which could then also refer to more extensive areas of the megaplasmids. This suggestion is supported by results from the fused SmABC strain, in which pSymB derived terminus-proximal sections of the replicon spatially occupy the same region in the cell, although they are widely separated along the fused replicon. A second possibility to explain the positioning of the terminus regions of the megaplasmids are inter-plasmid interactions. Here, with the help of the model, certain regions of the individual plasmids could be suggested whose interaction with each other would explain the observed organization of the replicons in the cell. These need to be tested in future Hi-C studies. Thus, the model provides an important starting point for further investigations, especially in the form of Hi-C analysis, to test the prediction of such contacts of the replicons. Remaining questions to be discussed here concern the range of possible inter-plasmid interactions (here $\approx$ 300nm was assumed, which corresponds to the magnitude of the experimental variances) and possible mechanisms that could lead to an enrichment zone.
The analysis of the two mutants provided further interesting insights. In the knock-out mutant $\triangle$ pSymA it was found that the basic organization of the two replicons remains unchanged from WT and can be correctly predicted by the model. The key difference compared to WT is a poleward shift of *terB*, which was also predicted in the model predictions based on an *terB* enrichment zone and inter-plasmid interactions, while this cannot be reproduced by a strictly local fixation of *terB*. Thereby, the study of the mutant provides evidence that spatial confinement of *terB* is indeed a rather "soft" mechanism, as opposed to strict fixation at a well-defined point in the cell.
The second mutant that was analyzed, SmABC, results from fusion of all three replicons. Although this effectively creates a single large chromosome, the fused chromosome does not arrange itself in the cell in its averaged configuration in a *C. crescentus* -like manner, but rather has an organization shaped by the spatially restricted oris and ters. Thereby, the investigation of SmABC strengthens the assumption of a fixation of oris and *terC* and provides indications for an *terB* enrichment zone since the position of *terA* seems to be determined by the adjacent *terB* -near loci.
In summary, the picture of genome organization in *S. meliloti* is as follows: By comparing experiments and modeling, we can assume that the organization of individual replicons is determined by spatial limitation of oris and ters and the mechanic properties of the compactified DNA (hypothesis 1). After the spatial location of oris and ters is defined, presumably by the action of partitioning proteins (ParAB and *repABC*) or interactions between the plasmids and with other cellular structures, the average location of the remaining genes of a replicon results from the excluded volume effects, confinement due to the cell and chain connectivity (hypothesis 2). In addition, the model allowed further hypotheses on inter-plasmid interactions to be formulated that could explain the observed organization of DNA in the cell (hypothesis 3).

Outlook At this point, there are many opportunities for further studies. Especially Hi-C measurements provide a valuable means of identifiying interactions between the plasmids as suggested by the model. Here, we could also further speculate whether possible inter-plasmid contacts are part of a checkpoint mechanism for spatiotemporal regulation of replication and segregation, as in *V. cholerae* [185]. This seems realistic in the context of previous studies that have described the strict spatiotemporal order of replication and segregation of plasmids [43] and the enhanced need for coordinaten of replication and segregation in multipartite bacteria. Furthermore, Hi-C contact maps could be calculated from the model data in order to compare them with experimental data. Other model extensions could be, for example, the implementation of different degrees of compactification at different sites of a replicon or implementation of the effect of SMC or other structure regulating proteins. In addition, the model can be applied to a variety of other species with any number of replicons.
Chromosome-membrane interactions offer another interesting application for the model developed here. These could be caused by different mechanisms in the cell. One possibility is transcription ([107], [141], [142]). A pioneering study by Libby *et al.* [107] on two loci of *E. coli* showed that induction of membrane protein expression rapidly results in a dramatic repositioning of the loci toward the membrane. It was further noticed that the positions of loci as far away as 90 kbp from the induced gene were changed. This indicated that transertion (= coupled transcription and translation) might lead to significant changes in the DNA configuration [141]. It would be an interesting task to use our model to explore the impact of gene expression state on the spatial organization of DNA. Within our model we could simulate the effect of transertion by restricting specific loci to subcellular positions near the membrane. Thus, additional constraints would be added to the model. Expression levels of the individual loci could also be represented. For this purpose the constraints would only be applied to a certain fraction of the calculated configurations within an ensemble. Thereby, these fractions would reflect the expression levels.
Another possibility would be to turn the question around and analyze whether certain positions along the chromosome are particularly suitable for transcription. This would mean using the model to search for loci that have a particularly high probability of being close to the membrane due to the global organization of the chromosome in the cell. To analyze this, an iterative approach could be taken. This would involve starting with some known loci and fixing them near the membrane. E.g., for *B. subtilis* the chromosomal positions of some membrane proteins have already been determined [111]. In the following, the model would be used to determine further loci that are close to the membrane and these would also be fixed in a second iteration step. This procedure can be continued until no more loci are found. Subsequently, one could determine which of the loci are already known experimentally and at which point the model makes predictions for further highly expressed genes.
A second mechanism that could lead to chromosome-membrane interactions is provided by proteins binding the chromosome to the membrane. An example for this is CadC in *E. coli*. In the context of the description with our model, the observed effect would also have to be implemented by additional constraints for some selected loci close to the membrane. These could then be compared experimentally with the distribution of labeled (CadC) proteins in the cell.
Finally, with the presented MC model it might also be possible to investigate an interesting hypothesis on membrane domain formation. In this context, there is a suggestion that transertion contributes to the formation of membrane domains ([124], [141], [142]).

Membrane domains are clusters of hyperstructures, i.e., spatially extended assemblies of molecules. In the case of membrane domains the hyperstructures consist of membrane-polysome DNA complexes. Within our model it would be possible to describe such hyperstructures as loci attached to the membrane. In the following one could analyze under which conditions two hyperstructures attract each other, thereby facilitating the formation of a membrane domain of multiple hyperstructures. This would involve a procedure of sampling different chromosome configurations, varying the distance between the two hyperstructures. From the ensemble of these configurations, entropy differences can then be calculated, e.g., using the hypothetical scanning method of White and Meirovitch [209]. Finally, an interaction potential can be calculated from the entropy differences. This interaction potential, influencing the position of the complexes in the membrane, would be mediated by the spatial configuration of the chromosome in the cell. Another interesting investigation of the results of the model concerns the quantitative analysis of the topological properties of the replicons of *S. meliloti*. From these, one could determine the extent to which the replicons mix and interwine. This could be done for example by calculation of the linking number with a procedure like the one proposed by Fourey *et al.* [48]. In this case, the degree of entanglement of the replicons could be determined as a function of the length of the replicons and the cell size. The assumption would be that a high degree of intermixing of replicons is a hindrance to their segregation. In this respect, quantitative information on the degree of mixing would be interesting to gain an estimate of how much multipartite bacteria are affected by the mixing of their replicons during segregation.

3. Segregation of DNA in bacteria

Having investigated the issue of spatial organization of DNA in bacteria in the previous project, we will address the dynamic processes of replication and segregation in the following two projects. First, results from another TRR174 collaboration with the Graumann lab will be presented. In this work, the segregation of oris in the model organism *B. subtilis* was studied. The background to the investigations is that no uniform mechanism for chromosome segregation has yet been discovered in bacteria, although such a mechanism exists in eukaryotes with the mitotic machinery ([9], [37], [55]). To determine whether chromosome segregation in *B. subtilis* follows recognizable patterns, the Graumann lab tracked the two oris at 10-s intervals during segregation. In parallel, a theoretical model of entropic chromosome segregation was developed according to the proposals of Arnold and Jun [84]. With this, MD simulations were performed. Thus, the experimental trajectories of the oris could be compared to the simulation results to evaluate whether an entropy-driven segregation mechanism is sufficient to ensure reliable transfer of genetic material to the daughter cell. The results of the project have been published and can be found at [37]. In the following, a short introduction of the model organism and the experimental data is given 3.1. Then the model for the MD simulations is presented in 3.2. Thereafter, the results of the work are presented and discussed in 3.3. Finally, an outlook on possible extensions of the model and open questions will be given in 3.4, some of which will be investigated in the last project.

3.1. Model organism and experimental data

The model organism of this study is the well known *B. subtilis*. It is a gram-positive rod-shaped bacterium. *B. subtilis* is found in soil and in the gastrointestinal tract of humans. It is known to tolerate extreme environmental conditions and while *B. subtilis* was long believed to be a strict aerobe, it has been shown that it can also grow anaerobically. *B. subtilis* is not only one of the most frequently used model organisms in research but it is also used in industrial applications for enzyme production ([55], [113], [137]). The chromosome of *B. subtilis* has a size of 4.2 Mbp. Concerning the orientation of the chromosome in the cell it was shown that the chromosome of *B. subtilis* alternates between the *ori-ter*-pattern and the left-*ori*-right-pattern. In sporulating *B. subtilis*, the chromosome adopts an *ori-ter* configuration while during vegetative growth, the chromosome alternates between the two patterns. The reason for this behaviour is still unclear. It was shown that the chromosome of *B. subtilis* is divided into supercoiled domains that are between 24 and 400 kbp ([55], [173], [197], [199]).
The replication period in *B. subtilis* was found to be fairly constant at 37°C with a duration of about 55min [37]. For *B. subtilis* a replication machinery positioned at midcell was proposed ([55], [103]). Later on, further experiments revealed that while the replication factory is relatively stationary positioned at midcell, the replication forks can still move within the cell center and show great mobility there. The chromosome is pulled through this replication factory at midcell while being replicated. The origin regions

start moving towards the cell poles soon after initiation of replication and are followed by the remaining duplicated chromosome regions ([55], [104], [129], [167]). Time-lapse fluorescence microscopy experiments revealed that the average velocity of the segregating origins is 0.17μm/min and thereby faster than the rate by which the cell increases of 0.011 to 0.025μm/min [203]. This movement of oris is even observed in the absence of ParAB. At the same time the separation of duplicated chromosomes in *B. subtilis* is very robust. Only 1 in 10,000 cell cycle events shows a failure in chromosome segregation [55]. This circumstance makes it particularly exciting to find out how such a low error rate can be maintained even though no sole segregation mechanism could be observed. As seen above, the oris segregate faster than the cell grows. Therefore, a tethering mechanism to the cell wall can be excluded as a possible mechanism. In the same way it was shown that deletion of ParA and ParB in *B. subtilis* causes very mild segregation defects and that the ParAB system consisting of *soj* and *spo0J* in *B. subtilis* is only essential during sporulation and not during vegetative growth ([37], [55], [78]). Another protein that has been suggested as an important component of the segregation mechanism is SMC. It was shown that the absence of SMC in *B. subtilis* leads to about 15% anucleate cells as well as to the loss of spatial arrangement of the chromosome. During fast growth, the origins cannot be segregated properly in the absence of SMC. However, slowing down the velocity of the replication forks can bypass the requirement for SMC for the separation of origins in rich medium while under slow growth conditions (22°C) origin regions are still segregated normally in absence of SMC. ([17], [55], [59], [198]).

There are quite a few other candidates for possible segregation mechanisms in bacteria. Among others it was also suggested that RNA polymerase could help segregating the chromosomes as it is a powerful molecular motor and it was shown that inhibition of RNA polymerase inhibits separation of newly duplicated DNAs near the origin of replication. Likewise, DNA polymerase could help segregate DNA in *B. subtilis*, pushing newly duplicated DNA from midcell towards the poles. However, since the persistence length of DNA is about 150bp, DNA cannot be pushed from the cell center to the poles without bending ([36], [55], [95]). Another idea is that transcription in concert with translation and insertion of membrane proteins could anchor the chromosome to the membrane and thereby facilitate chromosome segregation in a process termed transertion. This was supported by experiments showing that mebrane protein expression can affect the positioning of chromosomal loci and simulations proving the enhanced efficiency of chromosome segregation by membrane tethering of DNA for self-avoiding chromosomes ([32], [55], [95], [107], [212]). In addition to the multitude of possible mechanisms for chromosome segregation, another promising suggestion is that entropic repulsion of chromosomes is an important driving force [84]. The aim of the present work was to apply this concept to *B. subtilis* and to compare corresponding computer simulations with experimental data. First, we briefly describe the experimental design before moving on to the simulation model.

In the experiments fluorescence microscopy was used to track the dynamics of the origins of replication in *B. subtilis* for an entire cell cycle in 10-s intervals [37]. Therefore, fluorescent repressor/operator systems (FROS) as well as ParB/*parS* systems have been used. The FROS system has been used successfully in previous studies. However, it should be noted that the FROS arrays have been shown to hinder the progression of replication forks in some circumstances [202]. Keeping this in mind, in the here performed experiments the cells did not show pronounced negative effects due to the FROS system so that the measurements could be continued. Furthermore, the cells underwent a transition to adopt

to low-oxygen conditions. The final patterns observed were measured in cells that have grown exponentially under imaging conditions. Therefore, it can be assumed that the cells adopted to the growth conditions.
Observation of the origin dynamics revealed that the two oris remained at midcell for an initial period before they moved towards the quarter-cell positions. Thereafter, the oris remained at the new positions. Interestingly, it was not possible to describe a unified pattern of origin movement. Instead, the 80 segregation events that were tracked showed a rather random movement of the oris that nontheless lead to an overall separation.

3.2. Molecular Dynamics simulations of chromosome segregation

A widely used method to study the dynamics of biomolecules such as DNA in simulations are MD simulations. The basic idea behind MD simulations is to evaluate the time-dependent behavior and evolution of a biomolecule by integrating Newton's equation of motion. Thereby, they have proven useful for studies of functional mechanisms of proteins and polymers but can also be applied to model the dynamics of far bigger systems like stars or galaxies ([56], [63], [73]). The first MD simulations were simulations of simple gasses in the 1950s [6]. In 1967 Verlet proposed a procedure for time integration, the Velocity Verlet algorithm, that still represents the standard in recent MD simulations (for a brief description see appendix B.2.1). The first simulation of a protein was performed in 1977 [125] and the groundwork enabling these simulations was awarded the 2013 Noble Prize in Chemistry ([106], [109]). The main advantages of MD simulations are the fact that they capture the positions and velocities of all particles in the simulation at every time step and that one is able to completely control the simulation conditions. Both would be very difficult in any experiment. Thereby, the particles in the simulations can be seen as the building blocks of an abstract model so that it is up to the researcher to define the concrete setting of the simulation. The particles' trajectories throughout the simulation are determined by the forces acting on the particles. Therefore, a model force field must be implemented representing the various interactions of the particles ([56], [73], [143], [151]). In this study, we derived an idea for the particles of our simulation from the previously described properties of confined DNA. As discussed, the action of compaction proteins, supercoiling and macromolecular crowding divide the bacterial chromosome in a string of structural units. Furthermore, the DNA model had to reflect the property of chain connectivity. Therefore, we used the bead-spring model of DNA, in which the chromosome is modeled as a spring connected chain of spherical beads. In figure 3.1 a schematic depiction of a bead-spring chromosome is shown.

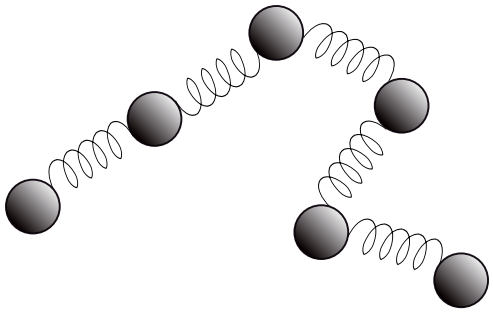

Figure 3.1.: Schematic depiction of the bead-spring model for the chromosome. The beads represent the structural units of the chromosome made up of supercoiled domains of compacted DNA. The springs conencting the beads ensure the chain connectivity and elastic properties of the chromosome.

Here, the beads represent the particles of the simulation. In accordance with the above mentioned numbers, we divided the chromosome of *B. subtilis* into 80 beads. Thus, the beads represented compacted units of $l \approx 52.5$kbp. Again, we determined the diameter d_B of a bead using the radius of gyration r_g and the length of one bp of DNA $b = 0.34$nm as

$$d_B = 2 \cdot r_g = 2 \cdot \frac{\sqrt{l} \cdot b}{\sqrt{6}} \approx 64\text{nm} \quad . \tag{3.1}$$

Furthermore, we had to define the different interactions of the particles with each other and introduce the constraint of the cell volume. To model the entropic repulsion between the beads we used the electrostatic Debye-Hueckel (DH) potential

$$V_{DH} = C \cdot \frac{q_1 q_2 \exp(-\kappa r)}{r} \text{ for } r < r_{cut} \quad . \tag{3.2}$$

For two beads of charge q_1, q_2 and at a distance of r. Furthermore, we set $\kappa = \frac{1}{d_B}$, $r_{cut} = 3d_B$ and the prefactor $C = l_{Bjerrum} \cdot k_B T/e^2$. Here, $l_{Bjerrum}$ is the Bjerrum length

$$l_{Bjerrum} = \frac{e^2}{4\pi\epsilon_0\epsilon_r k_B T} \quad , \tag{3.3}$$

which is the length at which the Coulomb energy between two unit charges is equal to the thermal energy $k_B T$. Here, e is the elementary charge, ϵ_0 is the vacuum permittivity, ϵ_r is the relative dielectric constant of the medium, k_B is the Boltzmann constant, and T is the temperature of the system. To implement the harmonic springs between the beads a Harmonic potential was used

$$V_H(r) = \frac{1}{2} k (r - r_0)^2 \quad , \tag{3.4}$$

where we set $r_0 = 0$. Importantly, the repulsive DH-potential and the attractive harmonic potential were adjusted to compensate each other at a distance of $r = d_B$.
The chromosomes were spatially constrained by the cell which we implemented as a cylinder of length $l_{cell} = 4\mu m$ radius $r_{cell} = 0.5\mu m$. The volume of the cell doubled during

the simulation. To prevent the chromosomes from leaving the cell we implemented a particle-membrane interaction using a Lennard-Jones (LJ) potential. It is defined as

$$V_{LJ} = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} + c_{shift} \right], \text{ if } r_{min} < r < r_{cut} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{, otherwise} \end{cases} . \tag{3.5}$$

Here, we set $\sigma = d_B$ as the diameter of a bead and $c_{shift} = 0.25$, and $r_{cut} = 2^{\frac{1}{6}}\sigma = 1.1225\sigma$, thereby turning the LJ potential into a purely repulsive Weeks-Chandler-Anderson (WCA) potential. In order to keep the system at constant temperature, a langevin thermostat was used. It is based on an extension of Newton's equation of motion to

$$m_i \dot{v}_i(t) = f_i\left(\{x_j\}, v_i, t\right) - \gamma v_i(t) + \sqrt{2\gamma k_B T}\eta_i(t) \quad . \tag{3.6}$$

Here, f_i are all deterministic forces from the interactions, γ is the bare friction (the friction term accounts for dissipation in a surrounding fluid) and η is a random, "thermal" force (mimics collisons of the particle with solvent molecules at temperature T). The random force η satisfies

$$\langle \eta(t) \rangle = 0, \langle \eta_i^{\alpha}(t)\eta_j^{\beta}(t') \rangle = \delta_{\alpha\beta}\delta_{ij}\delta(t-t') \quad . \tag{3.7}$$

Here, $\langle . \rangle$ denotes the ensemble average and α, β are spatial coordinates [205].
The MD simulations were carried out with simulation package `ESPResSo`. Here, the equations of motion are integrated with the Velocity Verlet integrator (see appendix B.2.1) with a fixed time step of $t = 0.01\tau$ [205]. This approach is also found in many of the before mentioned MD studies on the dynamics of confined polymers ([7], [82], [84], [89], [90], [130], [131], [147], [177]).
The simulations started with a circular chromosome which was replicated in the course of the simulation. The starting configuration of the chromosome was obtained as a random walk on a cubic lattice with the MC algorithm presented in chapter 2. Afterwards, the chromosome configuration was projected back onto the confining cylindrical compartment of the MD simulation. To ensure that no unphysical forces arise between beads in close proximity at the start of the simulation, an equilibration phase was established prior to the start of the actual simulation. In this 'warm-up' the chromosome configuration was integrated for some time with capped forces. In this process, occurring forces were artificially limited to a certain cap value, which was subsequently increased step by step. Thus, possible overlaps between beads could adjust and an equilibrated starting configuration was obtained. The process could be verified by measuring the energies during the equilibration phase as shown in figure 3.2. It can be seen that the energies settle very quickly at a relatively constant level and an equilibrated starting configuration was achieved.

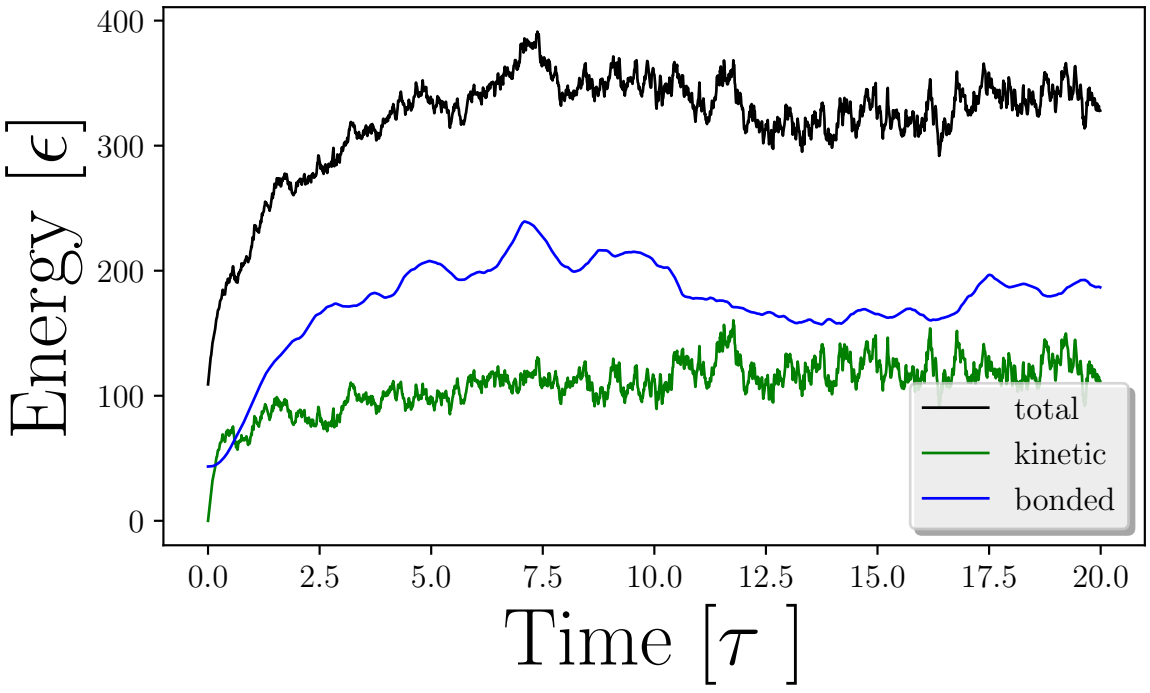

Figure 3.2.: Measurement of energies during equilibration phase. Both energies and time are shown in simulation units. The total energy is shown in black. The kinetic energy resulting from the particle velocities is shown in green and the energy of the harmonic bonds between the particles is depicted in blue.

After the equilibration phase the actual simulation could start with the joint replication and segregation of the chromosomes. Within the framework of the model, it was possible to implement the two replication models of the track model and the factory model discussed above. While the two replisomes move along the chromosome in the track model, replication is locally fixed in the factory model due to the replication factory positioned at midcell. In the simulations, the chromosome was duplicated bi-directionally, with two replisomes running in opposite directions starting from ori. The replication in the simulations was divided into individual duplication steps, during which one bead was duplicated in each direction. In the track model, the new beads were created in a random radius around the original position of the bead to be duplicated. In the factory model, an additional bead was fixed centrally in the middle of the cell and served as a replication factory. The mother chromosome was connected to this chromosome by additional springs and was thus pulled into the center of the cell. There, the beads closest to the replication factory were duplicated. The duplicated chromosomes then segregated from the center of the cell toward the poles. Thus, with this implementation of the factory model the replication was spatially fixed. A representation of the two replication models can be seen in figure 3.3.

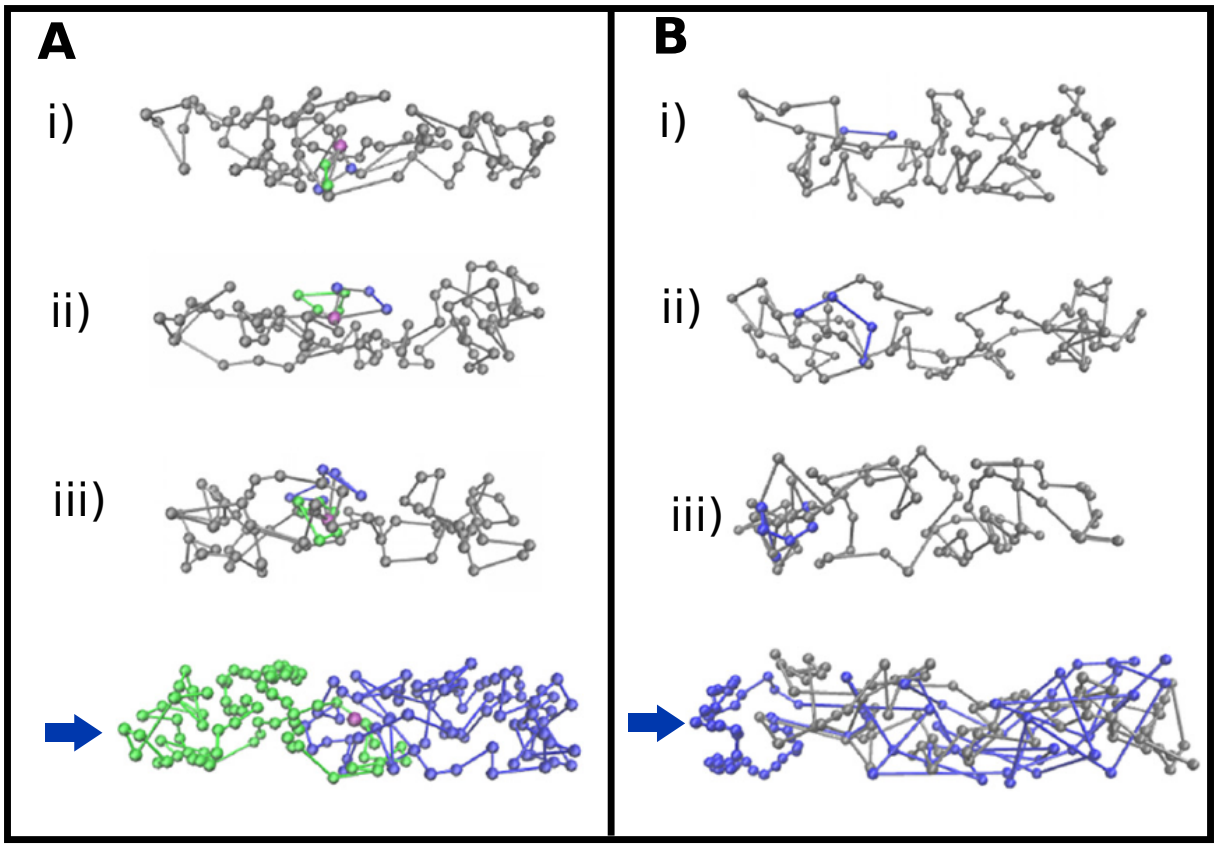

Figure 3.3.: Illustration of the two replication models in the MD simulations. Shown are the first three duplication events (i)-(iii) and the final configuration after termination of replication at the bottom. **A:** Within the factory model the old chromosome (shown in grey) is replicated at the replication factory (pink bead) in the center of the cell. After replication the new chromosome arms (depicted in green and blue) extend from midcell towards the poles. **B:** In the track model of replication the polymerases move along the old chromosome (shown in grey). Thereby, the new chromosome (shown in blue) emerges near the positions of the old chromosome beads.

In the simulations the newly created beads increased their size during a duplication period to the values of the old beads. It was assumed that newly build DNA is compacted right after synthesis. Furthermore, the entropic repulsion between the full chromosome and the partially replicated chromosome causes them to start separating. Therefore, replication and segregation occur simultaneously. In the literature, the factory model of replication is assumed for *B. subtilis* ([55], [104], [129], [167]). Simulations also showed better agreement with the factoy model, in which the chromosomes showed a more regular separation from the cell center to the poles. In contrast, "trapped" configurations of non-segregating chromosomes appeared more frequently in the track model as shown in the final snapshot of figure 3.3B.
It can be seen that while intermediate chromosome configurations of the factory model (steps i-iii in figure 3.3) are still mixed, the two chromosomes are almost completely separated at the end of replication. A more in-depth comparison of the two replication models will be made in the third project of this thesis in chapter 4. For the comparison with the model organism *B. subtilis*, we thus used the factory model in agreement with the literature.

With these implementations, our model was nearly complete. Finally, only the units for the simulation had to be defined in `ESPResSo`. Because `ESPResSo` does not predefine units, they must be specified by the user. For this, the length-, mass- and energy-scale have to be defined by the user and all remaining units are derived from these basic choices [205]. Here, the diameter of a bead, d_B, defined the basic length scale $[length] = 1d_B$. Also, the mass of a bead, m_B, defined the basic mass scale. For the latter we made use of the fact that one bp of DNA has a weight of $m_{1bp} = 650Da$ [4]. Thus, the mass of a bead is the produt of the amount of DNA per bead given by the loopsize and the weight of one bp. We received $[mass] = 1m_B = l \cdot m_{1bp}$. The energy scale was given by the thermal energy $[energy] = \epsilon = k_B T$ with the Boltzmann constant k_B and the temperature T. With these choices we obtained the basic time scale of the simulation as

$$[time] = \tau = [length]\sqrt{\frac{[mass]}{[energy]}} \quad . \tag{3.8}$$

If we insert values here and assume a temperature of $T = 300K$ and a length of the chromosome of 4.2 Mbp divided into 80 blobs we receive $\tau \approx 2.37 \cdot 10^{-7}s$. Similarly, we can determine other required quantities for the simulation in MD units. For example the value of the Bjerrum length at 300 K is $l_{Bjerrum} \approx 0.7095nm = 0.022d_B$.
An overview of values used in the simulation in SI units and in MD units is given in table 3.1

quantity	value (SI units)	value (simulation units)
diameter blob d_B	32 nm	1 $[d_B]$
thermal energy $k_B T$	$k_B \cdot 300K$	1 $[\epsilon]$
mass blob m_B	$5.665 \cdot 10^{-20}kg$	1 $[m_B]$
$l_{Bjerrum}(300K)$	$0.7095nm$	$0.022[d_B]$
time τ	$2.37 \cdot 10^{-7}s$	1 $[\tau]$

Table 3.1.: Table of used units in MD simulations. SI values and simulation values are shown.

3.3. Results

3.3.1. Analysis of time scales

A very common problem of MD simulations is that the biological events of interest take place on timescales that require a very long computation time for MD simulations. In addition, some interesting dynamic properties of biological molecules cannot be simulated directly because of nanosecond time scale limitations. A common problem is that the dynamic evolution of many molecular systems occurs through a series of rare events as the system moves from one potential energy basin to another. Thus, the task is to simulate a series of rare transitions between potential energy minima in order to perform realistic simulations of a molecular system. Thereby, it is possible that the question turns out to be a multiple timescale problem. This may be the result of special energy landscapes or may be due to the specific process of the problem in question ([11], [63], [73]). Such a problem could also be found in the work presented here. We see from the value of a simulation time step in table 3.1 that it is very small and thus a direct simulation of the replication time of *B. subtilis* of $\sim 55min$ was impossible. However, we used a preliminary consideration to justify accelerating the simulation time. The dynamics of the system consists of the rare duplication events of two beads, while between them there is only entropic 'equilibration' of partially replicated chromosomes. Accordingly, we are interested in correctly simulating the duplication events and the subsequent entropic equilibration phase during which the partially replicated chromosome arms segregate as much as possible within the limits of their still existing connections. After entropic equilibration is completed, a phase follows in which the still-connected chromosome arms remain in a state of equilibrium and cannot separate any further. Therefore, we can jump to the next duplication event at this point. Our task is thus to obtain an estimate for the phase of entropic equilibration following a duplication event. For this purpose, simulations were performed with completely intermingled chromosomes consisting of different numbers of beads. The theoretical time of equilibration of two beads can be derived from the results by interpolation as can be seen in figure 3.4.
The results from figure 3.4 show that the entropic separation of intermingled chromosomes is very fast and accomplished within the time scale of μs. Thus, the combined time scale of replication and segregation in bacterial cells is dominated by the replication time. This means that the time required for the duplication polymerases to migrate from one bead to the next is the determining factor for the duration of the combined process. To determine the time required for the new beads created in each duplication step to entropically equilibrate from the old chromosome we used the fit to the data from figure 3.4. We found that for two beads a separation time of $\approx 78\mu s$ is to be expected. This means that after this time, the newly formed beads were repelled from the mother chromosome as far as possible. Consequently, after that we can jump to the next duplication event. Thus, we enabled the representation of the complete replication phase while reducing the required computing time.

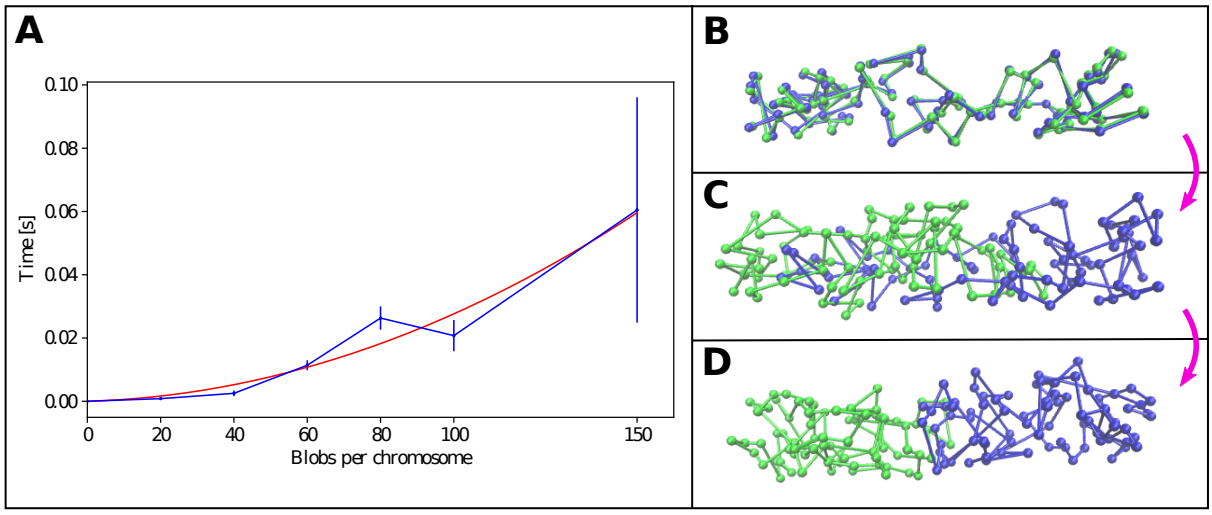

Figure 3.4.: Interpolation for entropic equilibration. **A:** Time needed for entropic separation of intermingled chromosomes of different numbers of beads. In red a polynomial fit for interpolation of the simulated separation times (in blue) is shown. **B - D:** Different stages of separating chromosomes. In **B** the starting configuration is shown which has already started to separate in **C** and is completely separated in **D**.

To verify this estimation we used another measure proposed by Thirumalai et al. [180] as well as Whitfield et al. [210], the ergodic measure. The question to be answered was at what point one can accept the ergodic hypothesis. This assumes that the average over the simulation trajectory is equal to an average over all states accessible to the system. Since the ergodic hypothesis is difficult to prove, one tests a necessary criterion of it instead: At equilibrium, independent simulations over an ergodic system must be self-averaging. The ergodic measure is used to estimate the simulation length needed to guarantee self-averaging. For this one calculates the mean-square difference between the average taken over a simulation α and the average taken over a simulation β, summed over all atoms of the system. The difference then provides a measure of the convergence of the two averages ([110], [180], [210]). One way to define the ergodic measure is to consider the energies of the particles.
In this case, the ergodic measure $\chi^2(t)$, could be defined as follows

$$\chi^2(t) = \frac{1}{N}\sum_{j=1}^{N}\left[\epsilon_{aj}(t) - \epsilon_{bj}(t)\right]^2 \quad . \tag{3.9}$$

Here, $\epsilon_{aj}(t)$ is the energy for the j-th particle in simulation a and $\epsilon_{bj}(t)$ is the corresponding quantity for simulation b. The total number of particles is N. The test for ergodicity with $\chi^2(t)$ is straightforward. If the system is ergodic at some point τ then $\chi^2(t)$ must vanish as the simulation time t approaches τ [180]. In figure 3.5 the results for the calculation of the ergodic measure for the MD simulations are shown.

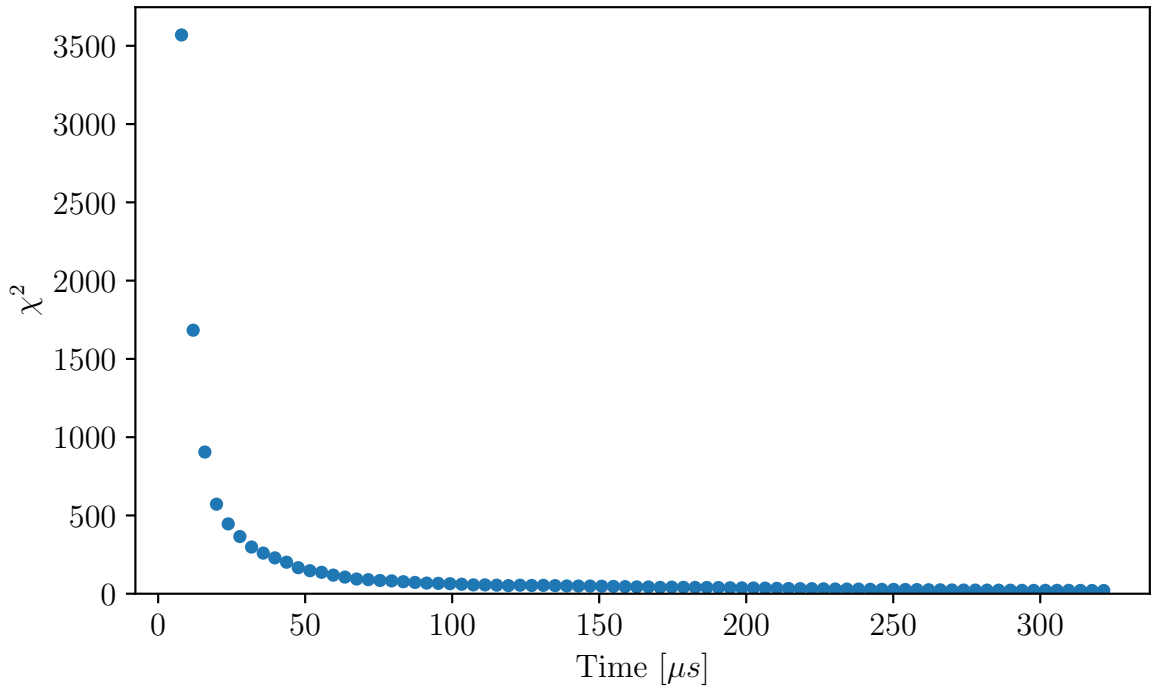

Figure 3.5.: Evaluation of the ergodic measure as defined in equation 3.9 for the MD simulations. For the plot the ergodic measure was calculated for 50 pairs of simulations and the results were averaged.

The time scale in the plot of figure 3.5 starts after duplication of a bead in a typical simulation of a chromosome consisting of 80 beads. For the calculation of the averaged entropic measure, 50 individual values of the entropic measure were calculated from 100 simulations and then averaged. The result of figure 3.5 confirms our previous estimate very well as the entropic measure decays to zero at $\tau \approx 75\mu s$. We therefore assumed that after the duplication of a new bead the system's average properties correspond to equilibrium averages after the time τ. Thus, we could jump from this point to the next duplication event.

3.3.2. Distance of oris over time

With the above considerations it was possible to speed up the simulations accordingly to cover a time of 55min. With this, we could now compare the results of the MD simulations for entropic segregation of oris with the experimental data. The quantitative analysis of experimental data included 80 separation events representing the typical separation patterns. Here, the separating oris were tracked at 10-s intervals and their distance was measured along the long axis and short axis of the cell. In figure 3.6 A, three examples of commonly found separation behaviors in the experiments are shown. Figure 3.6 B shows three corresponding examples of the separation of oris in the simulations.

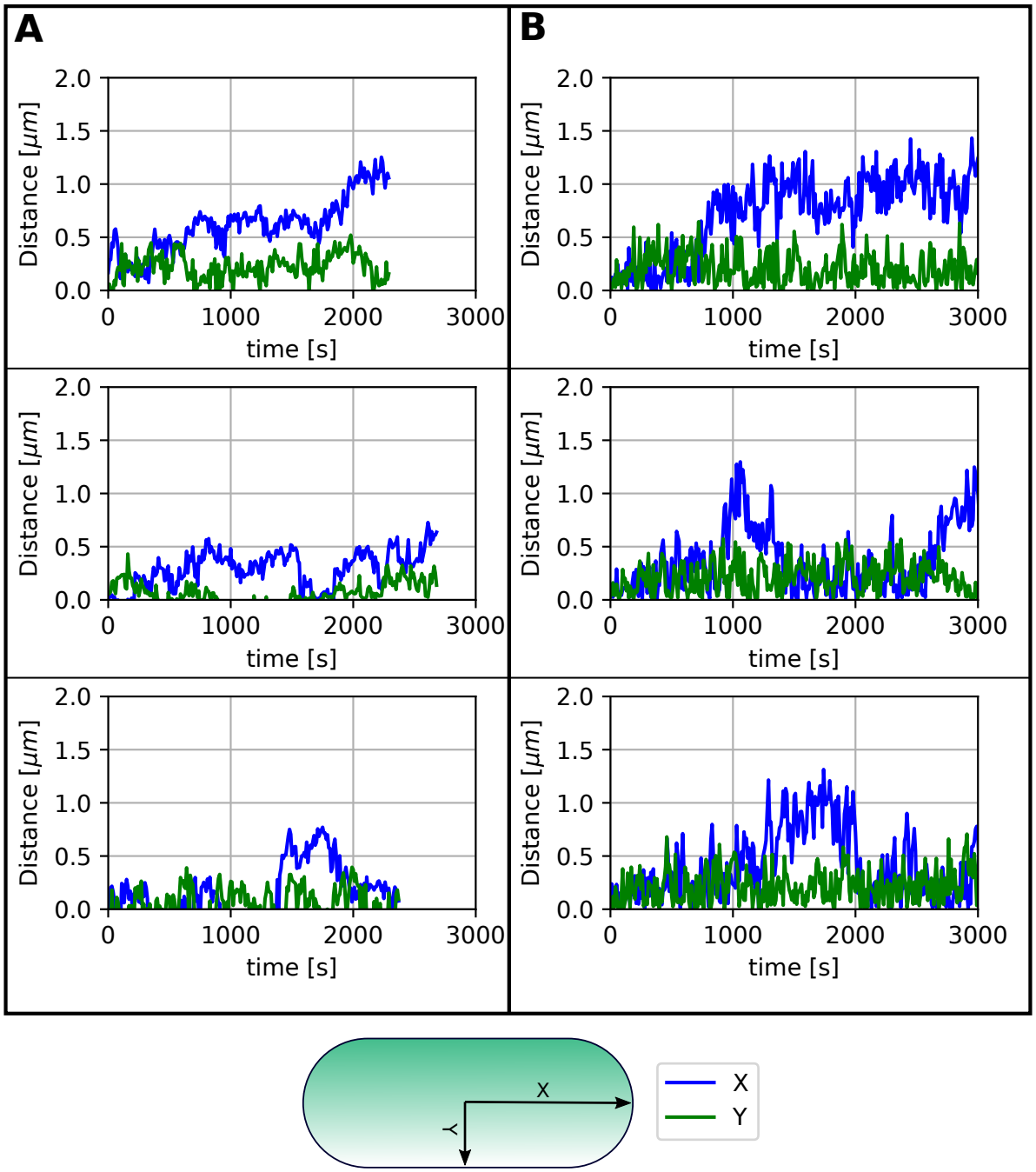

Figure 3.6.: Comparison of ori distances as measured in experiment and simulations. **A:** Experimental tracks for the separation of the oris over time as measured in the experiments. The distance along the long axis (X) of the cell is shown in blue and the distance along the short axis (Y) of the cell is shown in green. **B:** Example tracks of the separation of oris in the MD simulations. The simulations were performed for chromosomes consisting of 80 beads for a period of 60 min.

It can be seen from the plots of figure 3.6 that the separation of the oris is stochastic. The examples show very different separation patterns. The two upper plots in A and B show separation curves in which the separation is linear on average with a constantly increasing distance of the oris from each other. On the other hand, both the experiments and the simulations also showed patterns like the one shown in the middle row of figure 3.6. Here,

the distance between the oris initially increases, but then drops again in the meantime and only increases again afterwards. This could rather be described as an oscillating separation pattern in contrast to the linear patterns in the upper plots. Finally, there were also trajectories in which the oris are barely separated at the end of the tracked time. Examples for this are shown in the lower plots of figure 3.6. Based on these examples, we thus concluded that no uniform pattern for the separation of oris in *B. subtilis* can be established. Instead, one finds a pronounced heterogeneity of separation trajectories in both the experiments and the simulations. Therefore, we concluded that the model of entropic separation of the oris provides a good description for the observations. We additionally verified this by taking the average distance of the separating oris over the 80 experimental trajectories and comparing it to an average over 80 tracks from the simulations. The experimental tracks all showed a length of at least 1800s, so that the individual trajectories could be averaged over this period. The results are shown in figure 3.7.

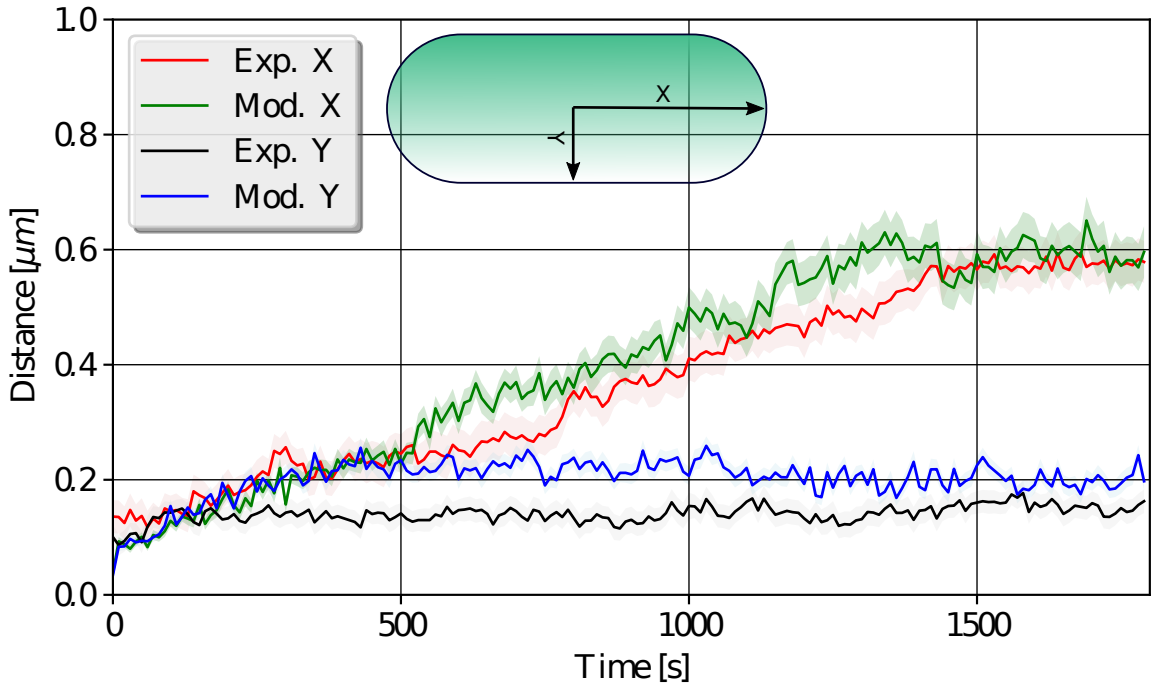

Figure 3.7.: Comparison of the mean distance of separating oris for the experimental data and the simulation data. The results of 80 trajectories were averaged and the distances along the long and short axis of the cell were compared. The shaded areas indicate the standard error of the mean. The distance of the oris along the longitudinal axis of the cell is shown in red for the experimental data and in green for the simulation data. Along the short axis of the cell, the experimental data are shown in black and the model data in blue.

The curves from figure 3.7 confirm that, on average, the oris show a linear separation along the longitudinal axis of the cell. At the same time, the distance of the oris along the short axis of the cell hardly changes. Furthermore, the overlap of the curves shows the good fit of the model with the experimental data. On average, the distance of the oris is about 0.6μ m after 1800s in both the model and experimental data.

3.3.3. Step size distribution

Another parameter with which we compared experimental data and simulations is the step size distribution for the indicidual oris. We expected a Gaussian distribution of the step sizes since the entropic force acting between the chromosomes should lead to a diffusion movement of the chromosomes [155]. In figure 3.8 the probability density function (PDF) for the step size distributions of the experimental tracks and the simulations are compared and probability plots for both data sets are shown.

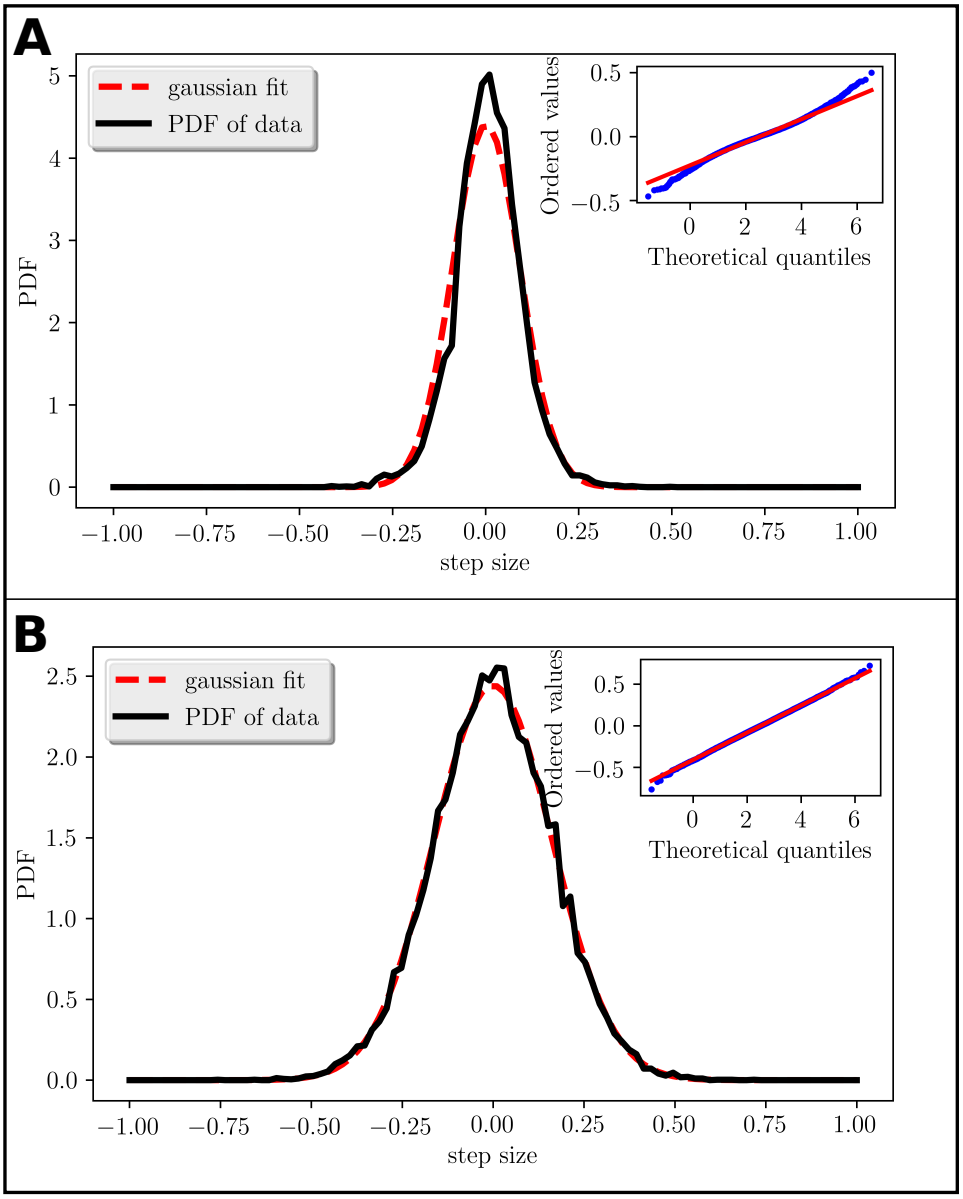

Figure 3.8.: Step size distribution of the ori movement. **A:** Analysis of experimental step size distribution. In the main plot the calculated PDF is shown in black and compared to a fitted normal distribution shown in red. The subplot is a probability plot for the experimental data. **B:** Analysis of the step size distribution from the simulation tracks. Again, in the main plot the calculated PDF is compared to a fitted normal distribution and a probability plot is shown in the subplot.

The plots of the PDFs in figure 3.8 show that the step size distributions for the experimental data and the simulation data are approximately Gaussian as depicted by the fitted normal distributions. In order to compare the probability distributions of the datasets to a normal distribution in both cases probability plots are shown. Here, the quantiles of the measured distributions (y-axis) are plotted versus the expected quantiles of a normal distribution (x-axis). For a perfect fit one would expect a linear relationship. However, we find that the experimental data shows a reasonably linear pattern in the middle of the probability plot but deviations at the edges. The deviations are mainly due to the high number of relatively small steps in the distribution. Presumably, these are a consequence of experimental difficulties in the tracking process. With this restriction we find a Gaussian step size distribution. This is in accordance with the expectation of a directed diffusion as a consequence of entropic segregation of the chromosomes.

3.3.4. Subcellular positioning of oris in the cell

The analyses presented so far suggested that the mechanism of entropic segregation is able to reproduce the basic separation dynamics in *B. subtilis*. However, in addition to the successful separation of the genetic material, the cell also depends on this being accompanied by an appropriate organization of the chromosomes in the cell. To examine at least one aspect of this organization we looked at the final positions of the oris after replication. As discussed above, the chromosome of *B. subtilis* adopts an *ori-ter* configuration in sporulating cells and alternates between the left-*ori*-right-pattern and the *ori-ter* pattern during vegetative growth ([55], [173], [197], [199]). Since the organization of the chromosome in the cell was not the focus of the publication, no corresponding experimental data were included for this purpose. However, from the simulations we could determine the relative positions of the two oris in the cell at the end of replication. The expectation was that we would also find one of the two patterns mentioned here. The results of this analysis can be seen in figure 3.9.
Figure 3.9 A shows the development of the two oris over time in the course of an example simulation. The positions of the oris are marked by the dots and the color code indicates the time. The oris are near the replication factory in the center of the cell at the beginning of replication. From there, they move with increasing time in the direction of the poles or the quarter cell positions. In figure 3.9 B the final positions of the oris from all 80 trajectories in the simulations are shown. The average positions along the long axis of the cells from these points are $x_1 = 0.38$ and $x_2 = 0.61$. In the case of a left-*ori*-right-pattern we would have expected average positions of $x_{exp,1} = 0.25$ and, $x_{exp,2} = 0.75$. For the *ori-ter* pattern even further poleward positions of the oris are expected. Thus, we find that the results of the simulations showed more similarity with the left-*ori*-right-pattern. However, the model could not exactly reproduce the organization of oris in the cell according to the known patterns. This indicates that additional mechanisms might be needed to realize exact positioning of the origins in the cell. This finding is in agreement with simulations on the organization of oris in *E. coli* where it was shown that entropic repulsion alone is not sufficient for an accurate chromosome organization-segregation. Instead, it was shown that in *E. coli* MukBEF and the oris act together as a self-organising system where preferential loading of MukBEF places the oris at the expected quarter cell position [72]. Due to the fact that a variety of different organizational systems have already been found in *E. coli* and *B. subtilis* and that entropic segregation results in a diffusive process, it is not surprising that our result above also indicates the need for additional organizational

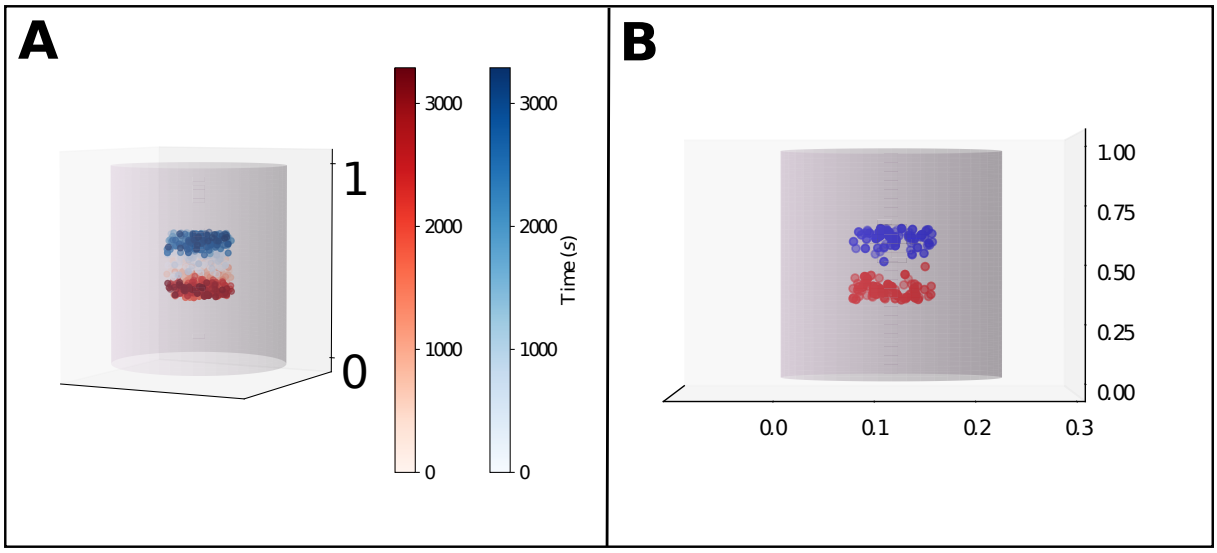

Figure 3.9.: Ori positions in MD simulations. **A:** Positions of the separating oris in the course of replication. The positions of the respective oris are depicted in red and blue while the time course is defined by the colorbars. The plot shows the positions of an example simulation. **B:** The respective end positions of the separating oris for the 80 simulations considered are shown in blue and red. As in A, the cell length is normalized to 1, so the expected quarter-cell positions of oris would be 0.25 and 0.75.

mechanisms to realize the highly ordered structures observed in bacteria. It would therefore be interesting to implement further mechanisms in the simulations. Such are presented in the third chapter 4, in which the MD simulations were further extended including the effects of ParAB and SMC.

3.4. Project summary and outlook

The project described here addresses the segregation of oris in *B. subtilis*. This is a particularly exciting field since no unique mechanism for chromosome segregation has been identified in bacteria so far ([9], [37], [55]). In the specific case of *B. subtilis*, experiments showed that even a deletion of the proteins ParAB and SMC, which are often mentioned in connection with chromosome segregation, does not necessarily prevent the oris from separating ([37], [55], [78]). Therefore, the entropic segregation of chromosomes was investigated as another possible mechanism for the separation of the genetic material in *B. subtilis* in this project ([7], [84], [130], [131]). In order to study the segregation dynamics of the oris the Graumann lab performed experiments in which the oris were tracked at 10-s intervals in the course of replication. The data from these experiments were then compared with MD simulations of entropic segregation of chromosomes.

Research approach For the simulations, a bead-spring model of the chromosome was assumed, in which the DNA is divided into supercoiled domains which represent the particles of the MD simulations. Start configurations with this model were produced with the MC scheme presented in the previous chapter. The simulations implement the entropic interaction between the chromosomes via a repulsive potential between the monomers of

the chromosomes. In combination with the confinement due to the cell wall this results in separation of the chromosomes. The potentials which were used to model interactions between the particles are comparable to the ones used in a number of similar studies ([7], [82], [84], [89], [90], [130], [131], [147], [177]).
Since replication and segregation occur simultaneously in bacteria, bi-directional replication was also implemented in the simulations in form of the track model and the factory model of replication. For *B. subtilis*, a large number of studies suggest that replication takes place in the form of the factory model, so this was also used as the basis for the results discussed here ([55], [103], [104], [129]). A more detailed comparison of the factory model and the track model is presented in the following chapter 4.
In order to compare the results of the MD simulations with the experimental data, an acceleration of the MD time was necessary. This could be achieved as simulations of the separation of complete chromosomes revealed a dominance of the replication time scale compared to the segregation time. Here, the entropic segregation of the chromosomes proved to be a very effective process, which runs significantly faster than the comparatively slow replication, the duration of which is determined by the speed of the duplication polymerases along the chromosome. These different time scales result in a relatively short equilibration time following the duplication of new beads after which the system is in a temporary equilibrium, so that at this point the simulation can be accelerated by transitioning to the next duplication event. These considerations were verified by the additional calculation of the entropic measure. The entropic measure is a measure of the self-similarity of the simulations whose decay to zero is a measure of reaching a thermodynamic equilibrium state. With this a MD scheme was developed that provides data comparable with experimental measurements in *B. subtilis*. Therefore, analogous to the experimental procedure, the distances of the oris from each other could be tracked in the following and compared with the *in vivo* data.

Key findings Comparison of individual measurements for the distance of the oris in the course of replication showed a large heterogeneity of the trajectories. Similar trajectories were found in the experiments and the simulations. Thereby, we found both approximately linear separation movements of the oris from each other as well as oscillatory movements of the oris or even no separation at all at the end of a measurement. However, the average separation of the oris in the ensemble showed an almost linear trend along the long axis of the cell in both experiment and simulations. Thus, these findings rather support the idea of separation due to entropic segregation instead of a coordinated movement which would be expected from any motor-like mechanism. At this point we can conclude that the analyses revealed a good agreement of the simulation model with the experimental data for the combined process of replication and segregation. The results confirmed the proposed factory model for replication as well as entropic segregation as a robust mechanism and driving force for separation of the oris. Thus, we can confirm the prediction that confined chromosomes entropically segregate due to their physical properties ([24], [25], [72], [85], [86], [146], [168], [216]).
Furthermore, our findings indicate that this mechanism can ensure separation of genetic material even in the absence of other separation mechanisms. Since entropic repulsion of the chromosomes is of purely physical origin, one can also speculate that this was the first and most important way for early life forms to segregate chromosomes before additional and more complex mechanisms developed during evolution. Moreover, it is certainly a mechanism that is involved in chromosome segregation in all bacteria, even if there are

additional more sophisticated mechanisms. For example, the efficiency and importance of entopic segregation was recently also shown in experiments with *E. coli*. It could be shown that the probability of successful chromosome segregation decreases with increasing cell width [81].
In addition to the important processes of replication and segregation, the question of the organization of genetic material in the cell must also be considered. Here, the simulations showed that entropic segregation does not reproduce the typical ori configurations for *B. subtilis* ([9], [197], [199]). Thus, these results indicate that additional mechanisms are at least needed for the spatial organization of the chromosomes. This result is understandable since pure entropic segregation does not provide a designated direction of separation. Therefore, it is unlikely that it is sufficient to ensure the complex organization of the chromosome in the cell. This goal could possibly be achieved through specialized proteins like ParAB ensuring the correct positioning of the oris in the cell. Thereafter, the organization of the rest of the chromosome could result from the mechanical properties and spatial confinement of the cell similarly to the discussed mechanism in chapter 2. Furthermore, SMC is known to be very important for the organization of the chromosome in *B. subtilis* by juxtaposition of the chromosomal arms which might even facilitate the segregation of the oris. Consequently, it can be assumed that several mechanisms orchestrate the complex interplay of replication, segregation, and organization of DNA in *B. subtilis* together. Thereby, proteins contribute to the compaction of DNA in the cell. Additionally, they might determine the position of individual loci in the cell and the arrangement of the chromosome arms.

Outlook In further studies, it would thus be interesting to extend the simulations with additional mechanisms such as the effects of the proteins mentioned. This is discussed in the following chapter, where ParAB and SMC were implemented in the MD simulations. Furthermore, the different segregation mechanisms were combined with the two competing models for replication, the factory model and the track model, to form different cell types. At this point another interesting question is analyzed: Is it possible to distinguish which mechanism underlies the movement of oris on the basis of their measured trajectories in the cell?
Besides that, it would be an interesting task for future investigations to extend the MD approach to multiple replicons. For example, the segregation of replicons in *S. meliloti* could be studied. For this, the start configurations could be obtained with the MC scheme of the first project. Thereafter, the MD scheme would allow testing whether the observed motion of oris and ters in *S. meliloti* with a strict temporal order [43] can also be explained by an entropy-driven segregation process. Presumably, however, additional segregation mechanisms such as the ParAB system and RepABC [145] will also be important in the segregation process here. These could be implemented in the MD model by additional forces.

4. Classification of segregation trajectories

In the previous project we already discussed that many different mechanisms are associated with the segregation of chromosomes in bacterial cells ([81], [84], [85], [108], [130], [173], [199], [198], [200], [187]). The successful modelling of the separation of oris in *B. subtilis* using MD simulations of entropic segregation of chromosomes serves as one example for a physical based mechanism. However, it is difficult to identify the relevant mechanisms for chromosome segregation in a given organism. For this, it would be desirable to infer the underlying mechanism of molecular motion from experimental data from SPT experiments. The experiments discussed in the previous project are a good example of the great progress that has been made in the field of SPT in recent years providing data at a temporal resolution of some seconds ([37], [54], [164], [203]). In the case of the previous project the movement of the ori through the cell was tracked.
The ori has a prominent role in the process of replication and segregation of bacterial chromosomes since bi-directional replication starts here. The ori is also important for the organization of the chromosome in the cell ([72], [198], [203]). In addition, the ori occupies a central role in the function of two of the proteins arguably most important for the organization and separation of genetic material. The first of these two protein complexes is the ParAB system, which is one of the best known partitioning systems of bacteria and is employed specifically for the segregation of oris. The second important protein complex, SMC, is also loaded at the ori ([72], [108], [198], [201]). SMC is known to be of great importance for chromosome organization and also supports segregation of oris. Thus, the ori and its movement through the cell are particularly important for understanding the separation of bacterial DNA.
Consequently, it seems natural to infer the underlying separation mechanism from the trajectory of the ori through the cell. Comparable successes have already been achieved in the similar problem of classifying different diffusion processes with ML algorithms ([80], [97], [135], [192]). However, so far it has not been possible to discriminate segregation mechanisms in bacteria using ML approaches. A main obstacle here is that one needs a large number of trajectories in order to train the ML models.
In this project the MD scheme from the previous chapter was therefore extended by implementing the effects of the two proteins SMC and the ParAB system in order to produce a large amount of data which can be used for an automated ML classification. This is especially promising since the previous project has already shown that the data from the MD simulation are comparable to experimental results. Thus, with the additional implementation of the effects of the ParAB system and SMC to the already discussed entropic segregation mechanism described in the previous project, three different drivers for chromosome segregation can now be studied. Furthermore, by switching the two protein systems on and off in the simulations, it is possible to simulate knock-out mutants. The classes to be distinguished are further extended by the two different replication mechanisms of the track model and the factory model.

Thus, the goal of the project presented here was to use MD simulations to produce trajectories of oris for various replication and segregation models corresponding to an SPT experiment, which would subsequently be classified using ML models. Here, a logistic regression (LR) classifier and a support vector machine (SVM) were used as linear classifiers. These were compared with the tree-based gradient boosting (GB) classifier and random forest (RF) classifier. The accuracy of the classifiers was tested for different inputs. In one approach, the classifiers were presented with high-dimensional input vectors consisting of the normalized trajectories. The second approach used low-dimensional input vectors constructed from eight statistical quantities computed for the original trajectories. Furthermore, the classifiers were challenged by presentation of very short trajectories of only a few seconds or by trajectories with a lower temporal resolution (i.e. fewer data points). This tested an application for corresponding experimental data, in which it is not always possible to measure with maximum temporal resolution over the complete replication phase.
The structure of this chapter is as follows. In section 4.1 the implementation of the additional segregation mechanisms in the MD simulations is described. Thereafter, a brief introduction into the used ML algorithms is given in section 4.2 followed by a description of the normalization procedure for the trajectories and the statistical features used for the classification in section 4.3. In section 4.4 the results of the various classification tasks are presented and discussed. A concluding summary of the project is given in section 4.5.

4.1. MD implementation of segregation mechanisms

In this section, the implementation of the additional segregation mechanisms by the proteins ParAB and SMC is described. Both proteins have already been mentioned briefly in section 1.2.4. At this point, we need to take a closer look at individual aspects to ensure that they can be implemented in the simulations.

4.1.1. ParAB implementation

As mentioned above, ParAB belongs to the partitioning systems used by bacteria to position specific loci in the cell. It plays a central role in segregating the oris towards the cell poles in many bacteria and thus is referred to as the closest analog to the mitotic apparatus in eukaryotes ([79], [108]). It was also shown by time-lapse microscopy that the ParAB system helps establish and maintain the *ori-ter* pattern of the newly replicated DNA in *B. subtilis*. The ParAB system appears to "pull" the duplicated origin region to the opposite cell pole, where it is anchored smilarly to its sibling that remains at the other cell pole ([14], [37], [42], [79], [108], [197], [199]). At the same time, it is interesting to note that although all partitioning systems of different bacteria function similarly, their contribution to origin segregation varies dramatically. For example, the ParAB system is absolutely critical for segregating oris in *C. crescentus*. Here, induction of dominant negative allele of ParA was shown to dramatically block origin segregation. On the other hand, *B. subtilis* mutants lacking ParA are able to segregate their chromosomes ([78], [182], [199]). Thus, in different species, one can assume different importance of the ParAB system for the segregation of oris, so it would be exciting to infer this by classifying the trajectories of oris.
The ParAB system consists of three components: the DNA sequence *parS*, the DNA-binding protein ParB, and the deviant Walker A-type ATPase ParA ([14], [42], [79]).

ParB specifically recognizes *parS* sequences, which are typically found near the ori of most bacterial chromosomes. Upon binding to *parS*, ParB is thought to spread on flanking sequences to form the so-called ParB/*parS* partition complex. ParA dimerizes upon ATP binding, which in turn promotes nonspecific DNA binding ([108], [197], [199]). A crucial question to the mechanism of ParAB-dependent transport is the origin of the translocation force. Lim et al. propose a model, where the translocation force is derived from the elastic property of the chromosome [108]. In their proposed DNA-relay mechanism the DNA-associated ParA-ATP dimers serve as transient tethers that harness the intrinsic dynamics of the chromosome to relay the partition complex from one DNA region to another. In this model the characteristic elastic force can thus be estimated by tracking single loci positions prior to replication and segregation. Using a Gaussian fit for the obtained distributions Lim et al. were able to estimate a force of $F \approx 0.06$ pN from the elastic property of the chromosome.
The same approach was used to implement ParAB in the MD simulations. For this purpose, the elastic force resulting from the dynamics of the chromosomal loci was estimated in the simulations as well. This could then be implemented as an external force on the ori in subsequent simulations. Thus, to determine the force, simulations of a chromosome in the cell were performed prior to replication and segregation. From these, the step size distributions of individual loci were calculated. The probability distributions for these are given by the Boltzmann distribution as

$$P(\triangle x) \sim \exp^{-\frac{E(\triangle x)}{k_B T}} \quad , \tag{4.1}$$

where $P(\triangle x)$ is the probability of a locus to fluctuate around its equilibrium point. Further, k_B is the Boltzmann constant, T is the absolute temperature, and $E(\triangle x)$ is the energy associated with the fluctuation. The idea now was to infer the energy potential $E(x)$ from the measurement of the distribution. Figure 4.1 shows the measured distribution in our simulations.

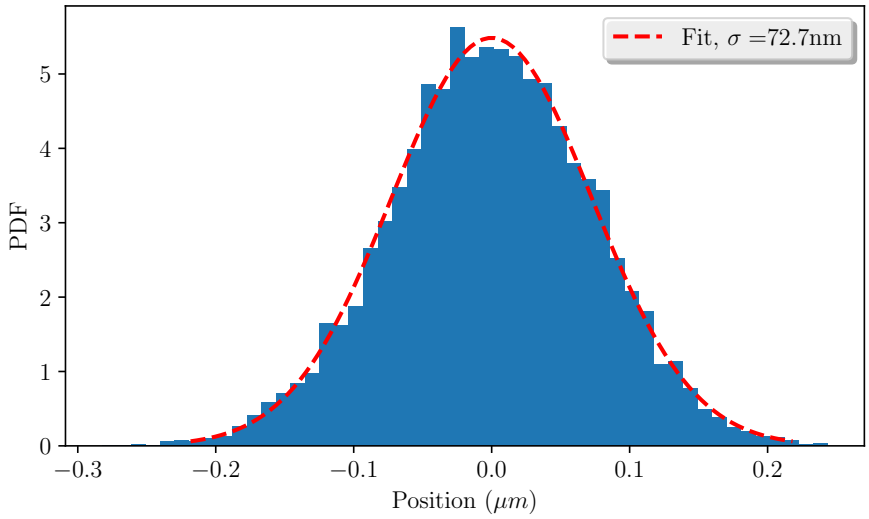

Figure 4.1.: Locus dynamics in MD simulations. Step size distribution of a locus of a freely diffusing chromosome in the cell. The red line is a Gaussian fit to the data shown in blue.

In figure 4.1 the red line depicts a Gaussian fit to the step size distributions. With this, the probability distribution can be written as

$$P(\triangle x) \sim \exp^{-\frac{(x-x_0)^2}{2\sigma^2}} \quad . \tag{4.2}$$

Thus, the chromosomal loci move in a harmonic potential of the form $E(x) = a\,(x - x_0)^2$ with $a = k_B T/2\sigma^2$. From this we can finally obtain the searched force $F(x)$ by differentiating

$$F(x) = 2a\,(x - x_0) \text{ with } k_{sp} = 2a = \frac{k_B T}{\sigma^2} \quad . \tag{4.3}$$

From the fit in figure 4.1 we found $\sigma = 72.7nm$. Using this, we can calculate the elastic force (for a temperature of 300K) as

$$F(x) = k_{sp} \cdot \sigma = \frac{kT}{\sigma} \approx 0.057\text{pN} \quad , \tag{4.4}$$

which is in very good agreement of the value obtained by Lim et al. from experimental measurements tracking single loci in the cell. Thus, for our simulations, we now have found the value of the effective force by which the newly duplicated ori is pulled to the pole.

4.1.2. SMC implementation

In addition to the active partitioning system ParAB, SMC proteins are also frequently associated with the segregation of bacterial DNA. However, it is assumed that these are primarily responsible for the compaction of the DNA in the cell and passively ensure the separation of the chromosomes through the topological separation of the oris ([17], [199]). SMC proteins are essential in many bacteria and are also beyond bacteria conserved in all domains of life ([14], [71], [173], [200], [201]). For example in eukaryotes, condensins act at the earliest stages of mitosis. They compact and resolve interphase chromosomes into rod-shaped structures that assemble at the metaphase plate [71].
In bacteria the SMC complex consists of the SMC protein (= kleisin) and ScpB [14]. In *E. coli* they are called MukB and MukF. The SMC proteins are characterized by interwined coiled-coil domains that have hinges at both ends enabling them to topologically embrace DNA helices. A model was proposed by Wang et al. in which the ring-shaped complexes of SMC encircle the DNA flanking their loading site and thus tethering the DNA duplexes together. Thereby, SMC plays an important role in the formation of topologically associated domains ([14], [17], [71], [199], [200], [201]).
Regarding the association of SMC with chromosome segregation, it was suggested that SMC and ParAB work together to segregate oris in *B. subtilis*. In doing so, one imagines a mechanism in which SMC is loaded at centromeric *parS* sites near the ori, where it encircles DNA and individualizes newly replicated origins by promoting the juxtaposition of DNA flanking *parS* sites, drawing sister origins away from each other. Thus, while SMC is loaded at the origin it is still present and acts along the compelte chromosome arms ([198], [200], [201]). A schematic depiction of this process was shown in figure 1.7.
Similar to the case of ParAB, different consequences of deleting the protein are found for SMC depending on the species considered. For example, the consequences of deleting SMC in *B. subtilis* depend on growth conditions. It was shown that during fast growth

the rapid inactivation of SMC leads to a failure in resolving newly replicated origins and chromosome segregation was blocked. On the other hand, during slow growth chromosome segregation was still possible in the absence of SMC. Here, it was suggested that the ParAB system provided enough origin segregation for the system even though the cells showed more heterogeneous nucleoid morphologies which might have resulted from a defect in the resolution of replicated oris. In *E. coli* it was reported that slow growing cells lacking SMC adopt an *ori-ter* configuration rather than their typical left-*ori*-right configuration ([198], [199]).
SMC proteins have already been implemented in various computer simulations. Goloborodko *et al.* performed polymer simulations of chromosome dynamics and showed that loop extrusion by condensins can robustly compact, segregate and disentangle chromosomes. In this simulations eukaryotic chromosomes were modeled as flexible polymers where each condensin complex was modeled as a dynamic bond between a pair of monomers [52]. Simulations for bacteria were performed by Wang et al. who showed that a limited number of $\sim$ 30 condensin complexes per replication origin can organize DNA in *B. subtilis* [201]. This number is in good agreement with single molecule tracking experiments in *B. subtilis* where the number of $\sim$ 30 SMC dimers moving throughout the chromosome was reported [165].
Thus, for the implementation in the MD simulations the SMC proteins were modeled as additional dynamic bonds between opposite beads on the chromosome as suggested by Goloborodko *et al.* [52]. Thereby, SMC is loaded at the ori-region by successively connecting the beads following the ori in the replication with harmonic bonds. Thus, the number of SMC proteins per chromosome (38) was also comparable to the above mentioned figures. An example snapshot of a chromosome with and without SMC in our simulations is shown in fig. 4.2.

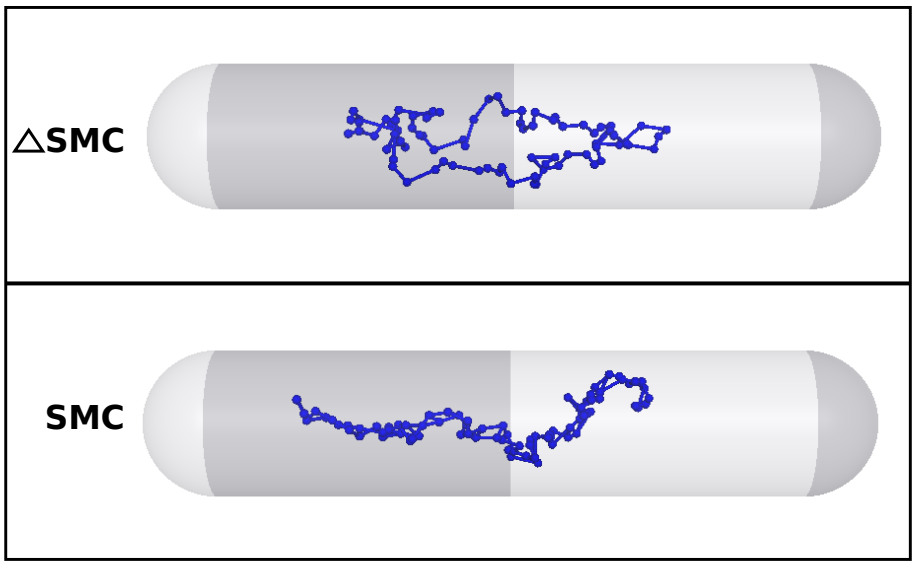

Figure 4.2.: Example snapshots for a chromosome without (upper picture) and with (lower picture) SMC bonds in MD simulations.

The snapshots in figure 4.2 show how the SMC proteins juxtapose the two chromosomal arms and thereby compact the chromosome in the simulation. In contrast, the chromosome is spatially more extended in the absence of SMC.
With the additional implementation of the two segregation mechanisms by the SMC and ParAB proteins, the classes (= cell types from here) to be distinguished by the ML

algorithms could be defined. On the one hand, the two different replication mechanisms of the track model and the factory model could be used. For the factory model, as seen in *B. subtilis*, a left-*ori*-right configuration of the chromosome was assumed, in which replication starts in the middle of the cell. In the track model, however, a *ori*-*ter* configuration was used at the beginning of replication. For segregation, the WT was defined as the case in which all three segregation mechanisms of entropic segregation and the two proteins were active. In addition, knock-out mutants were possible, in which one of the proteins or even both proteins were inactive. Thus, the mutant with inactivated SMC (dSMC) and the mutant with inactivated ParAB (dParAB) as well as the double-knockout mutant with inactivated SMC and dParAB (dSMCdParAB) emerged. Entropic interaction of chromosomes is the consequence of basic physical principles. Therefore, there is no sense in turning it off and it was thus activated in all simulations.
This resulted in a total of 8 different cell types presented to the ML algorithms. An overview of these is given in figure 4.3.

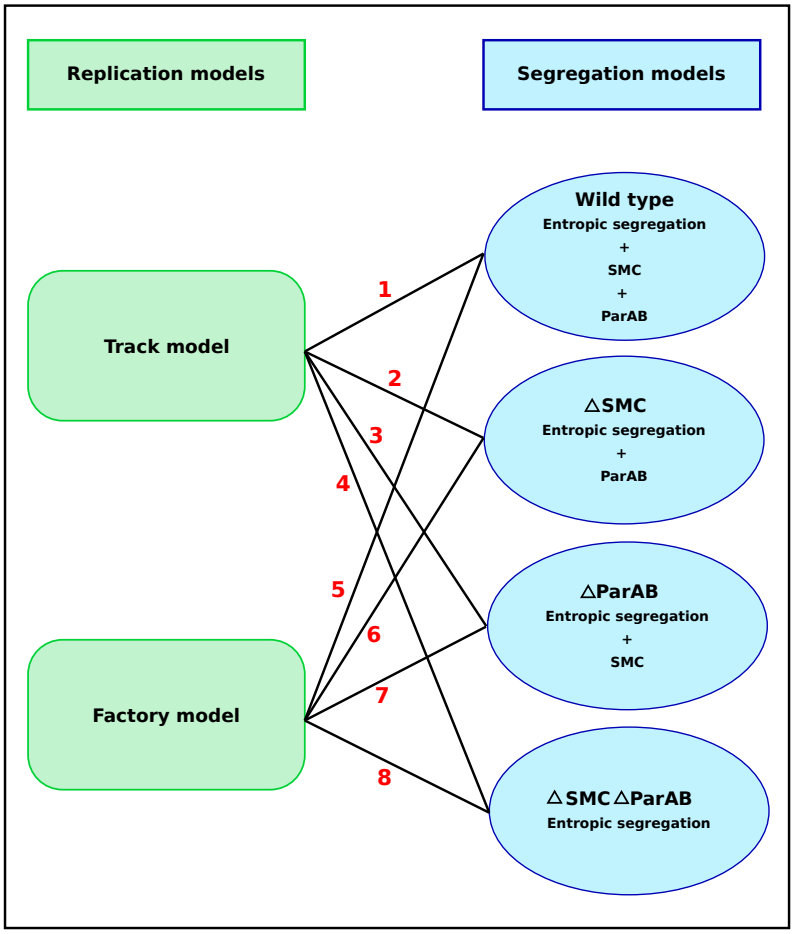

Figure 4.3.: Overview of replication and segregation mechanisms used to create different cell types. For replication, either the track model or the factory model were applied. For segregation, four different mechanisms were possible depending on whether or not the two proteins SMC and ParAB were aktiv. The combination of the two replication mechanisms with the four segregation mechanisms resulted in a total of 8 cell types (depicted by red numbers) to differentiate.

4.2. Machine learning algorithms

For the automated classifications of our trajectories of different segregation mechanisms produced with the MD simulations, various processing steps had to be performed. First, the produced data was divided into a training data set and a test data set (typically in a ratio of 70% training data to 30% test data). The training data was then used to train selected ML models for the classification task.
In order to classify a trajectory that was unseen by the models, it was first preprocessed. Two different approaches were used two preprocess the data. On the one hand, a scheme suggested by Muñoz *et al.* [135] was applied, in which the complete trajectory is normalized and thereafter used as a high-dimensional input vector for the ML models. The second approach was to reduce the dimensionality of the input vector by the calculation of statistical features from the trajectory to be classified. These features were then used to build an input vector for the ML models.
Four different types of classifiers were compared in this study. Two linear classifiers, namely the logistic regression (LR) classifier and a support vector machine (SVM) were compared to two tree-based classifiers, a gradient boosting (GB) classifier and a random forest (RF) classifier. For all classifiers the implementations available in the `scikit-learn` library from python were used. In figure 4.4 a schematic depiction of the applied workflow is shown.

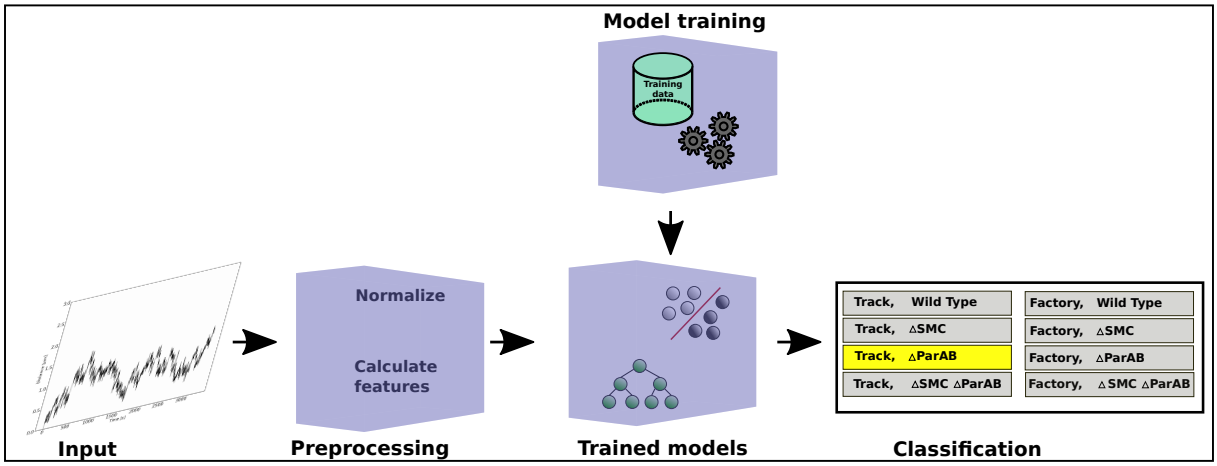

Figure 4.4.: Schematic depiction of the workflow for trajectory classification. An unknown trajectory is first preprocessed by either normalization of the trajectory according to the protocol from Muñoz *et al.* [135] or the calculation of statistical features from the trajectory from which a low-dimensional input vector is build. The input vector is fed into the previously trained ML models which assign it to one of the eight possible classes.

In the following, descriptions of the preprocessing procedures and the concepts of the ML classifiers are given.

4.2.1. Linear models

The LR classifier and the SVM belong to the group of linear classification methods. Such methods expect the target value to be a linear combination of the features. The general notation is

$$y(\beta, x) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \quad , \tag{4.5}$$

where the β_i are the coefficients and the x_i are the features (also called the explanatory variables). $y(\beta, x)$ is the target variable (or response variable) [67]. Upon introducing a virtual variable $x_0 = 1$ we can write $y(\beta, x) = \beta^T X$, with the two vectors $\beta = (\beta_0 \beta_1 ... \beta_n)$ and $X = (x_0 x_1 ... x_n)$.
One of the most widely used linear methods for supervised classification is logistic regression (despite the name regression). In supervised learning, one wants to predict the value of an outcome measure based on a number of input measures [67]. In the literature, the logistic regression classifier is also known as logit regression, maximum-entropy classification or the log-linear classifier. This method predicts the probability that an observation is part of a certain class. Thereby, the LR classifier in its standard form is a binary classifier whose target vector can only take two values ([62], [74]). The idea of a LR classifier is to include the linear model of equation 4.5 in a logistic (or sigmoid) function $\frac{1}{1+e^{-z}}$. Thereby, the logistic function constrains the values of the output between 0 and 1 so that it can be interpreted as a probability. The classifier then predicts class 1 for values greater than 0.5 and class 0 otherwise. Thereby, the probability for the occurance of an event is

$$P(y(\beta, x)) = \frac{1}{1 + e^{-y(\beta, x)}} \quad . \tag{4.6}$$

For the classification of more than two classes the LR classifier can be extended with two different schemes. In the one-vs-rest scheme (OVR) a separate model is trained for each class. Thereby, a binary problem can be solved for each class. In doing this, one assumes that the classification problems are independent of each other. The second scheme is the multinomial logistic regression (MLR). Here, the logistic function is replaced with a softmax function ([62], [74]). For the estimation of the coefficients $\beta_0, \beta_1, ..., \beta_n$ typically the maximum likelihood estimation method (MLE) is used. This method tries to maximize the log likelihood reflecting the odds that the observed values of the dependent variable may be predicted from the observed values of the independent variables ([5], [67]). Thus, the task at hand is an optimization problem. The algorithm starts with some estimate for the coefficients and tries to change them iteratively to increase the likelihood function. In order to reduce the variance of the trained model regularization is applied. In general, regularization procedures add a penalty term to the loss function that shall be minimized to penalize complex models. The most common penalties are the L1 and L2 penalties [67]:

$$\begin{aligned} &\text{L1: } \alpha \sum_{j=1}^{p} |\beta_j| \quad , \\ &\text{L2: } \alpha \sum_{j=1}^{p} \beta_j^2 \quad , \end{aligned} \tag{4.7}$$

where β_j are the parameters of the j-th of p features being learned and α is a hyperparameter denoting the regularization strength which is to be tuned to find the best model. Higher

values of α increase the penalty for more complex models ([5], [62], [67], [74]). In this work, the implementation of the LR classifier from the python library `scikit-learn` was used. Here, a "lbfgs" optimization algorithm is used with the L2 penalty to determine the direction and magnitude of change in the coefficients. The lbfgs algorithm belongs to the quasi-Newton methods and is used by default within the `scikit-learn` library because of its robustness.
As a second linear model we used a SVM, also known as the discriminative classifier. A SVM classifier aims to classify data by finding a hyperplane that divides the data into classes. A hyperplane is a $n-1$ dimensional subspace in an n-dimensional space. The distance from the separating hyperplane to the nearest expression vector is called the margin of the hyperplane. The SVM classifier tries to find the hyperplane that maximizes the margin between the classes in the data ([5], [67]). The name of SVM results from the fact that they use so called "support vectors", i.e. data points of the training dataset which are close to the hyperplane and include them in the decision function of the optimizer. In classifiaction problems the term support vector classifier (SVC) is also used. A typical proplem considering linear models in general is that real data will rarely be linearly separable. Thus, a SVC must be balanced between finding the maximal margin for the hyperplane and minimizing misclassified data points. This balancing task is controlled with a penalty, C, that is imposed on errors. Therefore, if a small value for C is chosen, the classifier will have a bigger bias but lower variance compared to a high value of C ([5], [67], [140]). Furthermore, it is possible to specify different kernel functions for the decision function. These kernel functions can be linear, polynomial or radial basis function. The kernel functions are used to transform the input data. Thereby, non-separable input data can be transformed to higher dimensional space where the problem is separable. One can even prove that for any given data with consistent labels a kernel function exists that will allow the data to be linearly separated [140]. However, this comes at a price which is often referred to as the curse of dimensionality. The problem is that the number of possible solutions increases exponentially with increasing number of variables under consideration. Thus, the algorithm will struggle to select the correct solution and furthermore the tendency to overfit increases [140]. Mathematically, a SVC can be represented as follows

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K\left(x_i, x_{i'}\right) \quad . \tag{4.8}$$

Here, β_0 is called the bias and S is the set of all support vector observations. The model parameters to be learned are termed α, and $(x_i, x_{i'})$ are pairs of two support vector observations, x_i and $x_{i'}$. The kernel function K compares the similarity between x_i and $x_{i'}$. The most commonly used kernels are the linear kernel, the polynomial kernel and the radial basis function kernel

$$\begin{aligned}
&\text{linear: } K\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} x_{ij} x_{i'j} \quad , \\
&\text{polynomial: } K\left(x_i, x_{i'}\right) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \quad , \\
&\text{radial basis: } K\left(x_i, x_{i'}\right) = e^{-\gamma \sum_{j=1}^{p} \left(x_{ij} x_{i'j}\right)^2} \quad .
\end{aligned} \tag{4.9}$$

Here, p is the number of features, d is the degree of the polynomial kernel function, and γ is a hyperparameter called the kernel coefficient which must be greater than zero ([5], [140]). For the simulations presented in this work again the implementation of the `scikit-learn` library was used with a linear kernel.

4.2.2. Tree-based models

In addition to the linear classification methods one can use classifiers build from decision trees. The key concept of a decision tree is to produce recursive binary splits of the input space, so that samples belonging to the same label are grouped together. The input constitutes the root node of the tree, while the subsets represent the successor children. The splitting process is repeated on each subset in a recursive manner. In this sense a decision tree consists of a series of chained decision rules ([5], [80], [97], [135]). Every decision rule occurs at a so-called decision node, with the rule creating branches leading to new nodes. If a branch has no decision rule at the end it is called a leaf of the decision tree. When no further splitting is possible or when the subset at a node has all samples belonging to the same class (i.e. the node is pure), a terminal tree node is reached, where the output is obtained ([5], [80], [97], [135]).
In order to perform the splits at each node, a criterion is needed. The typical measures for this purpose are the Gini impurity or the information gain. The gini impurity tells us how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in that set. It is given by

$$I_G = \sum_{i=1}^{J} p_i(1 - p_i) \quad , \tag{4.10}$$

where J is the number of classes and p_i is the fraction of items labeled with class i in the set. The information gain related to a split is simply the reduction of information entropy, calculated as the difference between the entropy of a parent node in the tree and a weighted sum of entropies of its children nodes. When a final decision tree is obtained, it classifies unseen data by passing it through the nodes of the tree, where each decision is made with respect to which direction to take. The benefits of decision trees are that they are easy to interpret and do not require data processing. However, they have the disadvantage that a small variation in the data may lead to a completely different tree. Furthermore, decision trees tend to overfit data. To circumvent these negative aspects decision trees are nowadays used as building blocks of advanced ensemble classifiers ([5], [80], [97], [135]). Ensemble learning methods are methods that generate many basic classifiers like decision trees and aggregate their results. There are different methods to combine the results of the individual classifiers. The two most prominent ones are bagging and boosting ([19], [20], [76]). The idea of bagging is to reduce the variance of a learning method by averaging the results of many classifiers. The background for this is that for a set of n independent observations $O_1, ..., O_n$, each with variance σ^2, the variance of the mean $\bar{O}$ of the observations is given by σ^2/n. Thus, we see that the variance is reduced through averaging. The bagging approach for ensemble methods uses bootstrapping to take multiple training data sets from the original training data. Then the same number of basic classifiers are trained on the separate bootstrapped data sets and their predictions are averaged.

Thus, the final prediction $\hat{f}_{bag}(x)$ of a machine learning algorithm using bagging can be written as

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(x) \quad , \tag{4.11}$$

where B is the number of different bootstrapped training data sets and $\hat{f}^{b}(x)$ is the prediction of the b-th basic classifier. In the case of decision trees as basic classifiers each individual tree has high variance, but low bias. Averaging these trees then reduces the variance and yields a better classifier. Typically, the overall prediction is determined by a majority vote ([5], [19], [20], [76], [97]).
Ensembles of decision trees might encounter a problem if the data contains one particularly strong predictor within the input variables. In this case all of the trees might use this stong predictor and become highly correlated. To overcome this problem, the RF classifier model was proposed by Breimann [20]. In a RF classifier the single decision trees are decorrelated. Therefore, while growing a decision tree in a RF at each split only a random subset of m predictors is chosen from the input variables as split candidates from the full set of p predictors. Typically, one uses $m \approx \sqrt{p}$ [5]. Thus, it is prevented that all trees use a possibly occurring strong predictor and thereby become highly correlated. Due to this randomization, the bias of the ensemble is slightly higher than that of a single tree, but the variance is decreased and the model is more robust to variations in the dataset ([20], [51], [76], [135]).
The second method for the aggregation of ensembles of decision trees is boosting. Here, the single trees are grown sequentially using information from the previously grown trees. One of the first boosting algorithms was the AdaBoost algorithm from Freund and Schapire [44]. Another generalization of boosting algorithms was presented by Friedman [45] with the invention of gradient boosting machines for both classification and regression. The idea of a GB classifier is that an algorithm, given a loss function and a basic weak learner like a decision tree, may be used to find an additive model that minimizes the loss function. For this, an iterative approach is chosen in which new weak basic classifiers are trained with respect to the error of the whole ensemble learnt so far. Implementations of the GB algorithm are typically initialized with a best guess and then a gradient (e.g. residual) is calculated. Thereafter, a model is fit to the residuals to minimize the loss function and the current model is added to the previous model. This process is repeated for a designated number of iterations. Within an ensemble of decision trees for the GB algorithm a single tree can be rather small with only a few terminal nodes. Such a small tree is sometimes called a stump. By adding small trees to the ensemble, which give a higher priority to observations the previous model predicted incorrectly, the averaged prediction of the model is improved in areas where it does not perform well ([5], [44], [45], [46], [67], [76], [97], [98], [136]).

4.3. Preprocessing protocols

An important step in every ML study is the preprocessing of the data. In this process, the original data is put into a form in which it can be understood by the classifier ([5], [67], [76]). Thereby, the original data is divided into features, which are used as input for the classifier. These features directly influence the results of the classification accuracy. In this work,

two different approaches were used to preprocess the segregation trajectories. These are presented below.

4.3.1. Rescale complete trajectories

The first preprocessing protocol applied to the trajectories is the one described by Muñoz *et al.* [135]. Here, the goal was to design a method that makes it possible to classify heterogeneous data from various spatiotemporal scales. Therefore, the original trajectories are rescaled. Thereby, a new trajectory is constructed via the normalized displacements of the original trajectory. As a consequence, the magnitudes of the resulting new trajectories are comparable, independent of their original values [135]. In detail, the preprocessing consists of the following steps:

1. The original trajectory is given as a vector of positions

$$\vec{X} = (\vec{x_1}, \vec{x_2}, ..., \vec{x_{t_{max}}}) \quad .$$

2. This vector is transformed into a vector of distances traveled in an interval of time T_{lag}, i.e

$$\vec{W} = (\triangle\vec{x_1}, \triangle\vec{x_2}, ..., \triangle\vec{x_{J-1}}) \quad ,$$

   where $J = t_{max}/T_{lag}$.
   Thereby, $\triangle\vec{x_i}$ is defined as

$$\triangle\vec{x_i} = |\vec{x_{iT_{lag}}} - \vec{x_{(i+1)T_{lag}}}| \quad .$$

3. To normalize the data, the vector $\vec{W}$ is divided by its standard deviation to get a new vector $\vec{W}'$.

4. Finally, a cumulative sum of $\vec{W}'$ is computed to construct the normalized trajectory $\vec{X}'$.

The normalized trajectory $\vec{X}'$ can be used as a high-dimensional input vector for the various classifiers. In their paper, Munoz *et al.* point out the high accuracy of the classifications of diffusion models with a RF classifier due to the preprocessing method, which gave good results even for short trajectories [135].

4.3.2. Trajectory features

The ML algorithms used in this work all belong to the class of feature-based methods. In this approach, the trajectories are characterized by certain features that serve as input to the classifiers and form the basis for their learning and predictions. In the approach of normalizing the complete trajectories described above, a very high-dimensional input vector is obtained, where the features are the individual points of the normalized trajectory. Another commonly used approach is to quantitatively compare the trajectories by a set of human-engineered statistical features which are used to construct a lower dimensional input

vector for the classifiers. In the literature there is a wide range of possible features for the quantitative comparison of trajectories from SPT experiments([8], [80], [97], [192], [207]). For this work, eight features were selected, which are described in the following.
Many of the features presented in the following have their origin in studies of distinguishing the different types of motion of particles in SPT experiments. According to Saxton [162], four basic movement types are distinguished here: normal diffusion (ND), anomalous diffusion (AD), directed motion with diffusion (DM), and confined diffusion (CD). The standard way to identify the different motion types is to analyze the mean squared displacement (MSD). For a trajectory of N consecutive positions $x_i (i = 1, ..., N)$ it is defined as follows

$$MSD = \langle r_n^2 \rangle = \frac{1}{N-n} \sum_{i=1}^{N-n} |x_{i+n} - x_i|^2 \quad . \tag{4.12}$$

The four basic motion types are commonly characterized by the shape of their MSD curve:

$$\begin{aligned} &\langle r_n^2 \rangle = 4Dn\triangle t && \text{(ND)}, \\ &\langle r_n^2 \rangle = 4D\,(n\triangle t)^{\alpha} && \text{(AD)}, \\ &\langle r_n^2 \rangle = 4Dn\triangle t + (\nu n\triangle t)^2 && \text{(DM)}, \\ &\langle r_n^2 \rangle \simeq r_c^2 \left[1 - A_1 \exp(-4A_2 Dn\triangle t / r_c^2)\right] && \text{(CD)}. \end{aligned} \tag{4.13}$$

In the equations above, α is the anomalous exponent, ν is the velocity in the directed motion and for confined diffusion the constants A_1 and A_2 characterize the shape of the confinement, while r_C is the confinement radius. We used the MSD along the trajectory as our first feature.
Besides the MSD one can also use the mean squared displacement ratio (MSDR). The MSDR characterizes the shape of the MSD curve. It is defined as

$$\langle r^2 \rangle_{n_1,n_2} = \frac{\langle r_{n_1}^2 \rangle}{\langle r_{n_2}^2 \rangle} - \frac{n_1}{n_2} \quad , \tag{4.14}$$

with $n_1 < n_2$. In order to calculate the MSDR we set $n_2 = n_1 + \triangle t$ and calculated an average ratio for every trajectory as proposed in [97].
Furthermore, one can use the anomalous exponent α (Alpha) as a separate feature. It can be calculated from

$$\langle r_n^2 \rangle = 4D\,(n\triangle t)^{\alpha} \quad . \tag{4.15}$$

For this, the MSD curves of the trajectories were fitted with equation 4.15 so that α could be obtained from the fit. Here, D is the diffusion coefficient and $\triangle t$ is the elapsed time. For AD one has $\alpha < 1$. For normal diffusion (ND) one finds $\alpha \approx 1$ [192].
A different measure for a trajectory is the fractal dimension (FD). It provides an index of complexity by comparing how detail in a pattern changes with the scale at which it is measured [93]. Thus, it can be seen as a measure of the space-filling capacity of a pattern. A curve with a fractal dimension close to 1 behaves quite like an ordinary line while a curve with fractal dimension close to 2 behaves almost like a surface in terms of space

filling capacity. It can be defined in several ways. Here the definition from Sevcik [166] was used:

$$FD = 1 + \frac{\log(L)}{\log(2N-2)} \quad , \tag{4.16}$$

where L is the contour length of the trajectory in the unit square and N is the number of points of the trajectory.
Another feature that was used is the radius of gyration (RG) of a trajectory. It is defined as

$$R_g = \frac{1}{N}\sum_{i=1}^{N} |r_i - r_S|^2 = \frac{1}{N}\sum_{i=1}^{N} |r_i - \bar{r}|^2 \quad . \tag{4.17}$$

Here, r_S is the focus (average position) of the trajectory.
The efficiency (E) is a measure for the linearity of a trajectory and relates the squared net displacement to the sum of the squared displacements

$$E = \frac{|x_{N-1} - x_0|^2}{(N-1)\sum_{i=1}^{N-1} |x_i - x_{i-1}|^2} \quad . \tag{4.18}$$

Thus, the E is a measure for linearity of a trajectory.
Similar to the E the straightness (S) of a trajectory relates the net displacement to the sum of step lengths:

$$S = \frac{|x_{N-1} - x_0|}{\sum_{i=1}^{N-1} |x_i - x_{i-1}|} \quad . \tag{4.19}$$

Finally, the gaussianity (G) was used as a feature for the comparison of the trajectories. The G was introduced by Ernst *et al.* [38] to check the Gaussian statistics on increments within the trajectory. It is defined as

$$g(n) = \frac{\langle r_n^4 \rangle}{2\langle r_n^2 \rangle^2} \quad , \tag{4.20}$$

with

$$\langle r_n^4 \rangle = \frac{1}{N-n}\sum_{i=1}^{N-n} |x_{i+n} - x_i|^4 \quad .$$

In table 4.1 the individual features are listed again and provided with literature references.

feature	references
MSD	([80], [97], [135], [162], [192], [204], [207])
MSDR	([97], [192], [204])
Alpha	([97], [192])
FD	([8], [93], [97], [166], [192])
RG	([8], [157], [179])
E	([97], [192])
S	([97], [192])
G	([38], [97], [192])

Table 4.1.: Table of features used to characterize the trajectories of the various segregation mechanisms. The input vectors for the classifiers are composed of these features.

4.4. Results

4.4.1. Hyperparameter tuning

The first task to successfully classify the different classes of trajectories was to fine tune the ML classifiers. This process is called hyperparameter tuning. In this context, a hyperparameter of a ML model is understood as a parameter that must be set before the actual learning process begins. The hyperparameter of a ML model define the model architecture. Their choice is crucial for the learning success of an ML model. Typically, it is not clear a priori what the optimal hyperparameters are for a given problem. Therefore, one uses an automatic search in the space of possible hyperparamters to find the best possible setting. After finding the ideal parameters for a model, the model can be trained with this architecture and the actual classifications can be performed. In this work, the workflow depicted in figure 4.5 was chosen for this task.
A central point in ML applications is that one wants to know the accuracy of a created model based on classifications of unknown data. Therefore, the original data (24,000 trajectories with 3,000 trajectories per class) were divided into training and test data in a ratio of 70% training data to 30% test data. The test data was put off to the side and it was pretended to never had been seen before. In the following, both the optimization of the hyperparameters for the models and the training of the same were carried out exclusively on the train set. The first step was to fine-tune the hyperparameters. To find the optimal parameters for the models the `RandomizedSearchCV` function from the `scikit-learn` library was used to perform a random search over possible hyperparameter settings. Thereby the goal was to find hyperparameters yielding the best possible accuracy and avoiding overfitting at the same time. Thus, the model should perform well not only on training data but also on unseen data. The typical technique to avoid overfitting is cross validation. Here, the training data is split into N subsets (called folds) and the model trained N times, whereby each time one only trains on $N-1$ of the folds and the Nth one is used for evaluation. Finally, the accuracy over all folds is averaged and final validation metrics are obtained. In the `RandomizedSearchCV` procedure many iterations of the cross validation process are performed with different model settings. The best model is chosen at the end. Here, three rounds of cross-validation on a set of 50 settings for each classifier were used. The advantages of such an approach are that the model for optimization is both trained and evaluated on the complete available data. In addition, the performance

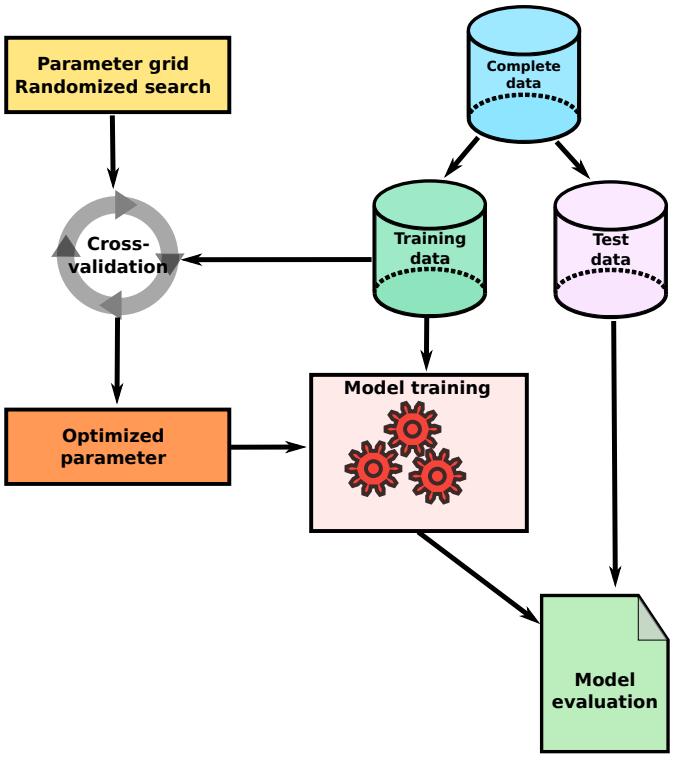

Figure 4.5.: Schematic depiction of hyperparameter tuning procedure. The synthetic data is split into training and test data. The hyperparameter tuning is performed on the training data using the `RandomizedSearchCV` function of the `scikit-learn` library. Thereafter, the optimal architecture is applied for the model which is trained on the training data again. The final model evaluation is performed on the separate set of test data which has never been seen by the model before.

of the model is thus less dependent on which particular set of data was used as the test set. Another important point for cross-validation is that the data is balanced (= same number of trajectories in each class), which was met here.
The hyperparameters to be optimized are different depending on the ML model used. The two ensemble methods based on decision trees have the same hyperparameters. Here, the following hyperparameters have been optimized: The number of trees in the ensemble, the maximum depth of a tree (i.e. the maximum number of levels in each decision tree), the minimal number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, the number of features to consider when looking for the best split (one typically chooses either the logarithm or the square root of the number of features) and a paramter to decide whether to use bootstrap samples when building trees or not (this parameter is only used for the RF classifier) ([5], [76], [80]).
We optimized each classifier two times: One time for the dataset of the complete trajectories with normalization and a second time on the dataset of the statistical features. In table 4.2 the parameters of our optimized tree-based classifiers for the complete trajectories are shown.

parameter	Random forest	Gradient boosting
Number of trees	800	700
Maximum depth of a single tree	20	10
Min. number of samples required to split an internal node	4	4
Min. number of samples required to be at a leaf node	10	6
Max. features	58	58
Bootstrap	True	-

Table 4.2.: Optimal hyperparameters for the tree-based classifiers trained on the complete trajectories. 50 different settings were tested wit 3 rounds of cross validation.

The optimized parameters for the feature dataset are shown in table 4.3

parameter	Random forest	Gradient boosting
Number of trees	700	700
Maximum depth of a single tree	90	70
Min. number of samples required to split an internal node	4	10
Min. number of samples required to be at a leaf node	12	4
Max. features	3	3
Bootstrap	True	-

Table 4.3.: Optimal hyperparameters for the tree-based classifiers trained on the feature dataset. 50 different settings were tested wit 3 rounds of cross validation.

To find the optimal hyperparameters for the two linear models, we used the `GridSearchCV` library implemented in `scikit-learn`. Here, one performs a search over a grid of parameter values. This made sense in the case of linear classifiers since they have fewer hyperparameters to optimize. The paramter that we optimized for both classifiers is the parameter C. It defines the strength of the regularization with a high C resulting in less regularization (= trying to fit the training data as best as possible) while a low value of C increases the generalization of the model.
For the SVM we also made the choice to use a linear kernel for the classifier. The reason for this is that we want to use the coefficients of the fitted classifier to compute feature importance. This is only possible with a linear kernel since here the fitted hyperplane and the coefficients are in the same dimensional space as the input vector of our features.
For the logistic regression classifier, one has to select the penalty function for the optimizer. Here, the l_2 penalty was chosen.

4.4.2. Classifiaction of rescaled trajectories

In total, the MD simulations created a dataset of 24,000 trajectories that were evenly distributed among the eight different cell types. In each case, the trajectory of the duplicated ori was tracked since this is the ori which is pulled to the opposite cell pole by the ParAB system. Thus, the largest changes were expected here by switching the ParAB system on and off. To get a first impression of the resulting data, example trajectories are shown in figure 4.6.

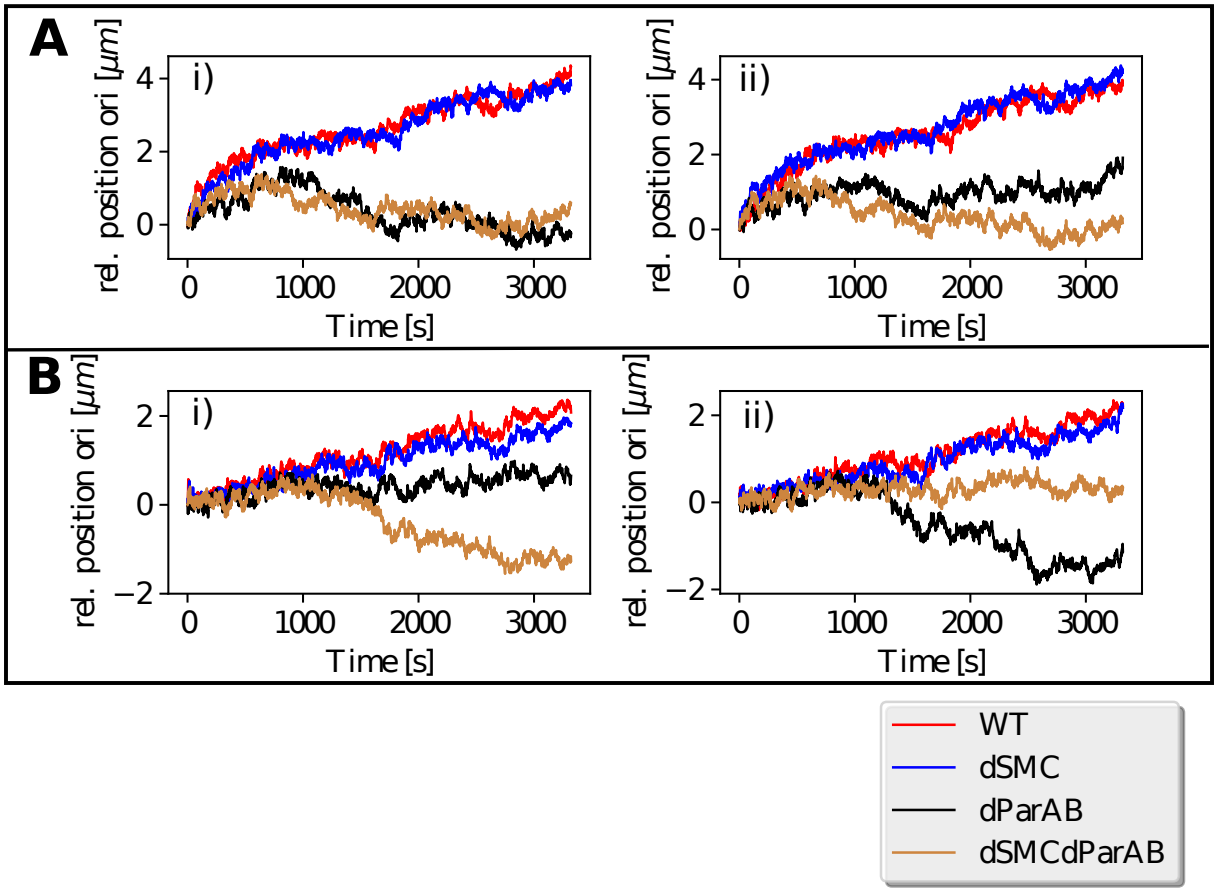

Figure 4.6.: Example trajectories of the duplicated ori as obtained by the MD simulations. The relative movement of the ori from its start position in the respective simulation are shown as a function of time. For every cell type two examples are shown in the plots on the left and right. **A:** The upper two plots show example trajectories of the four segregation mechanisms and the track model of segregation. **B:** The lower two plots show example trajectories of the four segregation mechanisms with the factory model of segregation.

One can see from the plots of figure 4.6 that the ParAB system acted as a strong segregation motor. In the cell types with activated ParAB (WT and dSMC) the ori moved in a directed way towards the cell pole. On the other hand, the cell types without ParAB show significantly higher variances and less directional motion of the ori. The different starting point of the ori in the track model (at the old cell pole) compared to the factory model (cell center) accounted for the fact that the ori in the track model traveled a longer distance to the opposite cell pole than in the factory model. At the same time, trajectories in the track model without ParAB become comparable to trajectories with ParAB in the

factory model in terms of the distance traveled by the ori. Another important fact to notice from the example plots of figure 4.6 is that in the cell types without ParAB the ori has no clear preference for a specific cell pole. Consequently, the simulations support the conjecture that the ParAB system makes an important contribution to organization of DNA in the cell by directing the duplicated ori to the new cell pole.
In addition to the influence of the ParAB system on the movement of the ori through the cell, the simulations also show a strong influence on the overall segregation of chromosomes. For this purpose, the degree of separation was defined as the ratio of the longitudinal overlap of the two chromosomes in the cell by the longitudinal extent of the shorter chromosome in the cell. In table C.1 it can be seen that the action of the ParAB system results in a much more effective separation of chromosomes. Furthermore, the histograms of the achieved degrees of separation in figure C.4 and C.5 after finished replication show that ParAB driven segregation also shows a significantly reduced variance in achieving the degrees of separation.
For the classification of cell types, at first the approach of a high-dimensional input vector from the normalized trajectories according to Muñoz *et al.* [135] was followed. Figure 4.7 shows the averaged courses of the oris in the different cell types over all trajectories. In addition, the averaged values for the rescaled trajectories are shown alongside for comparison.

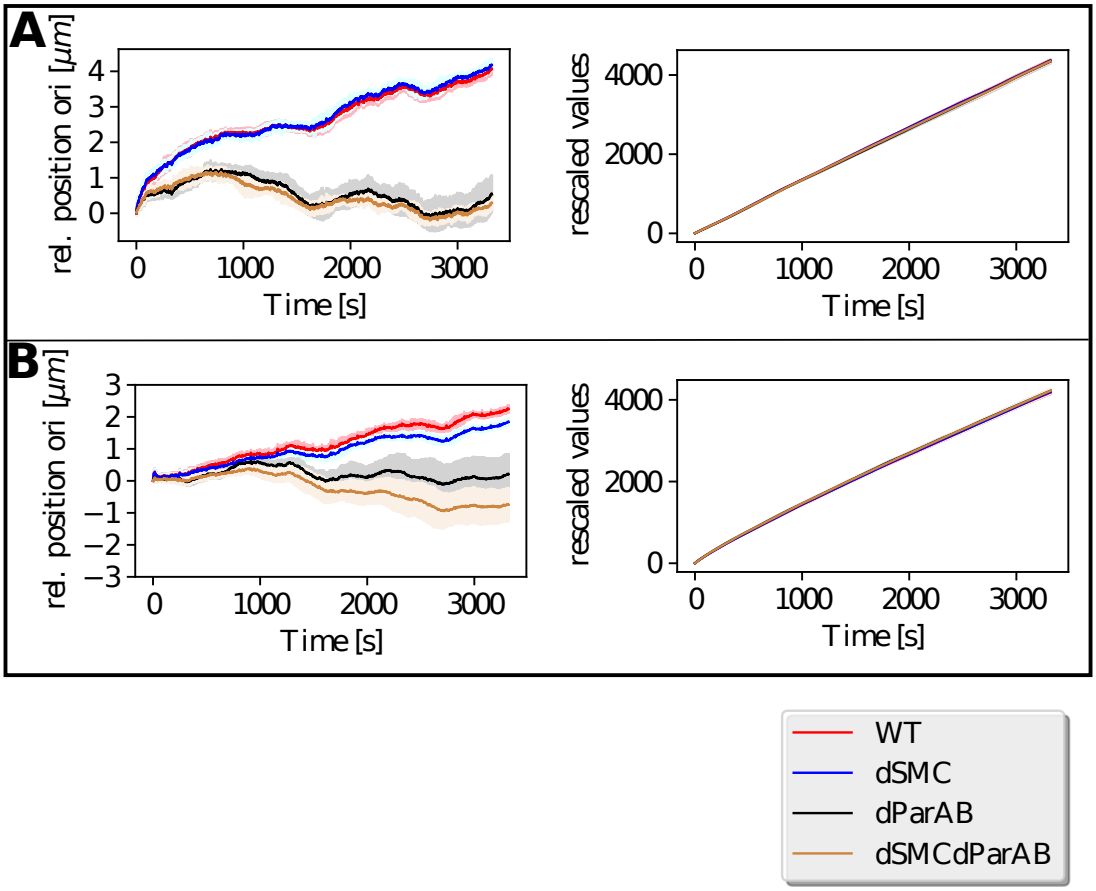

Figure 4.7.: Average trajectories of the ori in the different cell types. On the left side the average trajectories from the raw data are shown while on the right the average values of the rescaled trajectories are compared. **A:** Results for the track model of replication. **B:** Results for the factory model of replication.

Comparison of the raw data with the rescaled data showed that the preprocessing protocol indeed allowed comparison of the different spatial scales. At the same time, it was noted that differentiation of the cell types in the rescaled trajectories was impossible by eye. For the automatic classification of the trajectories, the classifiers were now trained on the training dataset and the overall prediction accuracies of the classifiers on the training and test set were evaluated.
The prediction accuracy of the classifiers is defined as the number of correct predictions divided by the total number of predictions. It is one of the basic measures to assess classification performance ([5], [67]). In table 4.4 the overall accuracies of our classifiers on both the train and the test data are shown.

Model	Acurracy (train set)	Acurracy (test set)
Random forest	0.994	0.915
Gradient Boosting	1.0	0.965
Logistic regression	0.950	0.932
SVM	0.949	0.925

Table 4.4.: Overall prediction accuracies of the classifiers on the data using high-dimensional input vectors.

The results of table 4.7 show excellent prediction accuracies for all classifiers. However, we can see that the RF classifier has a gap of 7.9% in prediction accuracy between the training and test set. This finding indicates overfitting, i.e. the classifier seems to have difficulties in generalizing from the training data. In contrast, the linear classifiers perform almost equally well on the test and on the training data.
For a better understanding of the performance of the classifiers, in fig. 4.8 the confusion matrices for each classifier on the test set are shown. The confusion matrix directly compares the predictions of the classifier with the actual labels of the data. Thereby, identification and visualization of the number of true and wrong predictions of the classifier is possible. With this information one gets a more detailed understanding of the strengths and weaknesses of the classifier. Furthermore, two additional measures can be calculated from the values of a confusion matrix: the precision value that gives us the fraction of correct predictions among all predictions of the selected class and the recall value that gives the fraction of correct predictions of a given class relative to the total number of members of this class. Thus, the precision score quantifies how often a classifier is correct if it predicts a certain class, while the recall value quantifies how often a class is predicted correctly. In tables 4.6 and 4.5 the values for all classifiers are shown.
The confusion matrices in figure 4.8 show that none of the classifiers confuses the two replication models. Among the cell types with the track model, most classification errors occur due to confusion of WT cells with dSMC cells. Within the factory model, the most common mistake is made in discriminating dParAB and dSMCdParAB cells. A further result is that the classifiers don't have difficulties in discriminating segregation with and without ParAB as we expected from the raw data, in which the strong influence of the ParAB system on segregation was already apparent.
We can further analyze the classification results with the precision and recall values of tables 4.6 and 4.5.
The analysis of tables 4.6 and 4.5 reveals very high precision and recall for all classifiers. For cells with the track model of replication, we find that cell types lacking ParAB both show highest precision and recall scores. Several conclusions can be drawn from this: First,

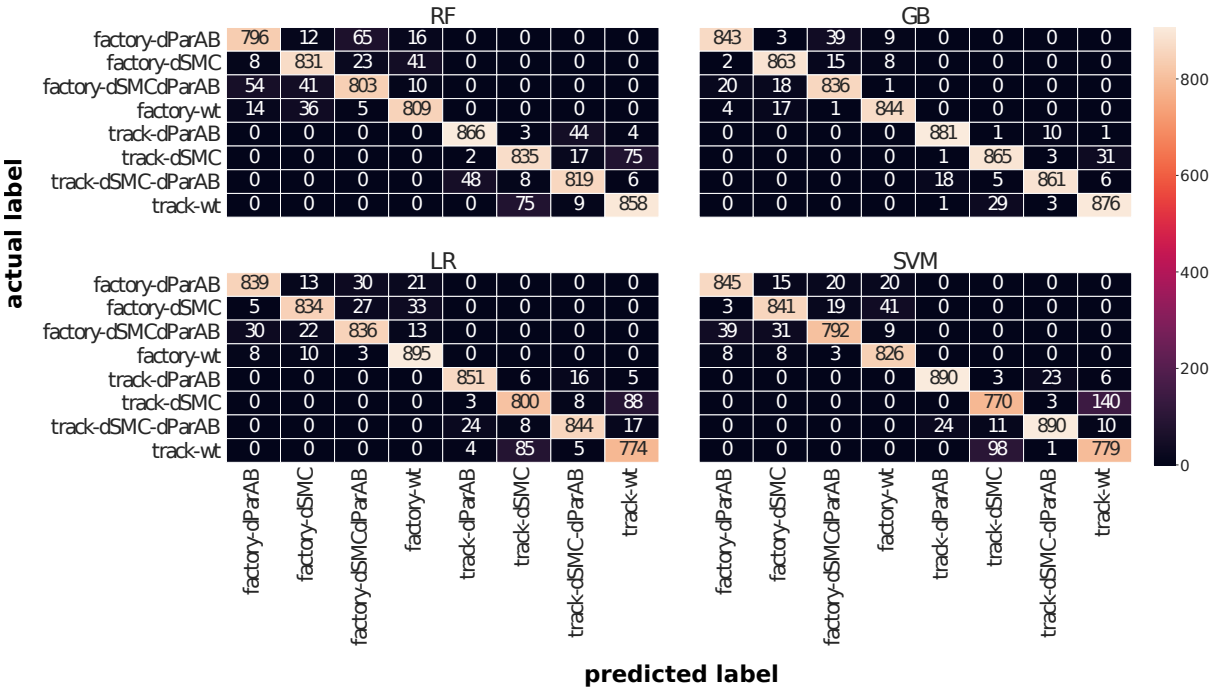

Figure 4.8.: Confusion matrices for the classifiers, comparing predicted labels with actual labels of the test data. On the horizontal axis the predicted labels are shown and compared with the actual labels on the vertical axis. Top left: Random forest. Top right: Gradient boosting. Bottom left: Logistic regression. Bottom right: SVM.

(a) Random forest

Cell	Precision	Recall
track-WT	0.910	0.911
track-dSMC	0.907	0.899
track-dParAB	0.950	0.944
track-dSMC-dParAB	0.921	0.930
factory-WT	0.924	0.936
factory-dSMC	0.903	0.920
factory-dParAB	0.913	0.895
factory-dSMC-dParAB	0.896	0.884

(b) Gradient boosting

Cell	Precision	Recall
track-WT	0.958	0.964
track-dSMC	0.961	0.961
track-dParAB	0.978	0.987
track-dSMC-dParAB	0.982	0.967
factory-WT	0.979	0.975
factory-dSMC	0.958	0.972
factory-dParAB	0.97	0.943
factory-dSMC-dParAB	0.938	0.955

Table 4.5.: Prediction and recall values for tree-based classifiers using high-dimensional input vectors.

the deactivation of ParAB is obviously clear to the classifiers, so that no confusions with cell types in which ParAB is active occur. Second, the cell types in which ParAB is active are more similar to each other than those in which ParAB has been deactivated. Therefore, the precision for the latter is greater than for the former.
In contrast, for cells with the factory model of replication both the precision and recall values tend to be higher for cells where ParAB is active. Thus, it appears that the opposite is true here, namely that the cell types in which ParAB is not active are more frequently confused with each other than those in which ParAB is active.

(a) Logistic regression)

Cell	Precision	Recall
track-WT	0.876	0.892
track-dSMC	0.890	0.890
track-dParAB	0.965	0.969
track-dSMC-dParAB	0.967	0.945
factory-WT	0.930	0.977
factory-dSMC	0.949	0.928
factory-dParAB	0.951	0.929
factory-dSMC-dParAB	0.933	0.928

(b) SVM

Cell	Precision	Recall
track-WT	0.833155	0.887244
track-dSMC	0.873016	0.843373
track-dParAB	0.973742	0.965293
track-dSMC-dParAB	0.970556	0.951872
factory-WT	0.921875	0.977515
factory-dSMC	0.939665	0.93031
factory-dParAB	0.944134	0.938889
factory-dSMC-dParAB	0.94964	0.9093

Table 4.6.: Prediction and recall values for linear classifiers using high-dimensional input vectors.

It can be concluded that in the track model of replication, in which the ori was transported over a long distance from one pole to the other, the effect of ParAB is particularly dominant and cells in which ParAB is active are therefore particularly similar. In contrast, the two cell types in which ParAB is active, WT and dSMC, are less likely to be confused by classifiers in the factory model of replication. The reason for this could be that in the factory model, the deactivation of SMC is more important because replication takes place in the middle of the cell, where it could be particularly important that SMC topologically divides the separating daughter chromosomes.
Together with the results from the confusion matrices we can state that the two most common errors are the confusion of WT with dSMC and dParAB with dSMCdParAB. It depends on the replication model which of the two errors is the more frequent. At the same time, there is a clear division into cell types with active ParAB and cell types in which ParAB is deactivated, which the classifiers can reliably distinguish.

4.4.3. Feature based classification approach

An alternative to using a high-dimensional input vector for the ML algorithms is to perform dimension reduction by combining selected features of the trajectories into a low-dimensional input vector. For this purpose, in a second approach, the statistical features of the trajectories described in the section 4.3.2 were calculated and an input vector was formed from these. This was to investigate whether the features could help deepen the understanding of both the classifiers and the data. In addition, it is hoped that the use of selected features will reduce overfitting by reducing the likelihood of fitting aspects of the data that cannot be generalized outside of the training data [80]. Furthermore, a reduced number of features makes the fitting procedure simpler and predictions faster. Therefore, one often aims to reduce the number of features by the use of feature selection analyses to identify the least important features which might be omitted.
Using the feature presented in section 4.3.2 resulted in the overall prediction accuracies of the classifiers shown in table 4.7.
The overall prediction accuracies of table 4.7 show increased accuracies for both the train and test set for the tree-based classifiers compared with the values found for the complete normalized trajectories in table 4.4. Another improvement is that the RF classifier shows a prediction accuracy of 97.4% on the training data and 96% on the test data using the features as input vector. Thus, the gap in the prediction accuracies between train and

Model	Acurracy (train set)	Acurracy (test set)
Random forest	0.974	0.960
Gradient Boosting	1.0	0.973
Logistic regression	0.889	0.860
SVM	0.891	0.872

Table 4.7.: Overall prediction accuracies of the classifiers on the test data using the statistical features as input.

test set drops from 7.9% in table 4.4 to 1.4% in the feature approach. We also note a reduced gap in prediction accuracies for the GB classifier on the two data sets. Thus, for both tree-based classifiers, we note both increased overall prediction accuracy and reduced overfitting.
At the same time, however, we note that the linear classifiers perform worse on the feature-based data compared with the analysis of the complete trajectories. Both linear classifiers show a clear drop in prediction accuracy.
We can investigate the causes of these changes using the confusion matrices in figure 4.9.

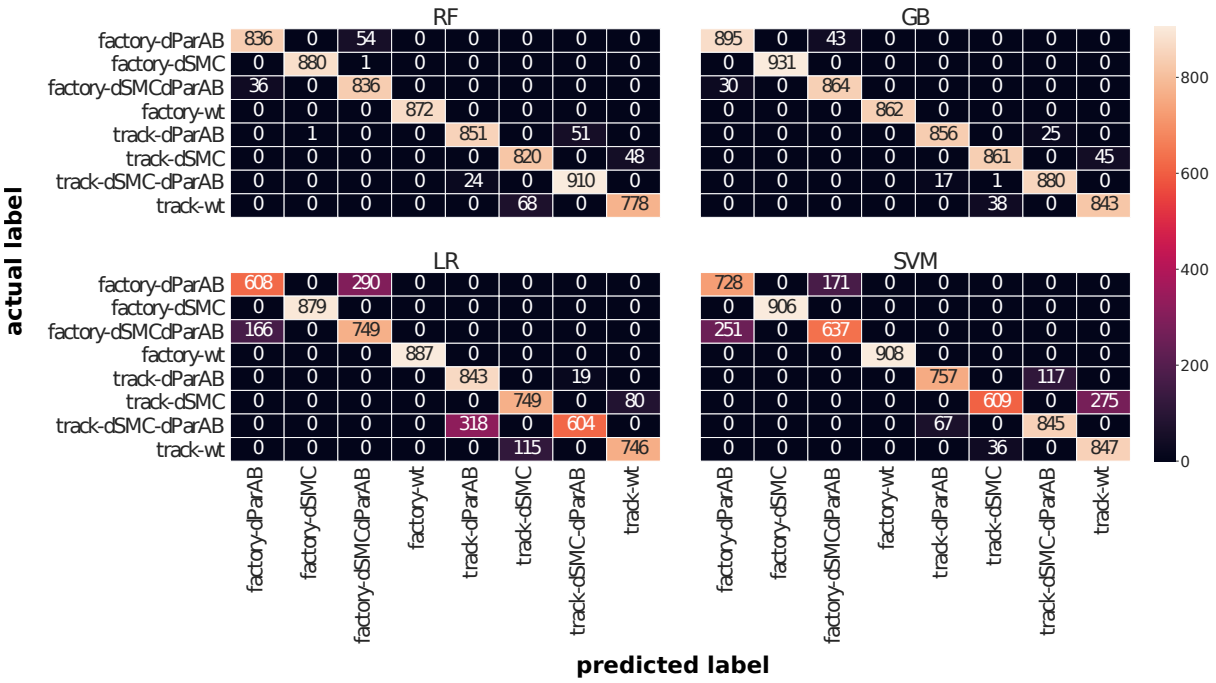

Figure 4.9.: Confusion matrices for the classifiers using statistical features of the trajectories as input vector. On the horizontal axis the predicted labels are shown and compared with the actual labels on the vertical axis. Top left: Random forest classifier. Top right: Gradient boosting classifier. Bottom left: Logistic regression classifier. Bottom right: SVM.

The confusion matrices of figure 4.9 show that the classifiers still clearly recognize whether ParAB is active or not. Consequently, misclassifications arise only between two cell types in which either ParAB is active or inactive.
The increased overall prediction accuracy of the tree-based classifiers is due to the fact that for the factory model, they only mix up the two cell types dParAB and dSMCdParAB.

The two cell types WT and dSMC are no longer incorrectly predicted by the tree-based classifiers in the factory model.
Also, the linear classifiers no longer make errors for the cells in which ParAB is active and replication is executed in the form of the factory model. However, since the error rate for cell types without ParAB increases significantly, the overall prediction accuracies decrease. In tables 4.8 and 4.9 the calculated precision and recall values are listed.

(a) Random forest

Cell	Precision	Recall
track-WT	0.942	0.92
track-dSMC	0.923	0.945
track-dParAB	0.973	0.942
track-dSMC-dParAB	0.947	0.974
factory-WT	1	1
factory-dSMC	0.999	0.999
factory-dParAB	0.959	0.939
factory-dSMC-dParAB	0.938	0.959

(b) Gradient boosting

Cell	Precision	Recall
track-WT	0.949	0.957
track-dSMC	0.957	0.95
track-dParAB	0.981	0.972
track-dSMC-dParAB	0.972	0.98
factory-WT	1	1
factory-dSMC	1	1
factory-dParAB	0.968	0.954
factory-dSMC-dParAB	0.953	0.966

Table 4.8.: Precision and recall values for tree-based classifiers using statistical features as input data.

(a) Logistic regression

Cell	Precision	Recall
track-WT	0.903	0.866
track-dSMC	0.867	0.903
track-dParAB	0.726	0.978
track-dSMC-dParAB	0.97	0.655
factory-WT	1	1
factory-dSMC	1	1
factory-dParAB	0.786	0.677
factory-dSMC-dParAB	0.721	0.819

(b) SVM

Cell	Precision	Recall
track-WT	0.755	0.959
track-dSMC	0.944	0.689
track-dParAB	0.919	0.866
track-dSMC-dParAB	0.878	0.927
factory-WT	1	1
factory-dSMC	1	1
factory-dParAB	0.744	0.81
factory-dSMC-dParAB	0.788	0.717

Table 4.9.: Precision and recall values for linear classifiers using statistical features as input data.

For the precision and recall values in tables 4.8 and 4.9, the behavior already discussed in the previous section for tables 4.5 and 4.6 is confirmed: Depending on the replication model used in the cell types, either the cells in which ParAB is active (if track model) or those in which ParAB is inactive (if factory model) are more often confused by the classifiers. This separation is even more evident in the feature-based approach, where the classifiers have perfect precision and recall values for the cells with active ParAB and replication in the form of the factory model. This could be due to the fact that in the factory model of replication, chromosome density in the middle of the cell is very high, especially at the start of replication. This could be a reason for why the effect of SMC, which ensures a juxtaposition of the chromosome arms, produces a particularly pronounced effect in the factory model of replication and facilitates the separation of the chromosomes. This effect could in turn be detected by the classifiers and be the reason for the improved accuracy.

Furthermore, the precision and recall values show that the tree-based classifiers further improve their performance in the case of low-dimensional input vectors. Both classifiers show very good scores for all cell types with values always above 90% and mostly even larger than 95%. However, this is not the case for the linear classifiers. Thus, the high-dimensional approach of normalized trajectories according to Muñoz *et al.* [135] is more recommended for the linear classifiers, while the tree-based classifiers achieve better results with the low-dimensional approach of the selected statistical feature.
Another interesting question that follows is which features the classifiers considered most important for classification. By identifying these features, it could be possible to perform a feature selection, which would further reduce the computational effort. For the analysis of feature importance for the tree-based classifiers the `scikit-learn` library offers a build-in function. It is based on a method proposed by Breiman [19] where the total decrease in node impurity caused by a given feature is calculated. To do this, the Gini impurities (see equation 4.10) are calculated before and after each split on a given feature and the total decrease in the impurity related to the respective feature is calculated. The outcome is finally averaged over all trees in the ensemble [80].
For the linear classifiers, feature importance values can be derived from the coefficients of the feature in the decision function. Here it is assumed that the higher the coefficient of a feature, the higher its importance. For a SVM classifier, however, one should do this only in the case of a linear kernel. For other kernel functions the size of the coefficients cannot be used to infer feature importance because the data is transformed into another space by the kernel and thus coefficients from the higher-dimensional space cannot be related to the input space. It is also important to scale the data for the fit so that the coefficients of the individual features are comparable. This was done with the `StandardScalar` function implemented in the `scikit-learn` library, which removes the mean and scales to unit variance.
Figure 4.10 shows the results of the feature importance analysis for the four classifiers.

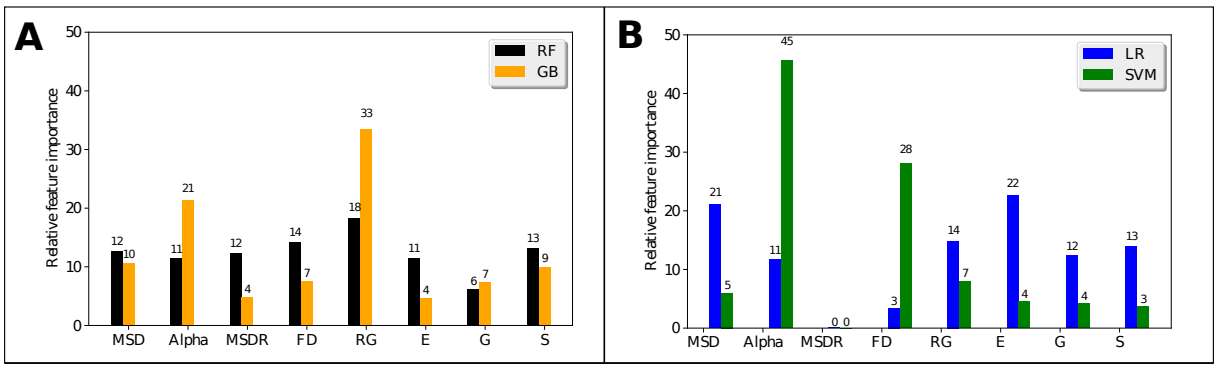

Figure 4.10.: **A**: Bar chart of the relative importance of the features for the predictions of the tree-based classifiers. **B**: Bar chart of the relative importance of the features for the predictions of the linear classifiers.

The tree-based classifiers show very similar values for the importance of the individual features. The four features that are considered most important are RG, Alpha, FD and MSD. However, the even distribution of the feature importance values indicates that it makes sense to provide all features as input to the classifiers.
In comparison with the feature importance values from figure 4.10B, it is noticeable that SVM assigns a significantly higher importance to Alpha and that FD as the second most important feature also stands out clearly from the other features. In contrast, the LR

classifier distributes feature importance more like the tree-based classifiers. However, both linear models agree on not needing the MSDR for classification.
To get an idea of what the classifiers can extract from the features, we can look at example scatterplots of the values of two features against each other. This is done in figure 4.11.

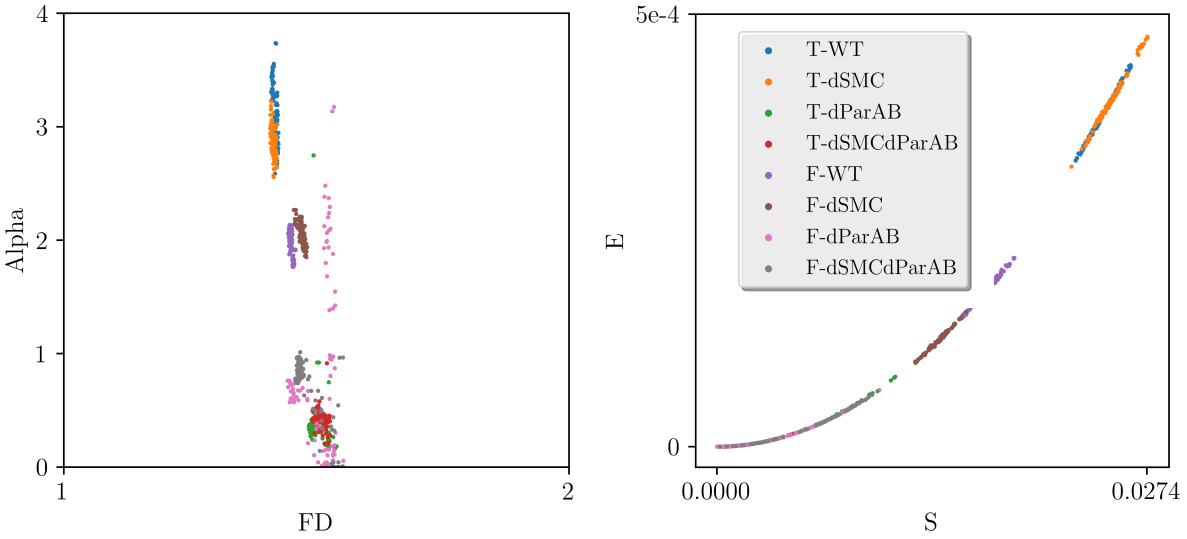

Figure 4.11.: Scatterplots of the feature values of 100 segregation trajectories per cell type. Left: Values of the exponent alpha for the trajectories plotted versus the values of the fractal dimension. Right: Gaussianity values plotted versus values for the mean-squared displacement ratio.

In the first plot of figure 4.11 the exponent alpha calculated for the trajectories is plotted versus their fractal dimension. For the exponent alpha one expects $\alpha \approx 1$ for normal distribution, $\alpha > 1$ for directed motion and $\alpha < 1$ for anomaleous diffusion ([97], [192]). One can see in the plot of figure 4.11 that the cell types with active ParAB result in segregation trajectories for which $\alpha > 1$ is calculated. Thus, the partitioning system ParAB ensures directed diffusion of the ori to the cell pole. In contrast, cell types without ParAB mostly show values of $1 \geq \alpha \geq 0$. Here, the ori performs normal diffusion or is in the subdiffusive regime. Consequently, we can see from these well-interpretable results why the exponent alpha is considered an important feature by the classifiers as it can be used to distinguish well between cell types with and without ParAB. The fractal dimension is a measure for the space-filling capacity of a trajectory. One expects values around 1 for straight trajectories and values around 2 for random trajectories [93]. In the plot of figure 4.11 we again find that the trajectories belonging to cell types with activated ParAB show lower values for the fractal dimension than the ones belonging to cell types in which ParAB is disabled. This can be interpreted by the fact that the more directed motion due to the effect of ParAB causes the trajectories to be more similar to a straight line than they are without the action of ParAB. Since the motion of the ori is still characterized by a thermal diffusion in all simulations, values of the fractal dimension of about $FD \approx 1.3 - 1.4$ are also obtained for the more directed trajectories. Nevertheless, it can be seen that the features Alpha and FD already allow a rough clustering of the cell types.
In the second plot of figure 4.11 the values of the efficiency and the straightness for the trajecotories are plotted against each other as a second example. The efficiency is a measure for the linearity of a trajectory by relating the square net displacement to the sum

of the squared displacements. Very similar, the straightness relates the net displacement to the sum of the step lengths ([97], [192]). Thus, it is not surprising to detect a quadratic dependence of the two values with each other. It can be seen that, due to the thermal fluctuation that underlies all trajectories, both values are close to 0 in each case. However, it can be observed that the effect of ParAB again is visible producing higher values for both efficiency and straightness due to the more directed motion of the ori in this cell types. However, the plot of efficiency versus straightness shows an effect that we have to keep in mind when considering feature importance values, especially for linear classifiers: Some of the statistical features correlate with each other, which makes it difficult to determine the weights. The reason for this is that in linear models the individual effects are added together, so that it eventually becomes indeterminable to which of two correlating features a particular effect is to be assigned.
One way to reduce the number of features is to look at the cumulative importance of the feature. For this purpose, in figure 4.12A the features were sorted according to their importance and the cumulative importance was plotted as a function of the number of most important features. One could use this graph as a tool for feature selection, i.e. define a threshold of accuracy to be reached and omit the remaining features which are not needed. Furthermore, the prediction accuracies of the classifiers were analyzed after they were trained with a reduced number of features. The results for this can be found in figure 4.12B.

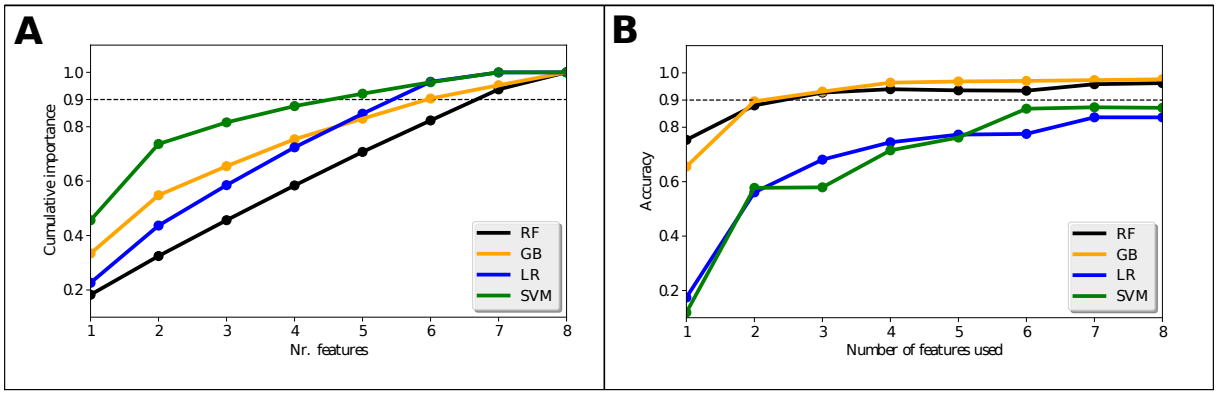

Figure 4.12.: **A**: Number of features required to reach a defined level of cumulative importance. The dashed line marks the threshold of 90 % cumulative importance. **B**: Accuracy of classifiers trained with reduced numbers of features. When reducing the number of features, those with the least importance were omitted in each step.

In figure 4.12A one can see that the SVM classifier already reaches a cumulative importance of 90% with the four most important features. At the same time, however, one can see from figure 4.12B that the prediction accuracy with these four most important features is still below 80%. In contrast, the tree-based classifiers reach a cumulative importance of 90% only after six and seven features, respectively. Nevertheless, even with the two most important features, they achieve a prediction accuracy of over 90%. This shows that the tree-based classifiers not only perform better overall in the feature-based approach than the linear models, but also achieve very good prediction accuracy with fewer features.

4.4.4. Classification of short trajectories

A final test for the classification capabilities of the classifiers, which is also important for a possible application on experimental data, is the classification of short trajectories. The practical background for this is that in real experiments it may not be possible to produce high resolution time-lapse data over 50-60min in every case. This raises the question of whether the classifiers are also capable of classifying the different cell types on the basis of significantly shorter recording periods. Another challenge from experimental data could be that different trajectories were recorded with different temporal resolution. Thus, one would need a protocol that produces comparable input vectors from these trajectories with a different number of measurement points and allows a classification. This case is also highlighted below.
Muñoz *et al.* have already demonstrated in their paper on the classification of different diffusion types that the normalization procedure of the trajectories allows a very good classification by a random forest classifier even for extremely short trajectories [135]. Therefore, this protocol was also used in the present work and the classifiers were trained with corresponding normalized trajectories of different lengths of down to 5s. The results of this analysis are depicted in figure 4.13.

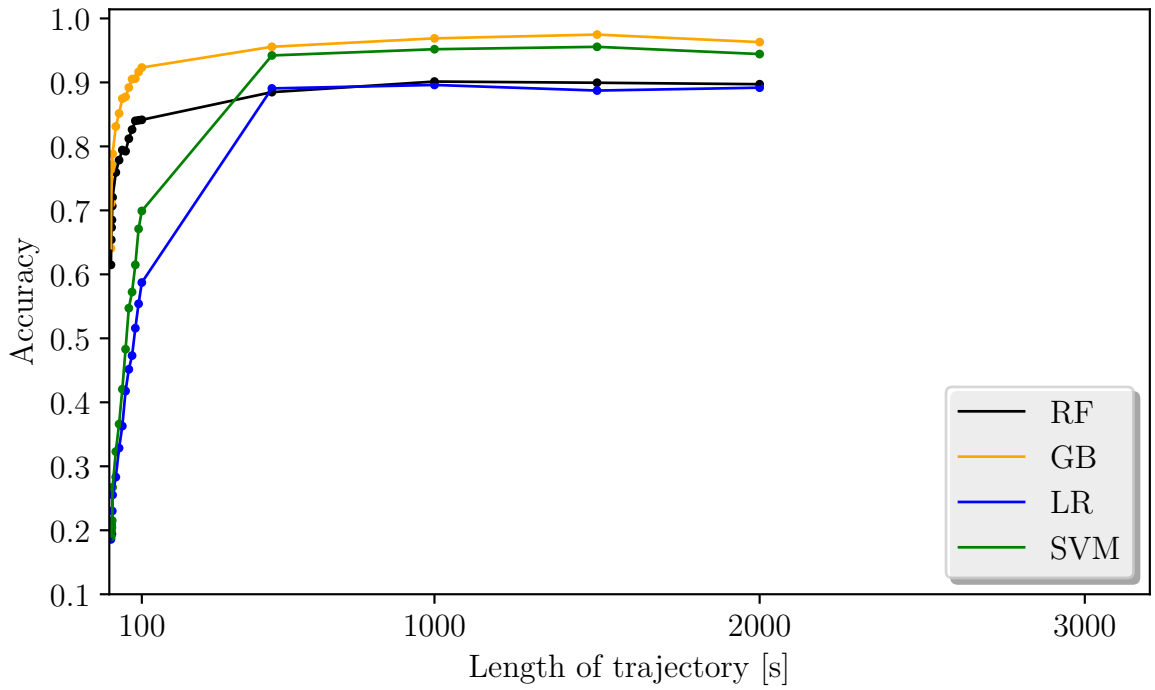

Figure 4.13.: Accuracy of the classifiers on trajectories of reduced length. The classifiers were trained and evaluated separatly on each length.

It can be seen from the plots of figure 4.13 that the classifiers reached very good prediction accuracies for very short trajectories up to a length of 10s. After $\sim$ 1000s all classifiers reached accuracies comparable to their scores on the complete trajectories. Furthermore we note, that the tree-based classifiers already reached prediction accuracies for trajectories shorter than 10s. In this area in particular, they clearly surpass not only the linear models but also the performance of the human eye, for which no classification of cell types is yet possible after such a short time. Thus, these results confirm the success of

the normalization protocol of Muñoz *et al.* also for trajectories of segregating oris and additional classifiers.
To test the case of a data set consisting of trajectories of different temporal resolution, it first had to be artificially created. For this purpose, the previously used data set could simply be used, but a new temporal resolution was randomly chosen for each trajectory. According to this new temporal resolution, only every xth point (for a temporal resolution of x seconds) of the trajectory was then retained. This is exemplified in figure 4.14.

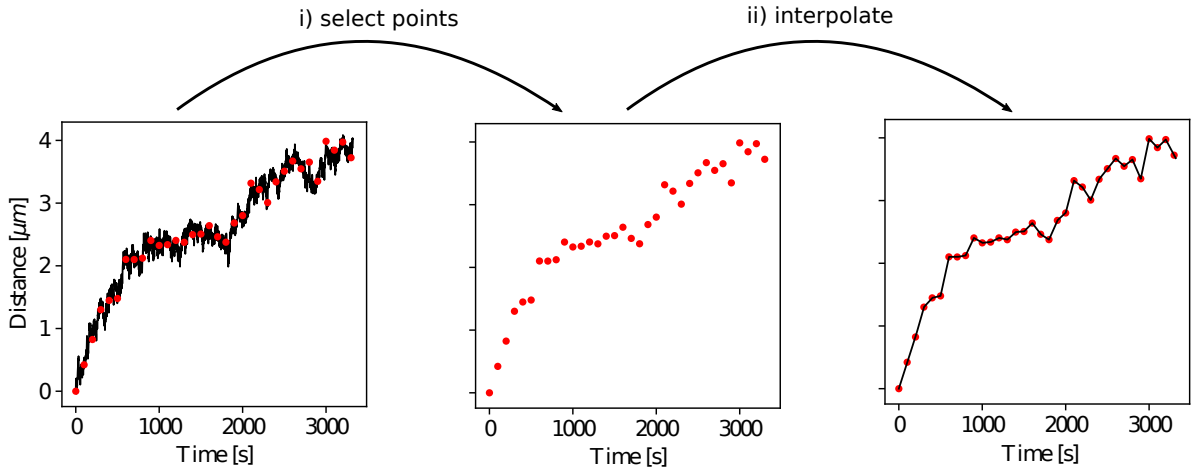

Figure 4.14.: Construction of a different temporal resolution for an example trajectory. In this example, every 100th point of the original trajectory is selected to mimic a temporal resolution of 100s instead of the original 1s. In the subsequent step, the resulting "experimental" trajectory with a temporal resolution of 100s is interpolated in order to compare it to trajectories of different temporal resolutions.

In the second preprocessing step shown in figure 4.14 the selected points of the new temporal resolution of the trajectory are interpolated again. This procedure allows comparison of trajectories from different temporal resolutions since we can harmonize them by the interpolation. This is necessary to construct input vectors of the same dimension for all trajectories. For the new dataset we used one of the following temporal resolutions per random for each trajectory: [1s, 2s, 3s, 4s, 5s, 10s, 15s, 20s, 30s, 40s, 50s, 100s].
After the new data set was produced, training and test sets were again created in a ratio of 70% (training) to 30% (test) and the classifiers were trained on the training set and finally the prediction accuracy was evaluated on both sets. In addition, the two approaches of a high-dimensional input vector from normalized trajectories and a low-dimensional input vector from the statistical features were again compared. The results of the prediciton accuracies are sumamrized in table 4.10.
Looking at the results from table 4.10 one finds that the prediction accuracies are reduced compared to the case of a dataset with an homogeneous temporal resolution. Nevertheless, the classifiers are still able to yield suprisingly good results. Thereby, it becomes clear that the feature-based approach yields higher prediction accuracies of the classifiers than the approach of high-dimensional input vectors does. Furthermore, the tree-based classifiers outperform the linear classifiers. The tree-based classifiers are still able to predict more than 90% of the cell types correctly. Thus, by using the tree-based classifiers and the statistical features as low-dimensional input vectors it is possible to also discriminate cells

	Normalized		Features	
	% (train set)	% (test set)	% (train set)	% (test set)
Random forest	0.988	0.72	0.946	0.917
Gradient Boosting	1.0	0.812	1.0	0.932
Logistic regression	0.763	0.697	0.663	0.652
SVM	0.749	0.702	0.795	0.783

Table 4.10.: Overall prediction accuracies of the classifiers on the data set containing trajectories of different temporal resolutions. The table displays both the results obtained by preprocessing the trajectories according to [135] and by using the statistical features as input vector for the classifiers. The reached accuracies on the train set and on the test set are shown for each case.

with different segregation mechanisms if the trajectories in the dataset show a variety of temporal resolutions.

4.5. Project summary and outlook

In the first two projects described in this paper, there was a direct relationship of data obtained from physical models and computer simulations to experimental results. In the third project, these results were built upon. For this purpose, the existing MD model was further extended with additional segregation mechanisms. Thereby, it was possible to produce data of oris segregating due to various mechanisms in a quantity beyond the current possibilities of experiments. The so obtained experimental data was used to provide a proof of principle for the automated classification of segregation trajectories with ML models. Thus, the aim of this work was to transfer a tool that has already been successfully used in other fields, such as diffusion classification ([80], [97], [135], [192], [207]), to the research field of bacterial chromosome segregation. Due to the increasing availability of high-resolution data from SPT experiments, it is hoped that the positive results of this work will pave the way for classifying experimental data with ML models in the future. To this end, first steps were taken to adapt to experimental data by testing the classification of very short trajectories or by classifying data of different temporal resolutions.

Research approach Within this project the MD model of entropic chromosome segregation presented in chapter 3 was extended by the implementation of the ParAB partitioning system and SMC proteins, which are both thought to play important roles in the process of chromosome segregation in bacteria. Besides this, the two replication schemes of the track and factory model of replication were used to create a total of eight different cell types. In the simulations, the duplicated ori was tracked according to a SPT and its trajectory through the cell was recorded. Thereby, the trajectories were labeled by the corresponding cell type. With this, two tree-based classifiers and two linear classifiers were trained to discriminate the various cell types based on the trajectories of the oris. The classifiers were presented with two kinds of input vectors: In a first approach a high-dimensional input vector was used according to a procedure proposed by Muñoz *et al.* [135]. Here, the complete trajecotry is normalized and presented to the classifiers. In the second approach, a set of eight statistical features was calculated from the trajectories and used to create a low-dimensional input vector. To test the application on possible

experimental data, the classifiers were also presented with shorter length trajectories. In addition, trajectories of different temporal resolution were produced, made comparable by interpolation of the measurement points, and presented to the classifiers. Thereby, the generalization ability of the ML models was tested.

Key findings Evaluation of the different segregation mechanisms showed that especially the ParAB system is a very strong segregation driver and is able to control the precise positioning of the oris in the cell. This is particularly interesting because the previous project showed that purely entropic segregation does not result in a clear positioning of the duplicated oris in the cell. In contrast, the ParAB system showed a clear segregation direction for the ori which reliably arrived at the new cell pole. In addition, the effect of the ParAB system also affected the segregation of chromosomes as a whole, which in the case of segregation by ParAB were reliably separated at the end of the replication phase, whereas without ParAB there was greater variation in the degree of separation. In this respect, the previously missing mechanism for maintaining the organization of DNA in the cell and further increasing the efficiency of chromosome segregation was identified with the ParAB system. Comparing the replication mechanisms, slightly higher degrees of separation were found for the factory model. In addition, the results of the prediction accuracies of the classifiers suggest that the effect of SMC is more important in the factory model of replication, because there is a higher DNA density in the middle of the cell. Thus, the juxtaposition of chromosome arms by SMC is particularly helpful in chromosome segregation here.
The results of the classification of the different segregation mechanisms showed very good overall prediction accuracies and thus strengthen the hope to be able to classify experimental trajectories with these methods in the future. In the case of classifying trajectories using high-dimensional input vectors, all classifiers achieved prediction accuracies of more than 90% on the test data. Thus, the good results of Muñoz *et al.* for the classification of diffusion trajectories could also be confirmed here for the case of different segregation trajectories of ori in bacterial cells. Furthermore, the results show that the preprocessing protocol also provides good results for classifiers other than the RF classifier. Thereby, the linear classifiers showed a slightly lower tendency to overfitting compared to the tree-based models. However, the highest accuracy was achieved with the GB classifier. A more detailed analysis of the classification results shows that all classifiers have no problems identifying if ParAB was active or not in a given trajectory. Thus, missclassifications only occurred between the WT and dSMC cells or between the dParAB and dSMCdParAB cells. Which of the two errors is more frequent depends on the replication mechanism. It was found that in the factory model the confusion of WT and dSMC was less frequent than in the track model. Therefore, it could be hypothesized that SMC is of particular importance in factory model as here chromosome density is very high at the cell center and SMC helps resolving the sister chromosomes.
In a second approach, low-dimensional input vectors were presented to the classifiers. For this purpose, eight statistical features were calculated from the trajectories and an input vector was constructed from these. This approach resulted in an increased speed for the fitting procedures of the classifiers and an overall increased predication accuracy of the tree-based classifiers. Furthermore, the tendency of overfitting was reduced for the tree-based classifiers. In contrast, the linear classifiers performed worse when presented with the lower-dimensional input vector. Thus, these results suggest to use the high-dimensional input vecors if one is interested in a linear model and apply the feature-based approach

for the tree-based classifiers. The analysis of the feature importance showed that the tree-based classifiers give very similar weights to the individual features, whereas the SVM classifier in particular already covers more than two-thirds of the feature importance with only two features. However, the features were not free of multicollinearity, so that the feature importance values of the linear models should be viewed with caution. At the same time, this could explain the weaker performance of the linear models when using the features. Analysis of the features revealed that they, too, were particularly helpful for the identification of the more linear trajectories due to the ParAB system. This was reflected for example in a smaller fractal dimension or higher exponent alpha compared to other cell types. Interestingly, the tree-based classifiers already reached an overall prediction accuracy of 90% using two out of eight features.
As a final challenge we tested our classifiers on trajectories consisting of less datapoints. Here, the classifiers proofed to be capable of discriminating trajectories of only some datapoints, i.e. some seconds of length, using the high-dimensional input vectors. This is a very promising result with respect to possible applications on experimental data which might also be of smaller length. Furthermore, it could be shown that it is also possible to simultaneously classify trajectories of different temporal resolutions using interpolation of the data points. In this case, however, the approach of low-dimensional input vectors from statistical features was more successful.
The presented results illustrate that ML models are able to classify segregation trajectories in bacterial cells according to the underlying mechanisms. This opens up the possibility of using a new tool in this field. Based on the results presented in this paper, suitable model architectures and input vectors can be proposed for this purpose. Thereby, it will be possible to autmatically classify a variety of microscopy data using previously trained ML models in future applications. In addition, for new organisms an estimate of the chromosome segregation mechanisms responsible for the tracked movement of loci can be made.

Outlook Taken together the results of the project demonstrate that classification of segregation mechanisms with ML methods is a promising approach. In further studies it might be possible to add experimental trajectories to the dataset produced with the MD simulations and thereby extend the application towards experimental data.
Another interesting application is the differentiation of replication models. In this project, both the track model and the factory model of replication were implemented in the MD simulations. These two models are also the central point of controversy in many fluorescence microscopy experiments of replication in bacteria. In the experimental observations often two optically resolvable replication foci are described. However, it is not clear whether these are two individually resolvable replisomes, as postulated by the track model of replication, or whether they are the two unresolvable replisomes of some sort of replication factory corresponding to a factory model of replication ([81], [117], [152]). Interestingly, also for the factory model of replication, it is reported that the subcellular localization of the replisomes varies, i.e. they are mobile, but their movement is caused by chromosomal re-arrangements rather than by the replisomes' own movement [117]. Within the MD scheme presented in this work one could generate a heat map of the duplication polymerases of the different replication models from the simulations to compare with experimental data. In addition, it would be possible to also track the trajectories of the replisomes in the MD simulations and classify them according to the ori trajectories with the ML models. Based on this, one could perform a characterization of the experimental data.

There are also some possibilities for further development of the MD simulations. Possible extensions include a variation of the pulling force of the ParAB system. Here, one could challenge the classifiers by turning the pulling force on and off during replication or by using pulling forces of different strengths in different trajectories. Moreover, the implementation of SMC could be extended and take into account that some SMC proteins exhibit diffusive dynamics. It has been shown that such dynamics can affect the organization of the chromosome by SMC [128]. Such dynamics could also be implemented in the model presented here by allowing the SMC bonds to change along the chromosome.

5. Conclusions

Object of research Within the scope of this work, the spatio-temporal organization and segregation of bacterial chromosomes was investigated. The underlying question was how the complex structure of living cells is maintained and what models we can use to understand and describe it. We have already seen in the introduction that bacteria are a suitable object of study for such questions for several reasons. The study of bacteria has not only led to some of the most important biological and medical discoveries that have had a lasting impact on our lives, but at the same time the simpler structure of bacteria compared to eukaryotes provides an advantageous model system for studying the mechanisms of life conserved across all organisms ([4], [138], [144]).

Biological physics approach Describing the structure of the complex system of the bacterial cell offers an exciting challenge for an interdisciplinary approach from physics and biology. The aim here is to combine the respective strengths of the approaches: The biological approach provides a variety of details of the complex living world. Here, the physical view can be used to decipher general laws and simple logics, that is, to see the forest formed by the trees [138]. To follow this combined approach was one of the central tasks of the present work. The starting point for this was the experimental data, which is a slice of the rich reality of life. These had to be abstracted in the following in order to create models as simple as possible that could be formulated mathematically (we remember Galileo's request from chapter 1.3). At the same time, this models still had to be close enough to reality to function as a realistic projection of reality into the conceptual space of the questions of interest. By analyzing the models, the hope was to discover the effects of general laws, to be able to make new predictions, or to open up new perspectives [138]. In this paper, three projects were presented that used this approach to investigate different research questions. In the following, we will consider how the results fit into the claim just formulated. A more detailed summary of the results of the individual projects can be found at the end of each project.

5.1. DNA organization

The first project of this thesis addressed the question of how bacteria manage to spatially organize their genetic material in the cell. For a long time, it was assumed that the genome was randomly arranged in bacterial cells. It was impossible to imagine how bacteria could compact their chromosomes by three orders of magnitude and at the same time achieve spatial sorting in the cell. All the more astonishing were the findings of recent years, which indeed revealed a complex organization of DNA in the cell of bacteria, which is also related to important functional and regulatory processes of the cell ([9], [40], [150], [173]). These new findings have changed the view of the spatial organization of DNA in the cell and highlighted its importance. At the same time, the new findings almost exclusively concern typical model organisms, so that the organization of the genome of multipartite bacteria

remains unclear. Therefore, the aim of this work was to broaden the research horizon in this direction.
In accordance with the approach to biophysics formulated above, the experiments performed in the laboratory provided us with a large amount of data for the model organism *S. meliloti*. Fluorescence microscopy was used to determine the positions of different loci along the entire genome in hundreds of cells per loci (more detailed information on the experimental procedure will be part of the papers [134], [193], which are under preparation). Thus, the task was to construct a simple model that would allow us to understand the mechanisms underlying the data and to test corresponding predictions using mutants. For this purpose, the proven concepts of polymer physics were used. We have seen in this work that it is possible to understand effects of compactification of DNA in the cell such as supercoiling and macromolecular crowding based on simple physical models. The effect of macromolecular crowding is entropy-driven. The loss of conformational entropy of the DNA (considered as a polymer) is overcompensated by the gain in accessible volume by the large number of crowding particles ([86], [119], [146], [168], [216], [173]). Accordingly, we can trace the formation of supercoiled domains using the topological properties of circular DNA. Here it can be understood that an energetic ground state results from the natural double-helix repeat of DNA, which we can express with the concept of the linking number. Changes in this state caused by proteins such as topoisomerases are balanced in the interplay of twist and writhe, with the substitution of writhe for twist leading to a compaction of DNA ([22], [23], [121], [122], [123], [175], [176]). These considerations allow us to describe DNA as a sequence of compacted monomers according to the FJC model. However, since a Gaussian probability distribution for the expansion of DNA can be calculated for such a model, it becomes possible at this point to use one of the most powerful ideas of science [144] for the simulation of DNA configurations: random walks. Thus, with the concept of the self-avoiding walk, a model has finally been found that is simple and can be described mathematically as requested above, but is at the same time sufficient to describe the real biopolymer DNA with adequate accuracy in the particular conceptual space, which is the global organization of the DNA in the cell.
As described in the project summary 2.4, the newly obtained model of DNA was used to investigate three basic hypotheses. These can also be formulated in a broader sense. Accordingly, the results of our studies show that the spatial organization of DNA in the cell is significantly influenced by (i) the mechanical properties of DNA as a polymer, (ii) by geometric constraints of individual loci and the complete DNA and (iii) interactions of individual replicons in the cell.
We have already discussed that the mechanical properties of DNA can be captured by the concept of conformational entropy. This finding can be extended to the whole field of polymer physics, where the conformational statistical properties of macromolecules determine the entire field of physical properties of polymers [57]. Furthermore, the results of this work have shown that, in addition to the global constraint of limiting DNA to the cell interior, the spatial fixation of individual loci (such as ori and ter) is sufficient to track the mean configuration of the entire genome in the cell. This statement already contains the insight that especially a statistical description of the averaged configuration over many cells makes sense. In this case we are able to recognize recurring patterns, whereas we discover a large cell-to-cell variability both in reality and in our modeling. Finally, model-based prediction of inter-replicon interactions from the simplified conceptual framework of our model could provide the basis for future experiments in the more complex reality. It could be shown that in the sense of the model used, the effects of inter-replicon interactions

can be understood as additional spatial constraints. Since the corresponding simulation results reflect the actual organization of the genome in the cell very well, they motivate future Hi-C experiments and already indicate possible loci for which interactions are to be expected. This is possible even though the chemical complexity of such interactions is not included in the simplified model. Another noteworthy finding from the modeling approach taken here concerns the interaction of the different levels of organization of the bacterial chromosome. In our model, a description of the chromosome at the level of supercoiled domains was used. At the same time, however, the modeling results have shown that this small-scale description, in combination with the aforementioned geometric constraints, produces the global organization of the chromosome in the cell. In this sense, the hypotheses formulated above form the bridge linking the different levels of organization of the genome in the cell.
Despite the already discussed findings in the context of this project, there are still a lot of further perspectives for future research and possibilities to develop the model. Among the limitations of the model is the fact that a constant degree of compaction is assumed along the entire genome. This is certainly a very strong simplification, so that in the future a variation of the degree of compaction is a useful extension, e.g. to account for different levels of gene expression [39]. In addition, the current model does not yet include the influences of various proteins, such as SMC, which affect the configuration of the chromosome in the cell. These could also be implemented in the form of additional spatial constraints. Using a similar logic, the scope of the model could also be extended to additional questions. Particularly noteworthy is the area of chromosome-membrane interactions, which could be caused by transcription or by specific proteins that bind the chromosome to the membrane. Such effects could also be described as additional constraints within the model. In this context it would be interesting to see to which degree the repositioning of specific loci due to transcription alters the global chromosome organization, and, vice versa, if specific loci are particularly suitable for transcription as a result of the global configuration of the chromosome in the cell. Furthermore, the model offers the possibility to investigate the topology of multiple replicons in the cell which might have an important influence on the processes of replication and segregation of the genomic material. These processes were part of the investigations of the second project of this thesis.

5.2. Replication and segregation of DNA

To address the issues of replication and segregation of DNA in the cell, the task was to answer the question of what effect the above regularities have on the dynamics of multiple chromosomes in the cell. In other words: while we previously considered static organization as a consequence of conformational entropy and geometric constraints for individual chromosomes, in the second project we had to address what these properties of DNA mean for the DNA's replication and segregation.
In general, cellular proliferation depends on successful replication and segregation of DNA. In reality, cells use a complex replication machine consisting of multiple proteins to ensure a high fidelity while copying their genetic material. The initiation and basic constituents of this process are conserved in all organisms ([4], [92], [102], [117], [152]). However, many basic questions still remain to be answered. For example, it is not known for many bacteria if bidirectional replication of the chromosome is spatially confined to the center of the cell (factory model) or if the replisomes are able to move independently along the chromosome

(track model) ([81], [117], [152]). Furthermore, no unique mechanism for chromosome segregation has been identified in bacteria so far ([9], [37], [55]). Of particular interest in the context of these questions is the origin of replication (ori). Not only does bidirectional replication start here, but the ori is also important for the organization of chromosomes in the cell, as well as being crucial for the action of important proteins like SMC and ParAB ([72], [108], [198], [201], [203]). At the same time, focusing on the segregation of the two oris provides a first step of simplification to approach the complex problem of replication and segregation of bacterial DNA. Thus, within the second project of the present work the Graumann lab provided time lapse data of segregating oris in *B. subtilis* as experimental input. For the theoretical study of this process in accordance with the above considerations, a model of DNA was chosen, in which it was again described as a compacted polymer under confinement (by the cell). Other characterizing properties of the polymer, such as its connectivity, were modeled by simple physical potentials holding the monomers together. However, since there are two chromosomes in the volume confined by the cell, the mechanical properties of the chromosomes yield further conclusions for their interaction in this case. In order to increase their conformational entropy, polymers repel each other even if there is no additional mechanism involved ([7], [85], [86], [130]). This effect could be represented by an effective repulsion of the individual monomers in our model. Similarly, the complex question of the mobility of replisomes during replication was simplified in the conceptual framework of our model. Within the simulations, it was possible to control whether the duplication of new DNA is fixed within the cell according to the factory model or whether new DNA was duplicated along the existing chromosome, as in the track model. It should be noted that this implementation did not consider the question of the origin of the force required for the mobility of the replisome, but merely considered its effect. Nevertheless, this modeling allowed quantitative comparisons with the real data of the experiments (compare project summary 3.4). As a result of the comparisons of model and reality, it could be stated that the observed separation of the oris can indeed be thought of as a result of entropic segregation of the chromosomes. Thus, this finding is another example for the fact that most cellular processes can be understood as the attempt to maximize entropy [144]. On the other hand, it was also found that entropic segregation alone is not sufficient to ensure the appropriate organization of genetic material in the cell. Here, it became apparent that the simplified model lacked certain ingredients without which reality cannot be described. Thus, a further development of the model was triggered as discussed in the third project with the additional implementation of the effects of SMC and ParAB. Moreover, since the experimental data were better fitted by modeling a factory model than by modeling a track model of replication, a further evidence could be derived with regard to the question of how replication is spatially organized in *B. subtilis*. At this point, however, there are still many opportunities for further investigation and more appropriate comparisons with experimental data. Thus, a first step would be to visualize the spatial distribution of duplication events, e.g., in the form of heat maps. These could then be compared with data from fluorescence microscopy experiments in which the replisomes could be labeled. The concept of entropic segregation also needs to be put to the test further. Here, the project on *S. meliloti* described above offers an interesting approach to test whether the segregation of the genetic material of multipartite bacteria can also be realized by entropic effects.

5.3. Trajectory classification

In the third part of this work, another requirement of the interdisciplinary approach should be addressed: the opening of new research perspectives through the application of new methods. A good way to do this is to transfer tools that have proven useful in one area to another. In our case, this concerns the automatic classification of trajectories using ML models. As information becomes available faster and in greater quantities, it becomes even more important to develop tools that allow us to process this information and derive predictions from it. It would be desirable to have an automated and standardized tool that recognizes and classifies patterns in the available data [98]. Once we have established such a method for classifying the data, it gives us the opportunity to ask more complex questions. In the case of trajectory classification such further questions might include whether it is possible to determine the underlying mechanism of a trajectory on the basis of a few datapoints. It would also be interesting to determine how many points are needed along a trajectory to classify it correctly.
In the context of the questions considered in this work, the data from SPT experiments offered themselves as an object of investigation as they are becoming more and more available in high resolutions ([37], [54], [164], [203]). At the same time, it has already been shown that ML algorithms are a promising tool for distinguishing different diffusion processes on the basis of SPT trajectories ([80], [97], [135], [192]). The aim of the present work was to test an application of this tool also for the classification of segregation mechanisms of bacterial chromosomes. A key obstacle to such an attempt in previous studies was the fact that a large number of trajectories are needed to train the ML models. The successful development of the MD model for combined replication and segregation of bacterial chromosomes in the previous project made it possible to produce trajectories of different segregation mechanisms in sufficient numbers to classify them with ML models for the first time. In this regard, the results of this work should act as a proof of principle that the tool of automatic classification of trajectories is also promising for the discrimination of trajectories of different loci in the bacterial cell. In addition, it was hoped that the analysis of the decision making of the ML models in their classification would also provide new insights into the interpretation of the individual models.
As described in the detailed project summary in 4.5, the additional implementation of further segregation mechanisms in the MD simulations was successful. In the process, the complex interactions of the proteins with the chromosome were broken down to their basic effects in the simplified conceptual framework of the model. In this sense, the effect of the interaction of SMC with the chromosome was considered as another topological constraint that could be implemented by installing harmonic potentials between the opposite chromosome arms. In the same way, the effect of the ParAB system can be understood as an effective traction force on the duplicated ori, which, according to Lim *et al.* [108], originates from the elastic properties of the chromosome and was calculated accordingly. In the process, comparison with experimental studies ([108], [165] [198], [200]) and further simulations ensured a correct description of the effects within the framework of the simulation ([52], [201]). A first interesting result of the following simulations was the identification of ParAB as the additional mechanism that ensures the maintenance of the spatial organization of DNA in the cell by providing designated direction of segregation for the oris. In the following, it was shown that it is indeed possible to use ML models to successfully classify the trajectories of ori in the cell based on the underlying mechanisms. Thus, in the form of ML approaches, another tool was identified to analyze the complexity

of the bacterial cell and the processes taking place within it. First of all, these methods provide a valuable opportunity to make automatic classifications and predictions as a "black box". This is an important aid, especially in biological or medical questions. In such it is not always possible to work out relationships that are comprehensible to the human brain due to the large number of interacting mechanisms. Nevertheless, one often needs reliable predictions or classifications. At the same time, however, the downstream analysis of the ML methods can give us a glimpse into the "black box" that can help us arrive at better interpretations of the underlying system. An example in the present work was the better distinction of cell types with active SMC from those without SMC in the factory model. From this it could be deduced that the topological organization of DNA by SMC must be particularly important in the factory model, in which a higher DNA density is present in the center of the cell. Moreover, the ML methods enable classification of very short trajectories and trajectories of different temporal resolutions at a scale at which this is hardly possible for the human eye. Thus, the results of the present work not only provide a proof of principle for the application of ML methods to segregation mechanisms in bacteria, but also show how a new understanding of the underlying processes can arise from the analysis of ML methods. In addition, suggestions are made for the ML methods to be used and the architectures of the individual models, as well as the preprocessing steps. The hope is that this will stimulate further applications of these methods to experimental data as well.
This offers a wide range of possibilities. For example, the trajectories do not necessarily have to describe the motion of the ori, but can be recorded for arbitrary loci of interest in the cell. In the context of *S. meliloti*, for example, a clear temporal order of the segregation movements of the terminus regions of the individual replicons [43] was described. Here it would be interesting to characterize the underlying mechanisms. In addition, the aforementioned question of the course of replication in track or factory model could be considered by tracking and classifying the trajectories of the replisomes during replication. Another challenge that the classifiers were not presented with yet is the possibility that segregation mechanisms might change within one trajectory. In this case, one would have to divide the trajectories into shorter sequences and try to classify these sequences independently of each other. In all these cases, it would be especially desirable to test transfer learning by testing the ML models on experimental data.

Final remark In conclusion, I hope to have provided some new perspectives on the spatio-temporal organization and segregation of DNA in the bacterial cell with this work. Physical modeling was used to gain a better understanding of the complex biological processes as well as new methods for their analysis. These may contribute to and inspire further investigations in the future.

Appendices

A. Polymer physics

A.1. Free energy of an ideal chain

The entropy S of an ideal chain is defined as

$$S = k_B \ln \Omega \quad , \tag{A.1}$$

with $\Omega(N, \vec{R})$ denoting the number of conformations of a freely jointed chain of N monomers with end-to-end vector $\vec{R}$ and k_B is the Boltzmann constant [157].
Thus, we need to find an expression for $\Omega(N, \vec{R})$ in order to estimate the entropy. For this purpose, we can write the probability distribution function of equation 1.9 as the fraction of all conformations that actually have an end-to-end vector $\vec{R}$ between $\vec{R}$ and $\vec{R} + d\vec{R}$ [157]

$$P(\vec{R}, N) = \frac{\Omega(N, \vec{R})}{\int \Omega(N, \vec{R}) d\vec{R}} \quad . \tag{A.2}$$

This relationship now allows us to wirte the entropy of our ideal chain as

$$\begin{aligned} S(N, \vec{R}) &= k_B \ln P(N, \vec{R}) + k_B \ln \left[\int \Omega(N, \vec{R}) d\vec{R} \right] \\ &= -\frac{3}{2} k_B \frac{\vec{R}}{Nb^2} + \underbrace{\frac{3}{2} k_B \ln \left(\frac{3}{2\pi N b^2} \right) + k_B \ln \left[\int \Omega(N, \vec{R}) d\vec{R} \right]}_{\text{independent of } \vec{R} \rightarrow S(N,0)} \quad . \end{aligned} \tag{A.3}$$

We find that in equation A.3 the last two terms don't depend on $\vec{R}$. Thus, we can denote them as $S(N, 0)$ and we receive [157]

$$S(N, \vec{R}) = -\frac{3}{2} k_B \frac{\vec{R}}{Nb^2} + S(N, 0) \quad . \tag{A.4}$$

Furthermore, we can now analyze the Helmholtz free energy $F(N, \vec{R}) = U(N, \vec{R}) - TS(N, \vec{R})$ of the ideal chain. Here, $U(N, \vec{R})$ denotes the internal energy of the system, T the absolute temperature of the system, and S is the entropy of the system which we just described. Since we describe an ideal chain, the monomers have no interaction energy and the internal energy $U(N, \vec{R})$ is independent of $\vec{R}$. Therefore, we can write [157]

$$F(N, 0) = U(N, 0) - TS(N, 0) \quad . \tag{A.5}$$

Inserting equation A.4 here yields

$$F(N,\vec{R}) = \frac{3}{2}k_BT\frac{\vec{R}^2}{Nb^2} + F(N,0) \quad , \tag{A.6}$$

as the free energy of the chain [157].
For the force needed to separate the ends of a FJC polymer by a distance $\vec{R}$ we find:

$$f = \frac{\partial F(N,\vec{R})}{\partial \vec{R}} = \frac{3k_BT}{Nb^2}\vec{R} \quad . \tag{A.7}$$

Thus, the force of the spring has an "entropic spring constant" of $3k_BT/Nb^2$ [157].

A.2. Frenet-Serret formulas

In this section we discuss the Frenet-Serret formulas which describe the geometric properties of a three-dimensional curve in $\mathbb{R}^3$ by describing the derivatives of the so-called tangent, normal and binormal unit vectors. We will first define the three vectors and then calculate their derivatives as done in [91]. By construction, the three vectors form an orthonormal basis for the $\mathbb{R}^3$ and the derivates allow us to define the curvature κ and the torsion τ of the curve. We will need the Frenet-Serret formulas later to calculate the Gauss linking integral.
To start our discussion, we describe a three-dimensional curve as

$$\vec{r}(s) = (x(s), y(s), z(s)) \quad , \tag{A.8}$$

with s being the arc length of the curve. Furthermore we assume the curve has a fixed length L. The *tangent vector* obviously is the vector tangent to the curve and defined as

$$\vec{t}(s) = \frac{\partial \vec{r}(s)}{\partial s} \quad . \tag{A.9}$$

Since we parameterized the curve by its arc length we know that the tangent vector is of unit length. We obtain the *normal unit vector* by calculating the derivative of the tangent vector with respect to the arc length and normalizing to unit length

$$\vec{n}(s) = \frac{\frac{\partial \vec{t}(s)}{\partial s}}{||\frac{\partial \vec{t}(s)}{\partial s}||} \quad . \tag{A.10}$$

At this point we can already define the **curvature** $\kappa(s)$ at point s as

$$\kappa(s) = ||\frac{\partial \vec{t}(s)}{\partial s}|| \quad . \tag{A.11}$$

The curvature describes the amount by which a curve deviates from a straight line. And we can write

$$\frac{\partial \vec{t}(s)}{\partial s} = \kappa(s)\vec{n}(s) \quad . \tag{A.12}$$

By construction we have

$$\vec{t}(s) \cdot \vec{n}(s) = 0 \quad , \tag{A.13}$$

ensuring that the first two vectors are perpendicular to each other. For our orthonormal basis of the $\mathbb{R}^3$ we only need one more vector perpendicular to $\vec{t}(s)$ and $\vec{n}(s)$. We can construct it using the cross product

$$\vec{b}(s) = \vec{t}(s) \times \vec{n}(s) \quad . \tag{A.14}$$

This vector $\vec{b}(s)$ is termed the *binormal vector*.
Now that we have obtained our orthonormal basis of the $\mathbb{R}^3$ we can write the derivative of $\vec{n}(s)$ as a linear combination of the other two basis vectors

$$\frac{\partial \vec{n}(s)}{\partial s} = \alpha(s)\vec{t}(s) + \tau(s)\vec{b}(s) \quad , \tag{A.15}$$

where $\alpha(s)$ is some function of s and $\tau(s)$ is called the **torsion** of the curve. Torsion describes how much the curve twists out of the plane of curvature.

To find a different expression for $\alpha(s)$, we differentiate the expression $\vec{t}(s) \cdot \vec{n}(s) = 0$. Doing this we find

$$\begin{aligned} 0 &= \frac{\partial}{\partial t}\left(\vec{t}(s) \cdot \vec{n}(s)\right) && \text{(A.16)} \\ &= \underbrace{\frac{\partial \vec{t}(s)}{\partial s}}_{\kappa(s)\vec{n}(s)} \cdot \vec{n}(s) + \underbrace{\frac{\partial \vec{n}(s)}{\partial s}}_{\alpha(s)\vec{t}(s)+\tau(s)\vec{b}(s)} \cdot \vec{t}(s) && \text{(A.17)} \\ &= \kappa(s) + \alpha(s) \quad , && \text{(A.18)} \end{aligned}$$

and thus $\alpha(s) = -\kappa(s)$. With this we have

$$\frac{\partial \vec{n}(s)}{\partial s} = -\kappa(s)\vec{t}(s) + \tau(s)\vec{b}(s) \quad , \tag{A.19}$$

for our second derivative. Last but not least we calculate the derivative of the binormal vector to

$$\begin{aligned} \frac{\partial \vec{b}(s)}{\partial s} &= \frac{\partial \vec{t}(s)}{\partial s} \times \vec{n}(s) + \vec{t}(s) \times \frac{\partial \vec{n}(s)}{\partial s} && \text{(A.20)} \\ &= \kappa(s)\vec{n}(s) \times \vec{n}(s) + \vec{t}(s) \times \Big(-\kappa(s)\vec{t}(s) + \tau(s)\vec{b}(s)\Big) && \text{(A.21)} \\ &= -\vec{n}(s)\tau(s) \quad . && \text{(A.22)} \end{aligned}$$

The equations A.12, A.19, and A.22 are the Frenet-Serret formulas and can be written in matrix notation as

$$\frac{\partial}{\partial s}\begin{bmatrix} \vec{t}(s) \\ \vec{n}(s) \\ \vec{b}(s) \end{bmatrix} = \begin{bmatrix} 0 & \kappa(s) & 0 \\ -\kappa(s) & 0 & \tau(s) \\ 0 & -\tau(s) & 0 \end{bmatrix} \begin{bmatrix} \vec{t}(s) \\ \vec{n}(s) \\ \vec{b}(s) \end{bmatrix} \tag{A.23}$$

A.3. Calculation of twist and writhe

In this section we derive expressions for the writhe and twist of two curves. We follow the approach from [91]. The most common way to calculate the linking number of two curves C and C' paramterized by $\vec{r}(s)$ and $\vec{r'}(s')$ is to use *Gauss Linking integral* ([91], [154], [191])

$$L_k = \frac{1}{4\pi} \oint_{C'} ds' \oint_C ds \frac{\vec{r'}(s') - \vec{r}(s)}{|\vec{r'}(s') - \vec{r}(s)|^3} \cdot \left[\frac{d\vec{r'}(s')}{ds'} \times \frac{d\vec{r}(s)}{ds} \right] \quad . \tag{A.24}$$

The linking integral can be derived in several ways for example by calculating Ampere's law with Biot-Savart's law ([31], [91]).
Since DNA appears as a single filament at long length scales, it is useful to recast the linking unmber in terms of the single polymer picture of DNA. For this purpose, in the following we assume a ribbon build up of the two curves $\vec{r}(s)$ and $\vec{r'}(s')$ [91]. We can assume the following relation

$$\vec{r'}(s') = \vec{r}(s') + \epsilon \vec{n}(s') \quad , \tag{A.25}$$

with a small number ϵ and the unit normal vector $\vec{n}(s')$ at $\vec{r}(s')$ pointing through $\vec{r'}(s')$ as in the Frenet frame above. We can furthermore calculate the tangent vector at $\vec{r'}(s')$ as

$$\frac{d\vec{r'}(s')}{ds'} = \frac{\vec{r}(s')}{ds'} + \epsilon \frac{d\vec{n}(\vec{s'})}{ds'} = \vec{t}(s') + \epsilon \frac{d\vec{n}(s')}{ds'} \quad . \tag{A.26}$$

Inserting this into equation A.24 we get

$$L_k = \frac{1}{4\pi} \oint_{C'} ds' \oint_C ds \frac{\vec{r}(s') + \epsilon\vec{n}(t') - \vec{r}(s)}{|\vec{r}(s') + \epsilon\vec{n}(t') - \vec{r}(s)|^3} \cdot \left[\left(\vec{t}(s') + \epsilon \frac{d\vec{n}(s')}{ds'} \right) \times \vec{t}(s) \right] \quad . \tag{A.27}$$

Now we want to calculate equation A.27 for $\epsilon \to 0$. In this limit we receive a singular part and a non-singular part separated at $s = s'$. For the evaluation we assume that there exists a $\delta \geq |s - s'|$ with $\epsilon << \delta$. For the non-singular part we receive

$$\frac{1}{4\pi} \oint_{C'} ds' \lim_{\delta \to 0^+} \left[\int_0^{s'-\delta} ds + \int_{s'+\delta}^{L} ds \right] \frac{\vec{r}(s') + \epsilon\vec{n}(t') - \vec{r}(s)}{|\vec{r}(s') + \epsilon\vec{n}(t') - \vec{r}(s)|^3} \cdot \left[\left(\vec{t}(s') + \epsilon \frac{d\vec{n}(s')}{ds'} \right) \times \vec{t}(s) \right] \equiv W_r. \tag{A.28}$$

Since $\epsilon << \delta$ and we take $\lim_{\delta \to 0}$, the two curves C and C' lie on top of each other and we can write

$$W_r = \frac{1}{4\pi} \oint_{C'} ds' \oint_C ds \frac{\vec{r}(s') - \vec{r}(s)}{|\vec{r}(s') - \vec{r}(s)|^3} \cdot \left(\vec{t}(s') \times \vec{t}(s) \right) \quad . \tag{A.29}$$

This is the writhing number W_r.
The singular part of equation A.27 is

$$\frac{1}{4\pi} \oint_{C'} ds' \int_{s'-\delta}^{s'+\delta} ds \frac{\vec{r}(s') + \epsilon\vec{n}(s') - \vec{r}(s)}{|\vec{r}(s') + \epsilon\vec{n}(s') - \vec{r}(s)|^3} \cdot \left[\left(\vec{t}(s') + \epsilon \frac{d\vec{n}(s')}{ds'} \right) \times \vec{t}(s) \right] \equiv T_W. \tag{A.30}$$

For small δ we can expand $\vec{r}(s)$ around s' and find in linear order

$$\vec{r}(s) \approx \vec{r}(s') + (s-s')\vec{t}(s') \tag{A.31}$$

$$\vec{t}(s) \approx \vec{t}(s') + (s-s')\frac{d\vec{t}(s')}{ds} \quad . \tag{A.32}$$

Inserting this yields

$$\begin{aligned}
T_W &= \frac{1}{4\pi}\oint_{C'} ds' \int_{s'-\delta}^{s'+\delta} ds \frac{\epsilon\vec{n}(s') - (s-s')\vec{t}(s')}{|\epsilon\vec{n}(s') - (s-s')\vec{t}(s')|^3} \cdot \left[\left(\vec{t}(s') + \epsilon\frac{d\vec{n}(s')}{ds'}\right) \times \left(\vec{t}(s') + (s-s')\frac{d\vec{t}(s')}{ds}\right)\right] \\
&\overset{\vec{t}(s')\times\vec{t}(s')=0}{=} \frac{1}{4\pi}\oint_{C'} ds' \int_{s'-\delta}^{s'+\delta} ds \frac{\epsilon\vec{n}(s') - (s-s')\vec{t}(s')}{|\epsilon\vec{n}(s') - (s-s')\vec{t}(s')|^3} \cdot \\
&\left[\epsilon\frac{d\vec{n}(s')}{ds'} \times \vec{t}(s') + \vec{t}(s') \times (s-s')\frac{d\vec{t}(s')}{ds} + \epsilon\frac{d\vec{n}(s')}{ds'} \times (s-s')\frac{d\vec{t}(s')}{ds}\right] \\
&= \frac{1}{4\pi}\oint_{C'} ds' \int_{s'-\delta}^{s'+\delta} ds \frac{1}{|\epsilon\vec{n}(s') - (s-s')\vec{t}(s')|^3} \cdot \\
&\{\epsilon\vec{n}(s') \cdot \left[\epsilon\frac{d\vec{n}(s')}{ds'} \times \vec{t}(s')\right] + \epsilon\vec{n}(s') \cdot \left[\vec{t}(s') \times (s-s')\frac{d\vec{t}(s')}{ds}\right] \\
&+ \left(\epsilon\vec{n}(s') - (s-s')\vec{t}(s')\right) \cdot \left[\epsilon\frac{d\vec{n}(s')}{ds'} \times (s-s')\frac{d\vec{t}(s')}{ds}\right]\} \quad .
\end{aligned} \tag{A.33}$$

We use further that $\vec{n}(s') \cdot \vec{t}(s') = 0$ and $\vec{n}(s') \cdot \vec{n}(s') = 1$ and this yields

$$\begin{aligned}
T_W &= \frac{1}{4\pi}\oint_{C'} ds' \int_{s'-\delta}^{s'+\delta} ds \frac{1}{\sqrt{(s-s')^2\vec{t}^2 + \epsilon^2}^3} \\
&\{\epsilon\vec{n}(s') \cdot \left[\epsilon\frac{d\vec{n}(s')}{ds'} \times \vec{t}(s')\right] + \epsilon\vec{n}(s') \cdot \left[\vec{t}(s') \times (s-s')\frac{d\vec{t}(s')}{ds}\right] + \\
&+ \epsilon\vec{n}(s') \cdot \left[\epsilon\frac{d\vec{n}(s')}{ds'} \times (s-s')\frac{d\vec{t}(s')}{ds}\right] - (s-s')\vec{t}(s') \cdot \left[\epsilon\frac{d\vec{n}(s')}{ds'} \times (s-s')\frac{dt(\vec{s'})}{ds}\right]\} \quad .
\end{aligned} \tag{A.34}$$

With a final integration and taking the limits by letting ϵ got to zero first and then letting δ go to zero one obtains

$$T_W = \frac{1}{2\pi}\oint_C ds\ \vec{t}(s) \cdot \left[\vec{n}(s) \times \frac{d\vec{n}(s)}{ds}\right] \quad . \tag{A.35}$$

Thus, inserting our results into equation A.27 yields the result optained by White and Fuller ([49], [208]).

$$L_k = W_r + T_W \quad . \tag{A.36}$$

It is also called Calugareanu's theorem [1].

A.4. Statistics of random walks

In this section a brief overview on the statistcs of random walks is given. Random walks (on a lattice) are frequently used to model flexible polymer chains. Thereby, one can either model ideal chains by allowing the random walk to visit the same site more than once or include self-avoidance and model a real chain by not allowing multiple occupancies of the same (lattice) site. The description of the basic statistics of random walks is based on the corresponding chapters in [179].
For simplicity, we start with the description of a one-dimensional random walk along x. We can define the N-step trajectory of the random walk with a step length b as a chain of length Nb [179]. In a one-dimensional random walk only "+" and "-" steps are possible. We can write the probability P_n to make n steps in one of the two directions out of a total of N steps as a binomial distribution

$$P_n = 2^{-N} \frac{N!}{n!(N-n)!} \quad . \tag{A.37}$$

Next, let us examine what changes in the probability distribution for large N. For this purpose, we take the natural logarithm on both sides of equation A.37 and use Sterling's formula $lnN! \cong N(lnN - 1)$. We receive

$$\ln P_n = -N \ln 2 + N(\ln N - 1) - n(\ln n - 1) - (N-n)[\ln(N-n) - 1] \tag{A.38}$$

$$= -N \ln 2 + N \ln N - n \ln n - (N-n) \ln(N-n) \quad . \tag{A.39}$$

In the end, we want to represent P as a function of x. That is, we want to find an expression for $P_n = P(x)$. For this we can use as expression for the distance x covered after N steps

$$x = b(2n - N) \quad . \tag{A.40}$$

From this follows $n = (N + x/b)/2$. Inserting this expression in equation A.39 yields

$$\begin{aligned}
\ln P \cong & -N \ln 2 + N \ln N - \frac{1}{2}(N - x/b) \ln[(N - x/b)/2] \\
& - \frac{1}{2}(N + x/b) \ln[(N + x/b)/2] \\
= & N \ln N - \frac{1}{2}(N - x/b) \ln(N - x/b) - \frac{1}{2}(N + x/b) \ln(N + x/b) \\
= & -\frac{1}{2} N \left[\left(1 - \frac{x}{Nb}\right) \ln\left(1 - \frac{x}{Nb}\right) + \left(1 + \frac{x}{Nb}\right) \ln\left(1 + \frac{x}{Nb}\right) \right] \\
\underbrace{\cong}_{\text{Taylor}} & -\frac{1}{2} N \left(\frac{x}{Nb}\right)^2 = -\frac{x^2}{2Nb^2} \quad .
\end{aligned} \tag{A.41}$$

In the last part we used the Taylor expansion up to the second order of $x/(Nb)$. This is possible because $P(x)$ is almost zero except at small $|x/(Nb)|$. Finally, one needs to normalize the result with the condition $\int P(x)dx = 1$ and we receive

$$P(x) = (2\pi N b^2)^{-1/2} \exp\left(-\frac{x^2}{2Nb^2}\right) \quad . \tag{A.42}$$

We find that a (one-dimensional) random walk results in a normal distribution with a zero mean and a variance of Nb^2.
We can now turn towards a three-dimensional case and consider a cubic lattice with lattice spacing b. Here we can define the displacement of one lattice step as $\triangle\vec{r}_1 = [\triangle x_1, \triangle y_1, \triangle z_1]$. Furthermore, we have

$$\begin{aligned} \langle \triangle\vec{r}_1 \rangle &= 0 \\ \langle \triangle x_1^2 \rangle = \langle \triangle y_1^2 \rangle = \langle \triangle z_1^2 \rangle &= b^2/3 \\ \langle \triangle r_1^2 \rangle &= b^2 \quad , \end{aligned} \tag{A.43}$$

for the displacements of a single lattice step. If we consider a total of N steps we find

$$\begin{aligned} \langle \triangle\vec{r} \rangle &= 0 \\ \langle \triangle x^2 \rangle = \langle \triangle y^2 \rangle = \langle \triangle z^2 \rangle &= Nb^2/3 \\ \langle \vec{r}^2 \rangle = \langle \triangle \vec{r}^2 \rangle &= Nb^2 \quad . \end{aligned} \tag{A.44}$$

The probability density for three dimensions can be written as $P(\vec{r}) = P_x(\vec{r})P_y(\vec{r})P_z(\vec{r})$. Thereby, the probability densities of the three components are only differ in the direction. For the x component we find that the random walks has a zero mean and a variance of $Nb^2/3$ after N steps. Thus, for large N we expect a normal distribution with the same mean and variance which can be written as

$$P_x(\vec{r}) = \left(\frac{2\pi N b^2}{3}\right)^{-\frac{1}{2}} \exp\left[-\frac{3x^2}{2Nb^2}\right] \quad . \tag{A.45}$$

Altogether, this gives us a probability density for three dimensions of

$$\begin{aligned} P(\vec{r}) &= P_x(\vec{r})P_y(\vec{r})P_z(\vec{r}) \\ &= (2\pi N b^2)^{-3/2} \exp\left(-\frac{3\vec{r}^2}{2Nb^2}\right) \quad . \end{aligned} \tag{A.46}$$

We see that equation A.46 is just identical to the probability distribution of a freely jointed chain as depicted in equation 1.9. Therefore, we see at this point that random walks are optimal models for polymers.

A.5. Overlapping polymers

In order to estimate the free energy cost of two overlapping polymers without confinement we consider two identical polymers consisting of N monomers each. The Flory theory estimates that the size of each of the two polymers scales as described in equation 1.25 $R_F \sim N^\nu$. Thus, the volume in which both polymers overlap might be estimated as

$$V \sim R_F^3 \sim N^{3\nu} \sim N^{9/5} \quad . \tag{A.47}$$

Furthermore, the Flory theory assumes independently distributed monomers in the overlap volume. Thus, the monomer concentration can be written as

$$\rho \approx \frac{N}{V} \approx N^{1-3\nu} \approx N^{-4/5} \quad . \tag{A.48}$$

In the next step one has to consider the number of contacts between the monomers of different polymers, $n_{contacts}$. Here, the Flory theory takes the monomer concentration as the contact probability, setting $p_{contact} = \rho$, and receives

$$n_{contacts} \approx N \cdot \frac{N}{V} \approx \frac{N^2}{N^{3\nu}} \approx N^{2-3\nu} \approx N^{1/5} \quad . \tag{A.49}$$

With this the free energy of the interacting polymers is estimated. The free energy scales with the number of monomer contacts, $n_{contacts}$, and thus we can write

$$\mathcal{F} \sim k_B T n_{contacts} \sim k_B T N^{1/5} \quad . \tag{A.50}$$

This result of the Flory theory predicts that long polymers should behave as mutually impenetrable hard spheres.
However, as discussed previously, the Flory theory makes some estimation errors. The main mistake of the assumptions made by the Flory theory is to assume independently distributed monomers. This is obviously wrong because monomer correlations are caused by the linkage of the monomers along the backbone of the polymer. It was shown later by Grosberg et al. [57] that in fact

$$\mathcal{F} \sim k_B T \quad . \tag{A.51}$$

This result now indicates that polymers in bulk can rather easy intermingle ([84], [157]).

B. Numerical implementation

B.1. Monte Carlo simulation

B.1.1. MOS algorithm

In this section a short description of the MOS algorithm as presented in [115] shall be given. The algorithm is defined for arbitrary dimensions d. Thus, we consider a d-dimensional space $\mathbb{R}^d$. Here, any point is defined by its coordinate $(x^{(1)}, x^{(2)}, ..., x^{(d)})$. Since we study SAWs on a lattice, the d-dimensional lattice is

$$\mathcal{L}^d = \{(x^{(1)}, ..., x^{(d)}) : x^{(i)} \text{ is an integer for } i = 1, ..., d\} \quad . \tag{B.1}$$

In this setting an N-step SAW w can be written as a sequence $w_0, w_1, ..., w_N$ where each point is a nearest neighbor of its predecessor. Thus, $|w_i - w_{i-1}| = 1$ for $i = 1, ..., N$. Furthermore, we want to fixate the endpoints, denoted as w_0 and w_N. With this it is possible to define the set of all N-step SAWs having $w_0 = A$ and $w_N = B$ for two points $A, B \in \mathcal{L}^d$, $S^N(A, B)$.
The MOS algorithm consists of a finite set $\mathcal{F} = \{F_1, ..., F_r\}$ of transformations of $S^N(A, B)$ into itself. If one is given any SAW, $w^{[0]}$, in $S^N(A, B)$ at time $t = 0$, one can iteratively come up with the next SAW at each successive integer time t, knowing $w^{[t-1]}$, by choosing a number $n(t)$ at random from $\{1, .., r\}$ accoring to a fixed probability distribution, and put $w^{[t]} = F_{n(t)}\left(w^{[t-1]}\right)$. The resulting sequence $w^{[0]}, w^{[1]}, ...$ of SAWs is a Markov chain on $S^N(A, B)$.
What remains is to define the transformations F_i. For this, in a space with $d \geq 3$ one needs three classes of transformations T_i, called *inversion, reflection*, and *interchange* as building blocks of the F_i transformations. A transformation T_i always tries to perturb a SAW into some other object. Subsequently, the deformation is accepted if the resulting object is a SAW and rejected otherwise. Therefore, a transformation F_i can be defined as follows

$$F_i(w) = \begin{cases} T_i(w) & \text{if } T_i(w) \in S^N(A, B) \\ w & \text{if } T_i(w) \notin S^N(A, B) \end{cases} \quad . \tag{B.2}$$

The first T_i is the inversion transformation. For a SAW $(w_0, ..., w_N)$ and integers k and l such that $0 \leq k < l \leq N$ the inversion transformation $T^{inv}_{k,l}(w)$ is defined as the sequence $w' = (w'_0, ..., w'_N)$ given by

$$w'_i = \begin{cases} w_k + w_l - w_{k+l-i} & \text{if } k \leq i \leq l \\ w_i & \text{otherwise} \end{cases} \quad . \tag{B.3}$$

Thus, the inversion transformation inverts the original curve $[w_l, ..., w_k]$ through the point $(w_k + w_l)/2$. Notably, upon setting $l = k + 2$ the inversions $T^{inv}_{k,k+2}(w)$ are exactly the

length-preserving BFACF moves.
The reflection transformation reflects a piece of a SAW through a hyperplane which makes angles of 45° with two of the coordinate hyperplanes. Thus, for a SAW $w \in S^N(A,B)$, and integers $0 \leq k < l \leq N, m \in \{-1,1\}$, and $1 \leq \alpha < \beta \leq d$, the reflection transformation $T^{ref,m}_{k,l;\alpha,\beta}(w)$ is defined as follows. The transformation is rejected if $w_l^{(\alpha)} - w_k^{(\alpha)} \neq m(w_l^{(\beta)} - w_k^{(\beta)})$ or if $w_l^{(\gamma)} \neq w_k^{(\gamma)}$ for some $\gamma \neq \alpha, \beta$. In the first case, the points w_l and w_k would be opposite corners of a square and in the second case the points would lie in the same plane. In both cases a reflection would not make sense and thus one puts $T^{ref,m}_{k,l;\alpha,\beta} = w$. However, if both conditions don't result in a rejection, one can perform the transformation by setting

$$(w_i^{'})^{(\gamma)} = \begin{cases} w_k^{(\gamma)} - m(w_{k+l-i}^{(\alpha+\beta-\gamma)} - w_l^{(\alpha+\beta-\gamma)}) & \text{if } k \leq i \leq l \text{ and } \gamma \text{ is } \alpha \text{ or } \beta \\ w_i^{(\gamma)} & \text{otherwise} \quad . \end{cases} \tag{B.4}$$

The last transformation we need to define to ensure ergodicity of the algorithm in three or more dimensions is the interchange transformation. Again, for $w \in S^N(A,B), 0 \leq k < l \leq N, m \in \{-1,1\}$, and $1 \leq \alpha < \beta \leq d$, define $T^{int,m}_{k,l;\alpha,\beta}(w)$ as follows. The transformation is also rejected if the points w_l and w_k lie on opposite corners of a square. If this is not the case, we define the transformation via the transformation of the sequence of steps that build the SAW. Thus, if the original walk w is defined by the sequence $s_1, s_s, ..., s_N$ of steps with $s_i = w_i - w_{i-1}$, the interchange transformation produces the walk $w^{'} = T^{int,m}_{k,l;\alpha,\beta}(w)$. Thereby, the new steps $s_i^{'} = w_i^{'} - w_{i-1}^{'}$ are

$$s_i^{'(\gamma)} = \begin{cases} m \cdot s_i^{\beta} & \text{if } k < i \leq l \text{ and } \gamma = \alpha \\ m \cdot s_i^{\alpha} & \text{if } k < i \leq l \text{ and } \gamma = \beta \\ s_i^{(\gamma)} & \text{otherwise} \quad . \end{cases} \tag{B.5}$$

Thus, the interchange transformation interchanges the α and β coordinates of the steps $s_{k+1}, ..., s_l$. The orientation of the interchanged coordinates is conserved for $m = +1$ and interchanged for $m = -1$.

B.1.2. A* algorithm

The A* algorithm is a prominent path search algorithm. The goal of the algorithm is to find the optimal path (i.e. the path with the lowest cost) from a given start node to a given end node [158]. The cost can be measured in arbitrary units like distance travelled or time spend. The A* algorithm is a modification of other path seach algorithms like Breadth First Search or Dijkstra's algorithm. While Breadth First Search and Dijkstra's algorithm explore equally in all directions, the A* algorithm tries to find a directed way following the smallest estimated cost. Therefore, the A* algorithm does not try to follow several paths at the same time in order to select a certain one at the end, but the algorithm decides on-the-fly which of the paths it will follow further. In order to do so, the algorithm maintains a tree of paths originating at the start node and decides at each iteration which path to extend further until the goal is reached. This is implemented using the concept of a frontier. The frontier can be imagined like an expanding ring around the start node and it describes the successive extension of the search radius over adjacent nodes. In each

iteration of the algorithm, the frontier is extended and the cost function is calculated for the new fields. The cost function is at the core of the algorithm [158]. It is defined as

$$f(n) = g(n) + h(n) \quad . \tag{B.6}$$

Here, n is the node on the path, $g(n)$ is the cost from the start node to the current node n, and $h(n)$ is a heuristic function that estimates the cost of the cheapest path from n to the goal node. On a lattice the logic choice for the heuristic function is the Manhatten distance. The use of the estimated cost to the target with the help of the heuristic is the central extension of the A* algorithm compared to the Dijkstra's algorithm and leads to a more goal oriented search. A schematic depiction of pathfinding with the A* algorithm on a two-dimensional grid is shown in figure B.1

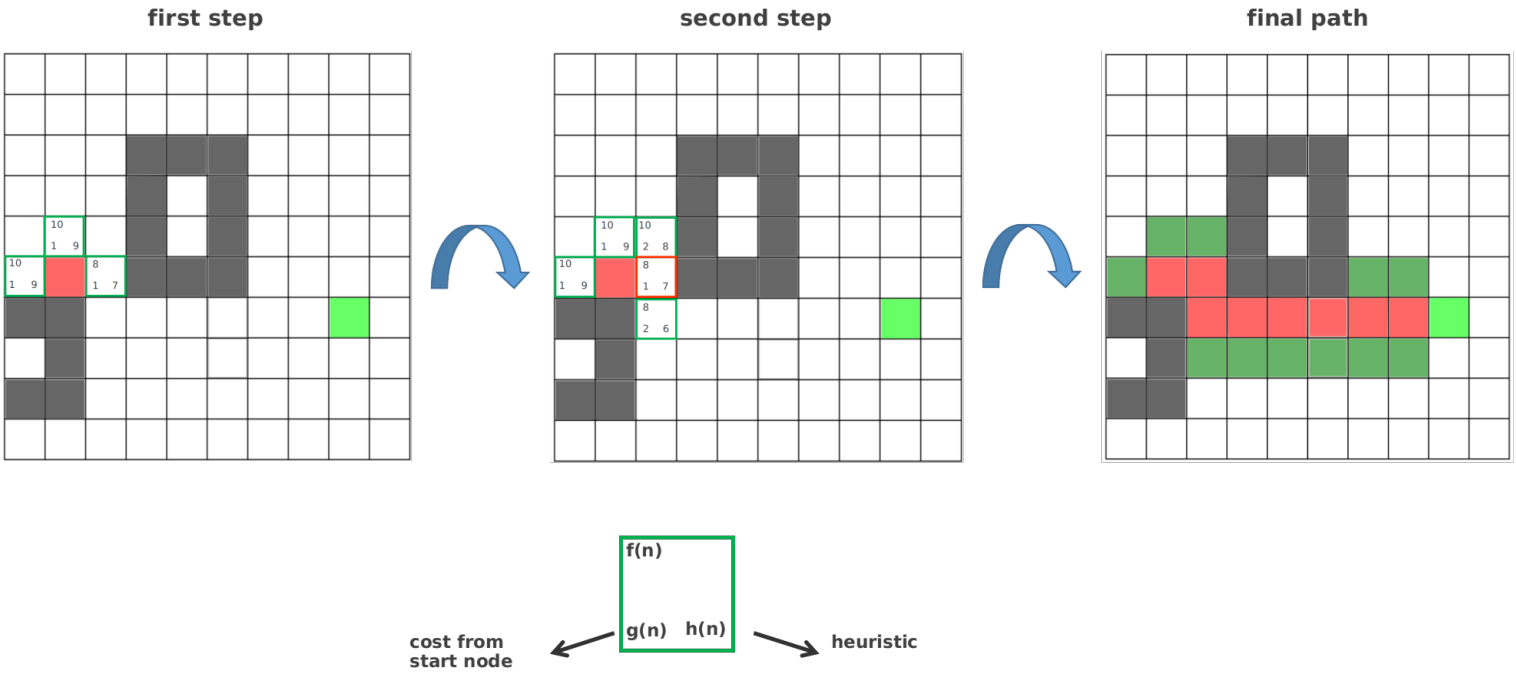

Figure B.1.: Path search with the A* algorithm. The start node is shown in red and the goal node is shown in green. The frontier is marked by the green border and the costs are noted in the squares. The gray areas represent obstacles.

As can be seen in figure B.1 the algorithm expands its frontier (squares with green borders) at every iteration. Thereby, the cost function is evaluated and the square within in the frontier with the lowest cost function is chosen to expand the path. If the goal node is reached by the frontier or if there are no paths eligible to be extended, the algorithm terminates [158].

B.2. MD implementation

B.2.1. Velocity verlet algorithm

In MD simulations the aim is to numerically solve Newton's equations of motion. Thereby, the time-dependent behavior of the system is evaluated. In order to do this, an efficient method for time integration is needed. The standard in Md simulations up to date is the Velocity Verlet algorithm introduced by Verlet in 1967 [189]. This algorithm shall here be briefly described. We start with the equations of motion for point-like particle

$$\begin{aligned} \dot{v}_i(t) &= \frac{F_i\left(\{x_i\}, v_i, t\right)}{m_i} \\ \dot{x}_i(t) &= v_i(t) \quad . \end{aligned} \tag{B.7}$$

Here, x_i, v_i, m_i are position, velocity and mass of the particle i and $F_i\left(\{x_i\}, v_i, t\right)$ are the forces acting on it as a result of interactions with other particles or external fields. A besic description of the Verlet formula can be obtained using Taylor expansions for $x(t)$ as

$$x(t+h) = x(t) + h\dot{x}(t) + (h^2/2)\ddot{x}(t) + O(h^3) \quad , \tag{B.8}$$

where t is the current time, and $h \equiv \triangle t$. Furthermore, we denote with $\dot{x}(t)$ the velocity and with $\ddot{x}(t)$ the acceleration of the particle. Analogue to equation B.8 we can write

$$x(t-h) = x(t) - h\dot{x}(t) + (h^2/2)\ddot{x}(t) + O(h^3) \quad . \tag{B.9}$$

The Verlet formula is obtained by adding equation B.9 to equation B.8

$$x(t+h) = 2x(t) - x(t-h) + h^2\ddot{x}(t) + O(h^4) \quad . \tag{B.10}$$

We realize that we can calculate the position at $t+h$ by using the information from the two previous time points t and $t-h$. This provides the opportunity to calculate a trajectory iteratively. Furthermore, the velocity can be calculated as

$$\dot{x}(t) = \frac{x(t+h) - x(t-h)}{2h} + O(h^2) \quad , \tag{B.11}$$

if needed ([56], [151]).
A variant of the Velocity Verlet method is the Leapfrog method which is used in the software package `ESPResSo` used in this work. The method is derived similarly. We write the Taylor expansion as

$$x(t+h) = x(t) + h\underbrace{\left[\dot{x}(t) + (h/2)\ddot{x}(t)\right]}_{=\dot{x}(t+h/2)} + O(h^3) \quad . \tag{B.12}$$

Thus, we can further write

$$\begin{aligned} \dot{x}(t+h/2) &= \dot{x}(t) + \frac{h}{2}\ddot{x}(t) \\ \dot{x}(t-h/2) &= \dot{x}(t) - \frac{h}{2}\ddot{x}(t) \quad . \end{aligned} \tag{B.13}$$

Combination of the two equations above yields

$$\dot{x}(t + h/2) = \dot{x}(t - h/2) + h\ddot{x}(t) \quad . \tag{B.14}$$

Together, the leapfrog integration formulae are [151]

$$\begin{aligned} \dot{x}(t + h/2) &= \dot{x}(t - h/2) + h\ddot{x}(t) \\ x(t + h) &= x(t) + h\dot{x}(t + h/2) \quad . \end{aligned} \tag{B.15}$$

The algorithm implemented in `ESPResSo` proceeds in the following four step procedure [205]

1. Calculate the velocity at the half step
2. Calculate the new position
3. Calculate the force based on the new position
4. Calculate the new velocity

It should be noted that in the first time step no forces are present yet in `ESPResSo`. Therefore, they are either computed before the first time step (for random forces) or are lacking in the first half time step (coupling forces).

C. Additional analyses

C.1. Outlier clearance for experimental data of S. meliloti

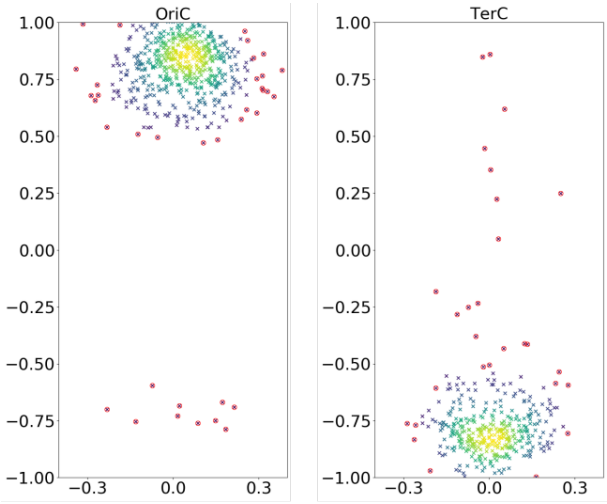

Figure C.1.: Outlier detection in experimental data. Shown are the representative heatmaps of the oris and ters in the *S. meliloti* WT strain. Analyzing the data, we realized that the automated foci detection in the MicrobeJ plugin sometimes exhibits errors in assigning the foci to their cell poles. Furthermore, we have also assumed a natural rate of measurement outliers of 10%. To eliminate such outliers, we calculated the probability density function (PDF) for each data sample and excluded all points with a probability below 10% of belonging to the distribution (red encircled points). This was done for all experimental data points.

C.2. Model results for corrected ori positions in S. meliloti

In the analyses of the experimental data on the WT and △ pSymA strain of *S. meliloti*, there was much variation in the positions of the oris. Assuming that the organization of plasmids in the cell is regulated by a spatial determination of the oris and that this works equally in all strains, we can adapt our model accordingly to suitable positions. It makes sense to choose those positions that do not result in a jump between the position of the respective ori and the following loci of a plasmid, which is difficult to explain. This is best satisfied by the position of *oriB* measured in WT and the position of *oriC* measured in △ pSymA. In addition, *oriA* is slightly offset within its standard deviation. In figure C.2 (A) we see the results for WT if we assume a spatial confinement in the model as before in consequence of an *terB* enrichment zone.

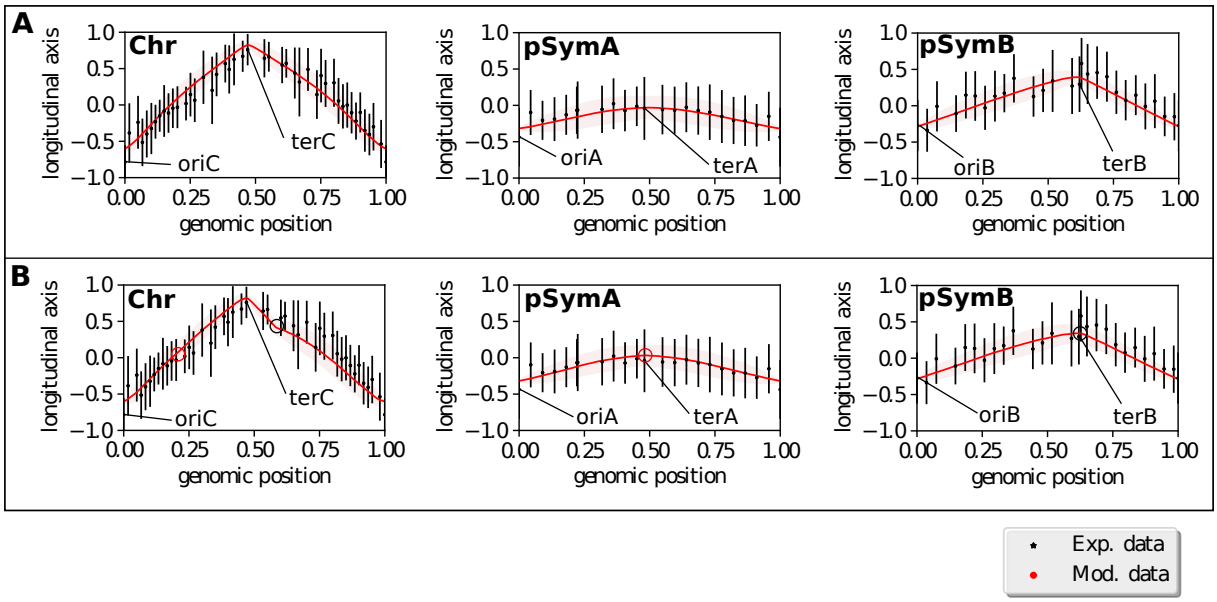

Figure C.2.: **A:** Spatial organization of the replicons in *S. meliloti* WT due to spatial confinement of *terA* and *terB* to enrichment zones. In black the experimental data is shown while the model results are shown in red (standard deviation as shaded area). *terA* and *terB* are restricted to an enrichment zone defined as 75 % of the experimental standard deviation for the markers. **B:** Spatial organization of the replicons in *S. meliloti* WT due to genomic fixation of *terA* and *terB* to the chromosome. The interacting loci are marked with corresponding circles. The three subplots in both A and B each show the organization of one of the three replicons in the same cell.

In figure C.2 (B) the model prediction for inter-replicon interactions of the plasmids with the chromosome is shown. We recognize that this assumption is not only more logical in terms of interpretation of the experimental data, but also results in an improved fit of the model to them. The same applies to △ pSymA. In figure C.3 the same coordinates of oris were used as for WT and we see an improvement of the fit here as well.

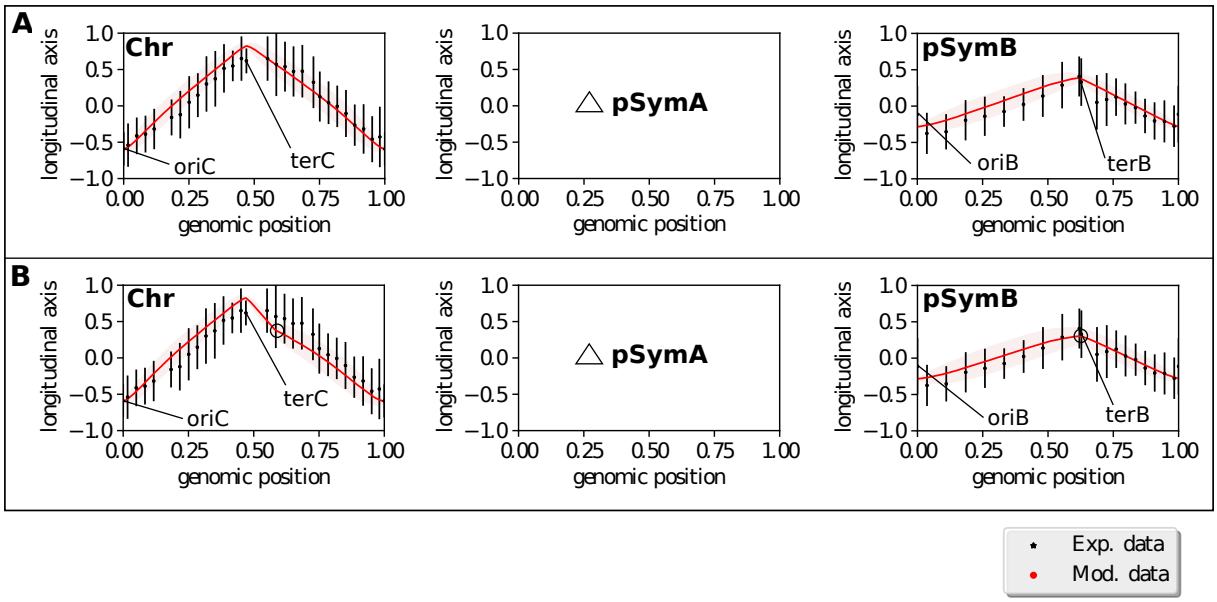

Figure C.3.: **A:** Model prediction of the spatial organization of the replicons in the knock-out mutant △ pSymA due to spatial confinement of the *terB* region. In black the experimental data is shown while the model results are shown in red (standard deviation as shaded area). **B:** Model prediction due to genomic fixation of the *terB* region to the chromosome.

C.3. Degree of separation after replication in MD simulations

In the simulations the degree of separation of the two chromosomes after replication was also measured. The results for the various cell types are shown in table C.1.

	Track model	**Factory model**
WT	82.16	86.66
dSMC	83.00	91.54
dParAB	40.88	46.1
dSMCdParAB	66.64	42.10

Table C.1.: Average degree of separation within the different cell types after replication. Results averaged over 3000 runs for each cell type.

For the calculations the degree of separation was defined as the longitudinal overlap of the chromosomes within the cell divided by the longitudinal elongation of the shorter chromosome in the cell.
The results from table C.1 show that for the cell types in which ParAB is active (WT and dSMC) a higher degree of separation of the chromosomes is reached after replication. Obviously, the ParAB system is both improtant for the partitioning of the oris and also has a strong influence on the segregation of the complete chromosomes.
The distribution of the achieved degrees of separation after replication per cell type is shown in the histograms of figure C.4 and figure C.5. Again, one finds a higher variability in the different degrees of separation after replication for cells lacking the ParAB system. Without the effect of the ParAB system cells still may achieve complete separation of

chromosomes during the replication phase, but not as reliably as in the case of additional help by ParAB.

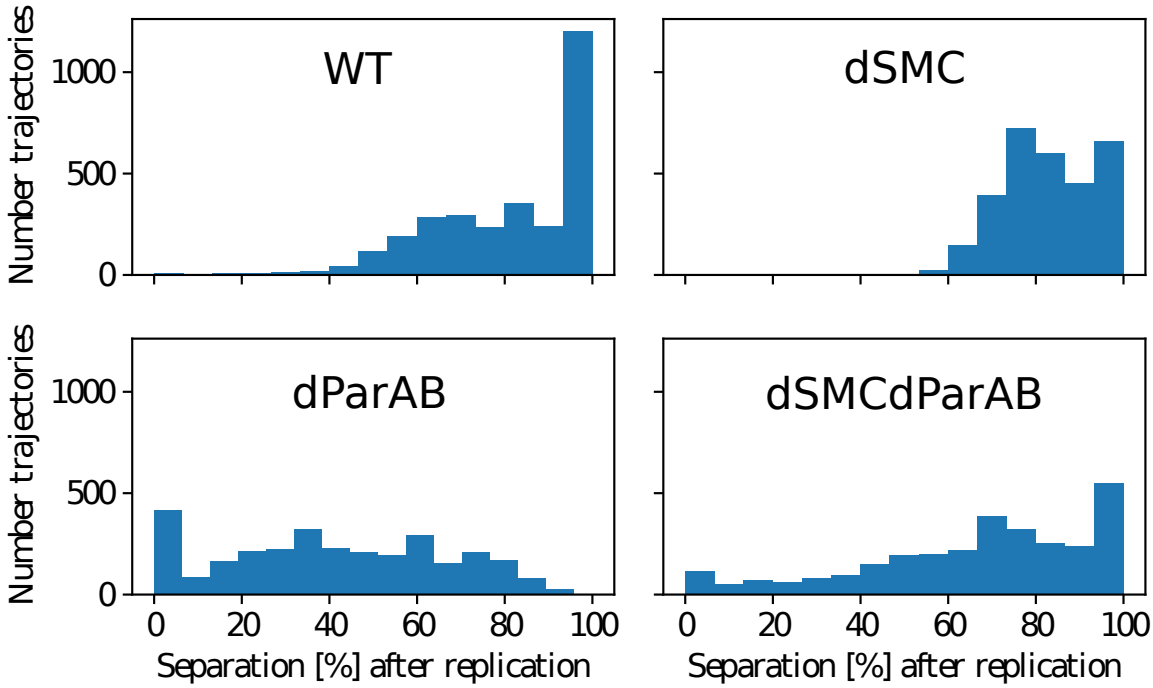

Figure C.4.: Histograms for the degree of separation of the chromosomes after replication with the track scheme. The four histograms display the results for the various segregation schemes.

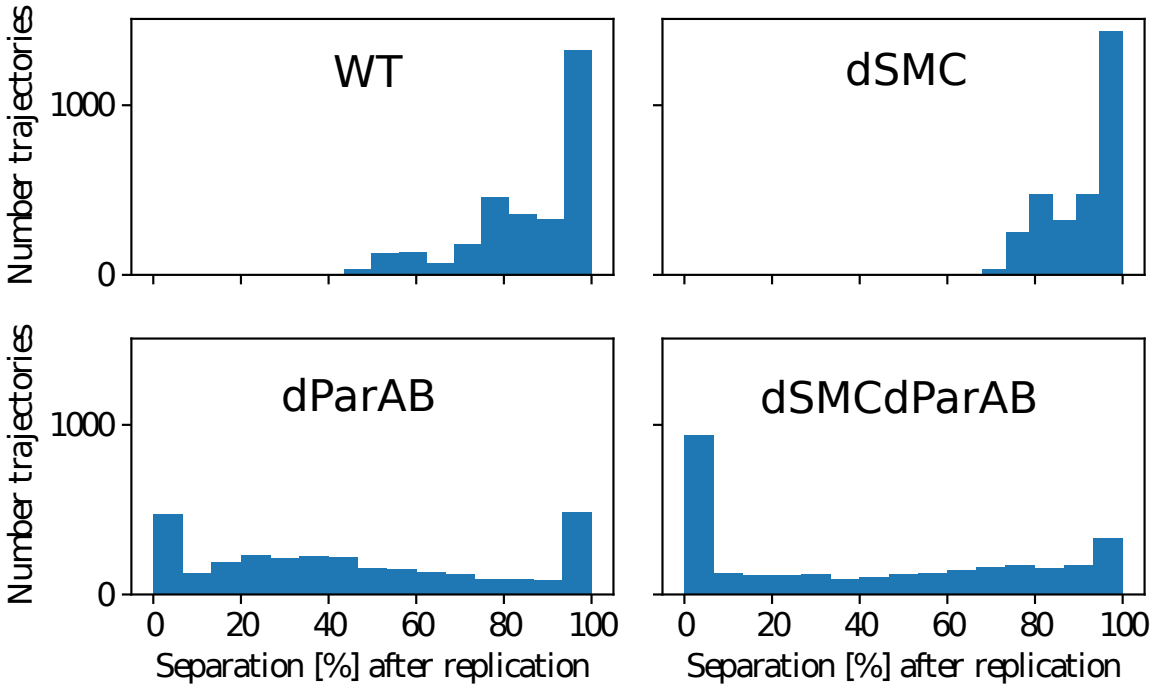

Figure C.5.: Histograms for the degree of separation of the chromosomes after replication with the factory scheme. The four histograms display the results for the various segregation schemes.

Bibliography

[1] Adams, C. C., & Brown, C. C. (1994). The Knot Book: An Elementary Introduction to the Mathematical Theory of Knots. Nature, 371(6498), 568-568.

[2] Agarwal, T., Manjunath, G. P., Habib, F., & Chatterji, A. (2019). Bacterial chromosome organization. I. Crucial role of release of topological constraints and molecular crowders. The Journal of chemical physics, 150(14), 144908.

[3] Agarwal, T., Manjunath, G. P., Habib, F., & Chatterji, A. (2019). Bacterial chromosome organization. II. Few special cross-links, cell confinement, and molecular crowders play the pivotal roles. The Journal of chemical physics, 150(14), 144909.

[4] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., & Keith Roberts, P. W. (2018). Molecular biology of the cell.

[5] Albon, C. (2018). Python machine learning cookbook.

[6] Alder, B. J., & Wainwright, T. E. (1957). Phase transition for a hard sphere system. The Journal of chemical physics, 27(5), 1208-1209.

[7] Arnold, A., & Jun, S. (2007). Time scale of entropic segregation of flexible polymers in confinement: implications for chromosome segregation in filamentous bacteria. Physical Review E, 76(3), 031901.

[8] Assmann, M. A., & Lenz, P. (2013). Characterization of bidirectional molecular motor-assisted transport models. Physical biology, 10(1), 016003.

[9] Badrinarayanan, A., Le, T. B., & Laub, M. T. (2015). Bacterial chromosome organization and segregation. Annual review of cell and developmental biology, 31, 171-199.

[10] Berg, B., & Foester, D. Random paths and random surfaces on a digital computer, 1981. Phys. Lett. B, 106, 323.

[11] Berne, B. J., & Straub, J. E. (1997). Novel methods of sampling phase space in the simulation of biological systems. Current Opinion in Structural Biology, 7(2), 181-189.

[12] Bhattacharjee, S. M., Giacometti, A., & Maritan, A. (2013). Flory theory for polymers. Journal of Physics: Condensed Matter, 25(50), 503101.

[13] Bianco, P. R., Brewer, L. R., Corzett, M., Balhorn, R., Yeh, Y., Kowalczykowski, S. C., & Baskin, R. J. (2001). Processive translocation and DNA unwinding by individual RecBCD enzyme molecules. Nature, 409(6818), 374-378.

[14] Bloom, K., & Joglekar, A. (2010). Towards building a chromosome segregation machine. Nature, 463(7280), 446-456.

[15] Bo, S., Schmidt, F., Eichhorn, R., & Volpe, G. (2019). Measurement of anomalous diffusion using recurrent neural networks. Physical Review E, 100(1), 010102.

[16] Bohn, M., & Heermann, D. W. (2011). Repulsive forces between looping chromosomes induce entropy-driven segregation. PloS one, 6(1), e14428.

[17] Borgmann, L. A. K., Hummel, H., Ulbrich, M. H., & Graumann, P. L. (2013). SMC condensation centers in Bacillus subtilis are dynamic structures. Journal of bacteriology, 195(10), 2136-2145.

[18] Borgmann, L. A. K., & Graumann, P. L. (2014). Structural maintenance of chromosome complex in bacteria. Journal of molecular microbiology and biotechnology, 24(5-6), 384-395.

[19] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

[20] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[21] Brocken, D. J., Tark-Dame, M., & Dame, R. T. (2018). The organization of bacterial genomes: Towards understanding the interplay between structure and function. Current Opinion in Systems Biology, 8, 137-143.

[22] Brouns, T., De Keersmaecker, H., Konrad, S. F., Kodera, N., Ando, T., Lipfert, J., ... & Vanderlinden, W. (2018). Free energy landscape and dynamics of supercoiled DNA by high-speed atomic force microscopy. ACS nano, 12(12), 11907-11916.

[23] Bryant, Z., Stone, M. D., Gore, J., Smith, S. B., Cozzarelli, N. R., & Bustamante, C. (2003). Structural transitions and elasticity from torque measurements on DNA. Nature, 424(6946), 338-341.

[24] Buenemann, M., & Lenz, P. (2010). A geometrical model for DNA organization in bacteria. PLoS One, 5(11), e13806.

[25] Buenemann, M., & Lenz, P. (2011). Geometrical ordering of DNA in bacteria. Communicative & Integrative Biology, 4(3), 291-293.

[26] Clisby, N. (2010). Efficient implementation of the pivot algorithm for self-avoiding walks. Journal of Statistical Physics, 140(2), 349-392.

[27] De Carvalho, C. A., Caracciolo, S., & Fröhlich, J. (1983). Polymers and $g|\phi|^4$ theory in four dimensions. Nuclear Physics B, 215(2), 209-248.

[28] Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2016). Transferable model for chromosome architecture. Proceedings of the National Academy of Sciences, 113(43), 12168-12173.

[29] Cook, P. R., & Marenduzzo, D. (2009). Entropic organization of interphase chromosomes. Journal of Cell Biology, 186(6), 825-834.

[30] De Gennes, P. G., & Gennes, P. G. (1979). Scaling concepts in polymer physics. Cornell university press.

[31] De Zela, F. (2004). Linking Maxwell, Helmholtz and Gauss through the linking integral. arXiv preprint physics/0406037.

[32] Di Ventura, B., Knecht, B., Andreas, H., Godinez, W. J., Fritsche, M., Rohr, K., ... & Sourjik, V. (2013). Chromosome segregation by the Escherichia coli Min system. Molecular systems biology, 9(1), 686.

[33] Dorier, J., & Stasiak, A. (2009). Topological origins of chromosomal territories. Nucleic acids research, 37(19), 6316-6322.

[34] Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science, 346(6213).

[35] Ducret, A., Quardokus, E. M., & Brun, Y. V. (2016). MicrobeJ, a tool for high throughput bacterial cell detection and quantitative analysis. Nature microbiology, 1(7), 1-7.

[36] Dworkin, J., & Losick, R. (2002). Does RNA polymerase help drive chromosome segregation in bacteria?. Proceedings of the National Academy of Sciences, 99(22), 14089-14094.

[37] El Najjar, N., Geisel, D., Schmidt, F., Dersch, S., Mayer, B., Hartmann, R., ... & Graumann, P. L. (2020). Chromosome Segregation in Bacillus subtilis Follows an Overall Pattern of Linear Movement and Is Highly Robust against Cell Cycle Perturbations. Msphere, 5(3).

[38] Ernst, D., Köhler, J., & Weiss, M. (2014). Probing the type of anomalous diffusion with single-particle tracking. Physical Chemistry Chemical Physics, 16(17), 7686-7691.

[39] Estévez-Torres, A., & Baigl, D. (2011). DNA compaction: fundamentals and applications. Soft Matter, 7(15), 6746-6756.

[40] Finan, T. M. (2017). The divided bacterial genome: structure, function, and evolution. Microbiology and molecular biology reviews, 81(3).

[41] Flory, P. J., & Volkenstein, M. (1969). Statistical mechanics of chain molecules.

[42] Fogel, M. A., & Waldor, M. K. (2006). A dynamic, mitotic-like mechanism for bacterial chromosome segregation. Genes & development, 20(23), 3269-3282.

[43] Frage, B., Döhlemann, J., Robledo, M., Lucena, D., Sobetzko, P., Graumann, P. L., & Becker, A. (2016). Spatiotemporal choreography of chromosome and megaplasmids in the Sinorhizobium meliloti cell cycle. Molecular microbiology, 100(5), 808-823.

[44] Freund, Y., & Schapire, R. E. (1999). Adaptive game playing using multiplicative weights. Games and Economic Behavior, 29(1-2), 79-103.

[45] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[46] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

[47] Fritsche, M., Li, S., Heermann, D. W., & Wiggins, P. A. (2012). A model for Escherichia coli chromosome packaging supports transcription factor-induced DNA domain formation. Nucleic acids research, 40(3), 972-980.

[48] Fourey, S., & Malgouyres, R. (2001). A digital linking number for discrete curves. International journal of pattern recognition and artificial intelligence, 15(07), 1053-1074.

[49] Fuller, F. B. (1971). The writhing number of a space curve. Proceedings of the National Academy of Sciences, 68(4), 815-819.

[50] Galibert, F., Finan, T. M., Long, S. R., Pühler, A., Abola, P., Ampe, F., ... & Bothe, G. (2001). The composite genome of the legume symbiont Sinorhizobium meliloti. Science, 293(5530), 668-672.

[51] Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. Pattern recognition letters, 31(14), 2225-2236.

[52] Goloborodko, A., Imakaev, M. V., Marko, J. F., & Mirny, L. (2016). Compaction and segregation of sister chromatids via active loop extrusion. Elife, 5, e14864.

[53] Grassberger, P. (1997). Pruned-enriched Rosenbluth method: Simulations of θ -polymers of chain length up to 1000000. Physical Review E, 56(3), 3682.

[54] Graumann, P. L. (2000). *Bacillus subtilis* SMC is required for proper arrangement of the chromosome and for efficient segregation of replication termini but not for bipolar movement of newly duplicated origin regions. Journal of bacteriology, 182(22), 6463-6471.

[55] Graumann, P. L. (2014). Chromosome architecture and segregation in prokaryotic cells. Journal of molecular microbiology and biotechnology, 24(5-6), 291-300.

[56] Griebel, M., Knapek, S., Zumbusch, G., & Caglar, A. (2013). Numerische Simulation in der Moleküldynamik: Numerik, Algorithmen, Parallelisierung, Anwendungen. Springer-Verlag.

[57] Grosberg, A. Y., Khalatur, P. G., & Khokhlov, A. R. (1982). Polymeric coils with excluded volume in dilute solution: The invalidity of the model of impenetrable spheres and the influence of excluded volume on the rates of diffusion-controlled intermacromolecular reactions. Die Makromolekulare Chemie, Rapid Communications, 3(10), 709-713.

[58] Grosberg, A. Y., & Khokhlov, A. R. (1994). Statistical Physics of Macromolecules (AIP, Woodbury, NY).

[59] Gruber, S., Veening, J. W., Bach, J., Blettinger, M., Bramkamp, M., & Errington, J. (2014). Interlinked sister chromosomes arise in the absence of condensin during fast replication in B. subtilis. Current biology, 24(3), 293-298.

[60] Ha, B. Y., & Jung, Y. (2015). Polymers under confinement: single polymers, how they interact, and as model chromosomes. Soft Matter, 11(12), 2333-2352.

[61] Hacker, W. C., Li, S., & Elcock, A. H. (2017). Features of genomic organization in a nucleotide-resolution molecular model of the Escherichia coli chromosome. Nucleic acids research, 45(13), 7541-7554.

[62] Hajmeer, M., & Basheer, I. (2003). Comparison of logistic regression and neural network-based classifiers for bacterial growth. Food Microbiology, 20(1), 43-55.

[63] Hamelberg, D., Mongan, J., & McCammon, J. A. (2004). Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. The Journal of chemical physics, 120(24), 11919-11929.

[64] Hamelberg, D., de Oliveira, C. A. F., & McCammon, J. A. (2007). Sampling of slow diffusive conformational transitions with accelerated molecular dynamics. The Journal of chemical physics, 127(15), 10B614.

[65] Hammersley, J. M. (1957, July). Percolation processes: II. The connective constant. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 53, No. 3, pp. 642-645). Cambridge University Press.

[66] Hammersley, J. M. (1961, July). The number of polygons on a lattice. In Mathematical Proceedings of the Cambridge Philosophical Society (Vol. 57, No. 3, pp. 516-523). Cambridge University Press.

[67] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

[68] Heermann, D. W. (2011). Physical nuclear organization: loops and entropy. Current opinion in cell biology, 23(3), 332-337.

[69] Helmuth, J. A., Burckhardt, C. J., Koumoutsakos, P., Greber, U. F., & Sbalzarini, I. F. (2007). A novel supervised trajectory segmentation algorithm identifies distinct types of human adenovirus motion in host cells. Journal of structural biology, 159(3), 347-358.

[70] Higgins, N. P., Yang, X., Fu, Q., & Roth, J. R. (1996). Surveying a supercoil domain by using the gamma delta resolution system in Salmonella typhimurium. Journal of bacteriology, 178(10), 2825-2835.

[71] Hirano, T. (2012). Condensins: universal organizers of chromosomes with diverse functions. Genes & development, 26(15), 1659-1678.

[72] Hofmann, A., Mäkelä, J., Sherratt, D. J., Heermann, D., & Murray, S. M. (2019). Self-organised segregation of bacterial chromosomal origins. Elife, 8, e46564.

[73] Hollingsworth, S. A., & Dror, R. O. (2018). Molecular dynamics simulation for all. Neuron, 99(6), 1129-1143.

[74] Hou, T., Zheng, G., Zhang, P., Jia, J., Li, J., Xie, L., ... & Li, Y. (2014). LAceP: lysine acetylation site prediction using logistic regression classifiers. PloS one, 9(2), e89575.

[75] Hsu, H. P., Paul, W., Rathgeber, S., & Binder, K. (2010). Characteristic length scales and radial monomer density profiles of molecular bottle-brushes: Simulation and experiment. Macromolecules, 43(3), 1592-1601.

[76] Huang, J. Z. (2014). An Introduction to Statistical Learning: With Applications in R By Gareth James, Trevor Hastie, Robert Tibshirani, Daniela Witten.

[77] Imakaev, M. V., Fudenberg, G., & Mirny, L. A. (2015). Modeling chromosomes: Beyond pretty pictures. FEBS letters, 589(20), 3031-3036.

[78] Ireton, K., Gunther, N. 4., & Grossman, A. D. (1994). spo0J is required for normal chromosome segregation as well as the initiation of sporulation in Bacillus subtilis. Journal of bacteriology, 176(17), 5320-5329.

[79] Jalal, A. S., Tran, N. T., & Le, T. B. (2020). ParB spreading on DNA requires cytidine triphosphate in vitro. Elife, 9, e53515.

[80] Janczura, J., Kowalek, P., Loch-Olszewska, H., Szwabiński, J., & Weron, A. (2020). Classification of particle trajectories in living cells: Machine learning versus statistical testing hypothesis for fractional anomalous diffusion. Physical Review E, 102(3), 032402.

[81] Japaridze, A., Gogou, C., Kerssemakers, J. W., Nguyen, H. M., & Dekker, C. (2020). Direct observation of independently moving replisomes in Escherichia coli. Nature communications, 11(1), 1-10.

[82] Jeon, C., Jung, Y., & Ha, B. Y. (2017). A ring-polymer model shows how macromolecular crowding controls chromosome-arm organization in Escherichia coli. Scientific reports, 7(1), 1-10.

[83] Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. science, 337(6096), 816-821.

[84] Jun, S., & Mulder, B. (2006). Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. Proceedings of the National Academy of Sciences, 103(33), 12388-12393.

[85] Jun, S., & Wright, A. (2010). Entropy as the driver of chromosome segregation. Nature Reviews Microbiology, 8(8), 600-607.

[86] Jun, S. (2015). Chromosome, cell cycle, and entropy. Biophysical journal, 108(4), 785.

[87] Junier, I., Martin, O., & Képès, F. (2010). Spatial and topological organization of DNA chains induced by gene co-localization. PLoS computational biology, 6(2).

[88] Junier, I., Boccard, F., & Espeli, O. (2014). Polymer modeling of the E. coli genome reveals the involvement of locus positioning and macrodomain structuring for the control of chromosome conformation and segregation. Nucleic acids research, 42(3), 1461-1473.

[89] Jung, Y., Jeon, C., Kim, J., Jeong, H., Jun, S., & Ha, B. Y. (2012). Ring polymers as model bacterial chromosomes: confinement, chain topology, single chain statistics, and how they interact. Soft Matter, 8(7), 2095-2102.

[90] Jung, Y., & Ha, B. Y. (2019). Confinement induces helical organization of chromosome-like polymers. Scientific reports, 9(1), 1-11.

[91] Kamien, R. D. (2002). The geometry of soft materials: a primer. Reviews of Modern physics, 74(4), 953.

[92] Katayama, T., Ozaki, S., Keyamura, K., & Fujimitsu, K. (2010). Regulation of the replication cycle: conserved and diverse regulatory systems for DnaA and oriC. Nature Reviews Microbiology, 8(3), 163-170.

[93] Katz, M. J., & George, E. B. (1985). Fractals and the analysis of growth paths. Bulletin of mathematical biology, 47(2), 273-286.

[94] Kahng, L. S., & Shapiro, L. (2003). Polar localization of replicon origins in the multipartite genomes of Agrobacterium tumefaciens and Sinorhizobium meliloti. Journal of bacteriology, 185(11), 3384-3391.

[95] Kjos, M., & Veening, J. W. (2014). Tracking of chromosome dynamics in live S treptococcus pneumoniae reveals that transcription promotes chromosome segregation. Molecular microbiology, 91(6), 1088-1105.

[96] Koch, R. (1982). The etiology of tuberculosis. Reviews of infectious diseases, 4(6), 1270-1274.

[97] Kowalek, P., Loch-Olszewska, H., & Szwabiński, J. (2019). Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. Physical Review E, 100(3), 032410.

[98] Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26). New York: Springer.

[99] Kwapien, J., & Drozdz, S. (2012). Physical approach to complex systems. Physics Reports, 515(3-4), 115-226.

[100] Lal, M. (1969). 'Monte Carlo'computer simulation of chain molecules. I. Molecular physics, 17(1), 57-64.

[101] Le, T. B., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-resolution mapping of the spatial organization of a bacterial chromosome. Science, 342(6159), 731-734.

[102] Leman, A. R., & Noguchi, E. (2013). The replication fork: understanding the eukaryotic replication machinery and the challenges to genome duplication. Genes, 4(1), 1-32.

[103] Lemon, K. P., & Grossman, A. D. (1998). Localization of bacterial DNA polymerase: evidence for a factory model of replication. Science, 282(5393), 1516-1519.

[104] Lemon, K. P., & Grossman, A. D. (2000). Movement of replicating DNA through a stationary replisome. Molecular cell, 6(6), 1321-1330.

[105] Lemon, K. P., & Grossman, A. D. (2001). The extrusion-capture model for chromosome partitioning in bacteria. Genes & development, 15(16), 2031-2041.

[106] Levitt, M., & Lifson, S. (1969). Refinement of protein conformations using a macromolecular energy minimization procedure. Journal of molecular biology, 46(2), 269-279.

[107] Libby, E. A., Roggiani, M., & Goulian, M. (2012). Membrane protein expression triggers chromosomal locus repositioning in bacteria. Proceedings of the National Academy of Sciences, 109(19), 7445-7450.

[108] Lim, H. C., Surovtsev, I. V., Beltran, B. G., Huang, F., Bewersdorf, J., & Jacobs-Wagner, C. (2014). Evidence for a DNA-relay mechanism in ParABS-mediated chromosome segregation. Elife, 3, e02758.

[109] Lifson, S., & Warshel, A. (1968). Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n - alkane molecules. The Journal of Chemical Physics, 49(11), 5116-5129.

[110] Liu, P., Kim, B., Friesner, R. A., & Berne, B. J. (2005). Replica exchange with solute tempering: A method for sampling biological systems in explicit water. Proceedings of the National Academy of Sciences, 102(39), 13749-13754.

[111] Lucena, D., Mauri, M., Schmidt, F., Eckhardt, B., & Graumann, P. L. (2018). Microdomain formation is a general property of bacterial membrane proteins and induces heterogeneity of diffusion patterns. BMC biology, 16(1), 1-17.

[112] MacLean, A. M., Milunovic, B., Golding, G. B., & Finan, T. M. (2014). Examination of prokaryotic multipartite genome evolution through experimental genome reduction. PLoS Genet, 10(10), e1004742.

[113] Madigan, M. T., & Martinko, J. (2005). Brock Biology of Microorganisms, 11th edn.

[114] Madras, N., & Sokal, A. D. (1988). The pivot algorithm: a highly efficient Monte Carlo method for the self-avoiding walk. Journal of Statistical Physics, 50(1-2), 109-186.

[115] Madras, N., Orlitsky, A., & Shepp, L. A. (1990). Monte Carlo generation of self-avoiding walks with fixed endpoints and fixed length. Journal of Statistical Physics, 58(1-2), 159-183.

[116] Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional Brownian motions, fractional noises and applications. SIAM review, 10(4), 422-437.

[117] Mangiameli, S. M., Cass, J. A., Merrikh, H., & Wiggins, P. A. (2018). The bacterial replisome has factory-like localization. Current genetics, 64(5), 1029-1036.

[118] Marbouty, M., Le Gall, A., Cattoni, D. I., Cournac, A., Koh, A., Fiche, J. B., ... & Nollmann, M. (2015). Condensin-and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. Molecular cell, 59(4), 588-602.

[119] Mardoum, W. M., Gorczyca, S. M., Regan, K. E., Wu, T. C., & Robertson-Anderson, R. M. (2018). Crowding induces entropically-driven changes to DNA dynamics that depend on crowder structure and ionic conditions. Frontiers in physics, 6, 53.

[120] Marenduzzo, D., & Orlandini, E. (2009). Topological and entropic repulsion in biopolymers. Journal of statistical mechanics: theory and experiment, 2009(09), L09002.

[121] Marko, J. F., & Siggia, E. D. (1994). Fluctuations and supercoiling of DNA. Science, 265(5171), 506-508.

[122] Marko, J. F., & Siggia, E. D. (1995). Stretching dna. Macromolecules, 28(26), 8759-8770.

[123] Marko, J. F., & Siggia, E. D. (1995). Statistical mechanics of supercoiled DNA. Physical Review E, 52(3), 2912.

[124] Matsumoto, K., Hara, H., Fishov, I., Mileykovskaya, E., & Norris, V. (2015). The membrane: transertion as an organizing principle in membrane heterogeneity. Frontiers in microbiology, 6, 572.

[125] McCammon, J. A., Gelin, B. R., & Karplus, M. (1977). Dynamics of folded proteins. Nature, 267(5612), 585-590.

[126] Messelink, J. J., Janssen, J., van Teeseling, M. C., Thanbichler, M., & Broedersz, C. P. (2020). Resolving the degree of order in the bacterial chromosome using a statistical physics approach. arXiv preprint arXiv:2002.03880.

[127] Micheletti, C., Marenduzzo, D., & Orlandini, E. (2011). Polymers with spatial or topological constraints: Theoretical and computational results. Physics Reports, 504(1), 1-73.

[128] Miermans, C. A., & Broedersz, C. P. (2018). Bacterial chromosome organization by collective dynamics of SMC condensins. Journal of the Royal Society Interface, 15(147), 20180495.

[129] Migocki, M. D., Lewis, P. J., Wake, R. G., & Harry, E. J. (2004). The midcell replication factory in Bacillus subtilis is highly mobile: implications for coordinating chromosome replication with other cell cycle events. Molecular microbiology, 54(2), 452-463.

[130] Minina, E., & Arnold, A. (2014). Induction of entropic segregation: the first step is the hardest. Soft Matter, 10(31), 5836-5841.

[131] Minina, E., & Arnold, A. (2015). Entropic segregation of ring polymers in cylindrical confinement. Macromolecules, 48(14), 4998-5005.

[132] Minina, E. (2016). Entropic segregation of polymers under confinement.

[133] Misra, H. S., Maurya, G. K., Kota, S., & Charaka, V. K. (2018). Maintenance of multipartite genome system and its functional significance in bacteria. Journal of genetics, 97(4), 1013-1038.

[134] Motnenko, A., Geisel, D., Wagner, M., ..., Lenz, P., & Becker, A. (2021). Spatial organization of the tripartite genome of Sinorhizobium meliloti [*Manuscript in preparation]

[135] Muñoz-Gil, G., Garcia-March, M. A., Manzo, C., Martín-Guerrero, J. D., & Lewenstein, M. (2019). Single trajectory characterization via machine learning. New Journal of Physics.

[136] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, 21.

[137] Nakano, M. M., & Zuber, P. (1998). Anaerobic growth of a "strict aerobe"(Bacillus subtilis). Annual review of microbiology, 52(1), 165-190.

[138] Nelson, P. (2004). Biological physics (pp. 315-332). New York: WH Freeman.

[139] Niki, H., Yamaichi, Y., & Hiraga, S. (2000). Dynamic organization of chromosomal DNA in Escherichia coli. Genes & Development, 14(2), 212-223.

[140] Noble, W. S. (2006). What is a support vector machine?. Nature biotechnology, 24(12), 1565-1567.

[141] Norris, V., & Madsen, M. S. (1995). Autocatalytic gene expression occurs via transertion and membrane domain formation and underlies differentiation in bacteria: a model. Journal of molecular biology, 253(5), 739-748.

[142] Norris, V., Den Blaauwen, T., Cabin-Flaman, A., Doi, R. H., Harshey, R., Janniere, L., ... & Skarstad, K. (2007). Functional taxonomy of bacterial hyperstructures. Microbiology and Molecular Biology Reviews, 71(1), 230-253.

[143] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., ... & Schulten, K. (2005). Scalable molecular dynamics with NAMD. Journal of computational chemistry, 26(16), 1781-1802.

[144] Phillips, R., Kondev, J., Theriot, J., & Garcia, H. (2012). Physical biology of the cell. Garland Science.

[145] Pinto, U. M., Pappas, K. M., & Winans, S. C. (2012). The ABCs of plasmid replication and segregation. Nature Reviews Microbiology, 10(11), 755.

[146] Pelletier, J., Halvorsen, K., Ha, B. Y., Paparcone, R., Sandler, S. J., Woldringh, C. L., ... & Jun, S. (2012). Physical manipulation of the Escherichia coli chromosome reveals its soft nature. Proceedings of the National Academy of Sciences, 109(40), E2649-E2656.

[147] Pereira, M. C. F., Brackley, C. A., Lintuvuori, J. S., Marenduzzo, D., & Orlandini, E. (2017). Entropic elasticity and dynamics of the bacterial chromosome: A simulation study. The Journal of chemical physics, 147(4), 044908.

[148] Polson, J. M., & Kerry, D. R. M. (2018). Segregation of polymers under cylindrical confinement: effects of polymer topology and crowding. Soft matter, 14(30), 6360-6373.

[149] Postow, L., Hardy, C. D., Arsuaga, J., & Cozzarelli, N. R. (2004). Topological domain structure of the Escherichia coli chromosome. Genes & development, 18(14), 1766-1779.

[150] Qi, Y., & Zhang, B. (2019). Predicting three-dimensional genome organization with chromatin states. PLoS computational biology, 15(6), e1007024.

[151] Rapaport, D. C. (2004). The art of molecular dynamics simulation. Cambridge university press.

[152] Reyes-Lamothe, R., & Sherratt, D. J. (2019). The bacterial cell cycle, chromosome inheritance and cell growth. Nature Reviews Microbiology, 17(8), 467-478.

[153] Rechnitzer, A., & van Rensburg, E. J. (2008). Generalized atmospheric Rosenbluth methods (GARM). Journal of Physics A: Mathematical and Theoretical, 41(44), 442002.

[154] Ricca, R. L., & Nipoti, B. (2011). Gauss linking number revisited. Journal of Knot Theory and Its Ramifications, 20(10), 1325-1343.

[155] Roos, N. (2014). Entropic forces in Brownian motion. American Journal of Physics, 82(12), 1161-1166.

[156] Rosenbluth, M. N., & Rosenbluth, A. W. (1955). Monte Carlo calculation of the average extension of molecular chains. The Journal of Chemical Physics, 23(2), 356-359.

[157] Rubinstein, M., & Colby, R. H. (2003). Polymer physics (Vol. 23). New York: Oxford university press.

[158] Russel, S., & Norvig, P. (2013). Artificial intelligence: a modern approach. London: Pearson Education Limited.

[159] Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science, 230(4732), 1350-1354.

[160] Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., ... & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science, 239(4839), 487-491.

[161] Salieb-Beugelaar, G. B., Dorfman, K. D., van den Berg, A., & Eijkel, J. C. (2009). Electrophoretic separation of DNA in gels and nanostructures. Lab on a Chip, 9(17), 2508-2523.

[162] Saxton, M. J. (1993). Lateral diffusion in an archipelago. Single-particle diffusion. Biophysical journal, 64(6), 1766-1780.

[163] Saxton, M. J., & Jacobson, K. (1997). Single-particle tracking: applications to membrane dynamics. Annual review of biophysics and biomolecular structure, 26(1), 373-399.

[164] Schenk, K., Hervás, A. B., Rösch, T. C., Eisemann, M., Schmitt, B. A., Dahlke, S., ... & Graumann, P. L. (2017). Rapid turnover of DnaA at replication origin regions contributes to initiation control of DNA replication. PLoS genetics, 13(2), e1006561.

[165] Schibany, S., Kleine Borgmann, L. A., Rösch, T. C., Knust, T., Ulbrich, M. H., & Graumann, P. L. (2018). Single molecule tracking reveals that the bacterial SMC complex moves slowly relative to the diffusion of the chromosome. Nucleic acids research, 46(15), 7805-7819.

[166] Sevcik, C. (2010). A procedure to estimate the fractal dimension of waveforms. arXiv preprint arXiv:1003.5266.

[167] Sharpe, M. E., & Errington, J. (1998). A fixed distance for separation of newly replicated copies of oriC in Bacillus subtilis: implications for coordination of chromosome segregation and cell division. Molecular microbiology, 28(5), 981-990.

[168] Shendruk, T. N., Bertrand, M., de Haan, H. W., Harden, J. L., & Slater, G. W. (2015). Simulating the entropic collapse of coarse-grained chromosomes. Biophysical journal, 108(4), 810-820.

[169] Sinden, R. R., & Pettijohn, D. E. (1981). Chromosomes in living Escherichia coli cells are segregated into domains of supercoiling. Proceedings of the National Academy of Sciences, 78(1), 224-228.

[170] Snyder, B. (2007). Why the NSF biology budget should be doubled. BioScience, 57(9), 727-728.

[171] Sokal, A. D. (1994). Monte Carlo methods for the self-avoiding walk. arXiv preprint hep-lat/9405016.

[172] Song, D., & Loparo, J. J. (2015). Building bridges within the bacterial chromosome. Trends in Genetics, 31(3), 164-173.

[173] Surovtsev, I. V., & Jacobs-Wagner, C. (2018). Subcellular organization: a critical feature of bacterial cell replication. Cell, 172(6), 1271-1293.

[174] Stavans, J., & Oppenheim, A. (2006). DNA–protein interactions and bacterial chromosome architecture. Physical biology, 3(4), R1.

[175] Strick, T. R., Allemand, J. F., Bensimon, D., Bensimon, A., & Croquette, V. (1996). The elasticity of a single supercoiled DNA molecule. Science, 271(5257), 1835-1837.

[176] Strick, T. R., Allemand, J. F., Bensimon, D., & Croquette, V. (1998). Behavior of supercoiled DNA. Biophysical journal, 74(4), 2016-2028.

[177] Swain, P., Mulder, B. M., & Chaudhuri, D. (2019). Confinement and crowding control the morphology and dynamics of a model bacterial chromosome. Soft matter, 15(12), 2677-2687.

[178] Swinger, K. K., & Rice, P. A. (2004). IHF and HU: flexible architects of bent DNA. Current opinion in structural biology, 14(1), 28-35.

[179] Teraoka, I. (2002). Polymer solutions. John Wiley & Sons, Inc.

[180] Thirumalai, D., Mountain, R. D., & Kirkpatrick, T. R. (1989). Ergodic behavior in supercooled liquids and in glasses. Physical Review A, 39(7), 3563.

[181] Tjong, H., Li, W., Kalhor, R., Dai, C., Hao, S., Gong, K., ... & Alber, F. (2016). Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. Proceedings of the National Academy of Sciences, 113(12), E1663-E1672.

[182] Toro, E., Hong, S. H., McAdams, H. H., & Shapiro, L. (2008). Caulobacter requires a dedicated mechanism to initiate chromosome segregation. Proceedings of the National Academy of Sciences, 105(40), 15435-15440.

[183] Umbarger, M. A., Toro, E., Wright, M. A., Porreca, G. J., Bau, D., Hong, S. H., ... & Shapiro, L. (2011). The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. Molecular cell, 44(2), 252-264.

[184] Val, M. E., Soler-Bistué, A., Bland, M. J., & Mazel, D. (2014). Management of multipartite genomes: the Vibrio cholerae model. Current opinion in microbiology, 22, 120-126.

[185] Val, M. E., Marbouty, M., de Lemos Martins, F., Kennedy, S. P., Kemble, H., Bland, M. J., ... & Mazel, D. (2016). A checkpoint control orchestrates the replication of the two chromosomes of Vibrio cholerae. Science advances, 2(4), e1501914.

[186] van Kuppevelt, D., Meijer, C., Huber, F., van der Ploeg, A., Georgievska, S., & van Hees, V. T. (2020). Mcfly: Automated deep learning on time series. SoftwareX, 12, 100548.

[187] van Raaphorst, R., Kjos, M., & Veening, J. W. (2017). Chromosome segregation drives division site selection in Streptococcus pneumoniae. Proceedings of the National Academy of Sciences, 114(29), E5959-E5968.

[188] van Rensburg, E. J. (2009). Monte Carlo methods for the self-avoiding walk. Journal of Physics A: Mathematical and Theoretical, 42(32), 323001.

[189] Verlet, L. (1967). Computer" experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. Physical review, 159(1), 98.

[190] Viollier, P. H., Thanbichler, M., McGrath, P. T., West, L., Meewan, M., McAdams, H. H., & Shapiro, L. (2004). Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. Proceedings of the National Academy of Sciences, 101(25), 9257-9262.

[191] Vologodskii, A. (2007). Monte carlo simulation of dna topological properties. In Topology in molecular biology (pp. 23-41). Springer, Berlin, Heidelberg.

[192] Wagner, T., Kroll, A., Haramagatti, C. R., Lipinski, H. G., & Wiemann, M. (2017). Classification and segmentation of nanoparticle diffusion trajectories in cellular micro environments. PloS one, 12(1).

[193] Wagner, M., Geisel, D., Motnenko, A., ..., Lenz, P., & Becker, A. (2021). Spatial organization of the fused SmABC mutant of Sinorhizobium meliloti [*Manuscript in preparation]

[194] Walter, J. C., Walliser, N. O., David, G., Dorignac, J., Geniet, F., Palmeri, J., ... & Broedersz, C. P. (2018). Looping and clustering model for the organization of protein-DNA complexes on the bacterial genome. New Journal of Physics, 20(3), 035002.

[195] Wang, M. D., Schnitzer, M. J., Yin, H., Landick, R., Gelles, J., & Block, S. M. (1998). Force and velocity measured for single molecules of RNA polymerase. Science, 282(5390), 902-907.

[196] Wang, X., Llopis, P. M., & Rudner, D. Z. (2013). Organization and segregation of bacterial chromosomes. Nature Reviews Genetics, 14(3), 191-203.

[197] Wang, X., Llopis, P. M., & Rudner, D. Z. (2014). Bacillus subtilis chromosome organization oscillates between two distinct patterns. Proceedings of the National Academy of Sciences, 111(35), 12877-12882.

[198] Wang, X., Tang, O. W., Riley, E. P., & Rudner, D. Z. (2014). The SMC condensin complex is required for origin segregation in Bacillus subtilis. Current Biology, 24(3), 287-292.

[199] Wang, X., & Rudner, D. Z. (2014). Spatial organization of bacterial chromosomes. Current opinion in microbiology, 22, 66-72.

[200] Wang, X., Le, T. B., Lajoie, B. R., Dekker, J., Laub, M. T., & Rudner, D. Z. (2015). Condensin promotes the juxtaposition of DNA flanking its loading site in Bacillus subtilis. Genes & development, 29(15), 1661-1675.

[201] Wang, X., Brandão, H. B., Le, T. B., Laub, M. T., & Rudner, D. Z. (2017). Bacillus subtilis SMC complexes juxtapose chromosome arms as they travel from origin to terminus. Science, 355(6324), 524-527.

[202] Weaver, G. M., Mettrick, K. A., Corocher, T. A., Graham, A., & Grainge, I. (2019). Replication fork collapse at a protein - DNA roadblock leads to fork reversal, promoted by the RecQ helicase. Molecular microbiology, 111(2), 455−472.

[203] Webb, C. D., Graumann, P. L., Kahana, J. A., Teleman, A. A., Silver, P. A., & Losick, R. (1998). Use of timelapse microscopy to visualize rapid movement of the replication origin region of the chromosome during the cell cycle in Bacillus subtilis. Molecular microbiology, 28(5), 883-892.

[204] Weber, S. C., Spakowitz, A. J., & Theriot, J. A. (2010). Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. Physical review letters, 104(23), 238102.

[205] Weik, F., Weeber, R., Szuttor, K., Breitsprecher, K., de Graaf, J., Kuron, M., ... & Holm, C. (2019). ESPResSo 4.0–an extensible software package for simulating soft matter systems. The European Physical Journal Special Topics, 227(14), 1789-1816.

[206] Weng, X., & Xiao, J. (2014). Spatial organization of transcription in bacterial cells. Trends in genetics, 30(7), 287-297.

[207] Weron, A., Janczura, J., Boryczka, E., Sungkaworn, T., & Calebiro, D. (2019). Statistical testing approach for fractional anomalous diffusion classification. Physical Review E, 99(4), 042149.

[208] White, J. H. (1969). Self-linking and the Gauss integral in higher dimensions. American journal of mathematics, 91(3), 693-728.

[209] White, R. P., & Meirovitch, H. (2005). Calculation of the entropy of random coil polymers with the hypothetical scanning Monte Carlo method. The Journal of chemical physics, 123(21), 214908.

[210] Whitfield, T. W., Bu, L., & Straub, J. E. (2002). Generalized parallel sampling. Physica A: Statistical Mechanics and its Applications, 305(1-2), 157-171.

[211] Wiggins, P. A., Cheveralls, K., & Kondev, J. (2010). Strong Intra-Nucleoid Interactions Organize the E. Coli Chromosome into a Nucleoid Filament. Biophysical Journal, 98(3), 658a-659a.

[212] Woldringh, C. L. (2002). The role of cotranscriptional translation and protein translocation (transertion) in bacterial chromosome segregation. Molecular microbiology, 45(1), 17-29.

[213] Wu, F., Japaridze, A., Zheng, X., Wiktor, J., Kerssemakers, J. W., & Dekker, C. (2019). Direct imaging of the circular chromosome in a live bacterium. Nature communications, 10(1), 1-9.

[214] Yildirim, A., & Feig, M. (2018). High-resolution 3D models of Caulobacter crescentus chromosome reveal genome structural variability and organization. Nucleic acids research, 46(8), 3937-3952.

[215] Yin, H., Wang, M. D., Svoboda, K., Landick, R., Block, S. M., & Gelles, J. (1995). Transcription against an applied force. Science, 270(5242), 1653-1657.

[216] Youngren, B., Nielsen, H. J., Jun, S., & Austin, S. (2014). The multifork Escherichia coli chromosome is a self-duplicating and self-segregating thermodynamic ring polymer. Genes & development, 28(1), 71-84.

[217] Zaburdaev, V., Denisov, S., & Klafter, J. (2015). Lévy walks. Reviews of Modern Physics, 87(2), 483.

[218] Zhang, B., & Wolynes, P. G. (2015). Topology, structures, and energy landscapes of human chromosomes. Proceedings of the National Academy of Sciences, 112(19), 6062-6067.

List of Figures

List of Tables

Acknowledgments

At the end of this thesis, there are a few people I would like to thank for their support over the past three years.
First, I would like to thank my PhD supervisor Prof. Dr. Peter Lenz for giving me the opportunity to get involved in the various projects. Your feedback and ideas have always been important impulses for the progress of my work. At the same time, through my participation in various meetings and conferences, you have given me the opportunity to develop personally and gain important insights into the organization of science.
I would also like to express special thanks to Dr. Myroslav Zapukhlyak and Dr. Alexander Orlov. Especially in the technical difficulties of my work, you have been patient and helpful teachers. At the same time, I will always have pleasant memories of our enjoyable office atmosphere and the coffee breaks we had together.
Another important support especially in the final phase of this work were the multiple discussions and the moral support by my friends and colleagues. Many thanks for this to Karl Kraft, Henning Krug, Martin Lellep, Benjamin Räthlein and Friederike Hartwig.
In addition, I would like to thank my parents and my sister for their everlasting love and support not only during this work but in all circumstances. I know that I can always rely on you.
Finally, many thanks to you, Katharina. Not only for your patient proofreading and the countless miles you took to support me, but also for every minute of our time together.

David Geisel - Scientific profile