Modeling Methods in Theoretical Biology

Pre-thesis report

Master's degree in Physics

Eric Behle

Supervisor: Prof. Dr Oliver Ebenhöh
Co-Supervisor: Dr Adélaïde Raguin

Institute for Quantitative and Theoretical Biology
Heinrich-Heine University
Düsseldorf, 29.10.2019

Abstract

Interdisciplinary approaches are of growing interest in Biology, as they allow to address very diverse questions, with new and complementary means. To face specific problems, many different modeling techniques are developed, a remarkable number of which are originated from Physics or applied Mathematics. This pre-thesis report proposes a brief overview over important modeling methods, inspired from current research carried out in the Institute for Quantitative and Theoretical Biology, at Heinrich-Heine University. Each of these techniques is discussed in terms of its domain of application, advantages and limits, together with examples. A selection of four well studied modeling approaches are exposed, ranging from data-driven procedures and static methods, such as flux balance analysis, to deterministic differential equation based models and stochastic algorithms. Following this overview, the main topic of the forthcoming Master thesis is introduced. This includes working with, and extending, a model for the enzymatic degradation of lignocellulosic agricultural residues, from maize plants. The approach will be based on stochastic simulations, precisely a Gillespie algorithm, to mimic, *in silico*, the saccharification of a three dimensional cell wall microfibril.

Keywords: theoretial Biology, modeling methods, stochastic simulations, cell wall, saccharification.

Contents

Part I
Introduction

The aim of research in theoretical Biology is to understand the inner workings of biological processes on a fundamental level, and from this to formulate accurate predictions leading to the design of experiments and applications. Hence, theoretical Biology tends to complement experimental research by offering new perspectives. Also known as modeling, this vivid topic may involve simulating biological systems *in silico*, analyzing large amounts of experimental data and building analytical theories to predict measurable phenomena.
Many established techniques originating from other scientific fields, like Physics or applied Mathematics, can be applied or even extended here. In fact, depending on the topic, the scientific questions, and the available data and biological knowledge, a model usually requires a very specific sort of approach. This document is meant to provide an overview over some of the most important and routinely used modeling techniques, which are presented here following four categories.
Models which mainly focus on the analysis of experimental data are investigated first. These are part of the field of Bioinformatics and include topics such as proteomics, transcriptomics and genome comparison between organisms.
Next, flux balance analysis, which is an important tool for the analysis of genome-scale metabolic networks, is considered, as an example of static models.
Some of the most useful modeling techniques, which are widely used across theoretical sciences in general, are deterministic algorithms, based on ordinary differential equations. They can be applied to a vast amount of problems and simulate the dynamics of a system according to equations defining its behavior.
The last types of models presented here are the ones based on stochastic algorithms. They allow to represent systems *in silico* and mimic their behavior without describing them with equations, but by instead evolving the system following a potentially very large number of steps.
In order to provide some further insight, each of these four modeling approaches is accompanied by examples.
The example presented for the stochastic modeling approach is the one of the forthcoming Master thesis, which will follow the completion of this report. The project, which will focus on simulating the degradation of a plant cell wall microfibril, is further described in the last part of this report. There, an overview of the scientific and industrial interest is provided, as well as the detailed structure of the plant material, a description of the enzymatic degrading agents, and a conclusion, which introduces the numerical model to be developed in the Master thesis.

Part II
Models

A model is usually a simplified representation of a process or a system. In essence it is created so that through it questions can be answered and predictions can be made. In this section, four different modeling approaches are devised.

1 Data-driven: Bioinformatics

The diversity of living systems is extreme. As of 2011, the number of different species of eukaryotes alone was estimated to be roughly (8.7 ± 1.3) million [1]. These, in turn, may each contain thousands of genes in their DNA. As a result of this diversity, when studying living systems, one often ends up with vast amounts of data. Analyzing them and retrieving useful information has become a field in of itself: Bioinformatics.

1.1 The method

Bioinformatics is applied in a wide range of areas. Some of the most important are the following [2]:

- **Phylogenetics:** the study of evolutionary relationships between genes and organisms.
- **Omics:** the characterization and quantification of biomolecules.
- **Systems biology:** the exploration of biological systems by studying the interactions between their components.
- **Functional annotation:** the characterization of genes and their function.
- **Protein structure prediction and homology modeling:** the study of the relationship between amino acid chain composition and the resulting three-dimensional conformation.
- **Sequence alignment:** the analysis of similarities between two or more DNA, RNA or protein sequences.

The methodology between these areas is diverse. However, a common theme in each approach is the processing of large amounts of experimental data into understandable quantities or structures. This can also be helpful if a system is not yet describable in terms of known properties, which is for instance apparent in the area of protein folding simulations. Currently, it is still an unresolved issue to accurately describe the folding process of an arbitrary amino acid chain and from this to predict its tertiary structure. Here, approaches using machine learning algorithms fed with large amounts of experimental data are gaining popularity [3]. In these, the amino acid chain of the unknown protein is compared with those of cataloged ones whose three-dimensional structure is known. From this, a probable tertiary structure is predicted. While these algorithms do not provide a mechanistic explanation for the folding process, they may be able to accurately predict the functional structure, and provide useful information in terms of protein engineering.

One of the main challenges when working with large datasets is to find patterns indicating possible correlations between phenomena. If the data originate from multiple sources, the aspect of processing them into a comparable form can be a source of further difficulty. This is an active area

of research and there are many projects, for example in genomics and proteomics, which focus on building centralized databases [2]. Their goal is for any researcher to be able to obtain data annotated in a standardized fashion. This ensures effortless sharing, discussion and extension of the data ressources.

The transition from Bioinformatics to other types of models originates in the fact that the data alone usually do not allow for a mechanistic model, which can explain observed phenomena. The field is meant to build a strong foundation and guidance from which to develop such types of models using other methods.

1.2 Example: analysis of genome similarity between phages, prophages and bacteria

An example of a data-driven approach can be found in the Master thesis project of Thomas Wenske [4]. It focused on the analysis of the genetic similarities in a large network of phages, prophages and prokaryotes. This type of research is motivated by the increasing finding of bacteria which are resistant to multiple antibiotics. Hence, phage therapies are investigated as a promising alternative treatment. To achieve this, it is necessary to fully understand the genetic correlations between viruses and bacterial cells.
The network presented in the Master thesis includes data obtained from more than 8 000 prokaryotic plasmids and over 10 000 chromosomes. Organisms are grouped into clusters based on genetic similarities. As a part of that project, the network can be visualized as a graph-representation (see Figure 1).

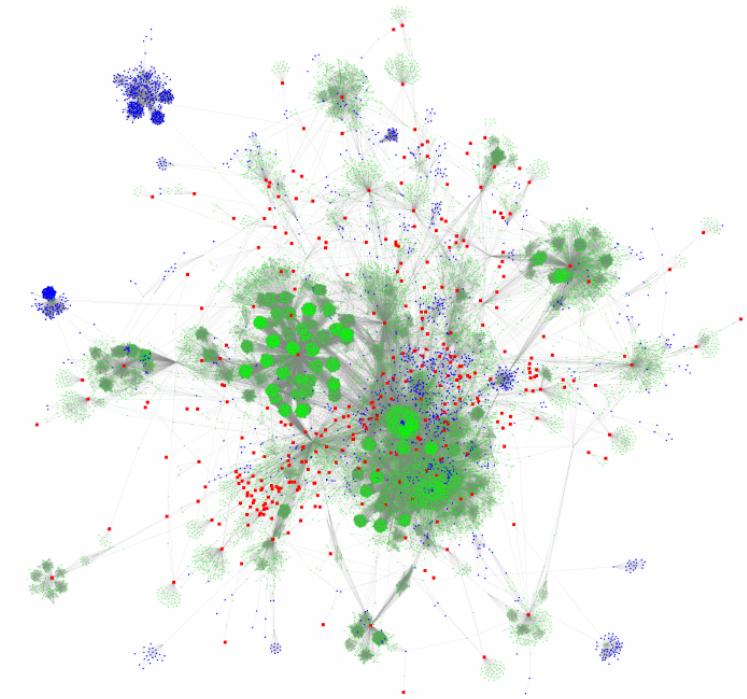

Figure 1: A graphical representation of the biological network from [4]. Genetic similarities between organisms are shown as clusters and edges. Phages are shown in blue, prophages in green and bacteria in red.

The analysis of these types of networks provides new insight into the relation between different organisms. It can also serve as a basis on which to formulate approaches towards influencing the interactions between them.

2 Static: flux balance analysis

Once there is enough data on a system, a model describing its properties can be built. However, the data available may at times not be sufficient yet for formulating this in a dynamic fashion. A common example for an alternative to such dynamic approaches, in the area of genome-scale metabolic models, is the so-called flux balance analysis method [5].

2.1 The method

Genome-scale models are meant to simulate a whole metabolic pathway or even an entire organism. As a result, they include many types of coupled chemical reactions and metabolites. These include so-called exchange reactions, which represent the import or export of metabolites, as well as reactions inside the system, which produce or consume the exchange metabolites and other intermediates (see Figure 2). As an example, a commonly used genome-scale model for *Escherichia coli* contains 1805 distinct metabolites and 2583 reactions [6].

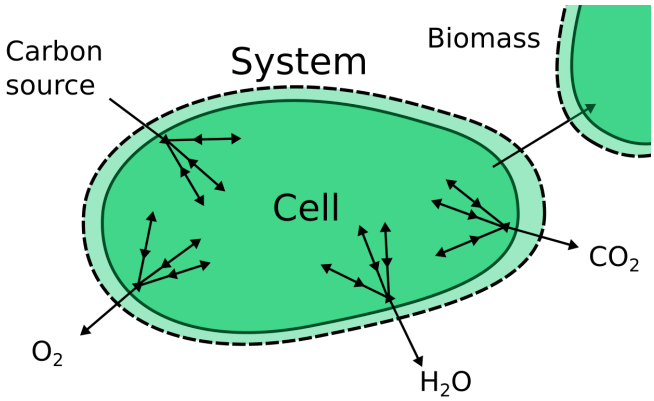

Figure 2: Sketch of a genome-scale system of a biological cell. The arrows originating from or leading outside of the cell represent exchange reactions, while the arrows, which are entirely inside the cell, correspond to intracellular reactions. Adapted from [7].

To perform dynamical simulations with systems of this scale and level of detail, one would encounter two substantial difficulties:

- First, computationally speaking, solving a partially coupled system of thousands of dynamic reaction equations is rather expensive.
- Second, all of the equations require empirical parameters, such as reaction rates, for many of which there is no available experimental data.

The flux balance analysis approach bypasses both of these problems by making two assumptions: steady-state conditions and optimization of the fluxes towards an objective. As a consequence, this formulation is meant to assume a static system and not to simulate dynamics.
The first assumption drastically simplifies the approach. Instead of looking at the reaction-driven time evolution of metabolite concentrations in the system, the overall metabolite concentrations are assumed to be constant, and the flux of each reaction at steady state conditions is investigated. This alleviates the need for knowing all kinetic parameters and makes it possible to formulate the problem as a system of linear equations:

$$A \cdot \vec{v} = 0$$

Here A denotes an $M \times N$ stoichometry matrix, which contains the stoichometric coefficients for all M metabolites regarding the N reactions of the system. $\vec{v}$ is a vector containing the (unknown) flux value for each reaction.
Since most reactions only contain a small number of metabolites ($m \ll M$), A is usually a so-called sparse matrix, which makes solving the system computationally relatively inexpensive. A remaining problem is the fact that the number of reactions N, and therefore the length of $\vec{v}$, is usually larger than the number of metabolites M. The system then contains more unknowns than equations and does not have a unique solution. While this is partially reduced by imposing lower and upper boundaries on the flux of each reaction, one still ends up with a range of solutions, named constrained solution space.
This is when the optimization assumption comes in. It is based on the postulate, that the biological system is optimized towards maximizing or minimizing certain reaction fluxes. A common example of this is the maximization of the so-called biomass function. This is a reaction included in genome-scale models of entire organisms, which acts as a measure of cellular growth. Imposing the constraint of maximizing the biomass reaction's flux significantly reduces the size of the solution space, such that discrete solutions can be selected.
Once a solution is found, it can be analyzed extensively: of particular interest are the exchange reactions, which define influxes and outfluxes of metabolites such as carbon sources, oxygen, carbon dioxide, etc. (see Figure 2).
Usually, solutions are generated many times using different constraints, such as different flux-bounds for certain influx reactions (see section II.2.2).

As mentioned above, the flux balance analysis method has a number of advantages, mainly the significant reduction of computational effort compared to dynamic models, and the reduced need for specific experimental data.
These advantages are however balanced by some trade-offs. The first one lies in the fact that the assumption of steady-state conditions does not always reflect experimental conditions. More significant than this is the assumption that one knows the goal towards which the system is optimized. For example, not all organisms behave in a way that maximizes their biomass production rate. Finding the form of the objective function, which best describes a specific system, is a significant difficulty when using flux balance analysis.

2.2 Example: analysis of a genome-scale model

To provide an example of the application of the flux balance analysis method, some data on my previous and preliminary project, regarding the analysis of a genome-scale model of *Escherichia coli* [6], are used.

First, the behavior of the model for different influxes of a carbon source, namely glucose, was studied. The glucose influx was increased incrementally and for each step the in- and outfluxes of the system were measured. Some examples of in- and outfluxes can be seen in Figure 3.

Carbon source	Carbon atoms	Degree of reduction per carbon
Glucose	6	4
Arabinose	5	4
Sorbitol	6	$4.\overline{3}$
Succinate	4	3

Table 1: The number of carbon atoms and the degree of reduction per carbon atom for four carbon sources, which were investigated as part of my analysis of a genome-scale model of *Escherichia coli* [6].

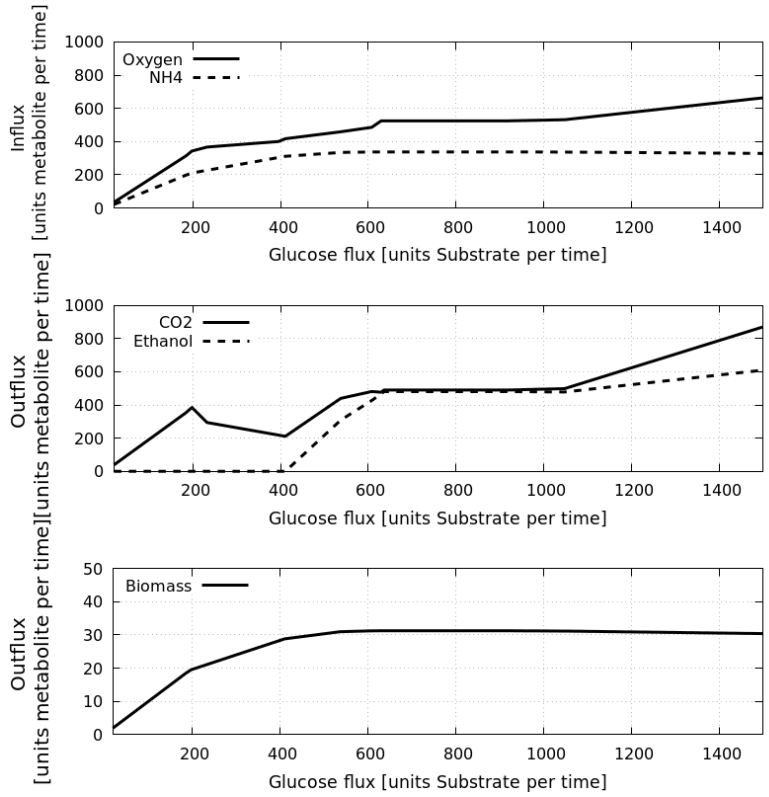

Figure 3: In- and outfluxes for some metabolites, as well as biomass flux in a genome-scale model of *Escherichia coli* [6] at varying glucose influx.

As can be seen in Figure 3, the relationship between incoming glucose and produced metabolites such as CO_2 or ethanol is linear only within specific glucose influx intervals.

The main goal of the project was to analyze thermodynamic properties of *Escherichia coli* cells with regard to the efficiency of substrate metabolism. As a part of this the ratio between carbon used for catabolism (energy supply) and anabolism (growth) was investigated for four different carbon sources, which are listed in table 1.
To analyze each carbon source, the fluxes of the exchange reactions were set such that carbon could not enter the system from other sources. Then the influx of the source investigated was incrementally increased similarily to the increase of glucose in Figure 3.

The comparision between catabolism and anabolism was done by assigning all carbon atoms entering via the carbon source to either one of the two sub metabolisms. The carbon used in the biomass reaction was attributed to anabolism. As the degree of reduction of biomass, roughly 4.9 per carbon atom, is usually higher than that of the carbon source, some of the available carbon has to be oxidized in order to produce biomass. Hence, to maintain the degree of reduction balance,

the carbon of part of the produced CO_2 was also attributed to anabolism. Any remaining carbon was attributed to catabolism, and the fraction used for anabolism for four different carbon sources was recorded in Figure 4.

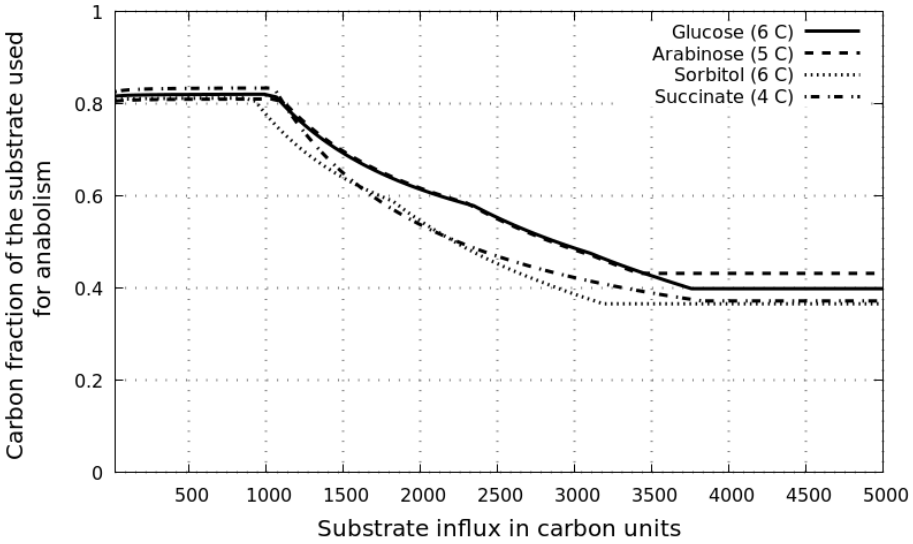

Figure 4: Fraction of carbon used for anabolism for four different sources, whose influx values are increased. For better comparability, the influx is shown in terms of carbon units.

As can be seen in Figure 4, the four carbon fractions behave similarily for low influx values, but slightly diverge for higher ones. In addition, for all sources investigated, the fraction of carbon used for anabolism is much lower at high influxes than at low ones.

3 Deterministic: ordinary differential equations

In contrast to the previous models, dynamic deterministic algorithms belong to the category of those which simulate the time evolution of a system. They often work on a basis of one or more ordinary differential equations (ODEs). Solving such a model's equations provides, usually, a continuum of solutions, which distinguish themselves through their sole dependence on initial conditions. Once these initial conditions are specified, the dynamics is completely defined, and a unique solution can be found. In particular, this means that, excluding digitization and numerical errors, any simulation run with the same starting parameters results in exactly the same outcome.

3.1 The method

As any model, equation based models may provide a simplified, coarse-grained, description of a process, for example in cellular growth laws [8]. In the latter example, detailed mechanisms are left out in order to reduce complexity and computational cost. In other cases however, exact interactions on the molecular level are of interest [9]. Then, the resolution, and as a result the complexity of the equations, is scaled up to atomic precision. Conversely, the size of the system itself is limited to very small systems. The complexity of many ODE-based models ranges between these two extremes.
Some simple systems of ODEs can be solved analytically. However, the complexity of the equation set very quickly increases as more realistic details are included, and numerical approaches are required in the vast majority of the research level cases. These numerical approaches need to be evaluated based on the relation between accuracy and computational cost. The importance of this aspect can be appreciated when examining the Euler method, one of the simplest numerical algorithms for solving an ODE. Consider the following ordinary differential equation:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = ax(t)$$

with $a \epsilon \mathbb{R}$. The Euler algorithm proposes discretizing the differential operation on the lefthand side via a finite difference approximation:

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} \approx \frac{x_{n+1} - x_n}{\delta t}$$

Here, $x_n = x(t)$ and $x_{n+1} = x(t + \delta t)$, with $\delta t \ll 1$. The resulting expression can be rearranged to provide the value of x_{n+1} if x_n is known:

$$x_{n+1} \approx x_n + a \cdot \delta t \cdot x_n$$

If initial conditions are known, the time evolution can, in principle, be simulated with low computational effort. The trade-off lies in precision and stability. During each time step, the finite difference approximation leads to an error on the order δt^2. As this accumulates, the numerical solution deviates further and further from the analytical one. This can be seen very well when comparing numerical solutions of the simple harmonic oscillator equation: $m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -kx$, with its analytical solution: $x(t) = A \cdot \cos(\sqrt{\frac{k}{m}} \cdot t + \phi)$, where $A = 1$, $k = 1$ and $m = 1$ in this example (see Figure 5).

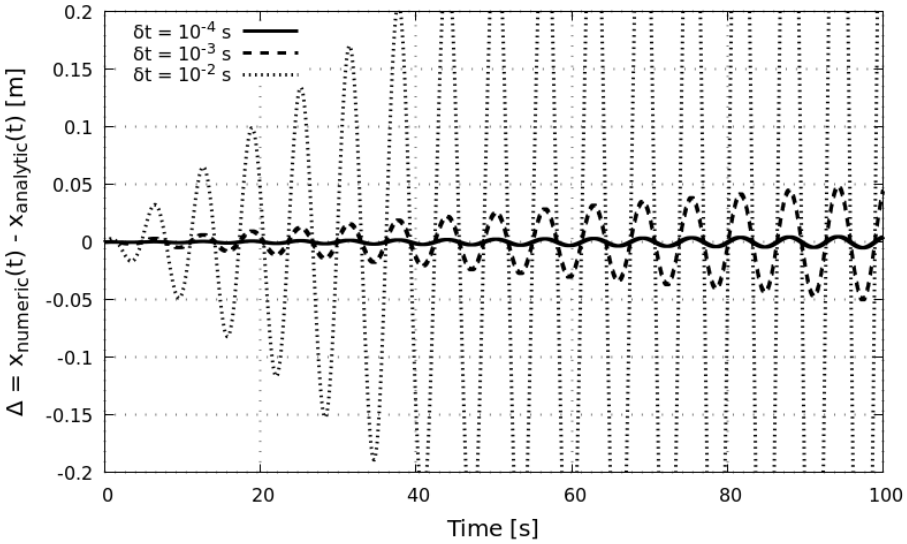

Figure 5: Difference Δ between the analytic solution and numerical solution of the simple harmonic oscillator generated via the Euler algorithm. Each curve represents a different size of integration time step.

Other numerical methods exhibit more precision, for example the Runge-kutta algorithms, which use multiple positions within the interval $[t, t + \delta t]$, leading to a more precise value for x_{n+1} [10]. These numerical algorithms are more expensive in computational cost however. Since the computational cost increases strongly with rising accuracy, the solution method for each system must be tuned to the available computational power and the research interest.

As mentioned earlier, ODE-based models are relevant to tackle very diverse real systems which can have a wide range of structure and complexity. To illustrate this, two examples are considered.

3.2 Example 1: the Lotka-Volterra equations

The Lotka-Volterra equations provide a simplified way to describe the population dynamics of predator-prey interactions, for example wolves and rabbits. They represent a continuous approximation of a discrete problem and read as follows:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha x - \beta xy$$

$$\frac{\mathrm{d}y}{\mathrm{d}t} = \delta xy - \gamma y$$

where x represents the number of prey and y the number of predators in the system. α parametrizes the growth rate of the prey, γ the starvation rate of the predators, and β and δ describe the effects of the interaction between the two species on the corresponding population. The solution for non-zero and non-negative parameters results in a periodic evolution of the two populations, where the predator population evolves phase-shifted in relation to the prey (see Figure 6).

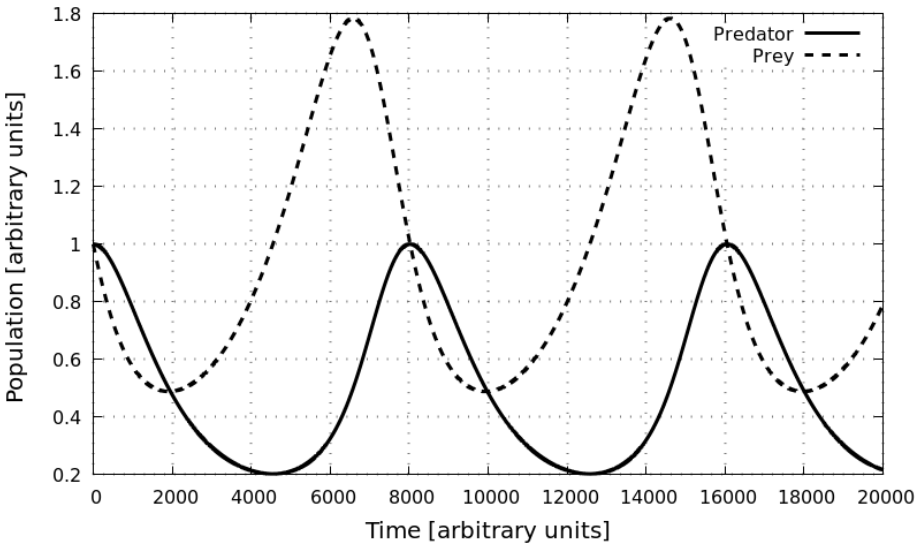

Figure 6: Time evolution of the prey and predator populations in the Lotka-Volterra scheme. Parameters were set as follows: $\alpha = \frac{2}{3}$, $\beta = \frac{4}{3}$, $\gamma = 1$, $\delta = 1$.

3.3 Example 2: molecular dynamics

Molecular dynamics (MD) simulations are of interest when high resolution is required. As the name suggests, they simulate the dynamics of a system by solving Newton's equations of motion at molecular, or even atomic, level. This technique enables very detailed analysis of molecular interaction processes, for instance in the structure of proteins (see Figure 7).

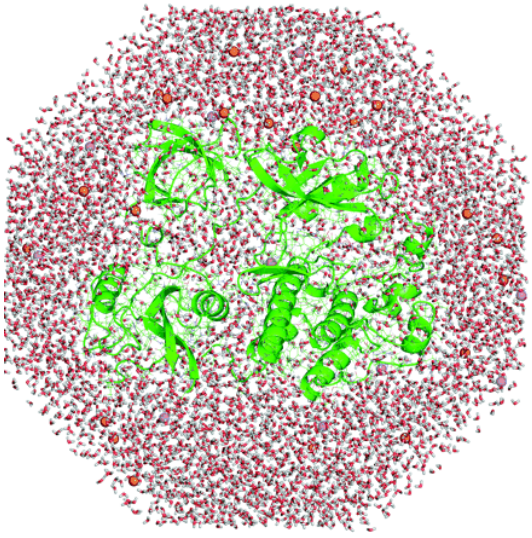

Figure 7: Visualization of an MD simulation of a protein in aqueous solution [9].

One of the downsides of MD simulations is their extremely high computational cost. The system from [9] contained roughly 50 000 atoms. Computation of only 1 ns of its dynamics took 4 days at the time of publication, corresponding to timescales different by 14 orders of magnitude. The relationship between the system size and the interval of real time which can be simulated is therefore bounded by the currently available computation power.
MD simulations are also limited when dealing with systems where the particle number is not constant. Since the dynamics is based on physical equations, the process of adding or removing a particle is often not physically valid, which leads to inconsistencies. When facing such difficulties, one could rather consider stochastic approaches, which will be discussed in the following section.

4 Stochastic

Some processes, for example the rolling of a die or the radioactive decay of an atom, cannot be described accurately by deterministic models, because they contain some inherent randomness. However, even though the exact dynamics of the system cannot be modeled, its average behavior is often of interest. It can also be that the computational cost of using an ODE-based approach is too high. These points support the use of stochastic models, and two common types of algorithms, Monte Carlo and Gillespie, are described below.

4.1 Monte Carlo algorithms

Monte Carlo algorithms are some of the first stochastic algorithms [11] and possibly the most common.

4.1.1 The method

The specific method varies, but the main point is stepwise sampling of random values from a distribution of system states.
To put this in less abstract terms, consider the problem of finding the minimum of a function. A simple Monte Carlo algorithm would entail randomly picking and comparing function values from a given interval. If at each step the minimum is set as the smaller of the previous and current function value, a value close to the actual minimum is reached after a sufficient number of steps. A more refined Monte Carlo algorithm can also include biases towards certain system properties. In the present example this could mean investigating the gradient of the function at the point picked and chosing the next point accordingly. While this may lead to finding a minimum faster, care needs to be taken to avoid local minima, i.e. by allowing the next value to be chosen in a sufficiently large interval.

4.1.2 Example: estimating π

A Monte Carlo algorithm can be used to obtain an estimate of the value of π. The basis of this is the ratio between the area of a circle (denoted A_{circle}) and the one of a square (denoted A_{square}), where the diameter of the circle is equal to the length of the sides of the square (see Figure 8).

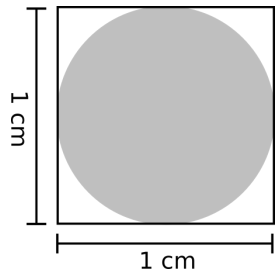

Figure 8: Sketch of the system used to calculate π.

This obeys the relation

$$\frac{A_{circle}}{A_{square}} = \frac{\pi}{4} \qquad (\star)$$

The lefthand side of equation $\star$ is also the probability of a randomly chosen point on the square to be inside the circle. If N points are randomly chosen, and N_c is the number of these which are inside the circle, then

$$\lim_{N\to\infty} \frac{N_c}{N} = \frac{\pi}{4} \qquad (\star\star)$$

Equation $\star\star$ can be used to approximate the value of π by sampling a large number of random points, checking whether they are in the circle and calculating the ratio $\frac{N_c}{N}$. N needs to be sufficiently large here, in order to increase the accuracy. Figure 9 shows results from a simulation during which N was increased in a domain from 100 to 500 000.

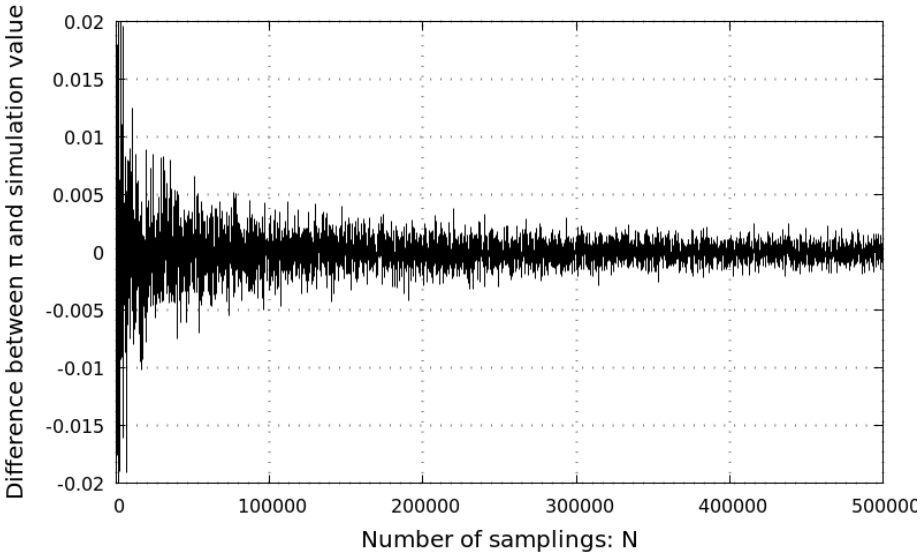

Figure 9: Difference between the value of π provided by the cmath C++ library and a value obtained from a Monte Carlo simulation for different numbers of samplings: N. Each simulation value of π is an average over 10 simulation runs.

For each number of samplings N, the value of π was calculated over 10 simulation runs. The resulting average value of the 10 runs was then compared to the π value saved in the cmath C++ library. As can be seen in Figure 9, initially the error drops quickly, but then it levels out. Since the computational cost also increases with N, this is a good reason for choosing a cutoff value, also allowing to estimate the error resulting from the numerical strategy.

Another important aspect of sampling also becomes apparent here: the generation of random numbers. Computationally generating a truly random sequence of numbers is impossible. Therefore, pseudo-random numbers are actually generated instead. There are algorithms, initialized by a seed value, that return a sequence of seemingly uncorrelated numbers. While these are not truly random, they are usually sufficient. From a practical standpoint, being able to use the same sequence of "random" numbers multiple times by using similar seeds can help investigate and eliminate errors.

4.2 The Gillespie algorithm

4.2.1 The method

The Gillespie algorithm can be viewed as an extension of Monte Carlo algorithms [12]. Yet, it contains two key differences.

The first one is the fact that at each step the list of changes (also called reactions), which are currently possible within the rules and restrictions of the model, is updated. The new change to

be made during the step is then randomly picked within this list. As a consequence, all changes selected can, for sure, be performed. This is different from the situation in regular Monte Carlo algorithms, where the randomly selected change is attempted but not always doable. A way to visualize the difference is to consider the so-called Totally Asymmetric Simple Exclusion Process (TASEP). It consists of a gas of hard sphere particles that move on a one-dimensional lattice. As depicted in Figure 10, particles enter the lattice at a rate α from the left and exit at a rate β to the right, while particles in the lattice move to the right at a rate γ. Movement towards the left is not permitted, and a particle can only move, if the site next to it is empty.

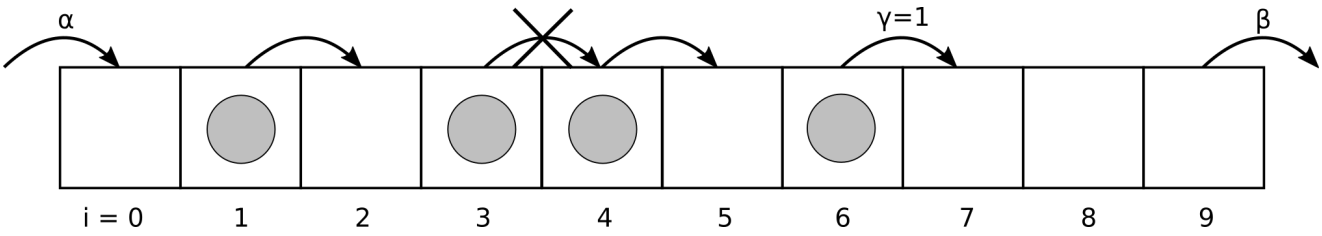

Figure 10: Sketch of a TASEP segment: a gas of hard spheres moving on a one-dimensional lattice.

When simulating the system's behavior using a Monte Carlo algorithm, there are two choices. In the simplest algorithm, a lattice site is sampled from the list of all sites at each step. Then, it is checked if the site contains a particle, and finally, if the particle can move. As there may be many empty lattice sites, this can lead to a large number of rejected steps. When applying the second, more refined algorithm, the sampling is done only from the list of particles on the lattice. This reduces the number of un-doable steps, because empty lattice sites are no longer sampled. However, it is still possible that the sampled particle is unable to move, in which case the step is also not accepted.
When using the Gillespie algorithm on the other hand, the sampling is done, not from the list of particles or lattice sites, but from the list of currently possible moves. In this way, any change made during a step is always possible and is therefore always accepted. This difference in sampling strategy can lead to drastic improvements in terms of computational effort. Using the example from Figure 10, the simple Monte Carlo algorithm would lead to sampling from a list of ten sites, seven of which would be rejected. The more refined algorithm would lead to sampling from a list of four particles, one of which would be rejected. Finally, in the Gillespie algorithm, the sampling would be done from a list of five moves (the un-crossed arrows in Figure 10), all of which would be accepted.

The second, and even more important difference, lies in the calculation of the "real" time. In contrast to regular Monte Carlo algorithms, in which each step lasts one unit of time, when using the Gillespie algorithm, at each step not only the change to the system, but also the time it takes, is calculated. This value is determined randomly, but the interval from which it is sampled is specified by using known parameters connected to the dynamics of the system. The consequence is that, on average, the "real" time inside the simulation corresponds to the time passing in the biological system [12]. This is a considerable advantage, as it improves the comparability between the simulation and corresponding experimental data.

4.2.2 Example: lignocellulose degradation

An application of the Gillespie algorithm will be developed during the Master project to which this pre-thesis leads. A detailed introduction to this project is provided in the next section.

Part III
Modeling lignocellulose degradation

The forthcoming project of the Master thesis is introduced here. It focuses on a stochastic model for the enzymatic degradation of lignocellulosic residues, more specifically for maize plants.

1 Scientific and industrial interest

Utilizing renewable ressources such as materials from our biosphere is an inviting alternative to face the worldwide challenge of energy supply. However, production capabilities of bioenergies are limited especially in light of growing global demand, and finding efficient ways of utilizing materials provided by industries such as agriculture is crucial. Substances which are currently considered waste, such as plant components not applicable for animal or human consumption, are of particular interest. They often contain large amounts of chemical energy, for example inside sugars bound in indigestible polymers. Hence, extracting these energy-carriers for use in industries like biofuel production is tempting.
Some efficient extraction methods can be found in microorganisms, which are specialized towards plant degradation [13]. In addition to the biological processes, diverse chemical treatments can be performed for the extraction of valuable sugars. Yet both biological and chemical options are so far costly, and the adaptation of these types of methods to large-scale industrial operation is an area of active research. To complement these experimental approaches, in the Master thesis a theoretical model based on numerical simulations will be developed, in order to mimic, *in silico*, the process of breaking down plant cell wall components into single glucose molecules, the so-called saccharification process.

2 Plant material structure

The extraction efficiency, more precisely the saccharification yield, for any type of plant material depends on both the material composition and structure. The transformation process from raw plant material to valuable sugars is made in several steps. The first main task is to break down big aggregates into smaller ones.

2.1 Multicellular

Plants consist of connected cells. The multicellular structure strongly impacts the properties of the material [14]. For example, in wood the cells are arranged in a honey-comb like fashion, while in softer tissue like apples or potatoes they produce a liquid-filled foam-like shape. A wide range of composites between these two arrangements exists. As a result, the mechanical properties of plant material may span many orders of magnitude. Young's modulus, for example, which is a measure for the response of a material to external strain, can range from 0.3 MPa for parenchyma tissue (stems, leaves, roots, etc.) to 30 GPa in the densest palms [14].
The separation of multicellular aggregates into smaller components can be done using mechanical methods and does not necessarily require chemicals or enzymes. The next task is then to break down the main structural component of the cells: the cell wall.

2.2 Cell wall

The cell wall consists of multiple layers [14], each of which contain different amounts of three main components: cellulose (C), hemicellulose (H) and lignin (L). A sketch of the structure can be found in Figure 11.

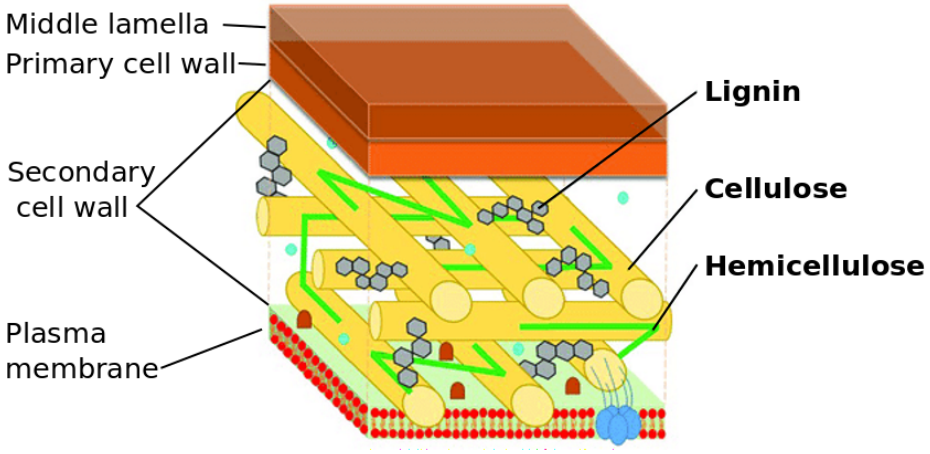

Figure 11: Sketch of the molecular structure of a typical cell wall fragment. Modified from [15]

.

The microstructure, meaning the arrangement of C, H and L, varies strongly between different cell types, leading to an even wider range of mechanical properties than from the multicellular arrangement alone. These three building blocks constitute the materials of interest as they can be further transformed into chemical valuables.

3 Cell wall composition

The mechanical properties of the cell wall originate from rigid cellulose fibres which are reinforced by a matrix of hemicellulose and lignin [14].

3.1 Cellulose

Cellulose is the most abundant organic polymer found on earth. It is usually composed of between 7 000 and 15 000 glucose molecules, which are alternately rotated by 180° [14]. Since a monomer denotes a structure which is exactly repeated many times in a polymer, as a result of the rotation a cellulose monomer technically consists of two glucose molecules, which are $\beta\,1 \rightarrow 4$ linked. The monomer is called cellobiose (see Figure 12).

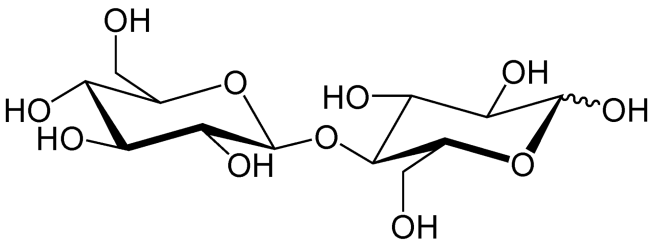

Figure 12: Molecular representation of a cellulose monomer: cellobiose.

Due to its strong mechanical properties, cellulose represents a favorable structural backbone. For further stabilization, multiple cellulose polymers are aligned to form wall-reinforcing microfibril

bundles. Microfibril bundles contain two types of domains: amorphous and crystalline. The amorphous regions show no pattern in the arrangement of the polymers, while the crystallized sections are highly ordered. These different arrangements directly impact the degradability of the material (for more on this see section III.4).

Cellulose is a plant material of very high interest and is already widely used in the paper industry.

3.2 Hemicellulose

Hemicellulose consists of short and amorphous chains of various sugar molecules [14]. It is therefore a polymer with multiple types of monomers, whose collective number per polymer lies between 500 and 3000. As a result of the heterogenity in monomers, hemicellulose is divided into four basic types [16]:

- xylans
- mannans
- β-glucans
- xyloglucans

The mannans are further separated into galactomannans and glucomannans. Each of the four basic types is characterized by different ratios of sugar monomers, as well as differences in their three-dimensional structure. Figure 13 provides a comparison.

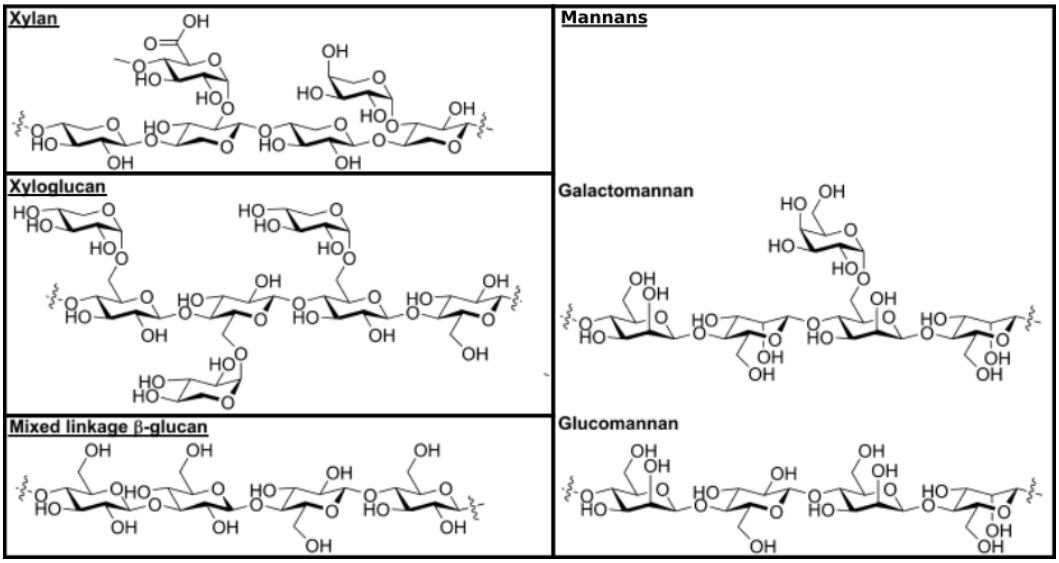

Figure 13: Strucural and compositional comparison between the main types of hemicellulose. The mannans have been further divided into galactomannans and glucomannans. Modified from [17].

Due to the diversity of hemicellulose structure and composition, it can have a wide range of properties. Its branched structure results in a progenity for entanglement with other polymers, making it a useful matrix polymer.
Hemicellulose is easier to degrade than cellulose, leading to higher saccharification yield. It is hence highly sought after in terms of sugar extraction.

3.3 Lignin

Lignin is the second most abundant biopolymer after cellulose [18]. It is a manifold branched polymer which is best known for being a strong contributor to the mechanical properties of wood. Lignin structure is mainly composed of three monomers (monolignols), although a multitude of other molecules have also been identified [18]. The most common monolignols are three alcohol molecules derived from phenylpropane, differing only in their degree of methoxylation (see Figure 14).

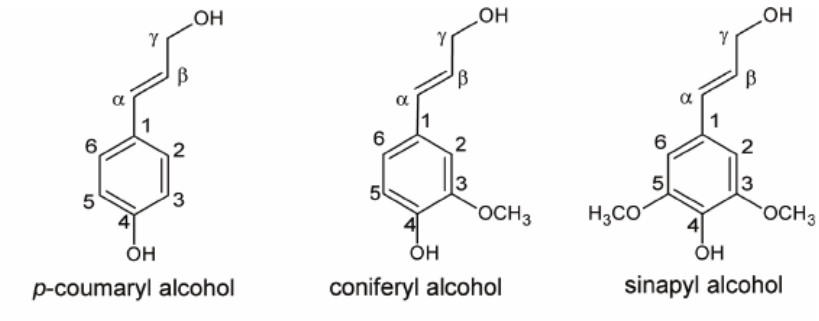

Figure 14: The three main monolignols contained in lignin polymers [19].

Depending on the population of the different monomers in the overall structure of the lignin polymer, the resulting material has different mechanical and chemical properties.
In terms of saccharification, lignin is a hinderance, because it inhibits the enzymatic degradation of cellulose and hemicellulose [20]. There are many other areas however, in which processed lignin may prove valuable [21].

4 Degradation and enzymatic activity

The degradation of the cell wall components is the overall process that consists in breaking down rigid polysaccharides like hemicellulose and cellulose, from which valuable chemicals (mainly sugars) can be extracted.

4.1 Lignin treatment

As written before, in the saccharification process lignin is a burden that needs to be treated to be removed. While some enzymes originating from fungi can degrade lignin [22], further research is required, before these can be scaled-up to industrial level. Therefore, so far, chemical treatments are the main method to suppress lignin.

4.2 Saccharification

Saccharification is the process of enzymatic hydrolysis of cellulose or hemicellulose into glucose monomers. It is a crucial processing step, both in nature and in industry.

The process is carried out by three categories of enzymes, which work in conjunction: exoglucanases, endoglucanases and β-glucosidases [13]. The exoglucanases hydrolyze monomer-bonds close to either the reducing or nonreducing end of a cellulose or hemicellulose polymer [23]. This leads to splitting off of either glucose or cellobiose. The endoglucanases in turn are able to hydrolyze bonds further along the polymer chain and can thereby split longer polymers into shorter

ones. This also provides exoglucanases with more attachment points. β-glucosidases split off single glucose molecules from very short chains containing only up to 7 monomers. They are however inhibited by glucose, thereby limiting efficiency. As mentioned above, lignin is an undesired component for saccharification purposes. This is attributed to its inhibitory action on the enzymes [20], which are inactivated and then cannot contribute further to the saccharification process.

Saccharification of cellulose also depends on its crystallinity, meaning the ratio between amorphous and crystallized regions in the microfibrils [24]. Lower yields were observed for higher crystallinity. Since a substantial amount of crystalline cellulose is found in cell wall materials, this needs to be considered. Hemicellulose in turn can be degraded more efficiently in this regard.

5 The model: stochastic simulations

The topic of the master thesis is the degradation of cell wall components, in particular from Maize plants. Towards this, a numerical model will be developed and theoretical predictions will be compared to experimental data.
In the model, the substrate consists of a cell wall complex composed of units of cellulose, hemicellulose and lignin, whose initial distribution can be varied. The substrate complex is a three-dimensional structure, surrounded by enzymes with variable concentrations.
Each simulation step consists in the random selection of an enzymatic reaction for which the diffusion of enzymes is considered to be infinitely fast and only the change on the substrate structure and composition is recorded (in three dimensions). Depending on the substrate (hemicellulose or cellulose), the enzyme kinetics can vary. Lignin only has an inhibitory effect. The amount of glucose released from the complex over time is tracked and can be compared to available experimental data. Additionally, predictions about the structure and composition of different plant materials, as well as favorable enzyme concentrations and ratios can be made.

Degradation of a subunit of the substrate complex not only depends on its nature or the enzyme concentration, but also on whether the unit is actually found on the outer surface of the strucure, or shielded by other, not yet degraded units. The model was therefore devised in terms of stochastic simulations, because it is difficult to take the structural composition of the cell wall into account when trying to use a deterministic approach. This would require a molecular dynamics type of model, where each enzyme would have to be tracked individually, leading to the expensive simulation of many irrelevant steps in which nothing except for the movement of the enzymes happens. Within the stochastic scheme, enzyme concentrations and resulting reaction probabilities can be used instead, without changing the average behavior of the system. Importantly, the difficulties arising from the three-dimensional structure of the substrate complex can also be adressed efficiently with the Gillespie algorithm. Using it has the further advantage of providing an estimate for the time passing in the biological system.

Part IV
Conclusion

Active improvement of established modeling techniques, as well as development of new ones is crucial for the advancement of theoretical science. It is also important to encourage the exchange of knowledge between different fields, as many forms of inspiration may arise from different viewpoints.
This is evident in the fact that, even though some of the models presented during the course of this pre-thesis originate from other scientific fields, they represent important tools for theoretical Biology as well. Knowing them not only provides a broad spectrum of potential applications but also enables designing new models, taylored to a system of interest. The applicability of these models is versatile: if a new phenomenon is observed, for which there are many data but no mechanistic hypotheses, a data-driven model can help to highlight correlations which may lead to a deeper understanding. In other cases, data on some parts of the dynamics may be missing. Then one may still be able to build a meaningful model by reformulating the approach into a static setting. When building dynamic models, both deterministic methods using ordinary differential equations and stochastic algorithms have specific domains of relevance. Here it can also be of interest to compare models which simulate the same system but are based on different approaches.

The model to be developed in the forthcoming Master thesis represents a promising situation for using a stochastic algorithm. In addition, the three-dimensional nature of the simulated cell wall structure can be simulated very efficiently with a Gillespie algorithm as compared to, for instance, a Monte Carlo algorithm. The project should also provide deeper insights into opportunities regarding utilization of plant material, which has both scientific and industrial implications.
Analyzing the saccharification yield for different microfibril configurations *in silico* and comparing it to experimental data may provide insight into the *in vivo* structure of these aggregates. In addition, the model will highlight the synergetic action of the three types of enzymes involved in saccharification and may reveal enzyme ratios which are ideal for maximizing saccharification yields. These, in turn, can be tested in future experiments, leading to a further test of the model's predictive power. Lastly, the model might provide further understanding of the role of lignin in terms of inhibiting the saccharification process.

References

[1] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on earth and in the ocean?," *PLoS biology*, vol. 9, p. e1001127, Aug 2011.

[2] W. J. S. Diniz and F. Canduri, "Review-article bioinformatics: an overview and its applications.," *Genetics and molecular research : GMR*, vol. 16, Mar 2017.

[3] L. Wei and Q. Zou, "Recent progress in machine learning-based methods for protein fold recognition.," *International journal of molecular sciences*, vol. 17, Dec 2016.

[4] T. Wenske, "Association network of phage and prophage protein families," Master's thesis, Heinrich Heine University, 2019.

[5] J. D. Orth, I. Thiele, and B. . Palsson, "What is flux balance analysis?," *Nature biotechnology*, vol. 28, pp. 245–8, Mar 2010.

[6] J. D. Orth, T. M. Conrad, J. Na, J. A. Lerman, H. Nam, A. M. Feist, and B. . Palsson, "A comprehensive genome-scale reconstruction of escherichia coli metabolism2011," *Molecular Systems Biology*, vol. 7, no. 1, p. 535, 2011.

[7] U. von Stockar and J.-S. Liu, "Does microbial life always feed on negative entropy? thermodynamic analysis of microbial growth," *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, vol. 1412, no. 3, pp. 191 – 211, 1999.

[8] A. Weisse, D. Oyarzn, V. Danos, and P. Swain, "Mechanistic links between cellular trade-offs, gene expression, and growth," *PNAS*, vol. 112, pp. 1416533112–, 02 2015.

[9] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function," *Proceedings of the National Academy of Sciences*, vol. 102, no. 19, pp. 6679–6685, 2005.

[10] G. Lehmann, "Integration of ordinary differential equations." Numerical simulations 1, Lecture, 2018.

[11] R. L. Harrison, "Introduction to monte carlo simulation.," *AIP conference proceedings*, vol. 1204, pp. 17–21, Jan 2010.

[12] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403 – 434, 1976.

[13] G. Beldman, A. G. J. Voragen, F. M. Rombouts, M. F. Searle-van Leeuwen, and W. Pilnik, "Adsorption and kinetic behavior of purified endoglucanases and exoglucanases from trichoderma viride," *Biotechnology and Bioengineering*, vol. 30, no. 2, pp. 251–257, 1987.

[14] L. J. Gibson, "The hierarchical structure and mechanics of plant materials," *Journal of The Royal Society Interface*, vol. 9, no. 76, pp. 2749–2766, 2012.

[15] C. Loix, M. Huybrechts, J. Vangronsveld, M. Gielen, E. Keunen, and A. Cuypers, "Reciprocal interactions between cadmium-induced cell wall responses and oxidative stress in plants," *Frontiers in Plant Science*, vol. 8, 10 2017.

[16] A. Ebringerová, Z. Hromádková, and T. Heinze, *Hemicellulose*, pp. 1–67. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.

[17] A. Shrotri, H. Kobayashi, and A. Fukuoka, "Chapter two - catalytic conversion of structural carbohydrates and lignin to chemicals," vol. 60 of *Advances in Catalysis*, pp. 59 – 123, Academic Press, 2017.

[18] W. Boerjan, J. Ralph, and M. Baucher, "Lignin biosynthesis," *Annual Review of Plant Biology*, vol. 54, no. 1, pp. 519–546, 2003. PMID: 14503002.

[19] S. Schoenherr, M. Ebrahimi, and P. Czermak, *Lignin Degradation Processes and the Purification of Valuable Products*. 03 2018.

[20] W. Ying, Z. Shi, H. Yang, G. Xu, Z. Zheng, and J. Yang, "Effect of alkaline lignin modification on cellulase-lignin interactions and enzymatic saccharification yield," *Biotechnology for Biofuels*, vol. 11, p. 214, Aug. 2018.

[21] H. Wang, Y. Pu, A. Ragauskas, and B. Yang, "From lignin to valuable productsstrategies, challenges, and prospects," *Bioresource Technology*, vol. 271, pp. 449 – 461, 2019.

[22] J.-C. Sigoillot, J.-G. Berrin, M. Bey, L. Lesage-Meessen, A. Levasseur, A. Lomascolo, E. Record, and E. Uzan-Boukhris, "Chapter 8 - fungal strategies for lignin degradation," in *Lignins* (L. Jouanin and C. Lapierre, eds.), vol. 61 of *Advances in Botanical Research*, pp. 263 – 308, Academic Press, 2012.

[23] K. Merklein, S. Fong, and Y. Deng, "Chapter 11 - biomass utilization," in *Biotechnology for Biofuel Production and Optimization* (C. A. Eckert and C. T. Trinh, eds.), pp. 291 – 324, Amsterdam: Elsevier, 2016.

[24] L. T. Fan, Y.-H. Lee, and D. H. Beardmore, "Mechanism of the enzymatic hydrolysis of cellulose: Effects of major structural features of cellulose on enzymatic hydrolysis," *Biotechnology and Bioengineering*, vol. 22, no. 1, pp. 177–199, 1980.