Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

Theoretical Optimization of Enzymatic Biomass Processes

dem Fachbereich Physik der Philipps-Universität Marburg vorgelegt von

Dipl.-Phys. Alexander Orlov aus Kyiv

am 20. Dezember 2018

Erstgutachter: Prof. Dr. Peter Lenz
Zweitgutachter: Prof. Dr. Martin Koch

Veröffentlicht in Marburg, 2019

Vom Fachbereich Physik der Philipps-Universität Marburg (Hochschulkennziffer 1180) als Dissertation angenommen am 2019-02-21

Erstgutachter: Prof. Dr. Peter Lenz
Zweitgutachter: Prof. Dr. Martin Koch

Tag der mündlichen Prüfung:
2019-03-01

Published in Marburg, 2019

Accepted as dissertation by the Department of Physics at the University of Marburg (University ID 1180) on
2019-02-21

Primary assessor: Prof. Dr. Peter Lenz
Secondary assessor: Prof. Dr. Martin Koch

Day of thesis defense:
2019-03-01

Philipps Universität Marburg

Dissertation zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

Theoretical Optimization of Enzymatic Biomass Processes

dem Fachbereich Physik der Philipps-Universität Marburg vorgelegt von

Dipl.-Phys. Alexander Orlov aus Kyiv

am 20. Dezember 2018

Erstgutachter: Prof. Dr. Peter Lenz
Zweitgutachter: Prof. Dr. Martin Koch

ERA IB

Bundesministerium für Bildung und Forschung

I dedicate this thesis to my parents.
For their confidence, support, encouragement and endless love.

Zusammenfassung

Diese Dissertation führt ein vollständiges, stochastisches Framework Cellulect ein, um die Hydrolyse-Vorgänge von lignozellulosehaltiger Biomasse zu untersuchen, optimieren und vorherzusagen.

Das Framework umfasst ein ausführliches geometrisches Modell für die Verfolgung von von kristallinen/amorphen Gebieten, einzelnen Monomeren, der Polymerverteilung und zugänglicher Oberfläche innerhalb eines lignozellulosehaltigen Substrates. Die Aktualisierungen geschehen reaktionsweise anhand eines seriellen effizienten Algorithmus. Dem Konzept der physikalischen Zeit wird genüge getan.

Verschiedene Arten von Enzymreaktionen (zufällige Schnitte, Schnitte an reduzierenden/nichtreduzierenden Enden und ein ggf. stationäres aktives Zentrum) sowie -struktur (Kohlenstoffbidungsmodul mit flexiblem Linker und einer katalytischer Domain) werden auf modulare Art berücksichtigt. Das Konzept des endlichen Zustandsautomaten wird verwendet, um die Enzyme zu modellieren. Dies ermöglicht eine zuverlässige, mächtige und wartbare Modellierung bereits bekannter Enzymeigenschaften, die mit zusätzlichen Eigenschaften erweitert werden kann, die in dieser Arbeit nicht berücksichtigt wurden.

Die verwendete allgemeine probabilistische Beschreibung der katalytischen Wirkungsweise umfasst weiterhin Adsorption, Desorption, kompetitive Inhibition mit gelösten Polymeren und dynamischen Bindungsbruchreaktionen, abhängig vom Zustand der Monomere und deren Polymeren innerhalb des Substrats. Alle integrierten Parameter nehmen Bezug auf systemspezifische Eigenschaften und liefern somit eine eineindeutige Beziehung zwischen den Freiheitsgraden des Systems und Eigenschaften des Modells.

Schließlich beruht der Zeitentwicklungsalgorithmus auf einem modifizierten stochastischen Algorithmus von Gillespie. Er stellt einen stochastisch exakten Zeit-Reaktions-Propagationsalgorithmus zur Verfügung, der sowohl die zufällige Eigenart der Reaktionen wie auch deren Auftreten berücksichtigt.

Das Framework ist bereitgestellt für eine Optimierung mit Nebenbedingungen mit empirischen Datensätzen von Produktkonzentrationsprofilen unter Verwendung der üblichen Optimierungsroutinen. Der Nachweis für verfügbare Daten für die häufigsten Enzymarten (EG, β-G, CBH) in der Literatur wurde erbracht.

Sensitivitätsanalyse der geschätzten Modellparameter wurde betrieben. Abhängigkeiten von verschiedenen experimentellen Eingaben wurde gezeigt. Optimierungsverhalten unter unterbestimmten Bedingungen wurde untersucht und dargestellt.

Ergebnisse und Vorhersagen von optimierten Enzymmischungen sowie ein praktischer Weg das Cellulect Framework zu implementieren und zu verwenden wurde ebenfalls zur Verfügung gestellt. Die erhaltenen Ergebnisse wurden mit experimentellen Daten aus der Literatur verglichen, um die Anpassungsfähigkeit, Effizienz und Genauigkeit des vorgelegten Frameworks für die Vorhersage der Biomassehydrolyse zu zeigen.

Abstract

This dissertation introduces a complete, stochastically-based algorithmic framework Cellulect to study, optimize and predict hydrolysis processes of the structured biomass cellulose.

The framework combines a comprehensive geometric model for the cellulosic substrate with micro-structured crystalline/amorphous regions distribution, distinctive monomers, polymer chain lengths distribution and free surface area tracking. An efficient tracking algorithm, formulated in a serial fashion, performs the updates of the system. The updates take place reaction-wise. The notion of real time is preserved.

Advanced types of enzyme actions (random cuts, reduced/non-reduced end cuts, orientation, and the possibility of a fixed position of active centers) and their modular structure (carbohydrate-binding module with a flexible linker and a catalytic domain) are taken into account within the framework. The concept of state machines is adopted to model enzyme entities. This provides a reliable, powerful and maintainable approach for modelling already known enzyme features and can be extended with additional features not taken into account in the present work.

The provided extensive probabilistic catalytic mechanism description further includes adsorption, desorption, competitive inhibition by soluble product polymers, and dynamical bond-breaking reactions with inclusive dependence on monomers and their polymers states within the substrate. All incorporated parameters refer to specific system properties, providing a one to one relationship between degrees of freedom and available features of the model.

Finally, time propagation of the system is based on the modified stochastic Gillespie algorithm. It provides an exact stochastic time-reaction propagation algorithm, taking into account the random nature of reaction events as well as its random occurrences.

The framework is ready for constrained input parameter estimation with empirical data sets of product concentration profiles by utilizing common optimization routines. Verification of the available data for the most common enzyme kinds (EG, β-G, CBH) in the literature has been accomplished.

Sensitivity analysis of estimated model parameters were carried out. Dependency of various experimental input is shown. Optimization behavior in underdetermined conditions is inspected and visualized.

Results and predictions for mixtures of optimized enzymes, as well as a practical way to implement and utilize the Cellulect framework are also provided. The obtained results were compared to experimental literature data demonstrate the high flexibility, efficiency and accuracy of the presented framework for the prediction of the cellulose hydrolysis process.

Contents

Symbols

k_i	rate constants of enzymes. For each distinguishable reaction a rate is defined. Depending on the reaction order n, the units are $$[k_i] = \frac{\text{mol}}{\text{t} \cdot \text{mol}^n} \quad (0.1)$$ There are five reaction rate constants derived from the chemical reactions 2.11 and 2.12: k_1: governing adsorption to the substrate k_2: governing desorption from the substrate k_3: governing hydrolization process k_4: governing desorption from the product k_5: governing adsorption to the product	13, 32, 33, 34, 60, 61, 66, 70, 73, 83
c_i	reaction propensities. Conditional probabilities of according reaction events in a system, given a system state. I.e., $$P_B(A) = \frac{P(A \cap B)}{P(B)} \quad (0.2)$$ where $P(A\|B)$ is the probability of a specific reaction event A and $P(B)$ is the probability, that an arbitrary reaction event occurs.	32, 33, 34
k_{cat}	Hydrolization rate constant. Identified from the reaction form to be k_3, the hydrolization rate plays a remarkable role amongst other reaction rates.	11, 12, 13, 44, 62, 63, 64, 83, 84, 102

v_{max} — Maximum hydrolyzation velocity for an enzyme mixture with concentration $[E]_{tot}$. If the hydrolyzation rate k_3 is $k_3 = k_{cat}$, then (44, 62, 70)

$$v_{max} = k_{cat} \cdot E_T \tag{0.3}$$

K_M — Michaelis constant. A measure of the substrate's affinity for the enzyme and corresponds at which substrate concentration the reaction rate is at half-maximum (12, 57, 62, 63, 64, 65, 67, 70, 72, 73, 74, 75, 76, 83, 84, 93, 94)

$$K_M = \frac{k_2 + k_3}{k_1} \tag{0.4}$$

K_E — Enzyme kinetic efficiency. A measure of how efficiently an enzyme converts a given substrate. (57, 72, 75, 76, 94)

$$K_E = \frac{k_{cat}}{K_M} \tag{0.5}$$

K_I — Inhibition constant. Similar to the Michaelis constant an inhibition constant is introduced, which shows the affinity of the enzyme to bind in a non-productive way to substrate or product. (57, 62, 64, 65, 73, 74, 75, 93)

$$K_I = \frac{k_4}{k_5} \tag{0.6}$$

K_C — Van Slyke-Cullen constant. Shows the propensity of hydrolization relative to adsorption. (57, 70, 72, 94)

$$K_C = \frac{k_3}{k_1} \tag{0.7}$$

$\widetilde{K}_C$ — Constant is defined in an analogous manner as the van Slyke-Cullen constant. Relates the propensity of hydrolization to the propensity of inhibition (73, 74, 94)

$$\widetilde{K}_C = \frac{k_3}{k_5} \tag{0.8}$$

K_D Dissociation equilibrium constant. Shows the propensity of desorption relative to adsorption. 57, 94

$$K_D = \frac{k_2}{k_1} \tag{0.9}$$

K_S Sampling parameter used for proper model parameter samplings in optimization routines. 60, 61, 66, 70, 78, 79

$$K_S = \sum_i k_i \tag{0.10}$$

N_A Avogadro constant, counting particles in a mole of substance. It is set to 34, 39, 41, 50

$$N_A = 84446888^3 \simeq 6.022 \cdot 10^{23}\text{mol}^{-1}, \tag{0.11}$$

as proposed in ref. [1]

S_T Total substrate concentration in the considered solution. Used as a shortcut for $[S]_{tot}$. As the amount of substrate is assumed to be in excess, compared to the enzyme concentration, this is used as the normalization value, if not otherwise stated. 13, 41, 62, 70, 72, 73, 74, 92, 93

E_T Total enzyme concentration in the considered solution. Used as a shortcut for $[E]_{tot}$. It is assumed, that the enzyme concentration is low compared to the concentration of substrate. Enzyme concentrations have to be defined and calculated on simulation input as they are measurable control parameters of the experiment, contributing to the ratio E_T/S_T. 11, 13, 39, 41, 44, 69, 70, 72, 73, 74, 92, 93

$\#_E$ number of enzyme entities, after discretization of the system 39, 41

$\#_S$ number of substrate particles, i.e., glucose units, after discretization of the system 41

$\#_T$ turn-over number. I.e., the number of conversions per time unit for an active (catalytic) site, equal to k_{cat} 44

CrI Crystallinity index. Statistical quantity of current crystallinity degree of the substrate. 26, 89

DP	degree of polymerization. I.e., the relation of the first to the zeroth moments of polymer length distribution $$DP = x_N = \frac{p^{(1)}}{p^{(0)}} = \mathbf{E}(x) \tag{0.12}$$ where the moments are defined by $$p^{(n)} = \int_0^\infty x^n p(x,t) dx \tag{0.13}$$	55
L_0	"Solubility barrier". Polymers with length smaller than this value are considered soluble and dissolve, if they are on accessible surface of the substrate	49, 55
T	Temperature within the reactor and the solution.	14, 88, 89
pH	Measure for acidity of an (aqueous) solution	88
M_E	molar mass of an enzyme	44
M_S	molar mass of substrate ($= 162 \frac{\text{g}}{\text{mol}}$), see ref. [2]	41
f/f_{min}	friction factor, defining the correction form factor for enzymes, used in diffusion factor calculation. Equal to sedimentation factor	40
D	Diffusion coefficient. In liquids defined as $$D = \frac{k_B T}{6\pi\eta R_0} \tag{0.14}$$ with k_B as the Boltzmann constant, T temperature, η dynamic viscosity of the solvent and R_0 the hydrodynamic radius of the diffusing particle. Another common definition considers the diffusion equation $$\partial_t f(\mathbf{x}, t) = D\Delta f(\mathbf{x}, t) \tag{0.15}$$ Assuming, the diffusion coefficient is constant.	40
l_D	characteristic diffusion length, defined by (A.6).	40
A	specific enzyme activity, given in $$[A] = \frac{IU}{\text{mg}} \tag{0.16}$$	44

1 Introduction

1.1 Enzymatic hydrolyzation of biomass

Biomass is the oldest energy source of humankind. Since the industrial revolution, combustion of fossile fuels was the most prominent way to gain energy. In the last century, however, costs for these energy sources grew considerably, see figure 1.1.

In times of increasing scarcity of resources other ways of biomass processing become more and more important. For example, biomass is not only the source of a direct energy production but also an important supplier of chemical products, such as glucose and other sugar species. Sugars for their part are a crucial precursor for bioethanol production.

There are two widespread feedstocks of biomass which can be converted to technical useful sugars [4, 5]. The more established one comes from sugar canes and starch crops. Here, the fermentable sugars are extracted by grinding or crushing followed by fermentation to ethanol. Production of ethanol from starch is performed by either dry grind or wet milling process [6]. However, the production of biofuels from food feedstocks has negative impacts on land use and food prices [7, 8]. This has led to an increasing interest in biofuel production from non-food biomass feedstocks, see figure 1.2 and ref. [9].

In the latter approach, glucose is separated from the biomass by hydrolytic enzymes. Such enzymes appear in different microorganisms as yeast (*Saccharomyces cerevisiae*), but also in bacteria like *Escherichia coli* or *Cellulomonas fimi* [10]. The cellulosic approach is by far less harmful to the environment and provides a possibility to further expand renewable energy production [11, 12]. Meanwhile, the sugar production from the biomass is difficult in terms of technological efficiency [13]. The production involves pretreatment steps of biomass to improve its accessibility to enzymes, see refs [14, 15]. The conversion reactions require highly specialized enzymes and the conversion reactions still remain slow.

A collaboration between experimental group at Edinburgh and Paris, industry partners and the theoretical group of Peter Lenz at Philipps University of Marburg was funded by BMBF for the Cellulect project. Within this project, the goal was to enhance cellulosic sugar extraction and to create a genetic platform technology by means of which mixtures of hydrolytic active enzymes can be optimized with regard to biomass conversion. To achieve this goal a detailed theoretical model of enzymes was developed, which can be analyzed through computer simulations.

The main challenge was to formulate the model using as little dynamic parameters as possible, but still to keep the model flexible and expandable enough to be able to incorporate features which are not considered in advance. Furthermore, to model the system is complex: There are many bacteria producing different enzyme types. The diversity of

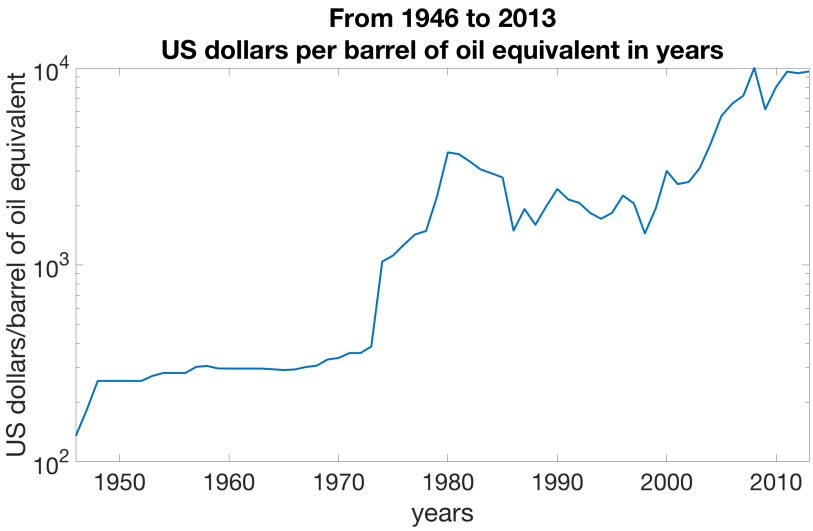


Figure 1.1: Oil prices from 1946 to 2013. x-axis in years, y-axis in US dollars per barrel of oil equivalent. The figure is the result of query "oil price 100 years" on the Wolfram|Alpha Knowledge base. Cf. [3]

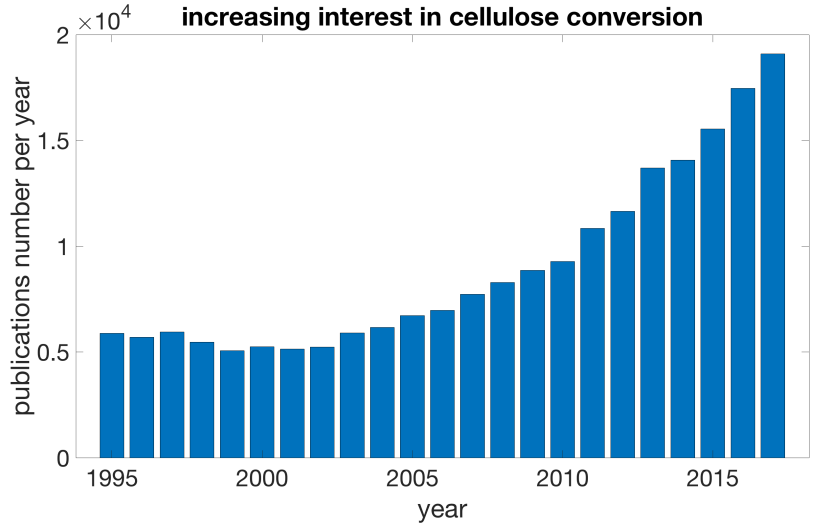


Figure 1.2: Amount of articles about enzymatic cellulose degradation for biofuel and bioethanol production. The figure is the result of the query "cellulosic production OR conversion" on the ScienceDirect literature database. The result was sorted and filtered by date to get new literature per year published.

enzymes reaches about 80 types, see ref. [16], and includes different types of Cellulases, classified by their typical reactions like *Endocellulases*, EC 3.2.1.4, *Endocellulases*, EC 3.2.1.91, *Cellobiases* EC 3.2.1.21. Also, extra cellular enzyme complexes *Cellusome* are known, see ref. [17].

Each enzyme has many biophysical parameters, of which values are partly not known in advance. Likewise the substrate has to be modeled as a geometrical object with a complex structure and should enclose multiple physical properties, like crystallinity, degree of polymerization and active surface. The initialization has to be performed with data provided by experimentalists. Updates during ongoing simulation have to be guaranteed to be at least feasible from physical, biological and chemical points of view.

The simulated system is coupled via the substrate in a nonlinear manner. Additionally, the interaction between enzymes should not be excluded a priori, which adds another nonlinear coupling. Due to these circumstances, the possible phase space has a high dimension of up to 11 continuous variables combined with up to 8 discrete and categorical degrees of freedom, making the system an interesting research object.

1.2 Recent trends

Over the last half of this century some models evolved, trying to predict the cellulose degradation in a qualitative and quantitative manner. Starting with Michaelis Menten theory, defined in ref. [18], the main approach to enzyme kinetics takes place via hidden states of an enzyme. Therein, one or more intermediate states of an enzyme are assumed. Then for each state of enzymes and substrate an appropriate master equation can be formulated, cf. refs. [19, 20]. See also section 2.2.

With the first approach of Michaelis-Menten kinetics, it is only possible to describe general chemical reactions, balances and equilibria. As soon as one is interested in a more detailed description of reactions going on and in examining specific features of the reactants, one has to extend descriptions of reactions under inspection.

Different approaches exist to capture extended dynamics of a biomass conversion system, some of them are mentioned in ref. [21]. The most common of them are

- A ”data driven” approach like in ref. [22] is an abstract way to model a complex system by a neural network. Therein no physical characteristics of the system are captured. An artificial regression model is trained by means of (experimental) data, providing abilities to make predictions about parameter regions, which are not known so far. However, this approach cannot model a specific behaviour change since there is no possibility to add features. Another drawback is that a lot of input data is needed to achieve accurate predictions. In general such data is not present.

- A ”mechanistic” approach, like in refs. [23, 24, 25] incorporates some properties of the substrate as well as enzymes, reformulating the problem in rate equations. This approach is capable to capture most individual properties like accessible surface of the substrate, as well as the behavior of different enzyme types. However, every feature requires an additional set of rate equations, each of those is not obvious to

formulate. As the overall set of rate equations grows fast with new features, the computational abilities of such approaches are limited. Also, there are problems to formulate cooperative action of enzymes by population-balance equations in use.

- Another promising approach is a "stochastic molecular" one, seen e.g. in ref. [26]. Therein features of enzymes and substrate are formulated. The algorithm itself works in an approximative manner. Updates to the environment due to enzyme activities propagate with a delay in comparison with the ongoing reactions. The assumption, that the environment does not suffer an appreciable change in its state on every reaction is not proven. Furthermore, some dynamic parameters are introduced, which cannot be interpreted as properties of reactants, but only as optimization parameters to suffice experimental results.

- A radically different approach is chosen by Kansou et al., in ref. [27]. "Qualitative reasoning" introduced in [28, 29, 30, 31] can be used, to model arbitrary precise causality effects, but without precise quantification. Therein, principles are motivated by human cognition, and only known qualitative values are included into the model. The model is also meant to provide ambiguous results instead of a unique one, which is natural for analysis of complex systems. While being able to compare various entities, like enzyme types or enzyme mixtures, this model fails to make exact numerical predictions by design. Because the goal of this approach is to manage incomplete knowledge, only a comparison is possible. On the contrary, a quantitative prediction is unfeasible as some dynamic parameters remain unknown.

From now on, the focus will be on quantitative models to be able to make predictions, rather that only comparisons. These models, including mechanistic and stochastic ones make heavy use of numerics. From the numerical methodology point of view also different approaches can be chosen, see ref. [32, 33]. The differences arise from to interpretation of chemical and kinetic equations of the ongoing reactions. The common interpretations include

- solving the governing master equation. Therein, a set of ordinary differential equations for each possible state of the system is defined. At time t the kth equation gives the probability of the system being in the kth state, see ref. [34]. As the overall system state is modeled, the number of states depends upon the total number of molecules. The resulting system of ordinary differential equations (ODEs) is therefore very large.

- Another approach comes with the chemical Langevin equation, see ref. [35]. Here, instead of solving the full set of ODEs to get a probability distribution over all possible states for each time t, a sample from these distribution is generated. This approach was derived by Gillespie, see ref. [36].

- A further approach is to neglect the stochastic nature of the ongoing process. This assumption is known as the thermodynamic limit, see ref. [37]. Then, the governing

equations becomes ordinary, nonlinear differential equations with reaction rates as constants. This approach is used in many simulation software packages, e.g., in the SimBiology toolbox in MATLAB [38].

The latter approach is not able to model single entities of interest in the system, but only the general behavior of chemical reactions in time. Therefore it is not possible to model specific properties of enzymes or properties of substrates which are known to be heterogeneous.

The first approach would enable these features. However, solving an ODE for every property of each entity of the system is only possible by adding more equations into the system. This is hardly affordable, if once interested in optimization, where each solution requires many thousands of system propagation steps with different input parameters.

Therefore the choice was made in favor of stochastically solving the chemical Langevin equation. The chemical Langevin and the master equations are related. The solution of the master equation yields all possible (stochastic) paths, which a system can undergo during evolution. A solution of the Langevin equation yields a single sample of this set.

This circumstance can be handled by enlarging the system size and the number of samples simulated. By enlarging the system in the model, the probability of a statistical outlier solution is lowered. By generating more samples of solution paths standard statistical analysis methods can be used to describe the quality of the model. Their mean serves as the reference solution, along with a deviation typical for a stochastic model.

Generally speaking, a lot of models of enzyme kinetics exist and their number is increasing rapidly, see ref. [9]. Bansal et al. in [33] mentions 73 kinetic models published from 1975 to 2009. This is an attempt to implement an extendable model, taking into account known features of substrate and enzymes, keeping the amount of unknown parameters for the simulation as small as possible. This results into a maintainable and extendible model so that introducing new features would not require to alter the behavior of existing ones.

The applications of such models can also be different. Qualitative models are subjected to show general behavior of its compartments and the influence of new features on their dynamic behavior. With quantitative models it is also possible to incorporate unknown dynamic variables, which then are subjects to an optimization process.

During the optimization process the model is ”trained” with some experimental results. Within this procedure the unknown parameters can be identified and saved. After this, the model is able to work and make predictions in parameter regions where no experimental results exist yet.

This approach is useful because it can be applied to various existent data. However, the quality of found parameters has to be verified. Since this optimization procedure in general has a heuristic nature, many approaches have to be tried out in order to achieve a significant statement.

However, the utilization of such models also has advantages in case all parameters are already known. Once the model is verified, it can be used to predict results in regions, where experiments are costly and take a long time to carry out.

After a short summary of symbols this review of recent, historical, environmental, technical trends and state of the art followed. In the next chapters the work will pursue the following structure:

Chapter 2 introduces the model, which was composed in the present work. The main components are the substrate and enzymes taking part in the hydrolization process. Special attention was paid to the time propagation algorithm. The algorithm and the discretized approach to the model components establish the stochastic nature of the model. Boundaries of validity are also discussed in this chapter.

In chapter 3 the described model is subjected to validation. This is done by means of experimental data and search of corresponding unknown model parameters. In that way the model can reproduce the given experimental data. An according optimization problem will be formulated and optimization results will follow. It will be shown, how difficulties arisen during this phase of work were tackled. The validity of found results will be discussed as well.

In chapter 4 some applications of the model will be demonstrated. Most of them are shown in context of experimental data. However, such data could not be found for all possible applications. Therefore, some of the results are of theoretical nature.

The last chapter concludes the thesis. It provides a short summary of found results, model features implemented throughout the work and encountered problems. An outlook and possible extension concludes this work.

2 Cellulect Model

2.1 Model overview

The developed model is based upon the concept of ”finite state machines” (FSM). FSM were first introduced in the context of theoretical informatics and comes from the theory of models of computation, see ref [39]. In this work FSMs are applied for the first time in an innovative manner for description of cellulosic hydrolysis.

The starting point of this work is the mechanistic model for enzymatic cellulose hydrolysis, derived by Levine et al., in ref. [23]. This model provides a mesoscopic description of the relevant processes that are characterized by many biochemical parameters (such as kinetic parameters for enzyme activities, Michaelis-Menten constants, inhibition constants and physical parameters characterizing the cellulose substrate). Additionally, the model separates the enzyme-substrate interaction into binding and complex-formation steps, and distinguishes between scission at the interior and at the end of the cellulose chain. The resulting rate equations are solved numerically. The drawback of this approach is the dependence on many biochemical parameters, most of which are unknown. Furthermore, this approach does not take into account any details of the distribution of substrate.

In refs. [24, 25] Griggs et al., introduced another mechanistic model. Features of the substrate, like separate crystalline and amorphous phases were not incorporated. Modifications for new enzyme types are difficult to implement and involve significant changes to the complete model.

The stochastic molecular formulation introduced in ref. [26] by Kumar et al., is able to simulate the synergistic action of multiple enzymes during the hydrolysis process and to capture quite a few important experimental observations (structural properties, inhibition, etc.). However, the reasoning for time propagating algorithm remains unclear, lacks flexibility and depends on properties of simulated enzyme kinds.

Based on the models and ideas of Griggs, Levine and Kumar a hybrid mechanistic stochastic model of the enzymatic hydrolysis process using an event-driven method of the enzyme-substrate interaction has been developed in this work. This is based on an explicit set of equations of chemical reactions between system components but does not require them. Instead rules for enzymes are defined, stating the mode of action of each enzyme type. The main advantage of this approach is the ability to incorporate substrate and enzyme properties into the model separately.

The model implements discrete-event enzyme-substrate interactions and includes initial environment properties, like temperature and pH, initial and dynamic substrate properties, pretreatment conditions, cellulose surface accessibility, crystallinity of cellulose fibrils, degree of polymerization. Currently, it can be used with at least three

different types of enzymes that hydrolyze the β-1,4-glycosidic bonds: cellobiohydrolases (CBH), endo-β-1,4-glucanases (EG) and β-glucosidase (β-G). More enzymes can be added using already prefabricated and prepared model training and verification with separate sets of experimental data for each enzyme.

The model includes the following distinct ingredients:

- The solid substrate which is represented as cellulose fibrils in the form of bundles of cellulose chains with predefined length distribution corresponding to initial degree of polymerization of the substrate. The substrate description also includes the initially active and accessible surface (only the chains at the surface are accessible to the enzymes), and a predefined crystallinity with a crystallinity distribution in the cellulose fibrils

- Multiple sets of enzymes of different types as finite-state machines

- An abstract reactor object for managing the back coupling of enzymes to the environment and dynamic tracking of the whole set of important system parameters during hydrolysis for evaluation, comparison and matching with experimental data.

- An independent time and reaction propagation algorithm based on the one defined by Gillespie in [36].

Cellulose in plants appears mostly in connection with hemicellulose and lignin, see ref. [40]. It has a complex, hierarchical structure, partially ordered in a distinguished direction into polymers, see figure 2.1.

Modelling the substrate is unavoidable, to achieve quantitative results. Without an explicit model, either an assumption of substrate in excess has to be made, or, the substrate model won't be able to store any dynamic properties, which are subject to change during the hydrolization process.

Some simplifying assumptions for modelling can be met without altering the main features of cellulose, most of them concern its geometrical structure. In figures 2.2 - 2.7 some common models are visualized, which are used for representing a substrate.

The solid cellulose substrate in the form of a 3D matrix contains the oriented fibril of a cellulose chain bundle with predefined location of the reducing/non-reducing ends. Each cellulose polymer chain consists of monomers with one glucan unit length. Only polymers at the surface of the matrix are accessible to the enzymes.

In order to mimic the complicated surface structure of the cellulose fibril and its enhanced surface accessibility for enzymes, the initial active surface of the cellulose fibril can be modified and made similar to patterns taken from microscopy images of the experimental cellulose fibers. This includes the substrate structure on the surface as well as on inner layers.

The length distribution of cellulose chains with narrow Gaussian-like shape corresponds to the initial experimental degree of polymerization of the substrate. Different efficiency of enzyme action in the crystalline and amorphous regions of the cellulose fibril is also included in the model. Such regions are automatically distributed along

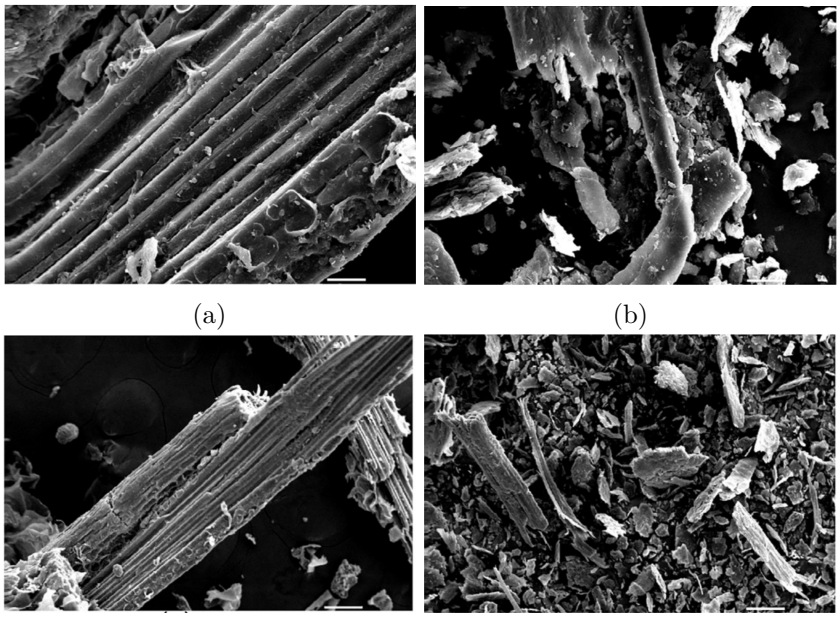


Figure 2.1: Scanning electron microscope images of unpretreated ((a), (c)) and pretreated ((b), (d)) corn stovers, as shown in ref. [41]. The pretreatment conditions were 0% (w/w) water input, 30 min of milling at 80°C. The bars indicate 20μm((a), (b)) and 100μm ((c), (d)).

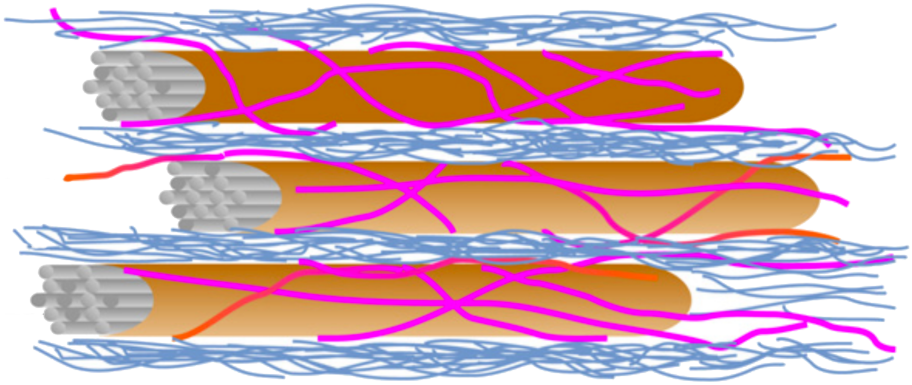


Figure 2.2: Microstructure model, as shown by Wei et al., in ref. [42]. The model shows cellulose polymer chains (grey), covered by hemicellulose (pink) and lignin (blue).

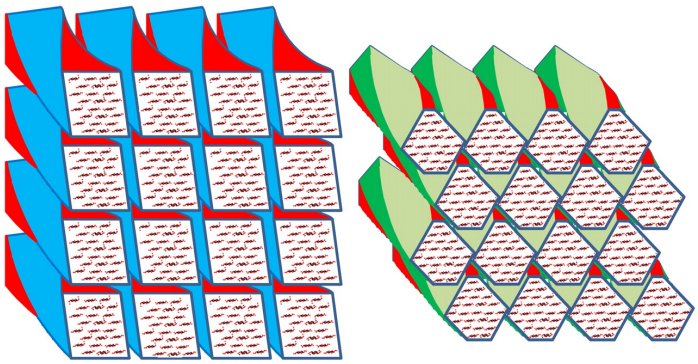


Figure 2.3: Schematic representation of chain and microfibril packing arrangements for a 4 × 4 aggregate of 24-chain spruce cellulose microfibrils with alternative shapes, as shown by Fernandes et al., in ref. [43]. Left: rectangular shape. Right: diamond shape. The rectangular shape is in closer accordance with the dimensions observed by X-ray scattering using the Scherrer equation corrected for disorder. Twisting of the microfibrils prevents their coalescence by ensuring that, even if, hypothetically, their crystal planes are aligned as in the transverse section shown, this alignment is quickly lost further along the microfibril aggregate.

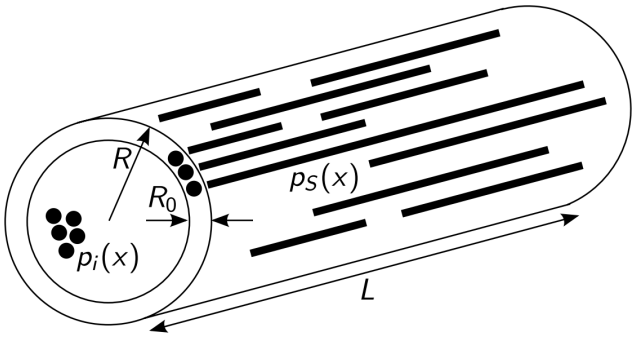


Figure 2.4: Schematic of a cylinder comprised of cellulose chains, as shown by Griggs et al., in ref. [44]. R is the microfibril radius, R_0 is the current radius of accessible substrate, p_i is the ratio of hidden substrate and p_s is the ratio of accessible substrate.

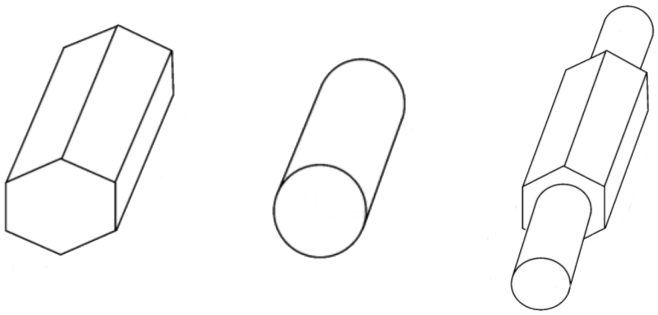


Figure 2.5: Different shapes and combinations, as shown by Levine in ref. [23] to represent polydisperse distribution of spheres of the substrate. The emphasis lies on the modelling of various surface to volume ratios and to maintain same ratios with different geometries.

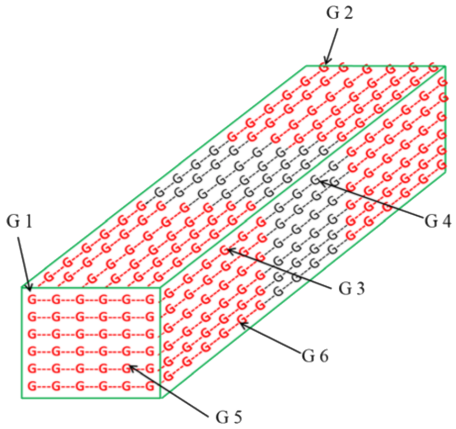


Figure 2.6: Structure of elementary fibril simulated in model of Kumar, see ref. [26]. The figure illustrates the arrangement of glucose molecules in an elementary fibril. Glucose molecules in red represents crystalline region and glucose molecules in black are in amorphous region.
$G1$: on surface; crystalline; reducing end;
$G2$: on surface; crystalline; non-reducing end;
$G3$: on surface; crystalline;
$G4$: on surface; amorph;
$G5$: on inside; crystalline; reducing end;
$G6$: on surface; crystalline

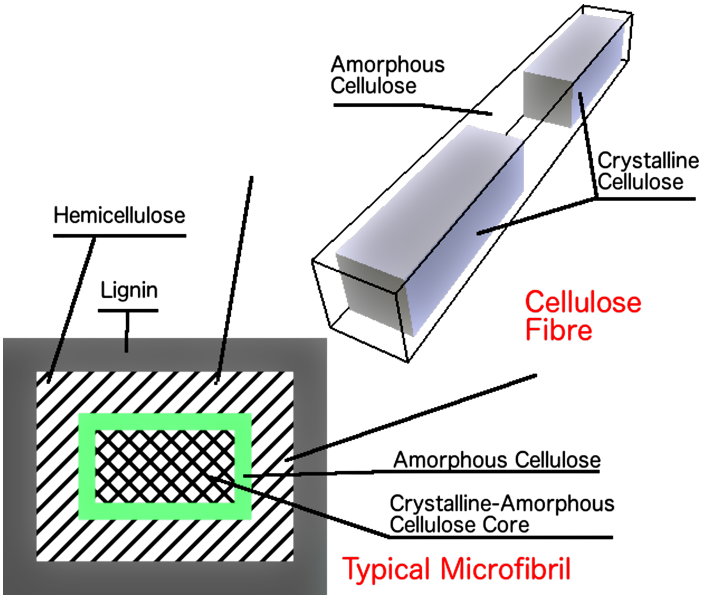


Figure 2.7: Representation of the cross-section of a micro fibril and cellulose as shown by Flores et. al. in [45].

the substrate matrix resulting in experimental crystallinity index CrI of the cellulose substrate.

The substrate is simplified in comparison with the model of Kumar, cf. [26], as that surfaces on elementary fibrils do no differ from surfaces on macrofibrils. However, this feature can be added easily. For a detailed description, see section 2.5.

The distinctive feature of the implemented model is the description of enzymes as finite-state machines (FSM), independent from substrate. I.e., in contrast to ref. [46], any assumptions on amount, concentration, or physical properties of the substrate were done, while formulating behavior of enzymes. The present model in particular does not assume substrate saturation to formulate enzyme actions.

Given a reaction graph, see e.g. [47], the states of the FSM for each enzyme type coincide with the nodes of this graph. Each reaction corresponds to a state transition, see section 2.2. This leads to minimization of the amount of parameters needed for a quantitative model, see ref. [48].

Enzymes detect their external environment, react in a stochastically (with predefined probabilities) or deterministically manner, and couple back to the environment (substrate and solution). The model is able to track dynamic changes of the substrate structural properties (average degree of polymerization, crystallinity, surface accessibility) and to provide time-dependent concentrations of soluble oligomers and cellulose chain length distributions.

In order to overcome the limitations of many previous studies the enzyme functionality was implemented as a FSM with finite and predefined states. In this description, enzymes

Figure 2.8: Quaternary structure of Endo-β-1,4-Glucanase. *Trichoderma reesei* cel12a P201C mutant. Image generated by Pymol, see ref. [49]. PDB image: 1OLQ.

undergo state changes during the hydrolysis of cellulose. Three states were identified and modeled:

- not bound to the substrate, i.e. "free" state,
- bound to a chain that currently cannot be hydrolyzed by the enzyme, i.e. "inhibited" (by product/intermediate) state and
- bound to a chain that can be hydrolyzed, i.e. "bound" state.

In general, an enzyme is a complex, bioactive macroprotein. As accelerators for chemical reactions, cells need enzymes to enable and to accelerate metabolic processes.

There are known many different enzyme types, from various bacteria. Enzymes differ in form, size, action mode, reaction rates, orientation, catalytic domain, carbohydrate binding module, processivity, dependency on substrate and product properties. Trying to model every single property of an enzyme can therefore be very difficult and does not promise to yield more significant results than with more simple assumptions.

In figures 2.8 - 2.10 there are parts of enzymes depicted as crystal structures. Even small parts demonstrate the complexity of these proteins.

Some simplifications are made in the current work to model enzymes properly. For example, an enzyme can bind to a single polymer chain only and cannot cover multiple polymers simultaneously in the transversal direction. Another simplification is induced by the difference between the size of an active center of an enzyme and the effective size, which an enzyme blocks on the substrate surface. The former defines the characteristic products of an enzyme, while the latter determines the maximal binding area of an enzyme. This discrepancy was solved by introducing a specific binding area (SBA) for

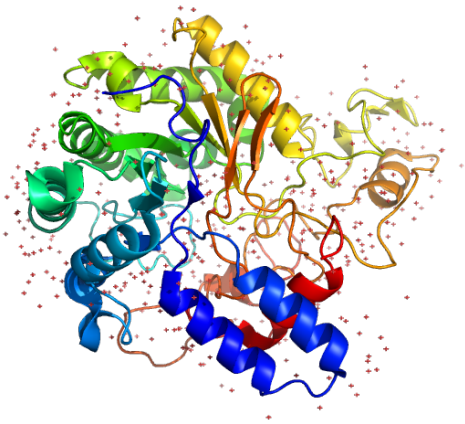


Figure 2.9: Crystal structure of catalytic domain of 1,4-beta-Cellobiosidase (CbsA) from *Xanthomonas oryzae pv. oryzae*. Grouped by color, based on identified atomic units. Image generated by Pymol, see ref. [49]. PDB image: 5XYH.

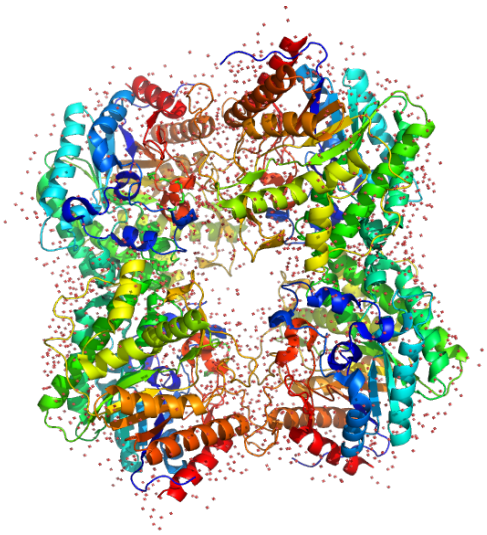


Figure 2.10: Crystal structure of beta-glucosidase A from bacterium *Clostridium cellulovorans*. Image generated by Pymol, see ref. [49]. PDB image: 3ahx.

an enzyme, keeping the characteristic product size as a distinct parameter. See section 2.4 for a detailed enzyme description.

Chemical reactions are modelled as a Poisson process. Therefore, the time between reaction events is also a random variable. In this context, the physical time is defined in terms of the so called stopping times. These are known to be exponentially distributed, see ref. [50] and sections 2.2 and A.3. Naturally, the time, when a reaction event occurs is not predictable by inspecting the former reaction event timings, see ref. [51]. It follows, that it is not possible to exactly predict, at which time a specific system state will arise.

The fact of randomized time has to be taken into account as a side effect of embedding a discrete time Markov chain (DTMC) into a continuous time Markov chain (CTMC). While technically being equivalent, the DTMC provides the same behavior as the CTMC. Though the DTMC does not contain any notion of time but an artificial one, called Markov moments. The CTMC provides the physical notion of time by defining transition rates. As the occuring events in the system are Poisson distributed, the interarrival times have an exponential distribution and in particular are not predictable. Section 2.3 describes, how this influences the precision of the model and its predictive power.

The algorithm is stochastically exact in the sense that it considers the random nature of reaction events sequence, as well as the random nature of their occurrences over time [52]. Further, it models Monte-Carlo simulations in a rejection-free manner, as a system change with a defined rate has to emerge. As such it provides the connection between artificial discrete time steps and the natural continuous time scale, without making any assumptions about possible reactions during the execution. The original algorithm of Gillespie was enhanced for performance, see section 2.2.

Sampling reactions and reaction timings simulateneously, the time algorithm is consistent, self contained and separates concerns in a clear way. This provides nice properties for reasoning and maintainability of the model in general and for the algorithm in particular.

Although the model requires optimized parameters from validated experimental data, the number of required parameters is smaller than in other mechanistic or data driven models. Additionally, the approach makes it possible to partially deduce the correspondence between the internal parameters of the FSM used as enzyme model and its kinetic parameters (e.g. activity, stability etc.). Provided input parameters intuitively trigger the features contained in the model. For example, providing an initial crystallinity index, the substrate will contain the notion of crystallinity. Providing inhibition rate constant for an enzyme, the enzyme will undergo inhibition reactions, like in figure 2.12, whereas omitting them leads to simplified dynamics, described by figure 2.11.

By means of this modelling approach it is possible to describe the relevant enzymes and the substrate in such detail, that the hydrolytic cellulose conversion can be examined under realistic conditions. As a first application it was used to find optimal enzyme mixtures concerning cellulosic degradation. The developed model is not limited to the description of this system. It is applicable to different enzyme driven reactions. As such it could provide a new theoretical approach to systems, where molecular details

$$\mathrm{E} + \mathrm{S} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C_s \xrightarrow{k_3} \mathrm{E} + \mathrm{P}$$

Figure 2.11: Basic Michaelis-Menten kinetics. k_1 is the adsorption rate, which forms the enzyme-substrate complex, $[k_1] = (\mathrm{mol} \cdot [t])^{-1}$. k_2 is the desorption rate and k_3 is the catalytic rate $[k_2] = [k_3] = [t]^{-1}$

$$\mathrm{E} + \mathrm{S} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} C_s \xrightarrow{k_3} C_p \underset{k_5}{\overset{k_4}{\rightleftharpoons}} \mathrm{E} + \mathrm{P}$$

Figure 2.12: Michaelis-Menten kinetics extended with (competitive) product inhibition. k_1, k_2, k_4 and k_5 are the adsorption and desorption rates, which form the enzyme-substrate and enzyme-product complexes, respectively. k_3 is the catalytic rate. $[k_1] = [k_5] = (\mathrm{mol} \cdot [t])^{-1}$. $[k_2] = [k_3] = [k_4] = [t]^{-1}$

of enzymes play a central role, different enzymes act together or where substrates have complex geometries or the above combined.

2.2 Time propagation algorithm

Starting from the chemical equations shown in figures 2.11 and 2.12 the underlying process is a process with constant rates and therefore a Poisson process, see ref. [53].

Knowing the governing property of the dynamics, the remaining question is how to convert the information contained in a reaction scheme into a usable and extendable simulation formulation.

Let's inspect the notion of time in a Poisson process. The process is interpreted as a counting process $\{N_t, t \geq 0\}$, with the properties

$$\begin{aligned} &N(t) \geq 0 \\ &N(t) \in \mathbb{N} \\ &s \leq t \Rightarrow N(s) \leq N(t) \\ &s \leq t \Rightarrow N(t) - N(s) = N_{s+t} - N_s, \end{aligned} \tag{2.1}$$

see ref. [54]. This definition is valid for observed events in the simulated system, which are identified as reactions in this work.

If $\{N_t, t \geq 0\}$ is a Poisson process with the intensity λ, see e.g., ref. [55], defined by

$$\mathbb{P}(N_t - N_s = k) = \frac{\lambda^k (t-s)^k}{k!} e^{-\lambda(t-s)}, \tag{2.2}$$

with $k \in \mathbb{N}$. Then, let $N_t^s = N_{s+t} - N_s$, $\forall s, t > 0$. As the increments are mutually independent and stationary, i.e,

- for all $n \geq 2, 0 \leq t_0 < t_1 < \cdots t_n$, the increments $\{N_{t_j} - N_{t_{j-1}}; 1 \leq j \leq n\}$ are mutually independent

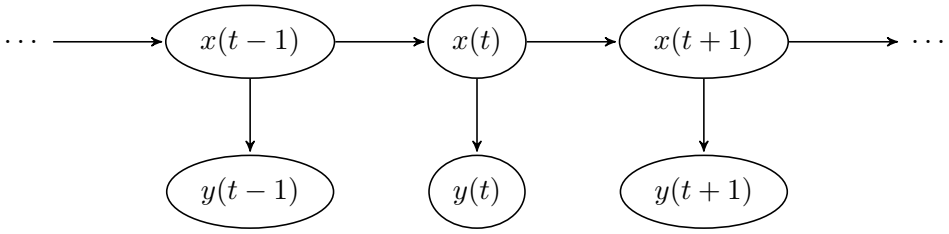


Figure 2.13: A general hidden Markov Model. While the states transits from one state x to the next in a discrete manner, the observable variable is y. The time is modelled either continuous: $t \in [0, \infty)$ or discrete: $t \in \mathbb{N}_0$. In either case, it is discretized and propagates *in concert* with the model state.

- for all $0 \leq s < t$, the law of $N_t - N_s$ depends upon the pair (s, t) only through the difference $t-s$, then $\{N_t^s, t \geq 0\}$ is a Poisson process with intensity λ, independent of $\{N_r, 0 \leq r \leq s\}$.

Since the increments $\{N_{s+t} - N_s, t \geq 0\}$ after s steps are independent of the past $\{N_t, 0 \leq t \leq s\}$, the future $\{N_{s+t}, t \geq 0\}$ after s steps depends upon the past $\{N_t, 0 \leq t \leq s\}$ only through the present value N_s. In other words, the past and the future are conditionally independent, given the present. This is the Markov property. It is also known, that the arrival times are exponentially distributed, as the underlying process is Poisson. Of course, this is only correct if the states are fine grained enough, i.e., the time scale is chosen in a way to be able to catch all present states, see ref. [56].

The choice of the right time scale has still to be solved at this point and is not trivial, even if the distribution of timings is known. The solution to this problem is derived with the concept of Poisson hidden Markov model, see ref [57].

Essentially, the idea of a hidden Markov model (HMM) is, that the system under inspection separates in two parts, a hidden and an observed one, see figure 2.13. The observed part depends on the hidden one only. Given that the global, measurable system is a Markovian process, the hidden process does not have to obey this restriction. Markovian in this context means, that the next observed state depends only on the current global state (which includes the state of the hidden process). As the hidden process is not observed, its inputs can be arbitrary, and depend e.g. on the global state. In this way, nonlinearities can be modelled, as long as global state changes are tracked with a time scale which is fine grained enough, see section A.3.

From the chemical reactions described in section A.2 rate equations [58, 59] can be formulated. From here two paths can be taken, one for the modelling approach, another for stochastic master equation formulation. While stating the master equation here,

$$\frac{dP_k}{dt} = \sum_k (T_{kl} P_l - T_{lk} P_k), \tag{2.3}$$

the derivation of it and the nomenclature used by this formulation is not needed in this work. Furthermore, its derivation tightly follows the algorithm described below, see ref. [60].

The attention should now be drawn to the model approach. For the processes described above many algorithms are known. Some of them use artificial Markov times, others work with natural time. Some of them are exact e.g., in ref [36], others are approximative, see refs [61, 62].

For the chosen algorithm of Gillespie, the rate constants k_i are converted to propensities c_i to achieve a proper reaction sampling, as described in the algorithm 2.14.

First ideas about propensity definition date back to de Donder in [63]. There, an affinity of the system to the next reaction is defined in terms of the free enthalpy change in a reaction.

In general, the propensities can be defined in terms of conditional probabilities, see ref. [64]. That is why they imply the normative notion of probability, being normalized by the probability of an arbitrary reaction. The sampling used to model reaction propagating is based upon this conditional, relative character of reaction probability.

However, the notion of continuous time, proved by Gillespie in [36] also uses the conditioning part of the propensities. In his algorithm this is done, by formulating the algorithm in a non-rejective manner, implying that the probabilities of described reactions form the probability of an arbitrary event exclusively. This approach is known as model marginalization [65].

In other words, the marginalization is not enforced for the notion of time. However, some normalization assumption for defining (reaction) event propensities has to be made. Modelling physical time can only be achieved, if the normalization assumption takes all reaction rates into account, which are present in the model. Moreover, if all reaction rates are taken equally into account, all assumptions that can be made are equivalent, see ref. [48].

The separation of concerns is emphasized here:

- The normalization constant, which is used to convert reaction propensities c_i into unconditioned probabilities is used for time calculation.
- The remaining probabilities are used to sample reactions in a correct manner.

Given (relative) reaction probabilities, the dynamics is pinned down. However, the time scale is only provided, if the normalization constant is recorded along with the the reaction probabilities.

The process of marginalization (and generally, normalization) can be inverted, however, by making some assumptions. This procedure is also known as embedding a DTMC into a CTMC, see ref. [66]. There are various possibilities of how to accomplish the embedding, see ref. [67], called as realization problem. Essentially, the question to solve is how to recover the normalization factor.

This is by design, as a normalization process, used for probability derivation, removes a common factor from the values being normalized. That is, if reaction propensities are given, the observed dynamics correspond to the dynamics described by reaction equations. However, the notion of time is lost. The (normalized) propensities do not carry any information about the original values of reaction rates.

The transition from rate equations to reaction propensities is technically solved by Gillespie in [36]. Therein an intensity vector $\overrightarrow{\lambda}$ with entries for each distinguished event is defined. In a reaction system, the entries in the vector take care of reaction rates, as well as reactant amounts and have the units of $1/[t]$. Then, the intensity

$$\lambda = \sum_i \lambda_i \tag{2.4}$$

is interpreted as the Poisson intensity for a reaction to happen in the time span $[t, t+dt]$, see also ref. [68]. The reaction is chosen in a random manner from the intensity vector $\overrightarrow{\lambda}$ weighted by its components. The time needed for the chosen reaction is the inverse of the intensity λ.

The inverse problem is much harder to solve and can generally only be tackled numerically. If experimental series are provided for model training, relative model parameters λ_i can be found, governing the proper reaction sampling. The time scale is left out. To find the proper time scale either a time scale parameter has to be provided along time series or a hierarchical, so called *hyperparameter* optimization has to be formulated. See e.g. ref. [69] and section A.3.

The present model works with natural time. To achieve this, the propagation algorithm has to couple time propagation with the Poisson intensity function, defined above. After this, the remaining algorithms break up in two groups: rejection kinetic Monte Carlo methods and those which are rejection-free. The difference is, that the former accept time going on without any change in the overall system, whereas the latter exclude such cases. Being defined in such way, the rejection-free algorithms build a proper subset of algorithms with rejection, namely those, where probabilities to undergo a self-transition of a system state are zero. Moreover, capabilities of the two groups of algorithms to model physical time are equal [48].

Now, as the choice across all known algorithms is arbitrary, it was made according to the knowledge and elaboration level. The Gillespie time propagation algorithm was defined in [52]. The algorithm is widely used and some enhancements for specific systems were made [70, 71, 72, 73] since its introduction. In this work the enhanced logarithmic direct method defined in [74] is applied.

In [75] a further simplification was proposed for exact propagation algorithms, which was also taken into account.

With the the enhancements above, the diagram of the Gillespie algorithm reads as in figure 2.14. For calculation of propensities there are two cases to differentiate

- reactions of first order, like

  $$X \xrightarrow{k_i} P \tag{2.5}$$

  For these reactions the propensity is calculated by

  $$c_i = k_i \cdot X \tag{2.6}$$

  where k_i denotes the reaction constant and X the concentrations of species for the reaction.

- reactions of second order, like

$$X + Y \xrightarrow{k_i} P \tag{2.7}$$

In this case the rate constant has to be normalized by the molar volume, to stay comparable with propensities of other reaction orders.

$$c_i = \frac{k_i}{N_A \cdot V} \cdot X \cdot Y \tag{2.8}$$

where k_i denotes the reaction constant, N_A the Avogadro constant, V the reaction volume and X, Y the amounts of species for the reaction, taken in their discretized form.

See also ref. [32] for propensity calculations for other reaction orders.

2.3 Boundaries of validity

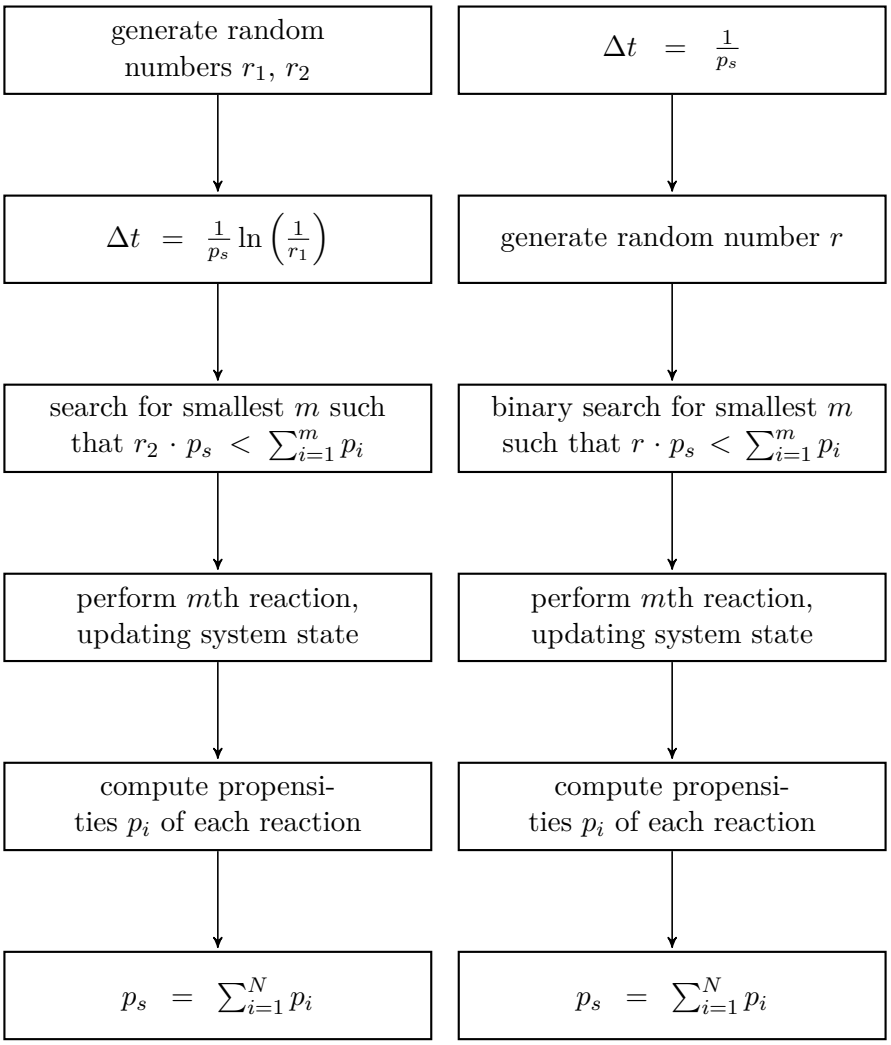
The model is a mean field model in the sense, that diffusion is neglected during simulation. Reaction rates drive essentially all possible reactions. Effects taken into account, like enzyme-enzyme interactions, substrate-surface properties, and implied biological enzyme properties or other coupling effects can be seen as perturbances to these rates. Reaction rate constants are assumed to be constant properties of enzymes. To implement this, a discretization of the system has to be defined. The total amount of enzymes and substrate remains constant during a single simulation run, see ref. [76].

After discretization the simulation is carried out by solving the according Langevin equations, as mentioned in section 1.2 This is done by constructing an appropriate Poisson hidden Markov model, see e.g., refs. [77, 78]. This has the desired distribution as its equilibrium distribution and produces a sample from this distribution for each time t, until the desired simulation time is reached. The distinct feature is, that the Markov chain in this model is constructed using substrate and enzyme properties only. Each step during a simulation run is carried out by using the time-reaction propagating algorithm, explained in section 2.2. The result of each step contains two statements.

- a time, sampled from an exponential distribution with the scale parameter $1/\lambda$, see section 2.2
- a reaction, sampled in a random manner, according to the reaction propensities vector $\overrightarrow{\lambda}$.

Both are saved. The reaction changes the global state of the system, the time contributes the according record to the time axis.

As the time is a result of sampling from a distribution, all informations have a distribution along the time axis. Moreover, the distribution is exponential, so the sequence of sampled times is in general totally inaccessible, see refs. [79]. I.e., knowing all times sampled so far, the next sample remains unknown and unpredictable.



(a) Original Gillespie algorithm.

(b) Gillespie algorithm with enhancements.

Figure 2.14: Gillespie algorithms, original, defined in ref. [36] is reproduced in (a) vs. enhanced one shown in (b). In comparison to the original, one less random number is generated, a logarithmic operation is omitted, see ref. [75] and a binary search is used for better performance at the expense of using more space for storing of the propensity array, see ref. [74].

Every reaction leads to a global state change, as the used algorithm is rejection-free. However, showing e.g., the time evolution of hydrolized products or the remaining substrate, does not reflect every reaction equally. This is, because not every reaction leads to a conversion of substate to product. For example, the reaction of adsorption or desorption leads to complexation or decomplexation, but does not change substrate amount. Nor does a hydrolysation of a polymer, if the hydrolization generates two insoluble polymers.

In this work, most plots concerning concentrations imply that reactions happened before being aggregated. This is the natural way of analyzing the effect enzymes as an active substance, due to the experimental accessibility. The aggregation happens over a time span, namely the time span since the beginning of measurement/simulation, as well as over different reactions, which have nominal character and are in general incomparable.

The sampling of a nominal variable has to be explained. This is done by assigning a propensity to each reaction. Being at least an ordinal variable, the propensities vector can be ordered, weighted and used as sampling outcomes.

Here, one has to keep in mind, that as reactions are the outcome of a random process, the aggregated value measured at some point also has a distribution. As the measured value also ignores some of the sampled reactions any direct relation between them is lost. I.e., even if any statistical information about the reactions is given, it is discarded during analysis of concentrations.

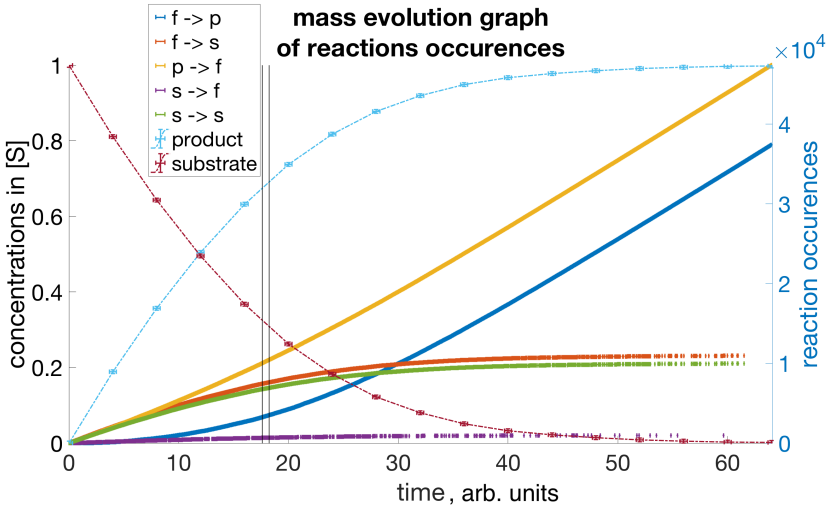
At this point however it is interesting to show how reactions are distributed in time, for an example system. Reactions and related entities are of different nature in general. I.e., they form categories, rather than comparable values. This is why methods for categorical variables should be applied accordingly, see ref. [57]. A new approach of depicting such data is the so called mass evolution graph, developed by Ribler in [80], see figure 2.15.

As seen in figure 2.15, there are time spans assigned to each reaction. The time spans are non-overlapping but contiguous. This is another interpretation of the exclusiveness of reaction sampling. Each time span corresponds to the exponential distribution of time samplings $1/\lambda$. I.e., the duration of a reaction and the error of time measurement of the manifestation of this very reaction is interpreted to be the same.

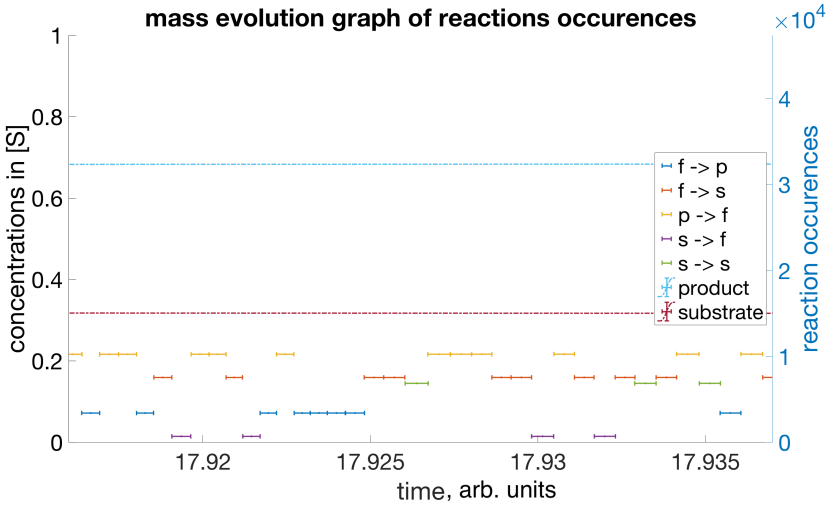
The mass evolution graph was chosen as a convenient method to represent the underlying simulation procedure of the model. It also allows us to apply statistical analysis methods to reaction arriving times, see figure 2.16. In the basic Michaelis-Menten kinetics, described by the model 2.11, the product related graphs are absent. With the presence of inhibition in the model, two new time evolution series arise.

The new time series depend on the product amount, however, they are independent from the former given time series. This fact allows to encode more information about the system state. For example, the velocities of substrate associated reactions can be related to the product associated for inspecting the product saturation level. This cannot be achieved without taking inhibition by product into account.

The identified parameter to control the standard error of a concentration measurement is the system size. The discretization ratio between reaction compartements is fixed

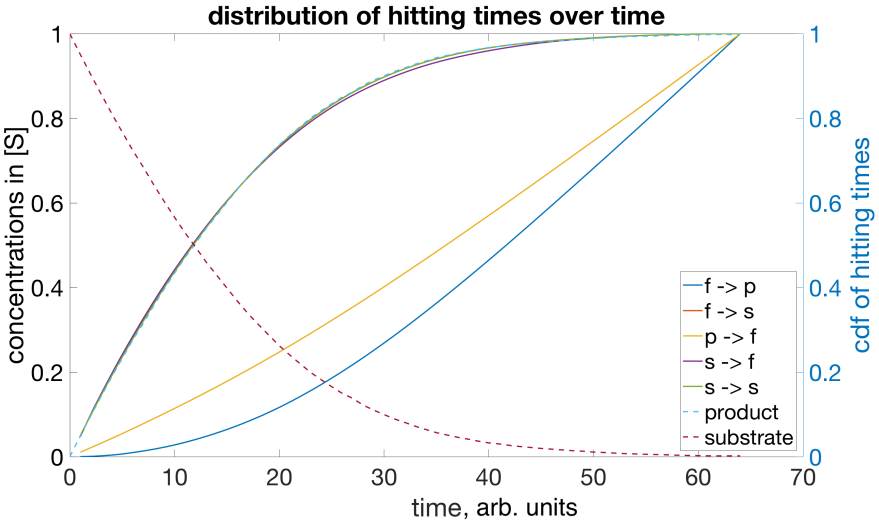


(a) Depicted are counted events of different reactions and concentrations of substrate and product vs. time. Different lines correspond to different reactions, according to the legend. In figure (b) the marked part is shown with a larger scale.

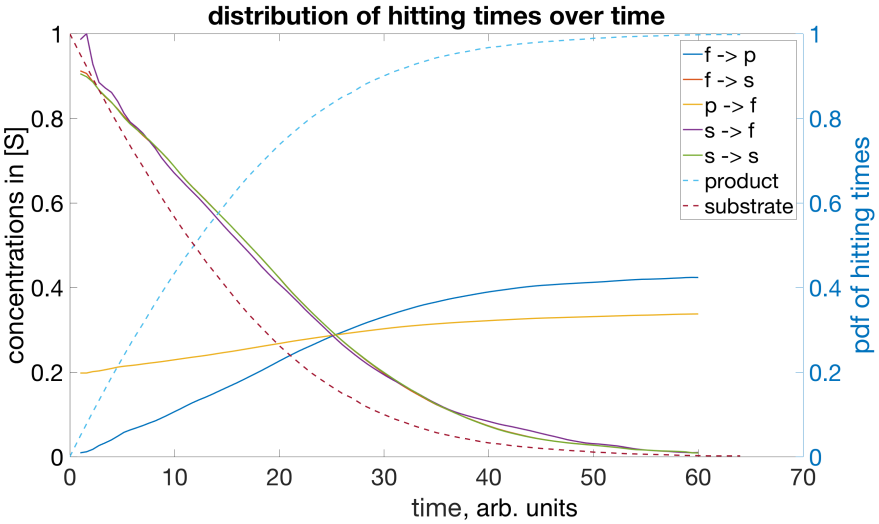


(b) Small time frame of reaction events shown figure (a). Given a time point, the concentrations of reactants are the result of some specified initial conditions and reactions until this time point.

Figure 2.15: Simulated data points of concentration as well the simulated reactions in between. Reactions are sampled at random, corresponding to their propensities and are mutually exclusive by design. The lines connecting the simulated concentrations are depicted for guidance purposes. Concentration data is distributed along both axes. Distribution along time is due the distribution of reaction arrival timings, as shown in figure (b). Distribution along the the ordinate axis is due the random reaction sampling.



(a) The sampled reaction arrival times, shown in figure 2.15 are now shown as cumulative distribution on the support of recorded time of the simulation. Reactions related to the substrate, especially the hydrolization reaction, depicted as $s \to s$ coincide with the product time evolution. The time evolution behavior is further analyzed in figure (b).



(b) From CDFs in figure (a) are taken derivatives and the according PDFs are shown. These are the velocities of reaction propensities evolutions.

Figure 2.16: Figure 2.15 contains essentially only reaction counts the interpretation of their occurrences can be achieved via the distribution of their arriving times. In figures are shown their CDFs and PDFs accordingly.

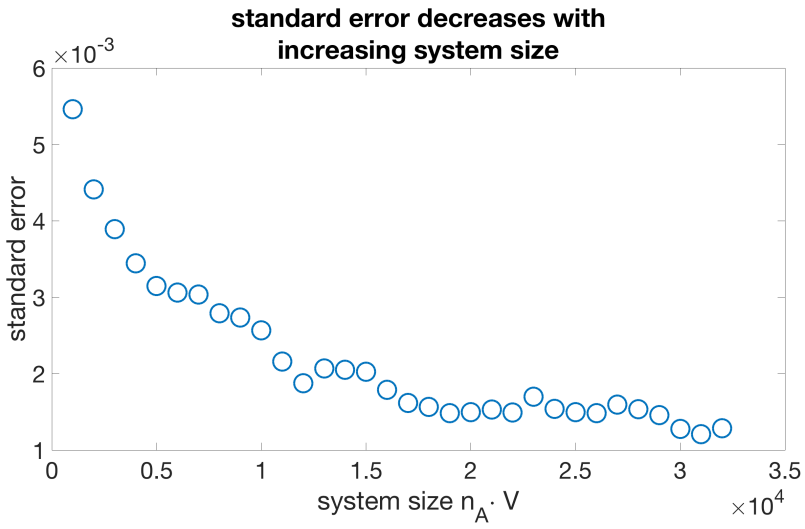


Figure 2.17: Standard error of the simulation diminishes with increasing number of enzymes. The number of enzymes is proportional to the system size $\#_E = N_A \cdot V \cdot E_T$. I.e., the control parameter of the model precision is the chosen system size.

by the experimental setup. So, the system size influences the number of entities in the simulation, while their ratios remain fixed. I.e., the amount of substrate and the amount of enzymes are sampled by means of the common discretization factor $N_A \cdot V$, where V is the system volume, or system size. The ratio of sampled units maintains some required ratio, predefined by the user. However, the system size remains a free parameter and can be renormalized to represent real systems of various sizes. In figure 2.17 the dependence of the standard error with respect to the system size is shown.

Being the square root of variance, the standard error is defined as the square root of

$$\sigma^2 = \sigma_t^2 + \sigma_c^2 \tag{2.9}$$

For each measured point of resulting time series of concentrations:

- σ_c^2 is the variance along the ordinate axis. Usually concentrations at a time point are measured on this axis.
- σ_t^2 is the variance along the time axis. As seen in figure 2.15, this part of the error is expected to be small in comparison to the other.

That is, every given time point cannot be predicted exactly and will also have a distribution with a given mean. This is the exponential scale parameter $\sigma_t = 1/\lambda$, discussed above and shown in the figure 2.15.

The question about how many samples are needed for significant statistics is also relevant. Here, the following qualitative argument is used: even if variance has a dependency

on the sample number:

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2 \tag{2.10}$$

where μ is the mean of samples taken, and x_i is a single sample, in general, variance has a saturation behavior [81, 82]. Therefore, any sample size beyond the saturation point can be taken. For figure 2.17 120 measuremens for each data point were made. Given this, the values of means and the values of variances, still remain statistical quantities, and depends also on the random number generation, which is in general unpredictable.

Diffusion

Diffusion is an important feature of many physical models. As such its role in a given system has to be inspected.

For this, beginning with the diffusion equation [83]

$$\frac{\partial C}{\partial t} = D\nabla^2 C \tag{2.11}$$

let's consider the fundamental solution in 1D

$$C(r,t) = \frac{1}{(4\pi Dt)} \exp\left(\frac{-x^2}{4Dt}\right) \tag{2.12}$$

and define the diffusion length to

$$l_D = 2\sqrt{Dt} \tag{2.13}$$

as an estimation of the characteristic diffusion length of an enzyme.

Now, using the diffusion coefficient, defined by Stokes-Einstein equations [84, p. 127]

$$D = \frac{k_B T}{6\pi\eta R} \tag{2.14}$$

with

$$\frac{m}{\rho} = V = \frac{4}{3}\pi R_0^3 \qquad \Rightarrow \qquad R_0 = \left(\frac{3m}{4\pi\rho}\right)^{1/3} \tag{2.15}$$

However, the dimensions based on the mass and density are not valid for an enzyme, as it is not a smooth sphere, so we take a factor of sedimentation respectively of friction of $f/f_{min} = 1.5$ into account, see ref. [85].

The estimated density of enzymes is about 1.37 $\frac{g}{cm^3}$ [86]

For $T = 300\ K$ and $\eta = 1 \cdot 10^{-3}\ Pa \cdot s$ we use standard conditions of water and the mass of an enzyme is taken as 50 kDa for a rough estimation. $k_B = 1.38 \cdot 10^{-23}\ \frac{J}{K}$

As the result we get

$$D = \frac{k_B T}{6\pi\eta \cdot f/f_{min} R_0} = 6 \cdot 10^4\ \frac{nm^2}{ms} \tag{2.16}$$

The extended calculation can be found in section A.1 Following eq. (2.13) a length of 1.5 μm of diffusion is achieved after about 10 ms. This is the typical length scale of a substrate.

As simulations are meant to be driven in the range of substrate saturation it is assumed that fluctuations across substrate don't play an important role and are therefore omitted. By neglecting diffusion, it is assumed, that every reaction is slow enough, to neglect concentration gradients.

System discretization

The system is being discretized on a grid with the unit cell spacing defined by experimental data. This happens while system size remains not fixed and free to choose. The grid size depends on the ratio of the active and passive species in the simulation. This ratio is given by the experimental data or user input and is not a free parameter. The system is therefore "discretized by monomers".

The essential value for discretization is the system size V [87] and the molar volume, defined by $N_A \cdot V$, where N_A is the Avogadro number.

Therefore, the number of entities taking part in the simulation is given by:

$$\#_S = N_A \cdot V \cdot S_T \tag{2.17}$$

$$\#_E = N_A \cdot V \cdot E_T \tag{2.18}$$

where S_T and E_T are the total concentrations of substrate and enzymes in the solution respectively.

Typically, the ratio of E_T/S_T is given in $\mu mol/g$ of substrate and the concentration of S_T in mol/vol. Taking into account, that the system size V remains a free parameter in the simulation, one can parametrize the total concentration of enzymes by the concentration of substrate, set $S_T = 1$ and choose an appropriate volume V to minimize the discretization error for the enzyme amount. The error is minimized, when the product for $\#_E$ is a natural number with minimized rounding error.

Taking this approach, total mass of substrate is calculated by

$$m_S = N_A V \cdot \frac{M_S}{N_A} \cdot S_T \tag{2.19}$$

where M_S is the molar mass of the substrate (= 162g/mol), see ref. [2].

If all concentrations are normalized by the substrate concentration, then enzyme type concentrations are automatically satisfied by plugging in either enzyme type concentrations itself or the number of enzyme samples into the simulation.

2.4 Enzyme description

In the chosen approach enzymes are described as 'finite state machines'. The concept was formalized by Harel in [88] and is a convenient way to represent a complex behavior, see refs. [89, 90, 91, 92, 93, 94, 95]. Also stochastic enzymatic processes can be functionally

described by the formalism, like in refs. [96, 97]. However, in these works it lacks a systematic application of the captured description.

Here, the approach facilitates a detailled description of the internal states of an enzyme. E.g., the enzyme can be adsorbed to substrate. After an adsorption a hydrolization step is possible. However, it can also be adsorbed in a non-productive way [98, 99]. After such a binding, the enzyme can only desorb. Enzyme states are mutually exclusive and a certain amount of time is needed for transitions to happen.

Some preliminary assumptions are made before discussing the behavior of enzymes. These concerns geometrical and logical constraints to guarantee operability of an active entity in the present system in general.

- As mentioned in section 2.1, there were three states for enzymes identified: ”free”, ”bound” ”inhibited”. This is done especially to adapt the model of enzymes to the underlying chemical reactions. With the equivalence of internal enzyme states and the underlying chemical model it is possible to directly relate all simulation parameters to substrate and enzyme properties.

- If an enzyme is bound to, or inhibited by a polymer chain, no other enzyme is able to bind to the same location. I. e., an enzyme exclusively ”owns” the substrate area it is bound to.

- The states an enzyme can take are exclusive. There are no mixed states in between.

- If a desorbed, (i.e., ”free”) enzyme is chosen for an action, an accessible substrate area is chosen for binding. If the binding site is located at a hydrolysable polymer chain, the enzyme transits to the bound state. Otherwise, the enzyme transits to the inhibited state.

- There can be substrate areas, where no binding is possible. These are excluded from considerations for an enzyme type while binding, until they become accessible. Due to this, the propensity function for reactions changes its magnitude, retaining the mutual probability ratios of reactions. Updates of substrate properties follow, among other things, geometrical considerations of the bulk used. Therefore, the propensity function updates in a nonlinear manner.

The rates to act on (bind to, desorb from and hydrolize) crystalline and amorphous substrate regions can be defined independently. It is also possible to optimize their difference based on one of the rates set. For the latter case, a one-parametric model was chosen and verified, see section 4.4.

Starting from the “free” state, the enzyme binds to a randomly chosen site at the substrate that is accessible to the enzyme. Additionally, the structural features of enzymes include their specific binding area. This area defines the maximal area of substrate surface, which becomes inaccessible to other enzymes after binding. If the binding site is located at a soluble polymer chain that cannot be cut, the enzyme undergoes a transition to the ”inhibited” state.

In the "inhibited" state the enzyme undergoes transition to the free state, determined by the type of the enzyme. The probability can further depend on the time spent in the current state, and on the soluble oligomer chain length it is blocked with.

In the "bound" state the hydrolysis behavior of the enzyme is also probabilistic and is defined by the enzyme type. It can either unbind or proceed with the hydrolysis. If the enzyme is of processive type, it remains adsorbed after a performed hydrolysis step. This continues until it reaches the end of the polymer chain or a blocked range of the chain, with non-zero probability to desorb underway. Addionally, the probabilities of actions depend on the crystallinity characteristics of the substrate, to which the enzyme is bound.

In particular, enzymes of various organisms differ in the way they bind to and hydrolyze the substrate, and how much time the transitions between different states need, see section 2.1. Each of these properties can be taken into account within the present approach. Furthermore, the geometrical structure plays an important role in cellulose degradation. Typically only the surface of the cellulosic bulk is accessible to enzymes. The difficulty here is that the surface changes over time. The information about the accessible surface has to be kept up-to-date for all enzymes in order to achieve correct behavior of the overall system. With the current approach the influence of the geometry on the enzymatic celluose degradation can be studied in more detail than with any other model until now.

In figure 2.18 the general state machine for considered enzymes is shown. There are three main exclusive states of an enzyme: being either adsorbed, inhibited or desorbed.

In general, there are n^2 transitions possible, if n is the number of states. However, taking the time propagation algorithm into account, some of the transitions have to be excluded. These are self transitions from state E_f and E_p as they would correspond to a rejection of any action. This contradicts the chosen time propagation algorithm, which excludes transitions, that do not change the global state. Furthermore, the transition from E_p to E_s has to be excluded, as this would imply a recombination of products to substate. Such recombination is unlikely to happen, as stated by e.g., Bisswanger in ref. [58]. Thus, there are six transitions left. They can be naturally interpreted in the context of Michaelis Menten kinetics as well as in the context of the Gillespie time-reaction propagation algorithm:

- $E_f \longrightarrow C_s$ corresponds to enzyme binding
- $E_f \longrightarrow C_p$ corresponds to enzyme inhibition
- $C_s \longrightarrow E_f$ and $C_p \longrightarrow E_f$ correspond to enzyme desorption
- both remaining transitions correspond to hydrolyzation processes. The path taken depends on dynamic properties of the enzyme and substrate, which are not depicted in this figure.

On the edges of the general state machine in figure 2.18 propensities for the according reaction are depicted. For propensity calculations see section 2.2.

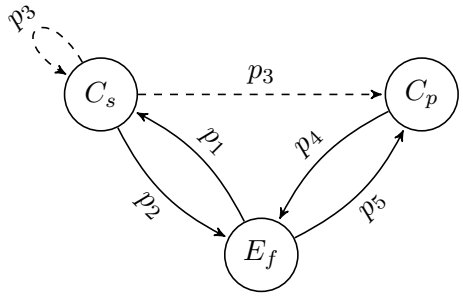


Figure 2.18: Common state machine for enzymes. The probabilities of each transition are depicted on the edges. Each transition changes the state of the whole system, by interaction of enzyme and substrate. Following reactions are shown:
adsorption: $E_f \longrightarrow C_s$,
inhibition: $E_f \longrightarrow C_p$,
desorption: $C_s \longrightarrow E_f$, $C_p \longrightarrow E_f$,
hydrolysis: remaining transitions starting from C_s are mutually exclusive, depending on enzyme type and dynamic substrate properties.

Common enzyme properties

Activity

The activity is the substantial chemical characteristics of an enzyme. There are some varying definitions of activity, which comprise enzyme units, volume activity, specific enzyme activity and enzyme velocity [100]. All of them represent the ability of an enzyme kind to convert a substrate to a product.

Starting with the specific enzyme activity A the turnover number $\#_T$ can be calculated to

$$\#_T = \frac{A \cdot M_E}{10^3 \cdot 60} \frac{1}{s} = \frac{v_{max}}{E_T} = k_{cat} \tag{2.20}$$

where A is the activity given in $[A] = IU/mg$, M_E is the molar mass of the according enzyme given in $M_E = mg/mol$, v_{max} is the maximum conversion speed, measured in initial enzyme assays. see ref [101].

The second part of the relation comes from the inspection of Michalis-Menten kinetics, see section A.2.

$\#_T$ is a measurement of frequency for every active site in conversion process. For simplicity, only enzymes with single active site are taken into account in the current model. However, enzymes with more active centers can be added without any significant difficulties.

Specific products

From various experimental data, it follows that different enzymes produce different products with different predispositions. I.e., the size of the active center of the enzyme define the lower possible limit of polymer chains, which can be hydrolyzed.

During the simulation for each polymer it has to be dynamically decided, whether it can be hydrolyzed and a reaction propensity has to be assigned as well. The size of the active center specifies the former result.

The measurement of the size of the active center can be inaccessible in general. It is meant to be replaced by the information about the smallest measurable product in this case.

Specific binding area

Beyond the size of the active center, the effective binding area of an enzyme has to be taken into account. They do not limit the reaction ability of an enzyme, but define the maximum area of the substrate that is going to be blocked after an adsorption. This property is not essential for carrying out a reaction as opposed to the information about specific products above. If the exact information about an enzyme is not known, it can be estimated. The value of the specific binding area is expected to be larger than the value of the according active center.

The specific binding is assumed to be the sum of the carbohydrate binding module extents and the linker if they are present, otherwise to be of the size of the catalytic domain. Each enzyme tries to block the current specific binding area of the chosen polymer. However, the adsorption or inhibition reaction are not dependent on the availability of this space.

Synergism effects

The model is meant to be driven in substrate excess mode. I.e., synergism effects are mainly indirect and stochastic, in the sense of coupling via the substrate and the appropriate propensity changes. However, direct enzyme-enzyme interactions cannot be excluded and are seen as rare events. For example, if an enzyme is inhibited and a nearby enzyme is desorbing, then the inhibited one has the ability to expand his adsorption area to perform a transition to the adsorbed state. In the tested regions of model parameters no influence of this feature could be observed.

Reaction rates

The number of reaction rates depends on the kinetics model taken into account. The basic Michaelis Menten kinetics assume three different reaction rates with rate constants k_1, k_2 and k_3, see figure 2.11. They are responsible for adsorption, desorption and hydrolysis processes respectively.

However, being somewhat unnatural, see ref. [97] the model in this work takes inhibition by product into account, see figure 2.12. With this extension the number of reaction

rates number increases by two, with rate constants k_4 and k_5. These are responsible for binding and desorption of products of the enzyme.

As a further approach to natural behavior a dependency of reaction rates on substrate regions can be modelled. For this purpose, more than a single set of reaction rates are stored in the enzyme model. For every distinct region of the substrate a specific set of reaction rates is stored.

For example, for modelling rate dependency on crystallinity, there are two reaction rate sets with rate constants k_{1a}, k_{2a}, k_{3a}, k_{1c}, k_{2c}, k_{3c} saved. Each set describes the enzyme behavior at amorphous and crystalline substrate regions. For modelling a product dependent inhibition for every product length a specific inhibition rate with rate constant k_{4l} can be stored.

It is also possible, to store only one set and imply some dependency model on substrate characteristics. Such an approach is shown in section 4.4.

Processivity

Some enzyme types are processive in nature, see ref. [102]. I.e., enzymes of these types remain adsorbed after a product is generated, move along the substrate and keep hydrolyzing as long as the adsorbed polymer meets the necessary criteria.

The unbinding probability is independent from this feature and remains non-zero, as described by the underlying model 2.12. However, the ability to stay adsorbed after a product generation cannot be expressed by a chemical equation. In this case a reaction graph like 2.18 is a more convenient model description.

Orientation

There is contradictory information about relevance of orientation of enzymes, cf. refs. [103, 104], and models of Kumar and Levine, refs. [26, 23], where orientation is omitted. This also includes modelling of a fixed position of the active center of an enzyme.

However, the microfibrils in the substrate clearly do have an orientation, see e.g., ref. [40]. An enzyme type therefore has at least a notion of orientation, however, not every modeled enzyme type has an elaborated propensity dependency on it.

Mode of action

It turns out, that all other features are not enough to describe the enzyme behavior with the needed detail. Another relevant property is the mode of action. Three different enzyme modes are taken into account: endoglucanasic, exoglucanasic and β-glucosidasic modes.

- Endoglucanases act on the interior of polymer chains, see figure 2.20. They are able to perform one or more hydrolizations before desorbing even if they are not processive, see ref. [26].

- Exoglucanases act on the exterior of polymer chains, see figure 2.19. I.e. they bind to reducing or non-reducing polymer chain ends, depending on the modelled kind and cleave one or more glucan units from the polymer.

- β-glucosidases bind only to already soluted polymer chains, see figure 2.21. I.e., they act on short polymer chains, which are known to be inhibiting for many other enzyme types and produce glucose from them.

Activity description

With enzyme behavior defined by a finite state machine, see figure 2.18, all possible actions for an enzyme are determined. The state machine is sufficiently general to catch all possible events, omitting all events with probability of zero.

Furthermore, the description allows to generalize and to formulate events across different enzymes and enzyme types in an abstract manner. The generalized description allows to formulate parts of enzyme dynamics preemptively and separated from enzyme type formulation itself and to reuse these parts across all enzyme types.

It turns out, that only the hydrolization process is unique to every enzyme type, whereas the binding and desorption can be formulated in a general manner, as follows.

- First, an enzyme is chosen for reaction, according to current propensities of the system.

- If the enzyme chosen for a reaction is desorbed, only adsorption is possible. The reaction partner is chosen, again according to the current propensities of the system. If the partner is a hydrolyzable polymer chain, the enzyme undergoes the transition to the adsorbed state, otherwise to the inhibited state. A part of the chosen polymer chain is blocked according to the specific binding area defined for the enzyme type.

- If the enzyme chosen for a reaction is inhibited, only desorption is possible. The enzyme undergoes the transition to the ”free” state, freeing up the polymer chain it is inhibited by.

- If the enzyme chosen for a reaction is adsorbed to a hydrolyzable polymer chain, it can desorb from or hydrolyze the polymer chain at its active center.
  - If the desorption reaction is chosen, the enzyme undergoes the transition to the ”free” state, freeing up the polymer chain it is inhibited by.
  - If the hydrolyzation reaction is chosen, the reaction is carried out according to the type specific description. Some of them were defined for the current model, see below for their description.

Modeled enzyme types

Michaelis Menten enzyme prototype

The archetype of an enzyme is an artificial type, that is supposed to strictly follow the Michaelis Menten kinetics. This type serves the goal to reproduce numerical solutions of differential equations (A.7) and (A.8).

This archetype builds the core of all other enzymes, which can be described in terms of this archetype. The modelling of different enzymes should be kept very fine-grained. To achieve this, only this simple archetype was defined, containing all the properties which are common to every enzyme kind. Additional properties are added as needed for any further developed type.

For this purpose the following common properties of the enzyme archetype are assumed:

- up to five reaction rates
- specific binding area is equal to the size of the active center, which is one glucose unit large.

After the choice of an enzyme entity, the reactions are carried out as described above. The reaction which remains to be described is the hydrolyzation reaction.

During the hydrolysis the enzyme cleaves the polymer and produces a new reducing and a new non-reducing polymer chain end. As the type is appointed to reproduce the Michaelis-Menten kinetics, it immediately undergoes the next transition. Depending which model is taken into account, the next state is the ”free” or ”inhibited” state, accordingly to figures 2.11 or 2.12.

Modelling results match experimental and artificial data quite well, see section 3.

The modelling of three reaction rates, like in the basic Michaelis Menten approach is also available. However, these dynamics seem unnatural [97] and incomparable to other more realistic enzyme types.

CBH(I) enzyme type

Additionally to the Michaelis Menten archetype, the CBH(I) enzyme type can have a carbohydrate binding module with a flexible linker. It is modelled as a processive enzyme type, which forms glucose as well as triose. However, its preferred product is cellobiose. The enzyme acts on reducing ends of polymers in a processive manner, staying attached to the polymer after a hydrolyzation reaction. This is an exoglucanase type, see figure 2.19.

This type can have various inhibition rates, depending on the product chain length. As both types of Exocellulases exist, acting either on reducing or non-reducing polymer chains, an orientation of enzymes is implied.

Figure 2.19: Schematic representation of the exoglucanase enzyme type. The enzyme exclusively acts on reducing or non-reducing polymer chains ends without differing between soluble and insoluble chains. The green tilde represents active center of the enzyme.

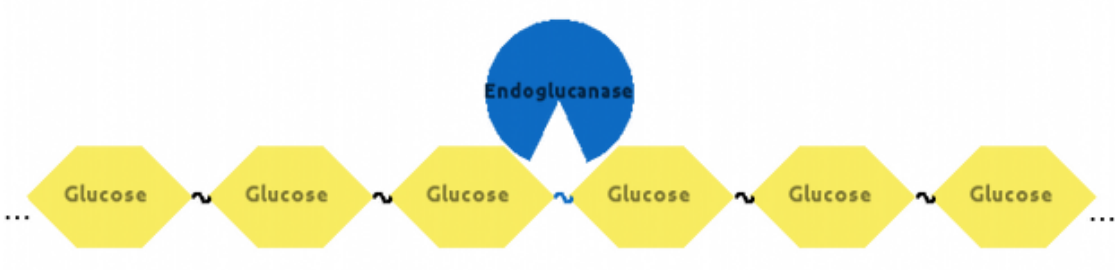


Figure 2.20: Endoglucanase enzyme type. The enzyme acts exclusively on the interior of polymer chains without differing between soluble and insoluble chains.

EG(II) enzyme type

Some variations of Endoglycanases exist. As an example, a type with a catalytic domain, but without a carbohydrate binding module, was implemented. It does not act in a processive manner, however, some polymer cuts are performed, before the enzyme either desorbs or inhibits. After a hydrolysis step is performed, depending on polymer properties where the enzyme is adsorbed to, desorption, inhibition and remaining in adsorbed state is available for this enzyme type.

This type can have various inhibition rates, depending on the product chain length. An orientation concept is also present for this enzyme type, however, it only influences the behavior during adsorption, rather than its propensity.

β-G enzyme type

The β-Glucosidase was modeled as a non-processive enzyme type, attacking only soluted polymer chains, see figure 2.21, These chains are assumed to have a length below a cutoff, assumed to be $L_0 < 6$, see section 2.5.

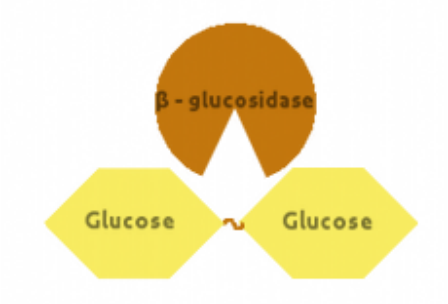


Figure 2.21: β-Glucanase enzyme type. Acts on soluted polymer chains only, in contrast to endo- and exoglucanases.

After a hydrolysis step is performed, depending on polymer properties which the enzyme is adsorbed to, desorption, inhibition and remaining in adsorbed state are available for this enzyme type.

In most cases, there are no appropriate polymer chains around at the beginning of simulation, as no products are present yet. This is of course a valid situation and is the reason why optimization of this enzyme type can only be performed in presence of another one. The other enzyme type should perform according actions to produce valid substrates for β-Glucosidase in this case.

Furthermore, this enzyme type cannot depend on crystallinity, as the only substrate chains it degrades does not belong to any substrate region any more. Also, no inhibition dependency on different products can be implied. The natural product for β-G was assumed to be glucose, so other polymer chains do not lead to any inhibition effects.

2.5 Substrate description

The Cellulect model contains an explicit model for the substrate. The substrate model concerns physical and chemical properties as well as a geometrical representation of the substrate object.

The first question to be solved was how to generate qualitative and quantitative applicable substrate to work with. As the application required to maintain a certain concentration ratio between enzymes and substrate, a common multiplier had to be found. This was already defined as the molar volume $N_A \cdot V$, see ref. [32]. Thus the discretization of the substrate can be stated as in equation 2.17.

A proper discretized substrate enables first quantitative simulations based on rates and propensities. However, it is not enough to catch all features of a more realistic substrate like its exposed surface, polymer distribution, lignin coverage, crystallinity regions. In order to do this, some geometry has to be assumed for the discretization units and for each of them has to be implied an identity. The geometry implicitly stores shape information about the substrate, all other features are uniquely mapped onto and stored with the identity of discretized units.

Further details about the features are described below. For now, the question about the identity and the form factor of each discretization unit is discussed.

To avoid an introduction of arbitrary units, the smallest units of a substrate were chosen as the discretization units. These are the glucan units comprising the polymers, which for their part form the substrate. This simultaneously solves the questions about the number of sampled substrate units, as well as the length and volume measurement units in the modelled system. However, this also introduces the limitations of sampling. The substrate can contain only an integral number of glucan units and every substrate property which should be expressed is forced to be saved across them.

The second question to be solved is the geometry of the discretization units. A realistic substrate can take almost any arbitrary form, see figure 2.1. Therefore, the form factor of the smallest unit has to be capable to tesselate, i.e., to fill the 3D-space. The structural properties of polymers also imply the geometric constraint of two dimensions behaving similarly, while the third, where the polymer is extended to, behaves differently.

The considerations above lead to inspection of some (regular) plane tilings, and how to add some volume onto it in form of a prism construction. The simplification made at this step is, that the chosen tiling should be moved along a symmetry axis without any rotation or other translation to construct the prism, cf. figure 2.3. In principle, at least every tiling which fulfills the Conway criteria, see ref. [105] and allows periodic tiling, see ref. [106] is sufficient. However, common computational techniques, see ref. [107] and applications, see ref. [108] suggest the use of simple cubic voxels as a base.

Hierarchical substrate representation

Glucose units are assumed to be the smallest substrate units in the model. The substrate is discretized by means of them. Glucose units were modeled as cubic units, with 6 distinguishable faces. The faces of each unit are interpreted as categorical variables. Other physicochemical properties were superposed to the modeled units independently of the geometrical form. Such approach is known especially from diverse imaging techniques, e.g., in ref. [109].

Glucose units build up polymer chains, which in turn form a substrate bulk. Each polymer chain is assumed to be a linear chain of glucose units, connected along a distinguished axis. It is assumed, that all polymers with the length less then a solubility barrier and a common exposed surface are soluble and dissolve immediately. The appropriate faces of the dissolved glucose units as well as their surroundings are kept up-to-date on every enzyme reaction.

The bulk is approximated as a dense block, where polymers with a given initial length distribution are located. Crystallinity and lignin coverage is superposed on other data independently, possibly following a given distribution.

Nevertheless, it is possible to model arbitrary defects by removing single glucose units from initial distribution. An example of such block can be seen in figure 2.22.

Several blocks can be modeled to achieve a more realistic picture of microfibrils which are combined to a macrofibril.

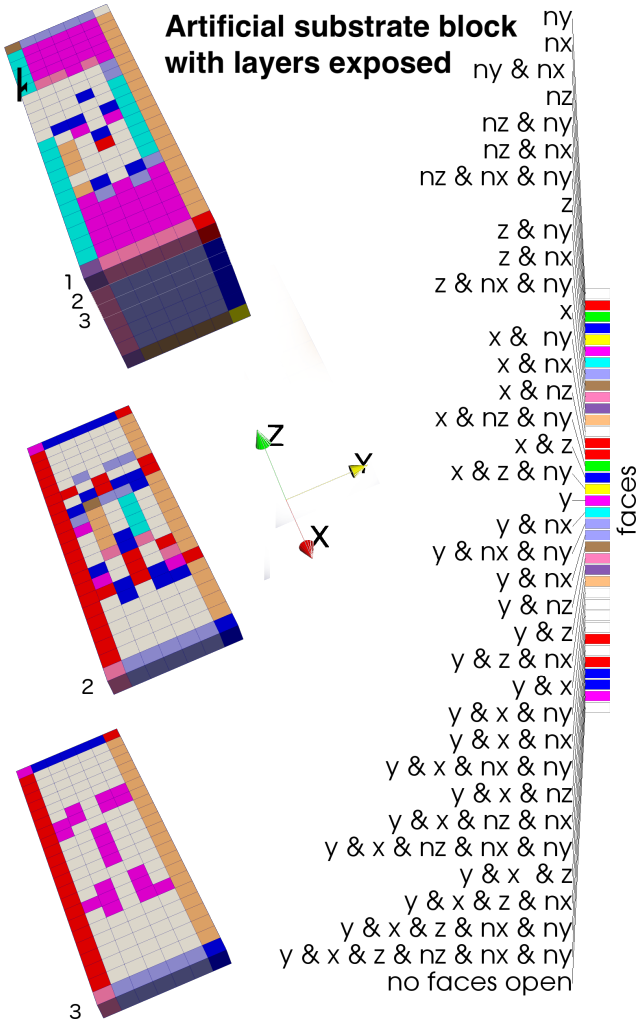


Figure 2.22: Artificial substrate block. Colors depict each distinguishable category of exposed faces of glucose units.
E.g. ny: the face of the monomer in negative y-direction is on suface.
x& nz: the faces of the monomer in positive x- and negative z-direction are on surface.
Polymers are connected along the x-axis. The solubility barrier was taken as 6. I.e., polymers with length < 6 are dissolved.
From top to bottom:
1. Artifical substrate block with dimensions of $7 \times 7 \times 23$ glucose units along the x-, y- and z-axis. The three top layers are numbered. Some polymer parts are removed from the top layer.
2. Second layer of the substrate block. Due to an artificial preparation of the shown block and removal of polymer parts from the top layer, here also some polymer parts are dissolved.
3. Third layer of the substrate block shows some tesselation pattern in terms of exposed glucose faces.

Accessible surface

An adsorption process can only take place on an accessible place of the substrate. If an enzyme is adsorbed to the substrate, any other enzyme is not able to adsorb the same location. Only after desorption, this surface area becomes accessible again. If a hydrolization process takes place, then a polymer is cut into two pieces. If a polymer becomes shorter than a given length, it is assumed to be soluble and dissolves from the main substrate block. That is, during the simulation more and more substrate becomes exposed due to its geometrical structure and the ongoing hydrolization processes.

Besides modelling arbitrary defects, the influence of accessible surface on the hydrolization process can be studied by initializing the substrate to different ratios of surface to volume. The results can be seen in chapter 4. As the substrate is modeled by a block, the volume is calculated by

$$V = l \cdot w \cdot h \tag{2.21}$$

where l is the overall length of the bulk, w the chosen width and h its height.

However, for the surface A only the lateral area is taken. I.e., the simplification is met, that front and back ends do not account for it. This is, because even for Exoglucnases to adsorb it, it is not enough to have only the end of a polymer chain accessible. Some interior glucose units also have to be accessible from another direction. This yields for the surface calculation to be

$$A = 2l \cdot (w + h) \tag{2.22}$$

Now, if a surface/volume ratio is given,

$$\frac{A}{V} = 2 \cdot \frac{w + h}{w \cdot h}, \qquad w, h \in \mathbb{N}, \qquad A/V \in \mathbb{R} \tag{2.23}$$

the cross section parameters, representing the width and the height of the block can be chosen, such that the ratio is met as exact as possible. If two equal choices are possible, the one, which minimizes the absolute difference of $|w - h|$ is chosen.

Crystallinity

Crystallinity is an important feature of cellulose regarding enzymatic hydrolization processes [110, 111, 112]. Crystallinity degrees are commonly given in percentage of the whole substrate. Usual crystallinity degrees lie between 0.5 and 0.9. The vast majority of enzymes act less effectively on crystalline regions compared to amorph ones. Due to this, the challenge in pretreatment stages is to lower the crystallinity degree. Naturally, cellulose forms different regions of crystallinity during its growth, however, the form and position of these regions is not well investigated. Two of the possible modelled distributions are shown in figures 2.23 and 2.24

Being local substrate property, crystallinity is mapped onto the modeled glucose units. During preparation, an initial distribution of crystallinity is build. Typical measurements of crystallinity take place by x-ray diffractometry, so initial distributions are less known. Nevertheless, some assumptions have to be made regarding it. Crystallinity changes during the hydrolyzation process, as degradation changes the accessibility of different

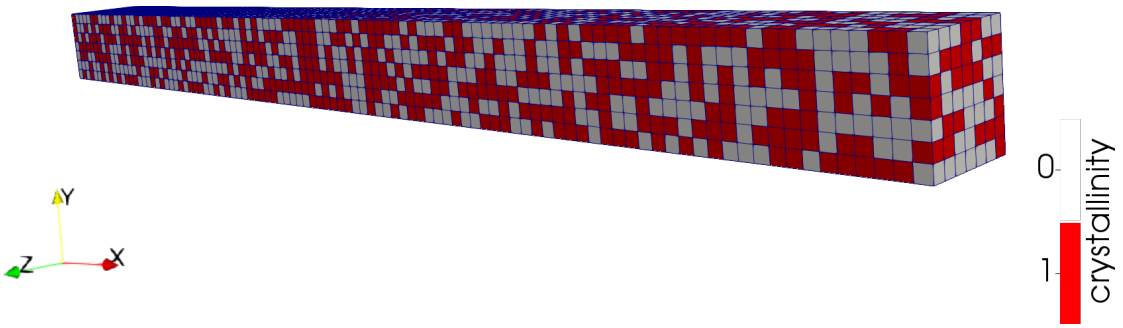


Figure 2.23: Random initial distribution of crystallinity regions across the substrate bulk. Crystallinity degree is set to 0.5. The colors show crystalline vs. amorph regions, mapped onto the monomers in a random manner.

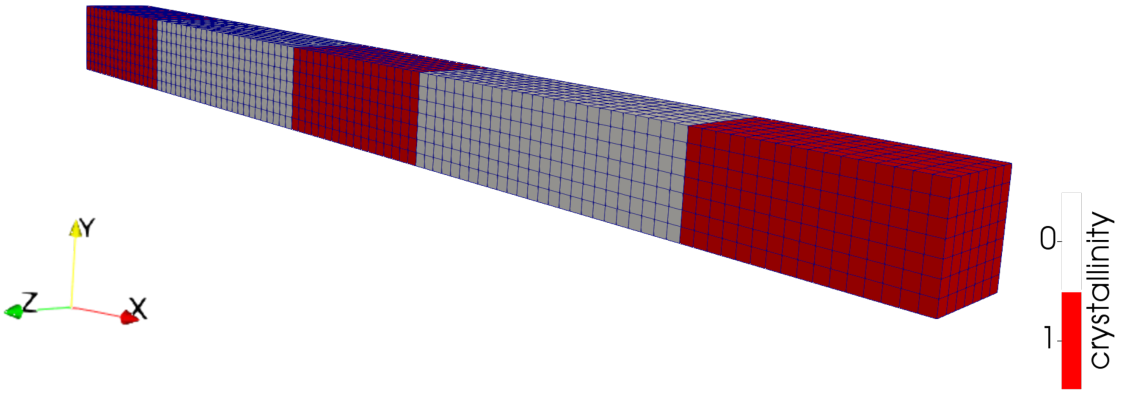


Figure 2.24: Band structure w.r.t. crystallinity, as assumed in [26]. Initial crystallinity degree is set to 0.5. Colors show crystalline vs. amorph regions, mapped onto the monomers in a band structure.

regions. This implies a coupling between crystallinity and the property of exposition of polymers. As a working hypothesis it was assumed, that polymers being exposed from all directions cannot be crystalline. This includes all soluted polymers. Otherwise, glucose chains are crystalline, if they lie in crystalline regions, constructed during initialization. The influence of crystallinity and its distribution on hydrolization processes can be seen in chapter 4.

Polymers as intermediate units

Polymer chains with a length $L_0 < 6$ are considered soluble products [26] and can lead to inhibition of enzymes. In contrast to the glucose units any polymer has an identity. Instead of this, a representative of the polymer chain is chosen. Via this representation the whole polymer can be constructed by inspecting the dynamic properties of each constituting glucose unit, see also sections B.2 and B.1.

Monomers as description units

Monomers are the smallest modeled units. Following [113, Chapter 14.7] the model is therefore volumetric and this property contributes to its efficiency. On every reaction the properties of monomers affected by the reaction are updated. This includes a recursive algorithm of surface updates, which properly transforms hidden surface of a microfibril into accessible one.

Degree of polymerization

The common way to describe the typical length of polymers in a substrate is by means of the degree of polymerization. This is done by observing moments of the distribution of polymer lengths following the definition in [25]

$$p^{(n)}(t) = \sum x^n p(x,t) \tag{2.24}$$

Here, the continuous distribution of polymer lenghts was discretized.

The degree of polymerization is then defined as

$$DP = \frac{p^{(1)}}{p^{(0)}} = \frac{\sum x p(x,t) dx}{\sum p(x,t) dx} \tag{2.25}$$

This is indeed the mean chain length of the substrate.

However, different substrates are prepared and pretreated in different ways [114]. Therefore, no special distribution is favored. Instead, the sensitivity to the distribution itself can be studied by modelling substrate with different polymer length distributions. Considered distributions were

$$\begin{aligned} p_C &= \begin{cases} 1, & x = DP \\ 0, & \text{else} \end{cases} \\ p_P &= \frac{\lambda^k}{k!} \exp(-\lambda), \qquad \text{with} \qquad \lambda = DP \end{aligned} \tag{2.26}$$

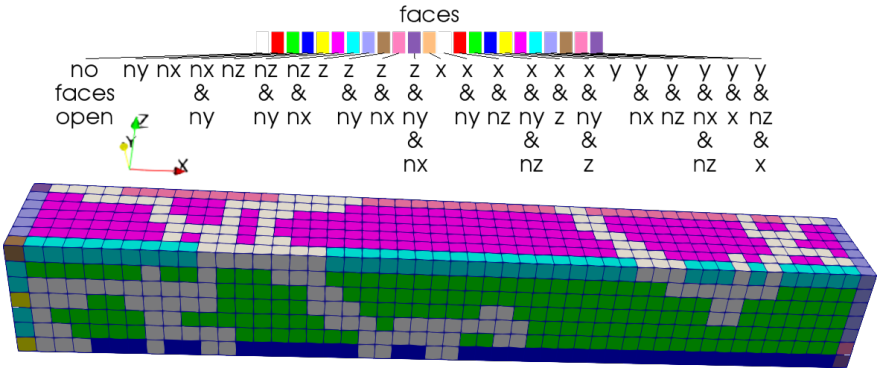


Figure 2.25: A substrate bulk partially covered by lignin. Percentage chosen: 20%. As in figure 2.22, the colors depict accessible faces of the substrate, exposed on the surface. The state of faces on the surface is interpreted as categorical variable and encoded by different colors.
If lignin covers a monomer, it is considered to be not on the surface and therefore not accessible. This state is permanent, until another face of the monomer, which initially could not be covered by lignin becomes accessible. This can only be achieved, if some monomers in the neighborhood dissolve.

The results can be seen in section 4.3.

Although only two polymer length distributions were considered, any input can be managed by the model, if a distribution can be provided. For this, Poisson distributions are chosen, which are suitable to represent the provided distribution. Then, the distributions are superposed and thinned out in a uniform manner. This yields again a valid Poisson distribution, which is available for the model to simulate [68, p. 160].

Lignin

Another available feature is hiding of accessible surface by lignin. Lignin is a cross-linked polymer incorporated into the cell walls of most cellulosic substrates. It is important for the cell wall rigidity and leads to a more persistent structure. As a side effect lignin forms a physical barrier shielding the substrate from enzyme attack. Due to its abundance on Earth, exceeded only by the cellulose itself, it is seen as another important feature of substrates among crystallinity, see ref. [115]. A surface partially covered by lignin can be seen on figure 2.25.

The surface once considered covered by lignin can neither be used as a position for enzyme adsorption, nor for hydrolization. However, if the polymer chain becomes accessible to an enzyme other than from the covered surface, it still can be hydrolyzed.

The lignin distribution is modeled by generating connected paths on the surface of a microfibril in a random manner. The surface hidden by lignin is not accessible and the hydrolization bonds between affected glucose units can only be broken from another side of the polymer chain.

3 Model validation

The model described in chapter 2 now has to be validated. Given that some parameters in the model are not known, they must be fixed using experimental data. Experimental data come in many facets. Time series can be provided along with initial values of concentrations. Results of multiple experiments after a fixed time can be given, again, with a description of the experimental setup. Another possibility is to provide some known system parameters, which were measured in advance.

The data in form of enzyme assays has the form of time series of substrate, product (and possibly enzyme) concentration profiles as functions of time. They are provided as data sets over some time, where notable concentration changes occur. Often, only concentrations of products and substrate are measurable, as enzymes are present in a small amounts compared to the substrate for the reasons mentioned in chapter 1.

The developed model of this dissertation has the ability to use experimental results as optimization target. The optimization behavior of the system, needed data, restrictions, etc. are the topic of this chapter.

If model parameters are given, they are often provided as ratios of reaction rates in the biochemistry community, cf. figure 2.12 and ref. [59]:

- $K_D = \frac{k_2}{k_1}$ is the dissociation equilibrium constant of the enzyme-substrate complex
- $K_I = \frac{k_4}{k_5}$ is the dissociation equilibrium constant of the enzyme-product complex
- $K_C = \frac{k_3}{k_1}$ is the Van Slyke-Cullen constant
- $K_E = \frac{k_3}{K_M}$ is the specifity constant or kinetic efficiency
- $K_M = \frac{k_2+k_3}{k_1}$ is the Michaelis-Menten constant. Note, $K_M = K_D + K_C$

In principle, far more relations can be defined, e.g., one can think of relating

- k_1/k_5 as a measure of enzyme affinity to substrate related to its affinity to the product when the same amounts of both presence conditions of they both
- $\frac{k_2+k_3}{k_4}$ as the activity relation of the two inspected enzyme complexes.

Some of them are dependent on others and therefore redundant, like it is the case for K_D, K_C and K_M. For definitions of such relations there are no restrictions given. However, the possibilities to measure them are rather limited.

Which of these constants are useful and how needed parameters can be derived for the optimization of the current model will be answered further below. The amount of required data for this model to be successfully optimized has to be answered as well.

3.1 Model optimization problem

The goal of model training is to fix unknown kinetic parameters, by providing known, measured data. For this purpose, the ability to use data coming in form of known parameters as well as any time series in form of data sets has to be ensured.

Model training is mainly the process of systematic search for the global minimum of the model function with respect to given experimental data. The experimental data set for this model consists of time series of concentrations of soluble polymers as hydrolization result of an enzyme mixture and known measured enzyme parameters of any kind. During the search for feasible model parameters they are varied to find the minimization objective. The problem which has to be solved is

$$\min_{\mathbf{x} \in S} \|M(\mathbf{x}, t) - E(t)\| \tag{3.1}$$

with

- $\mathbf{x}$ representing a valid model parameter set within a set S, defined below.
- $M(\mathbf{x}, t)$: the time series generated by the model and
- $E(t)$: the time series provided by the experiment, as training data and
- $\|\cdot\|$: a distance measure defined on the objective space, discussed below.

The set S is a valid parameter subspace, defined by different kind of restrictions:

- bounds, $\mathbf{lb} \leq \mathbf{x} \leq \mathbf{ub}$, including the possibility to provide ∞ as one or both of the boundaries for each of the components of $\mathbf{x}$.
- linear equality $A_{eq} \cdot \mathbf{x} = \mathbf{b}_{eq}$ and
- linear inequality $A_{ineq} \cdot \mathbf{x} \leq \mathbf{b}_{ineq}$ constraints
- nonlinear equality and inequality constraints: $\mathbf{c}_{eq}(\mathbf{x}) = \mathbf{0}$ and $\mathbf{c}_{ineq}(\mathbf{x}) \leq \mathbf{0}$

For all model parameters such constraints can be formulated.

The distance measure on the objective space is defined over the finite set of available timings set T from experimental data, point-wise to be one or a sum of the common distance definitions, like the maximum distance

$$\|f\|_{max} = \max_{t \in T} |f(t)| \tag{3.2}$$

the l^1 distance

$$\|f\|_1 = \sum_{t \in T} |f(t)| \tag{3.3}$$

the l^2 distance

$$\|f\|_2 = \sum_{t \in T} [f(t)]^2 \tag{3.4}$$

and so on. This notion of distance serves the purpose of measuring the quality of the chosen phase space point relating the outcome of the model to the given experimental data. The smaller this distance measure goes, the better the chosen candidate from phase space is found to be.

As the amount of available measurements is finite, the distance measures are equivalent, see e.g, ref. [116]. However, the distance definition in use is dependent upon the particular requirements demanded by the application. Furthermore, different optimization algorithms behave differently depending on the chosen measure, see e.g. the more conservative library in ref. [117] vs. recent usages for machine learning algorithms reviewed in ref. [118].

In general, optimization algorithms can be divided in two types: deterministic algorithms, which proceed until a minimum is found. Convergence criteria are provided along these algorithms and after a finite time an optimum can be found. However, there is no guarantee, that this is the global optimum in the phase space. Other algorithms are formulated in a stochastic manner. For such algorithms no convergence guarantees can be given. These algorithms yield a result when their abort criteria formulated in a more abstract way are fulfilled. The result of such algorithms is in general a collection of best points found so far.

In any case criteria can be formulated in a very general manner, and consider parameter candidates found so far, time spent on the search and other search statistics provided by the chosen algorithm which become available along the way.

In other words, the minimization problem itself is specified by experimental data, defined by means of a measure on the objective space. Convergence criteria or time bounds guarantee it is achievable in finite time, by defining a tolerance and time bounded. The minimization problem is therefore well defined both in technical as well as economical terms, see ref. [119].

In the present work some of well established optimization algorithms were tried out; the presented results derive from utilizing the l^1 distance measure defined on the objective space. These algorithms systematically trying to cover regions of the parameter space either deterministically or randomly and comparing the result of the model function to the required result.

Tested algorithms include

- controlled random search algorithms introduced by Price, see ref. [120]
- genetic algorithm formulated by Conn in ref. [121]
- pattern search developed by Audet in [122]
- particle swarm optimization described by Parsopoulos in ref. [123]
- simulated annealing of Kirkpatrick in ref. [124]
- alopex based algorithm mentioned in ref. [125]
- bayesian optimization developed by Bull in ref. [126]

- surrogate optimization described in ref. [127]
- multi-object optimization algorithms defined in refs. [128, 129, 130]

If neither the derivative nor the function to minimize is known, or if the model does not even has a closed form, as in the current case, most of the algorithms above serve only for finding a global minimum heuristically in the parameter space, see ref. [131]. If there are multiple valid solutions, the algorithms provide isolated, local solutions, depending on the choice of the initial starting point guess. This is the case for example, if there are multiple local minima of comparable magnitude or if the objective function is flat. The manifold of valid solutions remains uncovered and any trial to recover its structure fails.

3.2 Optimization procedure

For the optimization realization a convenient sampling method for generating parameter samplings has to be defined. The sampling method has to comply with assumptions about the optimization values, does not prefer any parameter region and is chosen preferably regarding the performance of the optimization routine.

The best approach is to choose first a single sampling parameter, which is then used to uniformly generate the remaining optimization parameters. The choice of the sampling, or "scale" parameter is rather arbitrary. However, there are hints, that the definition regarding the l^1 distance measure for the optimization parameter is appropriate:

- Once a distance measure in the objective space is chosen, it is known that many optimization algorithms perform better, when the optimization function has a similar overall structure.
- The l^1 norm is known to perform well in technical applications, see e.g., [132].
- The proposed propagation algorithm of Gillespie, makes heavy use of time calculation with the matrix exponential calculation. This is a further example of the trace class operator, see ref. [133] used as a time propagator.

Therefore, on the available phase space the l^1 metric was imposed, also known as Taxicab metric, see [134]. This lead to a diamond shaped norm perception of the phase space, see appendix B.3.

With such coordinate transformation, the phase space can be uniformly sampled via the sample parameter K_S, defined as

$$K_S = \sum_i k_i \tag{3.5}$$

Indeed, this is the radius in the Taxicab metric, forming a single diamond sphere around the coordinate center. With such sampling all rate constants can be transformed to be elements from the unit interval $(0, 1)$, multiplied by their common sum of sampled

K_S. The transformed rate constants are orthogonal to the chosen parameter K_S by definition of orthogonality of the coordinate system.

During optimization an inner optimizer is driven by the constraint of a constant K_S. Each solution lies therefore on a uniquely defined surface of the phase space $\mathbb{R}^n_+$, where n is the amount of k_i, dependent from the system under inspection. The outer sampling of K_S can either be driven by another, independent optimization algorithm or performed manually, in case of a brute force scanning of the phase space.

The inversion of the sampling process from "sample a k_i tuple, leading to a specific K_S" to "sample a K_S, leading to a specific k_i tuple" is well defined, as the probability of finding two different optimum sets with a common K_S value is zero almost surely, see ref. [135].

Performing samplings in this way implies that exactly one solution exists for each choice of K_S. The solution lies on the according diamond surface of dimension $n-1$, when there are n optimization parameters available. In case of optimization of reaction rate constants, the solution is defined to be the best found k_i tuple on this surface.

The common approach for global optimization is to use either optimization algorithms with an initial population of start points covering available phase space or to restart the optimization several times with different start points. This is done to ensure that the optimization results are independent from the choice of the starting point, which is typical for local optima.

A global optimization algorithm stops when the stopping criteria, formulated in terms of optimization parameters as well as minimization goal and further algorithm dependent specifics, are fulfilled. Algorithms guaranteed to find a minimum provide a convergence stopping criteria, however this is in general not the case.

3.3 Optimization results

If not otherwise stated, the following results shown in this chapter are obtained with artificially generated data, by using a simplified deterministic model, shown in figures 2.11 and 2.12. However, keeping in mind that the main model, introduced in chapter 2, is of stochastic nature, appropriate convergence criteria are formulated and emphasized at the appropriate places.

For a stochastic model, like in the current case, different outcomes are possible for the same input. If the optimization algorithm stopping criteria include any too narrowly formulated convergence criteria, the optimization algorithm will not converge. If they are missing, omitted or formulated too loosely, any found minimum is guaranteed to be global.

In other words, the global nature of optimization can be handled by choosing several starting points in the phase space, applying a combination of optimization algorithms and heuristics to systematically search the available phase space. The stochastic nature of the model and model input however prevent a formulation of any strict convergence criteria.

Under the mentioned conditions, the optimization was run on the model, depicted in figure 2.11. The results, see figure 3.1(a), show many solutions for k_{cat} and K_M. Which are located around their expected values.

At this place it is important to recall, that the knowledge of measurable constants is not enough for a model containing an independent time propagating algorithm, as already stated in section 2.2. For such a model, all rates of modeled events are needed.

To get every single rate constant, they were optimized independently, especially the rate constants for adsorption and desorption k_1 and k_2. The results, shown in figure 3.1(b), are placed along the line, defined by the Michaelis Menten constant K_M.

The problem, that many points are found, arises here for the first time. The reason does not lie in the microscopic nature of the problem. All the solutions are the consequence of the finite precision constraint on the optimizer algorithm. This is however unavoidable. By chance the dynamic defining constants are narrowly located and if the optimizer stops at any of it, similar results are achieved.

The found solutions are also consistent with the fit model

$$F(t) = \frac{S_T}{K_M} \exp\left(\frac{S_T}{K_M} - \frac{v_{max}}{K_M} t\right) \tag{3.6}$$

found in Goudar et. al, [136, 137]. However, while the fit model yields the same results for the time courses generated by means of the according parameter, the current model yields three individual reaction rate constants, and the fit model of Goudar yields only two.

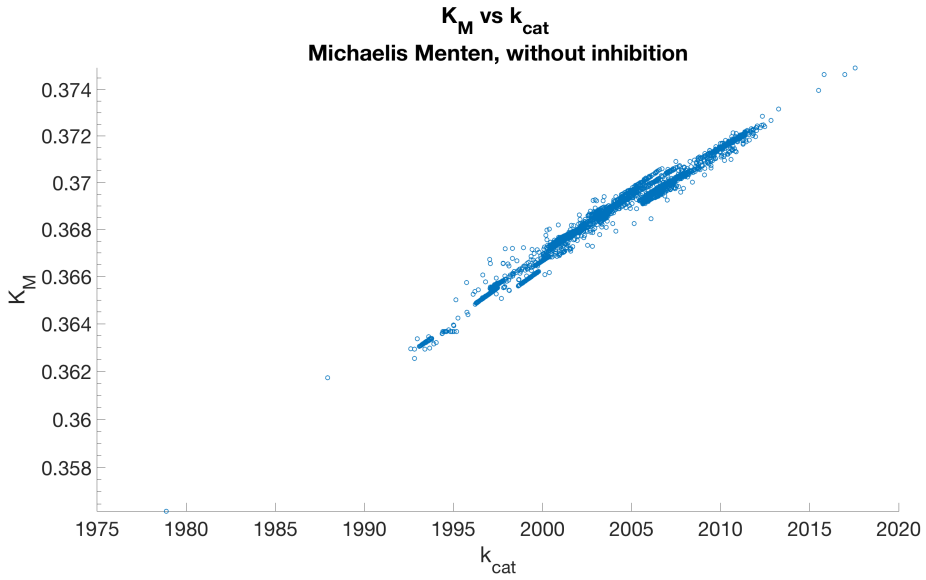
The state of affairs changes dramatically, if the system becomes more complex, containing inhibition. The optimization results for a system containing competitive inhibition are shown in figure 3.2. Here k_{cat}, K_M and most notably K_I are found in a large parameter space, covering in general many orders of magnitude.

Choosing a single optimum becomes more severe problem in this case. Depending on the optimization routine setup, very different macroscopic parameters can be found.

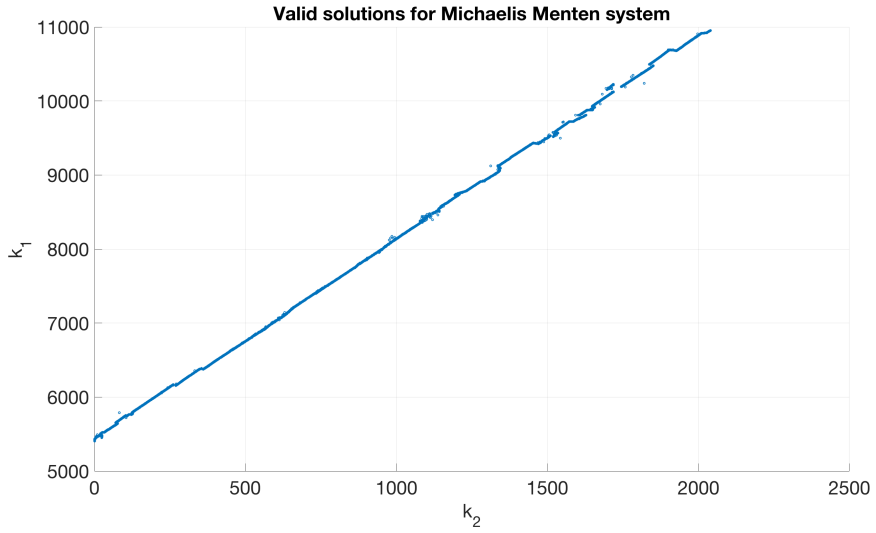
Any condition can be defined to reduce the solution set to a single value. As it was mentioned before, after defining a cutoff limit for optimization precision, every attempt to do so fails. As long as the found parameter reproduces the given time series in an acceptable manner, all of them are equivalent from this point of view. Statistical considerations like cluster analysis w.r.t. the found values and different averaging technics were tried out, however they did not achieve further narrowing of found results.

Such results are shown in figure 3.3. Even with the optimization precision set to the numerical limit of the computing machine, many solutions are still found.

As the fact of multiple solutions was discovered for the system with inhibition, the question has to be answered if the optimization itself is consistent. It turns out, that the number of solutions decreases smoothly as the precision is lowered. In figure 3.4 it is shown on logarithmic axes, that the number of optimal points is getting smaller in a well formed manner. For any fixed, finite precision there are however, still a multiplicity of possible parameter candidates left.

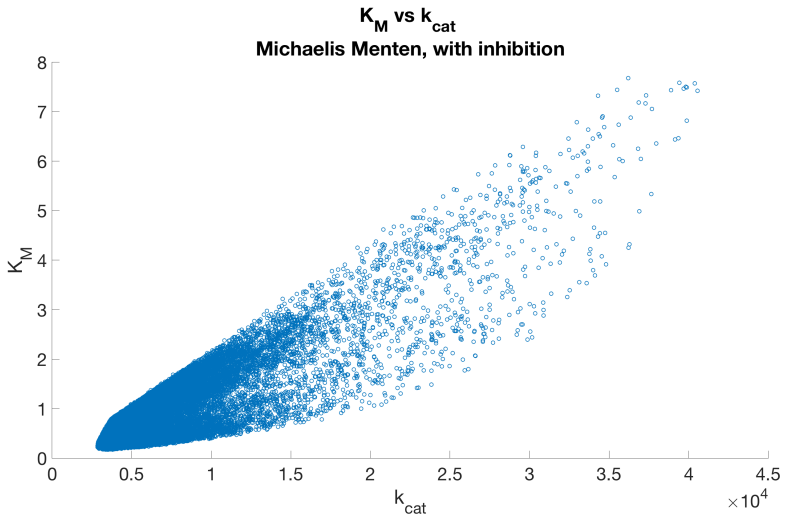


(a) Optimization of the simple model without inhibition also leads to ambiguous solutions, however, the solutions are located in a small region of the phase space. As the optimization was done with artificiall generated data, the expected values are known: $K_M = 0.3667$, $k_{cat} = 2000$ in renormalized units.

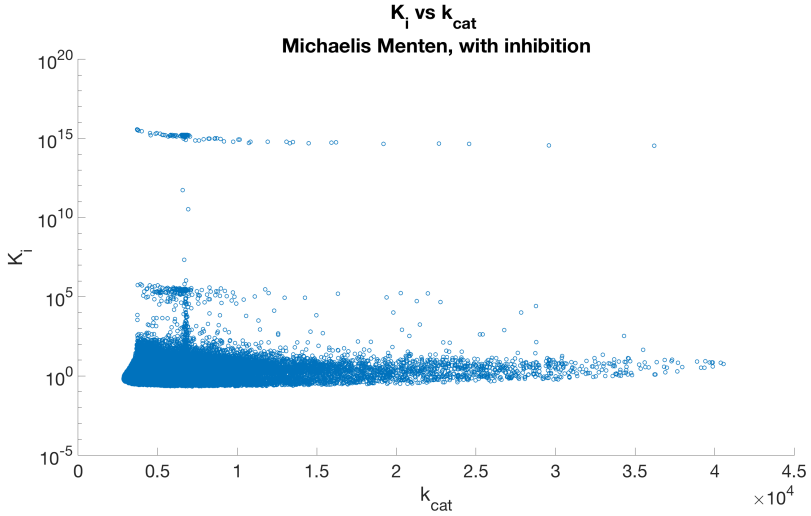


(b) In the k_1 k_2 projection the solutions are located on a line. Roughly speaking, the line fulfills the evident relation $k_1 = K_M k_2 + k_{cat}$, with a k_{cat} as depicted in figure (a), assumed to be constant. The both shown rates are not measurable, the exact solution remains uncovered. As the optimization was done with artificial generated data, the expected values are known: $k_1 = 6000$, $k_2 = 200$ in renormalized units.

Figure 3.1: Insufficient optimization input, and a finite precision constraint on the optimizer algorithm lead to many possible solutions. This is the case, even for the Michaelis Menten system without inhibition. While the values for K_M and k_{cat} are well localized, all rate constants are needed for quantitative simulation, see ref. [48]. These remain undetermined in general.



(a) Optimization results for the task (3.1) with measured concentrations of product and substrate over time. A finite acceptable precision was defined. Many feasible results emerge in the phase space projected onto k_{cat} and K_M axes. While values for K_M are between 0 and 10, feasible values for k_{cat} vary on the scale of 5000 and over 40000. As the optimization was done with artificial data, the expected values are known: ${k_{cat}}^{\text{exp}} = 6000$, ${K_M}^{\text{exp}} = 0.3666$ in renormalized units.



(b) Optimization results for the task (3.1) with measured concentrations of product and substrate over time. A finite acceptable precision was defined. Many feasible results emerge in the phase space projected onto k_{cat} and K_I axes. Values for K_I vary on the large scale of orders $10^0 - 10^{15}$. Values for k_{cat} vary on the scale of 5000 and over 40000. As the optimization was done with artificial data, the expected values are known: ${k_{cat}}^{\text{exp}} = 6000$, ${K_I}^{\text{exp}} = 0.75$ in renormalized units.

Figure 3.2: Optimization of the model providing time series as optimization input yields ambiguities on large scales for the quantities k_{cat}, K_M and K_I. The similarity to the given time series were used as stopping criteria. Therefore, the remaining, shown optima cannot be distinguished any further.

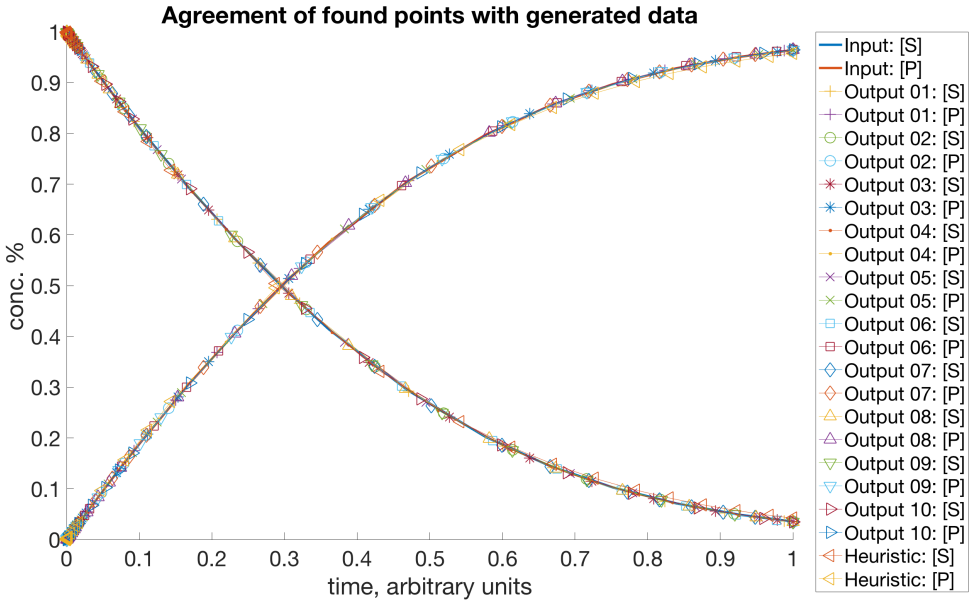


Figure 3.3: Optimization of current stochastic model leads to many results, all in a good agreement with the data, used as the optimization target. For all outputs the distance measure, defined as above, calculated with 10 equally distributed points is below 0.03. As the input data was generated artificially, all values can be compared in this case. The table below shows the input as well as a sample from the output values, in a unitless form. The last line shows a heuristic result found by applying another optimization approach as a try to identify a unique phase space point across all feasible points. As all other feasible parameter tuples, this value reproduces experimental results with good agreement. However, its values do not coincide with the values of the optimization input, nor do the resulting values of K_M or K_I.

	k_1	k_2	k_3	k_4	k_5	K_M	K_I
Input	12000	400	4000	6000	8000	0.36	0.75
Output 01:	12886.7	1946.6	13464.5	2893.2	4060.9	1.2	0.7
Output 02:	13811.4	1175.9	5988.9	3931.4	5838.4	0.5	0.7
Output 03:	19227.5	4790.7	3711.8	8350.4	6628.6	0.4	1.3
Output 04:	19998.9	1354.7	3198.0	8021.4	16088.9	0.2	0.5
Output 05:	18913.4	6151.4	4986.2	5285.6	4335.5	0.6	1.2
Output 06:	13587.3	401.3	2880.3	12923.6	20087.5	0.2	0.6
Output 07:	11701.3	0.0	5817.2	3991.0	6030.5	0.5	0.7
Output 08:	8438.1	162.2	36285.7	2762.8	2197.3	4.3	1.3
Output 09:	8874.6	410.0	3194.2	14677.0	9997.2	0.4	1.5
Heuristic:	10309.5	3400.6	6066.8	5793.1	3454.4	0.9	1.7

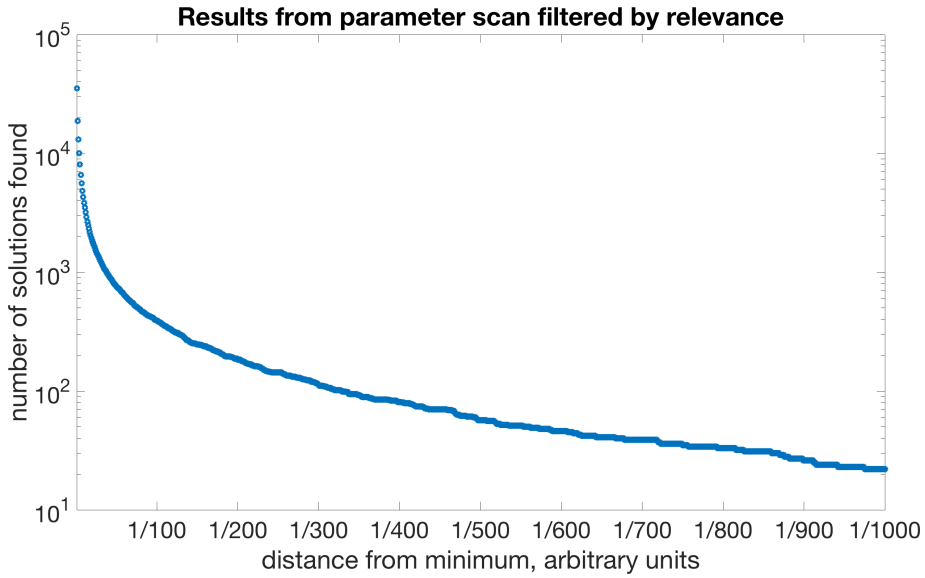


Figure 3.4: Optimization with experimental data only leads to many solutions, shown in figures 3.2 and 3.1. Their amount decreases by confining the optimization criteria smoothly, see figure 3.4. However a certain amount remains for a given optimum criteria.

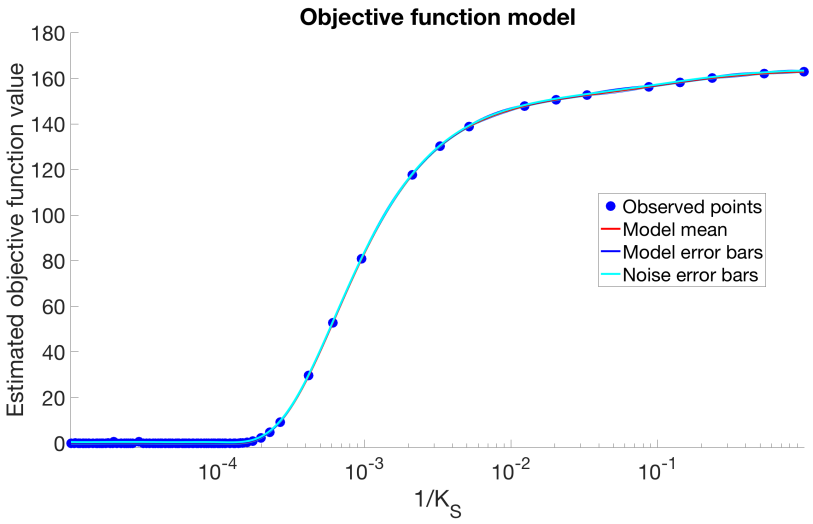
Evaluation of the parameter space

The presence of many equivalent optima, and therefore the underdetermined state of the system confirms the choice of l^1 distance measure for sampling generation, see section 3.2. This case is well known in theory, see e.g. ref. [138, 139, 140, 141], as well as in various applications, see e.g., refs. [142, 143, 144, 145].

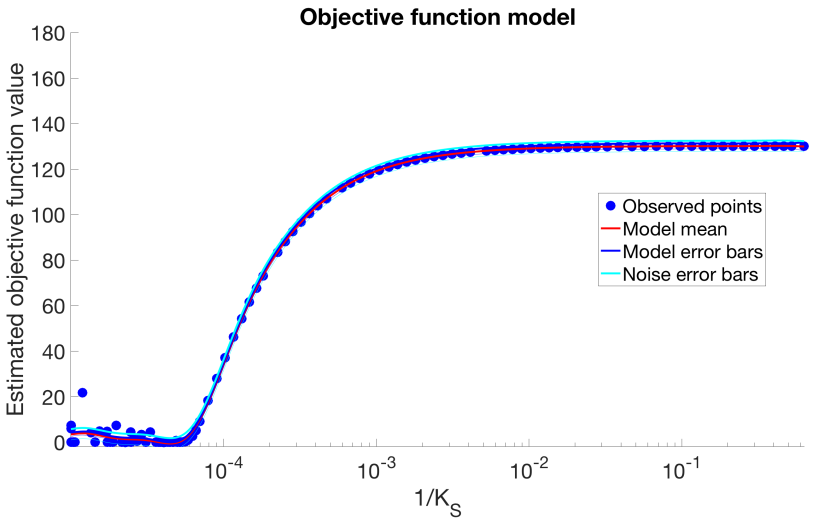
K_S is also a projection of k_i in terms of a radial basis function, see [146]. From this point of view it is now possible to inspect the achieved results along this projection. The bayesian and surrogate optimization algorithms, mentioned above are specialized to explore the phase space with as few function evaluations as possible. The results provided from bayesian optimization of systems (shown in figures 2.11 and 2.12) are shown in figures 3.5(a) and 3.5(b) respectively.

Therein one can see a separation in the k_i phase space. A region exists, where no k_i could be found to reproduce experimental measurements. However, there is another phase space region, where many k_i exist, which fulfill the requirements.

The results of the present work, shown in figures 3.1 and 3.2 lead to the conclusion that any valid model can be formulated based only on experimental enzyme assays. Any attempt to establish such a model, leads to not well-defined solutions. Therefore concluding on dynamic enzyme characteristics based only on their time assays is an incomplete approach, which is demonstrated here for the first time.



(a) The optimization without any additional parameters shows that a wide region of parameters is able to fit the provided data. Similar results are obtained if K_M is provided along the experimental data.



(b) A similar picture appears in case of optimization with the enhanced model with inhibition. There is an area, which can be excluded, as there isn't any parameter which can reproduce experimental data. Then phase space region exists, where many parameters are located, all of which can reproduce the optimization goal with a good approximation.

Figure 3.5: The pictures (a) and (b) show phase space estimations for numerical models with and without inhibition. Only experimental data is provided, in the first case even enzyme time series where provided to the optimization routine. A feasible region in the phase space can be identified, however no particular global minimum can be identified.

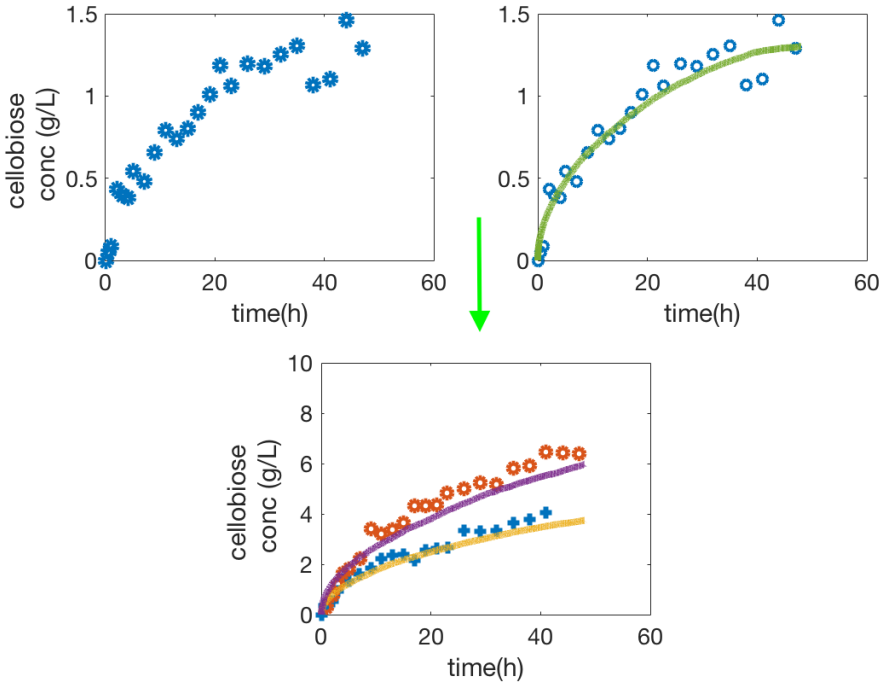


Figure 3.6: Optimizing result for CBHI. Left: Avicel (25g/L), various loadings of CBH I. Betzerra RMF, Dias AA, 2004; right: 4.1mg/g enzyme with model results after optimizing below: Avicel (25g/L), various loadings of CBH I. Bezerra RMF, Dias AA 2004 with results of model prediction. Correlation coefficient is $R = 0.9923$, showing a good prediction ability of the model w.r.t. experimental data.

Reusability and validity

In the last part of this section, the power of the model prediction after the optimization will be shown. After optimization utilizing some curves from the experiment, unknown model parameters are fixed. Here, either an arbitrary value is chosen from the available optimization minima or it is chosen by a heuristic, see e.g., ref. [147] and the caption of figure 3.3. In this state, the model is able to make prediction in parameter regions and time series not used during optimization.

In figure 3.6 experimental as well as modeled time curves are shown. One of the experimental curves is used for model training. Two others are used to check the prediction abilities of the model. Taking the both curves for the check, the resulting correlation coefficient is $R = 0.9923$, showing a striking match of experimental and modeled data.

In order to validate the EG model, there also was an optimization driven with the resulting correlation coefficient of $R = 0.91816$. However, simultaneous optimization of all product profiles was not achieved due the difficult behavior of EG type enzymes already known from literature, see refs. [148] and [26].

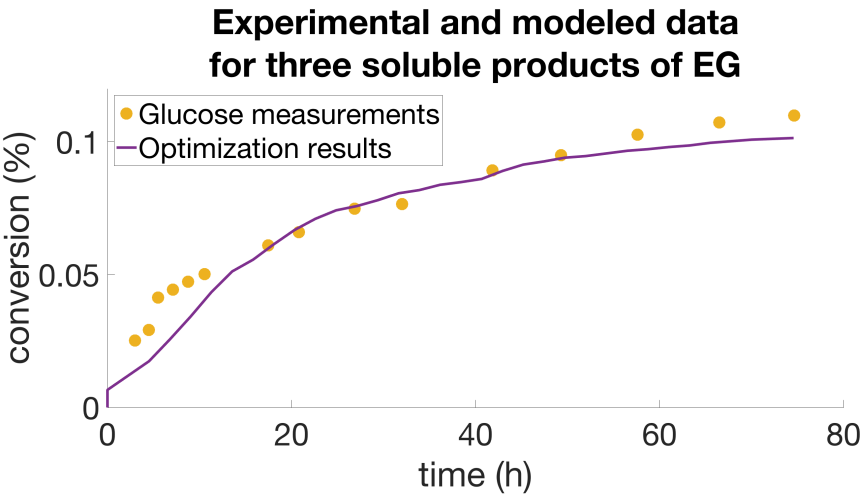


Figure 3.7: Optimization results for Glucose time series. Experimental data was taken from Medve et. al., see ref. [149]. While it was possible to find optimal parameters for the time series of glucose, the optimization of other sugar length failed. It is assumed therefore, that some action mode details remain uncovered in the model.

3.4 Identifying optimal Lagrange parameter

One of the consequences of the previous section is, that the minimization problem defined in (3.1) is usually underdetermined for this model. This is unexpected, as it is a common practice to use experimental data to solve the inverse problem of gaining model parameters.

Providing further independent time data is possible, but the mass conservation laws, present as algebraic equations in (A.7) and (A.7), remove one of the measurable time series from the independent set and thus from consideration. What remains are one or two respectively, enzyme concentration time series. These are however not measurable in most cases, due the small concentrations of total enzyme E_T.

Another possibility is to provide other measured data in form of constraints to the optimization algorithm. The optimization problem would then be formulated in terms of Lagrange multipliers method, see ref. [150] and becomes

$$\min_{\mathbf{x}\in S} \|M(\mathbf{x},t) - E(t) - G(\mathbf{x})\| \tag{3.7}$$

The additional parameter $G(\mathbf{x})$, also known as regularization term see ref. [151] or hyperparameter, see [152] has to be provided by the user. This is done based on existent knowledge of kinetic parameters, experience or other considerations, which are not topic of this work. For the rest of this chapter, the attention is drawn on the amount of hyperparameters needed and their definition.

Independent of the chosen Langrange function a parameter has to be chosen, to use for independent parameter samplings. This is seen as the complement to the regularization mentioned above, known as generalization, see ref. [153]. This task was already accomplished. K_S serves as generalization parameter for the present model.

Until a Lagrange parameter is defined, the generalization parameter serves as source of parameter samplings over the defined parameter space, good enough to estimate the behavior of the minimization problem.

For finding the Lagrange parameter (and generally, there are multiple of them needed) a common approach to analyze a complex system is chosen by converting it to an appropriate normal form, see ref. [154]. Then, information about the complex system can be retrieved by analyzing linearizations around specific points in the state space.

First, this is done on the simple case of Michelis Menten system without inhibition. This case serves as comparison for the actual system under inspection and as a proof of concept for the procedure. Then, the analysis is repeated for the case including inhibition.

Michaelis Menten kinetics without inhibition

After imposing the algebraic equations on the differential equations, cf. (A.7) one is left with only two linear independent differential equations for concentration quantities.

$$\begin{aligned} \partial_t C_s &= k_1[(E_T - C_s)(S_T - P - C_s) - K_M C_s] \\ \partial_t P &= k_1 K_C C_s \end{aligned} \tag{3.8}$$

where C_s is the concentration of the enzyme-substrate complex, P is the concentration of the product, $K_M = \frac{k_2+k_3}{k_1}$ is the Michaelis-Menten constant, $K_C = \frac{k_3}{k_1}$ the van Slyke-Cullen constant, E_T and S_T are the initial concentrations of enzyme and substrate in the system and k_i are the constants of enzyme reaction rates.

There are some other approaches of analyzing this system to consider, before the normal form analysis is done.

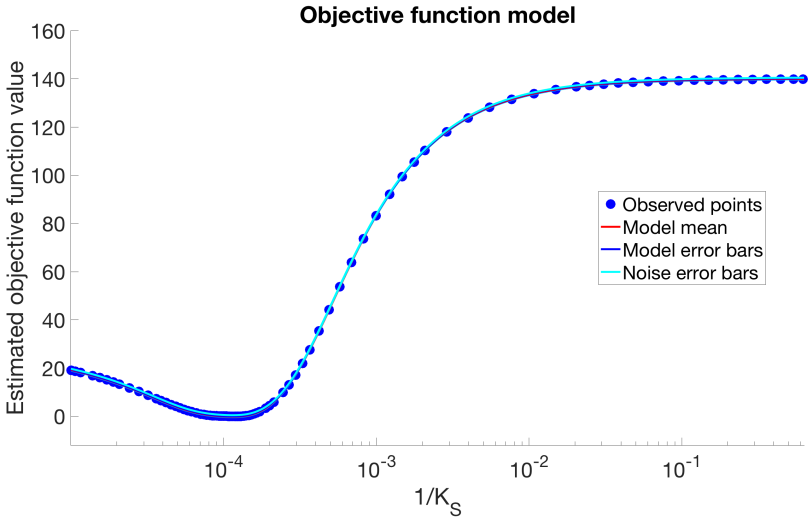
First, note the graphical approach. Analysing the initial rate of reaction

$$v = \frac{v_{max} S}{K_M + S} \tag{3.9}$$

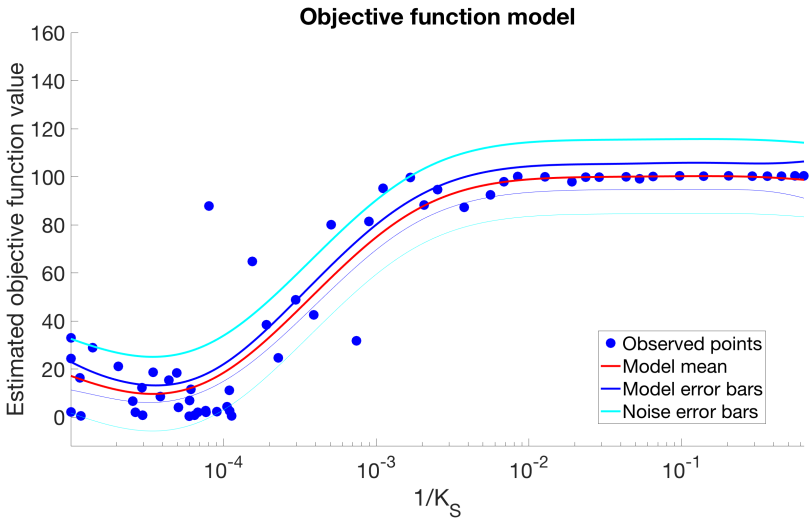
where v_{max} is the maximal conversion rate of the enzyme,one comes to the conclusion, that (3.9) is of the form of a hyperbola. By providing K_M and v_{max} one defines the hyperbola by defining its asymptotes in the parameter space. Another way of uniquely defining the hyperbola is by providing its eccentricity as a single parameter.

Another approach comes from the graph theory, see e.g. ref. [155]. Roughly speaking, therein one converts the reactions into a directed graph, which is then subjected to graph theory tools for examining for the essential dynamical properties. While this approach is the fastest for qualitative statements about the system, at the end, there are still differential equations to analyze for more precise predictions.

A different approach delivers perturbation theory. For the case of Michaelis Menten kinetics without inhibition, this approach can be found in [59]. While the system is



(a) By providing the optimal Lagrange parameter numerical errors are small, and the global minimum estimation is efficient.



(b) For the model with inhibition, there is more effort to estimate a proper minimum. However, by providing a proper Lagrange function this can still be solved, and the model for the hyperparameter can be found.

Figure 3.8: Optimization results with the optimal Lagrange parameter provided. The hyperparameter was found by linear stability analysis. In the simple case, without inhibition, (a) the model optimized values follow the model estimation almost exactly. In the case with inhibition the optimization is more noisy, however, a minimum can be identified, even with the single hyperparameter.

analyzed in detail therein, at this place only the first outcome is needed. Dingee et al. derive the ratio of two time scales present in the system to be

$$\epsilon = \frac{K_C E_T}{(E_T + S_T + K_M)^2} \tag{3.10}$$

Letting the initial parameters aside and neglecting the power in the denominator in (3.10) one arrives to the parabola eccentricity of the graphical approach. This value also coincide with the enzyme efficiency K_E.

While being a powerful tool for analyzing complex systems this approach fails for more complex cases. The reason for this is not the approach itself. It is rather the necessity to choose the number and form of the linearization parameters, i. e. an Ansatz at the beginning. However, for more complex cases the choice of an Ansatz is not apparent.

To the help comes another powerful tool of linear stability analysis, see e.g., ref. [156]. Formulating the normal form of the differential equations reveals bifurcations in the system, however in this work it is used to identify the relevant time scales of the system. Note, the system is two dimensional in the current case.

By formulating the Jacobian J defined as

$$J = \begin{pmatrix} \partial_{x_1} f_1 & \partial_{x_2} f_1 & \cdots & \partial_{x_n} f_1 \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{x_1} f_m & \partial_{x_2} f_m & \cdots & \partial_{x_n} f_m \end{pmatrix} \tag{3.11}$$

of the dynamic system at a fixpoint in the phase space of the reactants, and by searching for its eigenvalues one obtains

$$J|_{x_0} = k_1 \begin{pmatrix} -(E_T + S_T + K_M) & -E_T \\ K_C & 0 \end{pmatrix} \tag{3.12}$$

if x_0 is chosen to be

$$x_0 = (0, 0)^T \tag{3.13}$$

and

$$J|_{x_0} = k_1 \begin{pmatrix} -(E_T + K_M) & -E_T \\ K_C & 0 \end{pmatrix} \tag{3.14}$$

if x_0 is chosen to be

$$x_0 = (0, S_T)^T \tag{3.15}$$

one recognizes, that this is a common initial state of the inspected model, naturally being an unstable fixpoint of the system.

Solving for eigenvalues $\det(J|_{x_0} - \lambda \mathbb{I})$ reveals the discriminant Δ of the quadratic equation to be

$$\Delta|_{(0,0)^T} = 1 - \frac{4 \cdot K_C \cdot E_T}{(E_T + S_T + K_M)^2} \tag{3.16}$$

and

$$\Delta|_{(0,S_T)^T} = 1 - \frac{4 \cdot K_C \cdot E_T}{(E_T + K_M)^2} \tag{3.17}$$

which are up to constant terms identical to the time scales of Dingee et al., identified in ref. [59].

Michaelis Menten kinetics including inhibition

An analogous problem evolves in the more realistic case of kinetics extended by inhibition, cf. (A.8). Here, after incorporating algebraic equations into differential ones three linear independent equations are left.

$$\begin{aligned}\partial_t C_s &= k_1\{[(E_T-(C_s+C_p)][(S_T-P-(C_s+C_p)]-K_M C_s]\}\\ \partial_t C_p &= k_5\{[E_T-(C_s+C_p)]P+\widetilde{K}_C C_s - K_I C_p\}\\ \partial_t P &= k_5\{K_I C_p-[E_T-(C_s+C_p)]P\}\end{aligned} \tag{3.18}$$

where C_s is the concentration of the enzyme-substrate comples, C_p is the concentration of the enzyme-product comples, P is the concentration of the product, $K_M = \frac{k_2+k_3}{k_1}$ is the Michaelis-Menten constant, let $\widetilde{K}_C = \frac{k_3}{k_5}$, $K_I = \frac{k_4}{k_5}$ is the inhibition constant, E_T and S_T are the initial concentrations of enzyme and substrate in the system and k_i are the constants of enzyme reaction rates.

Compared to the Michaelis Menten case without inhibition, this case is three dimensional, containing five parameters. The characteristic polynomial will be of third degree and generates more different cases for inspection.

The velocity for product generation can't be formulated without assumptions about substrate or product, see ref. [157], not to mention any formal similarity to a hyperbola. This is due to the presence of a back reaction with the product. The hyperbolic association in the phase space is lost.

The reaction graph gives rise to the assumption, that three time scales are involved now: one for the substrate complexation, one for product complexation and one for the slow reaction of substrate conversion. However, a trial to derive them by means singular perturbation failed. What is left is the treatment of the system again by linear stability analysis applied to the corresponding normal form (3.18).

The according Jacobian has now the form

$$J|_{x_0} = \begin{pmatrix} -k_1\cdot(E_T+S_T+K_M-2C_p-P) & -k_1\cdot(E_T+S_T-2C_p-P) & -k_1\cdot(E_T-C_p)\\ -k_5\cdot(P-\widetilde{K}_C) & -k_5\cdot(P+K_I) & -k_5\cdot(C_p-E_T)\\ k_5\cdot P & k_5\cdot(P+K_I) & k_5\cdot(C_p-E_T)\end{pmatrix} \tag{3.19}$$

evaluated at $x_0 = (0,0,0)^T$ for the initial concentrations of (C_s, C_p, P) and $x_0 = (0, S_T - P^*, P^*)$ for the final concentrations of (C_s, C_p, P), with the positive branch solution for P:

$$P^* = (S_T - E_T - K_I)\left(\frac{1}{2}+\sqrt{1+\frac{4K_I S_T}{(S_T-E_T-K_I)^2}}\right) \tag{3.20}$$

The system is now three dimensional and the time scales observed from the normal form are defined by the solution of the charachteristic polynomial, which is now cubic. One more time scale is expected from graphical analysis.

The two time scales in the Michaelis Menten model without inhibition originated from the adsorbing/desorbing process w.r.t. the substrate and the hydrolization process. Now, a third one emerges additionally from the adsorbing/desorbing process w.r.t. the products.

The solution is to a certain extent more complex in comparison to the two dimensional case, however still tangible. It is also given in terms of a discriminant, however, now the discriminant has two obligatory terms which cannot be reduced to a single one without some effort.

The general form of the discriminant is defined by inspecting e.g., the Cardano's method, redefining the discriminant to

$$\Delta = \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3 \tag{3.21}$$

with

$$p = b - \frac{a^2}{3} \tag{3.22}$$

$$q = \frac{2a^3}{27} - \frac{ab}{3} + c \tag{3.23}$$

and

$$a_{(0,0,0)^T} = k_1(S_T + E_T + K_M) + k_5(E_T + K_I) \tag{3.24}$$

$$b_{(0,0,0)^T} = k_1 k_5[{E_T}^2 + K_I(S_T + E_T + K_M) + E_T(S_T + K_M + \widetilde{K}_C) + \tag{3.25}$$
$$+ \widetilde{K}_C S_T]$$

$$c_{(0,0,0)^T} = -k_1 k_5^2 \widetilde{K}_C E_T E_T + K_I + S_T \tag{3.26}$$

$$a_{(0,S_T-P^*,P^*)^T} = \frac{1}{2}(-E_T k_1 - 2k_5\sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2} + \tag{3.27}$$
$$+ k_1(K_I - 2K_M + S_T - \sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2}))$$

$$b_{(0,S_T-P^*,P^*)^T} = \frac{1}{2}k_1 k_5({E_T}^2 + {K_I}^2 + \tag{3.28}$$
$$+ 2K_M\sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2} -$$
$$- K_I(\widetilde{K}_C - 2S_T + \sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2}) +$$
$$+ E_T(2K_I + \widetilde{K}_C - 2S_T + \sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2}) -$$
$$- (-\widetilde{K}_C + S_T)(-S_T + \sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2}))$$

$$c_{(0,S_T-P^*,P^*)^T} = -\frac{1}{2}k_1 k_5^2 \widetilde{K}_C({E_T}^2 + E_T(2K_I - 2S_T + \tag{3.29}$$
$$+ \sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2}) -$$
$$- (K_I + S_T)(-K_I - S_T + \sqrt{(E_T + K_I)^2 + 2(-E_T + K_I)S_T + {S_T}^2}))$$

Now, there are four cases in general for the discriminant to evaluate, in the case of a cubic equation:

$\Delta = 0$: due $q = p = 0$. This is the case, when the solution of the cubic equation for λ has only a triple root. This case is the reason, why q cannot be factorized any further from the discriminant.

$\Delta = 0$: due $(q/2)^2 + (p/3)^3 = 0$, $q, p \neq 0$. This is the case, when the solution of the cubic equation for λ has two solutions, where one of it is of double kind.

$\Delta > 0$: This is the case, when the solution of the cubic equation for λ has one real and two complex solutions.

$\Delta < 0$: This is the case, when the solution of the cubic equation for λ has three distinct real solutions.

Exploring the first case more deeply, taking the parameters at initial state of the system $(0, 0, 0)$ one finds, that if $q = p = 0$, then the solution for the single available eigenvalue is $\lambda = -a/3$, while $a > 0$ for all possible model parameters. This implies, that for a system with parameters chosen such that $q = p = 0$ the eigenvalue belongs to a stable solution. The only valid solutions for k_1 and k_5 are obtained to be zero. This case is excluded during optimization, and therefore, the case of $p = q = 0$ cannot emerge. All other cases of dynamics are expressible by the single ratio of p/q, taken at one of the fix points.

Thus, the result of the stability analysis, is that a parameter exists, which is enough to provide a valid constraint, so that the dynamics of the system is uniquely described.

If the case $p = q = 0$ is excluded from further considerations, the discriminant governing the dynamics of the system can be further simplified by factorizing $q/2$ and one finds

$$\Delta = 1 - \frac{(p/3)^3}{(q/2)^2} \tag{3.30}$$

as the analogous time scale parameter for the case of Michaelis Menten kinetics with inhibition.

The experimental data reduced the dimension of the solution space already in the Michaelis Menten case without inhibition. This part is fully contained in the model of enzyme dynamics including inhibition. By removing the case of $p = q = 0$ the five dimensional solution space is reduced even more. This is done, however, in a highly nonlinear manner. For the discriminant all values are still possible, only the case of $\Delta = 0$ is not degenerated any more. A further implication is now, that even for the more complicated case of Michaelis Menten kinetics extended by inhibition still only a single parameter is needed along with experimental data to determine its behavior.

The other thing is, that the parameter

$$\epsilon = \frac{(p/3)^3}{(q/2)^2} \tag{3.31}$$

is not measurable in general, as opposed to the simple case without inhibition. That leaves only one with the possibility of providing more, measurable parameters. Encouraging results were achieved, when enzyme efficiency $K_E = k_3/K_M$ from the simplified model was complemented by $K_I = k_4/k_5$.

This requires two constraint parameters, however, a minimum is found without any problems, see fig. 3.9. The fact, that constraints can be formulated in many ways goes hand in hand with the fact, that the Lagrange parameter is also generally not unique. The method described in this chapter therefore defines a possible Lagrange parameter, however not necessarily a unique one.

Also this result won't be valid for a general nonlinear system, as the linearization does not yield meaningful results in every possible case. However, the numerical trials for this model did so, for the non-stochastic modelling approach with partial differential equations (A.7) and (A.8).

3.5 Sensitivity analysis of the model

The sensitivity analysis is a possible approach to confirm, that a minimum was found. However, without any additional constraints, no pronounced minimum is found, as seen in section 3.3. Due to this reason this analysis is postponed until now. After the section 3.4 a way is known, how to locate a minimum and a sensitivity analysis can be done.

Local and global sensitivity analysis can be carried out. The goal of the local sensitivity analysis is to show, how well a minimum is expressed under small perturbations of found parameters.

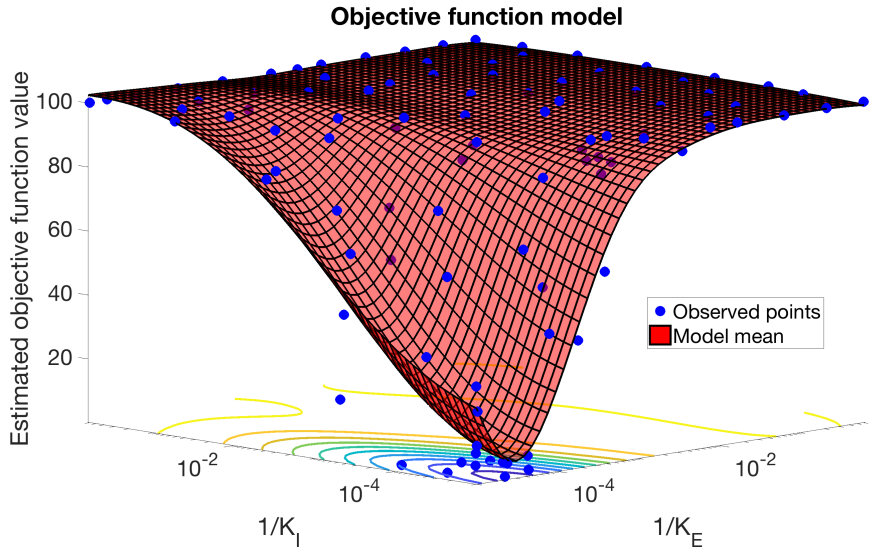


Figure 3.9: Optimization results with two constraints provided. Without being able to measure the single calculated time scale parameter, one is left with two hyperparameter, which can be passed to the optimization routine. This is of course also possible, resulting in a pronounced minimum in the parameter space. The plot shows a pronounced minimum in the phase space of $K_M \times K_E$

The term of global sensitivity analysis is not well defined in general. The sensitivity of the model on its parameters is defined locally, w.r.t. some distinguished parameter space point. Especially in the present case, where the overall system behavior has a pronounced dependency not only on kinetic parameters but also on initial parameter values, like initial concentrations of reactants.

However, following Leamer in [158, 159], global sensitivity can be seen as the investigation of model behavior under various assumptions variation, which enters the data generation.

Local sensitivity analysis of the model

A known method to analyze sensitivity of a found parameter set is to apply the common sensitivity analysis, described e.g., in ref. [160]. Here, the analysis is driven locally, by varying single parameter at once. However, the varied parameter can, but does not necessarily need to be one of the optimized model parameters, in general. Also, one of the provided constraints can be varied. The solution can be inspected w.r.t. stability of arbitrarily defined parameters in the aforementioned manner. Some correlation results for the found solutions are shown in figures 3.10 and 3.11.

Global sensitivity analysis of the model

For optimization very few assumptions were made explicitly. This was done deliberatively, to keep the optimization as general as possible. Therefore in the global analysis it will be shown, how the choice of distance measure affects the expressiveness of minimum found.

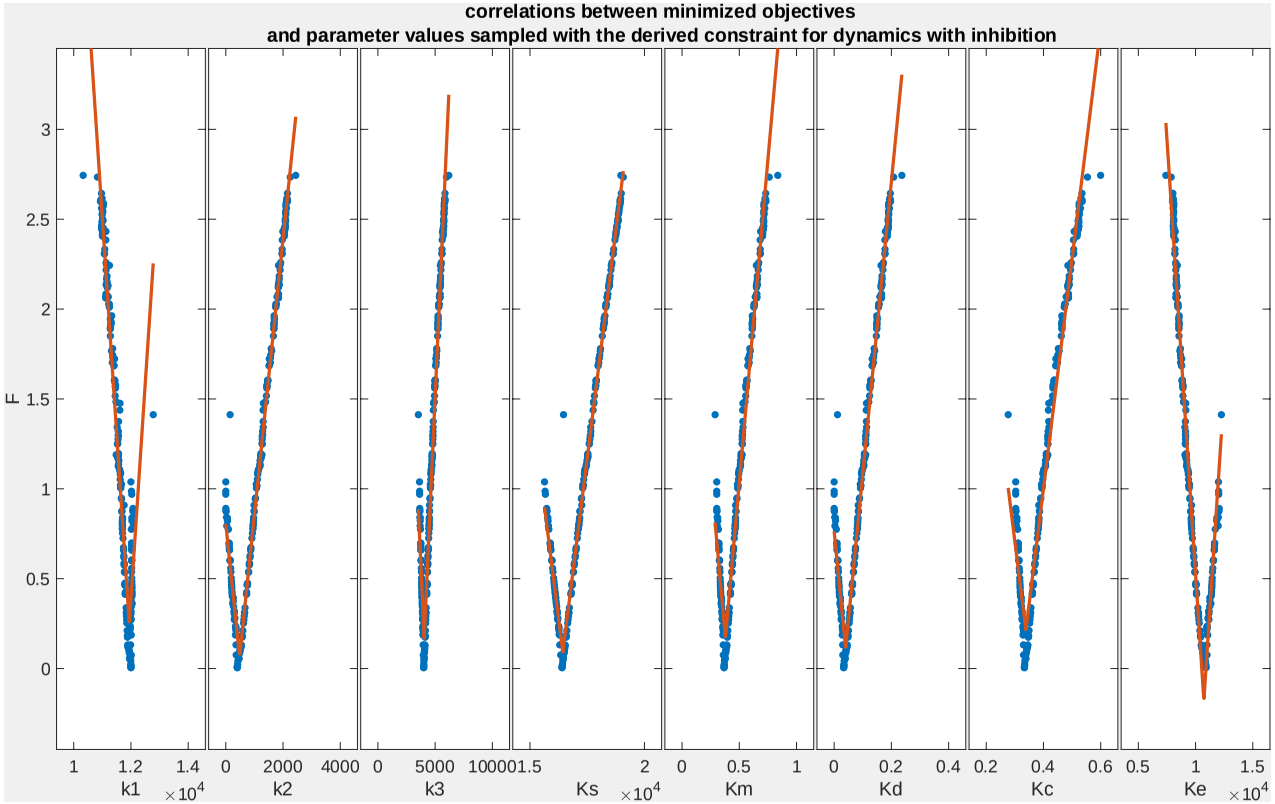


Figure 3.10: Correlations of found parameters with the optimization minimum are shown. A unique parameter set is found with the mean median absolute deviation of $4.55 \cdot 10^2$. The fit is done with an Ansatz $m|x - x_s| + t$, with the mean slope of 1.02.

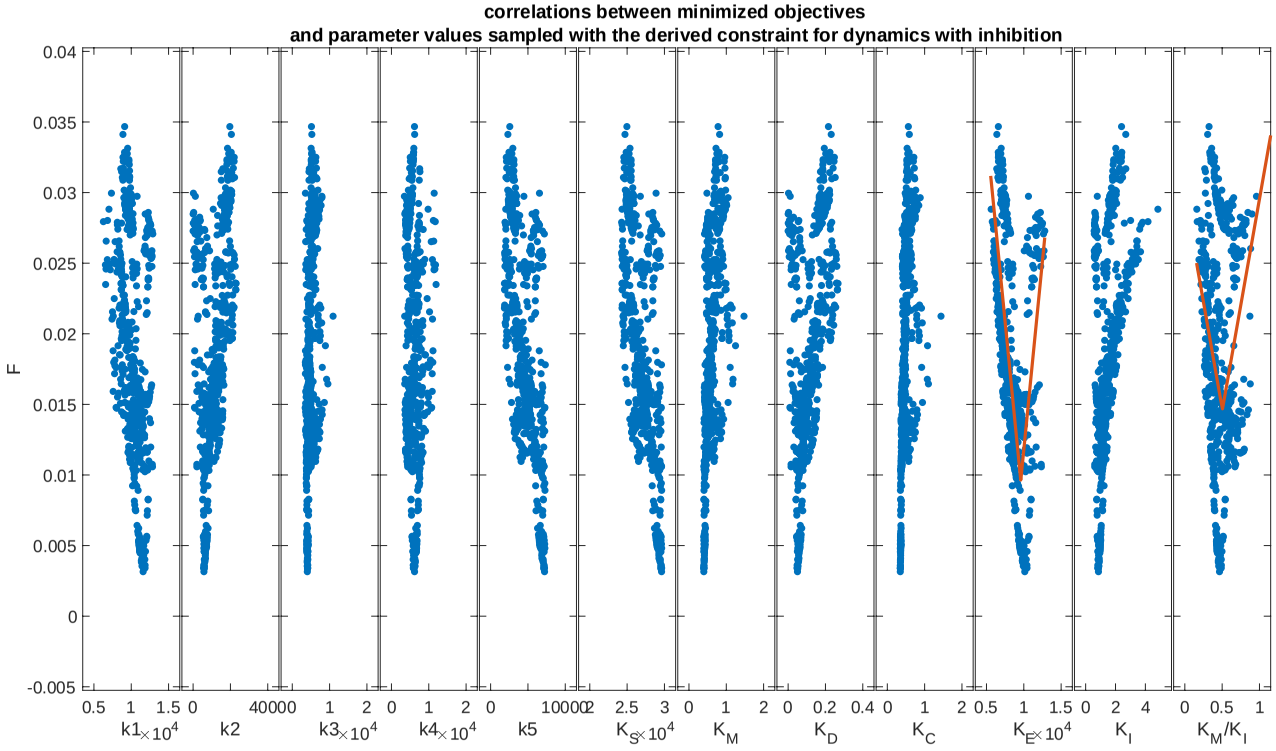


Figure 3.11: Correlations of found parameters with the optimization minimum, the sampling parameter K_S and the provided constraint found by linear stability analysis is shown. The uniqueness is less visible, compared to the simple dynamics case, however the minimum is apparent, as seen in figure 3.8(b). A unique parameter set is found with a mean median absolute deviation of $5.73 \cdot 10^2$

Again, the definitions of the sampling parameter and the constraint are kept fixed. The global analysis is therefore a rough test of optimization performance with the context of chosen sampling parameter, constraint and the optimization algorithm in use.

The unbound sampling parameter was chosen in such a way, that all other parameters are constrained into a box with finite dimensions. This can be accomplished in various ways, in particular four of them were inspected.

l^∞ norm

Choosing the l^∞ norm for the parameter vector of the model parameters, led to a box constraint of the sampling parameters with one of them being normalized to 1.

$$\begin{aligned} 0 \leq p_i \leq 1 \qquad & \forall i \\ \max_i p_i = 1 & \end{aligned} \tag{3.32}$$

l^2 norm

Choosing the l^2 norm for the parameter vector of the model parameters, led to a sphere constraint of the sampling parameters

$$\begin{aligned} 0 \leq p_i \leq 1 \qquad & \forall i \\ \sum_i \sqrt{p_i^2} = 1 & \end{aligned} \tag{3.33}$$

l^1 norm

Choosing the l^1 norm for the parameter vector of the model parameters, led to a diamond constraint of the sampling parameters

$$\begin{aligned} 0 \leq p_i \leq 1 \qquad & \forall i \\ \sum_i p_i = 1 & \end{aligned} \tag{3.34}$$

l^0 norm

Choosing the l^0 norm, defined by Banach, in ref. [161] for the parameter vector of the model parameters, led to an abstract F-Space defined by

$$\begin{aligned} 0 \quad \leq \quad p_i \quad & \leq 1 \qquad \forall i \\ k_i \quad = 2^i \frac{p_i}{1-p_i} & \\ \sum_i 2^{-i} \frac{k_i}{1+k_i} \quad & = \frac{1}{K_S} \end{aligned} \tag{3.35}$$

The constraint is nonlinear in this case, using all sampling parameters, including the normalization parameter for feasibility detection of model parameters. The sampling is associated with the l^0 norm due the normalization procedure, considered e.g., in ref. [162].

The normalization parameter K_S was sampled uniformly for direct phase space search and using its inverse for bayesian optimization.

	Michaelis Menten without inhibition		Michaelis Menten with inhibition	
	convexity	time per eval $[s]$	convexity	time per eval $[s]$
l^∞	$5.6 \cdot 10^{-1}$	$3.9 \cdot 10^{-1}$	$1.7 \cdot 10^{-1}$	7.3
l^2	$3.6 \cdot 10^{-5}$	$6.6 \cdot 10^{-1}$	$3.6 \cdot 10^{-3}$	6.9
l^1	$1.13 \cdot 10^{5}$	1.3	$9.5 \cdot 10^{4}$	3.3
l^0	$2.9 \cdot 10^{-5}$	1.5	$1.1 \cdot 10^{-5}$	1.2

Table 3.1: Different metrics are equivalent, however, the behavior of used optimization algorithm differs a lot, depending on the used metric definitions. As the convexity measure, the integral of $\int \min(F) - F$ over the sampled range was used. During the optimization a benchmark was done, which is then was divided by the number of samplings.
As larger the convexity measure is, as better: This is a rough measure of expressiveness of the found minimum.
As smaller the time per evalution is, as better: For a successful function phase space landscape approximation about 30 samplings were needed.

While the sampling procedure associated with l^1 norm shows the best results, others lead to less pronounced minimum during the optimization runs, see table 3.1 for minima shape characterization. The minimum could either not be found or the optimization time become huge, making the routine infeasible.

The optimization of the model given by partial differential equations is a smooth model, without any stochasticity. Therefore, for the hyperplane defined by the constraints above the local minimization routine fmincon was used. For phase space landscape of the problem bayesian optimization was used.

4 Model application

This chapter contains some demonstrations of the model abilities, in case dynamic parameters can be found.

Dependency on different properties of substrate can be studied, like dependency on the ratio of surface to volume or the presence of lignin and different polymer distributions, see sections 4.1 - 4.3. Further model refinements, like explicit modelling of crystallinity dependency can be tested, see section 4.4.

Last but not least, the model can be used to predict the best possible mixture of enzyme concentrations for a given substrate description, as it is done in section 4.5.

Some preliminary considerations can be taken beforehand prior discussing the influence of different substrate features on the enzyme dynamics. It is assumed, e.g., that the initial substrate is insoluble. This way, the shortest polymer length has to be chosen equal to the solubility barrier. Furthermore, realistic enzyme concentrations do not lie above 0.01mol/mol with respect to substrate concentration.

4.1 Ratio of active surface to volume

Keeping the enzyme concentration constant with respect to substrate concentration and varying substrate's initial accessible surface the dependency on it can be depicted. The results are shown in figure 4.1. The assumption still holds, that substrate is present in excess, that is, even with small accessible surface area all enzyme entities are able to bind at any time.

In order to achieve a constant concentration with tested surface/volume ratios, two different initial bulk configurations are tried.

$L_0 = 500$ glucose units, combined with 6×6 cross section is used first. This is compared against a single polymer chain with the length of 18000 glucose units. The enzyme concentration and the measured time is kept constant.

4.2 Dependency on lignin coverage

Modelling lignin is an attempt to model a more realistic substrate. Therefore also a more realistic enzyme has to be used for this demonstration.

Using an *EG* enzyme, two different substrates, which differ in initial lignin coverage were tried. The results can be seen in figure 4.2

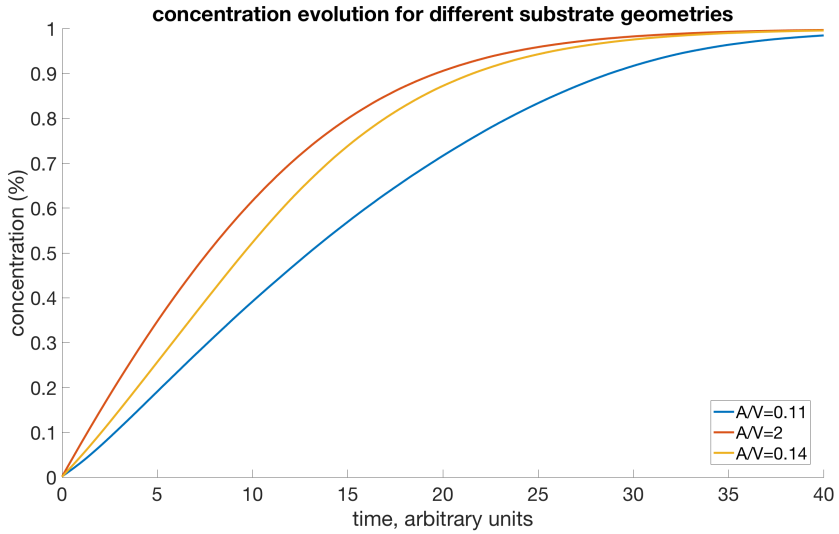


Figure 4.1: For substrates with different initial ratios of surface to volume time evolution of product concentration is shown. Even for the artificial enzyme kind, there evolves different dynamics. As the initial accessibility become larger, the substrate conversion also grows. This is expected behavior.

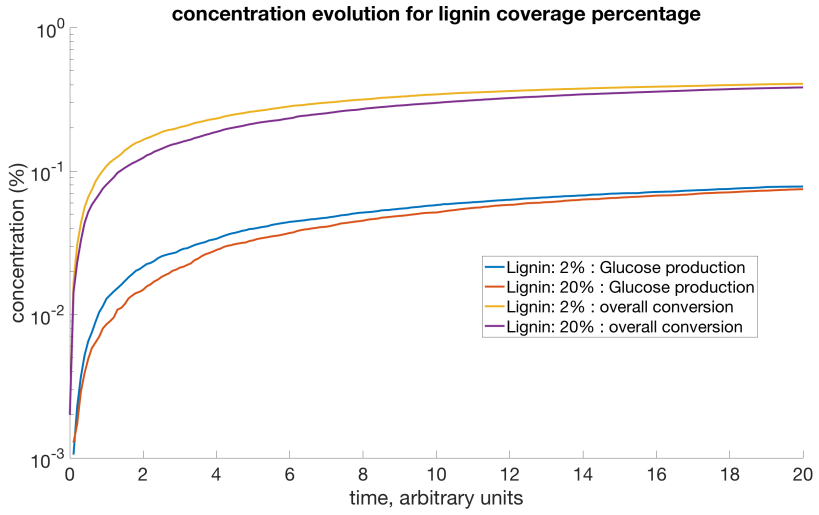


Figure 4.2: Two distinct lignin coverage percentages were tried. The overall conversion of substrate as well as glucose production depend on lignin coverage in general. The dependency is seen mostly in the initial and intermediate time span. At the end, substrate is expected to be degraded to the level, where lignin coverage plays a less important role.

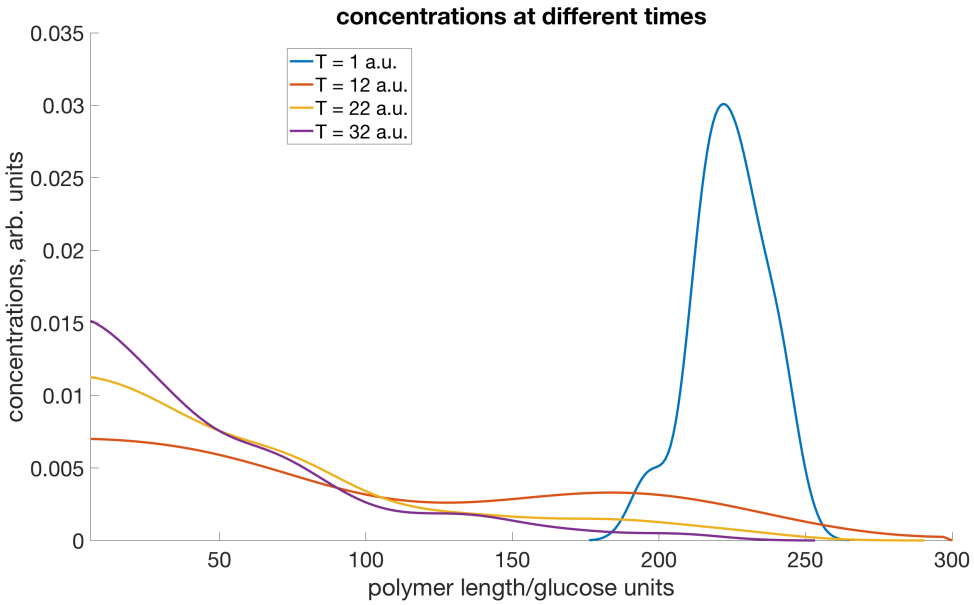


Figure 4.3: Model prediction of polymer lengths distribution during cellulose degradation at various times. The data was created by means of the stochastic model. Times, and concentrations are in arbitrary units. Polymer lengths are in glucose units.

4.3 Polymer distribution

The change of polymer distribution during the hydrolization process is a common heuristic shown in literature, see e.g., [163]. The results therein are provided by analytic or explicit numeric models. Here, the results shown in figure 4.3 are for the stochastic model derived in chapter 2.

4.4 Rate dependency on crystallinity

The hydrolytic ability of enzymes is known to be almost linearly dependent on the crystallinity degree of the substrate, see e.g., ref. [164].

Trials were done to introduce some relation between k_i to reproduce this behavior. However a parametric model was quickly dismissed.

Nevertheless, optimization of the rate constants to experimental data taken from Agarwal et. al., [164] was successful. The results show a clear relation between measurable enzyme characteristics and the measured data, see figure 4.4. Therein substrate hydrolysis percentage after a certain amount of time is shown on the left ordinate. The common x-coordinate is chosen to be initial crystallinity index of a given substrate. On the right ordinate optimization results of the ratio k_{cat}/K_M are shown.

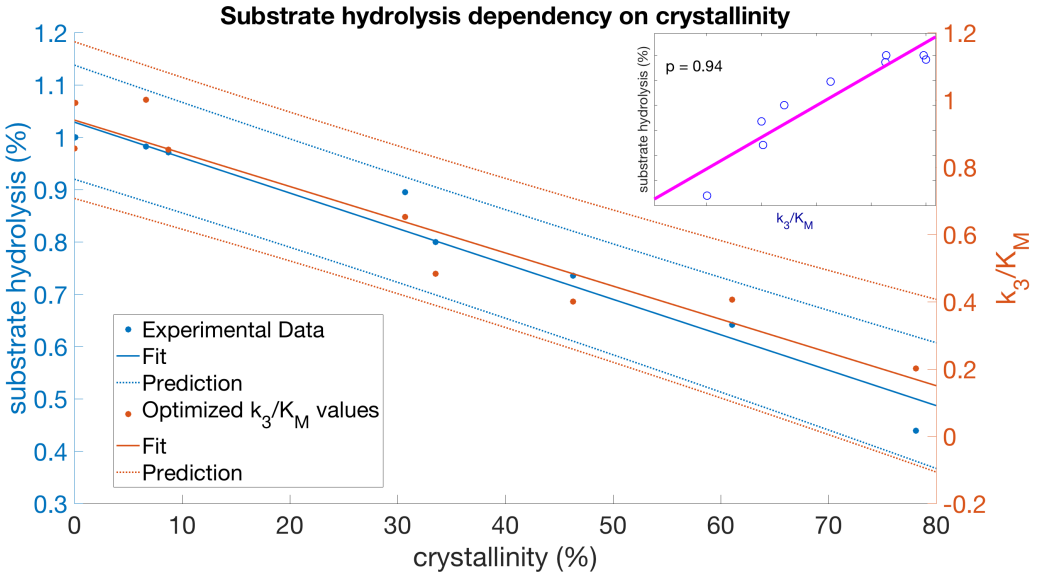


Figure 4.4: Experimental data taken from Agarwal et. al., [164] for hydrolysis after $72h$. It correlates clearly to the optimization results of rates, presented as the ratio of k_{cat} and K_M.
The rate constants k_1, k_2, k_3 were optimized independently for each experimental point under identical optimization conditions. Then, the results are compared with experimental data.

The prediction bands for further expected measurements are overlapping in a large area. The Pearson correlation coefficient is almost 1. The correlation coefficients of this experimental data vs. k_{cat} is 0.57, and vs. K_M −0.66.

4.5 Best possible mixture search

Predictions about best possible mixture are highly appreciated in enzyme model applications, see e.g. refs [165, 148]. The common way to depict the results are ternary plots.

For the best mixture search the optimization results from former sections were used for the CBH and EG enzymes. The action mode of β-G enzyme was assumed. The phase space was now scanned by varying the initial amounts of enzymes, whereas their total amount was kept constant.

The results are seen in figure 4.6.For comparison, results from Levine et. al, in ref. [148] are reused in figure 4.5.

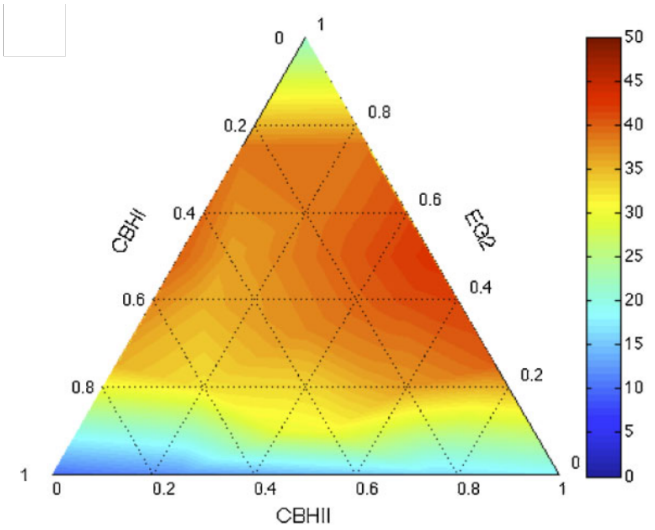


Figure 4.5: Three-dimensional plots of percent of cellulose at 24h of hydrolysis with mixture of EG2-CBH1-CBH2. See Levine, et. al, see ref. [148].Vertical axis shows total cellulose conversion (%). Total cellulase loadings is 10 $mg\ g^{-1}$, DP = 1500 and surface area = 47.6 $m^2\ g^{-1}$.

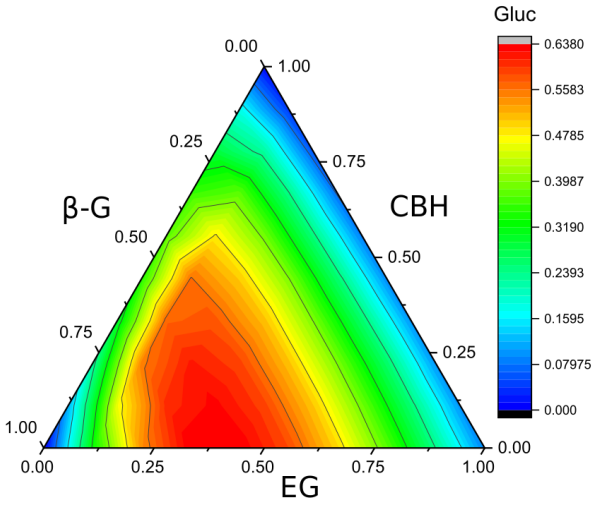


Figure 4.6: Simulation data of glucose yield after 24h of hydrolysis with mixtures of EG1-CBH1-β-G. Vertical axis shows percentage of glucose yield w. r. t. of initial substrate amount. Total cellulase loadings is $10 mmol/mmol$ substrate. DP= 200 and surface/volume ratio of substrate $2/3 m^{-1}$.

5 Conclusions

5.1 New model for enzymatic substrate hydrolization

In the present work, a new model for enzymatic substrate hydrolization was created. The novelty of the model lies in the rigour of its formulation. All features are designed in a modular way and are independent of each other. They can be switched on, off freely and tuned in different ways.

The Cellulect model contains a distinct formulation of the substrate and enzymes. Both contain numerous features, which can influence the behavior of the hydrolization process. New features can be added easily and their influence studied in connection with other features or independent of them.

Furthermore, the model contains a propagation algorithm formulated separately and independent of the underlying components. The algorithm is stochastic exact, paying tribute to the stochastic nature of biochemical reactions and being able to keep track of every of them.

The substrate is discretized at the level of product units. This approach keeps the discretization level lower than in many general purpose modelling frameworks. Furthermore, it allows to keep track of all conservation laws present in the modelled system.

All features of the substrate such as its crystallinity index, degree of polymerization, accessible surface, lignin coverage and the overall form factor is mapped and expressed by the discretized units. The discretization error is therefore predictable for all features and easy to control by unrestricted choice of the system size.

The initial state of the substrate can be freely described by experimentally measured means of its properties. However, the model also provides the possibility to specify initial distributions for them, such as initial crystallinity and polymerization distributions.

The main feature of current substrate modelling is the representation of its voluminous nature by voxel units. On one side, this approach is already known and widely used, e.g., in medicine. On the other, this allows to represent arbitrary structures, once the unit of measure is defined.

Modelled enzyme features contain, among other things, their size, a distinguishable position of its active center, orientation, processivity, different action modes, carbohydrate binding module, catalytic domain and a flexible linker. Experimental knowledge of typically produced products and binding partners can and is supposed to be plugged into their model. Different reaction propensities can be modelled dependent on encountered substrate regions.

Three different enzyme types were formulated and used for demonstration purposes. A further, artificial enzyme type was created for comparison reasons. The addition of further enzyme types can be done in a natural way, formulated by difference to

already modelled enzyme types. Extending common features leads to consistent behavior changes across defined enzyme types.

The enzyme model uses the same notion of discretization as the substrate. This allows an exact modelling of their actions on a molecular level. The model keeps track and updates the system state on a reaction base. The stochastic nature of the system is retained and demonstrated in the course of the work.

The heart of the model is a dedicated, well known time-reaction propagating algorithm, proved and used in diverse models since over forty years. The algorithm is kept completely independent from other parts of the model. This allows to easily reason about it, increases its maintainability and flexibility. Performance enhancements were applied to improve the overall simulation speed.

Separation the time propagation from other model content implies a normalization process of all quantities present in the model. This way, substantial quantities of the model could be defined and identified. First hints for the abstract analysis approach of the system appeared due to this modelling approach.

The used algorithm is optimal in the sense, that every simulation step contains a trackable state change. Idle time steps are excluded by definition. Every reaction is chosen stochastically, physical time scale is modelled in concert with the ongoing reactions.

Furthermore, an impressive agreement of the developed model with numerical solutions of differential equations was established. The numerical error is the result of system discretization only and can be taken under full control by enlarging the system. For deriving statistical surveys about simulation runs several runs are needed. This is common among stochastic models.

The purpose of the present work was to establish a simple yet powerful and extendible framework for enzymatic substrate degradation modelling. This was accomplished by tackling the biochemical reaction equations developed by Michaelis Menten. To get close to reality the model was complemented with competitive, reversible product inhibition. This is an important reaction observed in closed systems, preserving the conservation law of mass.

Diffusion was neglected, which imply a well-stirred system. Furthermore, the assumption about enzyme reaction rates far from diffusion limiting case goes hand in hand with this operational mode. The limiting time scale for enzyme reactions was derived. Another point is, that even if diffusion has to be be taken into account, the data cannot contradict the provided measurements. Therefore, all achievements in this work will remain valid within their own framework.

New features, which are not considered in the current model state can be added in a natural way, see section 5.3 for some ideas of model developing.

The validation of the model by means of experimental data was found to be underdetermined. In this regime, many valid solutions are found, fulfilling the requested optimization criteria. For convenience and phase space landscape estimation the bayesian hyperparameter optimization was used.

The hyperparameter was identified as the lacking Lagrange parameter missing in the underdetermined case. Linear stability analysis was used to identify the optimal parameter which then was used to receive numerical results. It was found, that only a

single parameter is enough to localize an optimization optimum however. More known parameters can be passed, in case the single one is not experimentally accessible.

Different random samplings, distance measures and constraint formulations were tried during optimization runs. However, a certain complexity degree seems to remain. A rough assessment of optimization problem formulations was done to compare optimization performance. One of the tested routines was chosen for best performance to make further quantitative analysis.

After the validation the model was used to reproduces experimental data. This was successful to some extent. So, the original model given by partial differential equation could be reproduced. Dynamic parameters of some of more realistic enzyme type could be obtained by optimization. However, as the experimental curves become more complex, the optimization of all experimental results simultaneously fails. Heuristics could be obtained and reproduces already known features of enzyme models.

All in all, most difficulties originated from the fact, that the description of enzyme action mode is not unified and in particular comes from a different source than the experimental data. This lead to the most discrepancies of the model and the experiments, trying to reuse the model on different measurements. I. e. as the dynamic properties of the enzymes are only loosely coupled to their structural properties, a description based on their dependency is most likely correct only for very rare cases.

It was tried to solve the uniformization task in this work, by regarding most different prevalent enzyme kinds of EG, CBH and β-G types. Many features are therefore included in the model and the observed dynamics could be explained by using only known biochemical terms. However, if further enzyme kinds are to be added, further properties are likely to be incorporated into the model.

5.2 Implemented model features

In figure 5.1 a sketch of implemented model features is shown. The code base is available under `git.physik.uni-marburg.de/alexander.orlov/cellulectlibrary.git`.

5.3 Outlook

Despite being a fully formulated model, there are still many features left out. For example,

- The acidity of solution pH during the reaction is assumed to be held on a constant level. This was done, assuming that all enzymes behave similarly under pH variation. However, to approach the reality even more, this effect can be taken into account, see ref. [166].
- The same assumption was made for the reactor temperature T. Here, the Arrhenius law 5.1 could be taken into account, see ref. [167]

$$k = k_0 e^{-\frac{E_a}{k_B T}} \tag{5.1}$$

with k_0 as a known reaction rate constants at some temperature, E_a the activation energy, k_B, the Boltzmann constant and T the current temperature.

While quantities above can be used for better modelling enzyme behavior under various reaction conditions, polarimetry measurements could also be used as input data, see refs. [168].

Crystallinity index CrI and lignin models could be unified by the recalcitrancy term, see ref. [169, 11, 149]. This would enhance the unification, integrity and consistency of the the substrate model even more. Also the difference of accessibility between microfibrils and macrofibrils would incorporated into this extension. The dependency of enzyme actions could be modelled on a finer scale.

Along recalcitrancy, enzyme behavior could be partly modelled in terms of fuzzy logic, see ref. [170]. This would reduce some of the logic cross products needed and contribute to model substrate accessibility in a uniform manner.

The Cellulect model is capable of modelling arbitrary complex substrate structures. This could also be used to model the substrate directly, from electron microscope pictures. Image recognition facilities can be involved for this purpose, see e.g., ref. [171] or [172].

Enzyme extents have to be expressed in the same units as the substrate. This is by design, as the substrate defines the smallest common unit of measure by the extent of products. However, in the present model all enzymes are assumed to have only one dimension, its length. A more elaborated model could be enhanced in this respect.

Another consideration concerns the form factor of the substrate. In the present model a rectangle cross section of the substrate was assumed. This was partly done for simplification and partly due to symmetry assumptions. However, another geometry could also be tested. The interplay of two or even tree dimensional enzymes with such geometries could be of interest.

The enzyme model can be extended in many ways. Beyond their geometry, mentioned above, more biological aspects could be involved. For example, several active centers and conformations could be included into the model, see ref. [173]. More complexation centers for a single substrate lead to hyperbolic propensity function [174].

Additional enzyme types could be forged. Enzyme types like Cellusomes, see ref. [175] were not considered in the present model. These were omitted in favor of well known enzyme types, which already build a complex system, which is not fully captured yet.

Hill equations, see ref. [176] were neglected to simplify the model even further. Using them as optimization constraints or taking them into account while comparing the model output could be a reasonable extension to the current model.

Diffusion treatment could be added. However, the model with diffusion should deliver the same numerical results, as seen in section 5.1. Nevertheless, the addition of diffusion could be used to compare results about the diffusion properties of various kinds of enzymes.

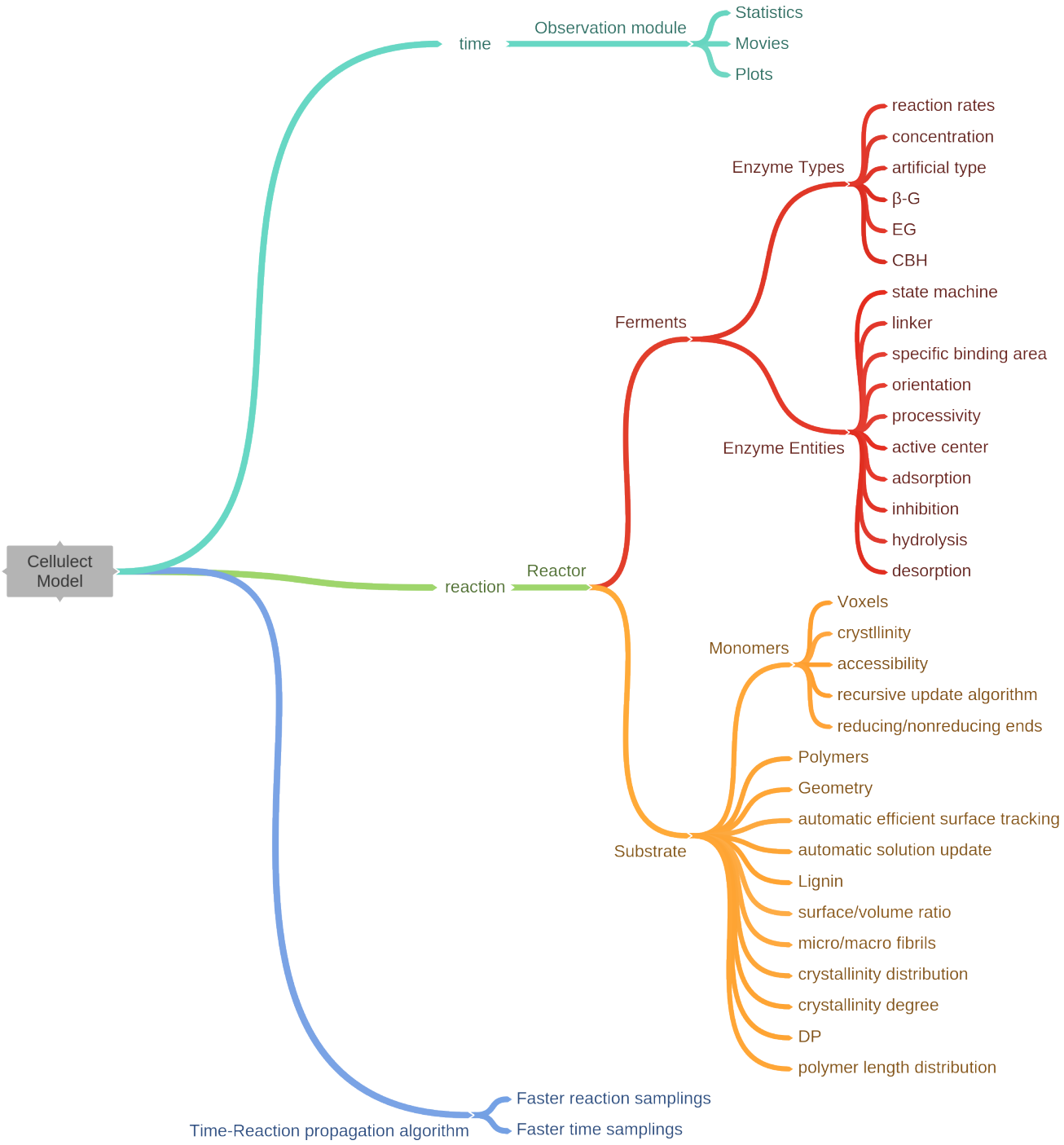


Figure 5.1: A sketch of implemented features in the cellulect model model.

A Support information

A.1 Diffusion coefficient calculation

Using the Stokes-Einstein equation the diffusion coefficient is derived to be

$$D = \frac{k_B T}{6\pi\eta \cdot f/f_{min} R_0} \tag{A.1}$$

see, ref. [177]. Where

- k_B: the Boltzmann constant
- T: temperature
- η: viscosity
- f/f_{min}: friction coefficient, taken into account, the particle is not a sphere
- R_0: radius of the particle

Now, first calculate R_0 to be

$$R_0 = \left(\frac{3m}{4\pi\rho}\right)^{1/3} = \left(\frac{3 \cdot 5}{4\pi \cdot 1.37 \cdot 6.022} \cdot 10^2\right)^{1/3} nm = 2.4367\ nm \tag{A.2}$$

here, is used

$$\begin{aligned} m &= 50\ kDa = 5 \cdot 10^4\ Da \\ \rho &= 1.37\ \frac{g}{cm^3} = 1.37 \cdot \frac{6.022 \cdot 10^{23}}{10^{21}}\ \frac{Da}{nm^3} \end{aligned} \tag{A.3}$$

see 2.3 for more details.

At the next step the constants are set, cf. [85].

$$\begin{aligned} k_B &= 1.38 \cdot 10^{-23}\ \frac{J}{K} = 1.38 \cdot 10^{-23} \cdot \frac{10^{18}\ nm^2 \cdot kg}{10^6\ ms^2 \cdot K} = 1.38 \cdot 10^{-11} \frac{nm^2 \cdot kg}{ms^2 \cdot K} \\ T &= 300\ K = 3 \cdot 10^2\ K \\ \eta &= 1 \cdot 10^{-3}\ Pa \cdot s = 1 \cdot 10^{-3} \frac{kg}{10^{12}\ nm \cdot ms} = 1 \cdot 10^{-15}\ \frac{kg}{nm \cdot ms} \\ f/f_{min} &= 1.5 \\ R_0 &= 2.4367\ nm \end{aligned} \tag{A.4}$$

This yields the diffusion coefficient to

$$D = \frac{1.38 \cdot 10^{-11} \cdot 3 \cdot 10^2}{6\pi \cdot 1 \cdot 10^{-15} \cdot 1.5 \cdot 2.4367} \frac{nm^2 \cdot kg \cdot K \cdot nm \cdot ms}{ms^2 \cdot K \cdot kg \cdot nm} = \\ = \frac{1.38 \cdot 3}{6\pi \cdot 1.5 \cdot 2.4367} \cdot 10^6 \frac{nm^2}{ms} = 6.0091 \cdot 10^4 \frac{nm^2}{ms} \tag{A.5}$$

The common definition of the diffusion length, depending on time is $l_D = \sqrt{2 \cdot n \cdot Dt}$. Here, a well stirred volume as well as only enzyme movements along the substrate are considered. Therefore, the diffusion length is set to $l_D = 2\sqrt{Dt}$; this yields

$$l_D = 2\sqrt{Dt} = 490,2673\ nm \cdot \sqrt{t} \tag{A.6}$$

This result is used in section 2.3 to derive a minimal time step cutoff, where neglecting diffusion remains valid.

A.2 Michaelis-Menten kinetics and beyond

Governing equations

In ref. [18] Michaelis and Menten introduced for the first time in the framework of enzyme kinetics. This theory is now known as Hidden Markov Model. The assumption was made, that some transient states exist, between that rates mediates the chemical conversion of reactants, see figure 2.11.

The basic reaction assumes a single intermediate state of adsorbtion between the substrate and the enzyme, after which the conversion of substrate to product by hydrolyzation happens. This initial idea already leads to a rich class of kinetics governed by nonlinear dynamic and algebraic equations, see equation (A.7) and ref. [59].

$$\begin{aligned} \partial_t S &= -k_1 SE + k_2 C_s \\ \partial_t P &= k_3 C_s \\ \partial_t E &= -k_1 SE + (k_2 + k_3) C_s \\ \partial_t C_s &= k_1 SE - (k_2 + k_3) C_s \\ E_T &= E + C_s \\ S_T &= S + P + C_s \end{aligned} \tag{A.7}$$

In case of a more elaborated model, it is possible to take more intermediate states into account. In this work, inhibition by product was taken into cosideration. This is a natural assumption, avoiding the system loosing all degrees of freedom at the end of cellulose conversion.

The according kinetic equations are as follows:

$$
\begin{aligned}
\partial_t S &= -k_1 SE + k_2 C_s \\
\partial_t P &= k_4 C_p - k_5 EP \\
\partial_t E &= -k_1 SE + k_2 C_s - k_5 EP + k_4 C_p \\
\partial_t C_s &= k_1 SE - (k_2 + k_3) C_s \\
\partial_t C_p &= k_3 C_s + k_5 EP - k_4 C_p \\
E_T &= E + C_s + C_p \\
S_T &= S + P + C_s + C_p
\end{aligned}
\tag{A.8}
$$

In both cases, there are two additional algebraic equations in the system, showing conservation of mass. They make it possible, to simplify two of the kinetic equations as they are linearly dependent via the constant of total enzyme E_T or substrate S_T concentration, respectively.

In the present work, the defined algebraic equations play by far a more important role than the kinetic equations. Rules for enzyme actions were added and imposed on the kinetic equations. The chosen modelling approach allows to define an arbitrary amount and variety of such rules for enzyme actions. This provides a fine grained control on the mode of action of enzymes and substrate properties updates, which are not trivially expressible by differential or integro-differential equations. However the given algebraic equations still play a role as stoichiometric mass conservation law, which is guaranteed to hold.

Macroscopic Constants

It is common to describe the substrate affinity by the Michaelis constant K_M, for which the following relation holds.

$$
K_M = \frac{k_2 + k_3}{k_1} \tag{A.9}
$$

where

- k_1 is the adsorption rate constant, which describes the formation of the enzyme substrate complex,
- k_2 is the desoption rate constant and
- k_3 is the catalytic rate constant, which denotes the conversion between substrate and product.

In case inhibition is modelled, it is convenient to introduce an equivalent inhibition constant K_I

$$
K_I = \frac{k_4}{k_5} \tag{A.10}
$$

where, additionally to the rate constants k_1, k_2, k_2

- k_4 is the desorption rate constant of an enzyme from the product

- k_5 is the adsorption rate constant, which corresponds to the forming of the enzyme product complex

Frequently, there are more macroscopic ratios introduced, for expressing measurable quantities at steady states, see ref. [59].

The Van Slyke-Cullen constant K_C, see [178]

$$K_C = \frac{k_3}{k_1} \tag{A.11}$$

The dissociation equilibrium constant K_D

$$K_D = \frac{k_2}{k_1} \tag{A.12}$$

The enzymatic specificity constant K_E

$$K_E = \frac{k_3}{K_M} \tag{A.13}$$

By analogy, an altered Van Slyke-Cullen constant $\widetilde{K}_C$, was defined as

$$\widetilde{K}_C = \frac{k_3}{k_5}. \tag{A.14}$$

State Machine and Graph represenation

As stated by Kemeny in [179] there is a natural connection between the general concept of state machines and Markov chains. Herein the connection between the representation of the process by means of state machines and computation by means of transition matrices is meant. In this work these two concepts are used in different contexts.

A Markov chain, is a stochastic process that moves successively through a set of states. If it is in state s_k it moves on to the next stop, state s_j, with probability p_{kj}. These probabilities can be exhibited in the form of a transition matrix. Here, it is important to note, that the whole system under inspection is modelled as a Markov chain. For each state, there is a defined constant rate for each transition. The process is therefore a Poisson process, as pointed out in section 2.1. On each transition the state of substrate changes unavoidable by using the time propagating algorithm, described in section 2.2.

On the other side, the concept of a finite state machine (FSM) is used to model enzymes. A FSM being either of Mealy-kind, see ref. [180] or Moore-kind, see ref. [181] and being indeed equivalent by the powerset construction, see ref. [182] is rather a technical notation based on conventions, see ref. [183] and does not have an unique definition.

It can be defined as a quintuple $\mathcal{A} = (S, I, A, \delta, \gamma)$. Where

- S is the set of automata states (or nodes),
- I the set of input set, called alphabet,

- A output set
- $\delta : S \times I \to S$ the next-state function (or transition function) and
- γ the output function.

cf. also [184]. S, I and therefore A are used to be finite.

The notion of automata contains general rules which lead to transitions from its state s_k to state s_j depending on the current state by means of an optional input (=event), optionally providing an output. This model approach is chosen to represent enzymes, identifying them as the active species of the model. With them as a helping tool, the probabilities for the whole system mentioned above are gained.

Constrained to chemical reactions 2.11 and 2.12 the reaction graphs and corresponding state machines for enzymes and substrates are straight forward to formulate, see fig. A.1.

There are two different problems here. On one side, the representation of enzymes as state machines. On the other side, the transition matrix representing the Markov process used to model the whole system of glucose units, polymers constituting the substrate and enzymes hydrolyzing the substrate. They must not be mixed. However the idea is, that the enzymes provide required probabilities to the governing transition matrix. As the simulation goes on, the states of enzymes and the substrate evolve, updating the possible inputs to the enzymes as state machines and providing in such a way the base of calculation of updated transition probabilities.

Now, for the purpose of the present work, following entities can be identified

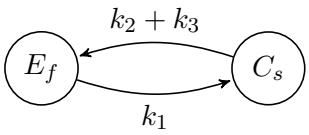
- S with the states of an enzyme in figure 2.11 or 2.12, respectively.
- I as the reaction being sampled by the time reaction evolution algorithm,
- the next-state function as the appropriate transition function according to the reaction graph

Inspecting the states of the reactants directly, the output function γ as well as the output A can be neglected. This leaves behind a semi-automat as defined by Medvedev in ref. [185, pp. 385-401]

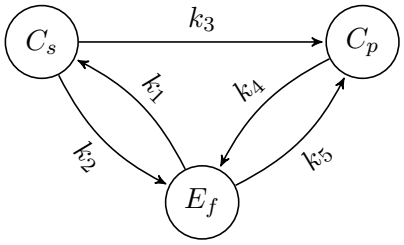
The goal is now, by inspecting the possible inputs to the enzymes as state machines, taking given reaction rates into account and the current state of enzymes to calculate every probability p_{ij} needed for the Markov process to make the next jump.

For this purpose the possible inputs are partitioned by means of the reaction rates. I.e., there is a 1:1 relation between reaction rates and reactions (= inputs for enzymes). Then, the substrate (and accordingly the product) is subdivided to match the given reaction rates. Similarly, enzymes are partitioned. This is done by counting sampled entities of the substrate and enzymes. The underlying process is a Poisson process and therefore a counting process, which allows the procedure.

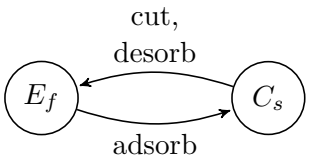
Multiplying the partitioned amounts of reactants with the according reaction rates w.r.t. their units yields propensities of each reaction, and their sum yields the intensity of the underlying Poisson process. This indeed coincides with formulation of the process



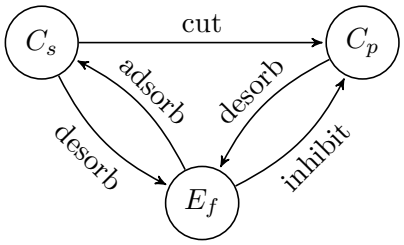
(a) Reaction graph for Michaelis-Menten kinetics without inhibition for enzymes



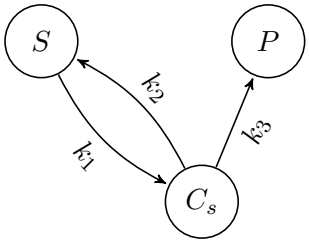
(b) Reaction graph for Michaelis-Menten kinetics with inhibitionfor enzymes



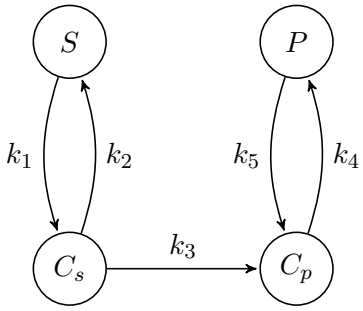
(c) State machine of an enzyme of Michaelis-Menten type without inhibition



(d) State machine of an enzyme of Michaelis-Menten type with inhibition



(e) Reaction graph for Michaelis-Menten kinetics without inhibtion for substrate and product



(f) Reaction graph for Michaelis-Menten kinetics with inhibtion for substrate and product

Figure A.1: Common state machine for enzymes, restricted to Michaelis Menten dynamics, see also figures 2.11, 2.12 for definitions of k_i and ref. [97].
Figures (a) and (b) show continuous time Markov chain (CTMC) representation with reaction rates on edges.
Figures (c) and (d) show general state machine representation with according events leading to state changes.
Figures (e) and (f) show the common reaction graphs for substrate and product.
As the descriptions are independent, they can be simultaneously achieved. In order for the simulation to be correct, this has to be guaranteed.

as a master equation, cf. [186]. By normalizing the propensities by their sum the probabilities of reactions are calculated. The reciprocal intensity yields the uniformization constant [75] and therefore the time which is needed for the next reaction.

A.3 Notion of time

To justify the time calculation for each reaction there is some work to be done. First, note, that enzymes are modelled as state machines. Though, the output of the state machines is not measured directly. This is influenced by the Michaelis-Menten kinetics as well as by the fact, that due to small enzyme amount, the measurement of their concentrations is generally not accessible.

The general framework governing this approach is the framework of hidden-Markov model (HMM). The first reference dates back to Baum & Petrie in [187], while a more elaborated work was done by Zucchini & MacDonald in [188].

HMMs assume a bivariate process (X_t, Q_t), $t \in T$ where X_t are the observable random variables, whereas Q_t is the hidden process, as defined in [57]. In this work Q_t is associated with the state machine of enzymes and X_t with observations made on the substrate.

However, in contrast to the definition in [57], the index set T is not taken to be necessarily discrete. It is either $T = \mathbb{N}$ or $T = [0, \infty)$ and $\mathcal{T}$ is the Borel σ-algebra of T. $(T, \mathcal{T})$ is then the measurable space of the time domain. The domain can be even extended by the one-point compactification of T to infinite times: $T_\infty = T \cup \{\infty\}$ and its σ-algebra in an analogous manner. So that $(T_\infty, \mathcal{T}_\infty)$ is also a measurable space. A random variable τ taking values in T_∞ is called a random time.

The hidden Markov chain is assumed to be time-homogeneous, i.e., the associated transition matrix Q for the hidden Markov model still has the Laplace property, see ref. [189]:

$$Q(G)_{ij} = \begin{cases} e_{ij} & \text{if } i \neq j \\ -\sum\limits_{v \neq j} e_{vj}, & \text{if } i = j \end{cases} \tag{A.15}$$

with G as the graph of the CTMC encoding the process Q_t and e_{ij} being the labels on the transitions of the graph. However, the entries e_{ij} are generalized to be dependent on the observed variables X_{t-1}: $e_{ij} \equiv e_{ij}(X_{t-1})$. The conditional distribution X_t also remains time-homogeneous: $P(X_t = x | Q_t = q) = p(x|q)\ \forall t$. The HMM defining property holds, too:

The ith conditional probability splits into

$$P(X_t, Q_t | X_{t-1}, \dots, Q_{t-1}, \dots) = P(X_t | X_{t-1}, \dots, Q_t, \dots) \cdot P(Q_t | X_{t-1}, \dots, Q_{t-1}, \dots) \tag{A.16}$$

The observation equation holds:

$$P(X_t | X_{t-1}, \dots, Q_t, \dots) = P(X_t | Q_t) \qquad \forall t \in T \tag{A.17}$$

stating, that the conditional distribution of X_t is completely determined by the current hidden state Q_t. This is the Markov property of the observed process X_t.

However, in the present case the process $(X_t, Q_t)_{t\in T}$ is neither purely parameter-driven:

$$P(Q_t|X_{t-1},\dots,Q_{t-1},\dots) \neq P(Q_t|Q_{t-1}) \tag{A.18}$$

nor purely observation-driven:

$$P(Q_t|X_{t-1},\dots,Q_{t-1},\dots) \neq P(Q_t|X_{t-1}) \tag{A.19}$$

Though, it holds:

$$P(Q_t|X_{t-1},\dots,Q_{t-1},\dots) = P(Q_t|Q_{t-1},X_{t-1}) \tag{A.20}$$

Now, as the Laplace property for the Q_t process still holds, assume X_t is a counting process $\{N(t), t \geq 0\}$ for the events of Q_t. It is valid to assume the Poisson distribution of X_t conditioned on $Q_t = q$ given for each time point as

$$P(X(t) = n|Q_t = q) = e^{-\lambda_q \cdot t} \cdot \frac{(\lambda_q \cdot t)^n}{n!} \tag{A.21}$$

So, the Poisson intensity for a given state Q_t is λ_q and time is exponential distributed.

During the simulation λ_q is calculated on the fly, taking the current state of concentrations as well as provided reaction rates into account. Then, after the current time step is determined, the reaction is sampled according to the provided propensities. Each propensity on their own has a one-to-one relation to the corresponding reaction rate.

The algorithm separates the concerns of time calculation and reaction sampling. During the time calculation the values of propensities are important, while during reaction sampling only their ratios are important. The optimization therefore has to be formulated in a hierarchical manner.

While the general optimization problem (3.1) remains well formulated, the hierarchical optimization is formulated as hyperparameter optimization, see refs. [190, 191, 192, 193, 194, 195, 196, 197, 198]. In general this problem remains hard to solve. Only randomization and estimation algorithms exist. Though, a proper sampling of variables can lead to some insights.

The full optimization procedure is defined as following:

- sample a hyperparameter, which corresponds to a time scale ratio in the system.
- given the hyperparameter, search for an optimum of other parameters present in the system optimizing a defined cost function.
- repeat the two steps (at random or by means of a dedicated estimator) until the global optimum of the cost function is found. The set of parameters yielded is valid in both respects. The hyperparameters are the time scale defining parameters. Intern parameters are set by optimization w.r.t. experimental data to fixed values.

Some possible hyperparameter definitions as well as optimization results are shown in chapter 3.

B Technical details

B.1 Van Emde Boas tree

One of the performance critical sections was the storing of polymer objects during simulation runtime. For this task a cache oblivious data structure, the van Emde Boas tree, with constant time for inserting, removing, searching of elements and neighbors actions was used, see refs. [199, 200].

Using this tree structure all operations are achievable in $loglogM$, where M is a constant universe size. As the law of mass conservation holds during a single simulation run, the maximum number of elements stays constant and is known since initialization.

A polymer is identified by the glucose unit at the reducing end. Each glucose unit has an identity. In such a way the look up for the corresponding polymer identity takes a single look up, starting at arbitrary position and performing a neighbors search in the appropriate direction.

By performing two searches in opposite directions the whole polymer can be constructed.

During a hydrolyzation step a new element is inserted into the tree structure. In this way a new polymer emerges and the list of them is kept up to date.

For convenience, the number of tree elements is also kept up-to-date. In such a way the look up for current polymer amount also takes $O(1)$.

B.2 Cut theorem

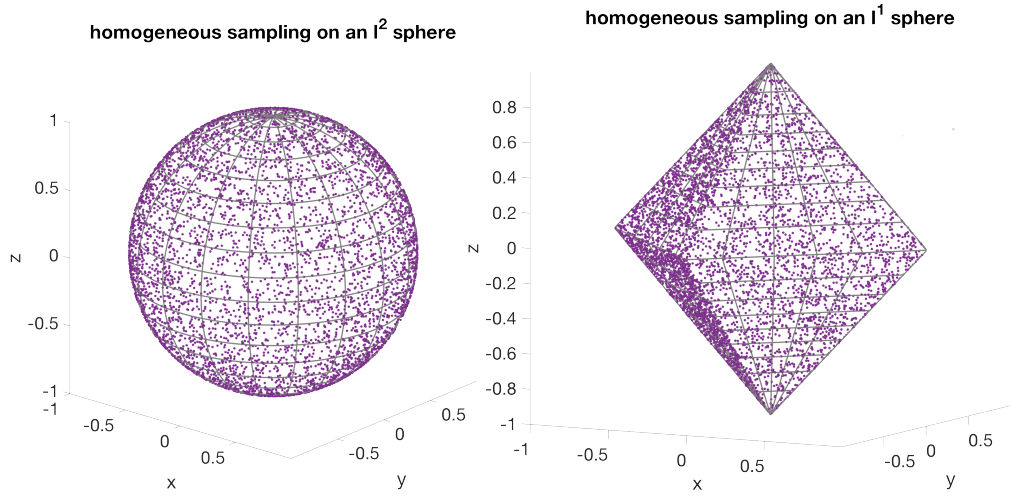
The cut elimination theorem is a central theorem in sequence calculus theory, see refs. [201, 202]. In its simple form the theorem states

$$\Gamma \vdash A \quad \wedge \quad A, \Delta \vdash B \quad \Rightarrow \quad \Gamma, \Delta \vdash B \tag{B.1}$$

A consequence of this theorem is, that for proof of consistency of a system it is enough to show, that A is consistent.

Now, lets identify Γ with some state of the substrate. A is the corresponding state of its polymers and Δ are actions performed by enzymes. The actions of enzymes transit the state of polymers A into state B.

With the cut elimination theorem it is easy to state now, that if the polymer objects behave consistent, then the whole system of substrate and enzymes is consistent. This approach was used during implementation to simplify reasoning about simulation runs.



(a) With l^2 metric a sphere arise as the unit cell of the natural space topology.

(b) In contrast to this, optimization of rates were driven using the l^1 distance metric. The unit cell deforms in this case to the shown form.

Figure B.1: Comparison of unit cells using l^2 and l^1 distance metrics in subfigures (a) and (b) respectively.

B.3 Random sampling within metric spaces

Random sampling is not a trivial task in general. Many efforts were made to achieve faster pseudo random number generation algorithms, longer periodicity and more uniformly samplings. For recent trends see e.g., refs. [203, 204, 205].

Additionally, the task becomes even more challenging, if constraints are added on the random samplings. For example, this has to be done to restrict the optimization routine to a feasible set of parameters, sampled from a more general set. Throughout this work it was a common task to sample random numbers in the positive metric space of $\mathbb{R}^n_+$, where either a scale parameter and/or an additional constraint was superposed.

To fulfill a proper sampling the scale parameter maps to its own linear subspace of $\mathbb{R}^n_+$. After this, the remaining problem becomes reduced by one dimension to $\mathbb{R}^{n-1}_+$. While the optimization algorithm takes over the sampling and phase space search inside the reduced subspace the separated $1-D$ radial component was used as a projection dimension either to represent the outcomes or to drive the samplings in the subspace.

It has to be guaranteed, that the samplings can be handled as on independent identically distributed random variables without any bias. In figure B.1 a comparison of such subspace separation for the l^2 and l^1 metric is shown. With the l^2 metric the natural space topology is euclidean and randomized points can be picked on a sphere, see refs. [206, 207]. For the l^1 metric the sphere deforms to a diamond.

However, in both cases a unit radius can be defined, which is orthogonal to the shown surfaces, and can therefore be independently sampled, optimized and analyzed.

List of Figures

Bibliography

[1] Ronald F. Fox and Theodore P. Hill. "A Proposed Exact Integer Value for Avogadro's Number". In: (Dec. 2006) (cit. on p. 13).

[2] M Okazaki and M Moo-Young. "Kinetics of enzymatic hydrolysis of cellulose: analytical description of a mechanistic model." In: *Biotechnol. Bioeng.* 20.5 (May 1978), pp. 637–63 (cit. on pp. 14, 41).

[3] *Wolfram statistics about historical barrel costs* (cit. on p. 16).

[4] Mamatha Devarapalli and Hasan K. Atiyeh. "A review of conversion processes for bioethanol production with a focus on syngas fermentation". In: *Biofuel Res. J.* 2.3 (2015), pp. 268–280 (cit. on p. 15).

[5] Mustafa Balat and Havva Balat. "Recent trends in global production and utilization of bio-ethanol fuel". In: *Appl. Energy* 86.11 (2009), pp. 2273–2282 (cit. on p. 15).

[6] Sims, R. Taylor M. Saddler J. Mabee W. "From 1st-to 2nd-generation biofuel technologies: An overview of current industry and RD&D activities". In: *Int. J. Veg. Sci.* 14.1 (2008), pp. 1–3 (cit. on p. 15).

[7] Kingsley Otulugbu. "Production of Ethanol from Cellulose (Sawdust)". In: *Plast. Technol.* (2012), pp. 1–46 (cit. on p. 15).

[8] Régis Rathmann, Alexandre Szklo, and Roberto Schaeffer. "Land use competition for production of food and liquid biofuels: An analysis of the arguments in the current debate". In: *Renew. Energy* 35.1 (2010), pp. 14–22 (cit. on p. 15).

[9] Kamal Kansou et al. "Testing scientific models using Qualitative Reasoning: Application to cellulose hydrolysis". In: *Sci. Rep.* 7.1 (2017), pp. 1–18 (cit. on pp. 15, 19).

[10] D J Whittle et al. "Molecular cloning of a Cellulomonas fimi cellulose gene in Escherichia coli". In: *Gene* 17.2 (1982), pp. 139–145 (cit. on p. 15).

[11] Xin-Qing Zhao, Li-Han Zi Feng-Wu Bai Hai-Long Lin Xiao-Ming Hao Guo-Jun Yue and Nancy W. Y. Ho. "Bioethanol from Lignocellulosic Biomass". In: *Adv. Biochem. Eng. Biotechnol.* 123.July 2015 (2010), pp. 127–141 (cit. on pp. 15, 89).

[12] Alicia Fernández Gómez. "Sustainability analysis of biofuels in Chile". PhD thesis. Lund University, 2007, pp. 1–50 (cit. on p. 15).

[13] Jadwiga Ziolkowska and Leo Simon. "Biomass ethanol production faces challenges". In: *Agric. Resour. Econ. Updat.* 14.6 (2011), pp. 5–8 (cit. on p. 15).

[14] Yi Zheng, Zhongli Pan, and Ruihong Zhang. “Overview of biomass pretreatment for cellulosic ethanol production”. In: *Int. J. Agric. Biol. Eng.* 2.3 (2009), pp. 51–68 (cit. on p. 15).

[15] Prasun Kumar et al. “Ecobiotechnological Strategy to Enhance Efficiency of Bioconversion of Wastes into Hydrogen and Methane”. In: *Indian J. Microbiol.* 54.3 (2014), pp. 262–267 (cit. on p. 15).

[16] Hoogland C. Ivanyi I. Appel R.D. Bairoch A. Gasteiger E. Gattiker A. *ExPASy: the proteomics server for in-depth protein knowledge and analysis*. 2018 (cit. on p. 17).

[17] Carlos M G A Fontes and Harry J Gilbert. “Cellulosomes: Highly Efficient Nanomachines Designed to Deconstruct Plant Cell Wall Complex Carbohydrates”. In: *Annu. Rev. Biochem.* 79.1 (2010), pp. 655–681 (cit. on p. 17).

[18] L Michaelis and M L Menten. “Die Kinetik der Invertinwirkung”. In: *Biochem Z* 49.February (1913), pp. 333–369 (cit. on pp. 17, 92).

[19] Jie Liang and Hong Qian. “Computational Cellular Dynamics Based on the Chemical Master Equation: A Challenge for Understanding Complexity”. In: *Comput Sci Technol* 6.1 (2010), pp. 247–253 (cit. on p. 17).

[20] Daniel a Beard and Hong Qian. “Chemical Biophysics - Quantitative Analysis of Cellular Systems”. In: *Cambridge Univ. Press* (2008) (cit. on p. 17).

[21] R Sousa Jr et al. “Recent trends in the modeling of cellulose hydrolysis”. In: *Brazilian J. Chem. Eng.* 28 (2011), pp. 545–564 (cit. on p. 17).

[22] E. Ccopa Rivera et al. “Enzymatic hydrolysis of sugarcane bagasse for bioethanol production : determining optimal enzyme loading using neural networks ˆ”. In: *Wiley Interscience* (2010), pp. 983–992 (cit. on p. 17).

[23] Seth E Levine et al. “A mechanistic model of the enzymatic hydrolysis of cellulose.” In: *Biotechnol. Bioeng.* 107.1 (Sept. 2010), pp. 37–51 (cit. on pp. 17, 21, 25, 46).

[24] Andrew J Griggs, Jonathan J Stickel, and James J Lischeske. “A mechanistic model for enzymatic saccharification of cellulose using continuous distribution kinetics I: depolymerization by EGI and CBHI.” In: *Biotechnol. Bioeng.* 109.3 (Mar. 2012), pp. 665–75 (cit. on pp. 17, 21).

[25] Andrew J Griggs, Jonathan J Stickel, and James J Lischeske. “A mechanistic model for enzymatic saccharification of cellulose using continuous distribution kinetics II: cooperative enzyme action, solution kinetics, and product inhibition.” In: *Biotechnol. Bioeng.* 109.3 (Mar. 2012), pp. 676–85 (cit. on pp. 17, 21, 55).

[26] D. Kumar and G. S Murthy. “Stochastic molecular model of enzymatic hydrolysis of cellulose for ethanol production”. In: *Biotechnol. Biofuels* 6.63 (2013), pp. 1–20 (cit. on pp. 18, 21, 25, 26, 46, 54, 55, 68).

[27] Bert Bredeweg et al. “Garp3 — Workbench for qualitative modelling and simulation”. In: *Ecol. Inform.* 4.5 (2009), pp. 263–281 (cit. on p. 18).

[28] K Forbus. “Qualitative Reasoning: Computer Science and Software Engineering”. In: *Comput. Handb.* Ed. by Teofilo Gonzalez, Jorge Diaz-Herrera, and Allen Tucker. 3rd. CRC Press, 2014 (cit. on p. 18).

[29] Yumi Iwasaki. “Real-World Applications of Qualitative Reasoning”. In: *IEEE Expert Intell. Syst. Their Appl.* 12.3 (1997), pp. 16–21 (cit. on p. 18).

[30] Johan de Kleer Daniel S. Weld, ed. *Readings in Qualitative Reasoning About Physical Systems*. Elsevier, 1990 (cit. on p. 18).

[31] Benjamin Kuipers. “Qualitative reasoning: Modeling and simulation with incomplete knowledge”. In: *Automatica* 25.4 (1989), pp. 571–585 (cit. on p. 18).

[32] Desmond J. Higham. “Strathprints Institutional Repository Modeling and Simulating Chemical Reactions”. In: 50 (2008), pp. 347–368 (cit. on pp. 18, 34, 50).

[33] Prabuddha B. et al. “Modeling cellulase kinetics on lignocellulosic substrates”. In: *Biotechnol. Adv.* 27.6 (2009), pp. 833–848 (cit. on pp. 18, 19).

[34] N.G. Van Kampen and N.G.Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Ltd, 1981 (cit. on p. 18).

[35] Franz Schwabl. *Statistische Mechanik*. 3., aktualisierte Aufl. Springer-Lehrbuch German Edition. Springer, 2006 (cit. on p. 18).

[36] Daniel T. Gillespie. “A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions”. In: *J. Comput. Phys.* 434 (1976), pp. 403–434 (cit. on pp. 18, 22, 32, 33, 35).

[37] T L Hill. *Thermodynamics of Small Systems*. unabridged. Dover Phoenix editions. Dover Publications, 2002 (cit. on p. 18).

[38] The Mathworks Inc. *SimBiology User’s Guide*. 2018 (cit. on p. 19).

[39] Jeffrey D Ullman John E. Hopcroft Rajeev Motwani. *Introduction to automata theory, languages, and computation*. Addison-Wesley, 2001 (cit. on p. 21).

[40] L. J. Gibson. “The hierarchical structure and mechanics of plant materials”. In: *J. R. Soc. Interface* 9.76 (2012), pp 2749–2766 (cit. on pp. 22, 46).

[41] Yang Mo Gu et al. “Effects of water content on ball milling pretreatment and the enzymatic digestibility of corn stover”. In: *Water-Energy Nexus* 1.1 (2018), pp. 61–65 (cit. on p. 23).

[42] Jianqiang Wei and Christian Meyer. “Degradation mechanisms of natural fiber in the matrix of cement composites”. In: *Cem. Concr. Res.* 73 (2015), pp. 1–16 (cit. on p. 23).

[43] Anwesha N Fernandes et al. “Nanostructure of cellulose microfibrils in spruce wood”. In: *Proc. Natl. Acad. Sci.* 108.47 (2011), pp. 18863–18864 (cit. on p. 24).

[44] Andrew J Griggs, Jonathan J Stickel, and James J Lischeske. “A mechanistic model for enzymatic saccharification of cellulose using continuous distribution kinetics I: depolymerization by EGI and CBHI.” In: *Biotechnol. Bioeng.* 109.3 (Mar. 2012), pp. 665–675 (cit. on p. 24).

[45] E. I. Saavedra Flores et al. “Mathematical modelling of the stochastic mechanical properties of wood and its extensibility at small scales”. In: *Appl. Math. Model.* 38.15-16 (2014), pp. 3958–3967 (cit. on p. 26).

[46] J.A. Meech. *Intelligent Applications in a Material World Select Papers from IPMM-2001*. Taylor & Francis, 2002 (cit. on p. 26).

[47] Ivanova A. N. Vol’pert A.I. “Mathematical models in chemical kinetics.” In: *Mathematical modelling: nonlinear differential equations of mathematical physics*. Samarskii, A.A., Kurdyumov, S. P., 1987 (cit. on p. 26).

[48] Santiago A. Serebrinsky. “Physical time scale in kinetic Monte Carlo simulations of continuous-time Markov chains”. In: *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.* 83.3 (2011), pp. 2010–2012 (cit. on pp. 26, 32, 33, 63).

[49] Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.8”. Nov. 2015 (cit. on pp. 27, 28).

[50] Fima C Klebner. *Introduction to stochastic calculus with applications*. 2nd ed. Imperial College Press, 2005 (cit. on p. 29).

[51] Philip E. Protter. *Stochastic Integration and Differential Equations*. Vol. 21. Springer, 2005 (cit. on p. 29).

[52] Daniel T Gillespie. “Exact Stochastic Simulation of Coupled Chemical Reactions”. In: *J. Comput. Phys.* 93555.1 (1977), pp. 2340–2361 (cit. on pp. 29, 33).

[53] Seongeun Yang et al. “Quantitative interpretation of the randomness in single enzyme turnover times”. In: *Biophys. J.* 101.3 (2011), pp. 519–524 (cit. on p. 30).

[54] Ross S. *Stochastic Processes*. Second Edi. 1995 (cit. on p. 30).

[55] S N Chiu et al. *Stochastic Geometry and Its Applications*. Wiley Series in Probability and Statistics. Wiley, 2013 (cit. on p. 30).

[56] J. Goutsias and G. Jenkinson. “Markovian dynamics on complex reaction networks”. In: *Phys. Rep.* 529.2 (2013), pp. 199–264 (cit. on p. 31).

[57] Christian H Weiss. *An Introduction to Discrete-Valued Time Series*. 1st ed. Wiley, 2018 (cit. on pp. 31, 36, 97).

[58] H Bisswanger. *Enzyme Kinetics: Principles and Methods*. Wiley, 2008 (cit. on pp. 31, 43).

[59] John W Dingee and A Brad Anton. “A new perturbation solution to the Michaelis-Menten problem”. In: *AIChE J.* 54.5 (2008), pp. 1344–1357 (cit. on pp. 31, 57, 70, 72, 92, 94).

[60] Daniel T. Gillespie. “A rigorous derivation of the chemical master equation”. In: 188 (1992), pp. 404–425 (cit. on p. 31).

[61] Carl Jason Morton-Firth and Dennis Bray. “Predicting Temporal Fluctuations in an Intracellular Signalling Pathway”. In: *J. theor. Biol* 192.January (1998), pp. 117–128 (cit. on p. 32).

[62] D. T. Gillespie. “Approximate accelerated stochastic simulation of chemically reacting systems”. In: *J. Chem. Phys.* 115.4 (2001), pp. 1716–1733 (cit. on p. 32).

[63] T de Donder and P Van Rysselberghe. *Thermodynamic theory of affinity*. Thermodynamic Theory of Affinity Bd. 1. Stanford university press, 1936 (cit. on p. 32).

[64] Analysing Probability. “21 Arguments Against Propensity Analyses of”. In: *Interpret. A J. Bible Theol.* 1975 (1996), pp. 1–36 (cit. on p. 32).

[65] Jan T A Koster. “Marginalizing and Conditioning in Graphical Models”. In: *Bernoulli* 8.6 (2002), pp. 817–840 (cit. on p. 32).

[66] Peter J Green. “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination”. In: *Biometrika* 82.4 (1995), pp. 711–732 (cit. on p. 32).

[67] Brian D. O. Anderson. “The Realization Problem for Hidden Markov Models”. In: *Math. Control. Signals, Syst.* 12 (1999), pp. 80–120 (cit. on p. 32).

[68] Sung Nok Chiu et al. *Stochastic Geometry and its Applications*. Wiley, 2013, p. 583 (cit. on pp. 33, 56).

[69] Anne-laure Boulesteix and Bernd Bischl. “Tunability : Importance of Hyperparameters of Machine Learning Algorithms”. In: (2018), pp. 1–22 (cit. on p. 33).

[70] I. Komarov and R. M. D’Souza. “Accelerating the Gillespie Exact Stochastic Simulation Algorithm Using Hybrid Parallel Execution on Graphics Processing Units”. In: *PLoS One* 7.11 (2012), pp. 1–9 (cit. on p. 33).

[71] Alexander Slepoy et al. “A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks”. In: *Journal of Chemical Physics* 205101.2008 (2010) (cit. on p. 33).

[72] Daniel T Gillespie. “Approximate accelerated stochastic simulation of chemically reacting systems Approximate accelerated stochastic simulation of chemically reacting systems”. In: *Journal of Chemical Physics* 1716.2001 (2006) (cit. on p. 33).

[73] Yang Cao, Dan Gillespie, and Linda Petzold. “Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems”. In: *Journal of Computational physics* 206 (2005), pp. 395–411 (cit. on p. 33).

[74] Hong Li and Linda R Petzold. “Logarithmic Direct Method for Discrete Stochastic Simulation of Chemically Reacting Systems”. In: *Tech. Rep.* (2006), pp. 1–11 (cit. on pp. 33, 35).

[75] Werner Sandmann. “Discrete-time stochastic modeling and simulation of biochemical networks”. In: *Comput. Biol. Chem. J.* 32 (2008), pp. 292–297 (cit. on pp. 33, 35, 97).

[76] Darren J. Wilkinson. *Stochastic Modelling for Systems Biology*. 2. CRC Press, 2012 (cit. on p. 34).

[77] Dieter Baum. *Grundlagen der Warteschlangentheorie*. 2013 (cit. on p. 34).

[78] Roberta Paroli, Giovanna Redaelli, and Luigi Spezia. "Poisson Hidden Markov Models for Time Series of Overdispersed Insurance Counts". In: *Casuality Actuar. Soc.* 1994 (2000), pp. 461–472 (cit. on p. 34).

[79] Anna Aksamit et al. "Predictable representation property for progressive enlargements of a Poisson filtration". In: *HAL* (2015) (cit. on p. 34).

[80] Randy Louis Ribler. "Visualizing Categorical Time Series Data with Applications to Computer and Communications Network Traces". PhD thesis. Blacksburg, VA, USA, 1997 (cit. on p. 36).

[81] Gerald Van Belle. *Statistical Rules of Thumb*. 2nd ed. Seattle, WA: Wiley, 2008 (cit. on p. 40).

[82] Robert A Parker and Nancy G Berman. "Sample Size: More Than Calculations". In: *Am. Stat.* 57.3 (2003), pp. 166–170 (cit. on p. 40).

[83] Meyer B. Jackson. *Molecular and Cellular Biophysics*. Cambridge University Press, 2006, p. 512 (cit. on p. 40).

[84] E L Cussler. *Fundamentals of Mass Transfer*. Vol. Second. Cambridge University Press, 1997, pp. 237–273 (cit. on p. 40).

[85] J Stahlberg, G Johansson, and G Pettersson. "A binding-site-deficient, catalytically active, core protein of endoglucanase III from the culture filtrate of Trichoderma reesei". In: *Eur J Biochem* 173.1 (1988), pp. 179–183 (cit. on pp. 40, 91).

[86] Harold P. Erickson. "Size and shape of protein molecules at the nanometer level determined by sedimentation, gel filtration, and electron microscopy". In: *Biol. Proced. Online* 11.1 (2009), pp. 32–51 (cit. on p. 40).

[87] D J Higham. "Modeling and simulating chemical reactions". In: *Psychother. Res.* 22.2 (2008), pp. 1753–1759 (cit. on p. 41).

[88] D Harel. "Statecharts: A visual formalism for complex systems". In: *Sci. Comput. Program.* 8 (1987), pp. 231–274 (cit. on p. 41).

[89] Mathworks. *Simulink: For Use with MATLAB; [user's Guide]*. 2017 (cit. on p. 41).

[90] Wojciech Skut. "Finite-State Machines for Mining Patterns in Very Large Text Repositories". In: *Proc. 2009 Conf. Finite-State Methods Nat. Lang. Process. Post-proceedings 7th Int. Work. FSMNLP 2008*. 2008, p. 23 (cit. on p. 41).

[91] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM - A library for support vector machines*. 2011 (cit. on p. 41).

[92] Dana Ron, Yoram Singer, and Naftali Tishby. "On the Learnability and Usage of Acyclic Probabilistic Finite Automata". In: *J. Comput. Syst. Sci.* 56.2 (1998), pp. 133–152 (cit. on p. 41).

[93] Borja Balle et al. "Spectral learning of weighted automata: A forward-backward perspective". In: *Mach. Learn.* 96.1-2 (2014), pp. 33–63 (cit. on p. 41).

[94] Frederick Jelinek. *Statistical Methods for Speech Recognition*. Bradford Book, 1998 (cit. on p. 41).

[95] J.E. Albus et al. *Syntactic pattern recognition, Applications*. Springer Berlin Heidelberg, 1977, pp. 1–278 (cit. on p. 41).

[96] Guy M Nicoletti. "Bio-molecular computing and memory systems - logic of certain enzyme kinetics". In: *Intell. Appl. a Mater. World*. Ed. by J.A. Meech. Taylor & Francis, 2001, pp. 241–249 (cit. on p. 42).

[97] Jeremy Gunawardena. "Time-scale separation–Michaelis and Menten's old idea, still bearing fruit." eng. In: *FEBS J.* 281.2 (Jan. 2014), pp. 473–488 (cit. on pp. 42, 45, 48, 96).

[98] Saikat Dutta and Kevin C.W. Wu. "Enzymatic breakdown of biomass: Enzyme active sites, immobilization, and biofuel production". In: *Green Chem.* 16.11 (2014), pp. 4615–4626 (cit. on p. 42).

[99] Deepak Kumar and Ganti S. Murthy. "Development and validation of a stochastic molecular model of cellulose hydrolysis by action of multiple cellulase enzymes". In: *Bioresour. Bioprocess.* 4.1 (2017), p. 54 (cit. on p. 42).

[100] Hans Bisswanger. "Enzyme assays". In: *Perspect. Sci.* 1.1-6 (2014), pp. 41–55 (cit. on p. 44).

[101] T. K. Harris and M. M. Keshwani. "Measurement of Enzyme Activity". In: *Methods Enzymol.* 1st ed. Vol. 463. C. Elsevier Inc., 2009, pp. 57–71 (cit. on p. 44).

[102] Tymoczko, John L Berg Jeremy M Gatto Gregory J Stryer Lubert. *Biochemistry*. W. H. Freeman and Company, 2015, p. 1053 (cit. on p. 46).

[103] D E Koshland. "Application of a Theory of Enzyme Specificity to Protein Synthesis". In: *Proc. Natl. Acad. Sci. U. S. A.* 44.2 (Feb. 1958), pp. 98–104 (cit. on p. 46).

[104] Joseph K Scott. "Stochastic Chemical Kinetics Introduction to Stochastic Chemical Kinetics". In: (2011), pp. 1–11 (cit. on p. 46).

[105] Glenn C. Rhoads. "Planar tilings by polyominoes, polyhexes, and polyiamonds". In: *J. Comput. Appl. Math.* 174.2 (2005), pp. 329–353 (cit. on p. 51).

[106] Joseph O'Rourke Jacob E. Goodman. *Handbook of discrete and computational geometry* (cit. on p. 51).

[107] R. Grossmann, N. Kiryati, and R. Kimmel. "Computational surface flattening: a voxel-based approach". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 24.4 (Apr. 2002), pp. 433–441 (cit. on p. 51).

[108] John Ashburner and Karl J. Friston. "Voxel-based morphometry - The methods". In: *Neuroimage* 11.6 I (2000), pp. 805–821 (cit. on p. 51).

[109] Chloe Hutton et al. "Voxel-based cortical thickness measurements in MRI". In: *Neuroimage* 40.4 (2008), pp. 1701–1710 (cit. on p. 51).

[110] James E. Mark. *Physical Properties of Polymers Handbook*. Vol. 199. Part_1. Springer, 1997, pp. 128–128 (cit. on p. 53).

[111] Yong-Hyun Lee and L. T. Fan. "Kinetic studies of enzymatic hydrolysis of insoluble cellulose: Analysis of the initial rates". In: *Biotechnology and Bioengineering* 24.11 (1982), pp. 2383–2406 (cit. on p. 53).

[112] L T Fan, Yong-hyun Lee, and H David. "Mechanism of the Enzymatic Hydrolysis of Cellulose : Effects of Major Structural Features of Cellulose on Enzymatic Hydrolysis". In: *Biotechnology and Bioengineering* XXII.1980 (1980) (cit. on p. 53).

[113] John F. Hughes et al. *Computer Graphics Principles and Practice*. Vol. 53. Addison-Wesley, 2014, pp. 1689–1699 (cit. on p. 55).

[114] Helena Pala, Manuel Mota, and Francisco Miguel Gama. "Enzymatic depolymerisation of cellulose". In: *Carbohydr. Polym.* 68.1 (2007), pp. 101–108 (cit. on p. 55).

[115] Makoto Y. et al. "Effects of Cellulose Crystallinity, Hemicellulose, and Lignin on the Enzymatic Hydrolysis of *Miscanthus sinensis* to Monosaccharides". In: *Biosci. Biotechnol. Biochem.* 72.3 (2008), pp. 805–810 (cit. on p. 56).

[116] Harold Jeffreys and Bertha Jeffreys. *Methods of Mathematical Physics*. 3rd ed. Cambridge Mathematical Library. Cambridge University Press, 1999 (cit. on p. 59).

[117] The Numerical Algorithms Group (NAG). *The NAG Library* (cit. on p. 59).

[118] Allen Y. Yang et al. "Fast 1-minimization algorithms for robust face recognition". In: *IEEE Trans. Image Process.* 22.8 (2013), pp. 3234–3246 (cit. on p. 59).

[119] George T. Doran. "There's a S.M.A.R.T. way to write managements's goals and objectives." In: *Management Review* 70.11 (1981), pp. 35–36 (cit. on p. 59).

[120] W. L. Price. "Global Optimization by Controlled Random Search". In: *J. Optim. Theory Appl.* 40.3 (1983), pp. 333–348 (cit. on p. 59).

[121] Andrew R. Conn, Nicholas I. M. Gould, and Philippe Toint. "A Globally Convergent Augmented Lagrangian Algorithm for Optimization with General Constraints and Simple Bounds". In: *SIAM J. Numer. Anal.* 28.2 (1991), pp. 545–572 (cit. on p. 59).

[122] Charles Audet and J. E. Dennis. "Analysis of Generalized Pattern Searches". In: *SIAM J. Optim.* 13.3 (2002), pp. 889–903 (cit. on p. 59).

[123] Konstantinos Parsopoulos and Michael Vrahatis. *Particle Swarm Optimization Method for Constrained Optimization Problem*. Vol. 76. IOS Press, Jan. 2002, pp. 214–220 (cit. on p. 59).

[124] S Kirkpatrick, C D Gelatt, and M P Vecch. "Optimization by Simulated Annealing". In: *Science* 220.4598 (2007), pp. 671–680 (cit. on p. 59).

[125] Shaojun Li and Fei Li. "Alopex-based evolutionary algorithm and its application to reaction kinetic parameter estimation". In: *Comput. Ind. Eng.* 60.2 (2011), pp. 341–348 (cit. on p. 59).

[126] Adam D. Bull. “Convergence rates of efficient global optimization algorithms”. In: *Journal of Machine Learning Research* 26.12 (2011), pp. 1–30 (cit. on p. 59).

[127] H.-M. Gutmann. “A Radial Basis Function Method for Global Optimization”. In: *J. Glob. Optim.* 19.3 (2001), pp. 201–227 (cit. on p. 60).

[128] A. L. Custódio et al. “Direct Multisearch for Multiobjective Optimization”. In: *SIAM J. Optim.* 21.3 (2011), pp. 1109–1140 (cit. on p. 60).

[129] M Fleischer. “The Measure of Pareto Optima Applications to Multi-objective Metaheuristics BT - Evolutionary Multi-Criterion Optimization”. In: ed. by Carlos M Fonseca et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 519–533 (cit. on p. 60).

[130] Thomas Hanne. “Multi-Objective Optimization Using Evolutionary Algorithms : An Introduction”. In: *J. Heuristics* 6.3 (2000), pp. 1–24 (cit. on p. 60).

[131] Won Young Yang et al. *Applied Numerical Methods Using MATLAB®*. Wiley Interscience, 2005, pp. 71–76 (cit. on p. 60).

[132] Shuang Wu et al. “L1-Norm Batch Normalization for Efficient Training of Deep Neural Networks”. In: (2018), pp. 1–8 (cit. on p. 60).

[133] Reed M. B. and Simon. *Methods of mathematical physics. Functional analysis*. Academic Press, Inc., 1980 (cit. on p. 60).

[134] S Özcan M.; Ekmekci and Bayar A. “A note on the variation of the taxicab lengths under rotations”. In: *Pi Mu Epsil. J.* 11.7 (2002), pp. 381–384 (cit. on p. 60).

[135] Jean Jacod and Philip Protter. *Probability Essentials*. Universitext. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004 (cit. on p. 61).

[136] Chetan T. Goudar et al. “Progress curve analysis for enzyme and microbial kinetic reactions using explicit solutions based on the Lambert W function”. In: *J. Microbiol. Methods* 59.3 (2004), pp. 317–326 (cit. on p. 62).

[137] Chetan T. Goudar, Jagadeesh R. Sonnad, and Ronald G. Duggleby. “Parameter estimation using a direct solution of the integrated Michaelis-Menten equation”. In: *Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol.* 1429.2 (1999), pp. 377–383 (cit. on p. 62).

[138] Y Shraga and R P Tewarson. “Parameter estimation in underdetermined problems”. In: *Comput. Math. with Appl.* 20.4 (1990), pp. 325–337 (cit. on p. 66).

[139] Jirí Blank, Pavel Exner, and Miloslav Havlícek. *Hilbert Space Operators in Quantum Physics*. Theoretical and Mathematical Physics. Dordrecht: Springer Netherlands, 2008 (cit. on p. 66).

[140] Hermann Minkowski. *Geometrie der Zahlen*. Teubner, 1896 (cit. on p. 66).

[141] Stephen M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press: Cambridge, 1986 (cit. on p. 66).

[142] Mohammad H Firooz, Student Member, and Sumit Roy. “Network Tomography via Compressed Sensing”. In: (), pp. 1–6 (cit. on p. 66).

[143] David L Donoho. “For Most Large Underdetermined Systems of Equations , Approximates the Sparsest Near-Solution”. In: *Comm. Pure Appl. Math* 40698 (2004), pp. 1–21 (cit. on p. 66).

[144] L. Kovarik et al. “Implementing an accurate and rapid sparse sampling approach for low-dose atomic resolution STEM imaging”. In: *Appl. Phys. Lett.* 109.16 (2016), pp. 1–5 (cit. on p. 66).

[145] Andrea Massa, Paolo Rocca, and Giacomo Oliveri. “Compressive Sensing in Electromagnetics - A Review”. In: *IEEE Antennas and Propagation Magazine* 57 (2015), pp. 224–238 (cit. on p. 66).

[146] Eric Ziegel et al. “Numerical Recipes: The Art of Scientific Computing”. In: *Technometrics* 29.4 (Nov. 1987), p. 501 (cit. on p. 66).

[147] G Anandalingam and T L Friesz. “Hierarchical optimization: An introduction”. In: *Ann. Oper. Res.* 34.1 (1992), pp. 1–11 (cit. on p. 68).

[148] Seth E Levine et al. “A mechanistic model for rational design of optimal cellulase mixtures.” In: *Biotechnol. Bioeng.* 108.11 (Nov. 2011), pp. 2561–70 (cit. on pp. 68, 84, 85).

[149] József Medve et al. “Hydrolysis of microcrystalline cellulose by cellobiohydrolase I and endoglucanase II from Trichoderma reesei: Adsorption, sugar production pattern, and synergism of the enzymes”. In: *Biotechnol. Bioeng.* 59.5 (1998), pp. 621–634 (cit. on pp. 69, 89).

[150] D P Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995 (cit. on p. 69).

[151] Christopher M Bishop. *Pattern Recognition and Machine Learning*. 2006 (cit. on p. 69).

[152] Alice Zheng. *Evaluating Machine Learning Models - O’Reilly Media*. 2015 (cit. on p. 69).

[153] Olivier Bousquet. “Stability and Generalization”. In: 2 (2002), pp. 499–526 (cit. on p. 70).

[154] Steven H Strogatz. *Nonlinear Dynamics and Chaos*. Perseus Books (cit. on p. 70).

[155] Maya Mincheva and Marc R Roussel. “Graph-theoretic methods for the analysis of chemical and biochemical networks. I. Multistability and oscillations in ordinary differential equation models”. In: *J. Math. Biol.* 55.1 (June 2007), pp. 61–86 (cit. on p. 70).

[156] James Murdock. *Normal Forms and Unfoldings for Local Dynamical Systems*. Springer Monographs in Mathematics. New York, NY: Springer New York, 2003 (cit. on p. 72).

[157] H Bisswanger. *Enzymkinetik*. Wiley, 2000 (cit. on p. 73).

[158] Edward Leamer. “Sensitivity Analyses Would Help”. In: *Am. Econ. Rev.* 75.3 (1985), pp. 308–313 (cit. on p. 76).

[159] Edward Leamer. “Let’s Take the Con Out of Econometrics”. In: *Am. Econ. Rev.* 73.1 (1983), pp. 31–43 (cit. on p. 76).

[160] Andrea Saltelli et al. *Global Sensitivity Analysis. The Primer*. Chichester, UK: John Wiley & Sons, Ltd, Dec. 2007, pp. 237–275 (cit. on p. 76).

[161] S Banach. *Theory of Linear Operations* (cit. on p. 79).

[162] Stefan Rolewicz. *Functional Analysis and Control Theory: Linear Systems* (cit. on p. 79).

[163] Jonathan J Stickel and Andrew J Griggs. “Mathematical modeling of chain-end scission using continuous distribution kinetics”. In: *Chem. Eng. Sci.* 68.1 (2012), pp. 656–659 (cit. on p. 83).

[164] U. P. Agarwal, J. Y. Zhu, and Sally A. Ralph. “Enzymatic hydrolysis of loblolly pine: Effects of cellulose crystallinity and delignification”. In: *Holzforschung* 67.4 (2013), pp. 371–377 (cit. on pp. 83, 84).

[165] Hélène Billard et al. “Optimization of a synthetic mixture composed of major Trichoderma reesei enzymes for the hydrolysis of steam-exploded wheat straw”. In: *Biotechnol. Biofuels* 5.1 (2012), p. 9 (cit. on p. 84).

[166] J A Yabefa, Y Ocholi, and G F et.al Odubo. “Effect of temperature and changes in medium pH on enzymatic hydrolysis of β (1-4) glycosidic bond in orange mesocarp”. In: *J. Plant Sci. Res.* 4.4 (2014), pp. 21–24 (cit. on p. 88).

[167] Mahmoudreza Ovissipour et al. “The effect of enzymatic hydrolysis time and temperature on the properties of protein hydrolysates from Persian sturgeon (Acipenser persicus) viscera”. In: *Food Chem.* 115.1 (2009), pp. 238–242 (cit. on p. 88).

[168] L Bowski et al. “Kinetic modeling of the hydrolysis of sucrose by invertase”. In: *Biotechnol. Bioeng.* 13.5 (1971), pp. 641–656 (cit. on p. 89).

[169] Anne Grethe et al. “Enzyme processivity changes with the extent of recalcitrant polysaccharide degradation”. In: *FEBS Lett.* 588.24 (2014), pp. 4620–4624 (cit. on p. 89).

[170] Burcin Cem Arabacioglu. “Using fuzzy inference system for architectural space analysis”. In: *Appl. Soft Comput.* 10.3 (2010), pp. 926–937 (cit. on p. 89).

[171] Christopher M Bishop. “Pattern Recognition and Machine Learning Springer Mathematical notation Ni”. In: () (cit. on p. 89).

[172] Curtis T. Rueden et al. “ImageJ2: ImageJ for the next generation of scientific image data”. In: *BMC Bioinformatics* 18.1 (Dec. 2017), p. 529 (cit. on p. 89).

[173] Mark R. Nimlos et al. “Molecular modeling suggests induced fit of Family I carbohydrate-binding modules with a broken-chain cellulose surface”. In: *Protein Eng. Des. Sel.* 20.4 (2007), pp. 179–187 (cit. on p. 89).

[174] Sarina Bromberg and Ken A. Dill. *Molecular Driving Forces Statistical Thermodynamics in Biology, Chemistry, Physics and Nanoscience*. Vol. 2. Garland Science, 2011 (cit. on p. 89).

[175] Yuval Shoham, Raphael Lamed, and Edward A. Bayer. “The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides”. In: *Trends Microbiol.* 7.7 (1999), pp. 275–281 (cit. on p. 89).

[176] Rudolf Gesztelyi et al. “The Hill equation and the origin of quantitative pharmacology”. In: *Arch. Hist. Exact Sci.* 66.4 (2012), pp. 427–438 (cit. on p. 89).

[177] E L Cussler. *Diffusion Mass Transfer in Fluid Systems*. 3rd ed. Cambridge Series in Chemical Engineering. Cambridge University Press, 2009 (cit. on p. 91).

[178] Cullen G Van Slyke D. “The mode of action of urease and enzymes in general”. In: *J. Biol. Chem.* 19.2 (Oct. 1914), pp. 141–180 (cit. on p. 94).

[179] Mirkil H. Snell J. L. Thompson G. L. Kemeny J. G. *Finite mathematical structures*. Prentice-Hall, 1959 (cit. on p. 94).

[180] G H Mealy. “A method for synthesizing sequential circuits”. In: *Bell Syst. Tech. J.* 34.5 (Sept. 1955), pp. 1045–1079 (cit. on p. 94).

[181] Edward F Moore. “Gedanken Experiments on Sequential Machines”. In: *Autom. Stud.* Princeton U., 1956, pp. 129–153 (cit. on p. 94).

[182] Michael O Rabin and Dana Scott. “Finite Automata and Their Decision Problems”. In: *IBM J. Res. Dev.* 3.2 (1959), pp. 114–125 (cit. on p. 94).

[183] Jochen Seemann and Jürgen Wolff von Gudenberg. “UML- Unified Modeling Language”. In: *Informatik-Spektrum* 21.2 (Apr. 1998), pp. 89–90 (cit. on p. 94).

[184] Herbert Fleischner. “On the equivalence of mealy-type and moore-type automata and a relation between reducibility and moore-reducibility”. In: *J. Comput. Syst. Sci.* 14.1 (1977), pp. 1–16 (cit. on p. 95).

[185] Yu. T. Medvedev. *On the class of events representable in a finite automaton. Sequential Machines*. Moskau, 1956 (cit. on p. 95).

[186] Daniel T. Gillespie. “Chemical Langevin equation”. In: *J. Chem. Phys.* 113.1 (2000), pp. 297–306 (cit. on p. 97).

[187] Leonard E Baum and Ted Petrie. “Statistical Inference for Probabilistic Functions of Finite State Markov Chains”. In: *Ann. Math. Stat.* 37.6 (1966), pp. 1554–1563 (cit. on p. 97).

[188] Iain L MacDonald Walter Zucchini. *Hidden Markov Models for Time Series: An Introduction Using R*. 1st. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman and Hall/CRC, 2009 (cit. on p. 97).

[189] Inomzhon Mirzaev and Jeremy Gunawardena. “Laplacian Dynamics on General Graphs”. In: *Bull. Math. Biol.* 75.11 (2013), pp. 2118–2149 (cit. on p. 97).

[190] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. “Bayesian Optimization with Unknown Constraints”. In: (), pp. 1–14 (cit. on p. 98).

[191] Adam D Bull and M L Oct. “Convergence rates of efficient global optimization algorithms”. In: 26 (), pp. 1–30 (cit. on p. 98).

[192] Ziyu Wang et al. “Bayesian optimization in a billion dimensions via random embeddings”. In: *J. Artif. Intell. Res.* 55 (2016), pp. 361–367 (cit. on p. 98).

[193] Venkataramana Ajjarapu. *Computational Techniques for Voltage Stability Assessment and Control*. Springer, 2007 (cit. on p. 98).

[194] Frank Hutter, Holger H Hoos, and Kevin Leyton-brown. “Datos Básicos de México - Edafología”. In: (2006), p. 2006 (cit. on p. 98).

[195] By Jasper Snoek, Hugo Larochelle, and Ryan P Adams. “Practical Bayesian Optimization of Machine Learning Algorithms”. In: (2001), pp. 1–12 (cit. on p. 98).

[196] Günther Schrack and Mark Choit. “Optimized relative step size random searches”. In: *Math. Program.* 10.1 (1976), pp. 230–244 (cit. on p. 98).

[197] Michael A. Schumer and Kenneth Steiglitz. “Adaptive Step Size Random Search”. In: *IEEE Trans. Automat. Contr.* 13.3 (June 1968), pp. 270–276 (cit. on p. 98).

[198] L. A. Rastrigin. “About Convergence of Random Search Method in Extremal Control of Multi-Parameter Systems”. In: *Avtomat. i Telemekh* 24.1 (1963), pp. 1467–1473 (cit. on p. 98).

[199] X. S. Wang et al. “Oblivious Data Structures”. In: *Proc. 2014 ACM SIGSAC Conf. Comput. Commun. Secur. - CCS '14* (2014), pp. 215–226 (cit. on p. 99).

[200] Thomas H. Cormen, Charles E. Leiserson Ronald L. Rivest Clifford Stein. *Introduction to Algorithms*. The MIT Press, 2009 (cit. on p. 99).

[201] Haskell B Curry. *Foundations of Mathematical Logic*. 2 Revised. Dover Publications, 2010 (cit. on p. 99).

[202] S C Kleene. *Introduction to Metamathematics*. 7th ed. Bibliotheca Mathematica. North Holland, 1980 (cit. on p. 99).

[203] François Panneton, Pierre L'ecuyer, and Makoto Matsumoto. “Improved long-period generators based on linear recurrences modulo 2”. In: *ACM Trans. Math. Softw.* 32.1 (2006), pp. 1–16 (cit. on p. 100).

[204] George Marsaglia. “Xorshift RNGs”. In: *J. Stat. Softw.* 8.14 (2003), pp. 1–6 (cit. on p. 100).

[205] Makoto Matsumoto and Takuji Nishimura. “A 623-dimensionally equidistributed uniform psudorandom number generator”. In: *ACM Trans. Model. Comput. Simulations* (1998) (cit. on p. 100).

[206] Eric W. Weisstein. *Sphere Point Picking*. (Cit. on p. 100).

[207] George Marsaglia. “Choosing a Point from the Surface of a Sphere”. In: *Ann. Math. Stat.* 43.2 (1972), pp. 645–646 (cit. on p. 100).

Acknowledgements

First and foremost, I would like to thank my PhD supervisor Prof. Dr. Peter Lenz for fruitful discussions and much helpful advice during my work. His ideas inspired me during the most Challenging times and his support was essential to the completion of this project. Thank you for this opportunity which has been a thrilling and enriching experience for me.

I am grateful to my colleagues David Geisel and Karl Kraft for our daily discussions, an amazing office atmosphere, and finally for helping me out with last minute corrections. Special thanks go to Dr. Myroslav Zapukhlyak who was patient enough to bear with me over the full project timespan of almost five years. His guidance led me through difficult times and was especially valuable to me, as I could rely upon him in every situation. Thank you for your feedback and calmness for all the recurring questions and endless discussions.

I would also like to thank my friends Dr. Philipp Schapotschnikow, Alexander Gouberman, Leo Khodos, Daniel Moesgen, Ana Sofia Ortega and Nicole Wochatz for discussions on the subject, grammatical corrections and many pleasant moments during my work.

Last but not least, I want to thank my parents for their continuous support in all my undertakings, endless love and encouragement. They stood by me all the time and gave a helping hand whenever I need it.

Alexander Orlov

scientific profile

university education

2019-03-01	Final doctoral examination and defense of this dissertation
Thesis	Theoretical Optimization of Enzymatic Biomass Processes
Supervisor	Prof. Dr. Peter Lenz, Complex Systems and Biophysics, Department of Physics, Philipps University of Marburg
Final Grade	cum laude
2010-09-13	Diploma in Physics
Thesis	Complex Ginzburg-Landau equation: Nonlinear global coupling and periodic 1:1 forcing
Supervisor	Prof. Dr. Katharina Krischer, Chemical Physics Beyond Equilibrium, Department of Physics, Technical University of Munich
Final Grade	good

scientific publications

[1] Vladimir García-Morales, Alexander Orlov, and Katharina Krischer. Subharmonic phase clusters in the complex ginzburg-landau equation with nonlinear global coupling. *Phys. Rev. E*, 82(6):065202, Dec 2010.