Computational Mass Spectrometry of Linear Binary Synthetic Copolymers

Dissertation
zur Erlangung des akademischen Grades
doctor rerum naturalium (Dr. rer. nat.)
vorgelegt dem Rat Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

von Dipl.-Bioinf. Martin S. Engler
geboren am 05.01.1984 in Gera

Gutacher

1. Prof. Dr. Sebastian Böcker, Friedrich-Schiller-Universität Jena
2. Prof. Dr. Gunnar W. Klau, Heinrich-Heine-Universität Düsseldorf
3. Prof. Dr. Peter Dittrich, Friedrich-Schiller-Universität Jena

Tag der öffentlichen Verteidigung: 04.05.2018

Abstract

The accurate characterization of synthetic polymer sequences represents a major challenge in polymer science. In this thesis, we present a computational approach to sequencing copolymers from mass spectrometry data, which enables the abundances of all sequences in a measured copolymer sample to be quantified.

The workflow presented in this thesis can be divided into two steps. The first step in our workflow is transforming mass spectra into copolymer fingerprints.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is frequently used for the characterization of copolymer samples. We present a method for computing copolymer fingerprints from mass spectra of the copolymer. Our method is based on linear programming and is capable of automatically resolving overlapping isotopes and isobaric ions. Using measured and simulated spectra, we demonstrate that our method is well suited for analyzing complex copolymer MS spectra.

Peak intensities in MALDI spectra are influenced by mass discrimination, *i.e.* mass- and composition-dependent ionization. We demonstrate a computational method to correct the abundance bias caused by the mass discrimination. We demonstrate our method using measured co- and homopolymers. First, the method is applied to homopolymer spectra. Subsequently, the copolymer fingerprint is computed from copolymer MALDI spectra and the correcting function applied. We find that the changes in the composition are plausible, indicating that the correction of copolymer abundances was reasonable. Our computational method may potentially help to avoid erroneous conclusions when analyzing copolymer MS spectra.

The second step in our workflow is interpreting the computed copolymer fingerprints using a new copolymerization model.

For many years copolymerization has been studied using mathematical and statistical models. We present new Markov chain models for copolymerization kinetics: The Bernoulli and Geometric models. They model copolymer synthesis as a random process and are based on a basic reaction scheme. In contrast to previous Markov chain approaches to copolymerization, both models take variable chain lengths and time-dependent monomer probabilities into account and allow the computation of sequence likelihoods and copolymer fingerprints. We compare both models against

Monte-Carlo simulations. We find that computing the models is fast and memory efficient.

Then, we focus on the Geometric copolymerization model with reactivity parameters and investigate its practicality. First, several approaches to identify the optimal model parameters from observed copolymer fingerprints are evaluated using Monte-Carlo simulated data. Directly optimizing the parameters is robust against noise but has impractically long running times. A compromise between robustness and running time is found by exploiting the relationship between monomer concentrations calculated by ordinary differential equations and the Geometric model. Second, we investigate the applicability of the model to copolymerizations beyond living polymerization and show that the model is useful for copolymerizations involving termination and depropagation reactions. We then compute several copolymer statistics using the Geometric model and compared them to the statistics obtained by counting in the copolymer chains computed by Monte-Carlo simulations.

Last but not least, we present our software framework COCONUT, which implements all algorithms presented in this thesis. Our software is freely available and provides a graphical user interface. COCONUT represents a step towards comprehensive computational support in polymer science.

Zusammenfassung

Die Analyse synthetischer Polymersequenzen ist eine große Herausforderung in der Polymerforschung. Diese Dissertation stellt einen neuen computergestützten Ansatz zur Polymersequenzierung vor, welcher die Quantifizierung aller Sequenzen eines gemessenen Copolymers ermöglicht.

Der präsentierte Ansatz kann in zwei Schritte eingeteilt werden. Der erste Schritt ist die Transformation von Massenspektren zu Copolymer-Fingerprints.

Matrix-assisted Laser Desorption / Ionization Time-of-Flight Massenspektrometrie (MALDI-TOF MS) ist eine gängige Methode zur Analyse von Copolymeren. Wir stellen eine neue Method zur Berechnung von Copolymer-Fingerprints aus Copolymer-Massenspektren vor, die auf Linearer Programmierung basiert und überlappende Isotopenmuster und Isobare aufklären kann. Anhand gemessener und simulierter Spektren zeigen wir, dass unsere Methode für komplexe Copolymer-Massenspektren geeignet ist.

Peak-Intensitäten in MALDI Spektren können durch differentielle Ionisierung, welche von der Masse und Zusammensetzung der Ionen abhängt, beeinträchtigt werden. Wir stellen eine computergestützte Methode zur Korrektur dieser Abweichung vor und wenden sie zunächst auf Homopolymerspektren an. Danach berechnen wir Fingerprints aus Copolymerspektren und wenden darauf die Korrekturfunktion an. Die errechneten Änderungen weisen auf ein sinnvolle Korrektur hin. Unsere computergestützte Korrekturmethode soll zukünftig helfen, Fehler in der Analyse von Copolymer-Massenspektren zu vermeiden.

Der zweite Schritt unseres Ansatzes ist die Interpretation der berechneten Copolymer-Fingerprints mit Hilfe neuer Modelle.

Mathematische und statistische Modelle der Copolymerisierung gibt es seit vielen Jahren. Wir präsentieren neue Modelle für die Copolymer Kinetik: Die Bernoulli und Geometrischen Modelle, welche auf einem einfachen Reaktionsschema basierend die Copolymersynthese als stochastischen Prozess mit Hilfe von Markovketten modellieren. Im Gegensatz zu bisherigen Modellansätzen mittels Markovketten modellieren beide neuen Modelle variable Längen der Polymerketten sowie zeitabhängige Wahrscheinlichkeiten der Monomere und ermöglichen die Berechnung von Sequenzwahrscheinlichkeiten und Copolymer-Fingerprints. Wir evaluaieren beide Modelle

mittels Monte-Carlo Simulationen. Die Berechnung der Modelle ist schnell und benötigt wenig Arbeitsspeicher.

Danach konzentrieren wir uns auf die – gemäß unserer Evaluation – beste Modellvariante: das Geometrische Copolymerisierungsmodell mit differenziellen Reaktionswahrscheinlichkeitsparametern. Zunächst evaluieren wir mittels Monte-Carlo Simulationen verschiedene Methoden zur Schätzung der optimalen Modellparameter aus gemessenen Copolymer-Fingerprints. Die direkte Optimierung der Parameter ist robust gegenüber Rauschen hat jedoch unpraktibel lange Laufzeiten. Wir finden einen Kompromiss zwischen Robustheit und Laufzeit durch Nutzung der Beziehung zwischen dem Geo-metrischen Modell und mittels Differentialgleichungen berechneter Monomerkonzentrationen. Danach untersuchen wir die Anwendbarkeit des Modells auf Copolymerisierungen jenseits von „Lebender Polymerisierung" und zeigen, dass das Modell nützlich ist für Copolymerisierungen die Terminations- und Depropagationsreaktionen beinhalten. Anschließend zeigen wir, wie man mit Hilfe des Modells verschiedene Statisken berechnet und vergleichen diese zu den Werten die wir durch einfaches Zählen in den mit Monte-Carlo Simulationen berechneten Polymerketten erhalten.

Schlussendlich präsentieren wir unser Softwareframework COCONUT, welches alle Algorithmen dieser Arbeit beinhaltet. Unsere Software ist frei verfügbar und stellt eine graphische Benutzeroberfläche bereit. COCONUT ist ein wesentlicher Schritt in Richtung umfassender informatischer Unterstützung der Polymerwissenschaften.

Acknowledgements

First and foremost I want to thank my supervisor Sebastian Böcker for giving me the opportunity to write this thesis. He always had an open door and helped guiding my research into the right direction. He also always organized support from various sources for this financially uncertain project.

I thank Ulrich S. Schubert for initiating and supporting the project. I thank Sarah Crotty for her helpful input and for taming the MS machine as well as Markus J. Barthel and Christian Pietsch for synthesizing the copolymers. I thank Gabriel Vivó-Truyols for providing the SBR source code. I also thank Peter Dittrich for his temporary financial support.

A special thanks goes to my co-worker Kerstin Scheubert for her helpful input and initial guidance. I thank my various office mates Tim White, Kai Dührkop, Florian Rasche and Florian Sikora for the great office atmosphere we had. Special thanks goes to fellow caffeine addict Sascha Winter for our motorcycle adventures. I also thank the rest of our group members Franziska Hufsky, Purva Kulkarni, Markus Fleischauer, Marcus Ludwig, Bertram Vogel and Marvin Meusel for our enjoyable time together. Additionally, I appreciate Tim White, Franziska Hufsky, Kerstin Scheubert, Marcus Ludwig and Marvin Meusel for proofreading parts of this thesis. A special thanks goes to Kathrin Schowtka for always helping me navigate around the pitfalls of bureaucracy.

I thank my family and friends for their enduring support, especially my parents Hans and Monika Engler for providing my necessary education.

Finally, I would like to thank my lovely fiancée and future wife Nathalie Lukajewski for always believing in me and supporting me.

In memory of my faithful dog Lucky.

Preface

This thesis covers most of my research in computational mass spectrometry of synthetic binary copolymers. During this work, I was associated with the bioinformatics group of Professor Sebastian Böcker at the Friedrich University Jena. My research was financed by the university's basic funding, the "Computer Supported Research" project (Thüringer Ministerium für Bildung, Wissenschaft und Kultur, grant no. 12038-514), the "Coordination of Biological and Chemical IT Research Activities" project (European 7th Framework Programme, project no. 270371), and others.

As teamwork is the foundation of science, the results presented within this thesis have been achieved in close cooperation with my supervisor Sebastian Böcker, our collaborators Ulrich S. Schubert, Sarah Crotty, Markus J. Barthel, and Christian Pietsch, and last but not least my colleague Kerstin Scheubert.

I started researching this topic while working on my Diploma thesis [24] and afterwards continued to research computational mass spectrometry of synthetic binary copolymers, which culminated in this thesis.

The main results of this thesis are presented in chapters 5-7. They tell a continuous story of our approach to analyzing copolymer mass spectra: from transforming the spectra into copolymer fingerprints to new models for the copolymerization process to the resulting software for the end user, *i.e.* the experimental chemist.

For chapters 5 and 6, Ulrich S. Schubert, Sebastian Böcker, Sarah Crotty, and I designed the experimental setup. Markus J. Barthel and Christian Pietsch performed the copolymerizations, Sarah Crotty recorded the mass spectra.

Chapter 5 introduces copolymer fingerprints. Section 5.1 describes how to transform mass spectra into fingerprints and how to overcome two major issues of this transformation: isobaric and overlapping isotope patterns [25]. Section 5.2 presents an approach to a well-known issue in mass spectrometry: peaks in (mostly) higher mass ranges being less pronounced than they theoretically should be [26]. I developed the linear program to compute the fingerprints and resolve overlapping isotopes, Sebastian Böcker and I developed the solution for the isobars and the abundance correcting method. I implemented the algorithms and performed the computations. Sarah Crotty and I evaluated the results on experimental data, I evaluated the results on simulated data.

Chapter 6 presents new models for copolymerization. Section 6.1 introduces and evaluates two new models, the Bernoulli and Geometric models, each in two different versions, with and without taking reactivity ratios into account [27]. The basic Bernoulli model (without reactivity ratio parameters) was first described in my Diploma thesis [24]. Section 6.3 describes how to estimate the model parameters and explores the limitations of the models for different polymerization types [28]. Section 6.4 presents several algorithms to compute useful statistical properties from the models. At the moment, the results of Section 6.4 are not published, the manuscript is in preparation. Sebastian Böcker and Kerstin Scheubert developed the basic Bernoulli model, I developed the three other models and the algorithms for generating statistics. I implemented the algorithms, performed the computations, and evaluated the results.

Chapter 7 presents the COCONUT (Copolymer Composition Numbering Tool) program. I designed and implemented the software. COCONUT includes, besides some preprocessing methods, all of the algorithms presented in chapters 5 and 6. COCONUT was briefly presented in our first two publications [25, 26]; here it is discussed in more detail.

Not included in this thesis is the work of Sebastian Böcker, Kerstin Scheubert and myself in cooperation with H. Martin Bücker, Vlad Dumitrel, and Emil Slusanschi on Jacobians and Automatic Differentiation for the basic Bernoulli model. The resulting manuscript is in preparation.

As usual in scientific literature, I will use “we” for the remainder of this thesis. The reader may choose to interpret this as “the reader and I”, “my colleagues and I”, or “my collaborators and I”.

Contents

1. Introduction

Polymers are macromolecules composed of monomer repeating units, typically joined by covalent bonds. Naturally occurring biopolymers like DNA, RNA or proteins are essential for life and might even precede life as we know it today [32]. They have been thoroughly studied in biology and its related interdisciplinary research areas of molecular biology, biochemistry, structural biology and bioinformatics. Recently, the transition from studying individual objects to characterizing heterogeneous collections of molecules sparked several new '-omics' fields: genomics, proteomics, transcriptomics, etc.

Compared to biopolymers, which have existed for millions of years, the history of synthetic polymers is rather short; with groundwork in the early 19th century and first practical results and the beginning of its industrial production in the early 20th century [87]. Today, polymer materials such as PVC, nylon, polyethylene, or silicone are essential for a large proportion of modern industrial products. Modern polymer science knows a wide range of polymer classes and architectures with very diverse applications, including for example organic batteries [68], self-healing materials [114] or drug delivery systems [52].

This thesis focuses on linear binary copolymers: non-branching polymer chains with monomer units from two different monomer species. Synthesizing (co-)polymers is usually a random process, as generating a fully sequence-controlled polymer is a challenging task [53, 57, 119]. Generally, the resulting copolymers are distributed both in terms of chain lengths and monomer sequences, and inferring structure-property relationships is difficult. Altuntaş and Schubert [2] outlined the necessity of computational support for characterizing (co-)polymers and introduced the term "polymeromics". In this thesis, our task is to develop computational methods for characterizing such heterogeneous collections of copolymer molecules in order to facilitate the design of optimized and application-specific polymer materials.

Sequencing copolymers has been named one of the last Holy Grails in polymer characterization [2]. Just as peptide sequencing contributed to the rise of high-throughput methods in proteomics [105], copolymer sequencing could speed up the development of new polymer materials significantly. However, unlike a separated peptide sample with – ideally – multiple copies of one sequence, a copolymer sample is a heterogeneous collection of polymer molecules, and thus, sequences. Statistically speaking, a copolymer sample contains all possible copolymer sequences, although with infinitesimal probability for the majority of sequences. Therefore, we do not want to determine just one copolymer sequence, but quantify the abundances of all sequences contained in the sample.

1.1 Structure of this Thesis

In this thesis, we present a computational approach to sequencing copolymers from mass spectrometry data, which enables the abundances of all sequences in a measured copolymer sample to be quantified. This thesis covers our whole workflow from transforming the spectra into copolymer fingerprints to new models for the copolymerization process to the resulting software for the end user, *i.e.* the experimental chemist. This thesis focuses mainly on the computational aspects of the workflow and results.

In Chapters 2 and 3 we briefly introduce the background and concepts necessary for understanding this thesis. Chapter 4 describes the used datasets. As this thesis focuses on the computational aspects, please refer to the corresponding publications for more information regarding materials, polymerization procedures, or instrumentation [25–27].

Chapter 5 introduces copolymer fingerprints. Section 5.1 describes how to transform mass spectra into fingerprints and how to overcome two major issues of this transformation: isobaric and overlapping isotope patterns [25]. Section 5.2 presents an approach to a well-known issue in mass spectrometry: peaks in (mostly) higher mass ranges being less pronounced than they theoretically should be [26].

Chapter 6 presents new models for copolymerization. Section 6.1 introduces and evaluates two new models, the Bernoulli and Geometric models, each in two different versions – with and without taking reactivity ratios into account [27]. Section 6.3 describes how to estimate the model parameters and explores the limitations of the models for different polymerization types [28]. Section 6.4 presents several algorithms to compute useful statistical properties from the models.

Chapter 7 briefly presents a software application with a graphical user interface, which we developed simultaneously with our copolymer sequencing approach: COCONUT (Copolymer Composition Numbering Tool). Finally, in Chapter 8 we conclude this thesis, discussing the main results and providing an outlook on further research questions.

2. Chemical and Computational Background

In this chapter, we briefly introduce the very basic concepts this thesis is built on. Readers familiar with the matter at hand may skip to the next chapter, where we describe the key concepts necessary to understand this work and the current state of art in computational approaches to analyzing copolymerization.

Here, on the one hand we discuss the experimental background: copolymers and the instrumental setup. On the other hand, we also discuss the computational background: concepts in numerical optimization and the measures we used for evaluating the computational results.

2.1 Copolymers

Polymers are macromolecules composed of monomer repeating units, small molecules typically joined by covalent bonds. They can be categorized into homopolymers, containing monomer units from a single monomer type, and copolymers, containing monomer units from two or more different monomer types. This thesis focuses on linear binary copolymers: Non-branching polymer chains with monomer units from two different monomer species. In the following we will refer to the monomer species as A and B, which can be replaced by arbitrary monomer types. Copolymers can be classified by the distribution of monomers on the chain (Fig. 2.1), the chemical classes of the monomers, or the synthesis type.

Polymerization

A polymer synthesis may occur if the following three conditions are met: The monomers have to be bifunctional (double or triple bonds, aromatic rings or functional groups) to form the polymer chain, the free enthalpy needs to be lower for

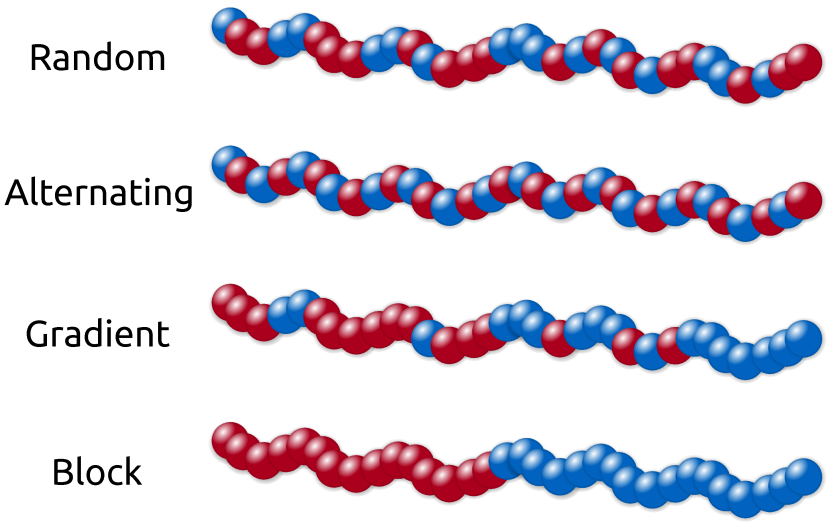


Figure 2.1 Different types of linear binary copolymers. In random copolymers, the probability of both monomers is uniformly distributed over the polymer chain length. In alternating copolymers, the monomers alternate between A and B. In gradient copolymers, the probability of monomer A gradually decreases with polymer length in a sigmoid or linear fashion. Block copolymers consist of two (or more) blocks, *i.e.* longer sequences of one monomer type.

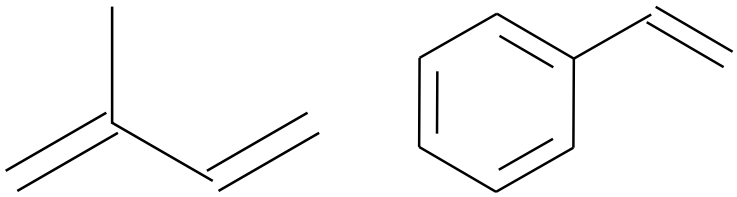
---

the polymer than for the monomers to form stable macromolecules, and the reaction rate has to be sufficiently high.

There are two major types of polymerization reactions, step-growth and living polymerization [89]. Step-growth polymerization is a controlled reaction. Naturally occurring polymers, such as DNA and RNA, are synthesized by a step-growth reaction. But also industrially produced polymers, such as polyethers, polyesters, and silicones can be synthesized with step-growth polymerization [89]. Such sequence-controlled polymers are not subject of this work.

Living polymerization is characterized by monomers attaching to the "living" reactive centers at the ends of the polymer chains, to which a free monomer can attach to in order to become the new reactive center. We can classify living polymerizations by the types of reactive centers: radical or ionic. Within a (free) radical polymerization, there are three reactions: initiation of the chain, propagation (elongation) of the chain by adding a monomer, and chain termination [89]. In controlled radical polymerization, there are additional (de-)activation reactions controlling the speed of initiation and propagation. Ionic (cationic or anionic) reactions usually correspond to the same basic reaction scheme of initiation, propagation, and termination. In some cases, it might be reversible or no termination occurs.

Another type of living polymerization is the ring-opening polymerization, which can be either radical or ionic. In ring-opening polymerizations, the reactive center inter-



(a) Isoprene (also known as 2-methyl-1,3-butadiene)

(b) Styrene (also known as ethenylbenzene)

Figure 2.2 Structural formulas of the isoprene and styrene monomers.

acts with cyclic monomers, which open their ring system to attach to the polymer chain.

Polystyrene and Polyisoprene

Polymers of isoprene and styrene are both naturally occurring compounds (Fig. 2.2). Isoprene was discovered to be the main component of organic rubber in 1860. Styrene was first isolated in 1839 [87]. However, it was not until 1920, that chemists realized that the properties of the investigated substances are caused by the compounds forming long polymer chains [87].

In this thesis, we use data obtained from copolymers of isoprene and styrene. The copolymers were synthesized by living anionic polymerization, a technique that is frequently used for polymeryizing other popular monomers such as ethylene oxide, allyl glycidyl ether, (meth)acrylate, etc. This polymerization technique produces well-defined polymers with a narrow distribution of polymer lengths, which is required for mass spectral analysis to ionize all polymer chains. Poly(isoprene styrene) copolymers have potential applications as porous membranes [83, 84], thin films [76], or micelles [16].

Copolymer Characterization

Polymers can be analyzed with various experimental methods. During the synthesis, monomer consumption can be measured by gas chromatography (GC) [58], high-performance liquid chromatography (HPLC) [7], ^{1}H nuclear magnetic resonance (NMR) spectroscopy [49], or size exclusion chromatography (SEC) analysis [77] to establish kinetic plots [48]. Simultaneous analysis of molar masses and chemical heterogeneities can be done with 2D-chromatographic methods such as HPLC (critical condition)-SEC coupling (2D-LC) [4, 29, 30, 72].

We are primarily interested in the distribution of polymer sequences. One of the oldest methods to identify the monomer content and parts of the sequence distribution of polymers is pyrolysis-gas chromatography [37, 104]. Modern methods for structural characterization of polymers are SEC [6, 97], HPLC [98], ^{1}H and ^{13}C

NMR spectroscopy [35], asymmetrical flow field-flow fractionation [70, 102], and viscosimetry [74]. However, all these methods have in common, that they are unable to characterize the full distribution of polymer sequences in a sample.

2.1.1 Copolymer Mass Spectrometry

Nowadays, mass spectrometry (MS) is frequently applied to characterize (co-) polymers [63], in particular using soft ionization techniques, such as *matrix-assisted laser desorption/ionization* (MALDI) [73] or *electrospray ionization* (ESI), in conjunction with *time-of-flight* (TOF) analyzers (Fig. 2.3).

MS techniques can highlight different features of polymers such as molecular weight distribution [36], or end-groups [17]. MS is frequently used to determine compositional drift [64], or the average composition [1, 47, 65, 67, 117], which then can be verified by other techniques, such as NMR.

MS Signal Processing

The output of the detector of a mass spectrometer is an analog signal. An analog/digital converter samples the analog signal at a certain rate and converts it to a digital signal, that the instrument software translates to a mass spectrum. Different types of mass spectrometers record different properties of the ions. For example, a TOF mass spectrometer records the time-of-flight of the ionized particles, which is roughly the square root of their mass over charge (m/z); thus, the instrument translates the time dimension of the signal. The result is a *raw spectrum* (Fig. 2.4): a list of m/z and intensity pairs. The mass is usually given in atomic mass units (u) or Dalton (Da), where 1 Da is $\frac{1}{12}$ of the mass of a ^{12}C atom (see below), while the intensity has no units.

Mass spectrometers are sensitive and small changes in the environment can result in large changes in the spectrum. For example, the TOF tube might expand or retract depending on the room temperature. This leads to uncertainties in the m/z dimension, which can be reduced by calibrating the mass spectrum by measuring a known standard. The calibration is either external – measuring the standard before the actual measurement – or internal – spiking the sample with the standard [118].

There may also be uncertainties in the intensity dimension. Noise caused by the detector can be reduced by smoothing the raw spectrum. Ubiquitous ions may lead to an amplified intensity in the lower m/z regions, which can be reduced by applying a baseline correction [5].

A raw mass spectrum is a sampled representation of a continuous signal. For computational processing of a spectrum, it is often necessary to convert it into a list of signal peaks with their m/z positions and intensities. *Peak picking* describes the process of transforming a raw spectrum into a peak list. An extensive range of methods is available for peak picking [78]. For example, peak positions can be determined by the weighted average of their datapoints or a wavelet approach [51]. The peak intensities can be determined by the maximum or area-under-curve of the peaks.

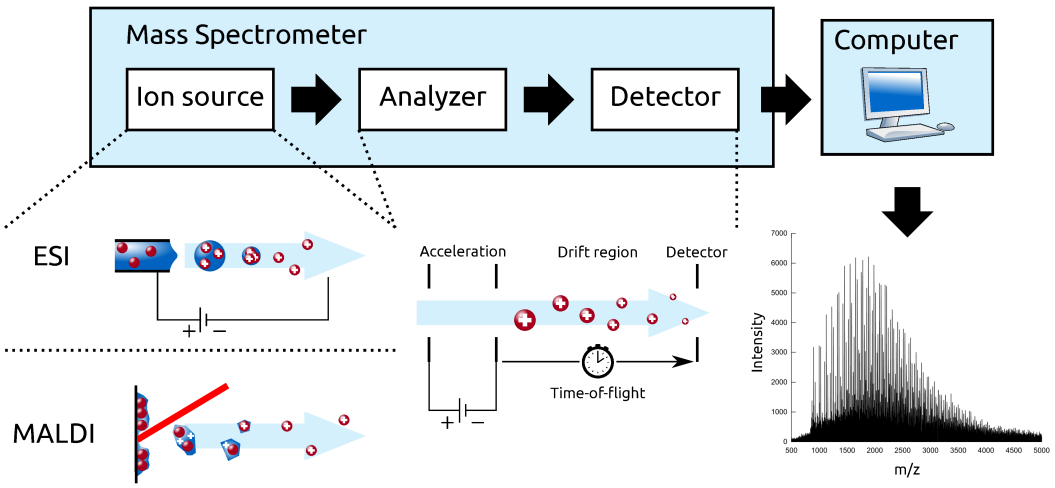


Figure 2.3 Schematical overview of a typical setup of a mass spectrometer in copolymer mass spectrometry: Ion source (ESI or MALDI), mass analyzer (TOF), and ion detector. An ESI ion source draws a liquid sample solution through a capillary tube and applies a high voltage to create charged droplets. Eventually the solution evaporates from the droplets until only gas phase ions remain. A MALDI ion source creates gas phase ions by firing a pulsed laser at a target plate with dried crystals of sample and matrix material. The matrix absorbs the laser energy, is desorbed from the plate and ionizes the sample. After ionization, the TOF analyzer accelerates the ions with an electric field and the detector measures their time-of-flight through the field-free drift region. Because lighter ions have higher velocities and fly faster than heavier ions, the ions get separated by mass. In a final step, the resulting change in current at the detector is translated by a computer to a spectrum of mass over charge (m/z).

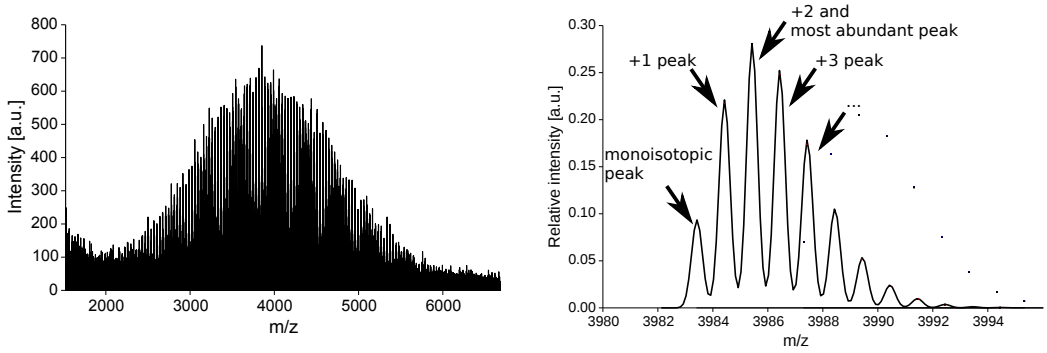


Figure 2.4 Left: Example of a raw copolymer MALDI-TOF mass spectrum with ubiquitous ions accumulated in the lower mass region. Right: Example of an isotopic pattern in a copolymer mass spectrum.

Isotopes and Isobars

For neutral atoms, the number of electrons equals the number of protons. However, the number of neutrons in the nucleus may vary. These variations are called *isotopes*. For example, a ^{13}C isotope has 6 protons and 7 neutrons. While most elements have only one known stable isotope, most the primary elements of organic chemistry – C, H, N, O, P, and S – have more than one naturally occuring isotope (except phosphorus). The probability of encountering an isotope is determined by the empirical *isotopic distribution* or *isotopic pattern* of each element. The isotopic distribution of a molecule (Fig. 2.4) is determined by the isotopic distributions of its atoms.

The *monoisotopic mass* of an element is its most abundant isotope. In mass spectrometry, the corresponding peak is called the *monoisotopic peak*, subsequent peaks the +1 peak, +2 peak, *etc.* The monoisotopic mass of a molecule is the sum of the monoisotopic masses of its atoms. For large molecules, including polymers, this often does not correspond to the most abundant peak, since the probability of containing heavy isotopes increases with the number of atoms per molecule.

Isobars are different molecules, whose monoisotopic masses are so close, that they are indistinguishable with current mass spectrometers. This is a challenge for copolymer MS, where certain masses of monomer types A and B may lead to isobaric copolymers. Additionally, the numbers of atoms of isobaric copolymers are most often too similar to be able to confidently distinguish the isobars based on the isotopic patterns.

2.1.2 Polymer Characteristics

Biopolymers formed by natural processes, such as DNA or proteins, are typically *monodisperse* [14]. That is, all polymers of a sample have the same chain length. On the other hand, synthetic polymers synthesized by living polymerization are

usually *polydisperse* with a wide range of polymer chains [14]. A mass spectrum of a polymer sample provides the *molecular weight distribution* (MWD), which can be used to calculate several key measures.

For homopolymers, the *chain length distribution* (CLD) corresponds to the MWD. For copolymers, where generally the mass of monomer A does not equal the mass of monomer B, we can calculate the CLD by summing up all peaks that correspond to the same number of monomers. In the literature, several distributions for modeling the CLD can be found: most probable (Schulz-Flory), gamma, Poisson, or hypergeometric distributions [13, 34, 94]. All these distributions are related. On the one hand, for large chain lengths the most probable distribution approximates the gamma distribution, while the gamma and binomial distributions approximate the Poisson distribution for large chain lengths. On the other hand, the gamma and Poisson distributions are the limiting cases of the hypergeometric chain length distribution [94].

The *peak molecular mass* M_p is the most abundant mass in a polymer sample. Let M_i be a *molar mass* (mass of a substance divided by the amount of substance) and N_i be the *number of moles* (amount of substance given in *mol*) of all polymer chains with molar mass M_i. Then the *number average molar mass* M_n is arithmetic mean over all molar masses:

$$M_n = \frac{\sum_i N_i M_i}{\sum_i N_i} \tag{2.1}$$

The *weight average molar mass* M_w is defined as:

$$M_w = \frac{\sum_i N_i M_i^2}{\sum_i N_i M_i} \tag{2.2}$$

The *degree of polymerization* (DP) is the average number of monomeric units in a macromolecule. For homopolymers with monomer mass M_0, it is defined as:

$$\mathrm{DP} = \frac{M_n}{M_0} \tag{2.3}$$

The *polydispersity index* (PDI) is a measure of distribution of the molar masses and is defined as:

$$\mathrm{PDI} = \frac{M_w}{M_n} \tag{2.4}$$

The PDI is related to the variance of the distribution. Monodisperse polymers have PDI of one, polydisperse polymers have a PDI > 1. In general, the larger the PDI is, the broader is the MWD.

2.2 Numerical Optimization

In its most general form, an *optimization problem* can be expressed by the following set of (in)equations

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq b_i, \quad i = 1, \ldots, m\,, \end{aligned} \tag{2.5}$$

where x is the vector of *optimization variables*, $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the *objective function*, functions $f_i : \mathbb{R}^n \to \mathbb{R}$ are the *constraint functions*, and b is a vector of known coefficients. The goal is to find an *optimal vector* x^*, that minimizes the value of the objective function while satisfying all constraints. The constraint functions confine the search space for x^* to a *feasible region*. We implicitely assume, that the feasible region is not empty. On a side note, optimization problems can of course also be formulated as maximization problems.

Optimization problems can be classified by the mathematical properties of the objective and constraint functions. In the following, we briefly give a description of the optimization types used in this thesis.

Linear Programming

An optimization problem is a *linear program* (LP), if both the objective function f_0 and the constraint functions $f_1, \ldots, f_m$ are linear, that is if they satisfy

$$f_i(\alpha x + \beta y) = \alpha f(x) + \beta f(y) \tag{2.6}$$

for all $x, y \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$. For linear programs, we can express Eq. 2.5 in the canonical form

$$\begin{aligned} \text{minimize} \quad & c^T x \\ \text{subject to} \quad & Ax \leq b\,, \end{aligned} \tag{2.7}$$

where b is a vector of constants, c a vector of coefficients, and A is a matrix of coefficients. The feasible region of a linear program is a convex polytype in $\mathbb{R}^n$, defined by the intersection of the constraints (Fig. 2.5). If we want to solve the LP, we need to find a point in the polytope, where the objective function has the minimal value.

There are two competing approaches to solve an LP: The simplex and interior point methods, which walk along the edges or interior of the feasible region, respectively. The simplex method has exponential worst case complexity while the interior point method is polynomial [81]. However, both methods are fast in practice and modern LP solvers can easily handle instances with hundreds of variables and thousands of constraints [11].

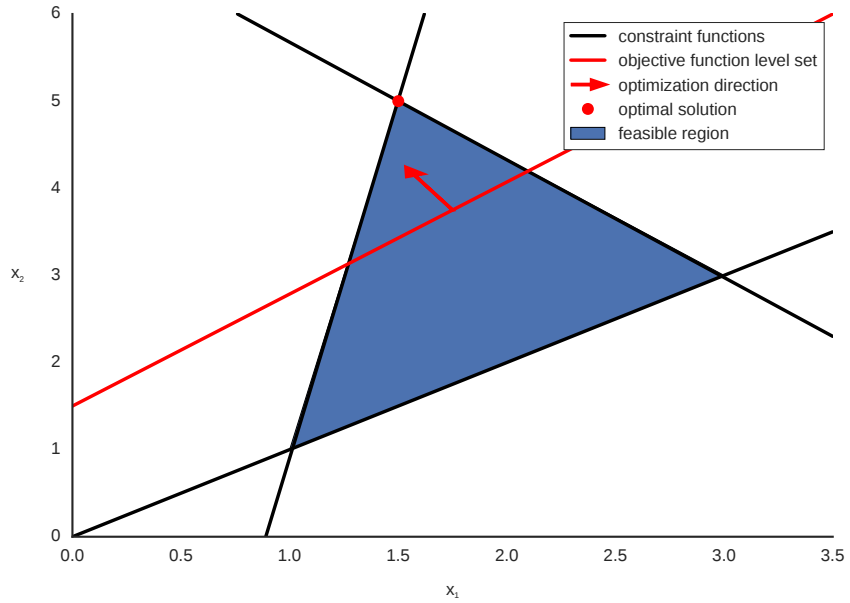


Figure 2.5 An example of an LP with two variables and three constraints. We see an example level set of the objective function, *i.e.* the set of all points with $f(x_1, x_2) = c$ for some constant $c \in \mathbb{R}$. To find the optimal solution x^*, we have to optimize in the direction indicated by the arrow.

General Purpose Optimization

While LPs can easily solve large optimization problems, they require the objective and constraints to be linear. But what if one or more of the objective and constraint functions are non-linear? In general, it is advisable to investigate if the functions are convex or if the problem can be formulated as a geometric program. Then, we can apply solvers for convex or geometric programs, respectively.

Unfortunately, most optimization problems are neither convex, geometric, nor linear [11]. However, the general optimization problem (Eq. 2.5) is surprisingly difficult to solve. In the recent years, the field of non-linear optimization has attracted many scientists and produced a large number of different algorithms [79]. The typical approach in practice is to simply try out a range of these methods on the problem in question [11].

Usually, non-linear optimizers are classified into local and global optimization [11, 81]. In this work we are primarily interested in finding global optima and we prefer to classify the optimizers into hillclimbers and evolutionary algorithms. Hillclimbers start with an arbitrary point and jump into the "correct" direction until they converge at a local minimum. Traditionally, they are local optimizers. However, hillclimbers can be used to find global optima, for example by introducing restarts with random starting points or random jumps [81]. Evolutionary algorithms use a population of points. The algorithms stochastically transform and then select points based

on a fitness score determined by the objective function [79]. Usually, evolutionary algorithms are computationally more demanding than hillclimbers, but are better suited for finding the global optimum. However, both approaches are not able to give a guarantee for finding the global optimum.

This work uses the algorithms implemented in the Optimization Algorithm Toolkit[1] [15] and Apache Math Commons 3.2 library[2].

2.3 Evaluation Criteria

In this section we briefly introduce the different measures used for evaluating the results in this thesis. In this work, there are two main evaluation types, checking matrices for similarity and evaluating the likelihood of a model.

Matrix Comparison

To compute the similarity of two matrices M and M' of size $n \times m$. Let $\overline{M}$ be the mean of a matrix M. Then, the *Pearson correlation coefficient* r is:

$$r = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(M_{i,j} - \overline{M})(M'_{i,j} - \overline{M'})}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(M_{i,j} - \overline{M})^2}\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(M'_{i,j} - \overline{M'})^2}} \tag{2.8}$$

If we think of the pairs $(M_{i,j}, M'_{i,j})$ as Cartesian coordinates, then a correlation coefficient of $r = 1$ means all points $(M_{i,j}, M'_{i,j})$ lie on a line with increasing slope, $r = -1$ means all points $(M_{i,j}, M'_{i,j})$ lie on a line with decreasing slope, while $r = 0$ indicates that there is no linear correlation between $M_{i,j}$ and $M'_{i,j}$. Unfortunately, having a perfect coefficient $r = 1$ does not imply that the matrices are equal. Given $M = M'$, then all our points lie on the angle bisecting line, but the coefficient does not capture the deviation from the angle bisector. Therefore, the Pearson correlation coefficient is a good measure to determine random errors, but not a linear bias.

To determine both error types, we resort to computing the distance between two matrices and use the *normalized root mean square error* ($NRMSE$). The $NRMSE$ of the matrices M and M' is given by

$$NRMSE(M, M') = 100 \cdot \frac{\sqrt{\frac{1}{n \cdot m}||M - M'||_2^2}}{\max(M')}, \tag{2.9}$$

where $||M - M'||_2$ is the $\mathcal{L}^2$-norm of $M - M'$.

[1] `https://sourceforge.net/projects/optalgtoolkit/`
[2] `http://commons.apache.org/proper/commons-math/`

Model Likelihood

We are given a large random sample of polymer chains D and a model, that is able to compute the likelihood of a single polymer chain $S \in D$. Then, we can evaluate the model by computing the likelihood of the whole dataset D. Let H be the *hypothesis*, that D was produced under the given assumptions and parameters of our model in question. Let $\mathbb{P}(S|H)$ be the *likelihood* of a single S given the hypothesis H. Then, the *likelihood* of a dataset D is:

$$\mathbb{P}(D|H) = \prod_{S \in D} \mathbb{P}(S|H) \tag{2.10}$$

Usually, the likelihoods are small and the dataset is large. This calculation is numerically challenging and prone to numerical underflow. To avoid these issues, we calculate the *log likelihood* instead by:

$$\log \mathbb{P}(D|H) = \sum_{S \in D} \log \mathbb{P}(S|H) \tag{2.11}$$

To further evaluate a model, we can compare the log likelihood of the data under the model to the log likelihood under the null hypothesis H_0. In the null model, all positions in the polymer chain are independent random variables. For each position over all chains in the dataset, we determine the frequencies f_{A} and f_{B} of A and B, respectively. Let $P(s_i)$ be the likelihood of monomer i in chain S. Then, the log likelihood of a dataset, assuming the null model, is:

$$\log \mathbb{P}(D|H_0) = \sum_{S \in D} \log \mathbb{P}(S|H_0) = \sum_{S \in D} \sum_{i=1}^{|S|} \log \mathbb{P}(s_i) = \sum_{S \in D} \sum_{i=1}^{|S|} \log \begin{cases} f_{\mathsf{A}}, & \text{if } s_i = \mathsf{A} \\ f_{\mathsf{B}}, & \text{if } s_b = \mathsf{B} \end{cases} \tag{2.12}$$

The *log likelihood ratio* is defined as:

$$\log \frac{\mathbb{P}(D|H)}{\mathbb{P}(D|H_0)} = \log \mathbb{P}(D|H) - \log \mathbb{P}(D|H_0) \tag{2.13}$$

The log likelihood ratio is a "sanity check" for statistical models. If the ratio is below zero, we dismiss our hypothesis, and accept it, if the ratio is above zero.

Given several models and a dataset, we can compare the models by comparing their log likelihoods. If we are also interested in comparing against the null hypothesis, we can use log likelihood ratios.

3. Computational Approaches to Copolymerization

In this chapter, we introduce the key concepts necessary for understanding this work and discuss the current state of art in computational approaches to analyzing copolymerizations. First, we introduce copolymer fingerprints, a convenient two-dimensional representation of copolymer mass spectra. Second, we discuss the state of art in modeling the copolymerization process.

3.1 Copolymer Fingerprints

Copolymer mass spectra can be transformed to copolymer fingerprints [25, 41, 101, 108], which represent the two-dimensional distribution of all copolymer chains. A copolymer fingerprint shows the abundance of each possible combination of monomer counts. In this thesis, we focus on copolymer fingerprints of linear binary copolymers.

Wilczek-Vera et al. [109] introduced copolymer fingerprints. Fingerprints provide information about the copolymer architecture [44, 45], the distribution of block lengths in block copolymers [109–111, 113], or the reactivity ratio of the consumed monomers [46]. They have been used to study degradation [47] and MALDI matrix effects [93]. Copolymer fingerprints are related to the bivariate distribution of monomer ratio and degree of polymerization, which can be used to highlight compositional drift [62, 66, 117].

3.1.1 Notation

This work focuses on linear binary copolymers: linear chains of monomer repeating units from two different monomer species, usually joined by covalent bonds. Throughout this thesis, we will denote the *monomers* as A and B. In computer science terms, our objects of interest are strings over the alphabet $\Sigma = \{\mathsf{A}, \mathsf{B}\}$.

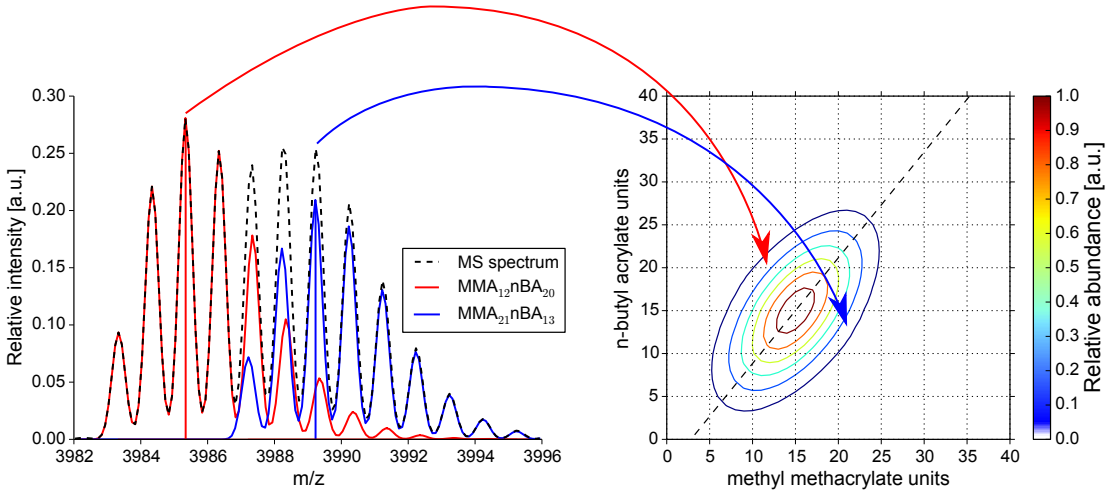


Figure 3.1 Example of the standard method to compute copolymer fingerprints using peak heights of the nearest observed peaks demonstrated with a poly(methyl methacrylate-*co*-*n*-butyl acrylate) (PMMA-*co*-P*n*BA) copolymer. We immediately see a major disadvantage: If we estimate the copolymer abundances using the peak height of the most abundant peak of each isotopic pattern, the overlap of both patterns leads to a false estimate of MMA_{21}-nBA_{13}.

In this work, a *copolymer fingerprint* is a matrix M, where $M_{i,j}$ holds the abundance of monomer composition $\mathsf{A}_i\mathsf{B}_j$, *i.e.* all copolymers with i monomers A and j monomers B.

3.1.2 State of the Art in Computing Fingerprints

In the past, several papers have been published on the straightforward method of transforming copolymer mass spectra to fingerprints [44–47, 93, 109–111, 113]. The abundance of each entry in the fingerprint is assigned to the height of some measured peak, which is closest to the most abundant theoretical isotopic peak of this copolymer (Fig. 3.1). However, this approach has certain drawbacks: [101, 110] First, since peak shapes change with increasing mass, abundance of the molecule is not correlated to the peak height but to the area of the peak. Using peak heights is only beneficial for very high masses (above the masses reported in this thesis), where peak resolution becomes poorer. Second, overlapping isotopes of different copolymers may result in imprecise polymer abundance assignments (Fig. 3.1). And third, isobaric molecules may prohibit entirely to correctly resolve copolymer abundances (Fig. 3.2) [41].

Strip-Based Regression

Vivó-Truyols et al. [101] presented a regression method to determine copolymer fingerprints from a single MS measurement. The method fits peak curves to the

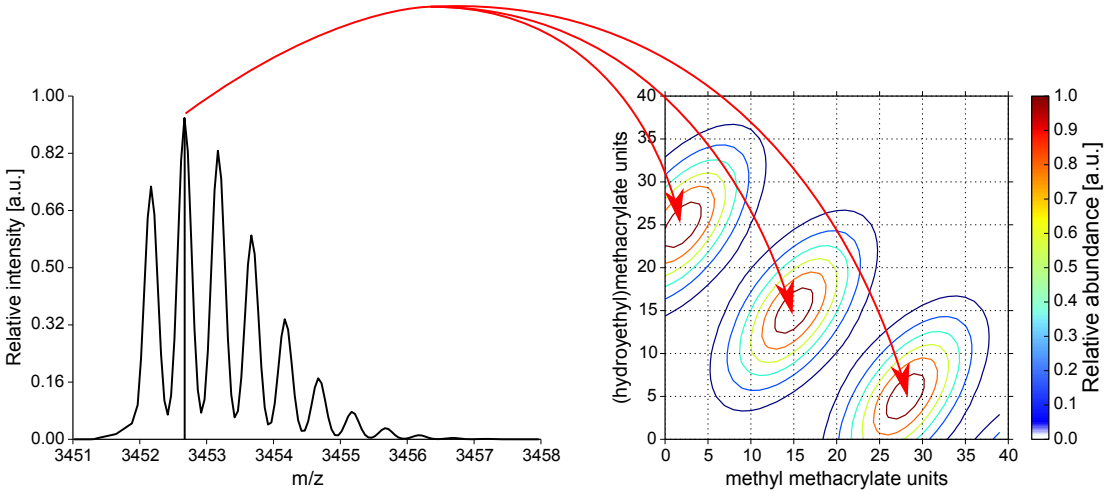


Figure 3.2 In this example, the standard method for computing copolymer fingerprints can not distinguish between three isobaric molecule candidates with almost equal monoisotopic masses. This results in several possible distributions of abundances in the copolymer fingerprints.

raw data, and can resolve overlapping isotopes. Because fitting the complete MS spectrum is computationally expensive, the method truncates the spectrum into strips. However, this truncation complicates quantification of isotopes on the strip borders.

MassChrom 2D

All methods mentioned above determine copolymer fingerprints from a single MS measurement, usually from MALDI-TOF MS data. A long-known issue of MALDI ionization is the non-linear relationship between MS signals and molecule abundances [42, 82, 85, 86]. Weidner et al.[106–108] presented MassChrom 2D, a method to determine copolymer fingerprints using liquid adsorption chromatography at critical conditions (LACCC) MS measurements. By using intensity information from chromatography, the authors evade the non-linear relationship between MS signals and molecule abundances. Fractions are separately analyzed and assembled *in silico* to form single copolymer fingerprints. Unfortunately, LACCC-MS is time-consuming, and critical conditions have to be known for at least one of the polymers.

3.2 Copolymerization Models

Several theoretical models for copolymerization were devised in the past, starting with Mayo and Lewis and their simple terminal model [59], which describes four propagation reactions, which append free monomers A and B to chain ends $\mathsf{A}^\bullet$ and $\mathsf{B}^\bullet$ according to the reaction rates k_{XY}, with $\mathsf{X}, \mathsf{Y} \in \{\mathsf{A}, \mathsf{B}\}$:

$$\begin{aligned} \mathrm{A}^\bullet + \mathrm{A} &\xrightarrow{k_{\mathsf{AA}}} \mathrm{AA}^\bullet \\ \mathrm{A}^\bullet + \mathrm{B} &\xrightarrow{k_{\mathsf{AB}}} \mathrm{AB}^\bullet \\ \mathrm{B}^\bullet + \mathrm{A} &\xrightarrow{k_{\mathsf{BA}}} \mathrm{BA}^\bullet \\ \mathrm{B}^\bullet + \mathrm{B} &\xrightarrow{k_{\mathsf{BB}}} \mathrm{BB}^\bullet \end{aligned} \tag{3.1}$$

Computational approaches to such a model can be categorized into three types: ordinary differential equations, Markov chains, and Monte-Carlo methods.

Ordinary Differential Equations

An ordinary differential equation (ODE) describes the change of one or more variables, in our case monomer concentrations [A] and [B], over time. Mayo and Lewis interpreted their reaction scheme as a set of ODEs, which are characterized by the reactivity ratios:

$$r_{\mathsf{A}} = \frac{k_{AA}}{k_{AB}} \tag{3.2}$$

$$r_{\mathsf{B}} = \frac{k_{BB}}{k_{BA}} \tag{3.3}$$

Using the reactivity ratios, they deduced the copolymer equation, which provides the theoretical change of composition, *i.e.* monomer ratio, of the copolymer at any time point during the synthesis:

$$\frac{d[\mathsf{A}]}{d[\mathsf{B}]} = \frac{[\mathsf{A}](r_{\mathsf{A}}[\mathsf{A}] + [\mathsf{B}])}{[\mathsf{B}]([\mathsf{A}] + r_{\mathsf{B}}[\mathsf{B}])} \tag{3.4}$$

This set of ODEs can be solved fast, but does not convey any information on the chain sequences. Kryven and Iedema advanced the ODE approach by applying population balance equations [50]. They showed the importance of recovering "distributions in a full form rather than averages, since average values may often be far from the most frequently occurring ones." [50, p. 305] They were able to extract simple sequence patterns, but not the full distribution of sequences.

Markov Models

A Markov chain is a stochastic process, where the next event depends only on the current and/or previous events. Several types of Markov chains are defined in literature. For our purposes, the simplest type is sufficient: Formally, a discrete, homogeneous Markov chain is a sequence of random variables $X_1, X_2, \ldots, X_t, \ldots$ defined at discrete time points $t \geq 1$ on the discrete probability space Ω with X_t :

$\Omega \to \mathbb{R}$. Let $x, y \in \Omega$, then the homogeneous property of the Markov chain is defined as:

$$P(y \to x; t) = P(X_{t+1} = x | X_t = y) \tag{3.5}$$

Informally, we can think of the Markov chain as an discrete automaton with the states Ω. At each time point t, the automaton randomly switches its state to a new or the same state. The state transition probability only depends on the current state.

The transformation of the traditional Mayo-Lewis model to a Markov chain is straightforward [13]. The transition probabilities can be simply deduced from the propagation reactions (Eq. 3.1):

$$\begin{aligned}
P(\mathsf{A}^\bullet \to \mathsf{A}^\bullet) &= \frac{k_{\mathsf{AA}}[\mathsf{A}^\bullet][\mathsf{A}]}{k_{\mathsf{AA}}[\mathsf{A}^\bullet][\mathsf{A}] + k_{\mathsf{AB}}[\mathsf{A}^\bullet][\mathsf{B}]} \\
P(\mathsf{A}^\bullet \to \mathsf{B}^\bullet) &= \frac{k_{\mathsf{AB}}[\mathsf{A}^\bullet][\mathsf{B}]}{k_{\mathsf{AA}}[\mathsf{A}^\bullet][\mathsf{A}] + k_{\mathsf{AB}}[\mathsf{A}^\bullet][\mathsf{B}]} \\
P(\mathsf{B}^\bullet \to \mathsf{A}^\bullet) &= \frac{k_{\mathsf{BA}}[\mathsf{B}^\bullet][\mathsf{A}]}{k_{\mathsf{BA}}[\mathsf{B}^\bullet][\mathsf{A}] + k_{\mathsf{BB}}[\mathsf{B}^\bullet][\mathsf{B}]} \\
P(\mathsf{B}^\bullet \to \mathsf{B}^\bullet) &= \frac{k_{\mathsf{BB}}[\mathsf{B}^\bullet][\mathsf{B}]}{k_{\mathsf{BA}}[\mathsf{B}^\bullet][\mathsf{A}] + k_{\mathsf{BB}}[\mathsf{B}^\bullet][\mathsf{B}]}
\end{aligned} \tag{3.6}$$

The resulting Markov chain is a simple model of the copolymerization. It can be used to compute the probability of a single copolymer chain, but not the distribution of all chains, as it does not take chain lengths nor time-steps without monomer additions into account.

Monte-Carlo Simulations

Classical probability theory uses probability density functions to describe a random process. But what about random processes, where the density function is computationally too expensive or even the true density function is unknown? One approach are Monte-Carlo simulations, which simply describes the process of repeated random sampling. The law of large numbers states that with increasing sample size the empirical distribution will likely converge to the true distribution.

Monte-Carlo methods can be used to simulate chemical reactions; Gillespie's algorithm [33] has been frequently used to simulate copolymerizations by randomly growing copolymer chains [12, 20, 60]. Several times Monte-Carlo simulations have been evaluated against experimental data [21, 23, 112] and it has been shown that Gillespie's algorithm can be used to compute copolymer fingerprints [22, 91, 99, 100].

Monte-Carlo simulations easily allow for more complex reactions schemes than the simple Mayo-Lewis model, for example by including initiation and termination reactions. However, Monte-Carlo simulations are time- and memory-intensive, in particular if an accurate representation of the distribution of copolymer chains is desired.

4. Datasets

In the following, all computations were performed in parallel on a compute cluster of four 2.4 GHz CPUs with 16 cores each and 6GB RAM per process.

4.1 Experimental Data

We briefly describe the data sets, concentrating on the computational focus of this thesis and omitting details of the chemical experiments. For more information regarding materials, polymerization procedures, or instrumentation please refer to the corresponding publications [25–27].

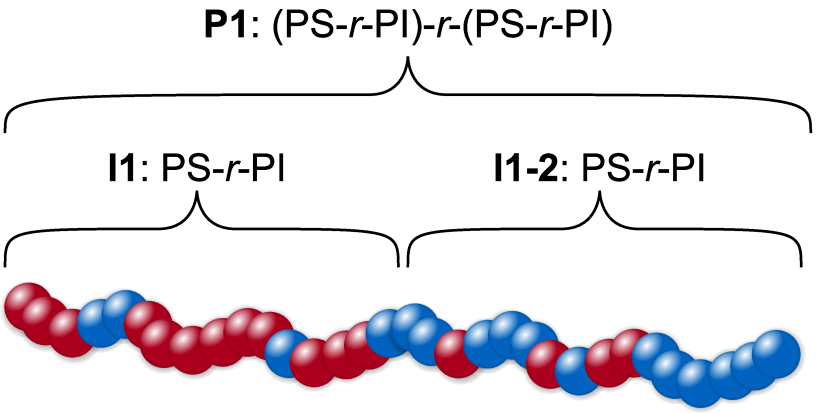


Figure 4.1 Schematic representation of the synthesized (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymer P1. P2 and P3 have the same architecture, but different PS to PI ratios.

4.1.1 (PS-r-PI)-r-(PS-r-PI) Copolymers

We synthesized three different random copolymers (Fig. 4.1), each consisting of two macromers with both a different ratio of styrene and isoprene (Tables 4.1 and 4.2).

	I1		I2		I3	
	PS	**PI**	**PS**	**PI**	**PS**	**PI**
Percent[%]	80	20	70	30	60	40
Degree of polymerization	19	7	17	11	14	15
Molar mass $[\,\mathbf{g \cdot mol^{-1}}\,]$	2000	500	1750	750	1500	1000
	I1-2		**I2-2**		**I3-2**	
	PS	**PI**	**PS**	**PI**	**PS**	**PI**
Percent[%]	20	80	30	70	40	60
Degree of polymerization	5	29	7	26	10	22
Molar mass $[\,\mathbf{g \cdot mol^{-1}}\,]$	500	2000	750	1750	1000	1500

Table 4.1 Summary of theoretical values of the first (I1 to I3) and second (I1-2 to I3-2) macromers.

	P1	**P2**	**P3**
PS	24	24	24
PI	36	37	37
Molar mass $[\,\mathbf{g \cdot mol^{-1}}\,]$	5,000	5,000	5,000

Table 4.2 Summary of theoretical values of the copolymers P1 to P3.

We measured the first poly(styrene-*rand*-isoprene) (PS-*r*-PI) macro-mers (I1 to I3) and the complete (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers (P1 to P3).

The first (I1 to I3) and second macromers (I1-2 to I3-2) are constituted of a random copolymer of styrene and isoprene. Theoretical molar masses of $5,000$ g mol^{-1} ($2,500$ g mol^{-1} for each macromer) were targeted for the copolymers P1 to P3. Differences between the theoretical and observed values for the DP in particular for isoprene can be explained by the difficult handling of the monomer, the related inaccurate added volume and the Ag cluster suppression in the MS spectra. All copolymers show PDI values lower than 1.1, indicating a living character of the polymerization.

4.1.2 PS and PI Homopolymer Mixtures

We synthesized polystyrene (PS) and polyisoprene (PI) homopolymers (Table 4.3). All homopolymers show PDI values lower than 1.1, indicating a living character

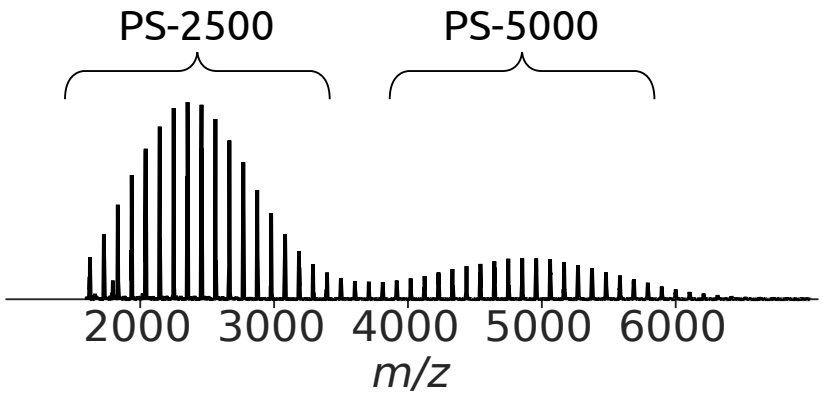


Figure 4.2 Exemplary MALDI-TOF MS of the PS-2500/PS-5000 homopolymer mixture.

	PS-2500	**PS-5000**	**PI-2500**	**PI-5000**
Degree of polymerization	19	48	29	73
Molar mass $[\,\mathbf{g \cdot mol^{-1}}\,]$	2,000	5,000	2,000	5,000

Table 4.3 Summary of theoretical values for PS and PI homopolymers.

of the polymerization. The M_n values of the homopolymers are near their theoretical molar masses. Equimolar mixtures of 2,500 $g \cdot mol^{-1}$ and 5,000 $g \cdot mol^{-1}$ homopolymers were prepared from PI and PS. Each homopolymer MS measurement was replicated three times (Figure 4.2 and Appendix Figure A.6). Additionally, we remeasured the copolymers P1 to P3.

4.1.3 Data Processing

MS data were processed using PolyTools 1.0 (Bruker Daltonics) and Data Explorer 4.0 (Applied Biosystems). The averaging of the homopolymer replicates was performed using in house built Groovy scripts. Spectral preprocessing, *i.e.* centroiding and baseline correction, and all other computations were performed using COCONUT [25].

4.2 Simulated Data

To assess the accuracy of our methods, we need to compare against a known ground truth. To this end, we use simulated mass spectra to evaluate our method for computing fingerprints (Section 5.1) and Monte-Carlo simulations to evaluate our copolymerization models (Chapter 6).

	$\mu_{\mathbf{A}}$	$\mu_{\mathbf{B}}$	$\sigma_{\mathbf{A}}$	$\sigma_{\mathbf{B}}$	ρ
1	11.0	9.0	2.4976131963448998	2.3899262596580195	-0.4501784953590977
2	6.0	12.0	2.532897843117028	4.349914949331601	0.40217536504947726
3	8.0	9.0	3.5328177424903933	2.8424627024892173	0.25925811791731745
4	19.0	13.0	5.192975793306746	4.12753685222484	0.2624095395236721
5	13.0	10.0	5.571620004044149	5.313466676925163	-0.05731275108645284

Table 4.4 Parameters of the simulated fingerprints used for evaluation. The five bivariate normal distributions were generated with randomly chosen parameters (means μ uniformly drawn from $[6, 22]$, variances σ uniformly drawn from $[2, 6]$, shape ρ uniformly drawn from $[-0.5, 0.5]$).

4.2.1 Simulated Mass Spectra

We simulated poly(methyl methacrylate-*co*-*n*-butyl acrylate) (PMMA-*co*-P*n*BA) and poly(methyl methacrylate-*co*-hydroxyethyl methacrylate) (PMMA-*co*-PHEMA) spectra as numerous overlapping isotopes and isobaric molecules appear in these copolymers. Although we can not simulate all aspects of the physical processes of an MS instrument, we tried to capture several fundamental aspects.

We start by generating a copolymer fingerprint. To this end, we use five bivariate normal distributions with randomly chosen parameters (Table 4.4). Given the copolymer fingerprint, we iterate over all monomer compositions: We add the appropriate end groups, and simulate the first 12 peaks of the isotope pattern [9], estimating both intensities and mean peak masses. We disturb each isotope peak by adding normally distributed noise with mean zero and variance $\frac{\sigma}{2}$ to the masses, and multiplying intensities by log-normal distributed random noise with mean zero and variance σ, with the noise parameter σ given below. For an isotope peak with mass m and intensity I, we add a Gaussian function with mean m, variance $\frac{1}{5}$, and height (multiplier) I to the simulated spectrum. To get a raw spectrum, we then sample the sum of Gaussians at sampling points with mass difference 0.1 Da. Finally, to simulate detector noise, this sampled (discretized) spectrum is again perturbed using multiplicative noise following a log-normal distribution with mean zero and variance $\frac{\sigma}{2}$.

The PMMA-*co*-P*n*BA mass spectra show a large number of overlapping isotope patterns, whereas the isotope patterns in the PMMA-*co*-PHEMA spectra have many possible isobaric molecule candidates. We used five noise levels σ with the values 0.0, 0.05, 0.1, 0.2, and 0.5. For each copolymer, all five copolymer fingerprints and all five noise levels, we simulated five mass spectra; resulting in 250 spectra in total.

Reaction type	LP	RLP	FRP	CRP
Initiation	×	×	×	×
Propagation	×	×	×	×
Depropagation		×		
Termination (Recomb. & Disprop.)			×	×
Initiator Decomposition			×	
(De-)Activation				×

Table 4.5 Overview of the modeled reactions types for the living polymerization (LP), reversible living polymerization (RLP), free radical polymerization (FRP), and controlled radical polymerization (CRP).

4.2.2 Monte-Carlo Simulations

Monte-Carlo Reaction Schemes

We performed Monte-Carlo simulations of different polymerization types (Table 4.5): Living polymerization (LP), reversible living polymerization (RLP), free radical polymerization (FRP), and controlled radical polymerization (CRP). For living polymerization, the following reaction scheme was used. An active center is donated as $X^\bullet$, and a polymer chain ending with X as $\sim$X, where X can be one of the monomers A or B, or initiator I. Two types of reactions, initiation and propagation reactions were modeled:

$$\left.\begin{aligned} \mathrm{I} + \mathrm{A} &\xrightarrow{k_{IA}} \sim\mathrm{A}^\bullet \\ \mathrm{I} + \mathrm{B} &\xrightarrow{k_{IB}} \sim\mathrm{B}^\bullet \end{aligned}\right\} \text{Initiation}$$

$$\left.\begin{aligned} \sim\mathrm{A}^\bullet + \mathrm{A} &\xrightarrow{k_{AA}} \sim\mathrm{A}^\bullet \\ \sim\mathrm{A}^\bullet + \mathrm{B} &\xrightarrow{k_{AB}} \sim\mathrm{B}^\bullet \\ \sim\mathrm{B}^\bullet + \mathrm{A} &\xrightarrow{k_{BA}} \sim\mathrm{A}^\bullet \\ \sim\mathrm{B}^\bullet + \mathrm{B} &\xrightarrow{k_{BB}} \sim\mathrm{B}^\bullet \end{aligned}\right\} \text{Propagation}$$

For reversible living polymerization, we use the initiation and propagation reactions of the living polymerization and additionally model depropagation reactions:

$$\left.\begin{aligned}
\sim\mathrm{IA}^\bullet &\xrightarrow{k^d_{IA}} \sim\mathrm{I}^\bullet + \mathrm{A} \\
\sim\mathrm{IB}^\bullet &\xrightarrow{k^d_{IB}} \sim\mathrm{I}^\bullet + \mathrm{B} \\
\sim\mathrm{AA}^\bullet &\xrightarrow{k^d_{AA}} \sim\mathrm{A}^\bullet + \mathrm{A} \\
\sim\mathrm{AB}^\bullet &\xrightarrow{k^d_{AB}} \sim\mathrm{A}^\bullet + \mathrm{B} \\
\sim\mathrm{BA}^\bullet &\xrightarrow{k^d_{BA}} \sim\mathrm{B}^\bullet + \mathrm{A} \\
\sim\mathrm{BB}^\bullet &\xrightarrow{k^d_{BB}} \sim\mathrm{B}^\bullet + \mathrm{B}
\end{aligned}\right\} \text{Depropagation}$$

For radical polymerization, we use the initiation and propagation reactions of the living polymerization and additionally model chain termination by recombination and disproportionation:

$$\left.\begin{aligned}
\sim\mathrm{A}^\bullet + \sim\mathrm{A}^\bullet &\xrightarrow{k^r_{AA}} \sim\mathrm{AA}\sim \\
\sim\mathrm{A}^\bullet + \sim\mathrm{B}^\bullet &\xrightarrow{k^r_{AB}} \sim\mathrm{AB}\sim \\
\sim\mathrm{B}^\bullet + \sim\mathrm{B}^\bullet &\xrightarrow{k^r_{BB}} \sim\mathrm{BB}\sim
\end{aligned}\right\} \text{Recombination}$$

$$\left.\begin{aligned}
\sim\mathrm{A}^\bullet + \sim\mathrm{A}^\bullet &\xrightarrow{k^{dp}_{AA}} \sim\mathrm{A} + \sim\mathrm{A} \\
\sim\mathrm{A}^\bullet + \sim\mathrm{B}^\bullet &\xrightarrow{k^{dp}_{AB}} \sim\mathrm{A} + \sim\mathrm{B} \\
\sim\mathrm{B}^\bullet + \sim\mathrm{B}^\bullet &\xrightarrow{k^{dp}_{BB}} \sim\mathrm{B} + \sim\mathrm{B}
\end{aligned}\right\} \text{Disproportionation}$$

For free radical polymerization, we use the initiation and propagation reactions of the living polymerization, chain termination by recombination and disproportionation, and the following additional initiation reaction to model a decomposing initiator complex:

$$\left.\mathrm{I}_2 \xrightarrow{k^{dec}} 2{\cdot}\mathrm{I}\right\} \text{Decomposition}$$

For controlled radical polymerization, we use the initiation and propagation reactions of the living polymerization, chain termination by recombination and disproportionation, and the following additional activation and deactivation reactions, where Z and L are (de-)activators:

$$\left.\begin{aligned} &\mathrm{IZ} + \mathrm{L} \xrightarrow{k_I^a} \mathrm{I} + \mathrm{LZ} \\ &\sim\mathrm{AZ} + \mathrm{L} \xrightarrow{k_A^a} \sim\mathrm{A}^\bullet + \mathrm{LZ} \\ &\sim\mathrm{BZ} + \mathrm{L} \xrightarrow{k_B^a} \sim\mathrm{B}^\bullet + \mathrm{LZ} \end{aligned}\right\} \text{Activation}$$

$$\left.\begin{aligned} &\mathrm{I} + \mathrm{LZ} \xrightarrow{k_I^{da}} \mathrm{IZ} + \mathrm{L} \\ &\sim\mathrm{A}^\bullet + \mathrm{LZ} \xrightarrow{k_A^{da}} \sim\mathrm{AZ} + \mathrm{L} \\ &\sim\mathrm{B}^\bullet + \mathrm{LZ} \xrightarrow{k_B^{da}} \sim\mathrm{BZ} + \mathrm{L} \end{aligned}\right\} \text{Deactivation}$$

Monte-Carlo Simulation Parameters

For the Monte-Carlo simulations, we use 10^2 to 10^6 polymer chains and 10 repetitions. The simulations are stopped at full conversion of A and B or if the simulated reaction time reaches 10^3 seconds. We implemented the Monte-Carlo simulation software in Java using the conventional Gillespie's algorithm [33] and computed fingerprints by calculating a histogram from the simulated chains.

For all datasets of living polymerizations, the reaction rates (table 4.6) were chosen such that $r_\mathsf{A} = \frac{1}{r_\mathsf{B}} = r$, with the reactivity ratios $r_\mathsf{A} = \frac{k_{AA}}{k_{AB}}$, $r_\mathsf{B} = \frac{k_{BB}}{k_{BA}}$, and the ratio of homopropagation rates $r = \frac{k_{AA}}{k_{BB}}$. For Monte-Carlo simulations of the other polymerization types, the parameters of the dataset with $\mathrm{DP}_n = 25, r_A = 2.0$ for initiation and propagation reactions were used. The reaction rates of the termination and depropagation rates k^d, k^r, k^{dp} varied over 0, 0.001, 0.01, and 0.1. For free radical polymerization, a decomposition rate $k^{DEC} = 10$ were used, and for controlled radical polymerization, activation rates $k^a = 100$ and deactivation rates $k^{da} = 0.01$ were used.

For the evaluation in Section 6.3, different noise levels were simulated by multiplying the fingerprint abundances by log-normal distributed random noise with mean zero and variance σ, where the noise parameter σ has the values 0, 0.05, 0.15, and 0.25.

Dataset		**Initial concentration**			**Reaction rates**			
DP_n	r_A	$[\mathrm{I}]_0$	$[\mathrm{A}]_0$	$[\mathrm{B}]_0$	k_AA	k_AB	k_BA	k_BB
3	0.01	1	1	2	0.01	1	0.01	1
3	0.05	1	1	2	0.05	1	0.05	1
3	0.1	1	1	2	0.1	1	0.1	1
3	0.25	1	1	2	0.25	1	0.25	1
3	0.5	1	1	2	0.5	1	0.5	1
3	0.75	1	1	2	0.75	1	0.75	1
3	1	1	1	2	1	1	1	1
3	1.25	1	1	2	1.25	1	1.25	1
3	1.5	1	1	2	1.5	1	1.5	1
3	1.75	1	1	2	1.75	1	1.75	1
3	2	1	1	2	2	1	2	1
25	2	1	10	15	2	1	2	1
45	2	1	20	25	2	1	2	1

Table 4.6 Initial concentrations (in $\mathrm{mol{\cdot}L^{-1}}$) and reaction rates of the Monte-Carlo simulations of living polymerizations.

5. From Mass Spectra to Copolymer Fingerprints

In this chapter we introduce copolymer fingerprints.[1] First, we discuss how to robustly transfrom mass spectra to fingerprints. Second, we propose a method to counteract the mass discrimination in MALDI-TOF mass spectra: peaks in (mostly) higher mass ranges of a spectrum being less pronounced than they theoretically should be.

5.1 Computing Copolymer Fingerprints

In this section, we propose a method to infer copolymer fingerprints from a single MS measurement. Our method uses peak areas instead of peak heights and can handle overlapping isotopes. We also propose an approach to resolve isobaric molecules, which is a frequently occurring issue in copolymer MS [41]. To the best of our knowledge, this has previously been possible only by using complementary measurements, such as NMR investigations.

We demonstrate the validity of our method using several synthesized copolymers measured with MALDI time-of-flight (TOF) MS. To evaluate our method's power to resolve isotope overlaps and isobaric molecules, we have simulated mass spectra for different monomers. We evaluate our software to the approach of Vivó-Truyols et al. [101], which is the most recent for this problem.

5.1.1 Computational Workflow

In the first step of our method, we centroid the spectra, that is, we identify peaks and their area-under-peak. We do not provide details for this approach, as it has

[1] In this chapter, parts of the sections 5.1.2, 5.2.1, and 5.2.4 where written in collaboration with Sarah Crotty. The focus of Sarah Crotty was on the chemical issues while my focus was on the computational aspects.

been discussed extensively in the literature [116]. For the following steps of our analysis, we will use the representation of the spectrum as a list of peaks and peak areas, as this allows us faster processing of the data. To reduce noise, we remove peaks below a certain threshold. We assume that all molecules in the MALDI spectrum are single-charged. The *mass range* is the interval from the smallest mass to the largest mass of any observed peak, but can be further restricted if required. Furthermore, we assume that the absolute mass error in the measured spectrum is at most $\Delta_{\mathrm{m}} < 0.5$ m/z; we will call this fixed Δ_{m} the *mass accuracy*. This implies that measured peaks can be uniquely assigned to one theoretical peak of an isotopic pattern. To simplify our presentation, we assume that the mass of initiating and terminating end-groups plus cationization agent is a constant which is ignored in our presentation: As a consequence, the mass of a monomer composition $\mathsf{A}_i\mathsf{B}_j$ is the sum of its monomer masses $m = i \cdot m_{\mathsf{A}} + j \cdot m_{\mathsf{B}}$.

Different compositions of monomer repeating units A and B can result in copolymers with similar monoisotopic masses. To this end, we often observe peaks with multiple potential explanations. We define two monomer compositions as *isobaric* if the difference of their monoisotopic masses is less than the mass accuracy. In this case, peak mass differences of the theoretical isotopic patterns for these two monomer compositions will usually be smaller than the mass accuracy, too. As the last step of our method, we present an approach for untangling the isotopic patterns of isobaric monomer compositions.

However, even if the difference of monoisotopic masses of two monomer compositions is above the mass accuracy, it is possible that some isotopic peaks of their theoretical isotopic patterns have mass difference below the mass accuracy. We say that two isotopic patterns are *overlapping*, if the difference of at least one peak in both two isotopic pattern is below mass accuracy.

Our method estimates relative abundances of all possible monomer compositions $\mathsf{A}_i\mathsf{B}_j$ in the MS spectrum. It proceeds in four steps:

1. Generate all candidate isotopic patterns;
2. assign candidate peaks to the MS spectrum;
3. compute the abundances and simultaneously resolve overlapping isotopes;
4. resolve isobaric molecules.

Candidate Generation

We first compute theoretical isotopic distributions for all monomer compositions $\mathsf{A}_i\mathsf{B}_j$ with monoisotopic mass within the mass range. We compute the first n peaks of each isotopic pattern by convolving the elemental isotopic distributions [9].

Next, we identify isobaric monomer compositions. Consider the monomer compositions $\mathsf{A}_i\mathsf{B}_j$ and $\mathsf{A}_{i-\Delta i}\mathsf{B}_{j+\Delta j}$ for natural numbers $i, j \geq 0$ and $\Delta i, \Delta j > 0$. Masses m_1 and m_2 of these two monomer compositions are:

$$\begin{aligned} m_1 &= i \cdot m_\mathsf{A} + j \cdot m_\mathsf{B}, \\ m_2 &= (i - \Delta i) \cdot m_\mathsf{A} + (j + \Delta j) \cdot m_\mathsf{B} \end{aligned} \tag{5.1}$$

Recall that two monomer compositions are isobaric, if their mass difference is less than the mass error $|m_1 - m_2| < \Delta_\mathrm{m}$. Substituting m_1 and m_2 using Eqn. (5.1) we infer $|\Delta i \cdot m_\mathsf{A} - \Delta j \cdot m_\mathsf{B}| < \Delta_\mathrm{m}$. Thus, given $\Delta j > 0$, any natural number $\Delta i > 0$ with

$$\frac{\Delta j \cdot m_\mathsf{B} - \Delta_\mathrm{m}}{m_\mathsf{A}} < \Delta i < \frac{\Delta j \cdot m_\mathsf{B} + \Delta_\mathrm{m}}{m_\mathsf{A}} \tag{5.2}$$

leads to isobaric monomer compositions $\mathsf{A}_i\mathsf{B}_j$ and $\mathsf{A}_{i-\Delta i}\mathsf{B}_{j+\Delta j}$. This is independent of the choice of $i, j \geq 0$. To this end, we call any such pair $(\Delta i, \Delta j)$ an *isobaric series*.

We determine all isobaric series; then, we use the isobaric series to arrange the monomer compositions (and, hence, the corresponding isotopic patterns) into isobaric sets. For each monomer composition $\mathsf{A}_i\mathsf{B}_j$ we iterate over all isobaric series $(\Delta i, \Delta j)$. If there is another monomer composition $\mathsf{A}_{i-\Delta i}\mathsf{B}_{j+\Delta j}$ within the mass range, these two are grouped into the same isobaric set. Note that an isobaric set can also contain only a single monomer composition. For each isobaric set, we compute an average isotopic pattern for all theoretical isotopic patterns of the monomer compositions in the isobaric set; this will be our *candidate* isotopic patterns. After computing the abundances for each isobaric series, we will distribute the abundances over all monomer compositions in each series in the last step.

Template Matching

In this step, we assign the candidate isotopic pattern peaks to the measured peaks in the experimental MS spectrum. However, measured peaks with a distance less than Δ_m can lead to ambiguous assignments: These peaks may be caused by overlapping raw peaks, or errors during the centroiding. Thus, we assume centroids with a distance less than Δ_m to originate from one continuous peak area, and merge them. The mass of a merged peak is the area-weighted average of its component peak masses. The area of a merged peak is the sum of its components peak areas. Naturally, we may accidentally merge two peaks, which are actually separate, or signal with noise peaks. However, the fingerprint estimation in the next step is robust towards this kind of error and noisy data in general.

Each measured peak is now assigned to zero, one, or several peaks of the candidate isotopic patterns. We match an isotopic pattern peak to a measured peak if their distance is less than Δ_m. Formally, let $m'_{i,j,k}$ be the mass and $I'_{i,j,k}$ the intensity of

the k th peak in the isotopic pattern of monomer composition $\mathsf{A}_i\mathsf{B}_j$. Let m_l and I_l be the mass and area under curve of the l th measured peak. Then, the set of matching peaks is:

$$S_l = \left\{(i,j,k) \,:\, |m_l - m'_{i,j,k}| < \Delta_\mathrm{m}\right\} \tag{5.3}$$

We define S_0 as the set of all unmatched candidate peaks:

$$S_0 = \{(i,j,k) \,:\, \text{there is no } l \text{ with } (i,j,k) \in S_l\} \tag{5.4}$$

These sets form a partition of all candidate isotope pattern peaks.

Fingerprint Estimation

We now describe how to estimate copolymer fingerprints. For each monomer composition $\mathsf{A}_i\mathsf{B}_j$ we want to find the matrix of relative abundances M, with $0 \leq M_{i,j} \leq 1$, which minimizes the distance of its theoretical isotopic pattern to the assigned measured peaks. Formally, we solve the following optimization problem:

$$\arg\min_M \sum_l \left| \sum_{(i,j,k)\in S_l} M_{i,j} \cdot I'_{i,j,k} - I_l \right| + \sum_{(i,j,k)\in S_0} M_{i,j} \cdot I'_{i,j,k} \tag{5.5}$$

The first term of Eqn. (5.5) tries to minimize the distance of the measured area under peak I_l to all its matching potentially overlapping candidate peaks, that is, the sum of polymer abundance times theoretical isotopic intensities $M_{i,j} \cdot I'_{i,j,k}$. The second term of Eqn. (5.5) considers all candidate isotope peaks that have no matching measured peak. Since these are not represented in the spectrum, we minimize the distance of the sum of their intensities times polymer abundance $M_{i,j} \cdot I'_{i,j,k}$ to a zero peak area.

The number of free parameters $M_{i,j}$ is determined by the number of possible template isotopic patterns, which increases quadratic in mass: There are $m+1$ compositions of two monomers for a given integer mass $m = i{\cdot}\mathsf{A}{+}j{\cdot}\mathsf{B}$ [8]. The sum of all compositions with integer mass at most m can be estimated by $\sum_{k=1}^{m}(k+1) = \frac{m(m+3)}{2} \in O(m^2)$.

We efficiently solve this high-dimensional optimization problem by transforming it to a Linear Program (LP). We introduce distance coefficients d_0 for the unmatched theoretical peaks and a distance coefficient d_l for each measured peak. Then, solving the Linear Program

$$
\begin{aligned}
\min \quad & \sum_{l} d_l \\
\text{s.t.} \quad & \sum_{(i,j,k)\in S_l} M_{i,j} \cdot I'_{i,j,k} + d_l \geq I_l \quad \forall l && \text{(5.6a)} \\
& \sum_{(i,j,k)\in S_l} M_{i,j} \cdot I'_{i,j,k} - d_l \leq I_l \quad \forall l && \text{(5.6b)} \\
& \sum_{(i,j,k)\in S_0} M_{i,j} \cdot I'_{i,j,k} + d_0 \geq 0 && \text{(5.6c)} \\
& \sum_{(i,j,k)\in S_0} M_{i,j} \cdot I'_{i,j,k} - d_0 \leq 0 && \text{(5.6d)}
\end{aligned}
$$

estimates the optimal abundances $M_{i,j}$. We omitted the upper and lower limit constraints for all coefficients. Constraint equations (5.6a) and (5.6b) correspond to the first term of Eqn. (5.5), and constraint equations (5.6c) and (5.6d) to the second term. In case there are isobaric monomer compositions with $M_{i,j} > 0$, we will resolve them in the next step.

Resolving Isobaric Molecules

Isobaric monomer compositions have almost identical monoisotopic mass, so there are competing possible explanations for certain measured peaks. Given any two isobaric monomer compositions, the differences in isotope abundances of the corresponding theoretical isotopic patterns are usually not significant enough to split the measured abundances. Therefore, we suggest an alternate approach to split corresponding entries in the copolymer fingerprint M. Obviously, this is not necessary if there are no isobaric monomer compositions present.

Our task is to split abundances $M_{i,j}$ that correspond to more than one monomer composition, *i.e.* that belong to isobaric sets with two or more elements. It has been suggested repeatedly that distributions of polymer abundances follow some common probability distribution such as Poisson distribution or Schulz-Zimm distribution. Wilczek-Vera et al. [109] suggested that monomer composition abundances can be modeled by a suitable bivariate distribution, and also suggested to use Poisson or Schulz-Zimm distributions as the marginal distributions. To simplify our computations, we further approximate this using a normal distribution: For example, the Poisson distribution $P(\lambda)$ with parameter λ can be approximated by a normal distribution $\mathcal{N}(\lambda, \sqrt{\lambda})$. The joint distribution of two normal distributions is a bivariate normal distribution. We now use the bivariate normal distribution to split abundances of isobaric sets with more than one monomer composition.

In principle, we may do this splitting by the following procedure:

1. Estimate the mean $\mu = (\mu_1, \mu_2)$ and covariance matrix Σ of the bivariate normal distribution $F = \mathcal{N}(\mu, \Sigma)$ of the fingerprint M. In the first iteration,

	Theoretical		M_n (^{1}H NMR)		M_n (COCONUT)		M_p (COCONUT)	
	PS	**PI**	**PS**	**PI**	**PS**	**PI**	**PS**	**PI**
I1	19	7	17	9	17.4	8.2	17	8
I2	17	11	12.5	11	13.7	8.3	11	8
I3	14	15	16	13	16.7	8.9	18	9
P1	24	36	21	35	23.6	26.6	25	26
P2	24	37	21	29	21.7	22.5	22	22
P3	24	37	22	33	23.1	26.0	24	26

Table 5.1 Summary of M_n and M_p values.

we consider only those entries of M where the corresponding isobaric set has cardinality one.

2. Do the following for each isobaric set B of cardinality two or more: Let r be the sum of abundances of all monomer compositions in B. Now, we distribute this abundance over all monomer compositions in B:

$$R_{i,j} := \frac{F(i,j)}{\sum_{(x,y)\in B} F(x,y)} \cdot r \tag{5.7}$$

Repeat this until M converges. We found that this approach is often too slow in practice; to this end, we instead use a general purpose optimizer [80] that combines both of these steps (estimating the bivariate normal and splitting the abundances) into one. We leave out the tedious technical details.

5.1.2 Experimental (PS-r-PI)-r-(PS-r-PI)

In this section, we demonstrate our method using a (PS-r-PI)-r-(PS-r-PI) copolymer (Section 4.1.1).

Copolymers were synthesized with two random macromers with different ratios of styrene and isoprene (Section 4.1.1), analyzed by MALDI-TOF MS (Appendix Fig. A.1) and the COCONUT software (Appendix Fig. A.2). The estimated copolymer fingerprints (Fig. 5.2) were transformed to distributions of chain sizes and compositions (Fig. 5.3) by calculating the isoprene ratios and interpolating them for each anti-diagonal of fingerprint. They show a compositional drift, indicating a high conversion rate, since the distribution is not symmetric with respect to the monomer fractions [66].

Table 5.1 shows the theoretical ratios between styrene and isoprene in the first macromer and the complete copolymer, the values obtained by ^{1}H NMR and the

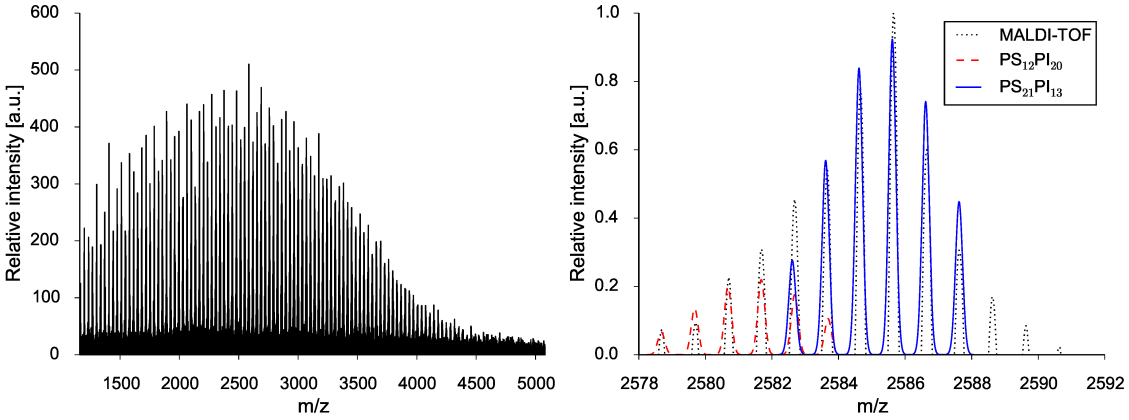


Figure 5.1 Left: MALDI-TOF spectrum of the (PS-*r*-PI) copolymer I1. Right: Detail of the spectrum overlayed with the estimated theoretical isotopes. We used six isotopic peaks per pattern to estimate the abundances.

ratios estimated from the copolymer fingerprints (Fig. 5.2). The maximal value in the fingerprint correlates to the highest intensity in the MS spectrum. It is thus the maximum of the copolymer distribution, the M_p value. We computed the M_n value by taking the average of the marginal distributions of the fingerprints (Appendix Fig. A.3). The COCONUT and ^{1}H NMR values are slightly lower than the theoretical values for both monomers, which may be due to some deactivation of the initiator by impurities in the solvent and also the challenging usage of isoprene. The M_n values of COCONUT and ^{1}H NMR are in a good correlation for the first macromer and are slightly shifted for the entire copolymers due to Ag^+ clusters. The clusters form when Ag^+ is used as cationization agent and thus ion suppression was used to have less interference with the polymer signal.

We did observe overlapping isotopes in the MS spectra and multiple isobaric distributions in the fingerprint, most likely due to added THF to act as a randomizer. As shown in Fig. 5.1 overlapping isotopes were resolved. Moreover, for each copolymer, one isobaric distribution was determined by our method, which we confirmed by comparing both average monomer composition from NMR and COCONUT (Table 5.1).

Huijser et al. [45], Staal [90] and Willemse [112] suggested a quick way to provide an indication of the microstructure from the slope of a line, fitted through the copolymer fingerprint. In reference to the fingerprints from I1 to I3 (Fig. 5.2), we can observe straight lines, which correlate to a block like structure. However, we expected a random copolymer, where the line should go through the origin with a constant slope. Possibly due to intensity deviations in the high m/z range the origin of the line could have a slight offset which explains the uncertainty in the microstructure determination. However, this deviation could also occur during the synthesis where

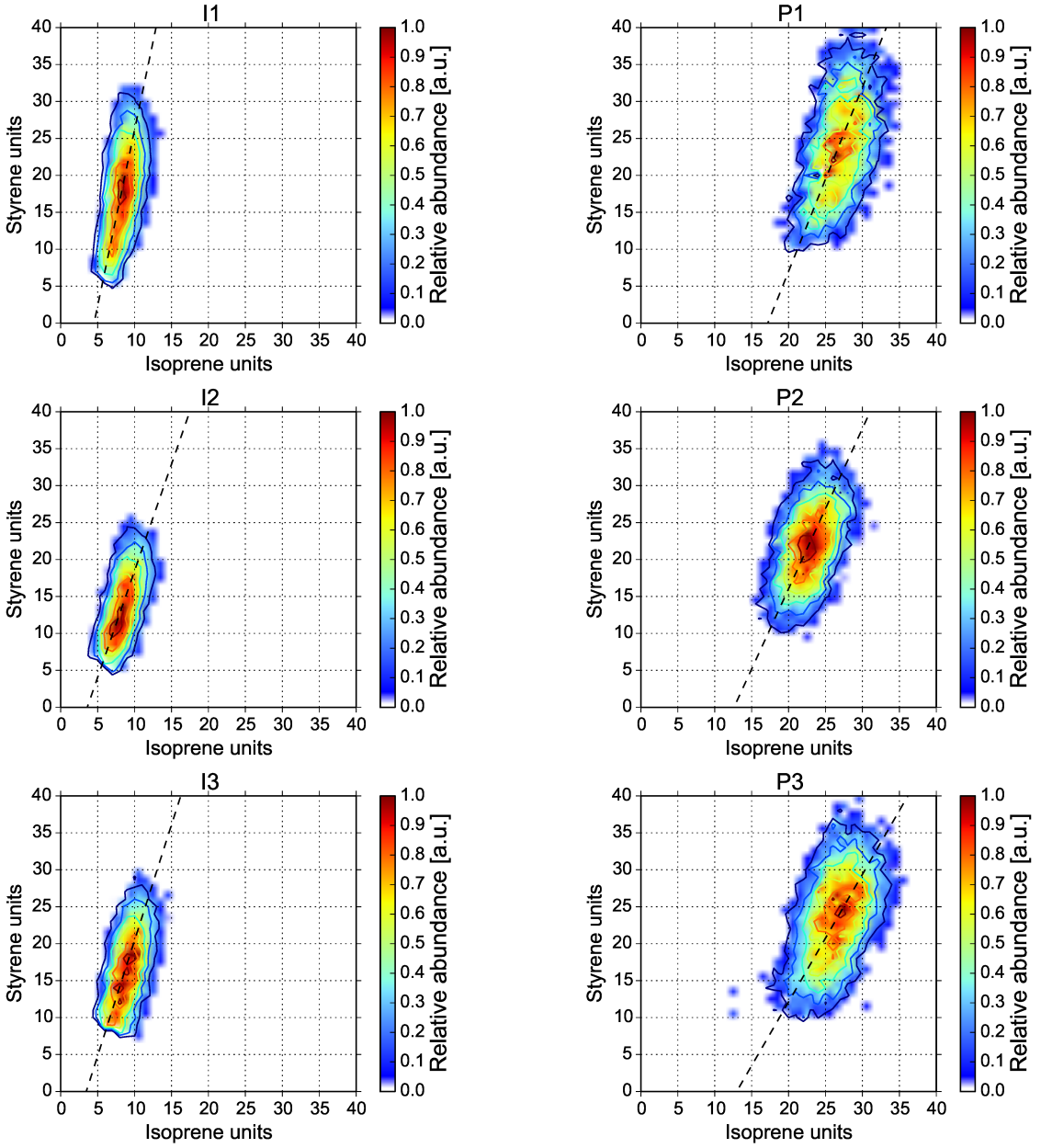


Figure 5.2 Copolymer fingerprint of the (PS-*r*-PI) macromers I1 to I3 (left) and the final (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers P1 to P3 (right).

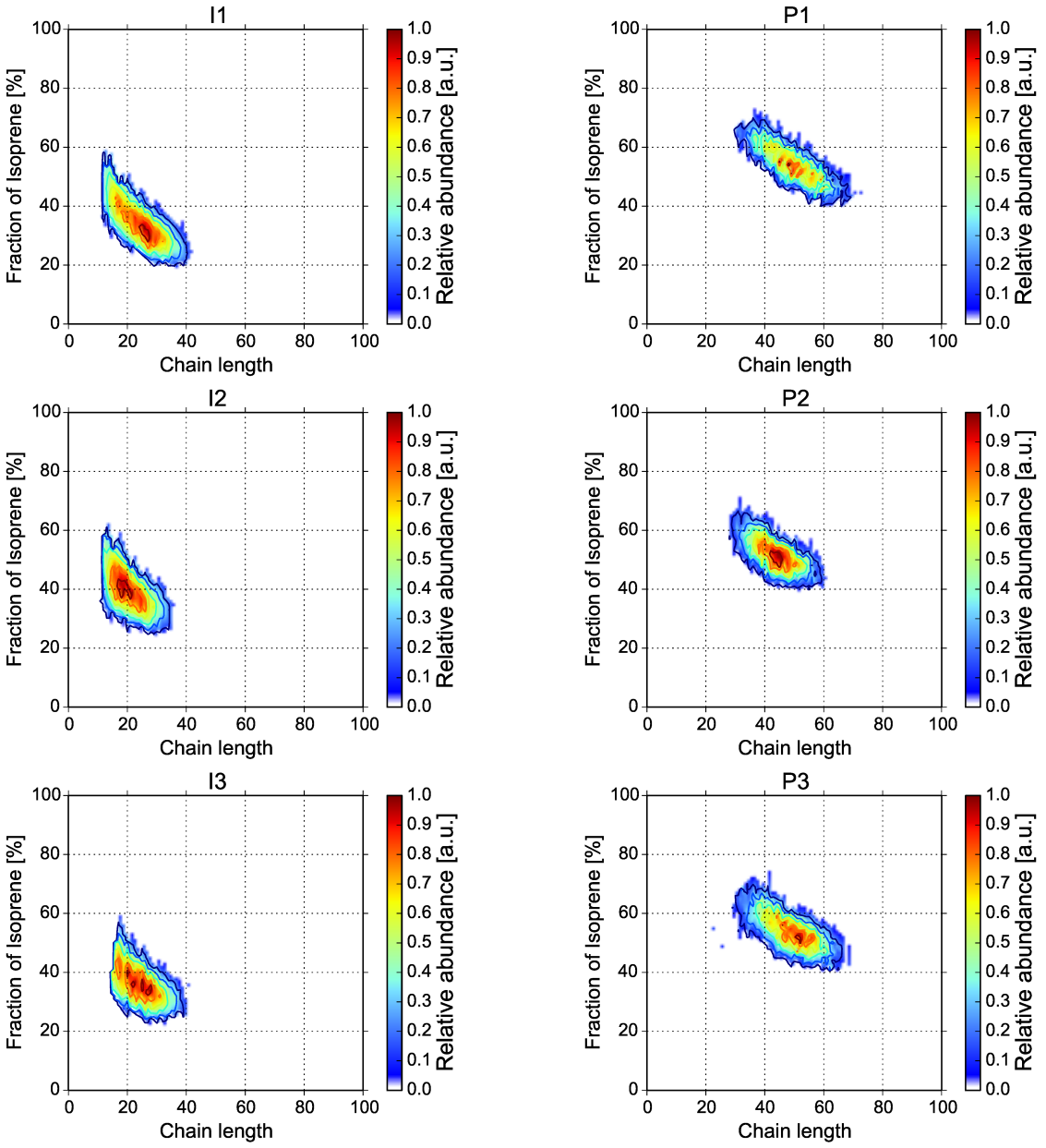


Figure 5.3 Copolymer composition as a function of degree of polymerization and the ratio of isoprene of the (PS-*r*-PI) macromers I1 to I3 (left) and the final (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers P1 to P3 (right).

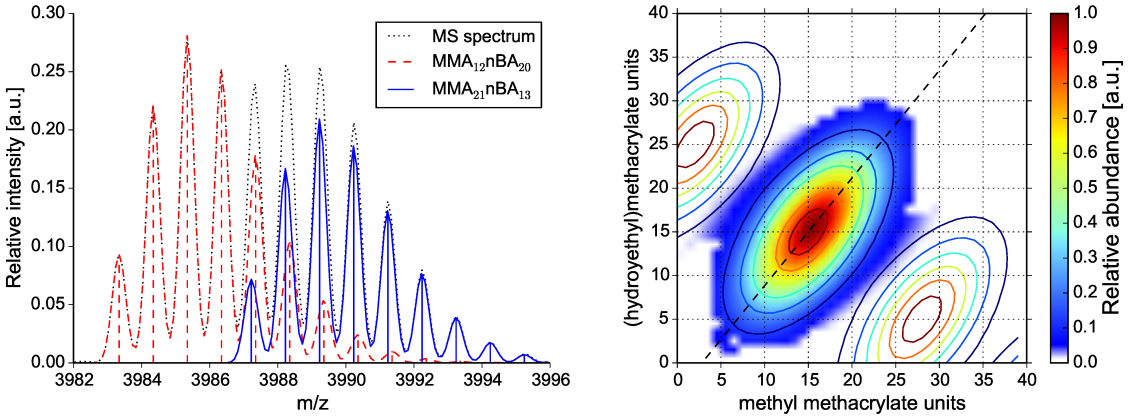


Figure 5.4 Left: Detail of the simulated MS spectrum of PMMA-*co*-P*n*BA showing overlapping isotopes. The relative molecule abundances estimated by COCONUT are represented by the centroid intensities. Right: Copolymer fingerprint estimated from a simulated MS spectrum of a PMMA-*co*-PHEMA copolymer overlayed with all isobaric distributions (contours).

THF is considered as randomizer. Nonetheless the P1 to P3 do correlate to block like structures as was desired.

5.1.3 Simulated PMMA-co-PnBA/PMMA-co-PHEMA

In this section, we evaluate our method using simulated datasets (Section 4.2.1).

First, we analyzed two noise-free spectra of PMMA-*co*-P*n*BA and PMMA-*co*-PHEMA using COCONUT with intensity threshold 0.05. The abundances of the overlapping isotopes in PMMA-*co*-P*n*BA spectrum were correctly calculated (Fig. 5.4). The distribution was almost perfectly reconstructed, only isotopes below the intensity threshold were not considered by our method and, thus, lost (Appendix Fig. A.4). In the simulated spectrum of PMMA-*co*-PHEMA (Appendix Fig. A.5), there exist three neighboring isobaric distributions that may explain the data; from these, COCONUT chose the correct distribution located in the center of the fingerprint (Fig. 5.4). Both simulations indicate that our method can reconstruct the true copolymer distribution, given that the input spectrum is free of noise.

To assess the robustness of our method we use the second simulated dataset with noise. We stress that for noise parameter $\sigma = 0.5$, resulting signal-to-noise ratios are below 50% on average, resulting in very challenging instances for any quantification method. We also applied the "strip-based regression" (SBR) method [101] to this simulated dataset. To the best of our knowledge, this is the only freely available software for this purpose; at the same time, it is the newest approach reported in the literature and, hence, arguably the most advanced to date.

We evaluated results by calculating the Pearson correlation coefficient of each estimated fingerprints against the original fingerprint (Fig. 5.5). For each method,

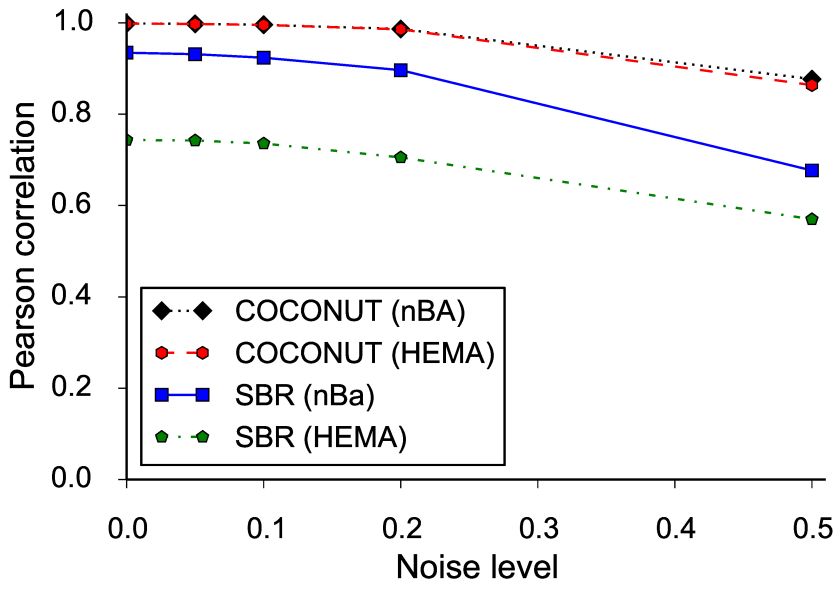


Figure 5.5 Median Pearson correlation coefficient for each method and copolymer dataset, PMMA-*co*-P*n*BA and PMMA-*co*-PHEMA, at five different noise levels.

noise level and dataset, we calculated the median over all coefficients. We find that for both datasets, our method is capable of reconstructing the correct fingerprint with very high accuracy (Pearson correlation close to one) for noise parameter up to 0.2. Only for noise parameter $\sigma = 0.5$, we observe a significant deviation between estimated and original fingerprint. We see a similar pattern for the SBR method, with no significant correlation differences for noise parameter between 0 and 0.2, and a pronounced drop for noise parameter $\sigma = 0.5$. But SBR reaches smaller Pearson correlation for both copolymers: for PMMA-*co*-P*n*BA correlation is between 0.89 and 0.93, and for PMMA-*co*-PHEMA it is between 0.70 and 0.74, leaving out noise parameter $\sigma = 0.5$. Examining the fingerprints calculated by SBR for individual spectra, it appears that SBR cannot redistribute abundances of isobaric monomer compositions, what explains the decreased Pearson correlation for PMMA-*co*-PHEMA copolymers.

On average, COCONUT required 8.7 seconds per PMMA-*co*-P*n*BA spectrum, and 46.0 seconds per PMMA-*co*-PHEMA spectrum. The difference was caused by the numerous isobaric isotopes, which had to be resolved in the second dataset. SBR required an average of 203.2 seconds per spectrum for both datasets.

5.2 Abundance Correction

In the previous Section 5.1, we propose a method to infer copolymer fingerprints matrix from a single MS measurement. However, the accuracy of the abundances depends highly on the accuracy of the mass spectral intensities. But the task of

estimating the entire copolymer fingerprint from MALDI-TOF spectra turns out to be only semi-quantitative due to the mass and composition-dependent ionization. The differential ionization leads to mass discrimination, *i.e.* peaks at certain m/z's being less intense than expected. This phenomenon is very pronounced at higher masses and it is best observed when peaks of the analyte ions span over a wide mass range [56, 82]. The mass discrimination depends on instrumental parameters such as the time-lag setting, the laser energy, and the wire-voltage setting [86]. Furthermore, mass discrimination depends on the polydispersity index (PDI) of the analyte, and the crystal homogeneity [85], as well as the monomer and matrix polarity [39]. In addition, it may be influenced by other factors, for example the matrix/salt ratio and matrix/analyte ratio [42], or the matrix solubility [108]. In consequence, many groups have used hyphenation techniques such as size exclusion chromatography (SEC), high pressure liquid chromatography (HPLC), 2D-LC, or ion mobility spectrometry to MS as methods for quantification [75, 95, 98, 108]. However, in our opinion MALDI-MS is a strong competitor, for example solvent-free MALDI-MS showed significant improvements in reliability and quantitation [96].

In this section, we describe a novel computational method to counteract the differential mass discrimination and investigate its limits with respect to the PDI. We demonstrate our new approach using two homopolymer mixtures and, for comparison, new MALDI-TOF MS measurements of the copolymers previously reported in Section 5.1.

First, we very briefly discuss the experimental results. For more information please refer to the corresponding publications [25, 26]. Second, we discuss in detail the new computational method: We show how to estimate the *molecular weight distribution* (MWD) of a homopolymer mixture, and investigate the limits of this approach with respect to the PDI. Next, we describe how to estimate an abundance correcting function from the MWD to counteract the mass discrimination in the homopolymer spectra. Thereafter, we apply the correction to the homopolymer spectra and describe a method to correct the copolymer spectra, based on the previously estimated homopolymer correction parameters.

5.2.1 Polymerization and MALDI-TOF MS

In Section 5.1, we use a (PS-r-PI)-r-(PS-r-PI) random copolymer dataset with two random macromers with different ratios of styrene and isoprene (Section 4.1.1). Compared to the previously reported spectra and Mn values, the newly measured spectra clearly show degradation products, (Appendix Fig. A.8), resulting in lower Mn values (Table 5.3). The degradation products accumulated in the lower mass regions of the spectra and have not been taken into account for computing the copolymer fingerprints (Fig. 5.9). The laser intensities of the MALDI laser where 46% and 49%.

Mixtures of two different molar masses of both homopolymer for PS and PI were measured at several laser intensities. Laser intensities for PS homopolymers where

48%, 51%, and 54%, for PI homopolymers 36%, 45%, and 54%. The intensities in the MALDI-TOF MS spectra reveal a dependency upon laser intensity and a discrimination of high molar masses particularly at low laser powers. The PS spectra were expected to show less discrimination of higher masses at the higher laser intensities [56, 86]. We did not observe this, which could be explained by the sample preparation: the dried droplet method being hindered from a matrix segregation and a coffee ring effect [31]. PS homopolymers were analyzed with dithranol and AgTFA whereas DCTB and AgTFA was used for the PI homopolymers to ionize both homopolymer mixtures. The change in the matrix was necessary, as no spectra with signals over the whole mass range could be obtained for the PI homopolymers mixtures when analyzed with dithranol. Both the solvent and cationization agent remained identical to reduce differences in the co-crystallization. We identified the baselines for each spectrum by fitting a Loess curve to the signal "valleys". The baselines were subtracted from the spectra. We identified isotope patterns and quantified the abundances of the oligomers using the average peak heights of the isotopic patterns (Appendix Fig. A.6). To reduce stochastic errors, the resulting peak lists were averaged over the three replicates for each laser intensity.

5.2.2 Molecular Weight Distribution

We suggest to apply an abundance correcting function f to mitigate mass discrimination effects. However, since the mass discrimination is an undetermined function, we propose a data-driven approach to estimate the correction parameters. To this end, we need to estimate the MWD of the homopolymer mixtures. Textbooks in polymer science [71] state that MWD of a homopolymer can be characterized by the Gamma distribution:

$$Gamma(\alpha, \beta) = \frac{\beta\alpha}{\Gamma(\alpha} x^{\alpha-1} e^{-\beta x} \tag{5.8}$$

The two dimensionless parameters can be transformed into parameters with dimensions of $\mathrm{M_n}$ and $\mathrm{M_w}$, *i.e.* g/mol. The parameter change is $\alpha = \frac{\mathrm{DP_w}}{\mathrm{DP_n}}$ and $\beta = \alpha - 1$, where $\mathrm{DP_n}$ and $\mathrm{DP_w}$ are proportional to $\mathrm{M_w}$ and $\mathrm{M_n}$, respectively. The resulting is known as the Schulz-Zimm distribution. Thus, given a mixture of homopolymers, the MWD can be described by a mixture of Gamma distributions. However, our goal is an estimation technique which is insensitive to small departures from the idealized assumptions. We use the symbols $N(x_0, \sigma)$ to indicate a Gaussian curve with variance σ, centered at x_0. It is universally accepted that, given a mixture distribution with well separated modes, estimating a mixture of normal (*i.e.* Gaussian) distributions is more robust than a mixture of Gamma distributions.

The question is, what is the error in the normal approximation to the Gamma distribution? We are interested in the cumulative distribution function (cdf), *i.e.* the integral. The upper and lower integration limits are usually $-\infty$ and x, respectively. But in our case there is no need for negative numbers and thus integration limits

are 0 and x. Let $F(x)$ be the error between the cdf of the Gamma distribution $Gamma(\alpha, \beta)$ and $\Phi(x)$ the cdf of the normal distribution $N(\frac{\alpha}{\beta}, \frac{\alpha}{\beta^2})$. The Gamma distribution is the sum of α exponential distributions. The central limit theorem tells us, the sum of any independent and identically distributed random variables converges in distribution to a normal distribution as the number of random variables approaches infinity. Thus, in general, the error in the approximation is decreasing, as α grows large. More specifically, according to the findings by Shevtsova [88], that the maximal error between both cdfs is:

$$\sup_{x \in \mathbb{R}} |F(x) - \Phi(x)| \leq \frac{0.3328(\rho + 0.429\sigma^3)}{\sigma^3 \sqrt{\alpha}} \tag{5.9}$$

Inserting $\sigma = \frac{\sqrt{\alpha}}{\beta}$ and the third absolute moment $\rho = \frac{\alpha(\alpha+1)(\alpha+2)}{\beta^3}$ of the Gamma distribution yields:

$$\sup_{x \in \mathbb{R}} |F(x) - \Phi(x)| \leq 0.3328(\alpha + \frac{2}{\alpha} + \frac{0.429}{\sqrt{\alpha}} + 3) \tag{5.10}$$

Since this a rather pessimistic upper limit, we also calculated the actual maximal error numerically by computing the maximum of $|F(x) - \Phi(x)|$ as a function of α. Figure 5.6 shows both error estimates as functions of the PDI $\frac{\mathrm{M_w}}{\mathrm{M_n}}$. In the following, we briefly recall the well-known relationship between PDI and the Gamma distribution. Let $\mathbb{E}$ be the expected value of the distribution of masses M. Then, the variance is:

$$\sigma^2 = \mathbb{E}(M^2) - \mathbb{E}(M)^2 = \mathrm{M_w} \cdot \mathrm{M_n} - \mathrm{M_n^2} \tag{5.11}$$

Thus, the PDI is:

$$\frac{\mathrm{M_w}}{\mathrm{M_n}} = \frac{\sigma^2}{\mathrm{M_n^2}} + 1 = \frac{\sigma^2}{\mu^2} + 1 = \frac{1}{\alpha} + 1 \tag{5.12}$$

Using the numerically calculated error in the normal approximation to the Gamma distribution (Fig. 5.6), we determined the limitations of this approach. The error is less than 4% for PDI ≤ 1.1, which is satisfied by the homopolymers we used. Also, the error is less than 10% for PDI ≤ 1.56, and less than 16% for PDI ≤ 2. Therefore, a normal approximation is applicable to most living polymerizations and the MWD of a mixture of such homopolymers can be described by a mixture of normal distributions $\sum_i w_i N(\mu_i, \sigma_i)$, with scaling factor w, mean μ and variance σ for each homopolymer in the mixture. We estimated the MWD of the PS and PI homopolymers mixtures for all laser intensities using least squares nonlinear regression. Formally, let I be homopolymers in the mixture, and K the indices of the observed abundances Y in the MS spectra. As usual, we assume that the observed

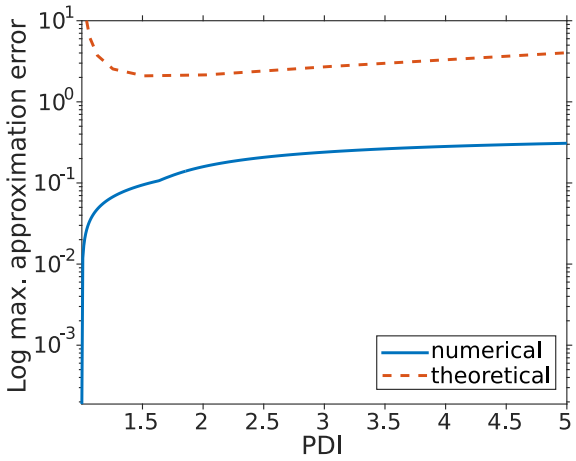


Figure 5.6 Numerically determined error in the normal approximation to the gamma distribution and its theoretical upper limit on a logarithmic scale as a function of the PDI.

abundances are normalized to one. The regression minimizes for each laser intensity j the squared error:

$$SE_j = |w_{i,j}N(\mu_i, \sigma_i) - Y_{j,k}|_2^2 \tag{5.13}$$

5.2.3 Abundance Correcting Function

In the previous section, we described how to estimate the MWD of the homopolymer mixtures using normal mixture distributions. In the following, we describe how to compute the abundance correcting function from the MWDs. Supposing that there were no mass discrimination effects, the areas under the curve of all homopolymers of each mixture should be equal, because the homopolymers in the mixtures are equimolar and the relationship between intensity and abundance is linear in homopolymers [103]. Thus, the ideal theoretical MWDs can be estimated by equalizing the areas of homopolymers with a normalizing factor. Let I be the homopolymers in the mixture, J the laser intensities. The theoretical MWD is

$$\sum_{i \in I} c_{i,j} w_{i,j} N(\mu_i, \sigma_i)\,, \tag{5.14}$$

with $j \in J$ and the normalizing factors $c_{i,j}$. We calculate the normalizing factor by taking the ratio of the largest area in the mixture to the area of the current homopolymer $i \in I$, such that:

$$c_{i,j} = \frac{\max\limits_{i' \in I} \int w_{i',j} N(\mu_{i'}, \sigma_{i'})}{\int w_{i,j} N(\mu_i, \sigma_i)} \tag{5.15}$$

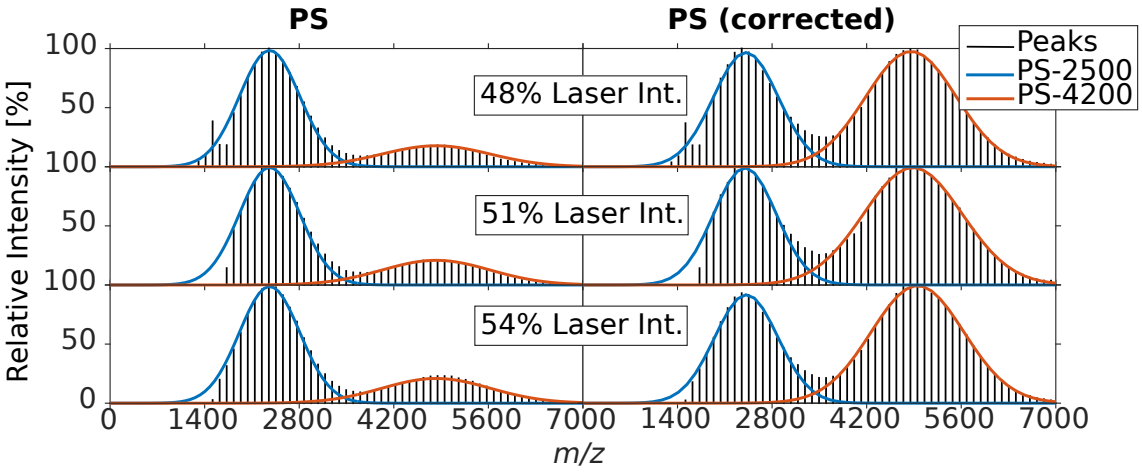


Figure 5.7 Measured (left) and corrected (right) peak lists of PS, with the estimated MWDs of PS-2500 and PS-4200.

The mass discrimination is an unknown function. Other important parameters, such as matrix/analyte and matrix/salt ratios, were supposed to be constant throughout our experiments. Thus, the observed mass discrimination depends on the laser intensity and mass. However, in principle, other parameters, such as matrix/analyte and matrix/salt ratios can be included by conducting more experiments with varying ratios.

To correct for the mass discrimination effects, the *correcting function* $f(m, l)$ (which takes different values as the mass m and laser intensity l change), can be calculated by dividing the ideal theoretical MWD by the observed MWD. We collected the sample points in the intervals $\mu_i \pm \frac{k\sigma_i}{c_{i,j}}$ for each component and laser intensity with the sample interval width $1 \leq k \leq 3$, which is automatically chosen with an hill climbing optimization to minimize the distance of area ratios to 1. We estimated the abundance correcting functions $f_{PS}(m, l)$ and $f_{PI}(m, l)$ for PS and PI by fitting a Thin Plate Spline (TPS) to the sample points (Appendix Fig. A.7) [10]. TPS is a standard technique for interpolating data with more than one dimension. It is able to provide a good fit to the sample points and avoids the oscillation problems that occur when interpolating using polynomials.

5.2.4 Abundance Correction

After calculating the correcting function, we apply it to the homopolymer spectra of PS and PI (Fig. 5.7 and Fig. 5.8, respectively). The areas under the curve of the homopolymers are now nearly equal (Table 5.2, Fig. 5.7 and Fig. 5.8, right). This indicates, that the spectra could be corrected for the contributions of both investigated parameters (mass and laser intensity) to the mass discrimination, which favored the low masses and underestimated the abundances of the higher masses.

Before correction the PS homopolymers show a slight mass discrimination for PS-4200, which is less pronounced at higher laser powers, which could be due to “hot

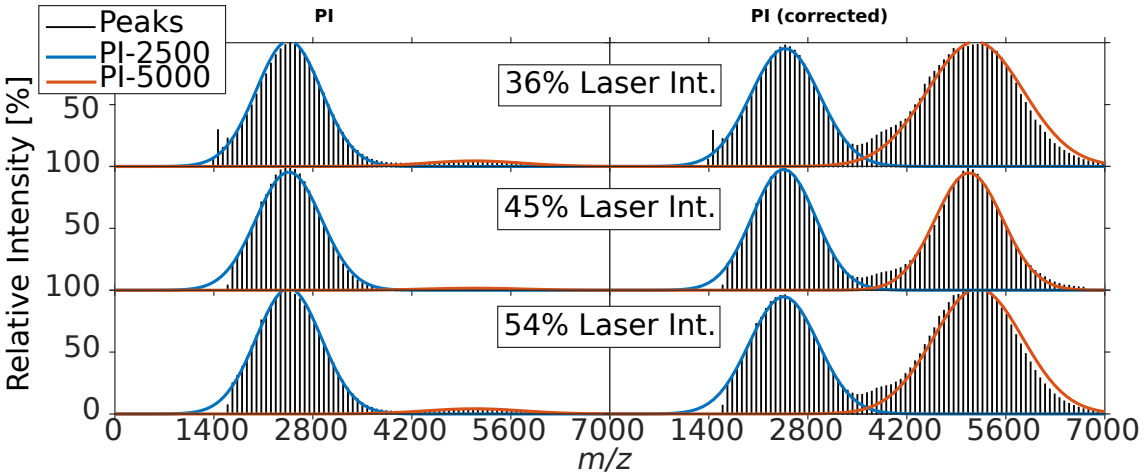


Figure 5.8 Measured (left) and corrected (right) peaks lists of PI, with the estimated MWDs of PI-2500 and PI-5000.

Laser Int.	**PS-4200/PS-2500**		**Laser Int.**	**PI-4200/PI-2500**	
[%]	**Uncorrected**	**Corrected**	[%]	**Uncorrected**	**Corrected**
48	0.1731	1.0634	**36**	0.0446	1.0594
51	0.2146	1.0006	**45**	0.0161	0.9474
54	0.2179	1.0542	**54**	0.0419	1.0484

Table 5.2 Ratios of the area under curve (AUC) of the MWDs of the homopolymers in the PS and PI mixtures before and after correction. With no mass discrimination, both AUCs should be equal and the ratio one, because the homopolymers are equimolar.

spots" of the analyte on the MALDI target plate. (Fig 5.7, left) As for the PI mixtures, the measured spectra show a strong mass discrimination even with DCTB as matrix. (Fig. 5.8, left) At first sight, this result may seem not in line with the results obtained by Yalcin and Schriemer [115]. In fact they used a copper salt with a different matrix and they measured less discrimination. However, at a second sight, the many changes in the ionization of the mixtures would affect significantly the intensity and, thus, the copolymer evaluations [54]. Nonetheless, a correction for the mass discrimination effects depending on mass and laser intensity was achieved with the PI mixtures, despite the strong mass discrimination favoring the low masses. Strong mass discrimination is more challenging for the estimation of the correcting function. Besides this, our approach is mainly limited by the mass spectrometer: The larger the mass range, the more peak intensities at higher mass might be suppressed or discriminated. [14] In case the signals are discriminated to the point of being indistinguishable from the noise, additional experiments like blanking out lower masses (*i.e.* suppressing intensities of lower masses) or fractionation could be performed.

Mass discrimination favoring low masses over high masses is a known effect in polymer MS and has been studied carefully for homopolymers [39, 42, 43, 85, 86]. Although the mass discrimination in copolymers has been experimentally observed, there is, to the best of our knowledge, no comprehensive theory of the mass discrimination phenomenon in copolymer MS. In the following, we assume that mass and monomer frequency are the predominant analyte factors for the copolymer ionization properties. In contrast, we assume that sequence plays only a subordinate role. We also assume that the influence of the three-dimensional structure is negligible, because this work focuses on linear polymers.

To account for the influence of the monomer frequency to the mass discrimination in copolymers, we indicate with $\#PS$ and $\#PI$ the copolymer composition (number of PS and PI monomer repeating units, respectively) and we propose to apply the correction in the simplest way, as a weighted sum according to their fraction of monomers in the chain:

$$f(m,l) = \frac{\#PS \cdot f_{PS}(m,l) + \#PI \cdot f_{PI}(m,l)}{\#PS + \#PI} \tag{5.16}$$

Applying the correction resulted in higher average numbers of PS and PI (Table 5.3). Instead of a compact circular shape, the distribution now shows a narrow oval shape (Fig. 5.9). However, the upper parts are "lost", as higher mass peaks dropped below the noise threshold and the distribution is less smooth due to the increased influence of the noise in the higher mass regions. The narrow shape is typical for living polymerizations [46], which is also supported by the PDI values of less than 1.1. Fitting a line through the most abundant oligomers before and after correction results in a straight line off center for both, which hints at the desired random-like structure [44].

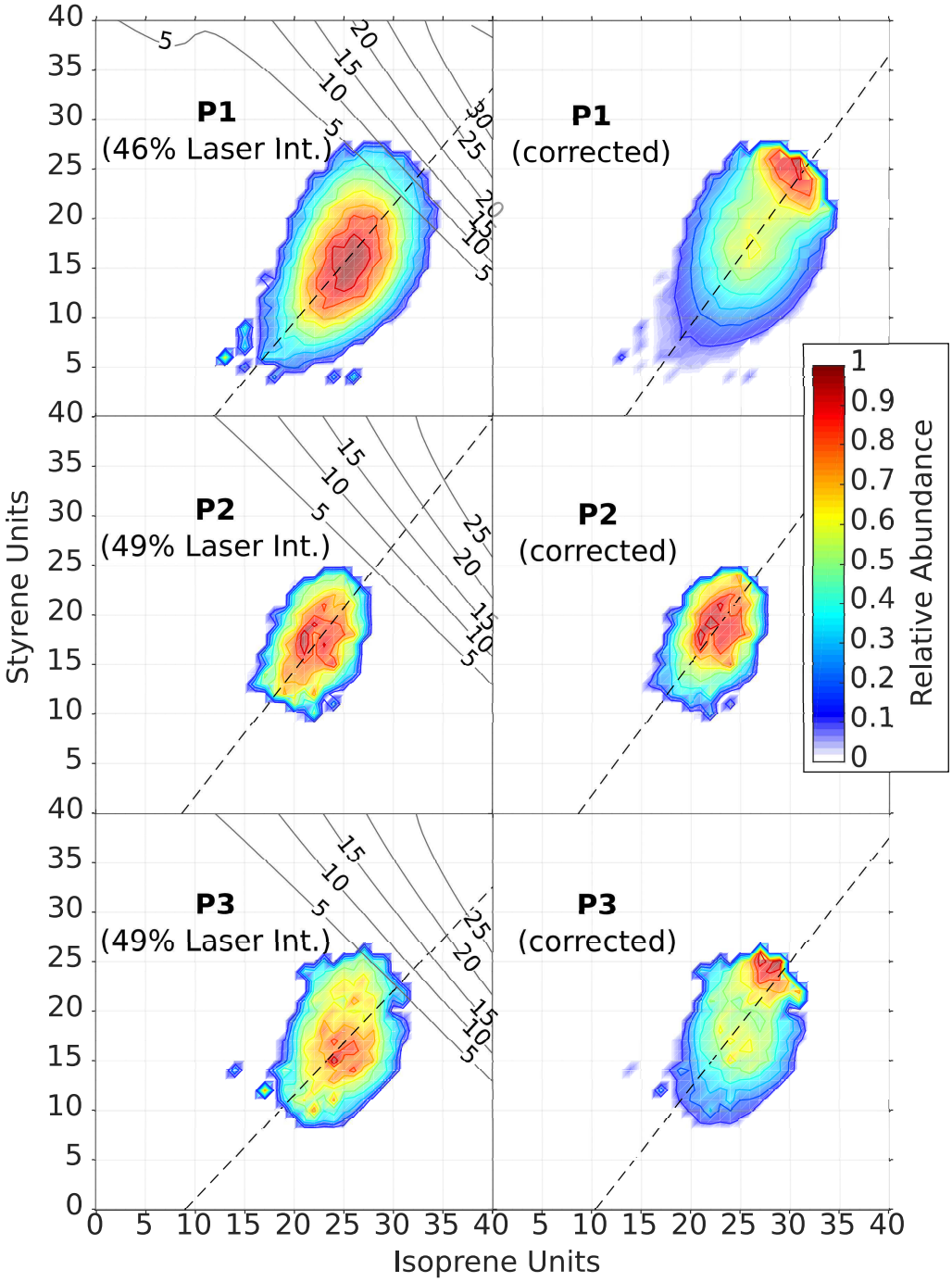


Figure 5.9 Measured (left) and corrected (right) copolymer fingerprints of P1 to P3. The overlaid contour lines on the left side represent the intensity correcting function, *i.e.* the correcting factor for each monomer combination of the fingerprint. Dashed lines represent the average compositions computed by fitting a line through the most abundant fingerprint entries.

	M_n (Uncorrected)			M_n (Corrected)		
	PS	**PI**	$\mathbf{g \cdot mol^{-1}}$	**PS**	**PI**	$\mathbf{g \cdot mol^{-1}}$
P1	15.9	25.2	3,533	18.0	26.5	3,838
P2	17.3	22.2	3,472	17.9	22.5	3,561
P3	17.1	24.7	3,623	18.3	25.3	3,787

Table 5.3 M_n values computed from the copolymer fingerprints of P1 to P3

However, due to the sharp slope of the correcting function for larger numbers of PI units, the measured copolymers are less affected in the PI dimension than the non-degraded copolymers would be. Also, there is the possible issue of underestimating PI even after correcting the abundances, due to the differences between copolymer and homopolymer MALDI matrix, which was necessary to obtain decent PI mass spectra due to the high mass discrimination in the homopolymer mixtures. The experimental setup for homo- and copolymers should be kept as similar as possible, because we assumed the mass discrimination affects them similarly. Generally, a new measurement of homopolymers and re-computation of the correcting function should be performed before major changes such as a change in monomers or MALDI matrixes. Also, as the instrument laser and detector degrade over time, homopolymer measurements should be repeated regularly.

6. New Copolymerization Models

In this chapter we present new models for copolymerization. First, we introduce and evaluate two models, the Bernoulli and Geometric model, each in two different versions, with and without taking reactivity ratios into account. Second, we describe how to estimate the model parameters and explore the model limitations with respect to different polymerization types. And third, we present several algorithms to compute useful statistical properties from the models.

6.1 The Bernoulli and Geometric Copolymerization Models

Copolymerization is a random process, where two or more monomer species are mixed to form polymer chains. In the past, several approaches to model copolymerization were proposed. The well-known terminal model by Mayo and Lewis describes four propagation reactions and is determined by the reactivity ratios of the monomers [59]. There are three different computational approaches to such a basic reaction scheme and each approach has certain disadvantages. The reaction scheme can be modeled as a set of ordinary differential equations (ODE), a discrete Markov chain or simulated with Monte-Carlo methods.

In this section, we propose two new Markov chain models for copolymerization kinetics, the Bernoulli and the Geometric model, based on a simple reaction scheme. Different to Mayo and Lewis [59], our model allows for variable chain lengths and time-dependent monomer probabilities. The accuracy of Monte-Carlo simulations depends on the number of simulated chains, the simulated distribution converges to the true distribution with an increasing number. This makes accurate computations time- and memory-intensive. In contrast to Monte-Carlo simulations, our models are exact and fast. We implement a simple copolymerization scheme using ODEs and Monte-Carlo simulations. We verify the Monte-Carlo simulations with the ODE

system. We evaluate our models against the fingerprints and copolymer chains computed by Monte-Carlo.

6.1.1 Bernoulli Model

Chain Lengths

Consider the synthesis of a single polymer chain. We divide the continuous reaction time into T discrete time steps, which we call synthesis steps. At each step, there are two mutually exclusive events: Adding a monomer or not. This random process is equivalent to conducting a series of T Bernoulli trials for every polymer chain and recording the chain lengths, *i.e.* how many monomers were added. Thus, the chain lengths are binomially distributed with parameters T, the number of trials, and p_M, the probability of adding a monomer.

Fingerprint Model

We extend the model to describe copolymer fingerprints. At each of the T discrete synthesis steps, three mutually exclusive events are possible: adding monomer A, monomer B, or nothing. However, in general, the ratio of A to B changes during the synthesis, therefore the probabilities of adding A or B change. We define the *monomer probability* parameters $p_\mathsf{A}(t)$ and $p_\mathsf{B}(t)$, with $p_\mathsf{A}(t) + p_\mathsf{B}(t) = 1$ for all $1 \leq t \leq T$. p_A and p_B are vectors of length T, describing the probability of encountering a monomer A or B at each synthesis step.

We model copolymerization as an inhomogeneous Markov chain and call this basic model the *Bernoulli model* (Fig. 6.1). We describe a copolymer fingerprint as a matrix M of size $n \times m$, in which entry $M_{a,b}$ gives the relative abundance of a copolymer with a monomers of type A and b monomers of type B. The states of the Markov chain correspond to the fingerprint entries. The transition probabilities correspond to the three possible events, append A, B, or nothing. The transition probability from state $M_{a,b}$ to $M_{a+1,b}$ is the probability of adding a monomer p_M times the probability of encountering an A at synthesis step t:

$$\mathbb{P}(M_{a,b} \to M_{a+1,b}; t) = p_M \cdot p_\mathsf{A}(t) \tag{6.1}$$

Analogously, the transition probability from $M_{a,b}$ to $M_{a,b+1}$ is the probability of adding a monomer times the probability of encountering a B:

$$\mathbb{P}(M_{a,b} \to M_{a,b+1}; t) = p_M \cdot p_\mathsf{B}(t) \tag{6.2}$$

The transition probability for staying in state $M_{a,b}$ is the probability of adding nothing:

$$\mathbb{P}(M_{a,b} \to M_{a,b}; t) = 1 - p_M \tag{6.3}$$

All other transition probabilities are zero.

The *starting distribution* $M(0)$ is a matrix of zeros, except for $M_{0,0}(0) = 1$. This means that before starting the synthesis all chains have zero monomer repeating units A and B. To conform to standard Markov chain notation, let M be a row vector. Let P be the *matrix of transition probabilities*. Starting with $M(0)$, the copolymer fingerprint at synthesis step t is:

$$M(t) = M(t-1) \cdot P(t) \tag{6.4}$$

We are interested in the fingerprint after the completed synthesis, that is the fingerprint at the last synthesis step $M(T)$. The transition matrix P is sparse, thus Eqn. 6.4 can be simplified for $a > 0$ and $b > 0$ to:

$$\begin{aligned} M_{a,b}(t) = \;& p_M \cdot p_{\mathsf{A}}(t) \cdot M_{a-1,b}(t-1) \\ + \;& p_M \cdot p_{\mathsf{B}}(t) \cdot M_{a,b-1}(t-1) \\ + \;& (1-p_M) \cdot M_{a,b}(t-1) \end{aligned} \tag{6.5}$$

If $a = 0$ or $b = 0$, one needs to delete from Eqn. 6.5 the first or second term, respectively. In each synthesis step $1 \leq t \leq T$ we compute $n \times m$ fingerprint entries in constant time for each entry. Because $n \leq T$ and $m \leq T$, the worst case running time is $O(T^3)$. It is not necessary to save the fingerprints for each synthesis step as $M(t)$ only depends on $M(t-1)$, therefore the memory requirement is $O(T^2)$.

Reactivity Ratios

So far, our model has not taken reactivity ratios into account. The probability of a reaction equals the probability of adding a certain monomer times the probability of encountering that monomer. However, the reactivity ratios are known to influence the copolymerization process. For example, if monomer A has a strong affinity for monomer B, a weak affinity for A, and monomer B has the reverse affinity, then the result will be an alternating copolymer. To this end, we define a new model, the *Bernoulli model with reactivity parameters* (Fig. 6.1).

We define the *reactivity parameters* p_{AA}, p_{AB}, p_{BA}, and p_{BB}, which describe the probabilities of the reactions between the four possible pairings of chain ends and monomers. To be able to distinguish between chains ends, we use two fingerprints: M^{A}, the distribution of chains ending with A, and M^{B}, the distribution of chains ending with B. We are interested in the fingerprint after the final synthesis step T. The final fingerprint can be calculated by adding the final distributions of chains ending with A and B:

$$M(T) = M^{\mathsf{A}}(T) + M^{\mathsf{B}}(T) \tag{6.6}$$

We define the transition probabilities for the four possible reactions of chain ends and monomers. For $\mathsf{X} \in \{\mathsf{A}, \mathsf{B}\}$, the transition probabilities for adding A are:

$$\mathbb{P}(M^{\mathsf{X}}_{a,b} \to M^{\mathsf{A}}_{a+1,b}; t) = p_M \cdot c_{\mathsf{X}} \cdot p_{\mathsf{XA}} \cdot p_{\mathsf{A}}(t) \tag{6.7}$$

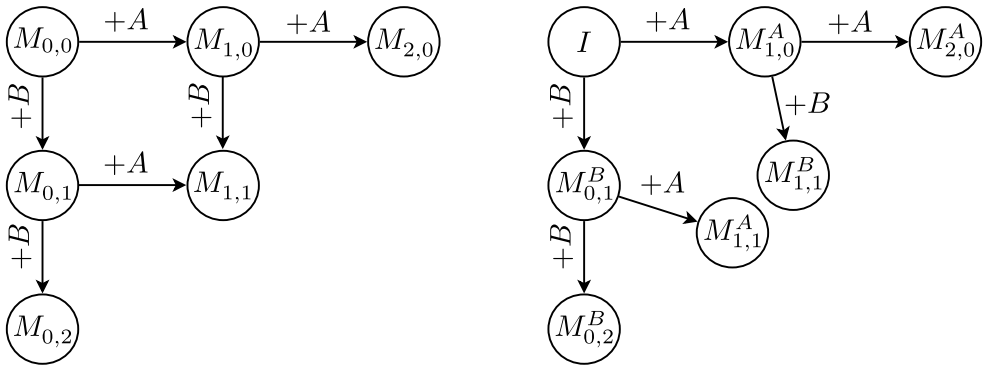


Figure 6.1 All possible transitions of the Bernoulli model without (left) and with (right) reactivity parameters for copolymer chain lengths ≤ 2. For example, the transition from the initiator state I to the state $M^{\mathsf{A}}_{1,0}$ (copolymer chains having one A-monomer and ending in A) corresponds to adding an A.

Analogously, the transition probabilities for adding B are:

$$\mathbb{P}(M^{\mathsf{X}}_{a,b} \to M^{\mathsf{B}}_{a,b+1}; t) = p_M \cdot c_{\mathsf{X}} \cdot p_{\mathsf{XB}} \cdot p_{\mathsf{B}}(t) \tag{6.8}$$

An important property of Markov chains is that the rows of the transition matrix sum to one. Introducing the reactivity parameters violated this property, therefore we use *normalization coefficients* c_{A} and c_{B} in the equations 6.7 and 6.8. The normalization coefficients are defined as:

$$c_{\mathsf{X}} = \frac{1}{p_{\mathsf{XA}} \cdot p_{\mathsf{A}}(t) + p_{\mathsf{XB}} \cdot p_{\mathsf{B}}(t)} \tag{6.9}$$

Because empty chains end neither with A nor with B, we define the *initiator state* I. The transition probabilities to start a chain are:

$$\begin{aligned} \mathbb{P}(\mathsf{I} \to M^{\mathsf{A}}_{1,0}; t) &= \; p_M \cdot p_{\mathsf{A}}(t) \\ \mathbb{P}(\mathsf{I} \to M^{\mathsf{B}}_{0,1}; t) &= \; p_M \cdot p_{\mathsf{B}}(t) \end{aligned} \tag{6.10}$$

The transition probabilities of the non-state-changing transitions are not affected by the reactivity parameters and are analogous to Eqn. 6.3. All other transition probabilities are zero. By applying the transition probabilities (equations 6.7, 6.8) and the normalization coefficients (Eqn. 6.9), the fingerprint M^{A} can be calculated for $a > 0$ and $b > 0$ by:

$$\begin{aligned} M^{\mathsf{A}}_{a,b}(t) = \;& p_M c_A p_{\mathsf{AA}} p_{\mathsf{A}}(t) \cdot M^{\mathsf{A}}_{a-1,b}(t-1) \\ + \;& p_M c_B p_{\mathsf{BA}} p_{\mathsf{A}}(t) \cdot M^{\mathsf{B}}_{a-1,b}(t-1) \\ + \;& (1-p_M) \cdot M^{\mathsf{A}}_{a,b}(t-1) \end{aligned} \tag{6.11}$$

Analogously, fingerprint M^{B} is:

$$\begin{aligned} M^{\mathsf{B}}_{a,b}(t) = \;& p_M c_{\mathsf{A}} p_{\mathsf{AB}} p_{\mathsf{B}}(t) \cdot M^{\mathsf{A}}_{a,b-1}(t-1) \\ + \;& p_M c_{\mathsf{B}} p_{\mathsf{BB}} p_{\mathsf{B}}(t) \cdot M^{\mathsf{B}}_{a,b-1}(t-1) \\ + \;& (1-p_M) \cdot M^{\mathsf{B}}_{a,b}(t-1) \end{aligned} \tag{6.12}$$

If $a = 0$ or $b = 0$, the appropriate terms can be deleted from equations 6.11 and 6.12. For $a = 1, b = 0$ or $b = 1, a = 0$ equations 6.11 and 6.12 change according to Eqn. 6.10.

The running time and memory requirements change by a constant factor, therefore the worst case running time is still $O(T^3)$ and memory is $O(T^2)$.

6.1.2 Geometric Model

Chain Length

The Bernoulli model we introduced above used T discrete synthesis steps to add monomers A, B or nothing. Adding a monomer or not is a Bernoulli trial and the resulting chain lengths are binomially distributed. However, in practice, polymer lengths often show a long-tailed distribution, which is usually modeled by a gamma distribution [13, 92, 109]. Here, we modify our discrete model for a long-tailed chain length distribution. The discrete equivalent to the continuous gamma distribution is the negative binomial distribution. A random variable following a negative binomial distribution with parameters T and p equals the sum of T independent geometrically distributed random variables with parameter $1-p$. To this end, we model the discrete steps using the geometric distribution.

Consider the synthesis of a single polymer chain. In each synthesis step, the number of monomers, which are added to the chain, is random. The probability of adding k monomers follows a geometric distribution with parameter p_ϵ, the "stop" probability:

$$p_G(k) = (1-p_\epsilon)^k p_\epsilon \tag{6.13}$$

We call this the *Geometric model*.

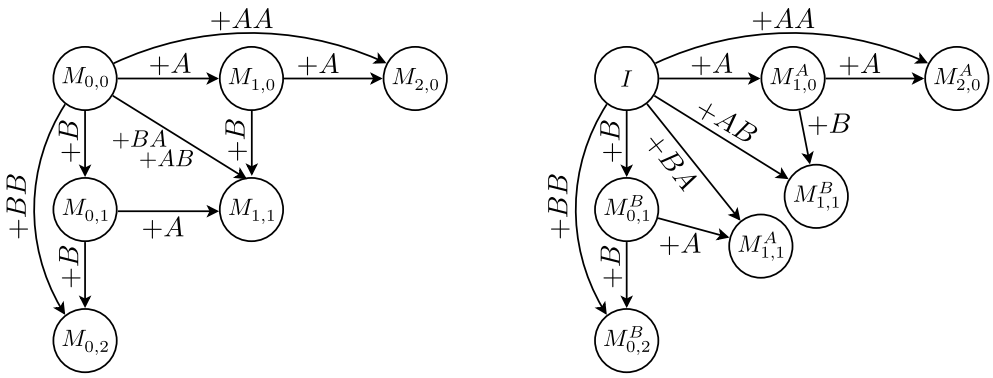


Figure 6.2 All possible transitions of the Geometric model without (left) and with (right) reactivity parameters for copolymer chain lengths ≤ 2. For example, the transition from the initiator state I to the state $M^{\mathsf{A}}_{2,0}$ (copolymer chains having two A-monomers and ending in A) corresponds to adding the sequence AA. Note that there are transitions, which correspond to multiple events. For example, the transition of I to $M^{\mathsf{A}}_{2,1}$ corresponds to adding the two sequences BAA and ABA.

Fingerprint Model

In the following, we describe the *Geometric model* (Fig. 6.2). Due the geometrically distributed number of monomers to add in each synthesis step, the number of possible transitions increases compared to the Bernoulli model. Given $i \geq 0$ and $j \geq 0$, the transition probability from $M_{a,b}$ to any state with equal or higher numbers of A and B is the number of combinations with i monomers of type A and j monomers of type B times the probability of adding $i + j$ monomers times the probabilities of encountering i monomers of type A and j monomers of type B:

$$\mathbb{P}(M_{a,b} \to M_{a+i,b+j}; t) = \binom{i+j}{j} \cdot p_G(i+j) \cdot p_{\mathsf{A}}(t)^i \cdot p_{\mathsf{B}}(t)^j \tag{6.14}$$

To save computation time, the number of combinations $\binom{i+j}{j}$ can be calculated using Pascal's triangle. As with the Bernoulli model, the memory requirements are $O(T^2)$. However, the running time increases to $O(T^5)$, because we need to iterate over all possible i and j.

Reactivity Ratios

Analogous to the Bernoulli model, we define a *Geometric model with reacitivity parameters* (Fig. 6.2). We use the reactivity parameters p_{AA}, p_{AB}, p_{BA}, and p_{BB} to model the reactivity ratios, the initiator state I, and two fingerprints M^{A} and M^{B} to describe the distributions of chains ending with A or B, respectively.

In contrast to the Bernoulli model, the Geometric model is able to add more than one monomer per synthesis step. We need to determine the reactivity parameters for all possible combinations of added A and B. Consider one synthesis step of the Markov chain: We say that we start in a state $\mathsf{X} \in \{\mathsf{I}, \mathsf{A}, \mathsf{B}\}$, if the last added monomer of all previous steps was nothing, A, or B, respectively. We stop in state $Y \in \{\mathsf{A}, \mathsf{B}\}$, if the last added monomer of this or any previous step is an A or B, respectively. To this end, we introduce the matrix R^{XY}. $R^{\mathsf{XY}}_{a,b}$ is the probability of starting in state X, adding a monomers A, b monomers B, and ending in state Y. We define R^{XY} as:

$$\begin{aligned} R^{\mathsf{XA}}_{a,b} &= R^{\mathsf{XA}}_{a-1,b} \cdot p_{\mathsf{AA}} + R^{\mathsf{XB}}_{a-1,b} \cdot p_{\mathsf{BA}} \\ R^{\mathsf{XB}}_{a,b} &= R^{\mathsf{XA}}_{a,b-1} \cdot p_{\mathsf{AB}} + R^{\mathsf{XB}}_{a,b-1} \cdot p_{\mathsf{BB}} \end{aligned} \tag{6.15}$$

To compute R^{XY} for each possible combination of X and Y, we need to know the initial values. If no monomer is added, we start and end in the same state:

$$\begin{aligned} R^{\mathsf{XX}}_{0,0} &= 1 \\ R^{\mathsf{XY}}_{0,0} &= 0 \text{ for } \mathsf{X} \neq \mathsf{Y} \end{aligned} \tag{6.16}$$

If we start in the initiator state I and add one monomer, it is independent of the reactivity parameters:

$$\begin{aligned} R^{\mathsf{IA}}_{1,0} &= 1 \\ R^{\mathsf{IB}}_{0,1} &= 1 \end{aligned} \tag{6.17}$$

Analogously to the Bernoulli model with reactivity parameters, the rows of the transition matrix need to sum to one. We therefore define normalization coefficients and normalize the transition probabilities for all transitions which add the same number of monomers:

$$c_{\mathsf{X}}(k) = \frac{1}{\sum\limits_{a+b=k} (R^{\mathsf{XA}}_{a,b} + R^{\mathsf{XB}}_{a,b}) \cdot p_{\mathsf{A}}(t)^a \cdot p_{\mathsf{B}}(t)^b} \tag{6.18}$$

We now combine equations 6.15 to 6.18 to specify the transition probabilities for the Geometric model. For $X \in \{\mathsf{A}, \mathsf{B}\}$:

$$\mathbb{P}(M^{\mathsf{X}}_{a,b} \to M^{\mathsf{Y}}_{a+i,b+j}; t) = c_{\mathsf{X}}(i+j) \cdot R^{\mathsf{XY}}_{i,j} \cdot p_G(i+j) \cdot p_{\mathsf{A}}(t)^i \cdot p_{\mathsf{B}}(t)^j \tag{6.19}$$

The transition probabilities from the initiator state I to any other state are given by:

$$\mathbb{P}(\mathsf{I} \to M^{\mathsf{Y}}_{i,j}; t) = c_{\mathsf{I}}(i+j) \cdot R^{\mathsf{IY}}_{i,j} \cdot p_G(i+j) \cdot p_{\mathsf{A}}(t)^i \cdot p_{\mathsf{B}}(t)^j \tag{6.20}$$

The transition probability to not start a chain and stay in state I is:

$$\mathbb{P}(\mathsf{I} \to \mathsf{I}) = p_G(0) \tag{6.21}$$

We are interested in the fingerprint after the final synthesis step T. Analogous to the Bernoulli model, the final fingerprint can be calculated by adding the final distributions $M^{\mathsf{A}}(T)$ and $M^{\mathsf{B}}(t)$. Compared to the Geometric model without reactivity parameters, the running time and memory requirements change by a constant factor, therefore the worst case running time is still $O(T^5)$ and memory is $O(T^2)$.

6.1.3 Polymer Chain Likelihood

The Bernoulli and Geometric models described above compute the copolymer fingerprints, the distribution of all chains over the numbers of monomer repeating units. However, an additional interesting question is: What is the likelihood of a single copolymer chain under a given model?

To compute the likelihood of a single chain, we only consider transitions which may lead to the chain in question and transitions which do not add a monomer, *i.e.* non-state-changing transitions. All other transition probabilities are zero. After progressing T synthesis steps, the likelihood of the chain is the probability of the last reachable state.

For example, let us compute the likelihood of the chain "ABB". In addition to the non-state-changing transitions, the Bernoulli model would allow $M_{0,0} \to M_{1,0}$, $M_{1,0} \to M_{1,1}$, and $M_{1,1} \to M_{1,2}$. The likelihood of "ABB" is the probability of the state $M_{1,2}$. The likelihood under the Bernoulli model with reactivity parameters and the Geometric models can be computed analogously.

6.1.4 Parameter Estimation

The Bernoulli and Geometric models fully characterize the distribution of copolymer chains. Unfortunately, the true underlying distribution of copolymer chains is unknown, therefore we want to evaluate our results by comparing them to Monte-Carlo simulations. However, Monte-Carlo simulated chains are random samples. But the larger the sample size is, the closer the empirical distribution is to the true distribution and the better we can use the sample to the evaluate our models.

The accuracy of Monte-Carlo simulations strongly depend on the number of simulated chains (Fig. 6.3, left). We choose several instances with low degree of polymerization $DP_n = 3$ and different reactivity ratios r_{A}, r_{B} and homopropagation ratios r. Please note that for all datasets $r_{\mathsf{A}} = \frac{1}{r_{\mathsf{B}}} = r$. For $r_A = 1.0$, we compute 10 fingerprints M with 10^2 to 10^6 chains and compare them to the fingerprint M_{total}, which we compute using all $10 \cdot \sum_{i=2}^{6} 10^i = 11,111,000$ chains (Fig. 6.3, right). For comparison we use the normalized root mean square error $NRMSE(M, M_{total})$.

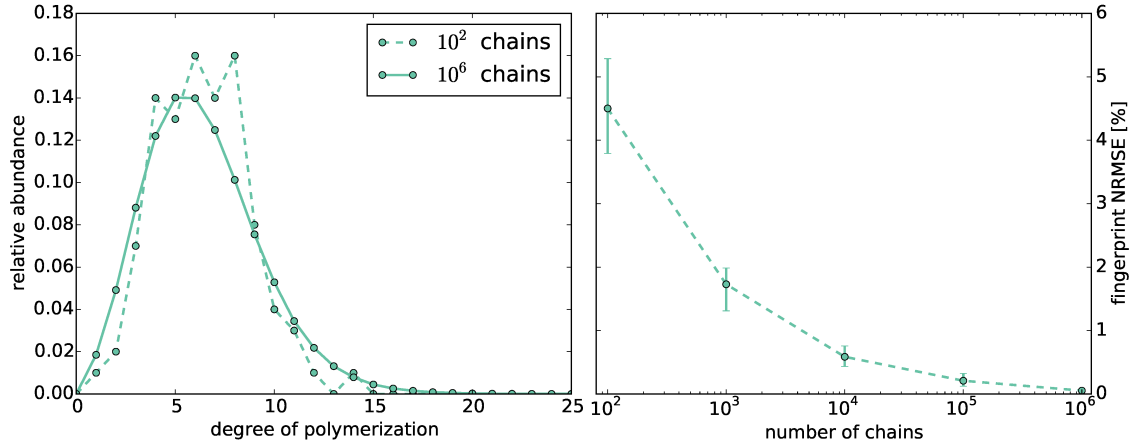


Figure 6.3 Left: Comparison of the distribution of chain lengths computed by the Monte-Carlo simulations with 10^2 *vs.* 10^6 chains at reactivity ratio $r_A = 1.0$. Right: Normalized root mean square error (NRSME) of the fingerprints computed by Monte-Carlo simulations with different numbers of chains compared to the fingerprint computed from all chains produced by all Monte-Carlo simulations at reactivity ratio $r_A = 1.0$.

The error decays with the number of chains. The lowest mean errors are $\sim$ 2% and $\sim$ 0.5% using 10^5 and 10^6 chains, respectively. We observe that the error for 10^5 is still significantly above zero. Thus, if not stated otherwise, we use 10^6 chains for Monte-Carlo simulations in the following.

For completeness, we evaluate the Monte-Carlo simulations by comparing the simulated concentrations to the concentrations computed by solving the ordinary differential equation model of the living copolymerization (Appendix Fig. A.9). The concentration curves are identical to the eye, strongly supporting the validity of the Monte-Carlo simulations.

We now compare the Bernoulli and Geometric models to the Monte-Carlo simulations. The reactivity parameters can be calculated from the reactivity ratios. For $\mathsf{X}, \mathsf{Y} \in \{\mathsf{A}, \mathsf{B}\}$, the reactivity parameters are:

$$p_{\mathsf{XY}} = \frac{k_{\mathsf{XY}}}{k_{\mathsf{XA}} + k_{\mathsf{XB}}} \tag{6.22}$$

Unfortunately, the other model parameters cannot be calculated intuitively from the Monte-Carlo simulation parameters. In principle, it is possible to estimate the parameters by fitting the model fingerprint to the Monte-Carlo fingerprint. However, to minimize the influence of the fitting algorithms, we apply a two-step estimation process.

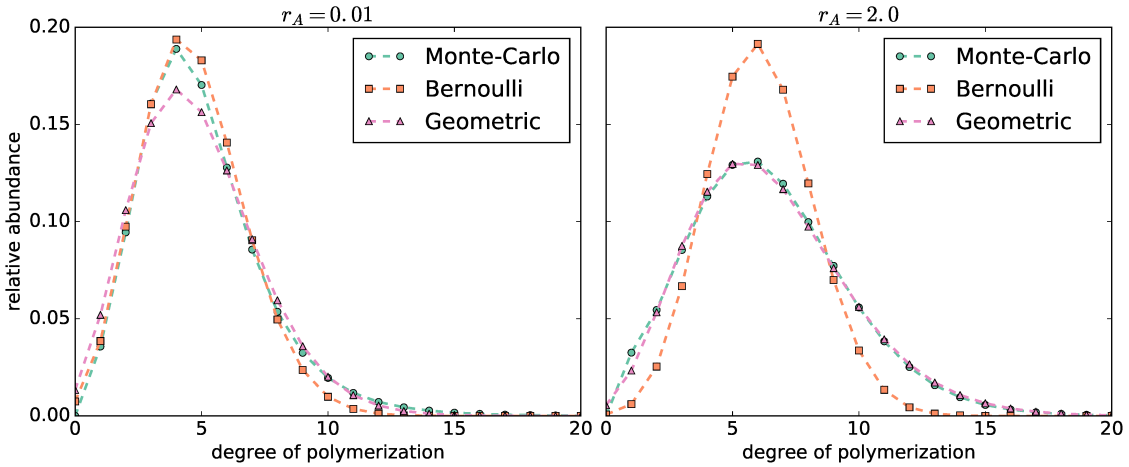


Figure 6.4 Comparison of the distribution of chain lengths computed by the Monte-Carlo simulations with $r_A = 0.01$ (left) and $r_A = 2.0$ (right) *vs.* the length distributions computed by the Bernoulli and Geometric models.

First, we estimate the number of synthesis steps and the probability of adding monomers. According to the Bernoulli and Geometric model, the chain lengths follow a binomial or negative binomial distribution, respectively.

We fit a binomial and a negative binomial probability mass function (pmf) to each copolymer length distribution (Fig. 6.4 and Appendix Fig. A.10). The length distributions become broader with increasing reactivity ratios r_A. The broader the distribution, the better it is approximated by a negative binomial pmf, and the worse by a binomial pmf. However, for narrow distributions ($r_\mathsf{A} =\leq 0.1$, which corresponds to a standard deviation $\sigma \leq 2.8$), we do not observe such a clear distinction: The mode of the distribution is better approximated by the binomial pmf. In contrast, the negative binomial pmf is better able to fit the long tail of the distribution.

Second, we estimate the monomer probabilities p_A and p_B. Because we defined $p_\mathsf{A} + p_\mathsf{B} = 1$, estimating p_A is sufficient. We divide the reaction time of the Monte-Carlo simulation into intervals. The number of intervals equals the number of synthesis steps T. We choose the left and right interval limits, such that the change in concentration is the same for each interval (Fig. 6.5 and Appendix Fig. A.11). We calculate the mean concentrations $\widetilde{[\mathsf{A}]}(t)$ and $\widetilde{[\mathsf{B}]}(t)$ for each interval $1 \leq t \leq T$. Then, the monomer probabilities $p_\mathsf{A}(t)$ can be calculated as:

$$p_\mathsf{A}(t) = \frac{\widetilde{[\mathsf{A}]}(t)}{\widetilde{[\mathsf{A}]}(t) + \widetilde{[\mathsf{B}]}(t)} \tag{6.23}$$

Please note that the parameter estimation using the concentrations from Monte-Carlo simulations is for evaluation purposes only. When applying the models to experimental data, the parameters can be estimated by fitting the computed fingerprint to the observed fingerprint.

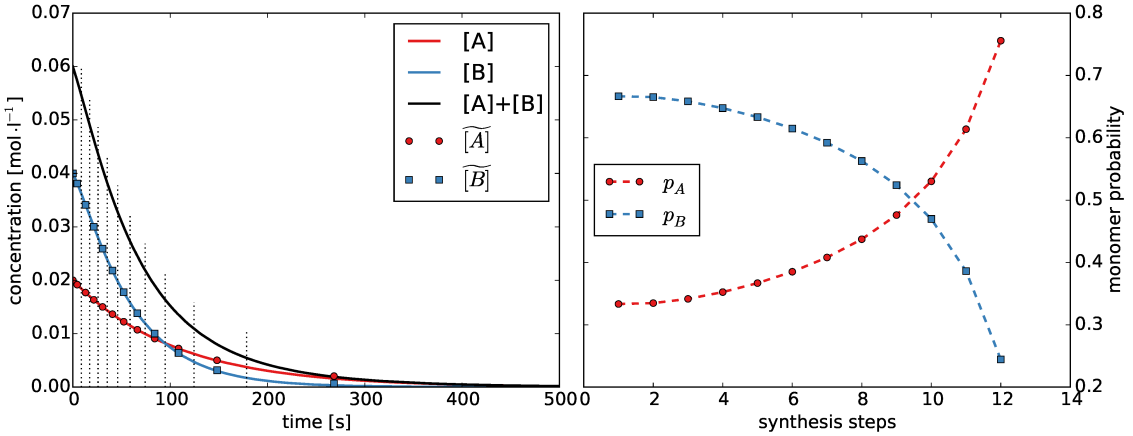


Figure 6.5 Left: Concentration of monomers [A] and [B] during the Monte-Carlo simulation with $r_A = 2.0$. We divided the time into discrete synthesis steps and determined the average concentrations $\widetilde{[\mathsf{A}]}$ and $\widetilde{[\mathsf{B}]}$. Right: Monomer probabilities p_A and p_B for each synthesis step calculated from the average concentrations.

6.1.5 Model Evaluation

Determining the model parameters allows us to compare the fingerprints computed by our models to the Monte-Carlo fingerprints (Fig. 6.6 and (Appendix Fig. A.12)). Additionally, we can compute the NRMSE of the Monte-Carlo fingerprints *vs.* the model fingerprints (Fig. 6.7, left).

Evidently, the reactivity parameters are crucial to model copolymerization. They determine the location and size of the distribution of abundances in the fingerprint. Both the Bernoulli and Geometric model fingerprints without the reactivity parameters have a significantly larger deviation than the models with reactivity parameters to the Monte-Carlo fingerprints, except for the instances with $r_\mathsf{A} = 1.0$. This is to be expected because in our setup this corresponds to reactivity parameters of $p_\mathsf{XY} = 1.0$ for all $\mathsf{A}, \mathsf{B} \in \{\mathsf{AB}\}$.

Overall, the Geometric model provides a better fit than the Bernoulli model for all fingerprints computed with $r_\mathsf{A} \geq 0.5$: The shapes of the distributions match closely and the deviations to the Monte-Carlo fingerprints are the lowest. For fingerprints computed with $r_\mathsf{A} < 0.5$, we observe the reverse: The Bernoulli model provides a better fit than the Geometric model for narrow distributions.

The Bernoulli and Geometric models are not only able to compute fingerprints, but also the likelihood of a single copolymer chain. Monte-Carlo simulations produce a large random sample of copolymer chains. This allows us to compute and compare the log likelihoods of the sampled data under the different models (Fig. 6.7, right) to further evaluate the models. A model that has a higher likelihood is "closer" to the sample.

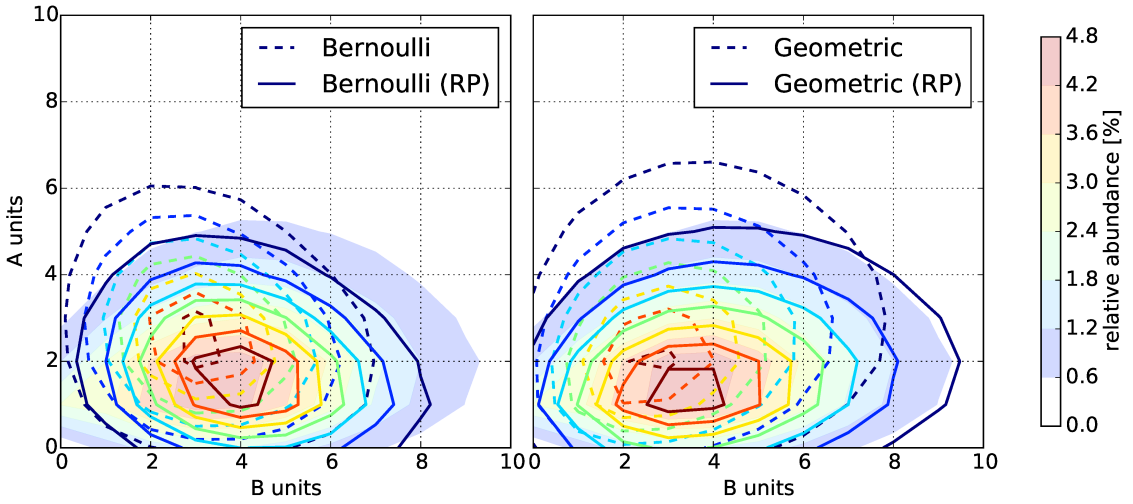


Figure 6.6 Copolymer fingerprint computed by the Monte-Carlo simulation with $r_A = 2.0$ (filled contours) compared to the fingerprints computed by the statistical models (solid and dashed contours). Left: Bernoulli model with and without reactivity parameters (RP). Right: Geometric model with and without reactivity parameters (RP).

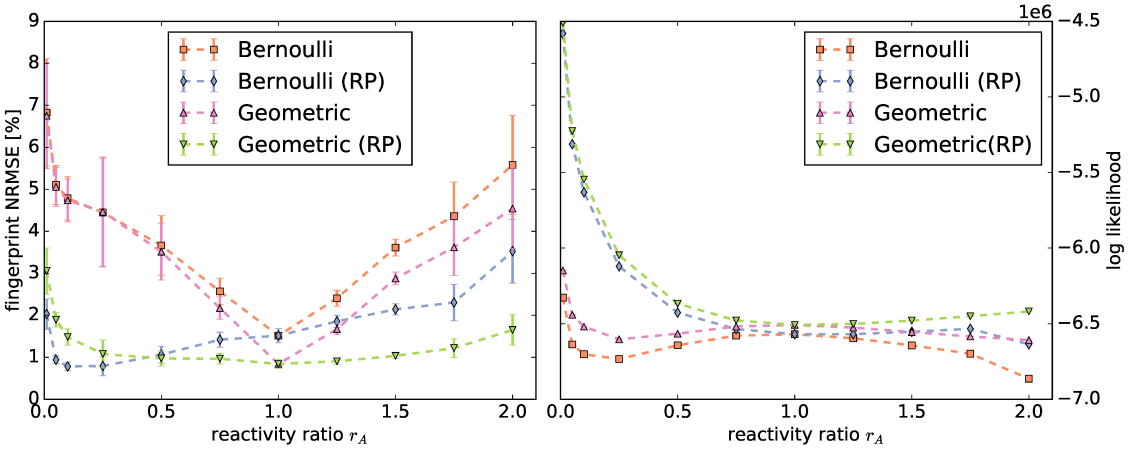


Figure 6.7 Left: Normalized root mean square error (NRMSE) of the copolymer fingerprints computed by Monte-Carlo simulations compared to the fingerprints computed by the statistical models. Right: Log likelihoods of the polymer chains produced by the Monte-Carlo simulations under the Bernoulli and Geometric models with and without reactivity parameters (RP). Note that the minimal and maximal log likelihoods are so close to the means, that the error bars are indiscernible.

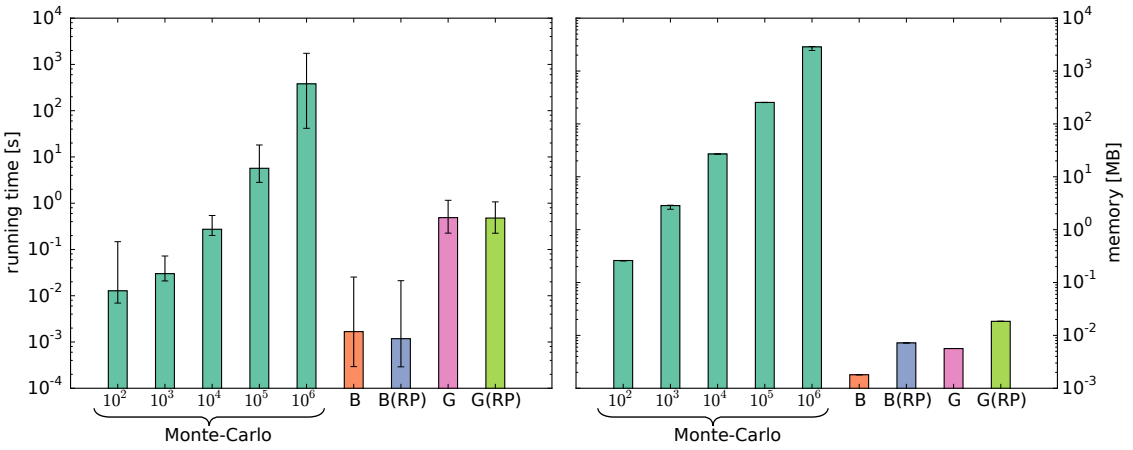


Figure 6.8 Comparison of the running time (left) and memory (right) measurements of the Monte-Carlo simulations using Gillespie's algorithm with 10^2 to 10^6 chains and the Bernoulli (B) and Geometric (G) models with and without reactivity parameters (RP).

Except for $r_{\mathsf{A}} = 1.0$, the log likelihood under the models without reactivity parameters are in all cases lower than their counterparts using reactivity parameters. This is consistent with the fingerprint comparisons. However, contrary to the fingerprint comparisons, the Geometric model has the best log likelihood for all instances.

The running time and memory requirements of a Monte-Carlo simulation increase with the number of simulated chains and for good accuracy the number should be high. The running time and memory of the Bernoulli and Geometric models are determined by the number of synthesis steps. Compared to the theoretical time complexity of $O(T^3)$ for the Bernoulli model, the Geometric model has a higher theoretical time complexity of $O(T^5)$. We measured running time (excluding I/O operations) and memory of the Monte-Carlo simulations with 10^2 to 10^6 chains and of our models (Fig. 6.8).

Computing the Bernoulli model is the fastest. As expected, the measured running time of the Geometric model is higher. However, computing the fingerprints with the Geometric model is still 11.8 and 788.5 times faster than the Monte-Carlo simulations with 10^5 and 10^6 chains, respectively. The reactivity parameters have no substantial impact on the running time. Both models require significantly less memory than the Monte-Carlo simulations. The additional matrices required for the reactivity parameters increase the memory consumption only slightly.

6.2 Independence of the Model Parameter Order

In the previous Section 6.1, we introduced a copolymerization model with several variants similar to a discrete Markov-chain, that append monomers in each synthesis

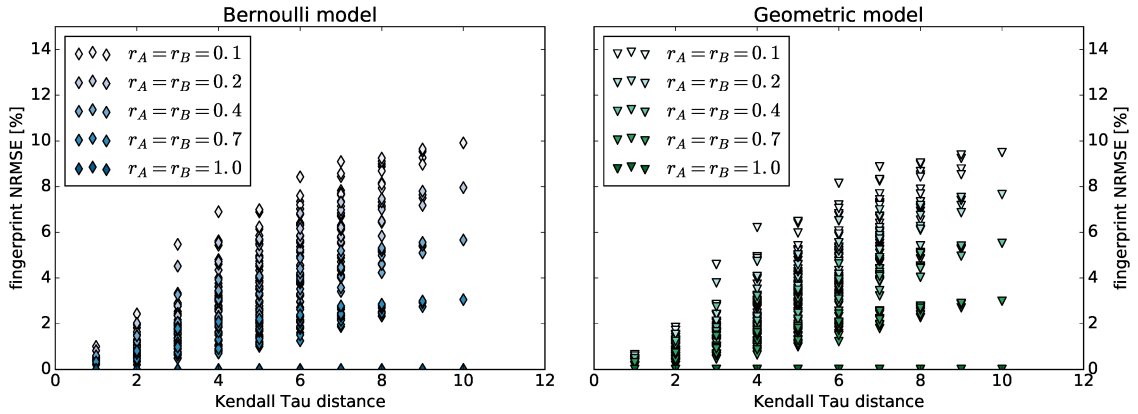


Figure 6.9 Normalized root mean square errors of the fingerprints for all permutations $\pi(p_\mathsf{A})$ compared to the fingerprint of the original p_A computed with the Bernoulli (left) and Geometric model (right). The Kendall Tau distance is the number of pairwise disagreements between two permutations.

(time) step with Bernoulli or geometrically distributed probability. Here, we will investigate if they are order-independent.

In the following, let the matrix M of size $n \times m$ be a copolymer fingerprint, in which entry $M_{a,b}$ gives the relative abundance of a copolymer with a monomers of type A and b monomers of type B. Let T be the number of synthesis steps. Let p_M be the probability of encountering a monomer, and let p_A be a vector of size T with the probabilities that the encountered monomer is an A for each synthesis step $1 \leq t \leq T$. Let $p_\mathsf{B}(t)$, the probability of encountering a monomer B be defined as $p_\mathsf{B}(t) = 1 - p_\mathsf{A}(t)$.

Let $\pi(x)$ be a permutation of some vector x. Let M^π be the resulting fingerprint of our model with input $\pi(p_\mathsf{A})$. We define a model to be *order-independent* if the resulting fingerprints are the same for any permutation of p_A, that is $M = M^\pi$ for any π.

We do a simple experiment to investigate the order-independence of our models. For both models, we compute a fingerprint with parameters $p_\mathsf{A} = [0, 0.1, 0.2, 0.4, 0.5]$, $p_M = 0.5$ and varying reactivity ratios. Subsequently, for all permutations $\pi(p_\mathsf{A})$ we compute a fingerprint and calculate the normalized root mean square error (NRMSE) in comparison to the first fingerprint (Fig. 6.9).

The distance between two fingerprints increases with the distance between p_A and $\pi(p_\mathsf{A})$. However, as the reactivity ratios approach one, the distance between the fingerprints decreases. In this experimental instance, we see that the models are order-independent if the reactivity ratios are one.

To verify that the models are order-independent for reactivity ratios of one, we investigate the model variants without reactivity parameters. For the Bernoulli

model without reactivity parameters, an entry $M_{a,b}$ in the fingerprint M at synthesis step t for $a > 0$, $b > 0$, and $1 \leq t \leq T$ is given by:

$$\begin{aligned} M_{a,b}(t) = &\quad p_M \cdot p_{\mathsf{A}}(t) \cdot M_{a-1,b}(t-1) \\ + &\quad p_M \cdot p_{\mathsf{B}}(t) \cdot M_{a,b-1}(t-1) \\ + &\quad (1-p_M) \cdot M_{a,b}(t-1) \end{aligned} \tag{6.24}$$

For the Geometric model without reactivity parameters, we first have to derive a closed form for $M_{a,b}$. Let $p_G(k)$ be the geometrically distributed probability of adding k monomers in one synthesis step. The probability of adding i monomers A and j monomers B to a copolymer chain is given for $a > 0$, $b > 0$, and $1 \leq t \leq T$ as:

$$P(M_{a,b} \to M_{a+i,b+j}|t) = \binom{i+j}{j} \cdot P_G(i+j) \cdot p_{\mathsf{A}}(t)^i \cdot p_{\mathsf{B}}(t)^j \tag{6.25}$$

We define $f_{i,j}^{a,b}$ as:

$$f_{i,j}^{a,b} = \binom{a+b-i-j}{b-j} \cdot P_G(a+b-i-j) \tag{6.26}$$

Now we apply Eq. 6.25 and Eq. 6.26 to find a closed form expression for a fingerprint entry $M_{a,b}$:

$$M_{a,b}(t) = \sum_{i=0}^{a} \sum_{j=0}^{b} f_{i,j}^{a,b} \cdot p_{\mathsf{A}}(t)^{a-i} \cdot p_{\mathsf{B}}(t)^{b-j} \cdot M_{i,j}(t-1) \tag{6.27}$$

Now that we are given the equations for computing an entry in the fingerprint at a specific synthesis step for both models without reactivity parameters, we can show that an inversion of neighboring values in p_{A} does not change the resulting fingerprint.

Lemma 1. *Given the Bernoulli model without reactivity parameters and a permutation $\pi(p_{\mathsf{A}})$ that swaps $p_{\mathsf{A}}(t)$ with $p_{\mathsf{A}}(t-1)$, then $M_{a,b}(t) = M_{a,b}^{\pi}(t)$ holds for all $a > 0$, $b > 0$, and $2 \leq t \leq T$.*

Proof. Inserting $M_{a,b}(t-1)$ into the recursive equation 6.24 yields:

$$
\begin{aligned}
M_{a,b}(t) = \quad & p_M^2 \cdot p_{\mathsf{A}}(t) \cdot p_{\mathsf{A}}(t-1) \cdot M_{a-2,b}(t-2) \\
+ \quad & p_M^2 \cdot p_{\mathsf{B}}(t) \cdot p_{\mathsf{B}}(t-1) \cdot M_{a,b-2}(t-2) \\
+ \quad & p_M^2 \cdot p_{\mathsf{A}}(t) \cdot p_{\mathsf{B}}(t-1) \cdot M_{a-1,b-1}(t-2) \\
+ \quad & p_M^2 \cdot p_{\mathsf{B}}(t) \cdot p_{\mathsf{A}}(t-1) \cdot M_{a-1,b-1}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot p_{\mathsf{A}}(t) \cdot M_{a-1,b}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot p_{\mathsf{A}}(t-1) \cdot M_{a-1,b}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot p_{\mathsf{B}}(t) \cdot M_{a,b-1}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot p_{\mathsf{B}}(t-1) \cdot M_{a,b-1}(t-2) \\
+ \quad & (1-p_M)^2 \cdot M_{a,b}(t-2)
\end{aligned}
\tag{6.28}
$$

We replace $p_{\mathsf{A}}(t-1)$ with $\pi(p_{\mathsf{A}})(t)$, $p_{\mathsf{A}}(t)$ with $\pi(p_{\mathsf{A}})(t-1)$, $p_{\mathsf{B}}(t-1)$ with $\pi(p_{\mathsf{B}})(t)$, and $p_{\mathsf{B}}(t)$ with $\pi(p_{\mathsf{B}})(t-1)$:

$$
\begin{aligned}
M_{a,b}(t) = \quad & p_M^2 \cdot \pi(p_{\mathsf{A}})(t) \cdot \pi(p_{\mathsf{A}})(t-1) \cdot M_{a-2,b}(t-2) \\
+ \quad & p_M^2 \cdot \pi(p_{\mathsf{B}})(t) \cdot \pi(p_{\mathsf{B}})(t-1) \cdot M_{a,b-2}(t-2) \\
+ \quad & p_M^2 \cdot \pi(p_{\mathsf{A}})(t) \cdot \pi(p_{\mathsf{B}})(t-1) \cdot M_{a-1,b-1}(t-2) \\
+ \quad & p_M^2 \cdot \pi(p_{\mathsf{B}})(t) \cdot \pi(p_{\mathsf{A}})(t-1) \cdot M_{a-1,b-1}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot \pi(p_{\mathsf{A}})(t) \cdot M_{a-1,b}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot \pi(p_{\mathsf{A}})(t-1) \cdot M_{a-1,b}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot \pi(p_{\mathsf{B}})(t) \cdot M_{a,b-1}(t-2) \\
+ \quad & p_M \cdot (1-p_M) \cdot \pi(p_{\mathsf{B}})(t-1) \cdot M_{a,b-1}(t-2) \\
+ \quad & (1-p_M)^2 \cdot M_{a,b}(t-2)
\end{aligned}
\tag{6.29}
$$

We simplify the equation to:

$$
\begin{aligned}
M_{a,b}(t) = \quad & p_M \cdot \pi(p_{\mathsf{A}})(t) \cdot M_{a-1,b}(t-1) \\
+ \quad & p_M \cdot \pi(p_{\mathsf{B}})(t) \cdot M_{a,b-1}(t-1) \\
+ \quad & (1-p_M) \cdot M_{a,b}(t-1)
\end{aligned}
\tag{6.30}
$$

Which can be further simplified to:

$$
M_{a,b}(t) = M^{\pi}_{a,b}(t) \tag{6.31}
$$

□

Lemma 2. *Given the Geometric model without reactivity parameters and a permutation $\pi(p_{\mathsf{A}})$ that swaps $p_{\mathsf{A}}(t)$ with $p_{\mathsf{A}}(t-1)$, then $M_{a,b}(t) = M^{\pi}_{a,b}(t)$ holds for all $a > 0$, $b > 0$, and $2 \leq t \leq T$.*

Proof Sketch. Inserting $M_{i,j}(t-1)$ into the recursive equation 6.27 yields:

$$\begin{aligned} M_{a,b}(t) = &\sum_{i=0}^{a}\sum_{j=0}^{b} f_{i,j}^{a,b} \cdot p_{\mathsf{A}}(t)^{a-i} \cdot p_{\mathsf{B}}(t)^{b-j} \\ &\cdot \sum_{k=0}^{i}\sum_{l=0}^{j} f_{k,l}^{i,j} p_{\mathsf{A}}(t-1)^{i-k} \cdot p_{\mathsf{B}}(t-1)^{j-l} \cdot M_{k,l}(t-2) \end{aligned} \tag{6.32}$$

Writing the terms of the sums explicitly yields a large equation of the following form:

$$\begin{aligned} M_{a,b}(t) = &f_{0,0}^{a,b} p_{\mathsf{A}}(t)^a p_{\mathsf{B}}(t)^b \\ &+ f_{0,1}^{a,b} p_{\mathsf{A}}(t)^a p_{\mathsf{B}}(t)^{b-1} \left(f_{0,0}^{0,1} M_{0,0}(t-2) + f_{0,1}^{0,1} p_{\mathsf{B}}(t) M_{0,1}(t-2)\right) \\ &+ f_{1,0}^{a,b} p_{\mathsf{A}}(t)^{a-1} p_{\mathsf{B}}(t)^b \left(f_{0,0}^{1,0} M_{0,0}(t-2) + f_{1,0}^{1,0} p_{\mathsf{A}}(t) M_{1,0}(t-2)\right) \\ &+ f_{1,1}^{a,b} p_{\mathsf{A}}(t)^{a-1} p_{\mathsf{B}}(t)^{b-1} \left(f_{0,0}^{1,1} M_{0,0}(t-2) + f_{1,0}^{1,1} p_{\mathsf{A}}(t) M_{1,0}(t-2)\right. \\ &\qquad \left. + f_{0,1}^{1,1} p_{\mathsf{B}}(t) M_{0,1}(t-2) + f_{1,1}^{1,1} p_{\mathsf{A}}(t) p_{\mathsf{B}}(t) M_{1,1}(t-2)\right) \\ &+ \dots \\ &+ f_{a,b}^{a,b} \left(f_{0,0}^{a,b} M_{0,0}(t-2) + \dots + f_{a,b}^{a,b} p_{\mathsf{A}}(t-1)^a p_{\mathsf{B}}(t-1)^b M_{a,b}(t-2)\right) \end{aligned} \tag{6.33}$$

If we now expand this equation, we see that for every term of the form $p_{\mathsf{A}}(t)^{\alpha} p_{\mathsf{B}}(t)^{\beta}$ $p_{\mathsf{A}}(t-1)^{\gamma} p_{\mathsf{B}}(t-1)^{\delta}$ there is a corresponding term $p_{\mathsf{A}}(t)^{\gamma} p_{\mathsf{B}}(t)^{\delta} p_{\mathsf{A}}(t-1)^{\alpha} p_{\mathsf{B}}(t)^{\beta}$ and we change equation 6.32 to:

$$\begin{aligned} M_{a,b}(t) = &\sum_{i=0}^{a}\sum_{j=0}^{b} f_{i,j}^{a,b} \cdot p_{\mathsf{A}}(t-1)^{a-i} \cdot p_{\mathsf{B}}(t-1)^{b-j} \\ &\cdot \sum_{k=0}^{i}\sum_{l=0}^{j} f_{k,l}^{i,j} p_{\mathsf{A}}(t)^{i-k} \cdot p_{\mathsf{B}}(t)^{j-l} \cdot M_{k,l}(t-2) \end{aligned} \tag{6.34}$$

We replace $p_{\mathsf{A}}(t-1)$ with $\pi(p_{\mathsf{A}})(t)$, $p_{\mathsf{A}}(t)$ with $\pi(p_{\mathsf{A}})(t-1)$, $p_{\mathsf{B}}(t-1)$ with $\pi(p_{\mathsf{B}})(t)$, and $p_{\mathsf{B}}(t)$ with $\pi(p_{\mathsf{B}})(t-1)$:

$$\begin{aligned} M_{a,b}(t) = &\sum_{i=0}^{a}\sum_{j=0}^{b} f_{i,j}^{a,b} \cdot \pi(p_{\mathsf{A}})(t)^{a-i} \cdot \pi(p_{\mathsf{B}})(t)^{b-j} \\ &\cdot \sum_{k=0}^{i}\sum_{l=0}^{j} f_{k,l}^{i,j} \pi(p_{\mathsf{A}})(t-1)^{i-k} \cdot \pi(p_{\mathsf{B}})(t-1)^{j-l} \cdot M_{k,l}(t-2) \end{aligned} \tag{6.35}$$

We simplify the equation to:

$$M_{a,b}(t) = M^{\pi}_{a,b}(t) \tag{6.36}$$

□

For the models without reactivity parameters, we know from Lemma 1 and 2 that no inversion of neighboring values in p_{A} changes the resulting fingerprint. Any permutation of a vector can be constructed by a sequence of inversions of neighboring elements. Therefore, for the Bernoulli and Geometric model without reactivity parameters, all permutations of a probability vector p_{A} have the same resulting fingerprint.

6.3 Exploring the Limits of the Geometric Copolymerization Model

In this section, we focus on the Geometric model with reactivity parameters (Section 6.1). We show that determining the model parameters from copolymer fingerprints is a challenging optimization problem. First, several methods are presented to increase the accuracy of the results and to decrease the running times. Several general purpose optimization algorithms and the robustness of the proposed methods against measurement noise are evaluated. Second, the accuracy of the Geometric model is evaluated using different copolymerization types beyond living polymerization: Reversible living polymerization, controlled radical polymerization, and free radical polymerization. The evaluation uses fingerprints and copolymer chains computed by Monte-Carlo simulations.

6.3.1 Objective Function

In the following, let the matrix M of size $n \times m$ be a copolymer fingerprint, in which entry $M_{a,b}$ gives the relative abundance of a copolymer with a monomers of type A and b monomers of type B. Let $f(p_{\mathsf{A}}, p_{\mathsf{AA}}, p_{\mathsf{AB}}, p_{\mathsf{BA}}, p_{\mathsf{BB}}) = M^c$ be the *fingerprint-generating function*, which uses the Geometric model with reactivity parameters to compute a fingerprint M^c. The model parameters are the monomer probability p_M, the reactivity probabilities p_{AA}, p_{AB}, p_{BA}, p_{BB}, and probability vector p_A of size T, which describes the probability of encountering an A-monomer for each synthesis step $1 \leq t \leq T$. The probability of encountering a B-monomer is implicitly given, because $p_{\mathsf{A}}(t)+p_{\mathsf{B}}(t) = 1$. The monomer probability p_M and the number of synthesis steps T can be easily computed from the copolymer length distribution.

Formally, the problem to solve is finding the parameters p_{AA}, p_{AB}, p_{BA}, p_{BB}, and the vector p_{A}, which minimize the distance of the computed fingerprint M^c to an observed fingerprint M^o. This corresponds to optimizing the following *objective function*:

$$\underset{p_{\mathsf{A}}, p_{\mathsf{AA}}, p_{\mathsf{AB}}, p_{\mathsf{BA}}, p_{\mathsf{BB}}}{\arg\min} \quad ||f(p_{\mathsf{A}}, p_{\mathsf{AA}}, p_{\mathsf{AB}}, p_{\mathsf{BA}}, p_{\mathsf{BB}}) - M^o||_2^2 \tag{6.37}$$

The objective function computes the difference between the computed and observed fingerprints according to Eq. 6.37. We use general purpose optimizers to identify the best parameters. The optimizers use different strategies and the running times vary greatly, in the small examples given in this work between 0.5 and 19 hours (Appendix Fig. A.13 to A.15). Generally, the optimization is challenging and its computation is time-demanding. First, the question needs to be answered: What are the main reasons for the long running time?

In Section 6.1, we introduced four variants of a discrete Markov chain copolymerization model. The models use either reactivity probabilities or not, and the number of added monomers per synthesis step either follows a Bernoulli or geometric distribution. A model is defined to be *order-independent* if the resulting fingerprints are the same for any permutation of its parameter p_A. The models are order-independent if the reactivity ratios are one (Section 6.2). Since there are $T!$ possible permutations, this results in $T!$ global optima. But for reactivity ratios of one, the ratios of monomers never change. As a consequence, p_A is constant and there is exactly one global optimum. However, for reactivity ratios near one, the objective values of all permutations are very similar. This is challenging for the optimization algorithms and certainly contributes to the long running time of the optimization.

Another contributing factor is the size T of the vector p_A, resulting in a T-dimensional search space. T can be computed from the observed copolymer length distribution. The length distribution of the Geometric model is a negative binomial distribution with the parameters T and p_M. In each of the T steps, the number of added monomers is geometrically distributed. Considering usual copolymer lengths, T can be expected to be between 10 and 100. Optimizing ~ 100 variables simultaneously with a general purpose optimizer is a challenging task and certainly contributes to the long running time of the optimization.

6.3.2 Parameter Space Reduction

The two main challenges for the optimization algorithms are the very similar objective values for reactivity ratios near one and – more importantly – the large search space defined by the length of the model parameter vector p_A. We focus on the second challenge and propose two approaches to change the fingerprint-generating function in order to speed up the optimization.

The first approach is to optimize only a fraction of the T values in p_A (25% in this work), and linearly interpolate all other values in between. Furthermore, we restrict the search space by forcing p_A to be either increasing or decreasing. To this end, a decreasing p_A is defined as:

$$\begin{aligned} p_\mathsf{A}(t) &= p(t) \cdot p_\mathsf{A}(t-1) \\ p_\mathsf{A}(1) &= p(1) \end{aligned} \tag{6.38}$$

And an increasing p_A as:

$$\begin{aligned} p_\mathsf{A}(t) &= p_\mathsf{A}(t-1) + p(t) \cdot (1 - p_\mathsf{A}(t-1)) \\ p_\mathsf{A}(1) &= p(1) \end{aligned} \tag{6.39}$$

The second approach is to exploit the relationship between p_{A} and monomer concentrations. We define T time intervals, such that the change in concentration is the same for each interval. Subsequently, the mean concentrations $\widetilde{[\mathsf{A}]}(t)$ and $\widetilde{[\mathsf{B}]}(t)$ are calculated for each interval $1 \leq t \leq T$. Then, the probability vector $p_{\mathsf{A}}(t)$ can be calculated as:

$$p_{\mathsf{A}}(t) = \frac{\widetilde{[\mathsf{A}]}(t)}{\widetilde{[\mathsf{A}]}(t) + \widetilde{[\mathsf{B}]}(t)} \tag{6.40}$$

There is also a relationship between the reaction rates and the reactivity model parameters. For $\mathsf{X}, \mathsf{Y} \in \{\mathsf{A}, \mathsf{B}\}$, the reactivity parameters are:

$$p_{\mathsf{XY}} = \frac{k_{\mathsf{XY}}}{k_{\mathsf{XA}} + k_{\mathsf{XB}}} \tag{6.41}$$

The second approach uses both relationships: First, an ODE system using the living copolymerization reaction scheme is solved. Second, the reactivity parameters are computed from the reaction rates and p_{A} from the concentration gradient. Then, the fingerprint M^c can be computed using the Geometric model. This allows us to optimize the ODE parameters (reaction rates and initial concentrations) according to Eqn. 6.37. Thus, the dimension of the search space is constant and independent of T.

6.3.3 Parameter Optimization

In the following, we compare three fingerprint-generating functions: Directly optimizing p_{A} (Direct), interpolating p_{A} (Spline), and optimizing the ODE parameters (ODE), with the Spline and ODE approaches as described above. All of the three functions use the Geometric model with reactivity parameters to compute the copolymer fingerprint. The transformation from the model parameters to the copolymer fingerprint is highly non-linear. To the best of our knowledge, no special purpose solvers exist for such a function. Therefore, we have to resort to general purpose optimization algorithms. We use the algorithms implemented in the Optimization Algorithm Toolkit[1] [15] and Apache Math Commons 3.2 library[2] with their default parameters. The algorithms use different strategies to find the best model parameters and do not require computing gradients. The performance of the optimizers is application-specific and depends on the selected fingerprint-generating function.

We choose several instances with low degree of polymerization $DP_n = 3$ and three different reactivity ratios r_{A}, r_{B} and homopropagation ratios r. Please note that for all datasets $r_{\mathsf{A}} = \frac{1}{r_{\mathsf{B}}} = r$. First, we choose the reactivity ratio $r_{\mathsf{A}} = 2.0$, for which the Geometric model can provide a good fit. Second, we choose $r_{\mathsf{A}} = 0.01$, since this results in a copolymer with binomial-like length distribution (in contrast to a more common Schulz-Zimm-like distribution), which should be more challenging for

[1] `https://sourceforge.net/projects/optalgtoolkit/`
[2] `http://commons.apache.org/proper/commons-math/`

	Abbrv.	Algorithm	#Ranks		
			1st	2nd	3rd
Direct	CLI	Cloning, Information Gain, Aging [18]	4	5	7
	PC	Probabilistic Crowding [61]	6	5	5
	RTS	Restricted Tournament Selection [40]	6	6	4
Spline	CMAES	Covariance Matrix Adaptation Evolution Strategy [38]	3	6	7
	DC	Deterministic Crowding [55]	3	8	5
	GEO	Generalized Extremal Optimization [19]	10	2	4
ODE	GA	Genetic Algorithm [3]	5	9	2
	GEO	Generalized Extremal Optimization [19]	8	0	8
	MHC	Mutation Hill Climber [69]	3	7	6

Table 6.1 Overview of the top three optimization algorithms for each fingerprint-generating function, selected based on Appendix Fig. A.13 to A.15. We ranked the results of the algorithms for each dataset based on the log likelihood ratios and counted the ranks.

the Geometric model. Third, we choose $r_\mathsf{A} = 1.0$. This results in constant monomer concentrations and, thus, the optimal p_A is also constant. That means the optimum lies on the parameter space limits when using the spline fingerprint-generating function, which should be a challenging task for the optimizers. Furthermore, we also select two instances with $r_\mathsf{A} = 2.0$ and higher degrees of polymerization $DP_n = 25$ and $DP_n = 45$, which are copolymer lengths to be expected in practice.

First, we choose the top three algorithms with highest log likelihood ratio for each fingerprint-generating function (Table 6.1). To this end, all algorithms are evaluated on the $DP_n = 3, r_\mathsf{A} = 2.0$ dataset without noise (Appendix Fig. A.13 to A.15) and the log likelihood ratios of the results are calculated. Additionally to comparing the log likelihoods, the ratio also acts as a "sanity check" for the model parameterizations. The ratio compares the likelihoods to the likelihood of a null hypothesis. The null hypothesis assumes, all positions are independent random variables. If the log likelihood ratio is below zero, the null model has a higher likelihood and the parameterization should be dismissed.

After selecting the top three algorithms for each fingerprint-generating function, we evaluate the robustness of the chosen algorithms. We run the top three algorithms for each function on the other dataset with increasing simulated noise. The highest noise level with $\sigma = 0.25$ results in strongly perturbed data (Fig. 6.10 and Appendix Fig. A.16 to A.18). For each resulting parameterization, we rank the top three algorithms by their log likelihood ratio and count the ranks for all instances (Table

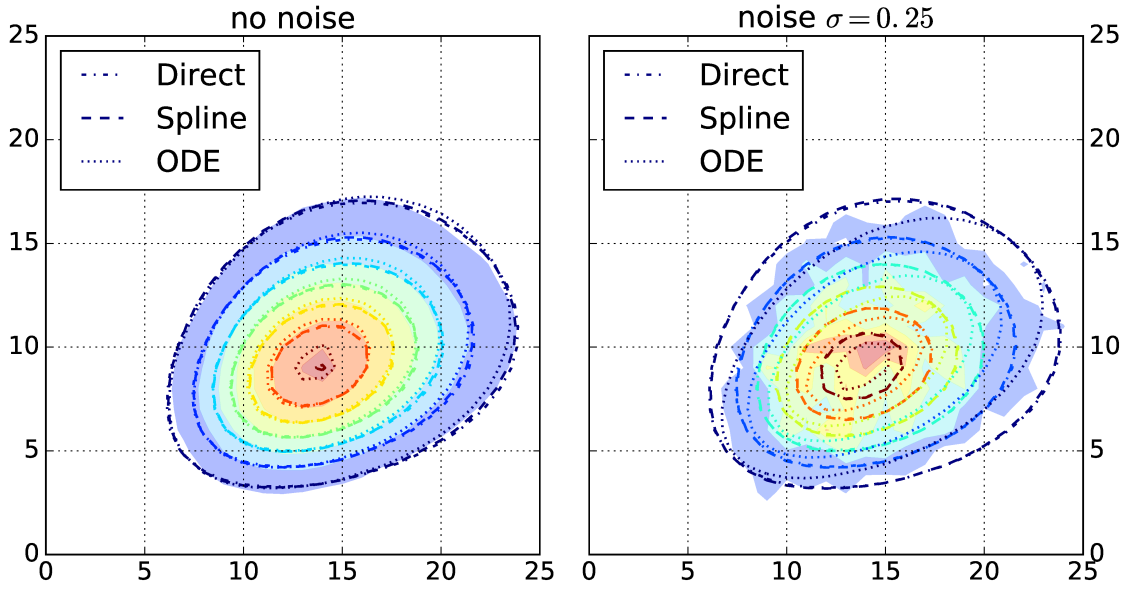


Figure 6.10 Filled contours: Copolymer fingerprints of $DP_n = 25$ computed by Monte-Carlo simulations with no (left) and high applied noise (right). Contours: Fingerprints computed by the Geometric model using the best parameters computed by the optimization algorithms for each of the fingerprint-generating functions (direct, spline, and ODE).

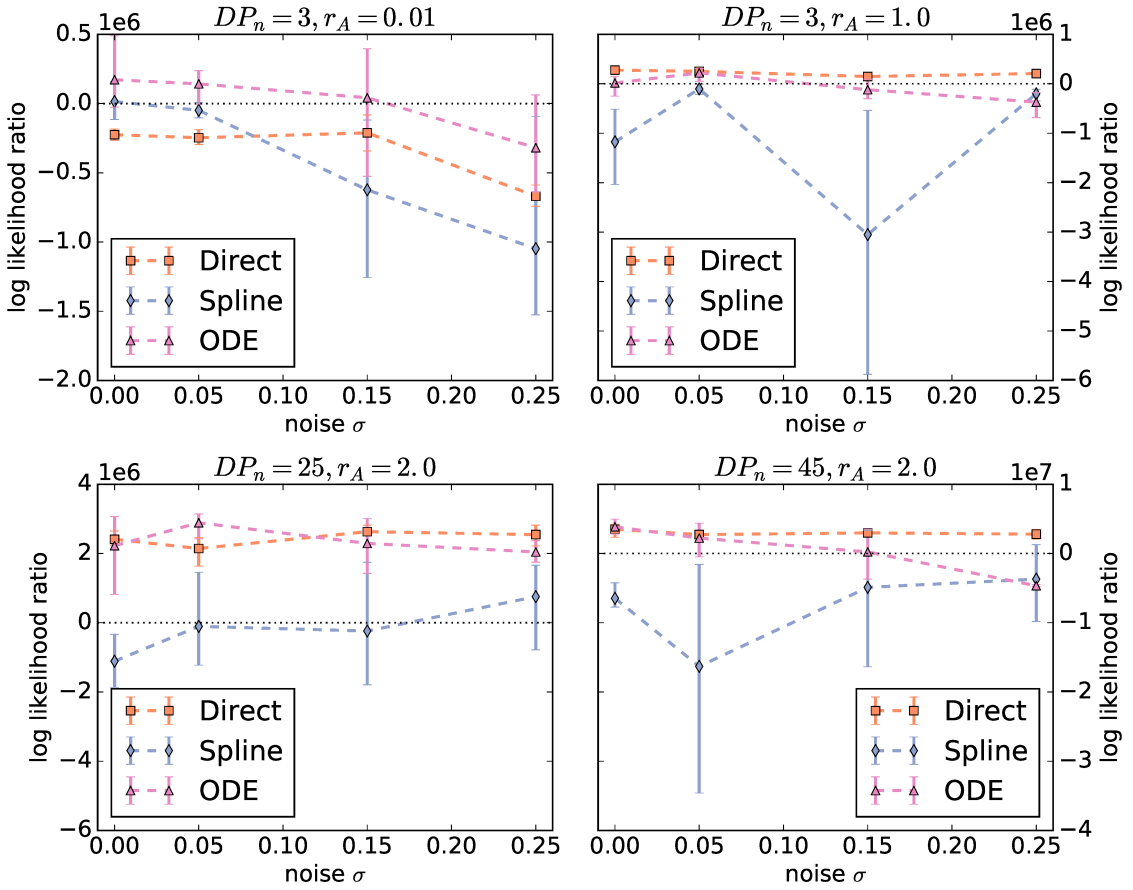


Figure 6.11 Log likelihood ratios of the results computed by the optimization algorithms as a function of noise. The ratios are averaged over all three algorithms for each fingerprint-generating function (direct, spline, ODE). The higher the ratios, the better the observed data is "explained" by the identified model parameterizations. If the ratio is below zero, the null model achieves a higher likelihood than the Geometric model with the given parameterization.

6.1). No algorithm outperforms its rivals. Therefore, in the following, we use all chosen algorithms.

To compare the three approaches (direct, spline, and ODE), we average the log likelihood ratios over all three algorithms for each fingerprint-generating function. Fig. 6.11 shows the averaged log likelihood ratios as a function of the noise level. There are two different behaviors for $r_\mathrm{A} = 0.01$ and the rest of the instances. For $r_\mathrm{A} = 0.01$ there is a significant decrease with increasing noise and only the ODE function is able to produce a good parameterization. For the Schulz-Zimm like copolymers with $r_\mathrm{A} > 0.01$, the behavior of the log likelihood ratios of the ODE and direct function is not significantly different. However, for the ODE function, the range between minimum and maximum log likelihood ratio is larger and the ratio

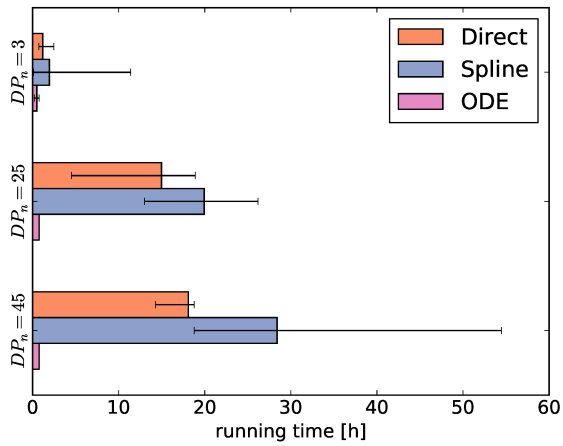


Figure 6.12 Running times until convergence of the optimizations averaged over all datasets with degree of polymerization $DP_n = 3$, 25, and 45 for each fingerprint-generating function (direct, spline, ODE).

decreases more with increasing noise. Thus, using the ODE function is less robust against noise than the direct method. Unexpectedly, the optimizers using the spline function fail on all instances and result in ratios below zero in almost all cases.

Then, we average the running times until convergence for each fingerprint-generating function for each degree of polymerization $DP_n = 3$, 25, and 45 (Fig. 6.12). As the running times largely depend on the selected optimization algorithms, the comparison of running times between the fingerprint-generating functions should be taken with a grain of salt. That means, using different optimizers may shift the numbers, but we can still infer general trends from Fig. 6.12.

The running times until convergence of the optimizers using the direct and ODE functions behave as expected. The running time using the direct function increases with the degree of polymerization, because the size of p_A increases. Thus, the number of parameters increases, the main reason for the long running time. In contrast, the ODE function always has the same number of parameters and therefore the running time is independent of the degree of polymerization. Different from our expectations, the using the spline function results in even higher running times than using the direct function, despite optimizing only a fraction of the p_A parameter values and using the generally fast optimizers CMAES and GEO (Appendix Fig. A.13 to A.15).

The optimization converges fastest using the ODE fingerprint-generating function. Now we further investigate its robustness. We run the top-three algorithms on four times on the $DP_n = 25$, $r_A = 2.0$ dataset and record the normalized scores of the best solutions over time (Fig. 6.13). We computed log likelihood ratios and normalized scores for the final solutions (Table 6.2). The normalized score is the objective function value (Eqn. 6.37) of a final solution divided by the objective function value of its starting point. Thus, the score is 1.0 if the optimization failed

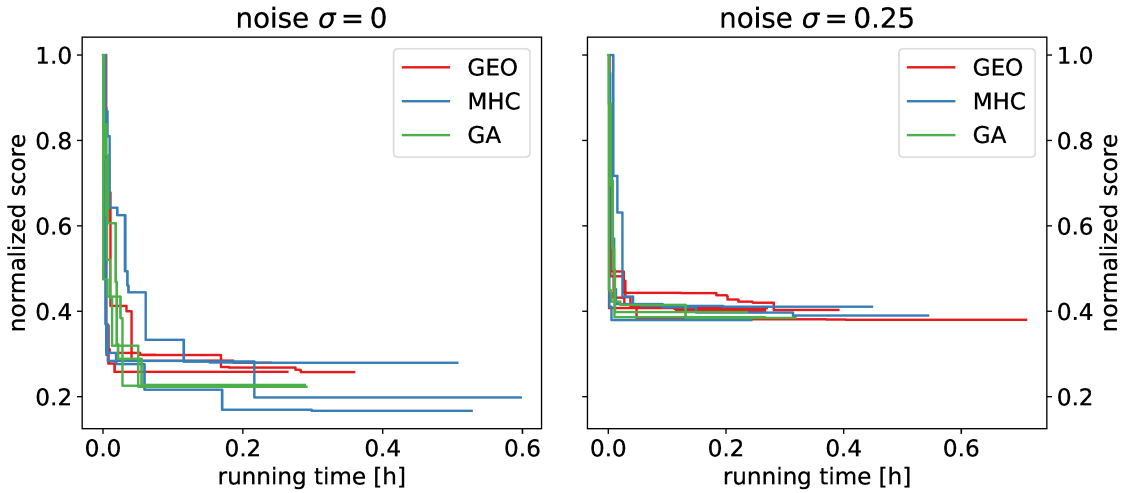


Figure 6.13 Normalized scores of the best solutions as a function of running time of four repeated runs for each top-three algorithm for the ODE fingerprint-generating function. The normalized score is the objective function value (Eqn. 6.37) of the current best solution divided by the objective function value of its starting point.

	Algorithm	**Normalized Score**		**Log likelihood ratio**	
		Mean	Standard dev.	Mean	Standard dev.
$\sigma = 0$	GEO	0.26	0.009	$4.99 \cdot 10^5$	$3.49 \cdot 10^4$
	GA	0.23	0.002	$4.86 \cdot 10^5$	$8.63e \cdot 10^4$
	MHC	0.23	0.050	$4.21 \cdot 10^5$	$7.67e \cdot 10^4$
$\sigma = 0.25$	GEO	0.4	0.011	$1.54 \cdot 10^5$	$2.63 \cdot 10^4$
	GA	0.39	0.006	$-4.28 \cdot 10^5$	$4.07 \cdot 10^5$
	MHC	0.40	0.014	$-4.11 \cdot 10^4$	$2.07 \cdot 10^5$

Table 6.2 Mean and standard deviation of the final solutions of the top three optimization algorithms for the ODE fingerprint-generating function (selected based on Appendix Fig. A.15) for no ($\sigma = 0$) and high noise ($\sigma = 0.25$).

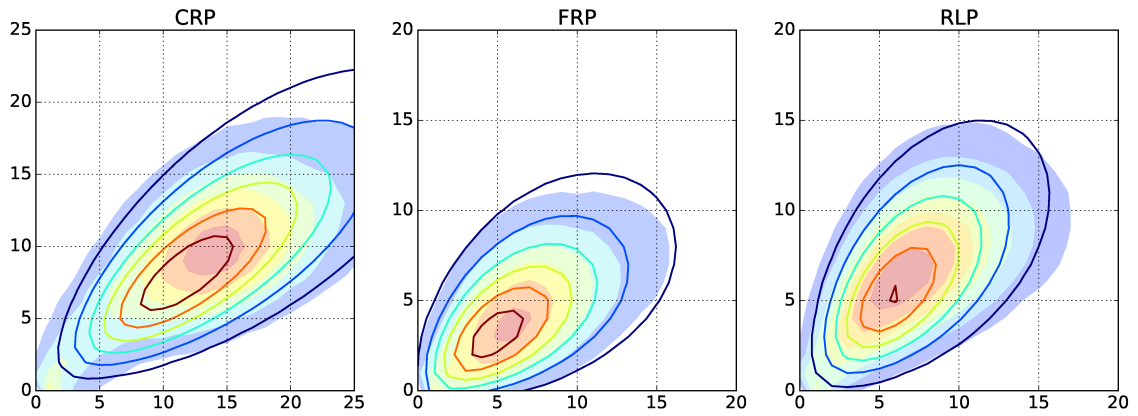


Figure 6.14 Filled contours: Copolymer fingerprints of the Monte-Carlo simulations of controlled radical polymerization (CRP, left), free radical polymerization (FRP, center), and reversible living polymerization (RLP, right) with the highest used termination and propagation reaction rates of 0.1. Contours: Fingerprints computed by the model with the best parameters resulting from the optimizations using the ODE fingerprint-generating function.

to improve the solution, and it approaches 0 if the solution matches the observed data perfectly.

For no noise we obtain good results. The scores and likelihood ratios are quite close with a low standard deviation. However, for high noise we obtain higher scores and lower log likelihood ratios. Additionally, the solutions of two of the three optimizers show for high noise a high standard deviation of log likelihood ratios and are unable to find an acceptable solution on average. Thus, we conclude that while the ODE fingerprint-generating function is fast and robust to some extend, it still shows to be a difficult objective function for the general purpose optimizers and optimization runs should be repeated several times.

6.3.4 Beyond Living Polymerization

Here, we investigate copolymerizations beyond a simple living polymerization. We select the $DP_n = 25$, $r_\mathrm{A} = 2.0$ instance and repeatedly run Monte-Carlo simulations with increasing termination and depropagation rates. For radical polymerizations, long and short length chains appear as a result of the termination by recombination and disproportionation, respectively. For free radical polymerization, the chosen decomposition rate of the initiator leads to lower average lengths. For reversible living polymerization, low length chains are appearing because of the depropagation reactions (Fig. 6.14 and Appendix Fig. A.19 to A.21).

We select the ODE method to identify the optimal model parameters. Fig. 6.15 shows the log likelihoods and log likelihood ratios averaged over the top three algorithms for the ODE method as a function of termination and depropagation reaction

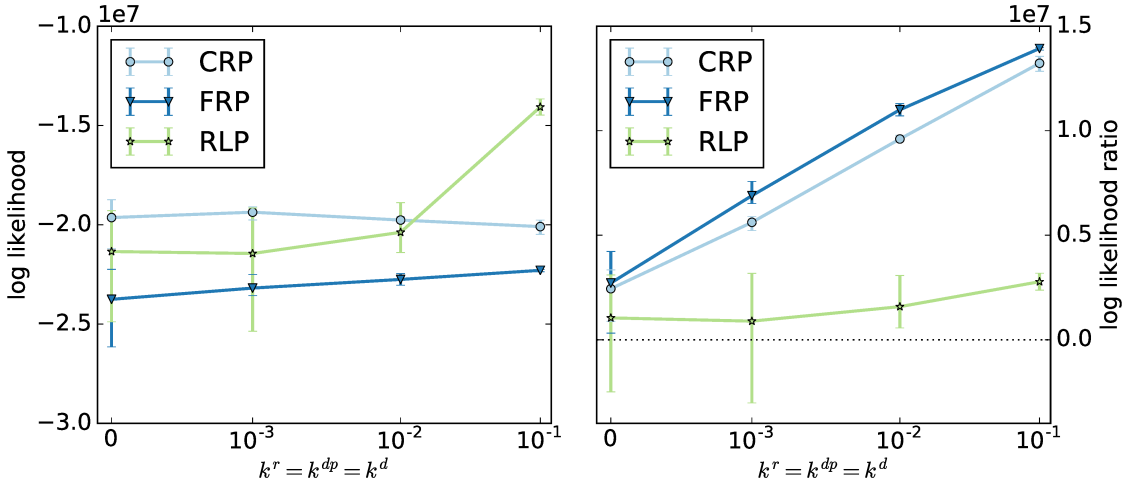


Figure 6.15 Log likelihoods (left) and log likelihood ratios (right) of the results from the optimizations using the ODE fingerprint-generating function for the controlled radical polymerization (CRP), free radical polymerization (FRP), and reversible living polymerization (RLP) as a function of termination and depropagation rates.

rates. The radical and reversible living polymerizations show different behaviors. For radical polymerization, the log likelihood is almost constant, but the ratio increases significantly. For reversible living polymerization, the likelihood increases significantly, but the ratio increases less.

Different from our expectations, the log likelihood ratios of all three copolymerization types increase with increasing termination and depropagation rates, due to a decreasing likelihood of the null model. We find that the Geometric model can be applied for systems involving termination and depropagation reactions, even though it was designed for living copolymerization.

6.4 Computing Copolymer Statistics

In this section, we focus on the Geometric model with reactivity parameters (Section 6.1). We show how the model may be used to compute interesting properties of the copolymer. We give three examples – the average sequence, dimer ratios and block length distributions – and compare them against Monte-Carlo simulations.

6.4.1 Average Sequence

In Section 6.1 we showed how to compute copolymer fingerprints using our model. A copolymer fingerprint represents a two-dimensional distribution of the abundances of each possible combination of monomer counts. However, it does not show the sequences of the polymer chains. Here, we show how to compute the average copolymer chain sequence from our model. Let $\tilde{S}$ be the *average sequence* and $p(\tilde{S}_k = \mathsf{X})$ the probability of observing monomer $\mathsf{X} \in \{\mathsf{A}, \mathsf{B}\}$ at position k in the average sequence $\tilde{S}$.

Let us recall the definition of the transition probability $\mathbb{P}(M^{\mathsf{X}}_{a,b} \to M^{\mathsf{Y}}_{a+i,b+j}; t)$ (Eqn. 6.19) as

$$\mathbb{P}(M^{\mathsf{X}}_{a,b} \to M^{\mathsf{Y}}_{a+i,b+j}; t) = c_{\mathsf{X}}(i+j) \cdot R^{\mathsf{XY}}_{i,j} \cdot p_G(i+j) \cdot p_{\mathsf{A}}(t)^i \cdot p_{\mathsf{B}}(t)^j, \tag{6.42}$$

where R^{XY} (Eqn. 6.15) is the matrix

$$\begin{aligned} R^{\mathsf{XA}}_{a,b} &= R^{\mathsf{XA}}_{a-1,b} \cdot p_{\mathsf{AA}} + R^{\mathsf{XB}}_{a-1,b} \cdot p_{\mathsf{BA}} \\ R^{\mathsf{XB}}_{a,b} &= R^{\mathsf{XA}}_{a,b-1} \cdot p_{\mathsf{AB}} + R^{\mathsf{XB}}_{a,b-1} \cdot p_{\mathsf{BB}}, \end{aligned} \tag{6.43}$$

$p_G(k)$ is the probability of adding k monomers, $p_{\mathsf{A}}(t)$, $p_{\mathsf{B}}(t)$ the monomer probabilities, p_{XY} the reactivity parameters and c_{X} a normalization factor.

To compute the probability of adding monomer X at position k in synthesis step t, we need to identify all transitions $\mathbb{P}(M^{\mathsf{X}}_{a,b} \to M^{\mathsf{Y}}_{a+i,b+j}; t)$ with $a + b < k$ and $a + i + b + j \geq k$, which add X to position k. We define $R^{\mathsf{XY};\mathsf{Z},k}$ with $\mathsf{Z} \in \{\mathsf{A}, \mathsf{B}\}$ as the matrix R^{XY} where from $R^{\mathsf{XY};\mathsf{Z},k}_{a,b}$ with $a + b = k$ to $R^{\mathsf{XY};\mathsf{Z},k}_{a+1,b+1}$ only the transition from Z to Z is allowed.

We define the probability of observing monomer Z at position k as

$$p(\tilde{S}_k = \mathsf{Z}) = p(|\tilde{S}| > k) * \sum_t \sum_{\mathsf{X},\mathsf{Y} \in \{\mathsf{A},\mathsf{B}\}} \sum_{(a,b),(i,j)}^{a+b<k, a+i+b+j \geq k} c_{\mathsf{X}}(i+j) \cdot R^{\mathsf{XY};\mathsf{Z},t}_{i,j} \cdot p_G(i+j) \cdot p_{\mathsf{A}}(t)^i \cdot p_{\mathsf{B}}(t)^j,$$

where $p(|\tilde{S}| > k)$ is the survival function of the negative binomial distribution.

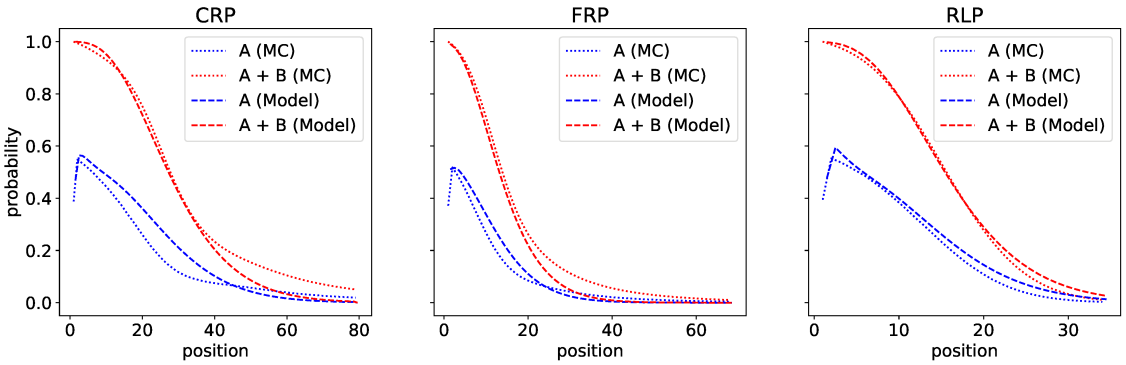


Figure 6.16 Comparison of the average sequences for the controlled radical polymerization (CRP), free radical polymerization (FRP), and reversible living polymerization (RLP) computed by Monte-Carlo (MC, dotted lines) and our model with the best parameters (dashed lines). A corresponds to $p(\tilde{S}_k = \mathsf{A})$ and B to $p(\tilde{S}_k = \mathsf{B})$ with k being the position.

We computed the average sequences for the three datasets from Sec. 6.3.4 and compared them to the average sequences computed by counting and normalizing the A and B frequencies for every position in all chains obtained by Monte-Carlo simulations (Fig. 6.16). We find that mostly the model agrees with the Monte-Carlo simulations. However, on some positions the model over- or underestimates the probabilities. This shows that estimating the parameter models from copolymer fingerprints is a difficult optimization problem.

6.4.2 Dimer Ratios

From the sequence probabilities we can compute the fraction of dimers in the sequence. The probability of observing a specific consecutive pair of monomers at a specific position k depends on the probability of observing the first monomer, the probability that the sequence is longer than k and the probability that the first monomer binds to the second. We define the *dimer probability* as

$$p^D_{\mathsf{XY}} = \sum_{k=1}^{|\tilde{S}|-1} p(\tilde{S}_k = \mathsf{X}) \cdot p(|\tilde{S}| > k) \cdot p_{\mathsf{XY}},$$

where $\mathsf{X}, \mathsf{Y} \in \{\mathsf{A}, \mathsf{B}\}$ are the monomers, p_{XY} is the reactivity probability, $\tilde{S}$ is the average sequence and $\tilde{S}_k$ the monomer at position k in the average sequence. We normalize the probabilities to obtain the *dimer ratio*:

$$D_{\mathsf{XY}} = \frac{p^D_{\mathsf{XY}}}{p^D_{\mathsf{XA}} + p^D_{\mathsf{XB}}}$$

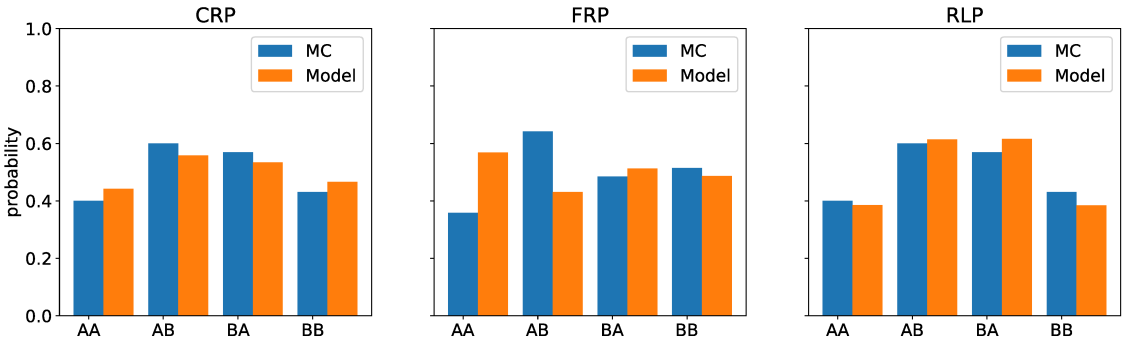


Figure 6.17 Comparison of the dimer probabilities for the controlled radical polymerization (CRP), free radical polymerization (FRP), and reversible living polymerization (RLP) predicted by Monte-Carlo (MC, dotted lines) and our model with the best parameters (dashed lines).

We computed the dimer ratios for the three copolymerization types as described above and compare it to the dimer ratios we obtained from counting all dimers in the copolymer chains computed by Monte-Carlo simulations (Fig. 6.17).

For FRP we observe that the fractions of AA and AB computed from Monte-Carlo simulations and the model differ significantly. This shows that the parameters of the model have been identified with insufficient accuracy, which is amplified by the fact that the model has many degrees of freedom. Many small numerical errors will then amount to large errors when accumulating model values to the four dimer fractions. This shows that the fitting the model to the observed fingerprint is a difficult optimization problem and in this case we most likely misidentified the reactivity parameters.

For CRP and RLP we see a significantly closer agreement between the dimer ratios computed from the Monte-Carlo simulations and the model. The fraction of AB is larger than the fraction of AA and the fraction of BA is larger than the fraction of BB. Thus, the copolymer sequences tend to be an alternating sequence of A and B. However the sequences are not strictly alternating, since the fraction of AA and BB is not zero. This agrees with the initial Monte-Carlo parameters.

6.4.3 Block Length Distribution

From the sequence probabilities we can compute the block length distribution in the sequence. We define the *block probability* as the probability of observing a specific consecutive block of monomers X of length l as:

$$B_{\mathsf{X}}(l) = \sum_{i=1}^{|\tilde{S}|} \prod_{k=i}^{i+l} p(\tilde{S}_k = \mathsf{X})$$

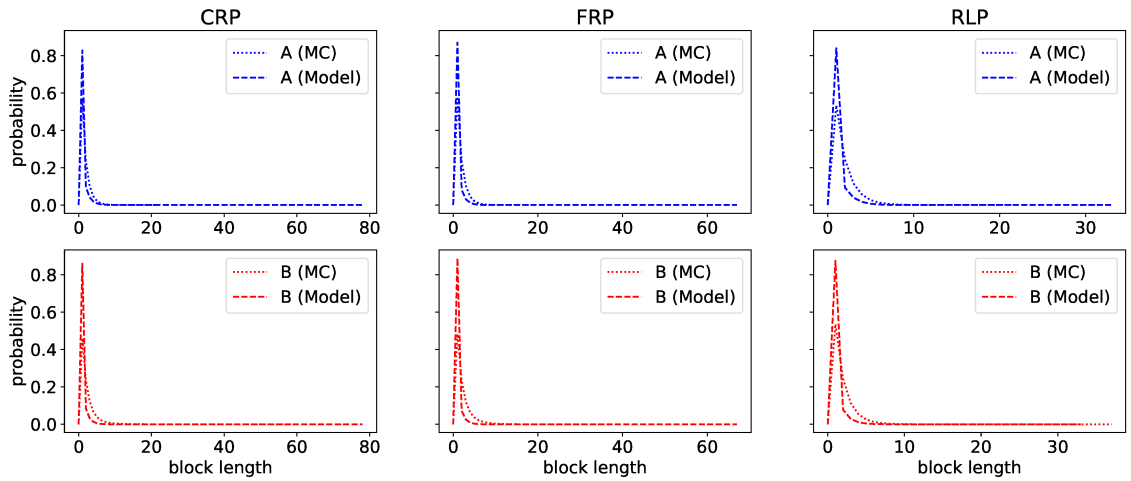


Figure 6.18 Comparison of the block length distributions for the controlled radical polymerization (CRP), free radical polymerization (FRP), and reversible living polymerization (RLP) predicted by Monte-Carlo (MC, dotted lines) and our model with the best parameters (dashed lines).

We computed the block length distributions for the three copolymerization types as described above and then normalize the B_{X} over all lengths l. We compared the results to the distributions we obtained from counting all blocks in the copolymer chains computed by Monte-Carlo simulations (Fig. 6.18).

There is a close agreement between the block length distributions computed from the Monte-Carlo simulations and the model for all three copolymerization types. The block length distributions show a sharp peak at length one, which suggests that the copolymer sequences tend to be a alternating sequences of A and B. This agrees with the initial Monte-Carlo parameters and the dimer ratios.

7. COCONUT - The Copolymer Composition Numbering Tool

We integrated the algorithms discussed in this work into the open source software COCONUT 2.0 (**Co**polymer **co**mposition **nu**mbering **t**ool).[1] [25] COCONUT 2.0 combines the algorithms with a user-friendly interface. The supported file formats include, amongst others, the open standards mzML and mzXML for mass spectra and the Open Document as well as the Excel format for copolymer fingerprints. Graphics can be exported as bitmaps, JPEG, or vector graphics.

7.1 Architecture

COCONUT 2.0 is implemented in Groovy and Java and runs on the JVM platform. The software is separated into a core library[2] and a graphical user interface. Both components are freely and openly available.

The core library includes all algorithms described in this thesis. It provides a common interface for the optimization algorithms of the Optimization Algorithm Toolkit[3] [15] and Apache Math Commons 3.2 library[4], that accepts any objective function as a Groovy closure, *i.e.* a function and the values of its free variables. The core library also includes in-/output routines for all supported file formats, and a custom visualization module using the Java 2D API for plotting spectra, copolymer fingerprints and other graphs. The visualization module is able to plot raw mass spectra with tens-of-thousands of data points as well as 2-dimensional density plots for the fingerprints.

[1] `http://www.bio.informatik.uni-jena.de/software/coconut`

[2] `https://bio.informatik.uni-jena.de/git/summary/?r+ms/polymer/polymer-library.git`

[3] `https://sourceforge.net/projects/optalgtoolkit/`

[4] `http://commons.apache.org/proper/commons-math/`

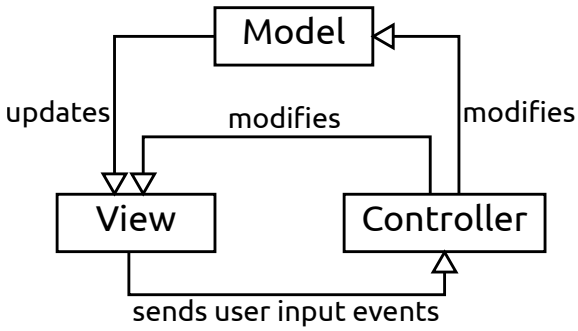


Figure 7.1 The model-view-controller (MVC) concept.

On top of the core library is a graphical user interface, realized as a plugin-framework based on the simple model-view-controller (MVC) scheme (Fig. 7.1). We determined several, sometimes conflicting, design goals for the interface. On the one hand it should be flexible, extensible and modular, on the other hand it should guide the user in his workflow, be easy to extend, and avoid too much overhead such as complex XML plugin configuration or unnecessary features such as loading modules at runtime or from a remote server.

There are two key features that help achieving our design goals. First, the COCONUT 2.0 framework detects and displays the specifically annotated plugins by simply iterating all classes in the package. The currently implemented plugins correspond to the three components described in Section 7.2. The three plugins and their layout are designed to guide the user workflow. Second, the framework achieves modularity and simplicity while adhering to the MVC concept of central controllers by two conventions (Fig. 7.2). First, every plugin defines its own actions, but all actions have the same parent class, which manages interactions with the view. Second, objects receiving the action results are singletons and the result distribution is managed by a central distributor which discovers and caches all receiving objects.

7.2 Components and User Interface

COCONUT 2.0 has three components realized as plugins: copolymer fingerprints, abundance correction, and copolymer statistics (Fig. 7.3). The three components are displayed as separate tabs in the main window. In the following, we briefly describe each component.

The core of the copolymer fingerprint component is formed by algorithms for calculating isotopic patterns, computing copolymer fingerprints and resolving isobaric species. It is distributed with the free open source linear program (LP) solver lp_solve[5] for computing the fingerprints. Our software also supports the efficient commercial Gurobi LP solver (Gurobi Optimization, Inc., Houston, USA). COCONUT 2.0 automatically detects and uses a Gurobi installation by searching for the

[5] `http://sourceforge.net/projects/lpsolve/`

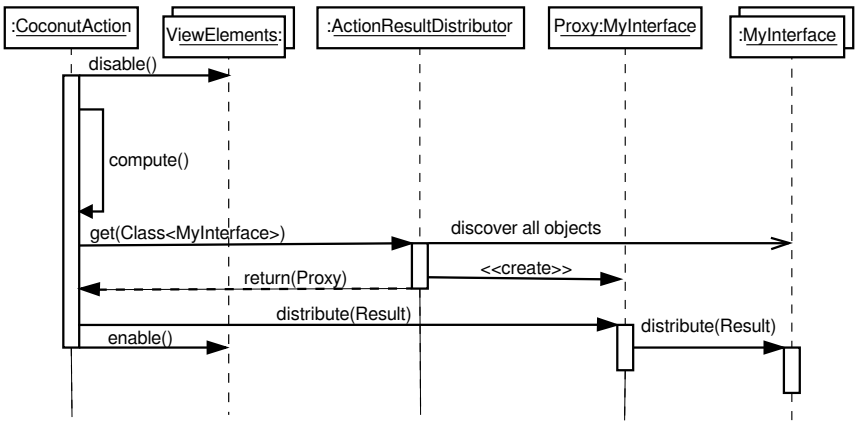
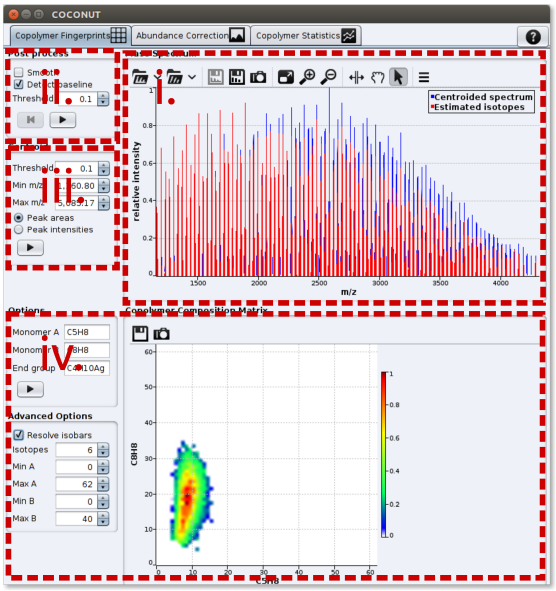
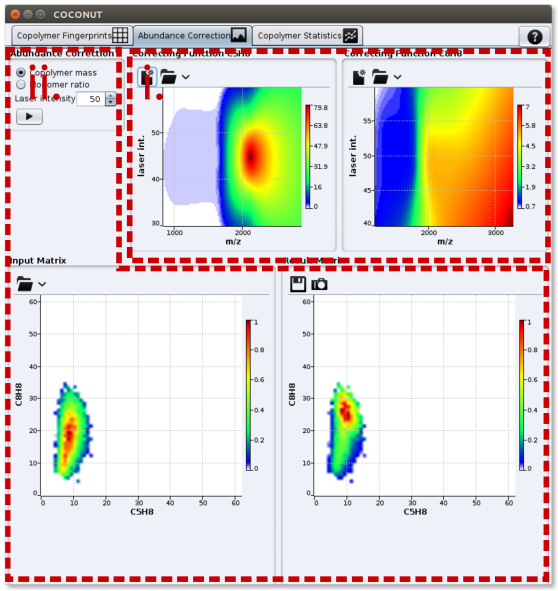


Figure 7.2 Sequence diagram of an action in the COCONUT 2.0 framework. The abstract parent class "CoconutAction" manages the interaction with the view elements. After computing a result, the action calls the "ActionResultDistributor", which discovers and caches all receiving singleton objects that implement the requested interface "MyInterface" by iterating all classes in the package. It then builds a proxy object with Groovy map coercion, and returns the proxy. To distribute the results, the action calls the proxy, which simply calls all receiving objects.

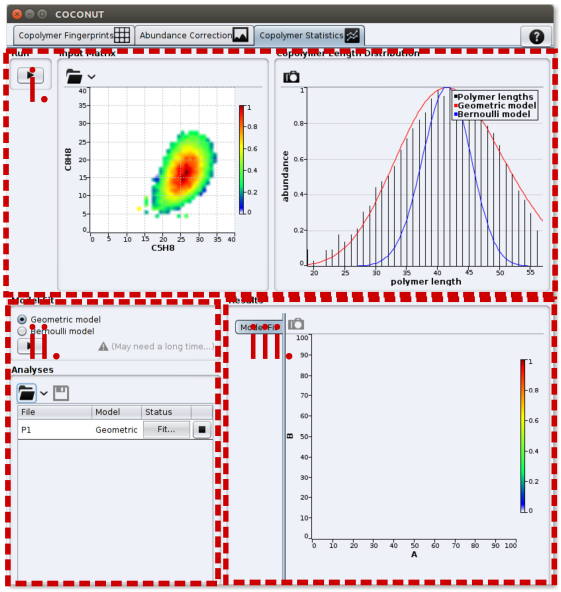
Gurobi JAR archive in the Java library path. Furthermore we included algorithms for spectral preprocessing (peak smoothing, centroiding and baseline correction) based on the routines implemented in the open source MS framework MzMine 2 [78]. The abundance correction component contains algorithms for estimating molecular weight distributions and the abundance correcting functions from homopolymer spectra, as well as for applying the correcting function to the copolymer fingerprint. The copolymer statistics component contains algorithms for estimating the copolymer length distribution, fitting a polymerization model to the copolymer fingerprint, and computing copolymer statistics with the parameterized model.



(a) Copolymer fingerprints



(b) Abundance correction



(c) Copolymer statistics

Figure 7.3 Overview of the user interface and workflow of a typical analysis: **a)** Copolymer fingerprints: i. Import of either a centroided or a raw MS spectrum. ii. Optional spectral pre-processing by smoothing raw peaks and baseline correction. iii. Centroiding of raw spectra by estimating the area under the curve of the detected peaks. iv. Copolymer fingerprint computation with optional automatic resolving of isobaric fingerprints. **b)** Abundance correction: i. Computation of the correcting functions from homopolymer spectra. ii. Applying the correcting functions to the fingerprint. **c)** Copolymer statistics: i. Estimating the copolymer length distribution from the fingerprint. ii. Starting a long-running model fitting and statistics computation. iii. Final resulting copolymer statistics.

8. Conclusions

Mass spectrometry has become an indispensable tool for analyzing copolymers. In this thesis, we present a computational approach to sequencing copolymers from mass spectrometry data, which enables the abundances of all sequences in a measured copolymer sample to be quantified.

The workflow presented in this thesis can be divided into two steps.The first step in our workflow is transforming mass spectra into copolymer fingerprints.

Copolymer spectra are highly complex and contain numerous peaks. Frequently occurring challenges include isobaric species, overlapping isotopes, background noise and peak shape perturbations. We have presented a robust algorithm to estimate copolymer fingerprints of linear binary copolymers from any type of MS spectra. Our approach is based on linear programming. We demonstrated it using several synthesized copolymers. In addition, we have evaluated our software on simulated datasets. Our method is swift and accurate for the simulated spectra. We argue that it is well suited for complex copolymer spectra, as we strove to incorporate their characteristic features in the simulated spectra.

Mass- and composition-dependent ionization – or mass discrimination – is a major challenge in computing copolymer fingerprints, especially with MALDI MS. We described an experimental protocol and a program to correct the measured abundances. Because our method uses a Gaussian approximation to the Gamma distribution to compute the molecular weight distributions (MWDs), it is applicable to narrowly distributed homopolymers up to PDI values of around 2. Approximating the MWDs is more robust using Gaussians, but if the need arises, in the future broader homopolymers could be analyzed by using Gamma distributions.

Crucial to advancing MALDI MS from a semi-quantitative to a quantitative technique for copolymers is a carefully planned experimental setup with the best possible matching conditions for homo- and copolymers. Most importantly, the MS instrument needs to be able to detect both homo- and copolymer signals over the whole

investigated mass and laser intensity ranges. Acquiring such data is challenging; the homopolymer spectra in this thesis did not perfectly conform to these stringent requirements. We invite all interested scientists to further evaluate our abundance-correcting method.

The second step in our workflow is interpreting the computed copolymer fingerprints using a new copolymerization model.

We introduced two new Markov chain models, the Bernoulli and Geometric models. The major differences to classical copolymer Markov chains based on the terminal model by Mayo and Lewis [59] are the variable number of added monomers per time step and the time-dependent monomer probabilities. The number of added monomers follows a Bernoulli or geometric distribution, respectively. The reactivity ratio has a major influence on synthesized copolymers and likewise the reactivity parameters of the models play a decisive role.

In our setup, the Geometric model is able to provide a good fit to the fingerprints of broad polymer distributions, while the fit of the Bernoulli model is particularly good to the mode but less good to the long tail of narrow polymer distributions. However, we observe that the likelihood of the copolymer chain sequences is always higher under the Geometric model. This shows that long chains play a major role in characterizing the distribution of copolymer chains.

Our models require less memory than Monte-Carlo simulations. The Bernoulli model is always significantly faster than Monte-Carlo simulations. The Geometric model is slower than the Bernoulli model, but still significantly faster (1–3 magnitudes) than Monte-Carlo for a high number of simulated chains, which is necessary for accurate Monte-Carlo simulations. Also, computing our models can be parallelized for multiple cores in a straightforward way, computing different rows of the matrices in parallel.

However, the main advantage of our models over Monte-Carlo simulations is that they do not produce just a random sample, but characterize the complete distribution of copolymer chains. Our computations are exact and deterministic. In particular, we can calculate the exact likelihood of any polymer chain. Although the Geometric model was more accurate in our setup, the Bernoulli model is a good characterization for copolymer distributions without a long tail and in general can be used as a rapid first estimate.

We then concentrated on the most accurate variant, the Geometric model with reactivity parameters.

First, the problem to solve was to find the optimal model parameters from observed data. To this end, three fingerprint-generating functions were compared, which all use the model to compute the fingerprint at the end, but differ in the number of parameters. General-purpose optimizers were used to find the optimal parameters for each function. Fitting the parameters using the model directly is the most robust method for copolymers with a Schulz-Zimm-like chain length distribution,

but has impractical running time. A simple approach to decrease the parameter search space using splines fails both in accuracy and in decreasing the running time. By exploiting the relationship between monomer concentration and the Geometric model, we find a compromise between running time and robustness against noise. For copolymers with a binomial-like chain length distribution, this approach performs best. For Schulz-Zimm-like copolymers, this method is slightly less robust against noise than the direct approach, requiring good input data. However, the running time is significantly shorter. More importantly, it is independent of the degree of polymerization and, therefore, can be used for long-chained copolymers. We recommend to use this method in practice.

On the theoretical side, the question whether the objective function is convex and smooth remains open. Additionally, the optimization is difficult and the question remains if different objective functions could make the optimization more robust. Another approach could be helping the optimization by taking more experimental parameters into account. Also of interest would be extending the current model to block copolymers in a two – or more – step process, with additional intermediate fingerprints for each synthesized block.

Second, we investigated polymerizations beyond living polymerization: controlled and free radical polymerization, and reversible living polymerization. We show that the Geometric model can be useful for copolymerization involving termination and depropagation reactions. It is yet to determine whether the model can be improved further by including termination and depropragation probabilities.

The usefulness of the model for copolymerizations beyond living polymerization is important, since these reaction systems are widely used in practice. Furthermore, termination and propagation reactions often occur accidentally in living polymerizations.

We then computed several copolymer statistics using the Geometric model and compared them to the statistics obtained by counting in the copolymer chains computed by Monte-Carlo simulations. The model agrees mostly with the Monte-Carlo simulations. However, there are some differences, which show that the model parameter estimation is a difficult optimization problem.

Last but not least we briefly discussed our software framework COCONUT, which implements all algorithms discussed in this thesis. COCONUT is freely available for polymer scientists to investigate synthesized linear binary copolymers for designing smart polymers. Our software fulfills chemists' demand for computational support in an efficient manner.

9. References

[1] A. M. Alhazmi and P. M. Mayer. Matrix effects on copolymer quantitation by matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom*, 21 (20):3392–3394, 2007.

[2] E. Altuntaş and U. S. Schubert. "Polymeromics": Mass spectrometry based strategies in polymer science toward complete sequencing approaches: A review. *Anal Chim Acta*, 808: 56–69, 2014.

[3] T. Back, D. B. Fogel, and Z. Michalwicz. *Evolutionary Computation 1 - Basic Algorithms and Operators*. Institute of Physics (IoP) Publishing, Bristol, UK, 2000.

[4] A. Baumgaertel, C. Weber, N. Fritz, G. Festag, E. Altuntaş, K. Kempe, R. Hoogenboom, and U. S. Schubert. Characterization of poly(2-oxazoline) homo- and copolymers by liquid chromatography at critical conditions. *J Chromatogr A*, 1218:8370–8378, 2011.

[5] M. Bellew, M. Coram, M. Fitzgibbon, M. Igra, T. Randolph, P. Wang, D. May, J. Eng, R. Fang, C. Lin, J. Chen, D. Goodlett, J. Whiteaker, A. Paulovich, and M. McIntosh. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, 22(15):1902–1909, 2006.

[6] D. Berek. Size exclusion chromatography - a blessing and a curse of science and technology of synthetic polymers. *J Sep Sci*, 33:315–335, 2010.

[7] S. Bernhardt, P. Glöckner, and H. Ritter. Cyclodextrins in polymer synthesis: Influence of methylated β-cyclodextrin as host on the free radical copolymerization reactivity ratios of hydrophobic acrylates as guest monomers in aqueous medium. *Polym Bull*, 46:153–157, 2001.

[8] S. Böcker and Zs. Lipták. A fast and simple algorithm for the Money Changing Problem. *Algorithmica*, 48(4):413–432, 2007.

[9] S. Böcker, M. Letzel, Zs. Lipták, and A. Pervukhin. SIRIUS: Decomposing isotope patterns for metabolite identification. *Bioinformatics*, 25(2):218–224, 2009.

[10] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE T Pattern Anal*, 11(6):567–585, 1989.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NJ, USA, 2004.

[12] A. L. T. Brandão, J. B. P. Soares, J. C. Pinto, and A. L. Alberton. When polymer reaction engineers play dice: Applications of Monte Carlo models in PRE. *Macromol React Eng*, 9 (3):141–185, 2015.

[13] J. Brandrup, E. H. Immergut, and E. A. Grulke, editors. *Polymer Handbook*. Wiley, Hoboken, NJ, USA, 4th edition, 1999.

[14] W. H. Brown, C. S. Foote, B. L. Iverson, and E. V. Anslyn. *Organic Chemistry*. Brooks/Cole, Belmont, CA, USA, 6th edition, 2012.

[15] J. Brownlee. OAT: The optimization algorithm toolkit. Technical report, Swinburne University of Technology, Victoria, Australia, 2007.

[16] G. Cheng, B. Hammouda, and D. Perahia. Polystyrene- block -polyisoprene diblock-copolymer micelles: Coupled pressure and temperature effects. *Macromol Chem Phys*, 215 (8):776–782, 2014.

[17] A. C. Crecelius and U. S. Schubert. Comprehensive copolymer characterization. In *Mass Spectrometry in Polymer Chemistry*, pages 281–318. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2011.

[18] V. Cutello and G. Nicosia. The clonal selection principle for in silico and in vitro computing. In *Recent Developments in Biologically Inspired Computing*, pages 104–147. Idea Group Publishing, Hershey, PA, USA, 2004.

[19] F. L. de Sousa, F. M. Ramos, R. L. Galski, and I. Muraoka. Generalized extremal optimization: A new meta-heuristic inspired by a model of natural evolution. In *Recent Developments in Biologically Inspired Computing*, pages 41–60. Idea Group Publishing, Hershey, PA, USA, 2005.

[20] D. R. D'hooge, P. H. Van Steenberge, P. Derboven, M.-F. Reyniers, and G. B. Marin. Model-based design of the polymer microstructure: bridging the gap between polymer chemistry and engineering. *Polym Chem*, 6(40):7081–7096, 2015.

[21] M. Drache. Modeling the Product Composition During Controlled Radical Polymerizations with Mono- and Bifunctional Alkoxyamines. *Macromol Symp*, 275-276(1):52–58, 2009.

[22] M. Drache and G. Drache. Simulating Controlled Radical Polymerizations with mcPolymer–A Monte Carlo Approach. *Polymers*, 4(3):1416–1442, 2012.

[23] M. Drache, G. Schmidt-Naake, M. Buback, and P. Vana. Modeling RAFT polymerization kinetics via Monte Carlo methods: cumyl dithiobenzoate mediated methyl acrylate polymerization. *Polymer*, 46(19):8483–8493, 2005.

[24] M. Engler. Analysing copolymer synthesis by fitting a stochastic model to mass spectrometry data. Diplomarbeit, Friedrich-Schiller-Universität Jena, 2011.

[25] M. S. Engler, S. Crotty, M. J. Barthel, C. Pietsch, K. Knop, U. S. Schubert, and S. Böcker. COCONUT – an efficient tool for estimating copolymer compositions from mass spectra. *Anal Chem*, 87(10):5223–5231, 2015.

[26] M. S. Engler, S. Crotty, M. J. Barthel, C. Pietsch, U. S. Schubert, and S. Böcker. Abundance correction for mass discrimination effects in polymer mass spectra. *Rapid Commun Mass Spectrom*, 30:1233–1241, 2016.

[27] M. S. Engler, K. Scheubert, U. S. Schubert, and S. Böcker. New statistical models for copolymerization. *Polymers*, 8(6):240, 2016.

[28] M. S. Engler, K. Scheubert, U. S. Schubert, and S. Böcker. Exploring the limits of the geometric copolymerization model. *Polymers*, 9(3):101, 2017.

[29] J. Falkenhagen and S. Weidner. Determination of critical conditions of adsorption for chromatography of polymers. *Anal Chem*, 81:282–287, 2009.

[30] J. Falkenhagen, J. F. Friedrich, G. Schulz, R.-P. Krüger, H. Much, and S. Weidner. Liquid adsorption chromatography near critical conditions of adsorption coupled with matrix-assisted laser desorption/ionization mass spectrometry. *Int J Polym Anal Charact*, 5:549–562, 2000.

[31] S. J. Gabriel, C. Schwarzinger, B. Schwarzinger, U. Panne, and S. M. Weidner. Matrix segregation as the major cause for sample inhomogeneity in MALDI dried droplet spots. *J Am Soc Mass Spectrom*, 25(8):1356–1363, 2014.

[32] W. Gilbert. Origin of life: The RNA world. *Nature*, 319(6055):618–618, 1986.

[33] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem*, 81 (25):2340–2361, 1977.

[34] G. Gody, P. B. Zetterlund, S. Perrier, and S. Harrisson. The limits of precision monomer placement in chain growth polymerization. *Nat Commun*, 7:10514, 2016.

[35] D. M. Grant and R. K. Harris. *Encyclopedia of NMR*. Wiley, Chichester, UK, 1996.

[36] C. M. Guttman, W. R. Blair, and P. O. Danis. Mass spectroscopy and SEC of SRM 1487, a low molecular weight poly(methyl methacrylate) standard. *J Polym Sci , Part B: Polym Phys*, 35(15):2409–2419, 1997.

[37] J. K. Haken. Pyrolysis gas chromatography of synthetic polymers - a bibliography. *J Chromatogr A*, 825:171–187, 1998.

[38] N. Hansen, S. D. Müller, and P. Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol Comput*, 11 (1):1–18, 2003.

[39] S. D. Hanton and K. G. Owens. Using MESIMS to analyze polymer MALDI matrix solubility. *J Am Soc Mass Spectrom*, 16(7):1172–1180, 2005.

[40] G. R. Harik. Finding multimodal solutions using restricted tournament selection. In *Proceedings of the Sixth International Conference on Genetic Algorithms*. Morgan Kaufmann, San Fransisco, CA, USA, 1995.

[41] J. Horský and Z. Walterová. Fingerprint multiplicity in MALDI-TOF mass spectrometry of copolymers. *Macromol Symp*, 339(1):9–16, 2014.

[42] A. J. Hoteling, W. J. Erb, R. J. Tyson, and K. G. Owens. Exploring the importance of the relative solubility of matrix and analyte in MALDI sample preparation using HPLC. *Anal Chem*, 76(17):5157–5164, 2004.

[43] A. J. Hoteling, T. H. Mourey, and K. G. Owens. Importance of solubility in the sample preparation of poly(ethylene terephthalate) for MALDI TOFMS. *Anal Chem*, 77(3):750–756, 2005.

[44] S. Huijser, B. B. P. Staal, J. Huang, R. Duchateau, and C. E. Koning. Chemical composition and topology of poly(lactide-co-glycolide) revealed by pushing MALDI-TOF MS to its limitations. *Angew Chem Int Ed Engl*, 45(25):1521–3773, 2006.

[45] S. Huijser, B. B. P. Staal, J. Huang, R. Duchateau, and C. E. Koning. Topology characterization by MALDI-ToF-MS of enzymatically synthesized poly(lactide-co-glycolide). *Biomacromolecules*, 7(9):2465–2469, 2006.

[46] S. Huijser, G. D. Mooiweer, R. van der Hofstad, B. B. P. Staal, J. Feenstra, A. M. van Herk, C. E. Koning, and R. Duchateau. Reactivity ratios of comonomers from a single MALDI-ToF-MS measurement at one feed composition. *Macromolecules*, 45(11):4500–4510, 2012.

[47] J. Kasperczyk, S. Li, J. Jaworska, P. Dobrzynski, and M. Vert. Degradation of copolymers obtained by ring-opening polymerization of glycolide and ϵ-caprolactone: A high resolution NMR and ESI-MS study. *Polym Degrad Stab*, 93(5):990–999, 2008.

[48] K. Kempe, S. Jacobs, H. M. L. Lambermont-Thijs, M. W. M. Fijten, R. Hoogenboom, and U. S. Schubert. Rational design of an amorphous poly(2-oxazoline) with a low glass-transition temperature: Monomer synthesis, copolymerization, and properties. *Macromolecules*, 43: 4098–4104, 2010.

[49] K. Knop, G. M. Pavlov, T. Rudolph, K. Martin, D. Pretzel, B. O. Jahn, D. H. Scharf, A. A. Brakhage, V. Makarov, U. Möllmann, F. H. Schacher, and U. S. Schubert. Amphiphilic star-shaped block copolymers as unimolecular drug delivery systems: investigations using a novel fungicid. *Soft Matter*, 9:715–726, 2013.

[50] I. Kryven and P. D. Iedema. Deterministic modeling of copolymer microstructure: Composition drift and sequence patterns. *Macromol React Eng*, 9(3):285–306, 2015.

[51] E. Lange, C. Gröpl, K. Reinert, O. Kohlbacher, and A. Hildebrandt. High-accuracy peak picking of proteomics data using wavelet techniques. In *Proceedings of Pacific Symposium on Biocomputing (PSB 2006)*, pages 243–254. World Scientific, Singapore, 2006.

[52] W. B. Liechty, D. R. Kryscio, B. V. Slaughter, and N. A. Peppas. Polymers for drug delivery systems. *Annu Rev Chem Biomol*, 1(1):149–173, 2010.

[53] J.-F. Lutz, M. Ouchi, D. R. Liu, and M. Sawamoto. Sequence-controlled polymers. *Science*, 341(6146):1238149, 2013.

[54] S. F. Macha and P. A. Limbach. Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry of polymers. *Curr Opin Solid St M*, 6(3):213–220, 2002.

[55] S. W. Mahfoud. Crowding and preselection revisited. In *Proceedings of the Second Conference on Parallel Problem Solving from Nature*, pages 27–36. Elsevier Science Inc., New York, NJ, USA, 1992.

[56] K. Martin, J. Spickermann, H. J. Räder, and K. Müllen. Why does matrix-assisted laser desorption/ionization time-of-flight mass spectrometry give incorrect results for broad polymer distributions? *Rapid Commun Mass Spectrom*, 10(12):1471–1474, 1996.

[57] K. Matyjaszewski. Architecturally complex polymers with controlled heterogeneity. *Science*, 333(6046):1104–1105, 2011.

[58] K. Matyjaszewski, T. E. Patten, and J. Xia. Controlled/"living" radical polymerization. kinetics of the homogeneous atom transfer radical polymerization of styrene. *J Am Chem Soc*, 119:674–680, 1997.

[59] F. R. Mayo and F. M. Lewis. Copolymerization. i. a basis for comparing the behavior of monomers in copolymerization; the copolymerization of styrene and methyl methacrylate. *J Am Chem Soc*, 66(9):1594–1601, 1944.

[60] D. Meimaroglou and C. Kiparissides. Review of monte carlo methods for the prediction of distributed molecular and morphological polymer properties. *Ind Eng Chem Res*, 53(22): 8963–8979, 2014.

[61] O. J. Menshoel and D. E. Goldberg. Probabilistic crowding: deterministic crowding with probabilistic replacement. Technical report, University of Illinois, 1999.

[62] M. S. Montaudo. Determination of the compositional distribution and compositional drift in styrene/maleic anhydride copolymers. *Macromolecules*, 34(9):2792–2797, 2001.

[63] M. S. Montaudo. Mass spectra of copolymers. *Mass Spectrom Rev*, 21(2):108–144, 2002.

[64] M. S. Montaudo. Mass spectra of copolymers which display compositional drifts or sequence constraints. *J Am Soc Mass Spectrom*, 15(3):374–384, 2004.

[65] M. S. Montaudo and G. Montaudo. Microstructure of copolymers by statistical modeling of their mass spectra. *Makromol Chem - M Symp*, 65(1):269–278, 1993.

[66] M. S. Montaudo and G. Montaudo. Bivariate distribution in PMMA/PBA copolymers by combined SEC/NMR and SEC/MALDI measurements. *Macromolecules*, 32(21):7015–7022, 1999.

[67] M. S. Montaudo, C. Puglisi, F. Samperi, and G. Montaudo. Structural characterization of multicomponent copolyesters by mass spectrometry. *Macromolecules*, 31(25):8666–8676, 1998.

[68] S. Muench, A. Wild, C. Friebe, B. Häupler, T. Janoschka, and U. S. Schubert. Polymer-based organic batteries. *Chem Rev*, 116(16):9438–9484, 2016.

[69] H. Mühlenbein. How genetic algorithms really work: Mutation and hillclimbing. In *Parallel Problem Solving from Nature 2*. Elsevier, Amsterdam, The Netherlands, 1992.

[70] T. Otte, H. Pasch, T. Macko, R. Brüll, F. J. Stadler, J. Kaschta, F. Becker, and M. Buback. Characterization of branched ultrahigh molar mass polymers by asymmetrical flow field-flow fractionation and size exclusion chromatography. *J Chromatogr A*, 1218:4257–4267, 2011.

[71] P. C. Painter and M. M. Coleman. *Essentials of polymer science and engineering*. DEStech Publications, Inc., Lancaster, PA, USA, 2008.

[72] H. Pasch. Hyphenated techniques in liquid chromatography of polymers. *Adv Polym Sci*, 150:1–66, 2000.

[73] H. Pasch and W. Schrepp, editors. *MALDI-TOF Mass Spectrometry of Synthetic Polymers*. Springer Berlin Heidelberg, 2003.

[74] A. Peterlin. Dynamic viscosity of polymer solutions. *Colloid Polym Sci*, 260:278–293, 1982.

[75] H. J. Philipsen. Determination of chemical composition distributions in synthetic polymers. *J Chromatogr A*, 1037(1-2):329–350, 2004.

[76] W. A. Phillip, R. M. Dorin, J. Werner, E. M. V. Hoek, U. Wiesner, and M. Elimelech. Tuning structure and properties of graded triblock terpolymer-based mesoporous and hybrid films. *Nano Lett*, 11(7):2892–2900, 2011.

[77] C. Pietsch, M. W. M. Fijten, H. M. L. Lambermont-Thijs, R. Hoogenboom, and U. S. Schubert. Unexpected reactivity for the RAFT copolymerization of oligo(ethylene glycol) methacrylates. *J Polym Sci, Part A: Polym Chem*, 47:2811–2820, 2009.

[78] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Oresic. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinf*, 11:395, 2010.

[79] R. Poli, W. B. Langdon, N. F. McPhee, and J. R. Koza. *A field guide to genetic programming*. Lulu.com, 2008.

[80] M. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical report, Centre for Mathematical Sciences, Cambridge, England, 2009.

[81] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, New York, NJ, USA, 3rd edition, 2007.

[82] H. Raeder and W. Schrepp. MALDI-TOF mass spectrometry in the analysis of synthetic polymers. *Acta Polym*, 49(6):272–293, 1998.

[83] F. Schacher, T. Rudolph, F. Wieberger, M. Ulbricht, and A. H. E. Müller. Double stimuli-responsive ultrafiltration membranes from polystyrene-block-poly(n,n-dimethylaminoethyl methacrylate) diblock copolymers. *ACS Appl Mater Interfaces*, 1(7):1492–1503, 2009.

[84] F. Schacher, M. Ulbricht, and A. H. E. Müller. Self-supporting, double stimuli-responsive porous membranes from polystyrene-block-poly(n,n-dimethylaminoethyl methacrylate) diblock copolymers. *Adv Funct Mater*, 19(7):1040–1045, 2009.

[85] D. C. Schriemer and L. Li. Mass discrimination in the analysis of polydisperse polymers by MALDI time-of-flight mass spectrometry. 1. sample preparation and desorption/ionization issues. *Anal Chem*, 69(20):4169–4175, 1997.

[86] D. C. Schriemer and L. Li. Mass discrimination in the analysis of polydisperse polymers by MALDI time-of-flight mass spectrometry. 2. instrumental issues. *Anal Chem*, 69(20): 4176–4183, 1997.

[87] R. B. Seymour, H. F. Mark, L. Pauling, C. H. Fisher, G. A. Stahl, L. H. Sperling, C. S. Marvel, and C. E. Carraher. *Raymond F. Boyer Thermoplastic Pioneer*, pages 123–126. Springer Netherlands, Dordrecht, 1989.

[88] I. Shevtsova. On the absolute constants in the berry-esseen type inequalities for identically distributed summands. Technical report, arXiv:1111.6554, 2011.

[89] L. H. Sperling. *Introduction to Physical Polymer Science*. John Wiley & Sons, Hoboken, NJ, USA, 2005.

[90] B. Staal. *Characterization of (co)polymers by MALDI-TOF-MS*. PhD thesis, University of Technology Eindhoven, 2005.

[91] R. Szymanski. On the determination of the ratios of the propagation rate constants on the basis of the MWD of copolymer chains: A new Monte Carlo algorithm. *e-Polymers*, 9(1): 538–552, 2009.

[92] I. Teraoka. *Polymer Solutions*. Wiley, New York, NY, 2002.

[93] P. Terrier, W. Buchmann, G. Cheguillaume, B. Desmazières, and J. Tortajada. Analysis of poly(oxyethylene) and poly(oxypropylene) triblock copolymers by MALDI-TOF mass spectrometry. *Anal Chem*, 77(10):3292–3300, 2005.

[94] H. Tobita. Molecular weight distribution of living radical polymers. *Macromol Theor Simul*, 15(1):12–22, 2006.

[95] S. Trimpin and D. E. Clemmer. Ion mobility spectrometry/mass spectrometry snapshots for assessing the molecular compositions of complex polymeric systems. *Anal Chem*, 80(23): 9073–9083, 2008.

[96] S. Trimpin, S. Keune, H. J. Räder, and K. Müllen. Solvent-free MALDI-MS: developmental improvements in the reliability and the potential of MALDI in the analysis of synthetic polymers and giant organic molecules. *J Am Soc Mass Spectrom*, 17(5):661–671, 2006.

[97] E. Uliyanchenko, P. J. Schoenmakers, and S. van der Wal. Fast and efficient size-based separations of polymers using ultra-high-pressure liquid chromatography. *J Chromatogr A*, 1218:1509–1518, 2011.

[98] E. Uliyanchenko, S. van der Wal, and P. J. Schoenmakers. Challenges in polymer analysis by liquid chromatography. *Polym Chem*, 3:2313–2335, 2012.

[99] P. H. M. Van Steenberge, D. R. D'hooge, Y. Wang, M. Zhong, M.-F. Reyniers, D. Konkolewicz, K. Matyjaszewski, and G. B. Marin. Linear Gradient Quality of ATRP Copolymers. *Macromolecules*, 45(21):8519–8531, 2012.

[100] P. H. M. Van Steenberge, D. R. D'hooge, M.-F. Reyniers, and G. B. Marin. Improved kinetic monte carlo simulation of chemical composition-chain length distributions in polymerization processes. *Chem Eng Sci*, 110:185–199, 2014.

[101] G. Vivó-Truyols, B. Staal, and P. J. Schoenmakers. Strip-based regression: A method to obtain comprehensive co-polymer architectures from matrix-assisted laser desorption ionisation-mass spectrometry data. *J Chromatogr A*, 1217:4150–4159, 2010.

[102] K.-G. Wahlund. Flow field-flow fractionation: Critical overview. *J Chromatogr A*, 1287: 97–112, 2013.

[103] Z. Walterová and J. Horský. Quantification in MALDI-TOF mass spectrometry of modified polymers. *Anal Chim Acta*, 693(1-2):82–88, 2011.

[104] F. C.-Y. Wang and A. D. Burleson. The development of pyrolysis-fast gas chromatography for analysis of synthetic polymers. *J Chromatogr A*, 833:111–119, 1999.

[105] B.-J. M. Webb-Robertson and W. R. Cannon. Current trends in computational inference from mass spectrometry-based proteomics. *Brief Bioinform*, 8(5):304–317, 2007.

[106] S. Weidner, J. Falkenhagen, R.-P. Krueger, and U. Just. Principle of two-dimensional characterization of copolymers. *Anal Chem*, 79:4814–4819, 2007.

[107] S. M. Weidner, J. Falkenhagen, S. Maltsev, V. Sauerland, and M. Rinken. A novel software tool for copolymer characterization by coupling of liquid chromatography with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*, 21(16):2750–2758, 2007.

[108] S. M. Weidner, J. Falkenhagen, and I. Bressler. Copolymer composition determined by LC-MALDI-TOF MS coupling and MassChrom2D data analysis. *Macromol Chem Phys*, 213 (22):1521–3935, 2012.

[109] G. Wilczek-Vera, P. O. Danis, and A. Eisenberg. Individual block length distributions of block copolymers of polystyrene-block-poly(r-methylstyrene) by MALDI/TOF mass spectrometry. *Macromolecules*, 29:4036–4044, 1996.

[110] G. Wilczek-Vera, Y. Yu, K. Waddell, P. O. Danis, and A. Eisenberg. Detailed structural analysis of diblock copolymers by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom*, 13:764–777, 1999.

[111] G. Wilczek-Vera, Y. Yu, K. Waddell, P. O. Danis, and A. Eisenberg. Analysis of diblock copolymers of poly(r-methylstyrene)-block-polystyrene by mass spectrometry. *Macro-*

molecules, 32:2180–2187, 1999.

[112] R. X. E. Willemse. *New Insights into Free-Radical (Co)Polymerization Kinetics*. PhD thesis, University of Technology Eindhoven, 2005.

[113] R. X. E. Willemse, B. B. P. Staal, E. H. D. Donkers, and A. M. van Herk. Copolymer fingerprints of polystyrene-block-polyisoprene by MALDI-ToF-MS. *Macromolecules*, 37:5717–5723, 2004.

[114] D. Y. Wu, S. Meure, and D. Solomon. Self-healing polymeric materials: A review of recent developments. *Prog Polym Sci*, 33(5):479–522, 2008.

[115] T. Yalcin, D. C. Schriemer, and L. Li. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for the analysis of polydienes. *J Am Soc Mass Spectrom*, 8(12): 1220–1229, 1997.

[116] C. Yang, Z. He, and W. Yu. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinf*, 10:4, 2009.

[117] E. Zagar, A. Krzan, G. Adamus, and M. Kowalczuk. Sequence distribution in microbial poly(3-hydroxybutyrate-co-3-hydroxyvalerate) co-polyesters determined by NMR and MS. *Biomacromolecules*, 7(7):2210–2216, 2006.

[118] R. Zubarev and M. Mann. On the proper use of mass accuracy in proteomics. *Mol Cell Proteomics*, 6(3):377–381, 2007.

[119] N. Zydziak, W. Konrad, F. Feist, S. Afonin, S. Weidner, and C. Barner-Kowollik. Coding and decoding libraries of sequence-defined functional copolymers synthesized via photoligation. *Nat Commun*, 7:13672, 2016.

A. Appendix

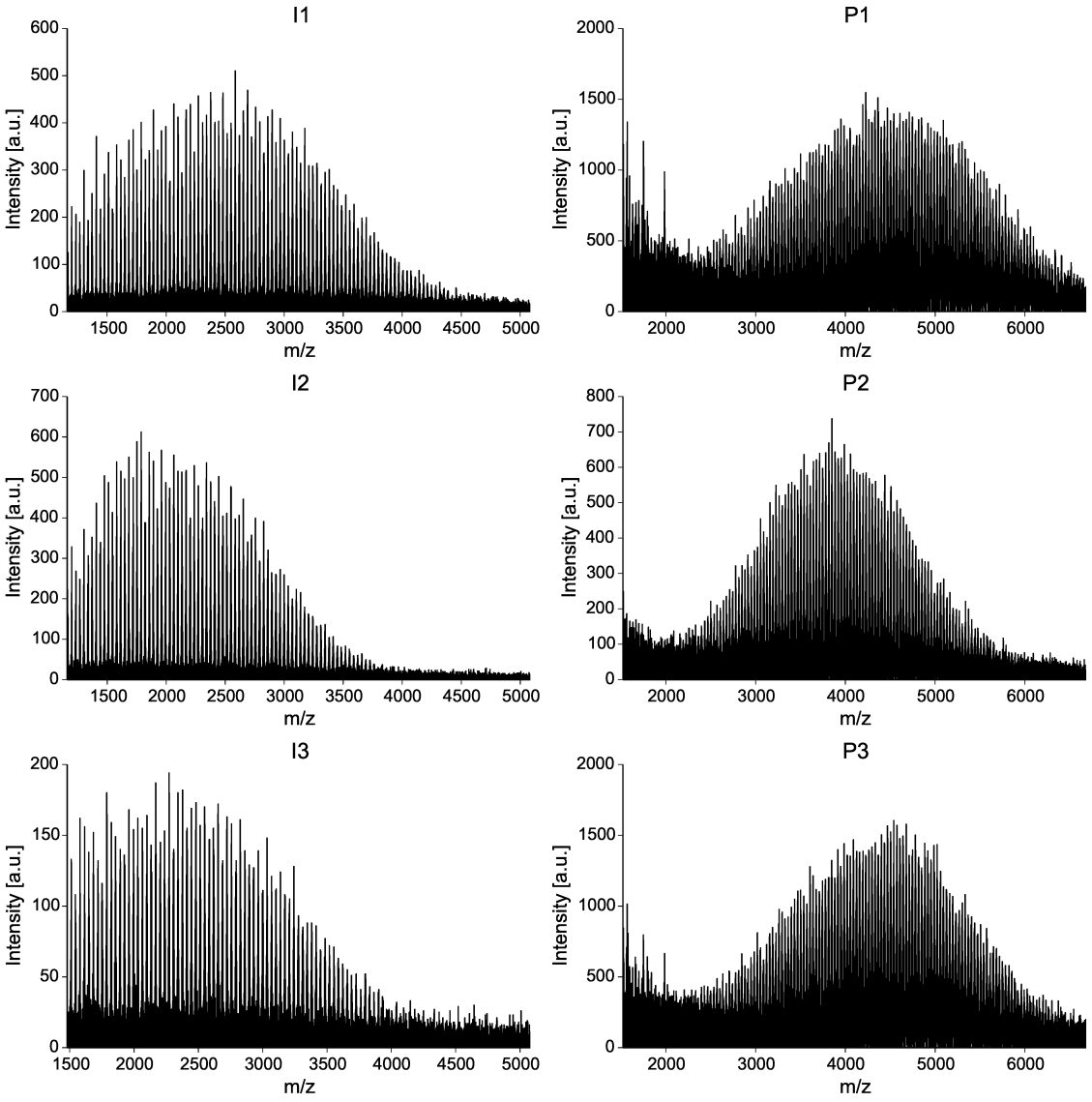


Figure A.1 MALDI-TOF spectra of the (PS-*r*-PI) macromers **I1 to I3** (left) and the final (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers **P1 to P3** (right).

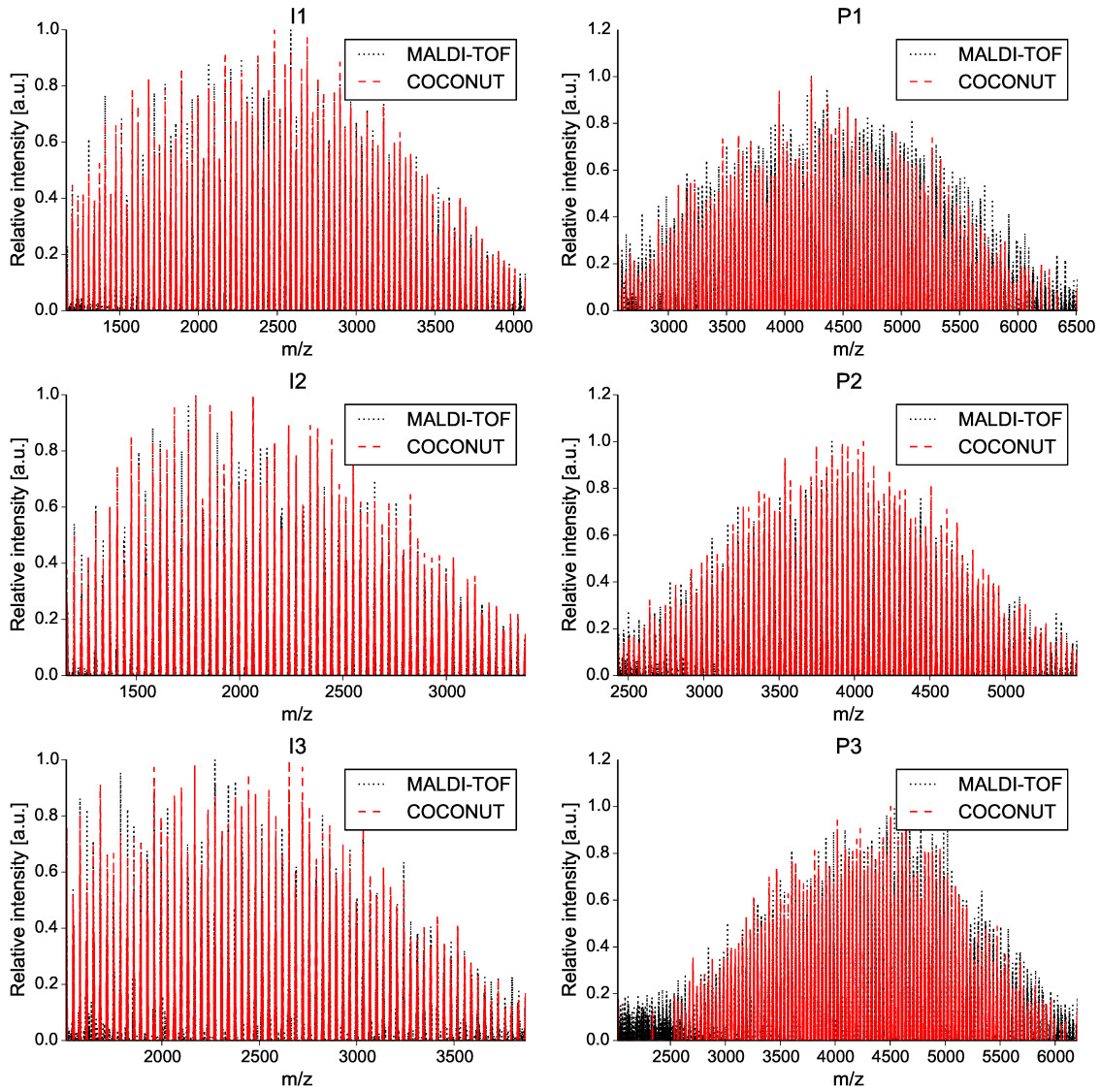


Figure A.2 MALDI-TOF spectra after baseline-correction overlayed with the isotopes estimated by our method of the (PS-r-PI) macromers **I1 to I3** (left) and the final (PS-r-PI)-r-(PS-r-PI) copolymers **P1 to P3** (right).

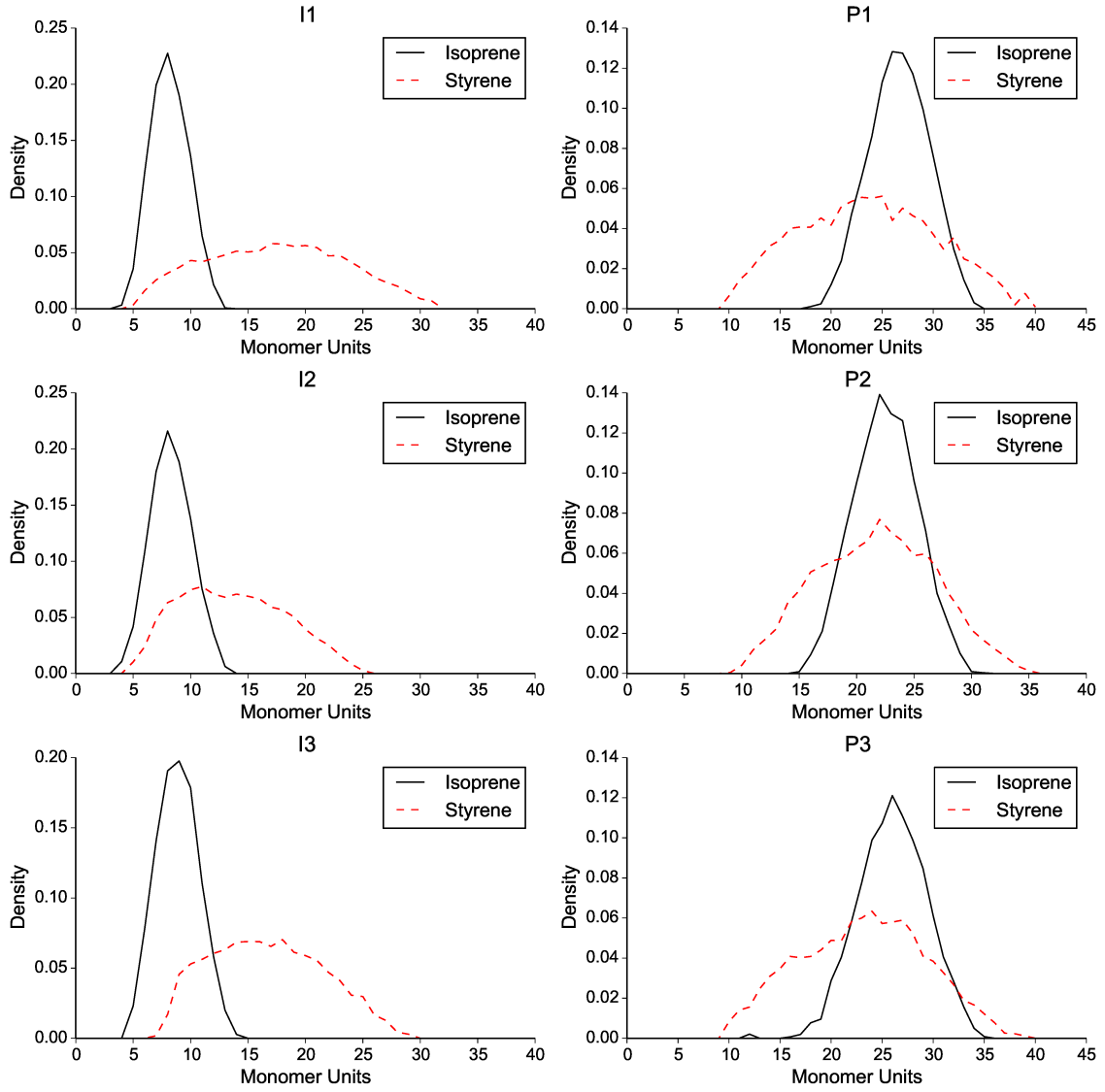


Figure A.3 Marginal distributions of the (PS-*r*-PI) macromers **I1 to I3** (left) and the final (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers **P1 to P3** (right).

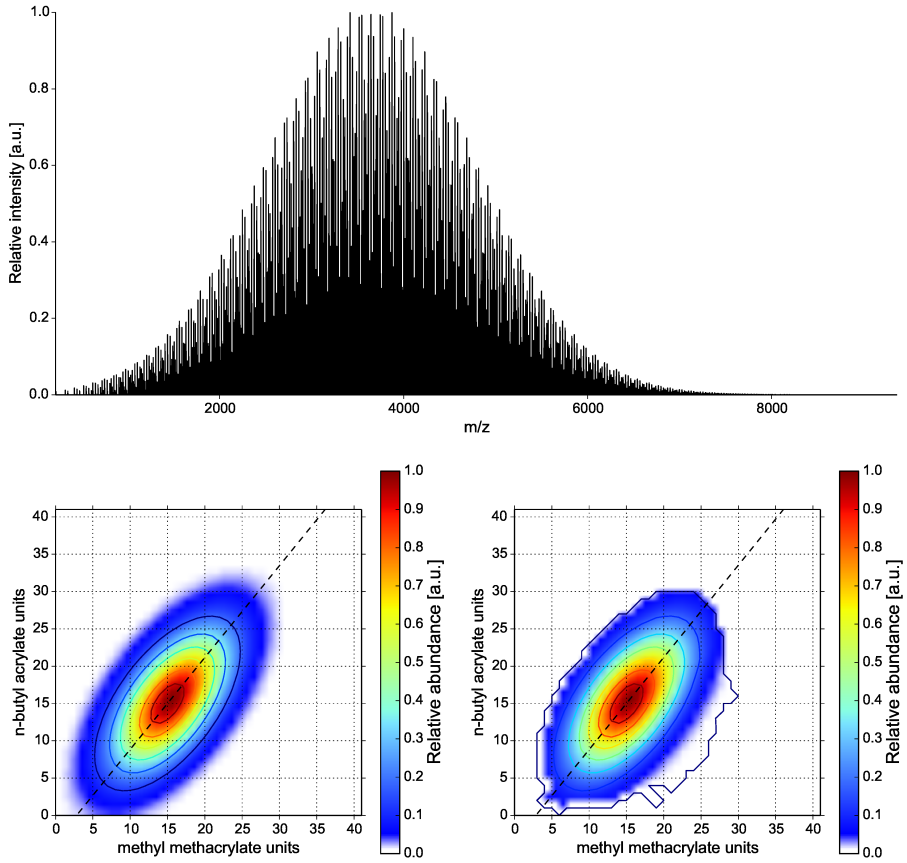


Figure A.4 Simulated copolymer fingerprint (bottom left), the resulting MS spectrum (top) of a PMMA-*co*-P*n*BA copolymer and the copolymer fingerprint estimated by our method (bottom right).

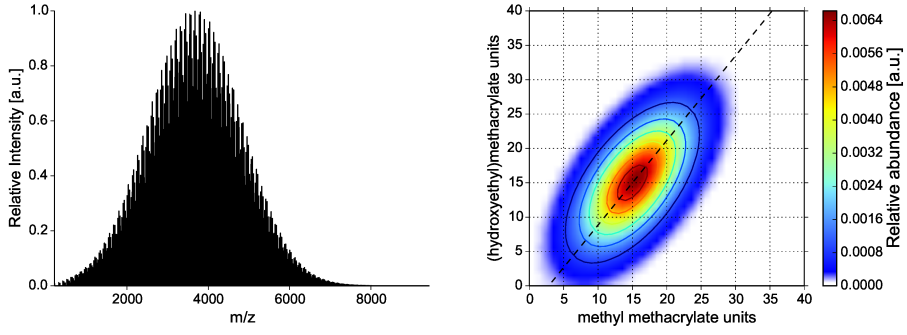


Figure A.5 Simulated copolymer fingerprint (right) and the resulting MS spectrum (left) of a PMMA-*co*-PHEMA copolymer.

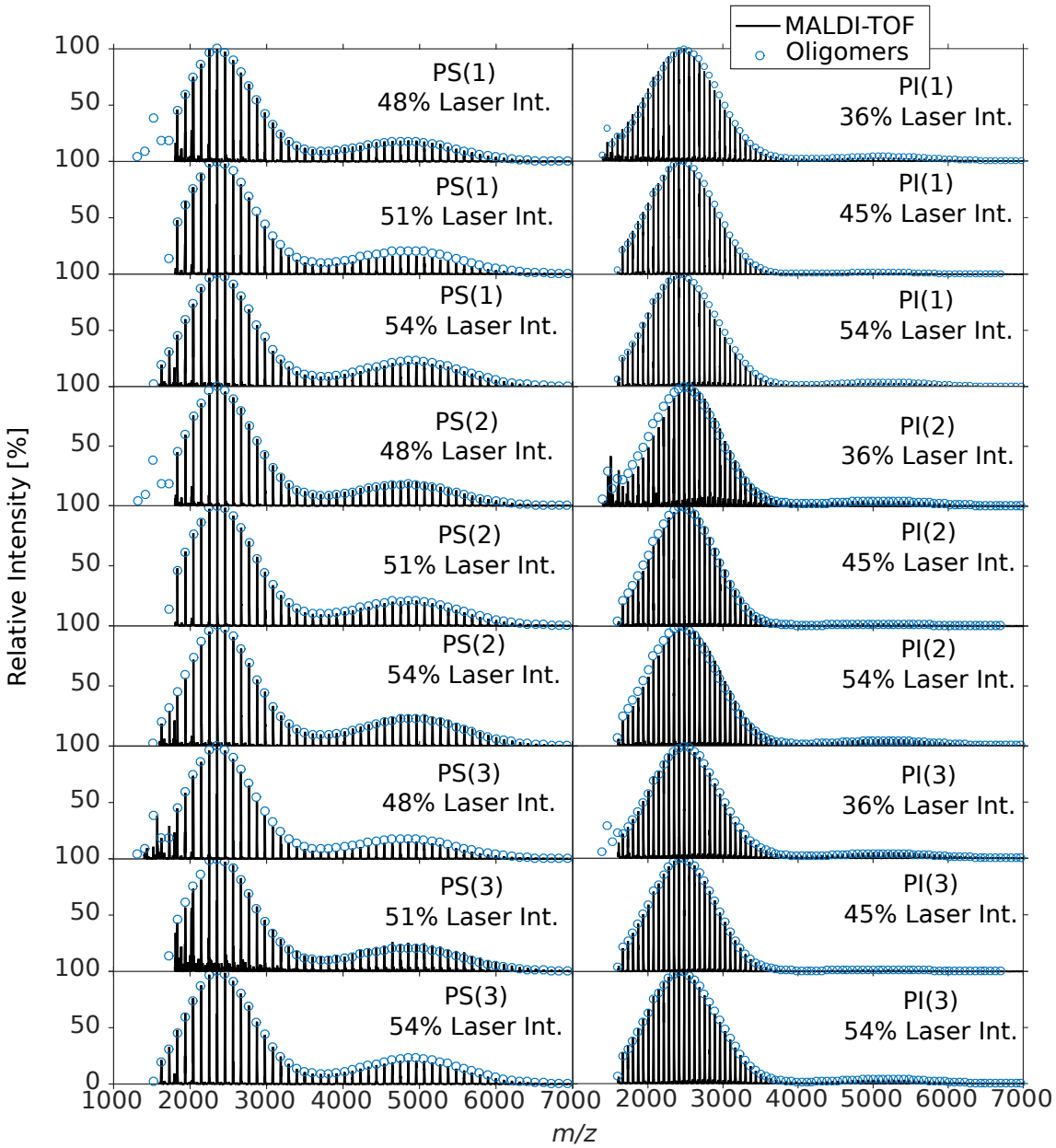


Figure A.6 MALDI-TOF mass spectra of PS and PI homopolymers overlaid with the oligomer peaks. The oligomer abundances were averaged over the three replicated spectra for each laser intensity.

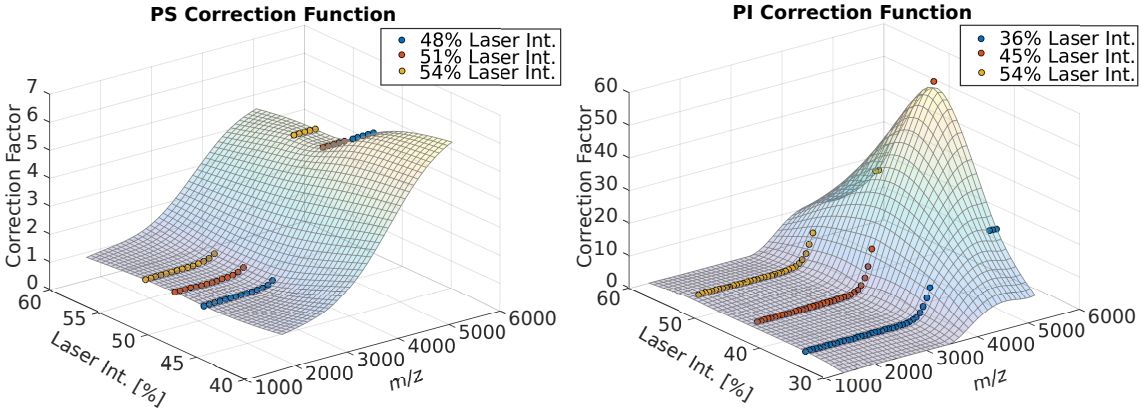


Figure A.7 Estimated correction factors for PS (left) and PI (right) as a bivariate function of mass and laser intensity.

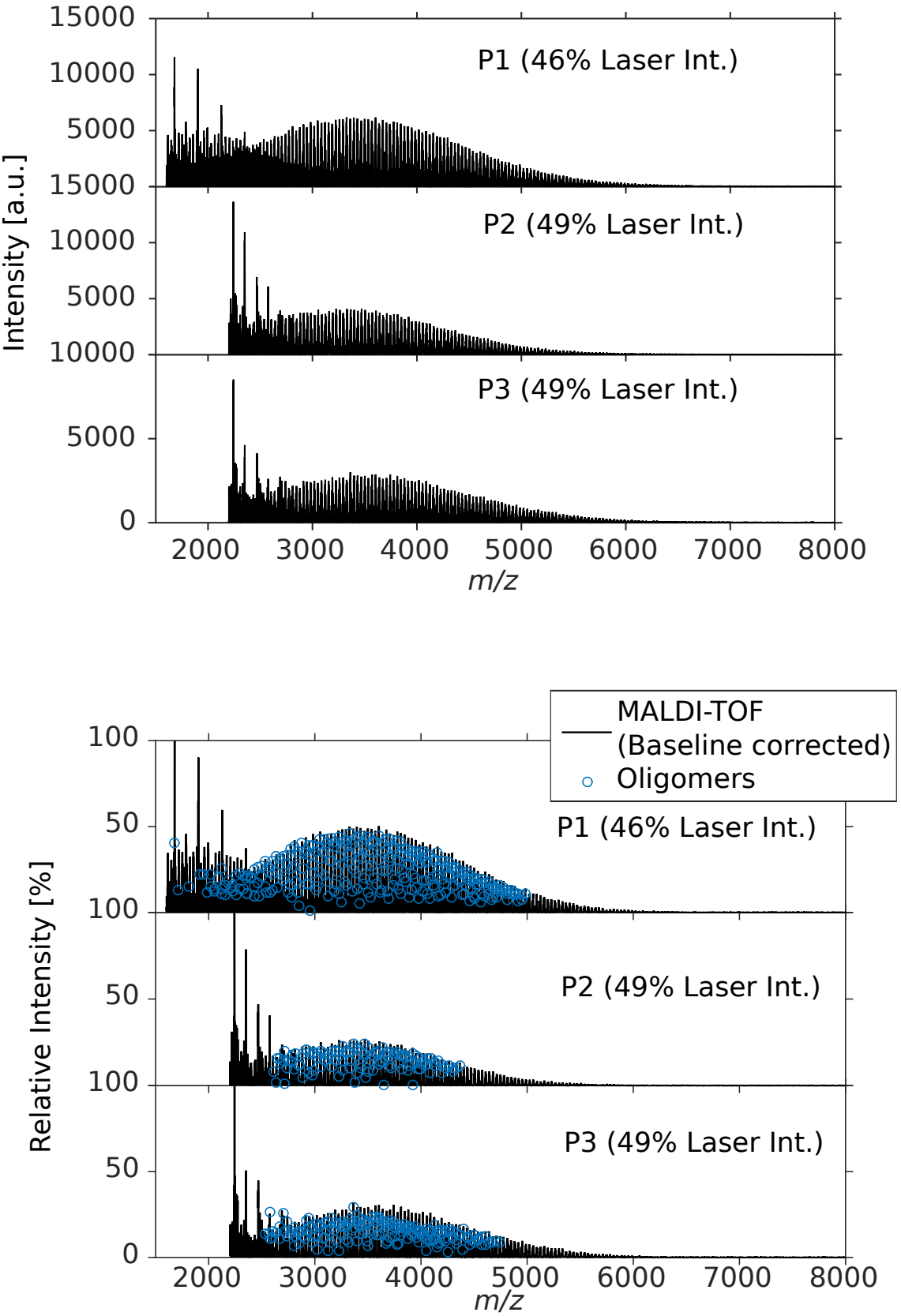


Figure A.8 Top: MALDI-TOF mass spectra of copolymers P1 to P3, bottom: mass spectra after baseline correction overlaid with oligomer peaks estimated by our method.

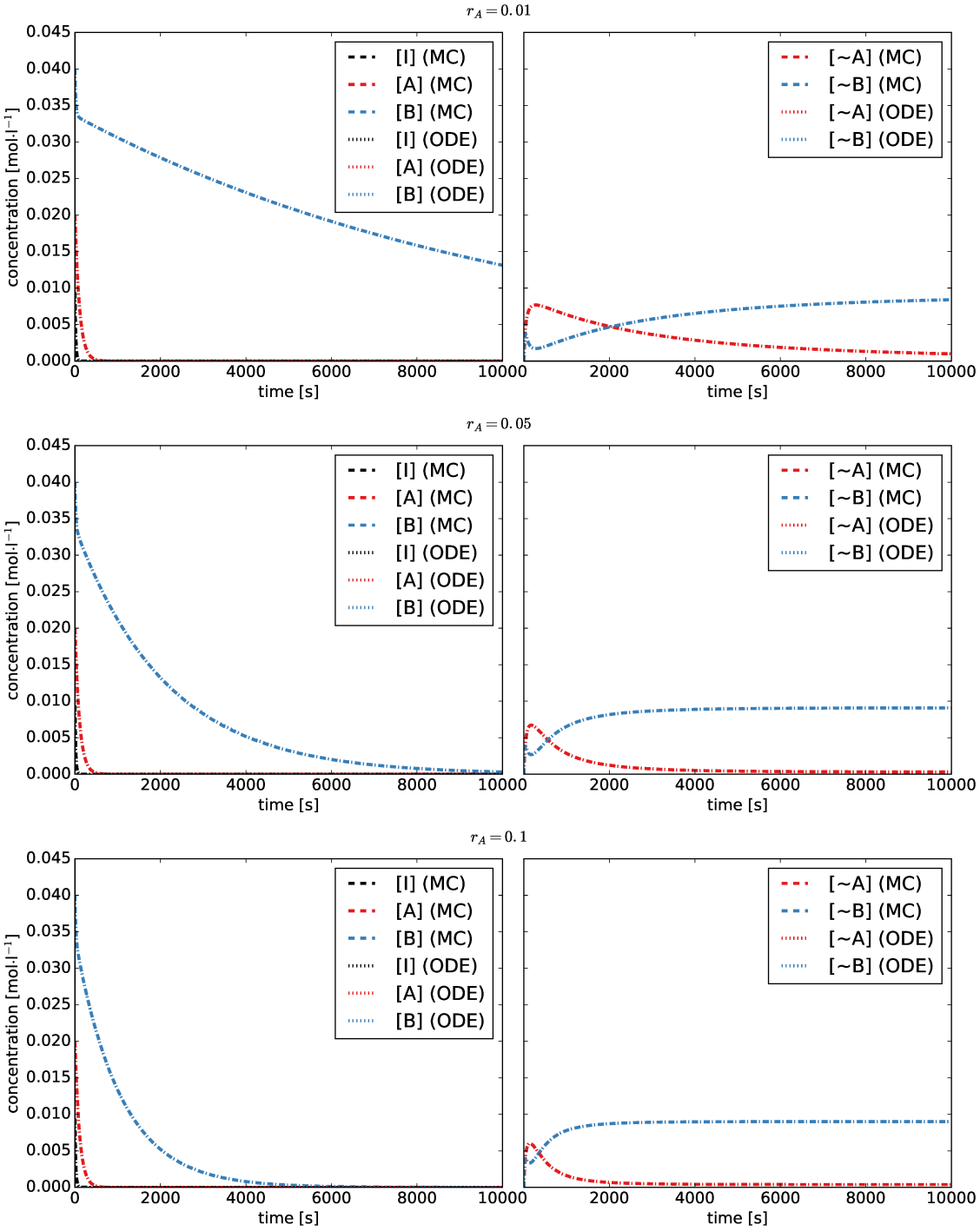


Figure A.9 Reactant (left) and product (right) concentrations simulated by Monte-Carlo (MC) compared to the concentrations computed by solving the ordinary differential equation (ODE) model of the living copolymerization.

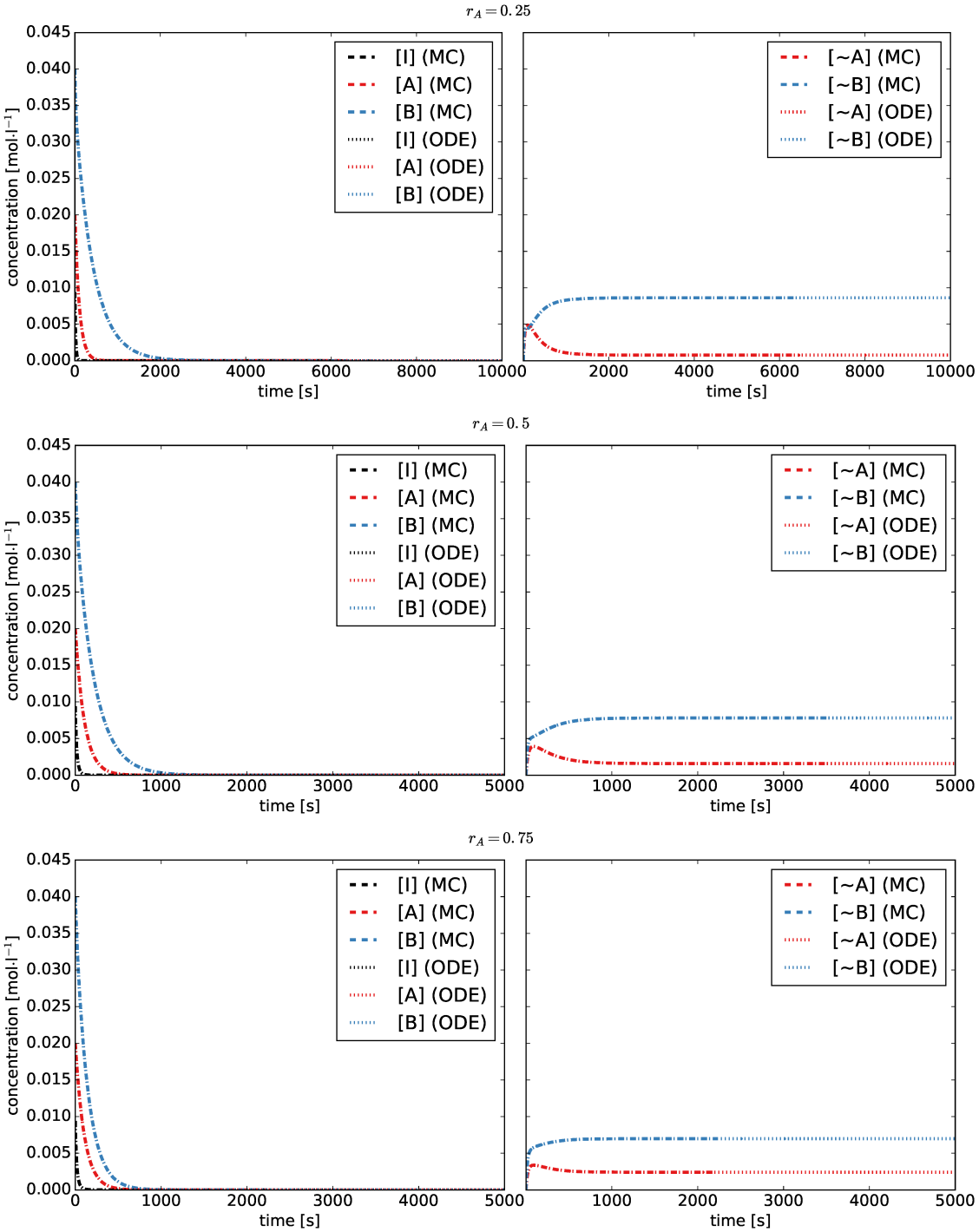


Figure A.9 (continued)

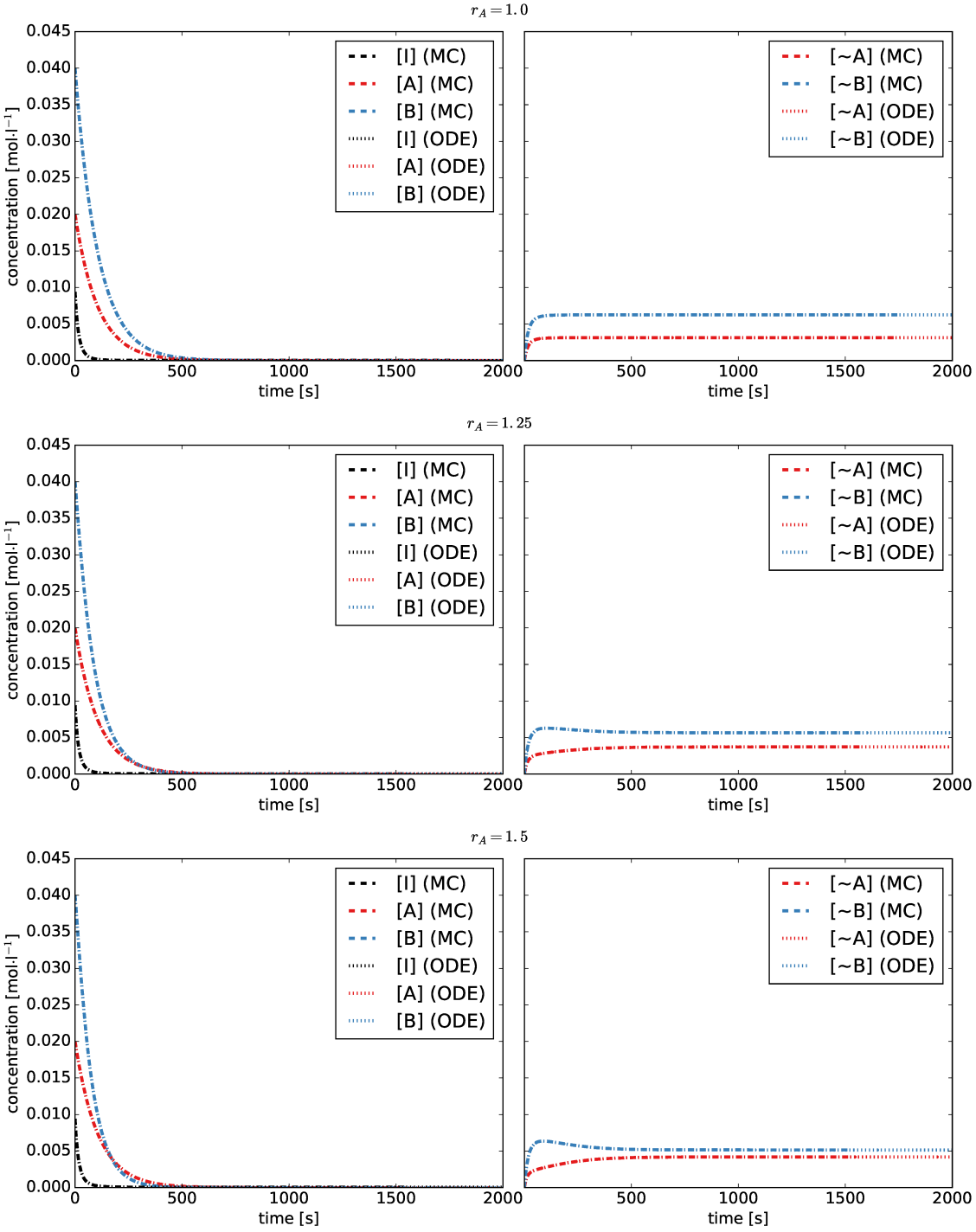


Figure A.9 (continued)

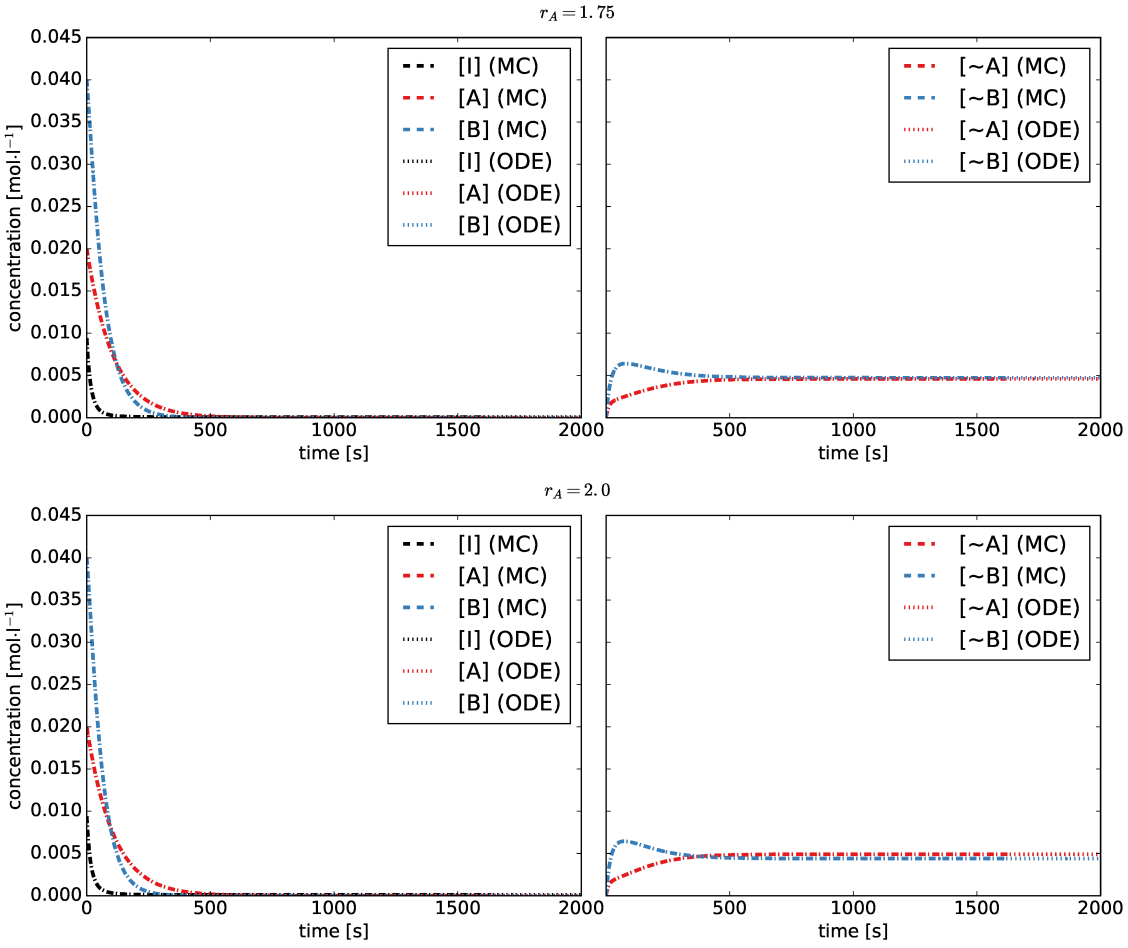


Figure A.9 (continued)

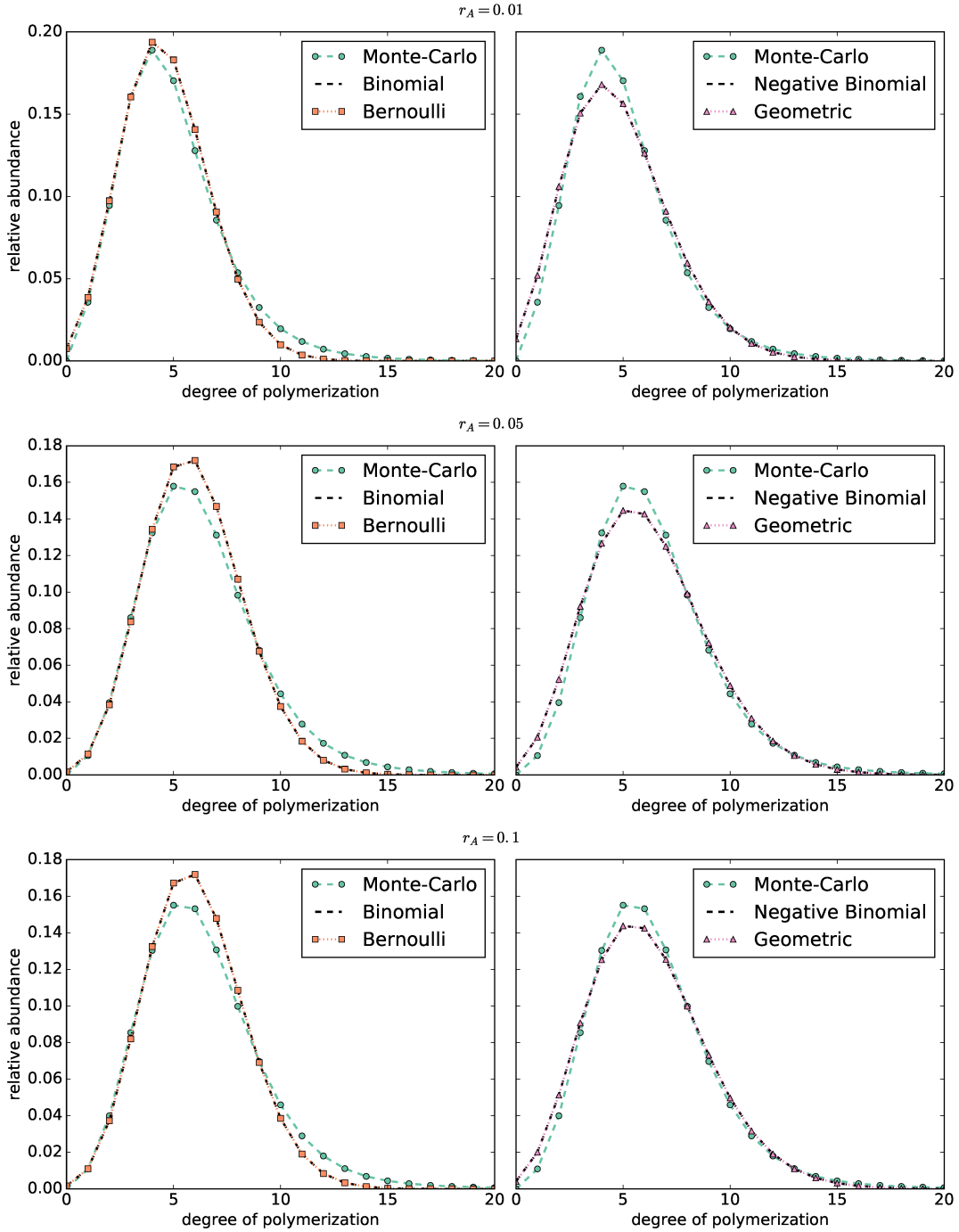


Figure A.10 Distribution of chain lengths computed by the Monte-Carlo simulations and the Bernoulli and Geometric models. Additionally, we plotted the binomial and negative binomial probability mass functions to show that the chain lengths computed by the models follow those distributions.

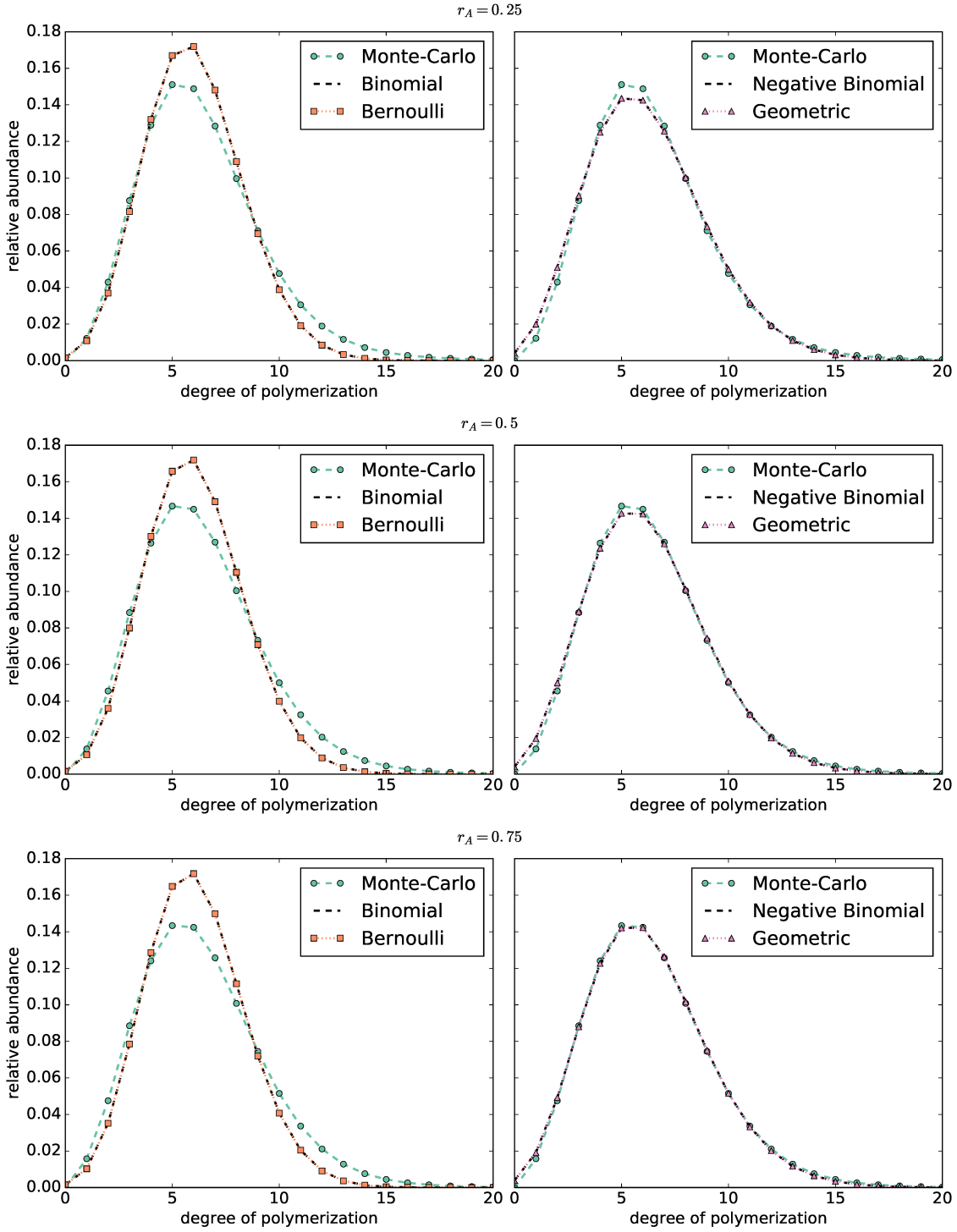


Figure A.10 (continued)

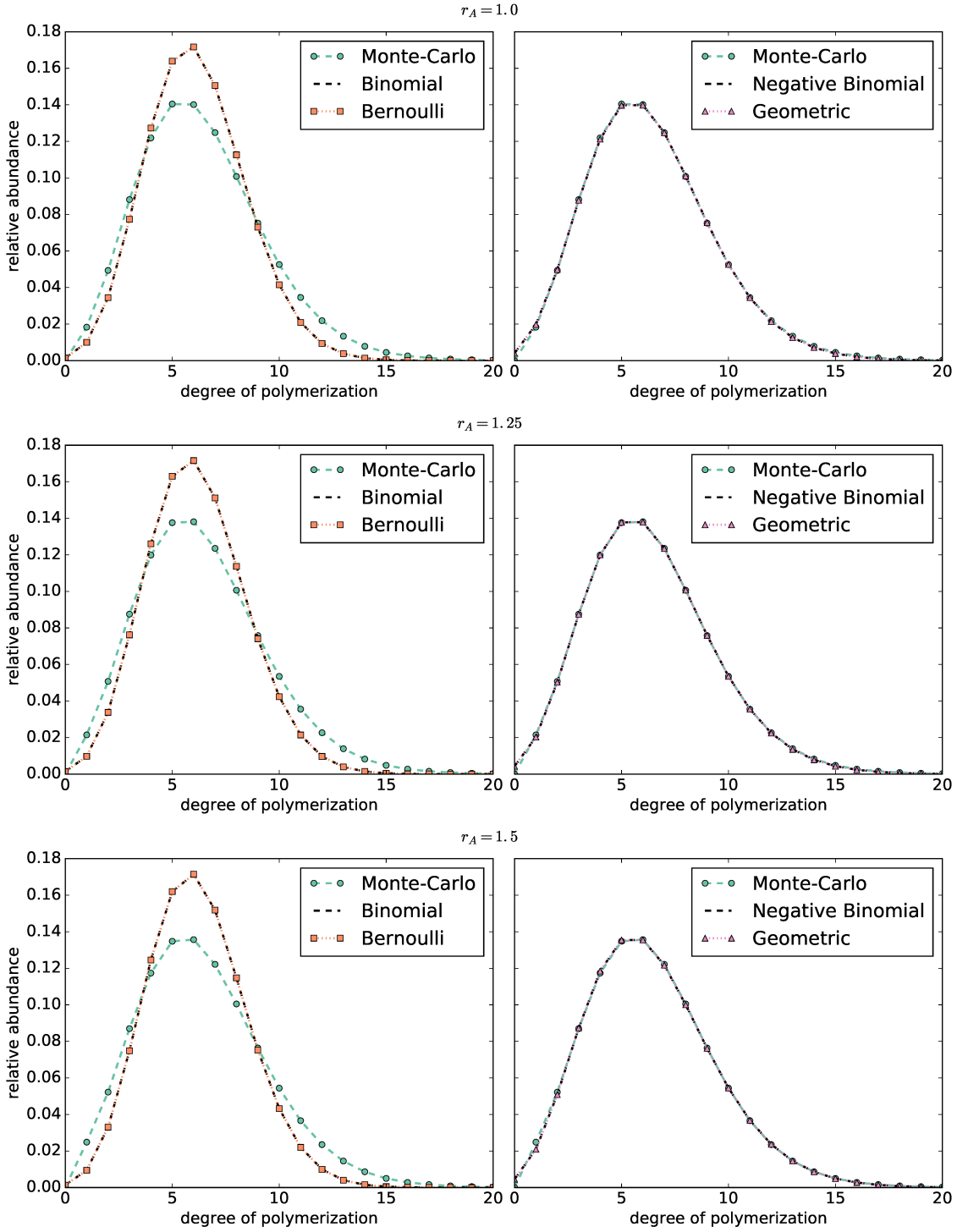


Figure A.10 (continued)

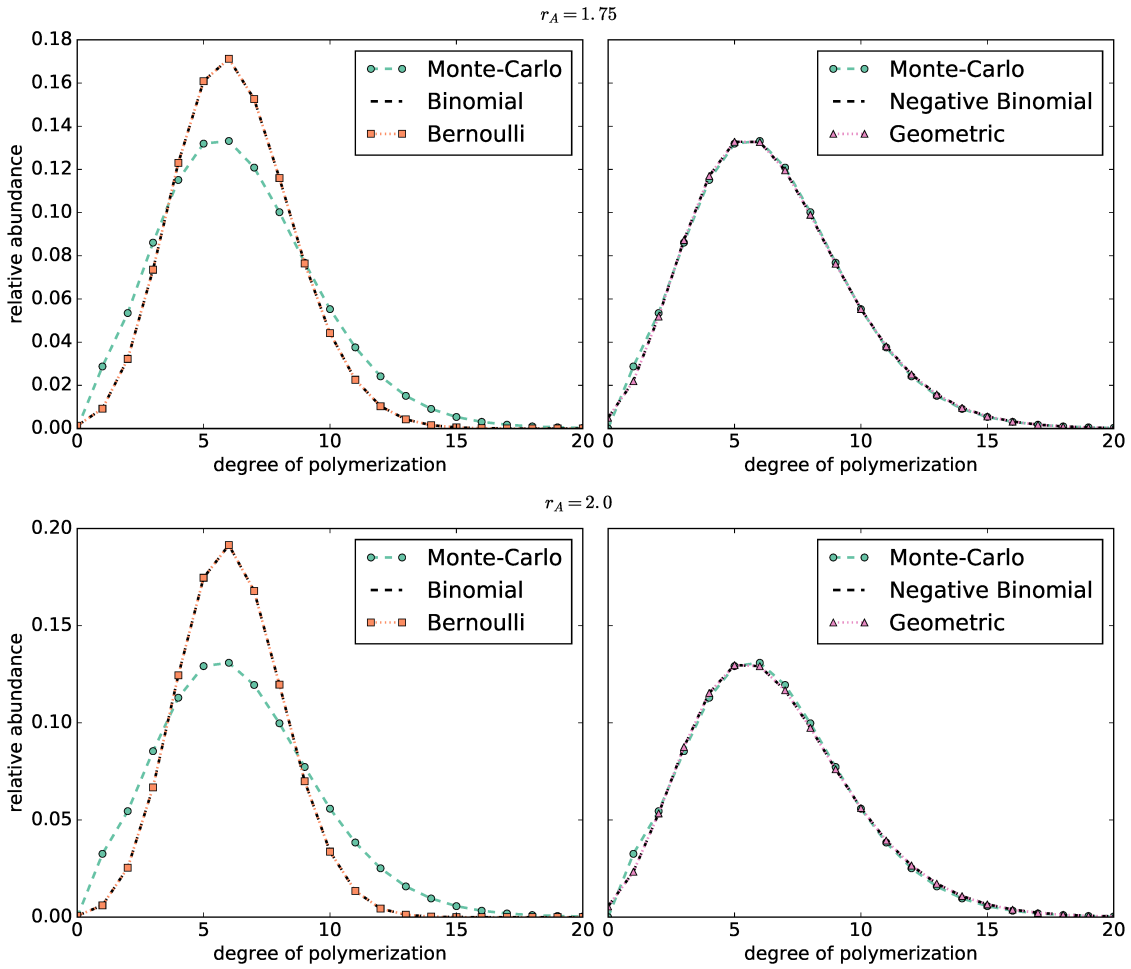


Figure A.10 (continued)

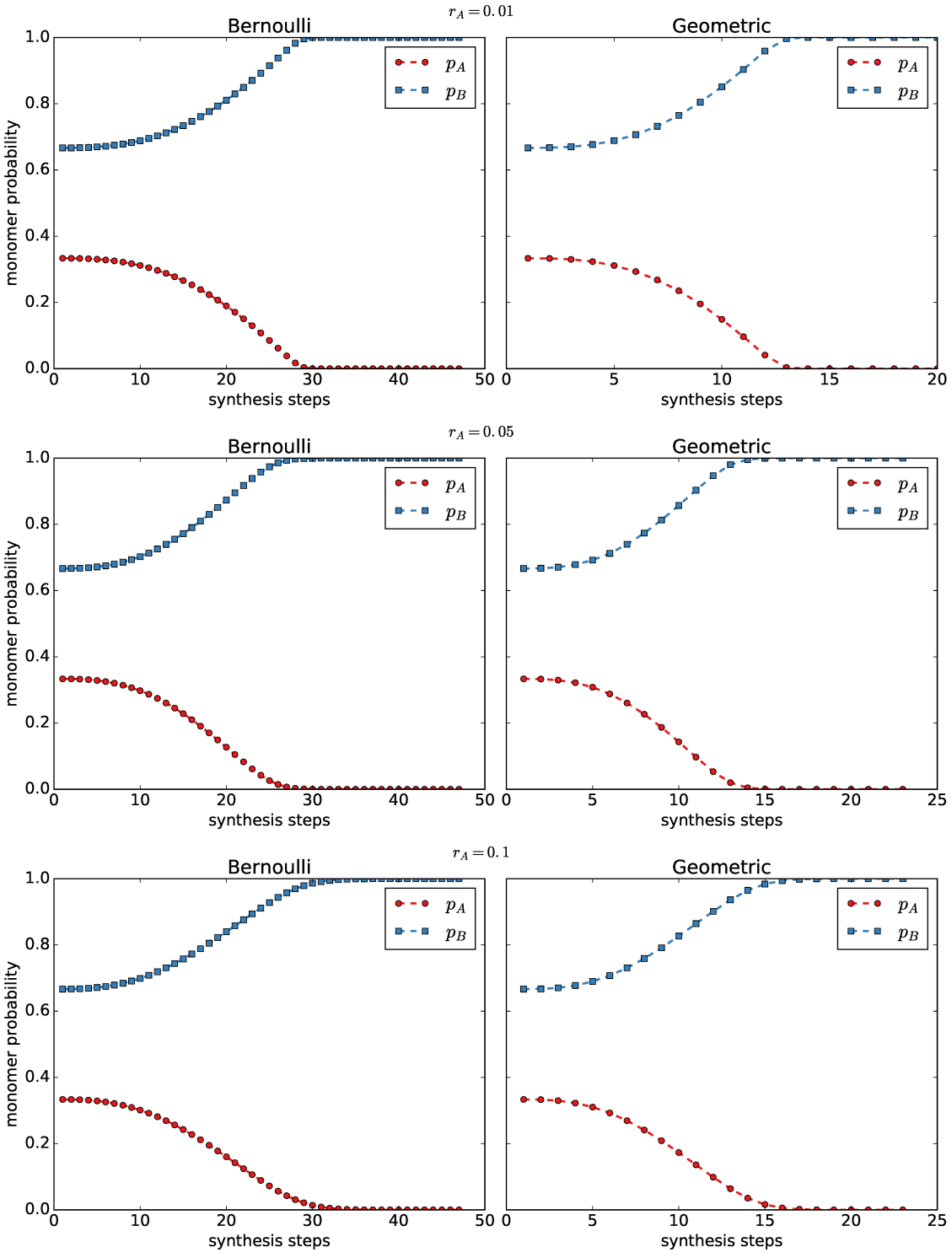


Figure A.11 Monomer probabilities p_A and p_B for the Bernoulli (left) and Geometric (right) models calculated from the average concentrations.

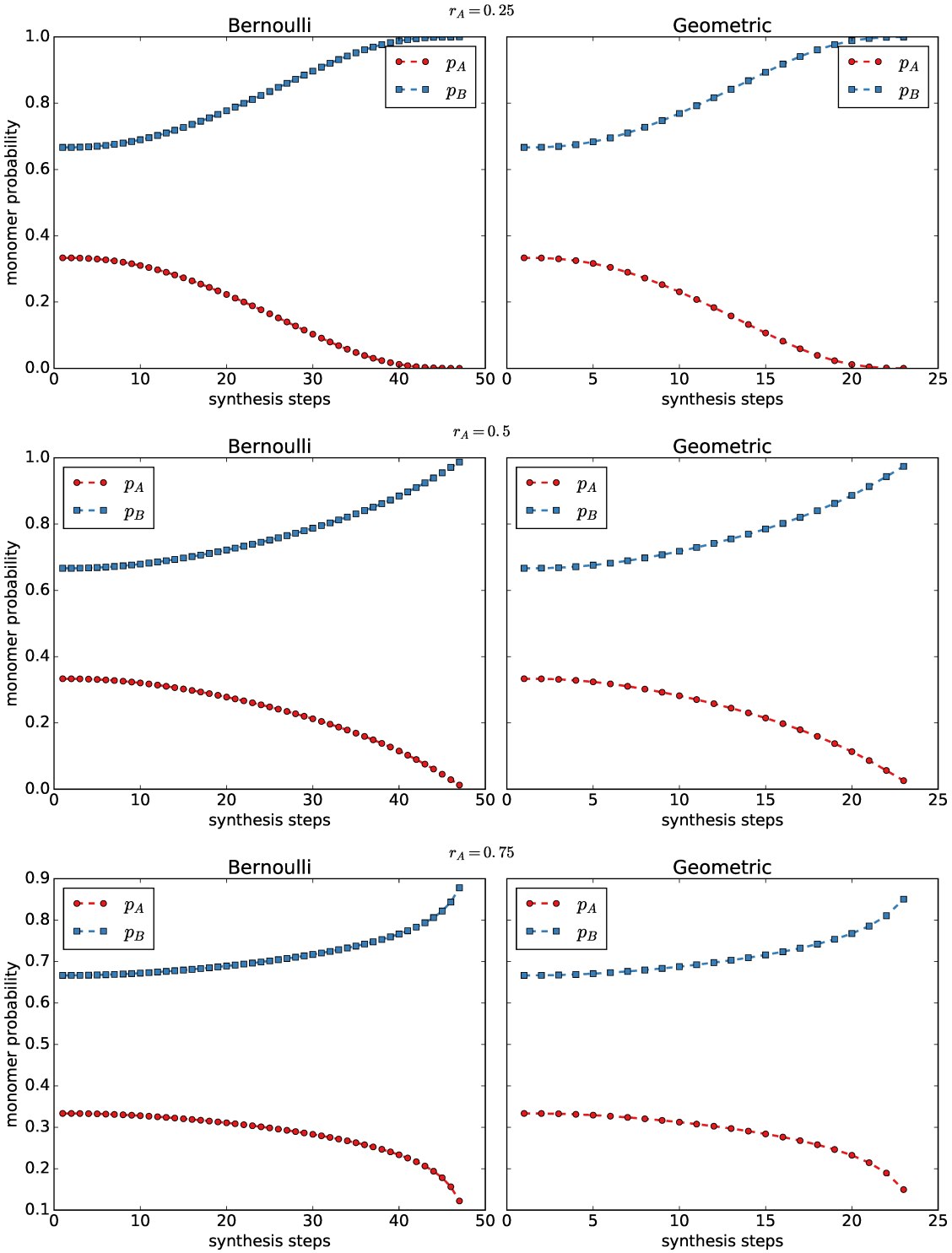


Figure A.11 (continued)

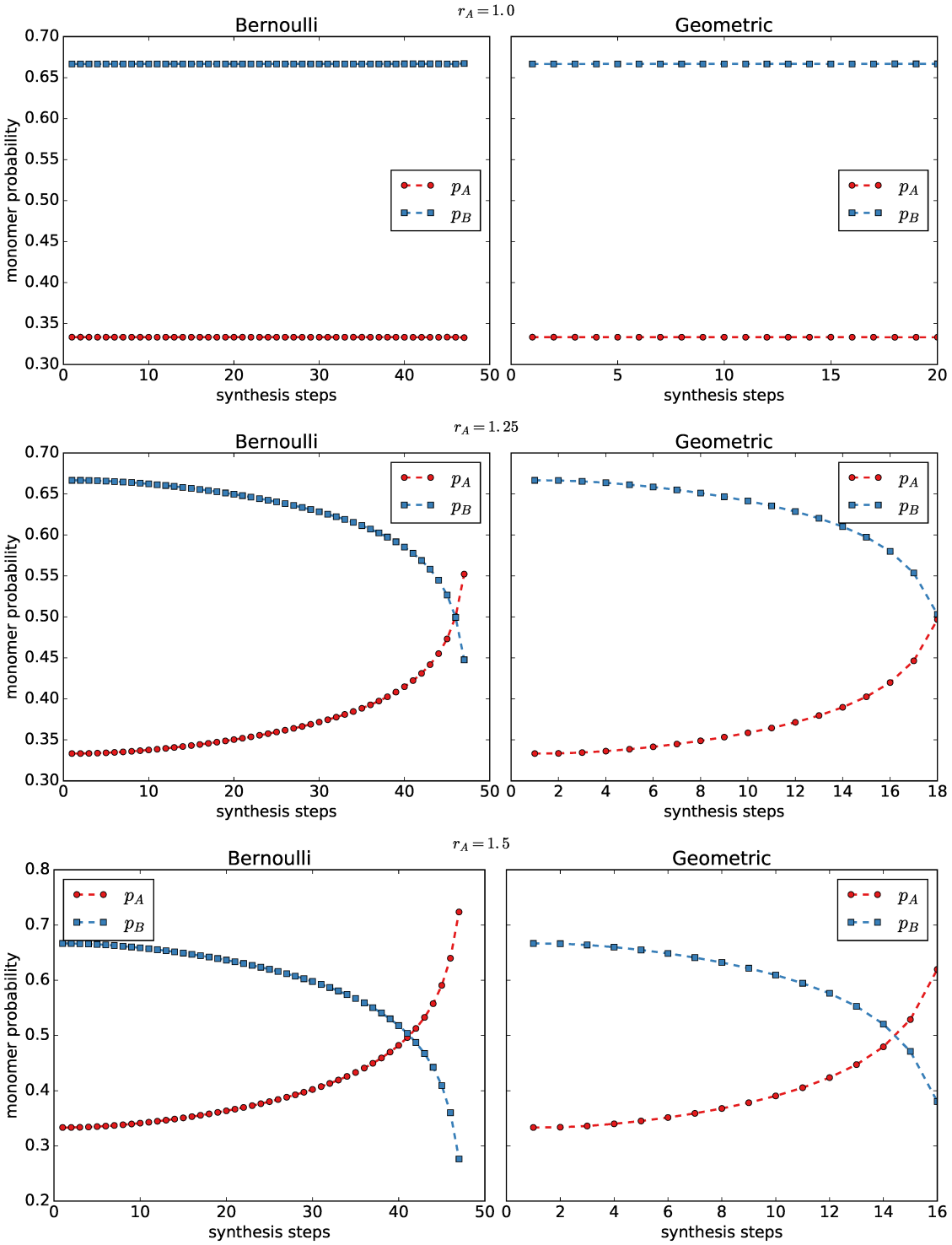


Figure A.11 (continued)

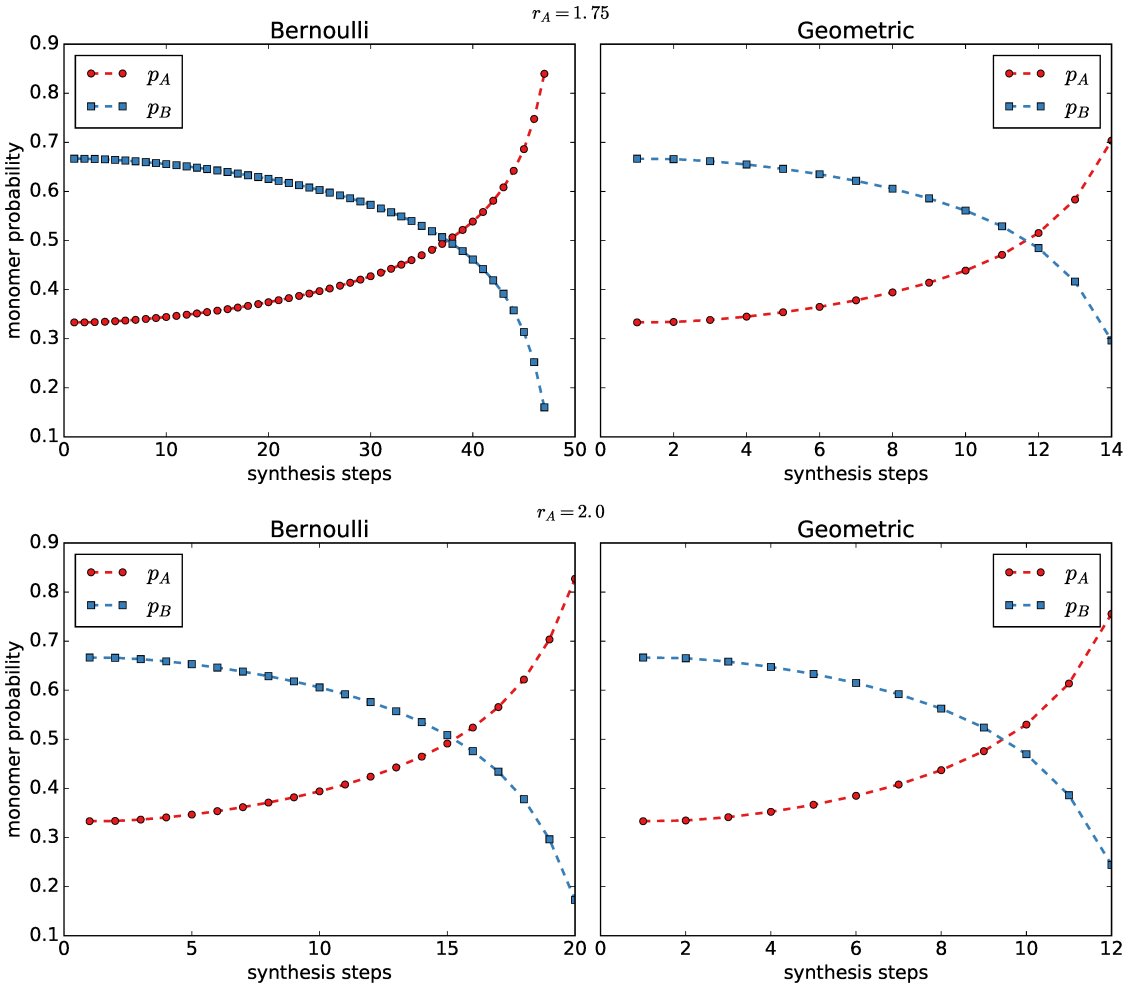


Figure A.11 (continued)

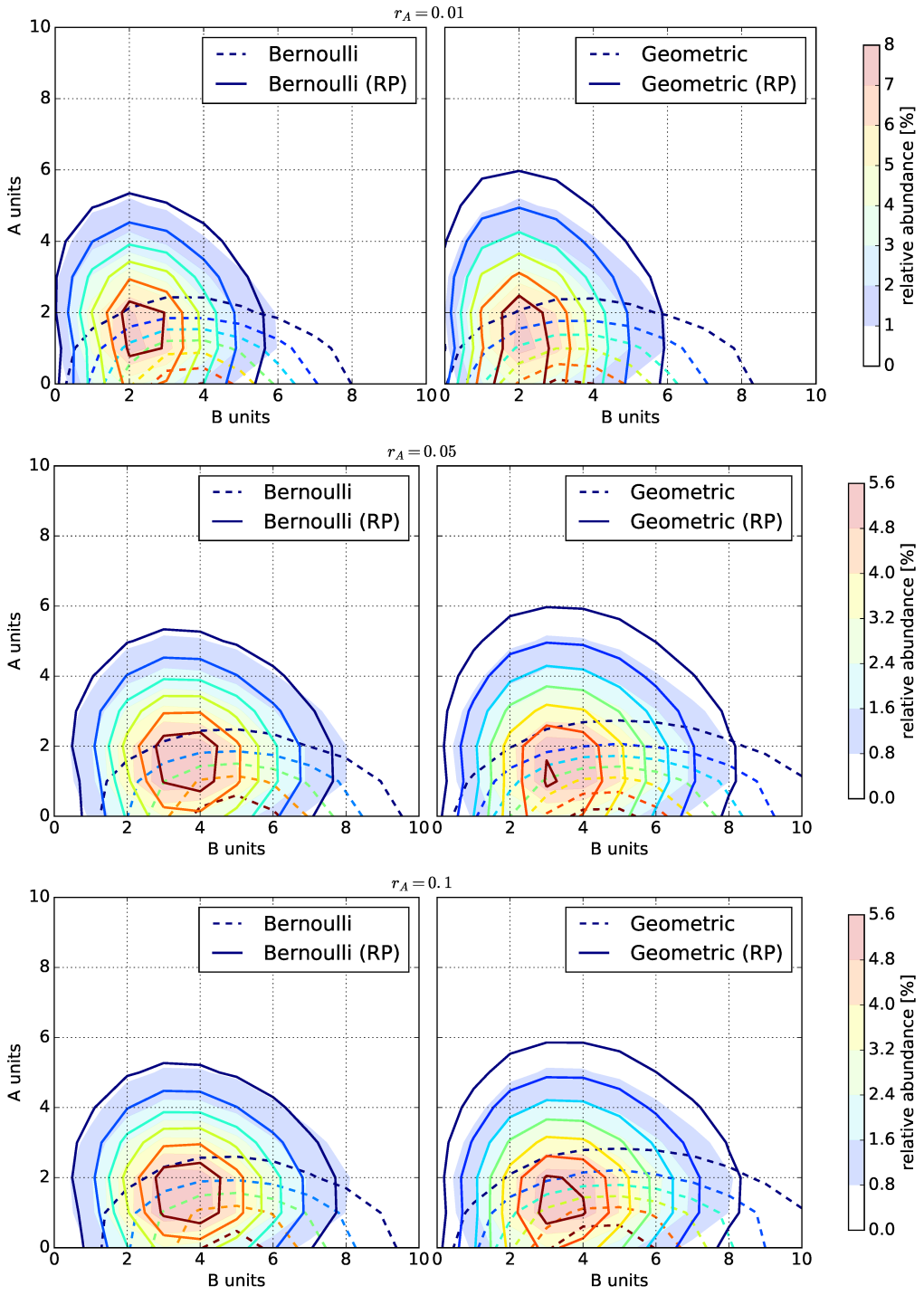


Figure A.12 Copolymer fingerprints computed by the Monte-Carlo simulation (filled contours) compared to the fingerprints computed by the statistical models (solid and dashed contours). Left: Bernoulli model with and without reactivity parameters (RP). Right: Geometric model with and without reactivity parameters (RP).

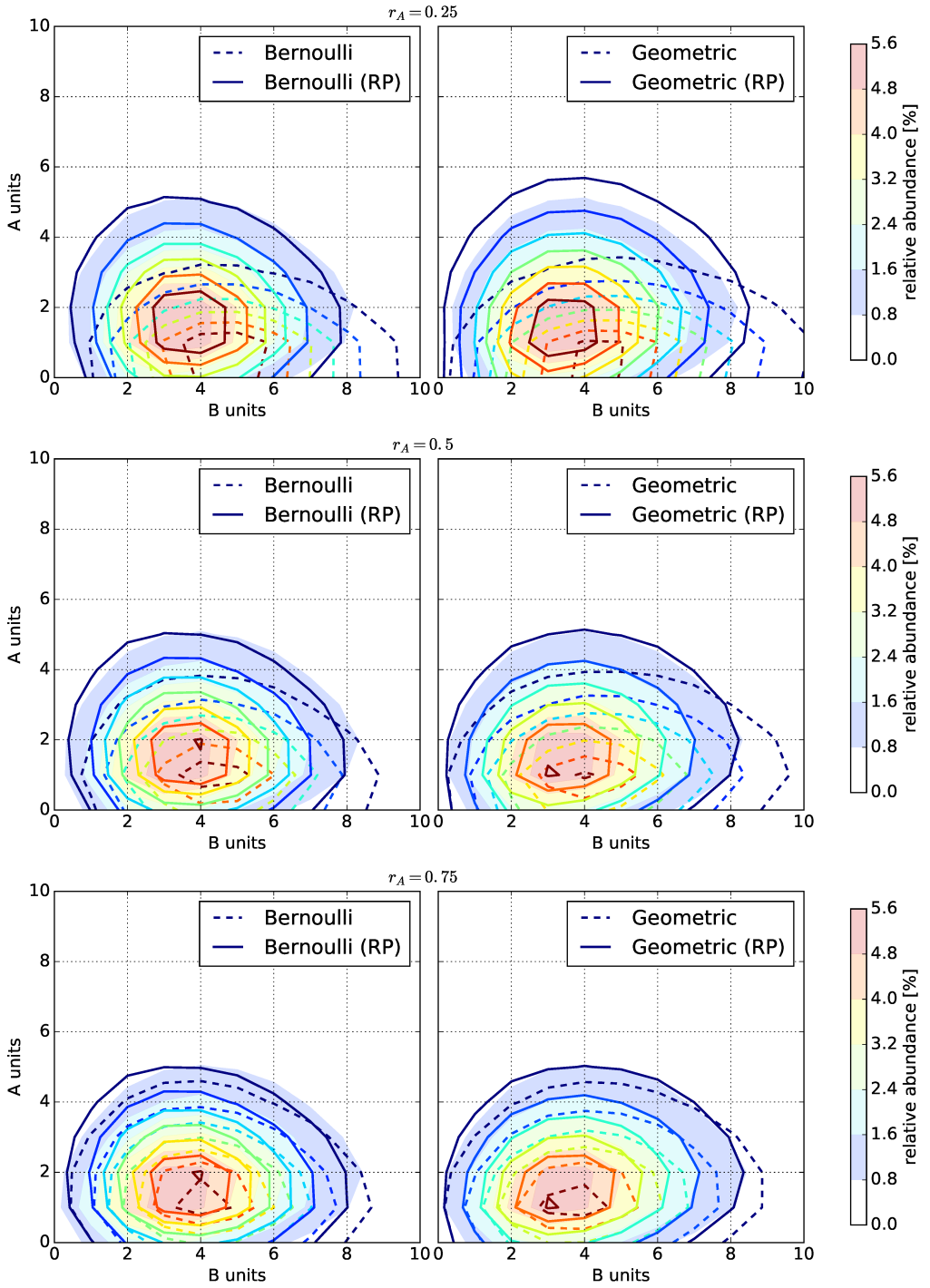


Figure A.12 (continued)

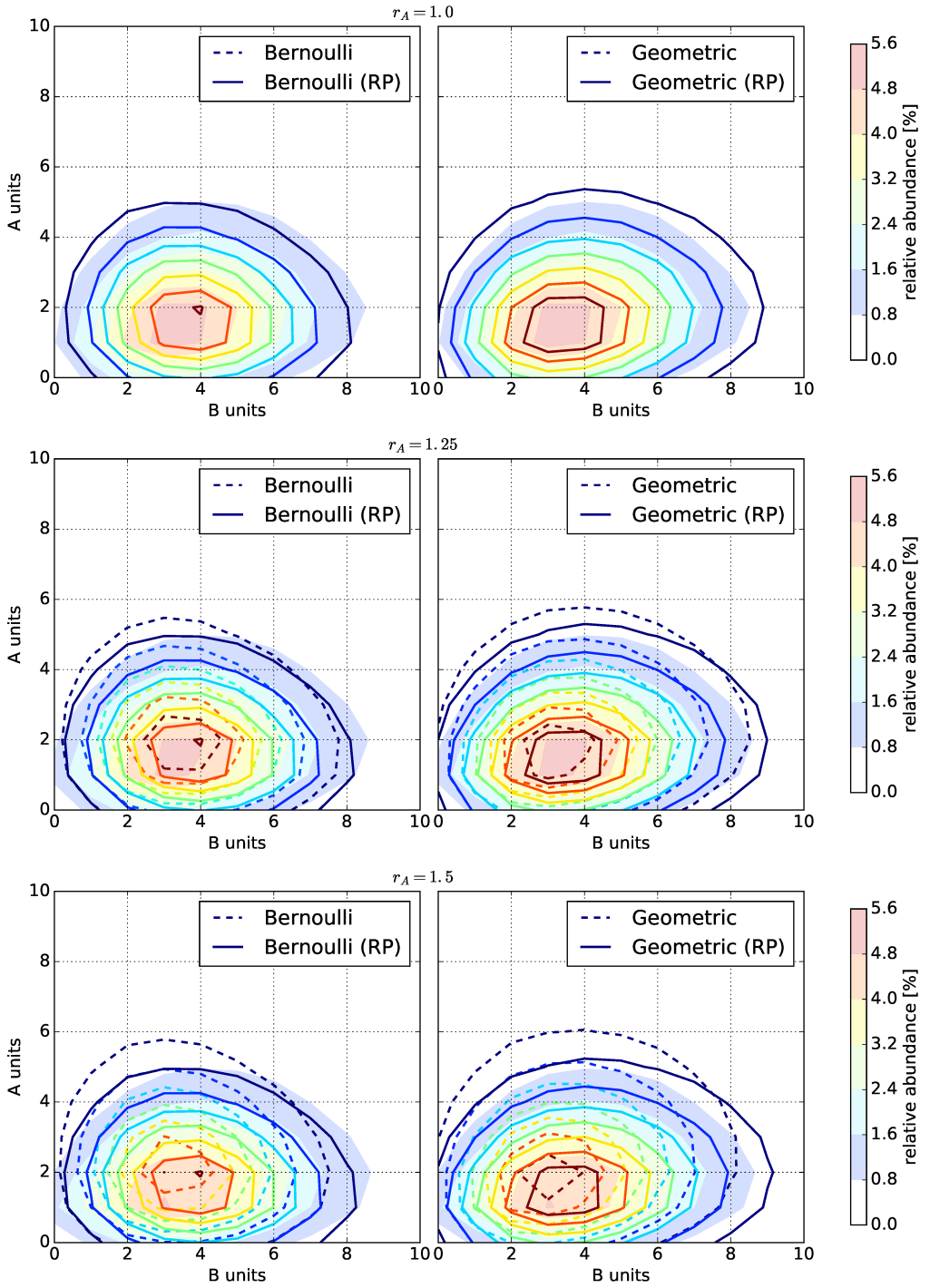


Figure A.12 (continued)

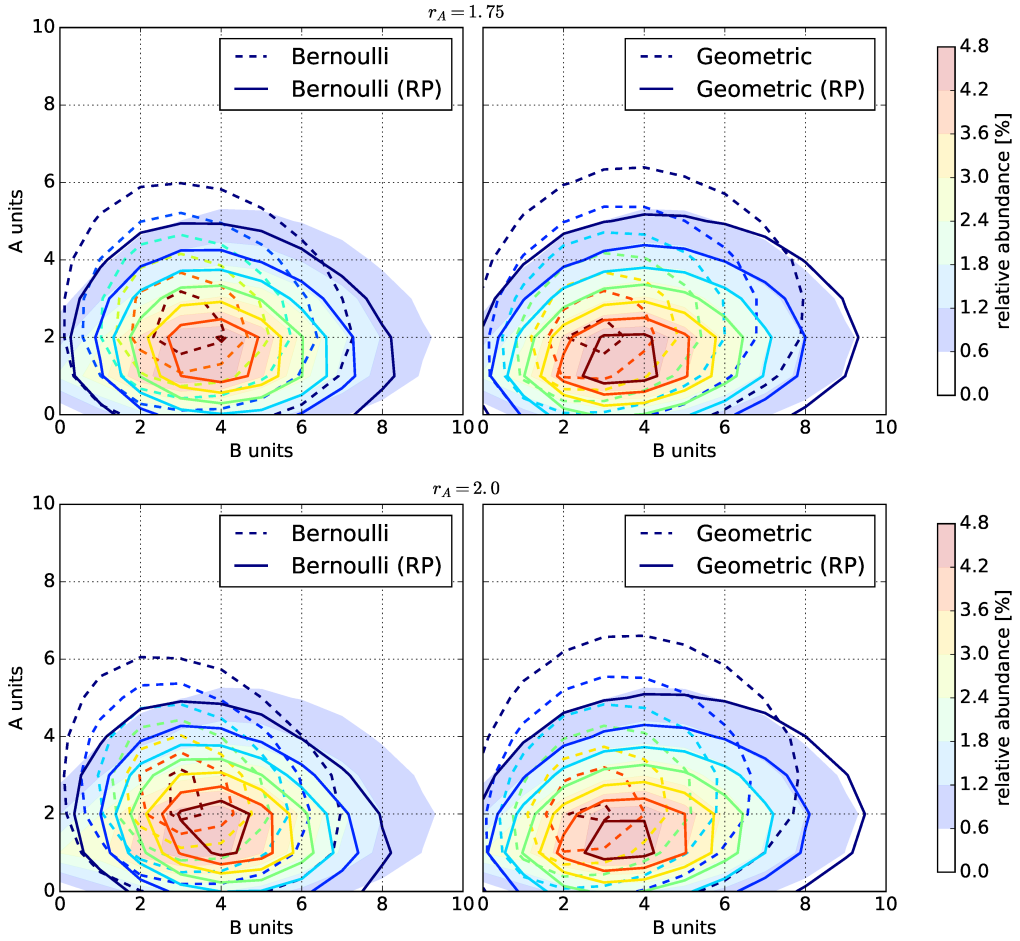


Figure A.12 (continued)

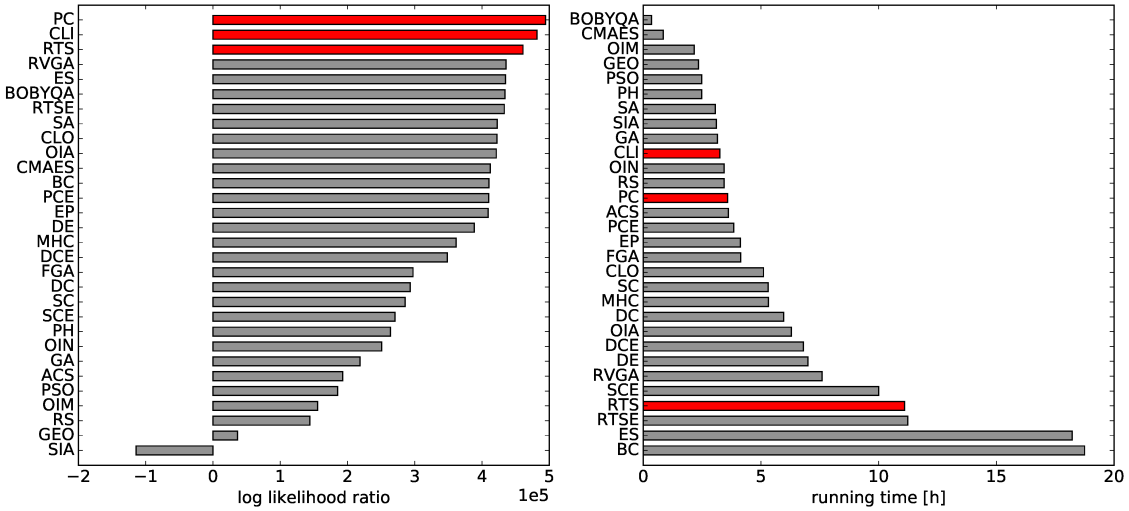


Figure A.13 Left: Log likelihood ratio of all optimization algorithms using the direct method on the $DP_n = 3, r_A = 2.0$ instance without noise. The top three algorithms are marked in red. Right: Running times of the algorithms.

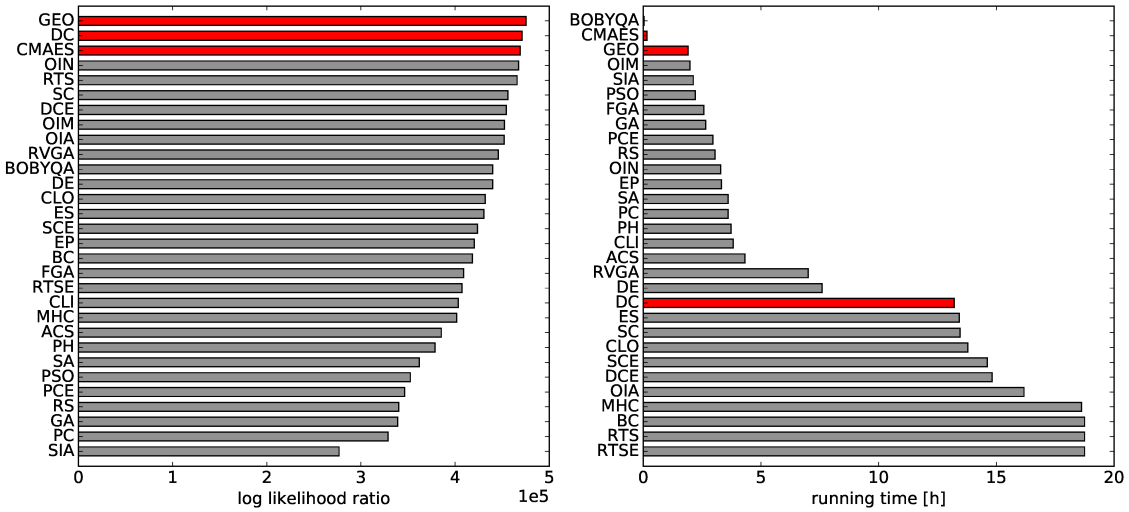


Figure A.14 Left: Log likelihood ratio of all optimization algorithms using the spline method on the $DP_n = 3, r_A = 2.0$ instance without noise. The top three algorithms are marked in red. Right: Running times of the algorithms.

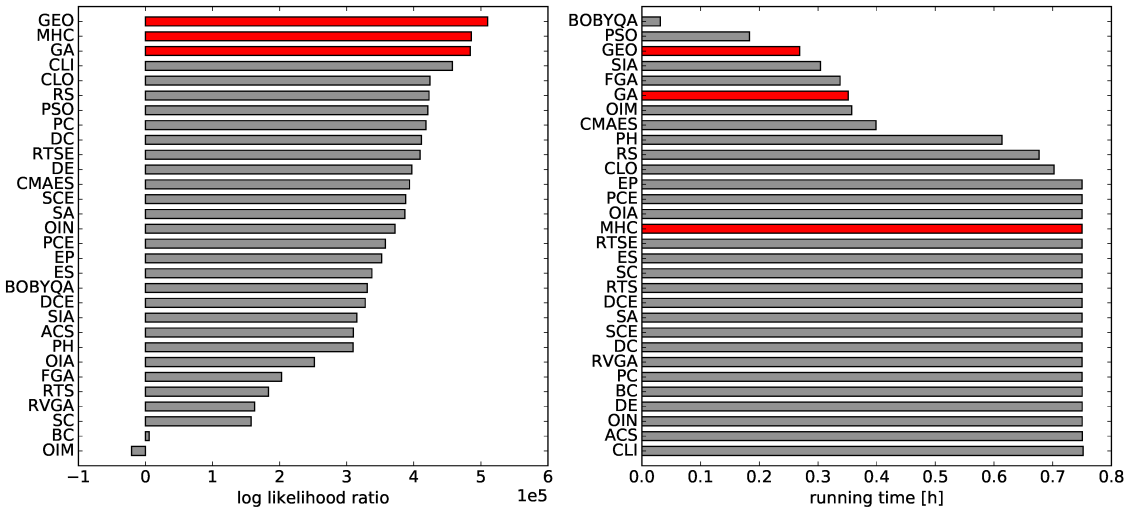


Figure A.15 Left: Log likelihood ratio of all optimization algorithms using the ODE method on the $DP_n = 3, r_A = 2.0$ instance without noise. The top three algorithms are marked in red. Right: Running times of the algorithms.

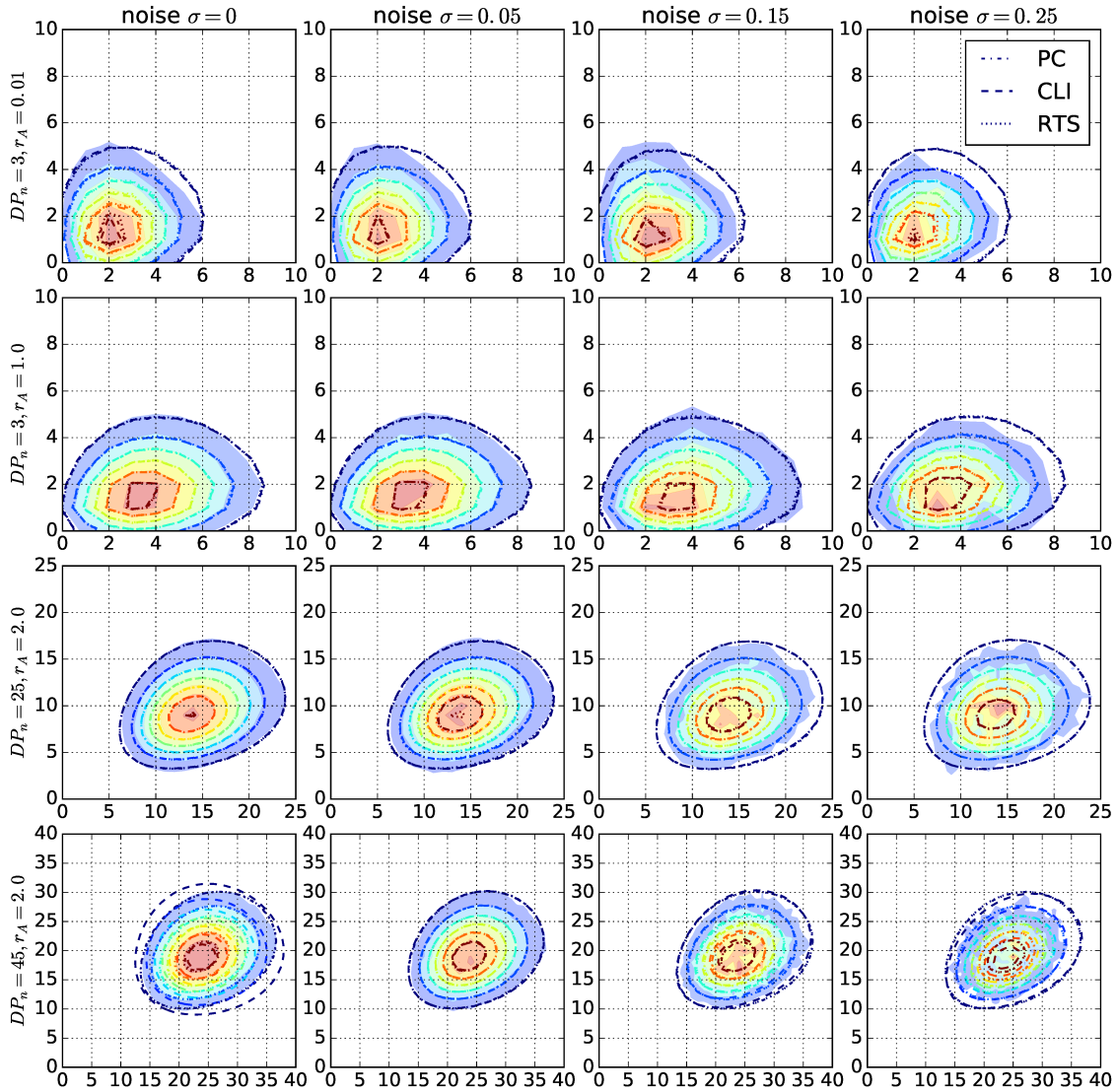


Figure A.16 Filled contours: Fingerprints of the datasets $DP_n = 3, r_\mathrm{A} = 0.01$, $DP_n = 3, r_\mathrm{A} = 1.0$, $DP_n = 25, r_\mathrm{A} = 2.0$, and $DP_n = 45, r_\mathrm{A} = 2.0$, (top to bottom) with increasing noise (left to right). Contours: Fingerprints computed by the model using the direct method.

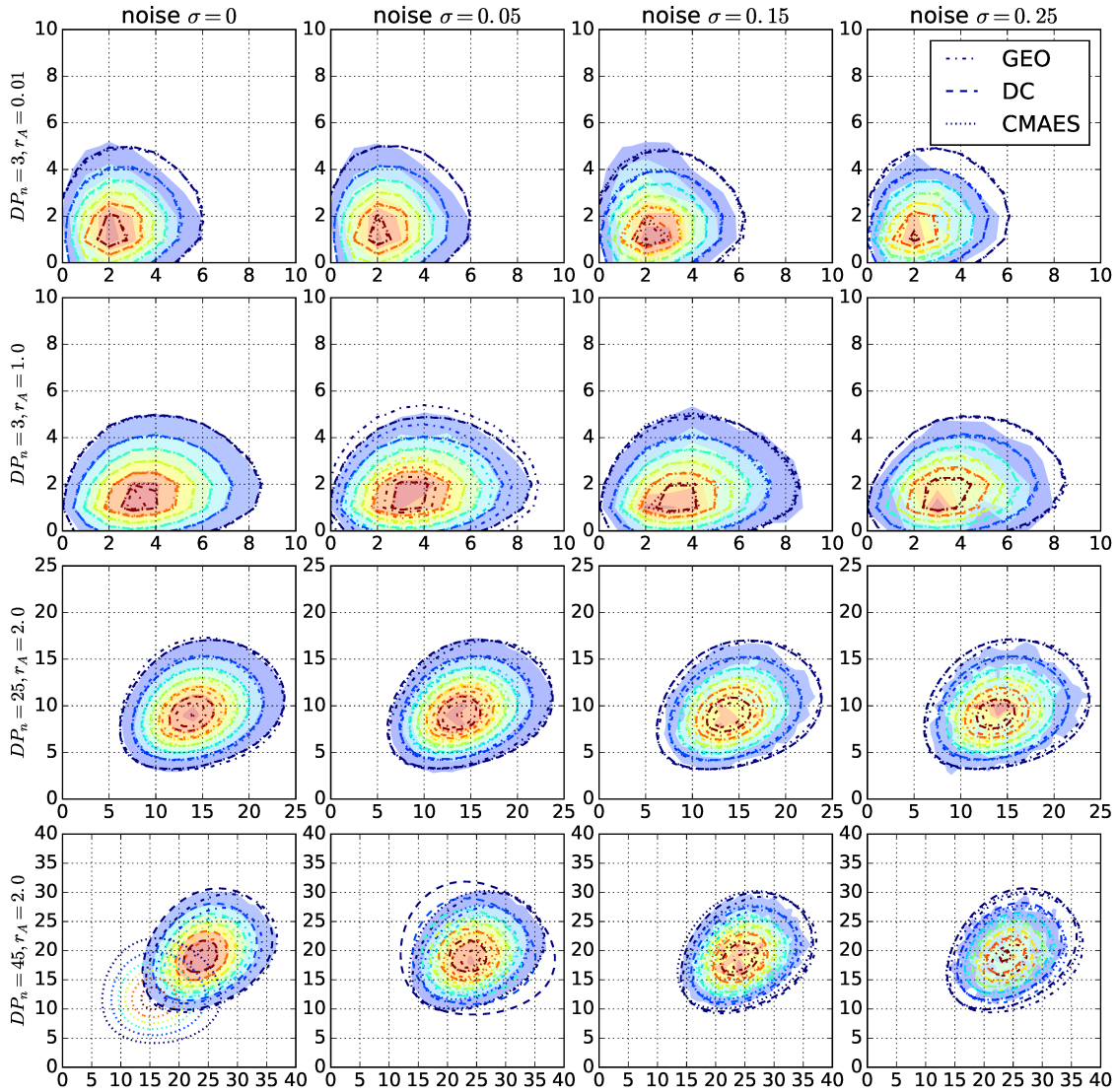


Figure A.17 Filled contours: Fingerprints of the datasets $DP_n = 3, r_\mathrm{A} = 0.01$, $DP_n = 3, r_\mathrm{A} = 1.0$, $DP_n = 25, r_\mathrm{A} = 2.0$, and $DP_n = 45, r_\mathrm{A} = 2.0$, (top to bottom) with increasing noise (left to right). Contours: Fingerprints computed by the model using the spline method.

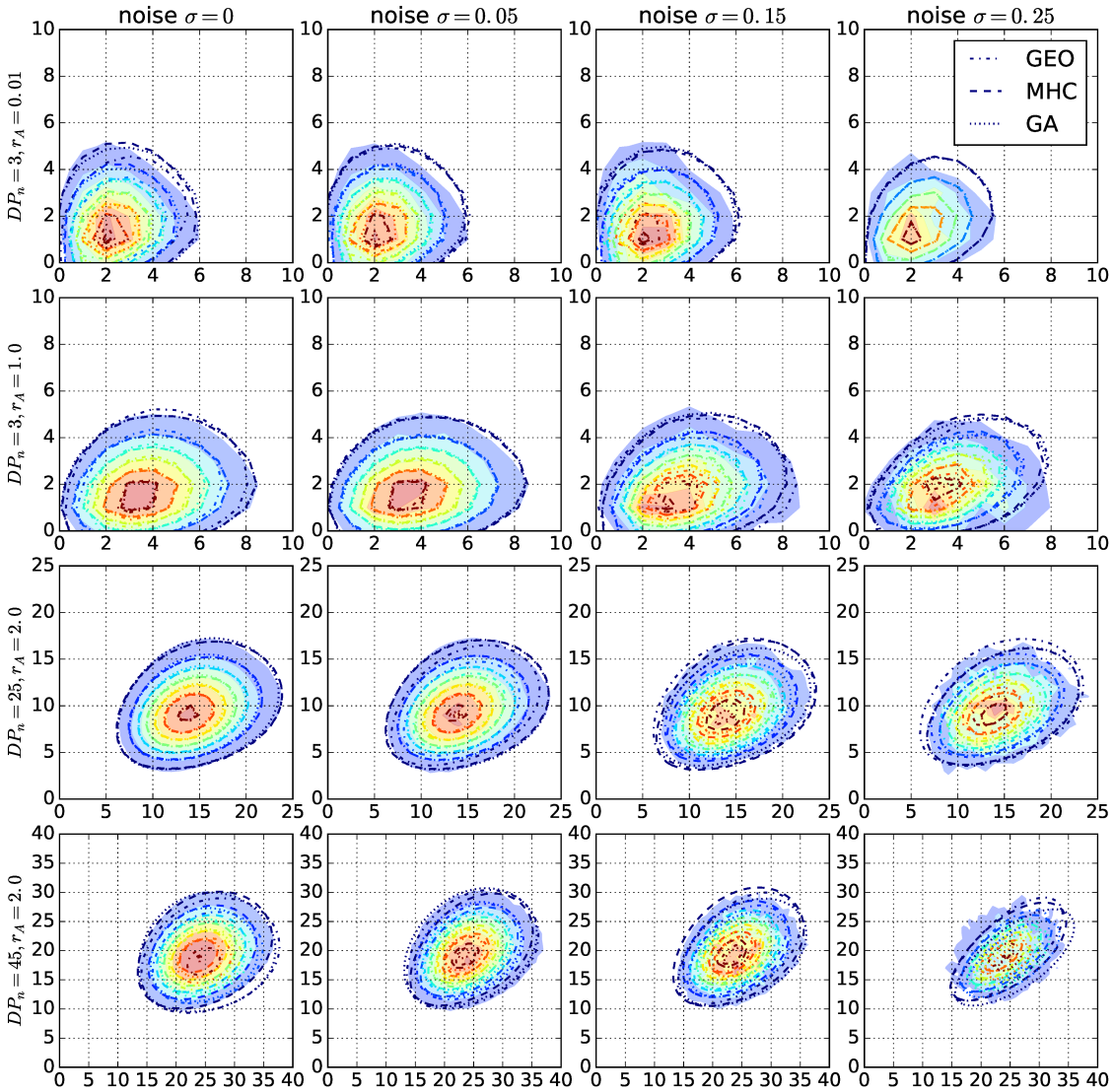


Figure A.18 Filled contours: Fingerprints of the datasets $DP_n = 3, r_\mathrm{A} = 0.01$, $DP_n = 3, r_\mathrm{A} = 1.0$, $DP_n = 25, r_\mathrm{A} = 2.0$, and $DP_n = 45, r_\mathrm{A} = 2.0$, (top to bottom) with increasing noise (left to right). Contours: Fingerprints computed by the model using the ODE method.

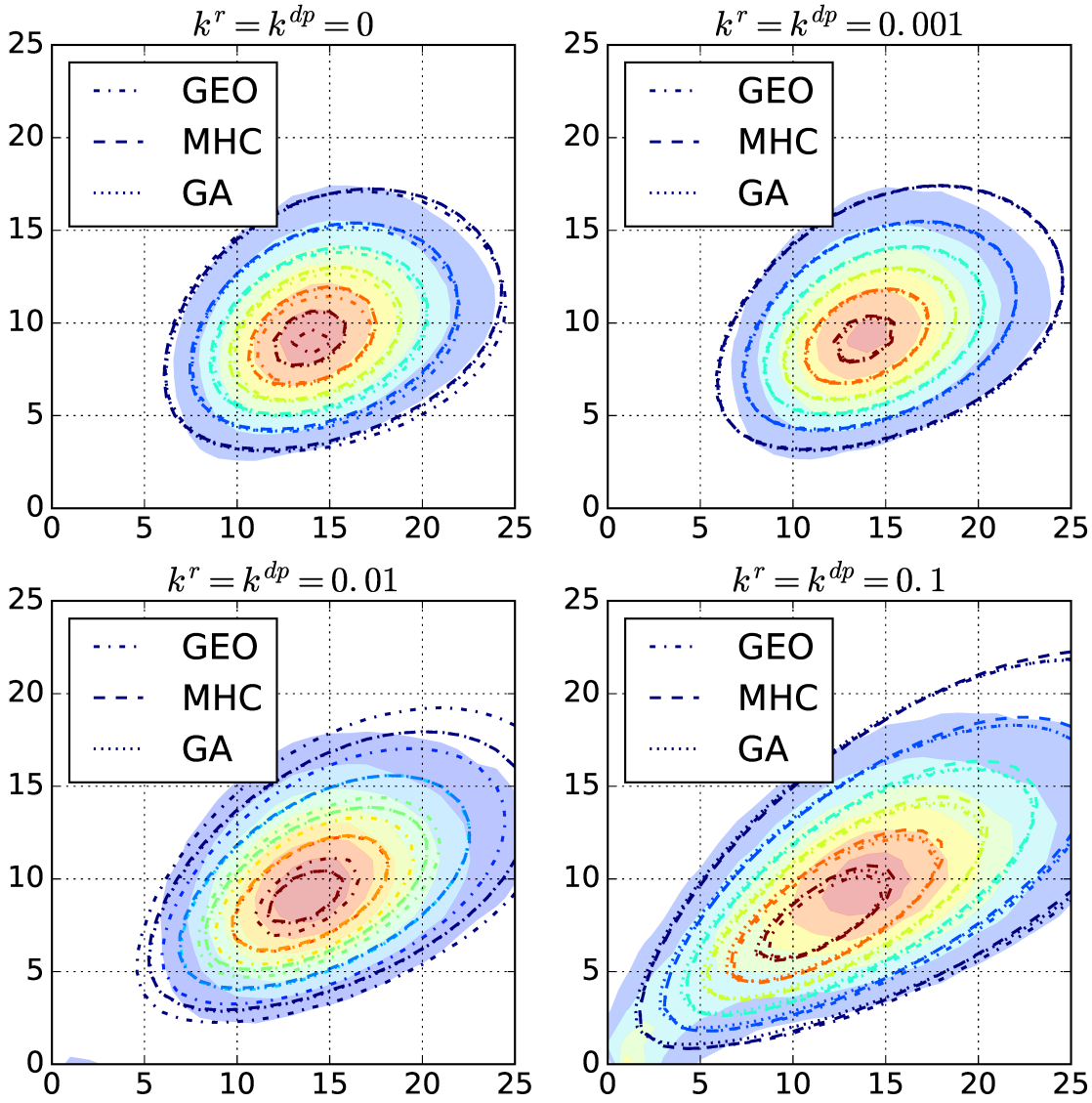


Figure A.19 Filled contours: Fingerprints of the controlled radical copolymerizations with increasing recombination k^r and disproportionation rates k^{dp}. Contours: Fingerprints computed by the model using the ODE method.

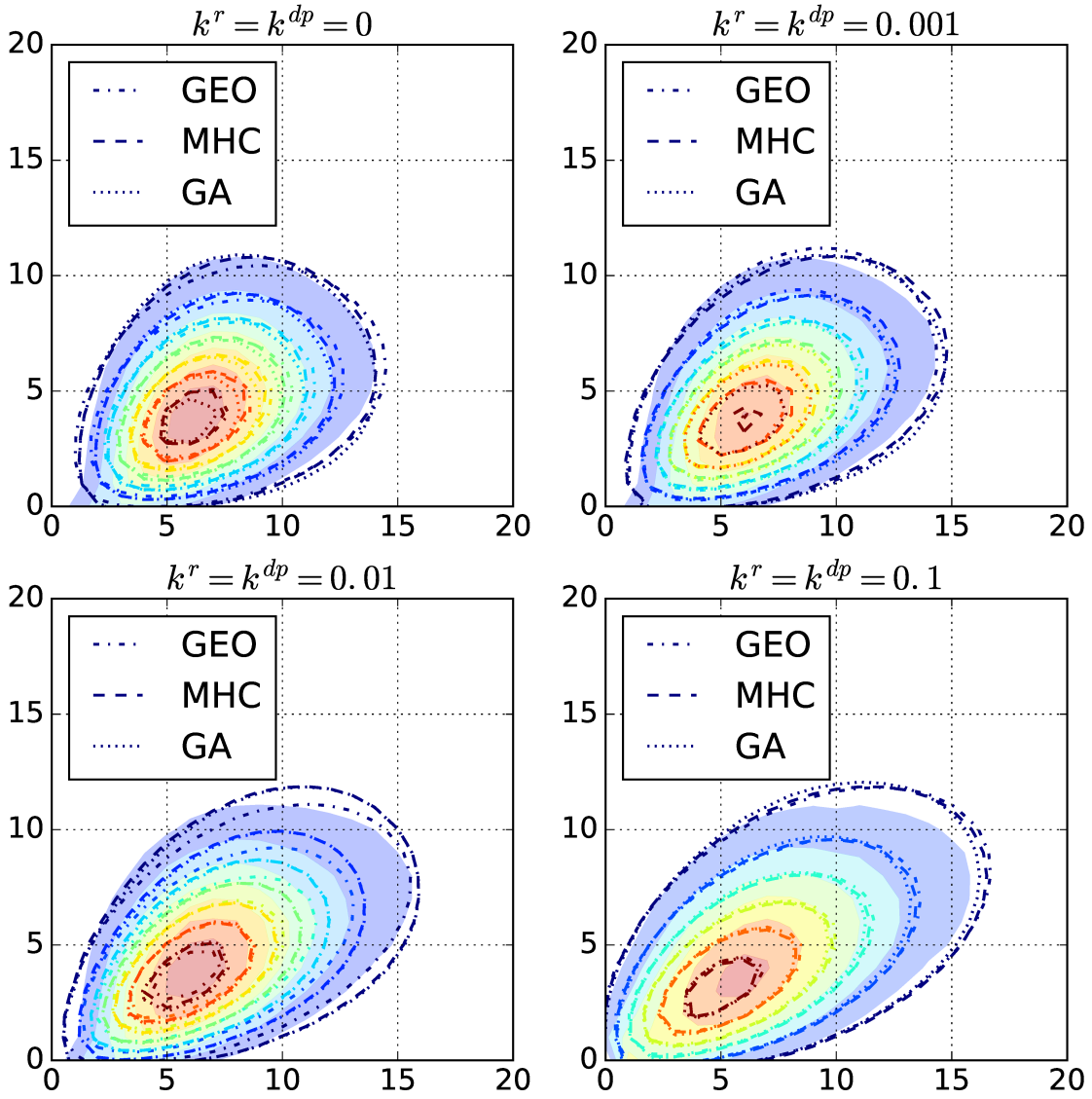


Figure A.20 Filled contours: Fingerprints of the free radical copolymerizations with increasing recombination k^r and disproportionation rates k^{dp}. Contours: Fingerprints computed by the model using the ODE method.

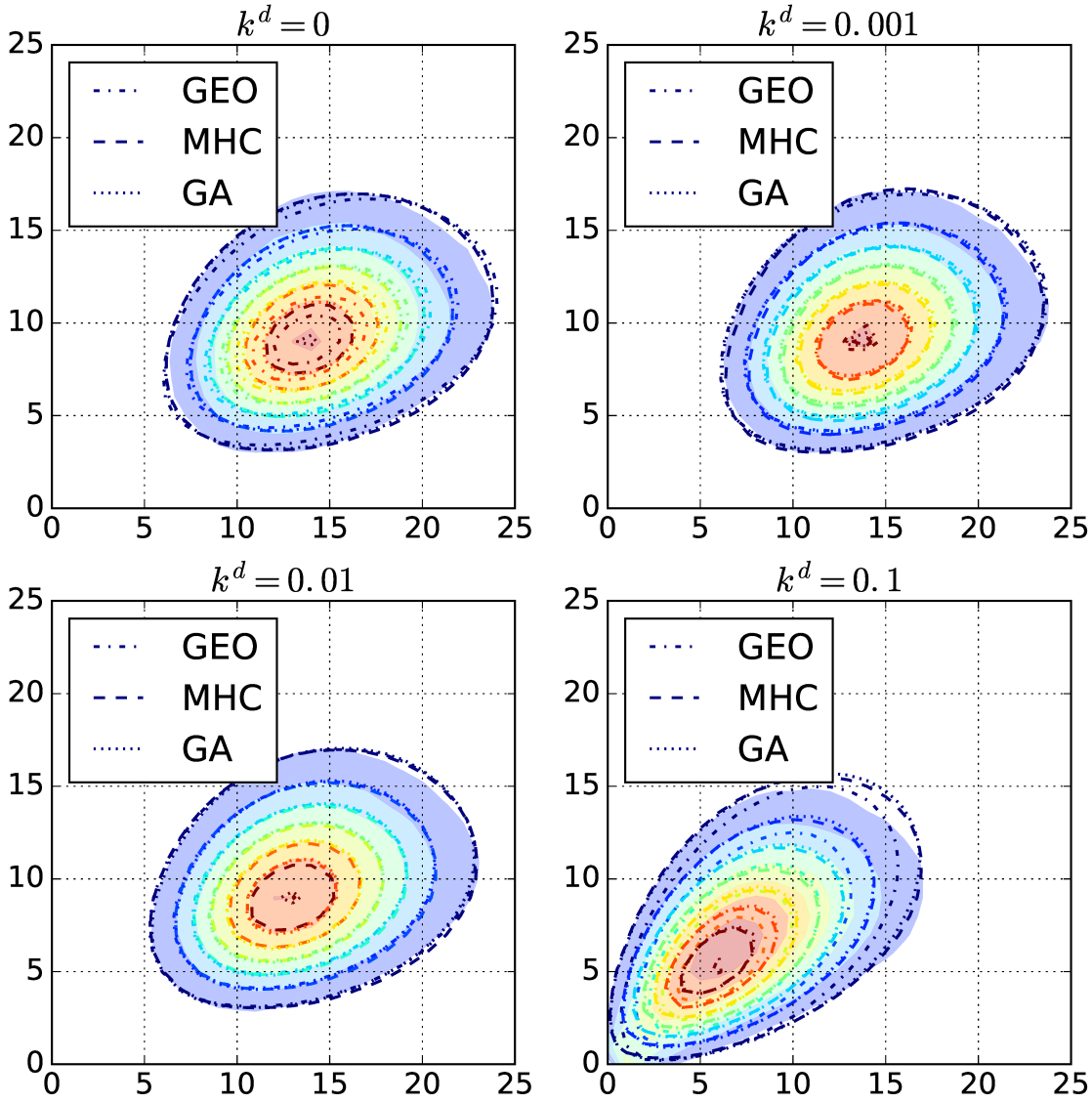


Figure A.21 Filled contours: Fingerprints of the reversible living copolymerizations with increasing depropagation rates k^d. Contours: Fingerprints computed by the model using the ODE method.

Ehrenwörtliche Erklärung

Hiermit erkläre ich

- dass mir die Promotionsordnung der Fakultät bekannt ist,
- dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönliche Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- dass ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts haben mich folgende Personen unterstützt:
Sebastian Böcker, Kerstin Scheubert, Sarah Crotty, Ulrich S. Schubert

Ich habe weder die gleiche, noch eine ähnliche oder eine andere Arbeit an einer anderen Hochschule als Dissertation eingereicht.

Jena, den 16.10.2017

Martin Engler