polymers

MDPI

*Article*

# Exploring the Limits of the Geometric Copolymerization Model

**Martin S. Engler [1], Kerstin Scheubert [2], Ulrich S. Schubert [3,4] and Sebastian Böcker [2,4,*]**

1 Life Sciences Group, Centrum Wiskunde & Informatica, Science Park 123, 1089XG Amsterdam, The Netherlands; martin.engler@cwi.nl
2 Chair of Bioinformatics, Friedrich Schiller University, Ernst-Abbe-Platz 2, 07743 Jena, Germany; kerstin.scheubert@uni-jena.de
3 Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstr. 10, 07743 Jena, Germany; ulrich.schubert@uni-jena.de
4 Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
* Correspondence: sebastian.boecker@uni-jena.de; Tel.: +49-3641-9-46451

Academic Editor: Sébastien Perrier
Received: 4 February 2017; Accepted: 8 March 2017; Published: 13 March 2017

**Abstract:** The geometric copolymerization model is a recently introduced statistical Markov chain model. Here, we 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.

**Keywords:** copolymer kinetics; copolymer fingerprint; Markov model; Monte Carlo simulations

## 1. Introduction

Theoretical models for copolymerization of linear binary copolymers of two monomer types A and B are well established in polymer science. Very recently, we introduced a new statistical model [1]. Here, we investigate its limits with regard to different polymerization types and evaluate several methods to determine the model parameters.

Mass spectrometry (MS) is frequently applied to characterize (*co*-)polymers, in particular matrix-assisted laser desorption/ionization time-of-flight MS [2,3]. Mass spectra can be transformed to copolymer fingerprints [4–7], which represent the two-dimensional distribution of all copolymer chains. A copolymer fingerprint shows the abundance of each possible combination of monomer numbers. This work focuses on copolymer fingerprints of linear binary copolymers.

Several theoretical models for copolymerization were devised in the past, starting with Mayo and Lewis and their terminal model [8], which describes four propagation reactions and is determined by the monomer reactivity ratios. Computational approaches to such a model can be categorized into three types: ordinary differential equations (ODEs), Markov chains, and Monte Carlo methods.

By applying ODEs to the terminal model, Mayo and Lewis deduced the copolymer equation [8], which provides the copolymer composition. Using population balance equations and ODE systems, Kryven and Iedema were able to extract simple sequence patterns, but not the full distribution of sequences [9]. Markov and Hidden Markov models are frequently applied in the field of synthetic

polymers and biopolymers [10–12]. The terminal model can be represented as a Markov chain in a straightforward way [13], which enables the probability of a single copolymer chain, but not the distribution of all chains, to be computed. Monte Carlo methods can be used to simulate chemical reactions [14]. In polymer science, Monte Carlo simulations have been evaluated against experimental data [15–17] and used to compute copolymer fingerprints [18–21]. However, Monte Carlo simulations can be time- and memory-intensive.

In a recent publication, we proposed several variants of a new Markov chain model to characterize the whole distribution of copolymer chains based on a simple living copolymerization scheme [1]. Unlike the traditional terminal model [8], the model allows for variable chain lengths and time-dependent monomer probabilities to model copolymer chain length distributions and differential monomer conversion rates, respectively. In contrast to Monte Carlo simulations, the model is exact and deterministic. In particular, it allows for calculating the exact likelihood of any polymer chain. Monte Carlo simulations provide a random sample, which converges to the true copolymer distribution with an increasing number of simulated chains. However, running times and memory requirements correlate with the number of simulated chains. We showed that the proposed model is faster and requires less memory [1].

The model is a Markov chain with discrete time steps, so-called synthesis steps. Let us consider a single polymer chain. In each synthesis step, the chain is propagated by some number $k \geq 0$ of monomers. To this end, the probabilities of all possible propagation events are calculated from three different probabilities.

First, the probability of adding $k$ monomers, which is constant for each synthesis step, has to be calculated using either a Bernoulli or geometric probability distribution. This leads to a binomial or negative binomial distribution of polymer lengths. The negative binomial is the discrete equivalent to the gamma (Schulz–Zimm) distribution. This is in agreement with the literature, where several distributions for modeling chain lengths can be found: most probable (Schulz–Flory), gamma, Poisson, or hypergeometric distributions [13,22,23]. 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 [23].

Second, the probability of the polymer chain colliding with $x$ A-monomers and $y$ B-monomers, where $x + y = k$ has to be calculated. This probability may change between synthesis steps to reflect differential monomer conversion rates.

Third, basic collision theory has to be considered. In order for a collision of two molecules to be successful, the collision energy needs to be higher than the activation energy. As a consequence, the geometric model needs to incorporate the probability of a successful collision between two monomers.

In our recent publication [1], Monte Carlo simulated data was used to evaluate four variants of the model: The number of added monomers following a Bernoulli or geometric distribution and either using reactivity parameters or not. This work focuses on the variant which was the most accurate, the Geometric model with reactivity parameters. In our previous paper, we suggested that it is possible to estimate the model parameters from the observed copolymer fingerprints [1]. In this paper, 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.

## 2. Methods

In the following, we briefly explain the Geometric copolymerization model and the experimental setup. First, copolymer fingerprints using Monte Carlo methods were computed (see below). Different noise levels were simulated by multiplying the fingerprint abundances by log-normal distributed random noise with mean zero and variance $\sigma$, where the noise parameter $\sigma$ has the values 0, 0.05, 0.15, and 0.25. Then, the model parameter optimizations were performed in parallel on a compute cluster of four 2.4 GHz CPUs with 16 cores each and 6 GB RAM per process. The optimization algorithms are single-threaded, thus the reported running times are similar to the running times to be expected when performing an analysis on a standard laptop. Finally, the log likelihoood ratios were computed as described below.

### *2.1. The Geometric Copolymerization Model*

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. The geometric copolymerization model is Markov chain with discrete time steps, so-called synthesis steps. The states of the Markov chain correspond to the fingerprint entries $M_{a,b}$. However, to incorporate the reactivity parameters, states $M_{a,b}$ have to be divided into $M^{\mathsf{A}}_{a,b}$, copolymer chains with $a$ A-monomers and $b$ B-monomers ending in A, and $M^{\mathsf{B}}_{a,b}$, copolymer chains with $a$ A-monomers and $b$ B-monomers ending in B, with $M_{a,b} = M^{\mathsf{A}}_{a,b} + M^{\mathsf{B}}_{a,b}$. In addition, the initiator state $I$ is defined as $I = M_{0,0}$.

In each synthesis step, the probabilities for all possible transitions $M^{\mathsf{X}}_{a,b}$ to $M^{\mathsf{Y}}_{a+x,b+y}$ and $I$ to $M^{\mathsf{Y}}_{x,y}$ are calculated for $\mathsf{X}, \mathsf{Y} \in \{\mathsf{A}, \mathsf{B}\}$ and all $a$, $b$, $x$ and $y$ (Figure 1). These probabilities are determined by the probability of adding $k = x + y$ monomers, the probabilities of adding $x$ A-monomers and $y$ B-monomers and the reactivity ratios. To compute a fingerprint using the model, the states of the Markov chain are initialized as $I = 1$ and $M^{\mathsf{X}}_{a,b} = 0$ for $\mathsf{X} \in \{\mathsf{A}, \mathsf{B}\}$ and all $a$ and $b$. Subsequently, all possible transitions are applied in each synthesis step $1 \leq t \leq T$. Finally, the fingerprint $M = M^{\mathsf{A}}(T) + M^{\mathsf{B}}(T)$ is computed.

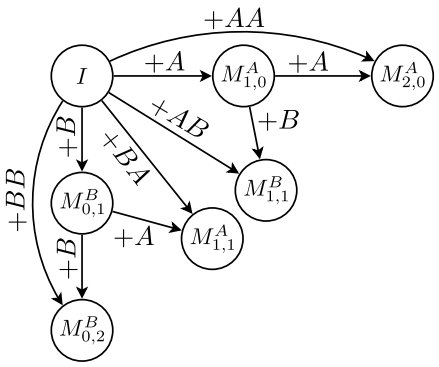


**Figure 1.** All possible transitions for copolymer chain lengths $\leq 2$. For example, the transition from the initiator state $I$ to the state $M^{\mathsf{A}}_{2,0}$ (copolymer chains having two A-monomers and ending in A) corresponds to adding the sequence AA. Note that transitions that add more than two monomers 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.

### 2.2. Monte Carlo Reaction Schemes

We evaluate our methods against Monte Carlo simulations of different polymerization types (Table 1): living polymerization (LP), reversible living polymerization (RLP), free radical polymerization (FRP), and controlled radical polymerization (CRP).

**Table 1.** Overview of the modeled reactions types for the living polymerization (LP), reversible living polymerization (RLP), free radical polymerization (FRP), and controlled radical polymerization (CRP).

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

For living polymerization, the following reaction scheme were 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:

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

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

For reversible living polymerization, the initiation and propagation reactions of the living polymerization and additionally the following depropagation reactions were used:

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

For free and controlled radical polymerization, the initiation and propagation reactions of the living polymerization and, additionally, model chain termination by recombination and disproportionation were used:

$$\left.\begin{array}{l} \sim A^{\bullet} + \sim A^{\bullet} \xrightarrow{k^{r}_{AA}} \sim AA\sim \\ \sim A^{\bullet} + \sim B^{\bullet} \xrightarrow{k^{r}_{AB}} \sim AB\sim \\ \sim B^{\bullet} + \sim B^{\bullet} \xrightarrow{k^{r}_{BB}} \sim BB\sim \end{array}\right\} \text{Recombination}$$

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

For free radical polymerization, 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 were used:

$$\left.I_2 \xrightarrow{k^{dec}} 2{\cdot}I\right\} \text{Initiator Decomposition.}$$

For controlled radical polymerization, the initiation and propagation reactions of the living polymerization, chain termination by recombination and disproportionation, and the following additional activation and deactivation reactions were used:

$$\left.\begin{array}{l} IX + L \xrightarrow{k^{a}_{I}} I + LX \\ \sim AX + L \xrightarrow{k^{a}_{A}} \sim A^{\bullet} + LX \\ \sim BX + L \xrightarrow{k^{a}_{B}} \sim B^{\bullet} + LX \end{array}\right\} \text{Activation}$$

$$\left.\begin{array}{l} I + LX \xrightarrow{k^{da}_{I}} IX + L \\ \sim A^{\bullet} + LX \xrightarrow{k^{da}_{A}} \sim AX + L \\ \sim B^{\bullet} + LX \xrightarrow{k^{da}_{B}} \sim BX + L \end{array}\right\} \text{Deactivation.}$$

### *2.3. Datasets and Monte Carlo Parameters*

For all datasets, the reaction rates (Supplementary Table S1) were chosen such that $r_A = \frac{1}{r_B} = r$, with the reactivity ratios $r_A = \frac{k_{AA}}{k_{AB}}$, $r_B = \frac{k_{BB}}{k_{BA}}$, and the ratio of homopropagation rates $r = \frac{k_{AA}}{k_{BB}}$.

For living polymerization, we use three Monte Carlo simulated datasets reported in our previous paper [1]: $r_A = 0.01$, $r_A = 1.0$, and $r_A = 2.0$. Two additional datasets were simulated with the same reactivity ratios $r_A = 2.0$, but different degrees of polymerization $DP_n = 25$ and $DP_n = 45$, respectively. For an overview of the initial concentrations and reaction rates, please see Supplementary Table S1.

For Monte Carlo simulations of the other polymerization types, the parameters of the dataset with $DP_n = 25, r_A = 2.0$ 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.

### *2.4. Log Likelihood Ratio*

Our model allows for computing the likelihood of a single polymer chain [24]. Let $S$ be a sequence (polymer chain) and $H$ a hypothesis, i.e., the geometric model. Let $P(S|H)$ be the likelihood of $S$, given model $H$. Then, the log likelihood of a dataset $D$ is:

$$\log P(D|H) = \sum_{S \in D} \log P(S|H). \tag{1}$$

To evaluate the models, we 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 $i$ over all chains in the dataset, we determine the frequencies $f_\mathsf{A}$ and $f_\mathsf{B}$ 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:

$$\log P(D|H_0) = \sum_{S \in D} \log P(S|H_0) = \sum_{S \in D} \sum_{i=1}^{|S|} \log P(s_i) = \sum_{S \in D} \sum_{i=1}^{|S|} \log \begin{cases} f_\mathsf{A}, \text{ if } s_i = \mathsf{A}, \\ f_\mathsf{B}, \text{ if } s_i = \mathsf{B}. \end{cases} \tag{2}$$

We compute the log likelihood ratio:

$$\log \frac{P(D|H)}{P(D|H_0)} = \log P(D|H) - \log P(D|H_0). \tag{3}$$

The log likelihood ratio is a "sanity check" for statistical models. If the ratio is below zero, the hypothesis should be dismissed and accepted if the ratio is above zero.

## 3. Results and Discussion

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 $M^c$ [1]. The model parameters are the monomer probability $p_M$, the reactivity probabilities $p_\mathsf{AA}$, $p_\mathsf{AB}$, $p_\mathsf{BA}$, $p_\mathsf{BB}$, 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 [1].

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

$$\underset{p_\mathsf{A}, p_\mathsf{AA}, p_\mathsf{AB}, p_\mathsf{BA}, p_\mathsf{BB}}{\arg\min} \; ||f(p_\mathsf{A}, p_\mathsf{AA}, p_\mathsf{AB}, p_\mathsf{BA}, p_\mathsf{BB}) - M^o||_2. \tag{4}$$

The objective function computes the difference between the computed and observed fingerprints according to Equation (4). 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 h (see Supplementary Figure S1). 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 our previous work [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 $p_\mathsf{A}$.

The models are order-independent if the reactivity ratios are one (see the Supplementary section). Since there are $T!$ possible

permutations, this results in $T!$ global optima. However, for reactivity ratios of one, the ratios of monomers never change. As a consequence, $p_{\mathsf{A}}$ 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 $p_{\mathsf{A}}$, 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 $p_M$ [1]. 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 $\sim$100 variables simultaneously with a general purpose optimizer is a challenging task and certainly contributes to the long running time of the optimization.

### *3.1. 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 $p_{\mathsf{A}}$ (25% in this work), and linearly interpolate all other values in between. Furthermore, we restrict the search space by forcing $p_{\mathsf{A}}$ to be either increasing or decreasing. To this end, a decreasing $p_{\mathsf{A}}$ is defined as:

$$\begin{aligned} p_{\mathsf{A}}(t) &= p(t) \cdot p_{\mathsf{A}}(t-1), \\ p_{\mathsf{A}}(1) &= p(1), \end{aligned} \tag{5}$$

and an increasing $p_{\mathsf{A}}$ as:

$$\begin{aligned} p_{\mathsf{A}}(t) &= p_{\mathsf{A}}(t-1) + p(t) \cdot (1 - p_{\mathsf{A}}(t-1)), \\ p_{\mathsf{A}}(1) &= p(1). \end{aligned} \tag{6}$$

The second approach is to exploit the relationship between $p_{\mathsf{A}}$ 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:

$$p_{\mathsf{A}}(t) = \frac{\widetilde{[\mathsf{A}]}(t)}{\widetilde{[\mathsf{A}]}(t) + \widetilde{[\mathsf{B}]}(t)}. \tag{7}$$

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:

$$p_{\mathsf{XY}} = \frac{k_{\mathsf{XY}}}{k_{\mathsf{XA}} + k_{\mathsf{XB}}}. \tag{8}$$

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 $p_{\mathsf{A}}$ 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 Equation (4). Thus, the dimension of the search space is constant and independent of $T$.

### *3.2. 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 three of the 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 nonlinear. 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 [25,26] and Apache Math Commons 3.2 library [27]. The algorithms use different strategies to find the best 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, $r_A = \frac{1}{r_B} = r$. First, we choose the reactivity ratio $r_A = 2.0$, for which the geometric model can provide a good fit [1]. Second, we choose $r_A = 0.01$, 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 the geometric model [1]. Third, we choose $r_A = 1.0$. This results in constant monomer concentrations and, thus, the optimal $p_A$ is also constant. This means that 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 $r_A = 2.0$ 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 2). To this end, all algorithms are evaluated on the $DP_n = 3, r_A = 2.0$ dataset without noise (see Supplementary Figures S1–S3) and the log likelihood ratios of the results are calculated. In addition 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 that 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.

**Table 2.** Overview of the top three optimization algorithms for each fingerprint-generating function, selected based on Supplementary Figures S1–S3. We ranked the results of the algorithms for each dataset based on the log likelihood ratios and counted the ranks.

| | **Algorithm** | **#Ranks** | | |
|---|---|---|---|---|
| | | **1st** | **2nd** | **3rd** |
| Direct | Cloning, Information Gain, Aging (CLI) [28] | 4 | 5 | 7 |
| | Probabilistic Crowding (PC) [29] | 6 | 5 | 5 |
| | Restricted Tournament Selection (RTS) [30] | 6 | 6 | 4 |
| Spline | Covariance Matrix Adaptation Evolution Strategy (CMAES) [31] | 3 | 6 | 7 |
| | Deterministic Crowding (DC) [32] | 3 | 8 | 5 |
| | Generalized Extremal Optimization (GEO) [33] | 10 | 2 | 4 |
| ODE | Genetic Algorithm (GA) [34] | 5 | 9 | 2 |
| | Generalized Extremal Optimization (GEO) [33] | 8 | 0 | 8 |
| | Mutation Hill Climber (MHC) [35] | 3 | 7 | 6 |

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 (Figure 2 and Supplementary Figures S4–S6). For each resulting parameterization,

we rank the top three algorithms by their log likelihood ratio and count the ranks for all instances (Table 2). No algorithm outperforms its rivals. Therefore, in the following, we use all chosen algorithms.

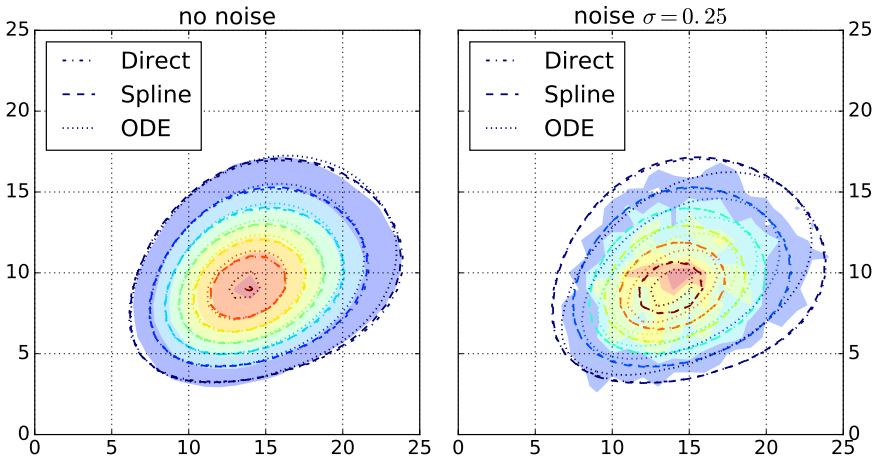


**Figure 2.** Filled contours: copolymer fingerprints of $DP_n = 25$ computed by Monte Carlo simulations with no (**left**) and high applied noise (**right**). Contours: fingerprints computed by the geometric model using the best parameters computed by the optimization algorithms for each of the fingerprint-generating functions (direct, spline, and ODE).

To compare the three approaches (direct, spline, and ODE), we average the log likelihood ratios over all three algorithms for each fingerprint-generating function. Figure 3 shows the averaged log likelihood ratios as a function of the noise level. There are two different behaviors for $r_A = 0.01$ and the rest of the instances. For $r_A = 0.01$, 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 $r_A > 0.01$, 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 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 for each fingerprint-generating function for each degree of polymerization $DP_n$ = 3, 25, and 45 (Figure 4). 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. This means that using different optimizers may shift the numbers, but we can still infer general trends from Figure 4.

The running times 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, 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 (see Supplementary Figures S1–S3).

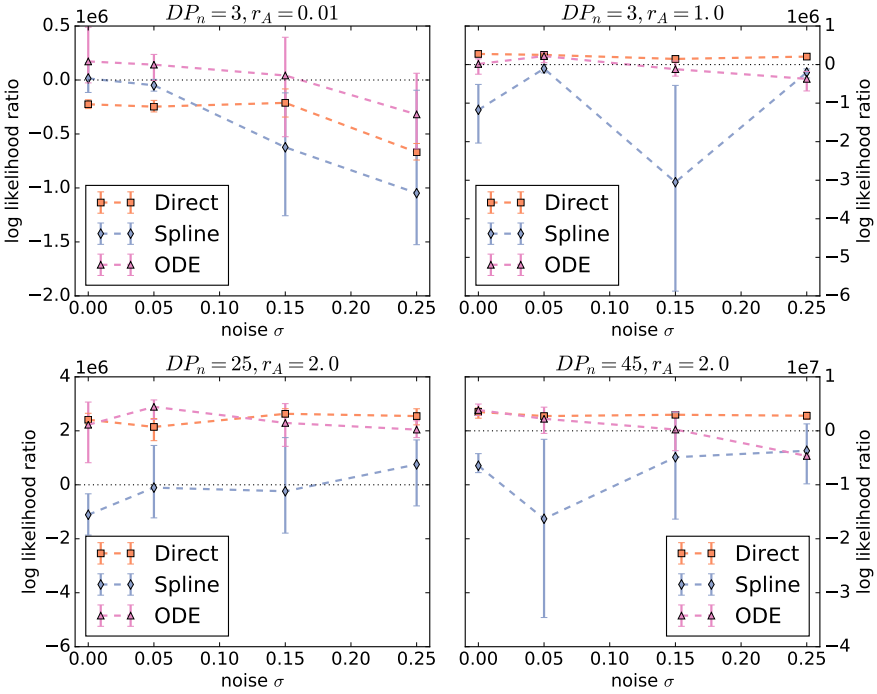


**Figure 3.** Log likelihood ratios of the results computed by the optimization algorithms as a function of noise. The ratios are averaged over all three algorithms for each fingerprint-generating function (direct, spline, ODE). The higher the ratios, the better the observed data is "explained" by the identified model parameterizations. If the ratio is below zero, the null model achieves a higher likelihood than the geometric model with the given parameterization.

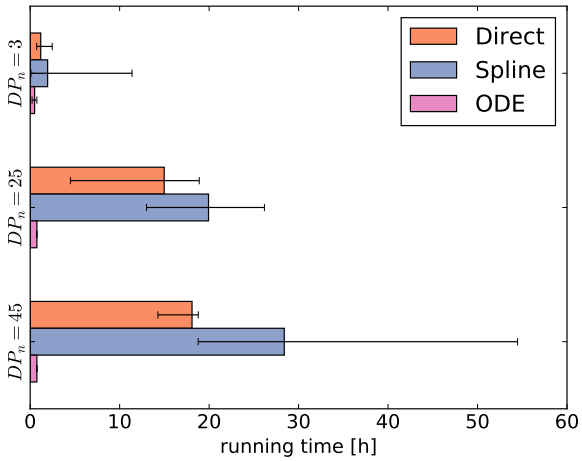


**Figure 4.** Running times of the optimizations averaged over all datasets with degree of polymerization $DP_n$ = 3, 25, and 45 for each fingerprint-generating function (direct, spline, ODE).

### *3.3. Beyond Living Polymerization*

Here, we investigate copolymerizations beyond a simple living polymerization. We select the $DP_n = 25$, $r_A = 2.0$ 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 (Figure 5 and Supplementary Figures S7–S9).

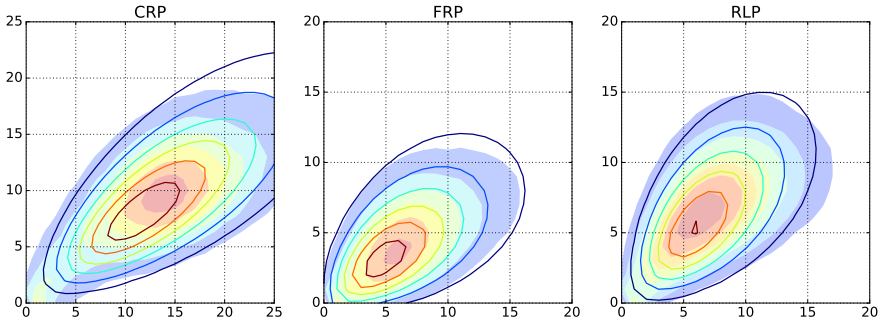


**Figure 5.** Filled contours: copolymer fingerprints of the Monte Carlo simulations of controlled radical polymerization (CRP, **left**), free radical polymerization (FRP, **center**), and reversible living polymerization (RLP, **right**) with the highest used termination and propagation reaction rates of 0.1. Contours: fingerprints computed by the model with the best parameters resulting from the optimizations using the ODE fingerprint-generating function.

We select the ODE method to identify the optimal model parameters. Figure 6 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 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.

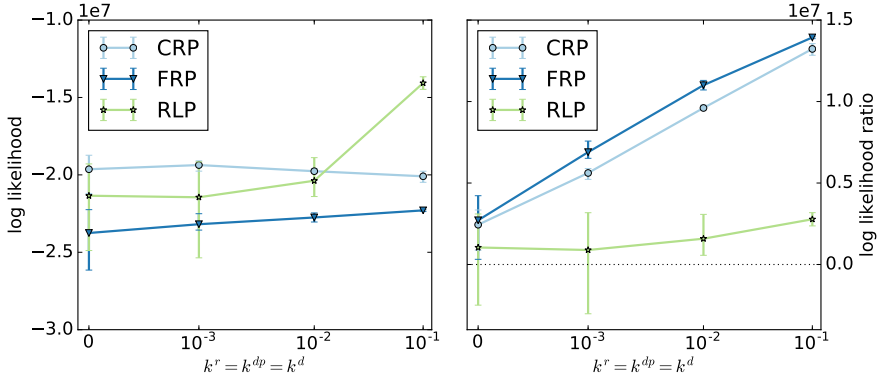


**Figure 6.** Log likelihoods (**left**) and log likelihood ratios (**right**) of the results from the optimizations using the ODE fingerprint-generating function for the controlled radical polymerization (CRP), free radical polymerization (FRP), and reversible living polymerization (RLP) as a function of termination and depropagation rates.

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.

## 4. Conclusions

In a previous publication, we evaluated four variants of a statistical copolymerization model [1]. Here, we concentrated on the variant, which was the most accurate, the geometric model with reactivity parameters. The model computes a copolymer fingerprint.

First, the problem solve is 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 length distribution, but has a long and 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 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.

For those interested in the theoretical aspects, the question remains open on whether the objective function is convex and smooth. More interesting from a practical viewpoint, the geometric model allows for computing previously inaccessible statistical properties of synthesized copolymers. This will be described in a forthcoming publication. 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. Still to determine is if 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.

**Supplementary Materials:** The following are available online at www.mdpi.com/2073-4360/9/3/101/s1.

**Acknowledgments:** We thank Sarah Crotty for fruitful discussions. Funding by the Thüringer Ministerium für Bildung, Wissenschaft und Kultur (Grant No. 12038-514).

**Author Contributions:** Martin S. Engler, Kerstin Scheubert, Ulrich S. Schubert, and Sebastian Böcker jointly contributed to model development; Martin S. Engler performed the experiments and drafted the paper; and Martin S. Engler, Kerstin Scheubert, Ulrich S. Schubert, and Sebastian Böcker finalized the paper.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MS | Mass spectrometry |
| ODE | Ordinary differential equation |
| LP | Living polymerization |
| RLP | Reversible living polymerization |
| FRP | Free radical polymerization |
| CRP | Controlled radical polymerization |

## References

1. Engler, M.S.; Scheubert, K.; Schubert, U.S.; Böcker, S. New Statistical Models for Copolymerization. *Polymers* **2016**, *8*, 240.
2. Montaudo, M.S. Mass spectra of copolymers. *Mass Spectrom. Rev.* **2002**, *21*, 108–144.
3. Pasch, H.; Schrepp, W. (Eds.) *MALDI-TOF Mass Spectrometry of Synthetic Polymers*; Springer: Berlin/Heidelberg, Germany, 2003.

4. Vivó-Truyols, G.; Staal, B.; Schoenmakers, P.J. Strip-based regression: A method to obtain comprehensive co-polymer architectures from matrix-assisted laser desorption ionisation-mass spectrometry data. *J. Chromatogr. A* **2010**, *1217*, 4150–4159.
5. Weidner, S.M.; Falkenhagen, J.; Bressler, I. Copolymer Composition Determined by LC-MALDI-TOF MS Coupling and MassChrom2D Data Analysis. *Macromol. Chem. Phys.* **2012**, *213*, 1521–3935.
6. Horský, J.; Walterová, Z. Fingerprint Multiplicity in MALDI-TOF Mass Spectrometry of Copolymers. *Macromol. Symp.* **2014**, *339*, 9–16.
7. Engler, M.S.; Crotty, S.; Barthel, M.J.; Pietsch, C.; Knop, K.; Schubert, U.S.; Böcker, S. COCONUT—An Efficient Tool for Estimating Copolymer Compositions from Mass Spectra. *Anal. Chem.* **2015**, *87*, 5223–5231.
8. Mayo, F.R.; Lewis, F.M. Copolymerization. I. A Basis for Comparing the Behavior of Monomers in Copolymerization; The Copolymerization of Styrene and Methyl Methacrylate. *J. Am. Chem. Soc.* **1944**, *66*, 1594–1601.
9. Kryven, I.; Iedema, P.D. Deterministic Modeling of Copolymer Microstructure: Composition Drift and Sequence Patterns. *Macromol. React. Eng.* **2015**, *9*, 285–306.
10. González Díaz, H.; Molina, R.; Uriarte, E. Stochastic molecular descriptors for polymers. 1. Modelling the properties of icosahedral viruses with 3D-Markovian negentropies. *Polymer* **2004**, *45*, 3845–3853.
11. González-Díaz, H.; Pérez-Bello, A.; Uriarte, E. Stochastic molecular descriptors for polymers. 3. Markov electrostatic moments as polymer 2D-folding descriptors: RNA-QSAR for mycobacterial promoters. *Polymer* **2005**, *46*, 6461–6473.
12. González-Díaz, H.; Saíz-Urra, L.; Molina, R.; Uriarte, E. Stochastic molecular descriptors for polymers. 2. Spherical truncation of electrostatic interactions on entropy based polymers 3D-QSAR. *Polymer* **2005**, *46*, 2791–2798.
13. Brandrup, J.; Immergut, E.H.; Grulke, E.A. (Eds.) *Polymer Handbook*, 4th ed.; Wiley: Hoboken, NJ, USA, 1999.
14. Gillespie, D.T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **1977**, *81*, 2340–2361.
15. Willemse, R.X.E. New Insights into Free-Radical (Co)Polymerization Kinetics. Ph.D. Thesis, University of Technology Eindhoven, Eindhoven, The Netherlands, 2005.
16. Drache, M.; Schmidt-Naake, G.; Buback, M.; Vana, P. Modeling RAFT polymerization kinetics via Monte Carlo methods: cumyl dithiobenzoate mediated methyl acrylate polymerization. *Polymer* **2005**, *46*, 8483–8493.
17. Drache, M. Modeling the Product Composition During Controlled Radical Polymerizations with Mono- and Bifunctional Alkoxyamines. *Macromol. Symp.* **2009**, *275-276*, 52–58.
18. Szymanski, R. 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* **2009**, *9*, 538–552.
19. Van Steenberge, P.H.M.; D'hooge, D.R.; Wang, Y.; Zhong, M.; Reyniers, M.F.; Konkolewicz, D.; Matyjaszewski, K.; Marin, G.B. Linear Gradient Quality of ATRP Copolymers. *Macromolecules* **2012**, *45*, 8519–8531.
20. Drache, M.; Drache, G. Simulating Controlled Radical Polymerizations with mcPolymer—A Monte Carlo Approach. *Polymers* **2012**, *4*, 1416–1442.
21. Van Steenberge, P.H.M.; D'hooge, D.R.; Reyniers, M.F.; Marin, G.B. Improved kinetic Monte Carlo simulation of chemical composition-chain length distributions in polymerization processes. *Chem. Eng. Sci.* **2014**, *110*, 185–199.
22. Gody, G.; Zetterlund, P.B.; Perrier, S.; Harrisson, S. The limits of precision monomer placement in chain growth polymerization. *Nat. Commun.* **2016**, *7*, 10514.
23. Tobita, H. Molecular Weight Distribution of Living Radical Polymers. *Macromol. Theory Simul.* **2006**, *15*, 12–22.
24. Engler, M.S.; Crotty, S.; Barthel, M.J.; Pietsch, C.; Schubert, U.S.; Böcker, S. Abundance correction for mass discrimination effects in polymer mass spectra. *Rapid Commun. Mass Spectrom.* **2016**, *30*, 1233–1241.
25. Brownlee, J. *OAT: The Optimization Algorithm Toolkit*; Technical Report; Swinburne University of Technology: Victoria, Australia, 2007.
26. The Optimization Algorithm Toolkit. Available online: https://sourceforge.net/projects/optalgtoolkit/ (accessed on 1 May 2014).
27. Apache Math Commons 3.2. Available online: http://commons.apache.org/proper/commons-math/ (accessed on 1 May 2014).

28. Cutello, V.; Nicosia, G. The clonal selection principle for in silico and in vitro computing. In *Recent Developments in Biologically Inspired Computing*; von Zuben, F.J., Ed.; Idea Group Publishing: Hershey, PA, USA, 2004; pp. 104–147.
29. Menshoel, O.J.; Goldberg, D.E. *Probabilistic Crowding: Deterministic Crowding with Probabilistic Replacement*; Technical Report; University of Illinois: Champaign, IL, USA, 1999.
30. Harik, G.R. Finding Multimodal Solutions Using Restricted Tournament Selection. In Proceedings of the Sixth International Conference on Genetic Algorithms, Pittsburgh, PA, USA, 15–19 July 1995; Morgan Kaufmann: San Fransisco, CA, USA, 1995.
31. Hansen, N.; Müller, S.D.; Koumoutsakos, P. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). *Evol. Comput.* **2003**, *11*, 1–18.
32. Mahfoud, S.W. Crowding and preselection revisited. In Proceedings of the Second Conference on Parallel Problem Solving from Nature, Brussels, Belgium, 28–30 September 1992; Elsevier Science Inc.: New York, NY, USA, 1992; pp. 27–36.
33. De Sousa, F.L.; Ramos, F.M.; Galski, R.L.; Muraoka, I. Generalized extremal optimization: A new meta-heuristic inspired by a model of natural evolution. In *Recent Developments in Biologically Inspired Computing*; de Castro, L.N., von Zuben, F.J., Eds.; Idea Group Publishing: Hershey, PA, USA, 2005; pp. 41–60.
34. Back, T.; Fogel, D.B.; Michalwicz, Z. *Evolutionary Computation 1—Basic Algorithms and Operators*; Institute of Physics (IoP) Publishing: Bristol, UK, 2000.
35. Mühlenbein, H. How Genetic Algorithms Really Work: Mutation and Hillclimbing. In *Proceedings of the Second Conference on Parallel Problem Solving from Nature;* Elsevier: Amsterdam, The Netherlands, 1992.