# COCONUT — An Efficient Tool for Estimating Copolymer Compositions from Mass Spectra

Martin S. Engler,[†,§] Sarah Crotty,[‡,¶,§] Markus J. Barthel,[‡,¶,∥] Christian Pietsch,[‡,¶] Katrin Knop,[‡,¶] Ulrich S. Schubert,[‡,¶] and Sebastian Böcker[*,†,¶]

*Chair of Bioinformatics, Laboratory of Organic and Macromolecular Chemistry, and Jena Center for Soft Matter, Friedrich Schiller University Jena, Germany*

E-mail: sebastian.boecker@uni-jena.de

This is the preprint version of Engler *et al.*: COCONUT — An Efficient Tool for Estimating Copolymer Compositions from Mass Spectra. *Anal. Chem.*, 2015, 87 (10), pp 5223–5231, doi:10.1021/acs.analchem.5b00146.
Changes resulting from the publishing process, such as peer review, corrections, editing, and structural formatting, are not reflected in this document.

[*]To whom correspondence should be addressed
[†]Chair of Bioinformatics, Friedrich Schiller University Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
[‡]Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstr. 10, 07743 Jena, Germany
[¶]Jena Center for Soft Matter (JCMS), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
[§]Contributed equally to this work
[∥]Present address: Drug Discovery and Development Department, Istituto Italiano di Tecnologia (IIT), via Morego 30, 16163 Genova, Italy

**Abstract**

The accurate characterization of synthetic polymer sequences represents a major challenge in polymer science. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) is frequently used for the characterization of copolymer samples. We present the COCONUT software for estimating the composition distribution of the copolymer. Our method is based on Linear Programming and is capable of automatically resolving overlapping isotopes and isobaric ions. We demonstrate that COCONUT is well suited for analyzing complex copolymer MS spectra. COCONUT is freely available and provides a graphical user interface.

# Introduction

Mass spectrometry (MS) is increasingly used for analyzing synthetic polymers,[1] utilizing soft ionization techniques such as matrix-assisted laser desorption/ionization (MALDI),[2] electrospray ionization, or atmospheric pressure chemical ionization. MS techniques can highlight different features of polymers such as molecular weight distribution,[3] or end-groups.[4] MS is frequently used to determine compositional drift,[5] or the average composition,[6–10] which then can be verified by other techniques, such as nucleic magnetic resonance (NMR).

Quantifying the relative abundances of copolymers in a sample provides insightful information: Wilczek-Vera et al.[11] introduced the copolymer composition matrix, representing the relative abundance of all compositions of monomers. The copolymer composition matrix provides information about the copolymer architecture,[12,13] the distribution of block lengths in block copolymers,[11,14–16] or the reactivity ratio of the consumed monomers.[17] It has been used to study degradation[10] and MALDI matrix effects.[18] The composition matrix is related to the bivariate distribution of monomer ratio and degree of polymerization, which can be used to highlight compositional drift.[9,19,20]

Here, we focus on linear copolymer architectures. Several assignment methods have been introduced to estimate the copolymer composition matrix from MS data.[10–18] For these methods, the abundance of each copolymer molecule is assigned to the height of some measured peak, being closest to the most abundant theoretical isotope peak for this copolymer. However, this approach has certain drawbacks:[14,21] 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. However, for very high masses above the reported masses in this publication, peak resolution becomes poorer. For such mass regions, peak intensities should be used. Second, overlapping isotopes of different copolymers may result in imprecise polymer abundance assignments. Third, isobaric molecules may prohibit to resolve copolymer abundances.

Weidner et al.[22,23] presented a method to determine the copolymer composition matrix 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 composition matrices. Unfortunately, LACCC-MS is time-consuming, and critical conditions have to be known for at least one of the polymers. Vivó-Truyols et al.[21] presented a regression method to determine the copolymer composition from a single MS measurement. The method fits peak curves to the raw data, and can resolve overlapping isotopes. Because fitting

the complete MS spectrum is computationally expensive, the method truncates the spectrum into strips. This truncation complicates quantification of isotopes on the strip borders.

In this contribution, we propose a method to infer the copolymer composition matrix 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. 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.[21], which is the most recent for this problem. Our method is implemented in the COCONUT (Copolymer Composition Numbering Tool) software, which is provided free and open-source, and offers a graphical user interface.

# Computational methods

**Overview** 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 been discussed extensively in the literature. 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. Further, 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 $\Delta_\mathrm{m}$ 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 the 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 $\mathsf{A}$ and $\mathsf{B}$ 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, mass differences of the peaks of the theoretical isotope 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 isotope patterns of isobaric monomer compositions. But even if the monoisotopic masses of two monomer compositions is above the mass accuracy, it is possible that some isotope peaks of their theoretical isotope patterns have mass difference below the mass accuracy. We say that two isotope patterns are *overlapping*, if there exist two peaks in the patterns with mass difference below the 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: (i) Generate all candidate isotopic patterns; (ii) assign candidate peaks to the MS spectrum; (iii) compute the abundances and simultaneously resolve overlapping isotopes; and (iv) resolve isobaric molecules.

**Candidate generation** We first compute theoretical isotope 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 isotope pattern by convolving the elemental isotopic distributions.[24]

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

$$\begin{aligned} m_1 &= i \cdot m_\mathsf{A} + j \cdot m_\mathsf{B}, \\ m_2 &= (i - \Delta i) \cdot m_\mathsf{A} + (j + \Delta j) \cdot m_\mathsf{B}. \end{aligned} \tag{1}$$

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 (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

$$\frac{\Delta j \cdot m_\mathsf{B} - \Delta_\mathrm{m}}{m_\mathsf{A}} < \Delta i < \frac{\Delta j \cdot m_\mathsf{B} + \Delta_\mathrm{m}}{m_\mathsf{A}} \tag{2}$$

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 tuple $(\Delta i, \Delta j)$ an *isobaric series*.

We determine all isobaric series; then, we use the isobaric species to arrange the monomer compositions (and, hence, the corresponding isotope 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 isotope pattern for all the theoretical isotope patterns of the monomer compositions in the isobaric set; this will be our *candidate* isotope patterns. In the following, we assume that for any isobaric set, abundances for all monomer compositions but one are set to zero during fitting the matrix (Sec. ). We will split abundances of these monomer compositions in Sec. .

**Template matching** In this step, we want to assign the candidate isotope pattern peaks to the measured peaks in the experimental MS spectrum. However, measured peaks with a distance less than $\Delta_\mathrm{m}$ can lead to ambiguous assignments: These peaks may be caused by overlapping raw peaks, or errors during the centroiding (usually caused by shoulder peaks) which have been falsely identified as separate peaks. Thus, we assume centroids with a distance less than $\Delta_\mathrm{m}$ to originate from one continuous peak area, and merge them. The mass of the merged peak is the area-weighted average of the masses of its component peaks. The area of the new peak is the sum of areas of its components. Naturally, we may accidentally merge two actually separate peaks or signal with noise peaks. However, the estimation of the composition 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 isotope patterns. We match an isotope pattern peak to a measured peak if their distance is less than $\Delta_\mathrm{m}$. 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

$$S_l = \left\{ (i,j,k) \,:\, |m_l - m'_{i,j,k}| < \Delta_\mathrm{m} \right\}. \tag{3}$$

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

$$S_0 = \{ (i,j,k) \,:\, \text{there is no } l \text{ with } (i,j,k) \in S_l \}. \tag{4}$$

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

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

$$\arg\min_R \sum_l \left| \sum_{(i,j,k)\in S_l} R_{i,j} \cdot I'_{i,j,k} - I_l \right| + \sum_{(i,j,k)\in S_0} R_{i,j} \cdot I'_{i,j,k} \tag{5}$$

The first term of (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 $R_{i,j} \cdot I'_{i,j,k}$. The second term of (5) considers all candidate isotope peaks that have no matching measured peak. Since these are not represented in the spectrum and, hence, should also not exist in the model, we minimize the distance of the sum of their intensities times polymer abundance $R_{i,j} \cdot I'_{i,j,k}$ to a zero peak area.

The number of free parameters $R_{i,j}$ is determined by the number of possible template isotope patterns, which increases quadratic in mass: There exist $m+1$ compositions of two monomers for a given integer mass $m = i \cdot \mathsf{A} + j \cdot \mathsf{B}$.[25] 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 coefficient $d_l$ for each measured peak. Then, solving the Linear Program

$$\begin{aligned}
\min \quad & \sum_l d_l \\
\text{s.t.} \quad & \sum_{(i,j,k)\in S_l} R_{i,j} \cdot I'_{i,j,k} + d_l \geq I_l \quad \forall l && \text{(6a)} \\
& \sum_{(i,j,k)\in S_l} R_{i,j} \cdot I'_{i,j,k} - d_l \leq I_l \quad \forall l && \text{(6b)} \\
& \sum_{(i,j,k)\in S_0} R_{i,j} \cdot I'_{i,j,k} + d_0 \geq 0 && \text{(6c)} \\
& \sum_{(i,j,k)\in S_0} R_{i,j} \cdot I'_{i,j,k} - d_0 \leq 0 && \text{(6d)}
\end{aligned}$$

estimates the optimal abundances $R_{i,j}$. We omitted the upper and lower limit constraints for all

coefficients. Constraints (6a) and (6b) correspond to the first term of (5), and constraints (6c) and (6d) to the second term. In case there are isobaric monomer compositions with $R_{i,j} > 0$, 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, then the differences in isotope abundances of the corresponding theoretical isotopic patterns are usually too small to split the measured abundances. Therefore, we suggest an alternate approach to split corresponding entries in the composition matrix $R$. Obviously, this is not necessary if there are no isobaric monomer compositions present.

Our task is to split abundances $R_{i,j}$ that correspond to more than one monomer composition, that is, 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.[11] 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 $\lambda$ 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 $\Sigma$ of the bivariate normal distribution $F = \mathcal{N}(\mu, \Sigma)$ from the matrix $R$. In the first round, we consider only those entries of $R$ 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 $B$:

$$R_{i,j} := \frac{F(i,j)}{\sum_{(x,y)\in B} F(x,y)} \cdot r \tag{7}$$

Repeat this until $R$ converges. We found that this approach is often too slow in practice; to this end, we instead use a general purpose optimizer[26] that combines both of these steps (estimating the bivariate normal and splitting the abundances) into one. We leave out the tedious technical details.

# Experimental methods

**Overview** We evaluated our method on two different datasets. First, we synthesized three different random copolymers (Fig. 1), consisting of two macromers with both a different ratio of styrene and isoprene (Tables 1, 2). We measured the first macromers (**I1 to I3**) and the complete (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers (**P1 to P3**). Second, to assess the accuracy of our method, we evaluated it with simulated datasets, as this is the only way to compare the result to a know ground

truth. We simulated PMMA-*co*-P*n*BA and PMMA-*co*-PHEMA spectra as numerous overlapping isotopes and isobaric molecules appear in these copolymers.

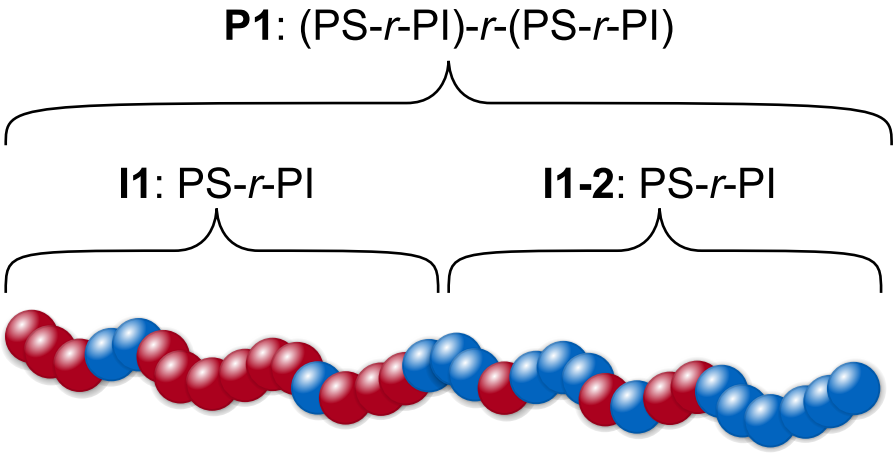


Figure 1: Schematic representation of the synthesized (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymer **P1**. **P2** and **P3** have the same architecture, but different PS to PI ratios.

**Materials and polymerization procedures** The first (**I1, I2, I3**) and second macromers (**I1-2, I2-2, I3-2**) are constituted of a random copolymer of styrene and isoprene. The copolymers (**P1, P2, P3**) were synthesized in Schlenk tubes under dry argon atmosphere. Solvents were dried over sodium/benzophenone and freshly distilled. Isoprene and styrene were dried over calcium hydride and both freshly distilled before reaction. Sec-butyllithium (1.4 mol in hexane) was used as received. All chemicals were obtained from Aldrich. The Schlenk flasks were heated and dried under vacuum and each filled with 10 mL cyclohexane and 0.09 mL (1.2 mmol) tetrahydrofuran as randomizer. To the solution 0.29 mL sec-butylithium solution (0.4 mmol) was added and allowed to stir for 15 minutes resulting in a slightly pink solution. Subsequently, each flask was heated to 40 °C and the monomer mixtures (Table 1) for the first macromer were added. After 1.5 hours stirring, the second monomer mixture (Table 2) was added for the formation of the second macromer.[27] Theoretical molar masses of 5,000 g mol$^{-1}$ (2,500 g mol$^{-1}$ for each macromer) were targeted and 2 g of final product were aimed for. 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 showed PDI values lower than 1.1, indicating a living character of the polymerization. All molar masses of **I1 to I3** were obtained in the range of 2,500 g mol$^{-1}$ and **P1 to P3** of 5,000 g mol$^{-1}$ using a polystyrene calibration. **I1**: $M_n$ = 2,310 g mol$^{-1}$, **I2**: $M_n$ = 1,960 g mol$^{-1}$, **I3**: $M_n$ = 2,153 g mol$^{-1}$, **P1**: $M_n$ = 4,546 g mol$^{-1}$, **P2**: $M_n$ = 4,058 g mol$^{-1}$, **P3**: $M_n$ = 4,380 g mol$^{-1}$.

**Instrumentation** $^1$H NMR spectra were recorded on a Bruker AC 300 MHz. Size exclusion chromatography was performed on either a Shimadzu SCL-10 A system (with a LC-10AD pump, a RID-10A refractive index detector, and a PL gel 5 $\mu$m mixed-D column at 25 °C) where the eluent

Table 1: Summary of theoretical values of each first macromer.

| | **I1** | | **I2** | | **I3** | |
|---|---|---|---|---|---|---|
| | PS | PI | PS | PI | PS | PI |
| Percent[%] | 80 | 20 | 70 | 30 | 60 | 40 |
| Molar mass [g mol$^{-1}$] | 2000 | 500 | 1750 | 750 | 1500 | 1000 |
| DP | 19 | 7 | 17 | 11 | 14 | 15 |
| Mass (monomer) [g] | 0.79 | 0.19 | 0.71 | 0.30 | 0.58 | 0.41 |
| Volume (monomer) [mL] | 0.87 | 0.28 | 0.78 | 0.44 | 0.64 | 0.60 |

Table 2: Summary of theoretical values of each second macromer.

| | **I1-2** | | **I2-2** | | **I3-2** | |
|---|---|---|---|---|---|---|
| | PS | PI | PS | PI | PS | PI |
| Percent[%] | 20 | 80 | 30 | 70 | 40 | 60 |
| Molar mass [g mol$^{-1}$] | 500 | 2000 | 750 | 1750 | 1000 | 1500 |
| DP | 5 | 29 | 7 | 26 | 10 | 22 |
| Mass (monomer) [g] | 0.21 | 0.79 | 0.29 | 0.71 | 0.42 | 0.60 |
| Volume (monomer) [mL] | 0.23 | 1.16 | 0.32 | 1.04 | 0.46 | 0.88 |

was a mixture of chloroform:triethylamine:isopropanol (94:4:2) with a flow rate of 1 mL/min. The system was calibrated with PS standards purchased from PSS Standard.

An Ultraflex III TOF/TOF (Bruker Daltonics, Bremen, Germany) was used for the MALDI-TOF-MS analysis. The instrument was equipped with a Nd:YAG laser and a collision cell. All spectra were measured in the positive reflector mode. The instrument was calibrated prior to each measurement with an external standard PMMA from PSS Polymer Standards Services GmbH (Mainz, Germany). MS data were processed using PolyTools 1.0 (Bruker Daltonics) and Data Explorer 4.0 (Applied Biosystems). Before applying our computational methods for estimating the copolymer composition, the spectra were centroided and baseline-corrected. The compositions were estimated using the COCONUT software.

**Sample preparation** For the sample preparation, all polymers (10 mg/mL) in chloroform, dithranol (50 mg/mL) in chloroform and silver trifluoroacetate (AgTFA) dissolved in chloroform at a concentration of 100 mg/mL were mixed and the dried-droplet sample preparation method was applied.

**Simulating mass spectra** To compare our results against some ground truth, we have to simulate mass spectra. Although we can not simulate all aspects of the physical processes of an MS instrument, we have tried to capture several fundamental aspects. We start by simulating a composition matrix; here, we use five bivariate normal distributions with randomly chosen parameters (Table S1). Given the composition matrix, we iterate over all monomer compositions: We add the appropriate end groups, and simulate the first 12 peaks of the isotope pattern, estimating both intensities and mean peak masses.[24] We disturb each isotope peak by adding normally distributed noise with mean zero and variance $\sigma/2$ to the masses, and multiplying intensities by log-normal distributed random noise with mean zero and variance $\sigma$, where the noise parameter $\sigma$ is given

below. For an isotope peak with mass $m$ and intensity $I$, we add a Gaussian function with mean $m$, variance $1/5$, and height (multiplier) $I$ to the simulated spectrum. We then sample this spectrum at sampling points with mass difference 0.1 Da. Finally, this sampled (discretized) spectrum is again perturbed using multiplicative noise following a log-normal distribution with mean zero and variance $\sigma/2$.

We simulated spectra for copolymers PMMA-*co*-P*n*BA and PMMA-*co*-PHEMA. Here, PMMA-*co*-P*n*BA results in a large number of overlapping isotope patterns, whereas PMMA-*co*-PHEMA results in many isobaric molecules. To demonstrate that our method can resolve overlapping isotopes and isobaric monomer compositions, we simulated noise-free spectra with $\sigma = 0$. To evaluate the robustness of our method, we additionally use four noise levels $\sigma = 0.05, 0.1, 0.2, 0.5$. For each copolymer, all five composition matrices and all five noise levels, we simulated five mass spectra; resulting in 250 spectra in total.

# Results and discussion

**Experimental (PS-*r*-PI)-*r*-(PS-*r*-PI)** The shown materials were synthesized by living anionic polymerization, which is widely used with other monomers such as ethylene oxide (EO), allyl glycidyl ether (AGE), (meth)acrylate, etc. This polymerization technique produces well-defined polymers with low polydispersity index values, which is required for MS analysis to ionize all polymer chains. The chosen copolymers have also been used as potential membranes applications when having high molar masses.[28–30]

Copolymers were synthesized with two random macromers with different ratios of styrene and isoprene (Fig. 1), analyzed by MALDI-TOF MS (Fig. S3) and the COCONUT software (Fig. S4). The estimated composition matrices (Fig. 3) were transformed to distributions of chain sizes and compositions (Fig. 4) by calculating the isoprene ratios and interpolating them for each anti-diagonal of the composition matrix. They show a compositional drift, indicating a high conversion rate, since the distribution is not symmetric with respect to the monomer fractions.[19]

Table 3: Summary of $M_n$ and $M_p$ values.

| | 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 3 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 ratios estimated from the composition matrices (Fig. 3). The maximal value in the matrix 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 composition matrices (Fig. S5). 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.

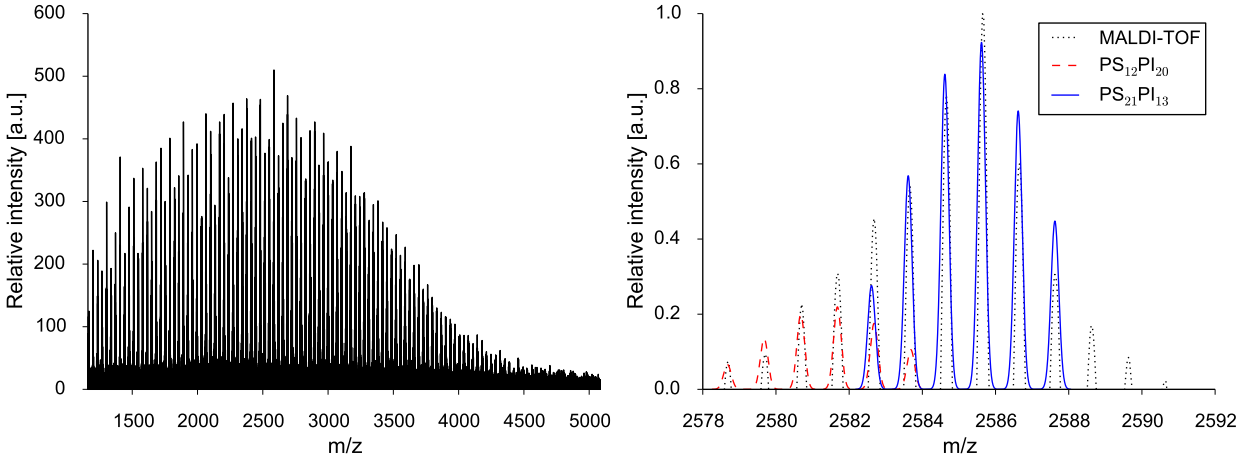


Figure 2: Left: MALDI-TOF spectrum of the (PS-*r*-PI) copolymer **I1**. Right: Detail of the spectrum overlayed with the estimated theoretical isotopes. We used six isotopic peaks per pattern to estimate the abundances.

Nevertheless, a living character of the polymerization can be assumed as well-defined polymers with a narrow molar mass distribution were obtained (Fig. S1). Isoprene as a monomer has three different microstructures (cis/trans: 1,4-:1,2-:3,4), where the 1,4 regiospecificity is mostly abundant. The different microstructures can induce slight errors in the NMR spectra (Fig. S2).[31] In addition, when THF was added to act as a randomizer, we did observe overlapping isotopes in the MS spectra and multiple isobaric distributions in the composition matrix. As shown in Fig. 2 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 3).

Huijser et al.[12], Staal[32] and Willemse[33] suggested a quick way to provide an indication of the microstructure from the slope of a line, fitted through the composition matrix. In reference to the composition matrices from **I1 to I3** (Fig. 3), 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 THF is considered as randomizer. Nonetheless the **P1 to P3** do correlate to block like structures as was desired.

**Simulated PMMA-*co*-P*n*BA/PMMA-*co*-PHEMA** 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

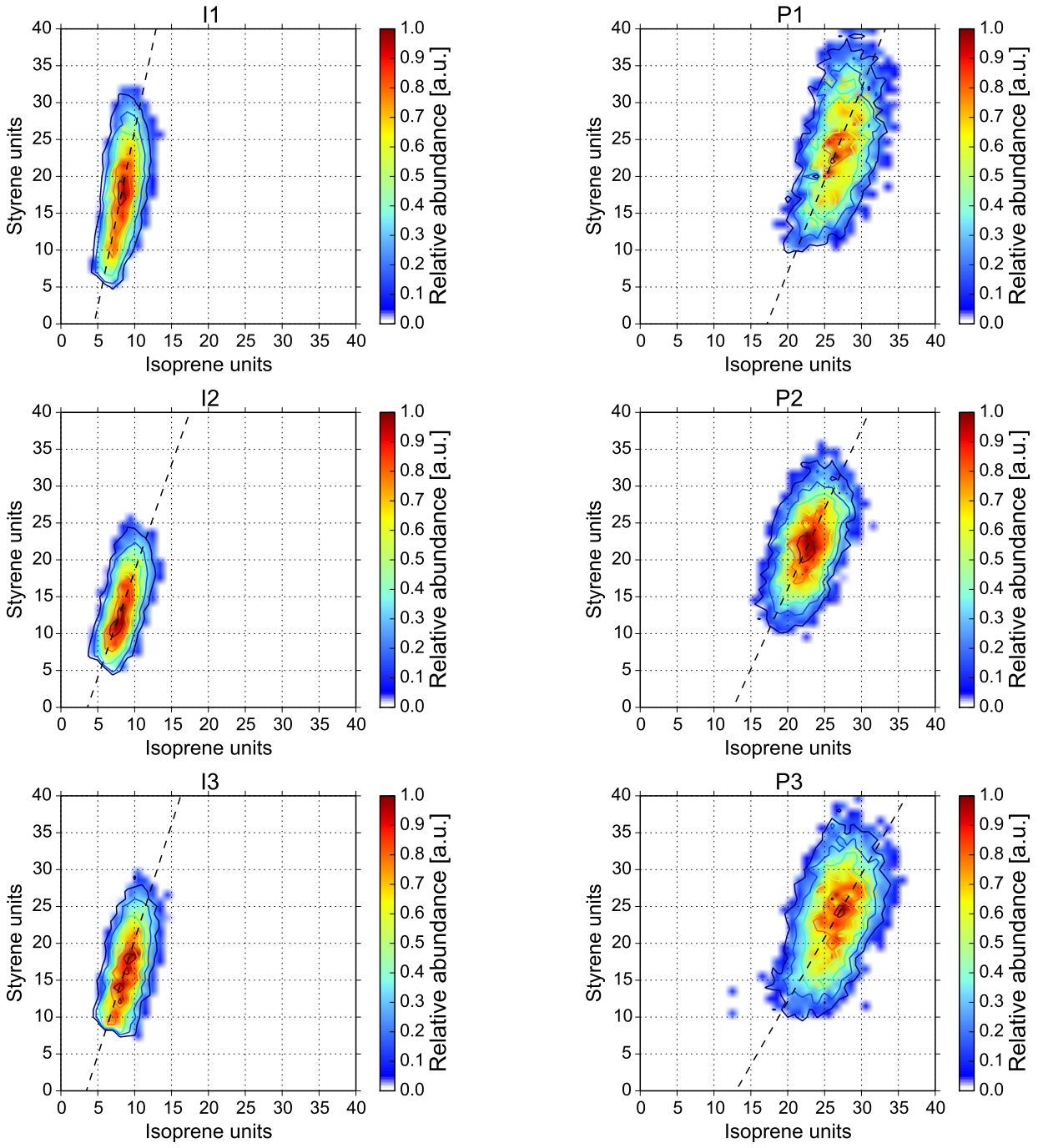


Figure 3: Copolymer composition matrix of the (PS-*r*-PI) macromers **I1 to I3** (left) and the final (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers **P1 to P3** (right).

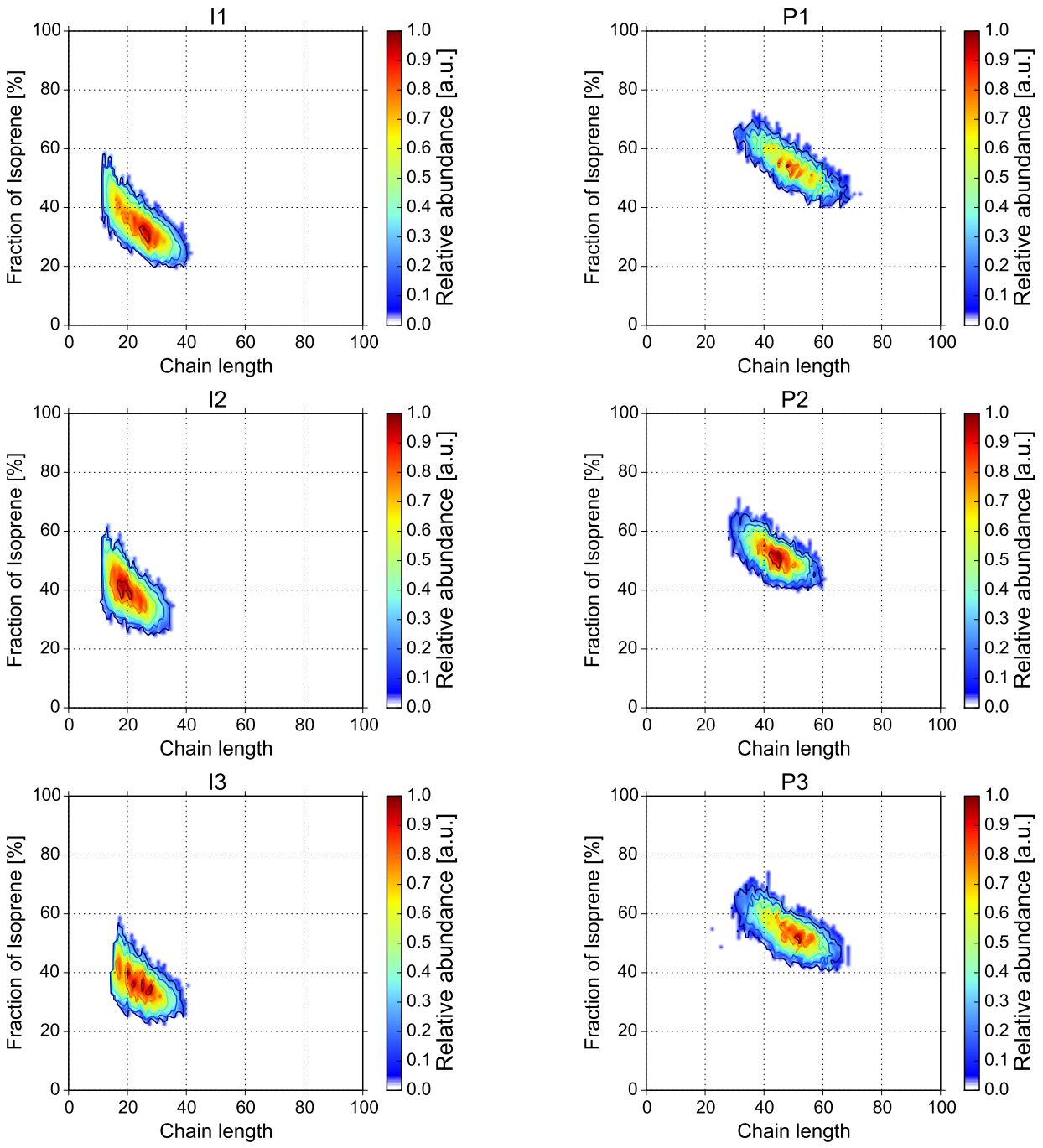


Figure 4: Copolymer composition as a function of degree of polymerization and the ratio of isoprene of the (PS-*r*-PI) macromers **I1 to I3** (left) and the final (PS-*r*-PI)-*r*-(PS-*r*-PI) copolymers **P1 to P3** (right).

(Fig. 5). The distribution was almost perfectly reconstructed, only isotopes below the intensity threshold were not considered by our method and, thus, lost (Fig. S6). In the simulated spectrum of PMMA-*co*-PHEMA (Fig. S7), there exist three neighboring isobaric distributions that may explain the data; from these, COCONUT chose the correct distribution located in the center of the composition matrix (Fig. 5). Both simulations indicate that our method can reconstruct the true copolymer distribution, given that the input spectrum is free of noise.

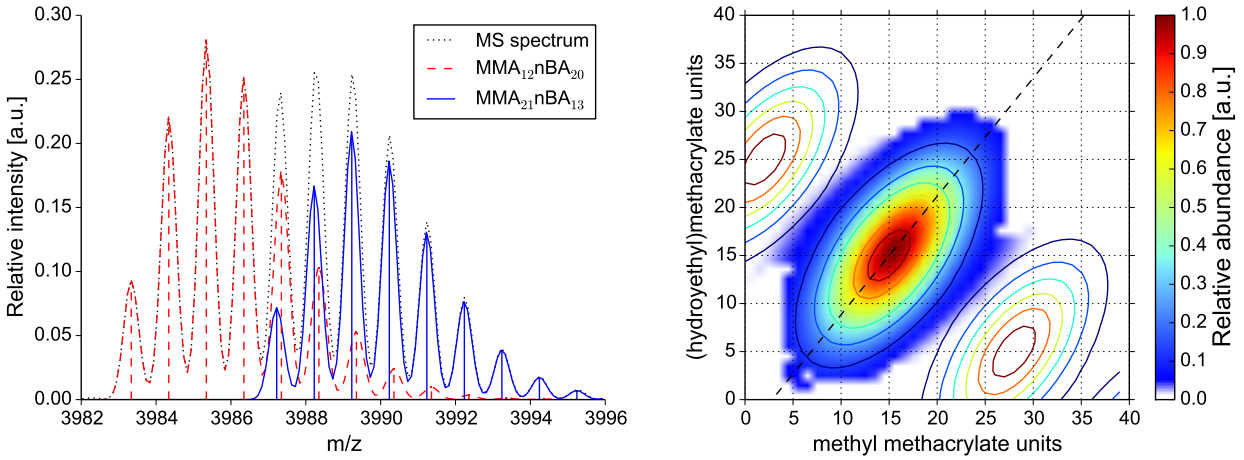


Figure 5: Left: Detail of the simulated MS spectrum of PMMA-*co*-P*n*BA showing overlapping isotopes. The relative molecule abundances estimated by COCONUT are represented by the centroid intensities. Right: Copolymer composition matrix estimated from a simulated MS spectrum of a PMMA-*co*-PHEMA copolymer overlayed with all isobaric distributions (contours).

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[21] 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 composition matrices against the original composition matrix (Fig. 6). For each method, 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 composition matrix 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 composition matrix. 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 composition matrices 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 sec-

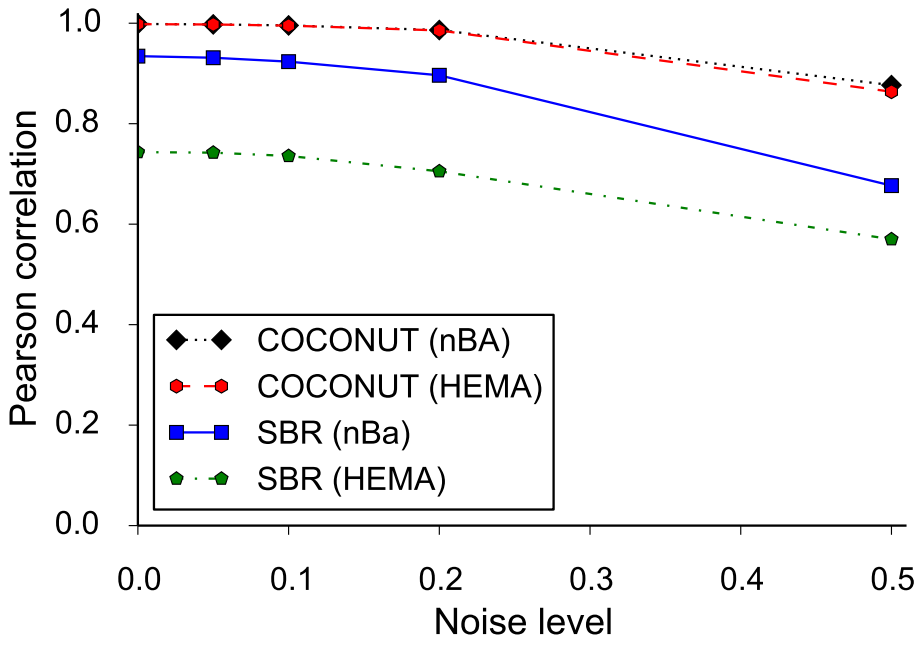


Figure 6: Median Pearson correlation coefficient for each method and copolymer dataset, PMMA-*co*-P*n*BA and PMMA-*co*-PHEMA, at five different noise levels.

onds 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.

**Software** Our software called COCONUT (Copolymer Composition Numbering Tool) was implemented in the Groovy language and runs on the Java platform. It is freely available for download at `http://bio.informatik.uni-jena.de/software/coconut`. The core is formed by efficient algorithms for calculating the isotope patterns, estimating the copolymer composition and resolving isobaric species. It is distributed with the free open source LP solver lp_solve (`http://sourceforge.net/projects/lpsolve/`). Our software also supports the efficient commercial Gurobi LP solver (Gurobi Optimization, Inc., Houston, USA). 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.[34]

COCONUT combines these algorithms with a user-friendly interface (Fig. S8). At the starting point of an analysis, the user can choose to import either a previously centroided or a raw MS spectrum. If necessary, noise in the raw signal peaks can be reduced by smoothing them with a Savitzky-Golay filter.[35] Baseline bias and noise peaks are filtered by a baseline correction and setting an intensity threshold. The raw spectrum is then centroided by estimating the area under the curve of the detected peaks. To calculate the copolymer composition, the molecular formulas of the monomers and initiating as well as terminating end-groups plus cationization agent are required. If there are isobaric species, the program resolves them automatically.

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 the copolymer compositions.

Graphics can be exported as bitmaps.

# Conclusions

Mass spectrometry has become an indispensable tool for analyzing copolymers. Copolymer spectra are highly complex and contain numerous peaks. Often occurring challenges include isobaric species, overlapping isotopes, background noise and peak shape perturbations. Computational methods have proven to be a remarkable efficient tool to counteract these recurring troublesome points. We have presented a robust algorithm to estimate composition matrices of linear copolymers from any type of MS spectra. A remaining open challenge in quantifying copolymers from single mass spectra is mass and composition-dependent ionization.

In this contribution we have demonstrated the power of our tool COCONUT using several synthesized copolymers. In addition, we have evaluated our software on simulated datasets, as this is the only way to compare the result to a known ground truth. We demonstrated COCONUT is swift and accurate for the simulated spectra. We argue that COCONUT is well suited for complex copolymer spectra, as we strove to incorporate their characteristic features in the simulated spectra.

COCONUT is freely available for polymer scientists to investigate composition and linear architectures for designing smart polymers. Our software fulfills chemists demand for computational support in an efficient manner.

## Supporting Information Available

Additional information as noted in text. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

## Acknowledgement

Funding by the Thüringer Ministerium für Bildung, Wissenschaft und Kultur (grants no. B515-07008, 12038-514) and the Ernst-Abbe Stiftung. We thank Bruker Daltonics for help and support, Dr. Gabriel Vivó-Truyols for providing the SBR software, and Alexander Meier for his help with the polymerization experiments.

## References

(1) Montaudo, M. S. *Mass Spectrom Rev* **2002**, *21*, 108–144.

(2) Pasch, H., Schrepp, W., Eds. *MALDI-TOF Mass Spectrometry of Synthetic Polymers*; Springer Berlin Heidelberg, 2003.

(3) Guttman, C. M.; Blair, W. R.; Danis, P. O. *Journal of Polymer Science Part B: Polymer Physics* **1997**, *35*, 2409–2419.

(4) Crecelius, A. C.; Schubert, U. S. *Mass Spectrometry in Polymer Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2011; pp 281–318.

(5) Montaudo, M. S. *J Am Soc Mass Spectrom* **2004**, *15*, 374–384.

(6) Montaudo, M. S.; Montaudo, G. *Makromolekulare Chemie. Macromolecular Symposia* **1993**, *65*, 269–278.

(7) Montaudo, M. S.; Puglisi, C.; Samperi, F.; Montaudo, G. *Macromolecules* **1998**, *31*, 8666–8676.

(8) Alhazmi, A. M.; Mayer, P. M. *Rapid Commun Mass Spectrom* **2007**, *21*, 3392–3394.

(9) Zagar, E.; Krzan, A.; Adamus, G.; Kowalczuk, M. *Biomacromolecules* **2006**, *7*, 2210–2216.

(10) Kasperczyk, J.; Li, S.; Jaworska, J.; Dobrzynski, P.; Vert, M. *Polymer Degradation and Stability* **2008**, *93*, 990 – 999.

(11) Wilczek-Vera, G.; Danis, P. O.; Eisenberg, A. *Macromolecules* **1996**, *29*, 4036–4044.

(12) Huijser, S.; Staal, B. B. P.; Huang, J.; Duchateau, R.; Koning, C. E. *Biomacromolecules* **2006**, *7*, 2465–2469.

(13) Huijser, S.; Staal, B. B. P.; Huang, J.; Duchateau, R.; Koning, C. E. *Angew. Chem. Int. Ed.* **2006**, *45*, 1521–3773.

(14) Wilczek-Vera, G.; Yu, Y.; Waddell, K.; Danis, P. O.; Eisenberg, A. *Rapid Commun. Mass Spectrom.* **1999**, *13*, 764–777.

(15) Wilczek-Vera, G.; Yu, Y.; Waddell, K.; Danis, P. O.; Eisenberg, A. *Macromolecules* **1999**, *32*, 2180–2187.

(16) Willemse, R. X. E.; Staal, B. B. P.; Donkers, E. H. D.; van Herk, A. M. *Macromolecules* **2004**, *37*, 5717–5723.

(17) Huijser, S.; Mooiweer, G. D.; van der Hofstad, R.; Staal, B. B. P.; Feenstra, J.; van Herk, A. M.; Koning, C. E.; Duchateau, R. *Macromolecules* **2012**, *45*, 4500–4510.

(18) Terrier, P.; Buchmann, W.; Cheguillaume, G.; Desmazières, B.; Tortajada, J. *Anal Chem* **2005**, *77*, 3292–3300.

(19) Montaudo, M. S.; Montaudo, G. *Macromolecules* **1999**, *32*, 7015–7022.

(20) Montaudo, M. S. *Macromolecules* **2001**, *34*, 2792–2797.

(21) Vivó-Truyols, G.; Staal, B.; Schoenmakers, P. J. *J. Chromatogr. A* **2010**, *1217*, 4150–4159.

(22) Weidner, S.; Falkenhagen, J.; Krueger, R.-P.; Just, U. *Anal. Chem.* **2007**, *79*, 4814–4819.

(23) Weidner, S. M.; Falkenhagen, J.; Maltsev, S.; Sauerland, V.; Rinken, M. *Rapid Commun. Mass Spectrom.* **2007**, *21*, 2750–2758.

(24) Böcker, S.; Letzel, M.; Lipták, Zs.; Pervukhin, A. *Bioinformatics* **2009**, *25*, 218–224.

(25) Böcker, S.; Lipták, Zs. *Algorithmica* **2007**, *48*, 413–432.

(26) Powell, M. *The BOBYQA algorithm for bound constrained optimization without derivatives*; 2009.

(27) Nakahara, A.; Satoh, K.; Kamigaito, M. *Polym. Chem.* **2012**, *3*, 190–197.

(28) Schacher, F.; Ulbricht, M.; Müller, A. H. E. *Adv. Funct. Mater.* **2009**, *19*, 1040–1045.

(29) Phillip, W. A.; Dorin, R. M.; Werner, J.; Hoek, E. M. V.; Wiesner, U.; Elimelech, M. *Nano Lett.* **2011**, *11*, 2892–2900.

(30) Schacher, F.; Rudolph, T.; Wieberger, F.; Ulbricht, M.; Müller, A. H. E. *ACS Appl Mater Interfaces* **2009**, *1*, 1492–1503.

(31) Worsfold, D. J.; Bywater, S. *Can. J. Chem.* **1964**, *42*, 2884–2892.

(32) Staal, B. Characterization of (co)polymers by MALDI-TOF-MS. Ph.D. thesis, University of Technology Eindhoven, 2005.

(33) Willemse, R. X. E. New Insights into Free-Radical (Co)Polymerization Kinetics. Ph.D. thesis, University of Technology Eindhoven, 2005.

(34) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics* **2010**, *11*, 395.

(35) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627–1639.

# Graphical TOC Entry