The properties/functions of biopolymers are encoded in their well-defined sequences. In contrast, in synthetic copolymers, such sequence–property correlations are obscured, since the properties are averaged over the polydisperse sequences.\(^1\) Nevertheless, outstanding functions originating from sequence-specific short segments stochastically generated via random copolymerization have been recently reported.\(^2\)–\(^4\) These discoveries imply that the properties of random copolymers are deemed to be encoded into their short sequences, herein called “codons”. From theoretical approaches, the impact of codon structures on the copolymer properties has been well studied using machine-learning and molecular dynamics simulations, where aperiodic and complex codons were designed as optimal repeating segments to achieve the target polymer properties.\(^5\)–\(^6\) However, owing to the lack of efficient methods for sequence distribution analysis, \(i.e.,\) sequencing\(^6\) (also see the ESI “Definition of sequencing”\( ^{\dagger}\)), a significant gap remains between theories and experiments: first, embedding such designed codons into the main chains is very challenging even with state-of-the-art sequence-controlled polymerization techniques;\(^1\) second, the codon–property correlations cannot be experimentally investigated, because conventional copolymer characterization using the monomer reactivity ratio can estimate the codon compositions only at the ground-state sequence distribution,\(^9\) but not at controlled sequence distribution. This means that the impact of the controlled sequence on material performances cannot be quantitatively evaluated although various sequence-controlled polymerization methods have been developed thanks to the continuous effort from the polymer synthesis community. The well-known sequencing method utilizes nuclear magnetic resonance spectroscopy (NMR); depending on the adjacent monomer-unit species, the electronic environment surrounding the central monomer unit varies, inducing peak shifts. However, such shifts are often subtle and coupled with tauticity in most monomer combinations, allowing only qualitative analysis, \(e.g.,\) see our previous attempts;\(^7\)\(^8\) therefore, NMR sequencing can only estimate the codon compositions of binary triads in very limited monomer combinations, which is insufficient and ineffective.\(^7\) As the complexity of codons increases with the length and monomer component number according to the system of interest, a data-driven approach accommodating this increasing complexity by expanding the dataset is attractive. Notably, if the codon composition of growing copolymers can be estimated in real time during the polymerization, the polymerization conditions could be autonomously tuned \(v i a\) reinforced learning\(^9\) so that the fraction of the theoretically designed codons would be maximized.

Toward practical sequence-engineering, herein, we propose a polymer sequencer quickly quantifying complex codon compositions in diverse multi-monomer systems based on ambient gas-phase mass spectrometry (MS)\(^{12}\) of pyrolyzed oligomeric fragments (Fig. 1A). The observed fragment pattern potentially reflects the codon compositions; however, direct
peak-wise characterization is ineffective, because the peak intensities are strongly biased owing to different ionization efficiencies and regioselective thermal cleavages of chemically inequivalent main chains. Our key strategy towards codon quantification is to interpret the observed fragment pattern as that generated from a certain mixture of K-basis sequence-defined copolymers repeating single-codon species (Fig. 1A, here K = 5 for quantifying the simplest binary triad codons). The actual copolymer ensemble is an intra-chain mixture of codons; yet, via fragmentation, they can be likened to an inter-chain mixture whose fraction (c1, ..., cK) represents the codon composition in the analyte (here, K = 5, an alternating-like copolymer ensemble is depicted). The basis spectra were not measured in reality but inferred via the RQMS algorithm from the spectral dataset of random copolymers (here, X: butyl acrylate; Y: styrene; the dataset size N = 21, and the sample information and spectral data are summarized in data S1). (B) The outline of RQMS sequencing.
spectra; binary triad, binary pentad and ternary triad codons were quantifiable with five, nine, and 13 basis spectra, respectively, inferred from 30, 80 and 84 samples of random copolymers (see Fig. S1† for the basis number selection). Therefore, the larger dataset is fed, the more complex codon information RQMS would output, which goes beyond the analytical limitation of the conventional NMR sequencing as later shown.

In the remaining part of this paper, we first theoretically construct an RQMS algorithm. We then verify the RQMS accuracy with a benchmark compositional analysis of ternary homopolymer films. This is because the verification of RQMS sequencing result itself is very challenging owing to the lack of alternative sequencing methods. Via the benchmark test, we demonstrate that RQMS can conduct accurate compositional analysis without using the pure constituent spectra nor any prior knowledge about the constituents. We then apply RQMS for the sequencing purpose. The inferred basis spectra of sequence-defined copolymers look like real measured spectra with the corresponding codon peaks at consistent peak positions. Finally, we compare the RQMS sequencing result to the NMR and theoretically predicted sequence for verifying the accuracy of RQMS sequencing.

Development of RQMS

In this section, a spectrum is represented by a D-dimensional row vector, $\mathbf{R}^{1 \times D}$ (D: channel number), storing the non-negative signal intensities at D-channels. The observed spectral set of the N-mixed samples is then represented by $\mathbf{X} \in \mathbb{R}_+^{N \times D}$, where N-spectra are vertically stacked. The spectrum of the $n$th sample is located at the $n$th row, and represented as $X_n$. The unmeasurable $K$-basis spectral set is then represented by $\mathbf{P} \in \mathbb{R}_+^{K \times D}$. Our mission is to infer the bases $\mathbf{P}$ from the observed spectral set $\mathbf{X}$ under the assumption that every observed spectrum $X_n$ can be approximated by linear combination of $K$-basis spectra:

$$X_n \approx \sum_{k=1}^{K} C_{nk} P_k, \text{ s.t. } \sum_{k=1}^{K} C_{nk} = 1, (n = 1, \ldots, N)$$

where $C \in \mathbb{R}_+^{N \times K}$ represents the basis fractions, whose $(n, k)$-element represents the fraction of the $k$th basis in the $n$th sample, satisfying the sum-to-one condition in every sample. This can be simply written as $\mathbf{X} \approx \mathbf{CP}$ in the matrix form, which is called non-negative matrix factorization (NMF). This non-negative constraint is requested because both fractionation and spectral intensity cannot be negative. In this context, NMF is a mathematical expression of compositional analysis that simultaneously identifies bases $\mathbf{P}$ and their quantities $C$. If the dataset does not contain all $K$-basis spectra $\mathbf{P}$, it corresponds to RQMS. The most significant difference between conventional quantitative MS and RQMS is that RQMS does not require the actual measurement of all basis spectra $\mathbf{P}$. To verify RQMS accuracy, we designed a ternary film system composed of poly(ethyl methacrylate) (E), poly(methyl methacrylate) (M), and polystyrene (S). The dataset (data S2) consisted of 24 binary and ternary mixtures, whose highest fractions did not exceed 80 wt% (therefore, reference-free). The accuracy of the RQMS algorithm can be evaluated by checking if the output $\mathbf{C}$ and $\mathbf{P}$ are well consistent with the true composition and the observed pure E/M/S spectra. This was actually achieved, as shown in Fig. 2A. Even a biased dataset without samples beyond 40 wt% $S$-fraction gave an accurate estimation (Fig. S2†). However, direct and single-step NMF, $\mathbf{X} \approx \mathbf{CP}$, yielded inaccurate results (Fig. S3†). This is because pyrolysis-MS does not measure polymers themselves, but measure their pyrolyzed fragments; therefore, the spectrally distinct and independent components are not $K$-polymers but their pyrolyzed $M$-fragments ($K \ll M$). This problem is particularly acute when the basis polymers contain common sub-structures such as the polymer backbone, which can often occur in practical polymer science. In our benchmark test, E and M have identical methacrylic backbones, which would generate the same fragments (the black spectra in Fig. 2B). To conduct fragment-based NMF, we assume a latent hierarchical structure (Fig. 2B), where the $k$th basis polymer generates the $m$th fragment spectrum $S_{mk}$ with a fragment abundance (FA) of $B_{km}$ ($k = 1, \ldots, K; m = 1, \ldots, M$), where $S \in \mathbb{R}_+^{M \times D}$ represents the $M$-fragment spectra, and $B \in \mathbb{R}_+^{K \times M}$ represents the FA of the basis polymers. The observed spectrum of the $n$th sample then can be written as a linear combination of $S$:

$$X_n \approx \sum_{m=1}^{M} \left( \sum_{k=1}^{K} C_{nk} B_{km} \right) S_{mk} = A_n S,$$

where $A = CB \in \mathbb{R}_+^{N \times M}$ represents the sample-wise FA. Our developed RQMS algorithm first conducts fragment-based NMF, $\mathbf{X} \approx \mathbf{AS}$, and subsequently second NMF, $A \approx CB$, summarized as $\mathbf{X} \approx \mathbf{AS} \approx \mathbf{CBS} \approx \mathbf{CP}$, outputting the sample-wise fraction $C$ and basis-polymer spectra $\mathbf{P}$. The actual implementation is slightly more complicated, as depicted in Fig. S4 and S5.†

Because both NMFs are low-rank approximations, the fundamental driving force toward the optimum solution is identical: e.g., $A$ and $S$ are determined so that the squared residuals of $\mathbf{X}$ and $\mathbf{AS}$ are minimized while adhering to the non-negative constraints, i.e., $\min_{A \succeq 0} \| \mathbf{X} - \mathbf{AS} \|_F^2$. A solution $(A^*, S^*)$ derived from this criterion is, however, not unique even for a given component number $M$, as any non-singular $Q \in \mathbb{R}_+^{M \times M}$ satisfying $A^* Q \succeq 0$, $Q^{-1} S^* \succeq 0$ gives another solution $(A^* Q, Q^{-1} S^*)$. Imposing additional constraints is thus necessary to approach a better solution. Soft orthogonal constraints14,15,18 on the fragment spectra $S$ are particularly effective for narrowing down the solution candidates, accounting for low yet non-zero possibilities that different fragments occupy common channels (full formulation in the ESI Computational Methods section†). NMF is called “unique” when $Q$ is limited to the identity or permutation matrices.16 Importantly, the first NMF, $\mathbf{X} \approx \mathbf{AS}$, is not unique because this is a spectral interpretation with no correct answer, whereas the second NMF, $A \approx CB$, should be unique to determine the polymer fraction $C$ with a single correct answer.19 The second NMF uniqueness can be ensured by minimizing the volume of the simplex, which is spanned by row-vectors of $B$ and contains all the datapoints, i.e.,
The connection between the two non-unique and unique NMFs is key to the RQMS algorithm. For robustly outputting \((C, B)\) from any “good enough” \(A\), which is a non-unique solution of the first NMF, the factorization residual of the second NMF is evaluated using Riemannian metrics,\(^{22,23}\) considering the non-orthogonality of \(S\) (Fig. S6†). The component number \(M\) for the first NMF is unknown and thus statistically determined via automatic relevance determination (ARD),\(^{14,24}\) whereas \(K\) for the second NMF should be appropriately given depending on the analytical purpose (Fig. S1†). Two formulations were thus separately derived, as presented in the ESI Computational Methods section.†

Basis inference for binary triad sequencing

RQMS sequencing consists of two steps: inferring \(K\)-basis spectra \(P\) of sequence-defined copolymers composed of the single-codon species and deconvoluting the targeted spectrum \(X\) into \(P\) (Fig. 1B). As forementioned, the simplest binary and triad sequencing requires five-basis spectra of \((XXX), (XXY), (XY), (YXX),\) and \((YYY)\). As two of the five bases are synthesizable and, thus, referenceable homopolymers, the inference of the other bases should be reliable even if the dataset is biased, as demonstrated in the benchmark test (see Fig. S2†). For sequencing, a “biased” dataset indicates that at least one of the \(K\)-basis codons does not occupy a sufficiently high fraction in any samples.

Since the accuracy of the RQMS algorithm has been ascertained in the benchmark test, the reliability of RQMS sequencing now depends on the validity of the linear combination model assuming that any copolymers can be approximated by linearly mixed sequence-defined copolymers (Fig. 1A). To verify this assumption, we began with M/S triad sequencing because the alternating copolymer \((MS)\) is exceptionally synthesizable (also see Fig. S7†), which allowed us to compare the inferred and measured \((MS)\) spectra. A dataset of 30 M/S random copolymers quickly prepared via free-radical copolymerization (data S3) was subjected to RQMS, outputting five-basis spectra (Fig. 3). They were attributable to \((MMM), (MMS), (MS), (SSM),\) and \((SSS)\), based on the proton-adducted triad peaks at 301, 305, 305/309, 309, and 313 \(m/z\) reflecting the molecular weight difference between \(M (100)\) and \(S (104)\). The peak distributions are rational; e.g., \((MMS)\) generates only MSM and MMS triads at 305 \(m/z\) with the absence of the other triad peaks at 301, 309 and 313 \(m/z\). Furthermore, the inferred and observed \((MS)\) spectra showed good consistency. (MS) theoretically consists of an equivalent number of MSM and SMS.
triads; nevertheless, regioselective thermal cleavages and individually different ionization efficiencies biased the peak intensities (SMS $\gg$ MSM). Such coefficients were inexplicitly learned and embedded into the basis spectrum by the RQMS algorithm, allowing the precise basis inference. Note that the observed (MS) spectrum was not included in the dataset and never used for the basis inference. This elucidated that the spectra of sequence-defined copolymers were accurately inferable only from a random copolymer dataset. No explicit instructions specifically oriented to sequencing were implemented in the RQMS algorithm, justifying the assumed linear combination model for reducing sequencing to compositional analysis (Fig. 1A).

**Basis inference for ternary triad sequencing**

For triad sequencing, $J$-multi ($J \geq 3$) monomer systems require the basis number $K = J C_3 + 3 J C_2 + \frac{1}{2} J C_1$. The three terms correspond to ternary, binary, and unary triad codons. We here demonstrate M/S/poly(butyl acrylate) (B) ternary triad sequencing, based on a dataset of 4 terpolymers and 80 binary copolymers (data S4). The output 13 basis spectra, including ternary-alternating (MSB) basis, were rational spectra with consistent peak positions (Fig. S8†). In contrast to NMR sequencing, which is limited to $J = 2$, RQMS sequencing has no limitations for $J$ thanks to the data-driven nature.

**Binary pentad sequencing**

S/B binary pentad sequencing necessitated a more careful dataset design based on the monomer-reactivity ratio (Fig. S9†). Based on the 80 S/B binary copolymers (data S5), RQMS output nine basis spectra of pentad codons with reasonable peaks (Fig. 4A). The output tetrad/triad distributions were also appropriate; e.g., (BBBSS) and (BBSBS) spectra showed a pentad peak at identical positions (593 m/z for the proton adduct and 607 m/z for the ammonium adduct), but showed triad peaks at different positions owing to the presence or absence of the BBB triad (385 m/z). As another example, (SB)$_2$ had two pentads of BSBSB (607 m/z) and SBSBS (569 m/z), a single tetrad of BSBS (479 m/z) and two triads of BSB (361 m/z) and SBS (337 m/z).

After the basis inference, spectral deconvolution is conducted by sequentially projecting the target spectrum $X$ onto the inferred basis spectra (see the ESI “Sequential projection of a target pyrolysis-MS spectrum”† for the detailed procedure). The basis spectra are peculiar to the monomer combinations and invariant to polymerization conditions such as catalytic systems. Therefore, a different polymerization technique can be employed for the target copolymer synthesis from the one used for dataset preparation; e.g., the dataset and target copolymers, respectively, can be synthesized by simple free-radical copolymerization and living copolymerization. Furthermore, impurity signals on the target spectrum would be automatically filtered when being projected onto the basis spectra, allowing semi-real-time direct sequencing from the polymerization solution without any chemical purification. By coupling with reversible addition–fragmentation chain-transfer (RAFT) polymerization using 2-(dodecylthiocarbonothioylthio)-2-methylpropionic acid (DDMAT) as a RAFT agent, varying codon distributions along the main chain were monitored (Fig. 4B and C). We selected S/B copolymers as targets, since their B-centered triad fractions can be obtained via NMR as well by decoupling their tacticity."
allowing result verification (Fig. S10†). Reflecting the lower monomer reactivity of B as compared to S, the fractions of B-rich codons such as BBBBB and BBBS increased as the copolymerization progressed (Fig. 4B). To verify this pentad sequencing result, the RQMS pentad fractions were downgraded to B-centered triad fractions (Fig. 4C, also see the ESI “Downgrading the S/B pentad-fraction”†), so that the RQMS sequencing result can be compared with NMR observations and theoretical predictions from the Alfrey–Mayo equation based on the monomer-reactivity ratio \( r_\text{S} / r_\text{B} = (0.70, 0.17)^* \) [see the ESI “Prediction of sequence distribution from the monomer reactivity ratio”†]. The modulation of the B-centered triad fraction throughout the polymerization was consistent between the RQMS and theoretical prediction (Fig. 4C, the gap was within 5%), suggesting the validity of RQMS sequencing. On the other hand, a significant gap between RQMS and NMR was observed for BBS and BBB codon fractions. We cannot explicitly conclude which method was more accurate; however, NMR sequencing seems to have failed to capture the increasing trend of the BBB codon at the final polymerization stage, which was predicted by theory reflecting the lower monomer reactivity of B as compared to S. This could be attributable to the large peak fitting error in the NMR spectrum as shown in Fig. S10,† which may overestimate the BBS fraction and underestimate the BBB fraction.

**Conclusions**

In this paper, we proposed a sequencer determining the codon compositions of synthetic copolymers where polymer properties/functions could be encoded. Despite the lack of pure constituent codon spectra, unsupervised learning of random copolymer spectra allowed identification and quantification of
the codons in the analyte. Unlike the conventional NMR sequencing, our method seems to have no restrictions on the applicable monomer combinations; nevertheless, sequencing of depolymerization-susceptible copolymers would be challenging. This is because a perfect depolymerization into single-monomer units would leave no clues of the original sequence. Such depolymerization would rarely occur in reality, since pyrolysis is conducted under ambient air. However, we already found that sequencing of methacrylic copolymer with a very bulky side-chain, such as adamantane groups, was challenging owing to the depolymerization issue even if pyrolysis was conducted under air. To aid such a problematic monomer chemistry, we are further developing an analytical system and algorithms. In this paper, we demonstrated that in versatile vinyl monomer systems the simplest codon compositions of binary triads were accurately determined with very small datasets ($N \sim 30$). Thanks to the data-driven feature, the compositions of higher complex codons, e.g., binary pentads and ternary triads, were also accessible with the expanded datasets ($N \sim 80$), which was beyond the analytical limit of conventional NMR sequencing. More complex codons would be accessible by further expanding the dataset. For stress-free preparation of a large dataset, feeding continuously changed monomer composition into a flow polymerization reactor seems very promising. Since random copolymers now become describable with their codon compositions, codon–property correlations can be quantitatively analyzed in the framework of material informatics. Furthermore, our semi-real-time sequencing directly applicable to growing copolymers in a polymerization solution would allow practical sequence-controlled polymerization in tandem with autonomous self-optimizing reactors. Overall, the proposed sequencer would facilitate sequence engineering towards innovative polymer materials.

Data availability

All spectral datasets used in this study are available at https://doi.org/10.26434/chemrxiv-2022-mw76d-v2.

Author contributions

Y. H. conceived the research, prepared the samples, developed the software, analyzed the data, and wrote the manuscript. S. U. conducted the pyrolysis-MS and NMR measurements. M. N. supervised the research.

Conflicts of interest

Y. H. and M. N. are the owners of patent applications on RQMS and RQMS sequencing.

Acknowledgements

The Core Research for Evolutional Science and Technology program of Japan Science and Technology Agency under Grant JPMJCR19J3 (to M. N.) is acknowledged.

References