Reference-free Quantitative Mass Spectrometry in the Presence of Nonlinear Distortion Caused by In-Situ Chemical Reactions among Constituents

Yusuke Hibi*

Data-driven Polymer Design Group, Research Center for Macromolecules and Biomaterials, National Institute for Materials Science (NIMS); 1-2-1, Sengen, Tsukuba, Ibaraki 305-0047, Japan.

*Corresponding authors. Emails: hibi.yusuke@nims.go.jp

Abstract

Materials performance is primarily influenced by chemical composition, making compositional analysis (CA) essential in materials science. Traditional quantitative mass spectrometry, which deconvolutes analyte spectra into reference spectra, struggles with reactive systems where spectral variations occur, such as peak shifts and new peak emergences. Additionally, obtaining reference spectra for all pure constituents is often impractical. To address these challenges, I propose nonlinear reference-free quantitative mass spectrometry (NL-RQMS). This method simultaneously determines composition, reference spectra, and nonlinear interaction effects directly from a spectral dataset of mixtures, eliminating the need for prior reference spectra. In a benchmark test on ternary reactive polymers of epoxy and amines, NL-RQMS inferred compositions with an error margin of just 3 wt%, significantly outperforming the 6 wt% error margin of linear RQMS. The inferred interaction terms clearly indicate in-situ reactions between epoxy and amine moieties. This framework enables accurate compositional analysis without prior knowledge of the constituents, even in systems with interactive components, and holds significant potential for applications such as grading recycled plastics, where pristine materials, degradation compounds, and stabilizers interact complexly, causing nonlinear spectral distortions.

Introduction

The comprehensive characterization of most chemical processes is achieved by monitoring the temporal changes in the concentrations of the system's constituents. In this context, compositional analysis (CA) is the most fundamental and decisive analytical technique for understanding chemical phenomena. However, as most spectroscopies except for nuclear magnetic resonance (NMR) are non-quantitative—where the intensity ratio of two peaks does not directly indicate the abundance ratio of the corresponding substances—CA typically involves two processes: preparing the reference spectra of all constituents and calculating the contribution of the reference spectra in an analyte spectrum.[1] The complexity of CA is primarily influenced by two factors: (1) the accessibility of reference spectra[2] and (2) the negligibility of interactions among the constituents.[3] In situations where the reference spectra of all components are accessible and the interactions among these components are spectrally negligible, CA is simply reduced to spectral deconvolution.[1] This means that an analyte spectrum can be expressed as a linear combination of the reference spectra, where the scalar coefficients directly reflect the proportions of each component. However, challenges arise when the constituents are unknown or non-isolable, and thus the reference spectra are unpreparable via direct measurements. This often happens in polymer sciences by their very nature; polymers constitute polydisperse ensembles, rendering pure reference compounds largely theoretical or nonexistent.[4] In such cases that fail to meet criterion (1), the reference spectra should be prepared via virtual inferences based on the spectral dataset of mixtures. Inferring reference spectra is an inverse problem of vector composition, which is an ill-posed problem whose solution cannot be uniquely determined.[2] Nevertheless, recent

advancements in unsupervised machine-learning—particularly contributed by spectral imaging community—have enabled accurate inference of reference spectra from mixed spectra via incorporating mathematical expressions to capture the inherent features of spectroscopy, such as non-negativity,[5] sparsity[6] and volume minimization of the simplex spanned by the reference spectra.[7–9] The reference inference allows reference-free CA, which is emerging as a key technology characterizing a polymer ensemble, e.g., monomer sequence distribution[4] and heat conductivity.[10] However, when criterion (2) is not met, the complexity of CA spikes dramatically due to the interactions among the constituents, which nonlinearly distort the mixed spectra.[3] Consequently, the analyte spectra cannot be linearly deconvoluted into the reference spectra. Such nonlinear distortion includes peak shifts or new peak emergences caused by the constituent interactions/reactions which are common occurrences in chemistry when mixing different compounds. Therefore, in polymer sciences, the difficulty of CA is compounded as criterion (1) is often unmet, and additionally, criterion (2) frequently remains unsatisfied. This imposes a significant demand for developing accurate CA methods applicable to highly interactive systems with the presence of nonlinear distortion.

To address the demand, herein I propose CA methods applicable to highly interactive and reference-free systems using pyrolysis mass-spectrometry (MS)[11,12], termed nonlinear and reference-free quantitative MS (NL-RQMS). Pyrolysis-MS analyzes the mass pattern of gaseous fragments generated by heat-decomposition of polymers, which is a very powerful spectrometry for chemical identification regardless of the solubility of analytes;[4,13–18] however, owing to individually different ionization efficiencies of the fragments, which are sensitive to their chemical structures, peak intensity ratio does not directly indicate the

abundance ratio of the corresponding substances. Therefore, a spectral set of all the references is necessary for quantitative analysis of MS, i.e., well-known label-free quantification technique in proteomics.[19] For cases where the complete reference set cannot be prepared, I recently developed RQMS framework that accurately infers reference spectra only from a spectral dataset of mixtures and demonstrated its application for polymer sequencer.[4] The reference-free nature of this approach provides great flexibility in selecting system constituents via the dataset design. This allows for analyzing not only chemical composition but also polymer properties[10] and geometric information.[4] However, as I already reported, RQMS could not execute accurate CA for highly interactive systems where the constituents react with each other.[12] This is because the products generated by in-situ reactions give rise to new peaks at positions unrelated to the reactants, significantly distorting the composition estimation. This distortion is illustrated in Fig. 1 using a simplified spectroscopy with only four-channels, where three components of A, B, and C occupy channel 1, 2 and 3 respectively. In non-interactive system, any mixed spectrum can be approximated by linear combination of the reference spectra with appropriate scalar coefficients, that represent the composition. However, in a highly reactive system, although none of A, B, nor C has a signal at the channel 4, it could be occupied in the mixed spectra by a new signal from the A-B cross-component generated by the reaction/interaction. This also reduces the signal intensities of A and B because they are consumed in the reaction, resulting in a mixed spectrum completely different from that of a non-interactive system. Evidently, linear deconvolution of the mixed spectra into the reference spectra, without considering such spectral variation due to mixing, results in incorrect compositions that

deviate significantly from the truth, even if all the reference spectra are known. Furthermore, the difficulty of CA drastically increased when the references are unknown, as the nonlinear distortion adversely affects the reference inference in linear RQMS; to explain the signal emergence at channel 4 in the mixed spectra, it would be mistakenly inferred that one of the reference spectra possesses the channel, leading to huge error in CA results. To the best of my knowledge, the proposed NL-RQMS would, for the first time, provide accurate CA results in the presence of such reactions commonly occurred in chemical spectroscopies. NL-RQMS is specialized for MS analysis, but the underlying idea of analyzing non-quantitative spectra quantitatively, even in the presence of chemical reactions, is applicable to a wide range of chemical spectra. This is because the NL-RQMS model was constructed based on general insights regarding law of mass action and the sparsity of reaction sites, as shown below.

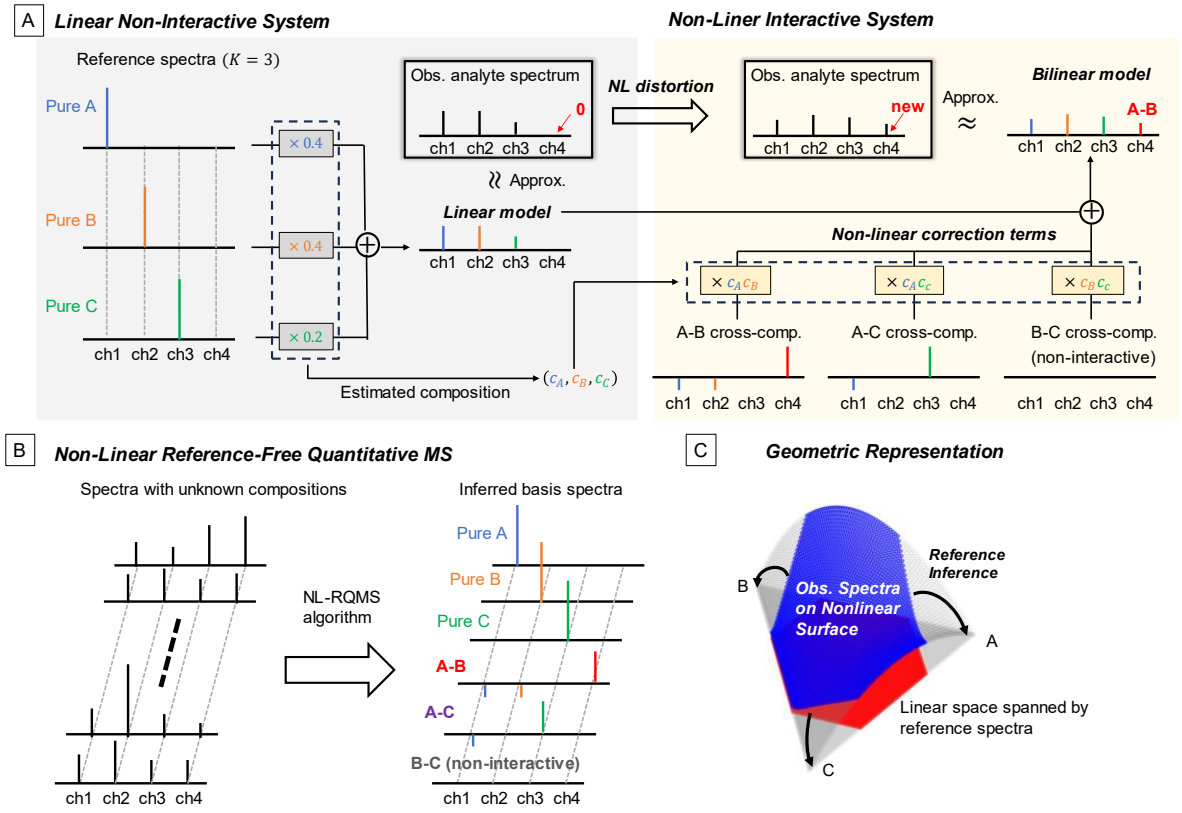

Fig. 1. Nonlinear distortion caused by interactions among the constituents. (A) A bilinear model for accounting for the interaction effect by introducing nonlinear correction terms to a linear model. (B) Inferring the reference and cross-component spectra by NL-RQMS and (C) its geometric representation in spectral space.

Results and discussion

Algorithm design principles

To begin with, let us consider the case where criterion (1) is satisfied, i.e., the reference spectra are all available (Fig. 1A). To account for nonlinear interactions described above, I herein introduce a bilinear model[3] where nonlinear correction terms are added to the linear combination model (Fig 1A, right). With the recent explosive advancements in deep learning, model-free nonlinear analysis using autoencoders is becoming mainstream in image spectral analysis, such as in remote sensing.[20] However, unlike image spectra, where a single image contains thousands to tens of thousands of spectra, i.e., each pixel having its own spectrum, in chemistry and materials science, it is standard to obtain only one spectrum per sample, making it extremely challenging to create spectral datasets with more than 100 samples. This makes the application of deep learning difficult, and model-based analysis grounded in general chemical knowledge remains effective. The bilinear model was originally proposed for considering the multi-scattering effect in spectral imaging analysis for remote sensing.[3] It defines the cross-component spectra by the element-wise product of two reference spectra, and their intensities by the element-wise product of their concentrations. Apparently, this original bilinear model cannot be directly applied to chemical spectroscopies, because cross-components generated by chemical interaction/reaction may exhibit new peaks unrelated to the corresponding references, which cannot be mathematically represented by the reference spectra. Due to this critical difference from the optical spectroscopies, a framework is needed to learn the cross-component spectra from an observed MS dataset of mixtures (Fig. 1B). Nevertheless, it is reasonable to assume

that the interaction degree between two components is proportional to the product of their abundances, in analogy to reaction kinetics where product concentrations are proportional to the products of reactant concentrations (Fig. 1A, right). Furthermore, this formulation captures a critical aspect of the nonlinearity caused by chemical interactions/reactions: if a sample lacks a specific component, the intensities of the corresponding cross-components in this sample becomes zero. This means that, even in a system with strong interactions, the reference spectra can match those of a system without interactions.[3] This can be geometrically represented in spectral space, where the reference spectra corresponding to the vertices define a surface that encompasses all the mixed spectra (Fig. 1C). The vertices of the nonlinear surface should match those of the linear space. This insight offers guidance on how to learn the cross-component spectra from the data; they should correspond to the distance between the linear space and the nonlinear surface.[21] Furthermore, the coincidence of these vertices provides an essential clue for inferring the reference spectra solely from the observed mixed spectra: by subtracting the interaction terms from the mixed spectra on the nonlinear surface and projecting them onto the linear space, linear RQMS can deduce the reference spectra at the vertices from the projected mixed spectra. Therefore, I designed the NL-RQMS so that the reference spectra, their abundances in the mixed spectra, and the cross-component spectra are cyclically updated through iterations, based on the theory of the block coordinated descent.[22] In the initial iterations, applying linear RQMS directly to the mixed spectra on the nonlinear surface leads to inaccurate reference spectra. However, after learning the cross-component spectra, the mixed spectra on the nonlinear surface can be appropriately transformed onto the linear space by subtracting the cross-component effects. Consequently,

the cyclic updates would eventually allow accurate inferences of the reference spectra and their abundances as well as the cross-component spectra.

Benchmark test system of reactive ternary polymers

The aim of this benchmarking CA test is to assess the precision of NL-RQMS algorithm. To achieve this, I here use a published dataset[12] of a ternary polymer system of diglycidyl ether of bisphenol A-based epoxy (Gly, M_w=1,650), poly(propylene glycol) diamine (Jeff, M_w=2,000), and polydimethylsiloxane diamine (Silox, M_w=2,500), where the glycidyl groups of Gly react with primary amines of Jeff and Silox. The dataset was prepared with predetermined compositions, allowing us to evaluate the accuracy of NL-RQMS through the discrepancy between the deduced and ground-truth compositions. Throughout the process of deducing compositions, I did not utilize any known mixture compositions nor the reference spectra to simulate CA for unknown polymer systems. I refrained from applying NL-RQMS to genuinely unknown systems, as verifying the accuracy of the inferred compositions would be infeasible. The dataset including 31 samples of binary or ternary mixtures had been prepared using thermogravimetry (TG)-MS setup as reported before, where no component exceeded an 80 wt% composition in any samples (reference-free condition). It was already reported that the estimated composition by conventional RQMS for the Gly-Jeff-Silox system contained errors of 6.1 wt%, which were significant and impractical compared to the estimation errors of 1-3 wt% for non-interactive systems.[12]

The origin of nonlinearity in pyrolysis-MS

MS with high resolution along *m/z* axis (full width at half maximum<0.01 *m/z*) often

contains thousands of peaks at different *m/z* channels (*D*), resulting in much more complex spectra than simplified spectra ($D = 4$) used for explanation in Fig. 1. Furthermore, peak overlapping, where the same *m/z* channel is occupied by different components with common substructures, becomes problematic, especially for small fragments generated by hard ionization methods like electron impact (EI). To address these issues, conventional linear RQMS employs a two-step non-negative matrix factorization (NMF).[4,5] The first NMF analyzes the entire spectral dataset and groups peaks that fluctuate with consistent intensity ratios. This allows, for instance, isotope peaks or smaller fragment peaks generated by hard EI-ionization, which are generated at a certain intensity ratio from their parent fragment, to be consolidated into a single spectrum, referred to as a fragment spectrum. The number of the fragment spectra, *M*, can be determined using an automatic method based on sparse modeling.[6] (see Fig. S1 for Gly-Jeff-Silox system where $M = 27$. The used hyperparameters were all consistent to previous work as presented in Table S1). The linear combination of *M*–fragment spectra should well approximate all the observed spectra in the dataset, and the combination coefficients for each fragment spectrum corresponds to the fragment abundances (FA) in each sample, represented as a non-negative matrix $\boldsymbol{A}$ with (*N, M*)-size (Fig. 2A; the numerical data is presented in Data S1). The benefits of the peak consolidation include not only dimensionality reduction–from the *D*-dimensional spectral space into *M*-dimensional FA space–but also increased robustness against peak overlapping. Distinct chemical structures produce unique fragmentation patterns with many peaks at different *m/z* values at specific intensity ratios. Even when some channels are co-occupied by two different fragments, the overlapped peaks can be accurately deconvoluted into their corresponding

fragment spectra, while maintaining consistency with other non-overlapped peaks. Importantly, this step can be handled by the linear algorithm, even in reactive or interactive systems, because MS peak intensities increase linearly with fragment abundance. The source of nonlinear distortion stems from interactions or reactions that modify fragment abundance.

Since any FA of a mixture should be generated by mixing the FAs of *K*–pure reference polymers (Fig. 2A top) according to their fraction, the reference FAs and compositions can be determined simultaneously in the second NMF. Formally, the two-step NMF can be expressed as follows:

$$\boldsymbol{X} \approx \boldsymbol{AS} \approx (\boldsymbol{CB})\boldsymbol{S},$$

where $\boldsymbol{X}$ is the observed spectral set represented by the (*N, D*)-matrix (a row of $\boldsymbol{X}$ corresponds to a spectrum with *D*–channels), $\boldsymbol{S}$ is the *M*–fragment spectra represented by the (*M*, *D*)-matrix, $\boldsymbol{C}$ is the sought composition of *K*–components in *N*–samples represented by the (*N, K*)-matrix and $\boldsymbol{B}$ is the FA of the *K*-pure polymers represented by the (*K*, *M*) matrix. Note that $\boldsymbol{X}, \boldsymbol{A}, \boldsymbol{S}, \boldsymbol{C}, \boldsymbol{B}$ are all non-negative matrices. Herein, some terms are used that may not be common in MS or analytical sciences, so please refer to the flowchart in Fig. S2 for their relationships and definitions as needed.

Figure 2A clearly elucidated why this conventional RQMS cannot find correct composition $\boldsymbol{C}$ for the interactive/reactive systems like a Gly-Jeff-Silox dataset. For an example, fragment 18 was not found in any references but abundant in Gly-Silox binary and ternary mixtures (see the column vertically). Note that the correct answer of the reference FA ($\boldsymbol{B}$) to be sought was obtained for explanation by projecting the reference spectra into the spectral space spanned by *M*–fragment spectra but never used for the analytical process, as

the reference cannot be prepared in the practical scenario. As described above, $\boldsymbol{B}$ should be determined so that any rows of $\boldsymbol{A}$ can be well-approximated by linearly mixing the rows of $\boldsymbol{B}$; however, this is impossible for fragment 18 as its abundance in $\boldsymbol{B}$ was zero. This consideration suggests that the nonlinearity in pyrolysis-MS is attributable to the interactions between hetero polymers in their melt states prior to the pyrolysis, which change the generation patterns of fragments from those of the pure reference polymers. Further careful observation of Fig. 2A revealed that each fragment can be classified into three types: those whose abundances change linearly with composition, and those whose abundances exhibit nonlinear effects (positive or negative) with composition. Representative fragments for these categories are highlighted with white (fragment 0), red (fragments 18-20), and green frames (fragments 12, 26), respectively. Fragment 0 was found only in Silox among pure samples, and its abundances linearly decreased as Gly composition increased in Gly-Silox binary samples (Fig. 2B; also see the white frame from bottom to top in Fig. 2A). Therefore, fragment 0 can be attributed to the non-interactive substructure of Silox, whose abundance directly reflects the Silox composition. In contrast, fragment 18, categorized as having positive nonlinearity, was found only in Gly-Silox binary and ternary samples and not in any references. The quadratic dependence of the FA on the Gly-Silox composition (Fig. 2B, red) clearly indicated that fragment 18 was generated through the interaction/reaction of Gly and Silox. Note that this does not necessarily mean that fragment 18 was derived from the new product of Gly-Silox, but it may mean the fragment was enhanced by their interaction, as will be later discussed. Fragments 12 and 26, attributable to the substructures of Gly and categorized as having negative nonlinearity, exhibited more complicated distributions; the

FAs decreased much more rapidly with the reduction in Gly composition in the Gly-Jeff binary system than in the Gly-Silox binary system. These nonlinear decreases in the FA in the Gly-Jeff binary were well represented by convex quadratic fitting (Fig. 2C), suggesting that fragments 12 and 26 were the reactive fragments consumed by the Gly-Jeff reaction, as will be later justified via fragment spectra analysis. The FAs of fragments 18-20 exhibited strong positive nonlinearities in the Gly-Jeff binary system, suggesting that they were enhanced or generated by Gly-Jeff interaction. Indeed, as described later, the *m/z* values of the fragment spectra indicated that they corresponded the Gly-Jeff joint moieties generated by the epoxy-amine reaction. To statistically estimate the nonlinear distortion effect in the Gly-Jeff-Silox system, the deviations of FA in each sample from the linear plane spanned by the references were integrated across all 27 fragments (Fig. S3). This analysis showed that 41 wt% of the fragments were affected by interaction effects from mixing, indicating significant nonlinear distortion. This suggests that the system weas well-suited for validating the nonlinear analysis algorithm.

It should be emphasized that such consideration was possible only because this was benchmark test of fully known system where the composition information and reference spectra are all available. In practical applications, it is impossible to draw composition-FA curves like those in Fig. 2B-2D, and thus, it is not feasible to determine which fragments have a linear/nonlinear distribution over the dataset. Therefore, the composition should be determined only based on the mixture FA (the matrix $\boldsymbol{A}$ in Fig. 2A) obtained through unsupervised learning of the first NMF without using any information about the system. Also note that neither the reference spectra nor their FAs ($\boldsymbol{B}$) were used in any learning steps. By

prohibiting the use of composition and reference information in the benchmark CA test, we can simulate the analysis of an unknown system; yet the known true composition and reference spectra allows us to validate the NL-RQMS algorithm.

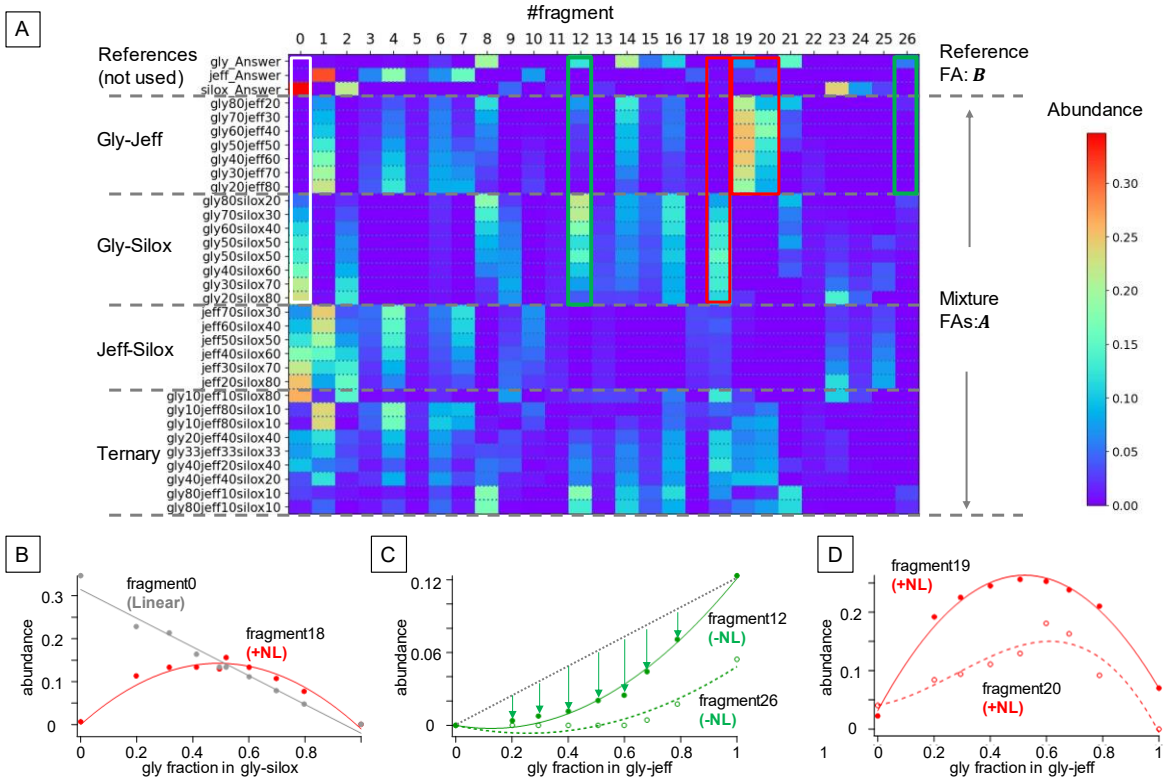

Fig. 2. The distribution of fragment abundance (FA) over the dataset. (A) FAs were represented as a heatmap. The representative fragments for linearity, positive nonlinearity and negative nonlinearity were marked as white, red, and green frames, respectively. Note that the three-top rows corresponding to the sought reference FAs ***B*** based on the mixtures FAs ***A***. (B-D) The relationship between FAs of the representative fragments and weight fraction of Gly in binary mixtures.

NL-RQMS Algorithm and Benchmark Test

NL-RQMS should find correct FAs of *K*–references $\boldsymbol{B}$ and composition $\boldsymbol{C}$ based on the nonlinearly distorted FAs of mixtures $\boldsymbol{A}$. To eliminate the nonlinearity caused by interactions/reactions from $\boldsymbol{A}$, *K'*–cross-components and their interaction degree are here introduced, where $K' = \binom{K}{2}$. The cross-components represent the binary interaction effects, which may alter the FA–composition curves positively or negatively compared to the linear change (Fig. 2B-D). Therefore, the FAs of the cross-components $\boldsymbol{B}'$ represented by a (*K', M*)-matrix may have both positive and negative values. In the Gly-Jeff-Silox system, $\boldsymbol{B}'$ describes Gly-Jeff, Gly-Silox and Jeff-Silox interactions at each row. The interaction degree in each sample can be calculated by production of the concentration of the corresponding components (see Fig. 1). Figure 3 shows a bilinear model of spectral generation, extended with a cross-component to account for the nonlinear interaction.

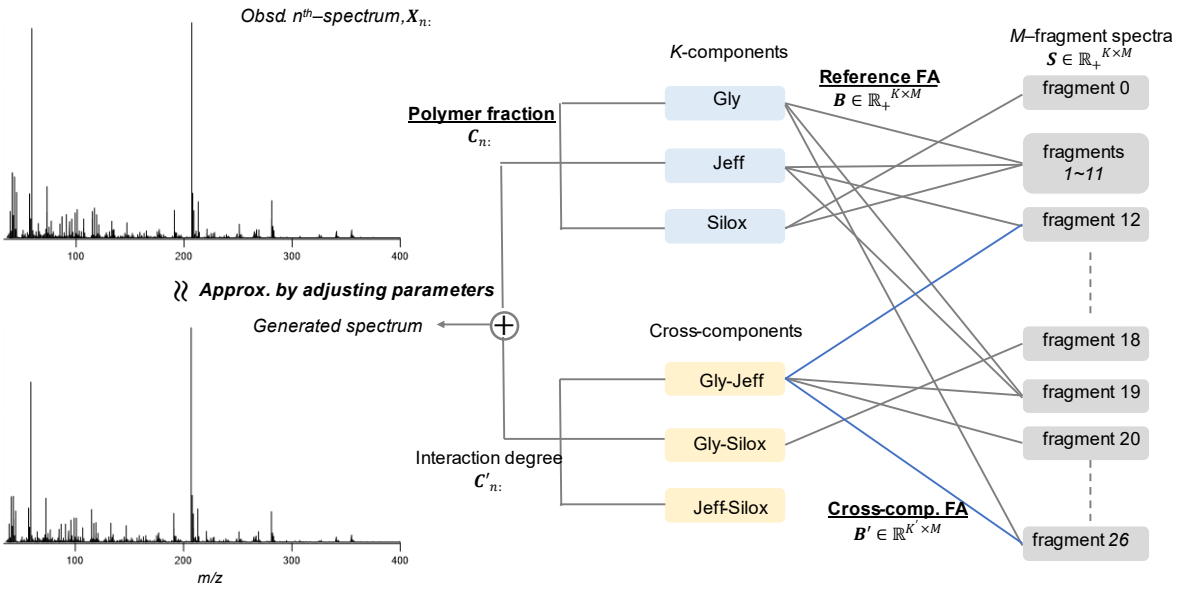

Fig. 3. Spectral generation model using the linear combination of 27 fragment spectra. The linear combination constants for each sample consist of the sum of the linear term, which mixes the FAs of *K*–reference polymers based on the polymer composition, and the nonlinear term, which mixes the FAs of *K′*–cross-components according to the interaction degree. The gray and blue lines represent positive and negative coefficients. Herein, $K = K' = 3$.

According to this model, an n^{th} observed spectrum stored in the n^{th} row of $\boldsymbol{X}$ (written as $\boldsymbol{X}_{n:}$) is approximated by linear combination of M–fragment spectra (Fig. 3):

$$\boldsymbol{X}_{n:} \approx \boldsymbol{A}_{n:}\boldsymbol{S} \approx (\boldsymbol{C}_{n:}\boldsymbol{B} + \boldsymbol{C'}_{n:}\boldsymbol{B'})\boldsymbol{S}, \qquad (n = 1, 2, \ldots, N).$$

Relating to Fig. 1C, $\boldsymbol{BS}$ corresponds to the K–reference spectra spanning linear space (red triangle) and $\boldsymbol{B'S}$ corresponds to the K'–cross-component spectra, which lift the datapoint of $\boldsymbol{X}_{n:}$ from the linear space to the nonlinear surface (blue surface). Therefore, by subtracting $\boldsymbol{C'}_{n:}\boldsymbol{B'S}$ from each spectrum, the nonlinearity can be removed. Then, linear RQMS algorithm applied to the projected datapoints onto the linear space would find the correct vertex points corresponding to $\boldsymbol{BS}$. The difficulty is that all $\boldsymbol{C}$, $\boldsymbol{B}$ and $\boldsymbol{B'}$ are unknown. To solute this situation, the nonlinearity is first ignored ($\boldsymbol{B'} \equiv \boldsymbol{0}$), and directly apply linear RQMS to the datapoints on the nonlinear surface, yielding temporal vertices and $\boldsymbol{C}$. This inaccurate inference of composition is depicted in Fig. 4A. The estimation errors for compositions were particularly large on the Gly-Jeff edge, the most reactive combination that causes significant nonlinear distortion. Since all data points on the curved surface cannot be arranged onto a linear plane, temporal $\boldsymbol{B'}$ are defined to minimize these deviations of datapoints from the plane. Here, it is crucial to consider the chemical meaning of $\boldsymbol{B'}$, which represents the reactive/interactive substructures of the polymers. In most case, it is chemically rational to predict that the variety of such substructures in a system is limited: i.e., most substructures are non-interactive. Furthermore, fragments derived from polymers unrelated to a specific cross-component cannot participate in these interactions/reactions, making it natural to expect $\boldsymbol{B'}$ to be a sparse matrix with many zero elements. Therefore, when optimizing $\boldsymbol{B'}$, L1 regularization was applied, which lead $\boldsymbol{B'}$ to be sparse.[23] After subtracting $\boldsymbol{C'}_{n:}\boldsymbol{B'S}$ from each

spectrum from the data points, linear RQMS was applied again to recalculate $\boldsymbol{B}$ and $\boldsymbol{C}_{n:}$. Iteration of this process eventually learned more suitable vertices and the nonlinear surface, resulting in a much better solution (Fig. 4B). This notably improved the estimation accuracy on the highly reactive Gly-Jeff edge compared to linear RQMS. The algorithm's accuracy was evaluated using root-mean-square errors (RMSE) between the ground-truth and estimated compositions, which was significantly reduced from 0.061 to 0.033 by introducing a nonlinear correlation term (average errors improved from 6.1wt% to 3.3wt%; the impact of this improvement is clearly illustrated in the numerical data in Table S2. The calculation method for accuracy is presented in the caption). Notably, NL-RQMS demonstrated strong robustness against input noise (Fig. S4). Even when random Gaussian noise with a 10% variance relative to the sample variance was added to the FA matrix $\boldsymbol{A}$, the RMSE remained at 3.8 wt%, indicating the reliability of the NL-RQMS method. The mathematical derivation of the new part of NL-RQMS for determining $\boldsymbol{C}$, $\boldsymbol{B}$ and $\boldsymbol{B}'$ based on $\boldsymbol{A}$ is detailed in Supporting Information. The reference spectra inferred by NL-RQMS were well consistent to the observed spectra (Fig. 4C). Furthermore, the interaction effects $\boldsymbol{B}'$ were chemically interpretable (Fig. 4D; see Fig. S5 for peak characterization of important fragment spectra specified as having positive or negative nonlinear effects) as follows. First, the fragment 12 and 26, which had negative nonlinear effects, were characterized as having terminal epoxy group. This is very reasonable as they were reactants and consumed by the in-situ reaction. Second, fragment 20 having positive non-linear effect was attributable to $C_4H_{10}N$. This was also very reasonable as neither the pristine Jeff nor Silox had the substructure of $C_4H_{10}N$ and this should be the product of the reaction with Gly. Based on the exact mass alone, it was not

possible to determine whether this was a secondary or tertiary amine. However, considering that the composition-abundance relationship of Fragment 20 in Fig. 2D was approximated by a cubic polynomial, it can be inferred that the tertiary amine should be the main product generated by a 1:2 reaction with epoxy. Note that NL-RQMS did not use the FA-composition information available only in benchmark tests. Therefore, the nonlinear relationship represented by a cubic polynomial was also fitted with a quadratic polynomial, as designed by the bilinear model. Although this is the limitation of the bilinear model, quadratic fitting also allowed for setting fragment 20 almost zeros both in both pure Gly and Jeff, allowing accurate CA. This is the clear advantage over linear RQMS; in Fig. 2D, without the data point where the Gly fraction is 1 under reference-free conditions, linear fitting would produce an upward-sloping line, leading to the incorrect conclusion that fragment 20 is derived from Gly. Fragment 18 and 19, which had positive nonlinear effects, were attributable to the main chains of Silox and Jeff, respectively. The interpretation why these fragments were enhanced by mixing with Gly cannot be easily rationalized, as these were not products of the in-situ reactions. A plausible explanation is that the generation of these cyclic compounds via back-biting depolymerization was accelerated by the existence of the proton on the side chain of Gly. Although such interactions rather than reactions are difficult to interpret only from MS data, the high resolution along *m/z* axis (full width at half maximum<0.01 *m/z*) would provide a clue for considering the plausible in-situ reaction (Fig. S5).

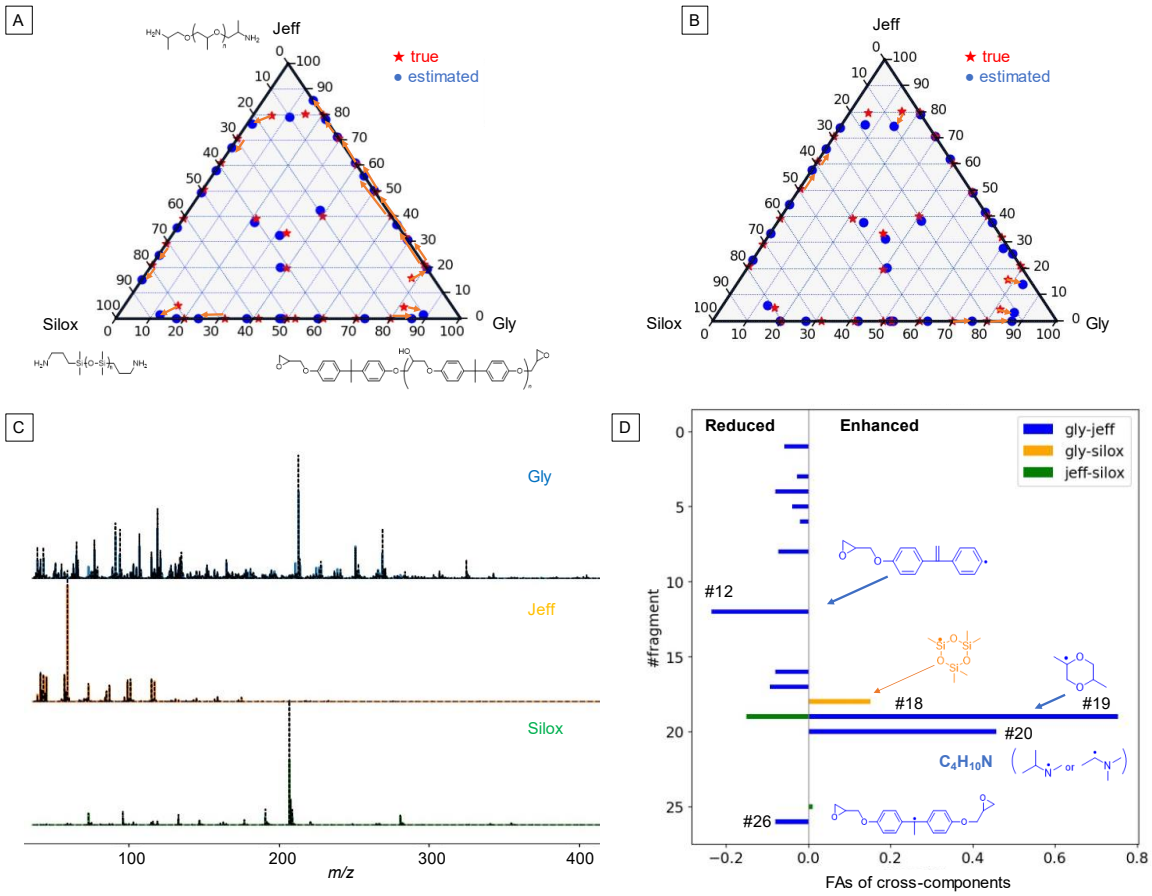

Fig. 4. Benchmark CA test of NL-RQMS. (A) The ground-truth (red stars) and linear RQMS-inferred (blue dots) compositions were depicted. For data with large discrepancies that make pairing difficult to discern, the combinations are indicated by orange lines (also see the numerical data for easy see of the estimation gaps on the Gly-Jeff edge in Table S2). RMSE: 0.061. (B) NL-RQMS–inferred compositions. RMSE: 0.033. (C) The reference spectra inferred by NL-RQMS. Observed reference spectra were super imposed to show the consistency. (D) Binary interaction effects on FAs.

Conclusion

This paper first demonstrates how to conduct compositional analysis for systems where in-situ chemical reactions and interactions occur among the components. The proposed NL-RQMS would be highly effective in grading polymer materials where weather-resistant agents, that act as radical scavengers and alter the pyrolysis mode of polymers through interactions, are included. It is also expected to be valuable for analyzing polymer systems with unstable sites where reactions frequently occur prior to pyrolysis. The proposed algorithm is built on fundamental chemical principles, such as the law of mass action and the sparsity of reactive substructures in polymers. This general foundation makes the nonlinear algorithm flexible and applicable to other chemical spectroscopies where quantitative analysis is challenging due to interaction-induced peak shifts and emergences, such as infrared absorption (IR), Raman, and ultraviolet-visible absorption spectroscopies. Lastly, I would like to highlight some key considerations when applying NL-RQMS to systems with a large number of components ($K >> 3$). The second step of matrix decomposition in RQMS relies on the so-called volume-minimization (volmin) algorithm,[9] which estimates the vertices of the smallest simplex enclosing all spectral data points. Although counterintuitive, the volmin developers have mathematically suggested that vertex estimation becomes easier as the number of vertices increases.[9] Therefore, as long as interaction effects are properly subtracted and the nonlinear surface is accurately projected into linear space, an increase in K should not cause significant error into the CA results. However, as K increases, the number of potential binary interactions grows rapidly, the estimation of these interaction effects become challenging with limited datasets. Errors in estimating interactions result in

projecting onto the wrong linear surface, leading to inaccurate vertex estimation. Therefore, it is critical to incorporate as much prior chemical knowledge of the system as possible by employing the following strategies.

1. Include any available measurable pure references in the dataset to anchor the vertices as much as possible.
2. Force interaction parameters to zero for known to be non-interactive.

These strategies would enhance the NL-RQMS's ability to analyze highly complex interaction systems.

Overall, NL-RQMS is anticipated to provide significant insights into the quantitative understanding of nonlinear systems in polymer science.

Supporting Information: Mathematical derivation for NL-RQMS, supplementary Fig. S1-S5, supplementary Table S1-S2, numerical data of ready-to-use starting matrix (Data S1). The raw MS dataset is available at previous report (Data S3)[12].

Conflicts of interests: The author declares no conflicts of interest.

Funding: This work was supported by JSPS KAKENHI Grant Number JP24K08520.

Reference List:

(1) Heinz, D. C.; Chang, C.-I. Fully Constrained Least Squares Linear Spectral Mixture Analysis Method for Material Quantification in Hyperspectral Imagery. *IEEE Trans. Geosci Remote Sens.* **2001**, *39* (3), 529–545.
(2) Fu, X.; Huang, K.; Sidiropoulos, N. D.; Ma, W. K. Nonnegative Matrix Factorization for Signal and Data Analytics: Identifiability, Algorithms, and Applications. *IEEE Signal Process Mag.* **2019**, *36* (2), 59–80.
(3) Dobigeon, N.; Tourneret, J. Y.; Richard, C.; Bermudez, J. C. M.; McLaughlin, S.; Hero,

A. O. Nonlinear Unmixing of Hyperspectral Images: Models and Algorithms. *IEEE Signal Process Mag.* **2014**, *31* (1), 82–94.
(4) Hibi, Y.; Uesaka, S.; Naito, M. A Data-Driven Sequencer That Unveils Latent "Codons" in Synthetic Copolymers. *Chem. Sci.* **2023**, *14*, 5619-5626.
(5) Lee D. D.; Seung H. S. Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature* **1999**, *401* (21), 788–791.
(6) Shiga, M.; Tatsumi, K.; Muto, S.; Tsuda, K.; Yamamoto, Y.; Mori, T.; Tanji, T. Sparse Modeling of EELS and EDX Spectral Imaging Data by Nonnegative Matrix Factorization. *Ultramicroscopy* **2016**, *170*, 43–59.
(7) Craig, M. D. Minimum-Volume Transforms for Remotely Sensed Data. *IEEE Trans. Geosci. Remote Sens.* **1994**, *32* (3), 542–552.
(8) Miao, L.; Qi, H. Endmember Extraction from Highly Mixed Data Using Minimum Volume Constrained Nonnegative Matrix Factorization. *IEEE Trans. Geosci. Remote Sens.* **2007**, *45* (3), 765–777. https://doi.org/10.1109/TGRS.2006.888466.
(9) Fu, X.; Huang, K.; Yang, B.; Ma, W. K.; Sidiropoulos, N. D. Robust Volume Minimization-Based Matrix Factorization for Remote Sensing and Document Clustering. *IEEE Trans. Signal Process.* **2016**, *64* (23), 6254–6268.
(10) Hibi, Y.; Tsuyuki, Y.; Ishii, S.; Ide, E.; Naito, M. Decoding Thermal Properties in Polymer-Inorganic Heat Dissipators: A Data-Driven Approach Using Pyrolysis Mass Spectrometry. *Sci. Technol. Adv. Mater.* **2024**, *25* (1). https://doi.org/10.1080/14686996.2024.2362125
(11) Zitomer F. Thermogravimetric-Mass Spectrometric Analysis. *Anal. Chem.* **1968**, *40* (7), 1091–1095.
(12) Hibi, Y.; Uesaka, S.; Naito, M. Thermogravimetry-Synchronized, Reference-Free Quantitative Mass Spectrometry for Accurate Compositional Analysis of Polymer Systems Without Prior Knowledge of Constituents. *Analyst* **2024**. https://doi.org/10.1039/D4AN00624K
(13) Marić, M.; Marano, J.; Cody, R. B.; Bridge, C. DART-MS: A New Analytical Technique for Forensic Paint Analysis. *Anal. Chem.* **2018**, *90* (11), 6877–6884.
(14) Liang, J.; Frazier, J.; Benefield, V.; Chong, N. S.; Zhang, M. Forensic Fiber Analysis by Thermal Desorption/Pyrolysis-Direct Analysis in Real Time-Mass Spectrometry. *Anal. Chem.* **2020**, *92* (2), 1925–1933.
(15) Yamane, S.; Nakamura, S.; Inoue, R.; Fouquet, T. N. J.; Satoh, T.; Kinoshita, K.; Sato, H. Determination of the Block Sequence of Linear Triblock Copolyethers Using Thermal Desorption/Pyrolysis Direct Analysis in Real-Time Mass Spectrometry. *Macromolecules* **2021**, *54* (22), 10388–10394.
(16) Nakamura, S.; Watanabe, R.; Yamane, S.; Sato, H. Kendrick Mass Defect Analysis-Based Data Mining Technique for Trace Components in Polyolefins Observed by Pyrolysis-Gas Chromatography/High-Resolution Mass Spectrometry. *J. Anal. Appl. Pyrolysis* **2023**, *170*.
(17) Watanabe, R.; Nakamura, S.; Sugahara, A.; Kishi, M.; Sato, H.; Hagihara, H.; Shinzawa, H. Revealing Molecular-Scale Structural Changes in Polymer Nanocomposites during Thermo-Oxidative Degradation Using Evolved Gas Analysis with High-Resolution Time-of-Flight Mass Spectrometry Combined with Principal

Component Analysis and Kendrick Mass Defect Analysis. *Anal. Chem.* **2024**, *96* (6), 2628–2636.
(18) Mase, C.; Maillard, J. F.; Piparo, M.; Friederici, L.; Rüger, C. P.; Marceau, S.; Paupy, B.; Hubert-Roux, M.; Afonso, C.; Giusti, P. GC-FTICR Mass Spectrometry with Dopant Assisted Atmospheric Pressure Photoionization: Application to the Characterization of Plastic Pyrolysis Oil. *Analyst* **2023**, *148* (20), 5221–5232.
(19) Zhu, W.; Smith, J. W.; Huang, C. M. Mass Spectrometry-Based Label-Free Quantitative Proteomics. *J. Biomed Biotechnol.* **2010**, *2010*. 840518. https://doi.org/10.1155/2010/840518.
(20) Su, Y.; Li, J.; Plaza, A.; Marinoni, A.; Gamba, P.; Chakravortty, S. DAEN: Deep Autoencoder Networks for Hyperspectral Unmixing. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57* (7), 4309–4321.
(21) Févotte, C.; Dobigeon, N. Nonlinear Hyperspectral Unmixing with Robust Nonnegative Matrix Factorization. *IEEE Trans. Image Process.* **2015**, *24* (12), 4810–4819.
(22) Kim, J.; He, Y.; Park, H. Algorithms for Nonnegative Matrix and Tensor Factorizations: A Unified View Based on Block Coordinate Descent Framework. *J. Global Optimization* **2014**, *58* (2), 285–319.
(23) Tibshiranit, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **1996**, *58* (1), 267–288.