*Journal of Surface Analysis, Vol. 25 No. 2 (2018) pp. 103 - 114*
*石倉航，他 多変量解析を利用した TOF-SIMS イメージデータフュージョンとスパースモデリングおよび機械学習による…*
*https://doi.org/10.1384/jsa.25.103*

*研究論文*

# 多変量解析を利用した TOF-SIMS イメージデータフュージョンとスパースモデリングおよび機械学習による TOF-SIMS スペクトル解析

石倉 航,[1] 高橋 一真,[1] 山岸 崇之,[1] 青木 弾,[2] 福島 和彦,[2] 志賀 元紀,[3,4,5] 青柳 里果 [1,*]
*[1] 成蹊大学理工学部 物質生命理工学科, 180-8633 武蔵野市吉祥寺北町 3-3-1*
*[2] 名古屋大学大学院 生命農学研究科, 464-8601 名古屋市千種区不老町*
*[3] 岐阜大学 工学部, 501-1193 岐阜市柳戸 1 番 1*
*[4] 科学技術振興機構 さきがけ, 332-0012 埼玉県川口市本町 4-1-8*
*[5] 理化学研究所 革新知能統合研究センター, 103-0027 中央区日本橋 1-4-1 日本橋一丁目三井ビルディング 15 階*
**aoyagi@st.seikei.ac.jp*

（2018 年 7 月 30 日受理; 2018 年 10 月 25 日掲載決定）

飛行時間形二次イオン質量分析法（Time-of-Flight secondary ion mass spectrometry: TOF-SIMS）は高空間分解能での化学イメージングが可能な手法であり，もっとも優れた化学イメージング法の 1 つである．しかし，ナノレベルでのより高い空間分解能でのイメージングが要求されているため，より空間分解能の高い SEM データとイメージフュージョンしたイメージデータを主成分分析することにより，化学情報を保ったまま TOF-SIMS 本来よりも高い空間分解能で表現した．また，TOF-SIMS スペクトルは解釈が難しい場合が多いため，スパースモデリングと機械学習を応用し，スペクトルの単純化や自動判別を試みた．

# TOF-SIMS Image Data Fusion by Multivariate Analysis and TOF-SIMS Spectrum Analysis by Sparse Modeling and Machine Learning

Wataru Ishikura,[1] Kazuma Takahashi,[1] Takayuki Yamagishi,[1] Dan Aoki,[2] Kazuhiko Fukushima,[2] Motoki Shiga,[3,4,5] and Satoka Aoyagi,[1,*]
*[1] Faculty of Science and Technology, Seikei University, 3-3-1 Kichijoji-kitamachi, Musashino, Tokyo 180-8633 Japan*
*[2] Graduate School of Bioagricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601*
*[3] 3 Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu City 501-1193, Japan*
*[4] JST PRESTO, 4-1-8, Honcho, Kawaguchi-shi, Saitama, 332-0012 Japan*
*[5] Center for Advanced Intelligence Project, RIKEN, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi,Chuo-ku, Tokyo 103-0027, Japan*
**aoyagi@st.seikei.ac.jp*

(Received: July 30, 2018; Accepted: October 25, 2018)

Time-of-Flight secondary ion mass spectrometry (TOF-SIMS) and scanning electron microscope (SEM) images were fused and then evaluated by means of principal component analysis. As a result, TOF-SIMS spatial resolution could be improved by adding SEM image information to TOF-SIMS data without drastic change of TOF-SIMS spectrum information. Sparse modeling and machine learning were applied to TOF-SIMS data to interpret complex TOF-SIMS spectra. Least Absolute Shrinkage and Selection Operator (LASSO) provided a simplified TOF-SIMS

spectrum with less noise. Machine learning using Random Forest and k-Nearest Neigbour appropriately predicted unknown test samples by learning TOF-SIMS data similar the test samples.

## １．はじめに

飛行時間形二次イオン分析（TOF-SIMS）は，高感度にサブミクロンでの分子分布が得られることから様々な分野に応用されている[1]が，フラグメントイオンからなるスペクトルの解釈が難しい[2]，共存物質によって二次イオン強度が変化するマトリックス効果[3,5]などの課題もある．スペクトルの解釈については，多変量解析などのデータ解析によって解決できる場合が多く，これまで多くの研究例が報告[1-3, 5-7]されている．ただし，近年では，Ar クラスターなどのガスクラスターイオンビーム（GCIB）によるスパッタリング[8]や，Orbitrap[9]や MSMS[10]などの TOF-SIMS 装置への導入によって，データ容量がより膨大となり，従来の多変量解析だけでは解析が困難な場合も出てきている．そこで，イメージング計測で実績のある機械学習法[11-15]や解釈可能性の高いスパースモデリング[16-18]の TOF-SIMS データへの応用が期待されている．本研究では，多変量解析からスパースモデリング及び機械学習までのデータ分析方法を活用した TOF-SIMS データ解析を示す．

すでに，光学顕微鏡像と TOF-SIMS 像とのフュージョンおよびピクセル削減による高輝度化データと高解像度データのイメージフュージョンに対する主成分分析の効果は検討[18]しているが，本研究では電子顕微鏡と TOF-SIMS データとのイメージフュージョンについて検討する．また，スパースモデリングに関しては，Robust PCA [19, 20]を適用することによって，主要な情報を失わずに二次イオンピーク数を削減し，少ないデータ数でも PCA を可能とした解析例[18]を報告した．本研究ではスパースモデリングの中で代表的な圧縮センシングの手法の一つである Least Absolute Shrinkage and Selection Operator（LASSO）[16, 17]を TOF-SIMS データに応用する．

## ２．実験方法

### 2.1 試料について

Si 基板上に leu-enkephalin（$C_{28}H_{37}N_5O_7$, Wako, Osaka）と 1,2-dioleoyl-*sn*-glycero-3-phosphocholine（DOPC, $C_{44}H_{84}NO_8P$, Avanti Polar Lipid Co. Ltd., Upsala, Sweden）を 75:25（wt%）で混合したペプチド脂質混合溶液 [21] を滴下し，真空デシケーター中で乾燥させた試料をイメージフュージョン用のデータ取得に用いた．

アルミコートガラス基板上の高分子 3 種類 polyethylene terephthalate（PET），polystyrene（PS）および polycarbonate（PC）の 4 層試料[2]をスパースモデリングと機械学習のモデル試料として採用した．また，PC を基板に 30 nm 程度の厚みとなるようにスピンコート（回転数 1000 rpm）した試料を機械学習の性能検証のために用意した．PC 単膜試料用の基板は，Si 基板および Si ウェハー上に 30 nm 程度の Au もしくは W 薄膜をコートした基板の３種類（PC/SI, PC/Au/Si, PC/W/Si）を用いた．

### 2.2 TOF-SIMS および SEM

高分子試料は，一次イオンビーム 54 keV $Bi_3^{++}$ の TOF-SIMS（PHI TRIFT V nano ToF, ULVAC-PHI, Chigasaki）で測定した．ペプチド脂質混合試料は，一次イオン源を 19 keV $Au^+$ とする TOF-SIMS（TRIFT III, ULVAC-PHI, Chigasaki）で 測定（Raster size; 100 × 100 μm$^2$ or 300 × 300 μm$^2$ , Pixel density; 256 × 256 pixels, Ion dose amount < $10^{12}$ ions/cm$^2$）し，さらに TOF-SIMS 測定部を含む位置を走査電子顕微鏡（SEM, S-3400N, Hitachi, Ltd., Tokyo）を用いて，加速電圧 1.5 kV，電流 40 mA，working distance 5 mm で測定した．

PC 単膜試料は，15 kV $Ga^+$を一次イオン源とする TOF-SIMS（TRIFT III）で測定（Raster size; 300 × 300 μm$^2$, Pixel density; 256 × 256 pixels, Ion dose amount < $10^{12}$ ions/cm$^2$, Ion dose amount < $10^{12}$ ions/cm$^2$）した．

測定した TOF-SIMS スペクトルについて，自動検索されたピーク全てについて，各ピクセルにおける強度をバイナリーファイルとして保存し，PLS Toolbox および MIA Toolbox（Eigenvector Research Inc., WA, USA）で，データを読み込んだ．

### 2.3 イメージフュージョンと PCA

TOF-SIMS スペクトル上の全ピークを自動検索して，各ピクセル上の各二次イオン強度をバイナリーファイルとして読み出したのち，MIA Toolbox の Image Manager を用いて行列形式（例えば，128×128 ピクセルに対する 1000 個の各二次イオン強度で，

**Table 1** Fused data of TOF-SIMS and SEM data.

| | A | | | | | | | B |
|---|---|---|---|---|---|---|---|---|
| | Peaks | | | | | | | |
| Pixels | 0 | 0 | 3 | 0 | … | 7 | 0 | 141 |
| | 0 | 0 | 3 | 0 | … | 11 | 0 | 150 |
| | 0 | 0 | 8 | 0 | … | 8 | 0 | 119 |
| | 0 | 0 | 2 | 0 | … | 15 | 0 | 103 |
| | 0 | 0 | 7 | 0 | … | 12 | 0 | 100 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 0 | 0 | 6 | 0 | … | 4 | 0 | 110 |
| | 0 | 0 | 2 | 0 | … | 10 | 0 | 69 |
| | 0 | 1 | 7 | 0 | … | 5 | 0 | 90 |

A: TOF-SIMS data, B: SEM data

16384×1000 行列となる）に変換した．Figure 1 に示すように，TOF-SIMS の測定位置に合わせて SEM イメージを切り出し，両イメージ画像のピクセル数を合わせた．各ピクセルにおける SEM イメージの強度を読みだし，TOF-SIMS データの二次イオン強度を示す変数に対して SEM イメージの強度を新たな変数として加えた．具体的には行（ピクセル数）×列（ピーク数）の TOF-SIMS の二次イオン強度のデータ行列の右横に，同じピクセル数だけある SEM 像のコントラストの数値の列ベクトルを追加し，その 1 変数追加された行列データで主成分分析を行った．データの例を Table 1 に示す．

TOF-SIMS データのみの場合と，SEM イメージとフュージョンした場合それぞれのデータを Matlab（Mathworks, MA, USA）上で作動する PLS Toolbox（Eigenvector Research Inc., WA, USA）を用いて主成分分析して，結果を比較した．

### 2.4 スパースモデリング

PET, PS, PC 三種類の高分子の 4 層試料[2]の TOF-SIMS 測定データに Least Absolute Shrinkage and Selection Operator（LASSO）[16,17]を適用した．Matlab の Statistics and Machine Learning Toolbox を用いた．ここで，行列 A は TOF-SIMS のスペクトル計測値であり，この行列の行数はピーク数（スペクトルチャネル数），列数は観測地点（ピクセル）数である．ベクトル $y$ は高分子由来二次イオン強度を表すベクトル，$x$ は高分子に起因する TOF-SIMS スペクトル波形を表すベクトルとすると次式が成り立つ．

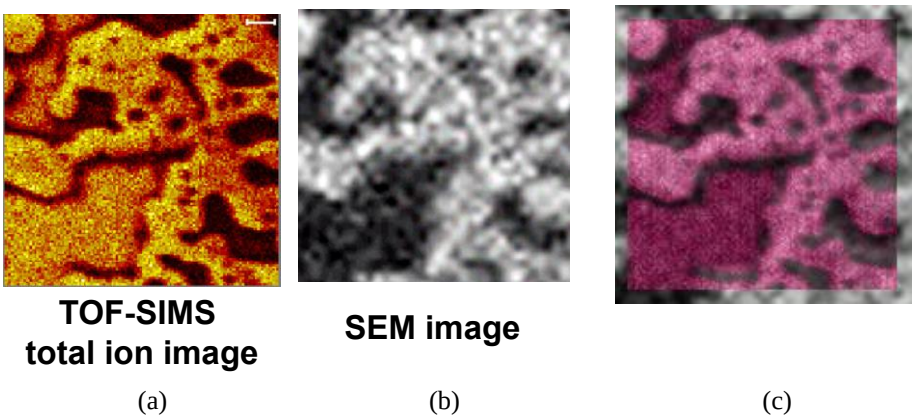



**Figure 1** Image data alignment: (a) TOF-SIMS, (b) SEM and (c) fusion image. (color online)

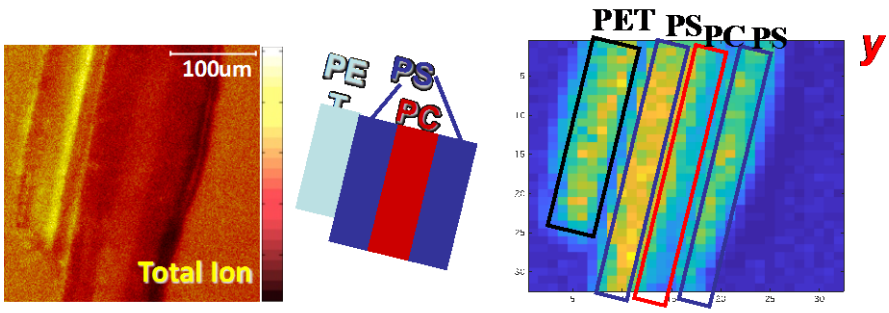



**Figure 2** TOF-SIMS total ion image (left), schematic of the sample (centre) and the integrated image of the secondary ions *m/z* 104, 91 and 135 (right). (color online)

$$y = Ax + \text{noise}$$

ノイズは無視できるとすると，$\|y - Ax\|^2$ は 0 に近づくはずであり，また $x$ がスパースであるとすると $|x|$も小さいはずである．そこで，次式の $E(x)$が最小となるように，$x$ を求めた．

$$E(x) = \|y - Ax\|^2 + \lambda \Sigma |x_i|$$

ただし，λは $L_1$ 正則化の寄与を決定するパラメーターである．高分子由来二次イオン強度を表すモデルとなる二次イオン $y$ は，高分子が存在する場所を示す 3 つのピーク（*m/z* 91, 104, 135）の和をしきい値処理して得られたベクトルとスペクトル行列 A との内積を計算することによって算出した．$y$ の算出後，LASSO によって，スペクトルと高分子由来二次イオン強度をスパース回帰することによって，高分子に起因する TOF-SIMS スペクトル波形を表すベクトル $x$ を学習した．Figure 2 に，高分子試料の TOF-SIMS データの Total ion 像と PET, PS, PC に由来する 3 つのピークが検出された位置を示す．

### 2.5 機械学習によるスペクトルデータの解析

k 最近傍法（k-Nearest Neighbor: kNN）と，決定木を利用する代表的な手法である Random Forest（RF）の二つの手法を用いて，TOF-SIMS スペクトルデータを学習し，未知と仮定したテストデータを正しく同定できるか評価した．各手法について簡単に説明する．

k 最近傍法（k-Nearest Neighbor: kNN）は特徴空間における最も近い訓練例に基づいた分類手法でパターン認識の一つであり，最近傍の鋳型を k 個とってきて，それらが最も多く所属するクラスに識別する方法である．k の値に識別性能が依存するので，適切な値を見つける必要がある．kNN は数あるアルゴリズムの中では計算が明確で簡単であるが，データが大きくなると計算時間が長くなるという特徴がある．

決定木とは説明変数と目的変数のデータから木の枝のような分類木を作成し，予測，分類を行うアルゴリズムである．中でも，RF は複数の決定木で集団学習させることで精度を高めた集団学習モデルである．RF はノイズに強く，データ量が多い場合でも高速で処理できる一方，学習データの説明変数をランダムに抽出するためデータと変数が少ないとうまく学習できないという特徴がある．

**Table 2** Training data example.

| Label | | | | | Descriptor | | | |
|---|---|---|---|---|---|---|---|---|
| A | B | C | D | … | 12 | … | 299.7 | 300.3 |
| 1 | 0 | 0 | 1 | 1 | 0.006 | … | 0.002 | 0.001 |
| 0 | 1 | 0 | 1 | 0 | 0.001 | … | 0.001 | 0 |
| 0 | 0 | 1 | 0 | 1 | 0.002 | … | 0.001 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

Label の化学構造の例

A: $C_2H_5^+$, B: $C_8H_7^+$, C: $C_7H_7O^+$, D: $C_7H_5O^+$, $C_8H_9^+$, E: $C_3H_5^+$, F: $C_{10}H_8^+$

使用したトレーニングデータは，試料の特徴を表す化学構造の有無を示すラベル部分と，スペクトルの二次イオンピーク強度（総二次イオンカウントで規格化した強度）を示す記述子のペアから成る．記述子（スペクトル）からラベル（化学構造の有無）を予測するモデルを構築するために，前述の 2 つの機械学習法を用いた．そして，学習されたモデルを用いて新しい観測位置に含まれる化学構造を予測した．

高分子試料については，各高分子が単独で存在する領域から関心領域（ROI: region of interest）を抽出した．高分子試料および PC 単膜試料データを全て同一のピークファイル（質量領域 *m/z* 12-300）を用いて各二次イオン強度を得たのち，各二次イオン強度を総二次イオン強度で規格化した．機械学習用のデータ数は全部で 367 で，内訳は PET 57, PS 111, PC 119（4 層高分子試料から 59，PC 膜試料から基板ごとに 20 ずつ）, PET，PS，PC を含む領域が 20 データ，PET と PS を含む領域が 20 データ，PET と PC を含む領域が 20 データ，PS+PC を含む領域が 20 データである．それぞれのデータごとにほぼ均等に 4 つに分け，そのうちの一つをテストデータ，残りを学習用データとして，交差検証した．Table 2 にトレーニングデータの例を示す．

学習手法として用いた kNN と RF のパラメータは，kNN は k=6，RF は決定木の数をデフォルトの 10，その他の値もデフォルトである．kNN で k=6 としたのは，最もトレーニングデータとして使われている数の少ない混合試料（PET+PS や PET+PC，PS+PC，PET+PS+PC）の数 15 を超えないようにするためであり，これを超えた場合，参照する近傍のトレーニングデータに別の種類の試料が含まれ，誤った結果につながるためである（しかし，k の値が小さすぎ

てもノイズだけを拾う危険性がある）．RF に関しては，パラメータを変えずに検討した．これらの学習の実行のために，Python の機械学習ライブラリー scikit-learn の kNeighborsClassifier と RandomForestClassifier を用いて，kNN，Random Forest の計算を行った．

ラベルは各高分子固有または共通のフラグメントイオンである．各サンプルがこれらのフラグメントイオンを含むかどうかを 1 と 0 の二値で表している．機械学習により，トレーニングデータを基にテストデータのスペクトルに対応するラベルの値を出力させた．この予測値と実際の値を比較し，正解率や精度を求めた．

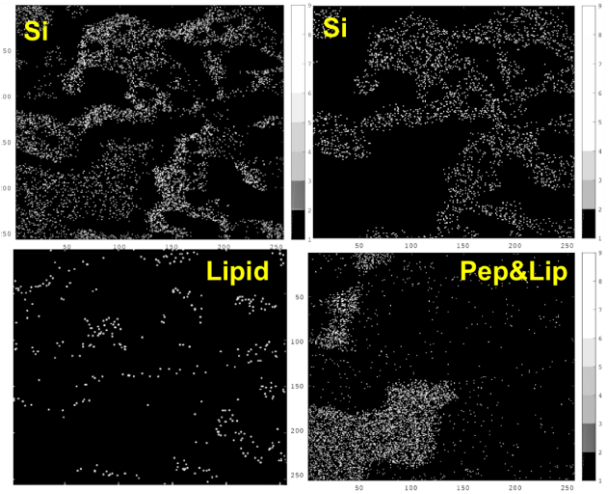
トレーニングデータは混合試料を含んだものと含んでいないものを用意し，結果を比較した．また，決定木では，ラベルのフラグメントイオンを各サンプルが含むかどうかを識別する際に，あるピークの強度情報を参照する．Random Forest ではこの決定木を複数用いており，それら決定木の多くで識別に用いられるピークを重要ピークと呼ぶ．Random Forest はこの重要ピークを抽出することができるため，適切なピークが選ばれているか（PS 固有ピークのラベルであれば，実際に PS 固有ピークが重要ピークとして選ばれているか）を確認した．

## ３．結果と考察

### 3.1 TOF-SIMS と SEM のデータフュージョン

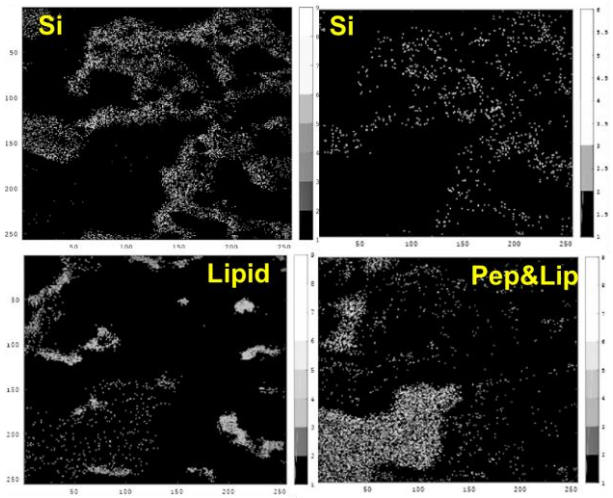
Figure 3 に TOF-SIMS データのみを主成分分析した結果の得点イメージと SEM データとフュージョンした後のデータを主成分分析した結果の得点イメージを示す．負荷量の結果と合わせて，主成分得点を解析すると，Figure 3(a)の PC1 の得点イメージでは，上段に示されている得点が正に大きい情報は，Si 基板と脂質分布の両方に対応しており，下段に示されている得点が負に大きい情報はおもに脂質に対応しているが Figure 4 に示す TOF-SIMS データの二次イオン像とあまり一致しない．一方で， Figure 3(b)に示すフュージョンしたデータの PC1 の正に大きい情報（上段）は適切に Si 基板の分布を示している．SEM データと融合することによって，正しい分布に補正された． Figure 3(b)の PC 1 で示される脂質（負方向）の得点分布イメージも，SEM 像の影響を受けて鮮明となっている．ただし，PC2 以下の主成分に関しては TOF-SIMS データとフュージョンデータから得られる情報に大きな差はなかった．

Figure 4 にもとの TOF-SIMS データとして，Si 基




PC1 PC2

(a) TOF-SIMS data




PC1 PC2

(b) SEM and TOF-SIMS fusion data

**Figure 3** PC score images of TOF-SIMS data (a), SEM and TOF-SIMS fusion data (b). (the upper part: positive direction, the lower part: negative direction).

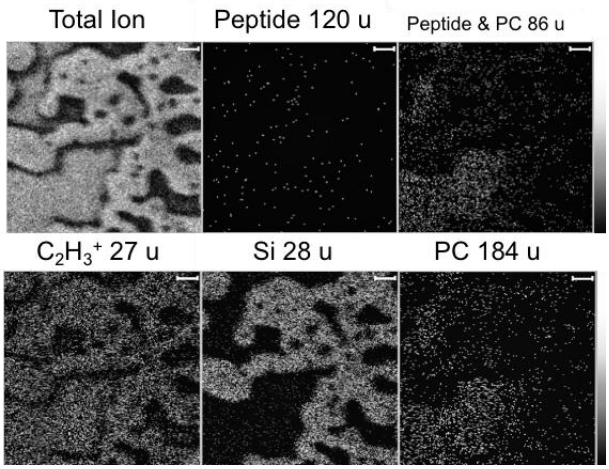
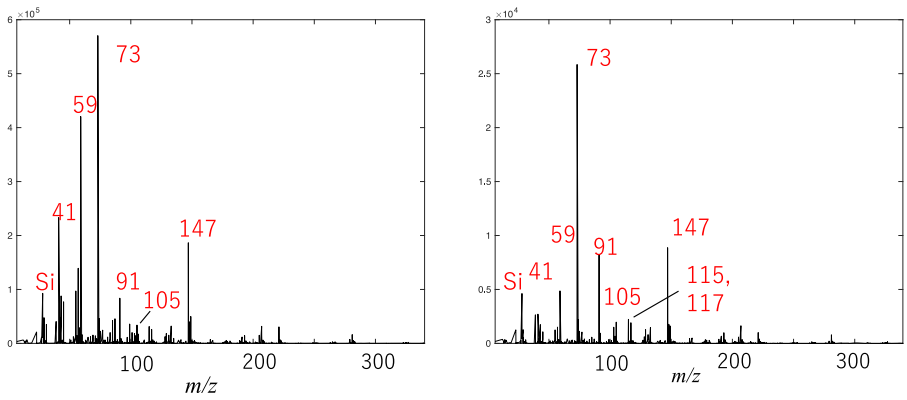



**Figure 4** TOF-SIMS secondary ion images.




(a) TOF-SIMS data　　(b) LASSO result

**Figure 5** TOF-SIMS spectrum analysis using LASSO.

板，ペプチドおよび脂質に関連する代表的な二次イオン像を示す．脂質のフォスフォコリン（PC）に由来する二次イオン 184 u，脂質とペプチド両方に由来する二次イオン 86 u およびペプチドに由来するアミノ酸フラグメントイオン 120 u の分布は Si 基板以外の領域に見られるが，二次イオン像だけでははっきりとは見られない．主成分分析によって，より明確な分布が見られるが，さらに SEM 像とのフュージョンによって，より明確になることが示された．なお，SEM 像とのフュージョンの有無による主成分負荷量の変化はほとんど見られなかったため，化学情報はほとんど変わらず，分布イメージのみ鮮明になった．

この手法では，TOF-SIMS 二次イオン像と SEM 像のおおよその位置を合わせたのち，画素数を同一として，同じピクセル上に TOF-SIMS で得られたスペクトル上の二次イオン強度情報（変数）に SEM 像のコントラストの数値情報を新たな変数として加えるだけで，イメージフュージョンができる． フュージョン前後の主成分分析結果を比較して，抽出された各主成分の得点分布図（イメージ図）の傾向と由来する物質を示す負荷量に大きな変化がなければ，適切なフュージョンが実施できたことがわかる．

### 3.2 スパースモデリング

Figure 2 に示した 3 種類の高分子の TOF-SIMS データを LASSO で解析した結果として，Figure 5 に示すように，元の TOF-SIMS スペクトル（Fig. 5(a)）で高強度に検出されていた *m/z* 73, 147（シロキサン由来）などの汚染ピークが相対的に抑えられ，高分子に由来する比較的高質量のフラグメントイオン，*m/z* 91（PS 由来），105（PS, PET 由来），115（PS 由来），117（PS 由来）が強調されたスペクトルが，LASSO の結果として得られた．Figure 2 右端の図で示すように，今回選んだ 3 つのピークでは，PS 部分の強度がもっとも強調されたため，LASSO 結果も PS 由来二次イオンがもっとも強調される結果となった．

このように，標的となる二次イオンの一部が分かっている場合や，試料に関する情報が一部でもあらかじめ得られている場合は，LASSO は TOF-SIMS スペクトルから汚染などの余分な情報を取り除き，解釈しやすくするために有効であることが示唆された．

### 3.3 機械学習によるスペクトルデータ解析

Table 3,4 に純粋試料のみのトレーニングデータを使用した結果，Table 5,6 に混合試料を加えたトレーニングデータを使用した結果を示す．表中の TPr, FNr, TNr, FPr はそれぞれ真陽性（true positive：TP）率= TP/(TP+FN)，偽陰性（false negative: FN）率= FN/(TP+FN)，真陰性（true negative: TN）率 = TN/(FP+TN)，偽陽性（false positive: FP）率 = FP/(FP+TN) であり [15]，Accuracy= (TP+TN)/(TP+FP+TN+FN)，Precision = TP/(TP+FP)である．

トレーニングデータに混合試料を加えることで kNN，Random Forest どちらの場合も Accuracy が向上した．このことから，正確な結果を得るには様々な条件での試料情報が必要であると示唆された．また，今回の検討では，Accuracy はどの場合でも 0.87 以上と比較的高かった．

**Table 3** Result using training data set with pure samples (k-NN (k=6))

| | TPr | FNr | TNr | FPr | Accuracy | Precision |
|---|---|---|---|---|---|---|
| A | 0.75 | 0.25 | 1 | 0 | 0.92 | 1 |
| B | 0.72 | 0.28 | 1 | 0 | 0.87 | 1 |
| C | 0.81 | 0.19 | 0.98 | 0.02 | 0.90 | 0.98 |
| D | 0.85 | 0.15 | 1 | 0 | 0.92 | 1 |
| E | 0.92 | 0.08 | 1 | 0 | 0.95 | 1 |
| F | 0.89 | 0.11 | 1 | 0 | 0.90 | 1 |

A: $C_2H_5^+$, B: $C_8H_7^+$, C: $C_7H_7O^+$, D: $C_7H_5O^+$, $C_8H_9^+$, E: $C_3H_5^+$, F: $C_{10}H_8^+$

**Table 4** Result using training data set with pure samples (RF)

| | TPr | FNr | TNr | FPr | Accuracy | Precision |
|---|---|---|---|---|---|---|
| A | 0.76 | 0.24 | 1 | 0 | 0.92 | 1 |
| B | 0.73 | 0.27 | 1 | 0 | 0.87 | 1 |
| C | 0.67 | 0.32 | 1 | 0 | 0.84 | 1 |
| D | 0.95 | 0.05 | 0.98 | 0.02 | 0.96 | 0.99 |
| E | 0.90 | 0.10 | 1 | 0 | 0.93 | 1 |
| F | 0.87 | 0.13 | 1 | 0 | 0.90 | 1 |

**Table 5** Results using mixed sample-included training data (k-NN (k=6))

| | TPr | FNr | TNr | FPr | Accuracy | Precision |
|---|---|---|---|---|---|---|
| A | 1 | 0 | 1 | 0 | 1 | 1 |
| B | 0.99 | 0.01 | 0.99 | 0.01 | 0.99 | 0.99 |
| C | 0.97 | 0.03 | 0.99 | 0.01 | 0.98 | 0.99 |
| D | 1 | 0 | 1 | 0 | 1 | 1 |
| E | 0.99 | 0.01 | 0.99 | 0.01 | 0.99 | 1 |
| F | 1 | 0 | 1 | 0 | 1 | 1 |

**Table 6** Result using mixed sample-included training data (RF)

| | TPr | FNr | TNr | FPr | Accuracy | Precision |
|---|---|---|---|---|---|---|
| A | 0.98 | 0.02 | 1 | 0 | 0.99 | 1 |
| B | 0.96 | 0.04 | 0.98 | 0.02 | 0.97 | 0.98 |
| C | 0.97 | 0.03 | 1 | 0 | 0.99 | 1 |
| D | 1 | 0 | 0.99 | 0.01 | 1 | 1 |
| E | 0.98 | 0.02 | 1 | 0 | 0.99 | 1 |
| F | 1 | 0 | 1 | 0 | 1 | 1 |

Table 7 にテストデータの各種試料ごとにラベルがどれだけ正しく出力されたかの確率を示す．kNN, Random Forest どちらも純粋試料のみのトレーニングデータを使用した場合は，テストデータの混合試料についてはラベルが完全に一致するものはなかった．トレーニングデータに混合試料を加えることで改善した．

ランダムフォレストは，トレーニングデータのランダム分割および決定木の学習を繰り返し，学習された多数の決定木の多数決によって予測するモデルである．個々の決定木の学習の際にスペクトルピークのモデル構築への寄与度によって，各ピークの重要度が数値化される．この重要度の指標によって，適切なピークが選ばれているか（PS 固有ピークのラベルであれば，実際に PS 固有ピークが重要ピークとして選ばれているか）を確認した．

Table 8 に PS 固有のフラグメントイオン *m/z* 103（$C_8H_7^+$）のラベルに対して，Random Forest で選ばれた重要ピークの一部の表を示す．左の表はデータセット 4 つを 3 回解析し，選ばれた回数が上位のピークを並べたもの，右の表はある 1 回の解析で重要度が高かったピークを並べたものである．

データベース上の PS 由来ピークの中で Table 8 の左の表にあるものは 91.3（$C_7H_7^+$），105.3（$C_8H_9^+$），115.3（$C_9H_7^+$），193.3（$C_{15}H_{13}^+$），103.3（$C_8H_7^+$）で，右の表にあるものは 103.3（$C_8H_7^+$），105.3（$C_8H_9^+$），141.3（$C_{11}H_9^+$）である．2 つの表のどちらも PS 由来のフラグメントイオンが含まれていたが，それ以外のピークも含まれていた．できる限り適切なピークだけを抽出するためには基板や試料作製方法などが異なる様々な条件の試料が必要になると考えられる．

**Table 7** Each polymer sample's result

| kNN | a | b | RF | a | b |
|---|---|---|---|---|---|
| PET | 1 | 1 | PET | 1 | 1 |
| PS | 1 | 1 | PS | 1 | 1 |
| PC | 1 | 1 | PC | 0.9664 | 0.975 |
| PET+PS | 0 | 0.95 | PET+PS | 0 | 0.85 |
| PET+PC | 0 | 0.9 | PET+PC | 0 | 0.8 |
| PS+PC | 0 | 0.85 | PS+PC | 0 | 0.75 |
| PET+PS+PC | 0 | 0.85 | PET+PS+PC | 0 | 0.75 |

a: result using training data set with pure samples, b: result using mixed sample-included training data

**Table 8** Important peaks for the label $C_8H_7^+$ in Random Forest.

| *m/z* | selected numbers | *m/z* | importance |
|---|---|---|---|
| 91.3 | 10 | 181.3 | 0.0673 |
| 105.3 | 9 | 116.3 | 0.0666 |
| 167.3 | 9 | 103.3 | 0.0666 |
| 129.3 | 9 | 105.3 | 0.0652 |
| 115.3 | 8 | 118.3 | 0.064 |
| 92.3 | 7 | 227.7 | 0.0617 |
| 16.3 | 7 | 119.3 | 0.0604 |
| 135.3 | 7 | 192.3 | 0.06 |
| 193.3 | 7 | 29.3 | 0.0544 |
| 103.3 | 7 | 229.3 | 0.05 |
| 130.3 | 7 | 141.3 | 0.0462 |
| 107.3 | 7 | 167.3 | 0.0452 |

## 4. 結論

SEM 像とのイメージフュージョンによって，TOF-SIMS イメージデータの解像度を高めることができた．また，もとの TOF-SIMS データとイメージフュージョンしたデータの主成分分析結果を比較することにより，化学情報がイメージフュージョンによって変化していないことが確認できる．

測定対象物質が分かっている場合や，試料に関する分布情報などがある場合は，スパースモデリングによって，ノイズを抑えて，TOF-SIMS スペクトルを単純化し，重要な二次イオンピークを強調できることが示された．

さらに，Random Forest や kNN などの機械学習手法を用いることにより，未知試料（テストデータ）の同定が可能となることが示唆された．ただし，テストデータと類似する試料データを含む学習データが必要である．

## 5. 謝辞

本研究において，polycarbonate（PC）単膜試料をご提供いただいた（社）研究産業・産業技術振興協会（JRIA）に深く感謝いたします．

## 6. 参考文献

[ 1] I. S. Gilmore, *J. Vac. Sci. Technol. A* **31**, 050819 (2013).
[ 2] Y. Yokoyama, T. Kawashima, M. Ohkawa, H. Iwai and S. Aoyagi, *Surf. Interface Anal.*, **47**, 439 (2015).
[ 3] K. Takahashi, S Aoyagi and T Kawashima, *Surf. Interface Anal.*, **49**, 721 (2017).
[ 4] A. G. Shard, et al., *J. Phys. Chem B*, **119**, 10784 (2015).
[ 5] J. L. S. Lee, I. S. Gilmore, I. W. Fletcher and M. P. Seah, *Surf. Interface Anal.*, **41**, 653 (2009).
[ 6] J. C. Vickerman and I.S. Gilmore (eds.), *Surface Analysis -The Principle Techniques*, John Wiley & Sons, (2009).
[ 7] 青柳里果，工藤正博，*現代表面科学シリーズ第 2 巻表面科学基礎*，日本表面科学会，pp.10-12, 共立出版 (2013).
[ 8] J. L. S. Lee, S. Ninomiya, J. Matsuo, I. S. Gilmore, M. P. Seah, and A. G. Shard, *Anal. Chem.*, **82**, 98 (2010).
[ 9] M. K. Passarelli, A. Pirkl, R. Moellers, D. Grinfeld, F. Kollmer, R. Havelund, C. F. Newman, P. S. Marshall, H. Arlinghaus, M. R. Alexander, A. West, S. Horning, E. Niehuis, A. Makarov, C. T Dollery and I.S Gilmore, *Nature Methods*, **14**, 1175 (2017).
[10] J. S. Fletcher, *Analyst*, **134**, 2204 (2009).
[11] “Using large-scale brain simulations for machine learning and A.I.”, June 26, 2012, Google Official Brog (https://googleblog.blogspot.jp/2012/06/using-large-scale-brain-simulations-for.html).
[12] M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, T. Mori, and T. Tanji, *Ultramicroscopy*, **170**, 43 (2016).
[13] H. M. Rostam1, P. M. Reynolds, M. R. Alexander, N. Gadegaard and A. M. Ghaemmaghami, *Sci. Rep.* **7**, 3521 (2017).
[14] S. A. Thomas, Y. Jin, J. Bunch and I. S. Gilmore, *2017 IEEE Symposium Series on Computational Intelligence* (DOI 10.1109/SSCI.2017.8285223).
[15] 平井有三，*はじめてのパターン認識*，pp. 30-31, 58-60 and 193-197，森北出版(2012).
[16] I. Rish and G. Grabarnik, *Sparse Modeling*, CRC Press (2014).
[17] 廣瀬慧，*電子情報通信学会誌*, **99**, 392 (2016).
[18] K. Takahashi, T. Yamagishi, S. Aoyagi, D. Aoki, K. Fukushima and Y. Kimura, *J. Vac Sci. Technol. B* **36**, 03F113 (2018).

[19] M. Hubert, P. J. Rousseeuw and K. V. Branden, *Technometrics*, **47**, 64 (2005).
[20] E. J. Candès, X. Li, Y. Ma and J. Wright, *J. Assoc. Compt. Math.*, **58**, 11 (2011).
[21] S. Nakano, T. Yamagishi, S. Aoyagi, A. Porty, M. Dürr, H. Iwai and T. Kawashima, *Biointerphases*, **13**, 03B403 (2018).

## 査読コメント，質疑応答

### 査読者１ 吉原 一紘（シエンタオミクロン）

信号処理の新しい技術を TOF-SIMS に適用した結果の報告であり，JSA の読者にとって重要な情報を提供している論文です．掲載を薦めます．しかし，JSA の読者は必ずしも信号処理の専門家ではないため，この論文だけでは内容を理解することは困難です．この論文は解説論文ではなく研究論文ですので，論文中に基礎的な解説をする必要はありませんが，TOF-SIMS の利用者に本論文が広く引用されるためにも，以下に述べる質問に対応できる程度の説明を付け加えていただけないでしょうか．

本論文で用いられたデータ処理技術は表面分析の分野で広く利用される可能性がありますので，本論文の貴重な情報を表面分析研究者や技術者に周知することが重要です．ここで述べたコメントは信号処理の素人の観点からのコメントですが，このコメントを参考にして信号処理の専門家ではない読者にも理解できる内容にしていただければ幸いです．

**[査読者 1-1]**

TOF-SIMS と SEM のデータフュージョン

「TOF-SIMS で得られたスペクトル上の二次イオン強度情報（変数）に SEM 像のコントラストの数値情報を新たな変数として加える」と記述されていますが，新たな変数として加えるというのはどのようなデータ行列を作成する作業なのでしょうか．そして，得られるデータ行列の各要素はどのように記述されるのでしょうか．

**[著者]**

適切なコメントをいただきありがとうございます．以下に回答いたします．また，ご指摘内容を受けて本文を修正いたしました．

TOF-SIMS は各ピクセルにおける各二次イオン強度として，行（ピクセル数）×列（ピーク数）の行列データに変換されます．データフュージョンに際しては，SEM データの解像度と TOF-SIMS データの解像度（ピクセル数）を合わせます．今回のデータでは，SEM 像の解像度を TOF-SIMS に合わせました．SEM 像は白黒のコントラストだけですので，1 変数だけですが，それを新たな変数として TOF-SIMS データの変数（ピーク）に追加することによって，フュージョンデータとしています．

**[査読者 1-2]**

スパースモデリング

「試料からの求める信号を x， TOF-SIMS による変換を行列 A とし，モデルとなる二次イオンデータを y とすると」と記述されていますが，試料からの求める信号[x]とは何を示すのでしょうか．また，行列[A]の各要素はどのように記述するのでしょうか．スパースモデリングは[x]ベクトルの要素がスパースであることを前提として，Lagrange の未定係数法から[x]を求める方法だと単純に理解していました．しかし，最終結果では，「元の TOF-SIMS データで高強度に検出されていたシロキサンなどに由来する汚染ピークが抑えられ，高分子に由来するフラグメントイオンが強調されたスペクトルが得られる．」とあり，汚染ピークが除去されたスペクトルが Figure 5 に示されていますが，その導出課程が明らかではありません．Figure 5 を求める過程を示していただきたく存じます．

**[著者]**

ベクトル x の L1 ノルム©$|x_i|$の大きさに関する制約付き最適問題を，ラグランジュの未定乗数法（未定係数法）によって目的関数 $E(x) = \|y\text{-}Ax\|^2 + \lambda\Sigma|x_i|$ に変換して，これを最小化する解 x を計算しております．これは吉原先生のご理解で良いと思います．ただし，x の最適値を解析的に導出できないので，勾配法によって逐次最適化しております．λ の値がある程度大きければ，最適な x はスパースになる（ほとんどの値がゼロになる）が一般的に示されているため，ここではスパースモデリングと呼んでいます．

ここで，行列 A は TOF-SIMS のスペクトル計測値であり，この行列の行数はピーク数（スペクトルチャネル数），列数は観測地点数です．ベクトル y は高分子の濃度を表すベクトルです．ベクトル y は，高分子が存在する場所を示す 3 つのピーク（m/z 91, 104, 135）の和をしきい値処理して得られるベクトルとスペクトル行列 A との内積を計算することによって算出しました．行列 A とベクトル y は固定され，データから最適化されません．一方，ベクトル x は高分子に起因する TOF-SIMS スペクトル波形を表すベクトルであり，データから学習されます．y の算出後，Lasso によって，スペクトルと高分子濃度をスパース回帰することによって，高分子に起因する TOF-SIMS スペクトル波形を表すベクトル x が得られます．

上記の内容は，本文のスパースモデリングの説明に加えました．また，説明するために，再度計算を行い，モデルも単純化したところ，最初に示した LASSO 結果と結果（Figure 5）が異なりました．こちらの方が一般的に得られる解に近いものとなったと思いますので，差し替えました．

**[査読者 1-3]**

機械学習によるスペクトルデータの解析

機械学習でトレーニングデータとして入力する情報は，Table 2 にある Label と Descriptor（規格化されたスペクトル強度のことですか）でしょうか．「機械学習により，トレーニングデータを基にテストデータのスペクトルに対応するラベルの値を出力させた．この予測値と実際の値を比較し，正解率や精度を求めた．」とありますが，k 最近傍法や決定木はクラス判別に使用する方法と理解していますが，これらクラス判別法がどのように使用されたかが分かりませんでした．Table 3~6 はラベルの検索性能の評価指標を示していると思われますが，情報技術の専門家以外には内容を理解することが困難です．分析技術者にとっては，未知試料のラベルが出力できたか否かを示す結果が表示されていれば分かりやすいと思いますので，ご一考いただければ幸いです．Table 8 に固有のフラグメントイオンのラベルに対して，ランダムフォレストで選ばれた重要ピークの一部の表が示されていますが，重要ピークとはランダムフォレストにより選ばれた決定木のことでしょうか．また，Table 8 はこれらピークのどれが判別に重要かということを示しているのでしょうか．「機械学習によるスペクトルデータの解析」の項の記述内容には情報技術の専門家以外には難解な表現が含まれていると思われますので，分析技術者に分かるような表現にされることを希望します．

**[著者]**

トレーニングデータはラベルと記述子で構成されています．具体的に，トレーニングデータの 1 標本は，規格化された TOF-SIM スペクトル（1 本）からなる記述子と試料の特徴を表す構造（分子）の有無を示す数値ベクトル（バイナリ要素のベクトル）からなるラベルのペアからなります．ここでは，記述子（スペクトル）からラベル（構造の有無）を予測するモデルを構築するために機械学習法を用い，学習されたモデルを用いて新しい観測位置に含まれる

構造を予測しています．「TOF-SIMS スペクトルデータは，試料の特徴を表す構造の有無を示すラベル部分と，スペクトルの二次イオンピーク強度（総二次イオンカウントで規格化）を示す記述子から成る．」と記述していましたが，以下のように訂正しました．

「使用したトレーニングデータは，試料の特徴を表す構造の有無を示すラベル部分と，スペクトルの二次イオンピーク強度（総二次イオンカウントで規格化した強度）を示す記述子のペアから成る．記述子（スペクトル）からラベル（構造の有無）を予測するモデルを構築するために，前述の２つの機械学習法を用いた．そして，学習されたモデルを用いて新しい観測位置に含まれる構造を予測した．」

k 最近傍法や決定木をどのように使用したかについてですが，パラメータに関しては，k 最近傍法は k=6，ランダムフォレストでは決定木の数を 10 としました．k 最近傍法で k=6 とした理由は，最もトレーニングデータとして使われている数の少ない混合試料（PET+PS や PET+PC，PS+PC，PET+PS+PC）の数 15 を超えないようにするためです．これを超えた場合，参照する近傍のトレーニングデータに別の種類の試料が含まれ，誤った結果につながります（k の値が小さすぎてもノイズだけを拾ってしまう可能性があり，よくないです）．ランダムフォレストでは，パラメータの値を変えずに様子を見ました．

また，機械学習の実行には Python の機械学習ライブラリ scikit-learn にあるクラス kNeighborsClassifier と RandomForestClassifier を用いて，kNN, Random Forest の計算を行いました．下記を本文に加えました．

「学習手法として用いた kNN と RF のパラメータは，kNN は k=6，RF は決定木の数をデフォルトの 10，その他の値もデフォルトである．kNN で k=6 としたのは，最もトレーニングデータとして使われている数の少ない混合試料（PET+PS や PET+PC，PS+PC，PET+PS+PC）の数 15 を超えないようにするためであり，これを超えた場合，参照する近傍のトレーニングデータに別の種類の試料が含まれ，誤った結果につながるためである（しかし，k の値が小さすぎてもノイズだけを拾う危険性がある）．RF に関しては，パラメータを変えずに検討した．これらの学習の実行のために，Python の機械学習ライブラリー scikit-learn の kNeighborsClassifier と RandomForestClassifier を用いて，kNN, Random Forest の計算を行った．」

真陽性率（true positive rate），偽陰性率（false negative rate），真陰性率（true negative rate），偽陽性率（false positive rate）については，求め方の式も表示しました．

未知試料のラベルが出力できたか否かを示す結果については，ご指摘の通り掲載すべきと考えましたので，Table 8 にその結果を，また以下の説明を加えました．「Table 8 にテストデータの各種試料ごとにラベルがどれだけ正しく出力されたかの確率を示す．kNN，Random Fores どちらも純粋試料のみのトレーニングデータを使用した場合は，テストデータの混合試料についてはラベルが完全に一致するものはなかった．トレーニングデータに混合試料を加えることで改善した．」

ランダムフォレストの重要ピークについてですが，ご指摘の通り説明不足であると考え，以下のように修正・追記いたしました．「ランダムフォレストは，トレーニングデータのランダム分割および決定木の学習を繰り返し，学習された多数の決定木の多数決によって予測するモデルである．個々の決定木の学習の際にスペクトルピークのモデル構築への寄与度によって，各ピークの重要度が数値化される．この重要度の指標によって，適切なピークが選ばれているか（PS 固有ピークのラベルであれば，実際に PS 固有ピークが重要ピークとして選ばれているか）を確認した．」

**査読者２ 村瀬 篤（豊田中央研究所）**

本論文は、データフュージョンなどデータ解析手法を用いることによって TOF-SIMS の空間分解能の限界を超えようとしたものであり、その試み自体は論文としての価値はあるものと考えます。ただし、以下の点で不備があり、掲載には改訂が必要と考えます。

**[査読者 2-1]**

最も重要な結論である「イメージデータの解像度を高めることができた」の根拠となる Fig. 2 ですが，私の眼には SEM 像とのフュージョンによって解像度が上がっているようにはどうしても見えません．掲載した図が問題であるならば，もっと明瞭にわかる図に差し替えるか，またはラインスキャンなどによって解像度を数値化する必要があると思います．

**[著者]**

適切なコメントをいただきありがとうございます．以下に回答いたします．また，ご指摘内容を受けて本文を修正いたしました．

グレイスケールとし，鮮明な図に修正しました．参考文献に載せた JVST B の論文では，ラインスキャン表示もしましたが，今回の分布はラインスキャンでの表示が難しいため，鮮明な図として，差を明確にしました．また，フュージョンによって分布図が明確になったのは PC1 のみで他の主成分では大きな差がないことも本文に記載しました．

**[査読者 2-2]**

Figs. 2，3 はカラーでは何とか見えますが，モノクロにした場合には全く識別できません．モノクロでも理解できる程度の画像にすべきと考えます．

**[著者]**

前述のようにグレイスケールでの表示に修正しました．

**[査読者 2-3]**

P.108 の下の方の Fig.5 に関する記述で「シロキサンなど」とありますが，Fig.5 を見る限り，シロキサンもあるかも知れませんがそれよりも含 O フラグメントが高強度で見られます．表現を変えた方がよいと考えます．

**[著者]**

ご指摘ありがとうございます．表現を変えました．