G-020

深層学習を用いた蛍光指紋分解による蛍光指紋解析手法の検討
Excitation-Emission Matrix Decomposition by Deep Learning

林田 純弥‡　柿下 容弓‡　長坂 晃朗‡
Junya Hayashida　Yasuki Kakishita　Akio Nagasaka

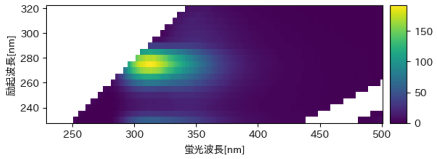

図 1　蛍光指紋の可視化例

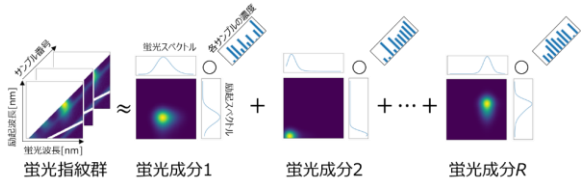

図 2　tri-linear モデル

1. はじめに

蛍光指紋 (Excitation Emission Matrix; EEM)とは、試料に励起光を照射した際の蛍光(試料が励起光を吸収し、試料が放出する光のこと)を分光スペクトルとして測定したデータであり、励起波長・蛍光(放出)波長の 2 つの波長軸を持った行列である。図 1 は蛍光指紋をヒートマップ形式で可視化した例であり、励起波長、蛍光波長の 2 軸に対応する光の強度を色で可視化している。図 1 において、ある励起波長における蛍光波長方向の信号は蛍光スペクトル、ある蛍光波長における励起波長方向の信号は励起スペクトルと称される。蛍光指紋は試料の蛍光特性を反映した信号情報であり、化学・医薬品、食品、工業製品等、様々な分野で、蛍光特性に関する試料の解析に用いられている。

蛍光指紋は、試料に含まれる各蛍光物質が放出するユニークな信号成分(蛍光成分)の和で観測される。蛍光指紋分解は、この性質に基づいて、どのような蛍光物質が含まれているのかを観察するための手法である。蛍光指紋分解によって、蛍光指紋を蛍光成分(蛍光物質が放出するユニークなスペクトル)と濃度(蛍光物質の混合比率)の情報に分解する。得られた各蛍光成分の形状を文献データと照合することで、試料にどのような蛍光物質が含まれているのかを推察することが可能となる。

蛍光指紋分解は PARAFAC (PARArell FACtor analysis)[1]と呼ばれるテンソル分解手法がデファクトスタンダードとなっている。PARAFAC は蛍光指紋の特性に基づいた分解として確立した手法であるが、①ノイズに対する頑健性が低く、②同じ蛍光物質を含む複数の試料を収集、蛍光指紋を測定、最適化する必要があり、③蛍光成分の数を仮定する必要がある。これらの課題に対して、本報告では、①ノイズに対する頑健性を有しつつ、②最適化を経ずに単独の蛍光指紋(一つの試料から得られる蛍光指紋)を、③蛍光成分の数も含めて end-to-end で分解する深層学習モデル SEEMD-Net (Single EEM Decomposition Network)を提案する。この深層学習モデルの学習は、大量の蛍光指紋分解パターンを教師情報として実現するが、試料の準備から測定までの作業コストを鑑みると、そのようなデータセットを準備することは困難である。そこで本稿では、蛍光指紋の特性をシミュレートした疑似的な蛍光指紋[2]を用いることで実現する。蛍光を放出する化合物の蛍光指紋を対象に、本提案手法を用いた蛍光指紋分解の有効性を検討した。

2. 従来手法: PARAFAC

まず、蛍光指紋分解手法として確立している PARAFAC について説明する。PARAFAC ではテンソル分解によって、複数の蛍光指紋をR組の(励起スペクトル、蛍光スペクトル、濃度)に分解する手法である。ここでRは対象とする試料群に含まれる蛍光成分の数であり、観測者が設定するハイパーパラメータである。励起・蛍光スペクトルは各蛍光指紋に共通して含まれる蛍光物質が放出する蛍光を表す情報であり、蛍光指紋の蛍光波長数、励起波長数を要素数とするベクトルである。濃度は蛍光物質が各蛍光指紋にどの程度含まれているのかを示す重み情報であり、PARAFAC に用いた蛍光指紋数を要素数とするベクトルである。

PARAFAC は式 1 に示すK個の蛍光指紋群$\mathbf{X}$を tri-linear モデルで近似するようにR組の(励起スペクトル、蛍光スペクトル、濃度)を最適化することで蛍光指紋分解を行う。式 1 において、$\mathbf{X}$は励起波長、蛍光波長、サンプル番号の 3 軸で構成される 3 階テンソルであり、$\boldsymbol{a}_i$、$\boldsymbol{b}_i$、$\boldsymbol{c}_i$はそれぞれi番目の蛍光物質の励起スペクトルと蛍光スペクトルと濃度を意味し、$\circ$は直積操作を意味する。図 2 に tri-linear モデルを表す概念図を示す。図 2 に示すように、各蛍光指紋は、蛍光物質毎にユニークな励起スペクトルと蛍光スペクトルと、サンプル毎の濃度の直積で表されるテンソルの和として表現される。

$$\mathbf{X} = \sum_{i=1}^{R} \boldsymbol{a}_i \circ \boldsymbol{b}_i \circ \boldsymbol{c}_i \qquad (1)$$

3. 提案手法: SEEMD-Net

深層学習モデルを用いた蛍光指紋分解手法を新たに提案する。蛍光指紋分解は PARAFAC がデファクトスタンダードとなっているものの、いくつかの課題が存在する。

①　ノイズに対する頑健性が低い

式 1 において左辺(複数の蛍光指紋で定義される 3 階テンソル)と右辺(分解結果)の誤差を最小にするように最小二乗法等を用いて最適化することで各分解結果が求まる。この際、複数の蛍光指紋に様々な蛍光以外の成分(ノイズ成分)も含まれるが、ノイズ成分は tri-linear モデルに含まれないため、ノイズ成分が大きいと正しい分解が難しいという課題がある。

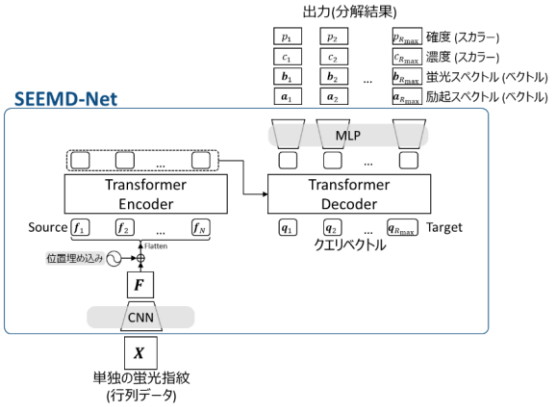

図 3　SEEMD-Net のネットワーク図

② **複数の蛍光指紋を用いた最適化を行う必要がある**

PARAFAC は複数の蛍光指紋を用いた最適化によって実行されるため、蛍光指紋分解実行の際に、複数の蛍光指紋を収集する必要がある。また、収集した蛍光指紋の一部にのみ存在する蛍光成分がある場合、その蛍光成分が正しく分解(抽出)できない。

③ **蛍光成分の数を仮定する必要がある**

式 1 にあるように、PARAFAC は蛍光成分数 R を仮定する必要がある。しかしこれは専門家の目視による判断となるため、専門家負担が増加するとともに、恣意性も発生する。

これらの課題に対し、単独の蛍光指紋を蛍光成分の数も含めて分解結果を最適化を経ずに end-to-end で予測する深層学習モデルとして、図 3 に示す SEEMD-Net (Single EEM Decomposition Network) を提案する。SEEMD-Net の大きな特徴は PARAFAC のような最適化手法(テンソル分解)によって蛍光指紋分解を行うのではなく、深層学習モデルの出力として分解結果を予測する点にある。具体的には、単独の蛍光指紋を入力すると、SEEMD-Net は$R_{\max}$組の出力を行う。蛍光指紋にR個の蛍光成分が含まれる場合、$R_{\max}$個の出力の内、R個は分解結果、$R_{\max}-R$個の出力はダミーの出力となっている。ただし$R_{\max} \gg R$である。各出力には分解結果かダミーかを示す指標である確度が備わっており、確度が高い出力を分解結果とすることで、end-to-end で蛍光成分数を含めて蛍光指紋分解を行う。このアプローチは物体検出モデル DETR (DEtection TRansformer)[3]のアルゴリズムを蛍光指紋分解に応用した方式であり、学習も DETR と類似した教師有学習によって行う。SEEMD-Net の教師有学習には、tri-linear モデルに基づく疑似的な蛍光指紋を用いる。疑似蛍光指紋は疑似的に蛍光成分を定義し、ノイズを加えて蛍光指紋の形状に構成することで作成するため、定義した疑似的な蛍光成分がそのまま教師情報となり、SEEMD-Net は教師有学習によって学習が可能となる。

3.1 疑似蛍光指紋を用いた学習データの作成

式 1 に示す tri-linear モデルのパラメタに様々なパターンを当てはめることで、蛍光指紋の特性をシミュレートした疑似蛍光指紋[2]を作成し、SEEMD-Net の教師データとする。蛍光成分の数Rと、対応する各濃度$\boldsymbol{c}_i$はランダムな値を当てはめる。しかし、励起・蛍光スペクトル$\boldsymbol{a}_i$、$\boldsymbol{b}_i$に関しては、滑らかな分光スペクトルであり、形状も様々であるため、ランダムな定義が難しい。ガウス過程のような生成モデルで波形を生成することも考えられるが、実際の蛍光指紋と形状が乖離する可能性があることから、実際の励起・蛍光スペクトル情報を使用する。具体的には、オープンデータセット PubSpectra[4]に含まれる蛍光色素の励起・蛍光スペクトルを用いた。PubSpectra には励起・蛍光スペクトル 267 組が含まれている。$\boldsymbol{a}_i$、$\boldsymbol{b}_i$を励起・蛍光方向にランダムにシフトさせつつ、濃度$\boldsymbol{c}_i$をランダムに定義することで大量の疑似蛍光指紋を生成しつつ、ノイズへの頑健性を獲得するために、ガウシアンノイズを付与した。

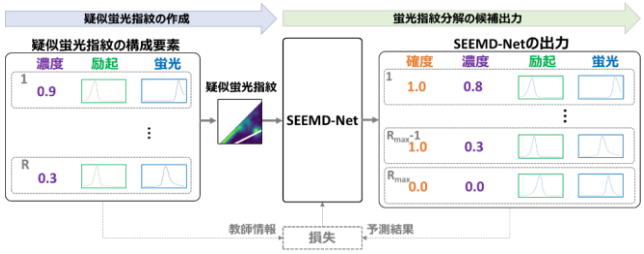

図 4　疑似蛍光指紋を用いた SEEMD-Net の学習

本来、実際の蛍光指紋はどのような蛍光成分が含まれているのかを利用することはできないが、疑似蛍光指紋によって深層学習モデルに大量の分解パターンの学習を可能とし、end-to-end で分解を行うモデルが構築可能となる。

3.2 SEEMD-Net の学習

SEEMD-Net の教師有学習のプロセスを図 4 に示す。疑似蛍光指紋を作成、SEEMD-Net に入力し、出力した結果と疑似蛍光指紋生成に用いた疑似的な蛍光成分情報を用いて損失を計算し、誤差逆伝播法によって SEEMD-Net を学習する。$R_{\max}$組の各出力は確度、濃度、励起スペクトル、蛍光スペクトルの 4 つが含まれる。確度は前述の通り、蛍光成分の出力かダミーの出力かを示すスカラー値である。

教師有学習のプロセスにおいて、誤差逆伝播に用いる損失については DETR と類似する方式で計算する。教師情報と出力を対応付けし、対応付けられた出力に蛍光成分用の損失、対応付けられなかった出力にダミー用の損失を計算する。対応付けはR個の教師情報と$R_{\max}$個の出力の 2 部マッチングによって行う。

i番目の教師情報をy_i、j番目の出力を$\hat{y}_j$とし、y_iに含まれる濃度をc_i、励起スペクトルを$\boldsymbol{a}_i$、蛍光スペクトルを$\boldsymbol{b}_i$、$\hat{y}_j$に含まれる確度を$\hat{p}_j$、濃度を$\hat{c}_j$、励起スペクトルを$\hat{\boldsymbol{a}}_j$、蛍光スペクトルを$\hat{\boldsymbol{b}}_j$と表記した場合、y_iとy_jの間の 2 部マッチングのコストは式 2 のように定義する。

$$C_{ij} = -\hat{p}_j - \mathcal{L}_{cos}(\boldsymbol{a}_i, \hat{\boldsymbol{a}}_j) - \mathcal{L}_{cos}(\boldsymbol{b}_i, \hat{\boldsymbol{b}}_j) \qquad (2)$$

式 2 において、$\mathcal{L}_{cos}$はコサイン類似度を表し、教師情報と出力の励起スペクトル、蛍光スペクトルの類似度を計算している。式 2 は確度が高く、励起スペクトルと蛍光スペクトルの類似度が高いときにコストが低くなるように設計している。教師情報y_1、y_2、…、y_Rおよび出力$\hat{y}_1$、$\hat{y}_2$、…、$\hat{y}_{R_{\max}}$の全組み合わせのコストからなるコスト行列を計算し、コスト行列に基づいて教師情報と出力の対応付けを行う。

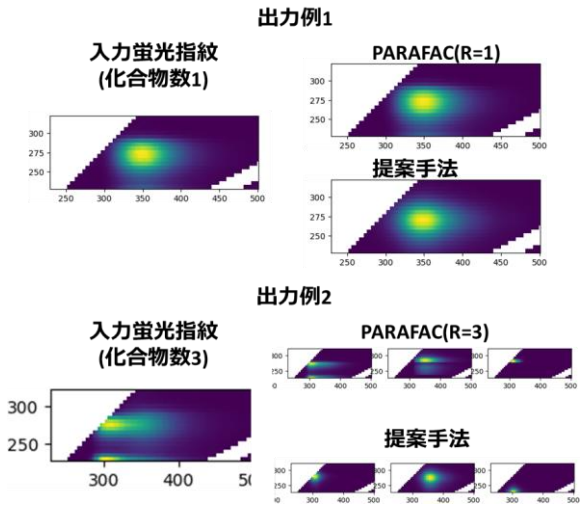

図 5　提案手法と PARAFAC の比較

2 部マッチングの結果、教師情報y_iと対応付けられた出力を$\hat{y}_{\sigma(i)}$ とした場合、対応付けられた出力に対する損失$\mathcal{L}_{\text{match}}(y_i、\ \hat{y}_{\sigma(i)})$を式 3 に示す。

$$\mathcal{L}_{match}(y_i、\ \hat{y}_{\sigma(i)}) = \mathcal{L}_{bce}(1、\ \hat{p}_{\sigma(i)}) + \mathcal{L}_{mse}(c_i、\ \hat{c}_{\sigma(i)}) + \mathcal{L}_{cos}(\boldsymbol{a}_i、\ \hat{\boldsymbol{a}}_{\sigma(i)}) + \mathcal{L}_{cos}(\boldsymbol{b}_i、\ \hat{\boldsymbol{b}}_{\sigma(i)}) \quad (3)$$

$\mathcal{L}_{bce}(1、\ \hat{p}_{\sigma(i)})$は二値交差エントロピーを示し、確度$\hat{p}_{\sigma(i)}$を 1 に近づけるための損失項である。$\mathcal{L}_{mse}(c_i、\ \hat{c}_{\sigma(i)})$は平均二乗誤差を示し、濃度$\hat{c}_{\sigma(i)}$を$c_i$に近づけるための損失項である。$\mathcal{L}_{cos}(\boldsymbol{a}_i、\ \hat{\boldsymbol{a}}_{\sigma(i)})$および$\mathcal{L}_{cos}(\boldsymbol{b}_i、\ \hat{\boldsymbol{b}}_{\sigma(i)})$は励起スペクトル$\hat{\boldsymbol{a}}_{\sigma(i)}$、蛍光スペクトル$\hat{\boldsymbol{b}}_{\sigma(i)}$をそれぞれ$\boldsymbol{a}_i$、$\boldsymbol{b}_i$に近づけるための損失項である。

教師情報と対応付けられなかった出力$\hat{y}_j$に対する損失$\mathcal{L}_{unmatch}(\hat{y}_j)$を式 4 に示す。式 4 は出力$\hat{y}_j$をダミーの出力とするために、確度$\hat{p}_j$および濃度$\hat{c}_j$を 0 に近づけるための損失として設計する。

$$\mathcal{L}_{unmatch}(\hat{y}_j) = \mathcal{L}_{bce}(\hat{p}_j、\ 0) + \mathcal{L}_{mse}(\hat{c}_j、0) \quad (4)$$

3.3 SEEMD-Net を用いた蛍光指紋分解

学習済みの SEEMD-Net を用いることで、単独の蛍光指紋を end-to-end で分解することが可能となる。単独の蛍光指紋を SEEMD-Net に入力し、得られた出力の内、確度が閾値以上の出力に含まれる濃度、励起スペクトル、蛍光スペクトルを分解結果とする。このように、確度の指標を用いることで、PARAFAC では観測者が指定する必要があった蛍光成分の数を含めて最適化を介さずに end-to-end での分解を実現する。また疑似蛍光指紋に tri-linear モデルに沿わないノイズを付与、SEEMD-Net が学習することで、ノイズへの頑健性を獲得することを可能とする。

4. 実験

疑似蛍光指紋を訓練用 2,670,000 件，検証用 26,700 件作成し，SEEMD-Net の学習データセットとした。疑似蛍光指紋に含まれる蛍光成分数Rは 1 から 7 の範囲でランダムに設定した。SEEMD-Net の出力数$R_{\max}$は最大蛍光成分数 7 より十分大きな数 30 とした。学習エポック数は 100 とし，検証用データに対する損失が最も小さいエポックのモデルを最終的な学習済み SEEMD-Net とした。

実験には蛍光分子を含んだ溶液から測定された蛍光指紋のオープンデータセット[6] を用いた。蛍光指紋は 6 種の蛍光分子から最大 3 つを混合した試料で構成される。

4.1 蛍光成分数の予測精度

各化合物がユニークな蛍光成分を発すると仮定し、試料に含まれる化合物数 0 から 3 までの蛍光指紋に対し、SEEMD-Net が正しく成分数を予測できるのかを評価した。表 1 にあるように、化合物数が多くなるにつれて、一致率が低下している。これは疑似蛍光指紋と実際の蛍光指紋との乖離や、ノイズパターン不足が原因と考えられる。

表 1　予測蛍光成分数と化合物数の一致率

化合物数	サンプル数	一致率
0	230	100%
1	260	93%
2	170	89%
3	62	80%

4.2 PARAFAC との定性比較

PARAFAC との分解結果の比較例を図 5 に示す。出力例 1 のように、入力との形状差は PARAFAC の方が少ない。PARAFAC は蛍光成分数を指定する必要があるものの、成分同士の重なりが少ない簡単なケースにおいては、PARAFAC の方が形状の正確性において有効であることが分かる。提案手法は蛍光指紋の詳細な形状までを保持しつつ分解が出来ていない一方、出力例 2 のように、複雑な構造の蛍光指紋の場合、PARAFAC と比較して自然な分解を実現している。このように、複雑なパターンでは PARAFAC では不十分な場合が多く、提案手法が有効であることが分かる。

5. 終わりに

本稿では、新たな蛍光指紋手法 SEEMD-Net を提案した。デファクトスタンダードである PARAFAC と比較して複雑な成分構造の蛍光指紋に対する優位性がある一方、出力形状が入力と異なってしまう課題が残る。今後は深層学習と従来の最適化方式を組み合わせることでの分解精度向上を検証する予定である。

参考文献

[1] R. Harshman, et al., “PARAFAC: Parallel factor analysis”, Computational Statistics & Data Analysis, Vol.18, (1994).
[2] J. Hayashida, et al., “Representation Learning using Pseudo-Excitation Emission Matrix”, FIT2024.
[3] N. Carion, et al. “End-to-End Object Detection with Transformers”, ECCV2020.
[4] G. McNamara, “PubSpectra - Open Data Access Fluorescence Spectra”, available at http://works.bepress.com/gmcnamara/9/ (2012)
[5] A. Vaswani, et al., “Attention is all you need”, NeurIPS2017
[6] B. Rasmus, et al., “Standard error of prediction for multilinear PLS 2. Practical implementation in fluorescence spectroscopy”, Chemometrics and Intelligent Laboratory Systems, Vol.75, (2005).