AI(人工知能)を用いた構造解析機能を搭載したJMS-T2000GC専用解析ソフトウェアmsFineAnalysis AI

久保 歩　日本電子(株) MS事業ユニット

msFineAnalysisは、日本電子製ガスクロマトグラフ飛行時間質量分析計(GC-TOFMS)専用解析ソフトウェアとして開発された。複数回のバージョンアップを経て、msFineAnalysisにはデコンボリューション検出や2検体差異分析等の機能が実装されてきた。今回、弊社はmsFineAnalysisの新バージョンとして、msFineAnalysis AIを開発した。msFineAnalysis AIには、人工知能(AI)を用いた構造解析手法"AI構造解析"が搭載されている。AI構造解析は、NISTライブラリーに未登録の化合物(未知化合物)の分子式だけでなく構造式まで決定することが可能である。AI構造解析のワークフローは次のようになる。

最初に、従来のmsFineAnalysisの特徴である統合解析が未知化合物の分子式を決定する。次に、1億を超える化合物が登録されているデータベース(PubChem)から決定された分子式を元に構造式候補が抽出される。抽出された構造式の電子イオン化(EI)法のマススペクトルを、AIが構造式から予測する。次に、予測されたマススペクトルと実測のマススペクトルの比較により構造式候補の順位付けが行われる。最後に、最上位の構造式候補が解析結果として採用される。

構造式からマススペクトルを予測するAIの学習と精度評価には、NIST20ライブラリーを使用した。精度評価の結果、未知化合物の構造解析においてAI構造解析が有用であることを確認した。本報告では、msFineAnalysis AIの特徴および評価結果を紹介する。

はじめに

ガスクロマトグラフ質量分析法(GC-MS)のイオン化法として、電子イオン化法(EI法)が広く採用されている。EI法で得られるマススペクトル(以後、EIマススペクトルと記述)では、フラグメントイオンが主として観測される。フラグメントイオンは化合物の構造を反映したものであり、そのパターンは化合物の構造特有のものである。そのため、GC-MSの定性解析には、標準品のEIマススペクトルが蓄積されたライブラリーとの比較が用いられている。最も広く使われている構造式とマススペクトルのライブラリーであるNISTライブラリーにおいては、登録されている化合物の数は約30万となっている。

一方で、代表的な化合物のデータベースであるPubChemは、2023年現在化合物の数は1億を超えている。PubChemには、EIマススペクトルが登録されていない。そのため、ごく一部のNISTライブラリーに重複して登録されているものを除いた大半の化合物はEIマススペクトルの情報を有していないこととなる。結果的に、それら化合物のEIマススペクトルをライブラリーサーチした場合は、定性解析結果が得られないだけでなく、別の化合物であると誤判定する可能性もある。このようなNISTライブラリーに無い化合物に対しては、電界イオン化法(FI法)を代表とするソフトイオン化法と精密質量を得ることができる質量分析計[1]の組み合わせ[2]が有用である。具体的には以下の手順となる。

① EI法とソフトイオン化法のマススペクトルを比較し、分子関連イオンピークが決定される。
② 決定されたピークの精密質量を元に分子式候補が獲得される。
③ 得られた分子式候補に対して、同位体パターン解析およびEIマススペクトルのフラグメントイオンに対する精密質量解析が行われる。2つの結果を元に、分子式が決定される。

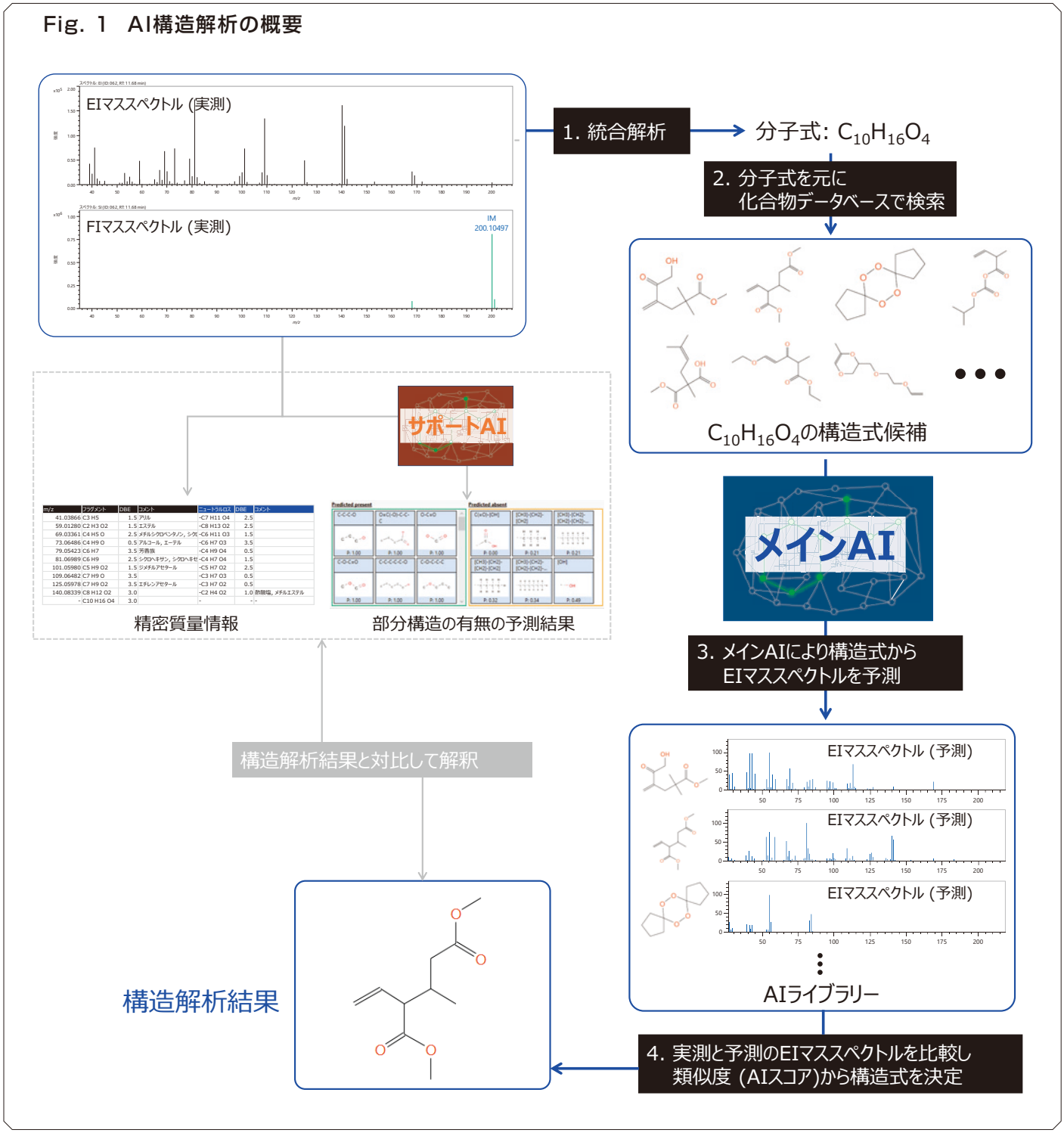
本手法を搭載したmsFineAnalysisは、自動で未知化合物の分子式を決定することが可能である。今回、未知化合物の分子式だけでなく構造式まで得られることを目指し、人工知能(AI)を活用した"AI構造解析"を開発した。"AI構造解析"を搭載したmsFineAnalysisの新バージョンである"msFineAnalysis AI"は、2023年1月から販売開始されている。本稿では、"AI構造解析"の概要と精度を評価した結果と、さらにNISTライブラリーに未登録の化合物に適用した結果を報告する。

AI構造解析

AI構造解析では、メインAIとサポートAIの2つのAIを使用している。**Fig. 1**にライブラリー未登録化合物に対する統合解析およびAI構造解析の手順を示す。msFineAnalysis AIでは、化合物検出および以下の手順1-4の処理は、全て自動で行われる。なお、2つのAIの詳細については、次節で紹介する。

〉〉 〒196-8558　東京都昭島市武蔵野3-1-2 ｜ E-mail: akubo@jeol.co.jp

Fig. 1　AI構造解析の概要



1. EIマススペクトルとソフトイオン化法であるFI法のマススペクトルを用いた統合解析が行われ、分子式が決定される。
2. 決定した分子式を元に1億の化合物が登録されているデータベース(PubChem)から構造式候補が抽出される。この際、抽出された構造式の数は多くとも1万となる。
3. 構造式の候補に対して、構造式からEIマススペクトルを予測するメインAIが、予測EIマススペクトルを与える。
4. 予測EIマススペクトルと実測のEIマススペクトルを比較し、AIスコア(コサイン類似度)が構造式候補を順位付けする。そして、最上位の構造式候補が解析結果として採用される。

※ 1-4で得られた構造解析結果と同時に、精密質量情報と共にサポートAIが予測した部分構造の結果が表示される。これらの情報と知見を元に、解析者が構造解析結果を解釈することができる。但し、本作業は独立しており、本作業を行わなくても自動で構造解析結果が得られる。

AI構造解析の特徴として、AIによるEIマススペクトルの予測だけでなく、統合解析により決定された分子式による絞り込みがある。AIにより予測されたEIマススペクトルとの比較の前に、統合解析により決定された分子式が構造式候補を絞り込む。これにより、1億の構造式候補を1万以下まで絞り込むことが可能であり、効率的かつ高精度な構造解析が可能となっている。

分子式を事前に決定していない場合は、化合物データベース全体に対して総当たりでの比較、あるいは化合物種などでの絞り込みが必要となる。前者の総当たりでの比較では、1億のEIマススペクトルに対する比較であるため、多大な時間を要し精度が低下する。精度が低下する要因としては、EIマススペクトルの情報だけでは区別が難しい化合物が存在することが挙げられる。**Fig. 2**に示す4つの化合物はいずれも構造式および分子式が異なるが、EIマススペクトルが非常に似通っている。そのため、EIマススペクトルの比較だけでは同定が難しく、誤った定性解析結果を導く可能性がある。一方、後者の化合物種の決定には、試料の情報や経験・知見が必要となる。試料の情報が十分にない場合には、化合物種を決定することが困難である。また、正しくない種の選択は誤った構造解析結果を導くことになる。結果として、解析作業が属人的となり再現性が低下する恐れがある。しかし、AI構造解析では、前述したように統合解析により決定された分子式により事前に構造式候補の絞り込みを行うため、Fig. 2に示された4つの化合物も正しく解析結果が示される。

msFineAnalysis AIには、メインAIそのものは搭載されていない。代わりに、PubChemから抽出した構造式と、構造式からメインAIが予測したマススペクトルが登録されたライブラリーである“AIライブラリー”が搭載されている。これにより、解析時にマススペクトル予測に関する時間が不要となり、解析のスループットが向上する。測定データを選択し実行ボタンを押すことで構造解析まで全てを自動で行い、100化合物の構造解析結果が10分以内に得られる。また、解析時に化合物データベースへのインターネット接続が不要となり、スタンドアローンで安定して解析することが可能となる。

AI構造解析のGUIを**Fig. 3**に示す。画面下部に構造式がAIスコア降順に一覧に並べて表示される。一覧で表示された中で、左上に位置している構造式が構造解析結果である。構造式の情報として、構造式以外にIUPAC名およびPubChem CID (PubChemデータベースでの識別番号)が表示される。また、画面右上に分子式に対する構造式の候補数やAIスコアを用いて作成されたヒストグラムが表示される。これらの情報が、構造解析結果を俯瞰して見ることを可能とする。

さらに、解析対象の化合物に関する知見がある場合は、解析者がベンゼン環やメチルエステル等の部分構造を用いて構造式をフィルタリングすることも可能である。フィルタリングを行った場合は、選択された部分構造を含む構造式のみが表示される。また、画面右端にあるボタンを押すことで、精密質量マススペクトルおよび精密質量情報だけでなく、サポートAIの部分構造予測結果が表示される。解析者が構造解析結果に対して確認・解釈することが可能となっている。

2つのAI

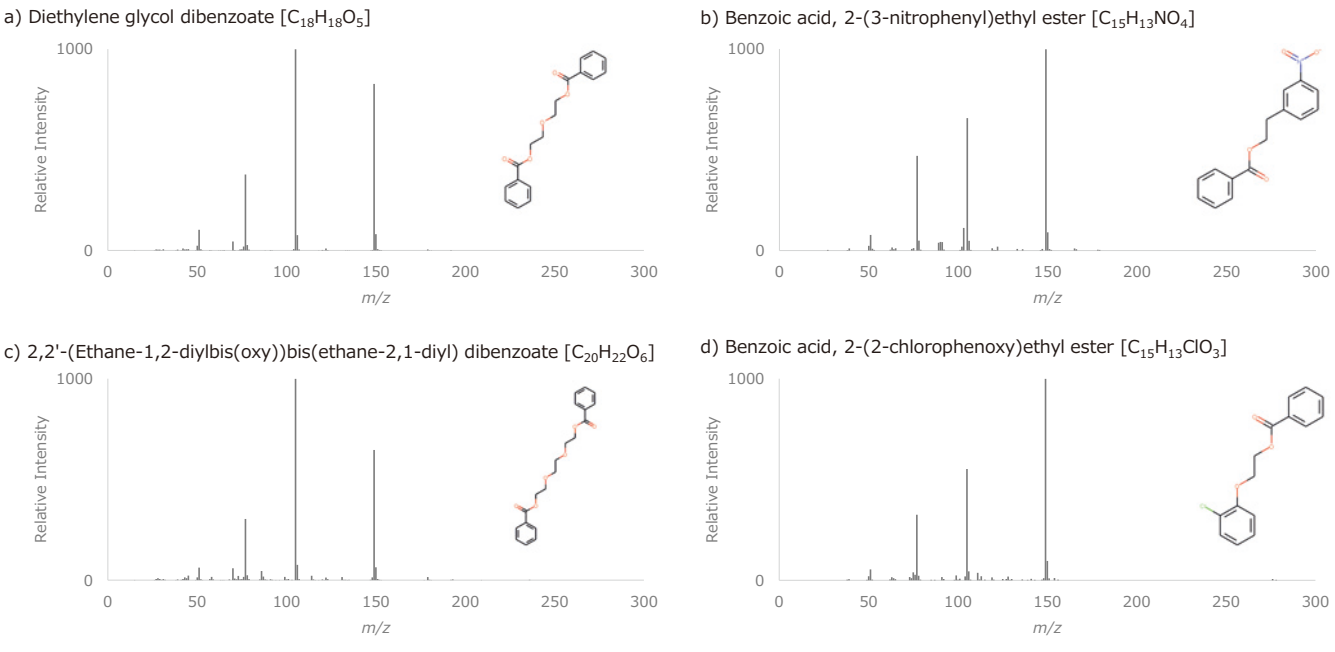
本節ではAI構造解析で用いている2つのAIについて説明する。

メインAIは、深層学習の1種であるグラフコンボリューショナルネットワークス(GCN)[3]をモデルとして採用している(**Fig. 4**上)。GCNは次のような動作をする。最初に、AIはマススペクトルに特徴的な信号を生み出す部分構造を構造式から探索し、部分構造を大量に生成する。そして、AIは生成した部分構造の情報を元にマススペクトルを予測する(Fig. 4下)。

具体的な処理としては次のようになる。最初に、GCNに入力する前に、構造式がグラフデータに変換される(**Fig. 5**)。グラフデータでは、構造式の原子はノードとして、結合はエッジとして扱われる。さらに、ノードが原子の元素種の情報を、エッジが結合の種類の情報を、それぞれ特徴ベクトルとして保持する。例えば、ノードは、元の原子が炭素原子であれば(1, 0, 0, …)、酸素原子であれば(0, 1, 0, …)、窒素原子であれば(0, 0, 1, …)という特徴ベクトルを持つ。

次に、グラフデータに変換された構造式に対して、AIはFig. 4上の左で描かれている畳み込み(コンボリューション; Convolution)を行う。畳み込みを行うことで、各ノードは隣接しているノードおよび

Fig. 2　NIST20に登録されている4つの化合物のEIマススペクトル



エッジの情報を取捨しながら獲得する。畳み込みを繰り返すことで、AIが原子の繋がりをブロックとして認識できるようになる。

そして、AIはFig. 4上の右で描かれている各原子の集約(プーリング; Pooling)を行う。これにより、構造式の特徴が捉えられ、AIがマススペクトルの予測を行うことが可能となる。

サポートAIは、深層学習ではなく従来の機械学習と呼ばれる手法(回帰)を採用している。AIが、精密質量マススペクトルを元に、48個の部分構造の有無をイオン・ニュートラルロスから予測する(**Fig. 6**)。サポートAIは係数の数が数十個とシンプルである。そのため、AIは予測結果だけでなくその特徴となるピークを同時に提示することができる。

AI構造解析の精度評価

—EIマススペクトル予測精度の評価—

AI構造解析では、メインAIが構造式から予測したマススペクトルを使用する。メインAIはNIST20ライブラリーの90%にあたる27万化合物の構造式とマススペクトルを元に学習した。学習時は、構造式から予測したマススペクトルとNIST20ライブラリーのマススペクトルのパターンが一致するようにメインAIの重みを最適化した。残りの3万化合物の内、1万化合物が過学習に陥らないようにするための進捗監視に、残りの2万化合物がEIマススペクトル予測精度の評価に割り当てられた。

学習に使用していない2万化合物を用い、メインAIのEIマススペクトル予測精度を評価した。対象の化合物に対して、学習済みのメインAIが構造式からEIマススペクトルを予測した。精度を評価する指標は、予測したEIマススペクトルとNIST20ライブラリーに登録されているEIマススペクトルのコサイン類似度とした。コサイン類似度は1が完全に一致しており、0に近づくほど一致していないという指標となる。

Fig. 7に2万化合物のコサイン類似度を元に作成したヒストグラムを示す。ヒストグラムを確認すると、90%を超える化合物でコサイン類似度が0.4を超えている。さらに、0.7～0.8の区間の化合物数が最も多い結果となり、全体の平均値が0.72となった。構造式からの予測により、メインAIが高い精度でマススペクトルを再現できることが確認された。

マススペクトルの例として、**Fig. 8**にコサイン類似度が平均より高い化合物、平均に近い化合物、平均より低い化合物の実測EIマススペクトルと予測EIマススペクトルの比較を示す。平均より高いBenzamide,3-methyl-N-decyl-を確認すると、強度の弱い

Fig. 3　AI構造解析のGUI

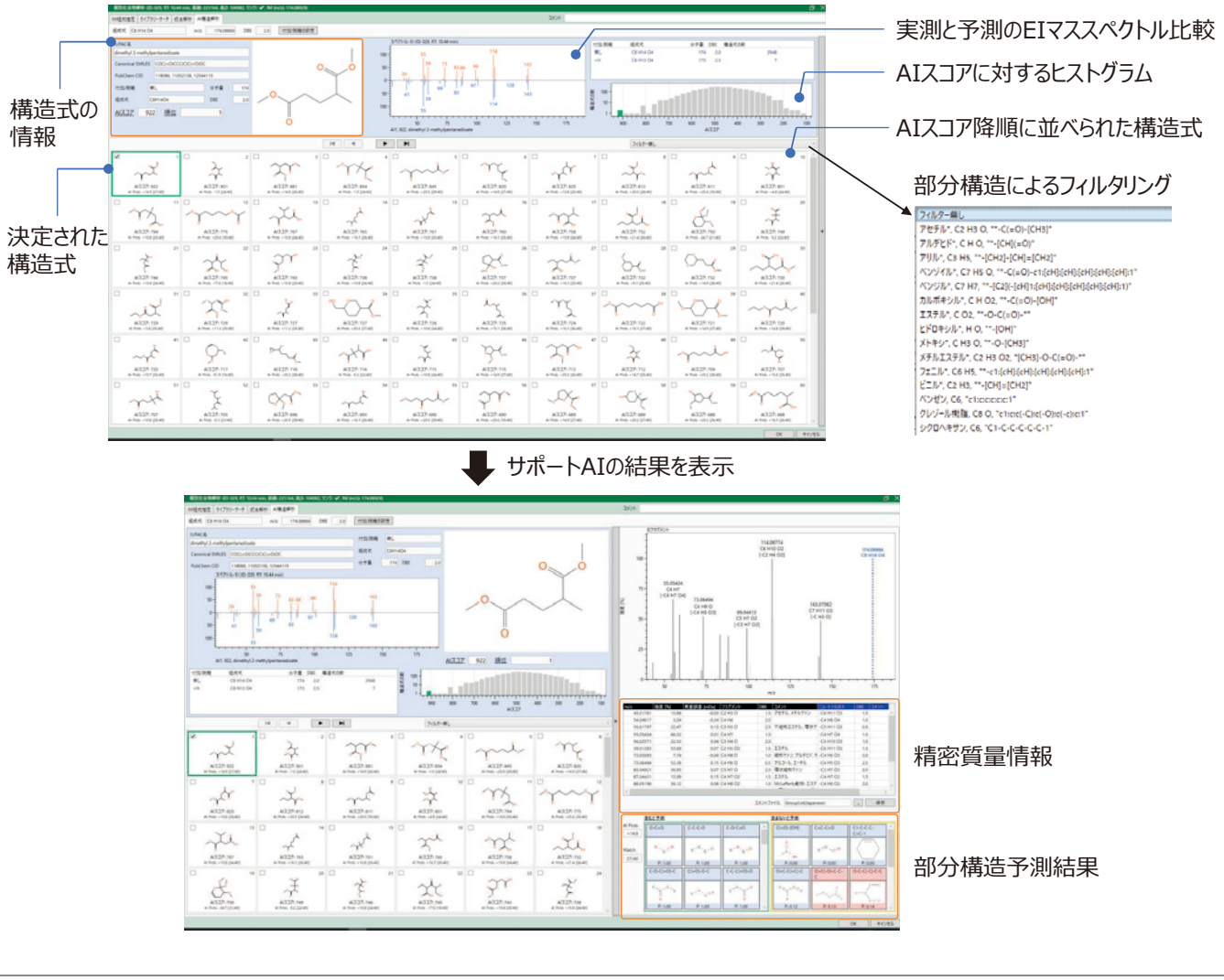


Fig. 4　メインAIで使用しているグラフコンボリューショナルネットワークス

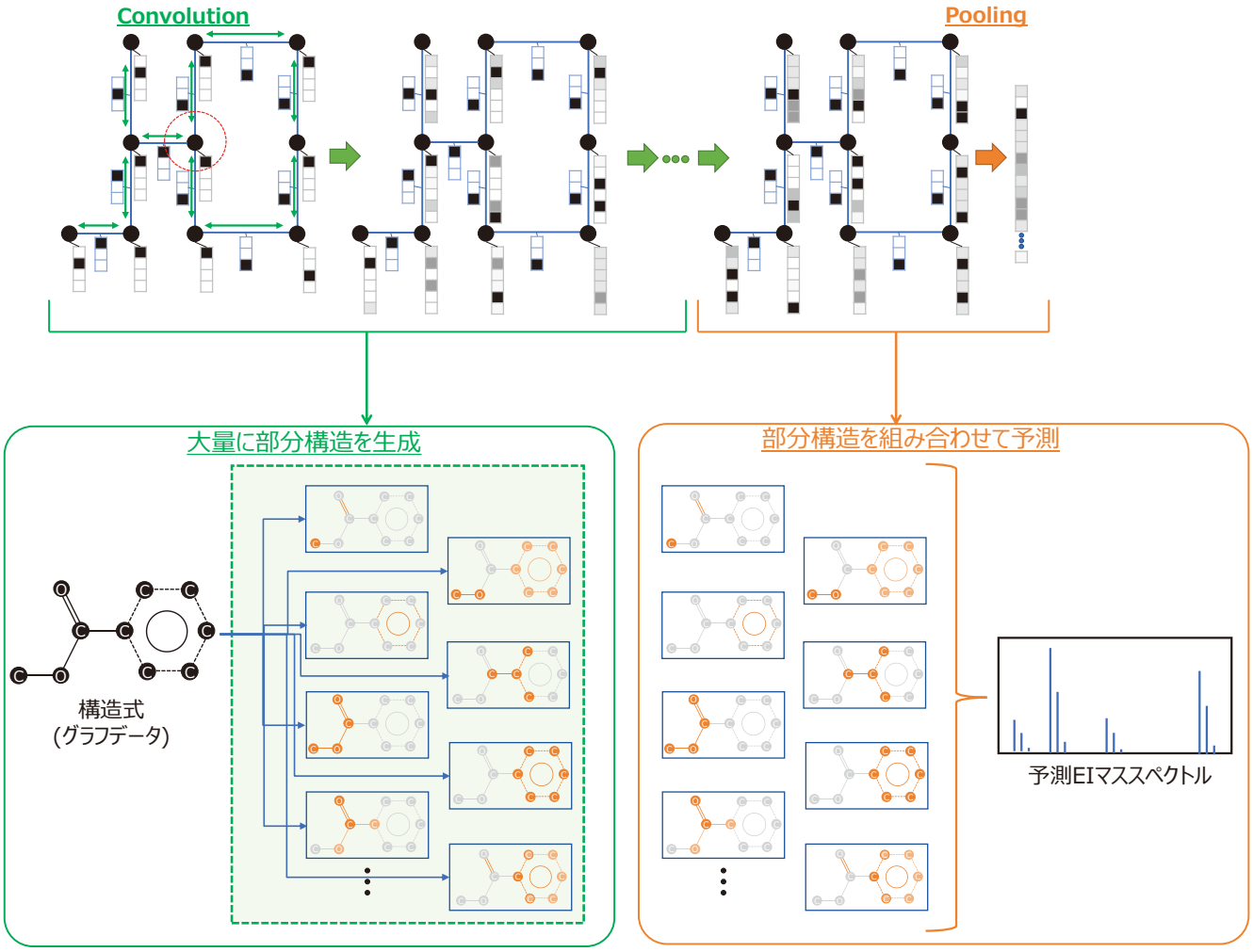


Fig. 5　構造式のグラフデータへの変換

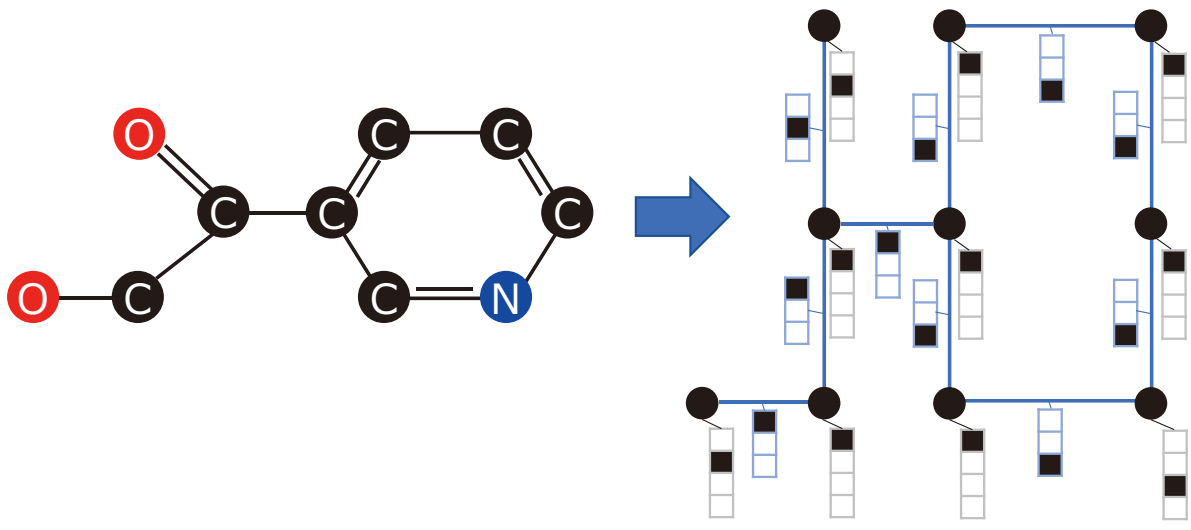


Fig. 6　サポートAIの概要

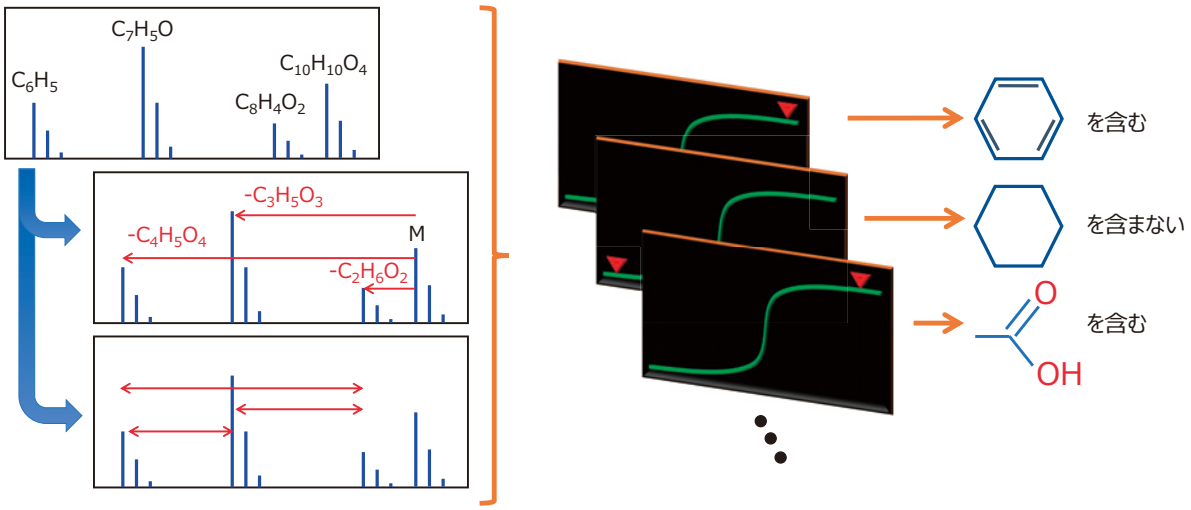
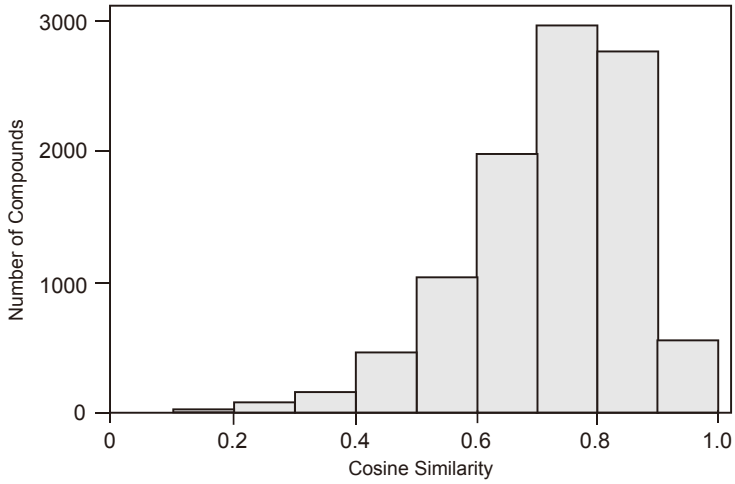


Fig. 7　学習に使用していない2万化合物のコサイン類似度に対するヒストグラム



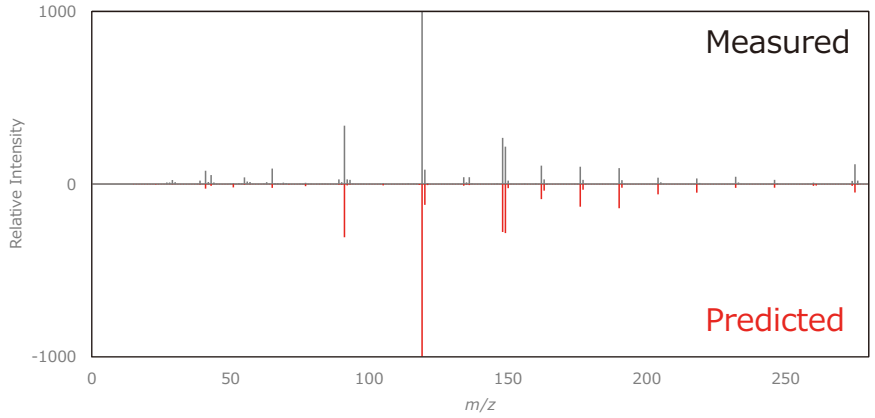
マスピークを含めほぼ完全にEIマススペクトルが再現できていることが分かる。要因として、この化合物がNIST20ライブラリー内で登録数が多いベンゼン環とアルカン鎖、アミド基のみで構成されているためと考えられる。平均に近いN-Acetyl-3-(3-formyl-4-methoxyphenyl)-d-alanine methyl esterを確認すると、比較的強度の強いマスピークが再現できており、全体的なパターンは一致している。Benzamide, 3-methyl-N-decyl- の構造式と比較すると、この化合物はベンゼン環に複数の側鎖が付いた幾分か複雑な構造をしており、そのために完全にはマススペクトルを再現できなかったと考えられる。平均より低いCyclododecane,1,5,9-tris(acetoxy)-を確認すると、パターンは全体的にあまり再現できない。これは、この化合物が12員環という巨大な環を含む化合物であり、NIST20ライブラリー内で12員環を含む化合物の登録数が少なく十分に学習できなかったことが原因と考えられる。それでも、最も強く観測されている *m/z* 43のマスピークを含む一部マスピークは再現できている。

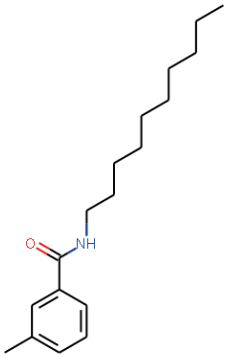
―構造解析精度の評価―

AI構造解析は、構造式から予測したEIマススペクトルと実測のEIマススペクトルとの比較により、構造式を決定する。この手法による構造式を決定する精度について評価した。評価方法としては、次の手順となる。最初に、学習に使用していないNIST20ライブラリーの化合物に対して、化合物データベースからその分子式と同じ分子式を持つ構造式(化合物)を抽出する。次に、正しい構造式と抽出した構造式それぞれに対し、学習済みのメインAIがEIマススペクトルを予測する。NIST20ライブラリーに登録されているEIマススペクトルと予測されたEIマススペクトルを比較し、そのコサイン類似度から正しい構造式を含めて全ての構造式が順位付けされる。精度を評価する指標は、全体に対する正しい構造式の順位とした。なお、本評価は、一定の基準を設けるために、化合物データベースで候補が100化合物以上抽出された分子式を対象とした。

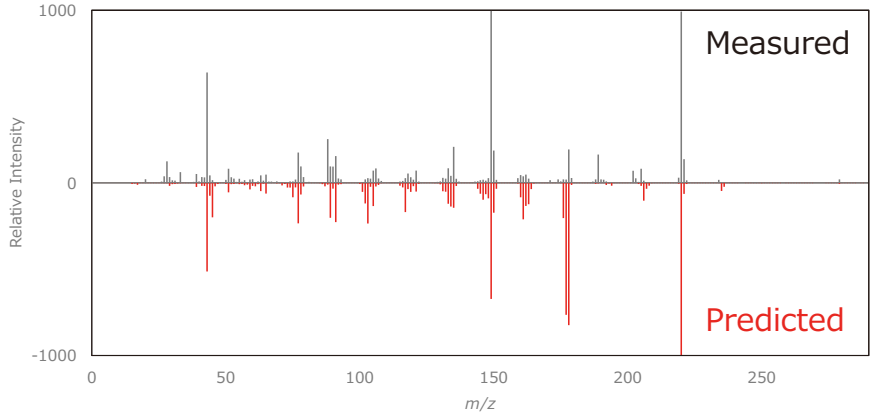
Fig. 8 実測と予測EIマススペクトルの比較

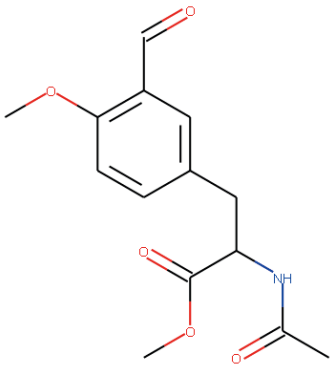
a) benzamide, 3-methyl-N-decyl-
[コサイン類似度: 0.95]



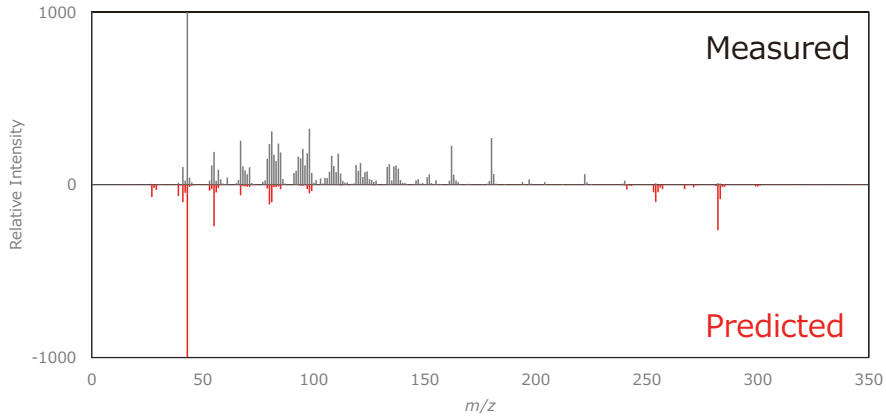


b) N-Acetyl-3-(3-formyl-4-methoxyphenyl)-d-alanine methyl ester
[コサイン類似度: 0.72]





c) cyclododecane, 1,5,9-tris(acetoxy)-
[コサイン類似度: 0.34]



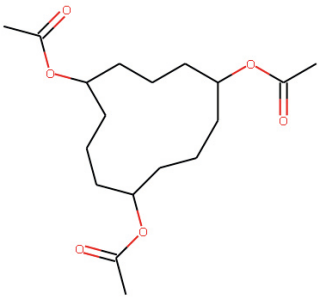


Table 1に、14,581化合物に対して順位を確認した結果を示す。22%の化合物が最上位に正しい構造式が順位付けられる結果となった。また、73%の化合物が上位1%以内に正しい構造式が入る結果となった。上位1%以内とは、1,000個の候補があった場合、上位10個以内に正しい構造式があったことを意味する。化合物データベースであるPubChemには、化合物によっては構造式がかなり近い化合物が多数登録されている。その点を考慮すると、本手法は高い精度であると言える。

次に、本手法が、NIST20ライブラリーに未登録の完全未知化合物に対しても、有効であることを評価した。評価は、NIST20ライブラリーに未登録のモデル化合物を用意して実施した。モデル化合物は、Cafenstrole（CAS: 125306-83-4, Wako）、MCPA-thioethyl（CAS: 25319-90-8, Wako）、Propaphos（CAS: 7292-16-2, Wako）、CNP-amino（CAS: 26306-61-6 Wako）、Butamifos oxon（CAS: 56362-05-1 Wako）、Isoxadifen-ethyl（CAS: 163520-33-0, Wako）である。

モデル化合物の実測EIマススペクトルは、標準品を測定することで用意した。**Table 2**に各化合物での正解の構造式の順位およびスコアとスコア降順で上位10個の構造式を示す。6個の内3個の化合物が、最上位に正解の構造式を位置している。最も正解の順位が低いIsoxadifen-ethylにおいても、正解の構造式が5,348個の構造式候補の中での22位という順位となった。すなわち、正解の構造式が上位1%以内には入っており、多数の候補の中から正しい構造式を絞り込む効果は認められる。CafenstroleやCNP-amino, Isoxadifen-ethylの最上位の構造式と正しい構造式を比べると、環のサイズや数は一致し、かなり類似していることが分かる。以上の6化合物の結果から考えても、本手法は構造解析において有用であると言える。**Fig. 9**に、実測のマススペクトルと予測マススペクトル

Table 1　14,581化合物に対する精度評価の結果

	Top	Within the top 1%	Within the top 5%	Within the top 10%
Number of Compounds	3215 (22 %)	10618 (73 %)	12934 (89 %)	13594 (93 %)

Table 2　構造解析の結果例

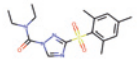
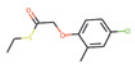
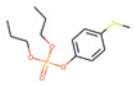
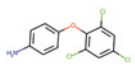
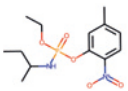
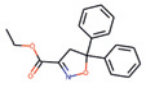
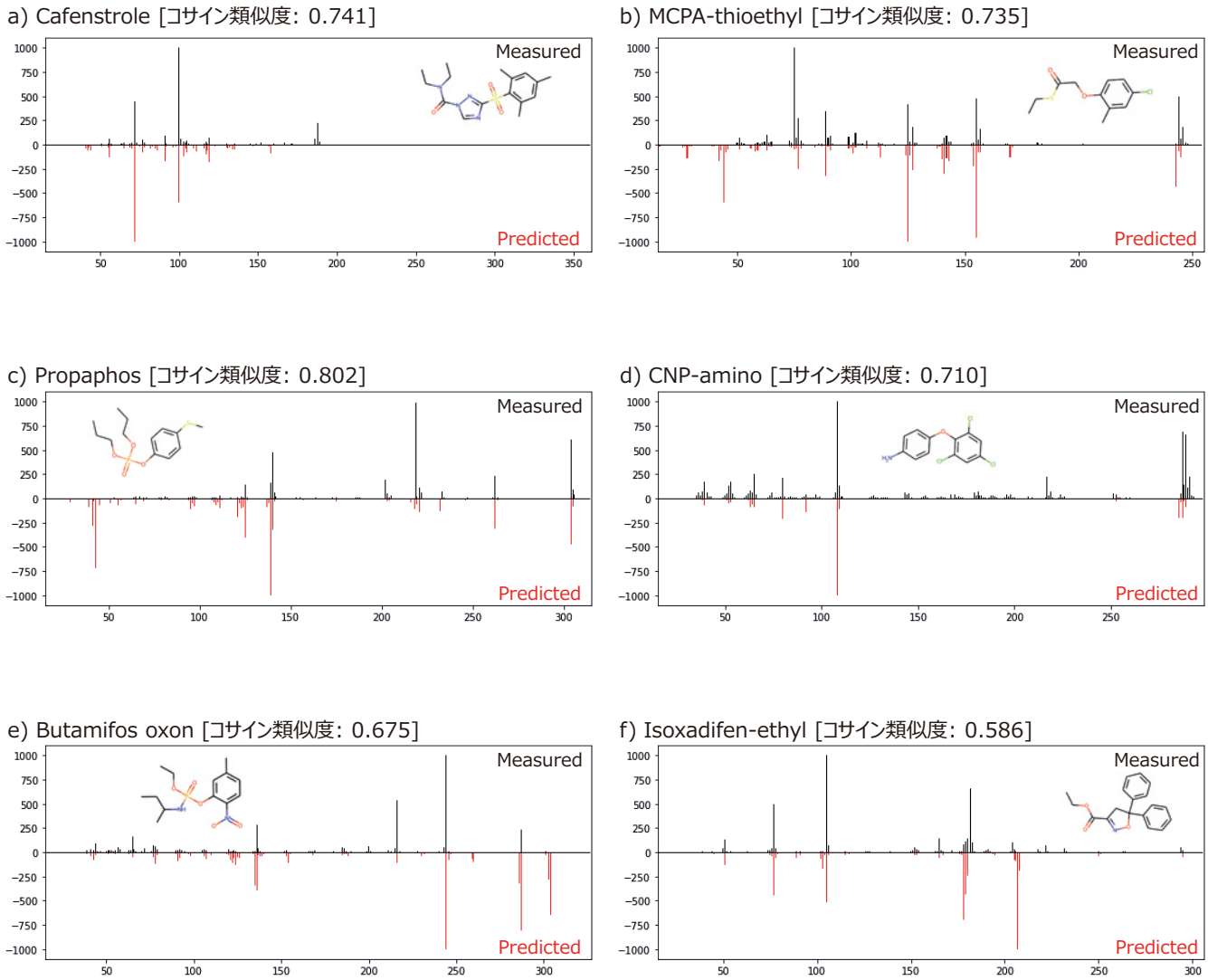
Compound Name	Structure	Similarity	Rank	Top 10 structures
Cafenstrole		0.741	3 (2933)	Similarity: 0.755, Similarity: 0.751, Similarity: 0.741, Similarity: 0.636, Similarity: 0.636, Similarity: 0.635, Similarity: 0.633, Similarity: 0.63, Similarity: 0.629, Similarity: 0.627
MCPA-thioethyl		0.735	1 (729)	Similarity: 0.735, Similarity: 0.708, Similarity: 0.689, Similarity: 0.67, Similarity: 0.661, Similarity: 0.659, Similarity: 0.654, Similarity: 0.654, Similarity: 0.638, Similarity: 0.623
Propaphos		0.802	1 (27)	Similarity: 0.802, Similarity: 0.783, Similarity: 0.779, Similarity: 0.736, Similarity: 0.587, Similarity: 0.532, Similarity: 0.418, Similarity: 0.4, Similarity: 0.398, Similarity: 0.395
CNP-amino		0.710	14 (618)	Similarity: 0.774, Similarity: 0.755, Similarity: 0.753, Similarity: 0.733, Similarity: 0.729, Similarity: 0.729, Similarity: 0.729, Similarity: 0.729, Similarity: 0.724, Similarity: 0.723
Butamifos oxon		0.675	1 (56)	Similarity: 0.675, Similarity: 0.581, Similarity: 0.562, Similarity: 0.553, Similarity: 0.506, Similarity: 0.498, Similarity: 0.468, Similarity: 0.437, Similarity: 0.436, Similarity: 0.431
Isoxadifen-ethyl		0.586	22 (5348)	Similarity: 0.767, Similarity: 0.743, Similarity: 0.729, Similarity: 0.725, Similarity: 0.718, Similarity: 0.691, Similarity: 0.654, Similarity: 0.649, Similarity: 0.644, Similarity: 0.642

Fig. 9 実測と予測マススペクトルの比較



を比較したものを示す。ピークの強度の大小や細かなピークの有無は実測と予測で異なるが、強度の強いピークは一致している。

以上の結果により、本手法が未知化合物の構造解析において有効であると確認できた。

おわりに

従来のmsFineAnalysisは、JMS-T2000GCの特徴である精密質量測定およびソフトなイオン化法による分子イオン観測に基づいた統合解析が特徴である。統合解析は、未知化合物に対しても分子式を決定することができる。新バージョンであるmsFineAnalysis AIは、AI(人工知能)を用いた構造解析により、分子式だけでなく構造式まで自動で得ることができる。msFineAnalysis AIは、統合解析により決定された分子式を元に構造式候補を抽出する。そして、AIが構造式から予測したEIマススペクトルを用いて構造式を決定する。統合解析とAIの組み合わせが、高効率かつ高精度の構造解析を可能とする。また、全ての解析が自動かつオフラインで動作するため、安定した分析業務を行うことが可能である。

参考文献

[1] 生方正章. マルチイオン化-未知物質解析システム新型ガスクロマトグラフ/高分解能飛行時間質量分析計 JMS-T2000GC “AccuTOF™ GC-Alpha”. 日本電子ニュース Vol. **53**(2021).

[2] 生方正章、上田祥久. 電子イオン化法およびソフトイオン化法を搭載した高質量分解能GC-TOFMSによる測定結果の統合解析手法の開発. 日本電子ニュース Vol. **51**(2019).

[3] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vynyals, G. E. Dahl. Neural message passing for Quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*. 2017; **70**: 1263-1272.