

msFineAnalysis AI Novel Qualitative Analysis Software for JMS-T2000GC with AI Structural Analysis

Ayumi Kubo MS Business Unit, JEOL Ltd.

JEOL developed msFineAnalysis as qualitative analysis software for our gas chromatograph time of flight mass spectrometer (GC-TOFMS). We implemented deconvolution detection, variance component analysis, and other features in the software through updates. We have recently developed a new version of the series called msFineAnalysis AI. msFineAnalysis AI is equipped with a structural analysis method using artificial intelligence (AI), called “AI structural analysis.” AI structural analysis enables the identification of molecular formulas as well as structural formulas of compounds that are not registered in the NIST 20 library (unknown compounds). The workflow of AI structural analysis is as stated below.

First, msFineAnalysis’s integrated analysis function identifies the molecular formula of an unknown compound. Next, based on the identified molecular formula, structural formula candidates are extracted from PubChem, the database containing over 100 million compounds. The AI predicts electron ionization (EI) mass spectra from the extracted structural formula candidates. Then, the structural formula candidates are ranked by comparing the predicted mass spectra with the measured mass spectrum. Finally, a candidate that ranks first is adopted as the analysis result.

Using the NIST 20 library, we trained the AI to predict mass spectra from structural formulas and evaluated its accuracy. From the results of accuracy evaluation, we confirmed that AI structural analysis is useful in the structural analysis of unknown compounds. In this report, we will introduce features of msFineAnalysis AI and provide our evaluation results.

Introduction

The electron ionization (EI) method is widely used as an ionization method for gas chromatograph mass spectrometry (GC-MS). Fragment ions are mainly observed in a mass spectrum obtained by the EI method (herein, an EI mass spectrum). Fragment ions reflect the structure of a compound and has a pattern unique to it. For this reason, in qualitative analysis of GC-MS, an EI mass spectrum is compared with libraries of EI mass spectra of reference compounds. The NIST library, the most widely used library of structural formulas and mass spectra, has about 300,000 registered compounds.

Meanwhile, PubChem, a major compound database, contains over 100 million substances as of 2023. However, EI mass spectra are not registered in PubChem. This means that most compounds in PubChem do not have EI mass spectral information, except for some also registered in the NIST library. When library searches are performed for EI mass spectra of such compounds, qualitative analysis results may not be obtained, or wrong compounds may possibly be identified. For these compounds that are not registered in the NIST 20 library, it is useful to combine [2] the field ionization (FI) and other soft

ionization methods with a mass spectrometer [1] that obtains accurate mass. The specific procedure is as follows:

1. The EI and soft ionization mass spectra are compared, and a molecular ion peak is determined.
2. Based on the accurate mass of the determined peak, molecular formula candidates are obtained.
3. For obtained molecular formula candidates, isotope pattern analysis and accurate mass analysis of fragment ions in the EI mass spectrum are performed. Based on the results of these two analyses, the molecular formula is determined.

The above method is implemented in msFineAnalysis, which enables the automated identification of the molecular formula of an unknown compound. We have newly developed a structural analysis method using artificial intelligence (AI), called “AI structural analysis,” with an aim to obtain not only molecular formulas but also structural formulas of unknown compounds. The new version of msFineAnalysis equipped with AI structural analysis, msFineAnalysis AI, was introduced to the market in January 2023. In this article, we will provide an overview of AI structural analysis and report the results of its accuracy evaluation. In addition, we will show the results of applying this function to compounds that are not registered in the NIST 20 library.

AI structural analysis

AI structural analysis uses two types of AI: main AI and support AI. **Figure 1** shows the procedures of integrated analysis and AI structural analysis for compounds that are not registered in the library. msFineAnalysis AI automatically performs the detection of a compound and steps 1 to 4 below. Details about two types of AI are described in the next section.

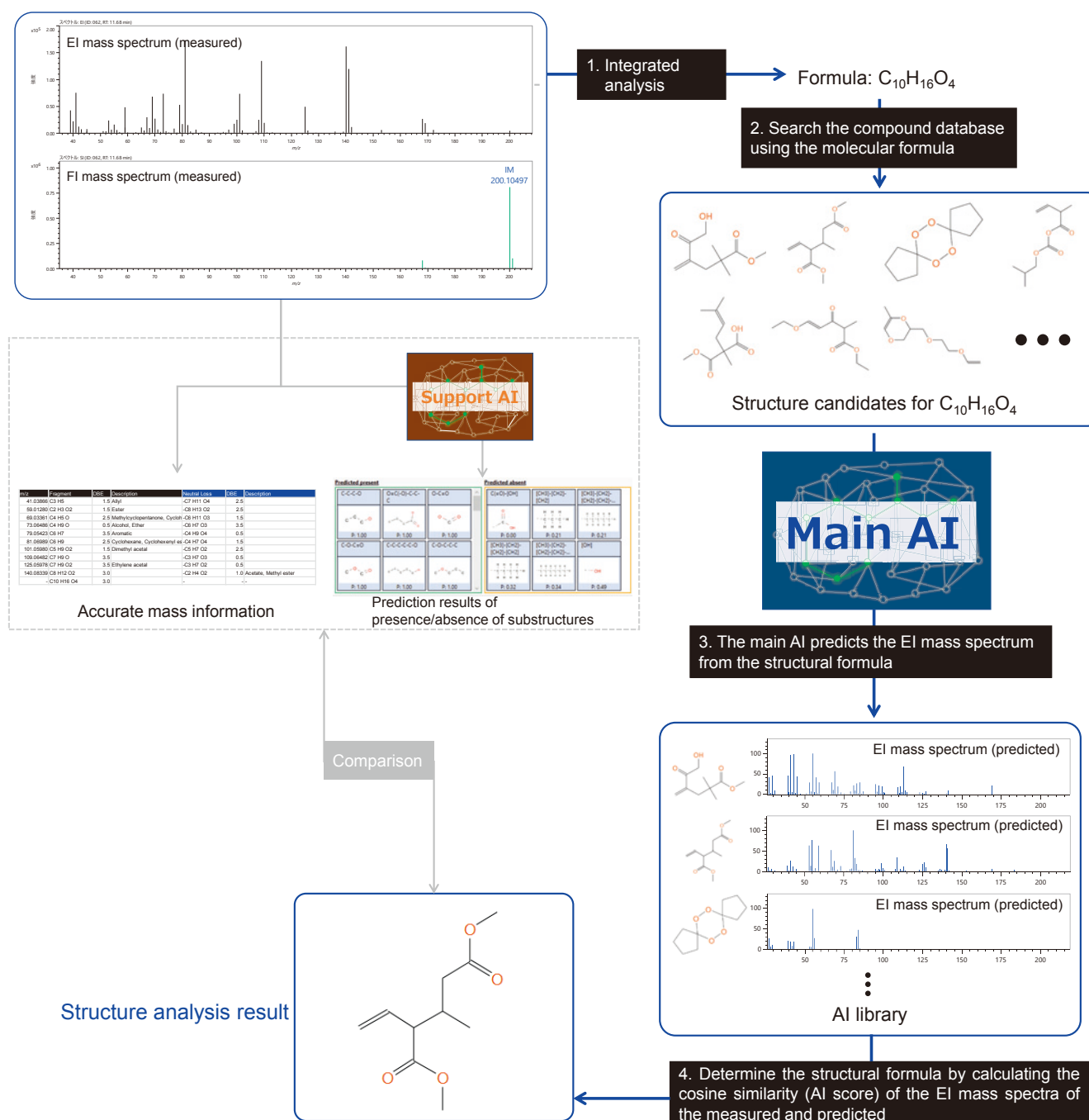
1. msFineAnalysis AI performs integrated analysis using the EI mass spectrum and the mass spectrum obtained by the FI method, a soft ionization method, to identify a molecular formula.
2. Based on the identified molecular formula, the software

extracts structural formula candidates from PubChem database that contains over 100 million compounds. Ten thousand or less candidates are extracted.

3. The main AI predicts EI mass spectra for the extracted structural formula candidates.
4. By comparing the predicted EI mass spectra with the actual measured EI mass spectrum, the software ranks the structural formula candidates using AI scores (cosine similarities). Finally, the candidate that ranks first is adopted as the analysis result.

*The software displays the structural analysis results obtained through steps 1 to 4, as well as accurate mass information and

Fig. 1 Overview of AI structural analysis.



the results of partial structure prediction by the support AI. Analysts can use this information and knowledge to interpret the structural analysis results. However, this process is performed independently, and the structural analysis results can be automatically obtained without it.

Features of AI structural analysis include the EI mass spectrum prediction by main AI, as well as narrowing down candidates based on a molecular formula identified with integrated analysis. Before the measured mass spectrum is compared with AI-predicted EI mass spectra, the molecular formula identified by integrated analysis helps narrow down structural formula candidates. This allows the scope of structural formula candidates to be narrowed from 100 million to 10,000 or less, making it possible to perform an efficient and highly accurate structural analysis.

If a molecular formula is not identified in advance, the measured EI mass spectrum must be compared against the entire compound database, or must be narrowed down using compound species. In comparison against the entire database, the measured spectrum must be compared with 100 million EI mass spectra, resulting in a time-consuming and less accurate analysis. The reason for a lower accuracy is that some compounds are difficult to distinguish from others based on EI mass spectral information alone. The four compounds shown in **Fig. 2** have different structural and molecular formulas, but exhibit highly similar EI mass spectra. Therefore, only comparing their EI mass spectra is not sufficient for identification and may lead to wrong qualitative analysis results. Meanwhile, to identify compound species, information about samples and analysts' experience and knowledge are required. If there is not enough sample information, identifying compound species will be difficult. Additionally, an incorrect selection of species can lead to wrong structural analysis results. Consequently, analysis might be dependent on individual skills of analysts, resulting in a low reproducibility. On the other hand, AI structural analysis

generates correct analysis results for the four compounds shown in Fig. 2, because it narrows down structural formula candidates beforehand using the molecular formula identified by integrated analysis as mentioned earlier.

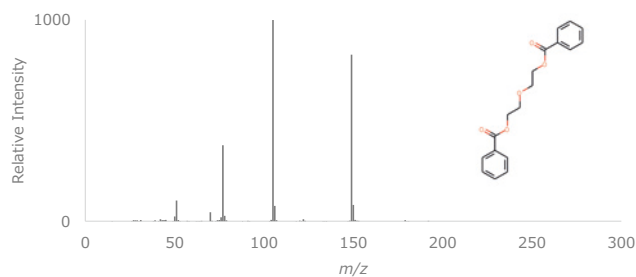
msFineAnalysis AI is not equipped with the main AI. Instead, it is equipped with the "AI library," which contains structural formulas extracted from PubChem and mass spectra predicted from the structural formulas by the main AI. The AI library helps eliminate the need for mass spectrum prediction during analysis, improving the analysis throughput. After an analyst selects measurement data and presses the button to start the analysis, msFineAnalysis AI automatically performs all the processing to complete the structural analysis. The analyst can obtain structural analysis results for 100 compounds within 10 minutes. The AI library also eliminates the need for connecting to the compound database via the Internet during analysis, enabling a stable and stand-alone analysis.

Figure 3 shows the graphical user interface (GUI) of AI structural analysis. Structural formulas are listed in descending order of AI score at the lower part of the window. On the top left corner of the list is the structural analysis result. As the information about the structural formula, its IUPAC name and PubChem CID (identification number in PubChem database) are also displayed. The number of structural formula candidates for the molecular formula and the histogram created using AI score are displayed on the upper right of the window. These various kinds of information help the analyst see the whole picture of the structural analysis results.

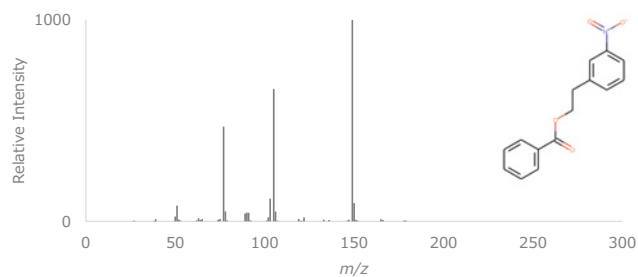
In addition, if there is knowledge about the target compound, the analyst can filter structural formulas using partial structures such as benzene ring and methyl ester. When the analyst presses the button on the right edge of the window, it displays the mass spectrum and information for accurate mass as well as the prediction results of partial structures performed by the support AI. The analyst can confirm and interpret the structural analysis results.

Fig. 2 EI mass spectra of four compounds registered in the NIST 20 library.

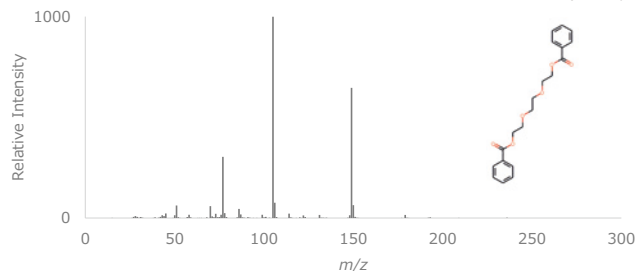
a) Diethylene glycol dibenzoate [$C_{18}H_{18}O_5$]



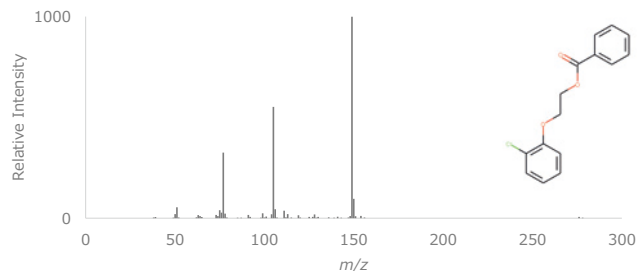
b) Benzoic acid, 2-(3-nitrophenyl)ethyl ester [$C_{15}H_{13}NO_4$]



c) 2,2'-(Ethane-1,2-diylbis(oxy))bis(ethane-2,1-diyl) dibenzoate [$C_{20}H_{22}O_6$]



d) Benzoic acid, 2-(2-chlorophenoxy)ethyl ester [$C_{15}H_{13}ClO_3$]



Two types of AI

This section describes two types of AI used in AI structural analysis.

The main AI employs Graph Convolutional Networks (GCN) [3], a type of deep learning, as its model (Fig. 4, top). GCN operates as follows: First, the machine searches structural formulas for partial structures that produce signals characteristic of a mass spectrum, and generates a lot of partial structures. Then, the machine predicts a mass spectrum based on the generated partial structural information (Fig. 4, bottom).

The specific processing is as follows: First, the structural formula is converted to graph data before being input into GCN (Fig. 5). In graph data, atoms and bonds in the structural formula are treated as nodes and edges, respectively. In addition, nodes hold information on the elemental species of atoms, and edges hold information on the type of bonds, as their feature vector. For example, a node for the carbon atom has the feature vector (1, 0, 0, ...), a node for the oxygen atom has the feature vector (0, 1, 0, ...), and a node for the nitrogen atom has the feature vector (0, 0, 1, ...).

Next, the machine performs convolutions on the structural formula that was converted to graph data as shown in the top left of Fig. 4. Through convolutions, each node sifts through and obtains information on neighboring nodes and edges. The machine learns to recognize the connection of atoms as a block by repeating convolutions.

Then, the machine performs pooling of each atom as shown in the top right of Fig. 4. This enables the machine to grasp the characteristics of the structural formula and predict a mass spectrum.

The support AI employs the traditional machine learning (regression) instead of deep learning. The machine predicts the presence or absence of 48 partial structures from ions and neutral loss based on the accurate-mass mass spectra (Fig. 6). The support AI is simple and uses dozens of coefficients. Therefore, the machine can provide prediction results and their characteristic peaks at the same time.

Accuracy evaluation of AI structural analysis

— Accuracy evaluation of EI mass spectrum prediction —

AI structural analysis uses mass spectra that are predicted from the structural formulas by the main AI. The main AI was trained using the structural formulas and mass spectra of 270,000 compounds, which account for 90% of the NIST 20 library data. During training, the weight of the main AI was optimized so that patterns of mass spectra predicted from the structural formulas match those of mass spectra in the NIST 20 library. Out of the remaining 30,000 compounds, 10,000 were allocated for validation to prevent overfitting, and 20,000 were used to evaluate the accuracy of EI mass spectrum prediction.

We evaluated the accuracy of the main AI's EI mass spectrum

Fig. 3 GUI of AI structural analysis.

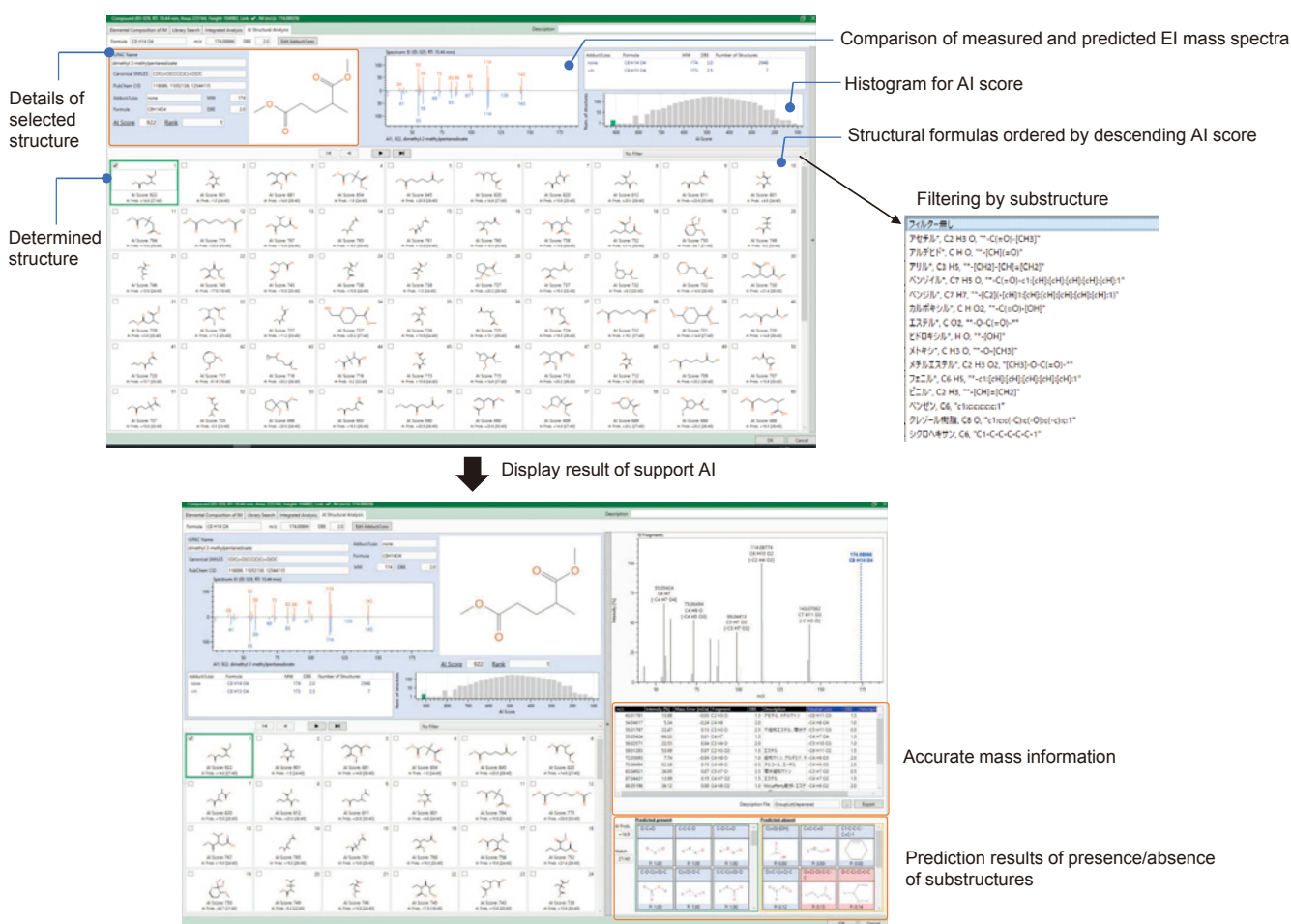


Fig. 4 Graph Convolutional Networks used in the main AI.

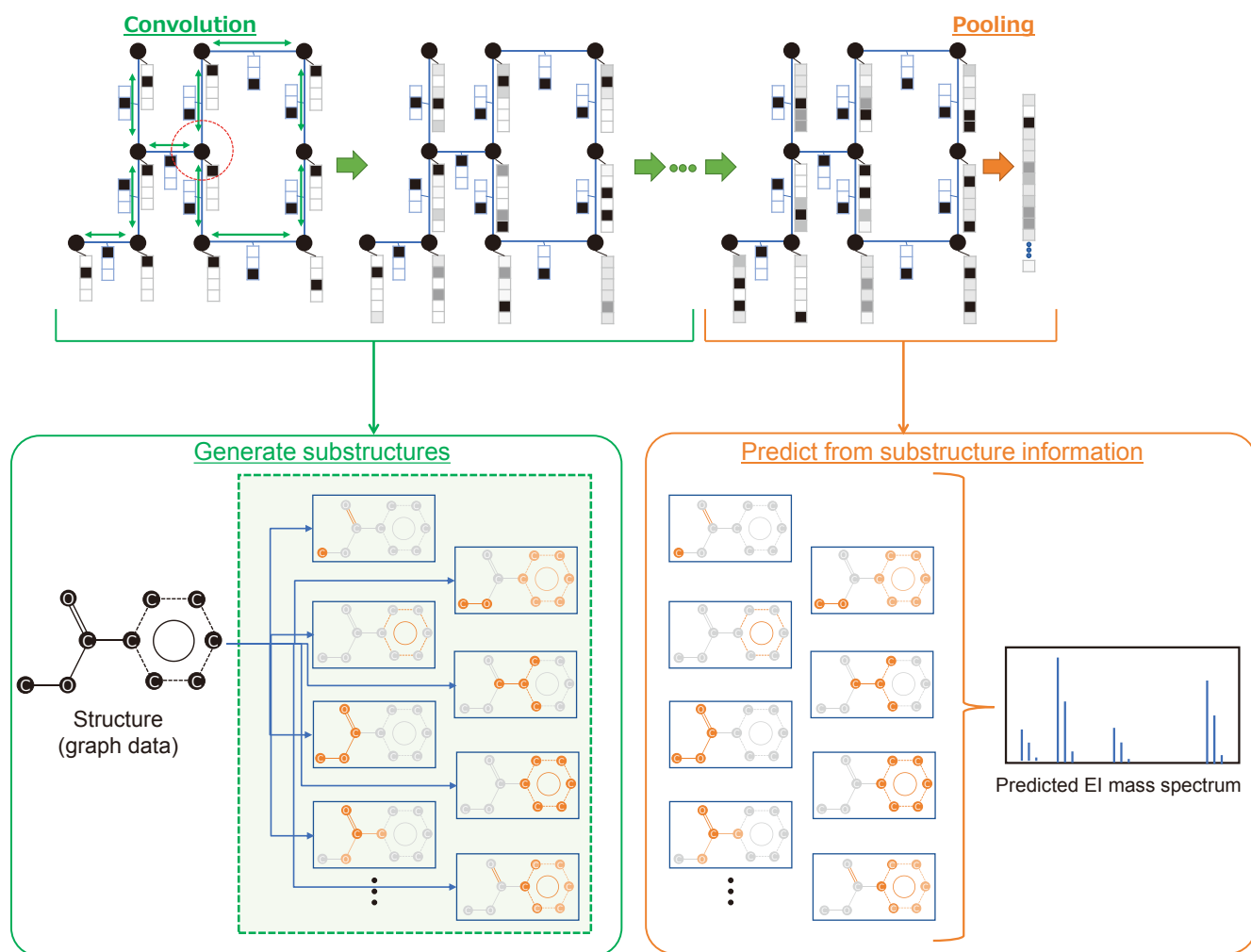
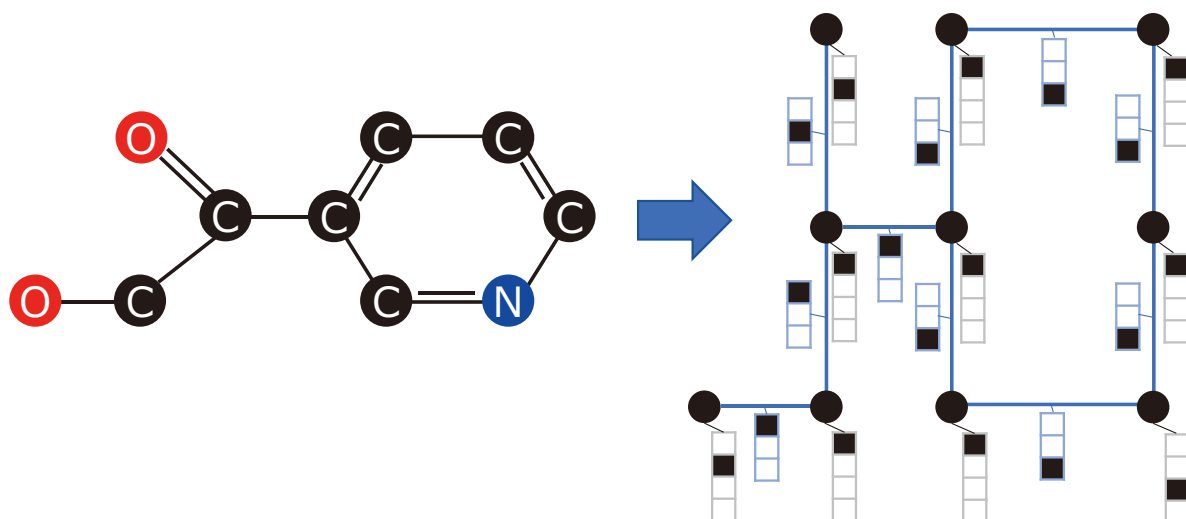


Fig. 5 Conversion of a structural formula to graph data.



prediction using 20,000 compounds that were not used in training. In the evaluation, the trained main AI predicted EI mass spectra from the structural formulas of the target compounds. We used the cosine similarity between the predicted EI mass spectrum and an EI mass spectrum registered in the NIST 20 library as the index of accuracy evaluation. A cosine similarity of 1 means the two EI mass spectra match perfectly. As the cosine similarity is closer to 0, they match less.

Figure 7 shows a histogram of cosine similarities calculated using 20,000 compounds. The histogram shows that more than 90% of the compounds had a cosine similarity of over 0.4. In addition, the 0.7-0.8 segment had the highest number of the compounds. The average cosine similarity was 0.72. We confirmed that the main AI can reproduce mass spectra with a high accuracy by predicting them from the structural formulas.

Figure 8 shows as examples the comparison between the measured and predicted EI mass spectra for each of the compounds with above-average, near-average, and below-average cosine similarities. For Benzamide, 3-methyl-N-decyl-, which had an above-average cosine similarity, the EI mass

spectrum was reproduced almost completely including mass peaks with low intensity. The reason is thought to be that this compound consists of only benzene rings, alkane chains, and amide groups, many of which are registered in the NIST 20 library. For N-Acetyl-3-(3-formyl-4-methoxyphenyl)-d-alanine methyl ester, which had a near-average cosine similarity, mass peaks with relatively high intensity were reproduced, and the overall patterns were similar. This compound has a somewhat complex structure, with multiple side chains attached to a benzene ring, compared with the structural formula of Benzamide, 3-methyl-N-decyl-. This is thought to be why a complete mass spectrum was not reproduced. For Cyclododecane, 1,5,9-tris(acetoxy)-, which had a below-average cosine similarity, the overall pattern was not well reproduced. A possible reason is that this compound includes a large 12-membered ring, and the NIST 20 library contains a small number of compounds that have this ring. This may have prevented the machine to be trained enough. However, some mass peaks, including the most intense one at m/z 43, were reproduced.

Fig. 6 Overview of the support AI.

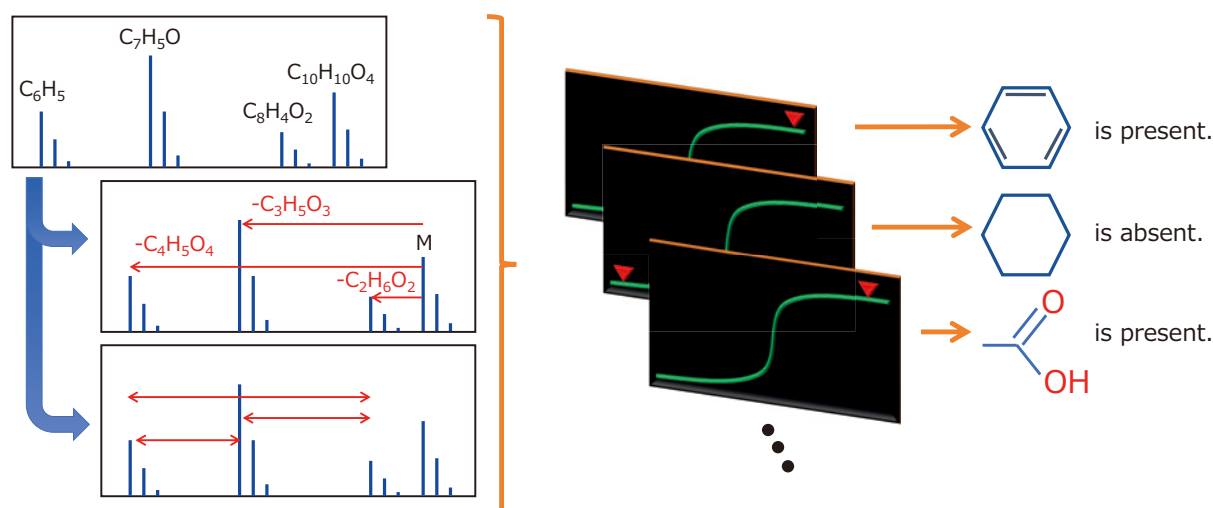


Fig. 7 Histogram of cosine similarities calculated using 20,000 compounds that were not used in training.

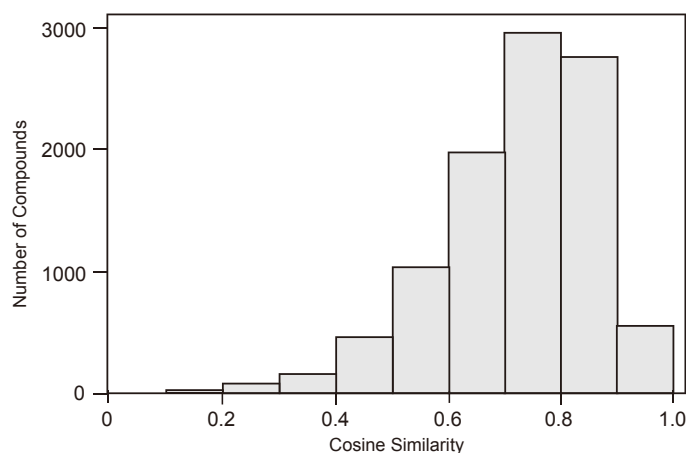
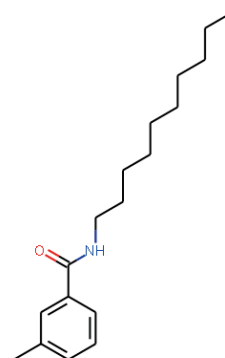
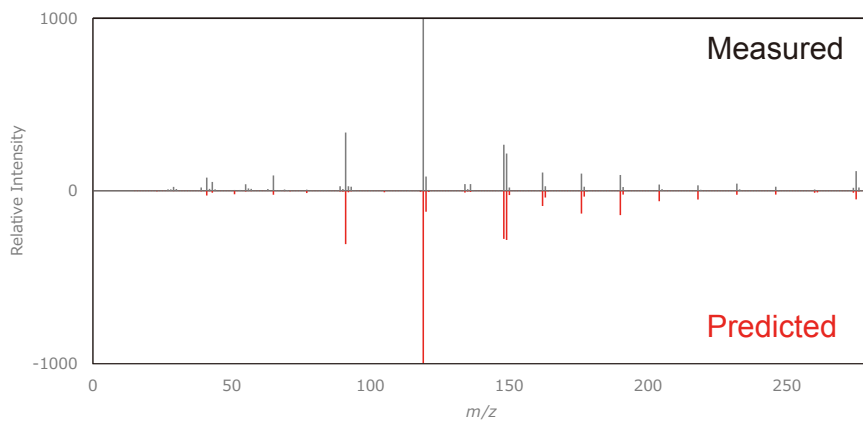
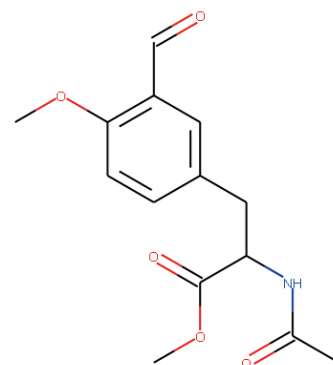
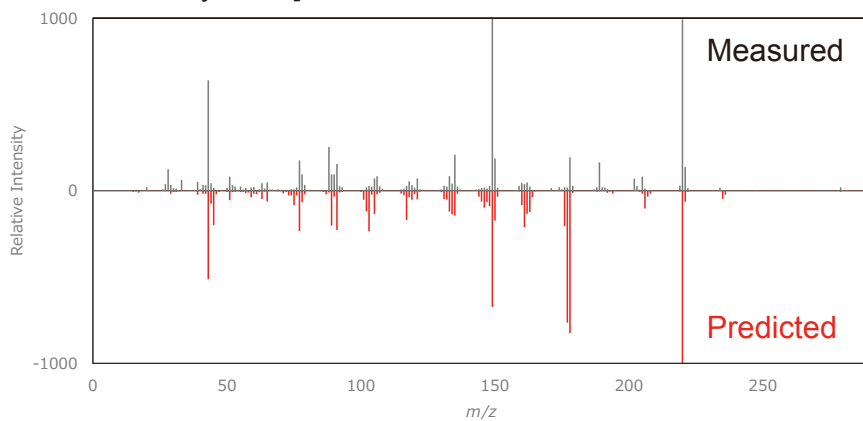


Fig. 8 Comparison between the measured and predicted EI mass spectra.

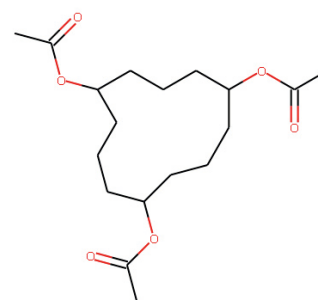
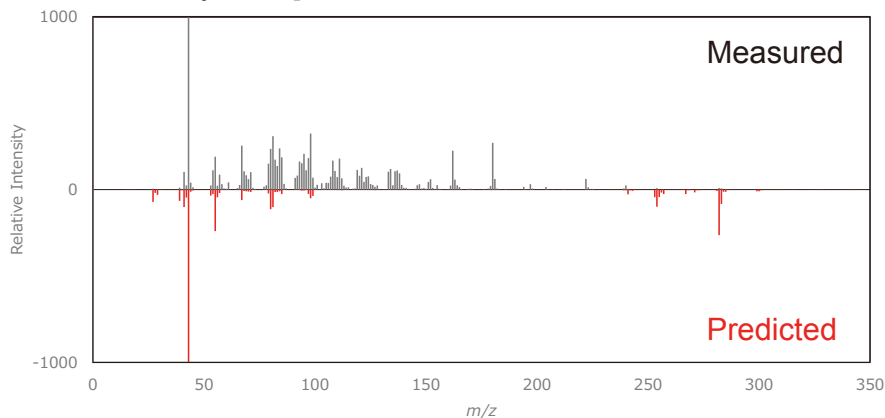
a) benzamide, 3-methyl-N-decyl-
[cosine similarity: 0.95]



b) N-Acetyl-3-(3-formyl-4-methoxyphenyl)-d-alanine methyl ester
[cosine similarity: 0.72]



c) cyclododecane, 1,5,9-tris(acetoxy)-
[cosine similarity: 0.34]



— Accuracy evaluation of structural analysis —

AI structural analysis compares EI mass spectra predicted from structural formula candidates with the actual measured EI mass spectrum to identify the structural formula. We evaluated the accuracy of this structural formula identification. The evaluation method is as follows: First, for the compounds in the NIST 20 library that were not used in training, structural formulas (compounds) that have the same molecular formula were extracted from the compound database. Next, the trained main AI predicted EI mass spectra for the correct structural formula and the extracted ones. The predicted EI mass spectra were compared with the ones registered in the NIST 20 library, and based on their cosine similarities, all the structural formulas, including the correct one, were ranked. We used the rank given by the correct structural formula among all the structural formulas as the index of accuracy evaluation. In this evaluation, to set certain criteria, we used only molecular formulas for which at least 100 compound candidates were extracted from the compound database.

Table 1 shows the results of ranking structural formulas for 14,581 compounds. The results indicate that the correct structural formula ranked top for 22% of the compounds. In addition, the correct structural formula ranked in the top 1% for 73% of the compounds. Ranking in the top 1% means that the

correct structural formula was placed within the top 10 out of 1,000 candidates. The PubChem compound database contains many compounds that have quite similar structural formulas. With taking this into consideration, this structural formula identification method is said to be highly accurate.

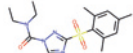
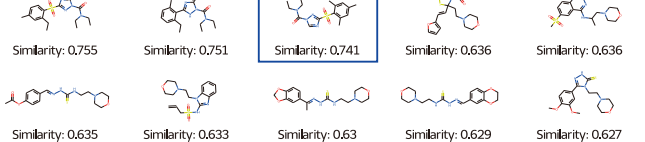
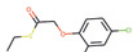
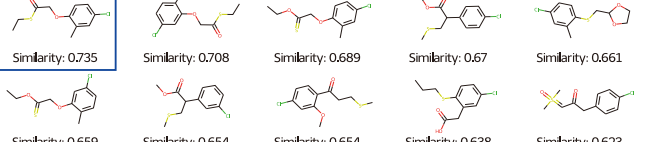

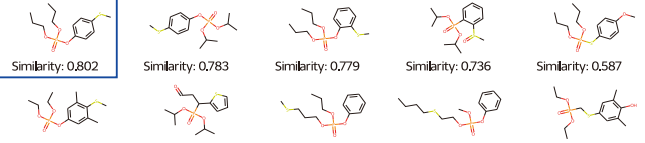
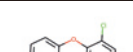
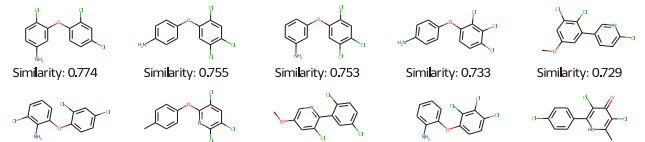

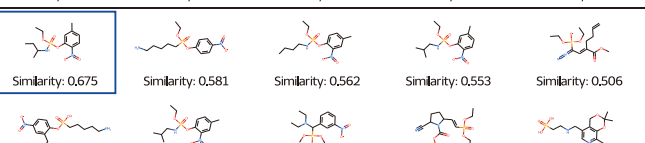
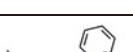
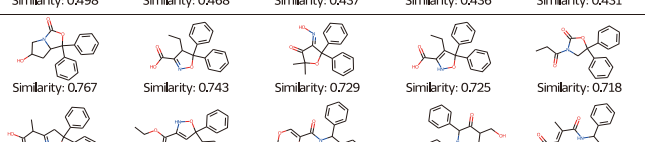
Next, we evaluated the effectiveness of this method for completely unknown compounds. We used model compounds that are not registered in the NIST 20 library to perform the evaluation. The following model compounds were used: Cafenstrole (CAS: 125306-83-4, Wako), MCPA-thioethyl (CAS: 25319-90-8, Wako), Propaphos (CAS: 7292-16-2, Wako), CNP-amino (CAS: 26306-61-6 Wako), Butamifos oxon (CAS: 56362-05-1 Wako), and Isoxadifen-ethyl (CAS: 163520-33-0, Wako).

The measured EI mass spectra for the model compounds were prepared by measuring standard samples. **Table 2** shows

Table 1 Results of accuracy evaluation on 14,581 compounds.

	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 Examples of structural analysis results.

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

the rank given by the correct structural formula, its score, and top 10 structural formulas in descending order of score for each model compound. For three compounds out of the six, the correct structural formula ranked top. For Isoxadifen-ethyl, the correct structural formula ranked lowest compared with the other five model compounds. However, it was placed 22nd out of 5,348 candidates, within top 1%. The result suggests that this structural formula identification method is effective in narrowing down the correct structural formula from many candidates. The top-ranked structural formulas for Cafenstrole, CNP-amino, and Isoxadifen-ethyl have the same size and number of rings as their correct structural formulas do, and they show considerable similarity. The results of our evaluation on these six compounds reveal that this identification method is useful in structural analysis. **Figure 9** shows the comparisons between the measured and predicted mass spectra. The measured and predicted mass spectra exhibit the same peaks with high intensity, although they are different in detailed peak intensities and distributions of mild peaks.

These results confirm that this method is effective in the structural analysis of unknown compounds.

Conclusions

Previous msFineAnalysis software features integrated analysis based on accurate mass measurement and molecular ion observation using the soft ionization method, which are

features of the JMS-T2000GC. Integrated analysis enables the identification of molecular formulas of unknown compounds. The new version, msFineAnalysis AI, is equipped with structural analysis using artificial intelligence (AI), which enables molecular formulas as well as structural formulas to be automatically obtained. msFineAnalysis AI extracts structural formula candidates based on the molecular formulas identified by integrated analysis. Then, it uses the EI mass spectra predicted from the structural formula candidates by the AI to identify the structural formula. The combination of integrated analysis and AI enables a highly efficient and accurate structural analysis. All the processes are performed automatically and offline, leading to a stable analysis.

References

- [1] Masaaki Ubukata. MultiAnalyzer – Unknown Compounds Analysis System New Gas Chromatograph Time-of-Flight Mass Spectrometer JMS-T2000GC “AccuTOF™ GC-Alpha”. *JEOL news* Vol. **56**.
- [2] Masaaki Ubukata, Yoshihisa Ueda. Development of an Integrated Analysis Method for the JMS-T200GC High Mass-Resolution GC-TOFMS by Electron Ionization and Soft Ionization Methods. *JEOL news* Vol. **54**.
- [3] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl. Neural message passing for Quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning*. 2017;70:1263-1272.

Fig. 9 Comparison between the measured and predicted mass spectra.

