Deep Learning and Mass Spectrum Based Analysis of Vocs Components

Lifeng Liu[1], Xuxia Zhao[2], Xin Lv[3], Jiankang Mu[3], Ping Cheng[1,*]

[1]School of Environmental and Chemical Engineering, Shanghai University, Shanghai 200444, China
[2]China Petroleum University (Beijing), Beijing 102200, China
[3]Zhuhai Comleader Information Science & Technology Co. Ltd., Zhuhai 519060, China

Abstract: The composition and concentration of Volatile Organic Compounds (VOCs) in the atmosphere can reflect the quality of the air, and the environmental quality changes with the quantity of these compounds. When unknown VOCs are encountered, researchers usually use gas chromatography-mass spectrometry (GC-MS) to measure and analyze them. This discriminative mode requires data analysts with a certain theoretical and practical foundation, is demanding and labor-intensive, and may also introduce errors due to the numerous steps. In order to solve these problems, we propose a deep learning and mass spectrum based method for the analysis of Vocs components. Using the deep learning technique, first, a high-quality mass spectral library is constructed as a reference library using molecular fingerprint information, and then, the sample data obtained in the GC-MS gas chromatography-mass spectrometer is preprocessed with data to extract mass spectra that can represent the VOCs components; finally, the selected candidate mass spectra are library matched with the reference library to return high matching VOCs components results. The experimental results show that the method can accurately and quickly discriminate the components of VOCs.

Keywords: Volatile Organic Compounds; Deep Learning; Mass Spectrum; Molecular Fingerprint; Library Matching

DOI: 10.57237/j.wjese.2023.02.006

1 Introduction

The unorganized emission of VOCs in industry is an important factor in polluting the air environment [1], and it is difficult to monitor and evaluate the unorganized emission of VOCs. GC-MS technology is widely used for the separation and identification of compound components, and researchers can obtain information on the molecular weight and chemical structure of compounds by using this method to analyze mass spectra [2]. However, the current GC-MS compound component identification technology has high standards for data personnel, requiring a certain theoretical and practical basis, and such personnel require a certain amount of time for training, which cannot be replicated by training in the short term. And the GC-MS instrumentation produces a large amount of data, according to incomplete statistics, a site produces a year the amount of data in 1 million, each data requires professional data analysts to spend 30s-2min to analyze, the workload is large.

Based on the above problems, we propose a deep learning and mass spectrum based analysis method of vocs components, The method first uses the molecular fingerprints of the VOCs components to construct a mass spectral library as a reference library through a neural network; Then use the mass spectrum matching model to match the similarity of the candidate mass spectrum obtained after sample pre-processing with the reference library and return the matching result sequence. To summarize, the main contributions of this work are as

*Corresponding author: Ping Cheng, chengping@shu.edu.cn

Received: July 11, 2023; Accepted: September 1, 2023; Published Online: September 8, 2023 http://www.wjese.com

follows:

1) For the problem that VOCs composition analysis requires professional data analysts and the analysis steps are cumbersome, this paper establishes a complete implementable process from data pre-processing to VOCs composition analysis based on deep learning, and proves through experiments that the method can significantly improve the efficiency of current VOCs composition monitoring while using deep learning methods no less effective than manual monitoring.
2) A lightweight mass spectrometry generation model is proposed, which can rapidly generate the corresponding mass spectra of molecules based on the molecular fingerprint information of small molecules and build a high-quality mass spectrometry library. Recall@10 achieves 90.2% on the NIST dataset.
3) Component analysis of candidate mass spectra using an MS2DeepScore twin neural network model that extracts potential feature information from the mass spectra to identify highly reliable structural matches and predicts the Tanimoto similarity scores of molecular pairs based on their fragment spectra, yielding a high-quality vector representation.

2 Related Work

2.1 Current Status of Component Monitoring of VOCs

In recent years, with the rising energy consumption, global VOCs emissions continue to grow, atmospheric environmental problems have become more and more prominent, especially ozone pollution, which has plagued atmospheric air quality in recent years, has become more and more serious [3]. Most VOCs carry an unpleasant and peculiar odor, especially the irritating VOCs such as benzene, formaldehyde and toluene can pose a great threat to people's health [4]. To improve the quality of the atmosphere and people's living environment, enhanced monitoring of VOCs is necessary.

Current VOCs monitoring requires data analysts with certain expertise to repeatedly review and confirm VOCs component information, which cannot be replicated in the short term due to the theoretical knowledge required by data analysts and the large volume of data generated by GC-MS. Therefore, automatic VOCs component analysis technology is particularly important.

2.2 Deep Learning in the Environmental Domain

With the booming development and popularity of deep learning [5], deep learning, as a new means of automatic extraction of high-dimensional nonlinear complex features, is becoming a new engine for data processing in academia and industry, and various industries are beginning to explore its applications.

In recent years, deep learning based methods have emerged in the field of environmental monitoring, and have become a new direction in the development of research in the field of environmental monitoring. For example, Xiang Li et al. [6] (2016) proposed a new spatiotemporal deep learning (STDL)-based air quality prediction method that considers spatio and temporal correlation of air quality data in the modeling process and automatically learns potential air quality features using a stacked self-encoder model, and uses the learned representations to construct regression models for air quality prediction. Zhendong Zhang et al. [7] proposed a hybrid deep learning model VMDBiLSTM, which combines variational mode decomposition (VMD) and bidirectional long short-term memory network (BiLSTM) to predict the PM2.5 in air. VMD decomposes the original PM2.5 complex time series data into multiple sub-signal components according to the frequency domain. Then, each sub-signal component is predicted separately using BiLSTM, which significantly improves the prediction accuracy.

3 Methodology

The VOCs component analysis method is divided into two sub-modules: Mass Spectrl Library Construction module and Library Matching module. The purpose of the Mass Spectrl Library Construction module is to build a library of mass spectra of VOCs in advance to serve the library matching module; the purpose of the library matching module is to search the mass spectra of VOCs to be analyzed in the already built library, and to use the compounds with good matches in the library as the component analysis results.

3.1 Mass Spectral Library Construction

The purpose of the Mass Spectral Library Construction module is to design a mass spectral map generation model that can accurately predict the GC-MS mass spectra of any volatile organic molecule, and use the model to generate a mass spectral map library as a reference library, and the generated mass spectral map library is used as part of the input for the subsequent library matching module.

The Mass Spectral Library Construction module can be divided into two steps: Molecular Fingerprint Extraction and Mass spectrometry Generation. Molecular Fingerprint Extraction is a method to convert the chemical expressions of molecules into computer-readable vectors; Mass Spectrometry Generation is a method to convert the extracted molecular fingerprints into mass spectrometry data using a mass spectrometry generation model. The architecture of Mass Spectrometry Construction is shown in Figure 1.

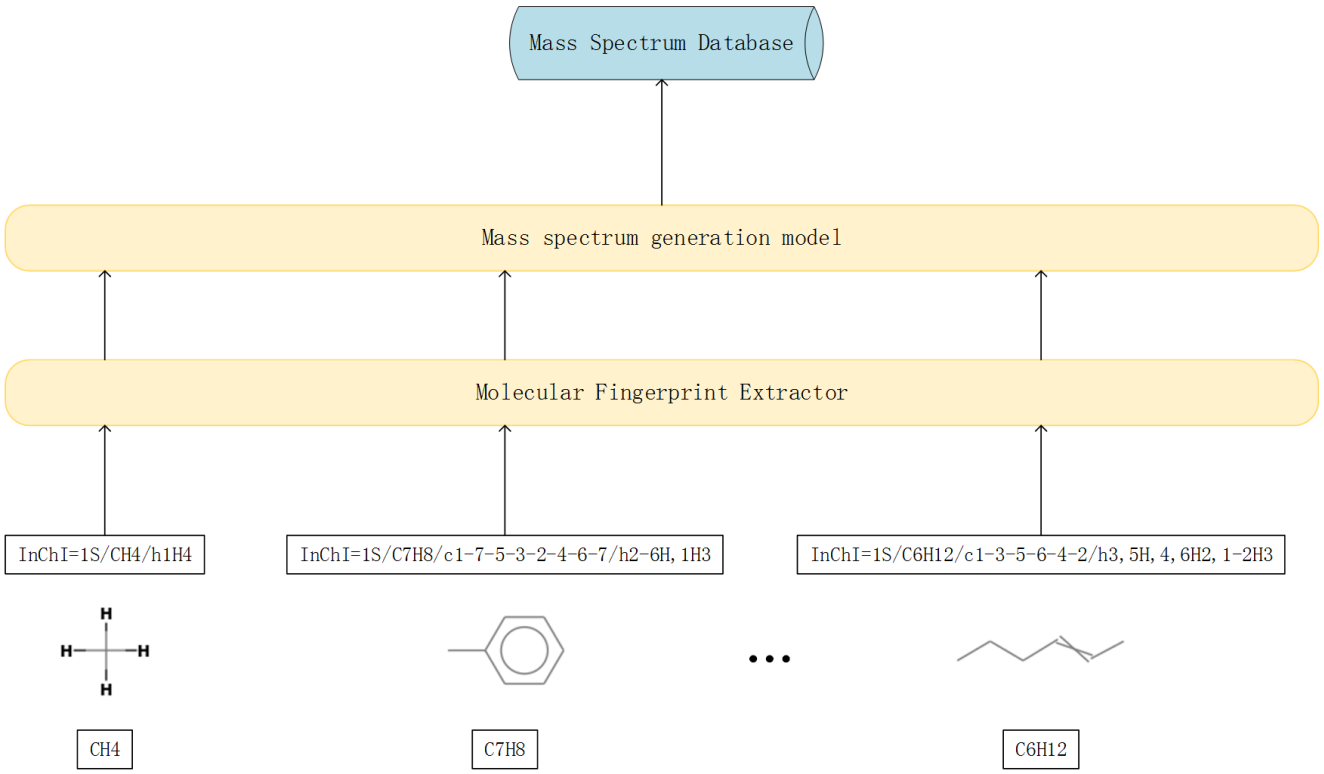

Figure 1 The architecture of mass spectrometry construction

3.1.1 Molecular Fingerprint Extraction

Molecular fingerprinting [8] refers to a method to describe the structural characteristics of molecules. It generates a set of numbers or binary coding sequences to represent the structural information of a molecule by encoding features such as chemical bonds, atomic species, and distances between atoms. Molecular fingerprinting can be used in fields such as computer simulation [9], chemoinformatics [10] and drug design [11]. Molecular fingerprints can be classified into various types [12], such as topology-based fingerprints [13], structural keys fingerprints [14], Circular fingerprints [15], etc. Based on these fingerprints, we can use various machine learning algorithms to perform molecular structure analysis and drug design tasks such as classification [16], clustering [17], and similarity comparison [18]. The following are general molecular fingerprinting extraction steps.

1) Representation of molecules: First, convert chemical molecules into a set of chemical descriptors, such as SMILES [19] (Simplified Molecular Input Line Entry System) or InChI [20] (International Chemical Identifier), etc.
2) Select fingerprinting algorithm: Select a suitable molecular fingerprinting algorithm, such as ECFP [21] (Extended-Connectivity Fingerprints), MACCS [22] (Molecular Access System), etc.
3) Fingerprint descriptor: Generate a fingerprint

descriptor based on the selected fingerprint algorithm. A fingerprint descriptor is a fixed-length binary or integer vector that represents the structure and characteristics of a molecule.

4) De-duplication: It is usually necessary to de-duplicate the fingerprint descriptors to reduce the dimensionality of the data and improve the computational efficiency.
5) Normalization: The fingerprint descriptors are normalized so that the values of each fingerprint descriptor are in the same range of values to avoid the differences between different fingerprint descriptors from adversely affecting the model training and prediction.

There are many molecular fingerprinting extraction toolkits available, such as RDKit [23], an open source package for molecular design and drug discovery, which provides a number of molecular descriptors and fingerprinting calculations, making it easier to implement the above molecular fingerprinting. In this paper, we use RDKit toolkit and choose ECFP molecular fingerprinting algorithm to extract molecular fingerprints of volatile organic molecules.

3.1.2 Mass Spectrogram Generation

Mass spectrometry [24] is an analytical technique widely used in chemistry, biology and medicine to identify and quantitatively analyze compounds. In mass spectrometry, molecules of compounds are ionized into charged ions by a mass spectrometer, and ions with different mass-to-charge ratios are separated by the action of a magnetic or electric field, and finally these ions are focused onto an ion detector for detection, and the image between ion signal intensity and mass-to-charge ratio is obtained as a mass spectrogram.

Mass spectra usually have two axes, one is the mass-to-charge ratio axis and the other is the ion signal axis. The mass-to-charge axis is the ratio of the mass to the charge of the ion, which is the relative molecular mass of the ion. The ion signal axis is the intensity of the signal produced by each ion on the ion detector.

The mass spectrometry generation is viewed as a multiple regression problem, where the molecular fingerprint vectors are fed into the model and the final output is the intensity value corresponding to each mass-to-charge ratio in the mass spectrometry. The model architecture is selected from a multilayer perceptron for feature extraction, and the model architecture diagram is shown in Figure 2.

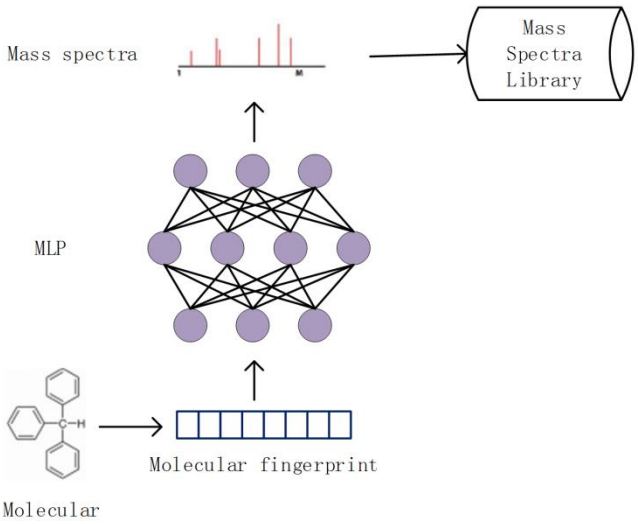

Figure 2 Mass spectrum generation model

Model Training. The mass spectrum generation model uses a modified mean squared loss[25] as a loss function to evaluate the difference between the predicted mass spectrum and the true mass spectrum. The loss function is formulated as follows:

$$L(I,\hat{I}) = \sum_{k=1}^{M(x)} \left(\frac{m_k I_k^{0.5}}{\left\| \sum_{k=1}^{M} (m_k I_k^{0.5})^2 \right\|} - \frac{m_k \hat{I}_k^{0.5}}{\left\| \sum_{k=1}^{M} (m_k \hat{I}_k^{0.5})^2 \right\|} \right)^2 \quad (1)$$

where I is the real mass spectrum data, and $\hat{I}$ is the predicted mass spectrum data, and $M(x)$ is the mass of the molecule. The parameters of the model were optimized using the Adam [26] optimizer using the stochastic gradient descent (SGD) method.

3.2 Library Matching

The purpose of the library matching module is to design a similarity matching model to retrieve the mass spectra with high matching scores from the mass spectral library, so as to obtain the components of the candidate mass spectra. The library matching module is divided into two sub-modules: data preprocessing module and mass spectrum matching module. The purpose of the data preprocessing module is to standardize the data collected from the instrument and extract the valid mass spectra from the standard data. The mass spectrum matching module uses a MS2DeepScore [27] model to retrieve the mass spectra obtained after data preprocessing in a reference library and return their component information.

3.2.1 Data Preprocessing

The raw data obtained from GCMS instrumentation usually has a lot of noise due to the environment and instrumentation, resulting in low quality raw data, and low quality data usually has a negative impact on data mining and model training [28]. GC-MS raw data are composed of two dimensions: mass-to-charge ratio (m/z) and retention time. Each row in the raw data represents the intensity of different mass-to-charge ratios at the same retention time, and the mass-to-charge ratio as the horizontal axis and the intensity of the mass-to-charge ratio as the vertical axis can be obtained for the mass spectrum at this retention time. Each column represents the change in intensity for different retention times at the same mass-to-charge ratio. Using retention time as the horizontal axis and intensity as the vertical axis one can obtain an ion chromatogram for that mass-to-charge ratio.

Typically, the instrument is scanned five times per second and the mass-to- charge ratio at the same moment is recorded and summed by the mass spectrometry software. The graph with the retention time on the horizontal axis and the sum of all mass-to-charge ratios on the vertical axis is called Total Ion Chromatography (TIC). Data preprocessing is mainly done by processing the TIC to obtain mass spectral data valid at a certain retention time for library matching tasks.

The data preprocessing includes the following steps: denoise, baseline correction, and find peak.

1) Denoise

The purpose of denoise is to reduce the influence of instrument state changes [29]. The commonly used denoise methods include Fourier transform, discrete wavelet transform, Savitzky-Golay filter, etc. The Fourier transform has good performance for stable signals and performs poorly for unstable signals. The discrete wavelet transform can decompose and reconstruct the signal at multiple scales. The wavelet function is suitable for removing homoskedastic noise but not heteroskedastic noise. Savitzky-Golay filter is a filtering method based on a local polynomial least squares fit in the time domain. The most important feature of this filter is that the shape and width of the signal can be ensured to be constant while filtering out the noise.

2) Baseline corrction

Baseline removal is extremely important because poor baseline correction may lead to corruption of the data, affecting quantification and data analysis, and it can also affect later work [30]. There are different baseline correction methods available, such as Alternating Least Squares (ALS) [31], Sensitive nonlinear iterative peak (SNIP) [32] and TopHat [33]. It is worth noting that a good baseline correction is a great improvement for the accuracy of the later peak finding algorithms.

3) Find peak

The mass spectral data that can be analyzed for components are determined by finding the peaks of the total ion chromatogram. Commonly used peak finding algorithms include the use of local maxima, continuous wavelet transform, and first-order differencing. With the development of deep learning in the field of environmental science and signal processing, neural network based peak finding algorithms are becoming popular [34].

3.2.2 Mass Spectral Matching

Mass spectral matching can be thought of as calculating the similarity between mass spectra. There are many ways to measure the similarity of mass spectra, but the current methods have a limitation: the method used to measure the similarity between two mass spectra is usually based on the structural similarity of the compounds. For example, in practice, molecular fingerprints of molecules are usually calculated using similarity functions such as cosine similarity, yet molecular fingerprints are calculated from the chemical structure, which only accounts for a fraction of the complexity of the compound. We therefore propose to use an end-to-end approach that relies only on deep learning models to predict the structural similarity of compounds directly from mass spectra, thus avoiding the traditional computational approach by comparing molecular fingerprint.

We choose a siamese network structured mass spectral matching model (MS2DeepScore) with the aim of predicting the structural similarity of the two input mass spectra, and the Tanimoto similarity is used as the label during training. The Tanimoto similarity is shown in equation (2).

$$T = c/(a + b - c) \tag{2}$$

where a and b denote the number of features in the two molecules, respectively, and c denotes the number of features common to these two molecules. Usually, the value of Tanimoto similarity ranges from 0 to 1, and higher values indicate that the two molecules are more similar.

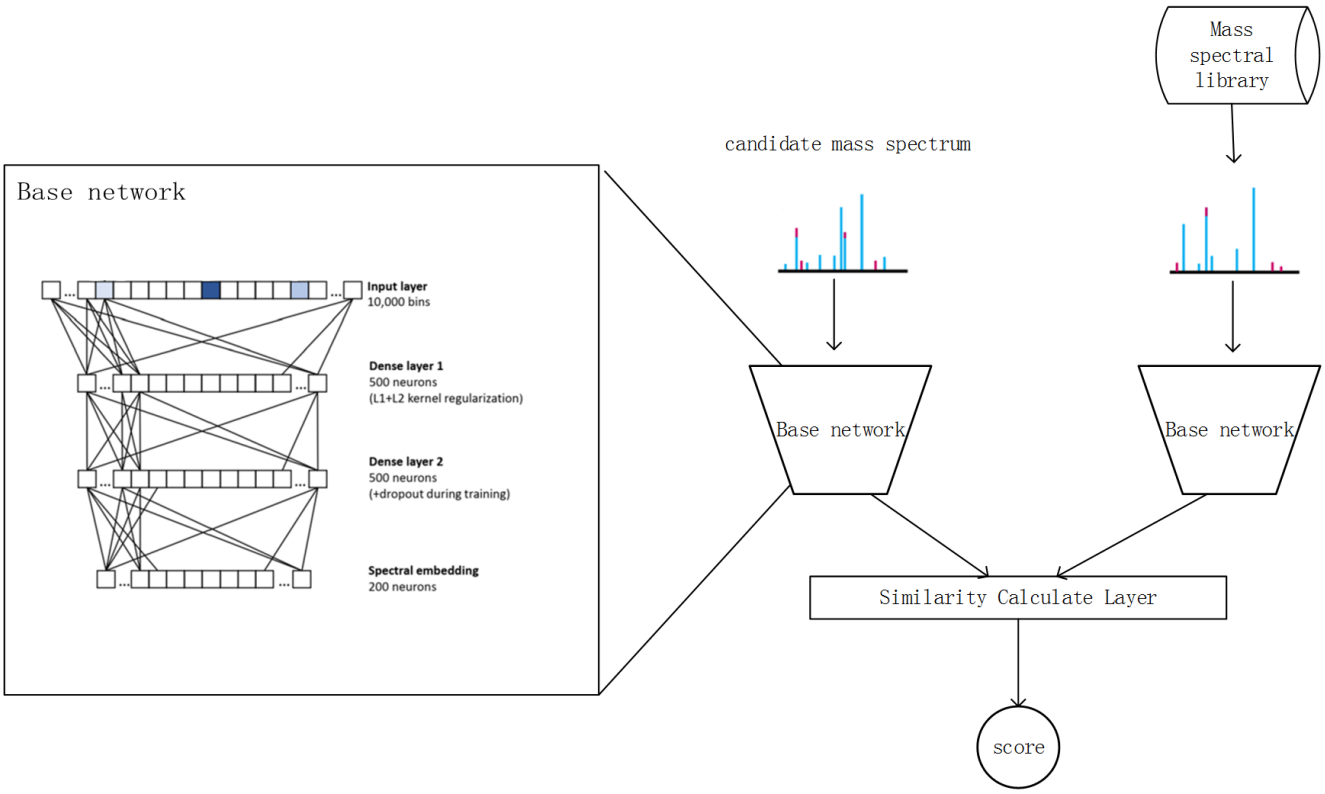

Figure 3 Mass spectrum matching model

The mass spectrum matching model is shown Figure 3. The input two mass spectra of the MS2DeepScore model are first passed through the two base networks separately to obtain their respective vector representations after feature extraction, and again, due to the simple structure of the mass spectral data, a fully connected layer is used as the network architecture of the base network. After obtaining the mass spectrum vector representations, the two mass spectrum vectors will interact in the similarity calculation layer to obtain the structural similarity of the corresponding molecules of the mass spectra.

Model training. The model training process uses mean square error (MSE) as the loss function, The loss function is formulated as follows:

$$\mathrm{MSE}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \quad (3)$$

Where Y is the real Tanimoto similarity, and $\hat{Y}$ is the similarity score predicted by the model. Stochastic gradient descent was used and Adam was chosen as the optimizer to optimize the mean squared loss function. We used a training batch is 32 and learning rate is 0.001.

4 Experiments

4.1 Experimental Environment

The experiments were conducted using the Tensorflow open source framework and CUDA-GPU acceleration scheme under Ubantu 20.04 operating system, with NVIDIA RTX A6000 for acceleration of the training and other major hardware such as AMD Ryzen Threadripper PRO 3995WX 64-Cores and 1T SSD.

4.2 Dataset

1) NIST Mass Spectral Library [35]. The NIST Mass Spectral Library is a fully evaluated collection of electron ionization (EI) and MS/MS mass spectra containing chemical and gas chromatography data. The NIST Mass Spectrometry Library contains more than 1 million mass spectra, including 306,000 EI spectra and 1,320,000 tandem MS/MS spectra for 350,000 compounds. The library is the result of more than 30 years of comprehensive evaluation and expansion of the world's most widely used mass

spectrometry reference library by a team of experienced mass spectrometry experts at the National Institute of Standards and Technology (NIST), and every spectrum has been tested for correctness. The mass spectrum generation model was trained using 240,942 mass spectra data from the NIST Mass Spectral Main Library.

2) GNPS spectral library [36]. GNPS (Global Natural Products Social) is a visualization network system for natural products based on secondary mass spectral data. The MS2DeepScore model was trained using 210407 MS/ MS spectra collected in GNPS.

4.3 Metric

1) Recall. The recall is calculated as the ratio between the number of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect Positive samples. The higher the recall, the more positive samples detected.

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}}$$

Where TP is true positive and FN is false negative.

2) Accuracy. The proportion of the number of correct predictions to the total number of positive and negative cases.

$$\text{ACC} = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

4.4 Analysis of Experiments and Results

In this section. The experimental results of the two steps of the VOCs components analysis method are described respectively.

1) Mass spectrum generation.

Several traditional machine learning models were selected on the NIST dataset to compare with the multilayer perceptron based mass spectrometry generation model, and the experimental results were evaluated using Recall@10 as an evaluation metric to prove the effectiveness of the mass spectrum generation model. The methods used for the experimental comparison are: linear regression [37], and decision tree regression [38]. The results are shown in the following Table 1.

Table 1 Experimental results of mass spectrum generation comparison

algorithm	**Recall@10**
linear regression	22.9
Decision tree regression	33.3
multi-layer perceptron	90.2

As can be seen from Table 1, the Recall@10 results of linear regression and decision tree regression are poor, However, the multi-layer perceptron model Recall@10 can reach 90.2% in the same situation. By analyzing the positive samples not recalled by linear regression and decision tree regression and the negative samples with wrong recall, the reasons for the model errors are as follows: the difference in features between the negative samples with wrong recall and the positive samples is not obvious, and there is some similarity in the extracted molecular fingerprint vectors, and the model cannot accurately capture the key feature information of the molecular fingerprint. In contrast, the multilayer perceptron model with strong feature representation can better capture the key feature information and thus perform better.

2) Library matching.

To test the effectiveness of the MS2DeepScore model, three model algorithms are used in this paper for training and testing under the same conditions: cosine similarity [39], modified cosine similarity, and Spec2vec [40].

The modified cosine similarity function is an enhancement of the cosine similarity function that alleviates the problem that the traditional cosine similarity only responds to directional information and not to positional information. Spec2Vec is a mass spectral similarity scoring algorithm based on the Word2Vec algorithm in natural language processing that learns mass fragmentation and neutral loss in MS/MS mass spectra.

Since the distribution of the data set collected from the GNPS spectral library is unbalanced and most of the mass spectra have low Tanimoto similaritys, to avoid that the mass spectral pairs with low Tanimoto similarity matching results will have a great influence on the results, a box splitting operation was used to divide all possible mass spectral pairs into 10 equally spaced boxes of Tanimoto fractions. For the definition of correlation, we considered that a mass spectrum pair is correlated when its valley principal similarity exceeds 0.6, otherwise it is not correlated. The check-all rate and check-accuracy rate of different algorithm models were also analyzed by the precision versus recall curves, and the precision versus

recall curves are shown in Figure 3.

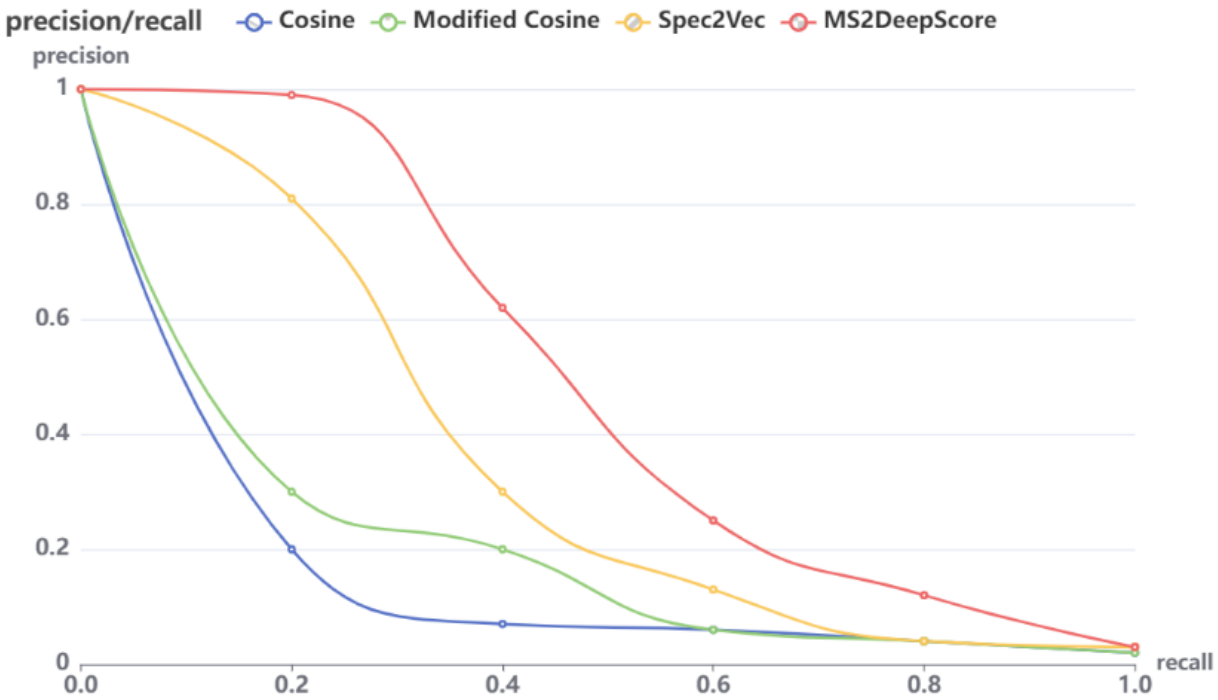

Figure 4 Precision/recall

From Figure 4, it can be seen that the MS2DeepScore model has better combined accuracy and recall results, followed by Spec2Vec and then modified cosine similarity, with the cosine similarity algorithm performing the worst. Based on the data analysis, it is found that the plain cosine similarity function and the modified cosine similarity function will perform better for mass spectrogram data with many identical peaks, while the results tend to be less satisfactory when the peaks of the two mass spectrogram data differ. Spec2Vec is an unsupervised machine learning method that has the advantage of being able to train without labeling and learn the distribution of the data from the data, i.e., the relationship between peaks in a mass spectrum from peak co-occurrence. This characteristic of Spec2Vec also leads to the inability to learn the structural information of the molecules corresponding to the mass spectra. MS2DeepScore is a supervised learning deep learning model that uses the predicted values of mass spectra on similarity with the corresponding molecule's Tanimoto similaritys to learn the distribution of mass spectral data while also uniting the structural information of molecules for training. As can be seen from Figure 3, the MS2DeepScore model that incorporates the use of deep learning network with molecular structure information has improved in precision and recall compared to the traditional machine learning methods of cosine similarity method and Spec2Vec.

3) Case Study

By gathering data in a real-world setting, this part puts the technique put forth in this paper to the test. The collection consists of 550 data obtained from meters bearing the Agilent brand, with a total of ten typical VOC components. Table 2 displays the dataset's details.

Table 2 Agilent meters collect VOCs data details

VOCs Component	Number
Propylene	59
chloroethylene	45
Butadiene	57
Acetone	52
n-Hexane	58
Benzene	49
n-Heptane	68
Chlorobenzene	64
Ethylbenzene	46
o-Xylene	52

The data collected above was preprocessed with data and then it was input to the trained MS2DeepScore model for library matching together with the constructed mass spectrometry library, and finally the predicted component results were obtained and then compared with the manually labeled real labels, and the evaluation metrics are shown in Table 3.

Table 3 Accuracy of VOCs components

VOCs Component	Number
Propylene	59
chloroethylene	45
Butadiene	57
Acetone	52
n-Hexane	58
Benzene	49
n-Heptane	68
Chlorobenzene	64
Ethylbenzene	46
o-Xylene	52

Table 3 shows that the method has a good recognition effect on different VOCs components, demonstrating that deep learning technology can effectively extract and retrieve the feature information of the original data from the VOCs components based on this feature information, and that this method has a stronger generalization ability. However, the method is not yet able to achieve fully correct recognition accuracy for the target data.

5 Conclusion

In this paper, we analyze the characteristics of VOCs data and use a deep learning network model to automatically identify VOCs components. The proposed method for component analysis of VOCs based on deep learning and mass spectrum firstly obtains component results by constructing a high-quality mass spectral library as a reference library for library matching, and then using a library matching model to search the mass spectra of the components to be predicted in the reference library. The experimental results show that the use of deep learning modeling has a good automatic identification capability for VOCs components, while achieving automated intelligent monitoring of atmospheric VOCs, which greatly reduces the time consumed by data reviewers and provides a new way of thinking for atmospheric environmental protection.

In the Mass Spectral Library Construction module, the MLP is used as the backbone of Mass Spectrogram Generation in consideration of the efficiency, but if there is too much information of the mass spectral library to be constructed, the problem of insufficient ability of this model for the generation of mass spectra will be slowly exposed, and find a generation model with both generation capability and efficiency will be the direction of our subsequent work. In addition, current methods still need to do some data preprocessing tasks and the data preprocessing phase is still time-consuming, so exploring a solution that does not require data preprocessing is also one of the future research directions.

References

[1] Simayi M, Shi Y, Xi Z, et al. Emission trends of industrial VOCs in China since the clean air action and future reduction perspectives [J]. Science of The Total Environment, 2022, 826: 153994.

[2] Gallego E, Roca F J, Perales J F, et al. Characterization and determination of the odorous charge in the indoor air of a waste treatment facility through the evaluation of volatile organic compounds (VOCs) using TD–GC/MS [J]. Waste management, 2012, 32 (12): 2469-2481.

[3] Wu W, Xue W, Zheng Y, et al. Diurnal regulation of VOCs may not be effective in controlling ozone pollution in China [J]. Atmospheric Environment, 2021, 256: 118442.

[4] Maung T Z, Bishop J E, Holt E, et al. Indoor air pollution and the health of vulnerable groups: a systematic review focused on particulate matter (PM), volatile organic compounds (VOCs) and their effects on children and people with pre-existing lung disease [J]. International Journal of Environmental Research and Public Health, 2022, 19 (14): 8752.

[5] Dong S, Wang P, Abbas K. A survey on deep learning and its applications [J]. Computer Science Review, 2021, 40: 100379.

[6] Li X, Peng L, Hu Y, et al. Deep learning architecture for air quality predictions [J]. Environmental Science and Pollution Research, 2016, 23 (22): 22408-22417.

[7] Zhang Z, Zeng Y, Yan K. A hybrid deep learning technology for PM2.5 air quality forecasting [J]. Environmental Science and Pollution Research, 2021, 28 (29): 39409-39422.

[8] Willett P. Similarity searching using 2D structural fingerprints [J]. Chemoinformatics and computational chemical biology, 2011: 133-158.

[9] Murray M. Sampling bias in the molecular epidemiology of tuberculosis [J]. Emerging infectious diseases, 2002, 8 (4): 363.

[10] Xue L, Bajorath J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening [J]. Combinatorial chemistry & high throughput screening, 2000, 3 (5): 363-372.

[11] Duch W, Swaminathan K, Meller J. Artificial intelligence approaches for rational drug design and discovery [J]. Current pharmaceutical design, 2007, 13 (14): 1497-1508.

[12] Yang J, Cai Y, Zhao K, et al. Concepts and applications of chemical fingerprint for hit and lead screening [J]. Drug Discovery Today, 2022: 103356.

[13] Yang W, Hu J, Stojmenovic M. NDTC: A novel topology-based fingerprint matching algorithm using N-layer Delaunay triangulation net check [C] // 2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2012: 866-870.

[14] Schietgat L, Cuissart B, Lepailleur A, et al. Comparing chemical fingerprints for ecotoxicology [J]. Proceedings of the 6i èmes Journ ées Nationales de Ch émoinformatique, 2013: 20-21.

[15] Glen R C, Bender A, Arnby C H, et al. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME [J]. IDrugs, 2006, 9 (3): 199.

[16] Li S, Zhang L, Feng H, et al. MutagenPred-GCNNs: a graph convolutional neural network-based classification model for mutagenicity prediction with data-driven molecular fingerprints [J]. Interdisciplinary Sciences: Computational Life Sciences, 2021, 13: 25-33.

[17] Butina D. Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets [J]. Journal of Chemical Information and Computer Sciences, 1999, 39 (4): 747-750.

[18] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening [J]. Expert opinion on drug discovery, 2016, 11 (2): 137-148.

[19] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules [J]. Journal of chemical information and computer sciences, 1988, 28 (1): 31-36.

[20] Heller S, McNaught A, Stein S, et al. InChI-the worldwide chemical structure identifier standard [J]. Journal of cheminformatics, 2013, 5 (1): 1-9.

[21] Rogers D, Hahn M. Extended-connectivity fingerprints [J]. Journal of chemical information and modeling, 2010, 50 (5): 742-754.

[22] Kuzminykh D, Polykovskiy D, Kadurin A, et al. 3D molecular representations based on the wave transform for convolutional neural networks [J]. Molecular pharmaceutics, 2018, 15 (10): 4378-4385.

[23] Bento A P, Hersey A, F élix E, et al. An open source chemical structure curation pipeline using RDKit [J]. Journal of Cheminformatics, 2020, 12: 1-16.

[24] De Hoffmann E, Stroobant V. Mass spectrometry: principles and applications [M]. John Wiley & Sons, 2007.

[25] Wei JN, Belanger D, Adams RP, Sculley D. Rapid Prediction of Electron-Ionization Mass Spectrometry Using Neural Networks. ACS Cent Sci. 2019 Apr 24; 5 (4): 700-708. doi: 10.1021/acscentsci.9b00085. Epub 2019 Mar 19. PMID: 31041390; PMCID: PMC6487538.

[26] Liu Z, Shen Z, Li S, et al. How do adam and training strategies help bnns optimization [C]//International Conference on Machine Learning. PMLR, 2021: 6936-6946.

[27] Huber F, van der Burg S, van der Hooft J J J, et al. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra [J]. Journal of cheminformatics, 2021, 13 (1): 1-14.

[28] Smolinska A, Hauschild A C, Fijten R R R, et al. Current breathomics—a review on data pre-processing techniques and machine learning in metabolomics breath analysis [J]. Journal of breath research, 2014, 8 (2): 027105.

[29] Bader S. Identification and quantification of peaks in spectrometric data [J]. Technical University of Dortmund: Dortmund, Germany, 2008.

[30] Bunkowski A. MCC-IMS data analysis using automated spectra processing and explorative visualisation methods [D]. Universit ätsbibliothek Bielefeld, Hochschulschriften, 2012.

[31] Gemperline P J, Cash E. Advantages of soft versus hard constraints in self-modeling curve resolution problems. Alternating least squares with penalty functions [J]. Analytical Chemistry, 2003, 75 (16): 4236-4243.

[32] Morháč M. An algorithm for determination of peak regions and baseline elimination in spectroscopic data [J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2009, 600 (2): 478-487.

[33] Statham P J. Deconvolution and background subtraction by least-squares fitting with prefiltering of spectra [J]. Analytical Chemistry, 1977, 49 (14): 2149-2154.

[34] Melnikov A D, Tsentalovich Y P, Yanshole V V. Deep learning for the precise peak detection in high-resolution LC–MS data [J]. Analytical chemistry, 2019, 92 (1): 588-592.

[35] Lemmon E W, Huber M L, McLinden M O. NIST standard reference database 23 [J]. Reference fluid thermodynamic and transport properties (REFPROP), version, 2010, 9.

[36] Wang M, Carver J J, Phelan V V, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking [J]. Nature biotechnology, 2016, 34 (8): 828-837.

[37] Montgomery D C, Peck E A, Vining G G. Introduction to linear regression analysis [M]. John Wiley & Sons, 2021.

[38] Timofeev R. Classification and regression trees (CART) theory and applications [J]. Humboldt University, Berlin, 2004, 54.

[39] Zhu Y, Lesch A, Li X, et al. Rapid Noninvasive Skin Monitoring by Surface Mass Recording and Data Learning [J]. Jacs Au, 2021, 1 (5): 598-611.

[40] Huber F, Ridder L, Verhoeven S, et al. Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships [J]. PLoS computational biology, 2021, 17 (2): e1008724.