


## Article

# Deep Learning-Based Method for Compound Identification in NMR Spectra of Mixtures

Weiwei Wei <sup>1,†</sup>, Yuxuan Liao <sup>2,†</sup>, Yufei Wang <sup>2</sup>, Shaoqi Wang <sup>2</sup>, Wen Du <sup>1</sup>, Hongmei Lu <sup>2</sup>, Bo Kong <sup>1,\*</sup>,  
Huawu Yang <sup>3,\*</sup> and Zhimin Zhang <sup>2,\*</sup> 

<sup>1</sup> Technology Center, China Tobacco Hunan Industrial Co., Ltd., Changsha 410014, China; weiw0216@hngyotobacco.com (W.W.); duwen0621@hngyotobacco.com (W.D.)

<sup>2</sup> College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, China; 212311021@csu.edu.cn (Y.L.); wyf991314@163.com (Y.W.); 8202180822@csu.edu.cn (S.W.); hongmeilu@csu.edu.cn (H.L.)

<sup>3</sup> Flavors and Fragrances Research Institute, Technology Center, China Tobacco Hunan Industrial Co., Ltd., Changsha 410014, China

\* Correspondence: kongb0706@hngyotobacco.com (B.K.); yanghw0918@hngyotobacco.com (H.Y.); zmzhang@csu.edu.cn (Z.Z.)

† These authors contributed equally to this work.

**Abstract:** Nuclear magnetic resonance (NMR) spectroscopy is highly unbiased and reproducible, which provides us a powerful tool to analyze mixtures consisting of small molecules. However, the compound identification in NMR spectra of mixtures is highly challenging because of chemical shift variations of the same compound in different mixtures and peak overlapping among molecules. Here, we present a pseudo-Siamese convolutional neural network method (pSCNN) to identify compounds in mixtures for NMR spectroscopy. A data augmentation method was implemented for the superposition of several NMR spectra sampled from a spectral database with random noises. The augmented dataset was split and used to train, validate and test the pSCNN model. Two experimental NMR datasets (flavor mixtures and additional flavor mixture) were acquired to benchmark its performance in real applications. The results show that the proposed method can achieve good performances in the augmented test set (ACC = 99.80%, TPR = 99.70% and FPR = 0.10%), the flavor mixtures dataset (ACC = 97.62%, TPR = 96.44% and FPR = 2.29%) and the additional flavor mixture dataset (ACC = 91.67%, TPR = 100.00% and FPR = 10.53%). We have demonstrated that the translational invariance of convolutional neural networks can solve the chemical shift variation problem in NMR spectra. In summary, pSCNN is an off-the-shelf method to identify compounds in mixtures for NMR spectroscopy because of its accuracy in compound identification and robustness to chemical shift variation.

**Keywords:** deep learning; identification; NMR; mixture analysis



**Citation:** Wei, W.; Liao, Y.; Wang, Y.; Wang, S.; Du, W.; Lu, H.; Kong, B.; Yang, H.; Zhang, Z. Deep Learning-Based Method for Compound Identification in NMR Spectra of Mixtures. *Molecules* **2022**, *27*, 3653. <https://doi.org/10.3390/molecules27123653>

Academic Editor: Lukasz Komsta

Received: 18 May 2022

Accepted: 5 June 2022

Published: 7 June 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The main technologies for analyzing mixtures consisting of small molecules include nuclear magnetic resonance (NMR) and mass spectrometry (MS). Each has its advantages and disadvantages [1] concerning sensitivity and reproducibility. NMR is reproducible and nondestructive, but its sensitivity is relatively poor, whereas MS is highly sensitive, it shows low reproducibility [2]. Recently, technologies such as low temperature probes and high-field NMR spectrometers have achieved large improvements in the sensitivity of NMR [3]. By probing local magnetic fields surrounding specific atomic nuclei, NMR can measure signals with the electronic structures and functional groups information from molecules. Therefore, NMR is particularly useful for identifying the structures of small molecules [4,5]. Furthermore, the advantages of nondestructive, unbiased and easy sample preparation make NMR spectroscopy widely used in many fields, including chemistry [6],

metabolomics [7–9], drug discovery [10–12], food [13,14], natural products [15,16], flavors [17], environments [18], forensic [19], cultural heritage [20], etc. It is mainly used for three tasks: identification, verification and quantification [21]. For the identification and verification tasks, it is necessary to evaluate the similarities between NMR spectra. Similarity methods for NMR spectra can be divided into two categories (chemical shift similarity and spectral similarity), according to their inputs [22]. The inputs of chemical shift similarity-based methods are peak tables, which are commonly used for searching NMR spectral databases [23–25]. The inputs of spectral similarity-based methods are full NMR spectra, which are used to calculate the vector-based similarity or distance [26]. The traditional similarity methods may fail in real applications because of chemical shift variations [27]. To avoid the chemical shift variation problem in NMR spectroscopy, the commonly used strategies are binning [28–30], shift alignment [31–33] and shift-insensitive similarities [34,35]. They are successfully used for the identification of pure substances and the verification of complex samples. In many cases, mixtures of small molecules are common in chemistry, and they can be analyzed by NMR spectroscopy directly without further separation and purification [36,37]. Due to the signal overlap and interferences in NMR spectra, the previously mentioned similarity methods may fail in analyzing the mixtures. Therefore, the identification of components in mixtures is highly challenging. One category of methods acquires 2D NMR spectra (e.g., TOCSY) and decomposes them using deconvolution methods to obtain the 1D NMR spectra of components for identification [38–42]. The other category of methods uses statistics, chemometrics and pattern recognition algorithms to identify components directly from NMR signals [43–47]. Furthermore, the combining MS, NMR and algorithm delivers good results for the reliable identification of the constituents in complex mixtures [48].

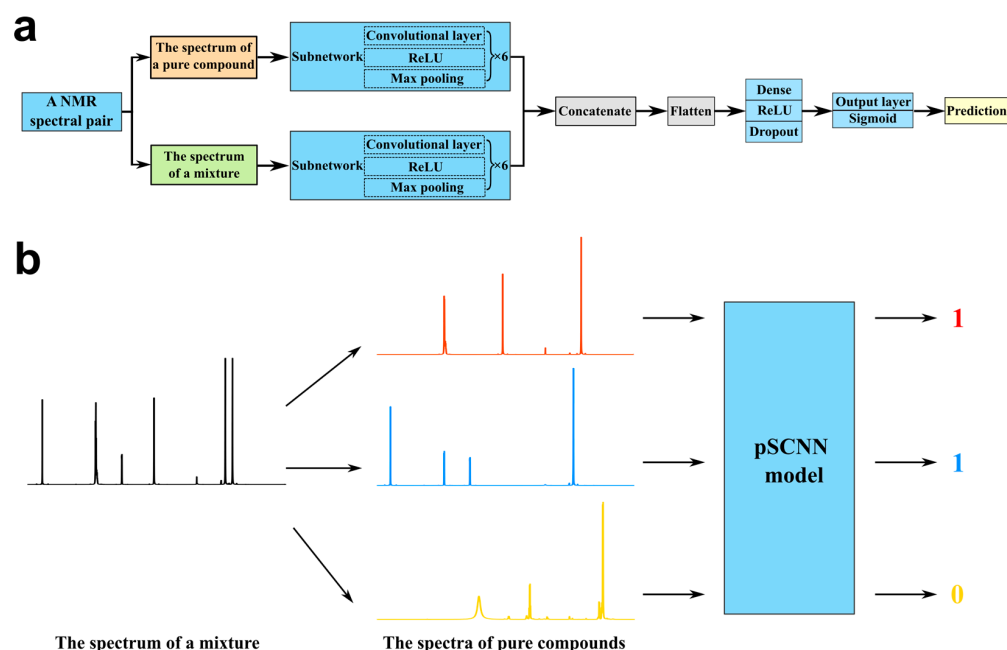
Deep learning is a category of flexible machine learning methods based on neural networks with multiple hidden layers to learn multilevel representation automatically for specific tasks [49]. It has three distinct advantages over the traditional learning methods. First, their network architectures are flexible enough to handle various kinds of raw inputs directly. For instance, there are convolutional neural networks (CNN) for computer vision [50], recurrent neural networks (RNN) [51], attention networks [52] for natural language processing and graph neural networks (GNN) for graph data structures [53]. Second, the multiple hidden layers can automatically transform the raw inputs into multilevel representations using a general purpose learning procedure [54]. Third, the deep learning methods have high expressive power and model capacity because of the depth efficiency, which can take full advantage of big data [55]. Due to these advantages, deep learning-based methods have achieved a state-of-the-art performance in numerous related fields of NMR spectroscopy [56,57], ranging from spectral reconstruction [58–60], denoising [61], peak picking [62,63], chemical shift prediction [64–68] and molecular recognition (the SMART method proposed by Zhang et al. in 2017) [69] to molecule identification [70–72]. It has shown unprecedented capabilities in solving difficult problems in NMR spectroscopy.

In this study, a pseudo-Siamese convolutional neural network (pSCNN) for NMR spectroscopy was developed to solve the chemical shift variation and the signal overlap problems in the component identification of mixtures inspired by Siamese neural networks [73,74] and the DeepCID method [75]. Each input of pSCNN is a spectral pair consisting of two full  $^1\text{H}$  NMR spectra: one is the spectrum of a pure compound from a spectral database, and the other is the spectrum of a mixture. The spectrum of a pure compound and the spectrum of a mixture are fed into two independent subnetworks consisting of convolutional layers, respectively. The translation-invariant representations can be learned for subsequent comparisons. The learned representations of the pure compound and the mixture are concatenated and fed into dense layers to predict their inclusion relationship (whether the mixture includes the pure compound or not). A data augmentation procedure was implemented to generate both positive and negative inputs for training, validating and testing the pSCNN model. The hyperparameters were optimized to obtain reasonable architecture and better performances. The compound identification procedure

was developed by predicting the inputs of each pure compound in the spectral database and the mixture with the model. The NMR spectra of known flavor mixtures and additional flavor mixture were acquired by an NMR spectrometer to benchmark the performance of the proposed method. Finally, the translational invariance of CNN was demonstrated to be suitable for solving the chemical shift variation problem when identifying compounds in a mixture with NMR spectroscopy. The main novelties of pSCNN compared to the SMART method are: (1) the two subnetworks of pSCNN have the same architecture, but their weights are trained separately, whereas the subnetworks of SMART have the same weights, and (2) the pSCNN method acquires 1D NMR spectra, while SMART needs 2D NMR spectra. The acquisition of 2D NMR spectra is time-consuming and highly expensive. To the best of our knowledge, this is the first work on the comparison of 1D NMR by the pseudo-Siamese convolutional neural network.

## 2. Method

The schematic diagram of the proposed pSCNN method is shown in Figure 1. Its source code is available at <https://www.github.com/yuxuanliao/pSCNN> (accessed on 4 May 2022). It mainly consists of two parts: a pseudo-Siamese convolutional neural network and model-based component identification. The data augmentation procedure is essentially the superposition several NMR spectra sampled from a spectral database at random ratios with random noises. Both positive and negative spectral pairs are generated and partitioned into training, validation and test sets. The pSCNN takes an NMR spectral pair as its input, extracts high-level representations using CNN layers and predicts the probability of a pure compound in a mixture using dense layers. For each compound in the database, the model-based component identification predicts the probability of it in a mixture. The possible components of the mixture can be obtained by filtering the predicted probabilities using a threshold value. In the subsequent subsections, we will elucidate the principle of each part of this method as clearly as possible.



**Figure 1.** Schematic diagram of the proposed pSCNN method. (a) The network architecture of the pSCNN model. pSCNN consists of two subnetworks. Each subnetwork consists of six convolutional layers. The extracted features are concatenated and fed into two dense layers for prediction. (b) pSCNN model-based component identification. The inclusion relationship between each compound in the database and a mixture is predicted by the pSCNN model.

### 2.1. Data Augmentation

The data augmentation and samples split are shown in Figure S1. Due to the good additivity in the NMR spectra between the mixture and its components, a data augmentation method has been developed, which can be used to generate enough spectra for training, validating and testing the pSCNN model. Essentially, data augmentation is the superposition of several flavor standards randomly sampled from the database at random ratios. Since spectra acquired by an NMR spectrometer contain noises, some random noises should be generated and added into the augmented spectra. Here, several NMR spectra were randomly selected from the spectral database, and the spectrum of a specific component was chosen as the pure spectrum in the spectral pair. The spectrum of this specific component was superposed with the other sampled ones at random ratios in a given range (e.g., 0.2–1.0) to generate positive ones. Negative ones were generated in the same way without the spectrum of this specific compound. Noises were added into the spectra of both positive and negative ones. The NMR spectral pair of each augmented sample was obtained by combining the augmented spectrum with the spectrum of the specific compound. A total of 22,000 augmented NMR spectral pairs were generated. The augmented dataset was split into the training set (18,000 spectral pairs), validation set (2000 spectral pairs) and test set (2000 spectral pairs) randomly. In summary, the augmented dataset was generated for training the pSCNN model, optimizing the hyperparameters of pSCNN and evaluating its accuracy, sensitivity and specificity.

### 2.2. Convolutional Neural Network

CNN is a popular category of deep learning methods in computer vision. Recently, it has also been widely used in chemistry [76], especially analytical chemistry [77]. It basically consists of an input layer, convolutional layers, pooling layers, dense layers and an output layer. Convolutional layers directly learn the multiple-level and translation-invariant representations from the raw inputs. Pooling layers reduce the dimensionalities of the data and improve the computational efficiency by combining several adjacent features. Dense layers perform nonlinear combinations of higher-level representations to achieve specific classification or regression tasks. The output layer is often a special dense layer, with its output size equaling the number of labels that each input has.

The core of the convolutional layer is convolution kernels, which are filters of a set of trainable parameters. Each convolution kernel scans the input with a given stride to extract features as its output. A feature map is generated by detecting a similar feature at different locations with the same convolution kernel. The most significant advantages of the convolutional layer are sparse connectivity and parameter sharing. Sparse connectivity is a learning process from local to global, which gradually enhances the understanding of global information with less parameters. Parameter sharing means that the same convolution kernel scans the feature map with a given stride, which can reduce the number of parameters significantly. The one-dimensional convolution layer is defined as follows:

$$out_{N_i, C_{out_j}} = f(bias_{C_{out_j}} + weight_{C_{out_j}} \otimes input_{N_i}) \quad (1)$$

where  $\otimes$  is the cross-correlation operation,  $C_{out_j}$  is the number of channels,  $N_i$  is the batch size,  $f()$  is the activation function and the rectified linear unit (ReLU,  $f(x) = \max(0, x)$ ) is used as the activation function in this study.

Pooling layers are often used to reduce the dimensions of feature maps by computing the maximal value or mean value of a small cluster. As a result, the number of parameters to be learned and the amount of computations are reduced significantly. It can also improve the translation invariance. Here, the max pooling layer was used in the pooling layers of the pSCNN model.

Each dense layer is connected to all the outputs of its previous layer. It mainly performs a nonlinear combination of higher-level features from convolutional layers to achieve a specific prediction task. Meanwhile, there is a dropout layer following each dense layer to

avoid overfitting and improve the generalization ability. The output layer is also a dense layer, with the output size equaling one. Since the output is the probability of a component in the mixture, Sigmoid is used as the activation function in the output layer.

### 2.3. Pseudo-Siamese Convolutional Neural Network

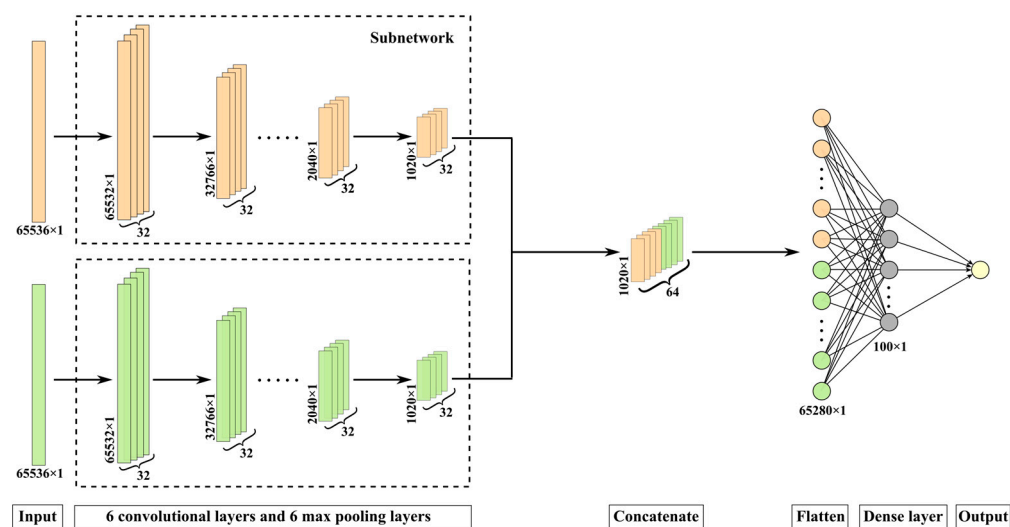
Compared with a traditional neural network with one input, a Siamese neural network uses two subnetworks with the same architecture and weights to extract two comparable feature vectors from two different inputs. For identifying compounds in mixture-based NMR spectroscopy, two feature vectors are extracted from the NMR spectrum of a mixture and the NMR spectrum of a given compound, respectively. Then, these feature vectors are compared by dense layers to determine their inclusion relationship. Since the NMR spectra are from the mixture and the pure compound, it is not appropriate to use subnetworks with the same weights for extracting their feature vectors. In this study, two subnetworks have the same architecture, but their weights are trained separately. This type of neural networks is often called pseudo-Siamese neural networks, because the subnetworks do not share weights. Furthermore, the feature extraction subnetworks employ convolutional layers, so this neural network is called a pseudo-Siamese convolutional neural network. The learned feature vectors are concatenated and fed into the dense layers to identify whether the two inputs have an inclusion relationship, thus achieving compound identification in the NMR spectra of mixtures.

### 2.4. Architecture of pSCNN for NMR

The detailed neural network architecture of pSCNN is shown in Figure 2. The input of pSCNN is an NMR spectral pair, which consists of one NMR spectrum of the mixture and one of the pure compound. The corresponding label of an NMR spectral pair is 0 or 1:1 if the mixture contains the pure compound and 0 otherwise. After feeding a spectral pair into pSCNN, their features are extracted by two subnetworks with the same architecture, respectively. Each subnetwork consists of 6 convolutional layers, each followed by a max pooling layer. The number of kernels for the convolutional layers is 32, and the kernel size is  $5 \times 1$ . The activation function of the convolutional layers is ReLU. They can learn the translation-invariant features from NMR spectra effectively, and the extracted features of two subnetworks are concatenated, flattened and fed into dense layers for comparison. The number of hidden units for the first dense layer is 100, and its activation function is also a ReLU. A dropout layer with the dropout rate equaling 0.2 is introduced to the dense layer for circumventing the overfitting problem. The output layer, the last dense layer, contains one hidden unit and uses the Sigmoid function as the activation function to form the final output. Binary cross entropy is chosen as the loss function, since it is suitable for binary classification problems. The Adam [78] optimizer is chosen as the optimizer because of its computational efficiency and little memory requirement.

### 2.5. Compound Identification with pSCNN

Given an NMR spectral database ( $\mathbf{D}$ ) and the NMR spectrum of a mixture ( $\mathbf{x}$ ), components in the mixture can be identified by the pSCNN model. The details of this identification method are described in the following procedure. Assuming that there are  $N$  NMR spectra of the standards in the NMR spectral database, the mixture spectrum is combined with these  $N$  spectra to form  $N$  spectral pairs  $(\mathbf{D}_1, \mathbf{x}), \dots, (\mathbf{D}_N, \mathbf{x})$ . For each spectral pair, its probability is predicted by the pSCNN model. After predicting all  $N$  spectral pairs, the probabilities of all compounds in the database are obtained. The probabilities of these compounds are filtered by setting a threshold value (e.g., 0.5). The components with a predicted probability greater than this threshold are regarded as candidates in the mixture.



**Figure 2.** The detailed neural network architecture of pSCNN.

### 2.6. Evaluation Metrics

To evaluate the performance of pSCNN on the mixture analysis, the metrics used in this study are accuracy (*ACC*), true positive rate (*TPR*, sensitivity) and false positive rate (*FPR*). The mathematical formulas for *ACC*, *TPR* and *FPR* are as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$TPR = \frac{TP}{TP + FN} \quad (3)$$

$$FPR = \frac{FP}{TN + FP} \quad (4)$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives, respectively. Samples were labeled as positive or negative in binary classification. If both the prediction value and actual value are positive, the sample is *TP*. If both the prediction value and actual value are negative, the sample is *TN*. The sample is *FP* if the prediction value is positive and the actual value is negative or *FN* if the prediction value is negative and the actual value is positive.

## 3. Experiments

### 3.1. Flavor Standards

Deuterated dimethyl sulfoxide (DMSO- $d_6$ , >99.8 atom% D, contains 0.03% (*v/v*) TMS) was purchased from Ningbo Cuiying Chemical Technology Co., Ltd., Ningbo, China. A total of 24 flavor standards were purchased from Guangzhou Huafang tobacco flavor Co., Ltd., Guangzhou, China. The information of each flavor standard is listed in Table S1. The sample solution was produced by dissolving 250  $\mu$ L of each flavor standard in 500  $\mu$ L of DMSO- $d_6$ . Then, 600  $\mu$ L of each sample solution were taken for NMR measurement. All  $^1\text{H}$  NMR spectra were acquired at 298 K on a Bruker AVANCE III 400 MHz NMR spectrometer (Bruker BioSpin, Rheinstetten, Germany). DMSO- $d_6$  was used for the NMR field lock. TMS was used as the internal standard. The pulse program was chosen as zg30, and the number of scans was 16.

### 3.2. Known Flavor Mixtures

Two, three, four or five flavor standards were mixed randomly to form 15 mixtures with known components. The information of each flavor mixture is listed in Table S2. Each flavor mixture was prepared by taking 100  $\mu\text{L}$  of each flavor standard and mixing them. The above prepared flavor mixture was added to 500  $\mu\text{L}$  of DMSO- $d_6$ , then vortex-mixed for 1 min at room temperature and, finally, transferred 600  $\mu\text{L}$  to an NMR tube for NMR measurements. The experimental conditions of the known flavor mixtures were set as those of the flavor standards.

### 3.3. Additional Flavor Mixture

The additional flavor mixture was provided by third-party personnel in the Technology Center of China Tobacco Hunan Industrial Co., Ltd., Changsha, China. The components of the additional flavor mixture were unknown when analyzing it with the pSCNN model. After submitting the predicted result to the Technology Center of China Tobacco Hunan Industrial Co., Ltd., we were informed of the corresponding components in this mixture. The information of the additional flavor mixture is listed in Table S3. The sample solution was prepared by dissolving 250  $\mu\text{L}$  of the additional flavor mixture in 500  $\mu\text{L}$  of DMSO- $d_6$ . Then, 600  $\mu\text{L}$  of the sample solution was transferred into an NMR tube. As previously described, the same experimental conditions were used for NMR measurement.

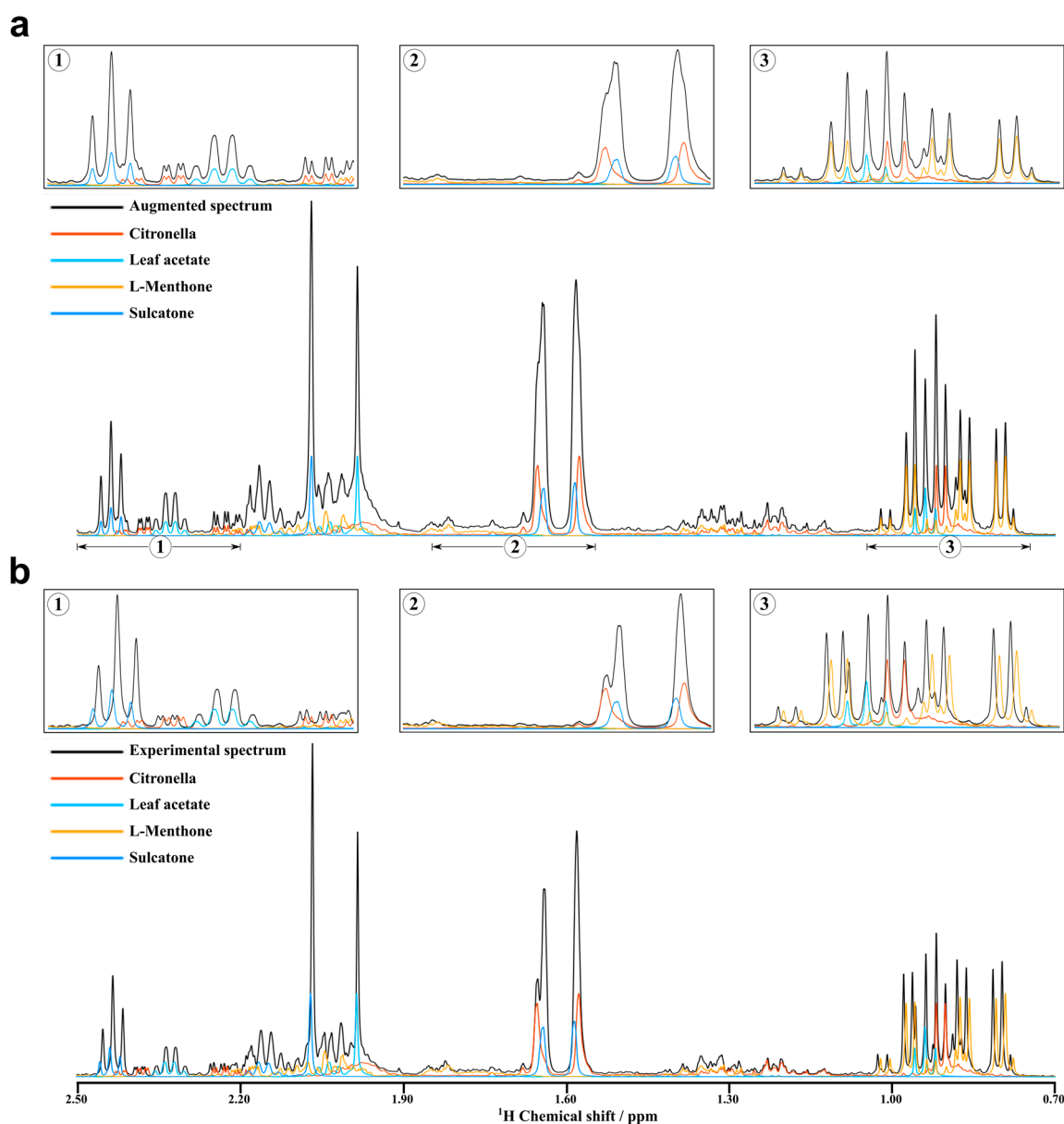
## 4. Results and Discussion

### 4.1. Implementation and Computing Resources

In this study, the neural network and related modules were implemented in Python (version 3.8.13), Tensorflow GPU package (version 2.5.0) and scikit-learn (version 1.0.0). The NMR spectra were read into Python using the nmrglue package (version 0.8.dev0). The computing tasks were submitted to the Inspur TS10000 high-performance computing (HPC) cluster of Central South University using the Slurm workload manager (version 20.02.3). This HPC cluster has 1022 central processing unit (CPU) nodes, 10 fat nodes and 26 graphics processing unit (GPU) nodes. For the training of pSCNN models, it was a GPU node with  $2 \times$  Intel(R) Xeon(R) Gold 6248R processors,  $2 \times$  Nvidia Tesla V100s, 384G DDR4 memory and a CentOS 7.5 operating system.

### 4.2. Validation of Data Augmentation

The augmented and experimental NMR spectra were compared to validate the rationality of the data augmentation. Here, the F10 mixture consisting of Sulcatone, L-Menthone, Citronellal and Leaf acetate was used as an example. An augmented spectrum was generated by the data augmentation method in Section 2.1 with the components of the F10 mixture and random ratios. Figure 3a,b show the augmented spectrum and the experimental spectrum of F10, respectively. The augmented spectrum is basically consistent with the experimental spectra from the local zoomed-in views, except for the chemical shift variations in the experimental spectra. Thus, it has shown that the data augmentation method can generate reasonable NMR spectra of mixtures from the NMR spectra of components.



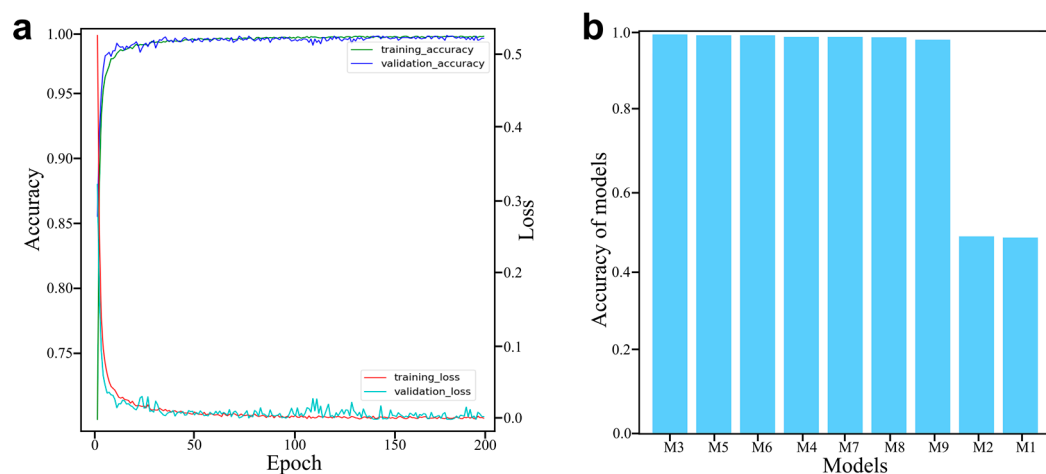
**Figure 3.** Augmented and experimental NMR spectra. (a) The spectrum of mixture obtained by data augmentation and its components spectra. (b) The experimental NMR spectra of the mixture and its components. (1–3) are the local zoomed-in views.

#### 4.3. Hyperparameters Optimization and Training

The optimization of the hyperparameters is crucial for establishing a high-performance model. For the pSCNN model used in this study, the key hyperparameters are the epoch, the learning rate and the number of convolutional layers, which should be optimized. Firstly, the epoch was set to 200, and the model was trained. The loss–epoch and accuracy–epoch curves are shown in Figure 4a. It can be seen that the model is basically stable after 100 epochs. Therefore, the epoch was set to 100. For the learning rate, the training is slow when too small, and the model does not converge when too large. Here, we investigated the learning rates in the range of  $10^{-2}$  to  $10^{-5}$ . Combining the results in Table 1 and Figure 4b, it can be concluded that the model fails to converge with the learning rates in a range from  $10^{-2}$  to  $10^{-3}$ . If the learning rate is set between  $10^{-4}$  and  $10^{-5}$ , the model can be successfully trained. For the number of convolutional layers, we tested inside the range of 5–10. As can be seen in Table 1 and Figure 4b, the accuracy of the validation set increases and then decreases as the number of convolutional layers increases. The



best performance was achieved when the number of layers was equal to 6. Therefore, the optimized epoch, learning rate and number of convolutional layers were 100,  $10^{-4}$  and 6, respectively. The final model was chosen as M3, which could achieve an accuracy of 0.9990 on the validation set.



**Figure 4.** Optimization of the pSCNN model. (a) The accuracy curves and loss curves of training set and validation set. (b) The accuracy of different models on the validation set.

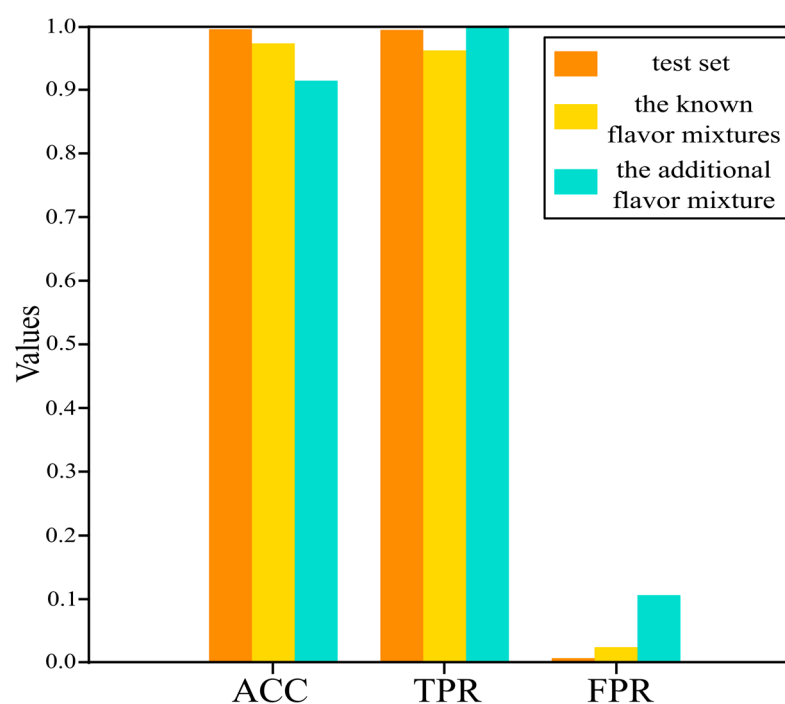
**Table 1.** The accuracy of different pSCNN models on the validation set.

Name of Models	Epoch	The Number of Convolutional Layers *	Learning Rate	ACC
M1	100	6	$10^{-2}$	0.4900
M2	100	6	$10^{-3}$	0.4935
M3	100	6	$10^{-4}$	0.9990
M4	100	6	$10^{-5}$	0.9935
M5	100	5	$10^{-4}$	0.9975
M6	100	7	$10^{-4}$	0.9975
M7	100	8	$10^{-4}$	0.9935
M8	100	9	$10^{-4}$	0.9925
M9	100	10	$10^{-4}$	0.9860

\* A max pooling layer whose stride is set to 2 follows a convolutional layer.

#### 4.4. Performance Evaluation

The training set and the validation set have already been used to update the parameters and adjust the hyperparameters, respectively. The performance evaluation metrics obtained on them are often overoptimistic. To test the true performance of the model on unknown samples, an independent test set is usually used for a performance evaluation. Here, the test set in the augmented dataset was used to evaluate the performance of the pSCNN model on unseen samples. Each spectral pair in the test set was fed into pSCNN, and their features were extracted and transformed into the learned representations to predict the possibility of the pure compound in the mixture. There were 2000 spectral pairs in the test set. As shown in Figure 5, its ACC, TPR and FPR are 99.80%, 99.70% and 0.10% respectively, which guarantee the performance of the pSCNN model on unseen samples.



**Figure 5.** Performance evaluation on the test set and application of pSCNN on the known flavor mixtures and the additional flavor mixture.

#### 4.5. Results of Mixture Analysis

Due to its excellent performance on the test set, the pSCNN model was used to identify the flavor standards in the known flavor mixtures and the additional flavor mixture. For the known flavor mixtures, the components in each mixture are known. Therefore, they were used to verify the identification performance of pSCNN on real NMR spectra. The NMR spectrum of each mixture in the known flavor mixtures dataset was combined with the NMR spectra of flavor standards to form its spectral pairs. These spectral pairs were fed into the pSCNN model to predict the probabilities of the flavor standards in this flavor mixture. The results of the known flavor mixtures are shown in Figure 5 and Table 2. The detailed results of all mixtures in the flavor mixtures dataset are listed in Table S4. It can be seen that the ACC, TPR and FPR are 97.62%, 96.44% and 2.29%, respectively. Therefore, the performance of pSCNN for component identification in the mixtures was verified by analyzing the known flavor mixtures dataset.

**Table 2.** The results of the pSCNN model on the experimental NMR datasets.

Datasets	ACC	TPR	FPR
flavor mixtures dataset	97.62%	96.44%	2.29%
additional flavor mixture dataset	91.67%	100.00%	10.53%

It was further applied to analyze the unknown flavor mixtures in the additional flavor mixture dataset. Since the components of U1 were unknown when analyzing it with the pSCNN model, it can test the accuracy, sensitivity and specificity of pSCNN for real-world applications. It was predicted in the same way as for the mixtures in the known flavor mixtures. For the additional flavor mixture, the model-based prediction probabilities for the U1 mixture were ranked from high to low as  $\beta$ -Ionone,  $\gamma$ -Decalactone,  $\gamma$ -Nonanoic lactone, Citral, Leaf alcohol, Isovaleric acid and 2-Methylbutyric acid. After submitting the predicted candidates to the Technology Center of China Tobacco Hunan Industrial Co., Ltd., they sent us the real formulation of U1. The formulation of U1 is listed in Table S5. It can be seen that the predicted results of pSCNN match well with the formulation provided by the Technology Centre of China Tobacco Hunan Industrial Co., Ltd. The

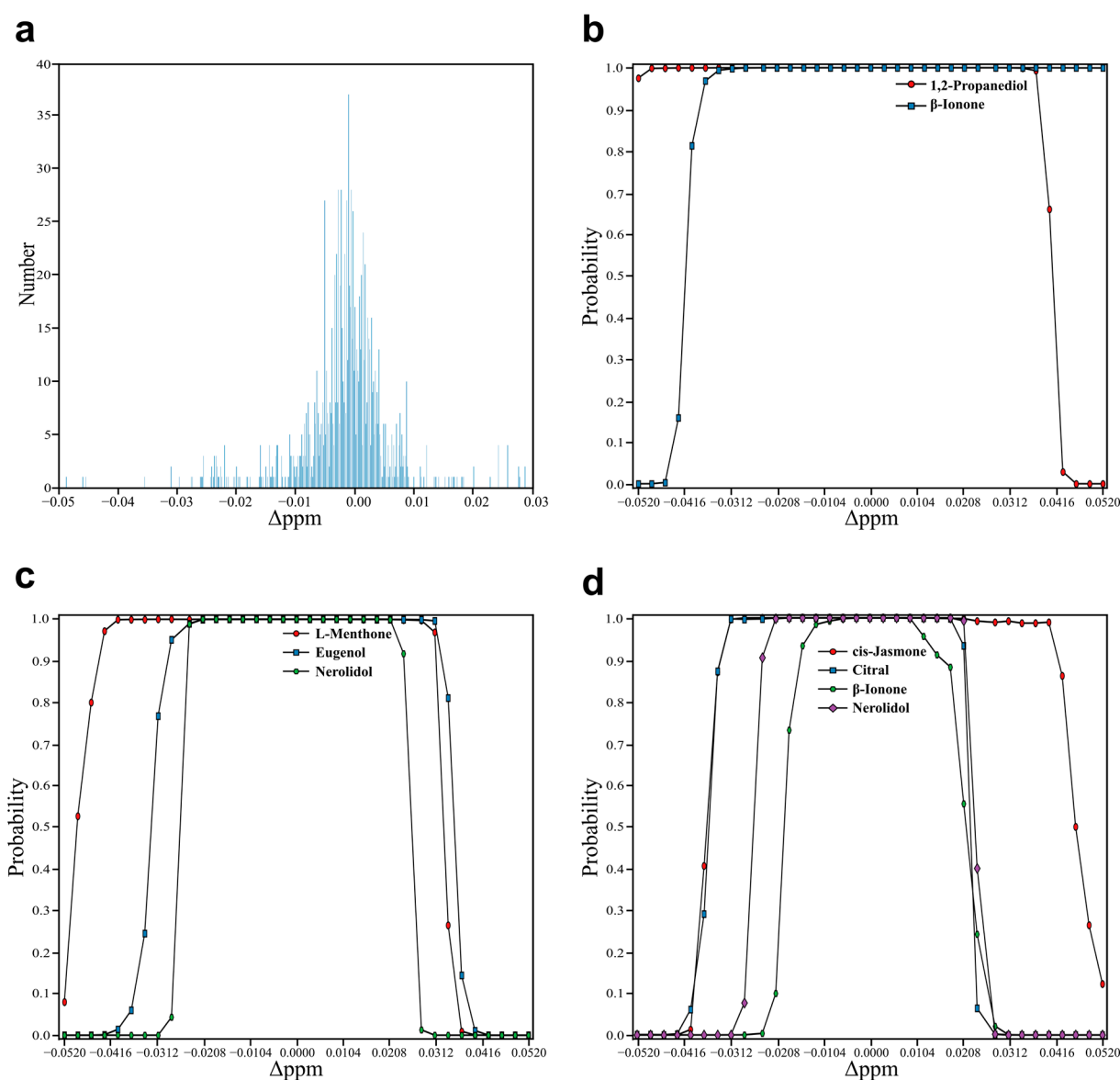
statistical results are also shown in Figure 5 and Table 2. The detailed results of the mixture in the additional flavor mixture dataset are listed in Table S5. The ACC, TPR and FPR on the additional flavor mixture dataset are 91.67%, 100.00% and 10.53%, respectively. The FPR in the additional flavor mixture (10.53%) is high, and the reasons are that Isovaleric acid and 2-Methylbutyric acid are structural isomers, and  $\gamma$ -Decalactone and  $\gamma$ -Nonanoic lactone are homologs, because their molecular structures differ by only one  $\text{CH}_2$  group. Actually, FPR is not an issue; this is because the false positives in the candidate components can be filtered out by further analysis. This shows that pSCNN is an efficient method for identifying compounds in real unknown mixtures.

#### 4.6. Translation Invariance for NMR Peaks

In NMR spectra, the chemical shifts of the same compound may vary in different samples because of influences from interactions between components, instruments or the environment. Convolutional neural networks have a translational invariance advantage due to the learned high-level representations from raw signals by their convolutional layers and pooling layers. Therefore, it would be interesting to investigate the relationship between chemical shift variations in NMR spectra and translation invariance of the convolutional neural networks.

First, the experimental NMR spectra are used to determine the interval of the chemical shift variations. The chemical shift of active hydrogen is highly correlated with the concentration of the compound because of the hydrogen-bonding interaction, so the chemical shift variations of active hydrogen in different NMR spectra are not taken into account. The chemical shift variations of the deuterated solvent (DMSO- $d_6$  2.50 ppm and HDO 3.33 ppm [79]) are also not considered, as the solvent signal is not the signal of interest. By observing the NMR spectra of all mixtures and their corresponding components, the chemical shift variation of each spectral pair was calculated by the deviation between the chemical shift of the peaks of the mixture and the chemical shift of the peaks for its component. As shown in Figure 6a, the mean value and the standard deviation of the chemical shift variation are  $-0.0016$  and  $0.0077$ , respectively. The obtained interval was  $(-0.013, 0.010)$  according to the mean value  $\pm 1.5 \times$  standard deviation. This interval contains the chemical shift variations of 90.52% of the characteristic peaks in the spectral pairs.

To generate NMR spectra with different chemical shift variations, the NMR spectra of the mixtures were obtained by the data augmentation method, and their corresponding components were retrieved from the spectral database. The number of components in these mixtures are 2, 3, 4 and 5. Two spectral pairs were generated under each number of component; thus, eight spectral pairs were generated. The information of the augmented spectral pairs is listed in Table S6. A total of 28 spectral pairs were obtained by combining each mixture with its components. For each spectral pair, chemical shifts of the NMR spectrum of its mixture were varied gradually within  $\pm 0.052$  ppm ranges, and the chemical shifts of the NMR spectra of its components were unchanged. A varied spectral pair was obtained after each chemical shift variation, and 35 varied spectral pairs were obtained for each spectral pair. Thus, 980 spectral pairs with different chemical shift variations were obtained to verify the translation invariance of the pSCNN model. The overall scheme for verifying the translation invariance is shown in Figure S1. For the spectral pair of a specific mixture and a specific component, its varied spectral pairs were fed into the pSCNN model to predict the probabilities of the component in the mixture under different chemical shift variations. The probabilities of the spectral pairs with different chemical shift variations are listed in Table S7. The results of representative spectral pairs are shown in Figure 6b–d, respectively. It can be seen that the minimum interval with probabilities higher than 0.5  $(-0.015, 0.015)$  exceeded the obtained interval  $(-0.013, 0.010)$ . The results showed that pSCNN can be directly used to compare the two NMR spectra without chemical shift alignment.



**Figure 6.** Demonstration of the translation invariance of pSCNN. (a) The number of chemical shift variations between all mixtures and their corresponding components. (b–d) The probabilities of the corresponding components in mixtures predicted by pSCNN for spectral pairs with different chemical shift variations. The results in (b–d) are the spectral pairs with two, three and four components, respectively.

## 5. Conclusions

In this study, an end-to-end method for compound identification in mixtures was developed based on a pseudo-Siamese convolutional neural network and  $^1\text{H}$  NMR spectroscopy. Two subnetworks consisting of convolutional layers were chosen to learn the representations from the spectra of pure compounds and the representations from the spectra of mixtures, respectively. The pure compound representation and the mixture representation were concatenated and fed into the dense layers to predict the probability of the compound in the mixture. The data augmentation method was used to generate 22,000 dual inputs from the NMR spectral database of flavor standards, which was randomly divided into the training set, validation set and test set. The performance of pSCNN was evaluated on the test set in terms of the ACC (99.80%), TPR (99.70%) and FPR (0.10%). Furthermore, the proposed method was applied in the flavor mixtures dataset and the additional flavor mixture dataset to benchmark its performance in real mixtures. The

performance metrics were ACC = 97.62%, TPR = 96.44% and FPR = 2.29% for the flavor mixtures dataset and ACC = 91.67%, TPR = 100.00% and FPR = 10.53% for the additional flavor mixture dataset. The results show that this method is able to identify components in mixtures accurately. Even in the chemical shift variations up to 0.015 ppm, the trained model can still identify the components in a mixture, which should be attributed to the translation invariance introduced by the convolutional layers and pooling layers in pSCNN. In conclusion, deep learning methods, especially pSCNN, are highly promising approaches to identify compounds in the mixture based on NMR spectroscopy. Due to the flexibility of CNN, the concept of pSCNN can be easily extended to NMR spectral library searching, the verification of complex samples and compound identification with multidimensional NMR spectroscopy. Further, high-field NMR spectrometers and low temperature probes can increase the sensitivity, which may help to analyze more challenging samples and improve the accuracy of the prediction results.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/molecules27123653/s1>: Figure S1: Figure of the data augmentation and samples split. Figure S2: Flowchart of the overall scheme for verifying the translation invariance. Table S1: Table of information of 24 flavor standards. Table S2: Table of information of the flavor mixtures. Table S3: Table of information of the additional flavor mixture. Table S4: Table of the detailed results of all mixtures in the flavor mixtures dataset. Table S5: Table of detailed results of the mixture in the additional flavor mixture dataset. Table S6: Table of information of the augmented mixtures. Table S7: Table of probabilities of spectral pairs with different chemical shift variations.

**Author Contributions:** Conceptualization, Z.Z. and H.Y.; methodology, W.W., Y.L., B.K., W.D. and H.L.; software, Y.L. and W.W.; validation, W.W., Y.L., B.K. and H.L.; formal analysis, Y.L.; investigation, Y.L., Y.W. and S.W.; resources, Y.L. and W.W.; data curation, Y.L. and Z.Z.; writing—original draft preparation, Y.L. and W.W.; writing—review and editing, Z.Z. and H.Y.; visualization, Y.L. and W.W.; supervision, Z.Z. and H.Y. and project administration, Y.L. and Z.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The source code, model, spectra, manual and tutorial are available at <https://www.github.com/yuxuanliao/pSCNN> (accessed on 4 May 2022).

**Acknowledgments:** We are grateful for the resources from the High-Performance Computing Center of Central South University. We are grateful to all employees of this institute for their encouragement and support of this research.

**Conflicts of Interest:** The authors declare no conflict of interest.

**Sample Availability:** Samples of the compounds are available from the authors.

## References

1. Akash, M.S.H.; Rehman, K. *Essentials of Pharmaceutical Analysis*; Springer: Singapore, 2020. [CrossRef]
2. Tsedilin, A.; Fakhrutdinov, A.N.; Eremin, D.; Zalesskiy, S.S.; Chizhov, A.O.; Kolotyrkina, N.G.; Ananikov, V. How sensitive and accurate are routine NMR and MS measurements? *Mendeleev Commun.* **2015**, *25*, 454–456. [CrossRef]
3. Kovacs, H.; Moskau, D.; Spraul, M. Cryogenically cooled probes—A leap in NMR technology. *Prog. Nucl. Magn. Reson. Spectrosc.* **2005**, *46*, 131–155. [CrossRef]
4. Elyashberg, M. Identification and structure elucidation by NMR spectroscopy. *TrAC Trends Anal. Chem.* **2015**, *69*, 88–97. [CrossRef]
5. Lodewyk, M.W.; Siebert, M.R.; Tantillo, D.J. Computational Prediction of <sup>1</sup>H and <sup>13</sup>C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.* **2012**, *112*, 1839–1862. [CrossRef]
6. Claridge, T.D.W. Chapter 2—Introducing High-Resolution NMR. In *High-Resolution NMR Techniques in Organic Chemistry*, 3rd ed.; Claridge, T.D.W., Ed.; Elsevier: Boston, MA, USA, 2016; pp. 11–59.
7. Edison, A.S.; Colonna, M.; Gouveia, G.J.; Holderman, N.R.; Judge, M.T.; Shen, X.; Zhang, S. NMR: Unique Strengths That Enhance Modern Metabolomics Research. *Anal. Chem.* **2021**, *93*, 478–499. [CrossRef]

8. Emwas, A.-H.; Roy, R.; McKay, R.T.; Tenori, L.; Saccenti, E.; Gowda, G.A.N.; Raftery, D.; Alahmari, F.; Jaremko, L.; Jaremko, M.; et al. NMR Spectroscopy for Metabolomics Research. *Metabolites* **2019**, *9*, 123. [[CrossRef](#)]
9. Wishart, D.S. Quantitative metabolomics using NMR. *TrAC Trends Anal. Chem.* **2008**, *27*, 228–237. [[CrossRef](#)]
10. Shi, L.; Zhang, N. Applications of Solution NMR in Drug Discovery. *Molecules* **2021**, *26*, 576. [[CrossRef](#)]
11. Soffley, C.A.; Bostock, M.J.; Popowicz, G.M.; Sattler, M. Paramagnetic NMR in drug discovery. *J. Biomol. NMR* **2020**, *74*, 287–309. [[CrossRef](#)] [[PubMed](#)]
12. Pellecchia, M.; Bertini, I.; Cowburn, D.; Dalvit, C.; Giralt, E.; Jahnke, W.; James, T.L.; Homans, S.W.; Kessler, H.; Luchinat, C.; et al. Perspectives on NMR in drug discovery: A technique comes of age. *Nat. Rev. Drug Discov.* **2008**, *7*, 738–745. [[CrossRef](#)] [[PubMed](#)]
13. Cao, R.; Liu, X.; Liu, Y.; Zhai, X.; Cao, T.; Wang, A.; Qiu, J. Applications of nuclear magnetic resonance spectroscopy to the evaluation of complex food constituents. *Food Chem.* **2021**, *342*, 128258. [[CrossRef](#)] [[PubMed](#)]
14. Santos, A.; Fonseca, F.; Lião, L.; Alcantara, G.; Barison, A. High-resolution magic angle spinning nuclear magnetic resonance in foodstuff analysis. *TrAC Trends Anal. Chem.* **2015**, *73*, 10–18. [[CrossRef](#)]
15. Wang, Z.-F.; You, Y.-L.; Li, F.-F.; Kong, W.-R.; Wang, S.-Q. Research Progress of NMR in Natural Product Quantification. *Molecules* **2021**, *26*, 6308. [[CrossRef](#)] [[PubMed](#)]
16. Robinette, S.L.; Brüscheweiler, R.; Schroeder, F.C.; Edison, A.S. NMR in Metabolomics and Natural Products Research: Two Sides of the Same Coin. *Acc. Chem. Res.* **2012**, *45*, 288–297. [[CrossRef](#)]
17. Martin, G.J.; Martin, M.L. Thirty Years of Flavor NMR. In *Flavor Chemistry: Thirty Years of Progress*; Teranishi, R., Wick, E.L., Hornstein, I., Eds.; Springer: Boston, MA, USA, 1999; pp. 19–30.
18. Singh, P.; Singh, M.K.; Beg, Y.R.; Nishad, G.R. A review on spectroscopic methods for determination of nitrite and nitrate in environmental samples. *Talanta* **2019**, *191*, 364–381. [[CrossRef](#)]
19. Santos, A.; Dutra, L.; Menezes, L.; Santos, M.; Barison, A. Forensic NMR spectroscopy: Just a beginning of a promising partnership. *TrAC Trends Anal. Chem.* **2018**, *107*, 31–42. [[CrossRef](#)]
20. Proietti, N.; Capitani, D.; Di Tullio, V. Nuclear Magnetic Resonance, a Powerful Tool in Cultural Heritage. *Magnetochemistry* **2018**, *4*, 11. [[CrossRef](#)]
21. Ebrahimi, P.; Viereck, N.; Bro, R.; Engelsen, S.B. Chemometric Analysis of NMR Spectra. In *Modern Magnetic Resonance*; Webb, G.A., Ed.; Springer International Publishing: Cham, Switzerland, 2017; pp. 1–20.
22. Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, S. Molecular search by NMR spectrum based on evaluation of matching between spectrum and molecule. *Sci. Rep.* **2021**, *11*, 20998. [[CrossRef](#)] [[PubMed](#)]
23. Steinbeck, C.; Krause, S.; Kuhn, S. NMRShiftDB Constructing a Free Chemical Information System with Open-Source Components. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1733–1739. [[CrossRef](#)]
24. Cui, Q.; Lewis, I.A.; Hegeman, A.D.; Anderson, M.E.; Li, J.; Schulte, C.F.; Westler, W.M.; Eghbalian, H.R.; Sussman, M.R.; Markley, J.L. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat. Biotechnol.* **2008**, *26*, 162–164. [[CrossRef](#)]
25. Wishart, D.S.; Feunang, Y.D.; Marcu, A.; Guo, A.C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617. [[CrossRef](#)] [[PubMed](#)]
26. Todeschini, R.; Ballabio, D.; Consonni, V. Distances and Similarity Measures in Chemometrics and Chemoinformatics. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: Hoboken, NJ, USA, 2020; pp. 1–40. [[CrossRef](#)]
27. Schaller, R.B.; Pretsch, E. A computer program for the automatic estimation of <sup>1</sup>H NMR chemical shifts. *Anal. Chim. Acta* **1994**, *290*, 295–302. [[CrossRef](#)]
28. De Meyer, T.; Sinnaeve, D.; Van Gasse, B.; Tsiportkova, E.; Rietzschel, E.R.; De Buyzere, M.L.; Gillebert, T.C.; Bekaert, S.; Martins, J.C.; Van Crieckinge, W. NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive, Intelligent Binning Algorithm. *Anal. Chem.* **2008**, *80*, 3783–3790. [[CrossRef](#)]
29. Åberg, K.M.; Alm, E.; Torgrip, R.J.O. The correspondence problem for metabolomics datasets. *Anal. Bioanal. Chem.* **2009**, *394*, 151–162. [[CrossRef](#)]
30. Worley, B.; Powers, R. Generalized adaptive intelligent binning of multiway data. *Chemom. Intell. Lab. Syst.* **2015**, *146*, 42–46. [[CrossRef](#)] [[PubMed](#)]
31. Vu, T.N.; Laukens, K. Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. *Metabolites* **2013**, *3*, 259–276. [[CrossRef](#)]
32. Savorani, F.; Tomasi, G.; Engelsen, S.B. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *J. Magn. Reson.* **2010**, *202*, 190–202. [[CrossRef](#)]
33. Veselkov, K.A.; Lindon, J.C.; Ebbels, T.M.D.; Crockford, D.; Volynkin, V.V.; Holmes, E.; Davies, D.B.; Nicholson, J.K. Recursive Segment-Wise Peak Alignment of Biological (1)H NMR Spectra for Improved Metabolic Biomarker Recovery. *Anal. Chem.* **2009**, *81*, 56–66. [[CrossRef](#)]
34. Castillo, A.M.; Uribe, L.; Patiny, L.; Wist, J. Fast and shift-insensitive similarity comparisons of NMR using a tree-representation of spectra. *Chemom. Intell. Lab. Syst.* **2013**, *127*, 1–6. [[CrossRef](#)]
35. Bodis, L.; Ross, A.; Pretsch, E. A novel spectra similarity measure. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 1–8. [[CrossRef](#)]
36. Mishra, R.; Marchand, A.; Jacquemmoz, C.; Dumez, J.-N. Ultrafast diffusion-based unmixing of <sup>1</sup>H NMR spectra. *Chem. Commun.* **2021**, *57*, 2384–2387. [[CrossRef](#)]
37. Lin, M.; Shapiro, M.J. Mixture Analysis by NMR Spectroscopy. *Anal. Chem.* **1997**, *69*, 4731–4733. [[CrossRef](#)]

38. Zhang, F.; Brüschweiler, R. Robust Deconvolution of Complex Mixtures by Covariance TOCSY Spectroscopy. *Angew. Chem. Int. Ed.* **2007**, *46*, 2639–2642. [[CrossRef](#)]
39. Castellanos, E.R.R.; Wist, J. Decomposition of mixtures' spectra by multivariate curve resolution of rapidly acquired TOCSY experiments. *Magn. Reson. Chem.* **2010**, *48*, 771–776. [[CrossRef](#)] [[PubMed](#)]
40. Bingol, K.; Brüschweiler, R. Deconvolution of Chemical Mixtures with High Complexity by NMR Consensus Trace Clustering. *Anal. Chem.* **2011**, *83*, 7412–7417. [[CrossRef](#)]
41. Toumi, I.; Caldarelli, S.; Torrèسانی, B. A review of blind source separation in NMR spectroscopy. *Prog. Nucl. Magn. Reson. Spectrosc.* **2014**, *81*, 37–64. [[CrossRef](#)] [[PubMed](#)]
42. Poggetto, G.D.; Castañar, L.; Adams, R.W.; Morris, G.A.; Nilsson, M. Dissect and Divide: Putting NMR Spectra of Mixtures under the Knife. *J. Am. Chem. Soc.* **2019**, *141*, 5766–5771. [[CrossRef](#)] [[PubMed](#)]
43. McKenzie, J.S.; Donarski, J.A.; Wilson, J.C.; Charlton, A.J. Analysis of complex mixtures using high-resolution nuclear magnetic resonance spectroscopy and chemometrics. *Prog. Nucl. Magn. Reson. Spectrosc.* **2011**, *59*, 336–359. [[CrossRef](#)]
44. Tulpan, D.; Léger, S.; Belliveau, L.; Culf, A.; Čuperlović-Culf, M. MetaboHunter: An automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinform.* **2011**, *12*, 400. [[CrossRef](#)]
45. Wei, S.; Zhang, J.; Liu, L.; Ye, T.; Gowda, G.A.N.; Tayyari, F.; Raftery, D. Ratio Analysis Nuclear Magnetic Resonance Spectroscopy for Selective Metabolite Identification in Complex Samples. *Anal. Chem.* **2011**, *83*, 7616–7623. [[CrossRef](#)]
46. Krishnamurthy, K. CRAFT (complete reduction to amplitude frequency table)—Robust and time-efficient Bayesian approach for quantitative mixture analysis by NMR. *Magn. Reson. Chem.* **2013**, *51*, 821–829. [[CrossRef](#)] [[PubMed](#)]
47. Hubert, J.; Nuzillard, J.-M.; Purson, S.; Hamzaoui, M.; Borie, N.; Reynaud, R.; Renault, J.-H. Identification of Natural Metabolites in Mixture: A Pattern Recognition Strategy Based on <sup>13</sup>C NMR. *Anal. Chem.* **2014**, *86*, 2955–2962. [[CrossRef](#)]
48. Kuhn, S.; Colreavy-Donnelly, S.; de Souza, J.S.; Borges, R.M. An integrated approach for mixture analysis using MS and NMR techniques. *Faraday Discuss.* **2019**, *218*, 339–353. [[CrossRef](#)]
49. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)] [[PubMed](#)]
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 2012, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
51. Graves, A.; Mohamed, A.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 6645–6649.
52. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–6 December 2017; pp. 5998–6008.
53. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [[CrossRef](#)]
54. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [[CrossRef](#)] [[PubMed](#)]
55. Lu, Z.; Pu, H.; Wang, F.; Hu, Z.; Wang, L. The expressive power of neural networks: A view from the width. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–6 December 2017; pp. 6232–6240.
56. Chen, D.; Wang, Z.; Guo, D.; Orekhov, V.; Qu, X. Review and prospect: Deep learning in nuclear magnetic resonance spectroscopy. *Chem. Eur. J.* **2020**, *26*, 10391–10401. [[CrossRef](#)]
57. Cobas, C. NMR signal processing, prediction, and structure verification with machine learning techniques. *Magn. Reson. Chem.* **2020**, *58*, 512–519. [[CrossRef](#)] [[PubMed](#)]
58. Qu, X.; Huang, Y.; Lu, H.; Qiu, T.; Guo, D.; Agback, T.; Orekhov, V.; Chen, Z. Accelerated Nuclear Magnetic Resonance Spectroscopy with Deep Learning. *Angew. Chem. Int. Ed.* **2020**, *59*, 10297–10300. [[CrossRef](#)] [[PubMed](#)]
59. Luo, J.; Zeng, Q.; Wu, K.; Lin, Y. Fast reconstruction of non-uniform sampling multidimensional NMR spectroscopy via a deep neural network. *J. Magn. Reson.* **2020**, *317*, 106772. [[CrossRef](#)] [[PubMed](#)]
60. Hansen, D.F. Using Deep Neural Networks to Reconstruct Non-uniformly Sampled NMR Spectra. *J. Biomol. NMR* **2019**, *73*, 577–585. [[CrossRef](#)] [[PubMed](#)]
61. Wu, K.; Luo, J.; Zeng, Q.; Dong, X.; Chen, J.; Zhan, C.; Chen, Z.; Lin, Y. Improvement in Signal-to-Noise Ratio of Liquid-State NMR Spectroscopy via a Deep Neural Network DN-Unet. *Anal. Chem.* **2021**, *93*, 1377–1382. [[CrossRef](#)]
62. Klukowski, P.; Augoff, M.; Zieba, M.; Drwal, M.; Gonczarek, A.; Walczak, M.J. NMRNet: A deep learning approach to automated peak picking of protein NMR spectra. *Bioinformatics* **2018**, *34*, 2590–2597. [[CrossRef](#)]
63. Li, D.-W.; Hansen, A.L.; Yuan, C.; Bruschiweiler-Li, L.; Brüschweiler, R. DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nat. Commun.* **2021**, *12*, 5229. [[CrossRef](#)]
64. Jonas, E.; Kuhn, S. Rapid prediction of NMR spectral properties with quantified uncertainty. *J. Cheminformatics* **2019**, *11*, 50. [[CrossRef](#)]
65. Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, M.; Kang, S. Neural Message Passing for NMR Chemical Shift Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2024–2030. [[CrossRef](#)]
66. Gerrard, W.; Bratholm, L.A.; Packer, M.J.; Mulholland, A.J.; Glowacki, D.R.; Butts, C.P. IMPRESSION—Prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **2020**, *11*, 508–515. [[CrossRef](#)] [[PubMed](#)]

67. Guan, Y.; Shree Sowndarya, S.V.; Gallegos, L.C.; John, P.C.S.; Paton, R.S. Real-time prediction of  $^1\text{H}$  and  $^{13}\text{C}$  chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.* **2021**, *12*, 12012–12026. [[CrossRef](#)]
68. Yang, Z.; Chakraborty, M.; White, A.D. Predicting chemical shifts with graph neural networks. *Chem. Sci.* **2021**, *12*, 10802–10809. [[CrossRef](#)] [[PubMed](#)]
69. Zhang, C.; Idelbayev, Y.; Roberts, N.; Tao, Y.; Nannapaneni, Y.; Duggan, B.M.; Min, J.; Lin, E.C.; Gerwick, E.C.; Cottrell, G.W.; et al. Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci. Rep.* **2017**, *7*, 14243. [[CrossRef](#)]
70. Zhang, J.; Terayama, K.; Sumita, M.; Yoshizoe, K.; Ito, K.; Kikuchi, J.; Tsuda, K. NMR-TS: De novo molecule identification from NMR spectra. *Sci. Technol. Adv. Mater.* **2020**, *21*, 552–561. [[CrossRef](#)] [[PubMed](#)]
71. Huang, Z.; Chen, M.S.; Woroch, C.P.P.; Markland, T.E.; Kanan, M.W. A framework for automated structure elucidation from routine NMR spectra. *Chem. Sci.* **2021**, *12*, 15329–15338. [[CrossRef](#)]
72. Kuhn, S.; Tumer, E.; Colreavy-Donnelly, S.; Borges, R.M. A Pilot Study for Fragment Identification Using 2D NMR and Deep Learning. *Magn. Reson. Chem.* **2021**. [[CrossRef](#)] [[PubMed](#)]
73. Chicco, D. Siamese Neural Networks: An Overview. In *Artificial Neural Networks*; Cartwright, H., Ed.; Springer: New York, NY, USA, 2021; pp. 73–94.
74. Huber, F.; van der Burg, S.; van der Hooft, J.J.J.; Ridder, L. MS2DeepScore: A novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminform.* **2021**, *13*, 84. [[CrossRef](#)]
75. Fan, X.; Ming, W.; Zeng, H.; Zhang, Z.; Lu, H. Deep learning-based component identification for the Raman spectra of mixtures. *Analyst* **2019**, *144*, 1789–1798. [[CrossRef](#)]
76. Mater, A.C.; Coote, M.L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559. [[CrossRef](#)]
77. Debus, B.; Parastar, H.; Harrington, P.; Kirsanov, D. Deep learning in analytical chemistry. *TrAC Trends Anal. Chem.* **2021**, *145*, 116459. [[CrossRef](#)]
78. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
79. Fulmer, G.R.; Miller, A.J.M.; Sherden, N.H.; Gottlieb, H.E.; Nudelman, A.; Stoltz, B.M.; Bercaw, J.E.; Goldberg, K.I. NMR Chemical Shifts of Trace Impurities: Common Laboratory Solvents, Organics, and Gases in Deuterated Solvents Relevant to the Organometallic Chemist. *Organometallics* **2010**, *29*, 2176–2179. [[CrossRef](#)]