ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Machine-Learning Approach to Identify Organic Functional Groups from FT-IR and NMR Spectral Data

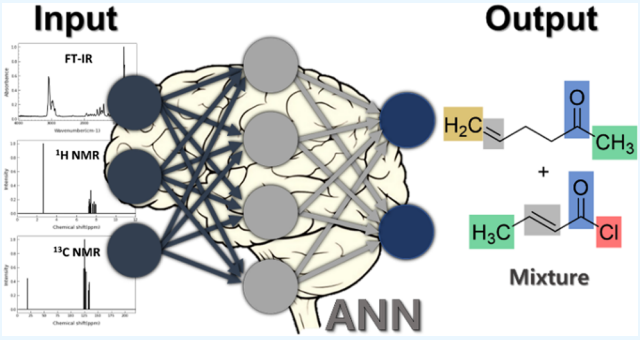
Gwanho Lee,[†] Hyekyoung Shim,[†] Juhyun Cho,[†] and Sang-Il Choi*

Cite This: *ACS Omega* 2025, 10, 12717−12723

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Interpreting spectral data to analyze the structure and properties of unknown chemicals requires a lot of time and effort. Herein, we developed a machine-learning model that simultaneously trains on multiple spectroscopic data to identify functional groups of compounds more accurately and quickly. An artificial neural network model trained on Fourier-transform infrared, proton nuclear magnetic resonance, and ^{13}C nuclear magnetic resonance together identified 17 functional groups with a macro-average F1 score of 0.93, outperforming the model using a single type of spectroscopy. The results indicated that training a machine-learning model with multiple spectral data can provide more accurate structural analysis when analyzing the structure of unknown chemicals, as can using multiple spectroscopy methods simultaneously.



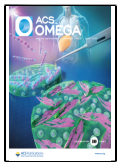
INTRODUCTION

A functional group is a specific group of atoms that determine the properties and reactivity of a compound. Therefore, the identification of functional groups is the fundamental and central task across all fields of chemistry to reveal the structure and properties of unknown chemicals.[1−7] In particular, monitoring changes in functional groups that occur during chemical reactions has contributed significantly to the discovery and design of reaction mechanisms.[4,8,9] As an example, in the pharmaceutical field, functional group identification has been critical in aiding the validation of drug formulations and ensuring quality control of incoming and outgoing materials.[10,11] Recently, in situ/operando functional group identification studies have been extensively conducted in the analysis of catalyst performances related to renewable energy applications, including hydrogen production/storage/utilization, carbon capture/utilization/storage, and nitrogen cycle.[12,13]

Spectroscopic instruments such as Fourier-transform infrared (FT-IR), nuclear magnetic resonance (NMR), and Raman spectroscopy are the best-known assays to identify the functional groups in compounds. However, interpretation of these spectral data is often challenging. For example, peak tables are generally used to interpret spectral data, but the peak ranges of some functional groups may vary depending on the structure of the compound or the experimental conditions.[14] In addition, the spectral interpretation is prone to error because it relies on the intuition and experience of the analyst. Computer programs have been developed that utilize spectral databases to quickly and accurately retrieve functional group analysis,[15,16] but building a database of numerous compounds is expensive and requires continuous updating for new compounds.

Machine learning is a computer program that improves the performance of tasks by learning through experience.[17] Collaborations using machine learning have been actively conducted in analytical chemistry fields. For example, various studies in spectroscopy analysis have been conducted utilizing machine learning to qualitatively and quantitatively analyze matrix[18,19] and noise reduction.[20] Research has also been done on combining FT-IR[21−23] or NMR[24−26] spectra with machine learning to predict molecular structure. However, it is difficult to predict the molecules with a single spectral-based model since each method provides only partial information about the molecule. Therefore, recent studies have reported combining FT-IR, NMR, and Mass spectrometry (MS) spectral data to improve the functional group prediction performance of machine-learning models.[27,28] These results showed that the machine-learning models using multiple spectral data can improve compound structure prediction, as experts would analyze multiple spectra simultaneously. However, the processes of obtaining multiple spectra using the series of analysis methods can be very time-consuming and difficult

Received: February 28, 2025
Revised: March 6, 2025
Accepted: March 12, 2025
Published: March 19, 2025



https://doi.org/10.1021/acsomega.5c01903
ACS Omega 2025, 10, 12717−12723

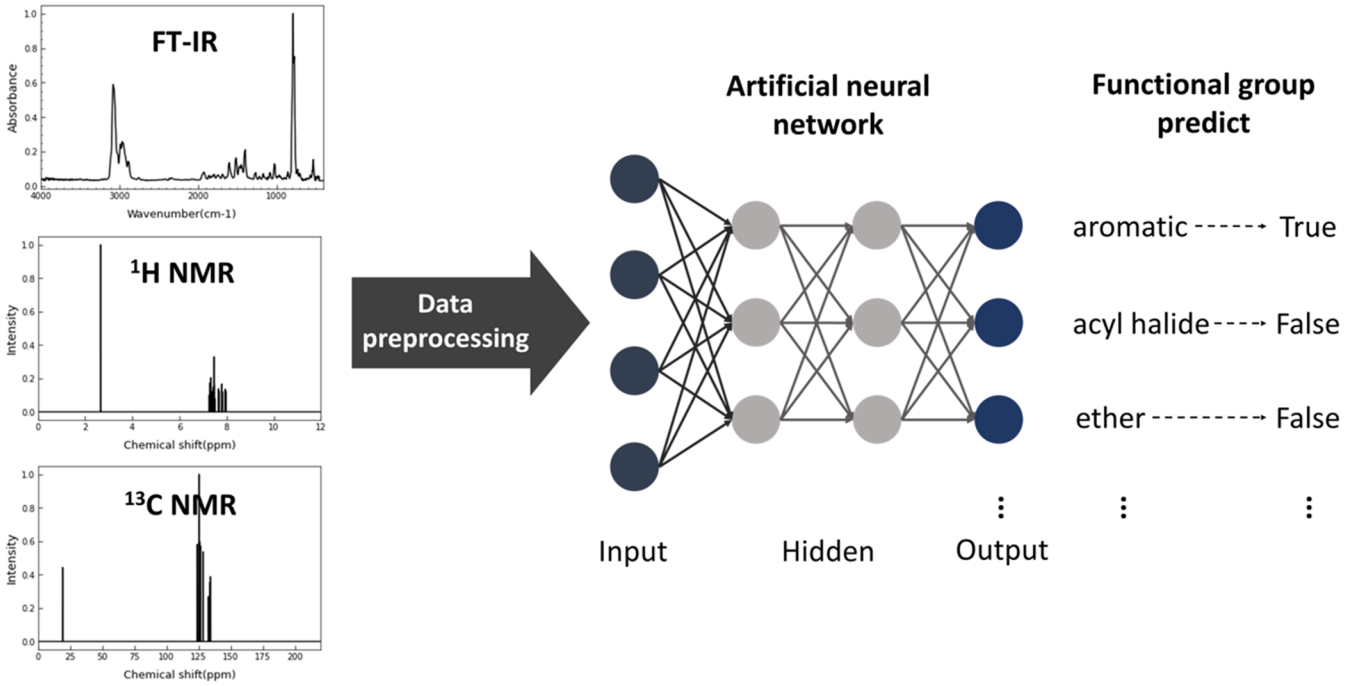


Figure 1. Overview of the machine-learning methodology for the classification of functional groups using FT-IR and NMR spectra.

depending on the research environment or the number of samples.[29,30] Therefore, increasing the level of functional group prediction with a minimum combination of analytical equipment is a pressing challenge in a related field.

Traditionally, FT-IR and NMR (^{1}H, ^{13}C) are minimal analytical combinations that have been used together complementarily in many fields to identify the structures of various organic molecules.[31−35] In addition, it is easy to build a machine-learning database with freely available FT-IR and NMR reference data elsewhere.[36,37] In this study, therefore, we present a machine-learning approach that integrates FT-IR, ^{1}H NMR, and ^{13}C NMR spectral data for the simultaneous identification of 17 functional groups in compounds (Figure 1). This model can analyze a wide range of functional groups, including nitrogen and halogen elements, and even mixtures, thereby outperforming existing models in the literature in terms of performance and applicability.[23,26−28] By applying an artificial neural network (ANN) model, the integrated spectral data identified functional groups with a macro-average F1 score of 0.93, which is an improvement over 0.88 for the model trained on FT-IR spectral data. These results demonstrate that combining FT-IR and NMR spectral data can not only significantly improve the performance of predicting various functional groups but also enable functional group prediction for certain spectra (such as nitriles, alkyl halides, ethers, etc.) that are typically difficult to analyze with FT-IR and NMR due to weak signals. The ultimate goal in this field is to determine the functional groups of unknown mixtures through machine learning of various NMR and FT-IR spectral data. Therefore, we believe that the results of predicting functional groups using machine learning can pave the way toward the research in the direction of predicting the analysis of intermediates during various chemical reactions.

METHODS

Data Collection and Preprocessing. For machine-learning approaches, FT-IR, ^{1}H NMR, and ^{13}C NMR spectra of 3027 compounds were collected. This data set consisted of compounds with a maximum molecular weight of 522 g/mol and up to 36 carbon atoms. FT-IR spectra of compounds in the gas phase were collected from the NIST chemistry WebBook.[38] Additional FT-IR spectra for external validation were also obtained from the NIST Chemistry WebBook. All FT-IR spectra were transformed into 1108 vectors representing wavelengths from 400 to 4000 cm^{-1} with 3.25 cm^{-1} resolution. Transmittance values were transformed into absorbance values to improve model training, and min-max normalization was applied by dividing each absorbance value by the maximum absorbance value of the data. Missing values resulting from this process were estimated by linear interpolation.

^{1}H and ^{13}C NMR spectra were obtained from the SDBS database.[37] Additional ^{1}H and ^{13}C NMR spectra for external validation were obtained from the SDBS database and CAS SciFinder.[39] Since chemical shifts can vary depending on the solvent used, NMR spectra recorded only in $CDCl_3$ solvent were considered to ensure consistency.[40] ^{1}H NMR spectra used in training of models was measured with a JEOL FX-90Q (89.56 MHz), a JEOL GX-400 (399.65 MHz), or a JEOL AL-400 (399.65 MHz). ^{13}C NMR spectra used in training of models was measured with a NEVA NV-14 (15.087 MHz), a JEOL FX-90Q (22.530 MHz), a Varian XL-100 (25.160 MHz), a Bruker AC-200 (50.323 MHz), a JEOL FX-200 (50.183 MHz), a JEOL GX-400 (100.535 MHz), or a JEOL AL-400 (100.40 MHz). ^{1}H NMR spectra used in predicting compounds was measured with a Bruker AVANCE III 500 (500.15 MHz). ^{13}C NMR spectra used in predicting compounds was measured with a Bruker AVANCE III 500 (125.775 MHz). NMR spectra were preprocessed by data binning, which is the generally used process of grouping continuous data into specific categories to reduce the data dimensionality. NMR spectral points were divided into specific bins and replaced with values representing those bins. For ^{1}H NMR, the range of 1−12 ppm was divided into 12 bins with 1 ppm interval, and for ^{13}C NMR, the range of 1−220 ppm was

https://doi.org/10.1021/acsomega.5c01903
ACS Omega 2025, 10, 12717−12723

divided into 44 bins with 5 ppm interval. ^{1}H NMR and ^{13}C NMR spectra were each divided at bins of a constant length, and the number of bins for each kind of spectra was labeled with N_H and N_C. The 1 or 0 was assigned depending on whether there is a peak in a particular bin or not. The intensity value was ignored. We tried to train using the intensity data of the NMR spectra but found that the model performed better when we excluded the data. The reason was that, unlike FT-IR where the spectral data is present in most of the wavenumber range, NMR spectral data is sparse, consisting mostly of zeros, which leads to the "curse of dimensionality" and can degrade model performance.[25] To address this issue, we trained the model using large bin intervals, which helped to alleviate the data sparsity but made it challenging to incorporate intensity information. Consequently, the model was trained to recognize only the presence or absence of peaks, which is expected to improve the performance of the model by incorporating the whole information on NMR spectral data without the curse of dimensionality.

Functional Group Assignment. FT-IR, ^{1}H NMR, and ^{13}C NMR spectra of 3027 compounds were collected. For each compound, the presence of 17 functional groups (aromatic, acyl halide, ether, alcohol, ester, methyl, nitro, alkane, carboxylic acid, amine, aldehyde, alkyne, ketone, alkyl halide, amide, alkene, and nitrile) was determined using SMARTS strings, which is a line notation that encodes molecular structures as short ASCII strings. We selected 17 functional groups based on previous work.[27] Table S1 shows the SMARTS strings of 17 functional groups. We generated target data for classification by assigning 0 or 1 depending on the absence or presence of a specific functional group. Figure S1 illustrates the distribution of compounds used in modeling, organized according to their specific functional groups. There are more than 1000 compounds containing aromatic, methyl, and alkane, while there are fewer than 100 compounds containing acyl halide, aldehyde, alkyne, and amide.

ANN Model. ANN is one of the most used machine-learning algorithms, which is inspired by the principle and structure of neurons. An ANN model consists of an input layer, one or several hidden layers, and an output layer. A layer consists of nodes, and a layer $A = [a_1, a_2, \cdots, a_n]$ with n nodes is calculated and connected to the node b_j ($j = 1, 2, \cdots$) of the next layer $B = [b_1, b_2, \cdots, b_n]$ as follows[41]

$$b_j = \sigma\left(\sum_i^n \omega_{ij} a_i + T\right)$$

where ω_{ij} is the weight, T is the threshold, and σ is the activation function. The weights and thresholds of the neural network were optimized to minimize the cost, which is the sum of the errors between the predicted and actual values.

Stratified *K*-Fold Cross Validation. Cross validation is a resampling method to avoid overfitting and to create more generalized models. We applied multilabel stratified 5-fold cross validation to our model. 20% of the collected data was used as test data for model performance evaluation, and multilabel stratified K-fold cross validation[42] was applied to the remaining 80% to train the model. In K-fold cross validation, the data was divided into K subsets of equal size. One subset was used as validation data for optimizing model parameters and the remaining K-1 subsets as training data for model training. The process was repeated K times.[43] The performance of the final model with the average of K model results was evaluated to avoid overfitting and to create a more generalized model. Also, multilabel stratified random sampling was applied to equally divide the class ratio of each subset to avoid assigning too much or too little data from a specific class.

Performance Evaluation Metrics. Precision, recall, and F1 score were used to evaluate the classification performance. Precision and recall were calculated as the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP, TN, FP, and FN are confusion matrices, which measure the predictive performance of the model by visualizing the predicted versus actual values. TP is the case when both the prediction and actual data are determined to be true. TN is the case when the prediction and actual data are determined to be false. FP is the case when the prediction is true, but it actually is false. FN is the case when the prediction is false but it actually is true (Table S2). TP indicates the result of accurately predicting the presence of a label, while TN indicates the result of accurately predicting the absence of a label. FP and FN represent the results of incorrectly predicting the presence and absence of labels, respectively. Precision is the proportion of predicted positives out of actual positives, and recall is the proportion of actual positives out of predicted positives, as shown in the equations below.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

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

Both metrics have values between 0 and 1, with higher values indicating better performance. The F1 score is a matric to reflect both precision and recall and is calculated as the harmonic average of precision and recall.

$$\text{F1 score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

For class imbalanced data, the performance of the minority class is almost ignored when using a microaverage since it can be dominated by the scores of majority class. Therefore, the performance of each model was compared using the macro-average to ensure that the performance of the minority class was also well represented.[44]

Training of ANN Model. The Python package Keras was used to create machine-learning models.[45] An ANN model was built with two hidden layers for multilabel classification of 17 functional groups in compounds. The output layer consists of 17 nodes that calculate the probability of the existence of each functional group. As the number of hidden layers increases, creating more complex models can improve performance but also increases the likelihood of overfitting problems. Moreover, as the size of the network increases, the training time becomes longer. To avoid this problem, dropout was applied that removes nodes from a layer with a certain probability.[46] Figure S2 shows the structure of the ANN model used for training (Figure 2).

In this experiment, the rectified linear unit (ReLU) and sigmoid were used as activation functions for the hidden layer and output layer, respectively.

$$\text{ReLU} = \max(0, x)$$

$$\text{sigmoid}(x) = 1/(1 + e^{-x})$$

https://doi.org/10.1021/acsomega.5c01903
ACS Omega 2025, 10, 12717−12723

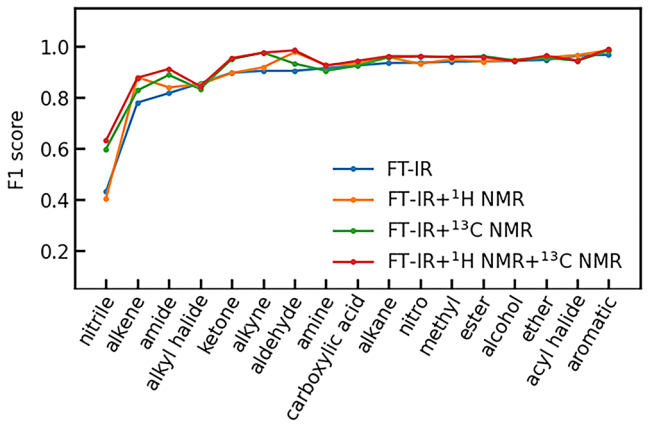


Figure 2. Comparison of compound functional groups F1 scores of models using FT-IR data, FT-IR + ^{1}H NMR data, FT-IR + ^{13}C NMR data, and FT-IR + ^{1}H NMR + ^{13}C NMR data.

For multilabel classification, binary cross-entropy (BCE) was used as a loss function to measure the error between actual and predicted values.

$$\mathrm{BCE} = -y \log \hat{y} + (1 - y) \log(1 - \hat{y})$$

$\hat{y}$ is the predicted value by the ANN model and y is the actual value.

RESULTS AND DISCUSSION

Identification of Functional Groups Using Combinations of FT-IR, ^{1}H NMR, and ^{13}C NMR Spectra. To identify 17 functional groups in unknown compounds, a machine-learning model was trained using FT-IR, ^{1}H NMR, and ^{13}C NMR spectra. We collected FT-IR, ^{1}H NMR, and ^{13}C NMR spectra data for 3027 compounds from the NIST chemistry Webbook and SDBS databases for machine learning for molecular structure prediction.[36,37] Machine learning was performed using FT-IR, ^{1}H NMR, ^{13}C NMR, FT-IR + ^{1}H NMR, FT-IR + ^{13}C NMR, ^{1}H NMR + ^{13}C NMR, and FT-IR + ^{1}H NMR + ^{13}C NMR combinations. The F1 scores for each feature group were obtained from the recall and precision values for FT-IR (Table S3), ^{1}H NMR (Table S4), ^{13}C NMR (Table S5), FT-IR + ^{1}H NMR (Table S6), FT-IR + ^{13}C NMR (Table S7), ^{1}H NMR + ^{13}C NMR (Table S8), and FT-IR + ^{1}H NMR + ^{13}C NMR spectra (Table S9). Comparing the macro-average F1 scores of the models trained on the data separately (Figure 2), we found that FT-IR, ^{1}H NMR, and ^{13}C NMR achieved 0.88, 0.41, and 0.68, respectively, indicating that FT-IR data are more informative for functional group identification. However, the F1 scores of alkyl halides, amides, and alkenes showed relatively low performances of 0.86, 0.82, and 0.78, respectively, and the nitrile group had a very low F1 score of 0.43. This is because class overlap problems occurred in areas where absorption bands of different functional groups are similar in the FT-IR spectrum, resulting in poor classification performance. The absorption bands of alkyl halides and alkenes appeared in the range of 1200−600 cm^{-1}, called the fingerprint region. Because of the high complexity and numerous peaks in the fingerprint region, identification of the functional groups is difficult. In the case of amides, the CO absorption bands can be confused with those of aldehydes, ketones, and esters. In addition, the N−H absorption band of amides largely overlaps with that of amines. The absorption band associated with the C≡N triple bond in nitriles may overlap with that of the C≡C triple bond in alkynes within the range of 2200−2000 cm^{-1}. This band often exhibits very weak intensity, complicating the observations.

NMR analysis provides more accurate and detailed information about molecule structure and atomic arrangement by observing local magnetic fields around atomic nuclei placed in a magnetic field. Therefore, we developed a new training model by combining FT-IR with NMR spectra to solve the class overlap problem that occurred in the FT-IR model solely. The macro-average F1 scores for the combined models of FT-IR + ^{1}H NMR, FT-IR + ^{13}C NMR, and FT-IR + ^{1}H NMR + ^{13}C NMR were 0.90, 0.91, and 0.93, respectively, outperforming models trained solely on FT-IR spectra. Notably, the improved identification of substances that can overlap in the FT-IR spectra, such as aldehydes, alkynes, ketones, amides, and alkenes, suggests that complementing FT-IR with NMR spectra addresses its limitations. The incorporation of a single ^{13}C NMR data set into the FT-IR model significantly enhanced the F1 score for detecting the nitrile group, elevating it from 0.43 to 0.60. Subsequent addition of ^{1}H NMR data into the augmented FT-IR + ^{13}C NMR model further improved the F1 score, achieving a value of 0.63. The F1 scores of the alkene and amide were also improved from 0.78 and 0.82 to 0.88 and 0.91, respectively. This suggests that ^{13}C NMR and ^{1}H NMR spectra data have complementary effects on model training. Considering that this FT-IR + ^{1}H NMR + ^{13}C NMR model has F1 scores for most functional groups that are higher than those of single FT-IR data models, we infer that adding NMR data to FT-IR data can improve performance without significant loss in F1 scores. In order to effectively predict the structure and properties of a compound, it is more important to accurately identify all functional groups in a molecule than to predict individual functional groups. Therefore, we introduced another metric called the exact match ratio (EMR) to evaluate our models. The EMR calculated the proportion of samples for which all labels (functional groups) were correctly classified. As shown in Figure S3, the FT-IR + ^{1}H NMR + ^{13}C NMR model achieved the highest score of 79.2, among other models. We introduced external validation technique[46] to evaluate our models and achieved a macro-average F1 score of 0.90, which is comparable to the model's macro-average F1 score of 0.93 (Table S10). This result demonstrates that our model can predict functional groups with similar accuracy on novel compound data not included in the training database. And we applied y-randomization[47] and observed a significant decrease in the macro-average F1 score to 0.17 (Table S11), confirming that there is a meaningful correlation between the spectra data and the functional groups learned by our model.

Attempting to Solve the Precision and Recall Imbalance Using Weighted Binary Cross-Entropy (WBCE). While real molecules typically have up to 5 functional groups, the model predicts 17 functional groups. This means that the number of nonexistent functional groups is bigger than the number of functional groups actually exist in the molecule. In this case, the more functional groups the model determines do not exist, the easier it is to improve the performance. This is known as a false negative (FN) error, and as the FN increases, the recall value and F1 score decrease, which is affected by recall and precision. There are three ways to solve the imbalances in these data sets: undersampling, oversampling, and weighting. Since the data applied to our model is a multilabel data set, oversampling may lead to overfitting, and

https://doi.org/10.1021/acsomega.5c01903
ACS Omega 2025, 10, 12717−12723

Table 1. Identification of Functional Groups in Unknown Compounds and Mixtures Is Based on Experimental Results

	Compound 1		Compound 2		Compound 3		Compound 4		Mixture 1		Mixture 2	
	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
Alkane	~~0~~	~~1~~	1	1	1	1	~~0~~	~~1~~	1	1	1	1
Methyl	0	0	0	0	~~0~~	~~1~~	1	1	~~0~~	~~1~~	~~1~~	~~0~~
Alkene	~~1~~	~~0~~	~~0~~	~~1~~	0	0	0	0	0	0	1	1
Alkyne	0	0	0	0	0	0	0	0	0	0	0	0
Alcohol	0	0	~~1~~	~~0~~	1	1	0	0	1	1	0	0
Amine	0	0	0	0	0	0	0	0	~~0~~	~~1~~	0	0
Nitrile	0	0	0	0	0	0	1	1	0	0	0	0
Aromatic	1	1	1	1	0	0	1	1	1	1	0	0
Alkyl halide	~~0~~	~~1~~	~~1~~	~~0~~	0	0	~~0~~	~~1~~	0	0	0	0
Ester	0	0	0	0	0	0	0	0	0	0	0	0
Ketone	1	1	0	0	0	0	0	0	0	0	1	1
Aldehyde	0	0	~~0~~	~~1~~	0	0	0	0	~~1~~	~~0~~	0	0
Carboxylic acid	0	0	~~1~~	~~0~~	0	0	0	0	0	0	0	0
Ether	0	0	0	0	0	0	1	1	0	0	0	0
Acyl halide	0	0	0	0	0	0	0	0	0	0	~~1~~	~~0~~
Amide	0	0	0	0	0	0	0	0	0	0	0	0
Nitro	0	0	0	0	0	0	0	0	~~1~~	~~0~~	0	0

undersampling may lead to data loss.[48] Therefore, we adopted weighted binary cross-entropy (WBCE), which assigns a higher weight for predictions indicating the presence of functional groups as follows (Figure S4).

$$\mathrm{WBCE} = w_{+} y \log \hat{y} + w_{-}(1 - y) \log (1 - \hat{y})$$

where, w_+ and w_- are the weights for the positive and negative classes, indicating the presence and absence of functional group. They are inversely proportional to the sample size of each class, meaning that the smaller the sample, the higher the weight to adjust for imbalance between classes. The weight for class was calculated as follows.

$$w_{+} = \frac{\text{total number of samples}}{2 \times \text{number of positive samples for class}}$$

$$w_{-} = \frac{\text{total number of samples}}{2 \times \text{number of negative samples for class}}$$

The Adam optimizer,[49] one of the gradient descent algorithms, was used for weight training of neural networks. The learning rate and batch size were fixed at 0.0005 and 64, respectively. We fixed the hyperparameters β_1, β_2, and epsilon at 0.9, 0.999, and 1e−7, respectively. Training of the neural network model was performed up to 500 epochs, with an early stop applied to stop training if the validation macro-average F1 score did not increase for 50 epochs. We recorded changes of F1 scores for training and validation of FT-IR + ^{1}H NMR + ^{13}C NMR data sets across the epochs with the k-fold cross-validation technique (k = 5) (Figure S5). The numbers of epochs for the five training and validation cycles were 170, 159, 151, 155, and 174. The blue line represents the F1 score of the training data set, and the yellow line represents the F1 score of the validation data set. As a result, we found that the F1 score for some functional groups increased slightly, but for functional groups such as acyl halide, ketone, alkyl halide, and alkene, the F1 score of the model without WBCE was actually higher. We speculated that this is due to the inherent trade-off relationship between recall and precision, where an increase in one typically results in a decrease in the other.[50,51] Therefore, we concluded that it is difficult to improve the performance of the model by addressing the imbalance as the F1 score does not change or decrease even with WBCE.

We validated the performance of the model using spectral data (Figures S6−S11) of unknown compounds and mixtures that were not in the database used to train the FT-IR + ^{1}H NMR + ^{13}C NMR model. Table 1 shows the results of identifying the unknown compounds and their mixtures by using the model. Unknown compounds 1,2,3,4 are trans-chalcone, α-bromophenylacetic acid, 1,2-cyclohexanediol, and anisonitrile, respectively. And mixtures 1,2 are 4-nitro-benzaldehyde + 1,2-cyclohexanediol and 5-hexene-2-one + crotonyl chloride, respectively. The blue colored rectangles represent successfully predicted functional groups, while the red-slashed lines represent failed predictions. In addition, 0 indicates that the functional group is not present, and 1 indicates that the functional group is present. The model successfully predicted 14, 12, 16, and 15 of the 17 functional groups in each compound, and 13 and 15 of the functional groups in each mixture, respectively. These results demonstrate practical applicability to the analysis of unknown compounds by machine-learning prediction, and that a deep learning

https://doi.org/10.1021/acsomega.5c01903
ACS Omega 2025, 10, 12717−12723

model trained on single compound spectra is effective at predicting functional groups in mixtures. In particular, the composition of mixtures 1 and 2 consisted of a combination of a cyclic hydrocarbon compound and an aromatic compound (mixture 1) and a combination of aliphatic compounds with similar structures containing a carbonyl group (mixture 2). It is surprising that the predictions were as good as those for single compounds. This result showed that pattern recognition inherent in the complex spectral data sets for mixtures was effectively achieved due to efficient parallel processing by each independent neuron of the ANN-based model.

CONCLUSIONS

In summary, we developed an ANN-based machine-learning model capable of pattern recognition of FT-IR and NMR spectral data to identify functional groups in compounds. The ANN model trained on FT-IR, ^{1}H NMR, and ^{13}C NMR recorded a macro-average F1 score of 0.93 for 17 functional groups and successfully predicted the functional groups in the mixtures of unknown chemicals. These results indicate that integrating NMR and FT-IR data can address the overlapping fingerprint region in FT-IR spectra, making the machine-learning approaches valuable for spectral interpretation and compound structure prediction. In particular, functional group prediction for mixtures is as good as functional group prediction performance for single compounds, suggesting that the ANN-based machine-learning model is effective in pattern recognition for complex spectral data. This is a significant step toward automatic functional group identification algorithms. We believe that improving the performance of the model through the development of more effective data preprocessing methods and the addition of spectral data will allow us to achieve accurate predictions of functional groups for unknown mixtures, which will lead to new analytical systems that support a wide range of research and industrial applications.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.5c01903.

Distribution of compounds used in modeling, organized according to their specific functional groups; structure of the final generated ANN model; comparison of EMR for the models; comparison of functional group F1 scores of ANN models using FT-IR spectra depending on whether class weight is applied or not; FT-IR, ^{1}H NMR, ^{13}C NMR spectral data of compounds and mixtures; SMARTS strings used to identify the presence of a functional group given the 2D topology of a molecule; confusion matrix; precision, recall, F1 score for functional group identification model (PDF)

AUTHOR INFORMATION

Corresponding Author

Sang-Il Choi − *Department of Chemistry and Green-Nano Materials Research Center, Kyungpook National University, Daegu 41566, Republic of Korea;* orcid.org/0000-0002-8280-3100; Email: sichoi@knu.ac.kr

Authors

Gwanho Lee − *Department of Chemistry and Green-Nano Materials Research Center, Kyungpook National University, Daegu 41566, Republic of Korea;* orcid.org/0009-0000-4420-6455

Hyekyoung Shim − *Department of Chemistry and Green-Nano Materials Research Center, Kyungpook National University, Daegu 41566, Republic of Korea;* orcid.org/0009-0003-2146-2757

Juhyun Cho − *Department of Chemistry and Green-Nano Materials Research Center, Kyungpook National University, Daegu 41566, Republic of Korea;* orcid.org/0000-0002-9475-3198

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.5c01903

Author Contributions

†G.L., H.S., and J.C. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was financially supported by the National Research Foundation of Korea (RS-2023-00207831 and RS-2024-00346153).

REFERENCES

(1) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559*, 377−381.

(2) Wen, L.; Ding, J.; Duan, L.; Wang, S.; An, Q.; Wang, H.; Zuo, Z. Multiplicative enhancement of stereoenrichment by a single catalyst for Deracemization of Alcohols. *Science* **2023**, *382*, 458−464.

(3) Wang, Y.; Zhang, Y.; Xin, X.; Yang, J.; Wang, M.; Wang, R.; Guo, P.; Huang, W.; Sobrido, A. J.; Wei, B.; Li, X. In situ photocatalytically enhanced thermogalvanic cells for electricity and hydrogen production. *Science* **2023**, *381*, 291−296.

(4) Qin, X.; Xu, M.; Guan, J.; Feng, L.; Xu, Y.; Zheng, L.; Wang, M.; Zhao, J.-W.; Chen, J.-L.; Zhang, J.; Xie, J.; Yu, Z.; Zhang, R.; Li, X.; Liu, X.; Liu, J.-X.; Zheng, J.; Ma, D. Direct conversion of CO and H2O to hydrocarbons at atmospheric pressure using a tio2−x/ni photothermal catalyst. *Nat. Energy* **2024**, *9*, 154−162.

(5) Aggarwal, S.; Vu, A.; Eremin, D. B.; Persaud, R.; Fokin, V. V. Arenes participate in 1,3-dipolar cycloaddition with in situ-generated diazoalkenes. *Nat. Chem.* **2023**, *15*, 764−772.

(6) Van Steenberge, P. H. M.; Sedlacek, O.; Hernández-Ortiz, J. C.; Verbraeken, B.; Reyniers, M.-F.; Hoogenboom, R.; D'hooge, D. R. Visualization and design of the functional group distribution during statistical copolymerization. *Nat. Commun.* **2019**, *10*, No. 3641.

(7) Yoo, J.; Lee, S. M.; Lee, K.; Lim, S. C.; Jeong, M. S.; Kim, J.; Lee, T. G. Functional group inhomogeneity in graphene oxide using correlative absorption spectroscopy. *Appl. Surf. Sci.* **2023**, *613*, 155885.

(8) Zhao, E. W.; Liu, T.; Jónsson, E.; Lee, J.; Temprano, I.; Jethwa, R. B.; Wang, A.; Smith, H.; Carretero-González, J.; Song, Q.; Grey, C. P. In situ NMR metrology reveals reaction mechanisms in redox flow batteries. *Nature* **2020**, *579*, 224−228.

(9) Shi, R.; Zhang, X.; Li, C.; Zhao, Y.; Li, R.; Waterhouse, G. I.; Zhang, T. Electrochemical oxidation of concentrated benzyl alcohol to high-purity benzaldehyde via superwetting organic-solid-water interfaces. *Sci. Adv.* **2024**, *10*, No. eadn0947.

(10) Bunaciu, A. A.; Aboul-Enein, H. Y. Adulterated drug analysis using FTIR spectroscopy. *Appl. Spectrosc. Rev.* **2021**, *56*, 423−437.

(11) Liu, Z.; Shen, T.; Zhang, J.; Li, Z.; Zhao, Y.; Zuo, Z.; Zhang, J.; Wang, Y. A novel multi-preprocessing integration method for the

qualitative and quantitative assessment of wild medicinal plants: Gentiana rigescens as an example. *Front. Plant Sci.* **2021**, *12*, 759248.
(12) Li, X.; Wang, H.; Yang, H.; Cai, W.; Liu, S.; Liu, B. In situ/Operando characterization techniques to probe the electrochemical reactions for energy conversion. *Small Methods* **2018**, *2*, 201700395.
(13) Liao, W.; Wang, S.; Su, H.; Zhang, Y. Application of in situ/Operando characterization techniques in heterostructure catalysts toward water electrolysis. *Nano Res.* **2023**, *16*, 1984−1991.
(14) Pretsch, E.; Bühlmann, P.; Affolter, C. Structure Determination of Organic Compounds, 2020; pp 9−15.
(15) Leung, A. K. M.; Chau, F.; Gao, J.; Shih, T. Application of wavelet transform in infrared spectrometry: Spectral Compression and library search. *Chemom. Intell. Lab. Syst.* **1998**, *43*, 69−88.
(16) Bremser, W. Hose — a novel substructure code. *Anal. Chim. Acta* **1978**, *103*, 355−365.
(17) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, Perspectives, and prospects. *Science* **2015**, *349*, 255−260.
(18) Conroy, J.; Ryder, A. G.; Leger, M. N.; Hennessey, K.; Madden, M. G. In *Qualitative and Quantitative Analysis of Chlorinated Solvents Using Raman Spectroscopy and Machine Learning*, SPIE, 2005; Vol. 5826, pp 131−142.
(19) Wang, Y.-T.; Li, B.; Xu, X.-J.; Ren, H.-B.; Yin, J.-Y.; Zhu, H.; Zhang, Y.-H. FTIR spectroscopy coupled with machine learning approaches as a rapid tool for identification and quantification of artificial sweeteners. *Food Chem.* **2020**, *303*, 125404.
(20) Henríquez, P. A.; Ruz, G. A. Noise reduction for near-infrared spectroscopy data using extreme learning machines. *Eng. Appl. Artif. Intell.* **2019**, *79*, 13−22.
(21) Wang, Z.; Feng, X.; Liu, J.; Lu, M.; Li, M. Functional groups prediction from infrared spectra based on computer-assist approaches. *Microchem. J.* **2020**, *159*, 105395.
(22) Robb, E. W.; Munk, M. E. A neural network approach to infrared spectrum interpretation. *Mikrochim. Acta* **1990**, *100*, 131−155.
(23) Wang, T.; Tan, Y.; Chen, Y. Z.; Tan, C. Infrared Spectral Analysis for prediction of functional groups based on feature-aggregated deep learning. *J. Chem. Inf. Model.* **2023**, *63*, 4615−4622.
(24) Wilkins, C. L.; Isenhour, T. L. Multiple discriminant function analysis of carbon-13 nuclear magnetic resonance spectra. Functional Group identification by Pattern Recognition. *Anal. Chem.* **1975**, *47*, 1849−1851.
(25) Li, C.; Cong, Y.; Deng, W. Identifying molecular functional groups of organic compounds by deep learning of NMR Data. *Magn. Reson. Chem.* **2022**, *60*, 1061−1069.
(26) Specht, T.; Münnemann, K.; Hasse, H.; Jirasek, F. Automated methods for identification and quantification of structural groups from nuclear magnetic resonance spectra using support vector classification. *J. Chem. Inf. Model.* **2021**, *61*, 143−155.
(27) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral deep learning for prediction and prospective validation of functional groups. *Chem. Sci.* **2020**, *11*, 4618−4630.
(28) Pesek, M.; Juvan, A.; Jakoš, J.; Košmrlj, J.; Marolt, M.; Gazvoda, M. Database independent automated structure elucidation of organic molecules based on IR, 1H NMR, 13C NMR, and MS data. *J. Chem. Inf. Model.* **2021**, *61*, 756−763.
(29) Vuckovic, D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography−mass spectrometry. *Anal. Bioanal. Chem.* **2012**, *403*, 1523−1548.
(30) Moon, H.; Wheeler, A. R.; Garrell, R. L.; Loo, J. A.; Kim, C.-J. An integrated digital microfluidic chip for multiplexed proteomic sample preparation and analysis by MALDI-MS. *Lab Chip* **2006**, *6*, 1213.
(31) Lyu, H.; Diercks, C. S.; Zhu, C.; Yaghi, O. M. Porous crystalline olefin-linked covalent organic frameworks. *J. Am. Chem. Soc.* **2019**, *141*, 6848−6852.
(32) Krishnan-Schmieden, M.; Konold, P. E.; Kennis, J. T.; Pandit, A. The molecular ph-response mechanism of the plant light-stress sensor PsbS. *Nat. Commun.* **2021**, *12*, No. 2291.
(33) Yang, X.; Li, Q.; Li, Z.; Xu, X.; Liu, H.; Shang, S.; Song, Z. Preparation and characterization of room-temperature-vulcanized silicone rubber using acrylpimaric acid-modified Aminopropyltriethoxysilane as a cross-linking agent. *ACS Sustainable Chem. Eng.* **2019**, *7*, 4964−4974.
(34) Ma, H.-X.; Li, J.-J.; Qiu, J.-J.; Liu, Y.; Liu, C.-M. Renewable cardanol-based star-shaped prepolymer containing a phosphazene core as a potential biobased green fire-retardant coating. *ACS Sustainable Chem. Eng.* **2017**, *5*, 350−359.
(35) Guan, Y.; Zhang, B.; Tan, X.; Qi, X.-M.; Bian, J.; Peng, F.; Sun, R.-C. Organic−inorganic composite films based on modified hemicelluloses with clay nanoplatelets. *ACS Sustainable Chem. Eng.* **2014**, *2*, 1811−1818.
(36) NIST Chemistry WebBook. *NIST Standard Reference Database Number 69*; Linstrom, P. J.; Mallard, W. G., Eds.;National Institute of Standards and Technology: Gaithersburg, MD, 20899, 2005.
(37) SDBSWeb. https://sdbs.db.aist.go.jp (National Institute of Advanced Industrial Science and Technology, date of access).
(38) Linstrom, P. J.; Mallard, W. G. The NIST chemistry webbook: A Chemical Data Resource on the internet. *J. Chem. Eng. Data* **2018**, *9*, 1−1951.
(39) Chemical Abstracts Service. SciFinder. https://scifinder.cas.org/ (accessed Nov 19, 2024).
(40) Dračínský, M.; Bouř, P. Computational analysis of solvent effects in NMR spectroscopy. *J. Chem. Theory Comput.* **2010**, *6*, 288−299.
(41) Wang, S.-C. Artificial Neural Network. In *Interdisciplinary Computing in Java Programming*, 2003; pp 81−100.
(42) Sechidis, K.; Tsoumakas, G.; Vlahavas, I. On the Stratification of Multi-Label Data. In *Lecture Notes in Computer Science*, 2011; pp 145−158.
(43) De Angeli, K.; Gao, S.; Danciu, I.; Durbin, E. B.; Wu, X.-C.; Stroup, A.; Doherty, J.; Schwartz, S.; Wiggins, C.; Damesyn, M.; Coyle, L.; Penberthy, L.; Tourassi, G. D.; Yoon, H.-J. Class imbalance in out-of-distribution datasets: Improving the robustness of the TEXTCNN for the classification of rare cancer types. *J. Biomed. Inf.* **2022**, *125*, No. 103957.
(44) Keras, C. F. *GitHub*, 2015, https://github.com/fchollet/keras.
(45) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929−1958.
(46) Kefeli, J.; Berkowitz, J.; Acitores Cortina, J. M.; Tsang, K. K.; Tatonetti, N. P. Generalizable and automated classification of TNM stage from pathology reports with external validation. *Nat. Commun.* **2024**, *15*, No. 15.
(47) Fu, L.; Liu, L.; Yang, Z. J.; Li, P.; Ding, J. J.; Yun, Y. H.; Lu, A. P.; Hou, T. J.; Cao, D. S. Systematic Modeling of log D 7.4 Based on Ensemble Machine Learning, Group Contribution, and Matched Molecular Pair Analysis. *J. Chem. Inf. Model.* **2020**, *60*, 63−76.
(48) Tarekegn, A. N.; Giacobini, M.; Michalak, K. A review of methods for imbalanced multi-label classification. *Pattern Recognit.* **2021**, *118*, 107965.
(49) Kingma, D. P.; Ba, J. 2014, arXiv:1412.6980. arXiv.org e-Print archive. https://arxiv.org/abs/1412.6980.
(50) Agaal, A.; Essgaer, M.; Alshareef, A.; Alkhadafe, H.; BenYahmed, Y. In *Application of Classification and Regression Tree and Spectral Clustering to Breast Cancer Prediction: Optimizing the Precision-Recall Trade-Off*, 2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), 2023; pp 311−317.
(51) Verine, A.; Negrevergne, B.; Pydi, M. S.; Chevaleyre, Y. 2023, arXiv:2302.00628. arXiv.org e-Print archive. https://arxiv.org/abs/2302.00628.

https://doi.org/10.1021/acsomega.5c01903
ACS Omega 2025, 10, 12717−12723