ACS OMEGA

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

Article

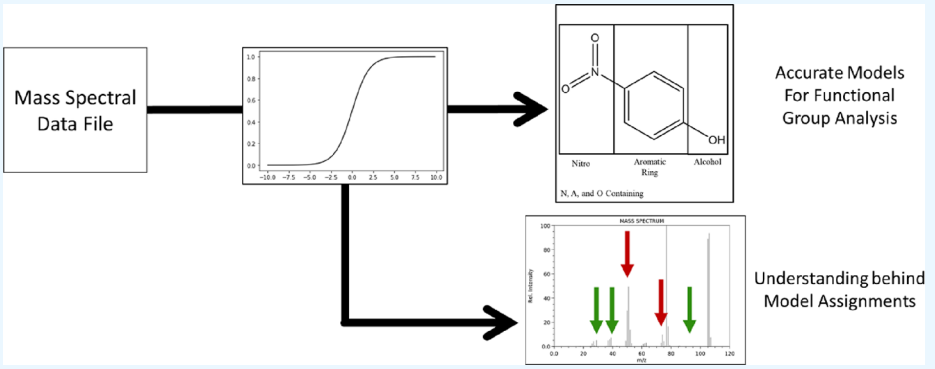
Array-Based Machine Learning for Functional Group Detection in Electron Ionization Mass Spectrometry

Nicole M. North, Abigail A. Enders, Morgan L. Cable, and Heather C. Allen*

Cite This: *ACS Omega* 2023, 8, 24341−24350

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



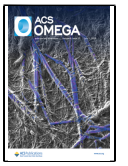
ABSTRACT: Mass spectrometry is a ubiquitous technique capable of complex chemical analysis. The fragmentation patterns that appear in mass spectrometry are an excellent target for artificial intelligence methods to automate and expedite the analysis of data to identify targets such as functional groups. To develop this approach, we trained models on electron ionization (a reproducible hard fragmentation technique) mass spectra so that not only the final model accuracies but also the reasoning behind model assignments could be evaluated. The convolutional neural network (CNN) models were trained on 2D images of the spectra using transfer learning of Inception V3, and the logistic regression models were trained using array-based data and Scikit Learn implementation in Python. Our training dataset consisted of 21,166 mass spectra from the United States' National Institute of Standards and Technology (NIST) Webbook. The data was used to train models to identify functional groups, both specific (e.g., amines, esters) and generalized classifications (aromatics, oxygen-containing functional groups, and nitrogen-containing functional groups). We found that the highest final accuracies on identifying new data were observed using logistic regression rather than transfer learning on CNN models. It was also determined that the mass range most beneficial for functional group analysis is 0−100 m/z. We also found success in correctly identifying functional groups of example molecules selected from both the NIST database and experimental data. Beyond functional group analysis, we also have developed a methodology to identify impactful fragments for the accurate detection of the models' targets. The results demonstrate a potential pathway for analyzing and screening substantial amounts of mass spectral data.

INTRODUCTION

Functional group identification is an important strategy for molecular structure analysis in analytical techniques such as mass spectrometry.[1−4] Mass spectrometry often looks at the fragmentation of molecules so that the original (parent) structure may be elucidated.[5−7] Such analyses can be challenging. The presence of functional groups can aid in predicting where fragments will occur; however, identifying specific fragments corresponding to the presence of functional groups proves difficult.[8] Machine learning (ML) methods aid in pattern recognition when supplied with large data sets. This couples nicely with mass spectrometry's fragmentation patterns, making ML a promising tool to identify functional groups and, thus, fragments of interest.[9−12]

Generally, mass spectrometry is not as commonly used for bulk functional group analysis without the use of extra sample preparation or tandem mass spectrometry techniques (MS/MS).[13,14] For example, in previous works, the analysis of amino acids has been aided by derivatization via ninhydrin prior to using high-performance liquid chromatography and tandem mass spectrometry for analysis.[15,16] It is also possible to use tandem mass spectrometry approaches including triple quadrupole mass spectrometry to preform precursor ion scanning to screen for functional groups.[17] These approaches

Received: March 13, 2023
Accepted: May 22, 2023
Published: June 29, 2023



are invaluable to the mass spectrometry community because they allow for the in-depth analysis of chemical compounds. In addition, these approaches have created a higher level of understanding of complex analyte mixtures inclusive of those containing high mass molecules, for example, in the field of proteomics. However, there are circumstances in which prior derivatization, separation, and tandem methods are not feasible; usually, time, resources, and/or location make such analysis impossible, such as with field-based analyses and planetary probes.

The employment of ML has the potential to overcome many of the challenges faced in analyzing mass spectra under limiting conditions. ML approaches have a strong backing in the literature regarding their ability to classify organic molecules through their fragmentation patterns. Examples include CANOPUS,[18] which works to predict thousands of classes of molecules using MS/MS data, or MSNovelist,[19] which was able to identify the structures of molecules that the model had never seen in the training phase. Similarly, CSI:FingerID[20] also utilizes MS/MS spectra to assist in searching a molecular structure database. Another application that takes advantage of the intersection of mass spectrometry and ML is in the understanding of metabolite chemistry.[21,22] There are also many papers utilizing ML with mass spectrometry to perform rapid screening methodologies for specific analytes of interest.[23,24] These ML methods have had powerful results and have been revolutionary in our implementation of mass spectral methods.

In this study, we aim to achieve meaningful fragment analysis using ML methods that do not require the use of tandem mass spectral techniques or controlled sample preprocessing. We generate a simplified method that can be applied in situations in which more sophisticated mass spectrometry techniques are not feasible, opening the door to many applications that have, to this point, been inaccessible with the current analytical techniques. We achieve this goal by only using single mass analyzer data, meaning that further fragmentation information on parent fragments is unavailable. By doing minimal preprocessing, particularly in not manually selecting peaks of interest, we generate models that need to develop their own understanding of fragmentation patterns, which we can evaluate. In doing so, we explore how a generalized method for analyzing mass spectra informs the interpretation of mass spectra for functional group analysis. Our methodology enables us to probe the model assignment mechanism, which further improves how we understand the functional group assignment and ML techniques.

Herein, we present a comparison of functional group analysis methods from electron ionization−mass spectrometry (EI-MS) spectra. We evaluate the success of two ML approaches, transfer learning on a previously trained convolutional neural network (CNN) and logistic regression (LR). Transfer learning has previously been successful in identifying functional groups from infrared (IR) spectral data;[25] therefore, its application to functional group analysis in mass spectra was evaluated. In contrast to transfer learning, LR provides a simpler architecture to allow for further analysis into the impact of the features themselves on the outcome of the models.

The transfer learning and LR algorithms were used with the same set of mass spectral data to identify specific functional groups (e.g., amines and esters) within molecules as well as place the molecules into generalized classifications based on these functional groups (aromatics, O-containing functional groups, and N-containing functional groups). We first explain the process of organizing the spectra obtained from the National Institute of Standards and Technology (NIST) Webbook through web scraping. We then show how the classifications of molecules are assigned prior to training followed by adjusting the different training parameters and how they affect both the final training and testing accuracies of the models. We then dive deeper into the LR-based models to explore how adjusting the mass ranges affects the model accuracies as well as the methods to quantify how the model is making its predictions.

METHODS

Spectral Preprocessing and Machine Learning Parameter Selection. Prior to training the CNN and LR models, the data was sorted and labeled. Jupyter notebooks describing these processes along with the model training will be available on our GitHub (https://github.com/Ohio-State-Allen-Lab/Mass_Spec_Functional_Group_ML). Data sorting and labeling was completed by identifying the functional groups that each molecule contained; this identification was done by looking at the InChiKeys. Certain segments of the InChiKey correlate with specific functional groups, allowing for the labeling of molecules. This process was tailored for our purposes from another publication.[26] After identifying the presence or absence of individual functional groups, the molecules were then sorted into the more generalized functional group classifications (e.g., alcohol, amine, etc.; Tables S1 and S2). After defining each of the functional groups, the number of available spectra for each functional group identification was determined. Figure 1 shows the distribution of the functional groups present in the NIST mass spectra.

All the mass spectra were normalized to their most intense fragment peak to ensure that all the y axes were scaled the same way. NIST mass spectra only reports intensities for mass fragments over a certain intensity. For these data to be able to be compared to each other, they all needed to have the same dimensionality. To match up the data, the unreported peaks were filled with a correlated 0 intensity. This was done based on the fact that the non-reported peaks were assumed to be in the noise of the instrument. This is a limitation because the addition of the zeroes, although necessary for the training of the models, does artificially inflate the signal-to-noise ratio of the data. This preprocessing was sufficient to prepare the data for the LR-based models. The CNN-based models required further preprocessing.

For the CNN-based models, the data was plotted. These plots were then used as the input data. All the spectra were saved with the same output parameters, so the resolution of the plots is consistent. However, further analysis of the pixel resolution of the exported plots showed that the plots are fewer pixels wide than there are mass values. This means that each pixel is not defining one mass channel as one would expect; this leads to an artificial reduction in mass resolution, which likely is the source of the lack of success for this approach. We do run into an artificial reduction in the resolution of the mass spectral data. The exported plots are 2D representations of the data and should not be confused with hyperspectral imaging, which would generate 3D data.

For both methodologies, it was necessary to scale the number of spectra that did not contain the model's functional

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

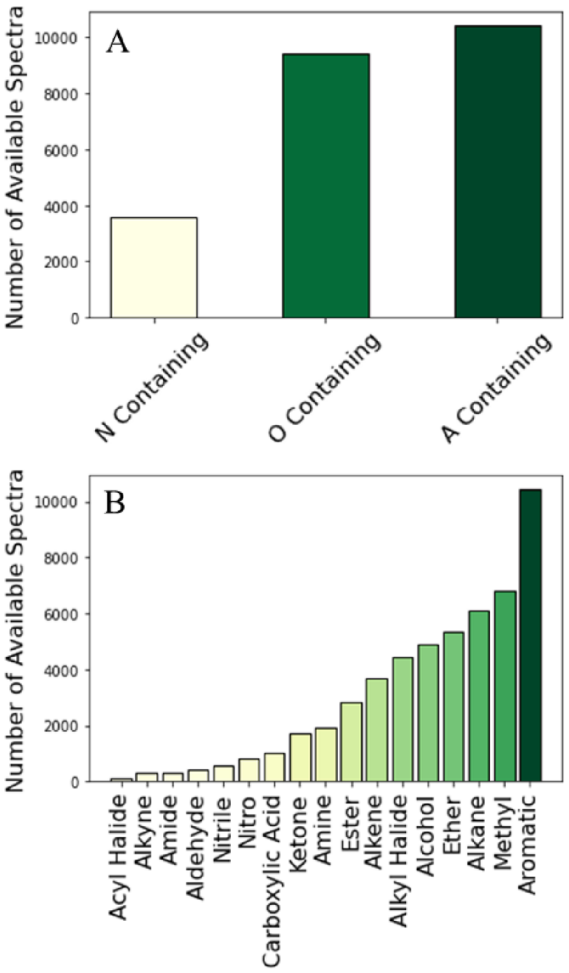


Figure 1. Distributions of available mass spectra from NIST included in this study. (A) Generalized functional group classifications. (B) Specific functional groups. Aromatics are listed as a specific functional group to help correlate the relative distribution between the generalized models and the functional group-specific ones.

group of interest. The number of spectra that did contain a given functional group or functional group classification was always outnumbered by the number of spectra that did not contain the given functional group or functional group classification. Because of this, spectra were randomly removed from the negative case in order to even out the classes preventing the models from always predicting the not-present class due to a disparity in the data. This also means that if a molecule had multiple unique functional groups, that spectra would be used in some way for each of the represented functional group models. Figure 2 shows a histogram demonstrating that the average number of unique functional groups present in molecules from the NIST database is three.

After preprocessing the data, it was separated into training and testing data sets (Figure S1). Once the model was trained, the test data was then used to determine how well the models performed on previously unseen data. The number of withheld test spectra was different for the two different parts of the project. When comparing the CNN- and LR-based approaches, only 10 spectra from each class were withheld. This was limited by computational expenses. The workstation that was utilized to run all the training and analysis was insufficient to run more than 10 test samples at a time. When focusing on using only the LR-based models, 50 test spectra were withheld

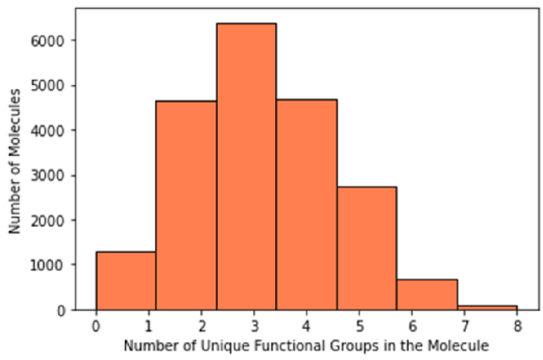


Figure 2. Histogram depicting the number of unique functional groups (duplicate functional groups within a molecule are not counted) present in each molecule from the NIST database. The largest distribution is molecules that contain three unique functional groups. Because the majority of molecules contain multiple functional groups, they can be used to represent the positive case for multiple functional group models.

from each class before training. This allowed for further analysis of the accuracies of the models. After the testing data had been removed, the remaining data was parsed into an 80:20 split of training and internal validation.[19]

Supplemental Experimental Data Collection. Mass spectra for multiple compounds were collected for further model analysis on experimental data outside of the NIST dataset. The data was collected from an Agilent 8890 GC coupled with a 5977B MSD.

Model Training and Testing. *CNN and Inception V3.* The architecture for our CNN in this work was a retraining of Inception V3,[27] a computer vision model. Inception V3 was trained on and has attained a greater than 78.1% accuracy on the ImageNet dataset (a large data set of millions of images with thousands of different words or word phrases labeled to them, a common test dataset in the computer vision realm[28−30]). ImageNet is certainly very different than a dataset consisting of 2D representations of mass spectra; however, the process of transfer learning on unrelated datasets has shown success in the literature.[31−33] The Inception architecture has been used explicitly in the past for spectral processing applications.[34] Image-processing CNNs have been used in other mass spec studies, for example, in 2019, Tran and colleagues developed DeepNovo-DIA, which utilized intensity vectors to train a model to identify peptides.[35] This history in the literature coupled with this approach's success with image-based IR data in our prior publication drove our decision to utilize the retraining of Inception V3.[25] These models were trained using a learning rate of 0.1 and training step ranges between 200 and 20,000 steps.

Logistic Regression through SciKit Learn. The LR models were developed using SciKit Learn's logistic regression classifier. In our utilization, we use the newton-cg solver. LR was chosen as our alternative ML approach due to its simplicity. Using a less computationally complex and specifically binary-classifying model allows for further analysis of where the inferences and assignments of the models are coming from. As we will show later, this simplicity allows us to adjust the dataset and evaluate how those changes affect the model outcomes.

LR is typically used as a binary classifier.[36,37] This is because of the mathematics behind the architecture; the training of the models is working to identify the classification by maximizing

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

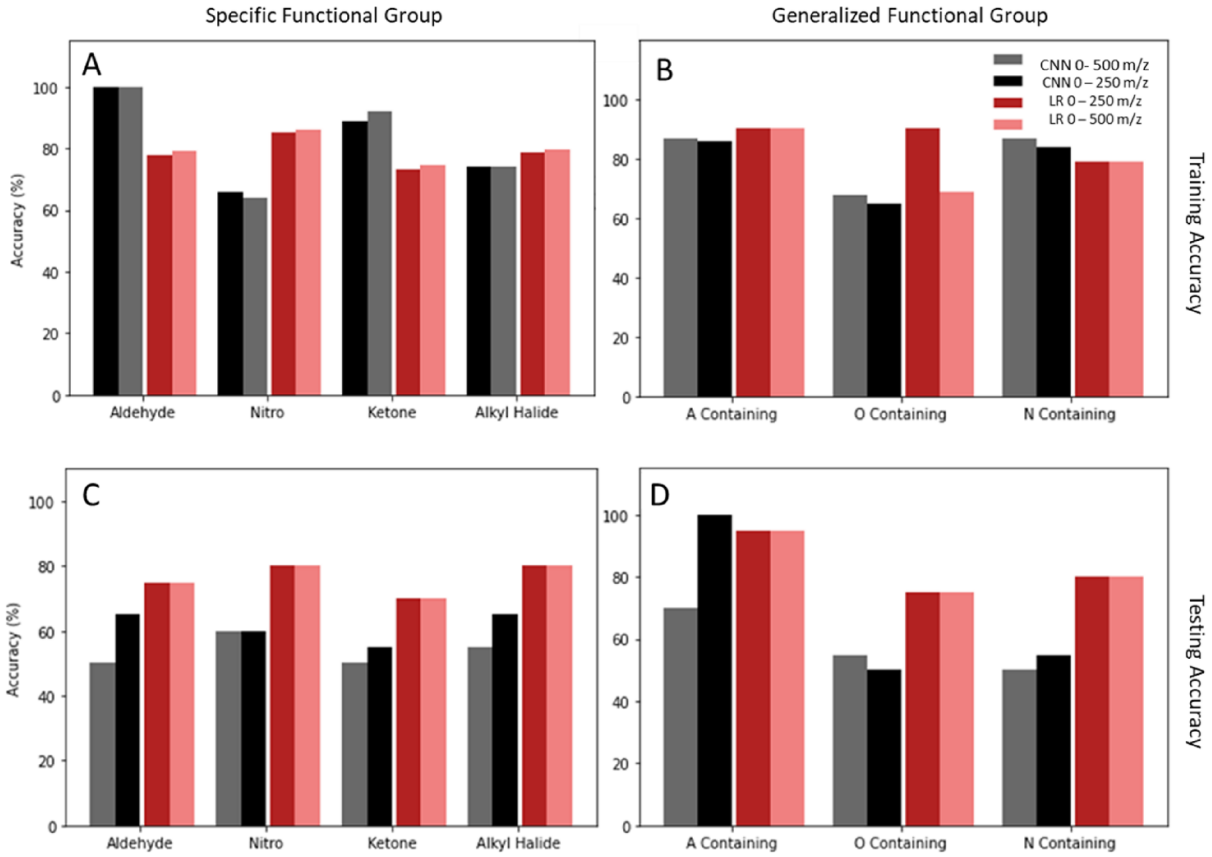


Figure 3. Results of the training and testing for four specific functional groups and the three functional group classifications. (A, B) Final training accuracy and accuracy of identifying the training portion of the data after the final training step has passed for both the functional group-specific and functional group-generalized models. For example, these plots would suggest that the CNN-based approach should be better at correctly identifying the aldehydes and ketones and that the LR-based approach should have an edge on the nitro group and the alkyl aldehydes. This, however, does not tell the full story. (C, D) Final test accuracies for the functional group-specific and functional group-generalized models. The testing accuracy of the models is the accuracy of the models when presented with new previously unseen data shows that a high training accuracy does not correlate necessarily with a high final testing accuracy. The final training and testing scores for each of the functional groups' models are presented in the SI (Tables S3 and S4 for specific functional groups and Tables S5 and S6 for the generalized functional groups).

the distance between the classes. This approach is very similar to support vector machines as both models maximize separation instead of minimizing an error function. This restriction of being a binary classifier coupled with using the entire mass spectrum as features are our reasoning for choosing to generate each model for the purpose of either identifying one specific functional group or one functional group classification. As we will explore later, specializing each of the models allows for the greatest model fit for that functional group as well as providing an avenue in which we can also describe how that greatest fit for each functional group was achieved.

RESULTS AND DISCUSSION

Acquiring the Dataset. The mass spectra were web-scraped from the NIST Webbook using a web-scraping implementation, the details of which are described in our previous publication.[25] In short, a web-scraping script was written to individually download the mass spectral files from each of the NIST Webbook pages that are labeled by Chemical Abstracts Service (CAS) number. We obtained a total of 21,166 mass spectra. The files that were downloaded were in a JCAMP-DX file format. These files were then converted from JCAMP-DX into CSV. The process does remove the associated metadata; however, this information was not necessary for our analyses. Once the files had been converted to CSV, further preprocessing could be completed. More information regarding the preprocessing steps are reported in the SI.

Comparing Convolutional Neural Networks and Logistic Regression Feasibility. Both CNN and LR architectures were used to train functional group-specific models and models to look at the functional group classifications. CNN was initially chosen due to its success in identifying functional groups using an IR dataset collected from the NIST database in our previous publication.[25] Both approaches were each trained on a unique dataset, which was a subset of all the data web-scraped from NIST. Once all the models were trained, it was possible to look at the final training accuracies to determine how well the final models fit the data sets.

There are two metrics that we utilized to describe how well the models were performing. The first of these is final training accuracy. This metric describes the final ability to fit a segmented subset of the testing data after all the training steps have been completed. For our models, we do a 80:20 split of training and internal validation data, which is cited as being the most beneficial split.[38] The training accuracy is a description of how well the data can fit the data that it has been trained with. The second metric of interest is the final testing accuracy. This

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

metric arises from how well the model can classify novel data. This metric is determined by analyzing previously withheld data using the models. Before generating the training datasets, certain spectra are removed from the total dataset and withheld for testing the final model accuracy. This metric is critical in understanding how we can expect our models to perform with data in the future. Figure 2 shows the final training and testing accuracies for four different functional groups' specific functional group models. Here, we compare the training accuracy (Table S3) and testing accuracy (Table S4) of the specific functional group models for mass ranges of 0−250 and 0−500 m/z. Model training and testing accuracies for all 19 functional groups explored are shown in the SI.

Based only on the training accuracies, it appears that the models generated through CNNs should show a greater final accuracy in a specific case than the LR-based models. The training accuracy values, however, do not tell the entire story. This highlights one of the main erroneous assumptions that is commonly made about ML. A model with an incredibly high fit of the training data is not necessarily better at describing novel samples. This metric is better described through the testing accuracy. This trend holds true for both the functional group-specific (Figure 3A,C) and functional group-generalized models (Figure 3B,D).

A question that arose during this analysis was why the transfer learning worked so well with the IR data and so poorly with the MS data.[25] The reasoning for this discrepancy likely falls under the differences in the atomic processes that are described with that technique. For the IR data, because it is vibrational spectroscopy, we see the signals taking broader peaks that are influenced by the bonding environment. This means that phase and having other molecular species in solution can lead to shifting those vibrational peaks. These broad and shifting peaks are both benefited by the transfer learning process. The broad peaks allow them to not be computationally removed when the mathematical convolutions occur. In fact, these convolutions make the model less sensitive to peak shifting on the range of tens of wavenumbers. These aspects make transfer learning promising for vibrational techniques. On the other hand, comparing mass peaks that are only a couple of mass units apart from each other are likely describing entirely different fragmentation patterns or isotopic ratios. MS data also has incredibly narrow peaks that can be missed entirely if they are low in intensity during the mathematical convolutions. These factors are likely why transfer learning using Inception V3 was successful with the IR based data and unsuccessful with the MS data. Upon the determination that the LR-based models performed better on correctly identifying new data compared to the CNN models, we decided to use the LR-based models for the remainder of this study.

Logistic Regression's Ability to Manage Specific Functional Group Classifications. The choice to switch to logistic regression arose from wanting to utilize binary classifiers. By simplifying each model to a binary classifier, it is more feasible to fully explain the model output. Given our dataset, it is easier to optimize one model per functional group than one model predicting on all functional groups. For example, there are thousands of aromatic-containing spectra and less than 400 amide samples. This would impart artificial bias that would have to be mathematically manufactured to avoid. Training one model would likely lead to the functional groups' penalization because of less examples and ultimately not being identified as consistently or frequently.

There is a large variation in the final training and testing accuracies for each of the different functional group models. This is to be expected due to the large variation between the fragment fingerprint for each functional group. A total of 17/20 of the models had a final testing accuracy of over 70%, and 13/20 of our models had a final test accuracy of over 75%. Figure 4 shows the final training and test accuracies of all the models.

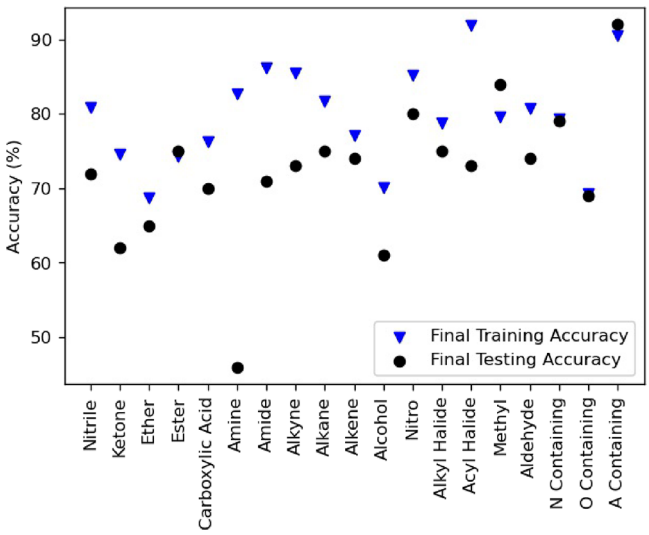


Figure 4. Scatter plot depicting all the final training and testing accuracies of each of the 20 different models. These final accuracies are highly variable with respect to the functional group that they are to be classifying.

The highest preforming models, in terms of final testing accuracy, were the nitro, methyl, and aromatic (A)-containing models. This makes sense because with each of these models, there are fragments that we can point to that would assist the model in its assignments. The nitro model could utilize the NO^+ and NO_2^+ fragments. The methyl model can look for the CH_3^+ ion, and the A-containing model can look for the loss of a benzene ring at 78 m/z.

Conversely, the poorest-performing models are those of ketone, amine, and alcohol. These models likely struggle since the current methods of identifying these functional groups rely on looking at mass losses and looking for the products of secondary processes including rearrangements and cleavages of certain areas of the molecule. These processes include α and β cleavages, McLafferty rearrangements, and radical losses, among others.

Similarly, when looking at the generalized functional groups, the N-containing and A-containing models both performed better than the O-containing model, albeit the O-containing model still had a final testing accuracy of approximately 70%. This likely has to do with the fact that there is clear logic for identifying odd numbers of both nitrogen and aromatics in mass spectra. For the odd nitrogen spectra, we can look for odd-numbered peaks suggesting the presence of nitrogen and we can look for a mass at 78 m/z to look for benzene, a common aromatic ring that shows up in organic molecules.

Identifying Mass Peaks that Guide Model Assignments. Feature selection and feature engineering are common

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

practices in the development of ML models, and there are a large variety of methods to determine which features generate the best model outcomes.[39−43] Feature selection differs from feature engineering; feature engineering works to reduce data dimensionality through convolving or creating statistical representations of the data through processes like principal component analysis or linear discriminant analysis, among others,[44] and feature selection works to reduce the raw data down to the most important features within.[43,45,46] Both processes can be done manually or automatically via a statistical method.[47] Using feature engineering and feature selection processes provides different benefits to the modeling process.

To evaluate and explain the logic behind the model's assignments, we looked at the coefficients that the model used in its final iteration. For each feature, in our case each mass, there is an associated coefficient describing the weight that that mass is used to determine if the functional group is present for that class. Positive peaks correlate to an increased likelihood that that functional group is present, and negative peaks correlate to the increased likelihood that that functional group is absent. The larger the intensity in either direction, the higher the correlation between that mass and the class that it is referring to. Figure 5 shows the overlapped final coefficients for both the generalized functional group models (Figure 5A) and the specific functional group models (Figure 5B).

When looking at all the aggregated coefficients, it looks like the most impactful mass region to the analysis is below 100 m/z. This suggests that the model would perform similarly well if those were the only features given to the training set. This is an important conclusion that suggests that for this kind of analysis, having a large mass range of available data is not necessary as long as the low mass range (less than 100 m/z) is thoroughly defined.

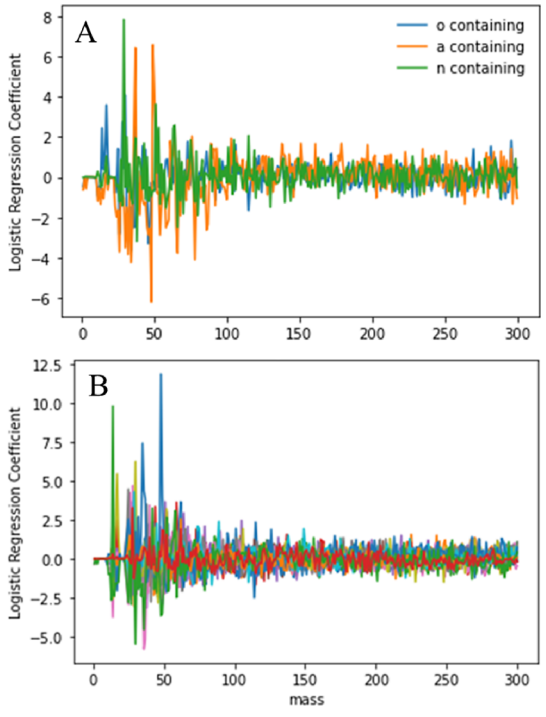


Figure 5. Model coefficients for each of the different trained models as a function of mass fragment. (A) Coefficients for the generalized functional group models and (B) coefficients for the specific functional group models. All the coefficient plots for the individual models are presented in the SI.

To further understand how the different models were doing their assignments, we developed a method to look at the impact of each peak on the final training and testing accuracies. To analyze the features, we trained each model 300 times, and in each iteration of training one mass was removed. These final training and testing accuracies were compared to the accuracy of the model when it had access to all 300 mass units. This was used to identify peaks that were beneficial to the model's ability to identify functional groups and those that were hindering the models in making their assignments. The peaks that were beneficial led to a decrease in model accuracy when removed; the larger the discrepancy, the more impactful the peak. On the other hand, peaks that were causing more false assignments, when removed, led to an increase in model accuracy. Looking at the most beneficial peaks for the generalized functional group classification models leads to some interesting and promising results. Table 1 has these values for the generalized functional group classifications. The beneficial peaks for the specific functional groups as well as the peaks that decreased the final accuracies (the peaks that confuse the inference) are presented in the SI (Table S8 for the beneficial peaks and Table S9 for the hindersome peaks.)

Table 1. Mass Values of the Top 5 Most Impactful Positively Correlated Peaks for Each of the Functional Group-Generalized Models[a]

	masses that reduce model accuracy when removed	
functional group classification	mass value (m/z)	% effect
A-containing	78	−0.5
	42	−0.4
	66	−0.3
	50	−0.2
	68	−0.2
N-containing	29	−2.5
	105	−0.7
	43	−0.6
	38	−0.5
	53	−0.4
O-containing	28	−0.4
	42	−0.4
	30	−0.6
	26	−0.8
	46	−1.0

[a]These were determined by comparing the testing accuracies of the model when it had access to all 300 mass units to when that mass unit of interest was removed. The % effect shown in the rightmost column is negative because when those masses were removed, the model experienced a reduction in the final testing accuracy. The mass values for the nitrogen-containing model are all odd mass values, and the mass values for the oxygen-containing and aromatic-containing spectra are all even, suggesting the utilization of the odd nitrogen rule without explicit training on that detail.

In Table 1, all the most impactful mass peaks for the N-containing model are odd mass values, whereas the most impactful peaks for the O- and A-containing models are even mass values. This suggests that even without explicitly "teaching" the model that there is an odd nitrogen rule, the model was able to come to that conclusion on its own. We can also look at the most impactful peak for the A-containing

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

model and see that it is 78 m/z, which can be attributed to the mass fragment of benzene. However, we can also see that removing 78 m/z only leads to a 0.5% reduction in the final training accuracy of the model. This means that although there may be peaks that are important for assigning functional groups, the model does not use a single peak or even a small set of peaks to make an assessment. The next step in our analysis shifted to the impacts of the number of available features on the final accuracies.

Effects of Mass Range on Model Accuracy. To evaluate whether more data leads to higher accuracies for these models, we adjusted the dataset. We trained the models with 100, 300, and 500 mass units. We decided to reduce this to 100 mass units because the majority of the previously identified impactful mass fragments occurred at less than 100 m/z. We also increased to 500 mass units so that we can encompass more of the high-mass range fragments. Both are compared to our 300 mass units' models for a basis.

For both decreasing the mass range from 300 to 100 m/z and increasing the mass range from 300 to 500 m/z, we see an inconsistent response in the final testing accuracy with respect to the different functional groups (Figure 6). In Figure 4, we observed no consistent trend in the mass range effect on the final test accuracy. These results are consistent with what we observed in our analysis of the model coefficients, suggesting that at mass ranges greater than 100 m/z, the features aren't being as heavily utilized as they are at smaller masses.

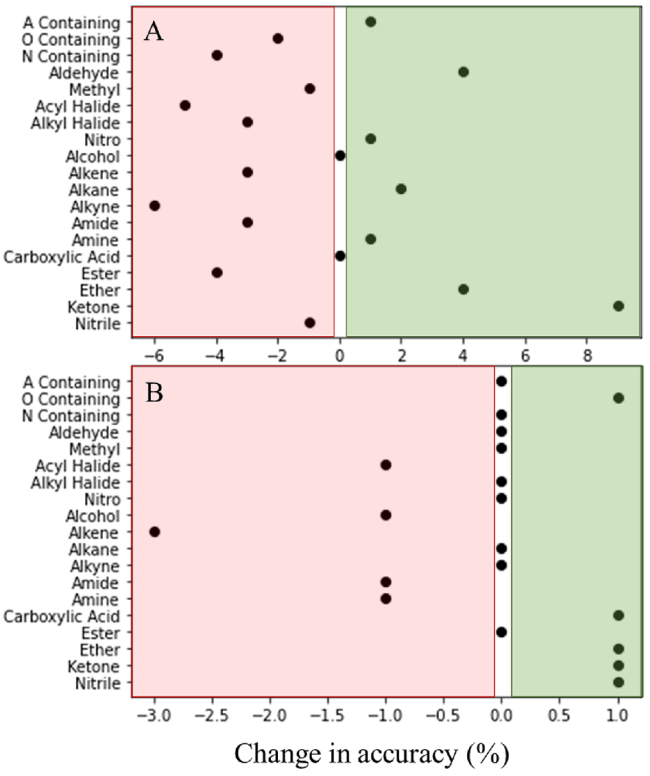
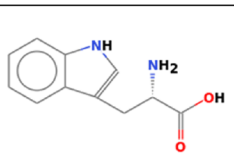


Figure 6. Scatter plots depicting the effect of (A) decreasing the utilized mass range from 300 to 100 mass units and (B) increasing the utilized mass range from 300 to 500 mass units on the final testing accuracy of the models. The presence of points that are positive on the x axis (shaded in green, rightmost box) show a net benefit in accuracy, whereas a negative x value (shaded in red, leftmost box) indicates a worsening accuracy.

Specific Examples of the Applications of this Approach. After exploring some of the parameters that affect the accuracy of these models, we then tested model success on a real-world application. When mass spectrometry data is returned, or downlinked, to Earth from planetary science missions, tens of thousands of mass spectra may have been collected. Yet only a small subset of spectra may be scientifically significant. For example, a common target to identification of life is amino acids. To mimic this process, we examined the NIST mass spectrum of tryptophan to evaluate the models' assignments. We also report the analysis of the mass spectra of histidine in Table S9. Table 2 shows the results of our analysis of tryptophan.

Table 2. Results of Selected Models on the Ability to Correctly Identify the NIST Spectra of Tryptophan[a]



	Tryptophan	
	Presence In Molecule	Model Predicted
Carboxylic Acid	Present	Correct
Amine	Present	Correct
Aromatic	Present	Correct
Alcohol	Absent	Correct
Ketone	Absent	Correct
A Containing	Present	Correct
N Containing	Present	Correct
O Containing	Present	Incorrect

	presence in molecule	model predicted
carboxylic acid	present	correct
amine	present	correct
aromatic	present	correct
alcohol	absent	correct
ketone	absent	correct
A-containing	present	correct
N-containing	present	correct
O-containing	present	incorrect

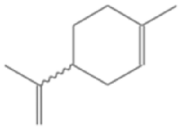
[a]This example works to show how these tandem models may be beneficial in screening large amounts of data to look for specific spectra of interest. Results of all of the models on NIST's spectra of tryptophan and histidine are presented in the SI (Table S9)

Aside from the O-containing model, the LR model correctly predicted the present functional groups and functional group classifications for the tryptophan mass spectrum. This shows promise in these models being a useful tool for screening large numbers of mass spectra. In the example of planetary science missions, this process can be done onboard the spacecraft to help assist in the process of prioritizing spectra to downlink. It could also be used on data after it has been transmitted to prioritize spectral analysis.

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

To further benchmark the success of our models, we also analyzed experimental data external to the NIST dataset. The spectra for limonene, pyridine, and 2 furanmethanol were preprocessed in the same way as the NIST data to ensure that every mass had an associated intensity. These spectra were then presented to each of the models. Table 3 shows the model

Table 3. Results of the Models on the Ability to Correctly Identify Experimental Spectra of Limonene[a]



	Limonene	
	Presence In Molecule	Model Predicted
Alkane	Present	Present
Alkene	Present	Present
Alkyne	Absent	**Present**
Methyl	Present	Present
Alcohol	Absent	**Present**
A Containing	Absent	Absent
N Containing	Absent	Absent
O Containing	Absent	**Present**

	presence in molecule	model predicted
alkane	present	present
alkane	present	present
alkyne	absent	present
methyl	present	present
alcohol	absent	present
A-containing	absent	absent
N-containing	absent	absent
O-containing	absent	present

[a]The experimental spectra were preprocessed in the same way as the NIST data used for training. The results for all the models on limonene, pyridine, and 2 furan methanol are presented in the SI (Table S10).

assignments for limonene. Pyridine and 2-furanmethanol are presented in the SI (Tables S9 and S10). Each of these compounds scored ∼80% accuracy for the identification of their functional groups. When making errors, the models tended to overestimate the number of functional groups rather than underestimate.

CONCLUSIONS

We present an investigation of multiple ML methods and parameters for mass spectral functional group analysis using minimal spectral preprocessing. Our results indicated that the CNN (Inception V3) did not perform as well as the LR models. We determined that the aromatic, nitro, and methyl functional groups are well defined though LR models, whereas the alcohol, ketone, and amine functional groups are more difficult for LR models to define based on their fragmentation patterns alone. The most impactful peaks affecting model accuracy were determined by iteratively training models and removing one mass value in each model, and these results were echoed in looking at the final model feature coefficients. We observed that nitrogen-containing functional group models independently learn the odd-nitrogen rule. We evaluated the effect of mass range to determine whether model accuracy is improved with additional spectral information; these results vary between functional groups. Electron ionization fragmentation of small molecules generally will result in mass values below 100 m/z. Our model coefficients suggest that a mass range of 0−100 m/z is most beneficial for describing functional groups. The application of LR models to new sample mass spectra is evaluated on an example target molecule of interest, tryptophan, as well as experimental data from outside of the NIST database. The success of these example analyses highlights the promise of ML approaches for screening a large volume of mass spectral data.

Future directions should further develop a methodology for developing an ideal ML approach. For example, feature optimization for each model would achieve the highest possible final testing accuracy. Further validation of the models on experimental data outside of the NIST database is also necessary. Exploration of the LR method applied to other fragmentation patterns would enable the implementation of generalizable ML more broadly in the field of mass spectrometry. The LR ML method explored herein provides a benchmark for applications to space exploration, ultimately improving analysis capabilities through the identification of chemically interesting spectra.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c01684.

Links to a GitHub repository in which the Jupyter notebooks for the code have been made available; information regarding web scraping and defining the functional group classifications; descriptions of the image-based CNN models in terms of troubleshooting the plotting; final accuracies for all of the image-based CNN models; final accuracies of all the array-based logistic regression models; model coefficients for each logistic regression model; top 5 most impactful (both positive and negative) mass fragments for each model; results for specific compound validation of the models both using withheld NIST spectra as well as experimental spectra (PDF)

AUTHOR INFORMATION

Corresponding Author

Heather C. Allen − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States;* orcid.org/0000-0003-3120-6784; Email: allen@chemistry.ohio-state.edu

Authors

Nicole M. North − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*

Abigail A. Enders − *Department of Chemistry & Biochemistry, The Ohio State University, Columbus, Ohio 43210, United States*

Morgan L. Cable − *NASA Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, United States;* orcid.org/0000-0002-3680-302X

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c01684

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

For funding this work, N.M.N. acknowledges NASA's Future Investigators of NASA Earth and Space Science Technology (FINESST) grant number 20-PLANET20-0067. A.A.E. acknowledges support by NSF through the Center for Aerosol Impacts on Chemistry of the Environment (CAICE), CHE-1801971. H.C.A. acknowledges support through the DOE-BES CPIMS Grant #DE-SC0016381. A portion of the research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government or the Jet Propulsion Laboratory, California Institute of Technology.

■ REFERENCES

(1) Llop, E.; Pinho, P.; Matos, P.; Pereira, M. J.; Branquinho, C. The Use of Lichen Functional Groups as Indicators of Air Quality in a Mediterranean Urban Environment. *Ecol. Indic.* **2012**, *13*, 215−221.

(2) Infantes, L.; Chisholm, J.; Motherwell, S. Extended Motifs from Water and Chemical Functional Groups in Organic Molecular Crystals. *CrystEngComm* **2003**, *5*, 480−486.

(3) Tsou, C. L. Relation between Modification of Functional Groups of Proteins and Their Biological Activity. I.A Graphical Method for the Determination of the Number and Type of Essential Groups. *Sci. Sin.* **1962**, *11*, 1535−1558.

(4) Coe, J. V.; Chen, Z.; Li, R.; Nystrom, S. V.; Butke, R.; Miller, B.; Hitchcock, C. L.; Allen, H. C.; Povoski, S. P.; Martin, Jr., E. W. Molecular Constituents of Colorectal Cancer Metastatic to the Liver by Imaging Infrared Spectroscopy. In *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XIII*; SPIE, 2015; Vol. *9328*, pp. 98−104. DOI: 10.1117/12.2079884.

(5) Kind, T.; Fiehn, O. Advances in Structure Elucidation of Small Molecules Using Mass Spectrometry. *Bioanal. Rev.* **2010**, *2*, 23−60.

(6) Levsen, K.; Schiebel, H.-M.; Behnke, B.; Dötzer, R.; Dreher, W.; Elend, M.; Thiele, H. Structure Elucidation of Phase II Metabolites by Tandem Mass Spectrometry: An Overview. *J. Chromatogr. A* **2005**, *1067*, 55−72.

(7) Winston, R. L.; Fitzgerald, M. C. Mass Spectrometry as a Readout of Protein Structure and Function. *Mass Spectrom. Rev.* **1997**, *16*, 165−179.

(8) Prabhudesai, V. S.; Kelkar, A. H.; Nandi, D.; Krishnakumar, E. Functional Group Dependent Site Specific Fragmentation of Molecules by Low Energy Electrons. *Phys. Rev. Lett.* **2005**, *95*, No. 143202.

(9) Ghojogh, B.; Ghodsi, A.; Karray, F.; Crowley, M. Generative Adversarial Networks and Adversarial Autoencoders: Tutorial and Survey. arXiv November 25, 2021 Cornell University. http://arxiv.org/abs/2111.13282 (accessed 2022-09-20).

(10) Weiss, S. M.; Kapouleas, I. An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods. In *Proceedings of the 11th international joint conference on Artificial intelligence - Volume 1*; IJCAI'89; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1989; pp. 781−787.

(11) Zhou, C.; Bowler, L. D.; Feng, J. A Machine Learning Approach to Explore the Spectra Intensity Pattern of Peptides Using Tandem Mass Spectrometry Data. *BMC Bioinf.* **2008**, *9*, 325.

(12) Ulintz, P. J.; Zhu, J.; Qin, Z. S.; Andrews, P. C. Improved Classification of Mass Spectrometry Database Search Results Using Newer Machine Learning Approaches. *Mol. Cell. Proteomics* **2006**, *5*, 497−509.

(13) Rivera, S. M.; Christou, P.; Canela-Garayoa, R. Identification of Carotenoids Using Mass Spectrometry. *Mass Spectrom. Rev.* **2014**, *33*, 353−372.

(14) Le Lacheur, R. M.; Sonnenberg, L. B.; Singer, P. C.; Christman, R. F.; Charles, M. J. Identification of Carbonyl Compounds in Environmental Samples. *Environ. Sci. Technol.* **1993**, *27*, 2745−2753.

(15) Shimbo, K.; Kubo, S.; Harada, Y.; Oonuki, T.; Yokokura, T.; Yoshida, H.; Amao, M.; Nakamura, M.; Kageyama, N.; Yamazaki, J.; Ozawa, S.; Hirayama, K.; Ando, T.; Miura, J.; Miyano, H. Automated Precolumn Derivatization System for Analyzing Physiological Amino Acids by Liquid Chromatography/Mass Spectrometry. *Biomed. Chromatogr.* **2010**, *24*, 683−691.

(16) Bidlingmeyer, B. A.; Cohen, S. A.; Tarvin, T. L. Rapid Analysis of Amino Acids Using Pre-Column Derivatization. *J. Chromatogr. B: Biomed. Sci. Appl.* **1984**, *336*, 93−104.

(17) Dron, J.; Abidi, E.; Haddad, I. E.; Marchand, N.; Wortham, H. Precursor Ion Scanning−Mass Spectrometry for the Determination of Nitro Functional Groups in Atmospheric Particulate Organic Matter. *Anal. Chim. Acta* **2008**, *618*, 184−195.

(18) Dührkop, K.; Nothias, L.-F.; Fleischauer, M.; Reher, R.; Ludwig, M.; Hoffmann, M. A.; Petras, D.; Gerwick, W. H.; Rousu, J.; Dorrestein, P. C.; Böcker, S. Systematic Classification of Unknown Metabolites Using High-Resolution Fragmentation Mass Spectra. *Nat. Biotechnol.* **2021**, *39*, 462−471.

(19) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nat. Methods* **2022**, *19*, 865−870.

(20) Dührkop, K.; Shen, H.; Meusel, M.; Rousu, J.; Böcker, S. Searching Molecular Structure Databases with Tandem Mass Spectra Using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12580−12585.

(21) Heinonen, M.; Shen, H.; Zamboni, N.; Rousu, J. Metabolite Identification and Molecular Fingerprint Prediction through Machine Learning. *Bioinformatics* **2012**, *28*, 2333−2341.

(22) Asef, C. K.; Rainey, M. A.; Garcia, B. M.; Gouveia, G. J.; Shaver, A. O.; Leach, F. E. I.; Morse, A. M.; Edison, A. S.; McIntyre, L. M.; Fernández, F. M. Unknown Metabolite Identification Using Machine Learning Collision Cross-Section Prediction and Tandem Mass Spectrometry. *Anal. Chem.* **2023**, *95*, 1047−1056.

(23) Feucherolles, M.; Nennig, M.; Becker, S. L.; Martiny, D.; Losch, S.; Penny, C.; Cauchie, H.-M.; Ragimbeau, C. Combination of MALDI-TOF Mass Spectrometry and Machine Learning for Rapid Antimicrobial Resistance Screening: The Case of Campylobacter Spp. *Front. Microbiol.* **2022**, *12*, No. 804484.

(24) Hao, Y.; Lynch, K.; Fan, P.; Jurtschenko, C.; Cid, M.; Zhao, Z.; Yang, H. S. Development of a Machine Learning Algorithm for Drug Screening Analysis on High-Resolution UPLC-MSE/QTOF Mass Spectrometry. *J. Appl. Lab. Med.* **2023**, *8*, 53−66.

(25) Enders, A. A.; North, N. M.; Fensore, C. M.; Velez-Alvarez, J.; Allen, H. C. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Anal. Chem.* **2021**, *93*, 9711−9718.

(26) 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.

(27) Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z *Rethinking the Inception Architecture for Computer Vision*; IEEE 2016; pp. 2818−2826.

(28) Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*; IEEE 2009; pp. 248−255. DOI: 10.1109/CVPR.2009.5206848.

(29) Recht, B.; Rolfs, R.; Schmidt, L.; Shankar, V. Do ImageNet Classifiers Generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*; PMLR, 2019; pp. 5389−5400.

(30) You, Y.; Zhang, Z.; Hsieh, C.-J.; Demmel, J.; Keutzer, K. ImageNet Training in Minutes. In *Proceedings of the 47th International Conference on Parallel Processing*; ICPP '18; Association for

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350

Computing Machinery: New York, NY, USA, 2018; pp. 1−10. DOI: 10.1145/3225058.3225069.
(31) Ahmed, E.; Jones, M.; Marks, T. K. *An Improved Deep Learning Architecture for Person Re-Identification*; IEEE 2015; pp. 3908−3916.
(32) Bantupalli, K.; Xie, Y. American Sign Language Recognition Using Deep Learning and Computer Vision. In *2018 IEEE International Conference on Big Data (Big Data)*; IEEE 2018; pp. 4896−4899. DOI: 10.1109/BigData.2018.8622141.
(33) Albatayneh, O.; Forslöf, L.; Ksaibati, K. Image Retraining Using TensorFlow Implementation of the Pretrained Inception-v3 Model for Evaluating Gravel Road Dust. *J. Infrastruct. Syst.* **2020**, *26*, No. 04020014.
(34) Zhang, X.; Lin, T.; Xu, J.; Luo, X.; Ying, Y. DeepSpectra: An End-to-End Deep Learning Approach for Quantitative Spectral Analysis. *Anal. Chim. Acta* **2019**, *1058*, 48−57.
(35) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep Learning Enables de Novo Peptide Sequencing from Data-Independent-Acquisition Mass Spectrometry. *Nat. Methods* **2019**, *16*, 63−66.
(36) Kirasich, K.; Smith, T.; Sadler, B. Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Sci. Rev.* **2018**, *1*, 9.
(37) Feng, J.; Xu, H.; Mannor, S.; Yan, S. Robust Logistic Regression and Classification. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2014; Vol. *27*.
(38) Rácz, A.; Bajusz, D.; Héberger, K. Effect of Dataset Size and Train/Test Split Ratios in QSAR/QSPR Multiclass Classification. *Molecules* **2021**, *26*, 1111.
(39) Khaire, U. M.; Dhanalakshmi, R. Stability of Feature Selection Algorithm: A Review. *J. King Saud Univ., Sci.* **2022**, *34*, 1060−1073.
(40) Kumar, V. Feature Selection: A Literature Review. *Smart Comput. Rev.* **2014**, *4*, 211−229.
(41) Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Comput. Surv.* **2017**, *50*, 1−45.
(42) Venkatesh, B.; Anuradha, J. A Review of Feature Selection and Its Methods. *Cybern. Inf. Technol.* **2019**, *19*, 3−26.
(43) Lu, Y.; Cohen, I.; Zhou, X. S.; Tian, Q. Feature Selection Using Principal Feature Analysis. In *Proceedings of the 15th ACM international conference on Multimedia*; MM '07; Association for Computing Machinery: New York, NY, USA, 2007; pp. 301−304. DOI: 10.1145/1291233.1291297.
(44) El-Amir, H.; Hamdy, M. *Deep Learning Pipeline: Building a Deep Learning Model with TensorFlow*; Apress, 2020.
(45) Sultana, N.; Chilamkurti, N.; Peng, W.; Alhadad, R. Survey on SDN Based Network Intrusion Detection System Using Machine Learning Approaches. *Peer Peer Netw. Appl.* **2019**, *12*, 493−501.
(46) Drgoňa, J.; Picard, D.; Kvasnica, M.; Helsen, L. Approximate Model Predictive Building Control via Machine Learning. *Appl. Energy* **2018**, *218*, 199−216.
(47) RM, S. P.; Maddikunta, P. K. R.; Parimala, M.; Koppu, S.; Gadekallu, T. R.; Chowdhary, C. L.; Alazab, M. An Effective Feature Engineering for DNN Using Hybrid PCA-GWO for Intrusion Detection in IoMT Architecture. *Comput. Commun.* **2020**, *160*, 139−149.

https://doi.org/10.1021/acsomega.3c01684
ACS Omega 2023, 8, 24341−24350