Accurate and automated *de novo* identification of molecular functional groups using deep learning architectures

Jonathan Fine[1, ‡], Anand AR[2, ‡], Gaurav Chopra[1,3,4,5,6,7 *]

[1]Department of Chemistry, Purdue University, 720 Clinic Drive, West Lafayette, IN 47906

[2]Indian Institute of Technology Madras

[3]Purdue Institute for Drug Discovery

[4]Purdue Center for Cancer Research

[5]Purdue Institute for Inflammation, Immunology and Infectious Disease

[6]Purdue Institute for Integrative Neuroscience

[7]Integrative Data Science Initiative

[‡]These authors share an equal contribution to this work.

*Corresponding Author

E-mail: gchopra@purdue.edu

Abstract

Functional groups (FGs) link analytical, physical, organic, and materials chemistry and are therefore central to the chemical sciences. In both analytical and organic chemistry, FGs are used to predict reactivity and other properties of molecules. The start-of-the-art approach to accurately identify all functional groups for unknown chemical entities involves manual or database dependent analysis of a Fourier Transform Infra-Red (FTIR) or Mass Spectroscopy (MS) spectrum using previously established rules and experience to match patterns by a skilled spectroscopist. For complex spectra, this process is time-consuming and error-prone for functional groups of chemical entities that are not characterized extensively in the literature. Herein, we present a fast machine learning based approach for identifying all the functional groups of an unknown compound using a combination of FTIR and MS spectra without the use of any database, pre-established rules, procedures, or peak-matching methods. We use Artificial Neural Networks to derive patterns and correlations directly from spectral data that benefit from reinforcement of multiple patterns, representing multiple functional groups. Our methodology is different from previous attempts to classify spectra as we treat the classification as a multi-class, multi-label problem compared to multiple binary classifiers. We train our method using the 7393 publicly available spectral dataset (FTIR and MS) from the NIST Webbook resulting in average cross-validated F1 score higher than 0.82 for 14 out of 17 defined functional groups. To achieve practical utility of our method, we introduce two new metrics (Molecular F1 score and Molecular Perfection rate) to measure the performance of identifying all functional groups on molecules in addition to the identification of functional group types. Our optimized model has a Molecular F1 score of 0.92 and a Molecular Perfection rate of 72%. Additionally, backpropagation of our model reveals IR patterns typically used by human chemists to identify standard groups suggesting "learning" of known spectral features. We further show that the introduction of new functional groups does not decrease the performance of our model. Finally, we show redundancy in FTIR and MS data by encoding all our features in a latent space that retains the accuracy of the original model. These results reveal the importance of using machine learning for the rapid identification of functional groups for small molecules as well as the importance of a large body of open data for training methods to achieve autonomous analytical processes in the future.

Introduction

The arrangement of atoms in a molecule gives rise to its emergent physical, chemical, and spectral properties. Small discrete, or large repeating arrangements of atoms that give rise to measurable changes in a molecule's reactivity[1–3], boiling point[4,5], melting point[6,7], and other characteristics are called functional groups. Using the structural formula of a molecule, a chemist can identify the functional groups present in the molecule and can postulate characteristics of the molecule as a "summation" of its functional groups. Identification of functional groups are essential to validate the formation of small molecules during synthetic reactions and to elucidate the structure of isolated natural products. Several techniques have been developed to identify functional groups and these techniques are applied in organic chemistry[8], metabolomics[9,10], and forensic sciences[11–13] based on matching profiles from known databases. Furthermore, continuous monitoring of functional group changes can be used to determine the progress of a reaction,[14] thereby identifying the components of complex mixtures for a reaction coordinate.

Organic chemists traditionally identify functional groups in a molecule using Fourier Transform Infrared spectroscopy (FTIR). This analytical method utilizes the frequencies associated with the bonds in a molecule, which typically vibrate around 4000 cm^{-1} to 400 cm^{-1}, known as the Infra-Red region of the electromagnetic spectrum[8]. FTIR uses Infra-Red light such that the bonds with specific frequencies change their oscillating patterns in the analyte. These frequencies are absorbed and do not transmit through the molecule, generating a unique pattern of frequencies that are recorded by the FTIR instrument and interpreted to obtain an FTIR spectrum[15]. Typically a spectroscopist manually analyzes this spectrum to identify patterns corresponding to a given functional group using previously established rules and principals[8], a time-consuming process

subject to human bias and interpretation. Alternatively, if the compound has previously been characterized, the spectroscopist can use software that match the peaks of the analyte to a database of known compounds for identification[16].

Mass spectroscopy (MS) is another technique commonly used by chemists for the identification of unknown compounds[8]. One of the first, and still a popular MS ionization technique is electron ionization (EI-MS)[17], a method performed by bombarding the analyte in the gas phase with high energy electrons (~70 eV) for molecular ionization. The resulting cationic radicals are energetically unstable and break apart, resulting in smaller charged particle fragments that are specific to the analyte. Such fragmentation patterns are dependent on molecular functional groups and their arrangements with other functional groups and motifs. The abundance of fragments with a given mass to charge ratio (m/z) is recorded and reported as the mass spectrum. Several MS-based methods such as Tandem MS/MS and multiple reaction monitoring (MRM) have also been used to identify mixtures of molecules by searching a database of MS peaks of known compounds, but large-scale automated identification of unknown molecules is still a major challenge[9,18–20].

Human intervention to analyze FTIR or MS spectrum is useful but to achieve autonomous instrumentation this should be done for problematic cases during the analysis of thousands of spectra for a given task, e.g., during high-throughput synthetic reaction screening. The current approaches to automate functional group identification are similar to those applied by humans, using a set of rules and pattern (peaks) matching to map spectra to a functional group[19,21]. Such methods do not use the entire spectrum of a molecule to identify functional groups, which results in the identification of a small percentage of confident predictions limited by their database of

known compounds[18]. Furthermore, to our knowledge, these methods can only incorporate data from a single spectral technique (i.e., either FTIR or MS) and ignore relationships between different spectral data for identification. Hence, there is a need for automated and accurate methods capable of multiple-spectra integration without the use of pre-established patterns on known databases. Such methods will need minimal-to-no human intervention towards the realization of automated synthetic robots that screen functional groups and combine spectral data to validate each step during reaction screening and multi-step automated synthesis[22]. The state-of-the-art robot for automated reaction detection currently employs different techniques to determine the presence of a reaction[23], but it is limited to identify the presence of pre-defined known compounds. It is a major challenge to develop fully-automated robots to discover new reactions that produce unexpected products. Our goal is to extend the capabilities of these automated synthetic robots by developing a fast, automated methodology for functional group determination that can be used in real-time during reaction screening to identify changes in functional groups in a database-free manner to discover new reactions.

Machine Learning (ML) is a set of techniques used by computers to perform a specific task without an explicit set of instructions provided by the user. ML techniques have been successfully applied to multiple chemical problems in recent years that have the potential to advance several areas of chemistry. Popular machine learning architectures, such as Random Forest[24–26], Multiple Layer Perception[27–29], Generalized Adversarial Networks[30–34], and Recurrent Neural Networks[35–37] have been used on chemical data for small-molecule design[38,39], metabolism[40,41], toxicology[40,42], photo-electric properties, solubility, and retrosynthesis[35,37]. While several applications of these methods take advantage of fingerprinting techniques to

represent the structure of a molecule, direct representation of a molecule as a subgraph of groups of atoms (i.e., functional groups) has distinct advantages over fingerprinting methods[43].

The representation of a molecule or dataset can be reduced to a lower dimensional latent space by using an autoencoder[39]. Here, we also used an encoder to create a corresponding latent space based on spectra to predict functional groups which may also be useful to design molecules for specific spectral properties. A few ML techniques to analyze spectra has been used previously[44–48] but such attempts for function group prediction used only one type of spectral data, the training data was specific to the application, and classified groups separately as a multiple binary classification problem[47,48]. Binary classifiers are not optimal for a large number of classes and sensitive to class imbalances during training resulting in problems to identify all functional groups in a molecule or mixtures[41,49]. In this work, we present the first ML method, to our knowledge, that integrates FTIR and MS data to obtain a combined set of features as a multi-class, multi-label classification methodology. Our method predicts multiple functional groups for a given molecule in a database-free manner, compared to identifying a molecule through peak matching or only identifying the major functional group in the molecule.

Methods

Figure 1 provides an overview of the final methodology used in this paper. First, we standardize all the IR and MS spectra and train a linear autoencoder to reduce the dimensionality of IR and MS spectra to a smaller latent space. Then, we train corresponding MLP models using this latent space to predict the functional groups of given molecules using these spectra using a multi-label classification network.

Standardization of IR spectra

All IR spectra obtained from NIST was truncated so that only peaks occurring from 400cm^{-1} to 4000cm^{-1} remain. The IR spectra available in the NIST webbook has varying degrees of resolution, which is problematic for a multilayer perceptron (MLP) network as this architecture requires a discrete and consistent number of input dimensions. To address this, we standardized all IR spectra so that each spectrum would have the same number of peaks by defining an IR dimension as being the percent transmittance in a wavenumber bin. For example, if a compound has a transmittance of 30% between 400cm^{-1} and 401cm^{-1}, then this dimension has a value of 0.30. Since the most common resolution present in the NIST data is approximately 3.25cm^{-1}, we decided to use this resolution to standardize all IR spectra in our dataset. However, other ranges may be better suited for the problem at hand, and the selection of an optimized resolution remains as to be done as future work that may be considered as a hyperparameter for the model. To standardize all the spectra, we performed linear interpolation on all the IR spectra and evaluated the fitted function at the same set of discrete points throughout each interpolated IR spectra. This process yields uniform IR spectra consisting of 1108 points, regardless of the resolution of the original IR data.

Standardization of MS data

Given the discrete nature of mass spectra, the standardization process for this type of spectra is straightforward. The bin size for these spectra was chosen to be 1 m/z unit, and the counts present in each bin were averaged together for spectra with a resolution less than 1 m/z. All the

bin counts in each spectrum were divided by the largest count in the same spectrum to yield the relative abundance for all the m/z peaks present in the dataset.

Assignment of functional groups

We obtained IUPAC InChI strings for all compounds of interest by resolving the CAS number associated with the molecule using the PubChem API[50]. Then, RDKit[51] performed substructure matching on each string via SMARTS[52,53] strings to test for the presence of a predefined molecular topology. If a match for a functional group's SMARTS was found, then the compound was classified as a member of the given functional group, and each SMARTS string was tested independently. Therefore multiple functional groups could be assigned to a single molecule. Initially, we picked functional groups that are commonly used as discussed in the previous works[44,47,48]. After training our initial model and analyzing the results (see **Guided back propagation of the 'learnt' model shows known IR and chemical patterns**), we decided to add more functional groups to our model in an attempt to improve model performance. The SMARTS strings used in this work are shown in **Table 1**.

Calculation of a Molecular F1 metric

Since the correct assignment of all functional groups in a single molecule is paramount to the analysis of organic reactions and realize minimal intervention during autonomous instrumentation, we introduce a single metric to quantify the predictive capability of our models versus the performance on individual functional groups. Therefore, the focus of our optimization methodology is to create a model that maximizes this overall accuracy measure as opposed to the accuracies of individual functional groups. Similar to the concept of an F1 measure, this

metric normalizes the performance when the classes (functional groups) are unbalanced. Hence we have termed this metric as the 'Molecular F1 score' as it describes the success of the model on the whole molecule. This number is calculated for **each molecule** in the validation set by calculating a 'Molecular Precision' and 'Molecular Recall' value for the functional groups predicted for a given molecule. Precision is the number of functional groups predicted correctly (true positives) divided by the total number of functional groups predicted to be present (the sum true positives and false positives). Molecular recall is the number of functional groups predicted correctly divided by the total number of actual functional groups present in the molecule (the sum of true positives and false negatives). Similar to the calculation of an F1 score for given functional groups, the Molecular F1 is the harmonic mean of the Molecular Precision and Molecular Recall. The overall Molecular F1 score for a given validation set is the arithmetic mean of all Molecular F1 scores. The difference between the Molecular F1 and Functional Group F1 is presented in **Figure 2**.

Calculation of a Molecular Perfection Rate metric

While the knowledge of overall Molecular F1 score is useful for comparing models to one another, it does not represent the more stringent criterion of whether a given method correctly predicts all functional groups of a given molecule without error. Therefore, we introduce a second metric termed 'Molecular Perfection Rate' to rigorously measure the accuracy of our model on a per molecule basis. To calculate this metric, we compare the known functional groups to the predicted functional groups. If the predicted functional groups perfectly match the defined functional groups of the target molecule, then the molecule prediction pair is assigned a

Molecular Perfection of 1; otherwise, it is assigned a Molecular Perfection of 0. The 'Molecular Perfection Rate' for each validation set is calculated as the sum of all individual 'Perfections' values divided by the total number of molecules. This metric can also represent the percentage of all molecules with a Molecular F1 score of 1.0.

Training and testing of neural networks

All Neural Networks reported in this work were created using the Keras Python Package[54]. All hidden layers were normalized using batch normalization and activated using rectified linear units, and a sigmoidal function is used to activate the final output layer. We used binary cross entropy as the loss function for training the neural network as this loss function is standard for multi-label classification problems. For each epoch of training, Keras calculated the loss of the training and validation sets and compared this loss to the loss of the previous epoch. Early stopping with a patience of 5 epochs was used to prevent the model from overtraining. All models were validated using 5-fold cross-validation, and sequential hyperparameter searching was used to optimize the final IR and MS model. The hyperparameters of the optimized model are given in the supporting information. The overall mathematical representation of this model can be represented with **Formula 1** given below.

$$\vec{y} = f\left(\overrightarrow{a^0}, \boldsymbol{W}, \vec{\boldsymbol{b}}\right) \qquad (1)$$

Here, $\vec{y}$ is the predicted functional groups from the model with a length equal to the number of functional groups defined in the previous section. Each component of this vector represents the probability that the corresponding functional group is present in the molecule. The vector $\overrightarrow{a^0}$ is

the input spectra and has a length equal to the number of components in the spectra. Matrix $\boldsymbol{W}$ is a weighting matrix and $\vec{\boldsymbol{b}}$ is bias a vector. All terms are applied in the following manner:

$$f\left(\overrightarrow{a^0}, \boldsymbol{W}, \vec{\boldsymbol{b}}\right) = \sigma\left(\boldsymbol{W}\overrightarrow{a^0} + \vec{\boldsymbol{b}}\right) \qquad (2)$$

This function can be applied. multiple times with matrices of varying length, producing hidden 'layers.' The optimal number of layers, as well as their respective lengths, are 'hyperparameters.' For each neuron *k* in a layer *l*:

$$a_j^l = \sigma\left(\sum_k \boldsymbol{w}_{jk}^l a_k^{l-1} + b_j^l\right) = \sigma\left(z_j^l\right) \quad (3)$$

Here, σ is an activation function. For hidden layers:

$$\sigma = ReLU(x) = max(0, x) \quad (4)$$

For the final layer:

$$\sigma = sigmoid(x) \qquad (5)$$

The cost function, or error in the model, is defined as follows using the binary cross entropy function:

$$C(\vec{y}, \tilde{y}) = \tilde{y} \log \vec{y} + (1 - \tilde{y})\tilde{y} \log(1 - \vec{y}) \qquad (6)$$

Where $\vec{y}$ is the predicted functional groups and $\tilde{y}$ are the true functional groups. The goal of back propagation is to minimize the cost function, thereby making the model increasingly accurate.

We define the error of any neuron to be

$$\delta_j^l = \frac{\delta C}{\delta z_j^l} \qquad (7)$$

For the final layer of the model, we can compute this value via the chain rule:

$$\delta_j^L = \frac{\delta C}{\delta a_j^L}\sigma'(z_j^L) \qquad (8)$$

This expression can be rewritten as the following for a matrix operation:

$$\delta^L = \nabla_a C \odot \sigma'(z^L) \quad (9)$$

Once the derivative of the final layer is obtained, the derivative of the penultimate layer can be calculated as follows:

$$\delta^l = \left(\left(\boldsymbol{W^{l+1}}\right)^{\boldsymbol{T}}\delta^{l+1}\right) \odot \sigma'(z^l) \qquad (10)$$

One can continue to 'backpropagate' this derivative until the derivative of the first layer is calculated. Now that δ_j^l can be calculated for all layers, the derivate of the total cost function with respect to a bias term can be written as

$$\frac{\delta C}{\delta b_j^l} = \delta_j^l \qquad (11)$$

Additionally, the following for the weighting terms:

$$\frac{\delta C}{\delta \boldsymbol{W}_{jk}^l} = a_k^{l-1}\delta_j^l \qquad (12)$$

Now that all backpropagation terms can be calculated for all layers, we can define a modified version of this procedure referred to as 'guided backpropagation.' This procedure begins with an already trained network consisting of $\boldsymbol{W}$ and $\vec{\boldsymbol{b}}$. The network is predicted in the forwards direction to obtain $\vec{y}$. Then equations (**9**) and (**10**) are used to calculate the weights of the input

vector $\overrightarrow{a^0}$. Note that all negative gradients are set to 0 during the application of (**10**). Details for the application of guided backpropogation in this work are given in the following section.

Application of guided backpropagation to identify patterns in the model

One of the most challenging problems of developing deep learning models is interpreting them in a chemical context[55], which results in many researchers treating these models as black box representations, thereby neglecting to understand how features are used to predict results. Solving this problem to understand the relative importance of various input features for a given result is a challenging problem. Many new methods are proposed to address this problem which can be broadly classified into perturbation-based methods and backpropagation-based methods. Back propagation-based methods compute the gradient of the activations concerning the feature space and identify the section of the input data that maximally activates that neuron. One such backpropagation based method, 'guided backpropagation,[56]' is similar to the well-known 'deconvnet' approach[57], but these two methods differ in the handling of backpropagation. The deconvnet approach computes the gradient based on top gradient signal and setting negative values to zero. Backpropagation algorithm sets the gradient of negative activations in the forward pass to zero, avoiding the 'flow' of negative gradients during backpropagation. We use this technique to identify the top 5 bins of an input IR spectral data responsible for predicting a particular functional group of a molecule and present these results in **Figure 9.**

Results and discussion

We obtained standard reference data published by the United States National Institute for Science and Technology[58] (see http:/webbook.nist.gov/). This database contains more than 16,000 Infrared spectra and over 33,000 mass spectra providing an open resource for training machine learning models. We filter and standardize these spectra (see **Methods** section) to obtain both IR and MS for 7,393 unique compounds. SMARTS definitions of 13 common functional groups given in **Table 1** were used to assign functional groups for all 7,393 unique compounds. Our classification results in a broad distribution of the number functional groups that are present in our selected dataset; a type of functional groups range from 50 to several thousands in numbers (**Figure S1**).

Multi layer perceptron neural networks outperforms Random Forest classifiers

We performed a first computational experiment to determine a choice of machine learning method (ensemble vs neural networks) with the best performance to identify functional groups without doing extensive model optimization. We selected Random Forest (RF) and Multi-layered Perceptron (MLP) to test on IR spectra to determine if there was a need for using neural networks (MLP) compared to ensemble methods (RF). An unoptimized MLP consistently outperfomed RF models based on the type of functional group (**Figure S2**) with an average F1-score of 0.771 for MLP compared to 0.650 for RF (see **Table S1, S2** for F1-score of each type). We trained the MLP to predict all functional groups simulaneously as one multilabel classifier. In order to evaluate the effect of transfer learning that has been previously done for MLP[40,41,49], we also evaluated 13 single task networks in addition to the 1 multitask network. The single task approach did not

improve the performanceof the MLP model significantly as the single task models only produced an improvement in functional group F-1 score of 0.006 over the aforementioned multitask model, suggesting that transfer learning is not a significant factor in the multitask network.

MS data addition improves prediction of functional group types

Our MLP model trained on IR data performs well on alkanes, ketones, arenes, carboxylic acids, and esters (average F1-score of 0.926) but it it did not perform at par to predict nitriles, amines, amides, and acyl halides with an average F1-score of 0.663 (**Figure S3a, Table S3**). We included the chemical features captured by mass spectrometry (MS) to augment the MLP-IR model (**Figure S3a**) to address these problematic functional groups. First, we trained a MLP model only on MS data to investigate its predictive capacity for functional groups (**Figure S3b, Table S4**). The difference between the F1 scores of the training set compared to the test set indicates that MS data needs other models to generalize for consistent performance compared to IR data using MLP (**Tables S3, S4**). Similar to the MLP-IR model, the MLP-MS model performed well with more data for a given functional group (e.g. alkanes, arenes, alkyl halides), and poorly when less data was available (e.g. acyl halides, amides, and amines). Next, we wanted to investigate if combining IR and MS data could improve *de novo* prediction of functional groups. We developed a combined IR+MS model by concatenating our combined spectral data features (see **Methods** section). **Table S5** shows the traing and validation F1-scores and **Table S6** show the 5-fold result of the MLP model. The improvement of the IR+MS model over the IR model is presented as **Figure S36c,** and the direct F-1 scores are shown in **Figure S3d** with an average improvement of 0.024 over all functional groups. However, combining IR and MS data results in a substantial increase

in F1 scores for the nitrile, alkene, and alkyl halide functional groups with improvements of 0.124, 0.048, and 0.061 respectively. The amide functional group remains unchanges as the F-1 score of 0.563 is the as the MLP-IR model. The improvement of alkyl halides (**Figure S36c**) may appear to match chemical intuition given the distinct pattern of halogen isotopes observed with MS. However, this conclusion is not supported by the architecture of an MLP model as each input neuron is independent. Future work incorporating the differences in abundance peaks instead of raw values may improve the performance of the MS only model.

Multiple functional groups prediction in a single compound present a second optimization problem

Analysis of the receiver operator characteristic (ROC) plots (**Figure S4**) show that at 1% of the false positive rate, the model identifies over 80% of the true positive functional groups. Therefore, we used a dynamic threshold for each functional group to determine the presence of a functional group in the molecule. This threshold is calculated to maximize the functional group F1 score for the training set after training is complete. While the ability of the model to predict the presence of a particular functional group is important for evaluating the performance of the model, a metric better suited for the study of chemistry and essential for autonomous instrumentation will be to measure the performance to prediction all functional groups in a given molecule. Therefore, we have introduced new metrics, such as, the 'Molecular F1 score (MF1)' and the 'Molecular Perfection Rate (MPR)' (see **Methods** section for more details) and optimized our models for the IR and IR+MS data. After optimization, the IR+MS model was able to perform on par or better than the optimized combined IR for the majority of functional groups (**Figure 3**). The resulting models have comparable average MPRs (72.5% vs 74.9%) and MF1s (0.923 vs 0.931)

for IR and IR+MS respectively (see **Tables S7, S8**). The hyper parameters for these models are given in **Details of the neural networks** section in **Supporting Information**.

Guided back propagation of the 'learnt' model shows known IR and chemical patterns

We performed guided backpropagation on the optimized MLP-IR model for molecules that were both predicted with a MPR of 1, and has the greatest activation in the neuron corresponding to the respective functional group (**Figure 4**). Several backpropagation plots reveal a known chemical association between peaks in IR spectroscopy and functional group assignment. This is encouraging as the model was trained without any 'expert' or chemical information about the location of the peaks corresponding to each functional group. Specifically, we discuss several functional group cases for our selected set of molecules. The alkane functional group backpropagation shows the use of peaks near 3000 cm^{-1}. This matches in the known location of alkane CH peaks tabulated in the literature. The remaining peaks, however, do not provide any additional chemical intuition with regards to the alkane functional group. Aromatic compounds are identified by a peak between 1400-1600 cm^{-1}, and the model selected peaks within this region. The model was able to identify the alkene bending motion around 900 cm^{-1} as well. A C-O stretch is typically observed around 1150 $cm^{-1,}$ and the backpropagation plots for carboxylic acids, alcohols, and esters indicate a peak in this region is used by our model for each of these functional groups. Additionally, a strong C=O peak is typically observed for carbonyl compounds near 1600 cm^{-1}, but the model only placed importance on this peak for the amide functional group. The example alcohol compound contained both an alcohol group and a carboxylic acid, and the model ignored the C=O in the prediction of the alcohol placing importance on peaks

corresponding to the O-H stretch near 3500 cm^{-1}. These results show that the model reproduces the 'known chemistry' of functional group features without explicitly giving such rules to predict.

However, from our chosen set of molecules with MPR of 1, none of the backpropagation plots revealed any chemically significant characteristics for alkynes, amines, ketones, alkyl halides, and acyl halides. Instead, it appears that these functional groups are identified by the lack of sharp peaks in various regions of the spectra. This observation is interesting as the functional group F-1 for these groups are relatively high. While nitrile groups have the lowest performance, the model was able to identify the 2210-2260 cm^{-1} band that is characteristic of this functional group. For the amine functional group, the model places high importance on a peak around 1550-1640 cm^{-1}. Although this may appear to indicate learned chemistry since the known N-H bending in this region, it also conflicts with the N-O bend of a nitro group. This observation may explain the reason our model misclassifies many nitro compounds as amides. Fortunately, there is a second N-O bend present which may rectify this issue if we include nitro groups to the model separately.

Next, we investigated the compounds with atleast one incorrect functional group prediction (MPR = 0) provided in **Listing S1** file. There are noticeable patterns of functional group types present in the set of failures. One example is nitro groups, which appear over 20 times in the failed compounds. This group is of interest as it is characterized by two strong bands which overlap with bending modes in alkane and amides functional groups. Many of these nitro compounds are misclassified as amides or alkanes. This observation partially explains the poor performance of amide functional groups shown in **Figure 3d**. Although it is discouraging to note that the model was unable to 'ignore' these peaks, the low count of amides present in the dataset may attribute to this poor performance.

Additional functional groups classification do not affect model performance of the original definitions

In the previous section, we show that some functional groups explicitly trained in the MLP model were incorrectly classified due to overlapping peaks belonging to functional groups that were not included in our original set of functional group types. We hypothesized that the separate classification of the "overlapping" functional groups could affect performance of our model. To test this hypothesis, we introduced the 'nitro,' 'ether,' and 'aldehyde' groups to the model. The 'nitro' group has significant overlap with the nitrile group (see the previous section), while the 'ether' group did not have peak values which overlapped with other functional groups in our previous definition. Another limitation of our model is the inability to distinguish methyl groups from other alkane functional groups. We propose that this is possible due to the lack of a C-C stretch in methyl groups and methyl groups contain characteristic peaks not present in other alkane groups (i.e. the CH3 bend). In the NIST dataset, many alkyl halides are present which do not contain any C-H bonds as all hydrogens in the molecule have been halogenated. Due to the large size of the alkane functional group in the training set, we hypothesize that splitting the alkane group into methyl and 'other' alkanes will not result in a large decrease in performance. Therefore, we decided to subdivide the 'alkane' group into 'methyl,' and 'other' alkanes as these groups performed the best out of all other groups in the original model.

Figures 5a-c show the results of these two hypotheses with details presented in **Figure S5** and **Tables S9, S10**. The relatively high F1 scores for the 'methyl' (0.932) and 'other' alkane (0.936) groups support our hypothesis that sub-division of the original alkane definition does not decrease performance. **Figures 5a-b** also suggest that our hypothesis to improve low

performance of functional groups by introduction of new functional groups for both IR and IR+MS MLP model is incorrect (compare **Tables S9, S10** with **Tables S3, S5**). Although the nitrile and amide groups do not show improvement after the introduction of the nitro and ether groups as the F-1 score for nitriles decreased by 0.019 and increased by 0.032, the new groups perform well as compared to the original problematic groups (0.932 for nitro groups and 0.923 for ethers). This suggests that the addition of new functional groups does not cause a significant loss in F1 score for other groups. Therefore, we speculate that more complex groups could be added to the model to provide detailed structural information, such as, a model to identify heterocyclic aromatic rings from rings comprised of only carbon. While further subdivision of functional groups is beyond the scope of this work, they present a potential extension of this work towards realization of autonomous instrumentation that result in minimal manual intervention.

Number of functional group predictions affects molecular perfection rate

We hypothesized that our stringent metric of MPR was affected by the increase in the number of functional group predictions for a given model. To test this hypothesis, we have a created synthetic models based on the accuracies of each functional group from the trained IR+MS model (see **Synthetic Models** in the supporting information for more details). The machine learning model outperforms these synthetic models (**Figures 5d and S5d**), indicating that increasing the number of functional groups does not decrease this metric more than what would be expected from the inclusion of additional functional groups alone. The overall conclusion of this section is encouraging as it suggests that more functional groups can be added to our model without affecting the model's performance to predict other functional groups. Values for the MPR and MF1 scores for the new functional group definitions are given in **Tables S11** and **S12**.

We were also interested to assess the performance of our model on molecules with different number of functional groups. We calculated the molecular perfection rate for compounds with one through six functional groups for both the original definition as well as the new definition of functional groups (results shown in **Figure 6**). Unfortunately, the result is inconclusive as the original versus new functional group definitions follow very different patterns. However, the original set of functional groups outperforms the new set of definitions. This observation is likely due to the reduced accuracy of the new alkane due to the split into methyl and non-methyl groups as both have accuracies of 91% where the previous model had an accuracy of 95% (**Figure 5c**). Since the result is biased on the set of molecules, we propose to still use new definitions of functional groups due to consistent performance over multiple groups.

Encoding spectra data in latent space retains functional group prediction performance

Given the success of our MLP model in predicting functional groups using complete standardized spectra, we wished to investigate the ability of an autoencoder to reduce the spectra into a latent-space. We trained a simple linear model for encoding the IR and MS spectra into a 256-length vector and decoding this vector back to the original spectra used to create the vector (see **Figure 1**). The 256-length vector was used to train a second network for multi-task functional group prediction. For individual functional groups, the autoencoder model performs similar to that of the original MLP model (F1 scores given in **Tables S13** and **S14**) The molecular performance of the autoencoder model is similar to that of the original MLP model (**Figure 7**) as the MPR for the autoencoder model is 62.6% and the MF1 score is 0.905 as compared to 65.2% and 0.912 for the original model (**Tables S15 and S16**). This reveals that the original spectra

contain redundant features that relates IR and mass spectra. We plan to explore the use of this latent space for inverse design of molecules with combined spectral properties in future works.

Conclusion

We present a first, to our knowledge, machine learning method for *de novo* prediction of functional groups using a combination of IR and MS data. We introduce two new metrics apart from functional group F1-score, namely, molecular F1-score and a stringent criteria as molecular perfection rate for practical use of our models. Our results show that, in general, the IR data is more consistent for the predicting functional groups than MS data, a conclusion backed by chemical intuition. However, several functional group predictions benefit from the inclusion of MS data. Additionally, our model architecture is more optimal for analysis of IR data due to the continuous nature of these spectra, and the mathematical structure of an MLP model. In future work, we hope to show that the use of a different architectures that are more suited for MS data may predict functional groups consistently. Our model's performance is not affected by the number of functional groups present in the training data and it predicted all the functional groups consistently across all metrics. This is essential for practical use in manual or automated chemical synthesis with a goal of minimal human intervention to fully realize autonomous instrumentation. Moreover, several known chemical patterns in the spectra were identified as features for the model to identify commonly occurring functional groups without any expert training of the system. We conclude that a *de novo* functional group identification problem is best set-up as a multi-class, multi-label problem for further studies to combine spectroscopic data types that may reveal unknown features useful for the identification of compounds. We show that our approach for functional group predictions is flexible as it can be extended to

introduce new or sub-divide existing functional groups into different classes without affecting performance of original functional group definitions. Furthermore, reducing chemical spectral data in a latent space does affect model performance to predict functional groups but can be used for inverse design of molecules based on combination of spectral properties.

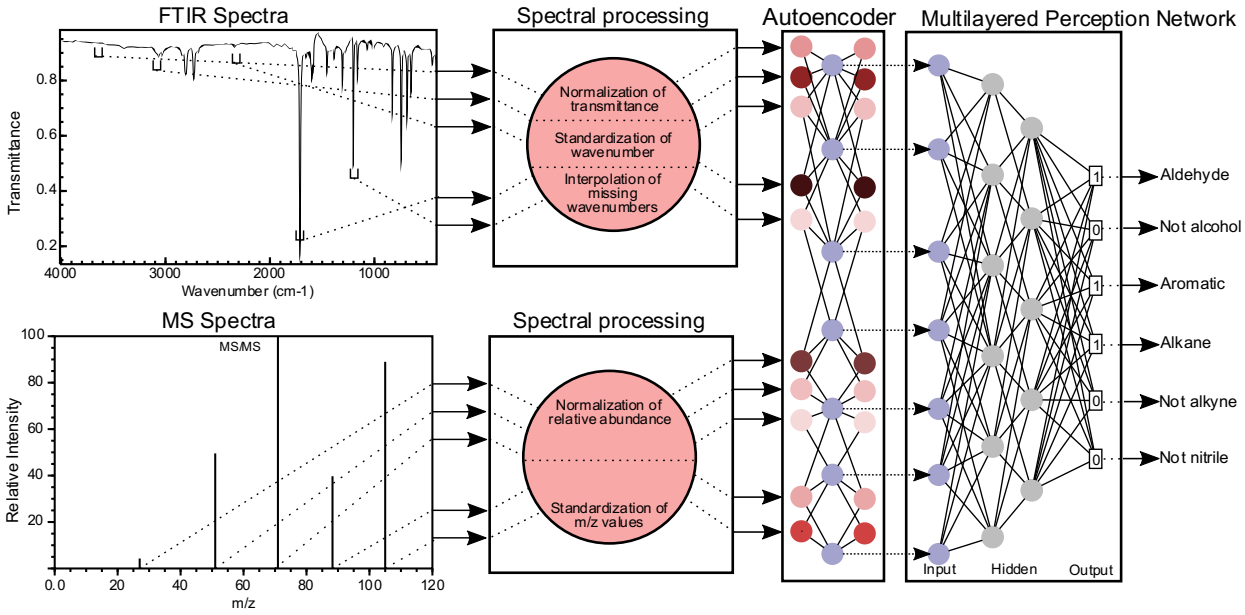

Figure 1: Overview of the MLP methodology for the classification of functional groups using FTIR and MS data. FTIR spectra are processed as to normalize the transmittance of the spectra and discretize the wavenumber numbers (creating wavenumber bins), thereby standardizing the wavenumbers for all FTIR spectra. Missing wavenumber bins in each spectrum are interpolated using B-Splines. A similar process is used for mass spectra data with the exception that no interpolation is performed. The normalized transmittance in all bins is encoded into a latent space by an autoencoder network and then used to predict the functional group of a molecule.

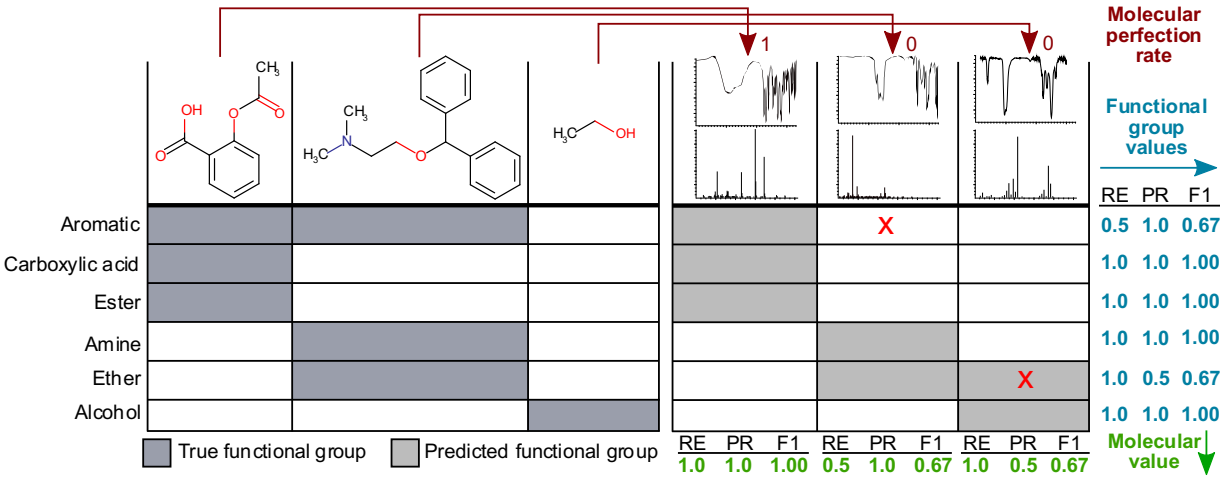

Figure 2: The left-hand side of the figure depicts the 'true' functional groups present in the example molecules, and the right-hand side are example predictions of the molecules functional groups given only their IR and MS spectra. Sample calculations for functional group F1, molecular F1, and molecular F1 score are given in the figure. Here, RE is short for 'recall' and PR is short for 'precision.'

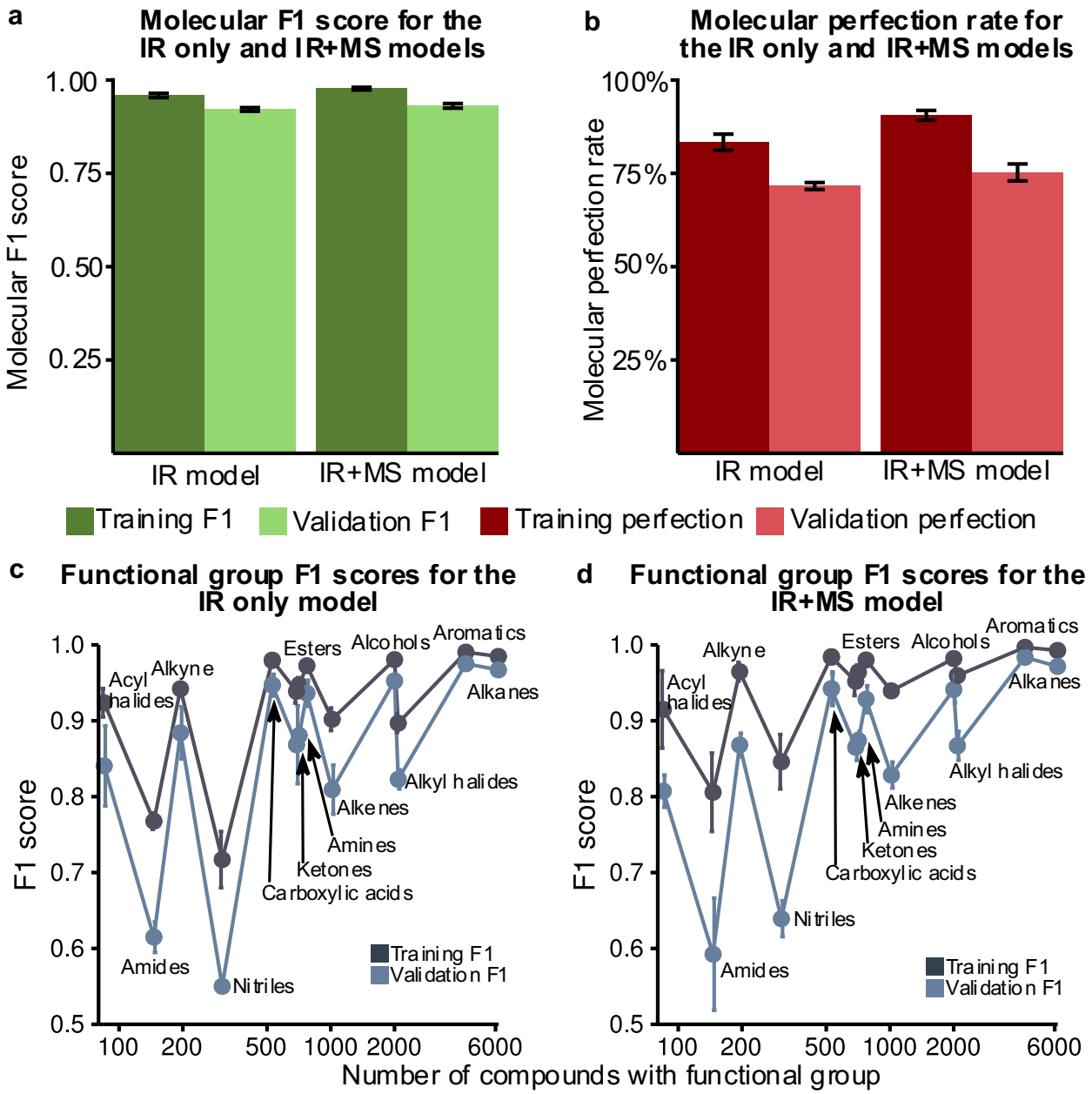

Figure 3: (**a**) The molecular F1 score for training and validation over the 5 folds is shown for both the optimized IR only and IR+MS models. The error bars indicate the standard deviation over the folds. (**b**) The molecular perfection for training and validation over 5 folds is shown for both the optimized IR only and IR+MS models. (**c**) The F1 score of the optimized IR only model plotted against the number of occurrences of that functional group. (**d**) The F1 score of the optimized IR+MS model plotted against the number of occurrences of that functional group.

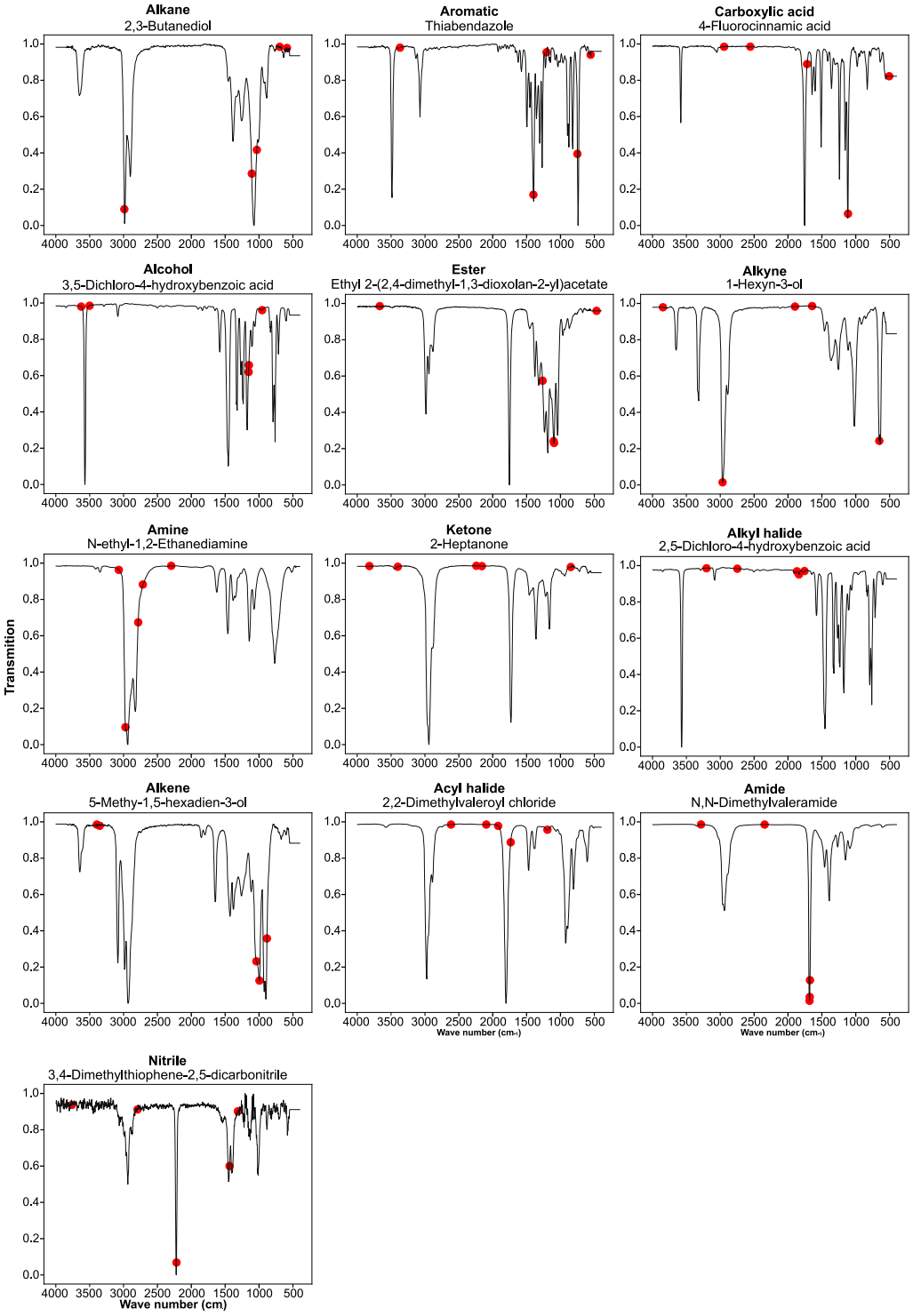

Figure 4: Backpropagation analysis for all 13 functional groups was performed to identify the regions of the spectra responsible for the result given. These plots are listed above in order of decreasing F1 score for the optimized IR+MS model.

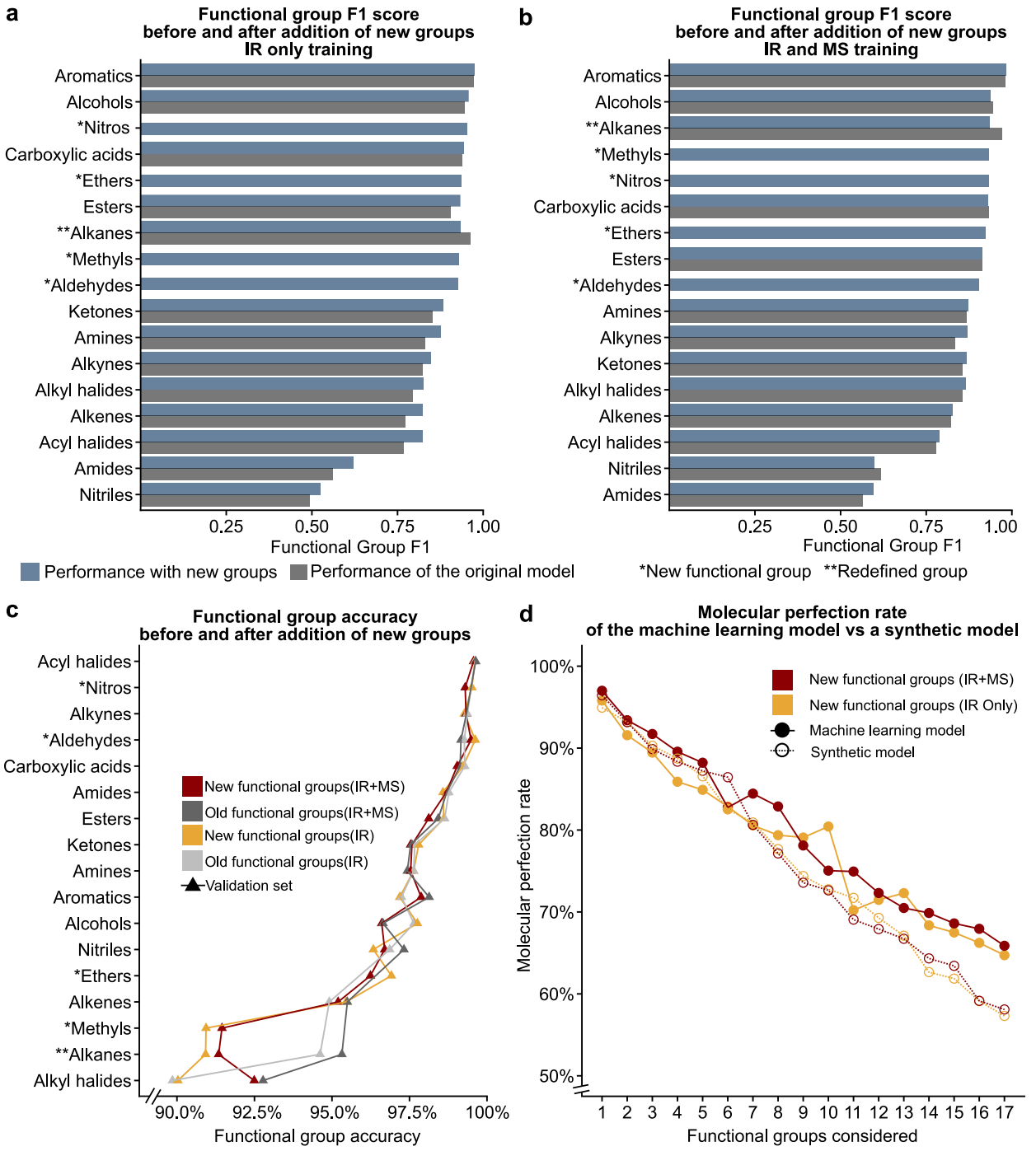

Figure 5: The bar plots given in (**a**) – (**b**) compare the functional group F1 scores for the original definitions of functional groups to the new definitions (see **Table 1**) showing that the addition of new additional functional groups does not have a significant impact on the previous functional groups. The line plot in (**c**) shows that the accuracy only decreases for the redefined functional group. The plot of molecular perfection rate in (**d**) compares the performance of the machine learning model to a synthetic model to show that the decrease in molecular perfection rate is expected as the number of functional groups increases.

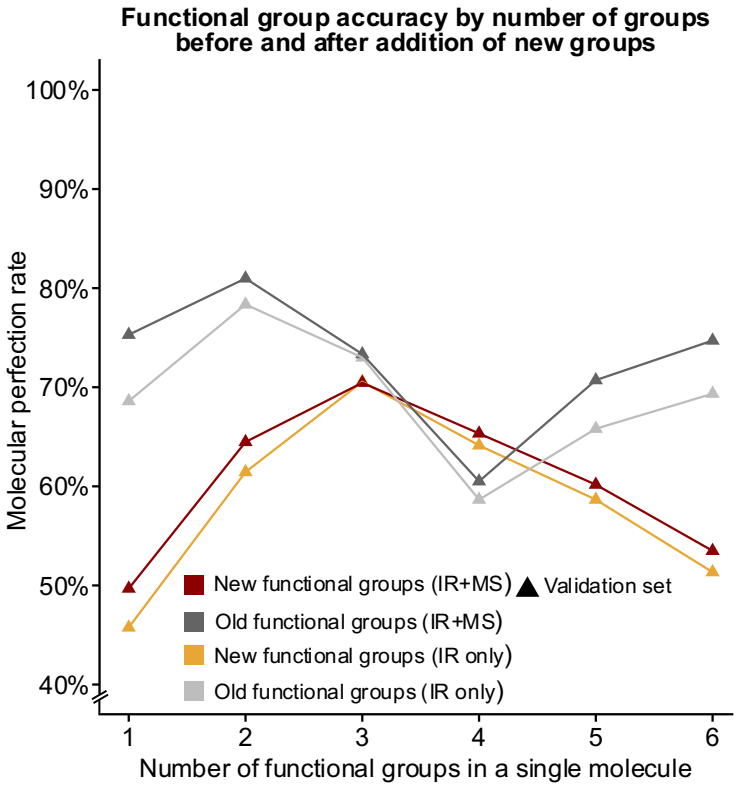

Figure 6: The molecular perfection rate calculated on molecules with a specific number of functional groups for both the original and new set of functional groups.

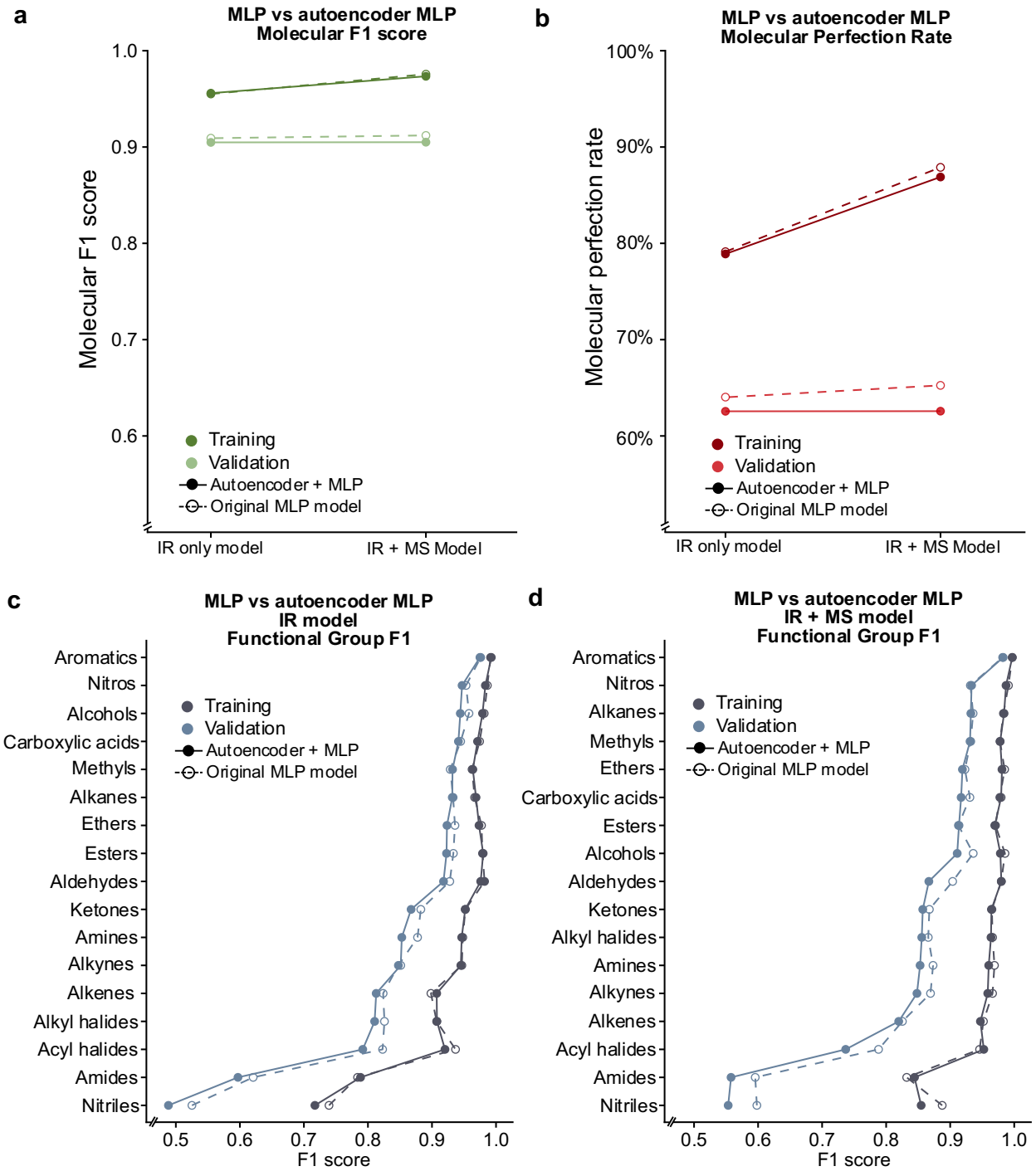

Figure 7: Comparison between the original MLP model and the autoencoder based model using the (**a**) molecular F1 metric and (**b**) molecular perfection rate are shown. Individual functional group F1 scores are provided for the IR only (**c**) and IR+MS (**d**) latent spaces.

Table 1: SMARTS strings used to identify the presence of a functional group given the 2D topology of a molecule.

Functional group	Smarts String
Alkane**	[CX4]
Alkene	[$([CX2]=[CX2])]
Alkyne	[$([CX2]#C)]
Arene	[c]
Ketone	[#6][CX3](=O)[#6]
Ester	[#6][CX3](=O)[OX2H0][#6]
Amide	[NX3][CX3](=[OX1])[#6]
Carboxylic acid	[CX3](=O)[OX2H1]
Alcohol	[CHX4][OX2H]
Amine	[NX3;H2,H1;!$(NC=O)]
Nitrile	[NX1]#[CX2]
Akyl halide	[CX4][F,Cl,Br,I]
Acyl halide	[CX3](=[OX1])[F,Cl,Br,I]
Ether*	[OD2]([#6])[#6]
Nitro*	[$([NX3](=O)=O),$([NX3+](=O)[O-])][!#8]
Methyl*	[CH3X4]
Alkane*	[CX4;H0,H1,H2]

** The alkane group is redefined in the second set of functional group definitions.

* Groups only present in the second set of functional group definitions.

References

(1) Kolb, H. C.; Sharpless, K. B. The Growing Impact of Click Chemistry on Drug Discovery. *Drug Discov. Today* **2003**, *8* (24), 1128–1137.

(2) Chatani, S.; Nair, D. P.; Bowman, C. N. Relative Reactivity and Selectivity of Vinyl Sulfones and Acrylates towards the Thiol–Michael Addition Reaction and Polymerization. *Polym. Chem.* **2013**.

(3) Freitas, V.; Ribeiro da Silva, M. Influence of Hydroxyl Functional Group on the Structure and Stability of Xanthone: A Computational Approach. *Molecules* **2018**, *23* (11), 2962.

(4) Marshall, B. D.; Bokis, C. P. A PC-SAFT Model for Hydrocarbons II: General Model Development. *Fluid Phase Equilib.* **2018**, *478*, 34–41.

(5) Dai, Y.; Zhu, Z.; Cao, Z.; Zhang, Y.; Zeng, J.; Li, X. Prediction of Boiling Points of Organic Compounds by QSPR Tools. *J. Mol. Graph. Model.* **2013**, *44*, 113–119.

(6) Withnall, M.; Chen, H.; Tetko, I. V. Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective. *ChemMedChem* **2018**, *13* (6), 599–606.

(7) Takei, T.; Nakada, M.; Yoshikawa, N.; Hiroe, Y.; Yoshida, H. Effect of Organic Functional Groups on the Phase Transition of Organic Liquids in Silica Mesopores. *J. Therm. Anal. Calorim.* **2016**, *123* (3), 1787–1794.

(8) Bruice, P. Y. *Essential Organic Chemistry*, 3rd ed.; Pearson: Upper Saddle Reiver, New Jersey, 2003.

(9) Cordeiro, F. B.; Ferreira, C. R.; Sobreira, T. J. P.; Yannell, K. E.; Jarmusch, A. K.; Cedenho, A. P.; Lo Turco, E. G.; Cooks, R. G. Multiple Reaction Monitoring (MRM)-Profiling for Biomarker Discovery Applied to Human Polycystic Ovarian Syndrome. *Rapid Commun. Mass Spectrom.* **2017**, *31* (17), 1462–1470.

(10) Minai-Tehrani, A.; Jafarzadeh, N.; Gilany, K. Metabolomics: A State-of-the-Art Technology for Better Understanding of Male Infertility. *Andrologia*. John Wiley & Sons, Ltd (10.1111) August 1, 2016, pp 609–616.

(11) Ewing, A. V.; Kazarian, S. G. Infrared Spectroscopy and Spectroscopic Imaging in Forensic Science. *Analyst*. Royal Society of Chemistry January 16, 2017, pp 257–272.

(12) Risoluti, R.; Materazzi, S.; Gregori, A.; Ripani, L. Early Detection of Emerging Street Drugs by near Infrared Spectroscopy and Chemometrics. *Talanta* **2016**, *153*, 407–413.

(13) Manheim, J.; Doty, K. C.; McLaughlin, G.; Lednev, I. K. Forensic Hair Differentiation Using Attenuated Total Reflection Fourier Transform Infrared (ATR FT-IR) Spectroscopy. *Appl. Spectrosc.* **2016**, *70* (7), 1109–1117.

(14) Anastas, P. T.; Fontalvo Gómez, M.; Johnson Restrepo, B.; Stelzer, T.; Romañach, R. J. Process Analytical Chemistry and Nondestructive Analytical Methods: The Green Chemistry Approach for Reaction Monitoring, Control, and Analysis. In *Handbook of Green Chemistry*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2019; pp

257–288.

(15) Baker, M. J.; Trevisan, J.; Bassan, P.; Bhargava, R.; Butler, H. J.; Dorling, K. M.; Fielden, P. R.; Fogarty, S. W.; Fullwood, N. J.; Heys, K. A.; et al. Using Fourier Transform IR Spectroscopy to Analyze Biological Materials. *Nat. Protoc.* **2014**, *9* (8), 1771–1791.

(16) Li, J.; Hibbert, D. B.; Fuller, S.; Vaughn, G. A Comparative Study of Point-to-Point Algorithms for Matching Spectra. *Chemom. Intell. Lab. Syst.* **2006**, *82* (1–2), 50–58.

(17) Griffiths, J. A Brief History of Mass Spectrometry. *Anal. Chem.* **2008**, *80* (15), 5678–5683.

(18) 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.* **2015**, *112* (41), 12580–12585.

(19) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; et al. Identification of Small Molecules Using Accurate Mass MS/MS Search. *Mass Spectrometry Reviews*. John Wiley & Sons, Ltd July 1, 2018, pp 513–532.

(20) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ. Sci. Technol.* **2014**, *48* (4), 2097–2098.

(21) Hufsky, F.; Böcker, S. Mining Molecular Structure Databases: Identification of Small Molecules Based on Fragmentation Mass Spectrometry Data. *Mass Spectrom. Rev.* **2017**, *36* (5), 624–633.

(22) Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M. E.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D. Automated Process. *Org. Synth.* **2015**, *347* (6227), 1221.

(23) 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* (7714), 377–381.

(24) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C–N Cross-Coupling Using Machine Learning. *Science (80-.).* **2018**, *360* (6385), 186–190.

(25) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein–ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38* (3), 169–177.

(26) Pereira, F.; Xiao, K.; Latino, D. A. R. S.; Wu, C.; Zhang, Q.; Aires-De-Sousa, J. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *J. Chem. Inf. Model.* **2017**, *57* (1), 11–21.

(27) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55* (2), 263–274.

(28) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; Von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13* (11), 5255–5264.

(29) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725–732.

(30) Sanchez-Lengeling, B.; Outeiral, C.; Guimaraes, G. L.; Aspuru-Guzik, A. Optimizing Distributions over Molecular Space. An Objective-Reinforced Generative Adversarial Network for Inverse-Design Chemistry (ORGANIC). *ChemRxiv* **2017**, 1–18.

(31) Benhenda, M. ChemGAN Challenge for Drug Discovery: Can AI Reproduce Natural Chemical Diversity? **2017**.

(32) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58* (6), 1194–1204.

(33) Guimaraes, G. L.; Sanchez-Lengeling, B.; Outeiral, C.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *ChemRxiv* **2017**, 1–7.

(34) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. DruGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharm.* **2017**, *14* (9).

(35) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models.

(36) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120–131.

(37) Xu, Z.; Wang, S.; Zhu, F.; Huang, J. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. *Proc. 8th ACM Int. Conf. Bioinformatics, Comput. Biol. Heal. Informatics - ACM-BCB '17* **2017**, 285–294.

(38) Jing, Y.; Bian, Y.; Hu, Z.; Wang, L.; Xie, X.-Q. S. Deep Learning for Drug Design: An Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS J.* **2018**, *20* (3).

(39) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276.

(40) Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Site of Reactivity Models Predict Molecular Reactivity of Diverse Chemicals with Glutathione. *Chem. Res. Toxicol.* **2015**, *28* (4), 797–

809.

(41) Hughes, T. B.; Le Dang, N.; Miller, G. P.; Swamidass, S. J. Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network. *ACS Cent. Sci.* **2016**, *2* (8), 529–537.

(42) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**.

(43) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving beyond Fingerprints. *J. Comput. Aided. Mol. Des.* **2016**, *30* (8), 595–608.

(44) Nalla, R.; Pinge, R.; Narwaria, M.; Chaudhury, B. Priority Based Functional Group Identification of Organic Molecules Using Machine Learning. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data - CoDS-COMAD '18*; 2018; pp 201–209.

(45) Barbon, S.; Costa Barbon, A. P. A. da; Mantovani, R. G.; Barbin, D. F. Machine Learning Applied to Near-Infrared Spectra for Chicken Meat Classification. *J. Spectrosc.* **2018**, *2018*, 1–12.

(46) Fu, W.; Hopkins, W. S. Applying Machine Learning to Vibrational Spectroscopy. *J. Phys. Chem. A* **2018**, *122* (1), 167–171.

(47) Fessenden, R. J.; Györgyi, L. Identifying Functional Groups in IR Spectra Using an Artificial Neural Network. *J. Chem. Soc., Perkin Trans. 2* **1991**, No. 11, 1755–1762.

(48) Robb, E. W.; Munk, M. E. A Neural Network Approach to Infrared Spectrum Interpretation. *Mikrochim. Acta* **1990**, *100* (3–4), 131–155.

(49) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22* (10), 1345–1359.

(50) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1).

(51) Landrum, G. A. RDKit: Open-Source Cheminformatics.

(52) Hanson, R. M. Jmol SMILES and Jmol SMARTS: Specifications and Applications. *J. Cheminform.* **2016**.

(53) DayLight. SMARTS.

(54) Chollet, F. Keras. **2015**.

(55) Gabel, J.; Desaphy, J.; Rognan, D. Beware of Machine Learning-Based Scoring Functions-on the Danger of Developing Black Boxes. *J. Chem. Inf. Model.* **2014**, *54* (10), 2807–2815.

(56) Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; Riedmiller, M. Striving for Simplicity: The All Convolutional Net. **2014**.

(57) Zeiler, M. D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; 2014; Vol. 8689 LNCS, pp 818–833.

(58) *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.