Electronic Supplementary Material (ESI) for Digital Discovery.

Supplements for Understanding the patterns that neural networks learn from chemical spectra

Laura Hannemose Rieger[1], Max Wilson[2], Tejs Vegge[1], and Eibar Flores[3]

[1]Department of Energy Conversion and Storage, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark.
[2]Technical University of Denmark, Department of Applied Mathematics and Computer Science, Kgs. Lyngby, Denmark
[3]Sustainable Energy Technology, SINTEF Industry, Sem Sælands Vei 12, Trondheim, 7034 Norway

1 Size and composition of molecules in the dataset

Fig. S1 illustrates the distribution of molecule sizes in the dataset. Out of the 14,346 molecules in the dataset, most of them have less than 40 non-hydrogen atoms. Fig. S2 shows the number of molecules containing at least one of the 10 most common atom types; the proportion of molecules with atoms other than those shown is negligible.

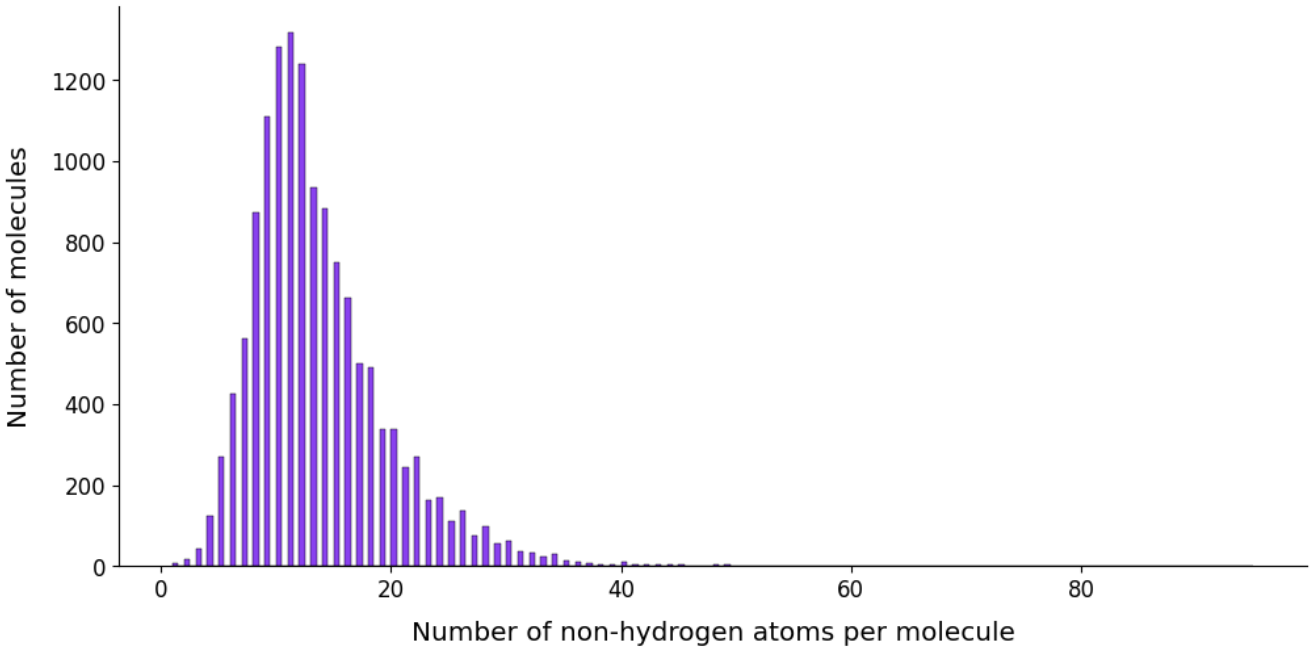

Figure S1: Distribution of molecular sizes: number of non-hydrogen atoms per molecule.

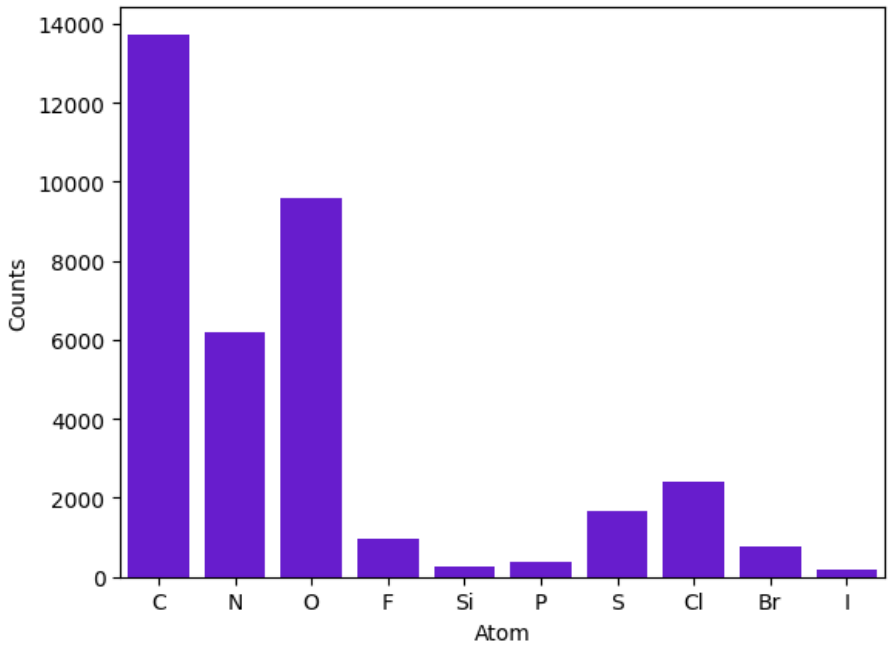

Figure S2: Distribution of atomic composition of molecules: number of molecules containing at least one of the 10 most common atom types in the dataset.

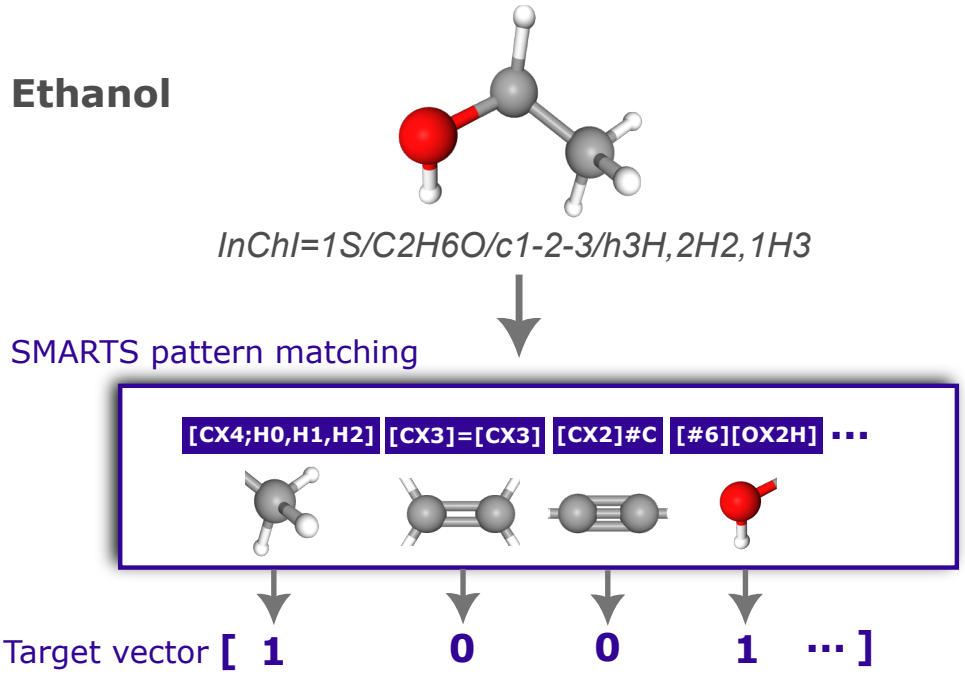

Figure S3: Pattern matching algorithm used to recognize the presence of functional groups, using ethanol as example.

2 Identification of functional groups

Fig. S3 illustrates the pattern matching algorithm used to recognize the presence of functional groups in ethanol. The algorithm takes as inputs an INChI string and a list of structural patterns in SMARTS notation (Section 2). Each SMARTS pattern is designed to recognize a functional group. The output is a vector encoding the presence of a functional group with a "1", or "0" otherwise. In the example of ethanol, the algorithm identifies the signature of an alkane and an alcohol group, and encodes both in the target vector with "1".; absent groups are encoded as "0". In contrast to Nalla et al, who assign a single dominant functional group per molecule [1], we keep all the functional groups identified in the molecule, since each functional group will result in a particular spectral signature. In the case of ethanol, both O-H and CH_3 bands are observed in its FTIR spectrum. The complete list of SMARTS sub-patterns used in this work can be found in Table 2.

Functional Group	SMARTS Pattern
Alkane	[CX4;H0,H1,H2]
Methyl	[CH3X4]
Alkene	[CX3]=[CX3]
Alkyne	[CX2]#C
Alcohols	[#6][OX2H]
Amines	[NX3;H2,H1;!$(NC=O)]
Nitriles	[NX1]#[CX2]
Aromatics	[$([cX3](:*):*),$([cX2+](:*):*)]
Alkyl halides	[#6][F,Cl,Br,I]
Esters	[#6][CX3](=O)[OX2H0][#6]
Ketones	[#6][CX3](=O)[#6]
Aldehydes	[CX3H1](=O)[#6]
Carboxylic acids	[CX3](=O)[OX2H1]
Ether	[OD2]([#6])[#6]
Acyl halides	[CX3](=[OX1])[F,Cl,Br,I]
Amides	[NX3][CX3](=[OX1])[#6]
Nitro	[$([NX3](=O)=O),$([NX3+](=O)[O-])][!#8]

Table S1: SMARTS patterns for the functional groups. The way functional groups are grouped adhere to the scheme employed in preceding literature, and thus enables us to compare between the classification accuracy of our model to references[2,3]

.

3 Repeatability

As explained in Section 2.3, we examine the learnt patterns of a specific convolutional neural network. Since neural networks learn slightly different features dependent on the starting randomly initialized weights, these learnt patterns do not necessarily generalize across neural networks. We therefore examine the generalisation across differently seeded neural networks. Since the patterns learnt in the convolutional layers will be permuted across neural networks it is not possible to compare them without strong further assumptions. Instead we focus on comparing the important positions for each functional group across neural networks.

In Fig. S4 we visualize the maximum weight across all channels for each neural network across neural networks trained with the same architecture and training procedure but different random seeds. If large values are consistent in location across neural networks, we can conclude that the learnt features are robust.

Evidenced by the tight bound of the standard deviation around the mean, we can conclude that important features are largely universal between neural networks.

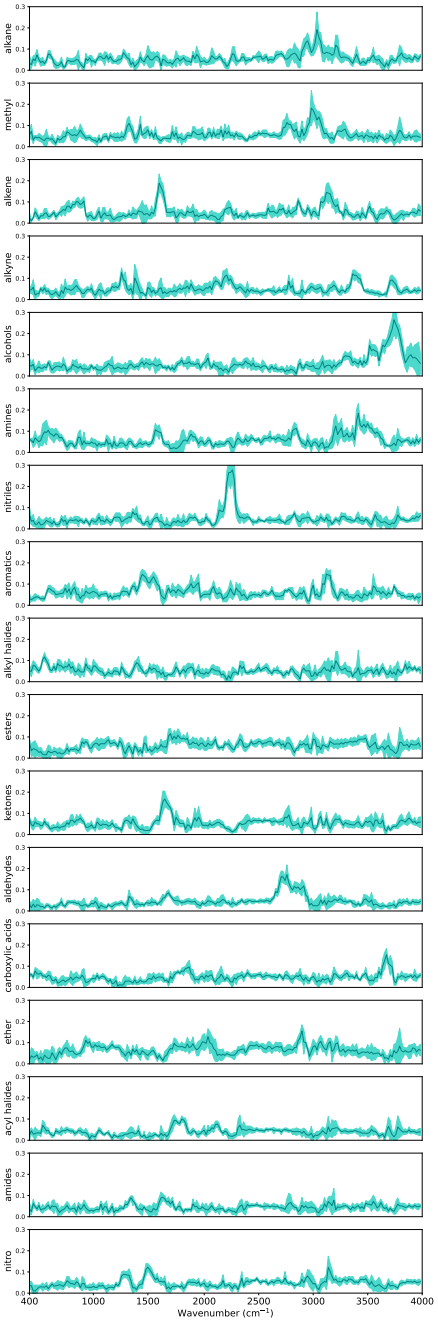

Figure S4: Weights of classifiers across randomly-initialized instances of the CNN. For each functional group we take the maximum weight (maximum taken across channels) across ten neural networks trained with different random seeds. We visualize the mean and standard deviation. Important regions (marked by high values) are constant across neural networks, indicating that similar features are learnt.

4 Weights of dense layer for all classifiers

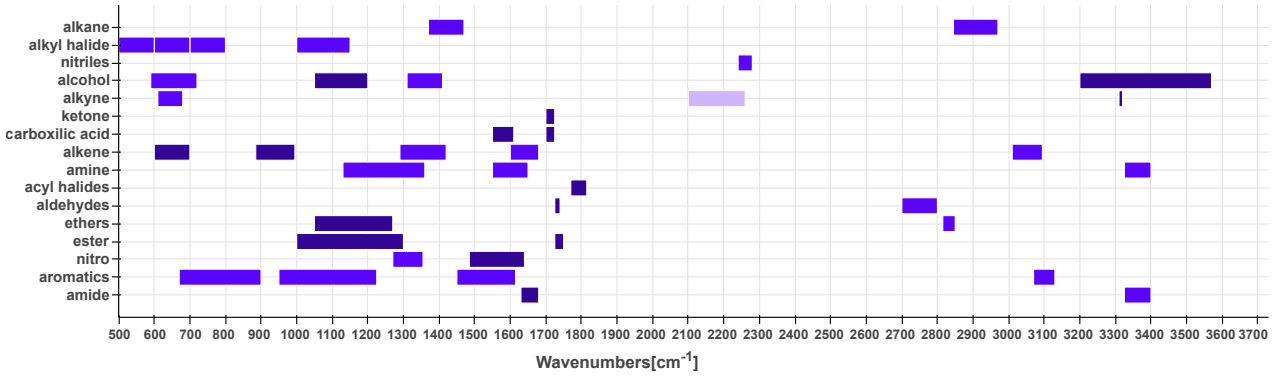

Figure S5: Group frequencies. Strong, mid and weak bands appear in dark, mid and light purple, respectively.

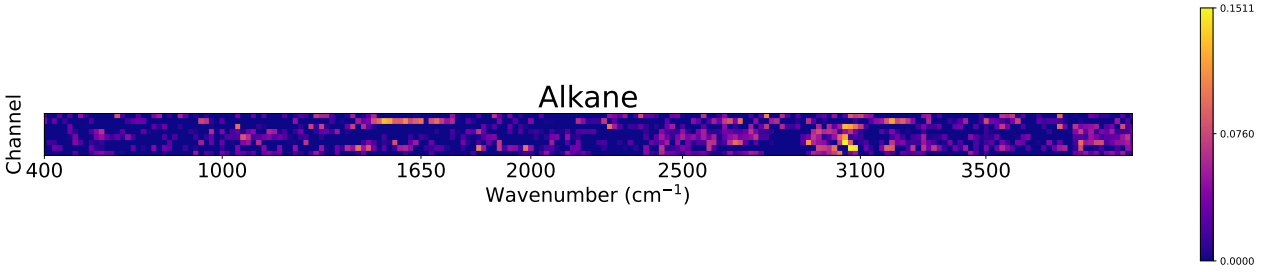

Figure S6: Weights of the alkane classifier

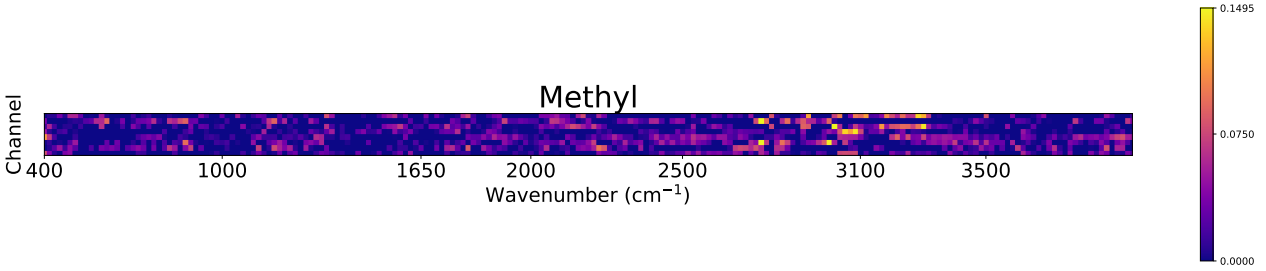

Figure S7: Weights of the methyl classifier

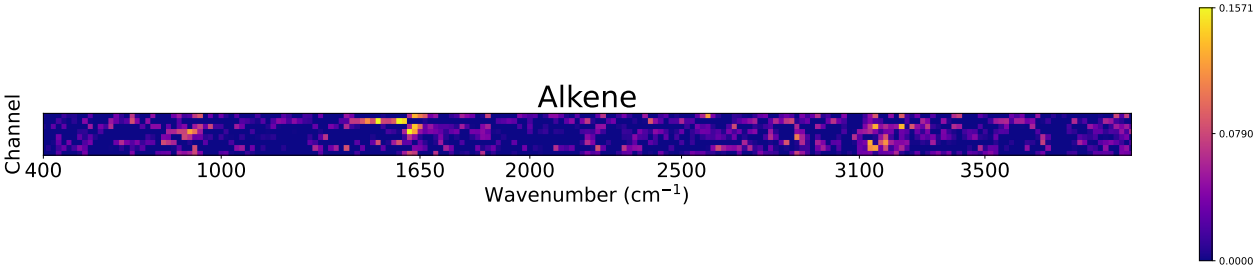

Figure S8: Weights of the alkene classifier

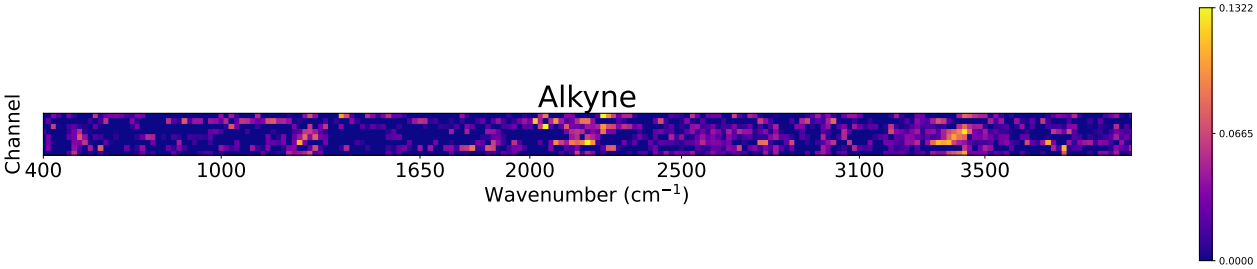

Figure S9: Weights of the alkyne classifier

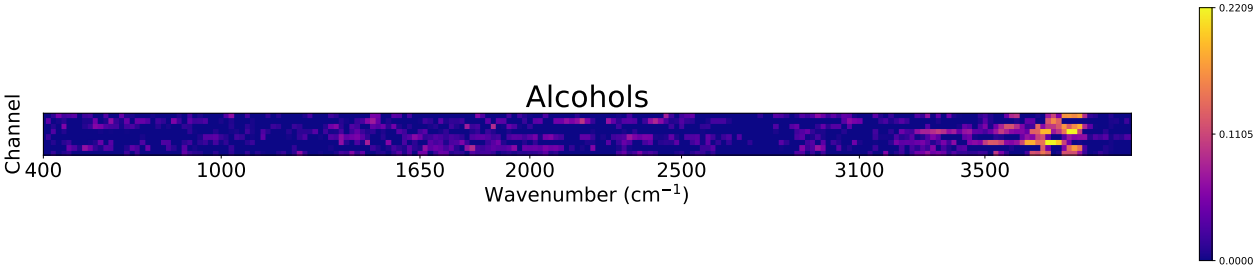

Figure S10: Weights of the alcohols classifier

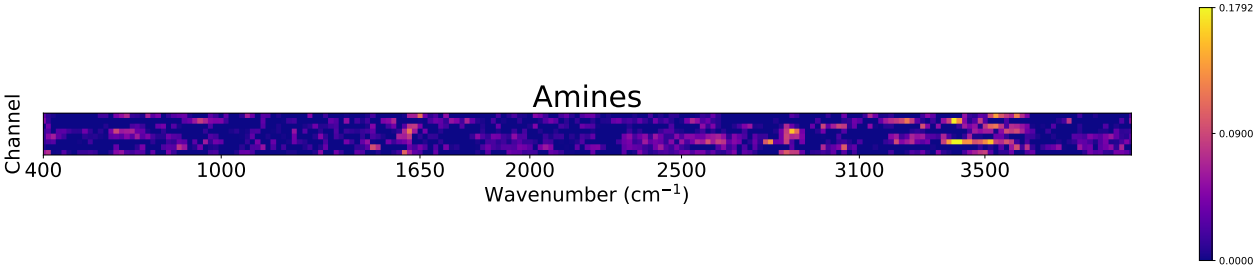

Figure S11: Weights of the amines classifier

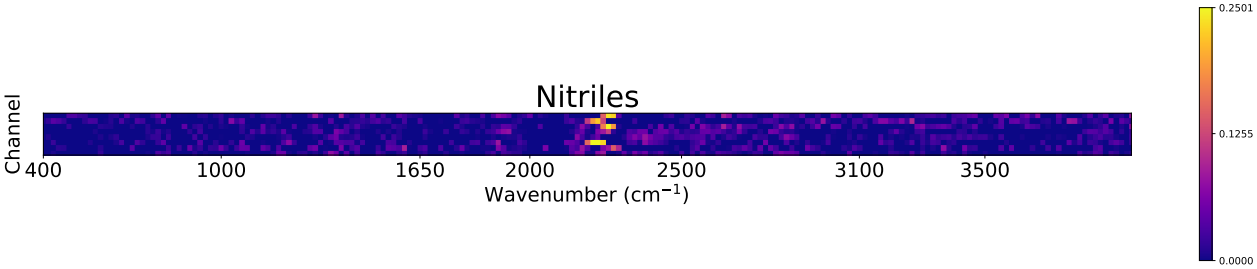

Figure S12: Weights of the nitriles classifier

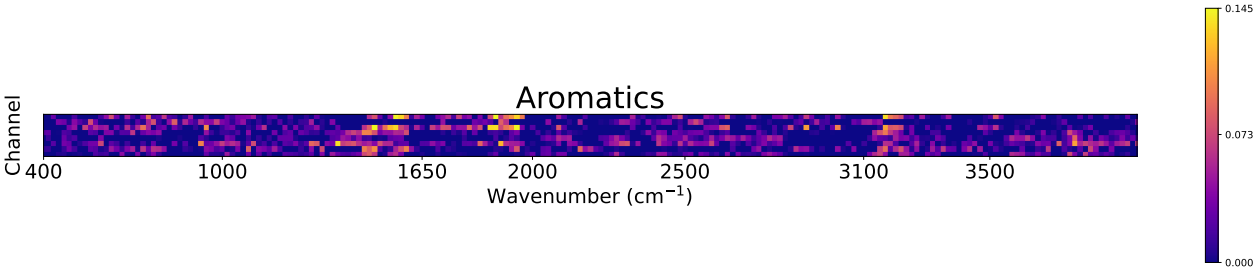

Figure S13: Weights of the aromatics classifier

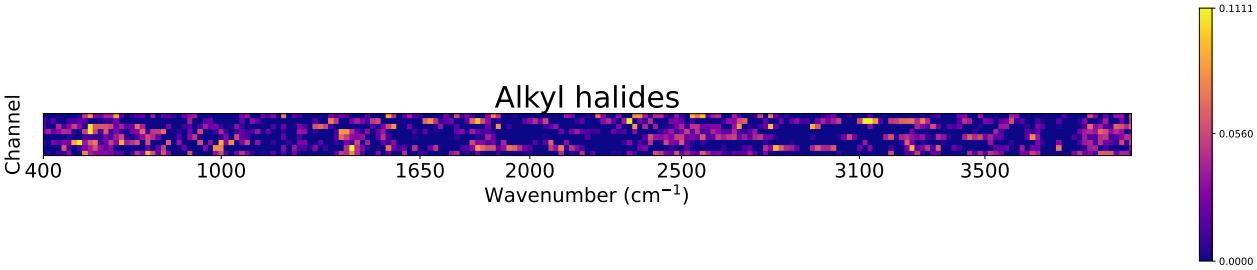

Figure S14: Weights of the alkyl halides classifier

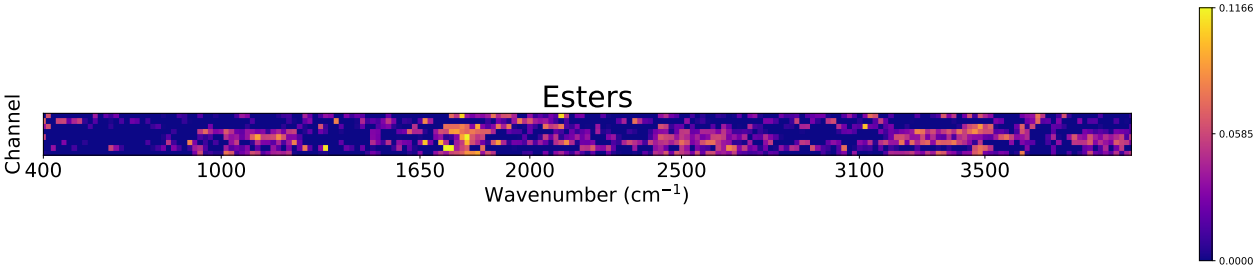

Figure S15: Weights of the ester classifiers

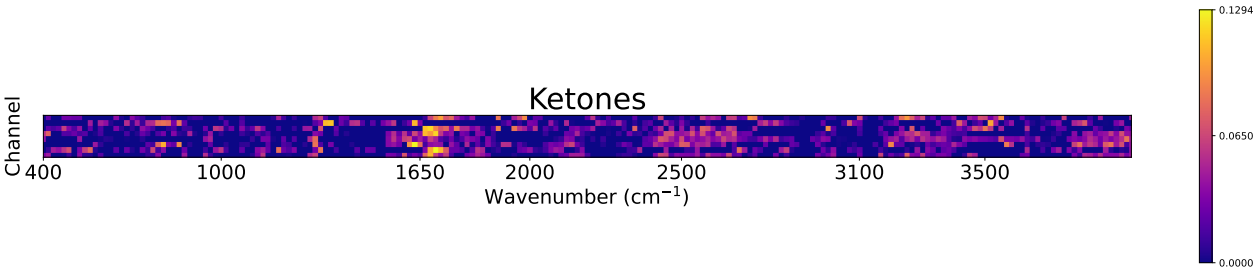

Figure S16: Weights of the ketones classifier

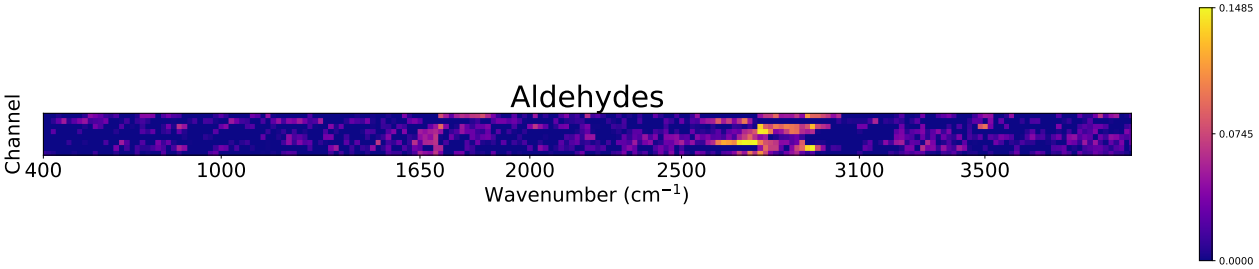

Figure S17: Weights of the aldehyde classifiers

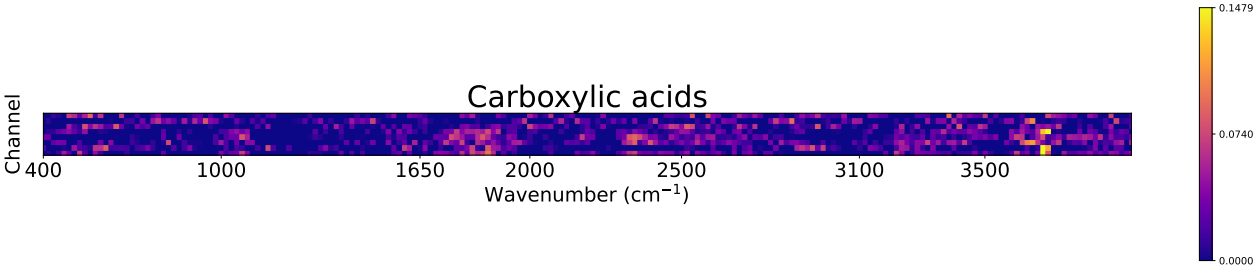

Figure S18: Weights of the carboxylic acids classifier

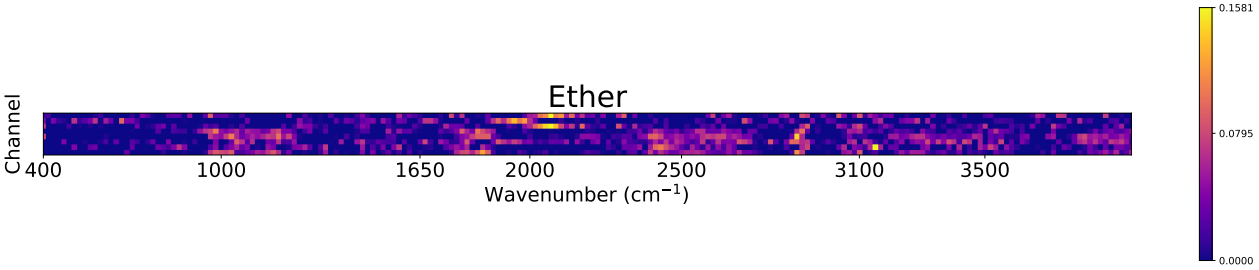

Figure S19: Weights of the ether classifier

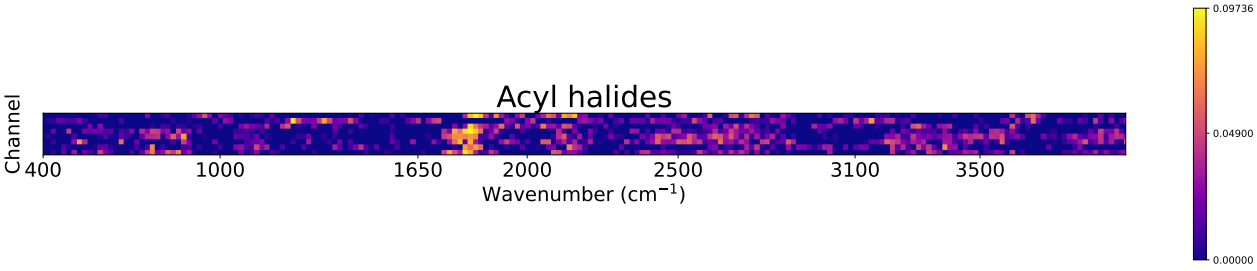

Figure S20: Weights of the acyl halides classifier

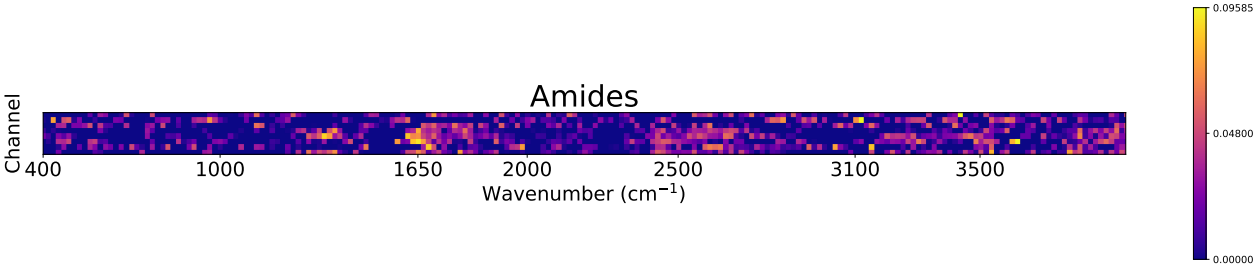

Figure S21: Weights of the amides classifier

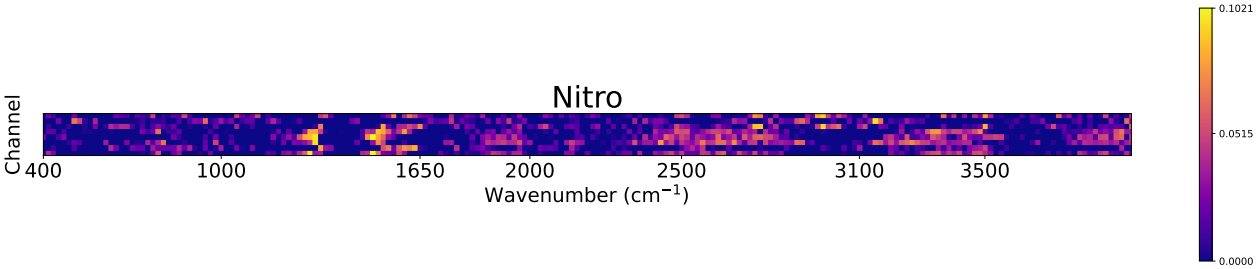

Figure S22: Weights of the nitro classifier

5 Class correlations

In Fig. S23 we visualize correlations between different functional groups being present. As can be seen, e.g. alkane and aromatics are less likely to be present in the same spectrum while ester and ether are likely to be present in the same spectrum.

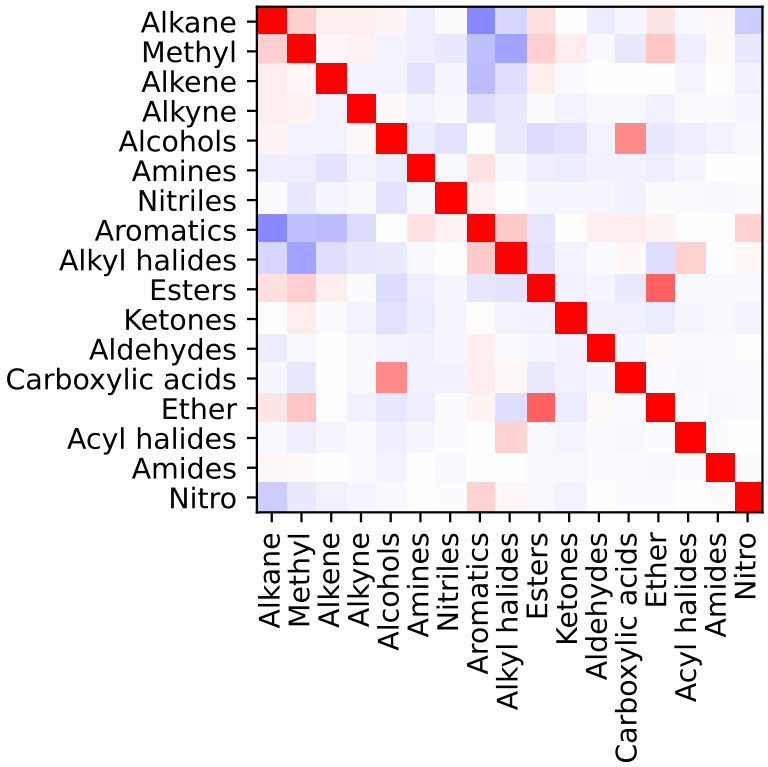

Figure S23: Correlation plot

6 Performance on out-of-distribution combinations

Some combinations of functional groups being present or not present are particularly common. We examine how the performance of the neural network for a particular functional groups changes dependent on other functional groups being present or not. Looking at Fig. S23, we identify the five most common combinations of functional groups being present or not present. The correlation coefficients are given in Table S2.

Note that the correlation coefficient for some pairings such as aromatics and alkane is negative, indicating that aromatics make it less likely for alkane to be prsent and vice versa. For each functional group in each pair we calculate the AUC metric if the other group is present and not present respectively.

We represent the results by showing the AUC score for a particular group when the paired group is not present on the left and the AUC score when the paired group is present on the right in Table S2.

Since we have identified the five pairings with the highest absolute correlation coefficient, we show the f1 score for 10 functional groups. For two of the groups there are no examples of the atypical pairing (there are no spectra with esters but no ether or carboxylic acid but no alcohol respectively).

We can see that the performance drop is relatively minor if a comparison can be made, i.e. if all possible combinations are in the test dataset.

Group 1	Group 2	Corr coeff	$AUC_{noGroup2}$	AUC_{Group2}
alkane	aromatics	-0.46	0.95	0.95
aromatics	alkane	-0.46	0.99	1.00
alkyl halides	methyl	-0.36	0.90	0.91
methyl	alkyl halides	-0.36	0.95	0.98
alkene	aromatics	-0.27	0.97	0.91
aromatics	alkene	-0.27	1.00	0.99
alcohols	carboxylic acids	0.46	0.98	N/A
carboxylic acids	alcohols	0.46	N/A	0.99
esters	ether	0.61	N/A	0.97
ether	esters	0.61	0.98	N/A

Table S2: Summary of the model performance for atypical and typical pairings of functional groups. Note that some AUC can not be calculated as one of the pairings (e.g. no alcohol but carboxylic acid) does not exist.

7 Spectra within the 2400-2700 cm^{-1} region

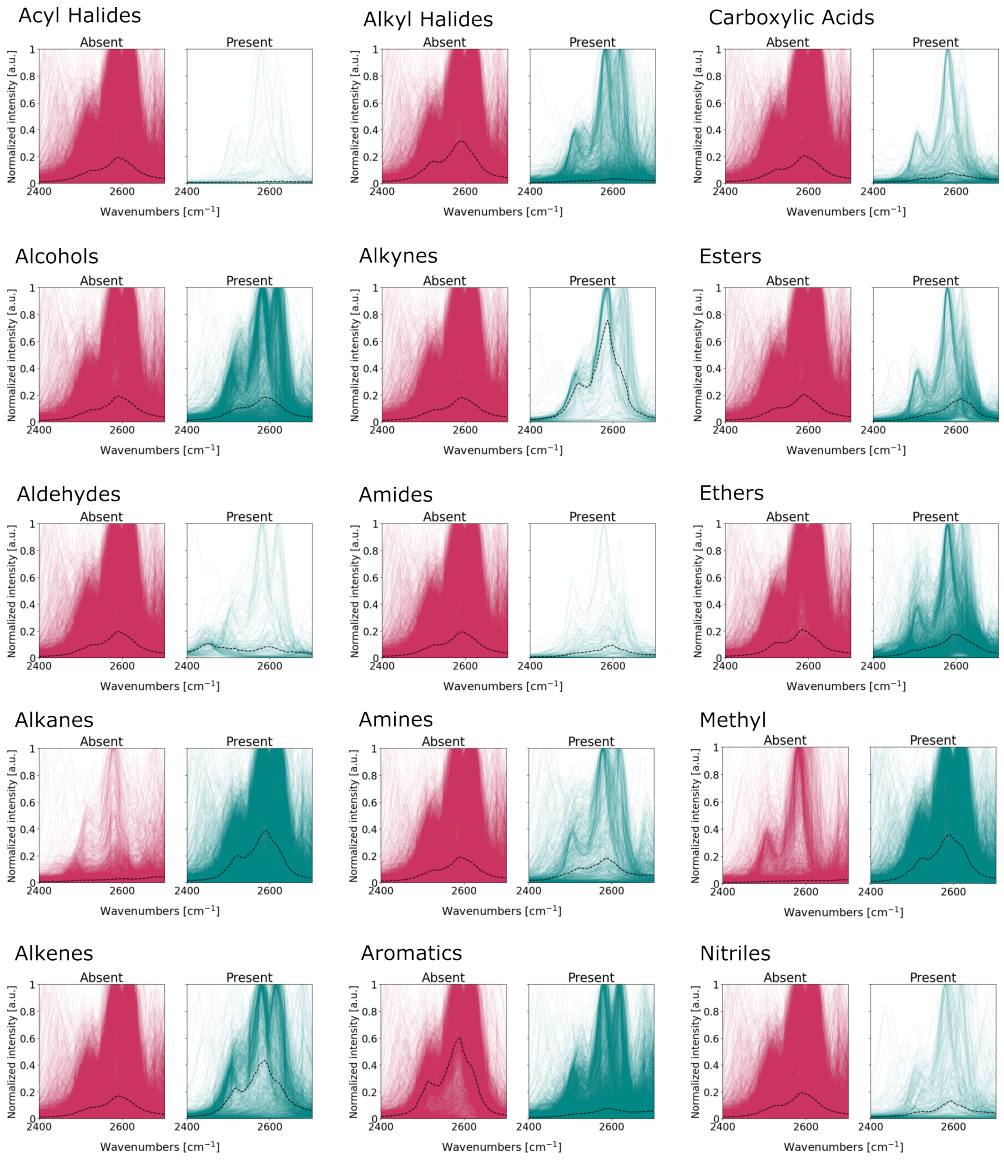

Figure S24: Sub-samples of infrared spectra from small organic molecules within the 2400-2700 cm^{-}1 region, for all studied functional groups. Each figure shows the overlap for spectra where the corresponding functional gropup is absent and present. The black dotted line indicates the median spectrum.

References

[1] Rushikesh Nalla, Rajdeep Pinge, Manish Narwaria, and Bhaskar Chaudhury. 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*, pages 201–209, 2018.

[2] Abigail A Enders, Nicole M North, Chase M Fensore, Juan Velez-Alvarez, and Heather C Allen. Functional group identification for ftir spectra using image-based machine learning models. *Analytical Chemistry*, 93(28): 9711–9718, 2021.

[3] Jonathan A Fine, Anand A Rajasekar, Krupal P Jethava, and Gaurav Chopra. Spectral deep learning for prediction and prospective validation of functional groups. *Chemical science*, 11(18):4618–4630, 2020.