

# Machine Learning Microplastic Characterisation Surpasses Human Performance and Uncovers Labelling Errors in Public FTIR Data

Frithjof Herb<sup>1,2\*</sup>, Mario Boley<sup>3</sup> and Wye-Khay Fong<sup>1,2\*</sup>

<sup>1</sup>Discipline of Chemistry, The University of Newcastle, University Drive,  
Newcastle, 2308, New South Wales, Australia.

<sup>2</sup>School of Chemistry, Monash University, Wellington Road, Melbourne,  
3800, Victoria, Australia.

<sup>3</sup>Department of Data Science and AI, Faculty of IT, Monash University,  
Wellington Road, Melbourne, 3800, Victoria, Australia.

\*Corresponding author(s). E-mail(s): [research@frithjofherb.com](mailto:research@frithjofherb.com);  
[khay.fong@monash.edu](mailto:khay.fong@monash.edu);

## 1 Supplementary Information

### 1.1 Source Code

The code used in this project can be found by following the below link to the GitHub repository. However, this repository is for code as it was at publication for archiving purposes. This repository will contain a link to a version of this project that is actively maintained and is receiving updates.

<https://github.com/Grimblewald/Neural-networks-for-microplastic-fourier-transform-infrared-spectroscopy-data>

### 1.2 Calculation of Performance Metrics

#### *Recall, Precision & F1*

Model performance is primarily evaluated using Precision and Recall metrics, and the relationship between these is captured through the F1 score. Further performance insights are gathered through AUC and ROC curves.

Precision is the proportion of true positive identifications among all positive identifications (both true and false positives). It is calculated as follows:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \quad (1)$$

Recall (also known as Sensitivity) is the proportion of true positive identifications among all actual positives (both true positives and false negatives). It is calculated as:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \quad (2)$$

These metrics are calculated on a one-vs-all basis, aggregating performance across all categories. This makes them suitable for imbalanced datasets or for an overall model assessment.

The  $F_\beta$  score is a single value metric that combines Precision and Recall. It is calculated as:

$$F_\beta = \frac{(1 + \beta^2) \cdot (Precision \cdot Recall)}{(\beta^2 \cdot Precision) + Recall} \quad (3)$$

When  $\beta = 1$ , Precision and Recall are valued equally, resulting in the  $F_1$  score:

$$F_1 = \frac{2 \cdot (Precision \cdot Recall)}{Precision + Recall} \quad (4)$$

### ***Receiver Operating Characteristic Curve & Area thereunder***

The Receiver Operating Characteristic Curve (ROC) is a graphical representation of the performance of a binary classifier system as its discrimination threshold is varied. It is created by plotting Recall against the False Positive Rate (FPR) at various threshold settings.

The FPR is calculated as:

$$FPR = \frac{FalsePositives}{TrueNegatives + FalsePositives} \quad (5)$$

The ROC curve provides a tool to select possibly optimal models and explore how distributions for True and False rates for both positives and negatives overlap, where a curve fitting more tightly into the upper right corner indicates less overlap between true/false pairs, which implies the model has more well-defined decision boundaries than for a curve which more closely resembles the diagonal.

The Area Under the Receiver Operating Characteristic Curve (AUC) aims to represent the information provided by the ROC as a single value metric. It is approximated using the trapezoidal rule given by:

$$AUC = \sum_{k=1}^{K-1} (FPR_{k+1} - FPR_k) \cdot \frac{(TPR_{k+1} + TPR_k)}{2} \quad (6)$$

where  $K$  is the total number of unique thresholds,  $FPR_k$  is the False Positive Rate, and  $TPR_k$  is the True Positive Rate at threshold  $t_k$ . AUC provides insight into the model's performance compared to random guessing, with 1.0 representing a perfect



model and 0.5 representing a model that is no better than random guessing. What this metric loses in its conciseness are measures of skew in the ROC curve, and so a model with an ROC indicating a propensity for false positives can have the same AUC as a model with a propensity for false negatives, therefore in cases of equivalent AUC it is wise to compare ROC and proceed in accordance with relative costs of each mistake.

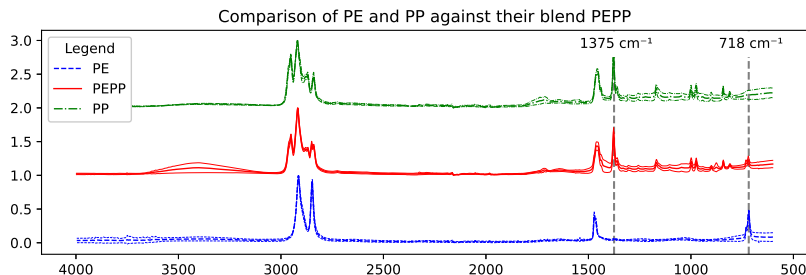
### **1.3 Table of Results for Per Dataset Model Performance Metrics**

| Dataset    | Categorical Accuracy |      |      | AUC    |      |      | Precision |      |      | Recall |      |      | F1     |      |      | Categorical Cross-entropy |      |      |
|------------|----------------------|------|------|--------|------|------|-----------|------|------|--------|------|------|--------|------|------|---------------------------|------|------|
|            | Median               | Min  | Max  | Median | Min  | Max  | Median    | Min  | Max  | Median | Min  | Max  | Median | Min  | Max  | Median                    | Min  | Max  |
| Kedzierski | 0.93                 | 0.91 | 0.93 | 0.00   | 0.99 | 1.00 | 0.94      | 0.93 | 0.95 | 0.00   | 0.91 | 0.93 | 0.00   | 0.91 | 0.94 | 0.00                      | 0.62 | 0.30 |
| Jung       | 0.97                 | 0.96 | 0.98 | 0.00   | 1.00 | 1.00 | 0.98      | 0.97 | 0.98 | 0.00   | 0.96 | 0.97 | 0.00   | 0.97 | 0.98 | 0.00                      | 0.42 | 0.30 |
| Carbery    | 0.90                 | 0.89 | 0.93 | 0.01   | 0.99 | 1.00 | 0.93      | 0.92 | 0.94 | 0.01   | 0.90 | 0.91 | 0.01   | 0.91 | 0.93 | 0.01                      | 0.55 | 0.35 |
| Brignac    | 0.94                 | 0.94 | 0.95 | 0.00   | 1.00 | 1.00 | 0.95      | 0.95 | 0.96 | 0.00   | 0.93 | 0.95 | 0.00   | 0.95 | 0.95 | 0.00                      | 0.70 | 0.23 |
| OpenSpecy  | 0.80                 | 0.66 | 0.87 | 0.03   | 0.94 | 0.92 | 0.97      | 0.84 | 0.90 | 0.03   | 0.76 | 0.84 | 0.03   | 0.79 | 0.87 | 0.04                      | 0.56 | 0.43 |
| Primiple   | 0.91                 | 0.90 | 0.93 | 0.01   | 0.99 | 0.98 | 0.99      | 0.92 | 0.96 | 0.01   | 0.89 | 0.86 | 0.91   | 0.92 | 0.90 | 0.94                      | 0.84 | 0.81 |
| Holdout    | 0.93                 | 0.92 | 0.94 | 0.00   | 1.00 | 1.00 | 0.95      | 0.93 | 0.95 | 0.00   | 0.91 | 0.90 | 0.93   | 0.91 | 0.92 | 0.94                      | 0.87 | 0.81 |

**Table 1** Summary of performance metrics for a deep learning model, showing the median and median absolute deviation (MAD) for each metric. These statistics provide insight into the model’s central tendency and variability, respectively, offering a comprehensive view of its performance.

## 1.4 Why we need the fingerprint region

Spectra were only included in the final dataset if they were collected within the minimum range, as the finger print region, as the name implies, is crucial for more precise identification of samples. For example, differentiating between PE, PP and PEPP is not easy without a signal at approximately  $718\text{cm}^{-1}$  (see Figure 1). So in order to improve outcomes, only data with this information is considered.



**Fig. 1** presents a comparison of ATR-FTIR spectra for PE, PP, and PEPP. Thicker lines show the mean for a subset of spectra with this label. The paired thin lines around means indicate the standard deviation around mean intensity at that wavenumber. A set of key distinctions are evident in classes, where only PE lacks a detectable peak at approximately  $1375\text{cm}^{-1}$ , which is associated with the C-H<sub>3</sub> 'umbrella' vibrational modes. Conversely, PP exhibits at most a weak signal from C-H<sub>2</sub> rocking at approximately  $718\text{cm}^{-1}$ . PEPP, however, displays clear signals in both these regions. Although there are other differences, such as those observed around  $2950 - 3050\text{cm}^{-1}$ , these can be more subtle, and become unreliable when samples are degraded or unknown additives are used. The features highlighted in the figure are the most prominent, easily identifiable, and reliable, even amidst potential degradation effects or differences due to additives. This is evident from the low amount of variation around these specific signals.

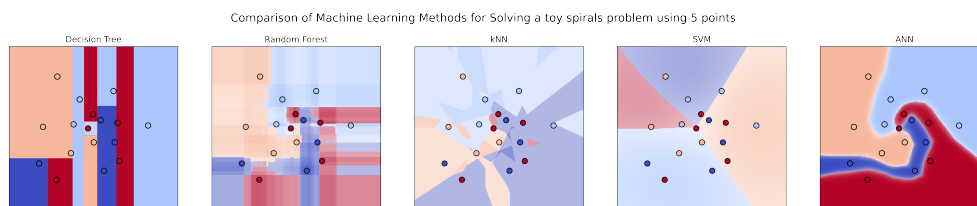
## 1.5 Decision Boundaries

The no free lunches theorems[1] support the intuition that any elevated performance of one model over one class of problems is offset by performance over another class of problems, and so we should consider the challenges a task presents and the relative suitability of a given machine learning model to that type of task.

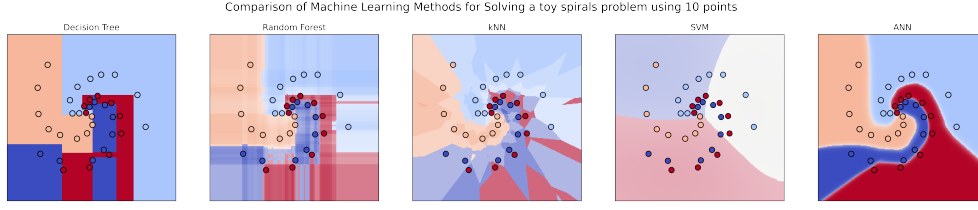
With sufficient data, poor choice of model can often be mitigated, for example when sampling a continuous space using 10 samples per class, performance of models that learn discrete spaces suffers greatly (figure 3). Taking the example of a classification tasks for 2D points, where classes are defined by which spiral arm they belong to (figures 2 to 5, we can see that for this type of problem discrete methods like decision trees or random forest, may learn boundaries that fail catastrophically, while models that handle spaces in a continuous way, while imprecise, learn realistic boundaries. For sparsely sampled spaces, like those represented by many of the poorly sampled classes in the microplastic FTIR problem, poor model choice can result in catastrophic failures similar to those shown in the figures 2 and 3. Catastrophic failure in this

context meaning decision boundaries are learned which may be entirely unrealistic rather than just imprecise.

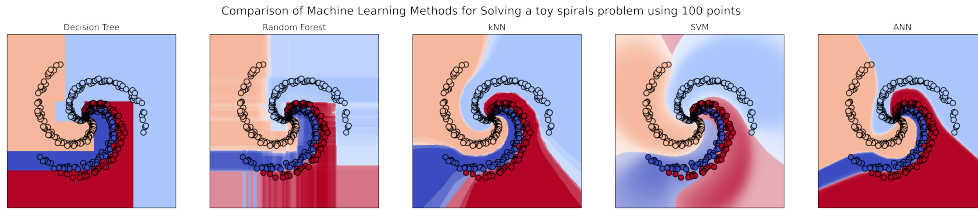
For this reason, choice of model is considered important, and special care should be given to the consideration of what type of decision boundaries are likely to exist in truth. If discrete boundaries are expected, a RF model is likely to be more appropriate while for continuous boundaries then SVM, given availability of an appropriate kernel, or an ANN might be a better choice. In case of FTIR data of microplastics, boundaries are not expected to be discrete neat polymers are frequently combined to marry beneficial traits of each polymer. This means samples may smoothly transition between any number of classes. This is important to consider, because blends of PE and PP (PEPP), is one of the most common sample types found in environmental samples. So a models tendency to learn continuous representations is therefore considered fundamental to reliably solving the problem of classifying environmental microplastic samples from their FTIR spectra, where data for specific samples is often sparse or incomplete.



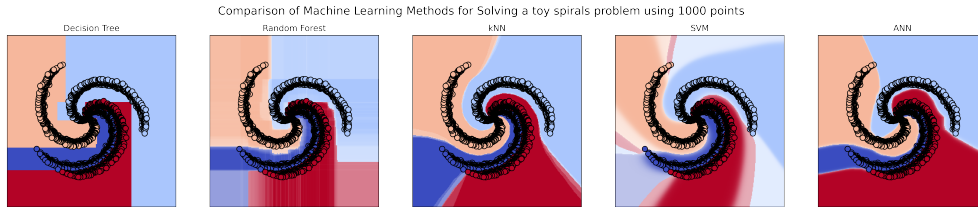
**Fig. 2** A comparison of decision boundaries as formed by different machine learning algorithms for a set of non-linearly separable classes. Shaded areas indicate what a point in that location would be predicted as, while colour intensity indicates the confidence with which it would be predicted. The data used was generated by creating interlocking spirals, where each point is uniformly distributed from 0 to 1 along the classes axis, where the class axis were spread unevenly around the origin, finally points were rotated around the origin relative to their magnitude. What this illustrates is how at extremely low sample counts (5 per class), an ANN can still learn a useful representation of this space, while other methods fail, typically in catastrophic fashion.



**Fig. 3** A comparison of decision boundaries as formed by different machine learning algorithms for a set of non-linearly separable classes. Shaded areas indicate what a point in that location would be predicted as, while colour intensity indicates the confidence with which it would be predicted. The data used was generated by creating interlocking spirals, where each point is uniformly distributed from 0 to 1 along the classes axis, where the class axis were spread unevenly around the origin, finally points were rotated around the origin relative to their magnitude. This shows that at relatively low sample counts (10 per class), this shows that most methods still perform catastrophically while ANN and SVM improve the accuracy of their boundaries



**Fig. 4** A comparison of decision boundaries as formed by different machine learning algorithms for a set of non-linearly separable classes. Shaded areas indicate what a point in that location would be predicted as, while colour intensity indicates the confidence with which it would be predicted.. The data used was generated by creating interlocking spirals, where each point is uniformly distributed from 0 to 1 along the classes axis, where the class axis were spread unevenly around the origin, finally points were rotated around the origin relative to their magnitude. At a reasonable sampling density for sample space, (100 per class), we can see that most methods learn reasonable, though not always great, decision boundaries.



**Fig. 5** A comparison of decision boundaries as formed by different machine learning algorithms for a set of non-linearly separable classes. Shaded areas indicate what a point in that location would be predicted as, while colour intensity indicates the confidence with which it would be predicted. The data used was generated by creating interlocking spirals, where each point is uniformly distributed from 0 to 1 along the classes axis, where the class axis were spread unevenly around the origin, finally points were rotated around the origin relative to their magnitude. For a relatively high sampling density (1000 per class), we can see that compared to 100 point sampling density, the additional samples have limited additional impact for improving the boundaries learned by methods such as random forest.

## 1.6 Per Dataset tSNE reduction of Predictions

This section showcases how errors in labelling are consistent across datasets, but also it showcases how certain plastic representations appear to have a geographic component. Taking PS space as an example, for some datasets, while the fringes appear well sampled, the central types appear to be absent, as in the Kedzierski set. This indicates a potential spatial or temporal component that is not captured adequately but just this dataset alone.



## 1.7 Disclosure Statement on Large Language Model Use

In the development of this manuscript, large language models (LLMs) were utilized to obtain feedback and suggestions for improvement. The LLMs employed include a series of models from openAI (gpt-3.5-turbo, and gpt-4)[2, 3] as well as a locally hosted model ("OpenLlama 13B" [4], and the quantized Mixtral-8x7B [5] variant "The-Bloke-Mixtral-SlimOrca-8x7B-GPTQ"). The text was provided in LaTeX format, and interactions with the models were facilitated through Python command line interface. Some local models were fine tuned using LoRA[6] on personal information, including project documents, experimental data, drafts, and other relevant texts. The following system prompt was used

Hey there! We're currently working on an exciting project focused on creating a comprehensive LaTeX syntax document. This document aims to cover the application of machine learning techniques in the analysis of microplastic FTIR data. As a Language Model, you are here to assist in exploring and evaluating various aspects of this project. You are encouraged to take the role of a critical and hard to please co-author. You are encouraged to critically evaluate your own replies, question design choices in work shared with you, or ask questions for clarification in decisions made. If questions or tasks are unclear, then ignore the initial prompt and ask clarifying questions first. Let's dive in with the first task!

Models were used to get feedback on draft material, or ease the burden of laborious but non critical tasks, such as converting a markdown formatted table to a latex formatted table. For example, in the methods section, a LLM was asked what questions someone might ask, if they were seeking to reproduce work, that are not already answered by in the document. The section was then updated such that answers to sensible questions the LLM's came up with were included.

It is important to note that not all feedback or suggestions generated by the LLM were implemented. The insights provided by the models were treated with the weight and credibility one might attribute to an individual who is well read and has proficient language skills, but holds limited comprehension of the knowledge they possess. Nonetheless, these tools proved valuable in refining the initial draft of the manuscript.

Another example, would be the use case of providing the following prompt

try write this out properly, expanding the acronyms "The aggregate data had a total of 16 classes after processing, where a Pareto plot (Figure ??) of the aggregate dataset highlights the significant imbalance in class representation. The final classes were PE, PP, PEPP, PS, EVA, PA, CA, PET, PVC, AF, PC, PU, Pest, PCL, NP, PMMA"

The output of which was mostly correct for non fine-tuned models, where for example AF being incorrectly, though reasonably, translated to "Aramid Fibre" or with Pest being predicted as Pesticides. The output was then edited to improve continuity and proof read for mistakes. Out of curiosity, the non tuned LLM's were asked "are there any of these acronyms where you are not certain of your guess?" the answer provided being:

I am confident about most of the acronyms, but "Pest" and "AF" are a bit uncertain. "Pest" is typically short for "Pesticides", but in the context of data classes, it could potentially stand for something else. Similarly, "AF" is often used for "Aramid Fiber", but it could also represent different terms depending on the specific context of your research. If



these acronyms have specific meanings in your field of study, I would recommend using those instead.

So, interestingly, the incorrect expansions were of 'low confidence', even before any corrections were made, or were indicated.

The final example prompt is as follows.

{python dictionary with model parameters} please re-write this code using plain language, using language suitable to an academic journal article. It should give the name of the parameter used and the value used for that parameter. "

The output was then edited to improve clarity, flow, and to fix syntactic errors.

The full extent to which LLM's were used in the creation of this document can be summarised as augmenting the proof reading of early drafts by providing feedback on text clarity and brevity, providing feedback on potential gaps in information provided, and alleviating workload on repetitive monotonous tasks like transcribing parameters used in code to plain English. LLM's proved valuable in the creation of this document, saving many hours of work. Of particularly noteworthy utility were open models, and the ability to readily fine tune these using tricks such as Low-Rank Adaptation.[6]

## 1.8 Label Consolidation rules

Labelling conventions varied significantly across different datasets. Rules used in this work are as follows

The general rule applied is that labels should be mapped to the same semantic representation, however there were two instances where significant alterations were made to the original label. Specifically, in the Kedzierski dataset, data labelled as 'morphotype1' and 'morphotype2' were relabelled as 'PEPP'. The Kedzierski group originally identified these as anthropogenic polymer samples, however their true identity was uncertain at the time of publication. We argue that there is evidence for these being PEPP, in part due to the relative abundance of the 'morphotypes' and that of PEPP across datasets. This is further supported by reference spectra drawn from Turku et al.'s work[7] which provides examples of PEPP spectra that closely match the averages for each of the morphotypes.

Additionally, some classes were ignored. For example, the Brignac dataset introduced a class labelled 'Phthalate', which represented spectra where the polymer signal was deemed to be significantly masked by phthalate additives. However, this class was not included in training or evaluation of the final model due its potential class heterogeneity. Classes labelled 'Unknown', or were under-represented ( $N \leq 4$ ) such as ABS were also ignored for training.

| Original                  | Relabel | Ignore |
|---------------------------|---------|--------|
| Unknown                   |         | true   |
| Non-plastic               |         | true   |
| N-MP                      |         | true   |
| Petroleum wax             |         | true   |
| Unidentifiable            |         | true   |
| Unidentifiable            |         | true   |
| ABS                       |         | true   |
| latex                     |         | true   |
| Latex                     |         | true   |
| noInfo                    |         | true   |
| Cellulose                 |         | true   |
| phthalate                 |         | true   |
| Phthalate                 |         | true   |
| HDPE                      | PE      |        |
| LDPE                      | PE      |        |
| LLDPE                     | PE      |        |
| Other PE                  | PE      |        |
| dPE                       | PE      |        |
| d-PE                      | PE      |        |
| PE_c                      | PE      |        |
| PE_d                      | PE      |        |
| PE_f                      | PE      |        |
| Unknown PE                | PE      |        |
| Unknown PE                | PE      |        |
| PP/PET mix                | PEPP    |        |
| Mixture                   | PEPP    |        |
| PP_w_HDPE                 | PEPP    |        |
| MT1                       | PEPP    |        |
| MT2                       | PEPP    |        |
| PE/PP mix                 | PEPP    |        |
| PP_d                      | PP      |        |
| EPS                       | PS      |        |
| PS                        | PS      |        |
| PEst                      | PEstr   |        |
| PETE                      | PET     |        |
| Poly(vinylidene fluoride) | PVC     |        |
| Cellulose Triacetate      | CA      |        |
| PP/PMMA mix               | PMMA    |        |
| "Nylon"                   | PA      |        |
| "Nylon "                  | PA      |        |
| EVA wax                   | EVA     |        |
| PEVA                      | EVA     |        |

## 1.9 ATR-FTIR spectral intensities

The reflectance  $R$  of a singal after  $N$  reflections is given by

$$R(N) = \frac{I}{I_0} = (1 - \alpha d_e)^N \quad (7)$$

Where  $R(N)$  is the reflectance after  $N$  reflections,  $I$  is beam intensity and  $I_0$  is intiial beam intensity,  $\alpha$  is the absorbtion coefficient and  $d_e$  is the effective depth of penetration. Here the depth of penetration can be calculated by

$$d_p = \frac{\lambda}{2\pi (n_1^2 \sin^2 \theta - n_2^2)^{\frac{1}{2}}} \quad (8)$$

Which describes the depth of penetration ( $d_p$ ) for the evanescent wave, which depends on the wavelength ( $\lambda$ ), beam incident angle ( $\theta$ ), and refractive indices of both the internal reflection element ( $n_1$ ) and the sample ( $n_2$ )[8]. Among these variables, only wavelength is expected to vary throughout spectrum collection. Consequently, variations in signal intensity arising from effective relative abundance of a bond relies solely on wavelength and a constant factor.

This would pose a problem if a sample is not sufficiently thick and homogeneous in composition. When a sample is too thin, the effects are obvious in the collected data where this presents as an upward curving baseline with a slope inversely proportional to wavelength, however inhomogeneity would not have immediately obvious impacts.

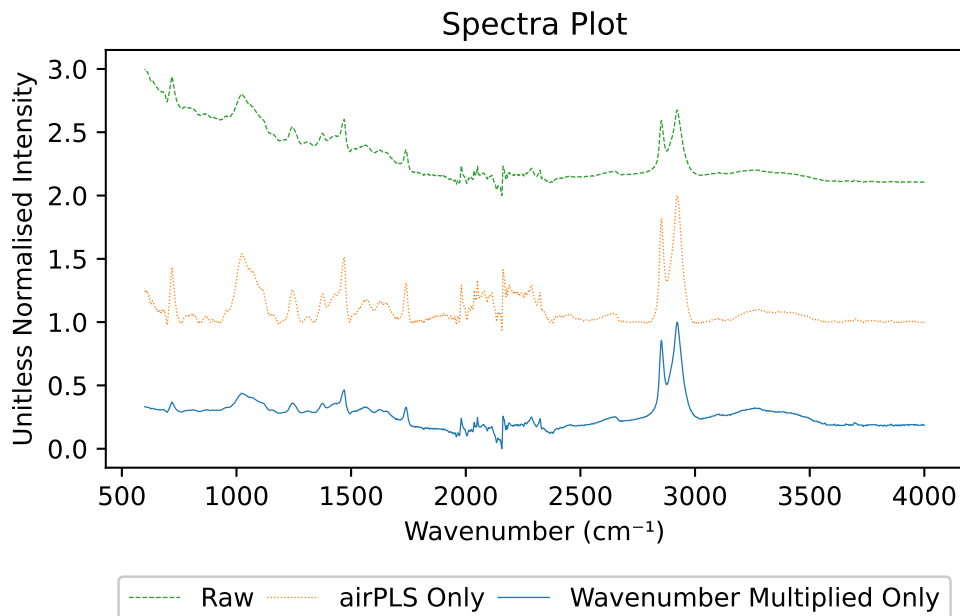
## 1.10 No Baseline Correction

Baseline correction is recommended for FTIR spectra,[9–11], however these algorithms carry the assumption that samples are sufficiently thick and have a homogeneous composition. If we consider the equation for calculating penetration depth for ATR-FTIR (refer (8)), then under typical conditions the penetration depth, and thus intensity of the signal, is proportional to the wavelength of the beam. Therefore, given the nature of MP data, it makes safe use of such corrections difficult for two reasons.

The first reason is that sample thickness can be thin enough that over-penetration will easily occur at lower wave-numbers. The second is that sample composition is not homogeneous, as such surface level composition will be over-represented at higher wave-numbers while at lower wave-numbers bulk composition is better represented. This is because degradation effects often lead to significant differences between surface and bulk composition, where surface degradation levels are also non-uniform across the surface.[12, 13] Therefore in samples with in-homogeneous composition where surface and bulk composition differ, this penetration effect could introduce a significant de-trend in relative signal intensities, independent of trends impacting the baseline, especially at lower wave-numbers. This is problematic, as if this is the case then algorithms like airPLS could exaggerate erroneous representations while removing clues of these errors existing. For this reason detrend effects, while an annoyance, may be too important and informative to remove.

These suspicions were investigated by removing effects due to wavelength by multiplying intensity by the reciprocal of the corresponding wavelength, the wave-numbers, before the normalisation step. This was compared against airPLS and a combined approach (wave-number multiplication first then airPLS). Results for a thin film sample can be found in figure 7, which shows how airPLS on it's own creates significant non-existent information. This is viewed as evidence supporting the position of using no baseline correction. While for some methods applying wave-numbers based correction might be beneficial, this is a transformation, which if even required, is readily learned by neural networks. Therefore, despite it being recommended practice - no baseline correction was used in the model presented in this work. However, if one wishes to employ baseline correction, then our results, which are partially reflected

in figure 7, indicate it may be better to simply multiply a deviant spectrum's intensities by the corresponding wavenumbers as the sole method of correcting baseline de-trends, as this produces more realistic results.



**Fig. 7** This highlights how airPLS might exaggerate some features that may not really exist, thus introducing noise to the dataset. If baseline de-trends are observed in ATR-FTIR data, it may be better to address these directly via their likely cause, which is over-penetration. This over-penetration de-trends the baseline as a function of wavelength, so multiplying by the reciprocal, the wave-number, this corrects the de-trending effect.

### 1.11 MinMax Normalisation vs. Z-Score Standardisation

MinMax normalisation arises as a natural choice in this study when one seeks to address non-sample originating variance in a sample's spectrum. In the context of ATR-FTIR the measured intensity, which correlates to the relative abundance of specific bond behaviours, is perfectly accurate. Inaccuracies arise from experimental, human and machine error. These uncontrolled acquisition variables introduce bias in scale and location, which is uniformly applied throughout the spectrum. This bias arises from sampling conditions including, but not limited to, variation in ambient lighting can shift the baseline and influence the signal to noise ratio.

In this study, we use MinMax normalisation to address the variance in a sample's spectrum that does not originate from the sample itself. This is particularly relevant in ATR-FTIR, where the measured intensity accurately reflects the relative abundance

of specific bond behaviours. However, inaccuracies can occur due to experimental, human, and machine errors. These uncontrolled variables introduce a bias in scale and location, uniformly applied throughout the spectrum. This bias can be influenced by various sampling conditions, including changes in ambient lighting.

To counter these issues, we first set the baseline to zero by subtracting the sample minimum. Next, determining the scaling factor, which arises from changes in the fraction of the IR beam exposed to the sample, is more challenging. Usually, reference samples with known quantities are used to address these issues during quantitative analysis. However, these references must be created at the time of acquisition, and this information is not available in public datasets. Therefore, we chose to scale by the maximum value as a practical solution.

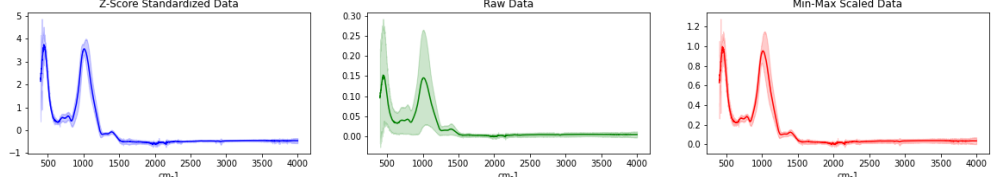
While this method simply shifts the bias rather than removing it, it improves precision as all samples of the same type will have the same maximum and therefore the same bias. This does not allow for quantitative comparisons, but it does enable comparison of relative abundance. This is considered reliable as the maximum for a class remains consistent over a broad range of impacts, including degradation effects. Therefore, small changes can be compared more readily, and can be compared indirectly to changes across classes, where changes are not relative to each position but rather relative to each class maximum.

This method has limitations. For example, changes in ambient lighting can alter the volume under the curve, causing some volume that should be under peaks to be under the baseline instead. This means scaling by min-max can make fluctuations between neighbouring values more extreme, proportional to the impact of unaccounted ambient lighting. However, if acquisition method follows recommendations, typically enforced by commercial software, then in practice this is not typically an issue since baselines are frequently recollected to account for shifts in background throughout an experiment.

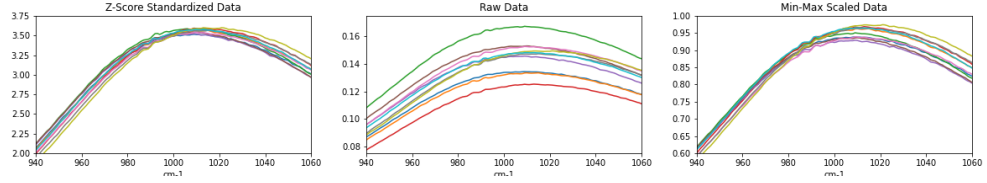
While z-score standardisation or similar is often considered better practice for an unknown problem, we propose that using our knowledge of how a spectrum is produced to guide our scaling approach can improve the informativeness of a resulting spectrum. For this problem, the Min-Max approach appears to be the natural choice for mitigating non-sample variation in a spectrum.

The following image shows a series of spectral averages (bold line) and variance (band around line) and showcases how Min-max scaling retains more information than z-score scaling would. Variance has been scaled up by a factor of 10 to highlight differences in behaviour between min-max and z-score methods.

The variance across samples should be very low, with the exception that the changes in mechanical pressure will strain the structure, causing shifts in some bonds, which should be seen in the Si-O-Si stretching vibrations at  $1000\text{cm}^{-1}$ , manifesting as a skew in this peak range.<sup>[14]</sup> We can see that there is no clear evidence of this trend in the raw data, or the z-score standardised data, but it is clear in the min-max scaled data. Zooming into the relevant region we can see how for Min-Max the left hand side of peaks remains relatively consistent across spectra, while the right hand side captures the change in skew for different pressures accurately.



**Fig. 8** The spectra shown are average (line) and 10x variance (band) for z-score standardised (left) and MinMax normalised (Right) spectra. Both scaling methods show a significant reduction in variance compared to the raw data (centre) however, only MinMax accurately captures subtle trends in the data. The spectra are of the same glass wool filter, each spectrum captured at the same location, varying only the contact pressure at which samples are acquired.



**Fig. 9** The z-score standardised (left) and Min-Max normalised (Right) spectra both show a lot less non-informative variance than the raw data (centre) however, how remaining variance is captured differs across the scaling methods, where min-max captures the appropriate expected trend.

## 1.12 Utility of Onehot Encoding

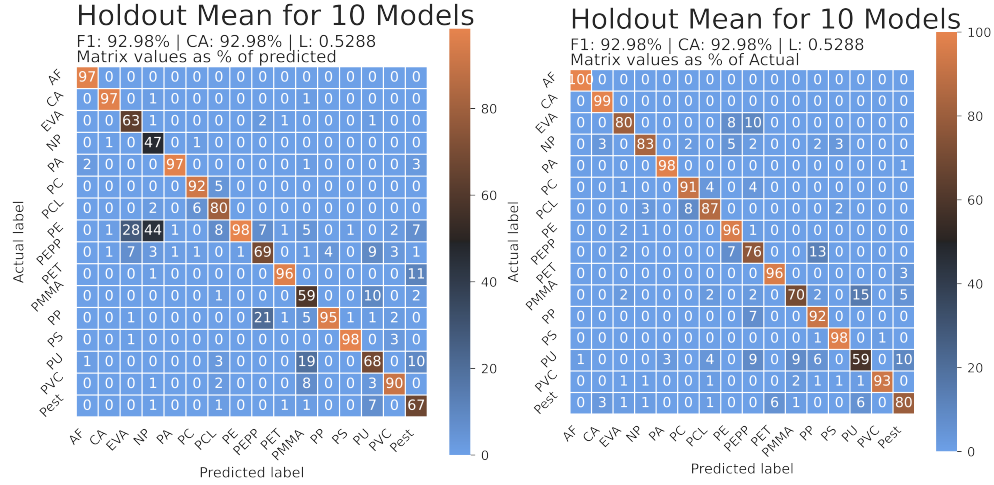
One-hot encoding is important to this problem as some classes will blend into one another, such as in the case of neat polymers and their blends. This is a problem as there is no set standard for interpretation. For this reason we expect tasking a model with learning to confidently predict only a single class for a sample to be potentially detrimental to model performance. By allowing a model to predict the relative probability of each class, even if expected to predict only one class, this allows the model to better handle the uncertainty arising from discord in labelling practices for neat and blended cases accross datasets. If the model aims to optimize categorical cross entropy for softmax predicitions, given by

$$\text{CCE} = - \sum_i^C t_i \log p_i \quad (9)$$

where  $C$  is the number of classes,  $t_i$  is the ground truth and  $P_i$  is the prediction, we can interpret this as describing the distance of a prediction's distribution from the ground truth distribution. From this it is clear that error equally distributed among all non target classes would yield the lowest loss, and so for predictions where error gathers in a specific class, it indicates the network sees noteworthy similarities between samples. Provided we do not over-train the network, this approach should give predictions that are crude, unintentional, yet informative fractional encodings of a sample.

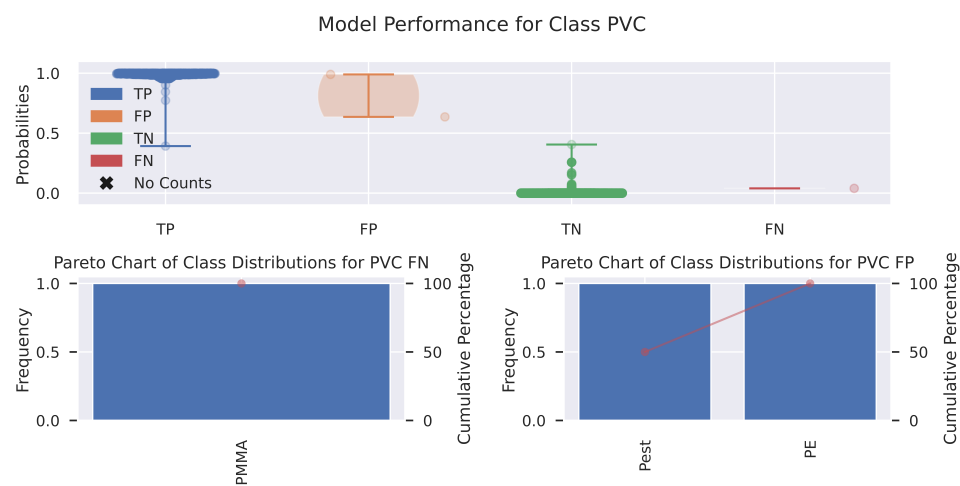
### 1.13 Prediction Behaviour Split by Class

A paired set of confusion matrices can give good superficial insight into model behaviour, these can be found in Figure 10, and highlight how predictions behave for both false positives and false negatives. The remaining section gives more in-depth explorations of the same information for interesting classes.

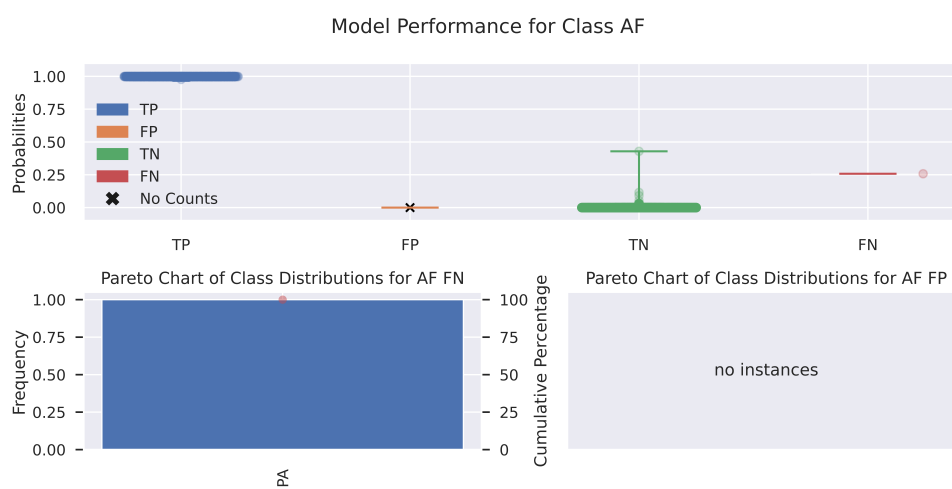


**Fig. 10** presents two confusion matrices (CMs) that illustrate the predictive performance across classes, with rows indicating the true labels and columns indicating the predicted labels. Values along the diagonal in both matrices represent correct predictions (true positives). The left matrix is normalized by row totals, which allows for an interpretation of predictive accuracy within each class. Specifically, a value in a cell shows the percentage of instances of a given class that are predicted as each possible class, highlighting the sensitivity or recall of the predictions. The right matrix is normalized by column totals, focusing on the precision of the predictions. Here, a cell's value indicates the likelihood that a prediction for a class corresponds to that actual class, addressing the question, 'When a class is predicted, how often is that prediction correct?' Off-diagonal values in this matrix reveal the proportion of false positives for each predicted class. Both matrices offer complementary insights: the left on the model's ability to identify all instances of a class (recall), and the right on the confidence in the model's predictions (precision).

These graphs show how confusion between classes occurs, reading top to bottom, left to right, the first panel shows the density for predictions, grouped by (true / false) and (positive / negative) combinations. The subsequent panels show prediction frequencies in terms of where false negative predictions end up, and where false positive predictions end up. In Figure 20 we can see prediction frequencies for PEPP that show a uniform density for false positives, indicating misclassification are unlikely to resolve with further training. Looking at what classes are chief among these confusions, we can see that PE and PP are the common culprits, which is reasonable given PEPP is a combination of these



**Fig. 11** PVC Prediction Density

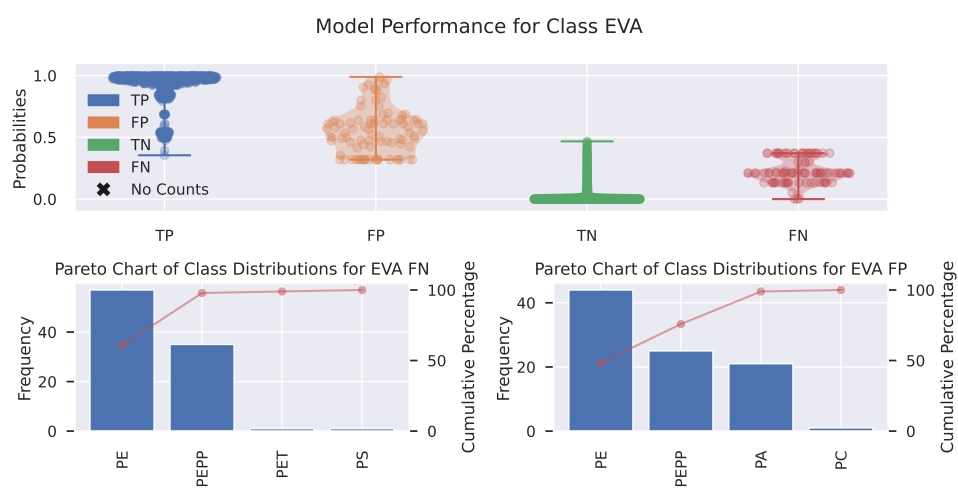


**Fig. 12** AF Prediction Density





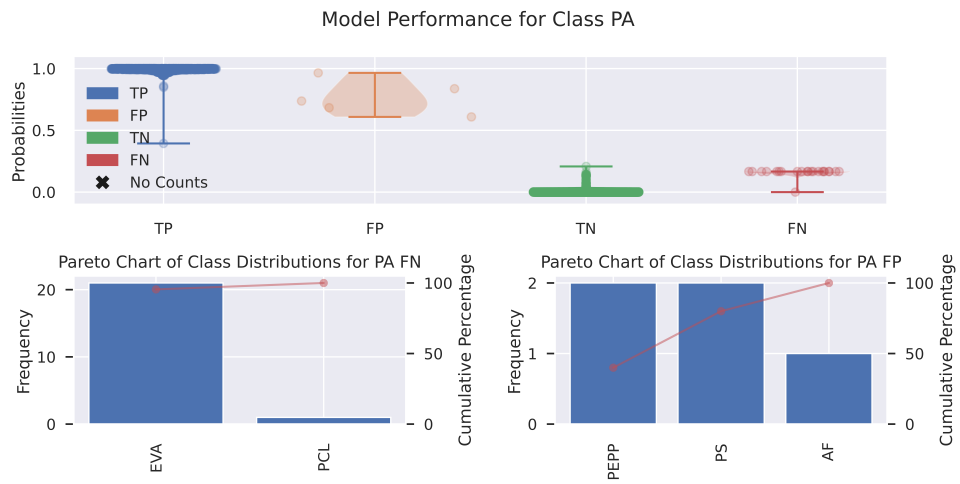
**Fig. 13** CA Prediction Density



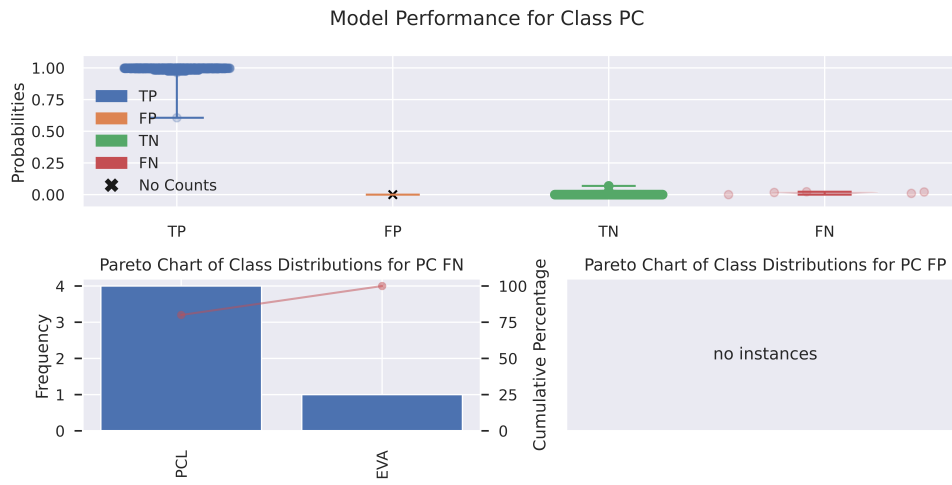
**Fig. 14** EVA Prediction Density



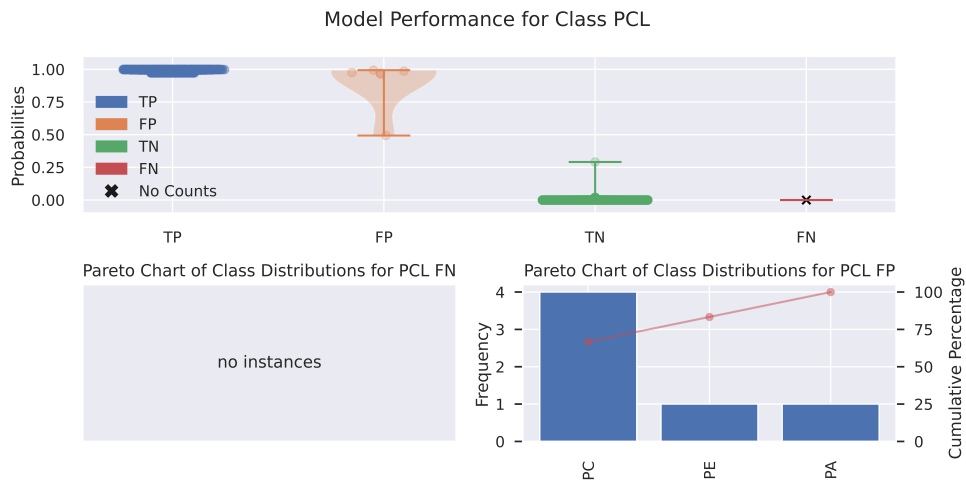
**Fig. 15** NP Prediction Density



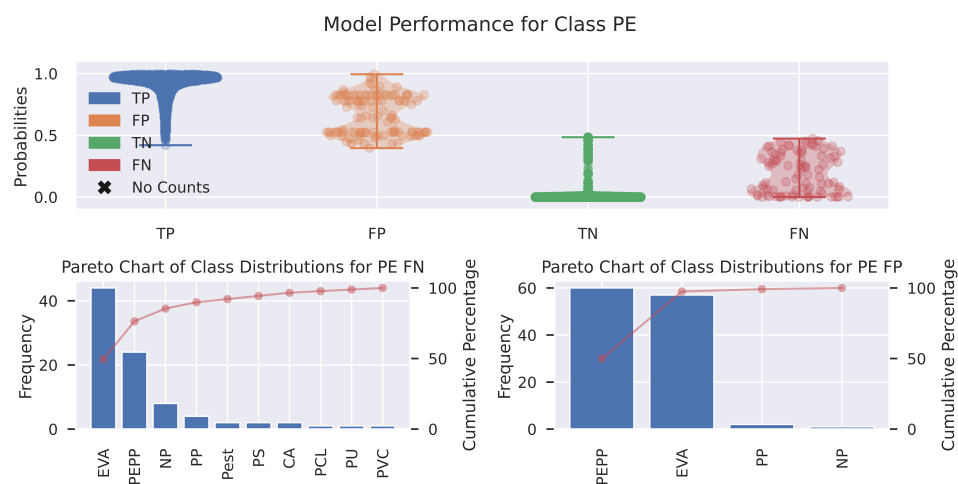
**Fig. 16** PA Prediction Density



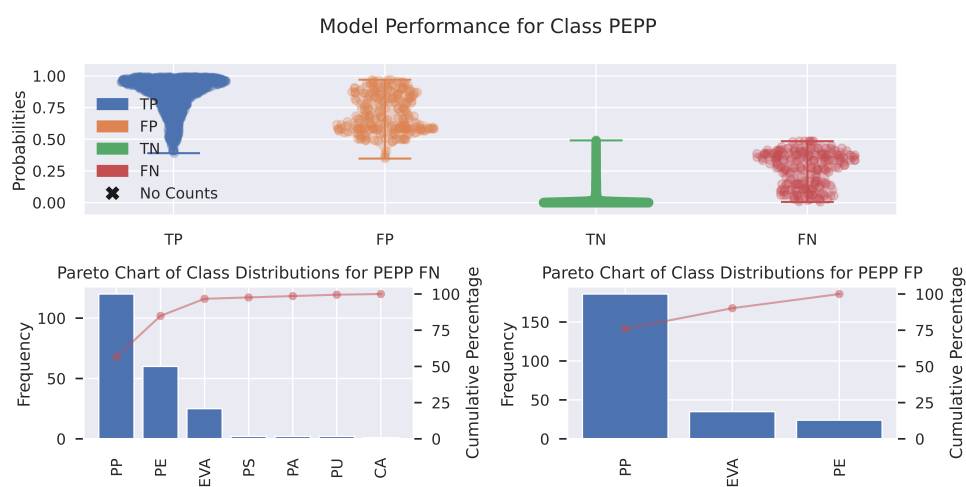
**Fig. 17** PC Prediction Density



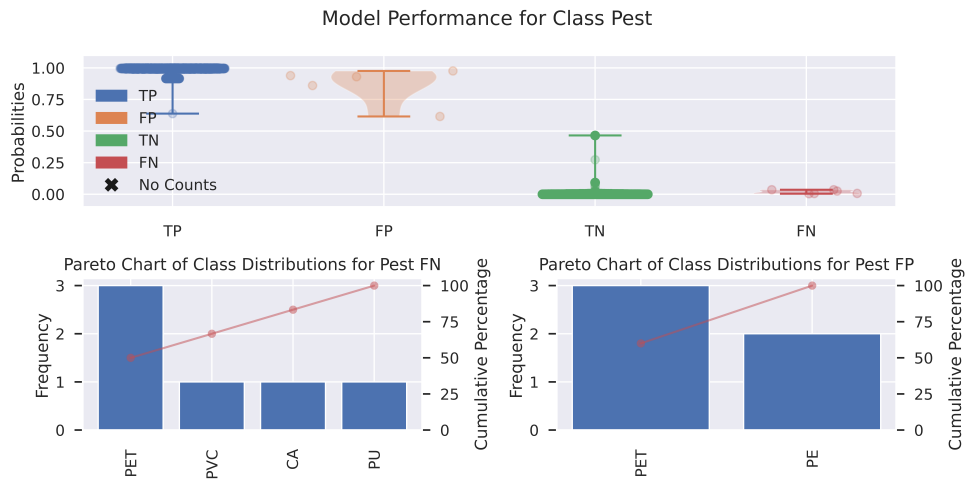
**Fig. 18** PCL Prediction Density



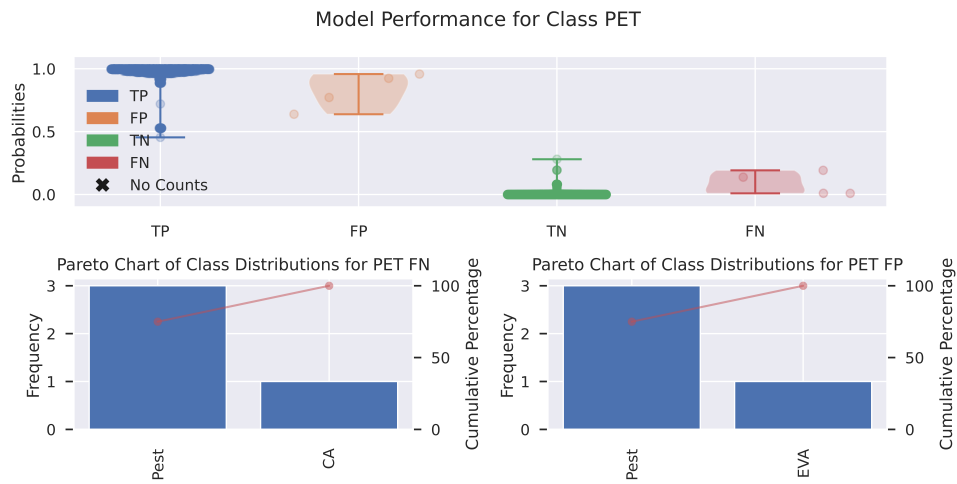
**Fig. 19** PE Prediction Density



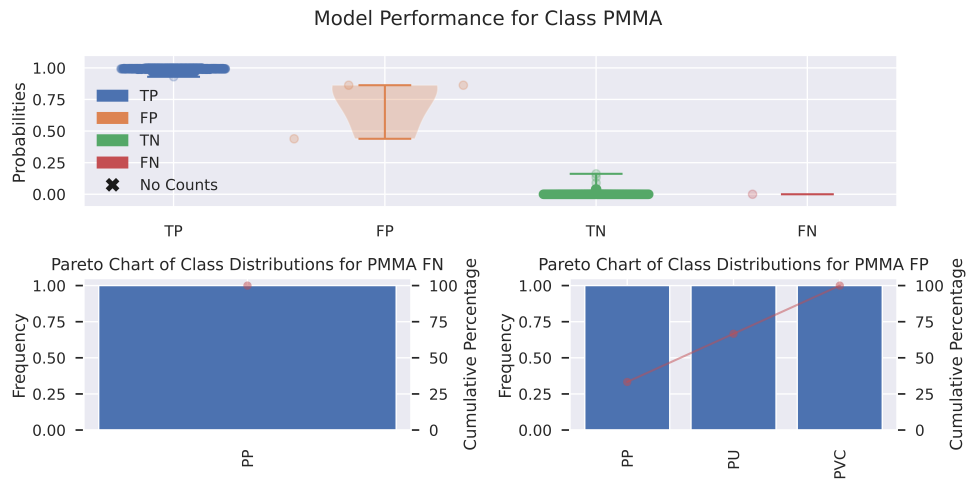
**Fig. 20** PEPP Prediction Density



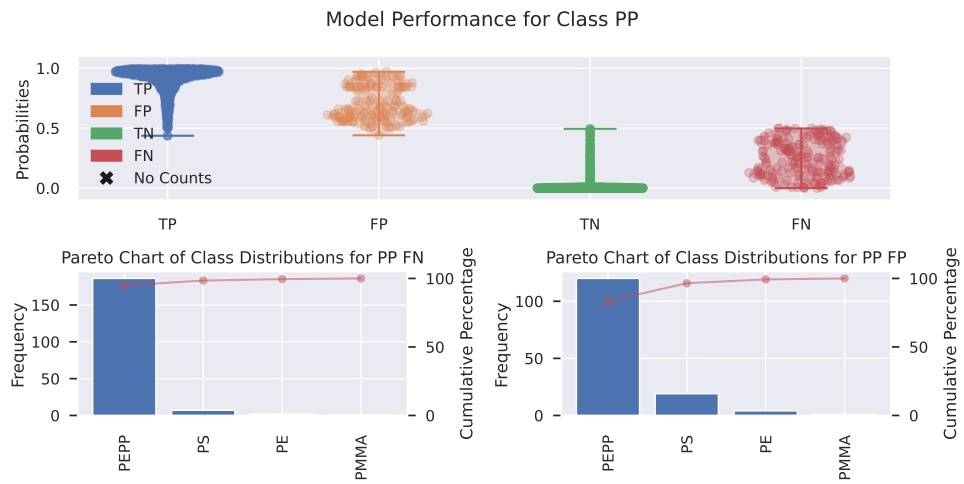
**Fig. 21** Pest Prediction Density



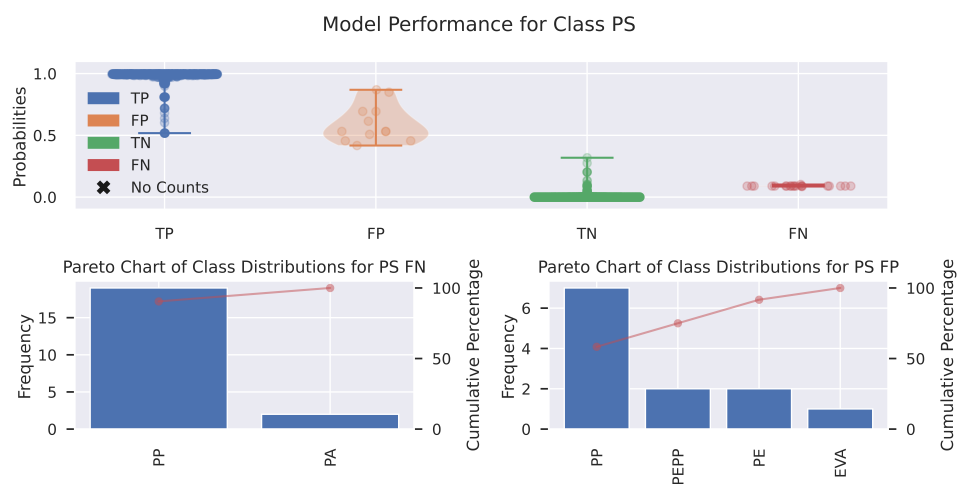
**Fig. 22** PET Prediction Density



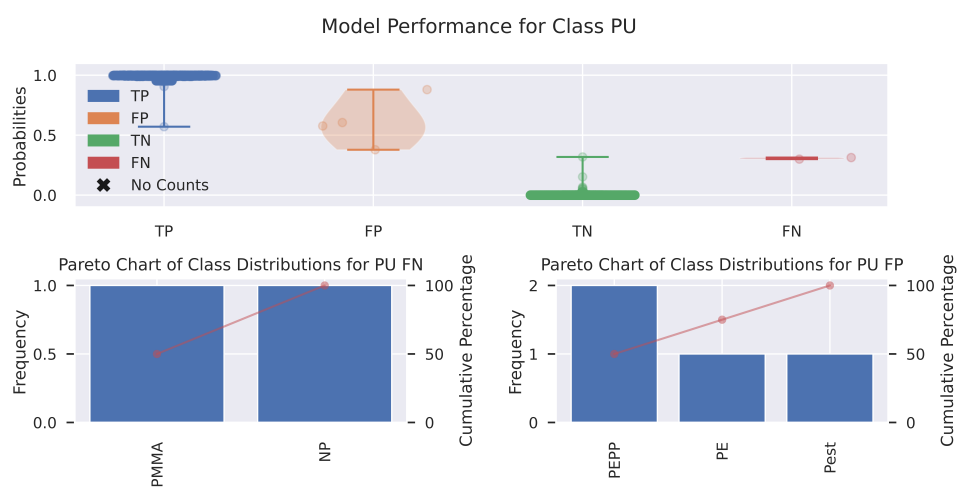
**Fig. 23** PMMA Prediction Density



**Fig. 24** PP Prediction Density



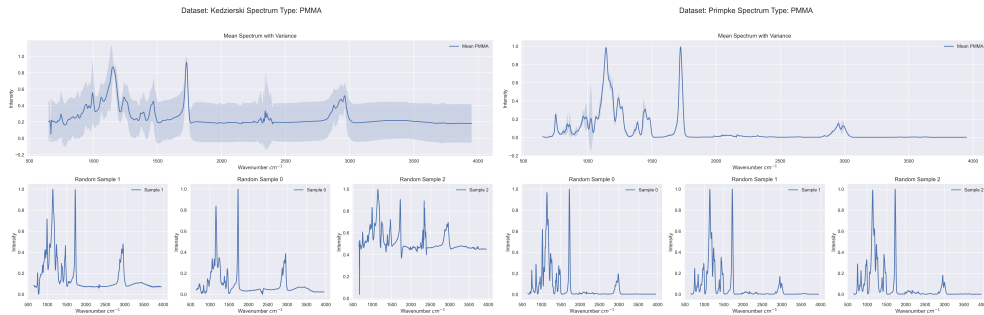
**Fig. 25** PS Prediction Density



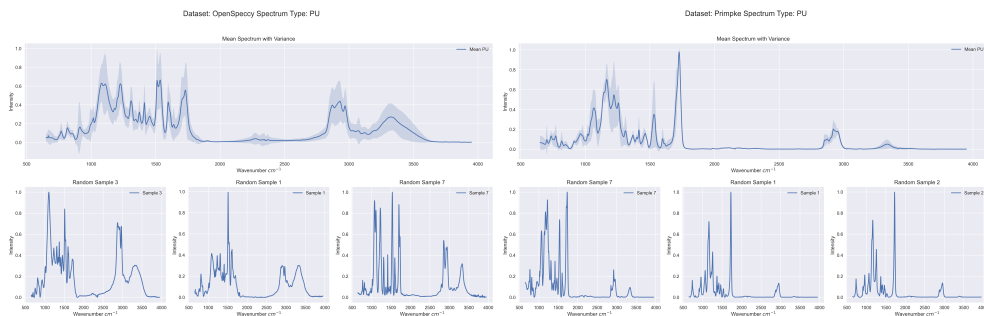
**Fig. 26** PU Prediction Density

## 1.14 Exploring Spectra in Data

When exploring the representation of certain spectra in datasets, poor classification accuracy for these becomes a little more apparent. In the case of PMMA as captured in the Kedzierski set (SI 27), we can see that there are some significant issues with how this data has been captured, reasonably leading to misclassification.



**Fig. 27** presents PMMA spectra averages and variance, as well as 3 random subsamples for the datasets Kedzierski (left) and Primpke (right)



**Fig. 28** presents PU spectra averages and variance, as well as 3 random subsamples for the datasets OpenSpecy (left) and Primpke (right)

## References

- [1] Wolpert, D.H., Macready, W.G.: No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1(1), 67–82 (1997) <https://doi.org/10.1109/4235.585893>
- [2] OpenAI: GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774v3> (2023)
- [3] OpenAI: Introducing ChatGPT. <https://openai.com/blog/chatgpt>



- [4] Geng, X., Liu, H.: OpenLLaMA: An Open Reproduction of LLaMA. OpenLM Research (2023)
- [5] Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of Experts. arXiv (2024). <https://doi.org/10.48550/arXiv.2401.04088>
- [6] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. arXiv (2021). <https://doi.org/10.48550/arXiv.2106.09685>
- [7] Turku, I., Kärki, T., Rinne, K., Puurtinen, A.: Characterization of plastic blends made from mixed plastics waste of different sources. <https://doi.org/10.1177/0734242X16678066> **35**(2), 200–206 (2016) <https://doi.org/10.1177/0734242X16678066>
- [8] Suëtaka, W., Yates, J.T.: Internal Reflection Spectroscopy. Surface Infrared and Raman Spectroscopy, 117–161 (1995) [https://doi.org/10.1007/978-1-4899-0942-8\\_3](https://doi.org/10.1007/978-1-4899-0942-8_3)
- [9] Renner, G., Nellessen, A., Schwiers, A., Wenzel, M., Schmidt, T.C., Schram, J.: Data preprocessing & evaluation used in the microplastics identification process: A critical review & practical guide. *TrAC Trends in Analytical Chemistry* **111**, 229–238 (2019) <https://doi.org/10.1016/J.TRAC.2018.12.004>
- [10] Primpke, S., Lorenz, C., Rascher-Friesenhausen, R., Gerdt, G.: An automated approach for microplastics analysis using focal plane array (FPA) FTIR microscopy and image analysis † (2017) <https://doi.org/10.1039/c6ay02476a>
- [11] Yan, X., Cao, Z., Murphy, A., Qiao, Y.: An ensemble machine learning method for microplastics identification with FTIR spectrum. *Journal of Environmental Chemical Engineering* **10**(4), 108130 (2022) <https://doi.org/10.1016/J.JECE.2022.108130>
- [12] Luo, H., Xiang, Y., Zhao, Y., Li, Y., Pan, X.: Nanoscale infrared, thermal and mechanical properties of aged microplastics revealed by an atomic force microscopy coupled with infrared spectroscopy (AFM-IR) technique. *Science of The Total Environment* **744**, 140944 (2020) <https://doi.org/10.1016/j.scitotenv.2020.140944>
- [13] Webb, H.K., Arnott, J., Crawford, R.J., Ivanova, E.P.: Plastic Degradation and Its Environmental Implications with Special Reference to Poly(ethylene terephthalate). *Polymers* **5**(1), 1–18 (2013) <https://doi.org/10.3390/polym5010001>

- [14] Kaya, H., Ngo, D., Gin, S., Kim, S.H.: Spectral changes in Si–O–Si stretching band of porous glass network upon ingress of water. *Journal of Non-Crystalline Solids* **527**, 119722 (2020) <https://doi.org/10.1016/j.jnoncrysol.2019.119722>