Analyst

PAPER

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2026, **151**, 2552

Classification of recycled plastics using sparse and imbalanced spectral data and data augmentation by the generative adversarial network

Xuan Liu,†[a] Xuerui Song,†[a] Yusuf Sulub,[b] David Zoller,[c] Zhenyu (James) Kong[a] and Blake N. Johnson *[a,d,e]

Accurate identification of post-consumer plastics is essential to establishing high-performance recycling processes and enabling a circular and sustainable economy and environment through effective recycling and remanufacturing. However, Fourier transform infrared (FTIR) spectra of recycled materials often exhibit noise, baseline shifts, and overlapping signatures from additives or contaminants, resulting in datasets that are both sparse and severely imbalanced. This data complexity, sparsity, and class imbalance can degrade conventional machine-learning classifiers, resulting in higher rates of misclassifying plastics. To address these challenges, we investigated if data augmentation using generative adversarial networks could enhance polymer classification performance. We implemented a Generative Adversarial Network (GAN) framework that integrates adversarial training with a classifier-guided feedback loop to synthesize realistic, class-discriminative FTIR spectra for six commonly recycled polymers, polyethylene (PE), polypropylene (PP), polystyrene (PS), polycarbonate (PC), polyethylene terephthalate (PET), and acrylonitrile butadiene styrene (ABS), and trained multilayer perceptron classifiers on datasets with varying ratios of synthetic data. The optimal balanced accuracy of 96.2% was achieved when synthetic spectra accounted for 50% of the training set, whereas including more than 90% synthetic data degraded generalization. Synthetic data augmentation using a GAN with the optimal augmentation ratio improved ABS classification accuracy, precision, and recall by 43%, 50%, and 33%, respectively, compared with no augmentation and replicate experimental measurements. These results demonstrate that GAN-based data augmentation can effectively mitigate data sparsity and class imbalance in spectral classification of common plastics, providing a practical foundation for creating robust online polymer classification systems.

Received 29th September 2025,
Accepted 10th March 2026

DOI: 10.1039/d5an01042j

rsc.li/analyst

1. Introduction

The concept of a circular economy has gained prominence in several industries, particularly in the plastics industry, as companies and consumers increasingly seek sustainable materials (*e.g.*, recyclable or renewable) and manufacturing processes and systems that minimize waste and environmental impact.[1,2] A circular production and economic model emphasizes the continual use of process outputs, such as previous products and waste (*e.g.*, scrap), as 'new' process inputs by recycling, remanufacturing, and material recovery (*e.g.*, waste and products at the end-of-life cycle), in contrast to the traditional approach of "take, make, dispose".[3] Within this framework, accurately predicting the composition of a recovered material to be reprocessed, such as recycled plastic,[4] is a critical pre-processing step that affects the performance and quality of both remanufacturing processes (*e.g.*, chemical or bioprocesses) and the resultant product. According to a report published by the U.S. Environmental Protection Agency (EPA) in 2018, a substantial 75.6% of plastic waste was disposed of in landfills, with only 8.7% being recycled and 15.8% incinerated for energy recovery.[5] This value is concerning, as low recycling rates may be exacerbating ecological and environmental challenges. In an effort to address these pressing issues, researchers are exploring innovative methods using mixed plastic waste as feedstocks, such as pyrolysis and mechanical recycling.[6,7] The effectiveness of these emerging remanufacturing methods relies heavily on understanding the composition of these feedstocks (*i.e.*, mixed recycled plastic). Thus, reliable and high-throughput characterization methods

[a]*Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061, USA. E-mail: bnj@vt.edu; Fax: +(540) 231-3322; Tel: +(540) 231-0755*
[b]*SABIC Technology & Innovation, Houston, TX 77042, USA*
[c]*SABIC Technology & Innovation, Mt. Vernon, IN, 47620, USA*
[d]*Department of Materials Science and Engineering, Virginia Tech, Blacksburg, VA 24061, USA*
[e]*Department of Chemical Engineering, Virginia Tech, Blacksburg, VA 24061, USA*
† Equally-contributing authors.

are essential for enabling the effective deployment of these recycling approaches and for supporting the broader transition to a circular plastics economy.[8] Quality-controlled manufacturing with recycled materials requires rapid, accurate, high-throughput methods for input material characterization and sorting to optimize production and minimize process downtime.[9,10] In contrast, incorrect identification of plastic composition (*e.g.*, misclassification) can affect the quality of resultant products and increase production cost, undermining the overall effectiveness of circular manufacturing systems based on recycled materials.[11]

The shift toward a circular plastics economy has created an urgent need for fast, accurate online polymer classification systems. Misclassification can contaminate recycling streams, lower material quality, and reduce resale value—directly threatening the efficiency and profitability of recycling operations. For example, misclassifying common plastics such as polyethylene terephthalate (PET), polystyrene (PS), polycarbonate (PC), or acrylonitrile butadiene styrene (ABS) as polyethylene (PE) or polypropylene (PP), which are the most widely used plastics, can introduce serious contamination into the recycling stream, resulting in poor product quality, reduced process efficiency, and increased operational burden. For example, PET and PC inclusion in the recycle stream of polyolefins will lead to incompatibilities due to differences in melting temperature. At PE/PP processing temperatures, PC & PET remain solid or partially melted, leading to poor product performance.[12,13] ABS and PS contribute aromatic and nitrogen-containing compounds that can foul reactors or poison catalysts in pyrolysis and cracking processes used for chemical recycling.[14] These misclassifications not only reduce the yield and market value of the recycled output but also raise operational costs through increased downtime and the need for additional purification. Ensuring high model accuracy, particularly minimizing false positives for PE and PP, is therefore essential to support high-throughput, economically viable recycling workflows. Conventionally, spectroscopic techniques, including Fourier-transform infrared (FTIR),[15] Raman spectroscopy,[16] and nuclear magnetic resonance (NMR)[17] spectroscopy, have been extensively used for off-line characterization[18] of polymer composition based on their unique spectral data features.[19] However, the characterization and sorting of recycled materials based on spectral data is challenging due to potential changes in spectral characteristics relative to virgin materials, including alterations in FTIR peak intensities, band broadening, baseline distortion, and the emergence of additional absorption features induced by polymer aging and degradation, as well as the presence of additives, surface contaminants, and degradation by-products commonly found in recycled plastics.[20–22] Moreover, the availability of representative sample sets is often severely limited due to the high cost and time-intensive nature of sample preparation and testing, creating a data sparsity challenge for leveraging machine learning (ML) models to predict polymer composition from spectral data.[23] This scarcity of experimental spectral data further complicates the creation and validation of robust ML models, such as deep learning models[24,25] for the classification of polymer composition and type, which often require large and diverse training datasets.[23] The heterogeneity of polymer types used in the plastics industry, combined with the prevalence of high- and low-density PE, creates a class-imbalance challenge for developing high-performance ML and deep learning models to predict polymer composition and type.

Data analytics and machine learning (ML) methods, capable of modeling complex nonlinear relationships, have emerged as essential methods for analyzing complex spectral data, such as polymer classification[26–28] and materials prediction.[29,30] For example, Carrera *et al.* integrate FTIR/NIR spectra with seven machine-learning models in an economic decision framework that aligns polymer profit tiers with the most cost-effective classifier, enabling high-throughput, revenue-optimized sorting of PE, PP, PET, PS, and PVC in an automated recycling line.[31] Pocheville *et al.* demonstrated that Raman spectroscopy combined with machine learning techniques, particularly discriminant analysis (DA) with 1064 nm laser data, enabled the classification of PS and ABS in Waste Electrical and Electronic Equipment (WEEE) plastics with up to 80% purity.[32] Liu *et al.* showed that a Convolutional neural network achieved an accuracy of 97.4% for raw FTIR spectra on a mixed microplastic dataset containing 4800 samples.[33] Despite the availability of large polymer and microplastic spectral databases, practical industrial datasets are often characterized by high noise levels, inconsistent preprocessing, and label uncertainty, which significantly constrain the reliability of supervised learning models.[34,35] Additionally, inherent data sparsity and class imbalance significantly challenge the performance and reliability of deep learning models for polymer classification, often resulting in major class-dominated misclassifications. A promising strategy to mitigate these limitations, particularly in noisy and class-imbalanced industrial spectral datasets, and enhance model robustness is data augmentation,[36–38] which expands and diversifies the training data without additional experimental costs. For example, Wu *et al.* proposed a GAN-based framework to augment Raman spectroscopic data for skin cancer tissue classification, demonstrating that synthetic samples can improve prediction accuracy in limited-data scenarios.[39] Chung *et al.* leveraged a conditional GAN-based data augmentation framework that significantly improved classification performance on imbalanced spectroscopic datasets, particularly for identifying phase transitions in hydrogels.[40] Platnick *et al.* presented GANsemble, a two-stage pipeline that first auto-selects an optimal augmentation strategy and then trains a conditional GAN to synthesize class-balanced microplastic spectra, establishing Flame Ionization Detector (FID)/Internal Standard (IS) baselines and markedly improving model performance on small, highly imbalanced datasets.[41] Motivated by the practical industrial need for robust classification of recycled plastic types and the aforementioned advances in deep learning with sparse, imbalanced, and complex spectral data, this study aims to evaluate the effectiveness of spectral data augmentation using the Generative Adversarial Network (GAN).[42]

In practical recycling streams, polymer identity cannot be reliably inferred from fragment appearance or morphology alone; therefore, this study uses FTIR spectra to identify polymer type. However, experimentally collected spectra of recycled plastics are often limited and strongly class-imbalanced, and polymers with subtle spectral differences can be difficult to separate when minority classes are underrepresented. Traditionally, polymer identification from FTIR spectra is performed by spectral library matching,[43] in which an unknown spectrum is compared against reference spectra and ranked using a similarity score (*e.g.*, Pearson correlation coefficient).[44] Although library matching is widely used and effective for well-characterized, pristine references, its reliability can decrease for real recycled or environmentally affected materials when spectra are altered by baseline variations, instrument-to-instrument differences, or mixtures, and therefore may require extensive preprocessing and expert review. To complement library matching in high-throughput sorting scenarios, particularly when labeled datasets are sparse and class-imbalance, GAN-based augmentation has demonstrated advantages over conventional data augmentation techniques for spectral classification.[45] Therefore, we propose an augmentation framework to generate minority-class spectra that are both realistic and informative for classification. Specifically, an adversarial loss aligns synthetic spectra with measured spectral distributions, while a classifier-guided signal encourages the preservation of class-discriminative features to mitigate majority-class bias, thereby improving classification performance. When distinguishing polymers with similar spectral signatures. We then evaluate how classification performance varies with the augmentation ratio (synthetic/total) and training sample size, and quantify their effects on metrics such as overall accuracy and balanced accuracy.

2. Materials and methods

2.1 Source and types of recycled plastics

Recycled plastic samples rich in PE, PS, ABS, PC, PP, or PET provided by Saudi Basic Industries Corporation (SABIC) were used for this study. Recycled plastics were obtained from several post-consumer recycled (PCR) plastics and industrial waste sources, including agricultural tubing, bottlecap regrind, PE agglomeration, waste material from machining waste, unsorted PE and PP mix regrind, and other unsorted plastic mixtures. FTIR spectra were collected from these recycled plastic samples to construct two experimental datasets. Dataset 1 comprises 150 FTIR spectra acquired from 50 physical plastic samples. A second replicated dataset was collected from the samples using the same measurement protocol, but with a different FTIR. In this replicate, spectra were obtained from newly selected individual plastic flakes or pellets drawn from the same material batches, rather than repeated measurements of the same physical specimens used for Dataset 1. Combining Dataset 1 and this replicate yields Dataset 2, totaling 300 spectra. Importantly, Dataset 2 preserves the same class composition and relative class ratios as Dataset 1, with each class contributing exactly twice the number of spectra reported for Dataset 1. Specifically, per measurement round, polyethylene (PE) contains 31 samples (93 spectra), polystyrene (PS) contained 6 samples (18 spectra), acrylonitrile–butadiene–styrene (ABS) contained 4 samples (12 spectra), and polycarbonate (PC), polypropylene (PP), and polyethylene terephthalate (PET) each contained 3 samples (9 spectra). The detailed class composition for Datasets 1 and 2 is summarized in Table 1.

2.2 Spectral data generation by Fourier-transform infrared spectroscopy

FTIR spectra were collected on a Shimadzu IR-Prestige-21 FTIR spectrometer with a diamond ATR stage. A pellet or flake from a given recycled plastic source was clamped onto the ATR crystal and scanned 32 times across the range 4000–400 cm^{-1} at 4 cm^{-1} resolution to generate transmittance spectra. The crystal was cleaned with acetone between measurements to avoid cross-sample contamination.

2.3 Data pre-processing

Raw spectral data, initially recorded in percent transmittance (%T), were first converted to absorbance (A) using the standard transformation:

$$A(\nu) = -\log\left(\frac{T(\nu)}{100}\right) \quad (1)$$

where $T(\nu)$ represents the percent transmittance at the wavenumber ν.

Following conversion to a common absorbance scale, spectra were min–max normalized to [0, 1] to mitigate global

Table 1 Description of the dataset properties used in this study

Dataset	# of spectra	# of points per spectra	Frequency range (cm^{-1})	Ratio of synthetic to total data (*i.e.*, augmentation ratio)	Polymer types
Dataset 1	150	1764	400–4000	PE: 0% PP, PC, PET: 40–93.5% PS: 0–86% ABS: 10–90.3%	PE, PS, ABS, PC, PP, PET
Dataset 2	300	1764	400–4000	PE: 0% PP, PC, PET: 40–93.5% PS: 0–86% ABS: 10–90.3%	PE, PS, ABS, PC, PP, PET

intensity variations arising from measurement conditions (*e.g.*, scattering and differences in crystal–sample contact), without altering peak locations or relative peak structure. For GAN/CNN training, the normalized 1D absorbance sequence (sampled on a fixed wavenumber grid) was converted to an image-like 2D matrix *via* deterministic reshaping using a fixed indexing order.[46,47] The wavenumber axis was not explicitly stored in the 2D matrix because it is identical across samples and is therefore implicitly defined by the shared sampling grid; the matrix entries directly encode the absorbance values. Importantly, this operation is a one-to-one re-indexing, so the original 1D spectrum can be recovered exactly by flattening the matrix back to the 1D sequence and pairing it with the same wavenumber vector. All processing was performed in Python.

2.4 Data augmentation

The framework of the GAN-based data augmentation method is shown in Fig. 1, which includes three components: a generator that creates realistic spectra, a discriminator that distinguishes whether spectra are real or generated, and a classifier that predicts the label of the spectral data (*i.e.*, polymer type). The generator synthesizes spectral data samples conditioned on random Gaussian noise and their corresponding class labels. Synthetic spectral data samples from minority classes were selected and integrated with an imbalanced spectral dataset containing only real (*i.e.*, experimental) spectral data to construct a new balanced augmented dataset that served as the training set for classifying real spectral data. The dataset was split into training (70%) and test (30%) sets using stratified random sampling pre-class, preserving class distributions and avoiding test set leakage. The training dataset is used to train the model, while the test dataset is reserved for performance evaluation. No augmentation was performed on the majority class (*i.e.*, no synthetic spectral data generated by the GAN for the majority class was used). Thus, integrated synthetic samples formed balanced mini batches for model training, which were fed into the classifier during each training iteration. Table S1 presents the hyperparameters of the GAN framework.

2.5 Machine learning classification

Convolutional neural network (CNN)-based feature extractors were used to directly learn from the image-like representations of the spectral data.[48] The model architecture comprises three convolutional layers (with 32, 128, and 256 dimensions). Each layer is followed by a LeakyReLU activation function and downsampling *via* stride convolutions. The resulting feature maps are flattened and passed through a 128-dimensional fully-connected latent layer, followed by an MLP classification head that outputs SoftMax-normalized class probabilities.

The classifier performance was characterized using multiple metrics, including accuracy, balanced accuracy, precision, recall, several variants of the F-score, and the area under the ROC curve (AUC-ROC). Let $\mathcal{T} = \{(x_j, y_j)\}_{j=1}^{M}$ denotes the test dataset of M samples with true labels y_j and predicted labels $\hat{y}_j$. Then the classification accuracy was computed as:

$$\text{Accuracy} = \frac{1}{M}\sum_{j=1}^{M} 1[y_j = \hat{y}_j] \tag{2}$$

where the indicator function $1[\cdot]$ returns 1 if the condition inside is true and 0 otherwise. The F1-score for the class i, which balances precision and recall, is given by the harmonic mean as:

$$\text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{3}$$

where the precision and recall for the class i are defined as:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \tag{4}$$

where TP_i, TN_i, FP_i and FN_i are the true positives, true negatives, false positives, and false negatives for each class i,

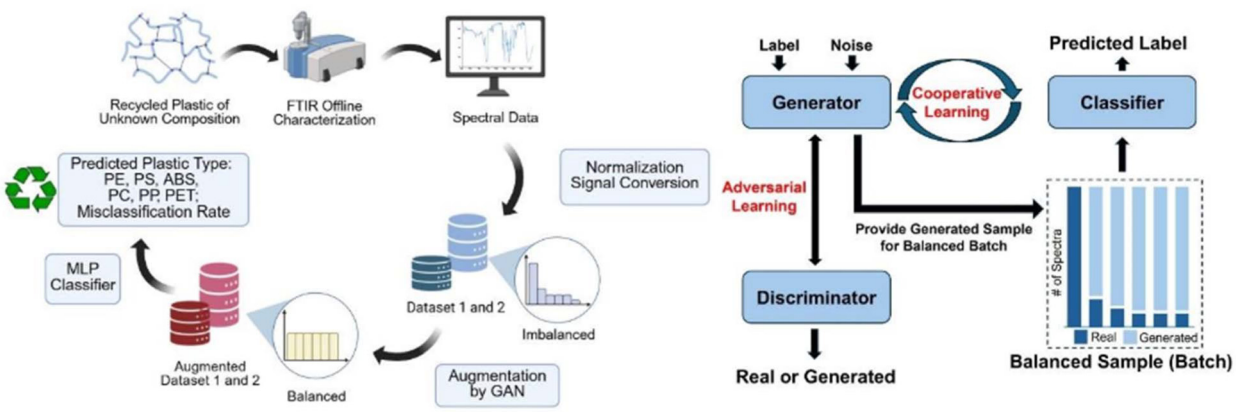

Fig. 1 (A) Workflow for plastic classification using FTIR spectral data and deep learning. Imbalanced spectral datasets are augmented with GAN-generated samples to achieve balanced class distributions for improved model training and plastic type prediction (created in BioRender. Johnson, B. (2025) https://BioRender.com/fezmfbc). (B) Schematic of the GAN-based augmentation process referenced in (A). The resulting balanced batches are then used to train the classifier.

respectively. To assess overall performance across all classes, macro-F1 is computed as the unweighted average of per-class F1 scores, treating all classes equally. Weighted-F1 averages per class F1 scores are weighted by the number of true instances per class, reflecting class distribution.

2.6 Characteristic peak identification

Peak identification on real (*i.e.*, experimental) and synthetic data from the different plastic classes was conducted using Python's find_peaks function to systematically extract critical spectral features, enabling direct assessment of how accurately the GAN reproduces domain-expert-relevant characteristics.

3. Results and discussion

3.1 Classification problem description and dataset characteristics

Fig. 1A illustrates a workflow for classifying recycled plastic materials based on their infrared spectral signatures. The pipeline uses GAN-based data augmentation to enhance underrepresented classes and improve classification accuracy on imbalanced datasets. The GAN framework and its hyperparameter values are provided in Fig. 1B and Table S1, respectively. The generator is linked to the discriminator through an adversarial loop that enforces spectral realism, while a parallel cooperative loop connects the generator to a task-specific classifier, encouraging the synthesis of class-balanced, decision-boundary-informative spectra. Merging these high-fidelity synthetic samples with imbalanced real (*i.e.*, experimental) data can create a uniform class distribution (see Fig. 1B) and equip the classifier with more discriminative training data, thereby potentially improving minority-class accuracy without compromising overall generalization.

The characteristics of the datasets used in this study are provided in Table 1 and Fig. 2 and 3. As shown in Table 1, Dataset 1 consisted of 150 experimental (*i.e.*, real) FTIR absorbance spectra collected from 50 unique post-consumer polymer samples, each containing six types of plastic (each sample measured in triplicate, yielding three spectra per sample). Each spectrum contained 1764 frequency values spanning the 400–4000 cm^{-1} range.

Dataset 2 was obtained by measuring an additional batch of recycled plastic samples with the same polymer class labels as Dataset 1. Specifically, two independent sets of heterogeneous recycled plastic samples were characterized using identical FTIR measurement protocols, yielding two experimental datasets (150 spectra per dataset) that together formed Dataset 2.

Fig. 2 shows representative real FTIR spectra for the six different recycled plastics (PE, PS, ABS, PC, PP, and PET). As shown in Fig. 2A, the representative spectra for each class highlight the complexity of the FTIR signal, such as noisy regions and regions with many adjacent and interacting resonant modes, as well as the similarity of characteristic peak locations across different polymer types. Replicate measurements acquired from a single sample provide an experimentally driven approach to augment the initial dataset with additional real spectral data to improve the classification performance (see Fig. 2B). FTIR spectra covering the full range of 400–4000 cm^{-1} were acquired on different instruments; therefore, noise characteristics may vary across samples due to inherent variability in acquisition conditions (*e.g.*, instrument drift and sample–contact variations), especially near ~400 cm^{-1}. All spectra are shown without baseline correction and were normalized to the [0, 1] range to ensure numerical stability for GAN training.

3.2 Augmentation of experimental spectral data *via* GAN and characteristics of GAN-generated synthetic data

As shown in Fig. 3, Dataset 1 exhibited an imbalanced class distribution consisting of a majority PE class (31 samples, 93

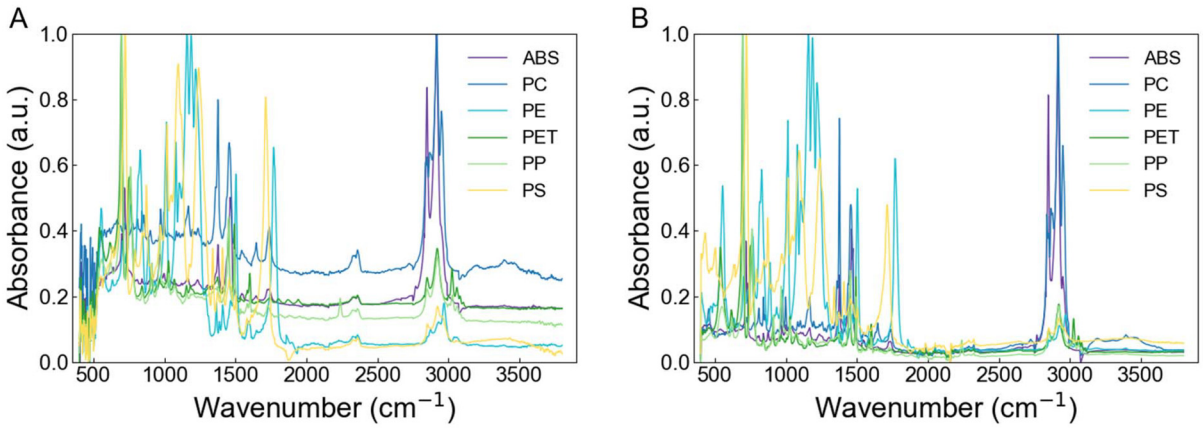

Fig. 2 Representative FTIR spectra for dataset 1 (A) and dataset 2 (B). Each color-coded group corresponds to one polymer type (PE, PS, ABS, PC, PP, and PET). Dataset 2 consists of replicate measurements from the same polymers used to generate dataset 1 but acquired using a different FTIR and randomly selected samples, as may occur in industrial settings and applications. All spectra are shown as absorbance data without baseline correction and were normalized to the [0, 1] range.

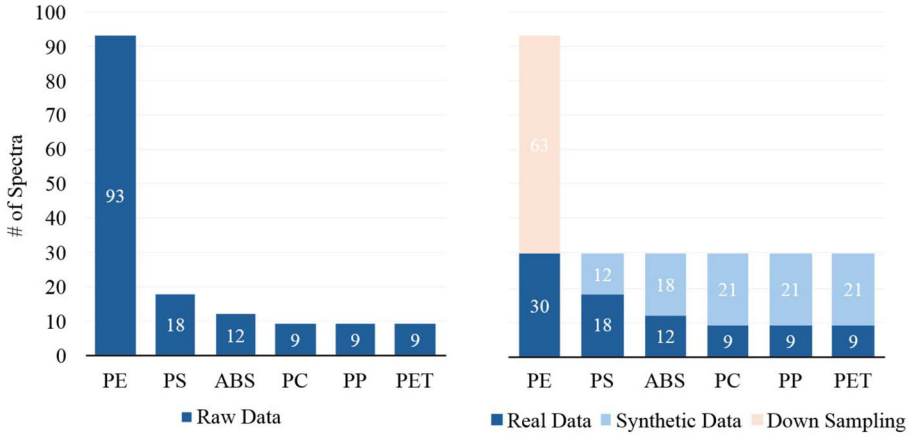

Fig. 3 Class distributions for six polymer types before (left panel) and after (right panel) dataset balancing for dataset 1. We note that dataset 2 exhibits the same class imbalance ratios as dataset 1 but contains more spectral data in each class (see Table 1).

spectra, 62%) and minority PS (6 samples, 18 spectra, 12%), ABS (4 samples, 12 spectra, 8%), PC (3 samples, 9 spectra, 6%), PET (3 samples, 9 spectra, 6%), and PP (3 samples, 9 spectra, 6%) classes. PE accounts for over 60% of all samples, while PS, ABS, PC, PP, and PET were under-represented. Given that there is a non-uniform distribution of class imbalance among the initial experimental datasets, there exists a range of synthetic-to-total spectra (*i.e.*, # synthetic spectra/# total spectra) across the multiple classes (*i.e.*, plastic types) in the balanced datasets (*i.e.*, Datasets 1 and 2). In other words, the augmentation ratio (*i.e.*, the ratio of synthetic to total spectra) in the training set was class-dependent, with a value of zero for the majority class. To evaluate the impact of the augmentation ratio on plastic type classification, we examined how balanced training datasets with varying augmentation ratios affect the machine learning model's performance (*e.g.*, misclassification rates). Random down-sampling was applied if the number of available real samples in a class exceeded the target size; synthetic samples were used to fill the gap if the number was insufficient.

To construct a balanced training set of 30 samples per polymer, we randomly down-sampled PE to the training sizes (*e.g.*, 30 samples) (see the PE column in Fig. 3 right panel) and then augmented each minority class with synthetic data generated using GAN until it reached the training size (see the PS, ABS, PC, PP, and PET columns in Fig. 3 right panel). Thus, no synthetic data was added to the majority PE class (see Fig. 4 and Fig. S1). As shown in the right panel of Fig. 3, this procedure yields uniform class representation and eliminates bias in downstream model training. Table 1 highlights the amount of synthetic data used for each polymer type and the augmentation ratio (*i.e.*, the ratio of synthetic to total data) for each class in both Dataset 1 and Dataset 2. As shown in Fig. 3 and Table 1, the majority PE class consists entirely of real spectra, while the PP, PC, and PET classes are augmented with 40–93.5% synthetic data. PS and ABS exhibit augmentation ratios ranging from 0–86% and 10–90.3%, respectively.

The locations of characteristic resonant modes (*i.e.*, peaks) in FTIR spectra of the different polymer types to be categorized and separated are typically used for classification. For example, peaks in the 2849–2915 cm^{-1} range and near 720 and 1455 cm^{-1}, associated with stretching and bending vibrations of methyl and methylene groups, are characteristic of PE[49] (see Fig. 2). We also note that in PE, methylene ($–CH_2–$) vibrations dominate, while any weak methyl-related features (*e.g.*, near ~1377 cm^{-1}) may indicate trace $–CH_3$ groups arising from either chain ends or short-chain branching and cannot be uniquely distinguished. To better understand the characteristics of GAN-generated synthetic data, we examined the locations of the 10 most prominent peaks in both the real (*i.e.*, experimental) and GAN-generated data. Fig. 4 compares the locations of the 10 most prominent modes in the experimental and GAN-generated spectral data for the six plastics. As shown in Fig. 4, the characteristic peak locations in the GAN-generated spectra agree well with the measured peaks for each minority-class plastic (recall that, in this study, no synthetic data were added for the majority class, PE) in Dataset 1. The comparison between real and GAN-generated spectra also reveals that augmentation performance varies across classes. For example, the location and amplitude of the primary characteristic peaks in PP, which was the largest minority class, were well predicted by GAN, while minority classes with fewer examples, such as PC, PP, and PET, exhibited relatively more deviation (*e.g.*, deviation near 2900 cm^{-1}), highlighting the challenges of synthetic data generation under data-sparse conditions. This deviation is likely due to imbalanced GAN training data: because the classifier focuses on the most informative peaks, less informative regions may be constrained less strongly, allowing some minority-class synthetic spectra to retain majority-class-like peak patterns. As shown in Fig. S1, the addition of replicate measurements enhanced the accuracy of the predicted spectral amplitudes in the ~2700–3000 cm^{-1} region of the spectra for the PC, PS, and PET classes. Overall, these results show that the synthetic data preserves traditional

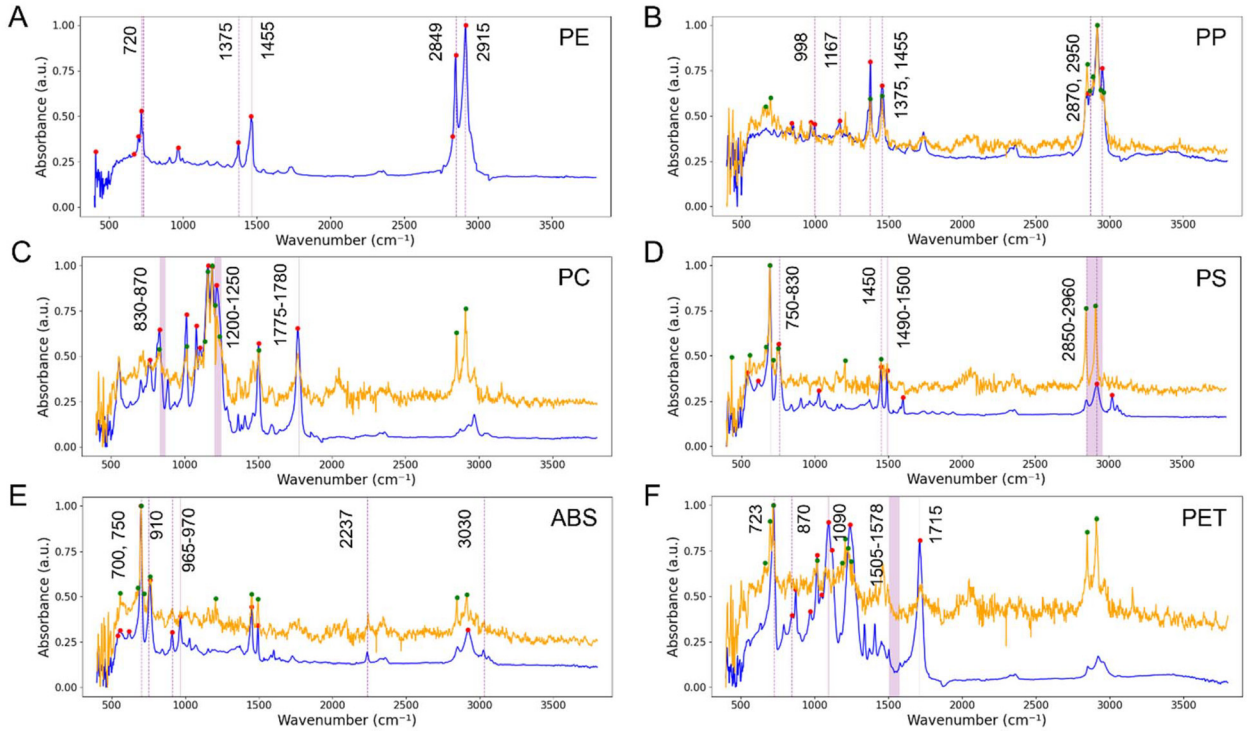

Fig. 4 Comparison between real and synthetic spectra for six plastic types (A–F: PE, PP, PC, PS, ABS, and PET, respectively) for dataset 1. The blue lines show real (*i.e.*, experimental) FTIR spectra, while the orange lines depict corresponding synthetic spectra generated *via* GAN. The largest six peaks (*i.e.*, modes) identified from real and synthetic spectra (red and green dots, respectively). Literature-reported characteristic peak positions and ranges are shown as purple dashed lines and shaded intervals (see Table S2).

features observed in the real samples, demonstrating the model's ability to generate realistic signals despite class imbalance. Additionally, we observe a correlation between the size of the experimental dataset (*i.e.*, number of spectra) and the accuracy with which the GAN approximates the real data distribution. This effect is particularly pronounced in minority classes. Although minor spectral peaks that do not correspond to any chemical vibrational mode are observed in some GAN-generated samples (see Fig. S1), classification performance is evaluated exclusively on held-out experimental spectra. As shown in Table 4 and the misclassification analysis in Section 3.3, no degradation in misclassification performance was observed, indicating that the classifier primarily relies on robust, class-diagnostic spectral features rather than spurious synthetic peaks.

3.3 Effect of data augmentation by replicate measurement and GAN on plastic classification and misclassification

Having discussed the characteristics of Dataset 1 and Dataset 2 before and after augmentation and the characteristics of the GAN-generated synthetic spectral used to balance the dataset, we next examined the performance of several classification models, particularly an MLP given its performance in our previous work with binary classification of sparse spectral data[40] to classify the six plastic type using the spectral data in Datasets 1 and 2. Table 2 reports the mean classification accuracy of the MLP classifier for six polymers using single *vs.* triplicate measurements. With only a single measurement per sample, PET, PS, PE, and ABS achieve high accuracies (100.0%, 95.2%, 90.9% and 90.0%, respectively), whereas PP fails (0%), and PC performs moderately (83.3%). Introducing three independent replicates markedly improves classification across all classes: PC and ABS reach 100%, PP recovers to 56.6%, PE rises to 98.9%, whereas PS and PET slightly decline to 90.0%. These results demonstrate that acquiring replicate measurements during characterization can improve the predictive power of sparse datasets and model robustness for multi-class classification tasks and mitigate class-specific failures. For example, replicate measurements can identify intrinsic variability (randomness) in spectral data, thereby enhancing the reliability and consistency of GAN-generated spectra and model predictions, especially for plastics such as PP, which were initially highly misclassified. Given that ABS is among

Table 2 Mean classification accuracy (%) for six polymers using single *versus* triple replicate measurements under 10-fold cross-validation (25 training samples per class)

Class (10-fold, training set: 25 spectra)	PE	PP	PC	PS	ABS	PET
Mean accuracy (1 repeat)	0.909	0.000	0.833	0.952	0.900	1.000
Mean accuracy (3 repeat)	0.989	0.566	1.000	0.900	1.000	0.900

the top ten produced plastics globally,[50] the ability to improve classification accuracy and recall (*i.e.*, reduce false-negative misclassifications) is a notable result. As shown in Table 2, using replicate experimental measurements improved ABS classification accuracy by 10%. Specifically, without replicate measurements, the precision and recall were 83.3% and 96.7%, respectively. The inclusion of additional replicate experimental measurements increased the precision and recall for the ABS class by 16.6% and 3.3%, respectively. These results demonstrate that integrating replicate measurements into the GAN-based data augmentation framework enables the generator to better capture intrinsic spectral variability, thereby synthesizing more realistic spectra and further reducing misclassification rates.

To examine the impact of data augmentation using GAN-generated synthetic data, we implemented a per-class augmentation strategy based on the size of the training set as listed in Table 3. The number of GAN-generated spectra per class (*i.e.*, plastic type) in the training set increases from 20 to 186. While PE was composed of all experimental (*i.e.*, real) spectra, all other plastic types increasingly rely on synthetic data to reach the target size. Specifically, the percentage of synthetic to total data for PP, PC, and PET increased from 40% at a training set size of 20 samples to over 93% at a training set size of 186 samples. The percentage of synthetic to total data in the PS and ABS classes increased from 0% to 86% and 10% to 90%, respectively, over the same ranges. Thus, the graduated augmentation strategy effectively balanced the class frequencies by supplementing underrepresented polymers with synthetic spectra.

Table 4 summarizes the classification results of the different augmentation levels per polymer type. Without augmentation, the model exhibited strong majority-class bias: overall accuracy remained relatively high (66.6%), while balanced accuracy dropped to 24.9%, indicating poor minority-class performance. With 20 spectra per class, the model achieved a balanced accuracy of 94.1%. Increasing the number of spectra in the training set from 40 to 50 spectra per class yielded the best performance, with balanced accuracies of 95.6% and 96.2%. However, further increases in training size by adding additional synthetic spectra (*e.g.*, 90 and 186 spectra per class) reduced balanced accuracy to 95.3% and 92.1%, indicating that moderate augmentation improves minority-class learning, whereas synthetic-heavy training reduces generalization to real spectra. In our experiments, performance peaks at 40–50 spectra per class, with the synthetic-to-total ratio in the most augmented classes at ~48% (Table 3).

A key limitation is that GAN-generated spectra may contain weak, non-physical local features (*i.e.*, peaks that do not correspond to chemical vibrations). At moderate augmentation levels, such minor wrong synthetic peaks do not lead to systematic misclassification because experimental spectra primarily guide the model. In contrast, when synthetic ratios become very high, these discrepancies can accumulate and shift the learned decision boundaries toward synthetic-specific patterns, consistent with the performance drop in Table 4.

Table 3 Description of real and synthetic spectra for each polymer class at varying per-class training set sizes

# of aug. (real + syn) spectra in training set (each class)	PE			PP			PC		
	Real (#)	Syn (#)	Syn/total (%)	Real (#)	Syn (#)	Syn/total (%)	Real (#)	Syn (#)	Syn/total (%)
20	20	0	0	12	8	40.0	12	8	40.0
30	30	0	0	12	18	60.0	12	18	60.0
40	40	0	0	12	28	70.0	12	28	70.0
50	**50**	**0**	**0**	**12**	**38**	**76.0**	**12**	**38**	**76.0**
90	90	0	0	12	78	86.7	12	78	86.7
186	186	0	0	12	174	93.5	12	174	93.5
# of aug. (real + syn) spectra in training set (each class)	PS			ABS			PET		
	Real (#)	Syn (#)	Syn/total (%)	Real (#)	Syn (#)	Syn/total (%)	Real (#)	Syn (#)	Syn/total (%)
20	20	0	0.0	18	2	10.0	12	8	40.0
30	26	4	13.3	18	12	40.0	12	18	60.0
40	26	14	35.0	18	22	55.0	12	28	70.0
50	**26**	**24**	**48.0**	**18**	**32**	**64.0**	**12**	**38**	**76.0**
90	26	64	71.1	18	72	80.0	12	78	86.7
186	26	160	86.0	18	168	90.3	12	174	93.5

Bold font indicates the augmentation ratios for each class corresponding to the best performing training set size.

Table 4 Classification performance after augmentation for varying augmentation size in the training set

# of aug. (real + syn) spectra in training set	Avg. accuracy	Balanced accuracy	Macro F1	Weighted F1	AUC-ROC
No aug. (benchmark)	0.666	0.249	0.204	0.542	0.555
20	0.975	0.941	0.952	0.972	0.966
30	0.966	0.934	0.942	0.964	0.961
40	0.977	0.956	0.964	0.977	0.974
50	**0.978**	**0.962**	**0.962**	**0.979**	**0.978**
90	0.975	0.953	0.958	0.975	0.973
186	0.966	0.921	0.935	0.963	0.954

The augmentation ratio is class-dependent and not shown (bold font indicates the training set size that exhibited the best performance).

In addition, without augmentation, the model exhibited severe class imbalance effects, collapsing toward the majority class and resulting in poor performance for ABS (50% precision, 67% recall, and 57% overall accuracy). In contrast, under optimal augmentation conditions, ABS classification reached 100% precision and 100% recall, demonstrating the effectiveness of the GAN-based augmentation in fully resolving misclassifications for this class.

Overall, these results suggest that GAN augmentation should be used as a supplement to experimental spectra rather than a replacement. In practice, the augmentation ratio should be selected based on validation performed on real (experimental) spectra, because synthetic spectra can occasionally contain false peaks that are not present in the measured data. When the synthetic portion is small, these false peaks are typically rare and weak, and the model is still mainly shaped by experimental spectra. However, if synthetic spectra occupy too large a share of the training set, the model may begin to treat these false peaks as stable, discriminative cues, which can shift decision boundaries and reduce performance on real spectra. Therefore, practical deployment should monitor whether synthetic spectra introduce peaks absent in the experimental class distributions and should limit augmentation accordingly. Finally, since this study is validated on industrial-recycled plastics, extending the approach to environmentally collected, highly weathered microplastics will require additional validation, as aging, contamination, and baseline variability may introduce spectral patterns not represented in the current training data.

4. Conclusions

The shift toward a circular plastics economy requires continuous characterization and sorting of mixed polymer waste. However, spectral data that exhibits high similarity, noise, and complexity, combined with class imbalance, can complicate accurate multi-class classification using sparse spectral data generated from off-line characterization. Conventional off-line plastic classification methods disrupt production and frequently misclassify commonly used polymers, such as PP and PE, compromising material quality and economic returns. These challenges highlight the need for rapid and accurate multi-class classification models for sparse, complex, and imbalanced spectral data. To address this, the present study employed a GAN-based data augmentation strategy. We found that replicating experimental measurements without synthetic data augmentation improved classification performance. Characteristic peak locations in GAN-generated synthetic spectral data matched the locations in experimental spectra. Using a combination of data augmentation by replicating experimental measurements (spectra) and GAN-generated synthetic spectral data (*i.e.*, Dataset 2), we achieved a balanced accuracy of over 96% for classifying six commonly recycled plastics and low misclassification of PP as PE. We also found that performance depended on the class-specific 'augmentation ratios' of the dataset, specifically the ratio of synthetic to total data in each class. These results demonstrate the potential of generative data augmentation for improving spectral classification performance in real-world polymer sorting applications. While GAN-based augmentation improves class balance, it may still fall short in capturing fine-grained spectral variations. In addition, no baseline correction was applied in this study; future work will investigate baseline-aware preprocessing strategies, as baseline variations may influence spectral fidelity and downstream classification performance.[51] Future research should investigate advanced generative models, such as diffusion models,[52] which offer higher fidelity and better capacity to capture subtle spectral features often missed by simpler methods. Systematic studies of the optimal synthetic-to-real data ratio are also essential for maximizing model performance without sacrificing generalization. In parallel, incorporating additional perturbation strategies—such as random spectral shifts, elastic distortions, and variable noise—may further enhance model robustness to real-world measurement variability. Finally, achieving continuous deployment in industrial settings may require data curation, database creation and management, and efficient online retraining mechanisms that incrementally update model parameters as new data becomes available, thereby ensuring sustained accuracy under dynamic operating conditions.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information is available. See DOI: https://doi.org/10.1039/d5an01042j.

Code for this article, including for the data augmentation and classification tasks, is available upon request from the corresponding author and SABIC.

Acknowledgements

The authors acknowledge the generous support of SABIC, which funded this project. BNJ and ZK also acknowledge the generous support of the National Science Foundation (NSF) (Grant – 1933525), which provided partial support for this work.

References

1 J.-G. Rosenboom, R. Langer and G. Traverso, Bioplastics for a circular economy, *Nat. Rev. Mater.*, 2022, **7**, 117–137.
2 A. Polyportis, R. Mugge and L. Magnier, Consumer acceptance of products made from recycled materials: A scoping review, *Resour., Conserv. Recycl.*, 2022, **186**, 106533.
3 P. K. Mallick, K. B. Salling, D. C. A. Pigosso and T. C. McAloone, Closing the loop: Establishing reverse logistics for a circular economy, a systematic review, *J. Environ. Manage.*, 2023, **328**, 117017.
4 M. Golkaram, R. Mehta, M. Taveau, A. Schwarz, H. Gankema, J. H. Urbanus, L. De Simon, S. Cakir-Benthem and T. van Harmelen, Quality model for recycled plastics (QMRP): An indicator for holistic and consistent quality assessment of recycled plastics using product functionality and material properties, *J. Cleaner Prod.*, 2022, **362**, 132311.
5 U.S. Environmental Protection Agency, Advancing Sustainable Materials Management: 2018 Fact Sheet: Assessing Trends in Materials Generation and Management in the United States, December, 2020.
6 M. S. Qureshi, A. Oasmaa, H. Pihkola, I. Deviatkin, A. Tenhunen, J. Mannila, H. Minkkinen, M. Pohjakallio and J. Laine-Ylijoki, Pyrolysis of plastic waste: Opportunities and challenges, *J. Anal. Appl. Pyrolysis*, 2020, **152**, 104804.
7 H. Jeswani, C. Krüger, M. Russ, M. Horlacher, F. Antony, S. Hann and A. Azapagic, Life cycle environmental impacts of chemical recycling via pyrolysis of mixed plastic waste in comparison with mechanical recycling and energy recovery, *Sci. Total Environ.*, 2021, **769**, 144483.
8 D. J. Langley, E. Rosca, M. Angelopoulos, O. Kamminga and C. Hooijer, Orchestrating a smart circular economy: Guiding principles for digital product passports, *J. Bus. Res.*, 2023, **169**, 114259.
9 F. A. C. Sanchez, H. Boudaoud, M. Camargo and J. M. Pearce, Plastic recycling in additive manufacturing: A systematic literature review and opportunities for the circular economy, *J. Cleaner Prod.*, 2020, **264**, 121602.
10 S. Zinchik, S. Jiang, S. Friis, F. Long, L. Høgstedt, V. M. Zavala and E. Bar-Ziv, Accurate Characterization of Mixed Plastic Waste Using Machine Learning and Fast Infrared Spectroscopy, *ACS Sustainable Chem. Eng.*, 2021, **9**, 14143–14151.
11 D. Peti, J. Dobránsky and P. Michalík, Recent Advances in Polymer Recycling: A Review of Chemical and Biological Processes for Sustainable Solutions, *Polymers*, 2025, **17**, 603.
12 K. Ragaert, L. Delva and K. Van Geem, Mechanical and chemical recycling of solid plastic waste, *Waste Manage.*, 2017, **69**, 24–58.
13 S. M. Al-Salem, P. Lettieri and J. Baeyens, Recycling and recovery routes of plastic solid waste (PSW): A review, *Waste Manage.*, 2009, **29**, 2625–2643.
14 G. Lopez, M. Artetxe, M. Amutio, J. Bilbao and M. Olazar, Thermochemical routes for the valorization of waste polyolefinic plastics to produce fuels and chemicals. A review, *Renewable Sustainable Energy Rev.*, 2017, **73**, 346–368.
15 P. R. Griffiths, Fourier Transform Infrared Spectrometry, *Science*, 1983, **222**, 297–302.
16 S. P. Mulvaney and C. D. Keating, Raman Spectroscopy, *Anal. Chem.*, 2000, **72**, 145–158.
17 J. H. F. Bothwell and J. L. Griffin, An introduction to biological nuclear magnetic resonance spectroscopy, *Biol. Rev.*, 2011, **86**, 493–510.
18 M. C. McCorry, K. F. Reardon, M. Black, C. Williams, G. Babakhanova, J. M. Halpern, S. Sarkar, N. S. Swami, K. A. Mirica, S. Boermeester and A. Underhill, Sensor technologies for quality control in engineered tissue manufacturing, *Biofabrication*, 2022, **15**, 012001.
19 S. B. Chakrapani, M. J. Minkler and B. S. Beckingham, Low-field ^{1}H-NMR spectroscopy for compositional analysis of multicomponent polymer systems, *Analyst*, 2019, **144**, 1679–1686.
20 T. Corrales, F. Catalina, C. Peinado, N. S. Allen and E. Fontan, Photooxidative and thermal degradation of polyethylenes: interrelationship by chemiluminescence, thermal gravimetric analysis and FTIR data, *J. Photochem. Photobiol., A*, 2002, **147**, 213–224.
21 M. R. Jung, F. D. Horgen, S. V. Orski, V. Rodriguez, C. K. L. Beers, G. H. Balazs, T. T. Jones, T. M. Work, K. C. Brignac, S.-J. Royer, K. D. Hyrenbach, B. A. Jensen and J. M. Lynch, Validation of ATR FT-IR to identify polymers of plastic marine debris, including those ingested by marine organisms, *Mar. Pollut. Bull.*, 2018, **127**, 704–716.
22 M. N. Miranda, M. J. Sampaio, P. B. Tavares, A. M. T. Silva and M. F. R. Pereira, Aging assessment of microplastics (LDPE, PET and uPVC) under urban environment stressors, *Sci. Total Environ.*, 2021, **796**, 148914.
23 N. Andraju, G. W. Curtzwiler, Y. Ji, E. Kozliak and P. Ranganathan, Machine-Learning-Based Predictions of

Polymer and Postconsumer Recycled Polymer Properties: A Comprehensive Review, *ACS Appl. Mater. Interfaces*, 2022, **14**, 42771–42790.

24 J. Bobulski and M. Kubanek, Deep Learning for Plastic Waste Classification System, *Appl. Comput. Intell. Soft Comput.*, 2021, **2021**, 1–7.

25 S. Lu and A. Jayaraman, Machine learning for analyses and automation of structural characterization of polymer materials, *Prog. Polym. Sci.*, 2024, **153**, 101828.

26 W. Ge, R. De Silva, Y. Fan, S. A. Sisson and M. H. Stenzel, Machine Learning in Polymer Research, *Adv. Mater.*, 2025, **37**, 2413695.

27 R. Houhou and T. Bocklitz, Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data, *Anal. Sci. Adv.*, 2021, **2**, 128–141.

28 S. Puthukulangara and D. N. P. Govindan, Deep Learning for Polymer Classification: Automating Categorization of Peptides, Plastics, and Oligosaccharides, *J. Adv. Future Res.*, 2024, **2**, 77–86.

29 B. Mortazavi, Recent Advances in Machine Learning-Assisted Multiscale Design of Energy Materials, *Adv. Energy Mater.*, 2025, **15**, 2403876.

30 Z. Zhuang and A. S. Barnard, Predicting battery applications for complex materials based on chemical composition and machine learning, *Comput. Mater. Sci.*, 2025, **246**, 113344.

31 B. Carrera, V. L. Piñol, J. B. Mata and K. Kim, A machine learning based classification models for plastic recycling using different wavelength range spectrums, *J. Cleaner Prod.*, 2022, **374**, 133883.

32 A. Pocheville, I. Uria, P. España and S. Arnaiz, Raman spectroscopy integrated with machine learning techniques to improve industrial sorting of Waste Electric and Electronic Equipment (wa) plastics, *J. Environ. Manage.*, 2025, **373**, 123897.

33 Y. Liu, W. Yao, F. Qin, L. Zhou and Y. Zheng, Spectral Classification of Large-Scale Blended (Micro)Plastics Using FT-IR Raw Spectra and Image-Based Machine Learning, *Environ. Sci. Technol.*, 2023, **57**, 6656–6663.

34 X. Yan, Z. Cao, A. Murphy and Y. Qiao, An ensemble machine learning method for microplastics identification with FTIR spectrum, *J. Environ. Chem. Eng.*, 2022, **10**, 108130.

35 J. Weisser, T. Pohl, M. Heinzinger, N. P. Ivleva, T. Hofmann and K. Glas, The identification of microplastics based on vibrational spectroscopy data – A critical review of data analysis routines, *TrAC, Trends Anal. Chem.*, 2022, **148**, 116535.

36 S. Di Frischia, P. Giammatteo, F. Angelini, V. Spizzichino, E. De Santis and L. Pomante, Enhanced data augmentation using GANs for Raman spectra classification, in 2020 *IEEE Int. Conf. Big Data Big Data*, 2020, pp. 2891–2898.

37 S. Lo, M. Seifrid, T. Gaudin and A. Aspuru-Guzik, Augmenting Polymer Datasets by Iterative Rearrangement, *J. Chem. Inf. Model.*, 2023, **63**, 4266–4276.

38 A. A. Khan, O. Chaudhari and R. Chandra, A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation, *Expert Syst. Appl.*, 2024, **244**, 122778.

39 M. Wu, S. Wang, S. Pan, A. C. Terentis, J. Strasswimmer and X. Zhu, Deep learning data augmentation for Raman spectroscopy cancer tissue classification, *Sci. Rep.*, 2021, **11**, 23842.

40 J. Chung, J. Zhang, A. I. Saimon, Y. Liu, B. N. Johnson and Z. Kong, Imbalanced spectral data analysis using data augmentation based on the generative adversarial network, *Sci. Rep.*, 2024, **14**, 13230.

41 D. Platnick, S. Khanzadeh, A. Sadeghian and R. A. Valenzano, GANsemble for Small and Imbalanced Data Sets: A Baseline for Synthetic Microplastics Data, *arXiv*, 2024, preprint, arXiv:2404.07356, DOI: 10.48550/arXiv:2404.07356.

42 I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial nets, *Adv. Neural Inf. Process. Syst.*, 2014, 2672–2680.

43 K. Munno, H. De Frond, B. O'Donnell and C. M. Rochman, Increasing the Accessibility for Characterizing Microplastics: Introducing New Application-Based and Spectral Libraries of Plastic Particles (SLoPP and SLoPP-E), *Anal. Chem.*, 2020, **92**, 2443–2451.

44 M. Meyns, S. Primpke and G. Gerdts, Library based identification and characterisation of polymers with nano-FTIR and IR-sSNOM imaging, *Anal. Methods*, 2019, **11**, 5195–5202.

45 S. Du, Y. Liao, R. Feng, F. Luo and Z. Li, FTIR-SpectralGAN: A Spectral Data Augmentation Generative Adversarial Network for Aero-Engine Hot Jet FTIR Spectral Classification, *Remote Sens.*, 2025, **17**, 1042.

46 V. Deev, V. Panchuk, E. Boichenko and D. Kirsanov, Spectrum is a picture: Feasibility study of two-dimensional convolutional neural networks in spectral processing, *Microchem. J.*, 2024, **205**, 111329.

47 R. G. McHardy, G. Antoniou, J. J. A. Conn, M. J. Baker and D. S. Palmer, Augmentation of FTIR spectral datasets using Wasserstein generative adversarial networks for cancer liquid biopsies, *Analyst*, 2023, **148**, 3860–3869.

48 X. Zhang, J. Xu, J. Yang, L. Chen, H. Zhou, X. Liu, H. Li, T. Lin and Y. Ying, Understanding the learning mechanism of convolutional neural networks in spectral analysis, *Anal. Chim. Acta*, 2020, **1119**, 41–51.

49 B. C. Smith, *Infrared Spectral Interpretation: A Systematic Approach*, CRC Press, Boca Raton, 2018.

50 D. Lithner, Å. Larsson and G. Dave, Environmental and health hazard ranking and assessment of plastic polymers based on chemical composition, *Sci. Total Environ.*, 2011, **409**, 3309–3324.

51 T. Chen, Y. Son, A. Park and S.-J. Baek, Baseline correction using a deep-learning model combining ResNet and UNet, *Analyst*, 2022, **147**, 4285–4292.

52 J. Ho, A. Jain and P. Abbeel, Denoising diffusion probabilistic models, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 6840–6851.

