PAPER

Check for updates

Computer vision for polymer characterisation using lasers

Seda Uyanik,[a] Sam Parkinson, [ID][ab] George Killick,[a] Biplab Dutta,[a] Rob Clowes,[a] Charlotte E. Boott [ID] *[a] and Andrew I. Cooper [ID] *[a]

Computer vision is a useful reaction monitoring and characterisation tool for scientists seeking to accelerate discovery processes using automation and machine learning (ML). Here we report a non-invasive laser-based method that combines computer vision and deep learning models to classify the solubility of different polymeric compounds across a range of solvents. Classifications were conducted using two to four solubility classes (soluble, soluble-colloidal, partially soluble, and insoluble), achieving high test accuracy rates ranging from 94.1% (2 classes), to 89.5% (4 classes). Using results from our solubility screening method, we also determined the Hansen Solubility Parameters (HSP) of the polymers using an optimisation algorithm. The calculated percentage Euclidean distance between the HSP values obtained from our dataset and the literature HSP values for the polymers, ranged from 11–32%. Finally, we developed the feature-wise linear modulation (FiLM) conditioned Convolutional Neural Network (CNN) regression model to estimate the size of polymeric nanoparticles between 20–440 nm and achieved a Mean Absolute Error (MAE) of 9.53 nm.

1. Introduction

Society needs the capability to discover new materials to tackle challenges including climate change, water scarcity, plastic waste, and disease. To accelerate this discovery process, scientists have turned to automation and high-throughput experimentation. This has in turn created a demand for quick and ideally non-invasive characterisation and data analysis methods, as the bottleneck shifts from conducting experiments to evaluating the large amounts of data produced. Artificial Intelligence (AI) systems enable effective learning of critical model features, potentially achieving a proficiency that parallels or surpasses conventional theories and methodologies.[1] In recent years, machine learning (ML) has become a valuable tool for materials discovery and in chemistry research.[2,3] ML has been applied in areas such as melting point prediction,[4,5] electrical and thermal conductivity prediction,[6,7] crystal structure representation,[8,9] particle size prediction,[10] and spectroscopy data analysis.[11] Additionally, Bayesian Optimisation (BO) and other optimisation algorithms have helped to accelerate chemical discovery by exploring large chemical spaces efficiently, thus lowering experimental costs, and discovering optimal materials with a limited amount of data.[12] While ML excels in pattern recognition, particularly in classification tasks

such as image analysis, it often falters in interpretive tasks that require a clear understanding of underlying mechanisms. This limitation constrains its use, although progress in areas such as transfer learning, distributional adaptation, and attention mechanisms may enhance the efficiency of ML models in the future, while also making models more robust to fluctuations in real-world data, allowing them focus on the most pertinent information.[13]

Computer vision is a rapidly advancing field of AI that leverages image processing, pattern recognition, ML and deep learning (DL) to enable machines to analyse, interpret, and extract meaningful information from visual data. By using deep learning models, such as Convolutional Neural Networks (CNNs)[14] and Vision Transformers,[15] computer vision systems can extract a wide range of characteristics from unregulated inputs[16] and can enhance precision and adaptability.[17] This ability to capture and analyse an extensive set of features in a non-invasive manner makes computer vision useful for chemistry applications. Current applications of computer vision in chemistry include the simultaneous tracking of multiple physical outputs (*e.g.*, liquid levels, solid formation, residue presence, and colour);[18] real-time monitoring of catalyst degradation kinetics and product formation using colourimetric data;[19] and viscosity estimation using semantic segmentation masks,[20] or a 3D Convolutional Neural Network (3D-CNN).[21] In addition, several computer vision methods have been developed for solubility screening of small molecules including the average brightness CV method used for HeinSight

[a]Department of Chemistry and Materials Innovation Factory, University of Liverpool, 51 Oxford Street, Liverpool, L7 3NY, UK. E-mail: c.boott@liverpool.ac.uk; aicooper@liverpool.ac.uk

*[b]Aston Institute for Membrane Excellence, Aston University, Birmingham, B4 7ET, UK*
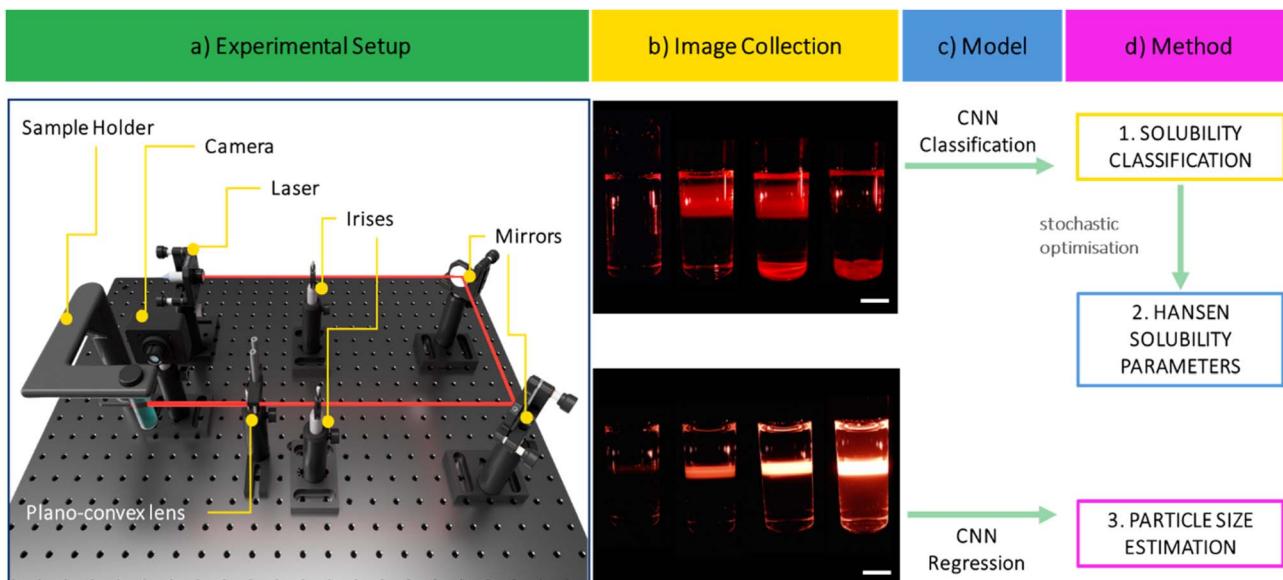
Fig. 1 Schematic representation of the laser scattering workflow: (a) experimental setup – illustration of the lab hardware used for the laser scattering experiments, (b) image collection – representative images collected for solubility classification (top) and particle size estimation (bottom); scale bar = 1 cm, (c) model – CNN models are used in the study, (d) method – the materials characterisation information obtained from the model.

turbidity,[22] mask R-CNN and image segmentation,[23] and deep neural networks (DNNs).[24]

We present here a laser-based platform that uses computer vision to characterise polymer solubility and other related properties (Fig. 1). ML techniques are still in the early stages of application in polymer science, whereas their development is more advanced in the fields of organic and inorganic materials.[25] Polymers present specific challenges for ML models, including the complex nature of the polymer structure—all polymers are mixtures of different molar mass molecules—as well as challenges related to sampling; for example, polymer solutions can become viscous and hard for liquid handlers to deal with.[26] Macromolecules exhibit different dissolution behaviour compared to small molecules.[27] For example, polymers chains exist as coils in solutions, and these coils are large enough to scatter a laser beam. Here, this allowed us to refine polymer solubility from two binary classes (soluble, insoluble) into four classes (soluble, soluble-colloidal, partially soluble, and insoluble) with an accuracy of 89.5%. Following these classifications, we used these solubility data along with an optimisation algorithm to determine the polymer Hansen Solubility Parameters (HSP), a widely used method for assessing solubility.[28] Finally, to demonstrate the versatility of this platform for characterising polymeric materials, we have created a regression model that uses the light scattering properties of polymer nanoparticles to estimate particle size between 20–440 nm and achieved a Mean Absolute Error (MAE) of 9.53 nm.

2. Methods

2.1 Experimental setup

A Logitech C930-E full HD webcam was used to capture images, with the focus set to 105, brightness adjusted to 0.55, a frame resolution of 1920 × 1080 pixels ($W \times H$), full list of settings (Table S1), and a distance between the sample and the camera of 5 cm. The laser source, CPS635 – Collimated Laser Diode Module (635 nm, 4.5 mW, $\varnothing$ 11 mm), was mounted using a 3R class laser mounting bracket purchased from Thorlabs. We selected the 635 nm wavelength due to its reduced susceptibility to scattering in the presence of solvent impurities, ensuring clearer and more consistent imaging results.[29,30] Two PF10-03-G01 – $\varnothing$ 1″ Protected Aluminium Mirrors were used to direct the laser beam. The beam path was further controlled using two ID12/M – Mounted Standard Irises with a $\varnothing$ 12.0 mm maximum aperture, TR75/M. To achieve a wider beam size and to further minimize the impact of solvent impurities on the sample, an LK1684L1-A – $f = -12.69$ mm, $H = 10$ mm, $L = 12$ mm, N-BK7 Plano-Convex Cylindrical Lens, ARC: 350–700 nm, was used (Fig. 1a). The sample holder, custom-designed, accommodated an 8 mL Chemspeed vial. A technical drawing is provided in Fig. S1. A TPS13 – 12″ × 6″ (305 mm × 152 mm) Straight Laser Safety Screen was used as a beam stop, all within a custom black hardboard enclosure measuring 525 mm × 375 mm × 300 mm ($L \times W \times H$).

Experimental setup for particle size estimation: as above, but without the plano-convex lens, this was removed due to the small size of the some of the particles measured.

2.2 Sample preparation for polymer solubility classification

The dataset used to train our classification model comprised 9 different solid polymers, 24 different solvents (Table S2) and solvent blends, and seven different polymer concentrations (0.1, 0.3, 0.5, 0.7, 1, 5, and 10% w/v), quantities used to prepare each sample concentration are shown in (Table S3), resulting in a total dataset of 911 images (note: not all solvents or their

blends were tested at all concentrations; see Table S4). All polymer samples were manually prepared in the laboratory, and solvents were filtered with a 0.2 μm PTFE filter before use. To enhance the robustness of our model, 167 out of 911 images in the dataset were captured without the plano-convex lens, providing additional variation compared to the lens-assisted images. To ensure dataset integrity, data cleaning was conducted by removing samples that exhibited image features that prevented their class from being identified (*e.g.*, excessive scattering and image artifacts), resulting in the exclusion of 30 images and low-concentration samples before model training (Fig. S2) An additional test dataset was prepared using liquid-form polymers, specifically polydimethylsiloxane (PDMS) and polyethylene glycol, $M_w = 600$ g mol^{-1} (PEG600) at five concentrations 0.1, 0.5, 1, 3, and 5% w/v, and 3 solvents per polymer (PDMS: dichloromethane, ethanol and heptane, PEG600: water, acetone and heptane) and images collect pre- and post-vortex giving a dataset of 60 images. This test dataset was not included in the main training dataset but was used exclusively to evaluate model performance in distinguishing effect of pre-vortex and post-vortex samples of these liquid polymer–solvent combinations. Details of the polymers and molecular weight (M_w) are given in SI Section S1.1.

2.3 Sample preparation for Hansen solubility parameter determination

The dataset for Hansen Solubility Parameter (HSP) optimisation was collected using 16 solvents that were selected as being well-distributed over the HSP space, as well as being commercially available and inexpensive (as detailed in Table S2). The distribution of solvents in three-dimensional system is provided in Fig. S3. Four common industrially relevant polymers were chosen, polystyrene (PS, $M_w = 192\,000$ g mol^{-1}), polymethyl methacrylate (PMMA, $M_w = 15\,000$ g mol^{-1}), poly-vinylpyrrolidone (PVP, $M_w = 55\,000$ g mol^{-1}) and poly-caprolactone (PCL, $M_w = 80\,000$ g mol^{-1}) and samples prepared in all 16 solvents at a concentration of 5% w/v (0.250 g in 4.750 mL) and left for 2 h before being analysed on the laser platform. To determine the HSP of polymers, datasets were created based on the 2-class classifications of soluble and insoluble substances derived from our solubility screening method. For the purposes of HSP estimation, soluble and soluble-colloidal classifications were both treated as soluble, labelled as 1, and partially soluble and insoluble classes are treated as insoluble, and labelled as 0 in the dataset.

2.4 Sample preparation for particle size determination

Commercial PS size standards were purchased with defined particle sizes of 20, 30, 50, 60, 80, 100, 150, 200, 240, 300, 350, and 400 nm. The concentrations of these size standards were adjusted to 0.01, 0.03, 0.05, 0.06, 0.07, 0.10, 0.13, 0.15, 0.20, and 0.40% v/v, respectively, in 5 mL aqueous solutions. A total of 120 samples were prepared manually in the laboratory, quantities used to prepare each sample concentration are shown in Table S3 and DLS analysis was conducted on each of them (Table S5). For all samples the particle size determined by DLS at 0.40% v/v

was used as the ground truth. The same procedure was applied to PMMA particles with sizes of 100 nm and 200 nm, silica particles with a size of 100 nm, and poly(dimethylacrylamide)-*b*-poly(diacetone acrylamide) (PDMAm-*b*-PDAAm) particles (Fig. S4) measuring 89 nm at a concentration of 0.40% v/v. These samples were used as a separate test set to evaluate the regression model's ability to predict the particle sizes of different substances based on the polystyrene model.

2.5 Model development for solubility classification

Our solubility classification models are deep convolutional neural networks. We trained and evaluated several deep convolutional neural networks for the solubility classification task. We considered three similar sized models by parameter count, namely: ResNet18,[31] EfficientNet_b0,[32] and ConvNeXt Tiny.[33] The final feature maps were average pooled into a single vector. The final layer of all models was a linear layer that predicts four classes. We use ImageNet pretrained weights. While the images from the solubility dataset differ significantly from the ImageNet dataset, we found that using pretrained weights nonetheless contributed to greater stability in metrics during training and validation. For each model, we used ImageNet pretrained weights and replaced the final linear layer with a four-output linear layer. We trained each model *via* gradient descent, using the Cross Entropy loss, the AdamW[34] optimizer, and a OneCycle[35] learning rate scheduler. Due to class imbalance, we used a weighted sampler to ensure that there was an approximately equal number of samples for each class in a training batch. Models were evaluated with 6-fold Stratified Group cross validation. We used stratification to ensure the ratio between the number of samples in each class was approximately the same between all folds. As multiple images exist for a given polymer solvent combination, we created a solvent and polymer identifier and used this as a group attribute. This ensures that images of the same solvent polymer combination are not split across training and test folds and allows us to more accurately assess the ability of the model to classify unseen polymer solvent combinations. We further split the training folds into train and validation sets at a 4 : 1 ratio, again using a Stratified Group split. Validation sets were used for selecting the best checkpoints over training as well as for hyperparameter optimization. We provide a detailed overview of the full training pipeline and hyperparameters in Table S6.

2.6 Model development for particle size prediction

For particle size prediction, we propose Polymer Particle Size Network (PPSNet). It consists of three convolutional blocks, followed by a fully connected network with one regression output. Each convolutional block comprises a convolution layer, ReLU activation, and max pooling. The amount of light scattering depends on both particle size and on concentration. To help the model disambiguate between these two factors, we also provided it with concentration information. This was implemented with feature-wise linear modulation (FiLM) layers,[36] which is a simple method for conditioning the internal representations of neural networks on other inputs. In our case,

we used a Multi-Layer Perceptron (MLP) to first encode the min–max scaled concentration into a higher dimensional space and used this encoding as the condition input to the FiLM layer. Our motivation for adopting an MLP is to allow conditioning to be based on a non-linear function of concentration and has the potential to allow more expressive behaviour than a simple linear projection. We compared both ReLU and sine activation for the MLP. Sine activations can facilitate the learning of high frequency functions over low dimensional inputs,[37] while ReLUs have a spectral bias to low frequency functions. In the case of the sine activated MLP, we took inspiration from Sitzmann *et al.*,[38] applying a scalar multiplier, omega, to the pre-activations of the hidden layer and is a hyperparameter we tuned. As the amount of light scattering is dependent on particle size, a reasonable hypothesis is that particle size can be predicted from average pixel intensity. To warrant the use of a CNN over simpler and more convenient models, we compared performance against a polynomial regression baseline. As input, we converted the RGB image to LAB and use the average L value as an image feature. Additionally, we provided the concentration value as is the case in the CNN model. We trained each model with mean squared error loss. We evaluated each model with 5-fold group cross validation. As with the solubility dataset, multiple images of the same sample exist within the dataset. Accordingly, we created a unique identifier based on particle size and concentration and use this as a group attribute to ensure these images are not spread over training and testing folds and allows us to better quantify the model's ability to generalize to unseen particle size and concentration combinations, details can be found in SI (Table S7).

3. Results and discussion

3.1 Polymer solubility classification

A total of 911 polymer–solvent samples were prepared and imaged using our automated laser platform to develop the training dataset for our polymer solubility classification. Before imaging, samples were classified visually as either soluble or insoluble based on the presence of residual undissolved polymer. When these samples were imaged using the laser platform, additional information was obtained that was undetectable to the human eye (Fig. 2a). When assessing the dataset, it became clear that the samples could be grouped into four distinct classes (Fig. 2 and S5). The first class, which we labelled as soluble, displays a homogenous solution that exhibits minimal laser scattering. For the second class, samples initially classified as soluble based on visual observation exhibited a distinct solid red band in the laser imaging, corresponding to the region
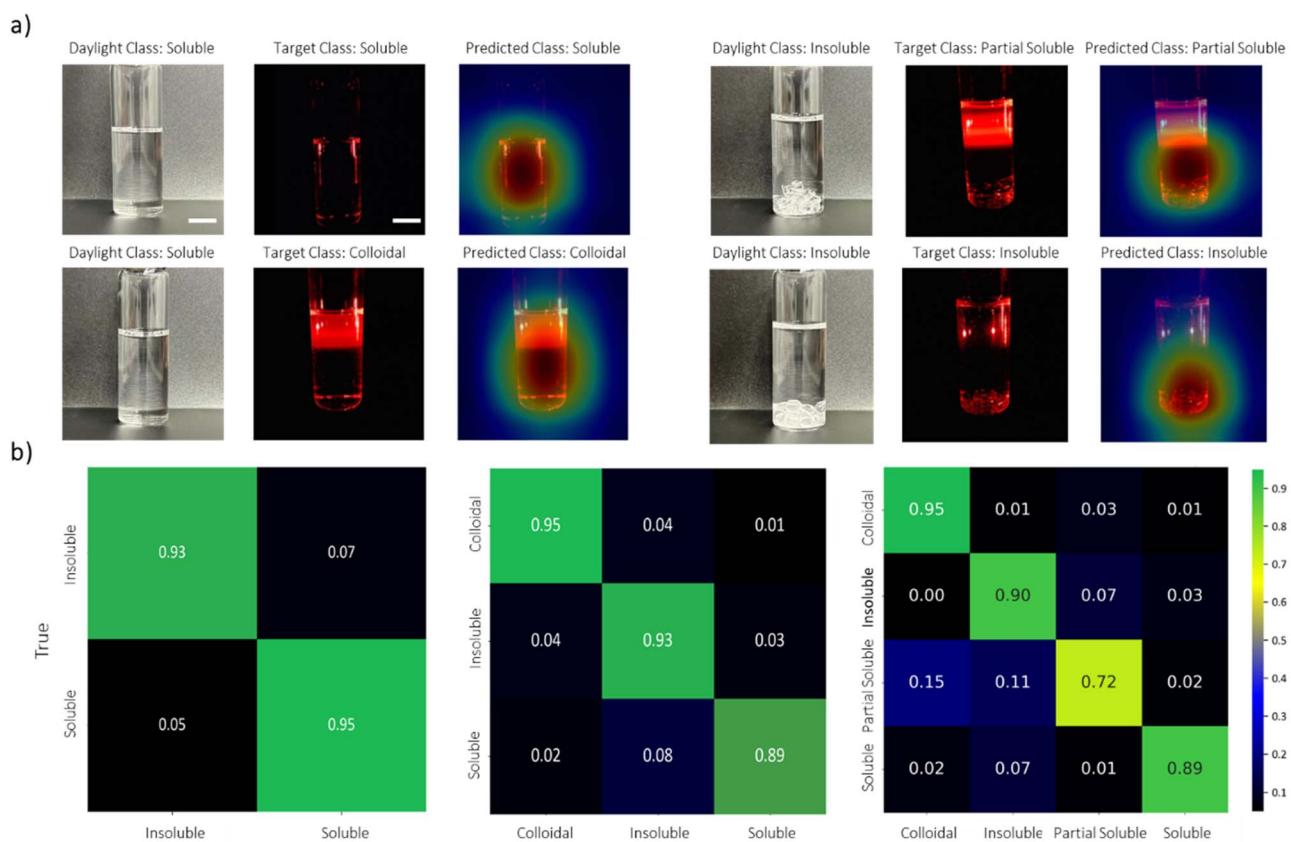

Fig. 2 Solubility classification model overview. (a) Illustration of four classes using natural light (daylight) and laser-scattering images, supported by Grad-CAM visualizations that explain the areas in the image that have key importance: soluble, soluble-colloidal, partially soluble, and insoluble (scale bar = 1 cm); (b) confusion matrix shows the classification distribution of 2-class, 3-class, and 4-class predictions *versus* true labels using the ResNet18 model.

Table 1 Classification differences observed using natural light and laser light sources; the laser source is more discriminatory

Solubility	Natural light	Laser
Soluble	442	87
Insoluble	469	351
Soluble-colloidal	0	355
Partial soluble	0	118

where the laser light passes through the solution; we labelled this class as soluble-colloidal since this band arises from scattering of the laser by polymer coils in solution. When comparing polymers in these first two classes, the size of polymer coils, as determined by DLS (Fig. S6) in the soluble-colloidal class were greater than those in the soluble class (soluble sizes < ~10 nm, whereas soluble-colloidal sizes > ~10 nm). Given that scattering intensity is known to be proportional to r^6 $(I \propto r^6)$,[39] where r is the radius of the scattering particle, it was clear that this difference in polymer coil size was the reason for the red-band in the soluble-colloidal class. The third class, though appearing insoluble by visual inspection, was reclassified as partially soluble based on laser imaging, which revealed the presence of both soluble and insoluble polymer. For this class, we separated the liquid and solid portions of the solution and performed gel permeation chromatography (GPC) measurements on each phase. The results revealed that the initial polymer distribution split into two fractions: the liquid portion contained polymers with lower molecular weight, while the solid portion contained polymers with higher molecular weight (Fig. S7), thus allowing us to confirm that the true class was partially soluble. The final class was identified as insoluble, with scattering attributed solely to the presence of undissolved polymer. Additionally, Fig. 2 highlights the visual differences of samples across each class under both laser and normal light conditions. An advantage of using laser illumination instead of natural light is its ability to increase the number of classification categories, allowing differentiation of up to four classes. Table 1 presents the data distribution under daylight conditions and demonstrates how the classification changes with laser illumination.

To classify the images in our dataset, we used a Convolutional Neural Network (CNN) architecture. CNNs. We evaluated the performance of three different publicly available CNN models (ConvNeXt Tiny, EfficientNet_b0, and ResNet18). We selected

these models because they are already pre-trained on large diverse object datasets and were then further trained with our dataset. Each model was trained and evaluated with 6-fold group cross validation where the polymer was used as the group attribute; that is, test folds contained polymer–solvent combinations not seen in the train folds. This allowed us to assess the model's ability to generalize to unseen polymers. We additionally split the training folds into a training and validation set with a 4 : 1 ratio. We used the validation set for checkpointing and hyperparameter tuning. Results were averaged over all test folds using the best checkpoints (based on validation loss). For the three models tested, ResNet18 achieved a consistently superior performance for all classification tasks achieving 94.1% accuracy for a two-class (soluble, insoluble) task. As the number of classes increased to three (soluble, partially soluble and insoluble) and four (soluble, soluble colloidal, partially soluble and insoluble), the accuracies fell to 93.5% and 89.5%, respectively. It is worth noting that the other models achieved accuracy that was also very close to ResNet18, (93.9% and 93.7% in the two-class task for EfficientNet_b0 and ConvNeXt Tiny, respectively) but ResNet18 was selected as the dedicated CNN model because it had the highest overall accuracy (Table 2). The confusion matrix indicates that model's performance decreases as the number of classification categories increases, with the partially soluble class being the hardest to predict (Fig. 2b). We used Grad-CAM to elucidate the features highlighted by our CNN model and to explain the rationale behind its predictions, these showed that the model primarily concentrates on the central and lower areas of the images to make the classification (Fig. 2a).

For liquid polymer samples, the initial classification based on visual estimation as soluble or insoluble was often inconsistent with the laser imaging results, particularly for samples with concentrations below 1% w/v. In these samples, the distinct liquid layers were not apparent, neither by eye nor by laser, leading to the incorrect classification of the samples as soluble (Fig. S8). To evaluated whether the model's performance for liquid–liquid polymer–solvent combinations could be improved by vortexing the samples prior to image collection, a separate test set was prepared using PEG600 and PDMS polymer samples and included both pre-vortex and post-vortex images (Fig. S9). As part of the procedure, all samples were vortexed prior to imaging, with visual classification refined based on whether an emulsion was observed. The soluble class exhibited minimal scattering, while the insoluble class

Table 2 Solubility classification models performance on 3 different classification number − 2, 3 and 4 classes

# of classes	Model	Precision ± std	Recall ± std	Accuracy ± std	F1 score ± std
2 classes	ResNet18	0.943 ± 0.021	0.934 ± 0.024	0.941 ± 0.021	0.938 ± 0.022
	EfficientNet_b0	0.935 ± 0.026	0.935 ± 0.022	0.939 ± 0.026	0.935 ± 0.024
	ConvNeXt Tiny	0.936 ± 0.022	0.931 ± 0.022	0.937 ± 0.018	0.933 ± 0.022
3 classes	ResNet18	0.907 ± 0.026	0.924 ± 0.040	0.935 ± 0.022	0.915 ± 0.030
	EfficientNet_b0	0.890 ± 0.045	0.922 ± 0.052	0.931 ± 0.020	0.905 ± 0.045
	ConvNeXt Tiny	0.901 ± 0.029	0.925 ± 0.036	0.933 ± 0.014	0.913 ± 0.028
4 classes	ResNet18	0.860 ± 0.046	0.882 ± 0.056	0.895 ± 0.035	0.870 ± 0.042
	EfficientNet_b0	0.843 ± 0.070	0.887 ± 0.070	0.892 ± 0.046	0.854 ± 0.078
	ConvNeXt Tiny	0.862 ± 0.058	0.883 ± 0.055	0.893 ± 0.045	0.871 ± 0.046

displayed scattering due to emulsions formed during vortex. We used the best weights from the ResNet18 model to evaluate the solubility of our liquid polymer samples. The model's accuracy was assessed on two sets of 30 images: pre-vortex and post-vortex. For pre-vortex images, the overall accuracy was 63.3%, while for post-vortex images, it improved to 73.3%. Notably, the model's accuracy was concentration-dependent; for example, at a low concentration of 0.1% w/v, the accuracy was only 50%, whereas for higher concentrations of 1, 3, and 5% w/v, the accuracy improved significantly to 83.3% for post-vortex images (Table S8). Based on these results, we recommend for liquid polymers to use a minimum sample concentration of 1% w/v and to vortex the samples before image collection to improve model performance.

3.2 Hansen solubility parameter determination

After developing our solubility classification model, we explored whether these results could be used to estimate the Hansen Solubility Parameters (HSPs) of the polymers. The model optimises solvent placement within a three-dimensional solubility sphere defined by dispersion forces (δD), polarity (δP), and hydrogen bonding (δH) parameters.[40] It uses a fitness function to minimise the Relative Energy Difference (RED), penalises solvents for incorrect positioning relative to the sphere, incorporates a size factor to regulate sphere size, and sets parameter bounds based on the mean HSP values of good solvents more details on the model development are provided in see SI section S2.2.[41–43] While HSPs can be predicted theoretically, for large macromolecules, such as polymers, experimental determination is preferred.[44] Moreover, in high-throughput experimentation settings, the composition of the polymer, which is needed for HSP prediction, may not be known without additional analysis (*e.g.*, in copolymerisation of more than one monomer). Typically, experimental methods determine the HSPs of polymers by assessing their solubility in solvents with known HSP values, enabling the estimation of the polymer's position within the solubility sphere.[28] Given that our solubility model inherently classifies samples in this format, we took solubility data directly from this model to determine polymer HSP values. We used a solubility classification model to determine polymer HSP values. This model was trained on a dataset of solubility data, but with a key restriction—for each polymer being tested, its corresponding images at 5% w/v were excluded from both the training and validation sets. Instead, these images were reserved

exclusively for testing. For example, in the case of PVP, all PVP-related images were omitted from training and validation, ensuring that the model had not seen them before testing. This same approach was applied consistently to the four different polymers to ensure an unbiased evaluation of their solubility classification. Experimentally determined HSP values are sensitive to the solvents selected for the experiment, therefore we selected 16 solvents (Table S2) that are well-distributed across three-dimensional HSP space to capture a comprehensive range of polymer–solvent interactions. We fixed the solvents for all polymers to minimize biases in solvent selection, and enhancing the generalizability of the HSP values derived from the dataset. In addition, we chose to use fewer solvents than normally used for HSP determination (16 *vs. ca.* 40 solvents), as we wanted to reduce the experimental cost and sample requirements.

First, we used a known dataset of polyether sulfone from the literature to evaluate the performance of different HSP optimisers reported in the literature (Table S9).[41–43] We determined the genetic optimiser to be one of the optimal as it provided the lowest percentage Euclidean distance of the optimisers tested at 1.2% similar values (18.84, 11.22 and 7.95 MPa$^{1/2}$ for δD, δP and δH, respectively) to the known HSP values (19, 11 and 8 MPa$^{1/2}$ for δD, δP and δH, respectively) for the polyether sulfone according to the Hansen solubility website.[45] We then applied our optimization algorithm to datasets collected using our solubility classification model for PS, PMMA, PVP and PCL. For datasets consisting of 16 single solvents, the optimizer achieved the following Euclidean distance (ED) and percentage Euclidean distance (PED) between the literature HSP values and the HSP values obtained from our optimizer: PMMA (ED = 2.4, PED = 11%), PS (ED = 2.9, PED = 15%), PVP (ED = 5.3, PED = 22%), and PCL (ED = 6.5, PED = 32%) in Table 3, solubility spheres can be seen in Fig S10. As expected, the HSP values obtained from the genetic optimiser were slightly different from those reported in the literature. We attribute this variation to differences in the selection and number of solvents used in the datasets. As previously reported, the accuracy of HSP estimation strongly depends on the variety and quality of the solvent dataset; insufficient or unbalanced datasets can significantly limit the predictive power of the model.[46]

3.3 Particle size estimation

Finally, to broaden applicability of our platform for polymer characterisation, we evaluated it for polymeric nanoparticle size

Table 3 Hansen solubility parameter optimisation model overview. The Hansen solubility parameters (HSP) for both original and optimized values of four different polymers: PMMA, PS, PCL, and PVP. ED = Euclidean Distance and PED = Percentage of Euclidean Distance

Polymer	Concentration (% w/v)	Hansen parameters (MPa$^{1/2}$)			Laser estimation parameter prediction (MPa$^{1/2}$)				ED	PED (%)
		δD	δP	δH	δD	δP	δH	R_0		
PMMA	5	18.6	10.5	5.1	17.4	10.4	3.1	9.2	2.4	11
PS	5	18.5	4.5	2.9	18.1	3.9	5.7	4.5	2.9	15
PVP	5	17.5	8	15	20.0	12.6	14.1	13.4	5.3	22
PCL	5	17.7	5	8.4	18.3	10.5	5.0	9.6	6.5	32

estimation, which is important for multiple applications. As mentioned previously for our solubility model, the intensity of scattered light is proportional to the scatterer size ($I \propto r^6$). We surmised that these data could be used to estimate the size of polymeric nanoparticles. To evaluate this hypothesis, polystyrene size standards (20–400 nm) were prepared over a range of concentrations (0.01–0.4% v/v) and imaged to provide the training data for our model (Fig. 3a and S11). Low size dispersity standards (PDI < 0.3) were used to build our initial model to ensure the greatest accuracy of prediction due to the heavy
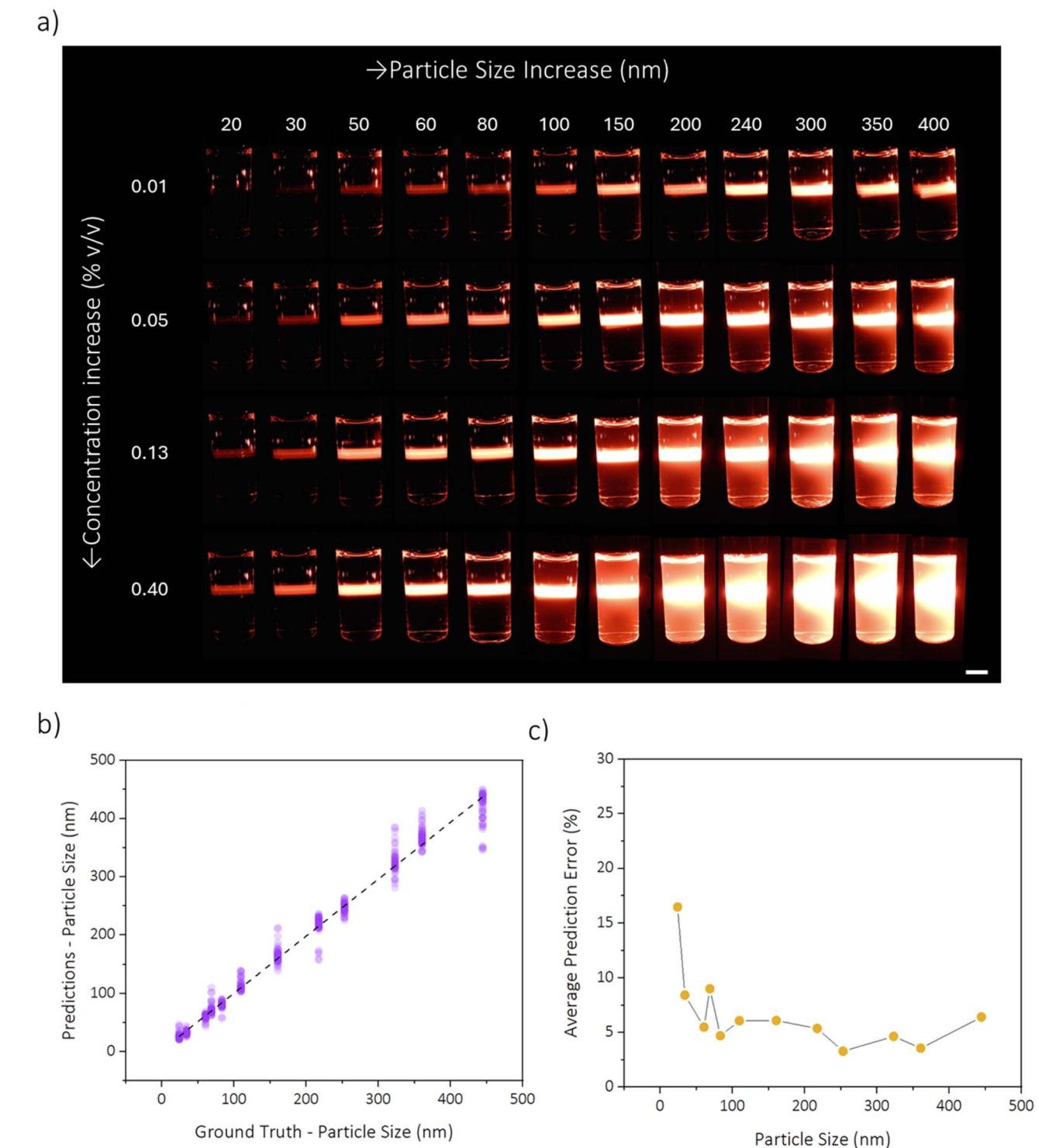
Fig. 3 Particle size estimation model overview. (a) Images of a subset of the dataset (concentrations = 0.05, 0.13, 0.40 % v/v) ordered by increasing particle size and concentration. (b) Predicted particle size *vs.* ground truth particle size for test data. (c) Mean absolute percentage error as a function of particle size for test data.

skewing of intensity-based measurements towards larger particles. Among the 120 samples analysed, most exhibited PDIs below 0.3, although a small subset had PDIs exceeding this threshold. These samples, which mostly exhibited higher PDI, had low concentrations, leading us to consider that this discrepancy arose from the limitations of DLS measurements, where low concentration samples pose challenges due to the inherently low signal-to-noise ratio (Table S5).[47] Additionally, we collected five images per sample to enhance the model's robustness against minor variations in scattering.

In Table 4, we report particle size prediction results for PPSNet, compared against an equivalent CNN with no concentration conditioning, an EfficientNet B0 with FiLM concentration conditioning, and a polynomial regression model using average image brightness and concentration as input. All models were trained with particle size images labelled using DLS measurements from high-concentration (0.4% v/v) samples. Results are averaged over 5 test folds. We show that PPSNet performs the best out of all models, with a MAE of 9.53 $\pm$ 4.27 nm and an R^2 of 0.99 $\pm$ 0.01. When comparing PPSNet to an equivalent CNN, but without concentration conditioning we observe an increase in prediction error (mean absolute error (MAE) of 9.53 nm *vs.* 22.25 nm) (Tables 4 and S10). This supports our hypothesis that incorporating concentration information is useful in building a particle size estimation model that is robust to different concentrations of polymers in solution. Interestingly, even without concentration information, the CNN performed relatively well and could find applications in scenarios where a higher prediction error is tolerable, and where no concentration information is available. Furthermore, we evaluated an Efficient Net B0 model with FiLM layers after each convolution block to test whether a larger model can reduce prediction error. Despite having significantly more

parameters than PPSNet, we find that performance is worse (MAE of 11.60 $\pm$ 3.07 nm). This can likely be attributed to overfitting, as the dataset is relatively small, and could be improved by collecting more data. Nonetheless, the ability of PPSNet to accurately estimate particle size despite its small size is appealing due to its lower computational overhead.

In contrast to PPSNet and other CNN models, the polynomial regression model performed relatively poorly with an MAE of 32.55 $\pm$ 6.67 nm, verifying that the adoption of more complex neural network-based architectures is warranted for this task. The failure of the polynomial regression model can be attributed to the fact that the function of image brightness from particle size and concentration is not injective as different particle sizes can be mapped to the same image brightness (Fig. S12). Therefore, it is not possible to find a function that maps image brightness to particle size. This non-injective behaviour is expected based on Mie scattering theory. In contrast, CNNs can automatically extract useful features from visual data, beyond just image brightness, which we speculate allows them to circumvent this problem. The model's predictions closely align with the ground truth particle sizes, with an observable trend of increasing absolute prediction error for larger particle sizes. We speculate this is due to a positive skewness (Fisher–Pearson skewness coefficient = 0.61) in the particle size distribution (Fig. 3b). Despite this, the average percentage error remains consistently low at around 5%, demonstrating that the model maintains high relative accuracy across the range of particle sizes (Fig. 3c).

Next, we tested our model's performance with different nanoparticles. Silica and PMMA size standards were purchased and poly(dimethylacrylamide)–poly(diacetone acrylamide) (PDMAm-*b*-PDAAm) spherical nanoparticles were synthesised and characterised *via* DLS.[48] Images were collected on our laser platform and passed to our model for particle size estimation (Fig. S13). For all nanoparticle systems, our model consistently underestimated the nanoparticle size (Table 5). This is not surprising, because the model was trained on PS samples only. In Mie theory, scattering intensity proportional to both particle size and the refractive index difference between the scatterer and solvent (Fig. S14). All nanoparticles we tested had lower refractive indexes than the PS standards (refractive index = 1.59) used to create our model. We attribute this as the key reason for the underestimation of the particle size. One possible approach to addressing this is to condition the model on both concentration and refractive index. It is trivial to extend the FiLM approach to

Table 4 Particle size estimation FiLM layer regression models' performance comparison with MAE, Root Mean Square Error (RMSE) and R^2

Method	MAE (nm) (Mean $\pm$ std)	RMSE (nm) (Mean $\pm$ std)	R^2 (Mean $\pm$ std)
PPSNet – MLP (ReLU)	9.53 $\pm$ 4.27	15.60 $\pm$ 7.58	0.99 $\pm$ 0.01
EfficientNet – MLP (sine)	11.60 $\pm$ 3.07	20.13 $\pm$ 5.95	0.98 $\pm$ 0.01
PPSNet (no conditioning)	22.25 $\pm$ 3.97	32.01 $\pm$ 6.95	0.93 $\pm$ 0.04
Polynomial regression	32.55 $\pm$ 6.67	47.81 $\pm$ 9.58	0.87 $\pm$ 0.03

Table 5 Predicted particle sizes with regression model and actual particle sizes, refractive indexes of polymers and silica, solutions concentrations are 0.4 (% v/v)

Polymer	Actual size (DLS – Av), nm	Polydispersity index	Predicted size (Av), nm	Refractive index
Silica	153	0.03	34	1.45
PDMAm$_{130}$-*b*-PDAAm$_{50}$	90	0.06	53	1.47
PDMAm$_{130}$-*b*-PDAAm$_{100}$	129	0.02	91	1.47
PDMAm$_{130}$-*b*-PDAAm$_{150}$	185	0.01	143	1.47
PMMA	127	0.04	45	1.49
PS	110	0.03	108	1.59

multiple inputs; however, our dataset has no variance in refractive index and thus would require data collection for polymeric nanoparticles with different refractive indexes to assess this claim. Nonetheless, this is an interesting area for future development and has the potential to improve the generalizability of the model and open it up to a wider range of use cases.

The molecular weight determination of the polymers using gel permeation chromatography (GPC) is often reported as a PS-equivalent molecular weight, but because of differences in hydrodynamic behaviour and Mark–Houwink parameters, these values can differ significantly from true values.[49] As our particle size estimation model was also trained on PS standards, it may be possible to adopt a similar approach to that used in GPC and report them as 'PS-equivalent particle sizes'.

We acknowledge that our method is unable to achieve the accuracy needed to calculate absolute particle size, polydispersity index or size distribution at the current state of the art, but it is not intended to replace DLS instruments. Rather, our main goal is to advance this technique as a non-invasive tool in systems that require quick estimation, using a simple low-cost setup within automation-driven laboratory environments.[50] Moreover, for industrial applications where the diversity of materials studied is low, it should be possible to build models trained on specific materials, when greater accuracy is required.

3.4 Limitations

The goal of this study was to explore the use of light scattering as a non-invasive tool to study polymer solubility and particle size. The underlying principle is light scattering, and therefore any parameter known to affect light scattering will also influence the accuracy of the predictions. Factors include refractive index (of both solvent and polymer), light polarization, absorption effects, and light wavelength. The type of scattering by the sample—Mie or Rayleigh scattering—is also important, particularly for the particle size estimation model. However, these are not the only factors that affect our model's performance, and we have tried to outline some of the key limitations below.

Hardware: the laser power, wavelength, camera selection, and experimental setup will influence the scattering intensity and/or the image brightness, thereby affecting the model's sensitivity and detection limits.

Samples:

• Scattering intensity: this depends on both sample concentration and polymer molecular weight/particle size, affecting detection limits and sensitivity (Fig. S15a and b). For the solubility classification, we found that using a concentration of ≥1% w/v improves the accuracy of the model, particularly for liquid-form polymers.

• Sample properties: fine powders produce extensive scattering and, potentially, confusion between colloidal and insoluble images (Fig. S15c and d). Liquid samples may be difficult to detect (Fig. S8), and when the density of the polymer is lower than the solvent, then the insoluble sample is located at the top of the vial rather than at the bottom (Fig. S15e).

• Polymer solubility: as polymers are large macromolecules, they can take a long time to solubilise/dissolve, so the time between sample preparation and data collection could have an impact on the results obtained. In this study we kept this time fixed at 2 hours, but it is also possible to collect at the data at different time points after mixing to understand the kinetic solubility behaviour of a particular system.

• Dissolution behaviour: gel formation at the bottom of the vial or bubble accumulation at the top present classification challenges (Fig. S15f).

• Sample colour: the impact of sample colour remains unknown, as all tested samples to date have been colourless. However, it is anticipated that colour may impose a maximum concentration limit.

Models:

• Solubility classification model: composition of training dataset could be improved by having more samples for the partially soluble class due to the diversity of samples in this class. Rare-case images, such as gelation and bubbles, should be increased as well as the samples currently predicted incorrectly by the model (Fig. S16, Tables S11 and S12). Furthermore, the scalability of this model to different vials size is also currently unexplored.

• HSP: though we wanted to keep the number of solvents used to less than 20 for practicality reasons, solvent selection could be investigated to improve accuracy of the HSP values determined.

• Particle size estimation: PPSNet was trained on particle sizes between 20 nm and 440 nm and it is unclear whether it will be able to extrapolate beyond this range. While it may be possible to address extrapolative generalization through careful regularization, ultimately, it may be necessary to collect more data if predicting a wider range of particle sizes is required. The training dataset was also limited by the PS particle size standards that were commercially available, collecting more data for samples in the 200–440 nm range, may also help to improve the performance of the model.

4. Conclusions

This study demonstrates the effective application of computer vision for classifying polymer solubility and particle size. We categorized polymer solubility and used these classifications in conjunction with a Genetic optimizer to determine the Hansen Solubility Parameters (HSP) for polymers. The approach facilitates the development of intelligent, user-friendly, and time-efficient systems for analysing polymer–solvent interactions. Despite the relatively small datasets, our robust system achieves accurate classifications and predictions with minimal manual input, underscoring the potential of machine learning to perform complex analytical tasks efficiently while reducing cognitive load for researchers. In future work, our particle size estimation method, based on a regression model for a particular polymer with low polydispersity indices, demonstrates a low prediction error, which provides a promising foundation for further research. We will also explore the implementation of these methods in fully automated robotic systems to enhance

efficiency and scalability. For example, because the approach is totally non-invasive, involving no sub-sampling, it could be easily retrofitted into existing automation workflows.

Author contributions

Conflicts of interest

There are no conflicts to declare.

Data availability

The code and data for Computer Vision for Polymer Characterisation using Lasers can be found at **https://doi.org/10.5281/zenodo.16536864**, Version v2. All other data supporting this article have been uploaded as part of the SI.

The SI provides additional details for the experimental and computational methods used in this work. See DOI: **https://doi.org/10.1039/d5dd00219b**.

Acknowledgements

Notes and references

1 A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila and F. Herrera, *Inf. Fusion*, 2020, **58**, 82–115.

2 X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski and T. Y.-J. Han, *npj Comput. Mater.*, 2022, **8**, 204.

3 I. Papadimitriou, I. Gialampoukidis, S. Vrochidis and I. Kompatsiaris, *Comput. Mater. Sci.*, 2024, **235**, 112793.

4 N. Qu, Y. Liu, M. Liao, Z. Lai, F. Zhou, P. Cui, T. Han, D. Yang and J. Zhu, *Ceram. Int.*, 2019, **45**, 18551–18555.

5 V. Venkatraman, S. Evjen, H. K. Knuutila, A. Fiksdahl and B. K. Alsberg, *J. Mol. Liq.*, 2018, **264**, 318–326.

6 H. Zhang, H. Fu, X. He, C. Wang, L. Jiang, L.-Q. Chen and J. Xie, *Acta Mater.*, 2020, **200**, 803–810.

7 L. Chen, H. Tran, R. Batra, C. Kim and R. Ramprasad, *Comput. Mater. Sci.*, 2019, **170**, 109155.

8 K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller and E. K. U. Gross, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2014, **89**, 205118.

9 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.

10 J. Youshia, M. E. Ali and A. Lamprecht, *Eur. J. Pharm. Biopharm.*, 2017, **119**, 333–342.

11 A. S. Anker, K. T. Butler, R. Selvan and K. M. Ø. Jensen, *Chem. Sci.*, 2023, **14**, 14003–14019.

12 S. Greenhill, S. Rana, S. Gupta, P. Vellanki and S. Venkatesh, *IEEE Access*, 2020, **8**, 13937–13948.

13 O. N. Oliveira and M. C. F. Oliveira, *Front. Chem.*, 2022, **10**, 930369.

14 R. Collobert and J. Weston, *presented in part at The Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008.

15 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, *arXiv*, 2020, preprint, arXiv:2010.11929, DOI: **10.48550/arXiv.2010.11929**, **https://arxiv.org/abs/2010.11929**.

16 A. Mohammed and R. Kora, *J. King Saud Univ. Comput. Inf. Sci.*, 2023, **35**, 757–774.

17 A. A. Khan, A. A. Laghari and S. A. Awan, *EAI Endorsed Trans. Scalable Inf. Syst.*, 2018, **8**, e4.

18 R. El-khawaldeh, M. Guy, F. Bork, N. Taherimakhsousi, K. N. Jones, J. M. Hawkins, L. Han, R. P. Pritchard, B. A. Cole, S. Monfette and J. E. Hein, *Chem. Sci.*, 2024, **15**, 1271–1282.

19 C. Yan, M. Cowie, C. Howcutt, K. M. P. Wheelhouse, N. S. Hodnett, M. Kollie, M. Gildea, M. H. Goodfellow and M. Reid, *Chem. Sci.*, 2023, **14**, 10304–10312.

20 J. H. Park, G. P. Dalwankar, A. Bartsch, A. George and A. B. Farimani, *Eng. Appl. Artif. Intell.*, 2024, **135**, 108603.

21 M. Walker, G. Pizzuto, H. Fakhruldeen and A. I. Cooper, *Digital Discovery*, 2023, **2**, 1540–1547.

22 P. Shiri, V. Lai, T. Zepel, D. Griffin, J. Reifman, S. Clark, S. Grunert, L. P. E. Yunker, S. Steiner, H. Situ, F. Yang, P. L. Prieto and J. E. Hein, *iScience*, 2020, **24**, 102176.

23 G. Pizzuto, J. de Berardinis, L. Longley, H. Fakhruldeen and A. I. Cooper, *presented in part at The 2022 International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, July, 2022.

24 M. Jeon, G. Yu, H. Choi, G. Kim and H. Hwang, *Sensors*, 2023, **23**, 5525.

25 S. Lu and A. Jayaraman, *Prog. Polym. Sci.*, 2024, **153**, 101828.

26 T. B. Martin and D. J. Audus, *ACS Polym. Au*, 2023, **3**, 239–258.

27 B. Narasimhan and N. A. Peppas, in *Polymer Analysis Polymer Physics*, Springer, Berlin, 1997, pp. 157–207.

28 C. M. Hansen, *Ind. Eng. Chem. Prod. Res. Dev.*, 1969, **8**(1), 2–11.

29 J. W. Strutt, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1871, **41**(271), 107–120.

30 G. Mie, *Ann. Phys.*, 1908, **330**, 377–445.

31 K. He, X. Zhang, S. Ren and J. Sun, *presented in part at The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, June, 2016.

32 M. Tan and Q. Le, *presented in part at The 36th International Conference on Machine Learning*, Long Beach, USA, June, 2019.

33 Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, *presented in part at The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June, 2022.

34 I. Loshchilov and F. Hutter, *arXiv*, 2017, preprint, arXiv:1711.05101, DOI: **10.48550/arXiv.1711.05101**, **https://arxiv.org/abs/1711.05101**.

35 L. N. Smith and N. Topin, 2017, *arXiv*, preprint, arXiv:1708.07120, DOI: **10.48550/arXiv.1708.07120**, **https://arxiv.org/abs/1708.07120**.

36 E. Perez, F. Strub, H. de Vries, V. Dumoulin and A. C. Courville, *arXiv*, 2017, preprint, arXiv:1709.07871, DOI: **10.48550/arXiv.1709.07871**.

37 M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron and R. Ng, *presented in part at the 34th Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, December, 2020.

38 V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell and G. Wetzstein, *presented in part at The 34th Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, December, 2020.

39 A. V. Malm and J. C. W. Corbett, *Sci. Rep.*, 2019, **9**, 13519.

40 C. M. Hansen, *On the Application of the Three Dimensional Solubility Parameter to the Prediction of Mutual Solubility and Compatibility*, Firg och Lack, 1967, vol. 13, no. 6, p. 132.

41 G. C. Vebber, P. Pranke and C. N. Pereira, *J. Appl. Polym. Sci.*, 2014, **131**, 39696.

42 M. Díaz de los Ríos and E. H. Ramos, *SN Appl. Sci.*, 2020, **2**, 676.

43 F. Gharagheizi, *J. Appl. Polym. Sci.*, 2007, **103**, 31–36.

44 K. G. Patel, R. K. Maynard, L. S. I. V. Ferguson, M. L. Broich II, J. C. Bledsoe, C. C. Wood, G. H. Crane, J. A. Bramhall, J. M. Rust, A. Williams-Rhaesa and J. J. Locklin, *ACS Sustain. Chem. Eng.*, 2024, **12**, 2386–2393.

45 *Hansen Solubility Parameters*, **https://www.hansen-solubility.com/**, accessed May 2025.

46 S. Venkatram, C. Kim, A. Chandrasekaran and R. Ramprasad, *J. Chem. Inf. Model.*, 2019, **59**, 4188–4194.

47 Z. Jia, J. Li, L. Gao, D. Yang and A. Kanaev, *Colloids Interfaces*, 2023, **7**, 45.

48 J. Bowman, C. Eades, M. Vratsanos, N. Gianneschi and B. Sumerlin, *Angew. Chem., Int. Ed.*, 2023, **62**, e202309951.

49 M. Netopilík and P. Kratochvíl, *Polymer*, 2003, **44**, 3431–3436.

50 P. A. Beaucage, D. R. Sutherland and T. B. Martin, *Macromolecules*, 2024, **57**, 8661–8670.