Progress in Polymer Science 153 (2024) 101828

Contents lists available at ScienceDirect

Progress in Polymer Science

journal homepage: www.elsevier.com/locate/progpolymsci

Machine learning for analyses and automation of structural characterization of polymer materials

Shizhao Lu[a,b], Arthi Jayaraman[a,b,c,*]

[a] Department of Chemical and Biomolecular Engineering, University of Delaware
[b] Data Science Institute, University of Delaware
[c] Department of Materials Science and Engineering, University of Delaware

ARTICLE INFO

Article history:
Received 31 January 2024
Revised 21 April 2024
Accepted 2 May 2024
Available online 3 May 2024

Keywords:
Machine learning
Microscopy
Small angle scattering
Automation
Polymer characterization
Structure data

ABSTRACT

Structural characterization of polymer materials is a major step in the process of creating materials' design-structural-property relationships. With growing interests in artificial intelligence (AI)-driven materials design and high-throughput synthesis and measurements, there is now a critical need for development of complementary data-driven approaches (e.g., machine learning models and workflows) to enable fast and automated interpretation of the characterization results. This review sets out with a description of the needs for machine learning specifically in the context of three commonly used structural characterization techniques for polymer materials: microscopy, scattering, and spectroscopy. Subsequently, a review of notable work done on development and application of machine learning models / workflows for these three types of measurements is provided. Definitions are provided for common machine learning terms to help readers who may be less familiar with the terminologies used in the context of machine learning. Finally, a perspective on the current challenges and potential opportunities to successfully integrate such data-driven methods in parallel/sequentially with the measurements is provided. The need for innovative interdisciplinary training programs for researchers regardless of their career path/employment in academia, national laboratories, or research and development in industry is highlighted as a strategy to overcome the challenge associated with the sharing and curation of data and unifying metadata.

1. Introduction

Establishing structure-property relationships for macromolecular materials (e.g., block copolymers [1-6], polymer blends [7,8], polymer nanocomposites [9-21]) has been the subject of active research within polymer science and engineering for many decades. It has been shown for various classes of polymeric materials that in addition to the choices of polymer chemistry and architecture, their assembled structures, which could be hierarchical and multiscale in many cases, dictate the ultimate properties of materials composed of these polymers. With recent advances in polymer synthesis and innovative polymer processing techniques, the variety of equilibrium and non-equilibrium structures accessible within the materials has grown exponentially. Identification of the optimal structure(s) that give rise to the desired properties and that require minimal costs and efforts to scale up to industrial level production drive the need for high-throughput experimentation.

Abbreviations: AFL, Autonomous formulation laboratory; AFM, Atomic force microscopy; AI, Artificial intelligence; CASGAP, Computational approach for structure generation of anisotropic particles; CDF, Cumulative distribution function; CNN, Convolutional neural network; CREASE, Computational reverse engineering analysis of scattering experiments; CRIPT, Community resource for innovation in polymer technology; CTAB, Cetyltrimethylammonium bromide; DKL, Deep kernel learning; FAIR, Findable, accessible, interoperable, reusable; FCN, Fully connected network; FTIR, Fourier transform infrared spectroscopy; GA, Genetic algorithm; GISAS, Grazing incidence small-angle scattering; GISAXS, Grazing incidence small-angle X-ray scattering; GNoME, Graph networks for material exploration; GRF, Gaussian random field; HAADF, High-angle annular dark-field; iPP, Isotactic polypropylene; LBNL, Lawrence Berkeley National Laboratory; LLM, Large language models; ML, Machine learning; MOF, Metal Organic Framework; MSDNets, Multi-scale dense networks; MSE, Mean squared error; NEXAFS, Near edge X-ray absorption fine structure spectroscopy; NLP, Natural language processing; NMR, Nuclear Magnetic Resonance; P3HT, Poly(3-hexylthiophene); P4VP, Poly(4-vinyl pyridine); PCA, Principle component analysis; PEO, Polyethylene oxide; PLS, Partial least squares; PMMA, Poly(methyl methacrylate); PS, Polystyrene; PVAc, Poly(vinyl acetate); R&D, Research and development; SANS, Small-angle neutron scattering; SAS, Small-angle scattering; SAXS, Small-angle X-ray scattering; SEM, Scanning electron microscopy; STEM, Scanning transmission electron microscopy; TEM, Transmission electron microscopy; t-SNE, t-distributed stochastic neighbor embedding; UMAP, Uniform manifold approximation and projection; UV, Ultraviolet; VAE, Variational autoencoder; WAXS, Wide-angle X-ray scattering; XANES, X-ray absorption near edge spectroscopy; XRD, X-ray diffraction.

* Corresponding author.
E-mail address: arthij@udel.edu (A. Jayaraman).

https://doi.org/10.1016/j.progpolymsci.2024.101828
0079-6700/

With high-throughput experiments comes the associated push for fast and preferably automated analyses of the synthesized materials and their characterization, and in turn, the need for development of appropriate machine learning (ML) workflows.

Development and use of ML workflows in (non-polymer) materials domains have already made great strides [22-25]. An example of the power of ML for materials design is the latest work from Google DeepMind which used graph networks for materials exploration (GNoME) [26] and predicted ~380,000 new theoretically stable inorganic crystalline materials [27]. Researchers from Lawrence Berkeley National Laboratory (LBNL) then synthesized more than 40 new materials in just 17 days in A-Lab [28]; A-Lab is a robotic (automated) laboratory with capability for synthesis, characterization, and analysis of synthesized inorganic crystalline samples. Similarly, significant advances have been made in ML-based analysis of inorganic/small molecule materials characterization results from microscopy [29-37], scattering [38-40] and spectroscopy [38,40-42]. In contrast to the many successes in the inorganic and small-molecule organic materials domains, the use of ML, high-throughput experimentation, and automation for synthesis and characterization within polymers and soft materials is still in initial stages. To achieve similar success for AI and ML in the field of polymer science and engineering requires investment into creating polymer databases that enable polymer informatics [43,44], high-throughput experimentation [45-47], characterization techniques [48-52], ML, and data science methods customized to visualize and analyze the type of data seen with polymer materials [46,53-58].

In this Review we present noteworthy studies aimed at development and application of ML models and ML based workflows specifically aimed at fast and automated analyses of polymer structural characterization data. We hope that these studies inspire the readers to either develop new ML models and methods or adapt these published methods for their own polymer characterization analyses. We also encourage the readers to look at previously written review articles and perspectives on other relevant subjects that we do not cover, within the broader topic of *machine learning for polymer science and engineering* namely, polymer informatics [43,44,59,60], featurization of polymers use in ML models [61], automation in polymer synthesis [62,63], and natural language processing for extracting polymer data from literature [64,65].

This review article is organized as follows: Section 2 presents relevant background information of three commonly used classes of experimental techniques for structural characterization of polymeric materials (microscopy, scattering and spectroscopy). Section 3 presents a review of ML methods applied for analyses, interpretation, and automated data acquisition of such characterization of polymeric materials. Section 4 provides current challenges and potential future directions for accelerating progress on the topic of this Review, with the aid of open-access database curation, unified metadata, and interdisciplinary education on relevant topics.

2. Common structural characterization in polymer materials

In this section we provide a high-level summary of the three commonly used classes of techniques for structural characterization of polymers, without many details of the instrumentation or the sophisticated protocols that researchers follow to use the instrumentation in these techniques correctly. Instead, our emphasis is on describing the types of data generated from these characterization techniques and the types of physical/chemical information one can obtain by analyzing that raw or processed data. We believe that the attention to the type of data and the information gathered from the data is necessary for the reader to consider suitable types of ML models for interpreting such data.

2.1. Microscopy

To characterize the morphology of polymer materials, the commonly used microscopy techniques are Optical Microscopy, Scanning Electron Microscopy (SEM), Transmission Electron Microscopy (TEM), Scanning Transmission Electron Microscopy (STEM), cryo-TEM, and Atomic Force Microscopy (AFM); we direct the reader to a recent review on electron microscopy for soft materials. [66] Briefly, SEM and TEM provide atomic-level to microstructure images by using electrons as the imaging radiation source. [67] The use of electrons as the source allows for images with spatial resolution as low as tens of picometers in contrast to the hundreds of nanometers resolution obtained using photons in optical microscopy. In contrast to electron microscopy, optical microscopy is used for imaging colloidal materials when the relevant lengths scales are between 100 s of nm and 100 s of microns. SEM and TEM differ in the way the techniques work (e.g., thickness of the sample required, sample preparation, cost, expertise for using the techniques) as well as the types of images they provide. SEM images provide information about the composition and roughness of the polymer film. Typically, in composites or blends, one of the components (polymers or additives) forms the continuous phase (the matrix) and the other minority component(s) forms the dispersed phases. Via SEM images one can delineate the shapes, sizes, and spatial distribution of these phases and in the case of thin films, the surface topography as well. TEM images provide higher resolution information as compared to SEM images and the types of structural details one can obtain from TEM images include molecular-level structure, dimensions, and shapes of the nanoscale objects in a polymer matrix, crystalline arrangements, and defects. In STEM, a mode of TEM, the incident focused beam scans across the specimen and the transmitted signal are collected as a function of the beam location as it rasters across the sample. One can obtain the atomic arrangement, orientation, and crystalline or semi-crystalline structure of polymer materials from STEM. Cryo-TEM is a subclass of TEM that enables imaging soft materials in solvated environment by rapidly freezing the solvated sample. Cryo-TEM reduces sample damage during preparation and from the electron beams and has gained attention as a valuable method for determining macromolecular solution structure. Another complementary technique to SEM and TEM is AFM which gives information of the surface topography (e.g., roughness) as well as the hardness/softness of the probed domains in the polymer material; typical length scales probed range from a few Angstroms to a few microns.

The type of data one obtains from these above microscopy techniques are two-dimensional images where the pixels intensity at various positions in the images conveys the intended physical information about the structure of the polymer sample being probed. Examples of raw data (colored or grayscale) from TEM, SEM and AFM are shown in Figs. 1a, 1b, and 1c, respectively. One has to remember that these images are 2D projections of complex structural features that may have irregular shapes, asymmetric surfaces, and heterogeneity in the 3D structure. The quality of microscopy images is subject to the different sample staining techniques, extent of in-focus vs. out-of-focus, and contrast between object of interest and the background.

2.2. Scattering

For investigating multi-scale 3D amorphous polymer structures ranging from 10 Å up to few microns, small-angle scattering (SAS) is a powerful technique. [48,68–76] Small-angle neutron scattering (SANS) and small-angle X-ray scattering (SAXS) have been used extensively in the polymer science and engineering community, for example, to study domains within microphase separated struc-

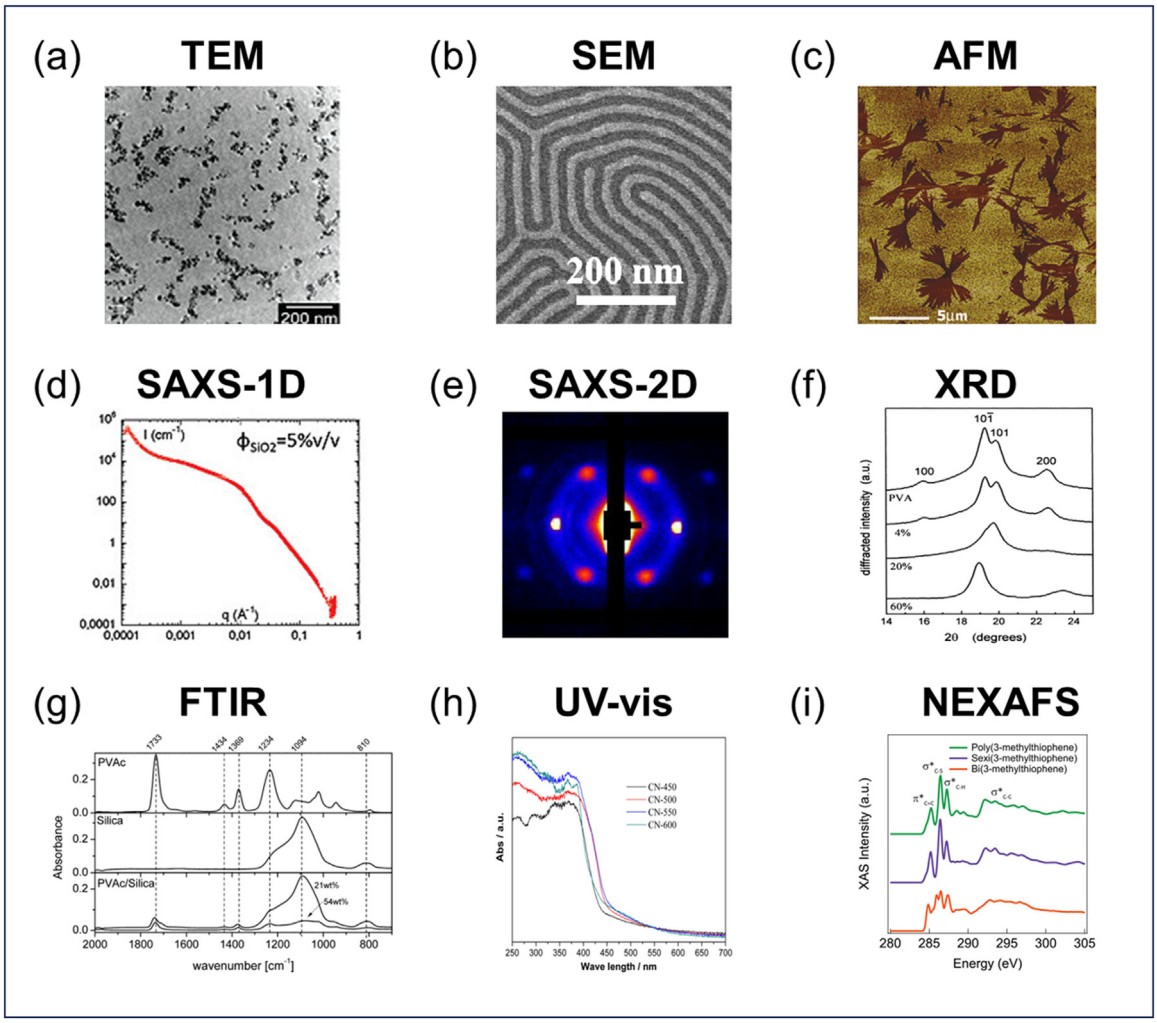

Fig. 1. Examples from polymer literature solely to depict the typical presentation of processed data from the various characterization techniques discussed in this review; the values, tick labels, and scale bars are specific to these examples and not general to all such measurements in the field of polymers. (a)TEM characterization of silica-polystyrene nanocomposite morphology. Reprinted with permission from ref. [88]. Copyright 2010 American Chemical Society. (b) SEM characterization of polystyrene−polydimethylsiloxane block copolymer thin film morphology. Reprinted with permission from ref. [89]. Copyright 2007 American Chemical Society. (c) AFM characterization of neat poly(vinyl alcohol) (PVA). Reprinted with permission from ref. [90]. Copyright 2001 American Chemical Society. (d) 1D SAXS profile of silica-polystyrene nanocomposite with 5 % volume fraction inclusion of silica nanoparticles. Reprinted with permission from ref. [88]. Copyright 2010 American Chemical Society. (e) 2D SAXS profile of polystyrene-b-poly(methyl methacrylate) (PS-b-PMMA) block copolymer thin film. Reprinted with permission from ref. [91] 2023 Creative Commons Attribution License. (f) XRD profile of neat PVA and PVA-sodium montmorillonite (MMT) nanocomposite with different MMT loading. Reprinted with permission from ref. [90]. Copyright 2001 American Chemical Society. (g) FTIR spectra of pure poly(vinyl acetate) (PVAc), pure silica nanoparticle, and PVAc absorbed onto silica nanoparticle surface. Reprinted with permission from ref. [92]. Copyright 2013 American Chemical Society. (h) UV–vis spectra of polymeric graphitic carbon nitride at different preparation temperatures. Reprinted with permission from ref. [93]. Copyright 2014 American Chemical Society. (i) NEXAFS spectra of dimer, oligomer, and poly(3-methylthiophene). Reprinted with permission from ref. [94]. Copyright 2017 American Chemical Society.

tures in block copolymers, dispersion or aggregated states of particles in polymer nanocomposites, and network structure in polymer gels. [48,51,77] The raw SAXS and SANS data is intensity (I) versus magnitude of scattering wavevector q and azimuthal angle. For polymer systems without any anisotropy in spatial arrangements, the scattering data is averaged azimuthally to create one-dimensional (1D) scattering profile, $I(q)$ vs. magnitude of wavevector q. (Fig. 1d) Traditionally, the $I(q)$ vs. q profile is analyzed in one of two ways: First, by fitting the data to analytical models (e.g., core-shell [78,79], core-multishell [80,81] etc.) developed for various canonical polymer structures, on user-friendly analysis packages like SASVIEW [https://www.sasview.org/] or SASFit [https://sasfit.org/], or second, by using shape-dependent analyses (e.g., Kratky plot, $q^2\ I(q)$ vs. q). [52,75,76]

For samples where one expects to see anisotropic structures (e.g., liquid crystalline order within one or more domains of the polymer blend, anisotropic aggregates of nanoparticles or extruded fibers within polymer nanocomposites), the steps taken during measurement as well as analyses is more complex than that for isotropic structures. In such cases, first the scattering measurements have to be made along carefully selected orientations (with some domain knowledge of which orientations are strategically better than others depending on direction of anisotropy) and then the resulting data has to be analyzed as a 2D scattering profile, without averaging over all or sections of azimuthal angles. (Fig. 1e) Ways to analyze such 2D profiles are quite complex as compared to 1D profiles. 2D profiles can be analyzed using packages like GRASP [https://www.ill.eu/users/support-labs-infrastructure/software-scientific-tools/grasp] and DAWN [https://dawnsci.org/]. The qualitative analyses of such images involve looking for diffuse halos (implying disorder or weak ordering), patterns of dots with high intensity like Fig. 1d (indicating ordered domains), and

asymmetry in the image (implying presence of some anisotropic arrangement along the direction of input beam). The quantitative interpretation of the 2D profiles traditionally relies on averaging sections of the 2D profile into 1D profiles that are then fit to shape dependent or independent models as described above.

While amorphous polymers usually lack precise order and spatial arrangement of atoms, semi-crystalline polymers exhibit periodic and precise atomic and polymer segment-level arrangements. To discern this atomic arrangement and extent of crystallinity in semi-crystalline polymer materials, X-ray diffraction (XRD) and wide-angle X-ray scattering (WAXS) techniques are used. [82,83] The plot in Fig. 1f shows the typical processed XRD data one analyzes to interpret structural information. The presence of peaks, their intensity and location along x-axis, are then interpreted to structural information about the type of crystalline order. 1D WAXS profiles look similar to the SAXS profiles of materials with more order (e.g., Fig. 1d), however, the length scales probed by WAXS are smaller than SAXS and related to atomic-level spatial arrangements.

2.3. Spectroscopy

Spectroscopy techniques such as Fourier Transform Infrared (FTIR) spectroscopy, UV–vis spectroscopy, and X-ray absorption spectroscopy are useful for studying the atomic and electronic composition of polymer materials. [84-86] Depending on the wavelength of the incident light wave, spectroscopy methods can detect or identify different chemical species or functional groups. FTIR spectroscopy is often used to identify functional groups present in molecules that have signature absorption peaks at tabulated vibrational frequencies in the FTIR spectrum. For UV–vis spectroscopy, researchers look for certain ratios of different absorption peaks present in the UV–vis spectrum for identification of the molecules or polymer materials. X-ray Absorption Near Edge spectroscopy (XANES) or Near-Edge X-ray Absorption Fine Structure (NEXAFS) can generate information of the electronic state, coordination environment, oxidation state of atoms or molecules from the X-ray absorption peaks.

Typical data from spectroscopic measurements are in the form of 1D vector array or 2D image containing a plot of the intensities of the measured physical property vs. the light beam wavelength (Figs. 1g and 1h) or energy (Fig. 1i). Analyses of these measurements require extensive expertise assigning features to known chemical species and in many cases comparison of the spectra of the new sample to a reference spectrum (or spectra) in a database. For example, in Ref. [87], focused on NEXAFS techniques for chemical analysis of polyurethanes, the authors show how spectra of model polymers provide reference standards for the quantitative analysis ('speciation') of polyurethane polymers (e.g., quantify the amounts of aromatic and aliphatic components of polyurethanes). We quote a section from Ref. [87]: "For accurate quantitation, well-characterized NEXAFS spectra of carefully chosen models of the polymer components is required. For a blend of two or more homopolymers (i.e., polystyrene/poly(methyl methacrylate)), the analytical models can simply be the individual homopolymers. For quantitation of components in a random block copolymer (i.e., styrene acrylonitrile), the spectra of the homopolymers (polystyrene and polyacrylonitrile) can be used as component models if the polymer and monomer spectra are additive. Polyurethane polymers are complex, and care must be taken in the choice of analytical models." The need for an established reference spectra database for polymer materials can be a barrier to analyses. Machine learning methods are deemed as attractive alternatives to not only circumvent the challenge of reference spectra but also accelerate the data analysis process.

3. Applications of ML models and workflows

In the polymer science and engineering community, traditional analyses of the results from microscopy, scattering, and spectroscopy relies on experts with knowledge and experience to interpret *qualitative* (e.g., identifying shapes of domains, types of ordered structure, or presence or absence or anisotropy in structures) and *quantitative* (e.g., domain sizes, signature peaks of functional groups) information of structural characterization data. We believe that when the dataset size is small and practical for manual analyses and interpretation, these traditional approaches will continue to remain the best option. When the amount of data being generated is too fast and too large for manual analyses, there is a clear need for development of ML workflows that enable fast and objective analysis of polymer structural characterization data. In the next sub-sections, we present noteworthy ML workflow developments in microscopy, scattering, and spectroscopy and divide these ML developments by the specific analysis tasks they accomplish. For readers with minimal experience with ML jargon, we provide in Table 1 basic definition of common ML terms (in alphabetical order) that we use extensively in the following sections. We choose to present these definitions in a table rather than explain each term as it is occurring in the following sections, to maintain smooth flow of information with minimal disruptions.

3.1. Applications of ML in microscopy

As the typical output of SEM, TEM, STEM, cryoTEM, and AFM measurements are images, it is logical to consider successful ML models that have been applied for automated analyses of images in other fields (e.g., biomedical images [95,96], facial recognition [97], autonomous driving [98,99]) and extend those models to learning features of polymer characterization images as well. Regardless of the tasks that one wishes to accomplish, the ML model has to understand the information (e.g., pattern shapes and sizes, textures, light vs. dark regions) present in the microscopy image to connect those patterns with a specific type of morphology/domain shape/size or physical/chemical feature(s) relevant to polymer science. Convolutional Neural Network (CNN) is one such deep learning approach that has been used successfully to learn, extract, and encode information from an image. [100] CNN typically consists of a hierarchical structure of convolutional layers placed in between an input layer and an output layer. [101] Each convolutional layer consists of multiple filters each of which encodes pieces of information by conducting convolution on a small perception field of the entire image. The input image is sequentially encoded into smaller and smaller feature maps through the hierarchical convolutional layers. The last feature map is used as input for training the classifier which is typically a couple layers of fully connected networks after the hierarchical convolutional layers. As one goes deeper into the convolutional network, the information learned in the feature maps goes from being specific and local to being abstract and global. The hierarchical convolutional architecture of CNN enables more generalizable learning of abstract and complex patterns in images and minimizes the risk of vanishing gradients and exploding gradients problems seen previously in deep fully connected neural networks. [100] Some of the established CNN architectures include: LeNet-5 [102], AlexNet [103], VGG16 [104], InceptionV3 [105], ResNet50 [106], Xception [107], Inception-ResNet-V2 [108], ResNeXt-50 [109], MobileNet [110], and EfficientNet [111]. These CNN architectures differ in the type, number of layers, and the size of trainable parameters. For additional information on concepts, architectures, applications of CNN we direct the reader to a recent review [112].

Table 1
Guide to Machine Learning Terms and Definitions.

Term	Definition
Autoencoder	An artificial neural network model used for learning abstract, condensed, low-dimensional representations of high-dimensional data. An autoencoder consists of an encoder and a decoder.
Classification	The action or process of grouping things according to shared qualities or characteristics.
Classifier	A classifier is any machine learning model that can tackle a classification task.
Decoder	An artificial neural network model for learning the forward relationship between a low-dimensional input and a high-dimensional output.
Dimensionality reduction	A category of methods that does transformation of high-dimensional data into low-dimensional representations / latent variables / feature maps while retaining key intrinsic properties of the original data.
Encoder	An artificial neural network model for learning the forward relationship between a high-dimensional input and a low-dimensional output.
Exploding gradient	The problem encountered in training neural network model when there is large or not-a-number (NaN) gradient updates to the neural network model weights. The trained model will produce erroneous predictions.
Feature engineering	The process of selecting, wrangling, and transforming raw / unprocessed data into features that can be used for more efficient learning / better training and testing performances in supervised learning, often require guidance from domain expertise.
Feature map / vector	A feature map / vector is a condensed / distilled set of features (or low-dimensional data representation) for the more complex and high-dimensional data.
Filter	A small matrix of learnable weights in a convolutional neural network that would slide over the input image data to apply matrix multiplications.
ImageNet dataset	An open-access dataset containing more than one million photographic images of everyday objects (like cats, dogs, houses, cars) split into 1000 categories.
k-nearest neighbors	A non-parametric, supervised learning method, which uses proximity and majority vote of k nearest neighbors to decide the label of unlabeled points in space.
Latent feature / variable / space	A latent feature / variable is the low-dimensional representation obtained through dimensionality reduction processes. A latent space is a collection of such latent features / variables. Latent feature / variable is synonymous with feature map / vector.
Layer	A layer is a building block of a neural network that is a collection of vectors or matrices containing learnable parameters known as weights.
Principle Component Analysis (PCA)	A linear dimensionality reduction method.
Reinforcement Learning	Training of machine learning models with a balance of exploration and exploitation for long-term reward optimization
Segmentation	In the context of image learning, segmentation is the task of delineation of objects of interest from the image background.
Self-supervised learning	Training of machine learning models without manually labeled data; machine-generated labels are used during training.
Semi-supervised learning	A data-efficient way to conduct supervised learning by combining self-supervised learning / unsupervised learning of the machine learning model with supervised learning using the trained model. The data used for the two parts of learning are often different datasets but can also be the same dataset.
Skip connection	Skip connections are connections made between non-adjacent layers (separated by one or more layers) in a residual convolutional neural network for learning of the residual (output vs. input) of the layer for improved performance.
Supervised learning	Training of machine learning models with manually labeled data.
t-distributed Stochastic Neighbor Embedding (t-SNE)	A non-linear dimensionality reduction method
Transfer learning	The use of a trained machine learning model to train and learn on a different dataset for improved performance and reduced amount of the training data needed.
Uniform Manifold Approximation and Projection (UMAP)	A non-linear dimensionality reduction method.
Unsupervised learning	Training of machine learning models without manually labeled data.
Vanishing gradient	The problem encountered in training of neural network model when there is exceedingly small or zero gradient updates to the neural network model weights. The model stops learning.
Variational Autoencoder (VAE)	A modified version of autoencoder that maps the low-dimensional representations to a probabilistic multi-dimensional Gaussian distribution.

In the following sub-sections, we describe the various analysis tasks one would want to accomplish using ML models when they are analyzing polymer characterization microscopy images.

3.1.1. Task: classification

Classification of microscopy images in the context of polymer science and engineering usually involves these types of tasks: labeling an image as corresponding to one or other type of morphology (e.g., lamellar vs. spherical), identifying shapes of one or more domains in the image (e.g., circular, elliptical, fractal, etc.), identifying particle or aggregated chains' orientations (isotropically arranged, dispersed, aggregated, orientationally aligned, fibrillar, etc.). The majority of the literature where ML models have been trained for classification tasks has been in the area of nanomaterials; next, we survey some of those recent works that we consider to be readily extendable to images from polymer materials as well.

Modarres et al. have used a CNN model that was pretrained on the ImageNet dataset [113] to classify SEM images belonging to different nanomaterial subcategories like nanoparticles, nanofibers, porous structures, films, coated surfaces, powder, etc. [114] The classification performances of individual subcategories has revealed some categories with significantly lower classification accuracy. The lower classification accuracy has been explained by one of these two reasons: some images from distinct categories were too similar (leading to poor classification) or some images contained elements of multiple categories (leading to difficulty in classification). The model achieved an overall accuracy of 90 % with their nanostructure classification workflow.

Xu et al. have used CNN to classify copolymer microstructures from AFM images. [115] In their study, they synthesized styrene-co-(n-butylacrylate) copolymers with five different architectures- random, diblock, triblock, linear-gradient and V-

shaped gradient- and used microscopy and differential scanning calorimetry to characterize microstructure and thermal properties (e.g., glass transition temperature, T_g) of spin-coated films of each of these polymers. In their quest to use ML to connect macroscopic (thermal) properties, microstructure, and copolymer architecture, in one step, they employed CNNs for classifying the AFM images into their respective microstructures. The authors discussed the difficulty in training CNNs (with many parameters to be learned) with small data sets; the smaller size of the data set is a universal problem within soft materials. So the authors took advantage of transfer learning; they used a model pre-trained on a much larger simulated microstructure image dataset [116] (what the authors call as "introducing domain knowledge into the ML model") to learn the specifics of the experimental data by fine-tuning the pre-trained model's parameters with the data from their own experiments. Transfer learning is a viable approach when dealing with small image data sets common to academic research labs without high-throughput instrumentation.

Another study that has used transfer learning in a semi-supervised manner to address the difficulty of training from small-size image datasets is the work by Lu et al. to classify TEM images to identify nanowire morphologies. Lu et al. have developed a semi-supervised transfer learning workflow to facilitate automatic and label-efficient classification of protein / peptide nanowire morphologies from TEM images (Fig. 2a). [117] Semi-supervised learning refers to ML workflow consisting of a self-supervised learning part (i.e., no manual labels required, target label can be derived by the machine from the input data) and a supervised learning part (i.e., process where images and their corresponding labels are provided by the user to the model). Lu et al.'s workflow has performed transfer learning using an image encoder of ResNet50 architecture [106] that was trained via self-supervised learning on generic microscopy images, as a feature encoder for their task-specific images

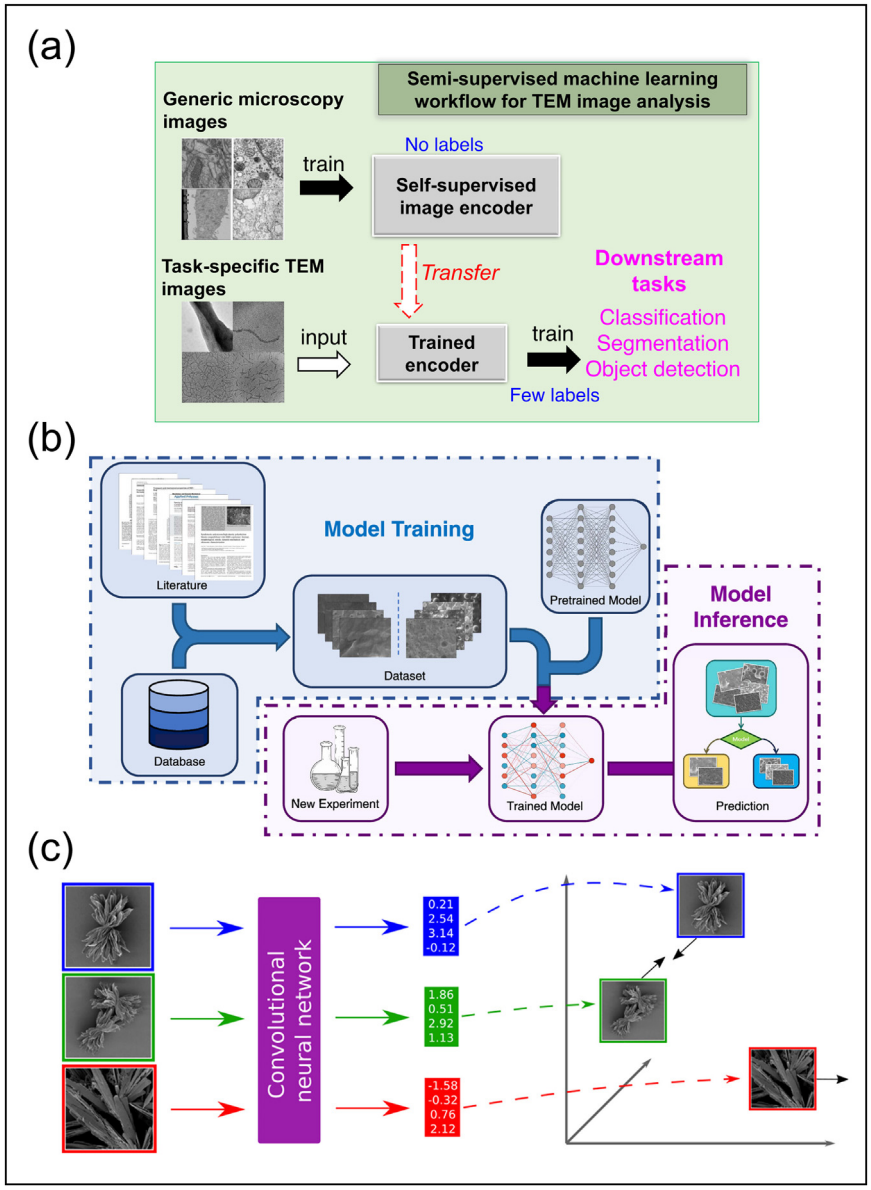

Fig. 2. Machine learning classification of phases or morphologies of polymer materials from microscopy data. (a) Semi-supervised machine learning workflow for automated classification of nanowire morphologies from TEM images. Reprinted with permission from ref. [117] 2022 Creative Commons Attribution License. (b) Machine learning workflow for classification of the miscibility of polymer blends. Reprinted with permission from ref. [120]. Copyright 2023 American Chemical Society. (c) Machine learning classification of chirality of nanoparticles from SEM using CNN. Reprinted with permission from ref. [121]. Copyright 2023 American Chemical Society.

containing nanowire morphologies. They showed classification accuracy above 90 % and their workflow requires fewer than ten labeled images for classification using the encoded features as input data for the downstream classifier. They also demonstrated the ability to generalize their workflow to classification of morphologies of nanoparticle assembly and identification of virus types from TEM images found in open-access datasets [118,119].

Liang et al. have utilized transfer learning methods to classify miscibility of polymer blends from SEM images in Fig. 2b. [120] More than five hundred SEM images depicting either miscible or immiscible polymer blends were collected from literature and used for training and testing their ML methods. They used and compared three different CNNs for learning and classification of SEM images, all of which showed better performance when applying transfer learning methodology. They achieved 94 % accuracy on the test dataset with the best classification model.

While not directly connected to polymers, a study by Visheratina et al. using Siamese learning protocol to create feature vectors of SEM images containing chiral nanoparticles (Fig. 2c) is worth sharing. [121] Chirality is an important aspect for characterization of hierarchical structure of semi-crystalline polymer assemblies. In Siamese learning, the feature vectors of images belonging to the same chirality (left-hand or right-hand) have similar values. Using this idea of Siamese learning, then the authors trained a nearest neighbor classifier to classify the chirality class of the nanoparticle in the SEM images using the feature vectors as input with 93 % accuracy.

3.1.2. Task: particle detection / segmentation / shape analysis

Identifying the spatial arrangement of nanoparticles in a polymer matrix/medium is a common structural characterization task in polymer nanocomposites sub-field because the physical properties of polymer nanocomposites have been shown to be dictated by the nanoparticles' spatial arrangement. [9,13,14,122,123] Understanding if the nanoparticles are in a dispersed state or if they are aggregated, and if so in what arrangement, requires ML models capable of particle detection, segmentation, and shape analysis. In this sub-section we describe ML model development and application, regardless of the materials science field, which have been successful specifically for these tasks – particle detection, segmentation, and shape analysis.

Ziatdinov et al. have demonstrated the potential of using deep learning for atomic detection in STEM images (Fig. 3a). [124] They have also shown that one can track atomic reorientation using their deep learning workflow; such ML methods could be extended for analysis of time series of electron microscopy images.

Qu et al. have used a deep learning model for detecting silica and iron oxide nanoparticles in a polymer (polymethylmethacrylate, PMMA or polyethylene oxide, PEO) matrix from TEM images. [125] They broke down a large TEM image into smaller images and extracted the nanoparticle positions and sizes automatically using their particle detection ML model. Using the location and size information of the nanoparticles, they quantified the assembly state of the nanoparticles in the polymer matrix by calculating a surrogate parameter termed as the particle number fluctuations. In a subsequent work, Qu et al. applied that same particle detection workflow to detect and track nanoparticles in semicrystalline polymer nanocomposites from AFM images [126].

Yao et al. have integrated real-time ML segmentation analysis models with liquid-phase TEM videos to study the diffusion, kinetics, and assembly of colloidal nanoparticles. [127] They used a popular neural network architecture U-Net [128] that in the past had been shown to do segmentation tasks for biomedical image analyses well. [129] Compared to conventional algorithms that require users to select *a priori* a threshold value in intensity to mark as pixel or region as belonging to one domain or another, U-Net does not require a threshold value to be selected manually as input. Instead, U-net automatically decides key features from the TEM image that determine what domain should be assigned to each pixel. These authors also showed that U-Net can deliver robust segmentation results even when the images are blurry or have low signal-noise ratio.

After particle detection and/or segmentation tasks are completed, one can also pursue automated shape analysis of the detected nanoparticles or segmented regions (Figs. 3b and 3c). [118,130-134] For readers interested in learning more about U-Net and use of the U-Net architecture for image analyses tasks, we recommend a recent comparison of model performances of U-Net [128] and other versions of U-Net by Saaim et al. [135] Five adaptations of U-Net including: R2U-Net [136], Attention U-Net [137], BDC U-Net [138], U-Net++ [139] and U-Net 3+ [140] have been compared alongside with U-Net on segmentation of nanoparticles from TEM images.

3.1.3. Task: automation in image acquisition and analyses

While the previous two sub-sections describe the ML model development for tasks to be done after the microscopy images have been obtained, one can also use computational methods to automate image acquisition prior to analyses. Here we highlight some of those types of studies and their accomplishments.

Touve et al. have reported a high-throughput TEM experiment to map the phase diagram of block copolymer amphiphile assisted by automated image analysis [141]. By varying the sample block copolymer compositions, they saw formation of spherical micelle, wormlike micelle, or vesicle morphologies. They used an automated image acquisition software SerialEM [142] to automate the high-throughput generation of high-resolution montages over a large area of 45 samples. Following that, statistical shape analysis was applied to quantify block copolymers' assembled structures' shapes and sizes. A robust image binarization method [143] was used for segmentation of micelle particles from the background. Subsequently, an elastic curve-based shape clustering algorithm [144] was used for categorizing different particle shapes into spherical micelle, wormlike micelle, and vesicles.

Krull et al. have developed a ML framework called DeepSPM for automating acquisition of high-quality scanning probe microscopy (SPM) images. [145] DeepSPM includes an active learning of regions in the image/sample that have points of interest. They trained a CNN model to assess the quality of SPM images in real-time. When the SPM image quality was assessed as poor by the CNN model, then a deep reinforcement learning agent would adjust the condition of the probe to obtain higher-quality images. Such models can be valuable in acquisition and classification of polymer film data continuously during a multi-day long experiment and for automatically correcting the probe as the experimental conditions (e.g., temperature, solvent composition) vary.

To develop a method friendly to beam-sensitive materials, Roccapriore et al. have applied Bayesian optimization methods to High-Angle Annular Dark-field (HAADF) STEM image acquisition that can map the domains through adaptive sampling. [146] They used small patches of atomic-resolution regions as input features and a Deep Kernel Learning (DKL) model [147] to predict experimental diffraction patterns acquired at these image locations. The DKL model combines the deterministic deep neural network and the stochastic Gaussian Process model; the Gaussian Process operates on low-dimensional representation of the microscopy image patches learned by the deep neural network. The DKL model predicts functional responses such as the diffraction patterns and gives uncertainty of the responses as part of the Gaussian Process model. For samples that are sensitive to electron beam dosage, they accomplished efficient sampling for different systems of ex-

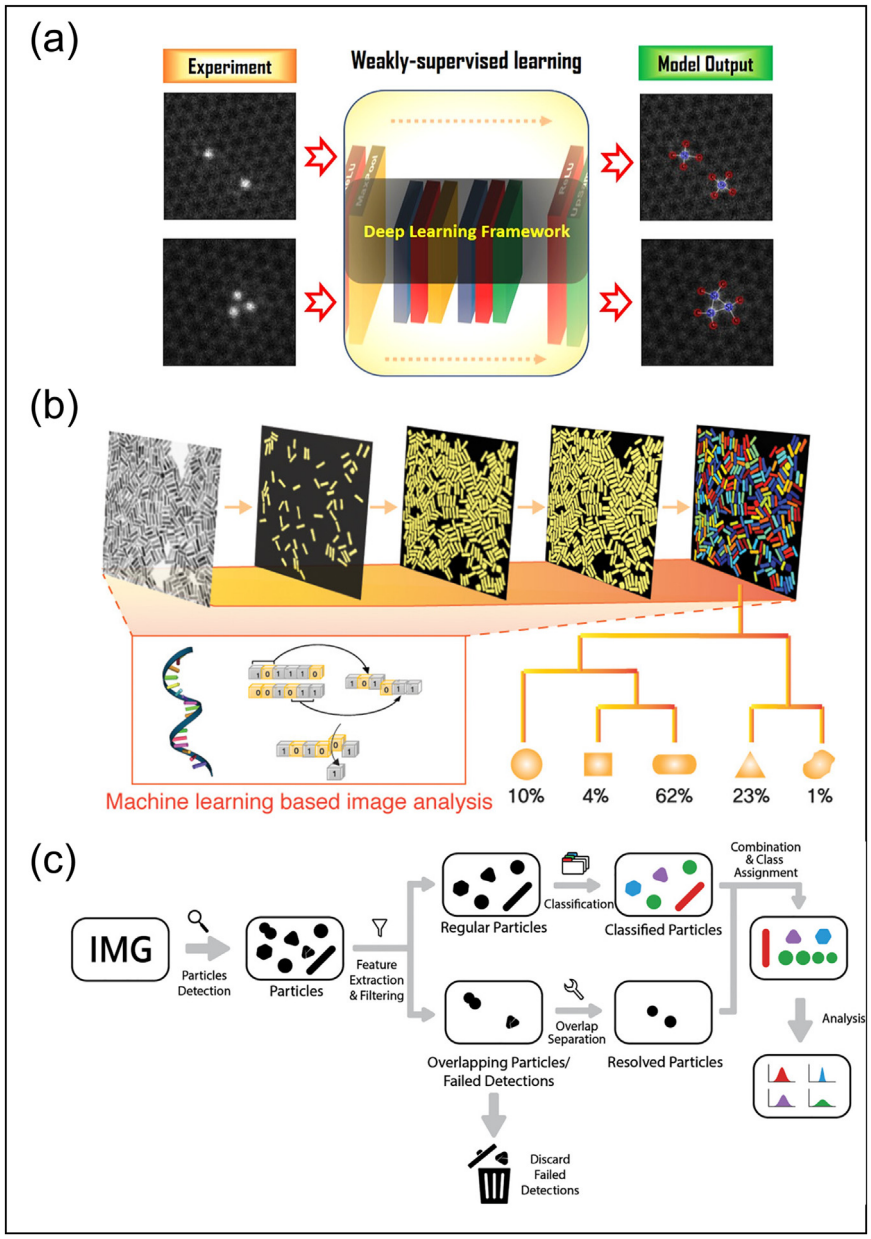

Fig. 3. Machine learning particle detection, segmentation, and shape analysis of nanoparticles from microscopy data. (a) Atomic resolution particle tracking machine learning workflow. Reprinted with permission from ref. [124]. Copyright 2017 American Chemical Society. (b) Machine learning segmentation and metrology analysis of nanoparticles. Reprinted with permission from ref. [131] 2020 Creative Commons Attribution License. (c) Automated machine learning workflow for morphological analysis of metal nanoparticles. Reprinted with permission from ref. [118] 2021 Creative Commons Attribution License.

perimental samples with low uncertainty in the responses (e.g., diffraction patterns).

3.1.4. Task: design-structure-property relationships

So far, we have described models for microscopy image acquisition and image interpretation (classification, segmentation, identification of components). In this section we review recent studies that have used data driven approaches linking information from microscopy images to predict physical properties of the polymer material or to create relationships between the materials design/condition and observed structures. These studies we describe below use ML modeling to directly address the holy grail in most polymer research activities – establishing design-structure-property relationships.

Using AFM images of various styrene-co-(n-butylacrylate) copolymer morphologies, Xu et al. have compared regression-based ML methods (linear regression, support vector regression, and random forest regression) vs. deep learning methods for predicting polymer property (e.g., glass transition temperature) from microscopy image data. [115] Visual features were extracted manually from AFM images and served as input to linear regression, support vector regression, and random forest regression. They also showed that for deep learning models (e.g., CNN) one does not need to manually extract visual features but rather the model selects visual features automatically from the AFM images. This makes deep learning methods better for *generalized* feature extraction tasks than conventional regression models.

Vargo et al. have applied various ML models such as random forest regressor, gradient boosting regressor, kernel ridge regressor and support vector regressor to identify structural features- periodicity, microdomain ratio, and grain size from an AFM image. The specific data they used were AFM images (output) of nanocom-

posite thin film comprised of polystyrene-b-poly(4-vinyl pyridine) or PS-*b*-P4VP, 3-pentadecylphenol, and iron oxide nanoparticles for varying (inputs) block copolymer molecular weight, ratio of the two blocks, nanoparticle size and loading, solvent fraction, and the nanocomposite thin film thickness [148]. The authors curated ~600 AFM images from past literature for ML model training and testing. They then performed feature engineering to create a 3-level hierarchical feature vector - whole-image, grain-scale, microdomain-scale - from an AFM image depicting block copolymer morphology as the prediction targets. By inspecting the correlation matrix between the inputs and outputs, they noted that the periodicity is heavily determined by the molecular weight, whereas the grain size was not correlated to any of the input variables. The microdomain ratio was deemed a more "interesting" target as it is non-linearly dependent on the different component ratios. With the random forest regressor, they satisfactorily predicted the microdomain ratio using experimental parameter inputs .

3.2. Applications of ML in scattering

The typical output of scattering measurements is a 1D profile of $I(q)$ (intensity) vs. q (magnitude of wave vector) or 2D profiles capturing intensity as a function of azimuthal angle and magnitude of wave vector, q. As shown in Figs. 1d, 1e and 1f, the 1D profiles are curves and the 2D profiles are two-dimensional images of intensity patterns that look like spots in a specific pattern, sharp or diffuse symmetric or asymmetric rings, and all of these patterns hold information about the sizes, orientations, and packing of atoms, molecules, and domains. Deep learning models, ensemble ML models, and optimization algorithms can be used for scattering data analyses tasks like automated classification of the measured data (e.g., connecting the measured profile to a specific class of morphology like spheres, rods, etc.) and interpretation of the scattering profile by quantifying relevant aspects of the structure (e.g., domain shapes, sizes, radius of gyration of the polymer). ML can also help in automating high throughput scattering measurements. In the following sub-sections, we review how ML developments have aided scattering tasks – classification, interpretation, and automation – in polymer science and engineering. As we did in the previous sections, here too we highlight work done in soft materials as well as promising ML work in other areas of materials sciences that could be extended to polymer sciences.

3.2.1. Task: classification

As one of the outputs of scattering experiments is a 2D image of intensities, we first describe ML model development for classification of scattering images based on the patterns of intensities they show. Wang et al. have used CNN and convolutional autoencoders to classify measured X-ray scattering patterns of self-assembled polymer films, nanoparticles, lithographic gratings, and organic semiconductors. [149] They trained their models to learn attributes of the scattering images, specifically an isotropic ring, anisotropic/isotropic halo, anisotropic/isotropic diffuse low q features. They noted that automatic attribute recognition can be a challenge because multiple images with the same attribute can be different morphologies. Upon inspection of classification precision score of individual labeling categories in their dataset, they found that for some cases where classification precision was low, either the target attribute had an unusual appearance or was highly localized or there was subtlety in some of the attributes, ambiguous labeling, and/or error happening during the measurements. To improve the classification accuracy of such atypical images, they suggest that researchers augment the training data with additional examples of the borderline cases; one way to augment the overall training data is by creating/generating synthetic (i.e., simulated) images exhibiting such atypical or marginal patterns.

While not directly related to polymers, it is worth noting that CNN models have also been used for classification of nanoparticle shapes from grazing incidence small-angle X-ray scattering (GISAXS) [150] and extraction of nanoparticle orientation distributions from grazing incidence small-angle scattering (GISAS) [151]. In the former study [150], the authors calculated 'synthetic' GISAXS intensities using the FitGISAXS code [152]. Specifically, they simulated the GISAXS patterns of tellurium nanoparticles on a Si substrate with a specific incident angle. Their data contained GISAXS patterns of eight classes: "capsule, spheroid, ellipsoid, truncated spheroid, hemispheroid, prism based on an equilateral triangle (prism3), prism based on a regular hexagon (prism6), and cylinder". These particle shapes were mathematically described by their diameter, aspect ratio in the vertical plane, and size dispersion; for ellipsoid and truncated spheroid, parameters more specific to those shapes were added (e.g., ratio of vertical diameter to height for the truncated spheroid and aspect ratio in the horizontal plane for the ellipsoid). After training and testing the CNN with simulated GISAXS images, they demonstrated that trained CNN could classify the shapes of Te nanoparticles in the GISAXS images from the experiments. Such models can be valuable for researchers working with polymer nanocomposites composed of inorganic nanoparticles embedded within a polymer matrix. Understanding how processing of such composites alters the spatial arrangement and orientation of the various nanoparticles (depending on their shape, size, filler fraction) can be valuable in predicting the resulting macroscopic properties of the composite.

Aty et al. have developed a computer vision-based ML pipeline for classification of lipid phases (e.g., lamellar, hexagonal, and cubic phases) from 2D SAXS patterns. [153] With a transfer learning approach, larger volumes of simulated (or 'synthetic') SAXS patterns were used for pretraining of their CNN model before finetuning on a small set of real experimental SAXS patterns. They used unsupervised clustering methods such as Principle Component Analysis (PCA) [154], t-distributed Stochastic Neighbor Embedding (t-SNE) [155] and Uniform Manifold Approximation and Project (UMAP) [156] to show that their simulated SAXS data are representative of the real experimental data. Synthetic data can be generated with a larger range of tunable parameters and smoother variations whilst constraining the noteworthy features of each phase. They achieved 99.6 % classification accuracy on the test samples with the transfer learning approach. Dealing with coexisting lipid phases, the authors suggested potential expansion of classification of individual lipid phases to classification of the composition of the coexisting phases. We feel that such ML approaches could be useful for polymer researchers who work with design parameters or conditions that lead to coexistence of two phases (e.g., perforated lamellar and cylinders or perforated lamellar and imperfect gyroids in block copolymer phase diagram) at equilibrium or due to intentional/unintentional kinetic trapping.

On that note of co-existence, many commercially used polymers (e.g., polyurea) and conducting polymers (e.g., poly(3- hexylthiophene) or P3HT) exhibit semi-crystalline morphologies where crystalline phases co-exist with amorphous phases. Relevant to crystallinity, we note that many studies have shown success in using ML for classification of *inorganic* crystalline phases from X-ray diffraction patterns [157-164]. These studies have greatly benefited from large databases of well-defined and characterized crystal structures (e.g. Inorganic Crystal Structure Database [165]). Even though XRD is often used to identify crystallinity in semi-crystalline polymers [166-169], unfortunately, there is a dearth of similarly well-curated databases comprising different semi-crystalline polymers. Additionally, unlike inorganic crystalline materials, for polymers it is vital to also store the processing history of the material along with this data as the extent of crystallinity measured is dependent on the processing steps and history. [170]

We circle back to this topic of ongoing challenges in polymer community in later sections to discuss what polymer researchers can do better so that we can transfer such successful ML models from other materials community for semi-crystalline polymer characterization purposes.

3.2.2. Task: interpretation of small-angle scattering profiles of polymers

Interpretation of small-angle scattering profiles (i.e., SAXS and SANS) from polymer materials can be qualitative – i.e., identifying the predominant shape of domains- or can be quantitative – i.e., calculating the distribution of shapes and sizes and other relevant structural parameters. For qualitative interpretation, ML models have been developed to take as input SAXS and SANS profiles and to directly output the most likely/closest shapes of the domain structures (e.g., spheres, rods, sheets) [40,171] Fig. 4a shows the use of transfer learning for re-training a convolutional neural network (CNN) to take as input a 2D SANS profile and output the shape of the structure(s) that is (are) in the material. [172] Alternately, ML models have also been used to identify the best choice of the shape-based analytical model one should use to fit the scattering profile to arrive at the quantitative structural information. [173] (Fig. 4b) While these methods are valuable in identifying standard shapes, they will not work for systems where the observed structures do not fit these standard shapes.

ML workflows have also been developed to directly obtain quantitative structural information from scattering profiles. Along these lines, Fig. 4c shows work by Franke et al. developing ML methods for analyzing SAXS data from protein solutions. [174] Franke et al. transform experimental SAXS patterns into feature vectors and then use k-nearest neighbor method to obtain not only the shape of the protein but also its maximal diameter as well as molecular mass. He et al. developed a deep learning method for model reconstruction from SAXS data [175]. They used an auto-encoder for protein 3D models to compress the information about the protein's 3D shape information into vectors of a 200-dimensional latent space. These vectors were optimized using genetic algorithms to build 3D models that are consistent with the input scattering data.

It is vital that we emphasize key differences between proteins and a variety of non-biological polymers (synthetic or bio-derived and functionalized) in this context of scattering analyses. Most proteins have the advantage of having precise secondary and tertiary structures and large protein databanks that contain the coordinates of thousands of such proteins' precise structures. In contrast, most synthetic polymers do not have precision in size (e.g., molecular weight distribution) or structures (amorphous, disorder structures with dispersity in sizes and shapes). As a result, most polymer structures are not stored in databases and thus, the polymer community lacks the advantage of having large protein databases that serve as training data and testing data for many ML models aimed at protein structure prediction. To specifically address the need to automate and accelerate interpretation of SAXS and SANS profiles obtained from polymer systems that are mostly amorphous (i.e., lack of secondary or tertiary structures as in proteins, zero to minimal crystalline arrangements) and have dispersity in most relevant structural dimensions, Jayaraman and coworkers developed CREASE – Computational Reverse Engineering Analysis of Scattering Experiments. [176-184] (Fig. 4d)

CREASE was developed to overcome some traditional challenges in scattering analyses in the field of polymer science and engineering. The scattering profiles in polymer systems have traditionally been interpreted using conventional analytical model-based fitting, as mentioned earlier and as described in many relevant review articles. [48,68-75] Conventional analytical scattering models involve assumptions about the 'primary particle' (i.e., macromolecule, micelle, coated nanoparticles, particles with unconventional shapes) and/or the interactions that lead to their spatial arrangement (e.g., sticky hard-sphere model). With significant advances in polymer synthesis and processing, polymer scientists are observing or deliberately achieving unconventional, novel structural arrangements that are not captured by the large library of existing analytical models. One can always develop new analytical models from scratch and manually fit the scattering profiles to arrive at the information they wish to learn. However, manual fitting is not conducive for analyzing high-throughput or time-series data. Hence, CREASE was developed to alleviate these problems, especially this reliance on analytical models, and to accelerate scattering analyses to enable automation.

CREASE [https://github.com/arthijayaraman-lab/crease_ga] uses a simple, easy to adopt optimization method - genetic algorithm. As shown in the schematic in Fig. 4d, CREASE's genetic algorithm (CREASE-GA) takes as input SAXS and/or SANS scattering profiles - I_{exp} (q) or I_{exp} (q,θ) where θ is the azimuthal angle. CREASE requires the user to choose their relevant structural features ('genes') based on their domain knowledge of the general shape of the assembled structure from other imaging techniques and/or subject matter expertise. Then, CREASE-GA starts with an initial 'generation' of many 'individuals,' where each individual has a unique set of 'genes.' CREASE-GA iterates towards identifying *all optimal individuals* whose structural features gives rise to a computed scattering profile, I_{comp} (q), that closely matches the input experimentally measured scattering profile, I_{exp} (q).

One important step in the CREASE-GA loop is the calculation of I_{comp} (q) for each individual in every generation. The traditional (physics-based) way to calculate the I_{comp} (q) is to create for each individual representative three-dimensional real space structures corresponding to the structural features (genes) of that individual. These real-space structures are then filled with point scatterers whose scattering length densities represent the constituents of the system, and using the Debye equation on the scatterer positions one arrives at the I_{comp} (q). This calculation can be computationally intensive. The faster way to calculate I_{comp} (q) is by using a ML surrogate model that links the structural features directly to I_{comp} (q). Jayaraman and coworkers have used both neural networks [178,181-183] and XG-Boost based model [184] to train on thousands of computed (or 'synthetic') scattering profiles calculated from the Debye method for various sets of genes and structures. Through ML enhancement, CREASE has been shown to be fast and suitable for identifying multiple real-space structures simultaneously, which is essential to the success of the proposed high-throughput screening experiments.

CREASE method has been used successfully to interpret experimental SAXS or SANS profiles from amphiphilic polymer solutions at dilute concentrations. CREASE identified structural features for a variety of assembled polymer structures in solution – spherical core-shell micelles, [185] polypeptide- based vesicles, [180] synthetic cylindrical micelles, [178,179,186] and bioderived polymer fibrils[182]. In the above studies, either CREASE was shown to outperform the analytical models (e.g., polydisperse vesicles [180], micelles composed of unique new polymer chemistries [185]) or performed just as well as analytical models (e.g., methylcellulose fibrils [182]). In some cases, CREASE enabled testing of hypothesized structures even if corresponding analytical models did not exist. [186] We note that in these studies, the shapes of the assembled structures were known from microscopy and the user chose only the relevant structural features for CREASE-GA to iterate over. If such shape information is not known, one potential new direction is to combine shape-classifying ML models described under classification and ML-CREASE. This way the classification ML model would identify the potential (closest) shapes, and then the user can

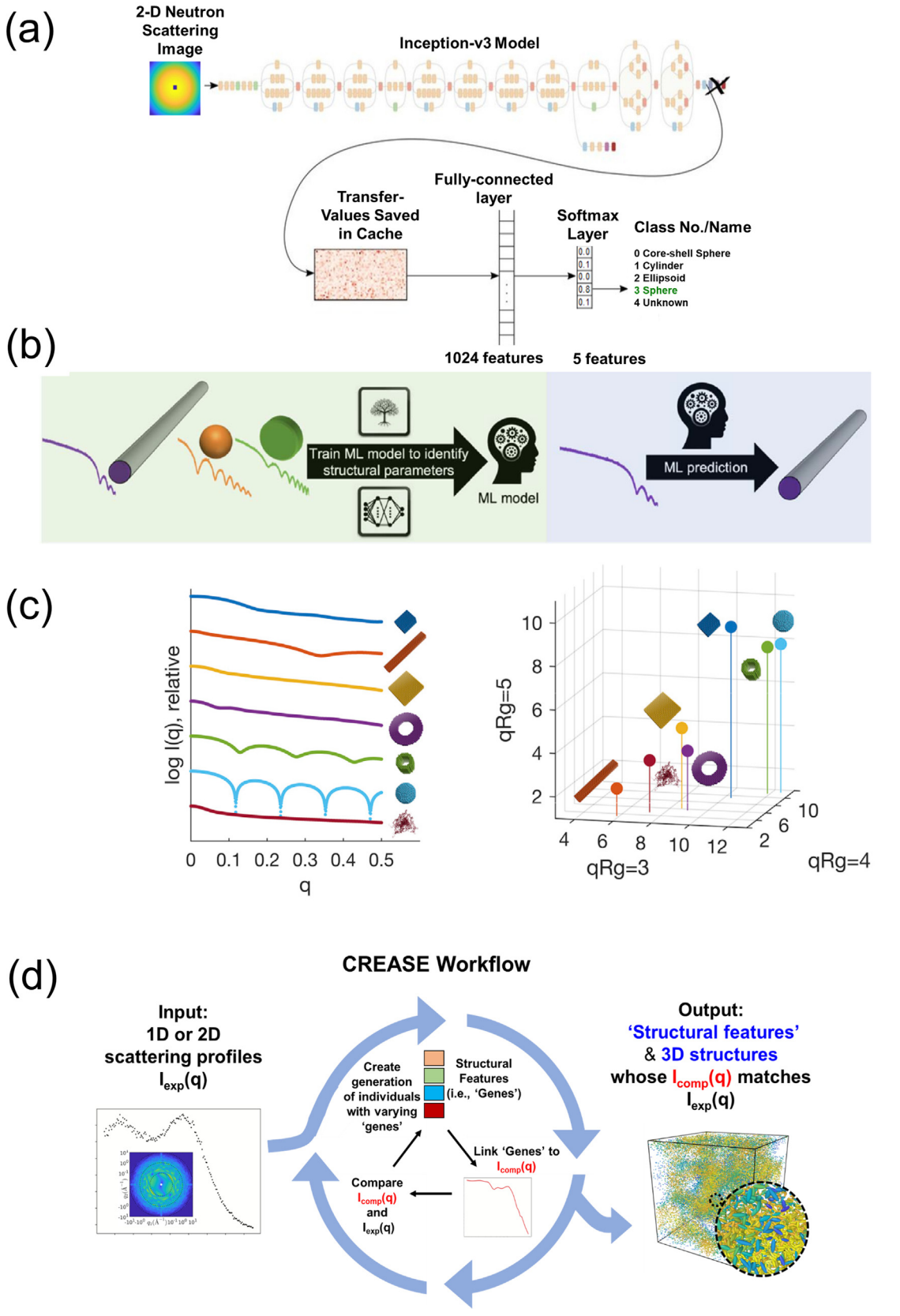

Fig. 4. a) Convolutional neural network taking as input 2D neutron scattering images of shapes of objects to be able to classify shapes of structures – sphere, ellipsoid, cylinder, etc. Adapted with permission from Ref. [172] 2020 Creative Commons Attribution License. b) Image adapted from a recent perspective on machine learning approaches for analysis of scattering and spectroscopy. This selected image conveys the idea of training ML models on analytical form factors to identify shapes of structures. Reprinted with permission from Ref. [174] 2023 Creative Commons Attribution License; c) Form factors of various canonical shapes – sphere, cube, ring, fractal chain, etc. Reprinted with permission from Ref [40]. Copyright 2018 with permission from Elsevier. d) ML-CREASE for interpretation of form and/or structure factors from SAXS and/or SANS profiles of macromolecular materials that may not have appropriate analytical models (form factors or structure factors).

customize ML-CREASE to optimize relevant structural features (e.g., dimensions) for each of those potential shapes.

CREASE has also been used to understand the extent of mixing and demixing in amorphous materials at high concentrations. For example, CREASE was used to analyze SAXS and SANS profiles from concentrated binary mixture of polydisperse spherical nanoparticles (i.e., P(q) is a sphere form factor) to determine the extent of segregation/mixing of the two types of nanoparticles – polydopamine and silica - and the precise mixture composition. [181,187-189] The extent of mixing/demixing and the composition of the mixture were relevant to predict structural colors. [187-189] Even though CREASE iterates over a lower-dimensional representation of the real-space structures, one of the advantages of CREASE is that at the end of the optimization loop it also outputs a 3D real-space structure(s) representative of those optimized structural features. These 3D real-space structures are valuable for follow-up simulations or calculations of properties (e.g., color, mechanics, rheology) using those real-space coordinates. Refs [187-189] demonstrate such calculations which serves to further confirm CREASE predictions.

Most recently. CREASE has been extended to overcome another scenario one may face when trying to interpret structure in polymer solutions where both the structure S(q) and form factors P(q) change with varying experimental conditions. [183,190] The reader is reminded that the measured I(q) has both P(q) and S(q) contributions, $I(q) = P(q)\,S(q)$. Traditionally, in many cases, researchers assume that the P(q) calculated at one condition (where $S(q) = 1$) does not change with concentration and use that calculated P(q) as is to interpret S(q) at higher concentration. This assumption is not necessarily valid if the 'primary particle' (e.g., micelles, vesicles) evolves with changing system/solution conditions. This drives the need for simultaneous identification of form and structure [i.e., P(q) and S(q)] and structural interpretation at the conditions of interest without assumptions about P(q) not changing and without using approximate analytical models that may be inapplicable for the system at hand. To address this need, in recent work, Jayaraman and coworkers 'P(q) and S(q) CREASE' that extends previous CREASE capabilities. [183]'P(q) and S(q) CREASE' can be used to analyze SAXS or SANS profiles from polymer materials to simultaneously obtain the form factor P(q) (e.g., dimensions of domains with unconventional shapes) and structure factor S(q) (e.g., spatial arrangement of those domains) without relying on any analytical models. They validated the approach by analyzing scattering from computationally generated structures for which the dimensions (form factor) and spatial arrangements (structure factor) are known. The validated method was then used to analyze SANS profile from experimental measurements of surfactant coated nanoparticle solutions with the goal to understand the surfactant coating/shell arrangement with changing salt concentration and temperature, without being limited by off-the-shelf approximate/incorrect analytical models. [190]

The above studies of CREASE took as input 1D SAXS profiles and/or SANS profiles, either (i) a single SAXS profile of the system, or (ii) one SAXS profile and a one SANS profile of the same system, or (iii) multiple SANS profiles with contrast matching one or the other component(s) in the system with the solvent. To extend CREASE to interpret 2D profiles for soft materials that show anisotropy in the assembled structure, Jayaraman and coworkers have now developed CREASE-2D [184]. CREASE-2D enables direct interpretation of 2D profile which is far more complex than analysis of 1D scattering profiles, I(q) vs. q, obtained by averaging along all azimuthal angles. Currently, researchers who study materials with any form of anisotropic structure (e.g., processed aligned synthetic conducting fibers, field-driven orientational alignment in polymers for sensing/electronics, sheared formulations during rheological measurements in personal care industry) need to interpret the entire 2D scattering profile. Yet analyses of such 2D profiles have traditionally only been done by fitting analytical models to 1D profiles obtained by averaging along all azimuthal angles or sections of the 2D profile. Such averaging schemes lose key information about the anisotropic structural arrangements that can drive the function of the materials. CREASE-2D method overcomes these current limitations and provides polymer researchers the speed (due to ML surrogate models) and accuracy (by avoiding any averaging of the 2D profile) to interpret quantitative structural information (e.g., domain shapes, sizes, orientation, volume fraction) from the entire 2D scattering profiles without any approximations. The surrogate model used to link structural features to 2D scattering profile was trained on 3D structures generated by a recent developed computational method - Computational Approach for Structure Generation of Anisotropic Particles (CASGAP) [191]. CASGAP generates representative 3D structures for input desired distribution of particle (representing domain) sizes and shapes and desired spatial orientations without particles overlapping at desired packing density. Using 2400 generated structures generated from CASGAP, Jayaraman and co-workers were able to train the surrogate XG-Boost ML model. Then, using 600 structures (unseen by the surrogate model) they validate the performance of the ML-model as well as the successful performance of the entire CREASE-2D workflow.

Another ML-based workflow developed by Röding et al. focused on interpreting 3D structures of disordered soft materials with two or three phases from their SAXS profiles [192] They considered model systems consisted of two phases that they label as "pore" and "solid" or three phases where the third phase could be an interface between the "pore" and "solid" phases; all phases have different electron densities. Even though they do not explicitly state the types of polymer systems where such structures are seen, one knows that binary polymer blends or multi-component polymer nanocomposites (e.g., binary blends with nanoparticles at interfaces) exhibit such bicontinuous structures making this method relevant. They used XG-Boost based model to estimate microstructural parameters from SAXS data. The microstructure model is restricted to a periodic Gaussian random field (GRF) with variable length scale. They process and threshold the GRF to yield two-phase (pore and solid) and three-phase (pore, interface, and solid) structures. They noted that for their method development artificial neural networks did not perform better than XGBoost for the purpose of predicting microstructure model parameters.

It is important to note that the challenge with analyzing scattering profiles from polymer systems is that the scattering profile in most cases does not correspond to one unique structure or one unique set of structural features. How well and correctly an ML model interprets the scattering profile as compared to an analytical model will depend on how the ML model is trained. Specifically, the quality and quantity of ML model's training data dictates how well the ML model can interpret the structural features from the 1D scattering profiles. How one generates the data used for training the ML model will dictate if the model is learning only physical realistic structural features or any set of structural features that numerically result in the scattering profile. Further, how the training data is sampled (not only values of the sampled structural parameters but also the structural parameters that lead to unique effects on scattering profiles) and how much training data is used to train the ML model, are important factors that dictate the success of the ML model in interpreting scattering profiles as compared to the traditional analytical models.

Besides all of the quantitative structural information, one may be interested in interpreting information about thermodynamics – e.g., effective interactions – from small-angle scattering profiles. In recent work, Chang et al. have used a ML (inversion) scheme for determining interactions from scattering profiles. [193] They used

a case study of colloidal suspensions and demonstrated that they can infer effective potentials from the scattering spectra without any restriction imposed by theoretical model assumptions. They demonstrated that their workflow could quantify effective interaction in highly correlated systems using data from scattering and diffraction experiments.

We end this sub-section on interpretation of small-angle scattering of polymers by highlighting recent work by Zhao et al. who have used a combination of ML methods to process 2D-SAXS datasets of isotactic polypropylene (iPP) films and directly create processing-structure mapping [194]. They took experimental 2D SAXS data into lower-dimensional representation using variational autoencoders (VAEs) and through that conversion they were able to extract the structural evolution features of iPP films. Their VAE was trained using SAXS dataset from their previous work where they stretched iPP films to break at temperatures from 30 to 160 °C and the maximum strain was ca. 400 %. They then used a hybrid neural network to create a processing-structure mapping of iPP and notably, generated 2D SAXS patterns at processing parameters that had not yet been experimented at.

3.2.3. Task: automation in scattering data acquisition

Doerk et al. have reported an automated phase exploration of thin film morphologies of blends of PS-*b*-PMMA block copolymers incorporating elements of chemical template for combinatory sampling, high-throughput SAXS measurements and Gaussian process-based active learning module in Fig. 5a [91]. A framework for Gaussian process guided autonomous experimental data acquisition called gpCAM [195,196] developed by Noack et al. was used as the active learning module. The automated workflow seamlessly integrated SAXS measurement, SAXS data analysis, and Bayesian modeling-based candidate suggestion of next sample in phase space. In their Bayesian modeling-based candidate suggestion model, they leveraged three acquisition functions for the selection of the promising next measurement candidate with the ability to choose between balanced random exploration, targeted exploration of rarely visited regions, and exploitation of regions deemed “interesting” by the experimenter for more efficient sampling. Through the use of their automated workflow, multiple novel morphologies have been discovered, visualized with top-down and cross-sectional SEM images, and the driving forces for these morphologies have been explained by physics-based coarse-grained molecular dynamics simulations.

Another noteworthy study in automation is that by Beaucage and Martin who have developed a state-of-the-art Autonomous Formulation Laboratory (AFL) - an adaptable platform for automated synthesis and characterization of complex polymer formulations using x-ray and neutron scattering techniques in Fig. 5b and 5c. [197] This platform incorporates hardware systems of robotic auto-pipetting, mixing, scattering characterization, and complementary software systems programmed in Python for control over the experiments with a simple and user-friendly interface. Beaucage and Martin showed three proof-of-concept examples using AFL: (i) classification of SAXS profiles of silica nanoparticle of various sizes in aqueous solution; (ii) study of the micellization of cetyltrimethylammonium bromide (CTAB) under the influence of salt and sodium salicylate; and (iii) the phase mapping of industrial polymer formulations where a model system containing Poloxamer F127, hexanes, water and salt was studied using AFL and SAXS as the measurement technique. Their AFL platform has the potential to incorporate other measurement modalities such as UV–vis-NIR spectroscopy and microscopy with rigorous control of the sample preparation. The facilitation of FAIR [198] data management also makes the AFL a promising automation platform for future integration with data-driven analysis.

3.3. Applications of ML in spectroscopy

The typical output of spectroscopy measurements is a 1D vector array or 2D image of the intensities of the measured physical property vs. the light beam wavelength or energy (as shown in Figs. 1g-i). Spectroscopy data often contain signature peaks for identification of specific molecules, functional groups only in a small part of the data. Dimensionality reduction methods are often used to preprocess the spectroscopy data into low-dimensional representations and clustering methods are used to classify the lower-dimensional data. In the following sub-sections, we review recent ML model development for – classification of molecules or functional groups from spectroscopic data and automation of material synthesis and screening using spectroscopic data as an optimization target. As in previous sections, we highlight promising ML workflows using spectroscopic data in other areas of materials sciences that we believe has promise for use in polymer materials.

3.3.1. Task: classification

Tetef et al. have demonstrated the use of unsupervised methods for the classification of both x-ray absorption spectra and X-ray emission spectra of sulphorganic molecules [199]. Unsupervised dimensionality reduction methods such as PCA, VAE and t-SNE were utilized to generate lower-dimensional representation of the X-ray spectra. This lower-dimensional representation was then used for classification of degree of oxidation state, aromaticity, and aromatic

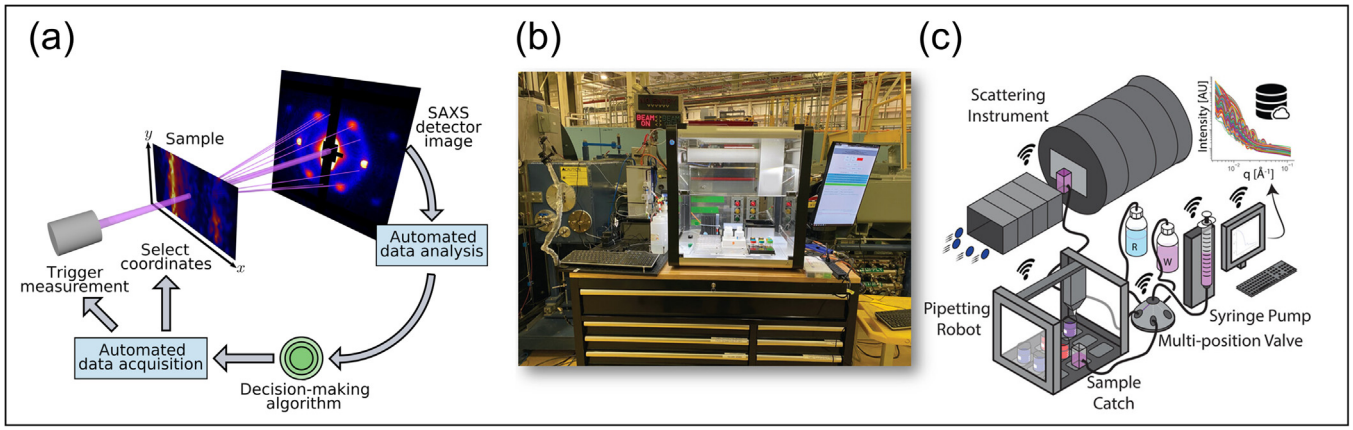

Fig. 5. Machine learning automation of scattering experiment. (a) Novel block copolymer morphology discovery for aided by automated SAXS experiments. Reprinted with permission from ref. [91] 2023 Creative Commons Attribution License. (b) Experimental setup and (c) schematic of automated experimentation of formulations using the Autonomous Formulations Laboratory. Reprinted with permission from ref. [197]. Copyright 2023 American Chemical Society.

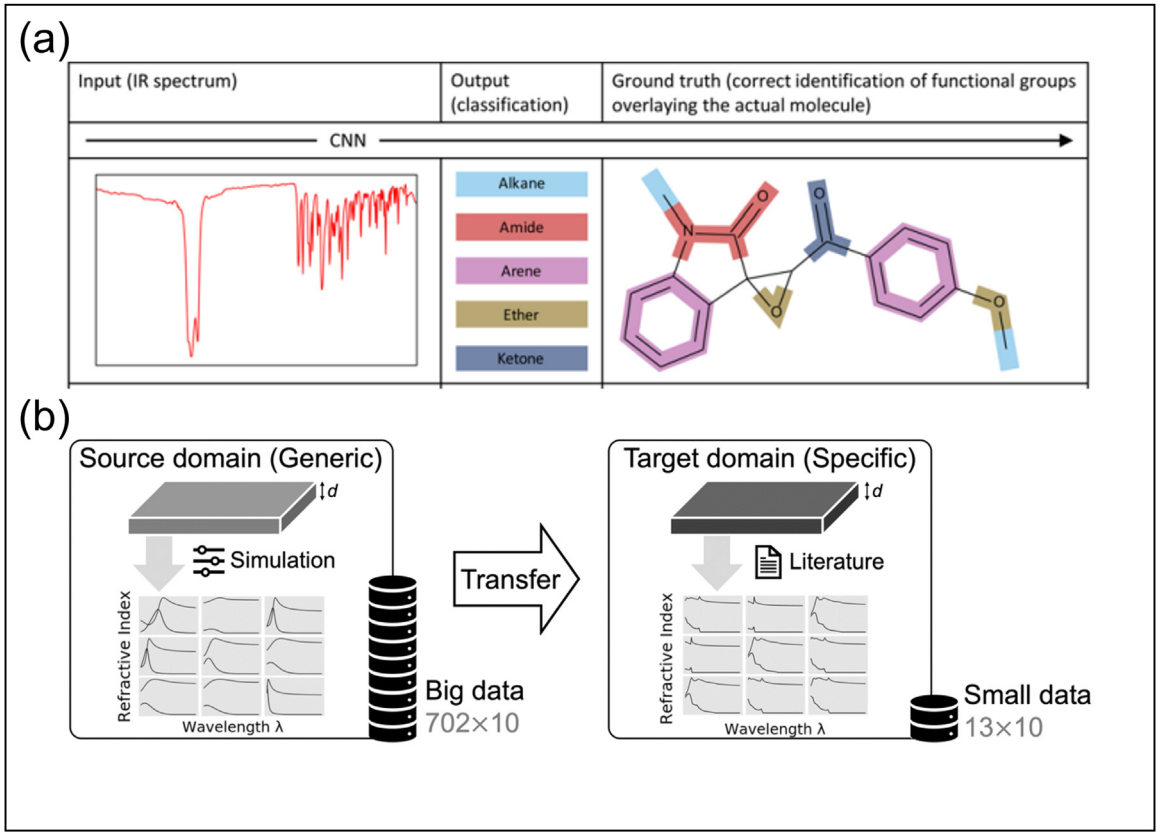

Fig. 6. Machine learning classification of spectroscopy data. (a) An automated machine learning workflow for classification of functional groups present in small molecules from IR spectra. Reprinted with permission from ref. [200] 2023 Creative Commons Attribution License. (b) Demonstration of transfer learning for data-efficient prediction of UV–vis spectra of perovskite thin film material. Reprinted with permission from ref. [202] 2023 Creative Commons Attribution License.

or aliphatic sub-categories with a supervised learning k-nearest neighbors method. The authors found that feature maps generated with t-SNE not only outperformed other unsupervised methods at classification, but also enabled discovery of new chemically relevant clusters (e.g., distinguishing nuances within a sub-category of aromaticity of the sulphorganic compounds) not seen by other dimensionality reduction methods.

Jung et al. have used CNN for the identification and classification of functional groups in molecules given the FTIR spectra as input (Fig. 6a) [200]. Trained on over 50,000 FTIR spectra and over 30,000 molecules, their model can classify 37 types of functional groups with good classification performance. Given that multiple functional groups can exist in the same molecule, they also evaluated their model for multi-label classification and observed a rate of exact matching of 0.72 for up to 9 functional groups in one molecule.

Different featurization approaches, dimensional reduction methods, and classification models have been examined by Chen et al. for the classification of oxidation states from XANES of metal oxides [201]. Originally trained on computed data, cumulative distribution function (CDF) featurization method showed higher robustness (compared to other types of featurization such as peak features or continuous wavelet transform) when the authors applied their workflow to classify experimental XANES data.

In physics-driven modeling, the optical response e.g., UV–vis spectra are determined by the refractive index as well as the thickness of the thin film material. Tian et al. have used CNN for the inverse design and interpretation of UV–vis spectra of perovskite thin films (Fig. 6b) [202]. The UV–vis spectra were instead used as input to predict the thickness as well as the real and imaginary part of the refractive index of the thin film. To tackle the data scarcity problem, they leveraged transfer learning to first train a generalist model on data from a generic source domain and then finetune the pretrained model on the data from target domain that contained only 18 spectra data achieving 92 % prediction accuracy on thickness.

All of the above studies are excellent examples of ML models applied to spectroscopy data from small molecules or metal oxides; we share these to show that they also have potential for analyzing FTIR and UV–vis spectra from polymer materials.

Similar to FTIR and UV–vis spectra, solid-state and solution phase nuclear magnetic resonance (NMR) spectroscopy data can also be analyzed using dimensionality reduction methods such as PCA, VAE and t-SNE. While there is a dearth of papers aimed at applying these methods to NMR data directly from polymeric materials, there are other noteworthy studies and reviews [42,203-206] that show how ML can be applied to NMR data from bioinformatics and protein structure characterization; these approaches could easily be extended to similar analyses of NMR data collected from polymer systems. For example, one can train ML models to identify NMR peaks (i.e., noted as the “peak picker” problem) by using a database of NMR spectra of polymers with peaks corresponding to known chemical composition; then these models can be used to identify NMR signals from new polymers. Analyses of patterns in 2D NMR data can be done in a manner similar to deep learning (e.g., neural networks) based 2D image analyses, described in prior sections. A recent perspective on the role of ML in analyses of NMR spectra from biomolecular systems could serve as a starting point for polymer researchers who wish to extend these ML techniques to polymer NMR data. [206]

Lastly, we share one noteworthy ML study on NMR data from metal-organic frameworks (MOFs) field that could be relevant for the nanoporous polymer materials. [204] In this study, the authors obtained data from NMR relaxometry that is highly sensitive to

the materials' nanoscale porosity. They used PCA and partial least squares (PLS) regression techniques on this NMR data to classify the studied MOFs into high and low surface area porous materials. The pore sizes for these materials were in the range of 0.5–100 nm which is similar to pore sizes found in polymeric membrane materials. Furthermore, their data set was similar in size to typical polymer experiments' outputs; specifically, they worked with 15 MOF materials with 20 solvent contents per material, "resulting in over 300 NMR time decays of 1000 points each". We find this study noteworthy not only for its PCA and PLS regression analyses but also because the authors highlight that this approach is "high-throughput and a non-destructive way to assess porosity in ~1 minute, resulting in a pore surface area estimate ~1440 times faster than a gas adsorption isotherm measurement which requires ~1 day to perform". Researchers working with high-throughput design and testing of porous polymer membranes would find the methods described in this study [204] valuable.

3.3.2. Task: automation

Pozzo and coworkers have applied Bayesian optimization methods for the automated synthesis of gold nanorods with target optical properties [207]. They also defined a new distance metric called amplitude-phase distance that works better than the conventional metrics such as Mean Squared Error (MSE) for similarity quantification of high-dimensional spectra data. The authors compared the performance of Bayesian optimization of UV–vis spectra of synthesized nanorods with surrogate Gaussian ML models using MSE or amplitude-phase distance as the similarity quantification. They demonstrated that the surrogate with amplitude-phase distance was better at mapping out the underlying phase diagram with identifications of multiple phases. The authors also extended the application of amplitude-phase distance criteria to phase mapping of polymer blends using SAXS and of metal alloys using XRD [208]. In a subsequent study, they also used a composite distance metric that included the amplitude-phase distance and distances of the intensity, peak position, and area under the curve of two spectra to quantify whether a UV–vis spectrum comes from a plate-like silver nanoparticle. Leveraging the high-throughput capabilities of UV–vis, they rapidly screened through the plate-like silver nanoparticles for more direct characterization of the nanoparticle structure using SAXS and obtain particle shape and size distributions [209].

Gormley and coworkers have developed an ML guided automation platform for high-throughput design and screening of polymer-enzyme hybrids [210]. Copolymers synthesized from various methacrylate monomers were used to stabilize the three model enzymes. The authors used a Bayesian optimization workflow for iterative design, selection, and testing of copolymer stabilized enzymes. A large design space of copolymers was explored, and new copolymer formulations were found to outperform enzyme stabilizing ability of the initial copolymer designs hypothesized with systematic variation of design parameters. We direct the readers to Gormley and coworkers' recently published tutorial/user guide on how to use ML in a step-by-step manner to accelerate design and testing of next generation biomacromolecules. They also provide a python script that provides the user a hands-on experience with the ML pipeline [211].

3.4. ML models for reconstruction and generation of characterization data

The previous three sub-sections (3.1, 3.2, and 3.3) Sections 3.1 focused on ML models aiding acquisition and analyses of data obtained from microscopy, scattering, and spectroscopy. In practical scenarios, researchers may not have access to some instruments or have challenges with sample preparation that serves as a barrier to doing certain measurements. In such cases, using the previously collected relevant data as training set for ML models, researchers may want to reconstruct or generate new data. This sub-section presents recent work with ML models for generation and reconstruction for microscopy, scattering, and/or spectroscopy. We note here that use of generated images can be controversial in publishing as it may constitute as 'fake data' unless declared explicitly as 'generated' data vs. a measured data. Furthermore, the authors in publications should be clear about their motivation for using 'generated' data. For example, image reconstruction/generation practices are valuable for other downstream calculation of physical properties that are dependent on the distribution of the structural features – in such cases, one may generate an ensemble of images from ML models trained on a few time-consuming intensive real measurements and use the ensemble of images to get the distribution of structural features. These calculations can then inform the researcher of how small variations in structural features manifest as variance in the calculated physical property. Researchers should follow all scientific ethical practices when reporting such data.

3.4.1. Reconstruction of microscopy images

Reconstruction of various modes of microscopy images [212-216] and other types of characterization data [213,215-223] have relied heavily on autoencoders. An autoencoder is a pair of neural networks -an encoder and a decoder - built to reconstruct the input in a fully unsupervised (i.e., without any manual labeling) mode. The encoder converts the information from the input image into a set of lower dimensional latent variables and the decoder deciphers information back to the original target space. To reconstruct images Long et al. have proposed an autoencoder called Fully Convolutional Network (FCN) with a CNN as the encoder part and a neural network consisting of transpose-convolutional layers as the decoder. [224] A drawback of autoencoder is that the set of latent variables is discrete and tends to do poor reconstruction when the test target is outside of the training set. To mitigate the discrete latent space problem, a modified version of autoencoder, a VAE [225] has been used. VAE projects the latent space to a continuous Gaussian random field and randomly samples a set of latent variables when reconstructing from the latent variables to the target space. Ronneberger et al. have proposed a modified version of fully connected network (FCN) called U-Net by establishing skip connections between the same hierarchical layers in both the encoder and decoders, feeding more information about local features in the image from earlier encoder layers directly to later decoder layers. [128] U-Net which was first developed to tackle reconstruction of biomedical images has now become more broadly applicable for other types of image reconstruction cases [127,226-234].

Li et al. have compared multiple methods for reconstruction of various material microstructures including polymer composites, sandstone, copolymer thin film morphology, and rubber composites from images [235]. The authors found that CNN autoencoders outperformed decision trees and Gaussian random fields at reconstruction of the material microstructures quantified by morphological metrics. Chavez et al. have made a comparison of several deep learning architectures for reconstructing of missing slices in experimental X-ray scattering profiles [236]. CNN autoencoders, tunable U-Net and multi-scale dense networks (MSDNets) [237] were compared with the baseline method – biharmonic functions for reconstruction of horizontal and vertical gaps in the X-ray scattering data.

3.4.2. Generation of cross-modal and multimodal characterization data

In all materials domain, researchers often use multiple complementary characterization techniques to elucidate different pieces of

information about the materials' structure, morphology, and properties. Such complementary, yet distinct, pairs or triplets of data obtained from the same system can be useful for ML modelers who wish to generate one form of characterization data from another form of characterization; we call this cross-modal or multi-modal generation or reconstruction. One may ask "*Why would such cross-modal or multi-modal reconstruction be useful in polymer science and engineering?*". Depending on the availability and accessibility of the different characterization techniques (e.g., scattering, microscopy, spectroscopy), each polymer research facility or academic research lab may have access to high-throughput capability in one technique but face limitations (sample preparation, resolution) with other technique(s). Researchers in smaller academic laboratories may have easier access to lower resolution microscopy instruments but not higher resolution imaging techniques or scattering time. Researchers in national laboratories may have access to high-throughput characterization techniques (e.g., SAXS) but have to deal with extensive sample preparation steps for a complementary technique (e.g., SEM). Some forms of characterization (e.g., microscopy images) are direct visuals and easier to interpret forms of data while other forms of characterization (e.g., scattering) being in Fourier space are harder to interpret with naked eye. In all of these scenarios, it would be valuable to have ML models that can generate one form of the characterization data (that is useful but either unavailable to a researcher or is not amenable to high throughput measurement) from another form of characterization data (that is easier to access or amenable to high-throughput measurements). Simultaneously, such cross-modal reconstruction can be used to convert harder-to-interpret characterization data to easier-to-interpret characterization data. For establishing design-structure-property relationships it would be valuable to have paired structural characterization from multiple techniques, regardless of ease or access.

We start with a couple of strong examples of such ML model developments in nanomaterials community. Stein et al. have used VAE to reconstruct optical microscopy images and UV–vis spectra of solution droplets containing various compositions of metal oxides (Fig. 7a). [213] A VAE model was trained to reconstruct the optical microscopy images and to obtain low-dimensional feature maps of optical microscopy images that were used to reconstruct UV–vis spectra. A conditional VAE model was trained to take both the optical microscopy image and the accompanying UV–vis spectra and reconstruct the optical microscopy image. They achieved good reconstruction performance supported by the underlying connection of the UV–vis spectra and the color exhibited by the imaged droplet.

Yaman et al. have used VAEs to enable cross-reconstruction of surface plasmonic spectra and SEM images of gold nanoparticles (Fig. 7b). [215] By matching the similarity of the latent space of spectra data and latent space of SEM images, they reconstructed one type of characterization from the other with particularly good accuracy for images containing a single nanoparticle.

Lu and Jayaraman have extended such cross-modal reconstruction capability to polymer data. They developed PairVAE [216] for universal cross-reconstruction of material characterization data with a proof-of-concept application to novel morphologies of PS-*b*-PMMA block copolymer thin film assemblies synthesized by Doerk et al. [91] (Fig. 7c). They used the openly available high-throughput SAXS data and SEM data from Doerk et al. to train the PairVAE. They demonstrate that one does not need much supervision during model training and that it works well even with a small amount of data (e.g., $\sim$ 50 pairs of characterization data). The trained PairVAE [https://github.com/arthijayaraman-lab/pairvae] successfully generated a SAXS profile from an SEM image as input or an SEM image from SAXS profile as input. They also found that by pairing the SEM latent space (relatively sparse) with the SAXS latent space (relatively clustered), the SEM latent space becomes more convergent, yielding morphologically closer reconstructions than seen with solo trained SEM-SEM reconstruction model; similarly, the paired training makes the SAXS latent space becomes less clustered, yielding more unphysical SAXS patterns than solo trained SAXS-SAXS reconstruction model. We note for interested readers that in this PairVAE implementation for SAXS–SEM reconstruction Lu and Jayaraman incorporated random cropping of larger (and few) SEM images during training as a means of data augmentation which helps mitigate the small size of data issue that we often face in polymer sciences.

4. Current barriers and future directions

To harness the power of all the ML methods and workflows we described in the sections above and to enable automation in synthesis, characterization, and manufacturing in polymer science and engineering, we still have a few barriers to overcome. The biggest barrier in the field of polymer science and engineering arises from the diversity of ways various laboratories record and store their characterization data. Laboratory practices range from storing data only locally on a computer connected to the measurement instrument or locally on an internal laboratory server. However, as many of the above workflow developments discussed in previous sections show, openly sharing data with researchers outside of an institution and laboratory can lead to impactful development of new and improved computational analyses methods. Thus, storing data only for sharing within a laboratory or an institution/facility hosting the instrument can be quite detrimental for progress in polymer science. With growing advances in cloud computing and storage platforms (e.g., Google drive, Amazon Web Services), measured data can be stored and shared within larger collaborations (with two to ten laboratories) using such platforms fairly easily. Furthermore, web-hosted open-data repositories like Zenodo and Figshare provide venues for anyone to share their scientific data on a published / working project that can lead to open-access and utilization by other researchers across the globe for model development and training.

In addition to the measured data itself, the context of the measured data should also be shared. In some cases, the context (e.g., processing history) impacts the measurement far more than the chemical composition of the polymer material. Such contextual information – 'metadata' – needs to be stored along with the measurement. However, researchers in many polymer laboratories are either unaware of the phrase "metadata" or do not follow uniform guidelines for recording metadata about the material processing history. We quote Pelkie and Pozzo from their recent perspective [238] that without a community-wide effort towards unified metadata and dedicated data management, we will continue to face roadblocks in our progress towards advancing automation.

With the growing popularity of large language models (LLM) in materials and chemistry fields, we expect to see a push towards research involving data collected by using LLM on the decades of scientific literature. [239,240] When LLM is used to extract data from publications that have specific phrases, again inclusion of the metadata with the measured data would be critical. Metadata including labeling of the systems, processing history of the synthesized and characterized material can also be used for learning by LLMs. Ensuring proper data collection with community standardization of metadata records will be a gold standard not only for adherence to FAIR [198] data principles, but also for ensuring reliable data source for LLMs training.

In the following sub-sections, we describe recent progress made to overcome the challenges we have described so far, as well as future directions for data-driven research in polymer science and engineering.

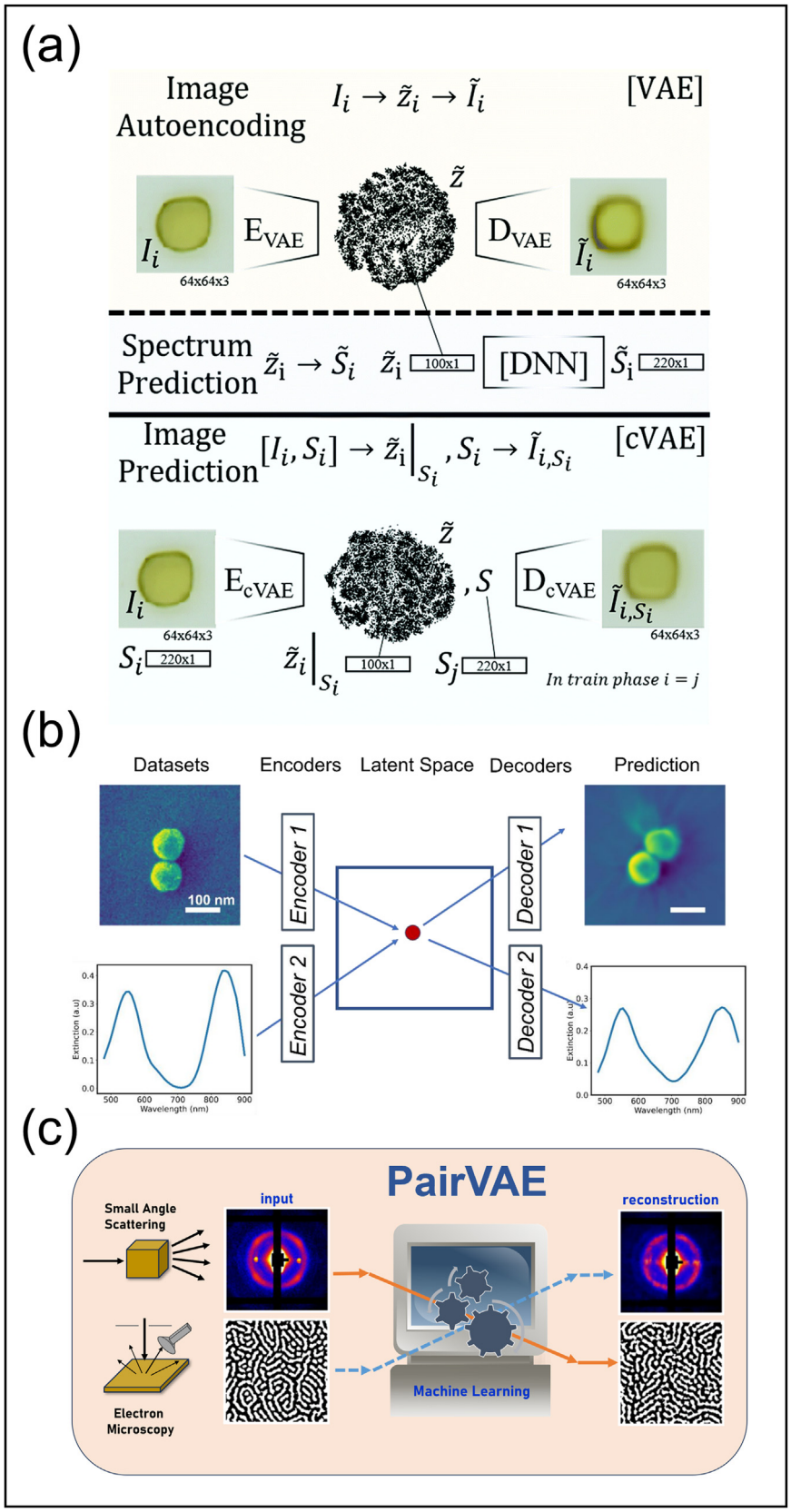

Fig. 7. Machine learning reconstruction of multimodal material characterization data. (a) Conditional variational autoencoder for reconstruction of UV–vis spectra or optical microscopy images of metal oxide solution droplets. Reprinted with permission from ref. [213] 2019 Creative Commons Attribution License. (b) Variational autoencoders for cross-reconstruction of surface plasmonic spectra and SEM images of gold nanoparticles. Reprinted with permission from ref. [215] 2022 Creative Commons Attribution License. (c) PairVAE for cross-reconstruction of SAXS and SEM images of block copolymer morphologies. Reprinted with permission from ref. [216] 2023 Creative Commons Attribution License.

4.1. Curation of polymer characterization data

Over the past decade, we have seen significant effort in community-curated databases of polymer properties adhering to the FAIR principles [198]: PolyInfo [241], PolymerGenome [242], MaterialsMine [243], etc. The macroscale physical property of a polymer material is influenced by multiple factors including the chemistry of the polymer(s) and additives, the processing conditions, and the resulting morphology. To capture the characterization relevant to all of these – chemistry and structure(s) - two or more complementary characterization techniques are used. The majority of open-access characterization data pertaining to polymer materials research are deposited with general purpose (i.e., not specifically for polymer research) databases such as Figshare, Zenodo, Dryad [244], MaterialsDataFacility [245,246], Globus [247], NOMAD [248], OSF [249], Harvard Dataverse, Materials Project [27], and ScienceDB [250]. Community Resource for Innovation in Polymer Technology (CRIPT) [251], a recent community-curated polymer database is an excellent example of an ongoing effort that pays attention to the heterogeneous types of data (material, chemistry, processing condition, and characterization) that exist during polymer material synthesis and characterization. CRIPT utilizes a graph structure for mapping of a particular material synthesis processing with material, characterization data, processing procedures and physical properties. We believe that curation of characterization data along with structure and property data of polymer materials can facilitate better understanding of how the processing history factor in the making of the material and enable multi-faceted data-driven research on structure-processing-property relationship of polymeric materials.

4.2. Uniform metadata of polymer characterization data and processing procedures

In a recent commentary [252] published in Scientific Data, Ghiringhelli et al. have described NOMAD Metainfo [253], a data schema for storing metadata about material and molecular properties obtainable from computational workflows such as electronic-structure theory, quantum chemistry, and molecular dynamics simulations. Examples by Ghiringhelli et al. demonstrate integrated computational workflows for electronic structure calculations supported by NOMAD Metainfo. Along these lines, there is a need for the development of a unified description of metadata for polymer material experimental characterization, and processing procedures. Individual deposits of polymer synthesis and characterization data in general purpose databases are like scattered gemstones. Unification of metadata of open-access datasets can help bring more standardization to polymer material synthesis, characterizations, and processing procedures, and provide more insights into the general challenges and common characteristics of different processes. We are witnessing more reports of harnessing natural language processing (NLP) for knowledge extraction of scientific material research [65,254,255], chemistry [256-258], reactions [259,260], material synthesis procedures [64,261-264], characterization data [130,265] and polymer properties [266,267] in recent years. Built on the inorganic material synthesis procedure datasets extracted from the Ceder group, Wang et al. have proposed a unified language for describing synthesis procedures of inorganic materials called ULSA [268]. Encompassing essential vocabulary of solid state, sol-gel, and solution-based inorganic material synthesis procedures, ULSA is a valuable effort towards unifying metadata describing material synthesis procedures. Similar efforts can be developed for synthesis schemes developed by researchers in the polymer science community.

4.3. Interdisciplinary training of workforce

With the current push for data-driven approaches for accelerating materials design and for AI-driven automation in chemical industries' research and development (R&D), there is a critical need for universities to invest in educational programs that train the workforce in an interdisciplinary manner. Higher education institutions usually only offer graduate degrees (Masters, Doctorates) in specific disciplines where the graduate students deepen the technical background knowledge they gained during their undergraduate education. However, it would be valuable to have students step out of that comfort zone of their core discipline and learn and collaborate with students from completely different technical backgrounds and associated cultures. In a real-life scenario, it takes a team composed of computer scientist(s), data scientist(s), polymer scientist(s), and electrical/electronics engineer(s) to build high-throughput characterization instrumentation for formulations and develop relevant ML methods to achieve the desired analyses tasks on characterization results. If researchers from each of these diverse disciplines remained in their own silos during their graduate training, they will not learn about other disciplinary cultures and technical jargon, which can hamper progress in real-life scenarios in industries and national laboratories. Interdisciplinary classes will also improve communication across disciplines and lead to the creation of customized ML models for the polymer science problem at hand. For example, for a computer scientist/data scientist to customize methods that suit the polymer scientists it would help if they knew how to express with minimal language barriers their needs (e.g., polymer scientist describing exactly what the model should accomplish) or their challenges (e.g., why the collected data is not leading to high-performance with the model and what can be done better). If academic institutions invest in personnel (e.g., faculty members) and resources (e.g., classroom space, laboratories) for creating new and practical interdisciplinary professional degree or certificate programs that complement existing pure disciplinary strengths, they will better prepare students for future careers in institutions that value collaboration and interdisciplinary competency. There are often barriers to investment for development of large degree/certificate programs without proven success in smaller pilot programs. Examples of pilot programs include project-based interdisciplinary courses that bring together graduate students from different degree programs within an institution. Teamwork in project-based classes forces students to practice effective communication across disciplines and experience real-life team dynamics that occur in larger collaborations or in industries.

5. Conclusion

We have provided a review of ML models and methods for analyzing results from three commonly used classes of structural characterization methods in polymer science and engineering: microscopy, scattering, and spectroscopy. We have highlighted recent developments and applications of ML models and workflows that have enabled automation, classification, segmentation, property prediction, and reconstruction of structural characterization data from these techniques. In some cases, we have shared developments and applications that occurred in fields outside of polymer science and engineering because we felt these approaches could be extended to polymer research. In the last section we have described some current barriers to wide-spread use of ML for analyzing polymer characterization and potential ways to address them so that we can advance the successful use of ML for structural characterization of polymer material.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The author Arthi Jayaraman is an Associate Editor for *[Macromolecules]* and was not involved in the editorial review or the decision to publish this article.

CRediT authorship contribution statement

Shizhao Lu: Conceptualization, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Arthi Jayaraman:** Conceptualization, Funding acquisition, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Data availability

No data was used for the research described in the article.

Acknowledgements

SL thanks the National Science Foundation (NSF) for financial support via the NSF-DMREF Grant 1921871. AJ is grateful to National Science Foundation grants (DMREF 1921871 and 1629156, CBET 1703402, and CMMT 2105744), Department of Energy Basic Energy Sciences (DE-SC0023264), and Air Force Office of Scientific Research (MURI-Melanin from AFOSR FA 9550-18-1-0142) for supporting her many collaborations with experimentalists - Karen Wooley, Darrin Pochan, Kristi Kiick, April Kloxin, Todd Emrick, Jessica Schiffman, Ali Dhinojwala, Nathan Gianneschi, Bhuvnesh Bharti, Quentin Michaudel, and Ryan Hayward. Thought-provoking discussions with these experimentalists and their research groups during these funded collaborations, along with their carefully collected experimental data has led to the development and use of machine learning (ML) workflows in the Jayaraman research group and the knowledge that AJ and SL share in this review.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.progpolymsci.2024.101828.

References

[1] Bockstaller MR, Mickiewicz RA, Thomas EL. Block copolymer nanocomposites: perspectives for tailored functional materials. Adv Mater 2005;17:1331–49.
[2] Orilall MC, Wiesner U. Block copolymer based composition and morphology control in nanostructured hybrid materials for energy conversion and storage: solar cells, batteries, and fuel cells. Chem Soc Rev 2011;40:520–35.
[3] Sarkar B, Alexandridis P. Block copolymer–nanoparticle composites: Structure, functional properties, and processing. Prog Polym Sci 2015;40:33–62.
[4] Doerk GS, Yager KG. Beyond native block copolymer morphologies. Mol Sys Des Eng 2017;2:518–38.
[5] Fasolka MJ, Mayes AM. Block copolymer thin films: Physics and applications. Annu Rev Mater Res 2001;31:323–55.
[6] Hamley I. Ordering in thin films of block copolymers: Fundamentals to potential applications. Prog Polym Sci 2009;34:1161–210.
[7] Yu L, Dean K, Li L. Polymer blends and composites from renewable resources. Prog Polym Sci 2006;31:576–602.
[8] Sionkowska A. Current research on the blends of natural and synthetic polymers as new biomaterials. Prog Polym Sci 2011;36:1254–76.
[9] Balazs AC, Emrick T, Russell TP. Nanoparticle polymer composites: Where two small worlds meet. Science (1979) 2006;314:1107–10.
[10] Moniruzzaman M, Winey KI. Polymer nanocomposites containing carbon nanotubes. Macromol 2006;39:5194–205.
[11] Vaia RA, Maguire JF. Polymer nanocomposites with prescribed morphology: going beyond nanoparticle-filled polymers. Chem Mater 2007;19:2736–51.
[12] Zeng QH, Yu AB, Lu GQ. Multiscale modeling and simulation of polymer nanocomposites. Prog Polym Sci 2008;33:191–269.
[13] Jancar J, Douglas JF, Starr FW, Kumar SK, Cassagnau P, Lesser AJ, et al. Current issues in research on structure-property relationships in polymer nanocomposites. Polymer (Guildf) 2010;51:3321–43.
[14] Kumar SK, Krishnamoorti R. Nanocomposites: Structure, phase behavior, and properties. Annu Rev Chem Biomol Eng 2010;1:37–58.
[15] Hore MJA, Composto RJ. Functional polymer nanocomposites enhanced by nanorods. Macromol 2014;47:875–87.
[16] Ganesan V, Jayaraman A. Theory and simulation studies of effective interactions, phase behavior and morphology in polymer nanocomposites. Soft Matter 2014;10:13–38.
[17] Liu M, Jia Z, Jia D, Zhou C. Recent advance in research on halloysite nanotubes-polymer nanocomposite. Prog Polym Sci 2014;39:1498–525.
[18] Kotal M, Bhowmick AK. Polymer nanocomposites from modified clays: Recent advances and challenges. Prog Polym Sci 2015;51:127–87.
[19] Kumar SK, Ganesan V, Riggleman RA. Perspective: Outstanding theoretical questions in polymer-nanoparticle hybrids. J Chem Phys 2017:147.
[20] Gartner TE, Jayaraman A. Modeling and Simulations of Polymers: A Roadmap. Macromol 2019;52:755–86.
[21] Li J, Liu X, Feng Y, Yin J. Recent progress in polymer/two-dimensional nanosheets composites with novel performances. Prog Polym Sci 2022;126:101505.
[22] Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res 2020;50:71–103.
[23] Kadulkar S, Sherman ZM, Ganesan V, Truskett TM. Machine Learning–Assisted Design of Material Properties. Annu Rev Chem Biomol Eng 2022;13:235–54.
[24] Duan C, Nandy A, Kulik HJ. Machine learning for the discovery, design, and engineering of materials. Annu Rev Chem Biomol Eng 2022;13:405–29.
[25] Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. Npj Comput Mater 2022;8:59.
[26] Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature 2023;624:80–5.
[27] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Mater 2013;1.
[28] Szymanski NJ, Rendy B, Fei Y, Kumar RE, He T, Milsted D, et al. An autonomous laboratory for the accelerated synthesis of novel materials. Nature 2023;624:86–91.
[29] Ge M, Su F, Zhao Z, Su D. Deep learning analysis on microscopic imaging in materials science. Materials Today Nano 2020;11:100087.
[30] Spurgeon SR, Ophus C, Jones L, Petford-Long A, Kalinin SV, Olszta MJ, et al. Towards data-driven next-generation transmission electron microscopy. Nat Mater 2021;20:274–9.
[31] Kalinin SV, Ziatdinov M, Hinkle J, Jesse S, Ghosh A, Kelley KP, et al. Automated and autonomous experiments in electron and scanning probe microscopy. ACS Nano 2021;15:12604–27.
[32] Baskaran A, Kautz EJ, Chowdhary A, Ma W, Yener B, Lewis DJ. Adoption of Image-Driven Machine Learning for Microstructure Characterization and Materials Design: A Perspective. JOM 2021;73:3639–57.
[33] Ede JM. Deep learning in electron microscopy. Mach Learn Sci Technol 2021;2:011004.
[34] Treder KP, Huang C, Kim JS, Kirkland AI. Applications of deep learning in electron microscopy. Microscopy 2022;71 i100-i15.
[35] Zhang L, Shao S. Image-based machine learning for materials science. J Appl Phys 2022;132:100701.
[36] Botifoll M, Pinto-Huguet I, Arbiol J. Machine learning in electron microscopy for advanced nanocharacterization: current developments, available tools and future outlook. Nanoscale Horiz 2022.
[37] Kalinin SV, Mukherjee D, Roccapriore K, Blaiszik BJ, Ghosh A, Ziatdinov MA, et al. Machine learning for automated experimentation in scanning transmission electron microscopy. Npj Comput Mater 2023;9:227.
[38] Chen Z, Andrejevic N, Drucker NC, Nguyen T, Xian RP, Smidt T, et al. Machine learning on neutron and x-ray scattering and spectroscopies. Chem Phys Rev 2021;2.
[39] Yager KG, Majewski PW, Noack MM, Fukuto M. Autonomous x-ray scattering. Nanotechnology 2023;34:322001.
[40] Anker AS, Butler KT, Selvan R, KMØ Jensen. Machine learning for analysis of experimental scattering and spectroscopy data in materials chemistry. Chem Sci 2023;14:14003–19.
[41] Unruh D, Kolluru VSC, Baskaran A, Chen Y, Chan MK. Theory+ AI/ML for microscopy and spectroscopy: Challenges and opportunities. MRS Bulletin 2022;47:1024–35.
[42] Guo D, Chen XJ, Lu ML, He WF, Luo SH, Lin YQ, et al. Review and Prospect: Applications of Exponential Signals with Machine Learning in Nuclear Magnetic Resonance. Spectroscopy 2023;38:22–32.
[43] Audus DJ, De Pablo JJ. Polymer Informatics: Opportunities and Challenges. ACS Macro Lett 2017;6:1078–82.
[44] Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. Npj Comput Mater 2017;3:54.
[45] Ferguson AL, Brown KA. Data-driven design and autonomous experimentation in soft and biological materials engineering. Annu Rev Chem Biomol Eng 2022;13:25–44.
[46] Day EC, Chittari SS, Bogen MP, Knight AS. Navigating the Expansive Landscapes of Soft Materials: A User Guide for High-Throughput Workflows. ACS Polym Au 2023;3:406–27.

[47] Ginige G, Song YD, Olsen BC, Luber EJ, Yavuz CT, Buriak JM. Solvent Vapor Annealing, Defect Analysis, and Optimization of Self-Assembly of Block Copolymers Using Machine Learning Approaches. ACS Appl Mater Inter 2021;13:28639–49.
[48] Genix AC, Oberdisse J. Structure and dynamics of polymer nanocomposites studied by X-ray and neutron scattering techniques. Curr Opin Colloid Interface Sci 2015;20:293–303.
[49] Son D, Cho S, Nam J, Lee H, Kim M. X-ray-based spectroscopic techniques for characterization of polymer nanocomposite materials at a molecular level. Polymers (Basel) 2020;12:1053.
[50] Morozova S, Hitimana E, Dhakal S, Wilcox KG, Estrin D. Scattering methods for determining structure and dynamics of polymer gels. J Appl Phys 2021;129:071101.
[51] Danielsen SPO, Beech HK, Wang S, El-Zaatari BM, Wang X, Sapir L, et al. Molecular Characterization of Polymer Networks. Chem Rev 2021;121:5042–92.
[52] Wei Y, Hore MJA. Characterizing polymer structure with small-angle neutron scattering: A Tutorial. J Appl Phys 2021;129:171101.
[53] Ferguson AL. Machine learning and data science in soft materials engineering. J Phys Condens Matter 2018;30:043002.
[54] Sattari K, Xie Y, Lin J. Data-driven algorithms for inverse design of polymers. Soft Matter 2021;17:7607–22.
[55] Patra TK. Data-driven methods for accelerating polymer design. ACS Polym Au 2021;2:8–26.
[56] Amamoto Y. Data-driven approaches for structure-property relationships in polymer science for prediction and understanding. Polym J 2022;54:957–67.
[57] Martin TB, Audus DJ. Emerging Trends in Machine Learning: A Polymer Perspective. ACS Polym Au 2023;3:239–58.
[58] Li D, Ru Y, Chen Z, Dong C, Dong Y, Liu J. Accelerating the design and development of polymeric materials via deep learning: Current status and future challenges. APL Machine Learning 2023;1.
[59] Chen L, Pilania G, Batra R, Huan TD, Kim C, Kuenneth C, et al. Polymer informatics: Current status and critical next steps. Materials Science and Engineering: R: Reports 2021;144:100595.
[60] Sha W, Li Y, Tang S, Tian J, Zhao Y, Guo Y, et al. Machine learning in polymer informatics. InfoMat 2021;3:353–61.
[61] Patel RA, Webb MA. Data-Driven Design of Polymer-Based Biomaterials: High-throughput Simulation, Experimentation, and Machine Learning. ACS Appl Bio Mater 2024;7:510–27.
[62] Gormley AJ, Webb MA. Machine learning in combinatorial polymer chemistry. Nat Rev Mater 2021;6:642–4.
[63] Upadhya R, Kosuri S, Tamasi M, Meyer TA, Atta S, Webb MA, et al. Automation and data-driven design of polymer therapeutics. Adv Drug Deliv Rev 2021;171:1–28.
[64] Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E. Materials synthesis insights from scientific literature via text extraction and machine learning. Chem Mater 2017;29:9436–44.
[65] Olivetti EA, Cole JM, Kim E, Kononova O, Ceder G, Han TYJ, et al. Data-driven materials research enabled by natural language processing and information extraction. Appl Phys Rev 2020;7.
[66] Lyu Z, Yao L, Chen W, Kalutantirige FC, Chen Q. Electron Microscopy Studies of Soft Nanomaterials. Chem Rev 2023;123:4051–145.
[67] Goodhew PJ, Humphreys J, Beanland R. Electron microscopy and analysis. CRC press; 2000.
[68] Mortensen K. Small-angle X-ray and neutron scattering studies from multiphase polymers. Curr Opin Solid State Mater Sci 1997;2:653–60.
[69] Pedersen JS. Analysis of small-angle scattering data from micelles and microemulsions: free-form approaches and model fitting. Curr Opin Colloid Interface Sci 1999;4:190–6.
[70] Castelletto V, Hamley IW. Modelling small-angle scattering data from micelles. Curr Opin Colloid Interface Sci 2002;7:167–72.
[71] Pedersen JS, Svaneborg C. Scattering from block copolymer micelles. Curr Opin Colloid Interface Sci 2002;7:158–66.
[72] Hamley IW, Castelletto V. Small-angle scattering of block copolymers in the melt, solution and crystal states. Prog Polym Sci 2004;29:909–48.
[73] Walker LM. Scattering from polymer-like micelles. Curr Opin Colloid Interface Sci 2009;14:451–4.
[74] Lund R., Willner L., Richter D. Kinetics of Block Copolymer Micelles Studied by Small-Angle Scattering Methods. In: Abe A, Lee KS, Leibler L, Kobayashi S, editors. Controlled polymerization and polymeric structures: flow microreactor polymerization, micelles kinetics, polypeptide ordering, light emitting nanostructures 2013. p. 51–158.
[75] Pokorski JK, Hore MJA. Structural Characterization of Protein-Polymer Conjugates for Biomedical Applications with Small-Angle Scattering. Curr Opin Colloid Interface Sci 2019.
[76] Jeffries CM, Ilavsky J, Martel A, Hinrichs S, Meyer A, Pedersen JS, et al. Small-angle X-ray and neutron scattering. Nat Rev Methods Primers 2021;1:1–39.
[77] Zemb T., Lindner P. Neutrons, X-rays and light: scattering methods applied to soft condensed matter. North-Holland; 2002.
[78] Guinier A, Fournet G, Yudowitch KL. Small-angle scattering of X-rays. New York: Wiley; 1955.
[79] Guinier A, Fournet G, Walker CB, Yudowitch KL. Small-angle scattering of X-rays. New York: Wiley; 1979.
[80] Feigin L, Svergun DI. Structure analysis by small-angle X-ray and neutron scattering. Springer; 1987.
[81] Kline SR. Reduction and analysis of SANS and USANS data using IGOR Pro. J Appl Crystallogr 2006;39:895–900.
[82] Widjonarko NE. Introduction to advanced X-ray diffraction techniques for polymeric thin films. Coatings 2016;6:54.
[83] Uzun İ. Methods of determining the degree of crystallinity of polymers with X-ray diffraction: A review. J Polym Res 2023;30:394.
[84] Riaz U, Ashraf SM. Characterization of Polymer Blends with FTIR Spectroscopy. In: Thomas S, Grohens Y, Jyotishkumar P, editors. Characterization of polymer blends: miscibility, morphology and interfaces. Wiley-VCH; 2014. p. 625–78.
[85] Ade H, Hitchcock AP. NEXAFS microscopy and resonant scattering: Composition and orientation probed in real and reciprocal space. Polymer (Guildf) 2008;49:643–75.
[86] Holland-moritz K, Siesler HW. Infrared Spectroscopy of Polymers. Appl Spectrosc Rev 1976;11:1–55.
[87] Urquhart SG, Hitchcock AP, Smith AP, Ade HW, Lidy W, Rightor EG, et al. NEXAFS spectromicroscopy of polymers: overview and quantitative analysis of polyurethane polymers. J Eletron Spectrosc Relat Phenomena 1999;100:119–35.
[88] Jouault N, Dalmas F, Sr Said, Di Cola E, Schweins R, Jestin J, et al. Direct measurement of polymer chain conformation in well-controlled model nanocomposites by combining SANS and SAXS. Macromol 2010;43:9881–91.
[89] Jung YS, Ross CA. Orientation-controlled self-assembled nanolithography using a polystyrene− polydimethylsiloxane block copolymer. Nano Lett 2007;7:2046–50.
[90] Strawhecker K, Manias E. AFM of poly (vinyl alcohol) crystals next to an inorganic surface. Macromol 2001;34:8475–82.
[91] Doerk GS, Stein A, Bae S, Noack MM, Fukuto M, Yager KG. Autonomous discovery of emergent morphologies in directed self-assembly of block copolymer blends. Sci Adv 2023;9:eadd3687.
[92] Füllbrandt M, Purohit PJ, Schönhals A. Combined FTIR and dielectric investigation of poly (vinyl acetate) adsorbed on silica particles. Macromol 2013;46:4626–32.
[93] Dong F, Wang Z, Li Y, Ho WK, Lee S. Immobilization of polymeric g-C3N4 on structured ceramic foam for efficient visible light photocatalytic air purification with real indoor illumination. Environ Sci Tech 2014;48:10345–53.
[94] Su GM, Patel SN, Pemmaraju C, Prendergast D, Chabinyc ML. First-principles predictions of near-edge X-ray absorption fine structure spectra of semiconducting polymers. J Phys Chem C 2017;121:9142–52.
[95] Shen D, Wu G, Suk HI. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017;19:221.
[96] Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: From machine learning to deep learning. Physica Medica 2021;83:9–24.
[97] Mellouk W, Handouzi W. Facial emotion recognition using deep learning: review and insights. Procedia Comput Sci 2020;175:689–94.
[98] Grigorescu S, Trasnea B, Cocias T, Macesanu G. A survey of deep learning techniques for autonomous driving. J Field Robotics 2020;37:362–86.
[99] Cui Y, Chen R, Chu W, Chen L, Tian D, Li Y, et al. Deep learning for image and point cloud fusion in autonomous driving: A review. IEEE T-ITS 2021;23:722–39.
[100] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 1995;3361:1995.
[101] Dhillon A, Verma GK. Convolutional neural network: a review of models, methodologies and applications to object detection. Progress in Artificial Intelligence 2020;9:85–112.
[102] LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86:2278–324.
[103] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60:84–90.
[104] Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint 2014.
[105] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2818–26.
[106] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.
[107] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1251–8.
[108] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence; 2017.
[109] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1492–500.
[110] Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint 2017.
[111] Tan M., Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International conference on machine learning: PMLR; 2019. p. 6105–14.
[112] Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8:1–74.

[113] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–55.
[114] Modarres MH, Aversa R, Cozzini S, Ciancio R, Leto A, Brandino GP. Neural network for nanoscience scanning electron microscope image recognition. Sci Rep 2017;7:1–12.
[115] Xu H, Ma S, Hou Y, Zhang Q, Wang R, Luo Y, et al. Machine learning-assisted identification of copolymer microstructures based on microscopic images. ACS Appl Mater Inter 2022;14:47157–66.
[116] Pokuri BSS, Ghosal S, Kokate A, Sarkar S, Ganapathysubramanian B. Interpretable deep learning for guided microstructure-property explorations in photovoltaics. Npj Comput Mater 2019;5:95.
[117] Lu S, Montz B, Emrick T, Jayaraman A. Semi-supervised machine learning workflow for analysis of nanowire morphologies from transmission electron microscopy images. Digi Discov 2022;1:816–33.
[118] Wang X, Li J, Ha HD, Dahl JC, Ondry JC, Moreno-Hernandez I, et al. AutoDetect-mNP: an unsupervised machine learning algorithm for automated analysis of transmission electron microscope images of metal nanoparticles. JACS Au 2021;1:316–27.
[119] Matuszewski DJ, Sintorn IM. TEM virus images: Benchmark dataset and deep learning classification. Comput Methods Programs Biomed 2021;209:106318.
[120] Liang Z, Tan Z, Hong R, Ouyang W, Yuan J, Zhang C. Automatically Predicting Material Properties with Microscopic Images: Polymer Miscibility as an Example. J Chem Inf Model 2023;63:5971–80.
[121] Visheratina A, Visheratin A, Kumar P, Veksler M, Kotov NA. Chirality Analysis of Complex Microparticles using Deep Learning on Realistic Sets of Microscopy Images. ACS Nano 2023;17:7431–42.
[122] Lu S, Wu Z, Jayaraman A. Molecular Modeling and Simulation of Polymer Nanocomposites with Nanorod Fillers. J Phys Chem B 2021;125:2435–49.
[123] Lu S, Jayaraman A. Effect of Nanorod Physical Roughness on the Aggregation and Percolation of Nanorods in Polymer Nanocomposites. ACS Macro Lett 2021;10:1416–22.
[124] Ziatdinov M, Dyck O, Maksov A, Li X, Sang X, Xiao K, et al. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. ACS Nano 2017;11:12742–52.
[125] Qu EZ, Jimenez AM, Kumar SK, Zhang K. Quantifying Nanoparticle Assembly States in a. Polymer Matrix through Deep Learning. Macromol 2021;54:3034–40.
[126] Bornani K, Mendez NF, Altorbaq AS, Müller AJ, Lin Y, Qu EZ, et al. Situ Atomic Force Microscopy Tracking of Nanoparticle Migration in Semicrystalline Polymers. ACS Macro Lett 2022;11:818–24.
[127] Yao L, Ou Z, Luo B, Xu C, Chen Q. Machine learning to reveal nanoparticle dynamics from liquid-phase TEM videos. ACS Cent Sci 2020;6:1421–30.
[128] Ronneberger O, Fischer P, U-net Brox T. Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer; 2015. p. 234–41.
[129] Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: A review of theory and applications. Ieee Access 2021;9:82031–57.
[130] Mukaddem KT, Beard EJ, Yildirim B, Cole JM. ImageDataExtractor: a tool to extract and quantify data from microscopy images. J Chem Inf Model 2019;60:2492–509.
[131] Lee B, Yoon S, Lee JW, Kim Y, Chang J, Yun J, et al. Statistical characterization of the morphologies of nanoparticles through machine learning based electron microscopy image analysis. ACS Nano 2020;14:17125–33.
[132] Wen H, Luna-Romera JM, Riquelme JC, Dwyer C, Chang SL. Statistically representative metrology of nanoparticles via unsupervised machine learning of TEM Images. Nanomaterials 2021;11:2706.
[133] Williamson EM, Ghrist AM, Karadaghi LR, Smock SR, Barim G, Brutchey RL. Creating ground truth for nanocrystal morphology: a fully automated pipeline for unbiased transmission electron microscopy analysis. Nanoscale 2022;14:15327–39.
[134] Yao L, An H, Zhou S, Kim A, Luijten E, Chen Q. Seeking regularity from irregularity: Unveiling the synthesis–nanomorphology relationships of heterogeneous nanomaterials using unsupervised machine learning. Nanoscale 2022;14:16479–89.
[135] Saaim KM, Afridi SK, Nisar M, Islam S. In search of best automated model: Explaining nanoparticle TEM image segmentation. Ultramicroscopy 2022;233:113437.
[136] Alom M.Z., Hasan M., Yakopcic C., Taha T.M., Asari V.K. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint 2018.
[137] Oktay O., Schlemper J., Folgoc L.L., Lee M., Heinrich M., Misawa K., et al. Attention u-net: Learning where to look for the pancreas. arXiv preprint 2018.
[138] Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Bi-directional ConvLSTM U-Net with densley connected convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision workshops; 2019. p. 0.
[139] Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J. Unet++: a nested u-net architecture for medical image segmentation. deep learning in medical image analysis and multimodal learning for clinical decision support. Springer; 2018. p. 3–11.
[140] Huang H, Lin L, Tong R, Hu H, Zhang Q, Iwamoto Y, et al. Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2020. p. 1055–9.
[141] Touve MA, Wright DB, Mu C, Sun H, Park C, Gianneschi NC. Block copolymer amphiphile phase diagrams by high-throughput transmission electron microscopy. Macromol 2019;52:5529–37.
[142] Mastronarde DN. Automated electron microscope tomography using robust prediction of specimen movements. J Struct Biol 2005;152:36–51.
[143] Vo GD, Park C. Robust regression for image binarization under heavy noise and nonuniform background. Pattern Recognit 2018;81:224–39.
[144] Srivastava A, Klassen E, Joshi SH, Jermyn IH. Shape analysis of elastic curves in euclidean spaces. IEEE Trans Pattern Anal Mach Intell 2010;33:1415–28.
[145] Krull A, Hirsch P, Rother C, Schiffrin A, Krull C. Artificial-intelligence-driven scanning probe microscopy. Comm Phys 2020;3:1–8.
[146] Roccapriore KM, Dyck O, Oxley MP, Ziatdinov M, Kalinin SV. Automated experiment in 4D-STEM: exploring emergent physics and structural behaviors. ACS Nano 2022;16:7605–14.
[147] Wilson AG, Hu Z, Salakhutdinov R, Xing EP. Deep kernel learning. Artificial intelligence and statistics: PMLR 2016:370–8.
[148] Vargo E, Dahl JC, Evans KM, Khan T, Alivisatos P, Xu T. Using Machine Learning to Predict and Understand Complex Self-Assembly Behaviors of a Multicomponent Nanocomposite. Adv Mater 2022;34:2203168.
[149] Wang B, Yager K, Yu D, Hoai M. X-ray scattering image classification using deep learning. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE; 2017. p. 697–704.
[150] Ikemoto H, Yamamoto K, Touyama H, Yamashita D, Nakamura M, Okuda H. Classification of grazing-incidence small-angle X-ray scattering patterns by convolutional neural network. J Synchrotron Radiat 2020;27:1069–73.
[151] Van Herck W, Fisher J, Ganeva M. Deep learning for x-ray or neutron scattering under grazing-incidence: extraction of distributions. Mater Res Express 2021;8:045015.
[152] Babonneau D. FitGISAXS: software package for modelling and analysis of GISAXS data using IGOR Pro. J Appl Crystallogr 2010;43:929–36.
[153] Aty HA, Strutt R, Mcintyre N, Allen M, Barlow NE, Páez-Pérez M, et al. Machine learning platform for determining experimental lipid phase behaviour from small angle X-ray scattering patterns by pre-training on synthetic data. Digi Discov 2022;1:98–107.
[154] Pearson KLIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science 1901;2:559–72.
[155] Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9.
[156] McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint 2018.
[157] Ziletti A, Kumar D, Scheffler M, Ghiringhelli LM. Insightful classification of crystal structures using deep learning. Nature Comm 2018;9:2775.
[158] Stanev V, Vesselinov VV, Kusne AG, Antoszewski G, Takeuchi I, Alexandrov BS. Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. Npj Comput Mater 2018;4:43.
[159] Aguiar J, Gong ML, Unocic R, Tasdizen T, Miller B. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. Sci Adv 2019;5:eaaw1949.
[160] Oviedo F, Ren Z, Sun S, Settens C, Liu Z, Hartono NTP, et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. Npj Comput Mater 2019;5:60.
[161] Lee JW, Park WB, Lee JH, Singh SP, Sohn KS. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. Nature Comm 2020;11:86.
[162] Kaufmann K, Zhu C, Rosengarten AS, Maryanovsky D, Harrington TJ, Marin E, et al. Crystal symmetry determination in electron diffraction using machine learning. Science (1979) 2020;367:564–8.
[163] Lim B, Bellec E, Dupraz M, Leake S, Resta A, Coati A, et al. A convolutional neural network for defect classification in Bragg coherent X-ray diffraction. Npj Comput Mater 2021;7:115.
[164] Schopmans H, Reiser P, Friederich P. Neural networks trained on synthetically generated crystals can extract structural information from ICSD powder X-ray diffractograms. Digi Discov 2023;2:1414–24.
[165] Bergerhoff G, Hundt R, Sievers R, Brown I. The inorganic crystal structure data base. J Chem Inf Comp Sci 1983;23:66–9.
[166] Shen X, Hu W, Russell TP. Measuring the degree of crystallinity in semicrystalline regioregular poly (3-hexylthiophene). Macromol 2016;49:4501–9.
[167] Mileva D, Tranchida D, Gahleitner M. Designing polymer crystallinity: An industrial perspective. Polym Cryst 2018;1:e10009.
[168] Doumeng M, Makhlouf L, Berthet F, Marsan O, Delbé K, Denape J, et al. A comparative study of the crystallinity of polyetheretherketone by using density, DSC, XRD, and Raman spectroscopy techniques. Polym Test 2021;93:106878.
[169] Wu Z, Wu JW, Michaudel Q, Jayaraman A. Investigating the Hydrogen Bond-Induced Self-Assembly of Polysulfamides Using Molecular Simulations and Experiments. Macromol 2023;56:5033–49.
[170] Venkatram S, McCollum J, Stingelin N, Brettmann B. A close look at polymer degree of crystallinity versus polymer crystalline quality. Polym Int 2023;72:855–60.
[171] Archibald RK, Doucet M, Johnston T, Young SR, Yang E, Heller WT. Classifying and analyzing small-angle scattering data using weighted knearest neighbors machine learning techniques. J Appl Crystallogr 2020;53:326–34.

[172] Song G, Porcar L, Boehm M, Cecillon F, Dewhurst C, Le Goc Y, et al. Deep learning methods on neutron scattering data. EPJ Web of Conferences: EDP Sciences; 2020:01004.
[173] Tomaszewski P, Yu S, Borg M, Rönnols J. Machine learning-assisted analysis of small angle x-ray scattering. In: 2021 Swedish Workshop on Data Science (SweDS). IEEE; 2021. p. 1–6.
[174] Franke D, Jeffries CM, Svergun DI. Machine learning methods for X-ray scattering data analysis from biomacromolecular solutions. Biophys J 2018;114:2485–92.
[175] He H, Liu C, Liu H. Model reconstruction from small-angle x-ray scattering data using deep learning methods. iScience 2020:23.
[176] Beltran-Villegas DJ, Intriago D, Kim KHC, Behabtu N, Londono JD, Jayaraman A. Coarse-grained molecular dynamics simulations of α-1,3-glucan. Soft Matter 2019;15:4669–81.
[177] Heil CM, Jayaraman A. Computational Reverse-Engineering Analysis for Scattering Experiments of Assembled Binary Mixture of Nanoparticles. ACS Mater Au 2021;1:140.
[178] Wessels MG, Jayaraman A. Machine learning enhanced computational reverse engineering analysis for scattering experiments (crease) to determine structures in amphiphilic polymer solutions. ACS Polym Au 2021:8581–91.
[179] Wessels MG, Jayaraman A. Computational Reverse-Engineering Analysis of Scattering Experiments (CREASE) on Amphiphilic Block Polymer Solutions: Cylindrical and Fibrillar Assembly. Macromol 2021;54:783–96.
[180] Ye ZY, Wu ZJ, Jayaraman A. Computational Reverse Engineering Analysis for Scattering Experiments (CREASE) on Vesicles Assembled from Amphiphilic Macromolecular Solutions. JACS Au 2021;1:1925–36.
[181] Heil CM, Patil A, Dhinojwala A, Jayaraman A. Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) with Machine Learning Enhancement to Determine Structure of Nanoparticle Mixtures and Solutions. ACS Cent Sci 2022;8:996–1007.
[182] Wu ZJ, Jayaraman A. Machine Learning-Enhanced Computational Reverse-Engineering Analysis for Scattering Experiments (CREASE) for Analyzing Fibrillar Structures in Polymer Solutions. Macromol 2022;55:11076–91.
[183] Heil CM, Ma YZ, Bharti B, Jayaraman A. Computational Reverse-Engineering Analysis for Scattering Experiments for Form Factor and Structure Factor Determination (?P(q) and S(q) CREASE?). JACS Au 2023;3:889–904.
[184] Akepati SVR, Gupta N, Jayaraman A. Computational Reverse Engineering Analysis of the Scattering Experiment Method for Interpretation of 2D Small-Angle Scattering Profiles (CREASE-2D). JACS Au 2024. doi:10.1021/jacsau.4c00068.
[185] Beltran-Villegas DJ, Wessels MG, Lee JY, Song Y, Wooley KL, Pochan DJ, et al. Computational reverse-engineering analysis for scattering experiments on amphiphilic block polymer solutions. JACS 2019;141:14916–30.
[186] Lee JY, Song Y, Wessels MG, Jayaraman A, Wooley KL, Pochan DJ. Hierarchical Self-Assembly of Poly (d-glucose carbonate) Amphiphilic Block Copolymers in Mixed Solvents. Macromol 2020;53:8581–91.
[187] Patil A, Heil CM, Vanthournout B, Bleuel M, Singla S, Hu ZY, et al. Structural Color Production in Melanin-Based Disordered Colloidal Nanoparticle Assemblies in Spherical Confinement. Adv Opt Mater 2022;10.
[188] Patil A, Heil CM, Vanthournout B, Singla S, Hu ZY, Ilavsky J, et al. Modeling Structural Colors from Disordered One-Component Colloidal Nanoparticle-Based Supraballs Using Combined Experimental and Simulation Techniques. ACS Mater Lett 2022;4:1848–54.
[189] Heil CM, Patil A, Vanthournout B, Singla S, Bleuel M, Song JJ, et al. Mechanism of structural colors in binary mixtures of nanoparticle-based supraballs. Sci Adv 2023;9.
[190] Ma YZ, Heil C, Nagy G, Heller WT, An YX, Jayaraman A, et al. Synergistic Role of Temperature and Salinity in Aggregation of Nonionic Surfactant-Coated Silica Nanoparticles. Langmuir 2023;39:5917–28.
[191] Gupta N, Jayaraman A. Computational approach for structure generation of anisotropic particles (CASGAP) with targeted distributions of particle design and orientational order. Nanoscale 2023;15:14958–70.
[192] Röding M, Tomaszewski P, Yu S, Borg M, Rönnols J. Machine learning-accelerated small-angle X-ray scattering analysis of disordered two-and three-phase materials. Front Mater 2022;9:956839.
[193] Chang MC, Tung CH, Chang SY, Carrillo JM, Wang Y, Sumpter BG, et al. A machine learning inversion scheme for determining interaction from scattering. Comm Phys 2022;5:46.
[194] Zhao C, Yu W, Li L. Visualization of small-angle X-ray scattering datasets and processing-structure mapping of isotactic polypropylene films by machine learning. Mater Des 2023;228:111828.
[195] Noack MM, Doerk GS, Li R, Streit JK, Vaia RA, Yager KG, et al. Autonomous materials discovery driven by Gaussian process regression with inhomogeneous measurement noise and anisotropic kernels. Sci Rep 2020;10:17663.
[196] Noack MM, Doerk GS, Li R, Fukuto M, Yager KG. Advances in kriging-based autonomous x-ray scattering experiments. Sci Rep 2020;10:1325.
[197] Beaucage PA, Martin TB. The Autonomous Formulation Laboratory: An Open Liquid Handling Platform for Formulation Discovery Using X-ray and Neutron Scattering. Chem Mater 2023;35:846–52.
[198] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.
[199] Tetef S, Govind N, Seidler GT. Unsupervised machine learning for unbiased chemical classification in X-ray absorption spectroscopy and X-ray emission spectroscopy. Phys Chem Chem Phys 2021;23:23586–601.
[200] Jung G, Jung SG, Cole JM. Automatic materials characterization from infrared spectra using convolutional neural networks. Chem Sci 2023;14:3600–9.
[201] Chen Y., Chen C., Hwang I., Davis M.J., Yang W., Sun C., et al. Robust Machine Learning Inference from X-ray Absorption Near Edge Spectra through Featurization. arXiv preprint 2023.
[202] Tian SIP, Ren Z, Venkataraj S, Cheng Y, Bash D, Oviedo F, et al. Tackling data scarcity with transfer learning: a case study of thickness characterization from optical spectra of perovskite thin films. Digi Discov 2023;2:1334–46.
[203] Cobas C. NMR signal processing, prediction, and structure verification with machine learning techniques. Magnetic Resonance in Chemistry 2020;58:512–19.
[204] Fricke SN, Salgado M, Menezes T, Santos KMC, Gallagher NB, Song AY, et al. Multivariate Machine Learning Models of Nanoscale Porosity from Ultrafast NMR Relaxometry. Angewandte Chemie-International Edition 2024;63:e202316664.
[205] Grazioli G, Martin RW, Butts CT. Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. Front Mol Biosci 2019;6.
[206] Li DW, Hansen AL, Bruschweiler-Li L, Yuan CH, Bruschweiler R. Fundamental and practical aspects of machine learning for the peak picking of biomolecular NMR spectra. J Biomol NMR 2022;76:49–57.
[207] Vaddi K, Chiang HT, Pozzo LD. Autonomous retrosynthesis of gold nanoparticles via spectral shape matching. Digi Discov 2022;1:502–10.
[208] Vaddi K, Li K, Pozzo LD. Metric geometry tools for automatic structure phase map generation. Digi Discov 2023;2:1471–83.
[209] Chiang HT, Vaddi K, Pozzo LD. Data-Driven Exploration of Silver Nanoplate Formation in Multidimensional Chemical Design Spaces. ChemRxiv 2023. doi:10.26434/chemrxiv-2023-60kd6.
[210] Tamasi MJ, Patel RA, Borca CH, Kosuri S, Mugnier H, Upadhya R, et al. Machine learning on a robotic platform for the design of polymer–protein hybrids. Adv Mater 2022;34:2201809.
[211] Meyer TA, Ramirez C, Tamasi MJ, Gormley AJ. A user's guide to machine learning for polymeric biomaterials. ACS Polym Au 2022;3:141–57.
[212] Cang R, Li H, Yao H, Jiao Y, Ren Y. Improving direct physical properties prediction of heterogeneous materials from imaging data via convolutional neural network and a morphology-aware generative model. Comput Mater Sci 2018;150:212–21.
[213] Stein HS, Guevarra D, Newhouse PF, Soedarmadji E, Gregoire JM. Machine learning of optical properties of materials–predicting spectra from images and images from spectra. Chem Sci 2019;10:47–55.
[214] Kim Y, Park HK, Jung J, Asghari-Rad P, Lee S, Kim JY, et al. Exploration of optimal microstructure and mechanical properties in continuous microstructure space using a variational autoencoder. Mater Des 2021;202:109544.
[215] Yaman MY, Kalinin SV, Guye KN, Ginger DS, Ziatdinov M. Learning and Predicting Photonic Responses of Plasmonic Nanoparticle Assemblies via Dual Variational Autoencoders. Small 2023;19:2205893.
[216] Lu S, Jayaraman A. Pair-Variational Autoencoders for Linking and Cross-Reconstruction of Characterization Data from Complementary Structural Characterization Techniques. JACS Au 2023;3:2510–21.
[217] Liu D, Tan Y, Khoram E, Yu Z. Training deep neural networks for the inverse design of nanophotonic structures. ACS Photonics 2018;5:1365–9.
[218] So S, Mun J, Rho J. Simultaneous inverse design of materials and structures via deep learning: demonstration of dipole resonance engineering using core–shell nanoparticles. ACS Appl Mater Inter 2019;11:24264–8.
[219] Zhou Q, Yong B, Lv Q, Shen J, Wang X. Deep autoencoder for mass spectrometry feature learning and cancer detection. IEEE Access 2020;8:45156–66.
[220] Banko L, Maffettone PM, Naujoks D, Olds D, Ludwig A. Deep learning for visualization and novelty detection in large X-ray diffraction datasets. Npj Comput Mater 2021;7:104.
[221] Grossutti M, D'Amico J, Quintal J, MacFarlane H, Quirk A, Dutcher JR. Deep Learning and Infrared Spectroscopy: Representation Learning with a β-Variational Autoencoder. J Phys Chem Lett 2022;13:5787–93.
[222] He C, Zhu S, Wu X, Zhou J, Chen Y, Qian X, et al. Accurate Tumor Subtype Detection with Raman Spectroscopy via Variational Autoencoder and Machine Learning. ACS Omega 2022;7:10458–68.
[223] Grossutti M, D'Amico J, Quintal J, MacFarlane H, Wareham WC, Quirk A, et al. Deep Generative Modeling of Infrared Images Provides Signature of Cracking in Cross-Linked Polyethylene Pipe. ACS Appl Mater Inter 2023;15:22532–42.
[224] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 3431–40.
[225] Kingma D.P., Welling M. Auto-encoding variational bayes. arXiv preprint 2013.
[226] Azimi SM, Britz D, Engstler M, Fritz M, Mücklich F. Advanced steel microstructural classification by deep learning methods. Sci Rep 2018;8:1–14.
[227] Ma B, Ban X, Huang H, Chen Y, Liu W, Zhi Y. Deep learning-based image segmentation for al-la alloy microscopic images. Symmetry (Basel) 2018;10:107.
[228] Madsen J, Liu P, Kling J, Wagner JB, Hansen TW, Winther O, et al. A deep learning approach to identify local structures in atomic-resolution transmission electron microscopy images. Adv Theory Simul 2018;1:1800037.
[229] Furat O, Wang M, Neumann M, Petrich L, Weber M, Krill CE III, et al. Machine learning techniques for the segmentation of tomographic image data of functional materials. Front Mater 2019;6:145.
[230] Roberts G, Haile SY, Sainju R, Edwards DJ, Hutchinson B, Zhu Y. Deep learning for semantic segmentation of defects in advanced STEM images of steels. Sci Rep 2019;9:12744.

[231] Strohmann T, Bugelnig K, Breitbarth E, Wilde F, Steffens T, Germann H, et al. Semantic segmentation of synchrotron tomography of multiphase Al-Si alloys using a convolutional neural network with a pixel-wise weighted loss function. Sci Rep 2019;9:19611.
[232] Tsopanidis S, Moreno RH, Osovski S. Toward quantitative fractography using convolutional neural networks. Eng Fract Mech 2020;231:106992.
[233] Hwang H, Choi SM, Oh J, Bae SM, Lee JH, Ahn JP, et al. Integrated application of semantic segmentation-assisted deep learning to quantitative multi-phased microstructural analysis in composite materials: Case study of cathode composite materials of solid oxide fuel cells. J Power Sources 2020;471:228458.
[234] Evsevleev S, Paciornik S, Bruno G. Advanced deep learning-based 3D microstructural characterization of multiphase metal matrix composites. Adv Eng Mater 2020;22:1901197.
[235] Li X, Zhang Y, Zhao H, Burkhart C, Brinson LC, Chen W. A transfer learning approach for microstructure reconstruction and structure-property predictions. Sci Rep 2018;8:1–13.
[236] Chavez T, Roberts EJ, Zwart PH, Hexemer A. A comparison of deep-learning-based inpainting techniques for experimental X-ray scattering. J Appl Crystallogr 2022;55.
[237] Huang G., Chen D., Li T., Wu F., Van Der Maaten L., Weinberger K.Q. Multi-scale dense networks for resource efficient image classification. arXiv preprint 2017.
[238] Pelkie BG, Pozzo LD. The laboratory of Babel: highlighting community needs for integrated materials data management. Digi Discov 2023;2:544–56.
[239] Jablonka KM, Ai Q, Al-Feghali A, Badhwar S, Bocarsly JD, Bran AM, et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. Digi Discov 2023;2:1233–50.
[240] Bran A.M., Cox S., White A.D., Schwaller P. ChemCrow: Augmenting large-language models with chemistry tools. arXiv preprint 2023.
[241] Otsuka S, Kuwajima I, Hosoya J, Xu Y, Yamazaki M. PoLyInfo: Polymer database for polymeric materials design. In: 2011 International Conference on Emerging Intelligent Data and Web Technologies. IEEE; 2011. p. 22–9.
[242] Kim C, Chandrasekaran A, Huan TD, Das D, Ramprasad R. Polymer genome: a data-powered polymer informatics platform for property predictions. J Phys Chem C 2018;122:17575–85.
[243] Brinson LC, Deagen M, Chen W, McCusker J, McGuinness DL, Schadler LS, et al. Polymer Nanocomposite Data: Curation, Frameworks, Access, and Potential for Discovery and Design. ACS Macro Lett 2020:1086–94.
[244] Isard M, Budiu M, Yu Y, Birrell A, Fetterly D. Dryad: distributed data-parallel programs from sequential building blocks. In: Proceedings of the 2nd ACM SIGOPS/EuroSys European Conference on Computer Systems 2007; 2007. p. 59–72.
[245] Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I. The materials data facility: data services to advance materials science research. Jom 2016;68:2045–52.
[246] Blaiszik B, Ward L, Schwarting M, Gaff J, Chard R, Pike D, et al. A data ecosystem to support machine learning in materials science. MRS Commun 2019;9:1125–33.
[247] Foster I. Globus Online: Accelerating and democratizing science through cloud-based services. IEEE Internet Comput 2011;15:70–3.
[248] Scheidgen M, Himanen L, Ladines AN, Sikter D, Nakhaee M, Fekete Á, et al. NOMAD: A distributed web-based platform for managing materials science research data. J Open Source Software 2023;8:5388.
[249] Foster ED, Deardorff A. Open science framework (OSF). Journal of the Medical Library Association: JMLA 2017;105:203.
[250] Chengzan L, Yanfei H, Jianhui L, Lili Z. ScienceDB: A Public Multidisciplinary Research Data Repository for eScience. In: 2017 IEEE 13th International Conference on e-Science (e-Science). IEEE; 2017. p. 248–55.
[251] Walsh DJ, Zou W, Schneider L, Mello R, Deagen ME, Mysona J, et al. Community Resource for Innovation in Polymer Technology (CRIPT): A Scalable Polymer Material Data Structure. ACS Cent Sci 2023;9:330–8.
[252] Ghiringhelli LM, Baldauf C, Bereau T, Brockhauser S, Carbogno C, Chamanara J, et al. Shared metadata for data-centric materials science. Sci Data 2023;10:626.
[253] Ghiringhelli LM, Carbogno C, Levchenko S, Mohamed F, Huhs G, Lüders M, et al. Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. Npj Comput Mater 2017;3:46.
[254] Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. Hong Kong, China: Association for Computational Linguistics; 2019. p. 3615–20.
[255] Gupta T, Zaki M, Krishnan NA, Mausam. MatSciBERT: A materials domain language model for text mining and information extraction. Npj Comput Mater 2022;8:102.
[256] Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. J Chem Inf Model 2016;56:1894–904.
[257] Beard EJ, Cole JM. ChemSchematicResolver: a toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities. J Chem Inf Model 2020;60:2059–72.
[258] Mavracic J, Court CJ, Isazawa T, Elliott SR, Cole JM. ChemDataExtractor 2.0: Autopopulated ontologies for materials science. J Chem Inf Model 2021;61:4280–9.
[259] Wilary DM, Cole JM. ReactionDataExtractor: A Tool for Automated Extraction of Information from Chemical Reaction Schemes. J Chem Inf Model 2021;61:4962–74.
[260] Guo J, Ibanez-Lopez AS, Gao H, Quach V, Coley CW, Jensen KF, et al. Automated chemical reaction extraction from scientific literature. J Chem Inf Model 2021;62:2035–45.
[261] Vaucher AC, Zipoli F, Geluykens J, Nair VH, Schwaller P, Laino T. Automated extraction of chemical synthesis actions from experimental procedures. Nature Comm 2020;11:3601.
[262] Wang Z, Kononova O, Cruse K, He T, Huo H, Fei Y, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. Sci Data 2022;9:231.
[263] Manning JR, Sarkisov L. Unveiling the synthesis patterns of nanomaterials: a text mining and meta-analysis approach with ZIF-8 as a case study. Digi Discov 2023;2:1783–96.
[264] Walker N, Lee S, Dagdelen J, Cruse K, Gleason S, Dunn A, et al. Extracting structured seed-mediated gold nanorod growth procedures from scientific text with LLMs. Digi Discov 2023;2:1768–82.
[265] Cruse K, Trewartha A, Lee S, Wang Z, Huo H, He T, et al. Text-mined dataset of gold nanoparticle synthesis procedures, morphologies, and size entities. Sci Data 2022;9:234.
[266] Shetty P, Ramprasad R. Automated knowledge extraction from polymer literature using natural language processing. iScience 2021:24.
[267] Shetty P, Rajan AC, Kuenneth C, Gupta S, Panchumarti LP, Holm L, et al. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. Npj Comput Mater 2023;9:52.
[268] Wang Z, Cruse K, Fei Y, Chia A, Zeng Y, Huo H, et al. ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols. Digi Discov 2022;1:313–24.

Shizhao Lu earned her bachelor's degree from Nanjing Tech University in 2017, earned her master's degree from University of Pennsylvania in 2019 and is working towards a PhD degree in chemical & biomolecular engineering at the University of Delaware. Her thesis focuses on molecular modeling and simulations of polymer nanocomposites containing nanorods and development of machine learning workflows for automation and acceleration of structural characterization of polymer materials.

Arthi Jayaraman is a professor of Chemical and Biomolecular Engineering and Materials Science and Engineering at the University of Delaware (UD), Newark. She received her Ph.D. in Chemical Engineering from North Carolina State University and was a postdoctoral researcher at the University of Illinois-Urbana Champaign. Her expertise is in the development and use of computational techniques to study macromolecular materials. She is the recipient of the AIChE COMSEF Impact Award (2021), American Physical Society (APS) Fellowship (2020), American Chemical Society PMSE Young Investigator (2014), AIChE COMSEF Young Investigator Award (2013), and Department of Energy (DOE) Early Career Research Award (2010).