Mini-review

Received: 12 August 2021 Revised: 30 November 2021 Accepted article published: 4 December 2021 Published online in Wiley Online Library: 21 December 2021

(wileyonlinelibrary.com) DOI 10.1002/pi.6345

Machine learning for polymeric materials: an introduction

Morgan M Cencer,[a,b,c] Jeffrey S Moore[a,c] and Rajeev S Assary[b*]

Abstract

Polymers are incredibly versatile materials and have become ubiquitous. Increasingly, researchers are using data science and polymer informatics to design new materials and understand their structure–property relationships. Polymer informatics is an emerging field. While there are many useful tools and databases available, many are not widely utilized. Herein, we introduce the field of polymer informatics and discuss some of the available databases and tools. We cover how to share polymer data, approaches for preparing a dataset for machine learning and recent applications of machine learning to polymer property prediction and polymer synthesis.
© 2021 Society of Industrial Chemistry.

Keywords: machine learning; polymers; informatics; inverse design

INTRODUCTION

Polymers are a critical material class due to their wide availability, range of properties and high tuneability. However, rational design of polymers is challenging due to the variety of aspects that influence their properties and performance.[1] For example, the monomer(s) structure, synthesis method and processing control the chemical structure, morphology and hence properties of the final polymer.[2] Additionally, researchers are increasingly considering sustainability of monomer sourcing, interactions between a polymer and its environment, polymer aging behavior and end of life (whether as waste, or recyclable).[3,4] These relationships are schematically shown in Fig. 1. These considerations – and more – mean any given monomer leads to a variety of properties, and desired properties may be accessed through a variety of monomers. For example, low-density polyethylene and high-density polyethylene have the same monomer but very different mechanical properties, and polyethylene, polypropylene and polyvinyl chloride are all used to make similar plastic bottles. As a result, traditional research methods using trial and error based on chemical intuition are often insufficient to fully design solutions to polymer innovation and discovery.[5] Data-driven and informatics-based approaches are needed to move the field forward faster.[6]

Recent advances in drug development and the successes of the Materials Genome Initiative[7–10] are examples illustrating the benefits of an informatics-based approach.[11] Data-driven research can dramatically accelerate discovery, and lead to improved performance.[12] Understanding which structures lead to specific properties (informatics) offers insights about underlying structure–function relationships.[13] Data-driven approaches also allow inverse design, where a desired property (or properties) is identified, and data are used to determine what structure(s) corresponds to that property.[9,14] Done properly, data-driven research allows researchers to move beyond their own intuition, experience and biases to discover connections that were previously unimagined.

Polymer informatics is a relatively new field, but one with rapidly growing importance. Polymer informatics has been applied to essentially every aspect of the polymer lifecycle. It has been used to design new monomers for various applications[12,15,16]; engineer reactions[17]; model processing conditions and parameters[18–20]; identify and predict polymer conformations and phases[21–26]; predict materials properties[27–35]; and finally offer insight into wear and end of life.[4,36–39] Most polymer informatics literature focuses on property prediction, but recently other aspects of polymer synthesis, processing and lifetime have been gaining attention.[14,17,40] There are still many areas ripe for an informatics approach, such as designing for longer term stability or circular economies.

In this mini-review we discuss necessary tools for polymer informatics. We aim to provide a starting point for the non-specialist to understand the tools and methods that currently exist in this rapidly evolving field. The data and databases section focuses on

* *Correspondence to: RS Assary, Materials Science Division, Argonne National Laboratory, Lemont, IL 60439, USA. E-mail: assary@anl.gov*

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science Laboratory, is operated under Contract no. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan. http://energy.gov/downloads/doe-public-access-plan

a *Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

b *Materials Science Division, Argonne National Laboratory, Lemont, IL, USA*

c *Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL, USA*

10970126, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pi.6345 by Argonne National Laboratory, Wiley Online Library on [11/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

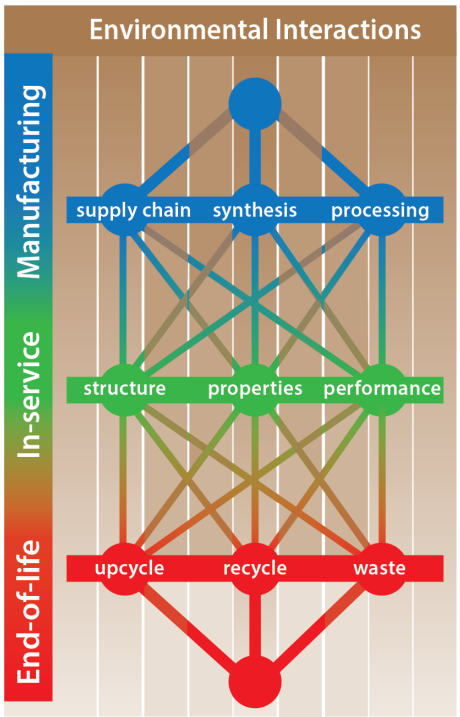

Figure 1. Designing for multigenerational lifecycles requires consideration of all three lifecycle stages (manufacturing, in service and end of life) and all the factors that contribute to each stage. Environmental interactions play a role in all aspects of the material lifecycle.

useful collections of information and specific tools to use for sharing data in the most accessible ways. Often, the best way to understand a dataset is through machine learning (ML) (i.e. regression and classification). To do this ML, we need accurate representations, so the polymer representation and featurization section focuses on popular approaches of developing ML input. This includes fingerprinting techniques for monomers and whole polymers, as well as alternative approaches such as graph-based methods. In the final section, ML approaches, we discuss commonly used methods, along with examples of each approach.

DATA AND DATABASES

Informatics is all about data, and as such, high-quality data are of paramount importance. ML is particularly sensitive to data quality, as it is very sensitive to artifacts,[41–43] and is poor at extrapolation.[44] Therefore, it is important to identify and account for any biases in a dataset, and gather large datasets.[45] Despite a wide array of available materials databases, it is often challenging to find a complete dataset relevant to a specific research question. In contrast to synthetic macromolecules, there are many small-molecule databases with millions of entries (ZINC,[46] ChemSpider,[47] PubChem,[48] ChEMBL,[49,50] and many more), with extensive property data for each entry. The number of high-quality materials databases is growing, but most databases only have hundreds or thousands of entries, representing a much smaller chemical space than the small-molecule databases. Additionally, initiatives to expand and create materials databases[10,51] are divided between inorganic materials and soft materials. For polymers, databases of interest are PolyInfo,[52] the extension of PolyInfo PI1M,[53] and the Khazana[54] databases. We note that to accelerate polymer informatics the community needs a validated and curated database and repository where researchers can deposit new polymer data, similar to the Cambridge Structural Database,[55] but containing property data as well as characterization data. An additional source of polymer data is handbooks. Polymer data handbooks have a broad array of data, and while most would require some effort to make the data accessible to a computer,[56,57] some are fully accessible online.[58]

The imprecision of polymer naming conventions is a hurdle to widespread polymer data sharing. Traditionally, polymer names indicate what repeat units are incorporated, and, if relevant, the relationship between comonomers (e.g. random or block) and tacticity. However, given the stochastic nature of polymers more precise information on structure is omitted. If a database of polymer information is going to be useful to a researcher who did not generate it, the new researcher must be able to understand the precise identity and nature of the entries in the database. This is especially important if multiple sources of data are being combined to develop a sufficiently large database for a specific problem. The polymer informatics community needs to settle on a standard method of detailing polymer structures and data. One approach to developing a standard schema for polymer data sharing is PolyDat.[59] PolyDat is designed to include all relevant data for a polymer, including characterization data, synthesis procedures and information on all relevant species and post-functionalization. An overview of the PolyDat schema is shown in Fig. 2(a). There are three key parts of the PolyDat schema: preamble, species and transformations. The preamble contains all the metadata, reaction network information and data provenance. It gives all the information needed to understand the other sections. The species section contains all the characterization data on all the species in the reaction network. This characterization data can be of any type. The transformations section includes information on all the reactions (both synthesis and any post-synthetic modifications), including the reaction procedures. Use of a standard data schema will greatly increase the ability of researchers to extract published results.

POLYMER REPRESENTATION AND FEATURIZATION

An accurate ML model requires inputs (features) that describe the system of interest. An accurate useful model depends on properly chosen and designed features.[60,61] Features are a wide range of items, from properties of atoms (e.g. partial charge, atomic number) in the molecule of interest, to calculated electronic properties (HOMO, LUMO, etc.), to measured experimental values (e.g. glass transition temperature, heat capacity), to reaction or processing conditions[24,62] (temperature, pH, etc.). The critical requirement for a feature set is that it accurately and uniquely describes each data point in a machine-readable format. Often, the lengthiest stage in a ML project is identifying which features are needed, which are superfluous and what is the best method to generate those features.

Some ML models for the prediction of polymer properties may achieve high accuracy solely using features based on monomers.[63] Monomer-based features range in complexity from constitutional descriptors (number of rings, number of heavy atoms, etc.), to two-dimensional representations (atomic connectivity, topological descriptors, molecular graphs, etc.), to three-dimensional geometric descriptors (computationally generated or crystallography based), to four-dimensional conformational ensemble descriptors.[64] Two common approaches for developing monomer-based features are fingerprinting[62,64,65] and graph-based methods.

© 2021 Society of Industrial Chemistry.

10970126, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pi.6345 by Argonne National Laboratory, Wiley Online Library on [11/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

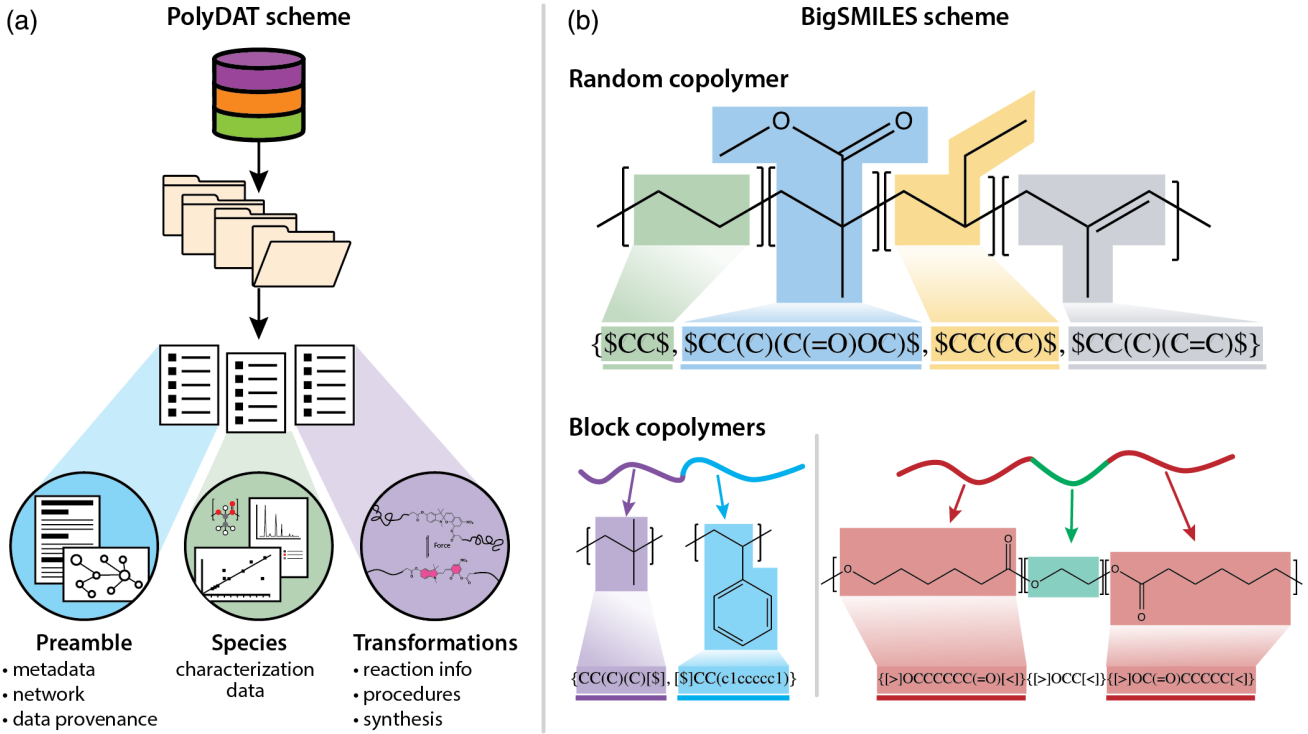

Figure 2. (a) The PolyDAT schema is a data-sharing layout that includes information on the polymer, characterization, synthesis, processing and any other measurements or relevant information. (b) BigSMILES is a text-based description of polymer structure using a variant of SMILES strings.

Fingerprinting involves converting geometric and chemical information to a numerical representation.[66] Most often, the numerical representation is a vector of fixed length, where each component in the vector represents a different characteristic of the monomer. A properly designed fingerprinting technique gives a unique fingerprint for each unique monomer. Fingerprints can be based on purely atomic neighborhoods,[65] or on the molecule as a whole.[62,64] Graph-based methods require a large number of data points and typically use neural nets to predict or classify polymers using a descriptor-free approach. Examples of this approach have been reported recently.[67,68] The selection of appropriate methods depends on the size of the available dataset and the chemical information available about each monomer.

A polymeric fingerprint[54] is appropriate when the behaviors or properties being modeled are dependent on the bulk structure of the polymer. Polymer fingerprints are created with a wide variety of details. The simplest method is to encode the identity of the building block, and the count of each type of building block. Additional complexity is added by including information about the relationships between types of building blocks, clearly identifying the difference between a random copolymer and an alternating or block copolymer.[54] However, a different approach is needed to include atomic and molecular properties. The Ramprasad group has developed a highly successful fingerprinting technique for polymers that includes information about every level of the molecule, from chain-specific values to atomic properties. This method starts with atomic-triple fingerprints,[69] adds molecular descriptors from RDKit,[70] then identifies commonly occurring substructures or blocks and finally adds polymer-chain-specific descriptors such as side-chain length.[71] This multilevel approach to fingerprinting performs well in predicting polymer properties.[54,71,72] However, most fingerprint approaches do not completely capture the stochastic nature of polymers,[73,74] especially for copolymers.[75] These compositions are complex mixtures and mixtures are fundamentally different from pure substances. Simple average values, while easy to measure, may not fully capture the richer complexity in the underlying distributions. For example, molecular weight, comonomer composition and comonomer sequence will differ from one chain to the next. How do the distributions of these structural characteristics correlate to properties?[76] Properly representing the dispersity and sequence variations inherent to polymers is an open question.[77,78]

In small-molecule research, fingerprinting and feature generation often use SMILES[79] (Simplified Molecular Input Line Entry System) strings as input. The SMILES notation system is widely used for small molecules as it is machine-readable and well suited for informatics purposes. However, the stochastic nature of polymers and their size make using SMILES for polymers inefficient and awkward. BigSMILES[74] adds the ability to define repeat units, copolymers and polymer structures (such as branched, star, etc.) clearly and easily to the SMILES system. Figure 2(b) shows a schematic representation of a few of the ways BigSMILES represents polymers. While there are other ongoing efforts to improve SMILES (i.e. SELFIES,[80] a self-referencing approach that is more robust than traditional SMILES), the BigSMILES approach is sufficiently flexible to still be one of the clearest and easiest methods of providing a polymer structural definition. Wide adoption of BigSMILES notation, especially within databases, will aid in making data fully accessible to all researchers.

ML APPROACHES

Two types of ML commonly used in polymer informatics are supervised and unsupervised learning (Fig. 3).[44,81] Supervised learning uses data where the label is known. For example, performing a

© 2021 Society of Industrial Chemistry.

10970126, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pi.6345 by Argonne National Laboratory, Wiley Online Library on [11/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

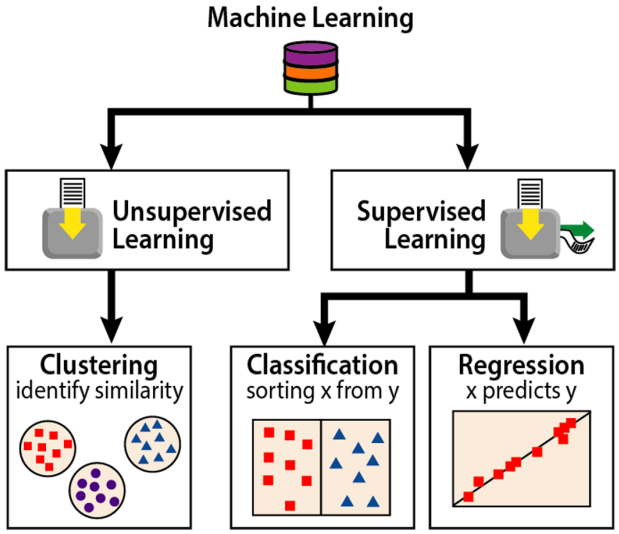

Figure 3. Unsupervised learning groups and interprets data based only on the input (i.e. *x* values). Supervised learning develops predictive models using both input and output (i.e. *x* and *y* values).

regression fit for predicting glass transition temperature, where all the training data have a known glass transition temperature.[82] Unsupervised learning uses unlabeled data. It is most commonly used to identify clusters or groups within data, such as autoidentifying nanocluster shapes in a molecular dynamics dataset.[83] Unsupervised learning can also be used for autoencoding. Autoencoders are a deep learning technique that independently learns how to encode or represent the training data.[80,84] Supervised and unsupervised learning are combined in semi-supervised and active learning. These techniques use a small subset of labeled data to assign labels[85] or predict outcomes for a larger, unlabeled, dataset.[86] Semi-supervised learning is very effective for labeling clusters and classification. Active learning iteratively identifies the unlabeled data that will most improve the model if the label was added, guiding experimental data collection. There are a wide variety of tools and packages available for polymer informatics. Some of the most commonly used tools include RDKit,[87,88] Pybel,[89] SciKit Learn[44] and Pymatgen.[90]

Supervised learning is the most common type of ML used in polymer informatics. Structure-to-property predictions are usually generated using supervised learning. The general process for this approach is to gather data (either from experimental/simulation data or from a database), verify that all data are comparable and have accurate labels, build a ML model to predict the property, before finally using the model on a new data point to predict the outcome. There are many examples in the literature using this approach. Supervised learning with a deep neural network has been used to predict solvents and non-solvents for a polymer[91] and polymer phase transitions.[25] Regression models can predict refractive indices of linear polymers[92] and erosion behavior of silicon carbide-reinforced polymer composites.[4] Supervised learning is most effective and useful when there is a large accurate dataset available for training, and directly measuring the predicted property is an intensive process. Ideally, all models of this type would be shared in a format that makes them useable to researchers without having to recreate the training process. One excellent example of sharing predictive models widely is the Polymer Genome,[54,72,93] which predicts a large number of properties from either the polymer name, the SMILES string for the repeat unit, or a drawing of the repeat unit.

Unsupervised learning has been used very effectively to solve inverse design problems in materials or polymer science.[15,22,25,63,94–96] In an inverse design problem, the desired properties are known, but the suitable molecule/polymer to achieve those properties is unknown.[77,97–99] In these problems a combination of autoencoders (unsupervised learning) and supervised learning often delivers accurate predictions. This approach is especially useful in situations where multiple properties must be optimized. Autoencoders in tandem with regression models have been used to predict polymers that are robust under high temperatures and high electric fields,[15] find polymers suited for solar cells[61,63] and predict polymer phases and phase transitions.[25,26] Unsupervised learning has also been applied to identify defects[94] and conformation states.[22,95] In these applications, self-organizing mapping[96] and clustering are used to identify subsets of data and determine which characteristics separate subclasses.

For small datasets, semi-supervised or active learning combines supervised and unsupervised learning to leverage a small starting dataset for large learning gains. While there are only a few examples in the literature of semi-supervised learning,[85,100] it is likely to grow in popularity. Active learning is relatively new; it is based around iterative data acquisition guided by Bayesian optimization.[101] Active learning is especially notable in how it utilizes a very small starting dataset (initial data can be as small as 10 samples), and guides data acquisition to obtain a desired outcome much faster than random sampling.[102] Active learning has been applied to discover redoxmers with a specific desired reduction potential,[86] high-glass-transition polymers,[82,103] ring polymer molecular dynamics[104] and epoxy adhesive strength,[105] among others. Active learning will become increasingly important and valuable, especially as high-throughput and robotic synthesis approaches are developed.

CONCLUSIONS

Moving forward, polymer informatics will be central to the genesis of new materials. As we design materials to solve increasingly difficult problems, we need data-driven design to make the most use of available knowledge. One of the greatest challenges in polymer design is developing polymers that need multiple properties optimized. Multi-property design (Fig. 1), especially when one of the properties is degradation behavior or recyclability, is increasingly necessary, and very difficult to do well, as maximizing one property often requires tradeoffs in other properties. Additional key challenges for polymer informatics include the need for polymer representations that capture stochasticity, larger datasets and more research into retrosynthetic design approaches.[77,78,99] An informatics-driven approach allows quantification of tradeoffs and expands the pool of possible materials, working from an inverse design approach. Whether it is designing a polymer that includes triggered deconstruction, one that responds to changing conditions, or is suitable for an extreme environment, data-driven approaches can shorten design cycles and open new avenues of research.

ACKNOWLEDGEMENTS

We acknowledge UChicago/Argonne, CDAC funding via AI for Electrochemistry program. The authors thank Dorothy Loudermilk for assistance in making figures.

© 2021 Society of Industrial Chemistry.

10970126, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pi.6345 by Argonne National Laboratory, Wiley Online Library on [11/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

GLOSSARY

A **curated** collection is one that is carefully managed and presented.

Features are the input for a machine learning model. **Featurization** is the process of generating features.

Inverse design is a process of determining the desired end properties then identifying the molecular structure needed to produce those properties.

Neural nets consist of densely linked processing nodes, modeled loosely on the neurons in a brain.

A **stochastic** feature is one that is best described by a random variable; for example, the distribution of each monomer in a copolymer.

Validated data have been checked and confirmed by a researcher uninvolved in generating the data.

REFERENCES

1 Audus DJ and De Pablo JJ, *ACS Macro Lett* **6**:1078–1082 (2017).
2 Lodge TP and Hiemenz PC, *Polymer Chemistry*. CRC Press, Boca Raton, Florida (2007).
3 Scaffaro R, Maio A, Sutera F, Gulino E and ortunato & Morreale, M., *Polymers (Basel)* **11**:651 (2019).
4 Kharb SS, Antil P, Singh S, Antil SK, Sihag P and Kumar A, *Silicon* **13**: 1113–1119 (2020). https://doi.org/10.1007/s12633-020-00497-z.
5 Zhou T, Song Z and Sundmacher K, *Engineering* **5**:1017–1026 (2019).
6 Mannodi-Kanakkithodi A, Chandrasekaran A, Kim C, Huan TD, Pilania G, Botu V *et al*., *Mater Today* **21**:785–796 (2018).
7 Khaira G, Doxastakis M, Bowen A, Ren J, Suh HS, Segal-Peretz T *et al*., *Macromolecules* **50**:7783–7793 (2017).
8 Arora A, Qin J, Morse DC, Delaney KT, Fredrickson GH, Bates FS *et al*., *Macromolecules* **49**:4675–4690 (2016).
9 Mulholland GJ and Paradiso SP, *APL Mater* **4**:053207 (2016).
10 de Pablo JJ, Jackson NE, Webb MA, Chen LQ, Moore JE, Morgan D *et al*., *NPJ Comput Mater* **5**:1–23 (2019).
11 Tripathi N, Goshisht MK, Sahu SK and Arora C, *Mol Divers* **25**:1643–1664 (2021). https://doi.org/10.1007/s11030-021-10237-z.
12 Chen G, Shen Z, Iyer A, Ghumman UF, Tang S, Bi J *et al*., *Polymers (Basel)* **12**:163 (2020).
13 Rickman JM, Lookman T and Kalinin SV, *Acta Mater* **168**:473–510 (2019).
14 Hong S, Liow CH, Yuk JM, Byon HR, Yang Y, Cho EA *et al*., *ACS Nano* **15**: 3971–3995 (2021). https://doi.org/10.1021/acsnano.1c00211.
15 Batra R, Dai H, Huan TD, Chen L, Kim C, Gutekunst WR *et al*., *Chem Mater* **32**:10489–10500 (2020).
16 Mannodi-Kanakkithodi A, Pilania G, Ramprasad R, Lookman T and Gubernatis JE, *Comput Mater Sci* **125**:92–99 (2016).
17 Lazzari S, Lischewski A, Orlov Y, Deglmann P, Daiss A, Schreiner E *et al*., *Adv Chem Eng* **56**:187–227 (2020).
18 Ibañez R, Casteran F, Argerich C, Ghnatios C, Hascoet N, Ammar A *et al*., *Fluids* **5**:1–23 (2020).
19 Abuomar O, Nouranian S, King R and Lacy TE, *Comput Mater Sci* **158**: 98–109 (2019).
20 Le TT, *J Compos Mater* **55**:787–811 (2021).
21 Tu KH, Huang H, Lee S, Lee W, Sun Z, Alexander-Katz A *et al*., *Adv Mater* **32**:1–8 (2020).
22 Sun LW, Li H, Zhang XQ, Gao HB and Luo MB, *Chinese J Polym Sci (Engl Ed)* **38**:1403–1408 (2020).
23 Venkatram S, Batra R, Chen L, Kim C, Shelton M and Ramprasad R, *J Phys Chem B* **124**:6046–6054 (2020).
24 Patra A, Batra R, Chandrasekaran A, Kim C, Huan TD and Ramprasad R, *Comput Mater Sci* **172**:109286 (2020).
25 Bhattacharya D and Patra TK, *Macromolecules* **54**:3065–3074 (2021).
26 Hiraide K, Hirayama K, Endo K and Muramatsu M, *Comput Mater Sci* **190**:110278 (2021).
27 Daghigh V, Lacy TE Jr, Daghigh H, Gu G, Baghaei KT, Horstemeyer MF *et al*., *J Reinf Plast Compos* **39**:587–598 (2020).
28 Massari L, Schena E, Massaroni C, Saccomandi P, Menciassi A, Sinibaldi E *et al*., *Soft Robot* **7**:409–420 (2020).
29 Zhang Y and Xu X, *Heliyon* **6**:e05055 (2020).
30 Gupta P, Schadler LS and Sundararaman R, *Mater Charact* **173**:110909 (2021).
31 Mikulskis P, Hook A, Dundas AA, Irvine D, Sanni O, Anderson D *et al*., *ACS Appl Mater Interfaces* **10**:139–149 (2018).
32 Rahman A, Deshpande P, Radue MS, Odegard GM, Gowtham S, Ghosh S *et al*., *Compos Sci Technol* **207**:108627 (2021).
33 Pilania G, Iverson CN, Lookman T and Marrone BL, *J Chem Inf Model* **59**:5013–5025 (2019).
34 Roch LM, Saikin SK, Häse F, Friederich P, Goldsmith RH, León S *et al*., *ACS Nano* **14**:6589–6598 (2020).
35 Epa VC, Yang J, Mei Y, Hook AL, Langer R, Anderson DG *et al*., *J Mater Chem* **22**:20902–20906 (2012).
36 Kojima T, Washio T, Hara S and Koishi M, *Sci Rep* **10**:1–11 (2020).
37 Yang J, Kang G, Liu Y, Chen K and Kan Q, *Int J Fatigue* **136**:105619 (2020).
38 Zhou X, Hsieh SJ, Peng B and Hsieh D, *Microelectron Reliab* **79**:48–58 (2017).
39 Prajna MR, Antony PJ and Jnanesh NA, *J Phys Conf Ser* **1142**:012007 (2018).
40 Peerless JS, Milliken NJB, Oweida TJ, Manning MD and Yingling YG, *Adv Theory Simul* **2**:1–12 (2019).
41 Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S *et al*., *Nat Mach Intell* **3**:199–217 (2021).
42 Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W and Müller KR, *Nat Commun* **10**:1–8 (2019).
43 Chuang KV and Keiser MJ, *Science* **362**:1–3 (2018).
44 Geron A, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Sebastopol, California (2019).
45 Halevy A, Norvig P and Pereira F, *IEEE Intell Syst* **24**:8–12 (2009).
46 Sterling T and Irwin JJ, *J Chem Inf Model* **55**:2324–2337 (2015).
47 ChemSpider. Royal Society of Chemistry. Available: chemspider.com.
48 Kim S, Chen J, Cheng T, Gindulyte A, He J, He S *et al*., *Nucleic Acids Res* **49**:D1388–D1395 (2021).
49 Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F *et al*., *Nucleic Acids Res* **43**:W612–W620 (2015).
50 Mendez D, Gaulton A, Bento AP, Chambers J, de Veij M, Félix E *et al*., *Nucleic Acids Res* **47**:D930–D940 (2019).
51 Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S *et al*., *APL Mater* **1**:011002 (2013).
52 Otsuka S, Kuwajima I, Hosoya J, Xu Y and Yamazaki M, *Proc 2011 Int Conf Emerg Intell Data Web Technol EIDWT* **2011**:22–29 (2011). https://doi.org/10.1109/EIDWT.2011.13.
53 Ma R and Luo T, *J Chem Inf Model* **60**:4684–4690 (2020).
54 Huan TD, Mannodi-Kanakkithodi A, Kim C, Sharma V, Pilania G and Ramprasad R, *Sci Data* **3**:1–10 (2016).
55 Groom CR, Bruno IJ, Lightfoot MP and Ward SC, *Acta Crystallogr* **B72**: 171–179 (2016).
56 Cheremisinoff NP, *Handbook of Polymer Science and Technology*. Taylor & Francis, Milton Park, United Kingdom (2019).
57 Brandrup J, Immergut EH and Grulke EA, *Polymer Handbook*. Wiley, New Jersey (1999).
58 *Polymer Data Handbook*. Oxford University Press, Oxford, United Kingdom (2009).
59 Lin TS, Rebello NJ, Beech HK, Wang Z, El-Zaatari B, Lundberg DJ *et al*., *J Chem Inf Model* **61**:1150–1163 (2021). https://doi.org/10.1021/acs.jcim.1c00028.
60 Zhao ZW, Del Cueto M, Geng Y and Troisi A, *Chem Mater* **32**:7777–7787 (2020).
61 Sun W, Zheng Y, Yang K, Zhang Q, Shah AA, Wu Z *et al*., *Sci Adv* **5**:1–8 (2019).
62 David L, Thakkar A, Mercado R and Engkvist O, *J Chem* **12**:1–22 (2020).
63 Jørgensen PB, Mesta M, Shil S, García Lastra JM, Jacobsen KW, Thygesen KS *et al*., *J Chem Phys* **148**:241735 (2018).
64 Gallegos LC, Luchini G, St. John PC, Kim S and Paton RS, *Acc Chem Res* **54**:827–836 (2021).
65 Batra R, Tran HD, Kim C, Chapman J, Chen L, Chandrasekaran A *et al*., *J Phys Chem C* **123**:15859–15866 (2019). https://doi.org/10.1021/acs.jpcc.9b03925.
66 Pattanaik L and Coley CW, *Chem* **6**:1204–1207 (2020).
67 Mercado R, Rastemo T, Lindelöf E, Klambauer G, Engkvist O, Chen H *et al*., *Mach Learn Sci Technol* **2**:025023 (2021).
68 Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D *et al*., *ACS Cent Sci* **4**:268–276 (2018).

© 2021 Society of Industrial Chemistry.

10970126, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pi.6345 by Argonne National Laboratory, Wiley Online Library on [11/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

69 Mannodi-Kanakkithodi A and Ramprasad R, *Comput Mater Discov*: 293–319 (2018). https://doi.org/10.1039/9781788010122-00293.
70 Landrum, G. *RDKit: Open-Source Chemiformatics*.
71 Doan Tran H, Kim C, Chen L, Chandrasekaran A, Batra R, Venkatram S *et al*., *J Appl Phys* **128**:171104 (2020).
72 Kim C, Chandrasekaran A, Huan TD, Das D and Ramprasad R, *J Phys Chem C* **122**:17575–17585 (2018).
73 Wu K, Sukumar N, Lanzillo NA, Wang C, "Rampi" Ramprasad R, Ma R *et al*., *J Polym Sci B Polym Phys* **54**:2082–2091 (2016).
74 Lin TS, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z *et al*., *ACS Cent Sci* **5**:1523–1531 (2019).
75 Webb MA, Jackson NE, Gil PS and de Pablo JJ, *Sci Adv* **6**:1–10 (2020).
76 Sifri RJ, Padilla-Vélez O, Coates GW and Fors BP, *J Am Chem Soc* **142**: 1443–1448 (2020).
77 Sattari K, Xie Y and Lin J, *Soft Matter* **17**:7607–7622 (2021).
78 Chen L, Pilania G, Batra R, Huan TD, Kim C, Kuenneth C *et al*., *Mater Sci Eng R Rep* **144**:100595 (2021).
79 Weininger D, *J Chem Inf Comput Sci* **28**:31–26 (1988).
80 Krenn M, Häse F, Nigam A, Friederich P and Aspuru-Guzik A, *Mach Learn Sci Technol* **1**:045024 (2020).
81 Muller AC and Guido S, *Introduction to Machine Learning with Python*. O'Reilly Media, Sebastopol, California (2016).
82 Kim C, Chandrasekaran A, Jha A and Ramprasad R, *MRS Commun* **9**: 860–866 (2019).
83 Zhu MX, Song HG, Yu QC, Chen JM and Zhang HY, *Int J Heat Mass Transf* **162**:120381 (2020).
84 Wetzel SJ, *Phys Rev E* **96**:1–11 (2017).
85 Ma W, Cheng F, Xu Y, Wen Q and Liu Y, *Adv Mater* **31**:1–9 (2019).
86 Doan HA, Agarwal G, Qian H, Counihan MJ, Rodríguez-López J, Moore JS *et al*., *Chem Mater* **32**:6338–6346 (2020).
87 Lovrić M, Molero JM and Kern R, *Mol Inform* **38**:4–7 (2019).
88 Landrum, G. *RDKit Documentation*. (2011). https://www.rdkit.org/docs/index.html#
89 O'Boyle NM, Morley C and Hutchison GR, *Chem Cent J* **2**:1–7 (2008).
90 Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S *et al*., *Comput Mater Sci* **68**:314–319 (2013).
91 Chandrasekaran A, Kim C, Venkatram S and Ramprasad R, *Macromolecules* **53**:4764–4769 (2020).
92 Minami T and Okuno Y, *MRS Adv* **3**:2975–2980 (2018).
93 Chandrasekaran A, Kim C and Ramprasad R, *Lect Notes Phys* **968**: 397–412 (2020).
94 Gasparotto P, Bochicchio D, Ceriotti M and Pavan GM, *J Phys Chem B* **124**:589–599 (2020).
95 Chen Z, Li D, Wan H, Liu M and Liu J, *Mol Simul* **46**:1509–1521 (2020). https://doi.org/10.1080/08927022.2020.1851028.
96 Huang Y, Zhang J, Jiang ES, Oya Y, Saeki A, Kikugawa G *et al*., *J Phys Chem C* **124**:12871–12882 (2020).
97 Jadrich RB, Lindquist BA and Truskett TM, *J Chem Phys* **146**:184103 (2017).
98 Patra TK, Loeffler TD and Sankaranarayanan SKRS, *Nanoscale* **12**: 23653–23662 (2020).
99 Park NH, Zubarev DY, Hedrick JL, Kiyek V, Corbet C and Lottier S, *Macromolecules* **53**:10847–10854 (2020).
100 Sivaraman G, Jackson NE, Sanchez-Lengeling B, Vázquez-Mayagoitia Á, Aspuru-Guzik A, Vishwanath V *et al*., *Mach Learn Sci Technol* **1**:025015 (2020).
101 Aggarwal CC, Kong X, Gu Q, Han J and Yu PS, *Active learning: a survey, in Data Classification*, ed. by Aggarwal CC. CRC Press, Landrum, G. RDKit Documentation. (2011).
102 Lookman T, Balachandran PV, Xue D and Yuan R, *NPJ Comput Mater* **5**: 21 (2019).
103 Jha A, Chandrasekaran A, Kim C and Ramprasad R, *Model Simul Mater Sci Eng* **27**:024002 (2019).
104 Novikov IS, Shapeev AV and Suleimanov YV, *J Chem Phys* **151**:224105 (2019).
105 Pruksawan S, Lambard G, Samitsu S, Sodeyama K and Naito M, *Sci Technol Adv Mater* **20**:1010–1021 (2019).

 © 2021 Society of Industrial Chemistry.

10970126, 2022, 5, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pi.6345 by Argonne National Laboratory, Wiley Online Library on [11/01/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License