Machine Learning for Polymeric Materials: An Introduction

Morgan M. Cencer†‡*, Jeffrey S. Moore†*, Rajeev S. Assary‡

†Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, United States
‡Materials Science Division, Argonne National Laboratory, Lemont, Illinois, 60439, United States
*Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, United States

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 to prepare a data set for machine learning, and recent applications of machine learning to polymer property prediction and polymer synthesis.

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 the polymer with its environment, polymer aging behavior, and end of life (whether as waste, or recyclable).[3,4] These relationships are schematically shown in Figure 1. These considerations – and more - mean any given monomer leads to a variety of properties, and desired properties may be

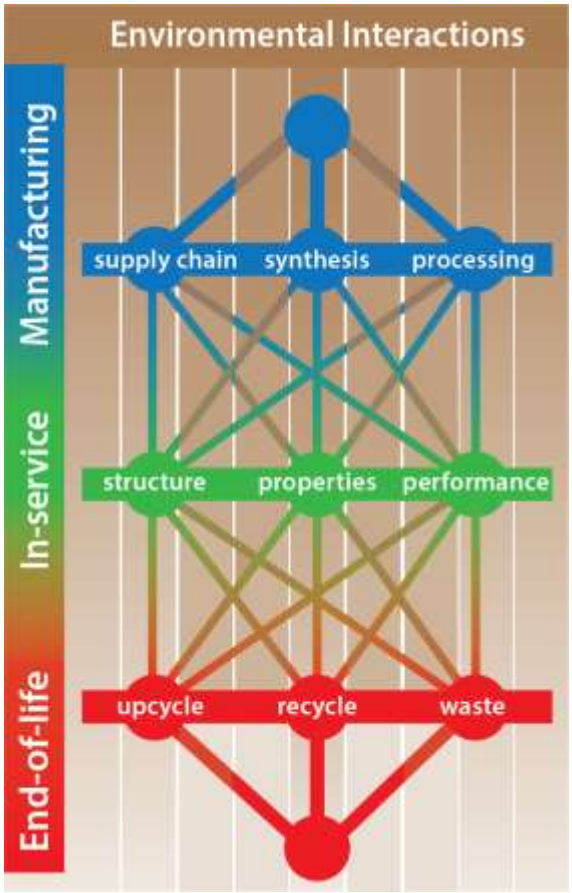

Figure 1: Designing for multigenerational lifecycles requires consideration of all three lifecycle stages (manufacturing, in-service, and end-of-life) and all the factors that contribute to each stage. Environmental interactions play a role in all aspects of the materials lifecycle.

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

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 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 is 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 have 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 are 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 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 (i.e., regression and classification). To do this machine learning (ML) we need accurate representations, so the polymer representation and featurization section focuses on popular approaches of developing machine learning input. This includes fingerprinting techniques for monomers and whole polymers, as well as alternative approaches such as graph-based methods. In the final section, machine learning 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 is of paramount importance. Machine learning 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 the data set, and gather large data sets.[45] Despite a wide array of available materials databases, it is often challenging to find a complete data set 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 computer accessible,[56,57] some are fully accessible online.[58]

The imprecision of polymer naming conventions is a hurdle to wide-spread polymer data sharing. Traditionally, polymer names indicate what repeat unit(s) are incorporated, and, if relevant, the relationship between comonomers (i.e., random or block) and tacticity. However, given the stochastic

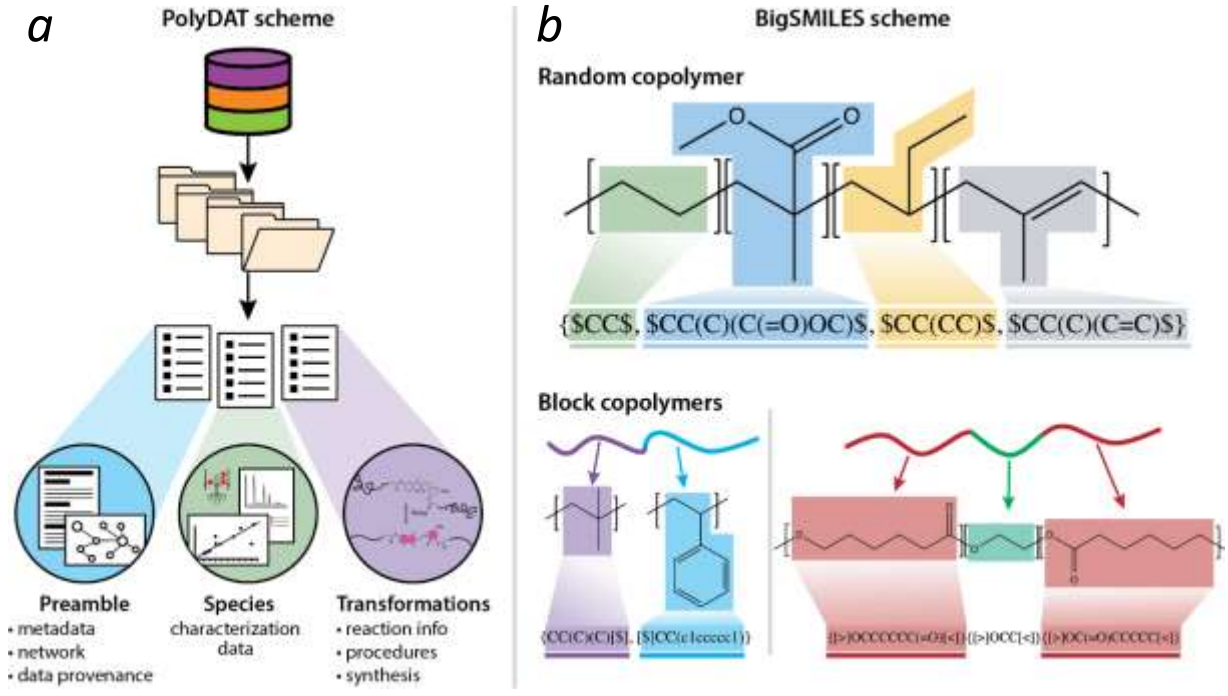

Figure 2: a) The PolyDAT schema is a data sharing layout that includes information on the polymer, characterization, synthesis, processing, and any other measurements or relevant information. b) BigSMILES is a text-based description of polymer structure using a variant of SMILES strings.

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 Figure 2a. 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 need 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 machine learning model requires inputs (features) that describes 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] (temp, pH, etc.). The critical requirement for a feature set is that it accurately and uniquely describe each data point in a machine-readable format. *Often, the lengthiest stage in a machine learning 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 2D representations (atomic connectivity, topological descriptors, molecular graphs, etc.), to 3D geometric descriptors (computationally generated or crystallography based), to 4D conformational ensemble descriptors.[64] Two common approaches to developing monomer-based features are fingerprint[62,64,65] and graph-based methods. Fingerprinting is converting the 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 large number of datapoints and typically use neural nets to predict or classify polymers using a descriptor-free approach. Examples of this approach are reported recently.[67,68] The selection of appropriate methods depends on the size of the available data set and the chemical information available about each monomer.

A polymeric fingerprint[54] is appropriate when the behavior or properties being modeled is 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 multi-level 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 than 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 corelate 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 2b 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.

Machine Learning (ML) Approaches

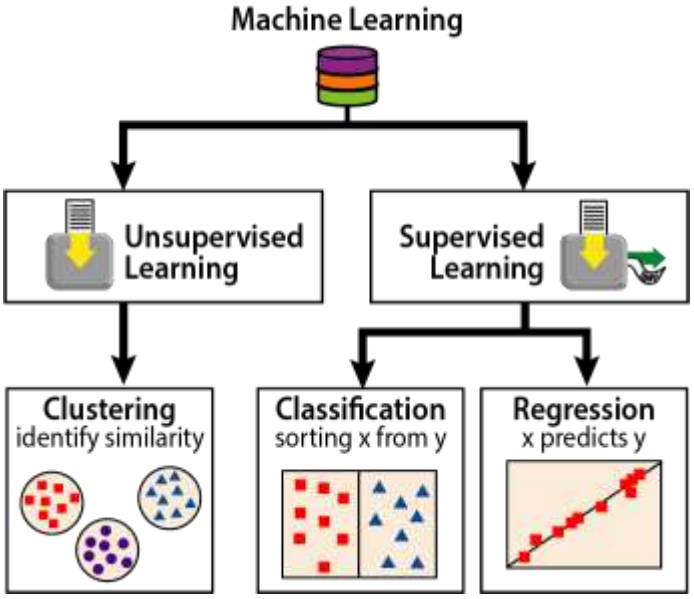

Figure 3: Unsupervised learning groups and interprets data based only on the input (i.e., x values). Supervised learning develops predictive models using both input and output (i.e., x and y values).

Two types of ML commonly used in polymer informatics are supervised and unsupervised learning (Figure 3).[44,81] Supervised learning uses data where the label is known. For example, performing a regression fit for predicting glass transition temperature, where all the training data has 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 auto-identifying nano-cluster shapes in a molecular dynamics dataset.[83] Unsupervised learning can also be used for autoencoding. Autoencoders are a deep learning technique that independently learns the 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 is comparable and has accurate labels. Build an 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 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 from 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 deliver 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 sub-classes.

For small datasets, semi-supervised or active learning combine supervised and unsupervised learning to leverage a small starting dataset for large learning gains. While there are only a few examples in 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 data set (initial data can be as small as ten 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.

Conclusion

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 (Figure 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 data sets, and more research into retrosynthetic design approaches.[76,77,98] 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 would like to thank Dorothy Loudermilk for assistance in making figures.

References

1. Audus, D. J. & De Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **6**, 1078–1082 (2017).

2. Lodge, T. P. & Hiemenz, P. C. *Polymer Chemistry*. (CRC Press, 2007).

3. Scaffaro, R., Maio, A., Sutera, F., Gulino, E. ortunato & Morreale, M. Degradation and recycling of films based on biodegradable polymers: A short review. *Polymers (Basel).* **11**, (2019).

4. Kharb, S. S. *et al.* Machine Learning-Based Erosion Behavior of Silicon Carbide Reinforced Polymer Composites. *Silicon* 1113–1119 (2020) doi:10.1007/s12633-020-00497-z.

5. Zhou, T., Song, Z. & Sundmacher, K. Big Data Creates New Opportunities for Materials Research: A Review on Methods and Applications of Machine Learning for Materials Design. *Engineering* **5**, 1017–1026 (2019).

6. Mannodi-Kanakkithodi, A. *et al.* Scoping the polymer genome: A roadmap for rational polymer dielectrics design and beyond. *Mater. Today* **21**, 785–796 (2018).

7. Khaira, G. *et al.* Derivation of Multiple Covarying Material and Process Parameters Using Physics-Based Modeling of X-ray Data. *Macromolecules* **50**, 7783–7793 (2017).

8. Arora, A. *et al.* Broadly Accessible Self-Consistent Field Theory for Block Polymer Materials Discovery. *Macromolecules* **49**, 4675–4690 (2016).

9. Mulholland, G. J. & Paradiso, S. P. Perspective: Materials informatics across the product lifecycle: Selection, manufacturing, and certification. *APL Mater.* **4**, (2016).

10. de Pablo, J. J. *et al.* New frontiers for the materials genome initiative. *npj Comput. Mater.* **5**, 1–23 (2019).

11. Tripathi, N., Goshisht, M. K., Sahu, S. K. & Arora, C. Applications of artificial intelligence to drug design and discovery in the big data era: a comprehensive review. *Mol. Divers.* (2021) doi:10.1007/s11030-021-10237-z.

12. Chen, G. *et al.* Machine-learning-assisted de novo design of organic molecules and polymers: Opportunities and challenges. *Polymers (Basel).* **12**, (2020).

13. Rickman, J. M., Lookman, T. & Kalinin, S. V. Materials informatics: From the atomic-level to the continuum. *Acta Mater.* **168**, 473–510 (2019).

14. Hong, S. *et al.* Reducing Time to Discovery: Materials and Molecular Modeling, Imaging, Informatics, and Integration. *ACS Nano* (2021) doi:10.1021/acsnano.1c00211.

15. Batra, R. *et al.* Polymers for Extreme Conditions Designed Using Syntax-Directed Variational Autoencoders. *Chem. Mater.* **32**, 10489–10500 (2020).

16. Mannodi-Kanakkithodi, A., Pilania, G., Ramprasad, R., Lookman, T. & Gubernatis, J. E. Multi-objective optimization techniques to design the Pareto front of organic dielectric polymers. *Comput. Mater. Sci.* **125**, 92–99 (2016).

17. Lazzari, S. *et al. Toward a digital polymer reaction engineering*. *Advances in Chemical Engineering* vol. 56 (Elsevier Inc., 2020).

18. Ibañez, R. *et al.* On the data-driven modeling of reactive extrusion. *Fluids* **5**, 1–23 (2020).

19. Abuomar, O., Nouranian, S., King, R. & Lacy, T. E. Application of materials informatics to vapor-grown carbon nanofiber/vinyl ester nanocomposites through self-organizing maps and clustering techniques. *Comput. Mater. Sci.* **158**, 98–109 (2019).

20. Le, T. T. Prediction of tensile strength of polymer carbon nanotube composites using practical

machine learning method. *J. Compos. Mater.* **55**, 787–811 (2021).

21. Tu, K. H. *et al.* Machine Learning Predictions of Block Copolymer Self-Assembly. *Adv. Mater.* **32**, 1–8 (2020).

22. Sun, L. W., Li, H., Zhang, X. Q., Gao, H. B. & Luo, M. B. Identifying Conformation States of Polymer through Unsupervised Machine Learning. *Chinese J. Polym. Sci. (English Ed.* **38**, 1403–1408 (2020).

23. Venkatram, S. *et al.* Predicting Crystallization Tendency of Polymers Using Multifidelity Information Fusion and Machine Learning. *J. Phys. Chem. B* **124**, 6046–6054 (2020).

24. Patra, A. *et al.* A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Comput. Mater. Sci.* **172**, 109286 (2020).

25. Bhattacharya, D. & Patra, T. K. dPOLY: Deep learning of polymer phases and phase transition. *Macromolecules* **54**, 3065–3074 (2021).

26. Hiraide, K., Hirayama, K., Endo, K. & Muramatsu, M. Application of deep learning to inverse design of phase separation structure in polymer alloy. *Comput. Mater. Sci.* **190**, 110278 (2021).

27. Daghigh, V. *et al.* Machine learning predictions on fracture toughness of multiscale bio-nano-composites. *J. Reinf. Plast. Compos.* **39**, 587–598 (2020).

28. Massari, L. *et al.* A Machine-Learning-Based Approach to Solve Both Contact Location and Force in Soft Material Tactile Sensors. *Soft Robot.* **7**, 409–420 (2020).

29. Zhang, Y. & Xu, X. Machine learning glass transition temperature of polymers. *Heliyon* **6**, e05055 (2020).

30. Gupta, P., Schadler, L. S. & Sundararaman, R. Dielectric properties of polymer nanocomposite interphases from electrostatic force microscopy using machine learning. *Mater. Charact.* **173**, 110909 (2021).

31. Mikulskis, P. *et al.* Prediction of Broad-Spectrum Pathogen Attachment to Coating Materials for Biomedical Devices. *ACS Appl. Mater. Interfaces* **10**, 139–149 (2018).

32. Rahman, A. *et al.* A machine learning framework for predicting the shear strength of carbon nanotube-polymer interfaces based on molecular dynamics simulation data. *Compos. Sci. Technol.* **207**, 108627 (2021).

33. Pilania, G., Iverson, C. N., Lookman, T. & Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **59**, 5013–5025 (2019).

34. Roch, L. M. *et al.* From Absorption Spectra to Charge Transfer in Nanoaggregates of Oligomers with Machine Learning. *ACS Nano* **14**, 6589–6598 (2020).

35. Epa, V. C. *et al.* Modelling human embryoid body cell adhesion to a combinatorial library of polymer surfaces. *J. Mater. Chem.* **22**, 20902–20906 (2012).

36. Kojima, T., Washio, T., Hara, S. & Koishi, M. Synthesis of computer simulation and machine learning for achieving the best material properties of filled rubber. *Sci. Rep.* **10**, 1–11 (2020).

37. Yang, J., Kang, G., Liu, Y., Chen, K. & Kan, Q. Life prediction for rate-dependent low-cycle fatigue

of PA6 polymer considering ratchetting: Semi-empirical model and neural network based approach. *Int. J. Fatigue* **136**, 105619 (2020).

38. Zhou, X., Hsieh, S. J., Peng, B. & Hsieh, D. Cycle life estimation of lithium-ion polymer batteries using artificial neural network and support vector machine with time-resolved thermography. *Microelectron. Reliab.* **79**, 48–58 (2017).

39. Prajna, M. R., Antony, P. J. & Jnanesh, N. A. Machine learning approach for flexural characterization of smart material. *J. Phys. Conf. Ser.* **1142**, (2018).

40. Peerless, J. S., Milliken, N. J. B., Oweida, T. J., Manning, M. D. & Yingling, Y. G. Soft matter informatics: Current progress and challenges. *Adv. Theory Simulations* **2**, 1–12 (2019).

41. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).

42. Lapuschkin, S. *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1–8 (2019).

43. Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science (80-.).* **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, 2019).

45. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**, 8–12 (2009).

46. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

47. ChemSpider. *Royal Society of Chemistry* chemspider.com.

48. Kim, S. *et al.* PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).

49. Davies, M. *et al.* ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **43**, W612–W620 (2015).

50. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

51. Jain, A. *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, (2013).

52. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. PoLyInfo: Polymer database for polymeric materials design. *Proc. - 2011 Int. Conf. Emerg. Intell. Data Web Technol. EIDWT 2011* 22–29 (2011) doi:10.1109/EIDWT.2011.13.

53. Ma, R. & Luo, T. PI1M: A benchmark database for polymer informatics. *J. Chem. Inf. Model.* **60**, 4684–4690 (2020).

54. Huan, T. D. *et al.* A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 1–10 (2016).

55. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Cryst.* **B72**, 171–179 (2016).

56. Cheremisinoff, N. P. *Handbook of Polymer Science and Technology*. (Taylor & Francis, 2019).

57. Brandrup, J., Immergut, E. H. & Grulke, E. A. *Polymer Handbook*. (1999).

58. *Polymer Data Handbook*. (Oxford University Press, 2009).

59. Lin, T. S. *et al.* PolyDAT: A Generic Data Schema for Polymer Characterization. *J. Chem. Inf. Model.* (2021) doi:10.1021/acs.jcim.1c00028.

60. Zhao, Z. W., Del Cueto, M., Geng, Y. & Troisi, A. Effect of Increasing the Descriptor Set on Machine Learning Prediction of Small Molecule-Based Organic Solar Cells. *Chem. Mater.* **32**, 7777–7787 (2020).

61. Sun, W. *et al.* Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Sci. Adv.* **5**, 1–8 (2019).

62. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminform.* **12**, 1–22 (2020).

63. Jørgensen, P. B. *et al.* Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **148**, (2018).

64. Gallegos, L. C., Luchini, G., St. John, P. C., Kim, S. & Paton, R. S. Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties. *Acc. Chem. Res.* **54**, 827–836 (2021).

65. Batra, R. *et al.* General Atomic Neighborhood Fingerprint for Machine Learning-Based Methods. *J. Phys. Chem. C* (2019) doi:10.1021/acs.jpcc.9b03925.

66. Pattanaik, L. & Coley, C. W. Molecular Representation: Going Long on Fingerprints. *Chem* **6**, 1204–1207 (2020).

67. Mercado, R. *et al.* Graph networks for molecular design. *Mach. Learn. Sci. Technol.* **2**, (2021).

68. Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).

69. Mannodi-Kanakkithodi, A. & Ramprasad, R. Chapter 9. Rational Design of Polymer Dielectrics: An Application of Density Functional Theory and Machine Learning. *Comput. Mater. Discov.* 293–319 (2018) doi:10.1039/9781788010122-00293.

70. Landrum, G. RDKit: Open-source chemiformatics.

71. Doan Tran, H. *et al.* Machine-learning predictions of polymer properties with Polymer Genome. *J. Appl. Phys.* **128**, (2020).

72. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D. & Ramprasad, R. Polymer Genome: A Data-Powered Polymer Informatics Platform for Property Predictions. *J. Phys. Chem. C* **122**, 17575–17585 (2018).

73. Wu, K. *et al.* Prediction of polymer properties using infinite chain descriptors (ICD) and machine learning: Toward optimized dielectric polymeric materials. *J. Polym. Sci. Part B Polym. Phys.* **54**,

2082–2091 (2016).

74. Lin, T. S. *et al.* BigSMILES: A Structurally-Based Line Notation for Describing Macromolecules. *ACS Cent. Sci.* **5**, 1523–1531 (2019).

75. Webb, M. A., Jackson, N. E., Gil, P. S. & de Pablo, J. J. Targeted sequence design within the coarse-grained polymer genome. *Sci. Adv.* **6**, (2020).

76. Sifri, R. J., Padilla-Vélez, O., Coates, G. W. & Fors, B. P. Controlling the Shape of Molecular Weight Distributions in Coordination Polymerization and Its Impact on Physical Properties. *J. Am. Chem. Soc.* **142**, 1443–1448 (2020).

77. Sattari, K., Xie, Y. & Lin, J. Data-driven algorithms for inverse design of polymers. *Soft Matter* **17**, 7607–7622 (2021).

78. Chen, L. *et al.* Polymer informatics: Current status and critical next steps. *Mater. Sci. Eng. R Reports* **144**, 100595 (2021).

79. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci* **28**, 31–26 (1988).

80. Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* **1**, 045024 (2020).

81. Muller, A. C. & Guido, S. *Introduction to Machine Learning with Python*. (O'Reilly Media, 2016).

82. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Commun.* **9**, 860–866 (2019).

83. Zhu, M. X., Song, H. G., Yu, Q. C., Chen, J. M. & Zhang, H. Y. Machine-learning-driven discovery of polymers molecular structures with high thermal conductivity. *Int. J. Heat Mass Transf.* **162**, (2020).

84. Wetzel, S. J. Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E* **96**, 1–11 (2017).

85. Ma, W., Cheng, F., Xu, Y., Wen, Q. & Liu, Y. Probabilistic Representation and Inverse Design of Metamaterials Based on a Deep Generative Model with Semi-Supervised Learning Strategy. *Adv. Mater.* **31**, 1–9 (2019).

86. Doan, H. A. *et al.* Quantum Chemistry-Informed Active Learning to Accelerate the Design and Discovery of Sustainable Energy Storage Materials. *Chem. Mater.* **32**, 6338–6346 (2020).

87. Lovrić, M., Molero, J. M. & Kern, R. PySpark and RDKit: Moving towards Big Data in Cheminformatics. *Mol. Inform.* **38**, 4–7 (2019).

88. Landrum, G. RDKit Documentation. *Read. Writ.* (2011).

89. O'Boyle, N. M., Morley, C. & Hutchison, G. R. Pybel: A Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* **2**, 1–7 (2008).

90. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

91. Chandrasekaran, A., Kim, C., Venkatram, S. & Ramprasad, R. A Deep Learning Solvent-Selection Paradigm Powered by a Massive Solvent/Nonsolvent Database for Polymers. *Macromolecules* **53**, 4764–4769 (2020).

92. Minami, T. & Okuno, Y. Number Density Descriptor on Extended-Connectivity Fingerprints Combined with Machine Learning Approaches for Predicting Polymer Properties. *MRS Adv.* **3**, 2975–2980 (2018).

93. Chandrasekaran, A., Kim, C. & Ramprasad, R. Polymer Genome: A Polymer Informatics Platform to Accelerate Polymer Discovery. *Lect. Notes Phys.* **968**, 397–412 (2020).

94. Gasparotto, P., Bochicchio, D., Ceriotti, M. & Pavan, G. M. Identifying and Tracking Defects in Dynamic Supramolecular Polymers. *J. Phys. Chem. B* **124**, 589–599 (2020).

95. Chen, Z., Li, D., Wan, H., Liu, M. & Liu, J. Unsupervised machine learning methods for polymer nanocomposites data via molecular dynamics simulation. *Mol. Simul.* (2020) doi:10.1080/08927022.2020.1851028.

96. Huang, Y. *et al.* Structure-Property Correlation Study for Organic Photovoltaic Polymer Materials Using Data Science Approach. *J. Phys. Chem. C* **124**, 12871–12882 (2020).

97. Jadrich, R. B., Lindquist, B. A. & Truskett, T. M. Probabilistic inverse design for self-assembling materials. *J. Chem. Phys.* **146**, (2017).

98. Patra, T. K., Loeffler, T. D. & Sankaranarayanan, S. K. R. S. Accelerating copolymer inverse design using monte carlo tree search. *Nanoscale* **12**, 23653–23662 (2020).

99. Park, N. H. *et al.* A Recommender System for Inverse Design of Polycarbonates and Polyesters. *Macromolecules* **53**, 10847–10854 (2020).

100. Sivaraman, G. *et al.* A machine learning workflow for molecular analysis: application to melting points. *Mach. Learn. Sci. Technol.* **1**, 025015 (2020).

101. Aggarwal, C. C., Kong, X., Gu, Q., Han, J. & Yu, P. S. Active Learning: A Survey. in *Data Classification* (ed. Aggarwal, C. C.) 571–605 (CRC Press, 2014). doi:10.1201/b17320.

102. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput. Mater.* **5**, (2019).

103. Jha, A., Chandrasekaran, A., Kim, C. & Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: The example of polymer glass transition temperatures. *Model. Simul. Mater. Sci. Eng.* **27**, (2019).

104. Novikov, I. S., Shapeev, A. V. & Suleimanov, Y. V. Ring polymer molecular dynamics and active learning of moment tensor potential for gas-phase barrierless reactions: Application to S + H2. *J. Chem. Phys.* **151**, (2019).

105. Pruksawan, S., Lambard, G., Samitsu, S., Sodeyama, K. & Naito, M. Prediction and optimization of epoxy adhesive strength from a small dataset through active learning. *Sci. Technol. Adv. Mater.* **20**, 1010–1021 (2019).