


REVIEW ARTICLE

Machine learning in polymer informatics

Wuxin Sha¹ | Yan Li² | Shun Tang¹ | Jie Tian² | Yuming Zhao² |
Yaqing Guo¹ | Weixin Zhang¹ | Xinfang Zhang³ | Songfeng Lu³ |
Yuan-Cheng Cao¹  | Shijie Cheng¹

¹State Key Laboratory of Advanced Electromagnetic Engineering and Technology, School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan, China

²Shenzhen Power Supply Co. Ltd., Shenzhen, China

³School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China

Correspondence

Yuan-Cheng Cao, State Key Laboratory of Advanced Electromagnetic Engineering and Technology, School of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.
Email: yccao@hust.edu.cn

Funding information

National Natural Science Foundation of China (52077096); China Southern Power research fund for fire safety for large scale grid energy storage system (090000KK52190179).

Abstract

Polymers have been widely used in energy storage, construction, medicine, aerospace, and so on. However, the complexity of chemical composition and morphology of polymers has brought challenges to their development. Thanks to the integration of machine learning algorithms and large data resources, the data-driven methods have opened up a new road for the development of polymer science and engineering. The emerging polymer informatics attempts to accelerate the performance prediction and process optimization of new polymers by using machine learning models based on reliable data. With the gradual supplement of currently available databases, the emergence of new databases and the continuous improvement of machine learning algorithms, the research paradigm of polymer informatics will be more efficient and widely used. Based on these points, this paper reviews the development trends of machine learning assisted polymer informatics and provides a simple introduction for researchers in materials, artificial intelligence, and other fields.

1 | INTRODUCTION

Polymers take up a necessary and investigative substance category in materials science. They are exploited in a wide range of applications, from daily products such as plastic packaging to the most advanced technologies such as lithium-ion batteries, solar cells, and 3D printing materials.^{1–4} However, the vast combination space of broad monomer atomic structures, complex arrangement of chains and various synthetic processes of polymers brings great obstacles to human researchers because of the

inherent limitations of human cognitive ability facing huge numbers of articles and high-dimensional data.⁵ Only a small part of data can be employed while the vast amount of polymer research data remains silent. The current polymer research is mainly the inefficient “trial-and-error method” based on lots of experiments guided by experience, which greatly hindered the innovation of polymer materials.

Fortunately, with the aid of super computing power and advanced algorithms, machine learning (ML), a sub-field of artificial intelligence (AI) has evolved rapidly in

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *InfoMat* published by UESTC and John Wiley & Sons Australia, Ltd

recent years and is naturally good at absorbing and processing the high-dimensional data. In fact, the rapid application of ML technology with powerful classification and regression abilities has brought new opportunities in many domains,⁶⁻¹⁷ including polymers, as showed in Figure 1.¹⁸⁻²²

Nevertheless, the development of polymer informatics is still in its infancy state. More research work is in progress for the continuous accumulation of data, the optimization of ML algorithms, and the deep integration of the two factors in specific applications. In the rest of this report, we introduce the nuts and bolts of machine learning, list some classic datasets for polymers, and review several typical applications of machine learning in polymer research. Finally, the future development directions are discussed to enable more wide-ranging impacts of data-driven methods in polymer science.

2 | BASICS OF MACHINE LEARNING

Machine learning is an important branch of AI. In the first few decades of AI research, the computer is only responsible for executing the hard-coded algorithm designed by domain experts which partly ignores the essence of intelligence. Different from traditional AI, machine learning enables the computer to discover rules from the data itself. In general, the data for ML consist of sets of input and associated output values. The learning process of a model is to adjust the internal parameters in a function that, given a specific set of input values, calculate the acceptable output values. After the parameter adjustment based on data is completed, the model can calculate and predict the output values of new samples. Primarily, as shown in Figure 2, machine learning is divided into supervised learning, unsupervised learning, and reinforcement learning. Supervised learning means

that the training data contains output labels to achieve data classification or regression, such as support vector machine and artificial neural. If there is no output label in the data, this kind of ML is called unsupervised learning which can realize the data clustering or dimensionality reduction, such as K-means clustering and principal component analysis. Reinforcement learning is an iteration method where an agent takes action to change its state and interact with the environment to maximize its target reward value, such as Markov decision process and active learning. Hence, the successful application of ML requires appropriate algorithms and sufficient data. We will discuss both of them in the next part.

2.1 | Algorithm

There are many categories of machine-learning algorithms. Appropriate algorithms should be selected based on the type and amount of available data. A brief introduction about these ML algorithms mentioned in the following examples will be provided in the following section, including graph neural networks (GNN), Gaussian process regression (GPR), active learning, and support vector machine (SVM).

GNN is presently a popular algorithm in material informatics by virtue of its suitability for representations of molecules and materials.²³⁻²⁷ Basically, GNN is a kind of artificial neural network (ANN) which consists of nodes connected by directed links. Each link between nodes has a corresponding numeric weight which determines the strength of the connection. These nodes and links make up different layers, commonly including the input layer, the output layer, and hidden layers. The input used in GNN, graph, a kind of common computer data structure, subtly overcomes the representation problem for polymers. Valid information representation for polymer molecules with branching, chirality, or other

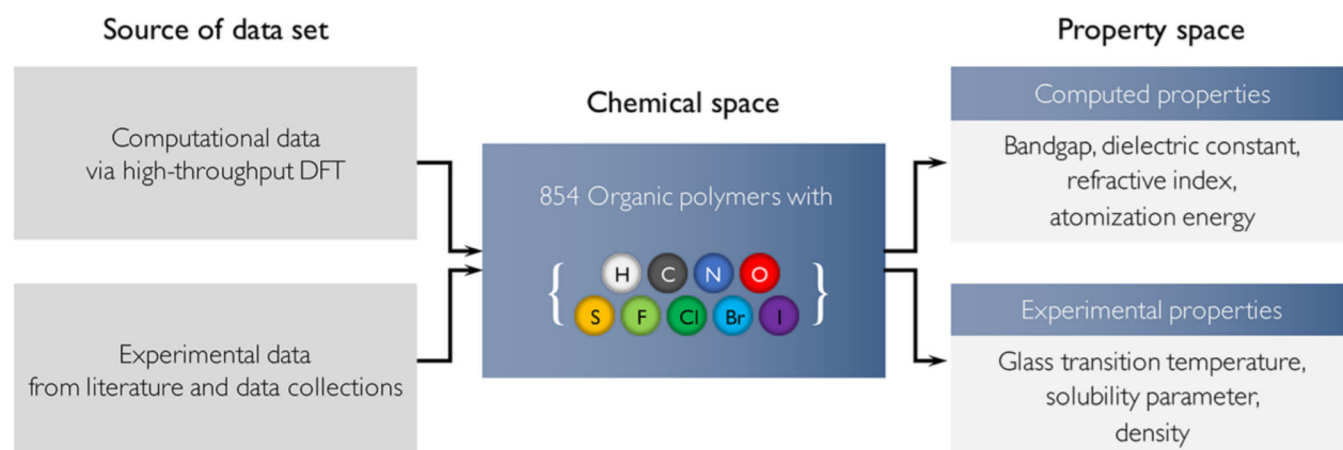


FIGURE 1 Overview of the polymer dataset used for the development of property prediction models⁵⁵

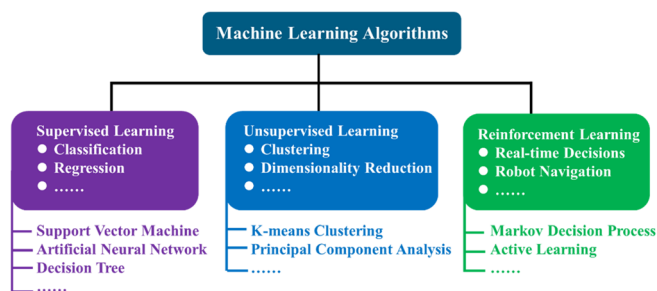


FIGURE 2 The overall categories of machine learning algorithms

intricate structures is more complex than other materials such as crystal that can be represented by a few crystallographic parameters. Graph uses different rows and columns in the high-dimensional matrix to represent atom types, the bond length, and bond angle, which lend themselves to encode the characteristics of polymers. The goal of GNN models is to find appropriate parameters to minimize the difference between calculated values and expected output values and finally identify trends, patterns, or clustering in the data.

GPR is a kind of regression algorithm to predict the properties of samples according to the similarity between samples. It has two crucial benefits. First, GPR learns a probabilistic model of the specific property and provides meaningful confidence intervals for the prediction. Second, there are only a few hyperparameters in GPR models, making the optimization efficiency relatively higher than normal ML models like neural networks with hundreds of parameters. The goal of GPR models is to predict the attributes of new data points in terms of the similarity between data.

Active learning may be of value if there is a chance for the model to catch new data as supplementary to datasets with a limited number of samples. In statistics, active learning is also called optimal experimental design. The main mission for the algorithm is to throw “query” (data without output values or unlabeled data) to the experts so that experts can conduct experiments to determine the label of the data. One may repeat this progress until the model can get better performance. Sometimes the “query” is the most difficult sample for the model to distinguish. Or it can be the sample with the largest model improvement. In general, the specific method for “query” selection needs to combine with a specific algorithm. The goal of active learning models is to obtain better performance with fewer new samples by paying more attention to “query.”

SVM is a kind of classification algorithm with the task to find a hyperplane. One may classify the samples according to the relative position between the data and the hyperplane. The linearly separable data can be directly optimized while the linearly inseparable data need to use the kernel function. The kernel function transforms the original data into a higher dimensional

representation and makes the problem linearly separable. The goal of SVM models is to optimize the position of the hyperplane and identify the category for each sample. Similarly, adapted from SVM, support vector regression (SVR) is to find the analytic expression of a hyperplane within an acceptable error. The goal of SVR models is to use the analytic expression to predict the interesting attribute values of new data points.

2.2 | Data

In addition, the databases also play a vital role in polymer informatics. Recently, many polymer-related databases are being established and improved, such as PoLyInfo.²⁸ Some related databases are listed in Table 1. The research on polymer materials by the Materials Project in the United States and the Novel Materials Discovery Laboratory in Europe has also led to the paradigm shift in the discovery of new functional polymers.²⁹⁻³¹ Nevertheless, some imperfections of current databases still hinder the widespread applications of polymer informatics. For instance, many polymer data resources are not fully accessible and tend to be small compared to the general fields of materials or AI research.³²⁻³⁴ And there is a lack of databases containing processing details or experimental information. Polymer communities need to encourage more researchers to make their code and data available for the public. Thus, drawing from the growing number of large-scale polymer databases, ML technology can discover more comprehensive and solid rules in polymers and continuously improve the polymer informatics framework.^{32,35-42}

3 | EXAMPLES OF MACHINE LEARNING APPLICATIONS

Because of the powerful linear and nonlinear fitting ability of machine learning, material informatics, which combines material datasets and machine learning algorithm, has developed rapidly in guiding the application of polymers. In this section, we will explore how data-driven machine learning can help polymer modeling, property prediction, synthesis, and characterization.

3.1 | Machine learning can assist in polymer modeling and analysis

For successful research projects, modeling and calculation are generally considered to be the necessary means to explain experimental phenomena. Molecular dynamics simulation, density functional theory, and other

TABLE 1 Publicly accessible databases for polymers

Name	Description	URL
Materials Project	Computed properties of known and hypothetical materials	https://materialsproject.org
Protein Data Bank (PDB)	3D structures of proteins, nucleic acids, and complex assemblies	http://www.wwpdb.org
Citrination	Computed and experimental properties of materials	https://citrination.com
Polymer Genome	An informatics platform for polymer property prediction and design	https://www.polymergenome.org
PoLyInfo	Various data required for polymeric material design	https://polymer.nims.go.jp
NanoMine	An open-source data resource for members of the nanocomposites community	https://materialsmine.org/nm#/
Polymer Property Predictor and Database	Flory–Huggins χ parameters and glass transition temperatures for various polymers	https://pppdb.uchicago.edu
Physical Properties of Polymers	Various physical properties and characterization techniques of polymers	by J. Mark, K. Ngai, W. Graessley, L. Mandelkern, E. Samulski, J. Koenig and G. Wignall
ACD/Labs NMR Databases	Polymer NMR spectra	https://www.acdlabs.com/products/dbs/nmr_db
Polymer Science Learning Center Spectral Database	Polymer IR and NMR spectra	https://pslc.uwsp.edu
NIST Synthetic Polymer MALDI Recipes Database	Matrix-assisted laser desorption ionization (MALDI) mass spectrometry on a wide variety of synthetic polymers	https://maldi.nist.gov
CROW Polymer Properties Database	A multitude of polymer properties	http://polymerdatabase.com
MATWEB Material Property Data	Material properties of thermoplastic and thermoset polymers	http://www.matweb.com
Material Properties Database	Engineering material properties that emphasize ease of comparison	https://www.makeitfrom.com

simulation methods have accumulated a lot of calculation results for the polymer field.^{43–47} Understanding the dynamical processes that govern the performance of polymers by exploring the simulation results becomes a big challenge in polymer modeling and analysis. As shown in Figure 3, Xie et al. used the MD trajectory file of poly(ethylene oxide) (PEO)/lithium bis-trifluoromethyl sulfonimide (LiTFSI) composite electrolytes as datasets to train the GNN models which then output specific vectors that contain information of the target atoms, including local configurations and bonding environments.⁴⁸ The GNN models were combined with the Koopman models which use a function to map the local configuration of target atoms into a lower-dimensional feature space, so the nonlinear dynamics can be approximated by a linear transition matrix. Then, the cluster analysis of these vectors is carried out to classify the possible four

coordination states of lithium ion (state 0, state 1, state 2, and state 3), calculate the conductivity ratio of each coordination state, and explain the transport of ionic clusters in PEO/LiTFSI system under high salt concentration. And AI-assisted modeling will have the potential to explore a wider range of polymer systems and understanding atomic dynamics and molecular structures that are crucial to their performances in the future.^{49,50}

ML also shows great potential in modeling and analyzing biomacromolecules. AlphaFold,⁵¹ the AI model created by DeepMind, can predict the folding spatial structure of protein segments from the sequence of amino acid residues. Although billions of base pairs have been sequenced by the Human Genome Project, there are still some challenges in inferring the function of specific gene segments because of the unknown folding mode of the polypeptide obtained from the transcription

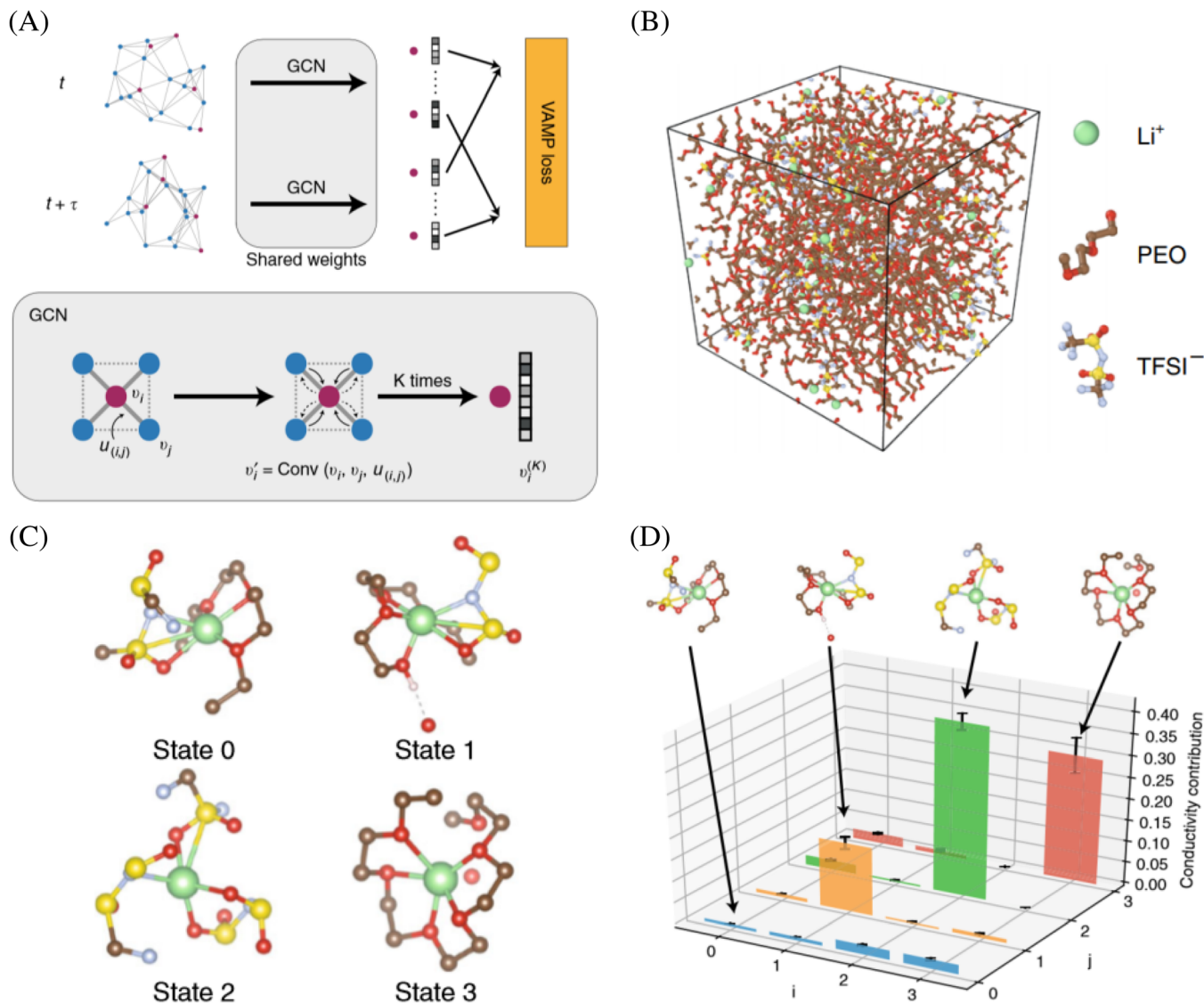


FIGURE 3 Machine learning for lithium ion dynamics in polymer electrolytes.⁴⁸ (A) Illustration of the graph dynamical networks architecture. (B) Structure of the PEO/LiTFSI polymer electrolyte. (C) Representative configurations of the four Li-ion states learned by the dynamical model. (D) Contribution from each transition to lithium ion conduction. Each bar denotes the percentage that the transition from state i to state j contributes to the overall lithium ion conduction. The error bars report the 95% confidence interval from four independent trajectories

and translation of the corresponding base sequence.^{52,53} The understanding of protein folding is also helpful for the design of proteins to meet specific functions. So Senior et al. developed the deep learning models which employed the amino acid sequences to predict the pairwise distances between residues and backbone torsion angles. After training with the data of nearly 30 000 proteins in PDB, the model won the 13th Critical Assessment of Protein Structure Prediction (CASP13).⁵⁴ In the future, with the improvement of various genome projects and the in-situ observation of protein spatial structure by frozen electron microscopy, more and more accurate and comprehensive data will come into existence, and the accuracy of machine learning model prediction will continue to improve.

3.2 | Machine learning can help to predict the properties of polymers

It has become an emerging area to apply machine learning methods to predict the properties of complex polymer-based materials and their composites. For example, Kim et al. built a database to summarize the characteristics and properties of 854 different polymers.⁵⁵ Its characteristics span multiple length scales, ranging from the type of atoms to the ratio of side-chain and main-chain atoms, so as to accurately describe the characteristics of specific polymers. This database includes DFT calculation results (atomization energy, bandgap, dielectric constant, etc.) and experimental data (such as permeability, glass transition temperature, and mechanical

properties). Then they created an information platform, which used the machine learning model based on the GPR algorithm to link the key features and properties of polymers, and used the database for training, achieving the purpose of predicting polymer performance. The above work has been opened on the related website (www.polymergenome.org) to guide the polymer selection for further research.

GPR algorithm has also been applied to predict the properties of polymer electrolytes. Hatakeyama-Sato et al. have constructed a large experimental database for lithium-conducting solid polymer electrolytes which contains more than 10^4 entries including properties like conductivity, transfer number, and chemical stability. Then, they employ GNN models to express the polymer molecules and GPR models to anticipate the conductivity of a new sample according to its similarity with historical samples in the database.⁵⁶ The accuracy of the models is satisfactory on a logarithmic scale (mean absolute error of less than 1). In the future, the property predictions generated by machine learning models can help researches to explore unknown polymer systems more efficiently.

3.3 | Machine learning can help to design and identify new polymers

Glass transition temperature (T_g) is one of the most important properties for polymers, which is related to the microscopic segment movement and macroscopic mechanical properties. One of the key challenges in guiding experiments to find polymers with required T_g is how to navigate efficiently in a wide composition and structure space. Kim et al. applied the active learning framework combined with the GPR algorithm to effectively select out the polymers posting a T_g greater than a certain threshold.⁵⁷ As shown in Figure 4, first, a model is trained based on the current knowledge of five randomly selected polymers. Using this model, predictions and associated uncertainties are obtained for new cases. Depending on the strategy that one wishes to employ, one may use the prediction, the uncertainty, or both the prediction and uncertainty to suggest the next-best case to be studied. Once the T_g of a new case has been tested, the results thus obtained are used to update the current knowledge of the system and the iteration is repeated until the desired T_g is achieved. They found that the ML model is more efficient in experimental design. Only several iterations are needed to find the polymers with high enough T_g . This idea can be widely used to guide the next step of synthesis in the polymer science fields lack of enough training data.⁵⁸⁻⁶⁰

3.4 | Machine learning can accelerate the characterization of polymers

The composition and structure of polymers are usually inferred by characterization methods, such as mass spectrometry, nuclear magnetic resonance and infrared spectroscopy. Machine learning methods have recently shown promise in accelerating the analysis of characterization data. For example, SVM has been applied to conduct thermal analysis of polymers.⁶¹ The dataset to train the SVM model is a matrix X of dimensions 293×215 collected by the technique of Dynamic Mechanical Analysis (DMA) which could record the stiffness of samples being heated. The stiffness of 293 samples with the temperature ranging from -51 to 270°C in increments of 1.5°C is learned by the SVM model. After maximizing the margins of SVM, these samples can be classified into two types (amorphous or semicrystalline) by their DMA data. Analogously, SVR algorithm is applied to predict the concentrations of specific polymers using data collected by UV/vis (ultraviolet-visible) spectra. The training dataset is a table recording the UV/vis spectra of a series of mixtures of 10 polyaromatic hydrocarbons (PAHs). The SVR model could manage to predict the concentrations of benzantracene (one of the 10 PAHs) after parameter optimization. With the further understanding of these intelligent algorithms by experts in various fields, machine learning will open up a new way in the more complex field of polymer characterization.

4 | SUMMARY AND OUTLOOK

Polymer informatics, including datasets, feature engineering and machine learning models, is still in its early stage. This paper reviews the representative research progress of polymer informatics in recent years, introduces the realization process of machine learning in polymer

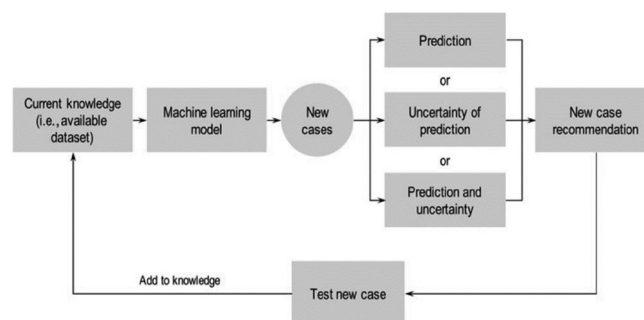


FIGURE 4 Overview of a typical active-learning framework.⁵⁷ Active-learning and materials design: the example of high glass transition temperature polymers

informatics, and emphasizes the advantages of data-driven method in the discovery of new excellent polymers. The following will look forward to the future development of polymer informatics from three aspects: data, algorithm, and computational power.

First of all, machine learning is a kind of data analysis method. The required data pursue authenticity, comprehensiveness, and objectivity. Therefore, it is necessary to develop high-throughput synthesis experiments coupled with high-throughput characterization, or high-throughput simulation to generate data.⁶²⁻⁶⁸ Meanwhile, the source website (such as ImageNet) that generates tag data is also a trend of data collection.

Second, in the fields lack of enough data, some advanced ML models can exhibit higher accuracy than conventional ML models.^{56,69-72} For example, transfer learning, a new kind of supervised learning method, makes full use of the similarity of different research problems and reduces the parameter space to be explored. Yamada et al. applied transfer learning models and succeeded to predict the specific heat capacity of polymers with excellent accuracy.⁶⁹ Besides, feature engineering, which is responsible for transforming the original data into key information input formats without redundancy, needs to cover a wider range of chemical space, fully capturing the characteristics of conformation (such as tacticity) and morphology (such as copolymerization, crystallization and side-chain).

Finally, in order to reduce the computing pressure of data centers and avoid the low efficiency of data upload and download, the idea of edge computing can be applied to polymer development. The sensor is coupled to the synthesis or characterization equipment, and the results of experimental or characterization are directly fed back to local AI processors to realize the timely optimization of experimental parameters. The relationship between ideal material properties and complex environmental parameters could be fast obtained by using ML algorithms.

AI represented by machine learning, with its larger storage space, broader datasets and faster computing power than human beings, has great potential to find complex basic knowledge in high-dimensional data and will become a powerful tool to accelerate the discovery process of polymer materials. Polymer researchers need learn to master this tool to find better polymers with faster speed and deeper understanding.

ACKNOWLEDGMENTS

This project was supported by the fund from National Natural Science Foundation of China (52077096) and China Southern Power research fund for fire safety for large scale grid energy storage system (090000KK52190179). Dr. Tang thanks the funds from Science and Technology Research Project of Hubei Education Department (B2017266).

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ORCID

Yuan-Cheng Cao  <https://orcid.org/0000-0001-5790-8608>

REFERENCES

1. Xu GL, Liu Q, Lau KS, et al. Building ultraconformal protective layers on both secondary and primary particles of layered lithium transition metal oxide cathodes. *Nat Energy*. 2019;4:484-494.
2. Tang S, Lan Q, Xu L, et al. A novel cross-linked nanocomposite solid-state electrolyte with super flexibility and performance for lithium metal battery. *Nano Energy*. 2020;71:104600.
3. Zhang J, Zhang Y, Fang J, et al. Conjugated polymer-small molecule alloy leads to high efficient ternary organic solar cells. *J Am Chem Soc*. 2015;137:8176-8183.
4. Tumbleston JR, Shirvanyants D, Ermoshkin N, et al. Continuous liquid interface production of 3D objects. *Science*. 2015;347:1349-1352.
5. Zahrt AF, Henle JJ, Rose BT, Wang Y, Darrow WT, Denmark SE. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science*. 2019;363:eaau5631.
6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84-90.
7. Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. *Sci Adv*. 2018;4:eaap7885.
8. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge. *Nature*. 2017;550:354-359.
9. Silver D, Huang A, Maddison CJ, et al. Mastering the game of go with deep neural networks and tree search. *Nature*. 2016;529:484-489.
10. Bory P. Deep new: the shifting narratives of artificial intelligence from deep blue to AlphaGo. *Convergence*. 2019;25:627-642.
11. Baughman AK, Chuang W, Dixon KR, Benz Z, Basilico J. DeepQA Jeopardy! Gamification: a machine-learning perspective. *IEEE Trans Comput Intell AI Games*. 2014;6:55-66.
12. Brown N, Sandholm T. Superhuman AI for multiplayer poker. *Science*. 2019;365:885-890.
13. Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature*. 2019;572:106-111.
14. Yao Y, Li X, Liu X, et al. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *Int J Geogr Inf Sci*. 2016;31:825-848.
15. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37:1038-1040.
16. Zhang D, Cao R, Wu S. Information fusion in visual question answering: a survey. *Inf Fusion*. 2019;52:268-280.
17. Liu Y, Han F, Li F, et al. Inkjet-printed unclonable quantum dot fluorescent anti-counterfeiting labels with artificial intelligence authentication. *Nat Commun*. 2019;10:2409.
18. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521:436-444.
19. Mannodi-Kanakkithodi A, Pilania G, Ramprasad R. Critical assessment of regression-based machine learning methods for polymer dielectrics. *Comput Mater Sci*. 2016;125:123-135.

20. Zhuang X, Zhou S. The prediction of self-healing capacity of bacteria-based concrete using machine learning approaches. *Comput Mater Contin.* 2019;59:57-77.
21. Wang W, Moreau NG, Yuan Y, Race PR, Pang W. Towards machine learning approaches for predicting the self-healing efficiency of materials. *Comput Mater Sci.* 2019;168:180-187.
22. Sha W, Guo Y, Yuan Q, et al. Artificial intelligence to power the future of materials science and engineering. *Adv Intell Syst.* 2020;2:2070320.
23. Duvenaud D, Maclaurin D, Aguilera-Iparraguirre J, et al. Convolutional networks on graphs for learning molecular fingerprints. *Adv Neural Inf Proces Syst.* 2015;28:2224-2232.
24. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des.* 2016;30:595-608.
25. Schutt KT, Arbabzadah F, Chmiela S, Muller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. *Nat Commun.* 2017;8:13890.
26. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett.* 2018;120:145301.
27. Schutt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Muller KR. SchNet—a deep learning architecture for molecules and materials. *J Chem Phys.* 2018;148:241722.
28. Otsuka S, Kuwajima I, Hosoya J, Xu Y, Yamazaki M. Presented at 2011 Int. Conf. On Emerging Intelligent Data and Web Technologies. Albania: Tirana; 2011:22-29.
29. Materials Project. <https://materialsproject.org> (accessed January 2021).
30. The Novel Materials Discovery (NOMAD) Laboratory. <https://nomad-coe.eu/> (accessed January, 2021).
31. Tolle KM, Tansley D, Hey AJ. The fourth paradigm: data-intensive scientific discovery point of view. *Proc IEEE.* 2011;99:1334-1337.
32. Le T, Epa VC, Burden FR, Winkler DA. Quantitative structure–property relationship modeling of diverse materials properties. *Chem Rev.* 2012;112:2889-2919.
33. de Pablo JJ, Jones B, Lind C, Ozolins V, Ramirez AP. The materials genome initiative, the interplay of experiment, theory and computation. *Curr Opin Solid State Mater Sci.* 2014;18:99-117.
34. Persson N, McBride M, Grover M, Reichmanis E. Silicon Valley meets the ivory tower: searchable data repositories for experimental nanomaterials research. *Curr Opin Solid State Mater Sci.* 2016;20:338-343.
35. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature.* 2018;559:547-555.
36. Peerless JS, Milliken NJB, Oweida TJ, et al. Soft matter informatics: current Progress and challenges. *Adv Theory Simul.* 2018;2:1800129.
37. Audus DJ, de Pablo JJ. Polymer informatics: opportunities and challenges. *ACS Macro Lett.* 2017;6:1078-1082.
38. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature.* 2018;555:604-610.
39. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. *Nat Methods.* 2019;16:687-694.
40. Lu W, Xiao R, Yang J, Li H, Zhang W. Data mining-aided materials discovery and optimization. *J Mater.* 2017;3:191-201.
41. 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.* 2019;5:1017-1026.
42. Rickman JM, Lookman T, Kalinin SV. Materials informatics: from the atomic-level to the continuum. *Acta Mater.* 2019;168:473-510.
43. Chen L, Venkatram S, Kim C, Batra R, Chandrasekaran A, Ramprasad R. Electrochemical stability window of polymeric electrolytes. *Chem Mater.* 2019;31:4598-4604.
44. Mardt A, Pasquali L, Wu H, Noé F. VAMPnets for deep learning of molecular kinetics. *Nat Commun.* 2018;9:4443.
45. Lusch B, Kutz JN, Brunton SL. Deep learning for universal linear embeddings of nonlinear dynamics. *Nat Commun.* 2018;9:4950.
46. Habershon S, Manolopoulos DE, Markland TE, Miller TF III. Ring-polymer molecular dynamics: quantum effects in chemical dynamics from classical trajectories in an extended phase space. *Annu Rev Phys Chem.* 2013;64:387-413.
47. Borodin O, Smith GD. Mechanism of ion transport in amorphous poly-ethylene oxide-LiTFSI from molecular dynamics simulations. *Macromolecules.* 2006;39:1620-1629.
48. Xie T, France-Lanord A, Wang Y, Shao-Horn Y, Grossman JC. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nat Commun.* 2019;10:2667.
49. Grzybowski BA, Bishop KJ, Kowalczyk B, et al. The ‘wired’ universe of organic chemistry. *Nat Chem.* 2009;1:31-36.
50. Huang L, Zhang H, Deng D, et al. LinearFold: linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search. *Bioinformatics.* 2019;35:i295-i304.
51. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. *Nature.* 2020;577:706-710.
52. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.* 2001;291:1304-1351.
53. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science.* 2012;338:1042-1046.
54. AlQuraishi M. AlphaFold at CASP13. *Bioinformatics.* 2019;35:4862-4865.
55. 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;122:17575-17585.
56. Hatakeyama-Sato K, Tezuka T, Umeki M, Oyaizu K. AI-assisted exploration of Superionic glass-type Li(+) conductors with aromatic structures. *J Am Chem Soc.* 2020;142:3301-3305.
57. Kim C, Chandrasekaran A, Jha A, Ramprasad R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Commun.* 2019;9:860-866.
58. Exploring Successful Parameter Region for Coarse-Grained Simulation of Biomolecules by Bayesian Optimization and Active Learning. *ACS Cent Sci.* 2018;4(2):268-276.
59. Shmilovich K, Mansbach RA, Sidky H, et al. Discovery of self-assembling pi-conjugated peptides by active learning-directed coarse-grained molecular simulation. *J Phys Chem B.* 2020;124:3873-3891.
60. Wang Y, Xie T, France-Lanord A, et al. Toward designing highly conductive polymer electrolytes by machine learning assisted coarse-grained molecular dynamics. *Chem Mater.* 2020;32:4144-4151.

61. Brereton RG, Lloyd GR. Support vector machines for classification and regression. *Analyst*. 2010;135:230-267.
62. Lin B, Hedrick JL, Park NH, Waymouth RM. Programmable high-throughput platform for the rapid and scalable synthesis of polyester and polycarbonate libraries. *J Am Chem Soc*. 2019;141:8921-8927.
63. Ren F, Ward L, Williams T, et al. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci Adv*. 2018;4:eaq1566.
64. Carrete J, Li W, Mingo N, Wang S, Curtarolo S. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys Rev X*. 2014;4:11019.
65. Gomez-Bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat Mater*. 2016;15:1120-1127.
66. Aykol M, Kim S, Hegde VI, et al. High-throughput computational design of cathode coatings for Li-ion batteries. *Nat Commun*. 2016;7:13779.
67. Yang X, Wang Z, Zhao X, Song J, Zhang M, Liu H. MatCloud: a high-throughput computational infrastructure for integrated management of materials simulation, data and resources. *Comput Mater Sci*. 2018;146:319-333.
68. Ong SP. Accelerating materials science with high-throughput computations and machine learning. *Comput Mater Sci*. 2019;161:143-150.
69. Yamada H, Liu C, Wu S, et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent Sci*. 2019;5:1717-1730.
70. Ma JYYW, Liang PW, Li C, Jiang JJ. FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inf Fusion*. 2019;48:11-26.
71. Kim E, Jensen Z, van Grootel A, et al. Inorganic materials synthesis planning with literature-trained neural networks. *J Chem Inf Model*. 2020;60:1194-1201.
72. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. *npj Comput Mater*. 2017;3:1-13.

AUTHOR BIOGRAPHIES



Wuxin Sha received his bachelor's degree in School of Materials Science and Engineering from Huazhong University of Science and Technology (HUST) in 2017. He is currently pursuing his PhD degree in School of Computer Science and Technology,

HUST. His research interests focus on artificial intelligent (AI) assisted materials genome, machine learning and solid-state electrolytes lithium batteries.



Yuan-Cheng Cao is currently a professor of State Key Laboratory of Advanced Electromagnetic Engineering and Technology at Huazhong University of Science and Technology (HUST, Wuhan). He received his PhD from HUST in 2006. Then he worked in Nottingham Trent University (UK, 2007-2010), Newcastle University (UK, 2010-2014), Jiangnan University (Wuhan, 2014-2018). He was selected as member of "Hundred Talents Program"(2018) and "Yellow-Crane Talent"(2015). His current research interests include solid state electrolytes in energy storage batteries, safety and extinguishing control for grid energy storage, eco-friendly recycling and regeneration of decommissioned batteries.



Shijie Cheng is the member of Chinese Academy of Sciences and professor of Huazhong University of Science and Technology. Prof. Cheng received his Bachelor's degree from Xi'an Jiaotong University in 1967, his Master's degree from HUST in 1981, and his PhD from the University of Calgary (Canada) in 1986, respectively, all in Electrical Engineering. In 2007, Prof. Cheng was elected as the member of the Chinese Academy of Sciences. He is currently engaged in the research on energy storage systems for electric power system stability and advanced materials for electrical engineering.

How to cite this article: Sha W, Li Y, Tang S, et al. Machine learning in polymer informatics. *InfoMat*. 2021;3:353-361. <https://doi.org/10.1002/inf2.12167>