Engineering 27 (2023) 31–36

Contents lists available at ScienceDirect

Engineering

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

Research
Smart Process Manufacturing toward Carbon Neutrality—Perspective

An Intelligent Manufacturing Platform of Polymers: Polymeric Material Genome Engineering

Liang Gao, Liquan Wang, Jiaping Lin *, Lei Du

Shanghai Key Laboratory of Advanced Polymeric Materials, Key Laboratory for Ultrafine Materials of Ministry of Education, Frontiers Science Center for Materiobiology and Dynamic Chemistry, School of Materials Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

ARTICLE INFO

Article history:
Received 8 October 2022
Revised 15 December 2022
Accepted 14 January 2023
Available online 1 August 2023

Keywords:
Polymeric materials
Materials genome approach
Machine learning
Property prediction
Rational design

ABSTRACT

Polymeric materials with excellent performance are the foundation for developing high-level technology and advanced manufacturing. Polymeric material genome engineering (PMGE) is becoming a vital platform for the intelligent manufacturing of polymeric materials. However, the development of PMGE is still in its infancy, and many issues remain to be addressed. In this perspective, we elaborate on the PMGE concepts, summarize the state-of-the-art research and achievements, and highlight the challenges and prospects in this field. In particular, we focus on property estimation approaches, including property proxy prediction and machine learning prediction of polymer properties. The potential engineering applications of PMGE are discussed, including the fields of advanced composites, polymeric materials for communications, and integrated circuits.

© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Materials with excellent performances are the foundation of the development of high-level technology and advanced manufacturing. Heretofore, materials science has undergone four research paradigms—namely, the experimental empirical, model-based theoretical, computational, and data-driven science paradigms [1–3]. As shown in Fig. 1, the first paradigm is based on the experimental trial-and-error approach. In the second paradigm, scientific laws are discovered by summarizing experimental experience and building physical models. In the third paradigm, the microscopic states of atoms or molecules are simulated by computers to obtain the macroscopic properties of materials. Both theories and simulations can provide accurate data from model-based theoretical science. With the advancement of information science and artificial intelligence (AI), the fourth paradigm emerged in the early 2000s. This paradigm is a research approach that utilizes algorithms to analyze large amounts of data and find the underlying rules. Unlike the second and third paradigms, the fourth paradigm can infer and predict unknown data based on existing experimental data. The combination of these four paradigms has resulted in the emergence of various advanced materials. However, the first research paradigm requires inevitable trial and error, resulting in long research cycles for discovering materials. The fourth paradigm based on data-driven aims to accelerate material research and reduce the cost through virtual synthesis, property prediction, and screening. It is evolving into a revolutionary paradigm [4–10].

Big data science is one of the foundations of interdisciplinary disciplines, including bioinformatics, chemoinformatics, and materials informatics. As a landmark achievement of bioinformatics, AlphaFold2 has partially surpassed human experts in predicting the sequences and three-dimensional structures of proteins [11]. In chemoinformatics, the use of AI to drive the discovery of new drugs is efficient and well-known. Unlike bioinformatics and chemoinformatics, which are now well established, materials informatics is still a rapidly growing field. As a pioneer, materials genome engineering (MGE) is becoming a vital platform for material intelligent manufacturing. With the development of MGE, the customized design and preparation of materials show advantages and potential.

2. Development of polymeric material genome engineering

The research paradigm of polymeric material genome engineering (PMGE) involves theoretical calculation, database technology, prediction and screening, and verification, with the aim of achieving rational design, virtual preparation, and intelligent

* Corresponding author.
E-mail address: jlin@ecust.edu.cn (J. Lin).

https://doi.org/10.1016/j.eng.2023.01.018
2095-8099/© 2023 THE AUTHORS. Published by Elsevier LTD on behalf of Chinese Academy of Engineering and Higher Education Press Limited Company.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

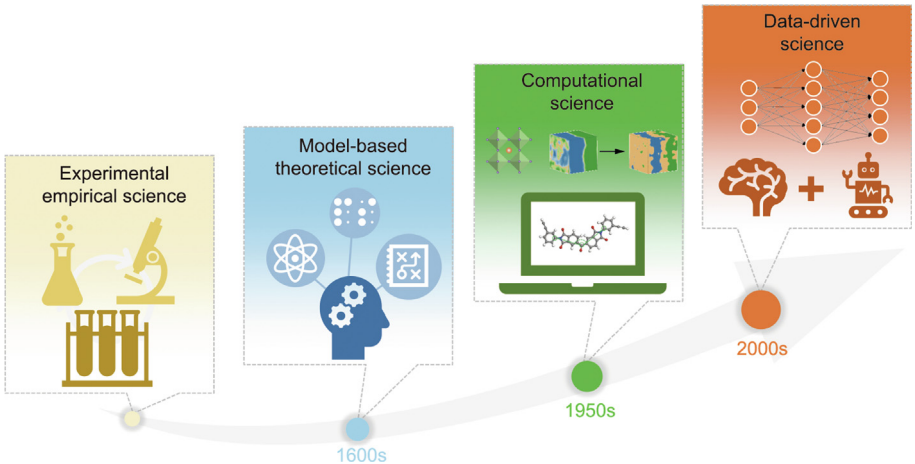

Fig. 1. The development of the four paradigms of materials research: experimental empirical, model-based theoretical, computational, and data-driven science paradigms. The first paradigm requires trial-and-error, resulting in long research cycles for discovering materials. Materials research has now entered a data-driven age (the fourth paradigm).

manufacturing, and accelerating the design and development of polymeric materials (Fig. 2) [1,2,12,13]. PMGE consists of the following three steps.

(1) **Definition of polymer "genes" and design of "virtual polymers."** According to certain rules based on analyses of existing chemical data and the experience of domain experts—that is, widely adopted theoretical models and empirical rules [13]—factors related to material properties, such as the chemical groups and elements comprising polymers, are defined as the so-called "genes" of polymers. Then, a series of "virtual polymers" can be designed by gene combining or editing (i.e., regulating the chain composition of the polymers).

(2) **High-throughput prediction and screening of polymer properties.** The quantitative structure–property relationship (QSPR) of the polymers is built based on experimental or simulation data to predict the properties of the designed "virtual polymers." Next, *in-silico* screening is conducted to obtain promising new polymers according to the performance requirements.

(3) **Verification.** The screened polymeric materials are synthesized and characterized to verify the reliability of the screening results and to optimize the prediction model. In addition, theoretical calculations with high accuracy can be used to verify the screening results. Furthermore, gene analysis based on PMGE can be conducted to deduce the underlying physics rules for inspiring future structural design of polymers.

Property estimation is the key to the rational design of materials. One type of prediction strategy involves finding the key features that can evaluate material properties through data mining. A calculable key feature is extracted as a proxy. Macroscopic properties that are difficult to be obtained accurately from the theoretical calculations are transformed into calculable proxy variables. Then, the polymeric materials can be screened by comparing the corresponding proxy variables. For example, Sharma et al. [14] utilized the band gap of polymer structures, which can easily be calculated using density functional theory (DFT), to represent the breakdown voltage and dielectric loss. Using the dielectric constant and the band gap as screening criteria, they obtained a series of promising all-organic polymer dielectrics.

The proxy variable strategy is sometimes empirical, however, the data-driven method can effectively eliminate subjective influences. For example, Zhu et al. [15] analyzed the existing experimental and computational data of more than 400 polymers from the PolyInfo database. They found that the 5% decomposition temperature (T_{d5}) of polymers depends on the bond dissociation energy (BDE) of the weakest bond in the polymer structure, where the Pearson correlation value is close to 0.7. Thus, the BDE can be considered a key feature for evaluating the thermal stability of polymeric materials. Then, they employed the polymer material genome to reconcile the contradiction between high thermal stability and low curing energy of resins [15]. The band gap calculated by DFT was considered to be the proxy of processability. Using the proposed prediction models of key features, two-step screening was performed to obtain the optimal poly(silane arylacetylene) (PSA) structure. Next, a promising PSA structure containing 2,7-diethynyl naphthalene was screened. The experimental verification indicated that the novel PSA resin exhibited a 5% thermal decomposition temperature of 655 °C and a curing enthalpy of 241.9 $J \cdot g^{-1}$, showing excellent comprehensive properties.

In addition, the ratio of bulk modulus to shear modulus (K/G) can be used to represent the toughness of polymers. After calculating and screening the proxy K/G for toughness and the proxy BDE for thermal stability, Gao et al. [16] obtained a novel acetylene-terminated polyimide (ATPI) that can be used to enhance the toughness of PSA resins through blending. For the copolymerized resin of ATPI and PSA, the toughness is significantly improved, while the heat resistance is maintained. As discussed above, it is

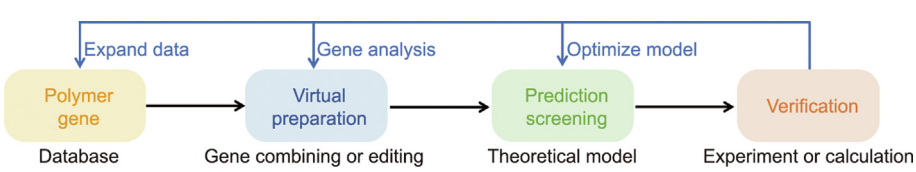

Fig. 2. The concept and steps of PMGE. Based on the database, polymer genes are defined and virtual polymers are designed. Then, the high-throughput prediction and screening of polymer properties are conducted via theoretical calculations or high-throughput experiments, and the screened results are verified through experiments or calculations.

effective and reliable to design and screen polymer materials using a prediction model of proxy variables. The key to property proxy prediction lies in mining the underlying relations between the target properties and the microscopic or macroscopic physical parameters.

Machine learning (ML) can mine underlying rules from historical data and predict, infer, or classify unknown data. This is another strategy to achieve high-throughput prediction and *in-silico* screening in PMGE [17–19]. The simplified molecular-input line-entry system (SMILES) provides a simple set of representations that are suitable as labels for chemical data [20]. SMILES can serve as an effective tool for translating chemistry knowledge into a machine-friendly form that fits many text-based ML algorithms [21]. Then, the QSPR between the inputs (e.g., SMILES, molecular graphs, molecular fingerprints, and other molecular descriptors) and the desired material properties can be constructed by training the existing data using various ML algorithms. An ML model trained on reliable experimental data can directly predict the material property. For example, based on open databases such as Polymer Genome and PubChem, Zhang et al. [17] utilized a multi-layer perceptron method to establish ML prediction models for a QSPR between the target properties (i.e., thermal decomposition temperature and viscosity) and polymer structures (Fig. 3). By gene combination, they obtained 368 candidate resins for screening. Using the two ML models, the properties of the candidate resins were predicted and screened with high throughput. Afterward, a series of resins with optimal processability and high heat resistance were obtained. The experimental verification demonstrated that the screened resin (PSNP-MV) has excellent comprehensive properties of processing and heat resistance.

When the experimental data is limited or low quality, theoretical calculation or simulation data can be utilized to train ML models directly. The obtained models can provide reliable property predictions. For example, based on DFT calculation results, Mannodi-Kanakkithodi et al. [18] built an ML model to predict band gap and permittivity by training the calculation data. When some data from the calculation, simulation, or database may be low fidelity, improving the data quality with a multi-fidelity surrogate model is an effective strategy [22]. Deviations between low-fidelity (e.g., simulation data) and high-fidelity (e.g., experimental) data can be trained, allowing an ML-based model to evaluate their differences and then improve the data quality. In addition, for the small-data conundrum, various advanced ML strategies can be utilized to avoid overfitting and improve the model generalization ability, such as physics-informed neural networks and the Bayesian method [23,24]. Using the above promising strategies, the problem of lack of experimental data is settled, and the design and screening of polymers can be realized.

Theoretical calculation can also be utilized to estimate polymer properties and screen the target properties, but sometimes it can be time-consuming. ML models can overcome the limitations of theoretical calculations, especially the time-consuming computational costs of a larger chemical structure space. When the number of polymer genes increases, the polymer structure space increases exponentially, and it is impractical to calculate the polymer properties. An ML model can achieve the prediction of these polymer properties in a short time. Overall, employing ML models brings the advantages of high prediction accuracy, a short development cycle, and broad applicability. These advantages well fit the requirements of material design and screening in PMGE.

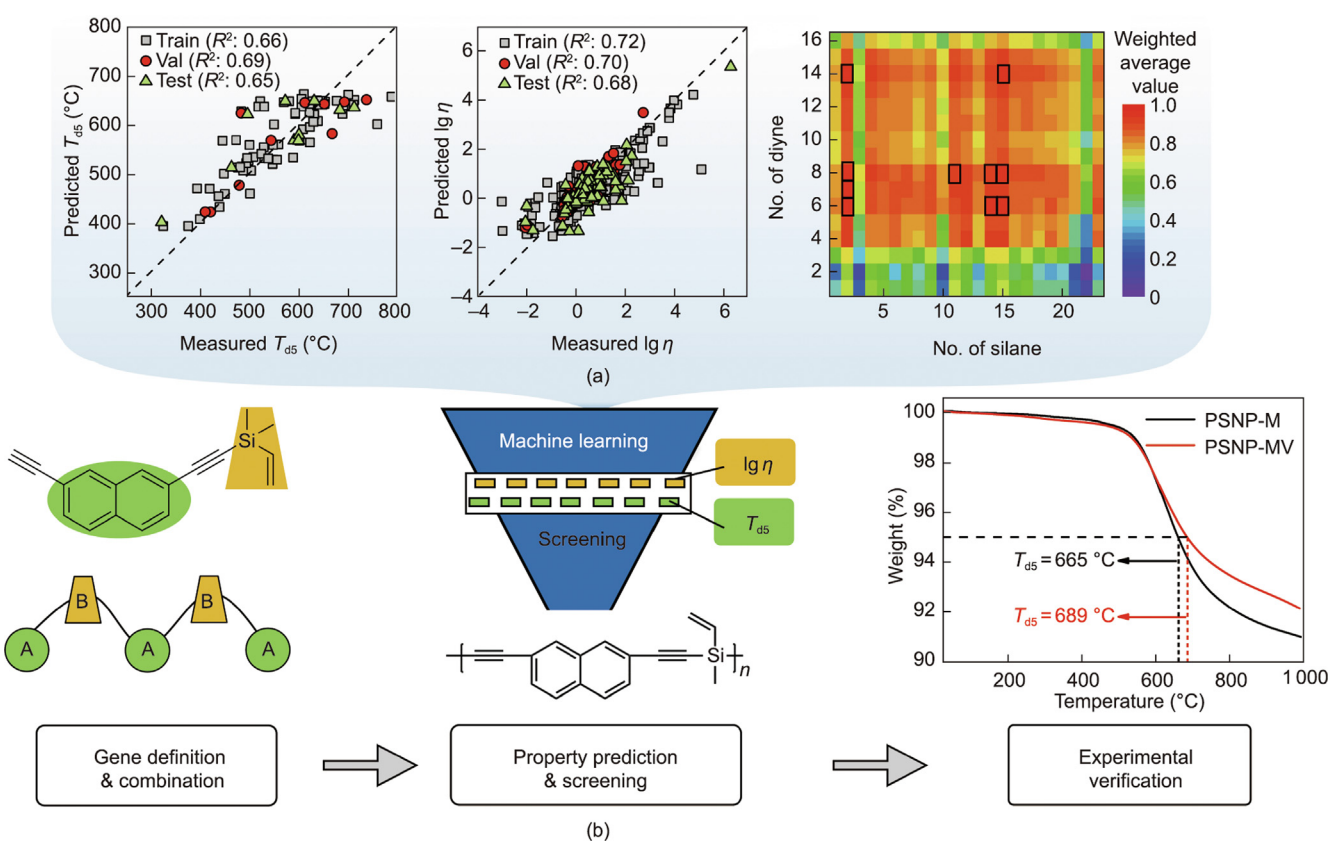

Fig. 3. ML prediction and screening of high-performance resin. (a) ML models of thermal decomposition temperature and viscosity, and the heat map of the comprehensive properties for 368 candidate resins. (b) Design of a novel PSNP-MV resin with excellent comprehensive properties aided by ML-enhanced materials genome approach. lg η: the common logarithm of the viscosity values. Reproduced from Ref. [17] with permission.

3. Challenges and prospects of PMGE

The development of the polymeric material genome is still in its infancy, and many issues remain to be addressed, as discussed below.

3.1. Gene definition and molecular structure description

The unique chain structures and complicated multiscale structures of polymers pose challenges to the gene definition and structural description of polymers. It is necessary to develop more advanced methods for describing the structural features of polymers. Existing methods, such as BigSMILES, graph representations, molecular fingerprints, and so forth, can be modified. In addition, new approaches from informatics or mathematics can be introduced. For polymer gene definition, to balance the flexibility of the structure design with synthesis accessibility in experiments, polymer genes can be defined according to the synthetic routes of the target polymer systems [13,15]. In BigSMILES, polymeric fragments are represented by a list of repeating units enclosed by curly brackets, making BigSMILES an excellent candidate for indexing identifiers in a polymer database system [25]. In addition, polymer genes should be systematically analyzed, classified, and labeled. The rules from data mining and the experience of domain experts should be combined to improve the accuracy and rationality of polymer gene definition and structure description.

Furthermore, multiscale characteristics should be considered in the structure description of polymers. For example, the chain information (e.g., conformation) and aggregation state (e.g., crystalline structure and cured crosslinked structure) can be obtained from theoretical calculations, simulations, and experiments. Recently, Hu et al. [26] utilized crosslinking density descriptors to predict the performance of cured epoxy resins. In addition, polymer polydispersity affects the multiscale structures of polymers, giving rise to variations in polymer properties. Polydispersity data can be identified, labeled, and included in the polymer database, and then served as one of the inputs when the QSPR is established to develop a reliable prediction model [27].

3.2. Prediction model of property proxy

It is necessary to find or establish more key features representing polymer properties, such as solvent resistance, wear resistance, impact resistance, and interface bonding property. In addition, to achieve the rapid prediction of polymer properties and multistep screening, more rapid calculation methods of proxies should be developed, such as the molecular connectivity method and group contribution method.

3.3. ML prediction of polymer properties

The current challenge is the lack of high-quality polymer structure–property data. In addition, the generalization ability of prediction models of polymer properties is not strong, and the multiscale structure–property relationship cannot be described precisely. All these issues limit the applications of ML prediction in PMGE. To address these challenges, data from open databases can be exploited and mined through natural language processing [28]. With the openness and sharing of the PMGE platform, more researchers will actively input data. Beyond that, massive data can also be obtained through high-throughput experiments and theoretical simulations. In addition, researchers should pay attention to the utilization of low-quality experimental data. In particular, all the data should be standardized to improve the data quality.

Moreover, advanced algorithms can be utilized to develop ML strategies for solving the problem of small amounts of data, such as transfer learning, supervised learning, and active learning [24,29]. Furthermore, introducing a prior algorithm and micro- or nano-structural information with molecular structure descriptors is promising for establishing ML prediction models with physical meanings. For example, the frequency-dependence mechanism for the polymer dielectric property can be introduced with a structural description to train the ML model. This could be beneficial in establishing accurate multiscale structure–property relationships.

3.4. High-throughput experiments

An experimental system for high-throughput polymer synthesis needs to be established, aiming to rapidly screen promising polymers, expand databases, and optimize prediction models. Current experimental techniques are developed from parallel synthesizers in other fields, and equipment for the high-throughput synthesis and characterization of polymers remains to be developed. Interdisciplinary research is an effective approach to this issue, involving various scientific and technological methods such as informatization, system control, and microfluidic technology.

In addition, the synthetic accessibility of polymers and the processing properties suitable for large-scale manufacturing should be considered in PMGE. Beyond forward prediction and screening, it is necessary to disassemble the performance demands of engineering applications and then develop a reversal design strategy to realize the double closed-loop design of PMGE. The reversal design of polymer structures will enrich the significance of PMGE and realize the rational design and intelligent manufacturing of polymers.

3.5. Engineering applications

As shown in Fig. 4, PMGE can accelerate the development of polymer materials in various engineering applications, especially when two or more properties are at odds with each other. For example, PMGE can be applicable in the following areas.

(1) **Advanced resin matrix composites.** In addition to the polymer resins, PMGE is being applied to the structural design and property improvement of polymer fibers with high strength and a high modulus. PMGE can also be employed to regulate the interface bonding between resin and fiber and to optimize the processability of composite materials. In addition, ML models can be trained based on the data from finite-element simulations and experiments of composites. With a trained ML model, the performance of composite materials can be rapidly predicted and screened, making it possible to realize the rational design of advanced composites.

(2) **Chemical engineering and catalysis.** The rational design and screening of various catalysts, including porous catalytic materials and polymerization catalysts, can be accelerated via an ML-enhanced material design strategy [30]. Catalysts determine the microstructure, macroscopic performance, and industrial efficiency of polyolefin. Thus, the structural design of catalysts is the key to advancing the polyolefin industry. For example, the rational design of the active sites of the Ziegler–Natta catalyst and the configurational selectivity prediction of the metallocene catalyst for propene polymerization are still challenging. The data-driven ML approach can provide a promising strategy for discovering and designing polymer catalysis.

(3) **Polymeric organic semiconductor materials.** Such polymer systems require high electron mobility, high luminescence efficiency, high spin characteristics, high conductivity, and so forth. It is difficult to obtain polymer materials with multiple excellent properties using a traditional trial-and-error approach. Designing conjugated polymers using PMGE can accelerate the research on

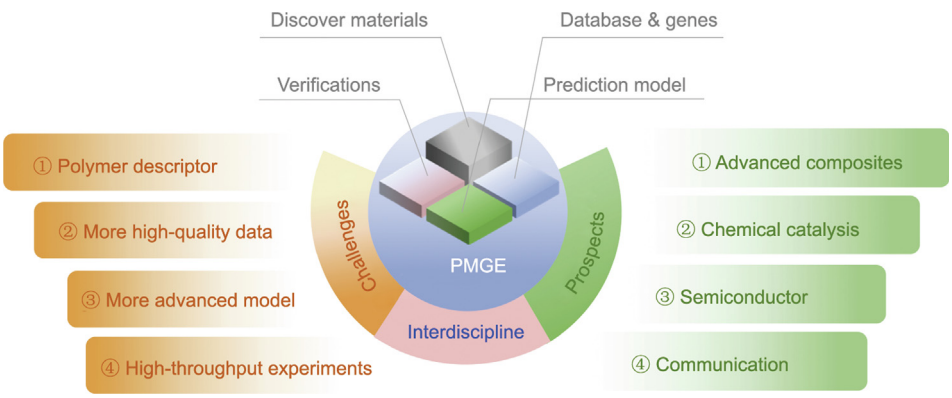

Fig. 4. The challenges and prospects of PMGE. Issues regarding polymer databases and genes, property prediction models, and verifications remain to be addressed through interdisciplinary research. The materials discovered by PMGE have potential for application in the fields of chemical engineering, semiconductors, communication, and more.

polymeric organic semiconductor materials with excellent comprehensive properties.

(4) **Communication and integrated circuit materials.** Polymers used in the field of high-frequency communication technology simultaneously require enhanced mechanical properties, heat resistance, and electromagnetic properties. For example, the polymer materials used in the sixth-generation (6G) communication equipment should have a relatively low dielectric constant and low dielectric loss. In addition, high-performance polymers used in chip packaging should have high heat resistance, low thermal expansion coefficient, high hardness, high toughness, high electrical insulation, and low dielectric constant. Therefore, the above engineering applications demand the discovery of advanced polymer materials with excellent comprehensive properties, and PMGE is undoubtedly the best choice. Through high-throughput prediction and screening, PMGE can meet the goal of discovering polymeric materials with excellent overall performances.

4. Summary

PMGE will fuel the innovation of the next generation of materials. It has the potential to lower the cost of materials research, balance the performance constraints, and even enable breakthroughs in polymeric materials. PMGE can revolutionize traditional polymer design methods and promote research progress in materials science. However, many issues remain to be addressed, as PMGE is still at an early stage. Interdisciplinary collaboration between the disciplines of information, mathematics, control engineering, and so forth can solve problems such as property prediction and experimental verification. In the future, the double closed loop of the forward prediction and screening, as well as the reversal design, may be realized. We envision PMGE as a sustainable public platform for polymer design and applications. Researchers can take advantage of PMGE for the rational design of the processing, composition, and performance of new polymeric materials in the fields of advanced composites, semiconductors, communication, and more.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (22103025, 51833003, 22173030, 21975073, and 51621002).

Compliance with ethics guidelines

Liang Gao, Liquan Wang, Jiaping Lin, and Lei Du declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] Yuan WL, He L, Tao GH, Shreeve JM. Materials-genome approach to energetic materials. Acc Mater Res 2021;2(9):692–6.
[2] Du S, Zhang S, Wang L, Lin J, Du L. Polymer genome approach: a new method for research and development of polymers. Acta Polym Sin 2022;53(6):592–607. Chinese.
[3] Xie J, Su Y, Zhang D, Feng Q. A vision of materials genome engineering in China. Engineering 2022;10:10–2.
[4] Doan Tran H, Kim C, Chen L, Chandrasekaran A, Batra R, Venkatram S, et al. Machine-learning predictions of polymer properties with polymer genome. J Appl Phys 2020;128(17):171104.
[5] Gao C, Min X, Fang M, Tao T, Zheng X, Liu Y, et al. Innovative materials science via machine learning. Adv Funct Mater 2022;32(1):2108044.
[6] Rizkin BA, Hartman RL. Supervised machine learning for prediction of zirconocene-catalyzed α-olefin polymerization. Chem Eng Sci 2019;210:115224.
[7] Xu P, Chen H, Li M, Lu W. New opportunity: machine learning for polymer materials design and discovery. Adv Theory Simul 2022;5(5):2100565.
[8] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater 2016;4(5):053208.
[9] Wang C, Fu H, Jiang L, Xue D, Xie J. A property-oriented design strategy for high performance copper alloys via machine learning. npj Comput Mater 2019;5:87.
[10] Xiong J, Shi SQ, Zhang TY. Machine learning of phases and mechanical properties in complex concentrated alloys. J Mater Sci Technol 2021;87:133–42.
[11] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.
[12] Zhao H, Li X, Zhang Y, Schadler LS, Chen W, Brinson LC. Perspective: NanoMine: a material genome approach for polymer nanocomposites analysis and design. APL Mater 2016;4(5):053204.
[13] Mannodi-Kanakkithodi A, Chandrasekaran A, Kim C, Huan TD, Pilania G, Botu V, et al. Scoping the polymer genome: a roadmap for rational polymer dielectrics design and beyond. Mater Today 2018;21(7):785–96.
[14] Sharma V, Wang C, Lorenzini RG, Ma R, Zhu Q, Sinkovits DW, et al. Rational design of all organic polymer dielectrics. Nat Commun 2014;5(1):4845.
[15] Zhu J, Chu M, Chen Z, Wang L, Lin J, Du L. Rational design of heat-resistant polymers with low curing energies by a materials genome approach. Chem Mater 2020;32(11):4527–35.
[16] Gao G, Zhang S, Wang L, Lin J, Qi H, Zhu J, et al. Developing highly tough, heat-resistant blend thermosets based on silicon-containing arylacetylene: a material genome approach. ACS Appl Mater Interfaces 2020;12(24):27587–97.
[17] Zhang S, Du S, Wang L, Lin J, Du L, Xu X, et al. Design of silicon-containing arylacetylene resins aided by machine learning enhanced materials genome approach. Chem Eng J 2022;448(15):137643.
[18] Mannodi-Kanakkithodi A, Pilania G, Huan TD, Lookman T, Ramprasad R. Machine learning strategy for accelerated design of polymer dielectrics. Sci Rep 2016;6(1):20952.
[19] Chen L, Kim C, Batra R, Lightstone JP, Wu C, Li Z, et al. Frequency-dependent dielectric constant prediction of polymers using machine learning. npj Comput Mater 2020;6:61.
[20] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28(1):31–6.
[21] Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018;9(2):513–30.
[22] Song X, Lv L, Sun W, Zhang J. A radial basis function-based multi-fidelity surrogate model: exploring correlation between high-fidelity and low-fidelity models. Struct Multidiscipl Optim 2019;60(3):965–81.

[23] Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys 2019;378:686–707.
[24] van de Schoot R, Depaoli S, King R, Kramer B, Märtens K, Tadesse MG, et al. Bayesian statistics and modelling. Nat Rev Methods Primers 2021;1(1):1.
[25] Lin TS, Coley CW, Mochigase H, Beech HK, Wang W, Wang Z, et al. BigSMILES: a structurally-based line notation for describing macromolecules. ACS Cent Sci 2019;5(9):1523–31.
[26] Hu Y, Zhao W, Wang L, Lin J, Du L. Machine-learning-assisted design of highly tough thermosetting polymers. ACS Appl Mater Interfaces 2022;14(49):55004–16.
[27] Ethier JG, Casukhela RK, Latimer JJ, Jacobsen MD, Shantz AB, Vaia RA. Deep learning of binary solution phase behavior of polystyrene. ACS Macro Lett 2021;10(6):749–54.
[28] Shetty P, Ramprasad R. Machine-guided polymer knowledge extraction using natural language processing: the example of named entity normalization. J Chem Inf Model 2021;61(11):5377–85.
[29] Wu S, Kondo Y, Kakimoto M, Yang B, Yamada H, Kuwajima I, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput Mater 2019;5:66.
[30] Boyd PG, Lee Y, Smit B. Computational development of the nanoporous materials genome. Nat Rev Mater 2017;2(8):17037.