Contents lists available at ScienceDirect

Engineering

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

Research
Smart Process Manufacturing toward Carbon Neutrality—Perspective

高分子材料的智能制造平台——高分子材料基因工程

高梁，王立权，林嘉平*，杜磊

Shanghai Key Laboratory of Advanced Polymeric Materials, Key Laboratory for Ultrafine Materials of Ministry of Education, Frontiers Science Center for Materiobiology andDynamic 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

关键词
高分子材料
材料基因组方法
机器学习
性能预测
理性设计

摘要

高性能高分子材料是高新科技和先进制造业的基石。高分子材料基因工程正在成为高分子材料智能制造的重要平台。然而，高分子材料基因工程的发展仍处于起步阶段，许多问题亟待解决。本文阐述了高分子材料基因工程的概念，总结了最新研究成果，并强调了该领域的重要挑战和发展前景。特别强调了高分子材料的性能预估方法，包括性能代理量预测和机器学习性能预测。最后，讨论了高分子材料基因工程在先进复合材料、通信和集成电路等领域所亟需的高性能高分子材料创制方面的潜在工程应用前景。

1. 引言

高性能材料是高科技和先进制造发展的基础。迄今为止，材料科学经历了四种研究范式：实验经验范式、基于模型的理论范式、计算模拟范式以及数据驱动范式[1–3]。如图1所示，第一个范式基于实验试错方法；在第二个范式中，通过总结实验经验和建立物理模型来发现科学定律；第三个范式通过计算机模拟原子或分子的微观状态来获得材料的宏观性质。理论范式和计算模拟范式都可以从基于模型的理论科学中得到准确的数据，随着信息科学和人工智能（artificial intelligence, AI）的发展，第四范式出现于2000年年初，该范式是一种利用算法分析大数据、寻找数据内在规律的研究方法。与第二范式和第三范式不同，第四范式可以根据现有的实验数据推断和预测未知数据。这四种范式的相互结合使用，能够开发出各种先进材料。然而，第一个研究范式不可避免地需要进行反复试验，导致材料的开发周期较长；而第四范式基于数据驱动，旨在通过虚拟地合成、性能预测和筛选来加速材料研究步伐并降低成本，它正在逐渐演变成为一种革命性的范式[4–10]。

大数据科学是生物信息学、化学信息学和材料信息学等跨学科研究的基础之一。AlphaFold2在预测蛋白质序列和三维结构方面已超越部分人类专家，成为生物信息学的里程碑式成就[11]。在化学信息学领域，使用AI驱动新型药物的发现也是高效且众所周知的。但材料信息学与生物信息学和化学信息学不同，仍是一个处于快速发展中的领

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

英文原文：Engineering 2023, 27(8): 31–36
引用本文：Liang Gao, Liquan Wang, Jiaping Lin, Lei Du. An Intelligent Manufacturing Platform of Polymers: Polymeric Material Genome Engineering. *Engineering*, https://doi.org/10.1016/j.eng.2023.01.018

域，材料基因组工程（material genome engineering, MGE）作为先驱，正在逐渐成为材料智能制造的重要平台。随着MGE的发展，其在材料的理性设计和制造方面显示出了独特的优势和潜力。

2. 高分子材料基因工程的进展

高分子材料基因工程（polymeric material genome engineering, PMGE）的研究范式涉及理论计算、数据库技术、预测筛选以及实验验证，旨在实现合理设计、虚拟制备和智能制造，加速高分子材料的设计和开发（图2）[1–2,12–13]。PMGE包括以下三个步骤。

（1）高分子“基因”的定义和“虚拟高分子”的设计。根据对现有化学数据的分析和领域专家的经验得出的规则，即广泛使用的理论模型和经验规则[13]，将与材料性能相关的因素，如构成高分子的化学基团和元素，定义为所谓的高分子“基因”。然后，可以通过基因组合或编辑（即调节高分子的链组成）来设计一系列“虚拟高分子”。

（2）高分子性能的高通量预测和筛选。高分子的定量结构-性能关系（quantitative structure-property relationship, QSPR）根据实验或模拟数据建立，以预测设计出的“虚拟高分子”的性能。然后根据性能要求进行计算机上的虚拟筛选以获得有前景的新型高分子。

（3）验证。对筛选出的高分子材料进行合成和表征，以验证筛选结果的可靠性并优化预测模型。此外，还可利用高精度的理论计算来验证筛选结果。进一步，基于PMGE的基因分析可用于推断潜在的物理规则，以启发未来高分子的结构设计。

性能预估是材料合理设计的关键。一种预测策略是通过数据挖掘找到可以评估材料特性的关键特征，提取可计算的关键特征作为代理，将理论计算难以准确获得的宏观性质转化为可计算的代理量，然后通过比较相应的代理量来筛选高分子材料。例如，Ramprasad等[14]使用密度泛函理论（density functional theory, DFT）能够轻松计算带隙，用来表示击穿强度和介电损耗等性质，然后使用介电常数和带隙作为筛选标准，获得了一系列有前景的全有机高分子电介质材料。

代理量有时是经验性的，但数据驱动的方法可以有效消除主观影响。例如，Zhu等[15]分析了PolyInfo数据库中400多种高分子的现有实验和计算数据。他们发现高分子的5%分解温度（T_{d5}）取决于高分子结构中最弱键的键解离能（bond dissociation energy, BDE），二者的皮尔逊相关系数接近0.7。因此，BDE可以被认为是评估高分子材

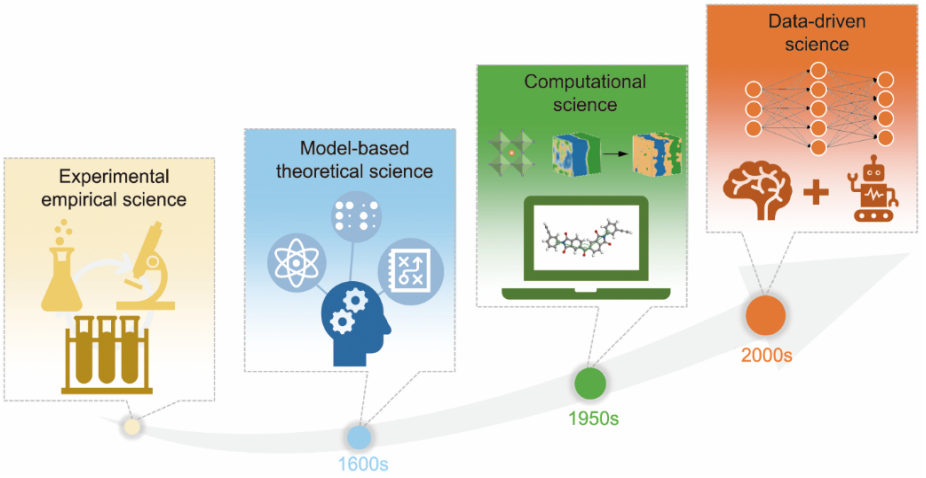

图1. 材料研究四种范式的发展：实验经验、基于模型的理论、计算模拟和数据驱动范式。第一个范式需要反复试验，导致发现材料的研究周期很长。当前，材料研究现已进入数据驱动时代，即第四范式。

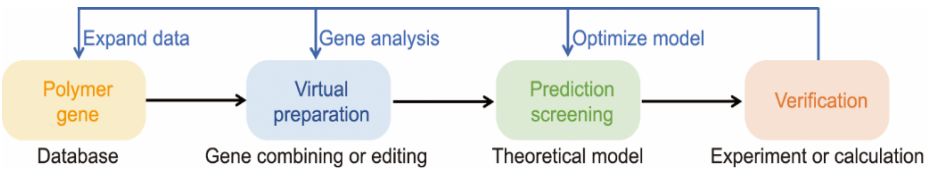

图2. 高分子材料基因工程的概念和步骤。基于数据库，定义高分子基因并设计虚拟高分子。然后，利用理论计算或高通量实验对高分子性能进行高通量预测和筛选，并通过实验或计算对筛选结果进行验证。

料热稳定性的一个关键特征。接着，他们利用高分子材料基因组来调和树脂的耐热性与低固化能之间的矛盾[15]。用通过DFT计算的带隙作为可加工性的代理量，使用所提出的代理量预测模型，进行两步筛选获得了优选的含硅芳炔[poly(silane arylacetylene), PSA]结构。最终，筛选出了一种有前景的含有2,7-二乙炔基萘的PSA结构。实验验证表明，该新型树脂的5%热分解温度为655 ℃，固化放热焓为241.9 $J \cdot g^{-1}$，表现出优异的综合性能。

此外，高分子的韧性可以用体积模量与剪切模量之比（*K/G*）来表示。计算筛选韧性代理量*K/G*和热稳定性代理量BDE后，Gao等[16]获得了一种新型乙炔封端聚酰亚胺（acetylene-terminated polyimide, ATPI），可通过共聚来增强PSA树脂的韧性。ATPI和PSA的共聚树脂，在保持耐热性的同时，韧性显著提高。如上所述，使用代理量预测模型来设计和筛选高分子材料是有效且可靠的。性能代理量预测的关键在于挖掘目标性质与微观或宏观物理参数之间的潜在关系。

机器学习（machine learning, ML）可以从历史数据中挖掘潜在规律，并对未知数据进行预测、推断或分类。这是PMGE实现高通量预测和虚拟筛选的另一种策略[17–19]。简化分子线性输入规范（simplified molecular-input line-entry system, SMILES）提供了一组简单且适合作为化学数据标签的表示方法[20]。SMILES可以将化学信息转化为计算机可接受的形式，适用于多种基于文本的机器学习算法，是一种有效的结构表示工具[21]。然后，可通过各种机器学习算法训练现有数据，以构建输入（如SMILES、分子图、分子指纹和其他分子描述符）和所需材料性能之间的QSPR模型。基于可靠实验数据训练的机器学习预测模型可以直接预测材料性能。例如，基于Polymer Genome、PubChem等开放数据库，Zhang等[17]利用多层感知机方法，建立了机器学习模型以预测目标性能（热分解温度和黏度）与高分子结构之间的定量关系（图3）。通过基因组合，他们获得了368个候选树脂进行筛选，使用两个机器学习模型高通量预测和筛选候选树脂的性能，随后获得了一系列具有优异加工性能和高耐热性的树脂。实验验证表明，筛选出的PSNP-MV树脂兼具易加工和耐高温的性能。

当实验数据有限或质量较低时，可以利用理论计算或模拟数据直接训练机器学习模型，获得的模型也可提供可靠的性能预测。例如，基于DFT计算结果，Ramprasad等[18]通过训练计算数据，建立了一个机器学习模型来预测带隙和介电常数。当计算、模拟或数据库中的某些数据保真度较低时，使用多保真代理模型可以有效地提高数据质量[22]。通过训练模型预测低保真度（如模拟数据）与高保真度（如实验）数据之间的偏差，使基于机器学习的模型能够评估它们的差异，从而提高数据质量。此外，针对数据量少的问题，可以利用各种先进的机器学习策略来避免过拟合并提高模型泛化能力，如基于物理信息的神经网络和贝叶斯方法[23–24]。利用上述策略，可以解决缺乏实验数据的问题，实现高分子的设计和筛选。

理论计算也可用于估计高分子性能并筛选目标性能，但有时可能非常耗时。机器学习模型能够克服理论计算的局限性，尤其是针对较大化学结构空间所需的昂贵计算时间成本。当高分子基因数量增加时，高分子结构空间呈指数增加，此时通过理论计算高分子性能是不切实际的，而机器学习模型可以在短时间内实现对这些高分子的性能预测。总而言之，运用机器学习模型具有预测精度高、开发周期短、适用性广等优点，这些优点使其非常适用于PMGE中的材料设计和筛选。

3. 高分子材料基因工程的挑战与展望

高分子材料基因组的研究仍处于起步阶段，还存在许多问题有待解决，如下所述。

3.1. 基因定义和分子结构描述

高分子独特的链结构和复杂的多尺度结构对高分子基因定义和结构描述提出了挑战，因此有必要开发更先进的方法来描述高分子结构特征。可进一步改进现有的方法，如BigSMILES、图表示、分子指纹等。此外，还可以引入信息学或数学的新方法。针对高分子基因定义，为了平衡结构设计的灵活性和实验合成的可行性，可根据目标高分子体系的合成路径来定义高分子基因[13,15]。在BigSMILES中，高分子片段由大括号括起来的重复单元列表表示，这使得BigSMILES成为高分子数据库系统中索引标识符的绝佳候选者[25]。另外，应对高分子基因进行系统分析、分类和标记，结合数据挖掘的规则和领域专家的经验，提高高分子基因定义和结构描述的准确性和合理性。

此外，在高分子的结构描述中应考虑多尺度特征。例如，可以通过理论计算、模拟和实验获得高分子链信息（如构象）和聚集态信息（如晶体结构和固化交联结构）。最近，Hu等[26]发展了交联密度描述符来更好地预测固化环氧树脂的性能。另外，高分子的多分散性会影响高分子的多尺度结构，进而导致高分子性能的变化，可以识别和标记多分散性数据，将其添加到高分子数据库中，在建立QSPR时作为输入之一，以开发可靠的预测模型[27]。

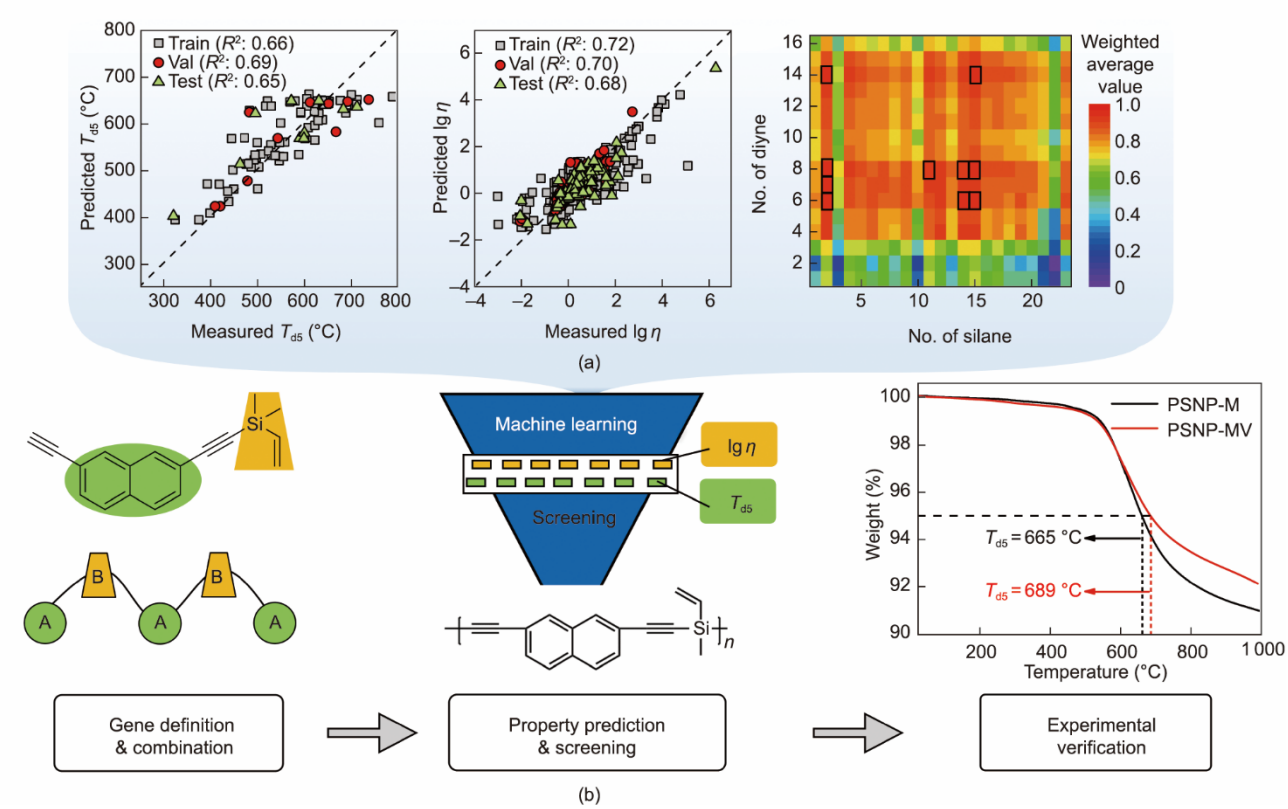

图3. 高性能树脂的机器学习预测和筛选。（a）368种候选树脂的热分解温度和黏度的机器学习模型以及综合性能的热图；（b）借助机器学习增强材料基因组方法设计具有优异综合性能的新型PSNP-MV树脂[17]。

3.2. 性能代理量预测模型

有必要寻找或建立更多代表高分子性能的关键特征，如耐溶剂性、耐磨性、抗冲击性和界面黏合性能。为实现高分子性能的快速预测和多步筛选，应开发更快速的代理计算方法，如分子连接法和基团贡献法。

3.3. 高分子性能的机器学习预测

当前面临的挑战是缺乏高质量的高分子结构-性能数据，并且高分子性能预测模型的泛化能力不强，无法精确描述多尺度结构-性能关系。以上问题都限制了机器学习预测在PMGE中的应用，可以通过自然语言处理和挖掘开放数据库中的数据来应对这些挑战[28]。随着PMGE平台的开放和共享，将有更多的研究人员积极地录入数据，还可通过高通量实验和理论模拟获得大量数据，与此同时，研究人员应注意低质量实验数据的利用，尤其是所有数据都应该标准化，以提高数据质量。

还可利用一些先进的算法开发机器学习策略来解决小数据的问题，如迁移学习、监督学习和主动学习[24,29]。引入先验算法和具有分子结构描述符的微纳米结构信息也有望建立具有物理意义的机器学习预测模型。例如，可以通过对结构的描述，考虑高分子介电性能的频率依赖机制，进而训练相应机器学习模型，这将有利于建立准确的多尺度结构-性能关系。

3.4. 高通量实验

需要建立高分子的高通量合成实验系统，用于快速筛选有前景的高分子、扩充数据库并优化预测模型。目前的实验技术是从其他领域的并行合成器发展而来的，用于高分子高通量合成和表征的设备仍有待开发。跨学科研究是解决这一问题的有效途径，这涉及信息化、系统控制、微流控技术等多学科的技术。

PMGE还应考虑高分子的合成可行性以及适合大规模制造的加工性能。除了正向预测和筛选之外，还需要拆解工程应用的性能需求，制定逆向设计的策略，以实现PMGE的双闭环设计。高分子结构的逆向设计将进一步丰富PMGE的意义，实现高分子的理性设计和智能制造。

3.5. 工程应用前景

如图4所示，PMGE可以加速高分子材料在各种工程应用中的发展，特别是当两种或多种性能彼此矛盾时。例如，PMGE可以应用于以下领域。

（1）**先进树脂基复合材料**。除高分子树脂外，PMGE还被应用于高强度、高模量高分子纤维的结构设计和性能

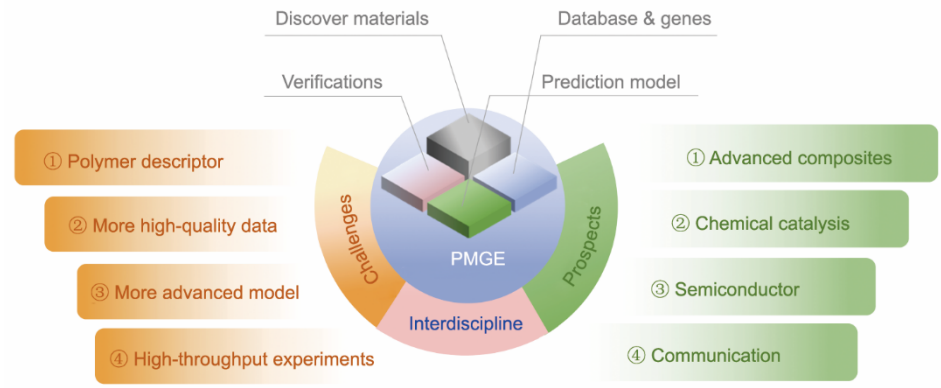

图4. PMGE的挑战与前景。高分子数据库和基因、性能预测模型以及验证等问题仍有待通过跨学科研究来共同解决。通过PMGE发现新材料，在化学工程、半导体、通信等领域具有应用潜力。

提升。PMGE还可用于调节树脂与纤维之间的界面结合性能并优化复合材料的加工性能。还可以基于复合材料有限元模拟数据和实验数据来训练机器学习模型，经过训练的机器学习模型可以快速预测和筛选最终复合材料的性能，从而实现先进复合材料的合理设计。

（2）化学工程与催化。通过机器学习增强的材料设计策略可以加速各种催化剂（包括多孔催化材料和聚合催化剂）的合理设计和筛选[30]。催化剂决定着聚烯烃的微观结构、宏观性能和工业效率，因此，催化剂的结构设计是推动聚烯烃工业发展的关键。例如，Ziegler-Natta催化剂活性位点的合理设计，以及丙烯聚合时茂金属催化剂的构型选择性预测仍然具有挑战性。数据驱动的机器学习方法为发现和设计高分子催化剂提供了一种有前景的途径。

（3）高分子有机半导体材料。此类高分子体系需要具备高电子迁移率、高发光效率、高自旋特性、高电导率等特点。采用传统的试错方法很难获得各性能都优异的高分子材料。利用PMGE设计共轭高分子可以加速具有优异综合性能的高分子有机半导体材料的研究。

（4）通信及集成电路材料。用于高频通信技术领域的高分子要求同时具备强机械性能、耐热性和电磁性能。例如，6G通信设备中使用的高分子材料应具有相对较低的介电常数和较低的介电损耗；用于芯片封装的高性能高分子应具有高耐热性、低热膨胀系数、高硬度、高韧性、高电绝缘性和低介电常数。因此，上述工程应用需要开发综合性能优异的先进高分子材料，使用PMGE无疑是最佳选择，通过高通量预测和筛选，PMGE能够开发出综合性能优异的高分子材料。

4. 总结

PMGE将推动下一代材料的创新，它有望降低材料研究成本，平衡性能限制，以实现高分子材料的突破。PMGE可以彻底变革传统的高分子设计方法，推动材料科学的研究进步。然而，由于PMGE仍处于早期阶段，许多问题还有待解决。信息、数学、控制工程等学科之间的跨学科合作可解决性能预测和实验验证等问题。未来有望实现正向预测和筛选以及逆向设计的双闭环。我们预计PMGE将成为高分子设计和应用的可持续公共平台，研究人员可以利用PMGE，对先进复合材料、半导体、通信等领域的新型高分子材料的加工、组分和性能进行合理设计。

致谢

本工作得到了国家自然科学基金项目（22103025、51833003、22173030、21975073和51621002）的支持。

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 a-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, heatresistant 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. Machinelearning- 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.