Contents lists available at ScienceDirect

Engineering

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

Research
Materials Genome Engineering—Article

数据中心设计——一种微结构材料体系设计新方法

Wei Chen [a,*], Akshay Iyer [a], Ramin Bostanabad [b]

[a] Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA
[b] Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA 92697, USA

ARTICLE INFO

Article history:
Received 10 August 2020
Revised 14 October 2020
Accepted 11 May 2021
Available online 18 February 2022

关键词
材料信息学
机器学习
微结构
重建
贝叶斯优化
混合变量建模
降维
材料设计

摘要

在高通量计算材料科学时代，材料基因组计划的核心是为计算材料设计建立数据处理、材料结构和材料属性（PSP）之间的关系。近年来，在数据获取和存储，微结构表征和重建（MCR），机器学习（ML），材料建模和仿真，数据处理、材料制造和实验方面取得的技术进步，显著提升了研究人员在PSP关系的建立和逆向材料设计方面的能力。本文将从设计研究的角度审视这些进步。特别介绍了一种数据中心设计方法，并从本质上将该方法分为三个方面：设计表征、设计评估和设计合成。每个方面的发展都由领域知识指导并从中受益。因此，针对每个方面，提出了一种应用广泛的计算方法，这些方法的集成实现了以数据为中心的材料发现和设计。

1. 引言

计算材料科学为人们可以更深入地了解不同规格的材料行为提供了一个平台。这一技术的进步对各个工业部门特别重要，可以使具有加工性能的材料实现符合成本效益的设计。材料基因组计划[1–4]以及材料设计[5–6]、微结构敏感设计[7]和集成计算材料设计[8]等工具和框架的出现也凸显了计算材料科学的重要性。因为材料的形态严重影响其属性[9–10]，所以这些框架的中心主题是逆向材料设计，其中阐明了数据处理、材料结构和材料属性（PSP）之间的关系，以便设计出独特的材料[5,11]。逆向PSP关系的非唯一性虽然提供了设计的灵活性，但却对PSP关系的未来发展提出了挑战[图1（a）]。

在20世纪，材料科学的研究和发展依赖于昂贵且耗时的Edisonian方法，该方法涉及多次尝试并出现很多错误。这种方法延缓了新兴材料在商业应用中的部署。为了实现材料设计的巨大飞跃，需要将材料研究的重点从简单地解释所观察到的现象转移到开发科学和可预测的模型，利用可控的定量因子来解释和预测材料行为，以满足工业应用的预期目标。为此，开发了所谓的高通量计算材料科学[12] [图1（b）]。本文的核心概念是创建一个用于存储材料微结构特征和性能的海量数据库。然后，将该数据库用于训练一个可以预测（或协助预测）PSP关系的机器学习（ML）模型。

PSP关系双向变化的整体设计策略依赖于解决一些关键挑战——具有经济效益的处理技术、微结构表征和重

* Corresponding author.
E-mail address: weichen@northwestern.edu (W. Chen).

英文原文：Engineering 2022, 10(3): 89–98
引用本文：Wei Chen, Akshay Iyer, Ramin Bostanabad. Data Centric Design: A New Approach to Design of Microstructural Material Systems. *Engineering*, https://doi.org/10.1016/j.eng.2021.05.022.

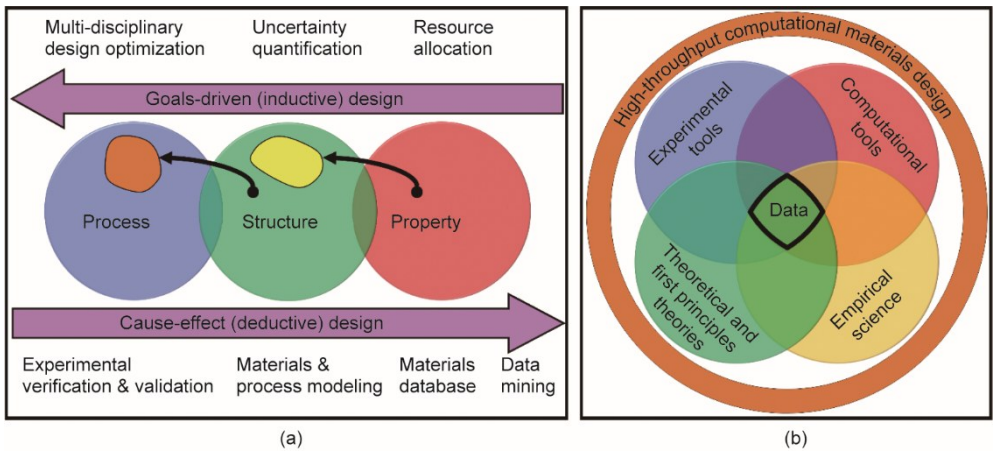

图1. （a）材料基因组计划中的正向和逆向PSP关系不是唯一的；（b）通过高通量仿真和实验进行数据驱动的材料设计。

建、降维和易于处理的优化方法。开源材料数据库[13–17]的出现以及ML技术[18]取得的最新进展，提高了使用以数据为中心的材料设计方法来应对其中一些挑战的能力（图2）。从设计研究的角度来看，这种方法可被分为设计表征、设计评估和设计合成三个方面，每个方面都以从存储在数据库中的PSP数据中获得的知识为指导。

• **设计表征。**包括描述设计中控制因素的方法，即影响材料行为的变量。这些因素取决于材料体系，因此，领域知识可以极大地帮助对材料进行识别。例如，无机化合物的带隙完全由其组成结构决定；因此，组成结构本身就是一种合适的表现形式。另一个例子是，聚合物纳米复合材料的电性能取决于其组成结构和微结构。由于这两个因素是高维的，因此必须使用谱密度函数（SDF）或物理描述符等微结构表示方法进行降维。

• **设计评估。**包括用于评估PSP关系所采用的方法论。所选择的方法在很大程度上取决于潜在现象发生所需的材料和时空尺度。例如，密度泛函理论（DFT）[19–20]计算方法可以捕获原子层级的属性，如带隙；分子动力学（MD）仿真方法能够对一组分子进行建模[21–23]；连续介质力学适用于发生在较长长度规模上的现象。每一种方法都需要校准嵌入的参数和验证预测的属性，这是通过数据库中包含的实验数据来完成的。在实验数据或仿真数据上进行训练的ML方法已被广泛用于构建替代模型，以取代基于物理的昂贵的仿真方法。

• **设计合成。**包括搜索设计空间以识别（可行的）满足目标性能的最优设计。优化方法的选择取决于设计变量的性质——是否存在定性和定量的设计变量、性能评估中是否存在不确定性或噪声，以及该方法需要的计算成本。为了考虑生产可行性以及与基本定律和已知材料行为的一致性，在优化过程中通常会施加约束和界限。

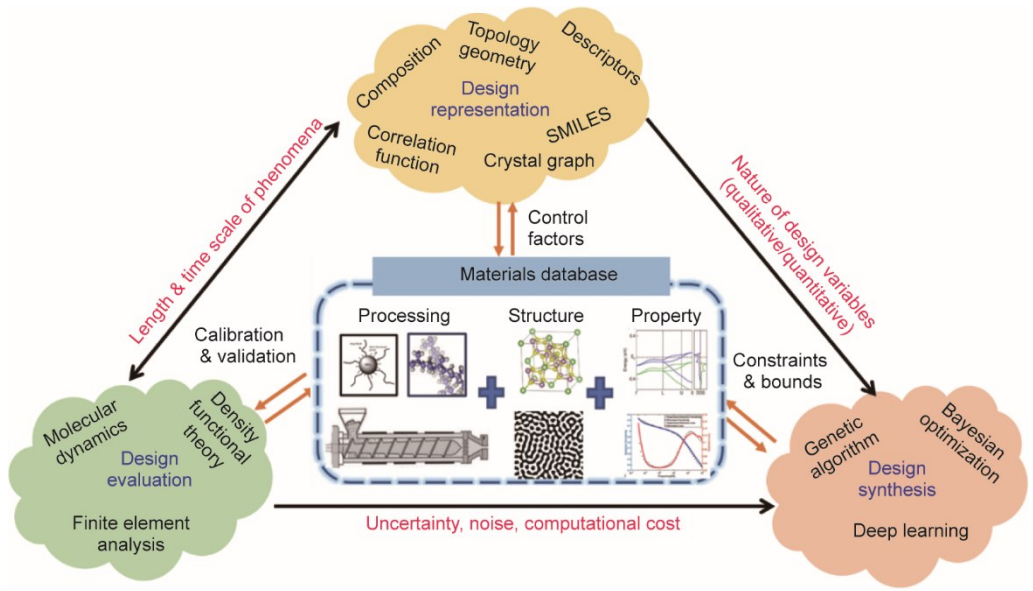

图2. 以数据为中心的材料设计框架。SMILES：简化分子线性输入规范。

值得注意的是，上述三个方面是相互关联的，如图2所示。例如，设计表征的选择——无论是混合变量（定性或定量）还是仅定量——将影响设计评估中ML的选择和设计合成中搜索算法的选择。本文首先概述了数据资源的作用，然后回顾了这三个方面各自面临的挑战和最先进的方法。

2. 材料数据资源

近年来，人们为构建大数据资源做出了很多努力以加速探索和设计材料。此类数据资源大多数都集中在金属材料体系和计算材料数据中，其中软件预测工具可以通过快速扫描组成空间来预测所研究的特定结构和属性。在最近发表的一篇文献[24]中可以找到这些数据资源的相关示例。我们的研究团队一直参与开发一种数据资源，即NanoMine，用于聚合物纳米复合材料领域的软材料设计[13–14,25]（图3）。NanoMine具有内置的数据管理、探索、可视化和分析功能，包含来自文献和各个实验室的2500多个样本的管理数据。原则上，NanoMine提供了一个可查找、可访问、可相互操作和可重复使用（FAIR）的平台，通过简单的搜索工具可以直接查找和访问论文中发布的数据，并且在开放元数据标准下可与更大型的材料数据注册表进行交互操作；此外，该平台还可以轻松地重复使用数据，例如，针对新结果进行基准测试。

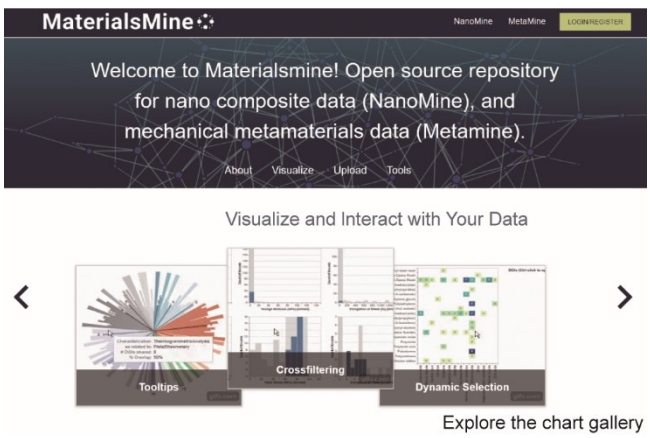

图3. NanoMine——聚合物纳米复合材料的在线数据资源（www.materialsmine.org）。

开发材料数据资源的核心在于针对所研究的领域专门创建一种数据模式。用于构建NanoMine元数据框架的材料词汇表是高级“聚合物数据核心”[26]的一部分，并且与储存于其他地方，如材料数据设施（MDF）的数据索引兼容[27–28]。在MDF的基础上，我们开发了一个基于本体的知识图谱框架[14]，帮助NanoMine建立以下6类数据之间的关系：

- **数据资源**。该类别中的数据来源于都柏林核心标准指导下的文献的元数据，包括被引来源的数字对象标识符（DOI）、作者、标题、关键词、时间和出版物来源。
- **材料**。该类别中的数据涉及材料的组成信息，包括填料颗粒、聚合物基体和表面处理。可以将纯基体和填料的性能，如聚合物化学结构、分子量和颗粒密度与结构（即体积/质量分数）一起输入。
- **数据处理**。该类别中的数据描述了化学合成和实验程序的顺序。目前的模板提供了三大类：溶液加工、熔融混合和原位聚合。对于每个处理步骤，可以输入温度、压力和时间等详细信息。
- **表征**。该类别中的数据提供有关材料表征设备、方法和使用条件的信息。这些信息包括有关常见显微成像（扫描电子显微镜、透射电子显微镜）、热力学和电测量以及纳米级光谱学的详细信息。
- **属性**。该类别中的数据是材料性能的测量数据，包括力学、电学、热学和体积性能。属性数据可以通过标量的形式或更高维度的形式，如二维（2D）光谱或三维（3D）图表示。
- **微结构**。该类别中的数据包括捕获纳米相分散状态的原始显微灰度图像，也包括几何描述符，用于描述微结构的统计特征。

NanoMine本体作为材料科学可扩展的知识表示平台，可以与我们为跨领域搜索、可视化和数据共享开发的工具一起使用，并与现有的科学元数据标准进行相互操作。除了物理数据外，一组模块化工具[用于微结构表征和重建（MCR）以及模拟体纳米复合材料响应的仿真软件]使实验产生的知识增加。整合这些不同的数据源以创建新知识，对于材料设计至关重要。然而，利用由成分、微结构形态和加工条件组成的无限组合所定义的广阔设计空间来生成实验或仿真数据是不切实际的。因此需要以数据为中心的方法，这些方法可以有效地查询现有数据，并在数据之间进行插值，以支持设计表征、设计评估和设计合成，以及发现新的高性能材料。

3. 设计表征——微结构的特征和重建

由于材料微结构的高维性，在微结构介导的设计中，微结构表征对于确保设计策略易于处理至关重要。良好的微结构表征可以显著降维；明显体现形态特征；具有物理意义，可以很容易地映射到加工条件中；提供计算效率高

的重建程序，以便创建统计学意义上相同的微结构，进而评估结构-性能关系，并量化与材料异质性相关的不确定性。

MCR与ML、材料建模和仿真相结合，是高通量计算材料科学时代探索PSP关系和逆向材料设计的重要组成部分。鉴于在工程材料中观察到多种多样的微结构，开发一种普遍适用的MCR技术是具有挑战性的。综述文章[29]对各种MCR技术进行了全面的综述，并详细说明了算法细节、计算成本以及它们如何适应PSP映射问题。其中，有兴趣的读者可以找到依赖于统计函数（如n点相关函数）、物理描述符、SDF、纹理合成和监督/非监督学习的多类MCR方法的详细描述。

应用于非均质微结构的MCR技术如图4所示。最著名的MCR方法是基于空间相关函数的[图4（b）] [30–31]，该方法提供了形态的概率表征，但依赖密集计算模拟退火（SA）算法进行重建。基于描述符的方法[图4（a）] [32–33]是使用一组能体现明显微结构细节但又不相关的描述符来表示微结构。重建涉及分层优化策略，将重建的微结构的描述符与目标值进行匹配。然而，常规几何特征的使用和椭球聚类假设阻碍了其在具有不规则几何形状的微结构中的应用。其他基于描述符的MCR版本已经在文献中进行报道，描述符的选择因材料体系而异，还取决于所研究的属性。最近邻算法的描述符在颗粒异质体系的传输过程[9]、在结晶中的微结构演变过程[34]、在颗粒粗化[35]和液相烧结[34]中起着重要作用。在纤维复合材料中，纤维的体积分数（VF）、尺寸、形状和空间分布会影响复合材料的力学性能，如杨氏模量、极限强度和断裂韧性[36–42]。在晶体结构中，晶间腐蚀对晶界很敏感[43]，因此必须将这些晶界用作准确设计表征的描述符。

ML和人工智能（AI）技术凭借从各向同性/各向异性

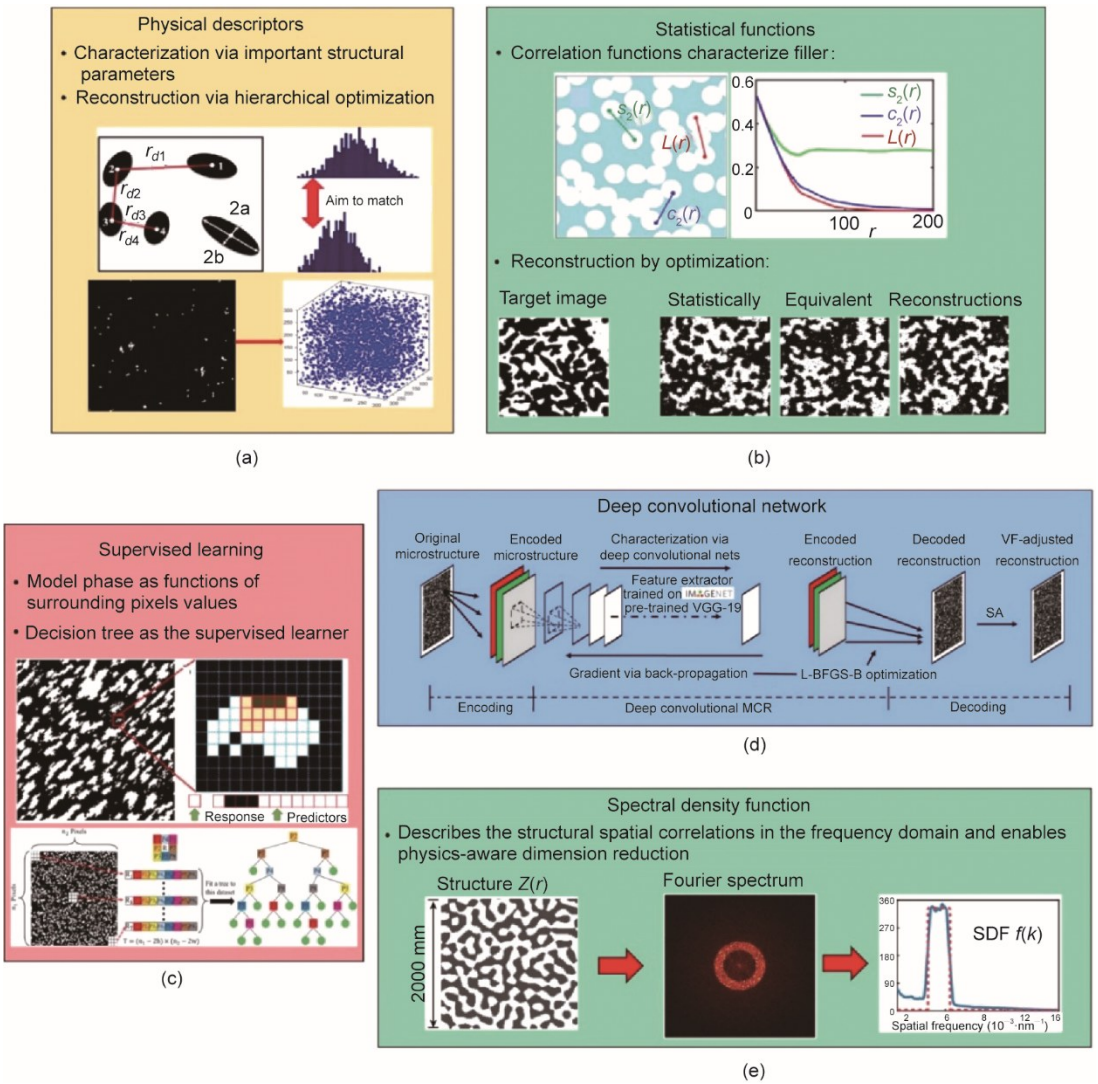

图4. 典型的MCR技术。（a）物理描述符；（b）统计函数；（c）监督学习；（d）深度卷积网络；（e）SDF。L-BFGS-B：有限内存拟牛顿代码，用于界限约束的优化；VGG-19：视觉几何组-19，一个19层深的卷积神经网络（CNN），在ImageNet数据库的100多万张图像上进行训练。

微结构中学习和重建复杂特征的卓越能力，成为一种广受欢迎的重建工具。基于实例进行学习的应用在处理复杂材料形态时已经显示出良好的重建准确性，这一应用使用支持向量机[44]、监督学习[图4（c）] [45–46]和迁移学习[图4（d）] [18,47]。特别是基于迁移学习的方法，通过利用预训练的深度卷积神经网络（CNN）、视觉几何组-19（VGG-19）[48]和损失函数（用于测量原始微结构与重建微结构之间的差异），仅对一个给定的目标微结构重建统计学上等效的微结构。然后利用在进行模型修剪过程中获得的信息来开发结构-属性预测模型，以确定网络架构和初始化条件。虽然基于深度学习的方法对于处理复杂的微结构形态非常有用，但这些方法通常不能提供微结构表征的物理意义，因此阻碍了它们在材料设计中的使用。诸如卷积深度信念网络[49]和生成对抗网络（GAN）[50]等深度学习方法被用于现有研究，以提供可用作设计变量的低维微结构表征。

SDF [图4（e）] [9,51–55]是一种频域微结构表示，能够提供具有复杂形态的准随机材料体系的低维和物理意义描述，因此受到广泛关注。对于各向同性材料，SDF是空间频率的一维（1D）函数，在频域上表示空间相关性。虽然SDF中包含的信息等同于两点自相关函数，但Yu等[51]已经表明，SDF提供了一种能够简单且明智地映射处理条件和属性的表示方法。然而，使用现有方法重建高分辨率的3D微结构[56–58]，其计算成本和时间仍然有很大挑战。此外，虽然现有的SDF技术仅限于各向同性材料体系，但在某些材料体系中对各向异性材料有很大的需求，特别是对于在潜在传输情况下表现出的性能，如有机光伏电池（OPVC）、电池、热电器件和用于水过滤的膜。在最近的工作[59]（图5）中，开发了一种各向异性微结构设计策略，通过被称为各向异性指数的无量纲标量变量，利用SDF以2D和3D方式快速重建高分辨率、两相、各向同性或各向异性微结构。应用于体异质结OPVC的有源层设计案例研究表明，具有较强各向异性的优化设计优于各向同性有源层设计。物理感知SDF方法还为理解PSP关系的设计评估提供了显著的降维。

4. 设计评估——PSP关系的机器学习

在基于物理的材料设计中，ML技术在很大程度上替代了昂贵的PSP仿真器。近年来，ML技术和AI技术在分子和聚合物体系[60]、金属体系[61–62]材料设计中的应用得到了广泛的研究。如图6所示，虽然有大量的统计模型，如神经网络（NN）、随机森林（RF）、树和高斯过程（GP）[63]可用于创建替代模型，但是特征识别在获取有

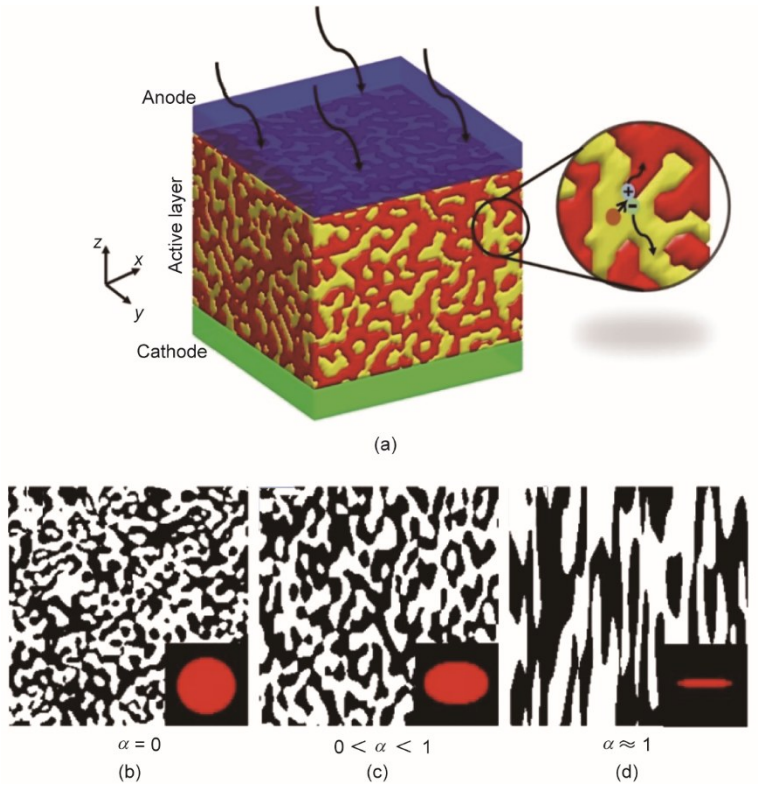

图5. （a）OPVC示意图。插图显示了激子（橙色）解离成质子（蓝色）和电子（绿色），分别游离到阳极和阴极。（b）~（d）使用各向异性指数α量化具有椭圆SDF的微结构的各向异性。

良好预测能力的可信赖统计模型方面发挥着关键作用。

“维度诅咒”（curse of dimensionality，即大量描述符或参数）使得构建具有中等样本数据规模的预测模型极具挑战性。因此，通过将这些ML方法与材料科学领域知识相融合，采用特征选择和特征提取相结合的方法来降维。一般来说，特征选择的目标有三个：提高预测性能、提供更具成本效益的预测器，以及促进发现数据生成的潜在概率原则[64]。变量排名是最常用的特征选择技术之一，可以识别信息量最大的特征以构建简约的预测模型。我们的研究团队开发了一系列用于微结构特征选择的技术。例如，徐等[65]采用两步特征选择过程，使用描述符成对相关性分析（仅基于图像的无监督学习）和回归浮雕（RReliefF）变量排序方法[66]（基于结构-属性关系的监督学习）选择最能控制聚合物复合材料阻尼性能的物理描述符。探索性因素分析[67]是另一种识别重要特征的技术，该技术通过将相关描述符组合在一起来构建一组潜在的共同因素。研究人员在结构方程建模方法中采用因子分析来设计介电聚合物复合材料[39]。简而言之，通过特征选择，可以删除冗余的统计特征，然后进行进一步的分析。

与特征选择不同，特征提取是将特征空间转换为减少物理解释的低维空间。虽然没有像特征选择方法那样保留尽可能多的物理解释，但特征提取技术有利于降低空间的维数，并且更易于训练以实现更高的预测准确性[68–69]。主成分分析（PCA）可能是最著名的线性降维方法，该方法可以将3D微结构图像的高维特征空间转换为低维近似值[70]。研究人员已经证明，PCA可以有效地将两点相关函数（常用于微观结构表征）的维数减少到几个参数[71–73]。近年来，由于ML技术的进步，用于材料设计中特征提取的非线性嵌入方法得到了广泛应用。第一种是自下而上的方法，其中假设非线性流形（嵌入在原始特征空间中）控制数据分布[74–75]；第二种是自上而下的方法，该方法试图在所有尺度上保留几何关系[76]。

可以选择使用广泛的ML技术来构建一个包含以下多种因素的统计模型，如①物理行为的性质（非线性和不规则性）；②输入变量的类型（定性、定量或混合的）；③研究的响应（连续或分类的）；④数据源（噪声实验、确定性仿真或随机仿真）；⑤数据量（大量或少量数据）。由于需要了解PSP映射中的因果关系，因此通常使用监督学习方法。虽然线性回归是应用和解释结果最直接的方法，但决策树[77]、k最近邻算法（k-NN）[78]、支持向量机[79–80]和RF [81]等方法更适合用于更复杂的行为和混合变量输入的情况，而且这些方法还可被用于灵活地创建回归和分类模型。

随着材料大数据变得越来越容易获得，近来ML和材料设计接口方面的研究呈指数级增长。神经网络由多层人工神经元连接，用于模仿人脑。单个神经元通过所谓的激活函数输出加权输入。深度神经网络（DNN）是一种特殊的神经网络，具有多个隐藏层和卓越的学习能力。对于无机材料，晶体图CNN [82]已被用于仿真高度非线性行为[使用从开放量子材料数据库（OQMD）中提取的DFT计算的热力学稳定性条目]，以加速材料的发现[83]。研究证明，对于纳米复合材料，虽然CNN提供了微结构重建和结构属性学习的能力[47]，但可以通过训练GAN来学习潜在变量（LV）和微结构之间的映射[50]。此后，将低维潜在变量作为设计变量，采用贝叶斯优化（BO）框架以获得具有所需材料性能的微结构。对于有机材料，简化分子线性输入规范（SMILES）[84]为大分子提供了有意义的表示，并可用于使用变分自动编码器[85]和强化学习[86]来设计合成分子。

对于存在少量数据的情况，尤其是来自确定性仿真（如DFT）的数据，并且这些数据需要数小时甚至数天来

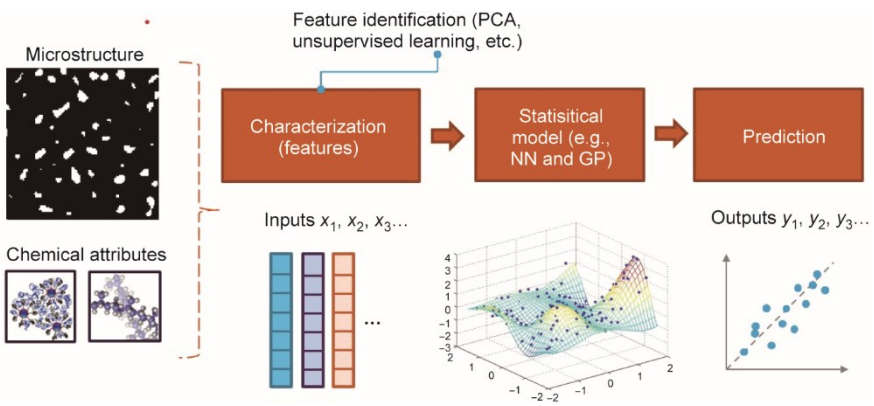

图6. 材料设计中的特征识别和ML。PCA：主成分分析。

计算一个材料设计，GP提供了一种非常可行的方法。图7是一个GP模型的1D示例，该模型拟合了$f(\cdot)$的收集数据。在每个输入x处，输出$f(x)$被视为一个正态分布的随机变量，GP模型预测其均值和方差。图中95%的预测区间反映了预测的置信区间[87–88]。

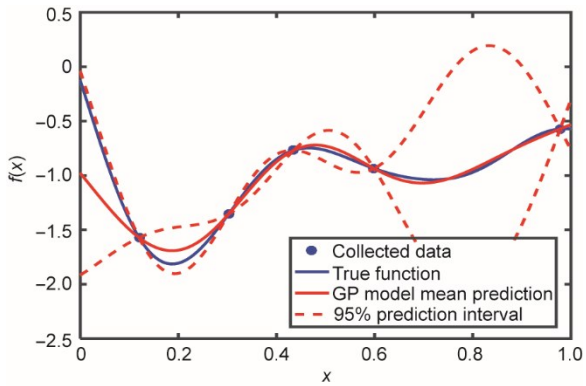

图7. 拟合$f(\cdot)$收集数据的GP模型对应的1D示例。

标准GP方法是在所有输入变量都是定量的前提下开发的，这在包含表示材料成分、微结构形态和加工条件的定性和定量设计变量的材料体系中并不成立。我们最近提出了一种潜在变量的高斯过程（LVGP）[89]建模方法，该方法将定性因子水平映射到一些潜在的不可观察的定量变量的一组数值中。换句话说，定性变量被“转换”为定量变量，然后可以应用传统的GP建模来获得所需的模型。定性因子的潜在变量映射为因子水平提供了固有的顺序和结构，从而可以深入了解定性因子的影响。与大多数监督ML方法不同，LVGP不需要手工制作的特征来描述定性变量。相反，LVGP通过最大化似然函数来学习影响响应（y）的潜在变量（Z）。

对特征工程需求的减少使得LVGP在材料设计应用中具有吸引力。如图8所示，M_2AX相家族中原子M的三个定性水平$t_j \in \{l_1, l_2, l_3\}$与潜在高维空间中的点$\{v_1, v_2, \ldots\}$相关，这个空间由原子半径、电离能和电子亲和力等物理参数定义。LVGP提供了从v到潜在空间Z的非线性流形映射$z(t) = g(v_1(t), v_2(t), \ldots)$，三个点之间的距离表明了三个水平对相关属性影响的差异。混合变量LVGP方法已在广泛的微结构体系（如用于优化准随机太阳能电池光吸收的并行材料选择和微结构优化）中得到测试和验证[90]。材料的组合搜索构成了最优的混合有机-无机钙钛矿设计[90]，以及纳米介电材料的并行组成和微结构设计[91]。材料发现和优化是通过将LVGP方法与BO集成来实现设计合成的，这将在下面进行介绍。

5. 设计合成——目标导向的贝叶斯优化

发现新材料通常需要耗费数年甚至数十年的时间，这与设计合成相关的几个挑战有关：①即使可使用大型数据集，但是已知材料的性能仍远未达到预期目标。使用现有数据创建的ML模型无法预测“外推”区域中的行为。②存在大量候选的设计组合。在有机材料的设计中，如在聚合物纳米复合材料设计中，材料成分（如填料和基体的类型）和加工条件（如表面处理的类型）的选择很多；每种组合都遵循截然不同的物理机制，这对整体性能将产生重大影响。在微电子等无机材料的设计中，有数百万个原子结构-组成变量空间可供选择；这是由不同的结构原型（晶体图）、组成（化学元素的选择）和化学计量（元素比值）导致的。③材料设计的定量和定性变量的存在导致在属性/性能空间中存在多个脱节区域。这种组合性质对材料建模和寻找最优解决方案提出了额外的挑战。

在过去的5年中，BO方法已成为最有效的材料设计合成方法[92–95]，该方法能够从数十到数百个目标函数（即材料性能）评估中找到高度线性函数的全局最优值。从一个小数据集开始，BO依靠自适应采样技术有效地接近全局最优值——这一特征在材料设计中具有吸引力。图9显示了本研究提出的按需目标驱动的数据增强框架，该框架将精心设计的材料数据库与材料性能仿真和ML集合在一起。该框架由数据库中实验数据和仿真数据启动，这些数据准确地描绘了材料性能。基于PSP关系，可以识

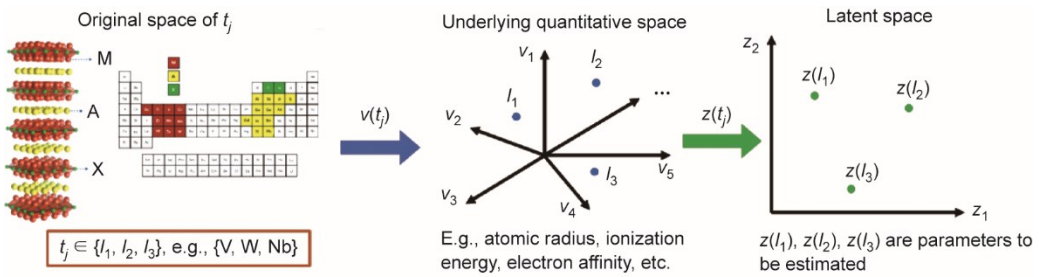

图8. 定性材料成分选择，使用从真正的高维潜在定量变量映射到2D潜在变量的Z进行建模。

别出一个已知会影响材料性能的属性子集，并在BO中充当设计变量。这些属性可以是定量的（如微观结构描述符或相间描述符）或定性的（如填料类型、聚合物或两者的组合）。

使用ML模型的预测结果和不确定性量化结果，贝叶斯推理确定了在材料性能方面显示出最大潜力的设计。有几个指标（通常称为采集函数）可用于评估潜在改进。采集函数在设计空间的探索（减少预测不确定性）与开发（优化设计目标）之间取得了平衡。最常用的采集函数是预期改进（EI）[96]和概率改进[97]。一旦采集函数确定了一个有前景的设计，就会根据“按需”实验、仿真或两者兼而有之对其材料性能进行评估。仿真的性质取决于所考虑的材料体系和性能，通常需要校准参数。例如，用于预测纳米复合材料介电性能的有限元素仿真需要校准相移参数[98]。性能评估完成后，就会将此设计添加到数据库中并重复上述步骤。终止标准通常是最大迭代次数，取决于仿真或实验所需的成本和时间。

在第4节，通过将混合变量LVGP模型和BO框架整合在一起，本研究成功地将BO方法应用于有机、无机和混合材料的设计。例如，在并行组合和微结构设计中[91]，电绝缘纳米复合材料的设计是一个多准则优化问题，其设计目标是在最大化介电击穿强度的同时将介电常数和介电损耗降到最小（图10）。选择SDF作为微结构表征，并根据实验图像识别底层函数类型。在数十次仿真中使用多响应LVGP方法，确定了帕累托边界（Pareto frontier）上的一组不同设计，表明介电性能之间的权衡。这种方法已被证明比使用遗传算法更有效。

通过对具有最佳溶剂结合能的ABX_3混合有机-无机钙钛矿进行组合搜索[90]，使用LVGP的BO的普遍性得到进一步验证。设计空间由A位点和X位点的各三种不同的组合以及8种溶剂类型组成，而B位点保持不变。此外，三个X可以独立选择。在648种可能的ABX_3溶剂组合中，有240种是稳定的，因此构成了BO的搜索空间。图11（a）显示，与迄今为止常用于定性变量的乘积的协方差

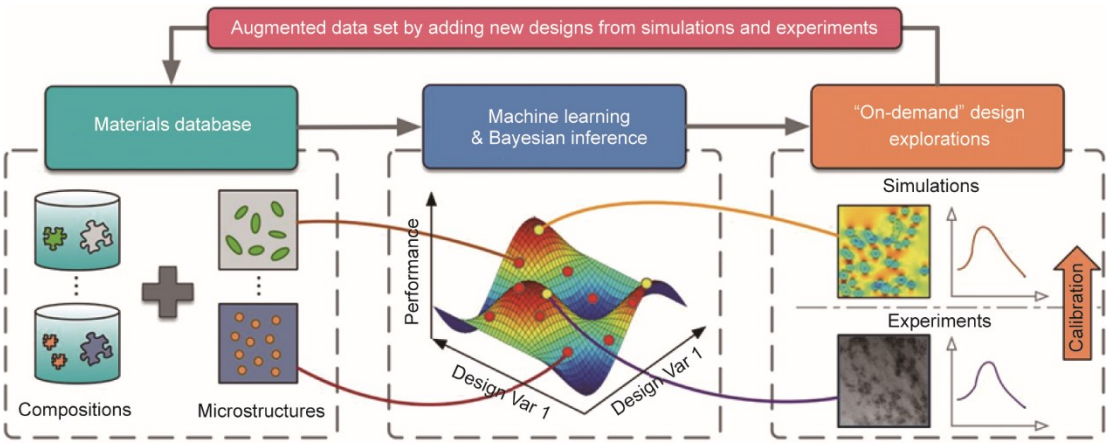

图9. BO方法将现有数据集当作先验知识，选择新样本，并使用设计的新的实验和仿真数据来构建ML模型，进而捕获PSP关系并进行优化。

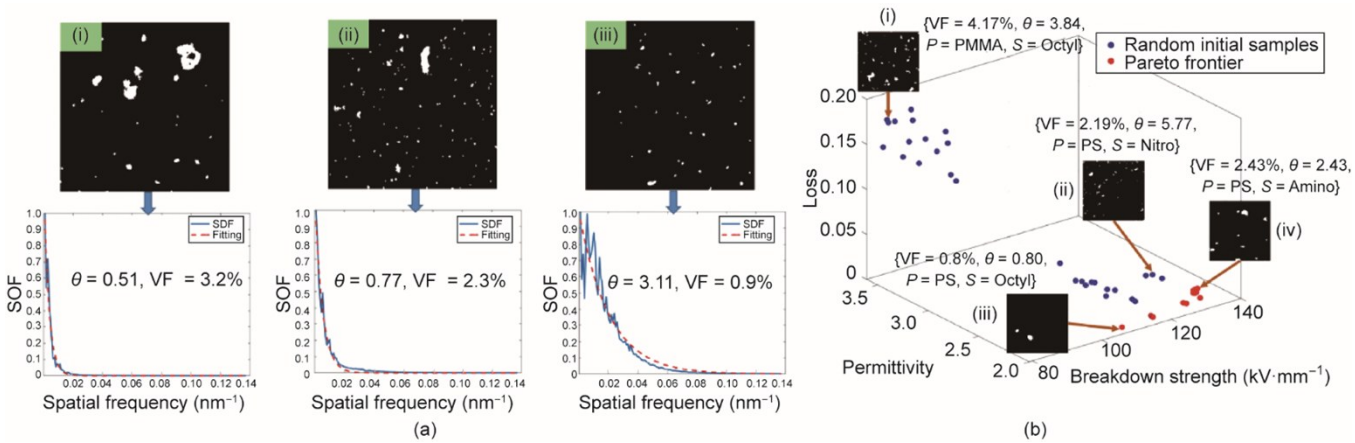

图10. 纳米复合材料的并行组合和微结构设计。（a）SDF使用参数θ表征纳米颗粒的分散，使用VF表征纳米颗粒的负载。（b）由使用LVGP的多准则混合变量BO确定帕累托边界，相对于随机选择的初始化样本有显著改善（*P*代表聚合物类型；*S*代表表面处理类型。PMMA：聚甲基丙烯酸甲酯；PS：聚苯乙烯）。

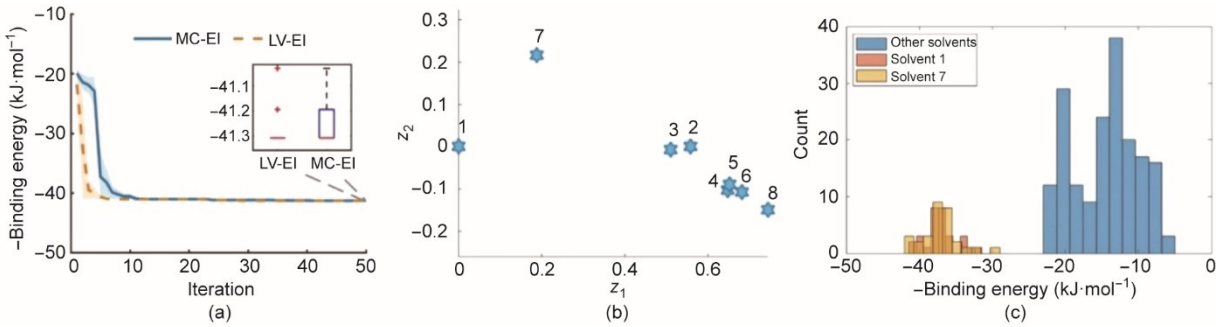

图11.（a）比较MC-EI和LV-EI GP的EI采集函数与BO收敛；（b）溶剂类型分类变量的潜在空间，具有8个级别；（c）按溶剂类型分类的结合能分布。

（MC）[99–100] GP模型相比，BO与LVGP能更快地收敛出最优组合。此外，LVGP估计的潜在空间为每个定性变量的水平性质提供了深入的见解。图11（b）中溶剂1和溶剂7与其他溶剂的位置相距甚远，表明它们对结合能的影响是不同的。通过分析图11（c）中的结合能可解释这一现象，表明与溶剂1和溶剂7的组合会产生更高的结合能。可以将几种材料设计应用程序作为一个组合优化问题。例如，最近的研究表明，使用基于LVGP的多准则BO可以更快速地搜索具有金属-绝缘体转变（MIT）特性[101]的功能性电子材料设计。这些发现表明，将混合变量LVGP模型与BO集成是工程材料体系设计中一种有效的设计合成方法。

6. 结论

本文提出了一种以数据为中心的材料设计方法，该方法集成了用于微结构分析和设计的最先进计算技术。这些技术涉及设计表征、设计评估和设计合成。这些方法的实现需要材料数据中心（如NanoMine）的支持。NanoMine涵盖了广泛的数据资源和工具，用于微结构分析和优化材料设计。正如本文所阐述的，该方法包括图像预处理、微结构表征、重建、降维、PSP关系的ML和多目标优化的系统集成。

要实现设计表征、设计评估和设计合成的无缝集成，存在一个关键的问题：对于所研究的材料体系来说，什么是合适的微结构表征？本文提出了一系列基于相关函数、物理描述符、SDF、监督学习和深度学习的微结构表征技术。虽然不同技术的优点因材料体系而异，但是随机性起着关键作用，必须在材料表示和性能预测中对此加以考虑。

对于设计评估而言，ML方法在知识发现和构建替代基于物理仿真模型方面发挥着越来越重要的作用。由于数据泛滥和数据缺乏在材料信息学中并存，因此必须注意确保所选的ML技术（如NN、RF或GP）与可用数据保持一致。随着材料数据生成的越来越多，深度学习在基于图像的材料信息学中越来越受欢迎，其中对学习的微结构特征的解释依赖于开发可解释的深度模型。

最后，ML不应被视为材料发现中的一个孤立组成部分。例如，可以将ML与BO等信息论方法结合起来显著提高材料发现的速度。由于材料的发现在本质上不是孤立的，因此需要可以处理定性和定量设计变量的LVGP等混合变量模型。这些模型基于对所需材料性能的影响，为不同材料概念提供了“距离”的定量测量。我们还需要更进一步的研究来扩展目前的方法，用于处理具有数百万或数十亿种组合的高维材料设计问题。相同的信息论框架可用于指导批量样本和高通量实验的设计。

Acknowledgements

The authors gratefully acknowledge support from the National Science Foundation (NSF) Cyberinfrastructure for Sustained Scientific Innovation program (OAC-1835782), the NSF Designing Materials to Revolutionize and Engineer Our Future program (CMMI-1729743), Center for Hierarchical Materials Design (NIST70NANB19H005) at Northwestern University, and the Advanced Research Projects Agency-Energy (APAR-E, DE-AR0001209). Collaborations from Drs. Daniel Apley, Catherine Brinson, and Linda Schadler and their students on the presented methods and materials design case studies are greatly appreciated.

Compliance with ethics guidelines

Wei Chen, Akshay Iyer, and Ramin Bostanabad declare

that they have no conflict of interest or financial conflicts to disclose.

References

[1] National Science and Technology Council (US). Materials genome initiative for global competitiveness [Internet]. Washington DC: Executive Office of the President, National Science and Technology Council; 2011 Jun 24. Available from: https://www.mgi.gov/sites/default/files/documents/materials_genome_initiative-final.pdf.

[2] Olson GB. Preface to the viewpoint set on: the materials genome. Scr Mater 2014;70:1–2.

[3] Ward C. Materials Genome Initiative for global competitiveness. In: Proceedings of the 23rd Advanced Aerospace Materials and Processes (AeroMat) Conference and Exposition; 2012 Jun 18–21; Charlotte, NC, USA; 2012.

[4] McDowell DL, Kalidindi SR. The materials innovation ecosystem: a key enabler for the materials genome initiative. MRS Bull 2016;41(4):326–37.

[5] Olson GB. Computational design of hierarchically structured materials. Science 1997;277(5330):1237–42.

[6] Olson GB. Designing a new material world. Science 2000;288(5468):993–8.

[7] Fullwood DT, Niezgoda SR, Adams BL, Kalidindi SR. Microstructure sensitive design for performance optimization. Prog Mater Sci 2010;55(6):477–562.

[8] Committee on Integrated Computational Materials Engineering. Integrated computational materials engineering: a transformational discipline for improved competitiveness and national security. Washington, DC: National Academies Press; 2008.

[9] Torquato S. Random heterogeneous materials: microstructure and macroscopic properties. New York: Springer-Verlag New York; 2002.

[10] Kumar H, Briant CL, Curtin WA. Using microstructure reconstruction to model mechanical behavior in complex microstructures. Mech Mater 2006;38(8–10):818–32.

[11] 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.

[12] Curtarolo S, Hart GLW, Nardelli MB, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. Nat Mater 2013;12(3):191–201.

[13] 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.

[14] Zhao H, Wang Y, Lin A, Hu B, Yan R, McCusker J, et al. NanoMine schema: an extensible data representation for polymer nanocomposites. APL Mater 2018;6(11):111108.

[15] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002.

[16] Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM 2013;65(11):1501–9.

[17] Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor RH, et al. AFLOWLIB. ORG: a distributed materials properties repository from high-throughput *ab initio* calculations. Comput Mater Sci 2012;58:227–35.

[18] Bostanabad R. Reconstruction of 3D microstructures from 2D images via transfer learning. Comput Aided Des 2020;128:102906.

[19] Koch W, Holthausen MC. A chemist's guide to density functional theory. 2nded. Medford: John Wiley & Sons; 2015.

[20] Parr RG. Density functional theory of atoms and molecules. In: Fukui K, Pullman A, editors. Horizons of quantum chemistry. Dordrecht: Springer; 1980. p. 5–15.

[21] Duan K, He Y, Li Y, Liu J, Zhang J, Hu Y, et al. Machine-learning assisted coarse-grained model for epoxies over wide ranges of temperatures and cross-linking degrees. Mater Des 2019;183:108130.

[22] Bejagam KK, Singh S, An Y, Deshmukh SA. Machine-learned coarse-grained models. J Phys Chem Lett 2018;9(16):4667–72.

[23] Wang W, Gómez-Bombarelli R. Coarse-graining auto-encoders for molecular dynamics. npj Comput Mater 2019;5(1):1–9.

[24] Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci 2019;6(21):1900808.

[25] Brinson LC, Deagen M, Chen W, McCusker J, McGuinness DL, Schadler LS, et al. Polymer nanocomposite data: curation, frameworks, access, and potential for discovery and design. ACS Macro Lett 2020;9(8):1086–94.

[26] Therneau T, Atkinson B, Ripley B. rpart: recursive partitioning and regression trees. Version 4.1-10 [software]. 2019 May 1. Available from: https://rdrr.io/cran/rpart/.

[27] Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I. The materials data facility: data services to advance materials science research. JOM 2016;68(8):2045–52.

[28] Blaiszik B, Ward L, Schwarting M, Gaff J, Chard R, Pike D, et al. A data ecosystem to support machine learning in materials science. MRS Commun 2019;9(4):1125–33.

[29] Bostanabad R, Zhang Y, Li X, Kearney T, Brinson LC, Apley DW, et al. Computational microstructure characterization and reconstruction: review of the state-of-the-art techniques. Prog Mater Sci 2018;95:1–41.

[30] Yeong CLY, Torquato S. Reconstructing random media. Phys Rev E 1998;57(1):495–506.

[31] Yeong CLY, Torquato S. Reconstructing random media. II. Three-dimensional media from two-dimensional cuts. Phys Rev E 1998;58(1):224–33.

[32] Xu H, Dikin DA, Burkhart C, Chen W. Descriptor-based methodology for statistical characterization and 3D reconstruction of microstructural materials. Comput Mater Sci 2014;85:206–16.

[33] Xu H, Li Y, Brinson C, Chen W. A descriptor-based design methodology for developing heterogeneous microstructural materials system. J Mech Des 2014;136(5):051007.

[34] Snyder VA, Alkemper J, Voorhees PW. The development of spatial correlations during Ostwald ripening: a test of theory. Acta Mater 2000;48(10):2689–701.

[35] DeHoff RT. A geometrically general-theory of diffusion controlled coarsening. Acta Metall Mater 1991;39(10):2349–60.

[36] Li M, Ghosh S, Richmond O, Weiland H, Rouns TN. Three dimensional characterization and modeling of particle reinforced metal matrix composites: part I: quantitative description of microstructural morphology. Mater Sci Eng A 1999;265(1–2):153–73.

[37] Nan CW, Clarke DR. The influence of particle size and particle fracture on the elastic/plastic deformation of metal matrix composites. Acta Mater 1996;44(9):3801–11.

[38] Breneman CM, Brinson LC, Schadler LS, Natarajan B, Krein M, Wu K, et al. Stalking the materials genome: a data-driven approach to the virtual design of nanostructured polymers. Adv Funct Mater 2013;23(46):5746–52.

[39] Zhang Y, Zhao H, Hassinger I, Brinson LC, Schadler LS, Chen W. Microstructure reconstruction and structural equation modeling for computational design of nanodielectrics. Integr Mater Manuf Innov 2015;4:209–34.

[40] Karásek L, Sumita M. Characterization of dispersion state of filler and polymer-filler interactions in rubber–carbon black composites. J Mater Sci 1996;31:281–9.

[41] Yuan M, Turng LS. Microstructure and mechanical properties of microcellular injection molded polyamide-6 nanocomposites. Polymer 2005;46(18):7273–92.

[42] Baghgar M, Barnes AM, Pentzer E, Wise AJ, Hammer BAG, Emrick T, et al. Morphology-dependent electronic properties in cross-linked (P3HT-*b*-P3MT) block copolymer nanostructures. ACS Nano 2014;8(8):8344–9.

[43] Rollett AD, Lee SB, Campman R, Rohrer GS. Three-dimensional characterization of microstructure by electron back-scatter diffraction. Annu Rev Mater Res 2007;37:627–58.

[44] Sundararaghavan V, Zabaras N. Classification and reconstruction of three-dimensional microstructures using support vector machines. Comput Mater Sci 2005;32(2):223–39.

[45] Bostanabad R, Bui AT, Xie W, Apley DW, Chen W. Stochastic microstructure characterization and reconstruction via supervised learning. Acta Mater 2016;103:89–102.

[46] Bostanabad R, Chen W, Apley DW. Characterization and reconstruction of 3D stochastic microstructures via supervised learning. J Microsc 2016;264(3):282–97.

[47] Li X, Zhang Y, Zhao H, Burkhart C, Brinson LC, Chen W. A transfer learning approach for microstructure reconstruction and structure–property predictions. Sci Rep 2018;8(1):13461.

[48] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.

[49] Cang R, Xu Y, Chen S, Liu Y, Jiao Y, Ren MY. Microstructure representation and reconstruction of heterogeneous materials via deep belief network for computational material design. J Mech Des 2017;139(7):071404.

[50] Yang Z, Li X, Catherine Brinson L, Choudhary AN, Chen W, Agrawal A. Microstructural materials design via deep adversarial learning methodology. J Mech Des 2018;140(11):111416.

[51] Yu S, Zhang Y, Wang C, Lee WK, Dong B, Odom TW, et al. Characterization and design of functional quasi-random nanostructured materials using spectral density function. J Mech Des 2017;139(7):071401.

[52] Uche OU, Stillinger FH, Torquato S. Constraints on collective density variables: two dimensions. Phys Rev E 2004;70(4):046122.

[53] Uche OU, Torquato S, Stillinger FH. Collective coordinate control of density distributions. Phys Rev E 2006;74(3):031104.

[54] Batten RD, Stillinger FH, Torquato S. Classical disordered ground states: super-ideal gases and stealth and equi-luminous materials. J Appl Phys 2008;104(3): 033504.

[55] Florescu M, Torquato S, Steinhardt PJ. Designer disordered materials with large, complete photonic band gaps. Proc Natl Acad Sci 2009;106(49):20658–63.

[56] Cahn JW. Phase separation by spinodal decomposition in isotropic systems. J Chem Phys 1965;42(1):93–9.

[57] Teubner M. Level surfaces of Gaussian random fields and microemulsions. EPL 1991;14(5):403–8.

[58] Chen D, Torquato S. Designing disordered hyperuniform two-phase materials with novel physical properties. Acta Mater 2018;142:152–61.

[59] Iyer A, Dulal R, Zhang Y, Ghumman UF, Chien T, Balasubramanian G, et al. Designing anisotropic microstructures with spectral density function. Comput Mater Sci 2020;179:109559.

[60] Chen G, Shen Z, Iyer A, Ghumman UF, Tang S, Bi J, et al. Machine-learning-assisted *de novo* design of organic molecules and polymers: opportunities and challenges. Polymers 2020;12(1):163.

[61] Johnson NS, Vulimiri PS, To AC, Zhang X, Brice CA, Kappes BB, et al. Invited review: machine learning for materials developments in metals additive manufacturing. Addit Manuf 2020;36:101641.

[62] Bock FE, Aydin RC, Cyron CJ, Huber N, Kalidindi SR, Klusemann B. A review of the application of machine learning and data mining approaches in continuum materials mechanics. Front Mater 2019;6:110.

[63] Bostanabad R, Chan YC, Wang LW, Zhu P, Chen W. Globally approximate Gaussian processes for big data with application to data-driven metamaterials design. J Mech Des 2019;141(11):111402.

[64] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[65] Xu H, Liu R, Choudhary A, Chen W. A machine learning-based design representation method for designing heterogeneous microstructures. J Mech Des 2015;137(5):051403.

[66] Robnik-Šikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression. In: Proceedings of the Fourteenth International Conference on Machine Learning; 1997 Jul 8–12; Nashville, TN, USA. San Francisco: Morgan Kaufmann Publishers, Inc.; 1997. p. 296–304.

[67] Fabrigar LR, Wegener DT, MacCallum RC, Strahan EJ. Evaluating the use of exploratory factor analysis in psychological research. Psychol Methods 1999;4 (3):272–99.

[68] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science 2000;290(5500):2323–6.

[69] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science 2000;290(5500):2319–23.

[70] Jolliffe IT. Principal component analysis. In: Everitt BS, Howell D, editors. Encyclopedia of statistics in behavioral science. Hoboken: John Wiley & Sons, Inc.; 2005.

[71] Yabansu YC, Steinmetz P, Hötzer J, Kalidindi SR, Nestler B. Extraction of reduced-order process-structure linkages from phase-field simulations. Acta Mater 2017;124:182–94.

[72] Popova E, Rodgers TM, Gong X, Cecen A, Madison JD, Kalidindi SR. Process-structure linkages using a data science approach: application to simulated additive manufacturing data. Integr Mater Manuf Innov 2017;6(1):54–68.

[73] Paulson NH, Priddy MW, McDowell DL, Kalidindi SR. Reduced-order structure–property linkages for polycrystalline microstructures based on 2-point statistics. Acta Mater 2017;129:428–38.

[74] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Dietterich T, Becker S, Ghahramani Z, editors. Advances in neural information processing systems. Cambridge: MIT Press; 2001. p. 585–91.

[75] Donoho DL, Grimes C. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. Proc Natl Acad Sci 2003;100(10):5591–6.

[76] Saxena A, Gupta A, Mukerjee A. Non-linear dimensionality reduction by locally linear isomaps. In: Proceedings of the 11th International Conference on Neural Information Processing; 2004 Nov 22–25; Calcutta, India. Berlin: Springer; 2004. p. 1038–43.

[77] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. London: Taylor & Francis Group; 1984.

[78] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory 1967;13(1):21–7.

[79] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.

[80] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory; 1992 Jul 27–29; Pittsburgh, PA, USA; 1992; p. 144–152.

[81] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[82] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018; 120(14):145301.

[83] Park CW, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. Phys Rev Mater 2020;4(6):063801.

[84] 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.

[85] Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018;4(2):268–76.

[86] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for *de novo* drug design. Sci Adv 2018;4(7):aap7885.

[87] Tao S, Shintani K, Bostanabad R, Chan YC, Yang G, Meingast H, et al. Enhanced Gaussian process metamodeling and collaborative optimization for vehicle suspension design optimization. In: Proceedings of the ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference; 2017 Aug 6–9; Cleveland, OH, USA; 2017.

[88] Bostanabad R, Kearney T, Tao S, Apley DW, Chen W. Leveraging the nugget parameter for efficient Gaussian process modeling. Int J Numer Methods Eng 2018;114(5):501–16.

[89] Zhang Y, Tao S, Chen W, Apley DW. A latent variable approach to Gaussian process modeling with qualitative and quantitative factors. Technometrics 2020; 62(3):291–302.

[90] Zhang Y, Apley DW, Chen W. Bayesian optimization for materials design with mixed quantitative and qualitative variables. Sci Rep 2020;10(1):4924.

[91] Iyer A, Zhang Y, Prasad A, Tao S, Wang Y, Schadler L, et al. Data centric mixed variable Bayesian optimization for materials design. In: Proceedings of the ASME International Design Engineering Technical Conference; 2019 Aug 18–21; Anaheim, CA, USA; 2019.

[92] Balachandran PV, Xue D, Theiler J, Hogden J, Lookman T. Adaptive strategies for materials design using uncertainties. Sci Rep 2016;6(1):19660.

[93] Li C, de Celis Leal DR, Rana S, Gupta S, Sutti A, Greenhill S, et al. Rapid Bayesian optimisation for synthesis of short polymer fiber materials. Sci Rep 2017;7(1):5683.

[94] Yamashita T, Sato N, Kino H, Miyake T, Tsuda K, Oguchi T. Crystal structure prediction accelerated by Bayesian optimization. Phys Rev Mater 2018; 2(1): 013803.

[95] Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. npj Comput Mater 2019;5(1):21.

[96] Mockus J, Tiesis V, Zilinskas A. The application of Bayesian methods for seeking the extremum. In: Dixon LCW, Szego GP, editors. Towards global optimization. Amsterdam: Elsevier; 1978. p. 117–29.

[97] Kushner HJ. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. J Basic Eng 1964;86(1):97–106.

[98] Wang Y, Zhang Y, Zhao H, Li X, Huang Y, Schadler LS, et al. Identifying interphase properties in polymer nanocomposites using adaptive optimization. Compos Sci Technol 2018;162:146–55.

[99] Zhang Y, Notz WI. Computer experiments with qualitative and quantitative variables: a review and reexamination. Qual Eng 2015;27(1):2–13.

[100] McMillan NJ, Sacks J, Welch WJ, Gao F. Analysis of protein activity data by Gaussian stochastic process models. J Biopharm Stat 1999;9(1):145–60.

[101] Wang Y, Iyer A, Chen W, Rondinelli JM. Featureless adaptive optimization accelerates functional electronic materials design. Appl Phys Rev 2020;7:041403.