第 54 卷第 1 期
2026 年 1 月

硅 酸 盐 学 报
JOURNAL OF THE CHINESE CERAMIC SOCIETY

Vol. 54，No. 1
January，2026

http://jccsoc.cbpt.cnki.net

DOI：10.14062/j.issn.0454-5648.20250867

材料领域知识嵌入的分而治之机器学习

杨正伟[1]，刘子图[1]，张王颖[2]，左 维[1]，余振垚[1]，邹欣欣[1]，袁浩栋[1]，苏玉凤[1]，施思齐[2,3]，刘 悦[1]

(1. 上海大学核电关键材料全国重点实验室&计算机工程与科学学院，上海 200444；2. 上海大学核电关键材料全国重点实验室&材料科学与工程学院，上海 200444；3. 上海大学材料基因组工程研究院，上海 200444)

摘 要：纯数据驱动的机器学习能够快速挖掘复杂材料数据隐含的科学规律。然而，“成分–结构–工艺–性能”组成的复杂材料构效关系使得机器学习难以准确剖析其内在/外在因素的作用机理，这导致模型精度、建模复杂程度及其可解释性面临严峻挑战。因此，本工作提出了一种材料领域知识嵌入的分而治之机器学习建模策略，旨在依据领域知识将复杂材料问题进行有效拆解并利用领域知识嵌入的机器学习对拆解后子问题进行针对性的自适应建模，以精准且具有一定可解释性的方式挖掘其内部隐藏作用机理。同时，利用变分推断验证该策略的科学性与有效性。进一步，总结了在材料科学领域中依据不同领域知识进行“分而治之”机器学习建模的应用示例。最后，展望了该策略未来所面临的挑战及对策。

关键词：材料科学；机器学习；分而治之；领域知识
中图分类号：TP181；TB3　　**文献标志码**：A　　**文章编号**：0454–5648(2026)01–0014–11
网络出版时间：2026–01–07

目前，以机器学习为主导的数据驱动研发范式因其能够从数据中提取、总结材料的“成分–结构–工艺–性能”构效关系或科学机理，被视为从根本上解决新材料研发效率、工程应用与成本控制方面核心难题的革命性突破口[1–3]，并已在材料构效关系预测、新材料发现及过程优化等方面取得显著进展[4–6]。

然而，材料构效关系具有高度复杂性。材料的微观结构会因不同类型的键合与价态之间的相互作用表现出多样而复杂的行为特征[7]，且可能触发结构重构或相变，从而致使材料内禀机制转变，产生积极促进/消极抑制效果。此外，材料性能并非由单一因素决定，而是成分、结构、制备工艺及环境条件等多影响因素协同作用的结果[8]，材料体系可能会出现多机理并存甚至竞争状态，呈现出具有强非线性耦合特征的内禀关系。由此可看出，材料构效关系的影响因素复杂多元，造成材料样本数据的分布规律高度复杂，其中所蕴含的反应机理不唯一，使得传统试图建立单一描述符集合与目标性能间映射的机器学习范式难以精准且全面刻画材料构效关系。

同时，通过分析材料数据特点和机器学习建模原理，本团队厘清了当前数据驱动的机器学习应用于材料领域面临的 3 大矛盾，即高维度与小样本数据的矛盾、模型准确性与易用性的矛盾、模型学习结果与领域专家知识的矛盾[9–10]。为有效缓解上述矛盾，本团队提出了领域知识嵌入的机器学习框架及系列新方法，将不同形式、来源的领域知识进行多方式表达，并有机融入机器学习 3 要素(即模型、算法、策略)[11]中，从而通过数据和知识的双向驱动来构建高精度且具有一定可解释性的机器学习新模型[12–14]。

进一步，通过分析材料研究问题特点以及材料数据蕴含的复杂多样构效关系特性，发现或可通过材料领域知识主导将复杂材料问题进行有效拆解并逐一解决，即将“分而治之”思想引入并与机器学习 3 要素紧密结合，以实现对复杂材料问题的建模。因此，本工作建立了一种材料领域知识指导的“分而治之”的机器学习建模策略，以此有效降低建模复杂

收稿日期：2025–11–26。　修订日期：2025–12–22。
基金项目：国家自然科学基金(92270124, 92472207, 52073169 和 52472223)。
第一作者：杨正伟(1994—)，男，博士研究生。
通信作者：刘 悦(1975—)，女，博士，教授；
施思齐(1978—)，男，博士，教授。

Received date: 2025–11–26.　**Revised date:** 2025–12–22.
First author: YANG Zhengwei (1994–), male, Doctoral candidate.
E-mail: zhw_yann@shu.edu.cn
Correspondent author: LIU Yue (1975–), female, Ph.D., Professor.
SHI Siqi (1978–), male, Ph.D., Professor.
E-mail: yueliu@shu.edu.cn; sqshi@shu.edu.cn

程度，并通过领域知识的嵌入实现模型学习结果的精准、可信，是调和上述 3 大矛盾的有效途径。

本工作旨在阐明构建材料领域知识嵌入的分而治之机器学习建模新策略的必要性、可行性和潜在探索方向。首先，阐述了分而治之机器学习建模策略全流程，并尝试通过变分推断对“分而治之”建模思想进行科学性验证；其次，总结了针对不同材料体系开展的一系列分而治之建模策略应用示范；最后，展望了材料分而治之机器学习建模的未来发展。

1　分而治之机器学习建模策略

分而治之(divide and conquer)可以看作分治法作用于机器学习建模，也可以看作一种新的集成学习策略，其基本思路如图 1 所示。该策略包含以下部分：1)“分”(问题拆解)，即在材料领域知识的引导下，通过解析材料物理化学机制，将复杂的材料问题依据多角度拆解为若干子问题，从而识别不同分布的数据子空间；2)“治”(问题求解)，即针对若干子问题(空间)分别进行自适应建模，同时将领域特有的知识注入到机器学习子模型中，精准建立不同材料物理化学机制与目标属性的映射；3)“合”(问题求解)，即设计决策融合机制整合子模型的提取特征/决策结果，进而得到对复杂材料问题更加全面、合理的多角度分析结果，以此帮助领域专家探究材料内禀机理演化过程。

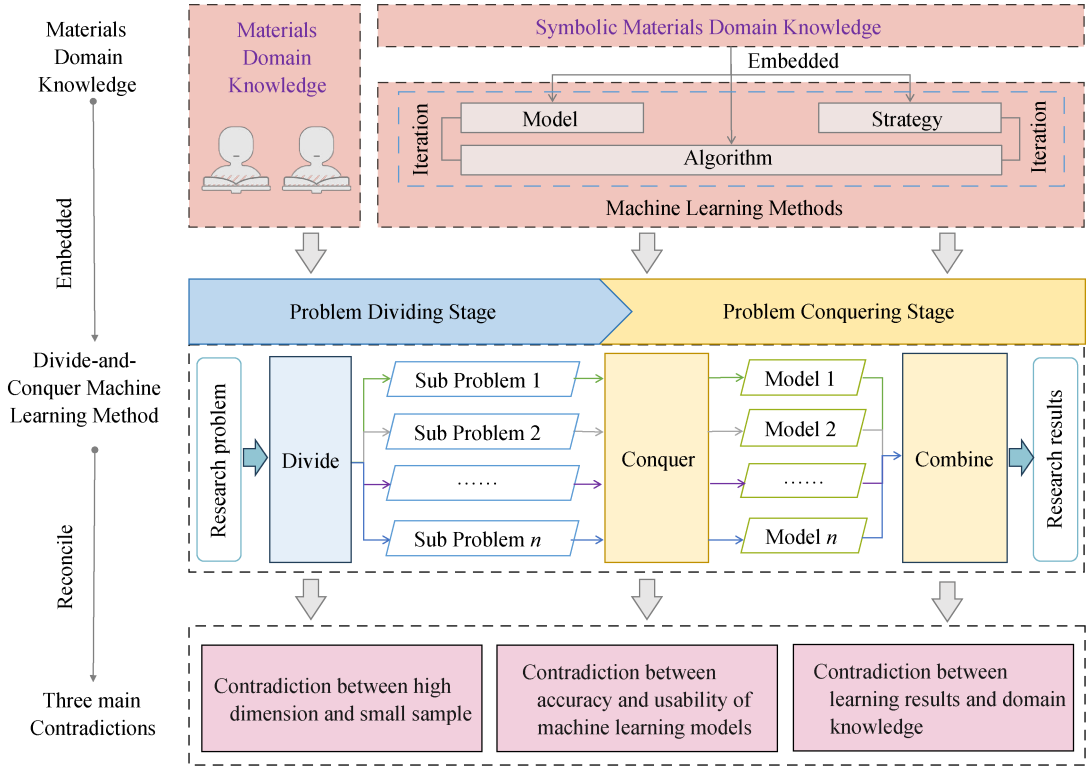


图 1　材料领域知识嵌入的分而治之机器学习

Fig. 1　Divide-and-conquer machine learning embedded with materials domain knowledge

1.1　领域知识引导下的问题拆解阶段

直接针对具有高度复杂构效关系的材料体系构建模型易使其复杂程度呈指数级上升[15]。在问题拆解(problem dividing, *PD*)阶段，基于材料领域知识的物理化学解析，识别表征不同性能影响机制的样本/特征子空间，进而将一个宏观、复杂的材料问题解构为一系列可建模且符合物化机理的子问题，进而降低机器学习建模难度。其本质并非是对复杂问题的被动分解，而是材料科学知识体系在方法论上的能动的体现。

本节提出的 *PD* 可形式化表示为一个三元组，如式(1)所示：

$$PD = \text{optimal} < D, B, T > \tag{1}$$

式中：D 是从数据分布特征维度出发，刻画数据集或特征空间的统计特性、结构异质性以及潜在噪声分布，有助于识别不同数据/特征子集之间的差异；B 是从材料物理化学行为维度出发，反映材料在不同结构、成分、工艺及环境条件下表现出的性能及其驱动机制差异；T 是从目标任务维度出发，表示材料研究目标或学习子任务的差异性。

具体地，结合领域知识对现有机器学习适用的数据集(machine-learning-oriented dataset)、材料体系及其研究任务进行系统而深入的剖析，以明确材料数据的同质性与异质性特征，全面认知材料体系的物化机制复杂性和不同物化规律之间的差异，以及推测潜在的材料目标任务性能驱动机制等。随后，依托前期分析结果，确定拆解方式，即从材料数据、材料行为以及目标任务维度中选择最能反映材料体系关键特征、物化机制差异及性能驱动规律的分析角度。以材料数据角度为例，将整体数据集按照这些维度进行系统划分，生成 n 个子集，每个子集承载特定类型的信息描述(如成分、结构、缺陷、工艺或性能特征等)。每个子集相对独立，既能突出其内部的同质性，又能保留与其他子集的逻辑关联。通过这种方式分解，可为后续针对性建模提供明确的输入，使每个子集能够应用最合适的特征表示、数据处理和建模方法，从而提升模型的预测准确性、可解释性和计算效率。

1.2 领域知识嵌入的问题求解阶段

在对子问题进行求解时(即机器学习建模阶段)，研究人员依据各子空间的数据特性或凭借自身对材料潜在“构效关系”的理解，为各子空间选择候选机器学习模型。依据各子空间内涵的材料特性，从材料领域知识中抽取相关理论或经验准则，并依据数据在机器学习建模全流程中的作用机理，将其进行针对性表示(如目标函数或约束条件)，以满足指导机器学习精准拟合数据的需求，从而约束模型学习到与物理规律一致的内在映射关系[9]。接着，采用多种评估指标用于量化模型学习能力，并以此选择最佳模型及其最优超参数组合，为子空间找到最适宜的构效关系分析模型。随后，融合领域知识与集成建模思想，设计决策融合机制，将各子模型的预测结果重构为整体问题的综合解答，确保不同角度的模型分析结果能够被系统地整合，从而形成对复杂材料问题的精准且全面的理解。

2 分而治之机器学习建模理论分析

尽管“分而治之”的理念已被广泛应用[16–17]，但均未尝试对其进行理论推导验证。本节将通过变分推断对“分而治之”建模这一思想进行符号化表示并进行合理性与科学性验证。具体为：

令 $X=\{(x_i,y_i)\}_{i=1}^{n}$，表示一个未知分布 P 对应的独立同分布数据集 X，其中 $x_i\in\mathbb{R}$，$y_i\in Y$，Y 对应真实标签或模拟的真实值。“分而治之”的目标是：学习一个函数 F，以最小化其期望风险 $R(F)=E_{(x,y)\sim P}[l(F(x),y)]$，其中 $l(\cdot)$ 为损失函数。由于分布 P 未知，通常根据最小化经验误差寻找模型的最优解：

$$\hat{R}(F)=\frac{1}{n}\sum_{i=1}^{n}l(F(x_i),y_i) \tag{2}$$

针对数据集 X 对应的函数 F 可以将其分解为有限个不同的子函数，表示为：

$$F(X)=f_1(M_1,\theta_1)+f_2(M_2,\theta_2)+\cdots+f_s(M_s,\theta_s) \tag{3}$$

因此，数据集被有限特征集 $T=\{t_1,t_2,\cdots,t_j\}$ 表示，j 为特征数量。将目标函数 $F(X)$ 分解为有限个子函数 $f(x)$，分别对应不同的描述集合 $M=\{M_1,M_2,\cdots,M_s\}$，其中不同描述 M_s 包含数据中有限个样本，表示为 $M_s=\{x_1,x_2,\cdots,x_l\}$，不同描述满足 $X=M_1\cup M_2\cup\cdots\cup M_s$。对于目标任务存在模型最优解，其对应的函数 $G(X)$ 满足：

$$G(X)\approx F(X)+o(\eta,\lambda) \tag{4}$$

其中：$F(X)$ 表示目标任务完成相应的任务的函数；f 表示不同描述 M 中的模型；$o(\eta,\lambda)$ 表示由不同描述和泛化误差等累加的非模型预测经验误差的其他误差。

对于数据满足：

$$p(X)=\int p(X,Z)\mathrm{d}z=\sum_{i=1}^{s}\int p(z_i,M_i)\mathrm{d}z \tag{5}$$

其中：Z 表示数据 X 对应的空间信息；M_i 表示不同的描述；z_i 表示不同描述下对应的局部空间信息；$p(z_i,M_i)$ 表示不同描述与局部空间信息构成的分布。不同描述模型给出的分布表示为 $q(z_i\mid M_i)$。

$$\ln p(X)=\sum_{i=1}^{s}\ln\left(\int p(z_i,M_i)\mathrm{d}z\right)\geqslant \tag{6}$$

$$\sum_{i=1}^{s}E_{q(z_i|M_i)}\left[\ln\frac{p(z_i,M_i)}{q(z_i|M_i)}\right]= \tag{7}$$

$$L_{\mathrm{ELBO}}(X)=\sum_{i=1}^{s}L_{\mathrm{ELBO}}(M_i) \tag{8}$$

因此，为保证其变分下界，进一步推导其对应的分布信息：

$$\sum_{i=1}^{s}E_{q(z_i|M_i)}\left[\ln\frac{p(z_i,M_i)}{q(z_i|M_i)}\right]= \tag{9}$$

$$\sum_{i=1}^{s}E_{q(z_i|M_i)}\left[\ln\frac{p(z_i\mid M_i).p(M_i)}{q(z_i|M_i)}\right] \tag{10}$$

$$=\sum_{i=1}^{s}E_{q(z_i|M_i)}\left[\ln\frac{p(z_i\mid M_i)}{q(z_i|M_i)}\right]+E_{(z_i|M_i)}\ln p(M_i) \quad (11)$$

$$=\sum_{i=1}^{s}-D_{\mathrm{KL}}(q(z_i\mid M_i)\|\,p(z_i\mid M_i))+\ln p(M_i) \quad (12)$$

其中 D_{KL} 表示 KL 散度(Kullback-Leibler Divergence)，用于衡量 $q(z_i\mid M_i)$ 和 $p(z_i\mid M_i)$ 两分布之间的差异，分而治之给出的证据下界可以表示为：

$$p(X)=\sum_{i=1}^{s}-D_{\mathrm{KL}}(q(z_i\mid M_i)\|\,p(z_i\mid M_i))+\ln p(M_i) \quad (13)$$

数据集 X 包含不同的样本 $X=\{x_1,x_2,\cdots,x_n\}$，并且数据集被有限特征集 $T=\{t_1,t_2,\cdots,t_j\}$ 表示，j 表示特征数量，特征表示为随机变量。进一步，给定一个描述 $M_s=\{x_1,x_2,\cdots,x_l\}$，其中包含不同样本，对于函数 $f(x)$ 满足：

$$G(M_s)\approx f(M_s)+o(\omega_s) \quad (14)$$

其中：$o(\omega)$ 表示函数拟合过程产生的误差。

$$F(D)\approx f(M_1)+f(M_2)+\cdots+f(M_s)+o(\omega) \quad (15)$$

该过程产生的函数和最优函数之间的关系满足：

$$G(D)\approx F(D)+o(\omega) \quad (16)$$

受限于描述 $f(M_s)$ 对应的概率分布与原始数据的概率分布之间的差异，使得 $D_{\mathrm{KL}}(q(z_i\mid M_i)\|\,p(z_i|M_i))$ 发生变化，影响预测结果。受限于局部模型 $f(M_s)$ 无法完全拟合最优模型对应的分布，而这种差异来源于：

$$D_{\mathrm{KL}}(q(z_i\mid M_i)\|p(z_i|M_i))\geqslant\varepsilon>0 \quad (17)$$

其中：$D_{\mathrm{KL}}(\bullet\|\bullet)$ 用于衡量分布之间的差异，ε 是一个不可避免的误差。数据分布差异，直接导致了局部模型对后验的估计可能发生偏差。具体表现为，局部模型的分布 $q(z_i\mid M_i)$ 与(在全局数据下才能得到的)真实后验 $p(z_i|M_i)$ 之间的 KL 散度增大。

$$\Delta D_{\mathrm{KL}}=D_{\mathrm{KL}}(q(z_i\mid M_i)\|p(z_i|M_i))-D_{\mathrm{KL}}(q^*(z_i)\|p(z_i|M_i)) \quad (18)$$

其中：$q^*(z_i)$ 表示在数据中可以得到的最优分布。由于存在误差导致 $\Delta D_{\mathrm{KL}}>0$，进而导致最终预测产生一定偏差，损害模型的预测能力。

3　分而治之机器学习建模应用示例

材料复杂特性导致机器学习模型难以对数据中的隐含机理进行精准拟合。目前，研究人员开始尝试依据领域知识拆解材料问题或数据，并以此衍生出不同类型的“分而治之”建模策略。我们将其归结为样本空间“分治”与特征表示“分治”(如表 1 所示)，具体解释如下。

表 1　材料领域知识嵌入分而治之机器学习的多样性分类

Table 1　Diversity classification of divide-and-conquer machine learning embedded with materials domain knowledge

Perspective	Type	Key issue	Involved technique	Example
Sample space	Stage	Turning point identification	Curve recognition method, Structural analysis method, …	Complex materials problems with multi-stage turning characteristics[18]
	Pos./Neg. examples	Pos./Neg. example division	Threshold partition, Weighting, Clustering, …	Electrolyte formulation optimization and design
	Category	Category division	Rule-based method, Clustering, Expert construction, …	Creep fracture life prediction of Ni-based single-crystal superalloys[19]
Feature representation	Attribute	Core feature identification / Feature cluster determination	Rough set theory …	NASICON-type solid-state electrolyte activation energy prediction[20]
	Modality	Modality fusion	Explicit fusion: Early/ Late fusion; Implicit fusion: Attention-based, GAN-based, …	Prediction of ionic conductivity, band gap, *etc.*, of inorganic solid-state electrolytes[21]

3.1　样本空间视角下的分而治之

样本空间“分治”旨在通过领域知识分析数据特点并按照某一材料特性或机理对样本空间进行解构(分)，进而开展针对性机器学习建模(治)。其中，阶段分而治之侧重于从时间/过程的角度“分”和“治”，正反例分而治之侧重于从性能优劣的角度“分”和“治”，类别分而治之侧重于从异质机制的角度“分”和“治”，这 3 个视角相互独立，能够协同解决材料

数据中存在的非线性、多目标冲突及异质性难题。

1) 阶段分而治之。材料的成分比例、内部复杂反应及外部环境影响变化可能引起其结构剧变，如触发重构或相变等，这种微观结构的骤变导致材料性能在时间或反应维度上呈现出显著的非平稳多阶段演化特征。针对该现象，需将长周期、高复杂的演化过程，解构为若干个物理化学机制相对单一且稳定的子阶段，从而“化繁为简”。其核心在于，建立基于临界特征的自适应划分机制，如基于领域知识的划分方法(如结构分析法)、基于统计学的划分方法(如变点检测、梯度突变分析)，以及基于机器模型拟合的划分方法(如曲线识别法)，以精准识别演化曲线中的关键拐点与机制切换边界；进而针对不同子阶段构建适配的机器学习子模型，并自适应加权融合，有效避免单一模型对不同物化机制的强制“平均化”或过平滑处理，从而实现对材料全生命周期或全反应路径下非线性行为的高保真表征。

以 Ge_xSe_{1-x} 玻璃结构为例，当锗含量增加时，其结构会从链状相向网状相转变，导致其属性呈现出显著的分阶段变化特征。结合这一特点，Liu 等[18]提出了一种两阶段支持向量机回归模型(FSTS-SVR)，实现了对其玻璃化转变温度(T_g)的精准预测。具体而言，FSTS-SVR 首先通过皮尔逊相关系数(Pearson correlation coefficient，PCC)从材料体系候选特征中筛选出与材料属性高度相关的特征，并结合曲线识别法与结构分析法，精准识别了属性变化过程中的转折点，将其划分为 2 个物理阶段。然后，采用支持向量回归分别构建各阶段预测模型，并通过最小误差模型将两者整合，从而实现对 T_g 的预测。与拓扑学方法及其他基于 SVR 的方法相比，FSTS-SVR 实现最高的 T_g 预测精度，展现出显著优势。

此外，这种“阶段分而治之”的建模思想也在其他研究中被广泛采用并得到验证。例如，在锂离子电池中，正极材料在深度充放电时易发生晶格畸变，颗粒内部裂纹使锂离子传输受阻，从而导致电池容量突然快速衰退，出现“跳水点”，该转折点前后是性能差异显著的 2 个阶段。结合这一特点，Greenbank 等[22]提出了基于自动特征选择的分段线性建模方法，通过在退化曲线的多阶段分别训练独立的子网络，并设计自适应权重机制，取得了容量预测误差仅为 1.1%，寿命终点预测误差为 1.6%的优异表现。相较于复杂深度模型，该方法在快速性、可解释性及内存效率上更具优势。

此类“阶段分而治之”机器学习建模策略适用于具有多阶段转折特征的复杂材料问题，为揭示材料因其成分比例、内部复杂反应及外部环境影响变化而可能发生的阶段性转变提供了系统化、可解释的分析方式。

2) 正反例分而治之。符合预期目标的样本(正例)能够为探究材料内禀关系提供坚实基础，但预期以外的样本(反例)中往往也包含具有参考价值的信息，且二者所蕴含的构效关系存在异质性。例如，物理化学因素在材料体系中的作用机制存在显著的异质性，特定的成分配方或工艺参数可能在反应动力学或热力学层面产生非线性的正向增益或反向抑制作用。针对该现象，需从正、反 2 个维度解析涵盖性能表现优异至性能劣化的完备数据，从而定向剪裁设计空间，并以此驱动模型学习二者样本空间的差异性，进而挖掘二者潜在的一般模式。其关键在于，建立基于目标特性的动态划分机制：当仅有一个性能指标时，依据该性能指标的区间阈值划分正反例；当存在多个性能目标且相互协同变化时(即同向或反向)，采用加权融合算法构建综合评价优值，从而通过量化分级精确界定优劣阈值并划分正反例；当存在多个性能目标且相互拮抗约束时(即此消彼长)，引入聚类算法识别并分离出不同性能倾向的样本簇并划分正反例。进一步，分正反例定向挖掘构效关系规则，剖析正例集合中的关键参数共现模式，并识别反例集合中的性能致败因子，从而实现材料设计空间的精准剪裁与定向优化。

在研发锂离子电池电解液时，正反例配方在溶剂选择、添加剂类型及浓度等方面的轻微变化，可能导致其电化学性能表现出显著差异。结合这一特点，Liu 等提出了基于正反例的锂离子电池电解液配方优化设计方法，旨在解决电解液多项性能指标(如容量特性、阻抗特性、循环寿命等)之间相互制约的协同优化难题。具体而言，针对构造的电解液“配方–性能”完备数据集进行特征选择和降维，并通过集成 *K*-means、谱聚类、层次聚类的方式，实现了在多维性能冲突中对“正向增益正例”与“负向抑制反例”的自动化识别。随后，通过相关性、随机森林和关联分析，提取出可用于指导电解液配方设计的定量规则与优化方案。该研究成功指导某全球领先的锂离子电池企业优化电解液配方，通过规避诱

发界面不稳定的负面组分，使 37 A·h 等多款商用电池的循环寿命提升最高达 30%。

在无机–有机杂化材料的水热合成研发中，有机模板剂的物理化学性质、无机反应物的选择及反应条件(如 pH 值、温度、时间)的轻微变化，就可能导致其最终产物在结晶度、相纯度及结构维度上表现出显著差异(即“成功结晶”与“失败/无定形沉淀”的界限)。结合这一特点，Raccuglia 等[23]提出了基于“黑暗反应”(失败实验)数据的机器学习辅助材料发现方法，旨在解决新材料合成中组合空间巨大且传统“试错法”效率低下的难题。具体地，针对从实验历史记录本中提取的 3955 条完备反应数据(包含大量未发表的失败案例)进行特征工程，利用化学信息学软件计算有机胺的分子描述符(如极化率、氢键供体数)及无机反应物的原子属性(如电负性、原子半径)，构造了高维的“反应物属性–结晶结果”特征集。利用支持向量机实现了在复杂的化学反应空间中对“导致结晶的正例”与“导致失败的反例”的自动化高精度识别，其对未知反应的预测成功率达到 89%，超越了专家经验判断。

此类“正反例分而治之”机器学习建模策略不仅突破了复杂材料体系中正负样本界定模糊的瓶颈，更有效揭示了多物理场耦合作用下各物理化学参数的深层贡献机制。

3) 类别分而治之。在不同的材料体系、制备工艺(如热处理)及服役环境条件下，材料微观结构演化及其性能影响机制往往表现出显著的跨体系分布异质性。针对该现象，可利用显式标签或隐式分布特征，将异质数据拆解为若干物理化学机制相对一致的“同质子域”，从而降低全局建模的复杂度。其核心要义在于建立基于多源样本的动态分域与适配机制。例如，基于显式元数据的规则划分(依据材料代系、工艺类型等先验标签进行物理域解构)、基于隐式特征的聚类挖掘(利用无监督算法捕捉样本空间中的潜在分布结构以识别数据子域)，以及基于差异化适配的局部专家构建(针对不同子域的数据特性自适应匹配最优模型架构)，以实现对复杂数据分布的精准划分与表示。基于此，构建各子域的专用回归器或分类器，在保证局部拟合精度的同时，有效提升模型面对跨体系复杂数据的全局泛化能力与鲁棒性。

以镍基单晶高温合金为例，依据合金成分不同(如铼、钌等元素的含量)其可被划分为 4 个代系[24]，每个代系所具备的性能不尽相同。结合这一特点，Liu 等[19]提出了一种结合多种材料描述符的分治自适应学习方法(divide-and-conquer self-adaptive, DCSA)(如图 2 所示)，实现了对镍基单晶高温合金蠕变寿命的精准预测。首先，利用 K-Means 聚类将将 266 个合金样本划分为具有相似蠕变机制的 8 个簇。在每个簇中，采用适应度函数自动选择最适合该类样本的最优预测模型，实现针对性预测。为验证方法有效性，采用 6 种回归模型来预测蠕变断裂寿命，结果表明 DCSA 在预测准确性方面明显优于其他 5 个单一回归模型。

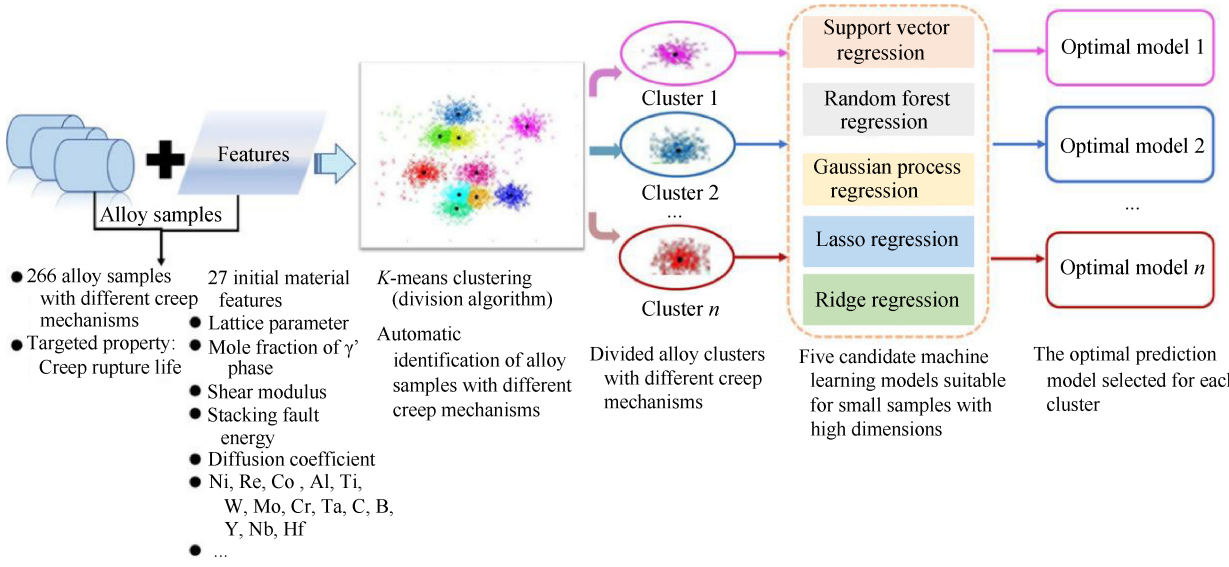


图 2　DCSA 用于预测蠕变断裂寿命的流程[19]

Fig. 2　DCSA workflow for creep fracture life prediction[19]

这种“类别分而治之”的思想也在其他研究中被广泛采用并得到验证。例如，Feng 等[25]通过构建一个基于合金强化机制的分而治之机器学习框架(DCLF)，对多合金混合小样本数据集进行有效处理与划分，实现了钢硬度预测的高效性与鲁棒性；针对材料体系中不同结构类型或相带来的复杂性，Zeni 等[26]提出了分而治之的线性机器学习势函数，通过自动聚类特定的结构类型或相，显著提升了对原子尺度力和能量的预测精度；Wei 等[27]基于数据点在构效关系上的统计分布一致性，提出了高斯过程回归树分类器(TCGPR)，将原始异质数据集自动划分为 3 个内部同质子集，从而显著提升小规模数据集上的预测精度，并成功设计、制备并验证了 3 种兼具高强度与高延展性的新型无铅焊料合金。此外，Guo 等[28]提出了一种基于质谱成像(MSI)数据空间分布特性的 dc-DeepMSI 模型，用于在无须组织学先验知识的情况下识别潜在的代谢异质性。从小鼠胎儿数据上得出的实验结果表明，dc-DeepMSI 优于现有最先进的质谱成像分割方法。

此类“类别分而治之”机器学习建模策略不仅适用于包含多源异质样本的复杂材料问题，还能在一定程度上探寻成分、结构、工艺等不同因素对多体系材料性能的影响机理，进而揭示各因素之间可能存在的协同效应或拮抗关系，为高效优化材料设计提供可解释性指导。

3.2 特征表示视角下的分而治之

特征表示“分治”旨在通过领域知识对用于刻画复杂构效关系的描述符组合或数据表示进行拆解，从而达到消解对高精度复杂模型构建的依赖，破除建模知识壁垒的目的。

1) 属性分而治之。物理化学因素对材料性能的作用不仅体现在正向或负向影响上，不同物化因素对材料性能的影响机制不尽相同。这使得基于单一的描述符组合进行机器学习建模难以从数据中揭示潜在的多样化作用机制。针对这一问题，需要摒弃以往寻找唯一的描述符子集的建模方式，寻求一种有效方法筛选出多组最优描述符组合，进而有针对性地构建多个高精度模型，并从这些模型中提取多样化的规则，以此挖掘隐藏在材料数据中的多重机制，既保证了模型的预测准确性，又提供了多维度的物理化学解释。

为此，Liu 等[20]提出了描述符分而治之机器学习方法(descriptors divide-and-conquer machine learning, descriptors-DCML)，如图 3 所示。为实现多描述符组合寻优，提出核特征(候选特征集中与机器学习模型预测性能相关性最强的特征或最小特征组合)与特征团(包含核特征在内的特征组合，能够使模型达到更高预测性能的特征集合)的概念，用于区分隐藏在数据中的不同影响机制。进一步，将粗糙集理论(rough set theory, RST)[29–30]引入并结合材料领域知识，旨在从数据内部提取特征间的关联模式并识别对目标属性有重大影响的关键描述符。具体地，通过构建不同特征组合与其在机器学习模型上预测性能之间的特征决策表，并与关键描述符相关的材料领域知识和粗糙集的求核算法相结合，识别出一组核特征。进而，利用领域知识嵌入的特征选择算法(NCOR-PSO)[31]获得特征团，从而构建多个高精度的机器学习模型。最后，从所构建机器学习模型的学习结果中提取规则，并整合与影响机制相关的材料领域知识，为复杂的构效关系提供多维度且可解释的分析。该方法通过 NASICON 型无机固态电解质、镍基单晶高温合金、纳米复合固体聚合物、陶

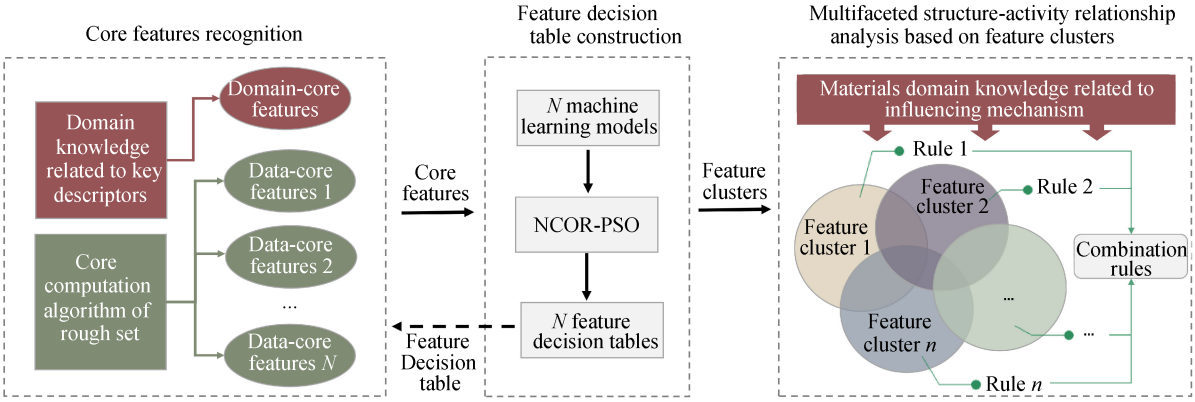


图 3 descriptors-DCML 的总体示意图[20]

Fig. 3 Overall schematic of descriptors-DCML[20]

瓷材料、无序钙钛矿氧化物、立方锂氩罗石六个不同材料体系的数据集进行实验验证。实验结果表明，该方法相较于已报道的预测精度，模型性能的提升最高达到 13%。同时，以 NASICON 型无机固态电解质为例，抽取出了 9 条相关规则。该方法不仅能够对材料特性进行高精度预测，还能够实现规则的提取以及构效关系的可解释分析。

基于此类“属性分而治之”建模策略，不仅能够有效提取可解释的构效关系，还可以探寻不同描述符对单体系材料性能的作用机制，这在推进高通量材料筛选、提升材料性质预测方面具有巨大潜力。

2) 模态分而治之。为应对复杂材料构效关系，研究人员采用不同计算/实验手段从不同角度对其进行剖析，并衍生出大量的异构数据。然而，材料异构数据所蕴含的信息各有差异且数据表示方法仍存在自身局限性，如基于描述符组合的结构化数据表示方法难以精准刻画晶体的几何拓扑信息。因此，当前将某一种结构数据转化为单一模态并进行机器学习建模的方式难以满足对全面、准确挖掘材料中隐含的内禀关系的需求。

针对这一局限性，研究人员开始采用多模态学习对异构数据进行拆解，进而表示成多种模态数据以全面刻画构效关系。例如 Ren 等[32]构建包含双相钢的成分数据和多源微观结构图像的多模态学习框架，实现极限抗拉强度和均匀延伸率的精准预测，并通过可视化热图技术，解释微观结构中对力学性能起关键作用的区域特征；Moro 等[33]提出一种多模态基础模型框架，通过自监督策略联合建模晶体结构、电子结构与文本信息，显著提升泛化与迁移能力，并通过对隐空间进行可视化，初步解释部分结构–性质分布规律；Wang 等[21]提出了数值型成分–图结构双模态网络 COSNet，在预测锂离子电导率、带隙等材料特性方面表现出显著提升的性能。

此类“模态分而治之”建模策略能够有效摆脱单一模态表示的局限性，实现各模态数据的优势互补，进而可以更全面表征材料复杂内禀关系，准确获取材料内部作用机理的调控方式。

4　总结与展望

针对材料数据及任务复杂导致的机器学习模型建模复杂问题，本工作提出了一种材料领域知识嵌入的分而治之机器学习建模策略，建立不同机器学习模型到不同材料物理化学机制的映射。值得注意的是，该策略并非替代现有主流机器学习模型(如图神经网络、物理信息神经网络)，而是将其视为面向不同子任务的定制化模型选择，充分发挥模型优势以最佳匹配各子任务获得最优解。这有效解决了复杂材料任务中模型准确性与易用性的平衡问题，更全面揭示了材料复杂构效关系。

材料领域知识嵌入的分而治之机器学习建模策略的核心涵盖问题拆解与问题求解两大模块。设计其需要解决以下关键问题：1) 如何保证数据质量以满足分而治之对于数据的多样性与均衡性需求？2) 如何进行更加全面的问题拆解？ 3) 如何实现更加智能化的问题求解？4) 如何保证模型的可解释性以满足构效关系显性揭示的需求？

针对上述问题，未来的研究工作可重点围绕以下思路展开，以期深入探索有效的解决方案：

1) 数据质量是驱动高精度模型构建的燃料。在执行分而治之建模策略时，需要为拆解的子问题收集充足的高质量数据，以保证数据表达的全面性以及模型学习的准确性，同时也要保证数据的均衡性，以避免在模型学习时，模型更倾向于学习多数样本，降低模型的泛化能力[12–13]。

2) 随着材料问题日趋复杂，可引入更多的分而治之建模策略以全面覆盖复杂材料问题的划分维度。以无机固态电解质为例，与以往仅采用基于晶体结构的图表示(graph representation)方法不同，本研究团队提出了一种基于间隙网络(IN)的图表示方法[34]，从晶体内部视角刻画离子输运机理，进而实现对激活能的精准预测。若能将晶体结构与间隙网络图表示方法进行结合，有望实现更全面的构效关系解析，为探究针对材料结构的分而治之提供了思路，同时也为模态分而治之提供了新的数据表示视角。

3) 当前，分而治之机器学习依赖于材料专家经验进行手动建模与调参，可引入自动机器学习辅助以降低其建模复杂度。例如，Liu 等[35]提出的一种基于元学习的自动聚类集成学习方法(AutoCluster)可用于实现复杂材料问题的自动化拆解与划分。此外，通过所提出的一种基于元学习的材料属性预测自动建模方法(Auto-MatRegressor)可实现对各子问题的智能化治理与精准建模[36]。

4) 分而治之建模策略有效地实现了复杂任务的解构，显著降低了单个模块的学习难度。这种“化繁为简”的转变，使得研究者能够针对特定问题引入决策树、线性回归等具备可解释特性的机器学习模型，同时也可引入面向机器学习模型的内部解释方法(如 Cloud-VAE[37]、DWVAE[38])，从而在保持系统

整体性能的同时，增加了关键决策环节的透明度。

未来随着材料科学研究范式的数字化、智能化深刻变革，分而治之建模策略将被赋予新的生命力。通过深度融合专业知识的科学解构，“分而治之”有望推动人们更透彻、更系统地理解材料的复杂行为，深入揭示其内禀关系与作用机制，进而加速高性能材料的精准设计与研发进程。

参考文献:

[1] HORTON M K, HUCK P, YANG R X, et al. Accelerated data-driven materials science with the Materials Project[J]. Nat Mater, 2025, 24(10): 1522–1532.

[2] LIU Y, YANG Z W, YU Z Y, et al. Generative artificial intelligence and its applications in materials science: Current situation and future perspectives[J]. J Materiomics, 2023, 9(4): 798–816.

[3] LIU Y J, JOVANOVIC M, MALLAYYA K, et al. Materials expert-artificial intelligence for materials discovery[J]. Commun Mater, 2025, 6: 212.

[4] SCHMIDT J, MARQUES M R G, BOTTI S, et al. Recent advances and applications of machine learning in solid-state materials science[J]. NPJ Comput Mater, 2019, 5: 83.

[5] GUO Y B, WANG Y H, ZHANG W, et al. From algorithms to applications: A comprehensive review of machine learning in computational materials science[J]. Arch Comput Meth Eng, 2025, https://doi.org/10.1007/s11831-025-10342-4.

[6] LIU Y, ZHAO T L, JU W W, et al. Materials discovery and design using machine learning[J]. J Materiomics, 2017, 3(3): 159–177.

[7] WU Y S, ZHOU B J, LIU L, et al. Structural characteristics and recent advances in thermoelectric binary indium chalcogenides[J]. Research, 2025, 8: 0727.

[8] LIU Y, WU J M, AVDEEV M, et al. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties[J]. Adv Theory Simul, 2020, 3(2): 1900215.

[9] 刘悦, 邹欣欣, 杨正伟, 等. 材料领域知识嵌入的机器学习[J]. 硅酸盐学报, 2022, 50(3): 863–876.
LIU Yue, ZOU Xinxin, YANG Zhengwei, et al. J Chin Ceram Soc, 2022, 50(3): 863–876.

[10] LIU Y, GUO B R, ZOU X X, et al. Machine learning assisted materials design and discovery for rechargeable batteries[J]. Energy Storage Mater, 2020, 31: 434–450.

[11] 李航. 统计学习方法[M]. 2 版. 北京: 清华大学出版社, 2019.

[12] LIU Y, YANG Z W, ZOU X X, et al. A general framework to govern machine learning oriented materials data quality[J]. Mater Sci Eng R Rep, 2025, 166: 101050.

[13] LIU Y, YANG Z W, ZOU X X, et al. Data quantity governance for machine learning in materials science[J]. Natl Sci Rev, 2023, 10(7): nwad125.

[14] 刘悦, 马舒畅, 杨正伟, 等. 面向材料领域机器学习的数据质量治理[J]. 硅酸盐学报, 2023, 51(2): 427–437.
LIU Yue, MA Shuchang, YANG Zhengwei, et al. J Chin Ceram Soc, 2023, 51(2): 427–437.

[15] 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[J]. Engineering, 2019, 5(6): 1017–1026.

[16] WU Z F, BAI R H, ZHANG A N, et al. Divide-or-conquer which part should you distill your LLM[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. Miami, Florida, USA. Stroudsburg, PA, USA: ACL, 2024: 2572–2585.

[17] ALLMAN E S, BAÑOS H, RHODES J A, et al. NANUQ^{+}: A divide-and-conquer approach to network estimation[J]. Algorithms Mol Biol, 2025, 20(1): 14.

[18] LIU Y, WU J M, YANG G, et al. Predicting the onset temperature (T_g) of Ge_xSe_{1-x} glass transition: A feature selection based two-stage support vector regression method[J]. Sci Bull, 2019, 64(16): 1195–1203.

[19] LIU Y, WU J M, WANG Z C, et al. Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning[J]. Acta Mater, 2020, 195: 454–467.

[20] LIU Y, WU L H, YANG Z W, et al. Descriptors divide-and-conquer enables multifaceted and interpretable materials structure–activity relationship analysis[J]. Adv Funct Mater, 2025, 35(26): 2421621.

[21] WANG S, GONG S, BÖGER T, et al. Multimodal machine learning for materials science: Discovery of novel Li-ion solid electrolytes[J]. Chem Mater, 2024, 36(23): 11541–11550.

[22] GREENBANK S, HOWEY D A. Piecewise-linear modelling with automated feature selection for Li-ion battery end-of-life prognosis[J]. Mech Syst Signal Process, 2023, 184: 109612.

[23] RACCUGLIA P, ELBERT K C, ADLER P D F, et al. Machine-learning-assisted materials discovery using failed experiments[J]. Nature, 2016, 533(7601): 73–76.

[24] 张健, 王莉, 谢光, 等. 镍基单晶高温合金的研发进展[J]. 金属学报, 2023, 59(9): 1109–1124.
ZHANG Jian, WANG Li, XIE Guang, et al. Acta Metall Sin, 2023, 59(9): 1109–1124.

[25] FENG Y, KAI S, LIANG G, et al. DCLF: A divide-and-conquer learning framework for the predictions of steel hardness using multiple alloy datasets[J]. Mater Today Commun, 2022, 30: 103195.

[26] ZENI C, ANELLI A, GLIELMO A, et al. Divide-and-conquer potentials enable scalable and accurate predictions of forces and energies in atomistic systems[J]. Digit Discov, 2024, 3(1): 113–121.

[27] WEI Q H, CAO B, YUAN H, et al. Divide and conquer: Machine learning accelerated design of lead-free solder alloys with high strength and high ductility[J]. NPJ Comput Mater, 2023, 9: 201.

[28] GUO L, DONG J Y, XU X N, et al. Divide and conquer: A flexible deep learning strategy for exploring metabolic heterogeneity from mass spectrometry imaging data[J]. Anal Chem, 2023, 95(3): 1924–1932.

[29] ZHOU Q L, MO H M, DENG Y. A new divergence measure of Pythagorean fuzzy sets based on belief function and its application in medical diagnosis[J]. Mathematics, 2020, 8(1): 142.

[30] GAETA A, LOIA V, LOMASTO L, et al. A novel approach based on rough set theory for analyzing information disorder[J]. Appl Intell, 2023, 53(12): 15993–16014.

[31] LIU Y, ZOU X X, MA S C, et al. Feature selection method reducing correlations among features by embedding domain knowledge[J]. Acta Mater, 2022, 238: 118195.

[32] REN D, WANG C C, WEI X L, et al. Building a quantitative composition-microstructure-property relationship of dual-phase steels *via* multimodal data mining[J]. Acta Mater, 2023, 252: 118954.

[33] MORO V, LOH C, DANGOVSKI R, et al. Multimodal foundation models for material property prediction and discovery[J]. Newton, 2025,

1(1): 100016.
[34] SHI S Q, LIN H L, WU L H, et al. Predicting Li transport activation energy with graph convolutional neural network[M]//Intelligent Computers, Algorithms, and Applications. Singapore: Springer Nature Singapore, 2024: 153–164.
[35] LIU Y, LI S, TIAN W J. AutoCluster: Meta-learning based ensemble method for automated unsupervised clustering[M]//Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2021: 246–258.
[36] LIU Y, WANG S Y, YANG Z W, et al. Auto-MatRegressor: Liberating machine learning alchemists[J]. Sci Bull, 2023, 68(12): 1259–1270.
[37] LIU Y, LIU Z T, LI S, et al. Cloud-VAE: Variational autoencoder with concepts embedded[J]. Pattern Recognit, 2023, 140: 109530.
[38] LIU Y, YU Z Y, LIU Z T, et al. Improving disentanglement in variational auto-encoders *via* feature imbalance-informed dimension weighting[J]. Knowl Based Syst, 2024, 296: 111818.

Divide-and-Conquer Machine Learning Embedded with Materials Domain Knowledge

YANG Zhengwei[1], *LIU Zitu*[1], *ZHANG Wangying*[2], *ZUO Wei*[1], *YU Zhenyao*[1], *ZOU Xinxin*[1], *YUAN Haodong*[1], *SU Yufeng*[1], *SHI Siqi*[2,3], *LIU Yue*[1]

(1. State Key Laboratory of Materials for Advanced Nuclear Energy & School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; 2. State Key Laboratory of Materials for Advanced Nuclear Energy & School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China; 3. Materials Genome Institute, Shanghai University, Shanghai 200444, China)

Extended Abstract

Data-driven machine learning (ML) has made significant progress in predicting structure-property relationships, discovering new materials, and optimizing processes within materials science. However, the factors influencing materials structure-property relationships are highly complex and multifaceted, resulting in extremely intricate distribution patterns in materials data. This complexity makes it difficult for traditional data-driven approaches, which attempt to establish a single mapping between a fixed set of descriptors and target properties, to accurately and comprehensively unravel these relationships. Through the analysis of the characteristics of materials data and the modeling principle of data-driven ML methods, we have clarified the three main contradictions occurring to the application of ML in materials science, *i.e.*, the contradictions between high dimension and small sample, accuracy and usability of models, learning results and domain knowledge. To reconcile these three contradictions, we propose the framework of machine learning embedded with materials domain knowledge and a series of novel methods. Then, by analyzing the characteristics of materials research problems and the complex and diverse structure-property relationships inherent in materials data, we believe that it is possible to effectively decompose complex material problems and address them individually by leveraging materials domain knowledge. To this end, we here propose a divide-and-conquer ML modeling strategy embedded with materials domain knowledge.

Specifically, the “Divide and Conquer” strategy aims to leverage materials domain knowledge to decompose complicated material problems and construct ML models towards each sub-problem, which can effectively reduce the complexity of ML models and ensures that the learning results of models are both accurate and credible through the integration of domain knowledge. It can be regarded as an application of the algorithmic paradigm of divide-and-conquer to ML modeling, or alternatively, as a novel form of ensemble learning. This strategy consists of: i) “Divide”: Guided by materials domain knowledge, a complicated materials problem can be decomposed into several sub-problems from multiple perspectives based on an analysis of underlying physicochemical mechanisms. ii) “Conquer”: Individual models are constructed for each sub-problem, where domain-specific knowledge is injected into the sub-models, establishing mappings between different ML models and distinct physicochemical mechanisms. iii) “Combine”: A fusion mechanism is designed to integrate the feature extraction or decision outputs from the sub-models. As a result, multi-faceted analysis from diverse perspectives yields more comprehensive and rational insights into complex materials problems. This aids domain experts in identifying the critical thresholds and transition boundaries of the intrinsic mechanisms governing material property variations, reduces the complexity of individual ML models within each subspace, and enhances both the predictive accuracy and interpretability of the models. A further investigation reveals that, although the concept of “divide and conquer” has been widely adopted, no attempt has yet been made to theoretically derive and validate its content. In this study, we employ variational inference to provide a formal symbolic representation for the divide-and-conquer ML modeling strategy and verify its rationality and scientific validity.

Currently, researchers have begun to leverage domain knowledge to decompose materials problems or data, giving rise to diverse “divide-and-conquer” ML modeling strategies. Specific implementations include: i) Stage divide-and-conquer. Deconstructing long-term, highly complex evolutionary processes into several substages with relatively singular and stable physicochemical mechanisms, thereby simplifying the intricate. This strategy can be applied to complex materials issues featuring multi-step transition characteristics, such as the glass transition temperature (T_g) and the battery rollover point, to enable accurate prediction. ii) Positive/negative examples divide-and-conquer. Samples are categorized into distinct subspace according to their positive or negative impact on the target performance. This strategy can be applied to complex materials problems that encompass samples from superior

performance to degradation, such as in the optimal design of lithium-ion battery electrolytes and the hydrothermal synthesis of inorganic-organic hybrid materials. iii) Category divide-and-conquer. By leveraging explicit labels or implicit distributional features, heterogeneous data is decomposed into several isotropic subdomains with relatively consistent physicochemical mechanisms, thereby reducing the complexity of global modeling. This strategy can be applied to the precise prediction of heterogeneous samples, such as the creep life of nickel-based single-crystal superalloys, the hardness of steel, and the strength and ductility of lead-free solder alloys. iv) Attribute divide-and-conquer. Multiple sets of optimal descriptor combinations are screened out, thereby enabling the targeted construction of several high-precision models. This strategy is validated on six distinct material systems, such as NASICON-type inorganic solid-state electrolytes, demonstrating high predictive accuracy and interpretative capabilities. v) Modality divide-and-conquer. Decomposing heterogeneous data and representing it as multi-modal data to comprehensively unravel structure-property relationships. This strategy is applicable to complex materials problems with multi-modal information, such as dual-phase steels and lithium-ion solid-state electrolytes.

Summary and prospects In this study, we propose a divide-and-conquer ML modeling strategy embedded with materials domain knowledge, which establishes mappings between different ML models and the distinct physicochemical mechanisms of materials. Future research efforts can focus on the following directions to further advance this strategy: i) Data quality remains the fundamental fuel driving the construction of high-accuracy models. When implementing this strategy, it is necessary to collect sufficient high-quality data for each sub-problem. ii) As material problems become increasingly complex, more and more divide-and-conquer ML modeling strategies can be introduced to comprehensively cover the dimensions of decomposing complex material problems. iii) Currently, divide-and-conquer ML relies on manual model selection and hyperparameter optimization based on expert experience. Automated machine learning (AutoML) can be introduced to eliminating the knowledge barriers to modeling. iv) The divide-and-conquer ML modeling strategy allows researchers to introduce inherently interpretable models for specific sub-problems and incorporate internal interpretation methods for models. In the future, as the paradigm of materials science research undergoes a profound shift towards digitalization and intelligence, this strategy is poised to gain renewed relevance and vitality. Through the deep integration of scientific deconstruction guided by domain expertise, the “divide and conquer” strategy is expected to propel a more thorough and systematic understanding of complex material behaviors. It will facilitate deeper revelation of intrinsic relationships and underlying mechanisms, thereby accelerating the process of precise design and development of high-performance materials.

Keywords materials science; machine learning; divide-and-conquer; domain knowledge