

机器学习在热电材料领域中的应用

盛 晔¹, 宁金妍¹, 杨 炯^{1,2}

(1. 上海大学材料基因组工程研究院, 上海 200444; 2. 之江实验室, 杭州 311100)

摘 要: 热电材料是环境友好型能源转换材料, 涉及的体系十分多样。其性能优化是一个多参数协调的复杂问题, 一直是研究者们关注的热点。虽然热电的计算模拟方法和实验方法发展迅速, 但是热电材料的搜索效率仍需要进一步提高。机器学习具有计算成本低和预测速度高的优势, 可以快速缩小搜索空间, 加快对热电材料结构和性能优化的研究。本综述从数据类型的角度出发, 介绍了热电材料中的小样本数值数据(数据量约为 10^2), 大样本数值数据(数据量大于 10^4)及图片数据中机器学习的应用和研究进展, 进一步详细地讨论了在不同的数据类型中研究热电材料的结构和性能所使用的不同的机器学习算法模型, 并对其未来的发展趋势和应用方向进行了展望。

关键词: 热电材料; 机器学习; 图卷积神经网络

中图分类号: TP181; O738; O736 文献标志码: A 文章编号: 0454-5648(2023)02-0499-11

网络出版时间: 2022-12-27



Applications of Machine Learning in Thermoelectric Materials

SHENG Ye¹, NING Jinyan¹, YANG Jiong^{1,2}

(1. Materials Genome Institute, Shanghai University, Shanghai 200444, China; 2. Zhejiang Laboratory, Hangzhou 311100, China)

Abstract: Thermoelectric materials are environmental-friendly energy conversion materials. Their performance optimization is a complex issue of multi-parameter coordination, which becomes a challenge. Although the computational simulation and experimental methods for thermoelectric materials have developed rapidly, the efficiency of searching thermoelectric materials still needs to be further improved. Machine learning has some advantages of low computational cost and high prediction speed, which can shorten the search process and accelerate the corresponding studies on the structure and performance optimization of thermoelectric materials. This review introduced the research progress on machine learning for small sample numerical data (data volume is about 10^2), large sample numerical data (data volume $>10^4$) and image data in thermoelectric materials from the perspective of data types. Moreover, different machine learning algorithm models used for the structure and performance of thermoelectric materials in different data types were discussed. In addition, the future development and application direction were also prospected.

Keywords: thermoelectric materials; machine learning; graph convolutional neural networks

热电转换技术可利用材料实现电能和热能直接相互转换, 在制冷或发电领域具有很大的应用潜力。热电转换过程中没有气体的排放和其他污染, 具有绿色环保特点, 因此加强热电材料的研究可以为减少全球二氧化碳和温室气体排放做出重要贡献, 助力中国实现“碳达峰, 碳中和”的目的。目前热电转换技术已经逐步实现了商业化应用, 如汽车座椅冷

却器/加热器、激光二极管的冷却器、以及汽车和卡车余热回收发电设备等。热电器件是固态材料, 以电流为工作液, 与使用液态工作液(蒸汽/弗利昂等)的传统热机系统相比, 没有活动部件, 具有运行无噪声、无震动的优点。但是, 热电装置的转换效率相比传统装置有待提升^[1]。热电装置的效率主要由其所使用的半导体材料的热电性能决定。热电材料

收稿日期: 2022-10-12。 修订日期: 2022-11-05。

基金项目: 之江实验室科研攻关项目(2021PE0AC02); 国家自然科学基金(52172216)。

第一作者: 盛 晔(1994—), 女, 博士研究生。

通信作者: 杨 炯(1981—)男, 博士, 教授;

宁金妍(1985—), 女, 博士, 实验师。

Received date: 2022-10-12. Revised date: 2022-11-05.

First author: SHENG Ye (1994—), female, Doctoral candidate.

E-mail: yesheng@shu.edu.cn

Correspondent author: YANG Jiong (1981—), male, Ph.D., Professor.

NING Jinyan (1985—), female, Ph.D., Experimentalist.

E-mail: jiongy@t.shu.edu.cn; jyning@t.shu.edu.cn

性能可以用无量纲的热电优值 ZT 来量化,

$$ZT = S^2 \sigma T / \kappa \quad (1)$$

其中: S 为材料的 Seebeck 系数; σ 为材料的电导率; $S^2 \sigma$ 是材料的功率因子; κ 为材料的热导率; T 为绝对温度。要提高材料的 ZT 值, 就要实现高的 Seebeck 系数和电导率的同时, 获得低的热导率。但是 Seebeck 系数、电导率及热导率之间变化趋势不一, 无法对单一参数单独调控, 所以热电性能优化需要多个参数协同控制^[2-3]。

从历史维度来看, 热电材料性能有 2 次重大的飞跃。第 1 次的性能飞跃是在 20 世纪 60 年代初期, 随着半导体理论物理快速发展, 研究者们发现部分半导体材料的 Seebeck 系数大致是发现热电效应的金属材料的 10 倍之多, 因此半导体材料的热电性能研究成为热点。在此期间, 一些热电材料如 Bi_2Te_3 、 PbTe 、 SiGe 等脱颖而出^[4], 其中 Bi_2Te_3 基化合物依旧是当今商业应用中的主流热电材料。第 2 次的性能飞跃是在 2000 年左右, 随着声子玻璃-电子晶体概念的复兴, 一些低热导的热电材料备受关注, 如填充方钴矿^[5-8]、笼状化合物^[9-11]、类金刚石结构材料^[12]、Half-Heusler 合金^[13-18]和 $\text{Cu}_2(\text{Se}, \text{S})$ 类液态材料^[19]等, 其中性能优越的热电材料的 ZT 值在 1~2 的区间范围内。纳米技术进一步提高了热电材料的性能, 使得部分材料的 ZT 值可以突破 2^[20], 如 $(\text{PbTe})_m(\text{AgSbTe}_2)^{[21]}$ 、 $\text{Bi}_2\text{Te}_3/\text{Sb}_2\text{Te}_3$ 低维超晶格等^[22] (见图 1), 更多有关热电发展的细节参考综述^[23-24]。

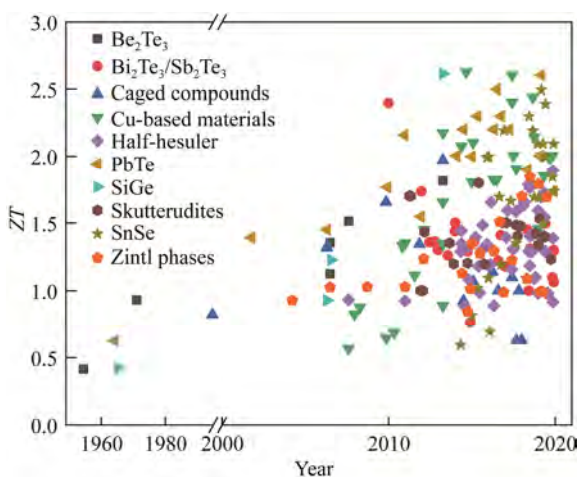


图 1 热电材料发展^[5-24]

Fig. 1 Development of thermoelectric materials^[5-24]

自材料基因组计划启动以来, 高通量技术在材料领域发挥了极其重要的作用, 改变了材料的研究模式, 打开了材料研发的新思路。在热电材料领域

中, 相关的高通量计算^[25-26]和高通量实验^[27-31]方面的报道也是层出不穷。机器学习应用到热电材料研究中, 可以成为实现热电材料筛选的有力工具。这主要因为: 1) 机器学习的算法多样可以针对不同的数据类型; 2) 机器学习的模型预测速度相比第一性原理计算或实验要快, 且模型训练本身的计算成本很低。因此, 高通量技术大大增加了热电材料数据的积累速度, 数据驱动的机器学习兼顾了研发效率和成本, 为热电材料的再一次性能飞跃提供了新的可能。

本综述从热电数据的类型角度出发, 将热电材料数据分为小样本数值数据、大样本数值数据及图像数据(图 2), 并由此介绍机器学习在各类型热电数据中的应用案例及研究进展, 并对未来的发展趋势和应用方向进行了展望。

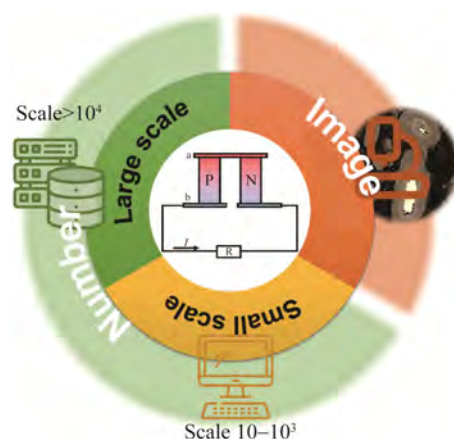


图 2 热电材料数据分类

Fig. 2 Classification of thermoelectric material data

1 数值热电数据

1.1 小样本数值热电数据的机器学习

在材料科学中, 无论是通过计算还是实验的方法, 获得样本的成本通常都很高, 因此材料的数据集往往是小样本数据集, 热电材料中亦是如此。同时, 热电材料涉及的性质参数多样, 按获取的难易程度来排序分别有: 晶体结构基础数据, 电输运性质相关数据(带隙、Seebeck 系数、功率因子、载流子浓度、电子弛豫时间等), 热输运性质(热导率)。后文将依次介绍各类热电性质的机器学习工作。

1.1.1 晶体结构基础性质的机器学习 早期研究的热电材料性质主要集中在晶体结构基础数据, 如晶体结构类型、晶格常数等, 这些性质容易获取, 并与化合物是否稳定密切相关。研究者们尝试建立这些性质与对应的化合物包含基础元素性质之间的

关系。例如, Oliynyk^[32]使用聚类分辨率特征选择(CR-FS)和支持向量分类(SVC)算法,建立了元素价电子数、主量子数和电负性等与目标三元多晶型化合物晶体结构类型的关系模型。Li 等^[33]基于 157 个 GeFeO_3 型钙钛矿化合物的晶格常数、离子半径、电负性和价电子数等 5 个特征,运用人工神经网络算法得到了误差小于 2% 的晶格常数预测模型。随后, Javed 等^[34], 基于同一数据集,用支持向量回归(SVR)算法,在特征参数相同的情况下将预测模型误差降低到 1%。

1.1.2 电输运性质的机器学习 随着第一性原理和热电电输运计算的发展,针对热电材料电输运性质的研究逐渐变多。金属、半导体和绝缘体之间的电输运性质有着巨大的差异。金属的电导率虽然很高,但 Seebeck 系数很低,而绝缘体电导率为 0。半导体材料中电导率和 Seebeck 系数相对平衡。因此决定金属、半导体和绝缘体分类的带隙性质就成为了筛选良好的热电材料的先决条件。Lee 等^[35]用第一性原理计算了 270 个二元/三元化合物的带隙(G_0W_0 方法),以化合物所包含的元素的基本物理性质以及其晶体结构作为特征,对 SVR、普通最小二乘回归(OLSR)、最小绝对收缩和选择算子(LASSO) 3 种机器学习算法得到的模型进行了比较。除了不同算法的精度比较外,也可以采用集成学习,结合多种算法来提高机器学习的精度,如 Xu 等^[36]以晶格参数、平均价电子数、原子序数等 42 个特征,用

集成学习结合了 LASSO、SVR、梯度提升树(GBDT) 3 种模型,对类金刚石化合物的带隙进行预测。

除带隙这一基础性质外,电输运的其他相关性质,如 Seebeck 系数,功率因子、载流子浓度、电子弛豫时间等也通过机器学习得以进一步研究。电输运的数据来源有 2 类,一类是基于实验数据,譬如, Furmanchuk 等^[37]的训练集来源于实验数据库 UCSB 和部分补充体系,以温度范围在 300 K 至 1 000 K 晶体材料的 Seebeck 系数为目标性质,用随机森林方法建立了组成元素基本物理属性在内的 452 个特征参数和目标为 Seebeck 系数的模型。另一类数据由第一性原理计算产生,如 Sheng 等^[38]以类金刚石硫族和磷族化合物的 p 型功率因子(PF)为目标性质,采用了主动学习的框架。主动学习循环包含 3 个部分:机器学习候选者选择、候选者的电输运计算验证、验证数据入库并更新模型(图 3),这一框架被证明可以提高机器学习模型的外推能力。在测试完主动学习中的 3 种候选者选择策略后,候选者选择为委员会策略的梯度提升回归(GBR)模型具有最高的精度(Pearson $R=0.95$)。使用 GBR 模型预测了整个搜索空间的化合物的 PF 后,发现了 3 种具有特殊结构的类金刚石结构材料可能具有更高功率因子,分别为二元磷族化合物、具有 VCC 结构的 IIB1: IIIA2: VIA4 化合物及含有小原子半径元素的材料。随后, Yang 等^[39]基于相同的体系,通过堆叠技术结合多种机器学习方法,得到了一个高精

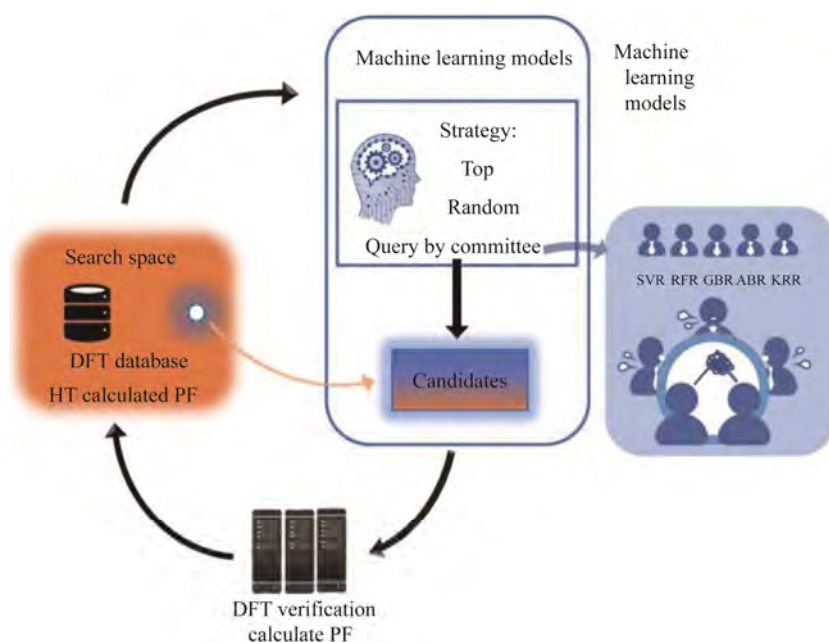


图 3 主动学习循环的工作流程^[38]

Fig. 3 Workflow of active learning cycle^[38]

度(决定系数大于 0.95)类金刚石化合物 PF 预测模型, 并且用 Shapley 补充解释(SHAP)方法分析了目标和描述符之间相关性。

1.1.3 热输运性质的机器学习 材料热导率的精确计算会涉及高阶的力常数的计算, 这对于结构复杂的体系来说往往非常困难。因此, 多数热导率数值由实验产生。研究者尝试建立元素或结构的基础性质和热导率之间的关系。如 Liu 等以辉砷铋矿体系, 基于构型熵的概念实现了对具有任意化学配比体系晶格热导率的快速准确预测, 这一方法在 half-Heusler 体系也同样表现优异^[40]。同时, 也可以通过拟合力常数和机器学习原子间势(ML-IPs)来加速热导率的计算, 如 Yang 等^[41]采用主动学习的双自适应采样(DAS)方法来自动选取 ML-IPs 的训练数据集, 以实现在宽温度范围内合理的热输运模拟。并且基于拟合得到的 ML-IPs 计算了 Diamond、BAs、CoSb₃ 以及 Mg₃Sb₂ 的晶格热导率, ML-IPs 在热输运模拟在 4 个体系中都有很高的准确度。

1.2 大样本热电数值数据的机器学习

2011 年材料基因组计划的提出, 催生了一系列材料数据库的产生, 如材料计划(MP)^[42–43]、自动物料发现流程(AFLOW)^[44–46]、开放量子材料数据库(OQMD)^[47]及新材料发掘(NOMAD)^[48]等。随着材料数据库的发展, 材料数据的体量达到了 10⁴ 的量级。在深度学习方法未引入时, 大样本的数据都是以结合材料分析经验的统计分析为主。Chen 等^[25]将按阴离子分类统计了化合物的功率因子的小提琴图, 发现氧化物往往比硫化物、硒化物和碲化物有更低的功率因数; 使用聚类方法, 基于 58 个描述符, 包括第一性原理计算衍生的特征、带隙、最近邻键长、元素电负性和原子半径等, 对化合物进行聚类, 并分析了各簇的电输运性质的规律。

大样本数据的机器学习一大困难在于特征的生成。数据库中化合物结构各异难以用一个化学通式来表示, 因此由元素性质生成的特征向量的维度无法统一, 而且采用加权平均或其他统计手段的方法又会损失部分信息。材料结构图表示的方法可以解决这一难题, 结合图卷积神经网络能实现大样本数据的机器学习。2017 年, 信息传递神经网络(MPNN)^[49]提出了可以描述多数图神经网络的结构架构, 其将图神经网络分为 2 个部分: 信息传输部分和读取部分。信息传递包括 2 个函数: 信息传递函数和节点更新函数。读取部分输入为信息传递后图中所有节点的特征, 输出为整个图的一个整体

性质。由于热电材料多数是无机的晶体化合物, 所以着重关注在无晶体结构的性质上有所应用的图神经网络。在研究晶体结构的图神经网络中, 图卷积神经网络是主流, 图卷积的架构中有图卷积与池化两部分。图卷积, 事实上是将信息传递函数和节点更新函数结合成一个函数, 两者的目的都是融合局域节点信息, 生成高阶的节点特征; 读取部分从功能上对应于图卷积神经网络的池化层, 都是聚合各节点的特征输出一个全图的性质。因此, 本文采用图卷积神经网络划分方式, 分别从图表示、图卷积和池化 3 个部分来介绍可应用于无机的晶体化合物的图卷积神经网络。

1.2.1 图表示 图表示中的图的定义是节点集合和节点之间边的集合, 通常表示为 $G(V, E)$ 。其中: G 表示一个图; V 表示图中节点的集合; E 是图中边的集合。材料图表示中, 图中的节点对应晶体中的原子, 图中的边对应晶体结构中原子间的相互作用。在有机分子的研究中图表示的应用相对广泛, 因为在分子体系中原子之间的化学键有明确的分类(单键、双键、三键), 使得图和分子结构双向映射易于实现^[50–51]。

近期, 也有研究报道将图表示方法应用在晶体结构中^[52–53]。晶体结构的图表示主要思想是将晶体结构编码成为图, 如图 4 所示。晶体图 G 是一个无向多重图, 它由表示原子的节点和表示晶体中原子之间作用的边来定义。在晶体图中, 由于晶体的周期性, 在同一对节点之间可能有多条边。每个节点 i 由一个特征向量 $\mathbf{v}_i$ 表示, $\mathbf{v}_i$ 中的值是节点 i 的原子相关的属性。类似地, 每条边 $(i, j)_k$ 由特征向量 $\mathbf{u}(i, j)_k$ 表示, 对应于连接原子 i 和原子 j 的第 k 条键的相关性质。

1) 晶体图卷积神经网络(CGCNN)^[53]

CGCNN 中使用的图表示方法见图 5。晶胞包含 N 个原子, 对应产生 N 个节点特征向量(用 01 向量表示), 特征向量包含的原子性质有: 原子的主族数、在元素周期表中的周期数、电负性、共价半径、价电子数、第一电离能、电子亲和能、原子体积和 *spdf* 轨道排布。 N 个原子的特征向量组成了节点矩阵。CGCNN 中边特征向量仅考虑原子与近邻原子距离, N 个原子的边向量组合成为边矩阵。此外边特征向量的每一个元素都由 Gaussian 展开为一长度为 $\mathbf{M}$ 向量。由于不同的原子的近邻原子不同, 因此还有一个序号矩阵, 记录近邻原子的序号。

在 2020 年, Wolverton^[52]基于 CGCNN 的基础上提出了改良 CGCNN (iCGCNN), 其图表示在边的

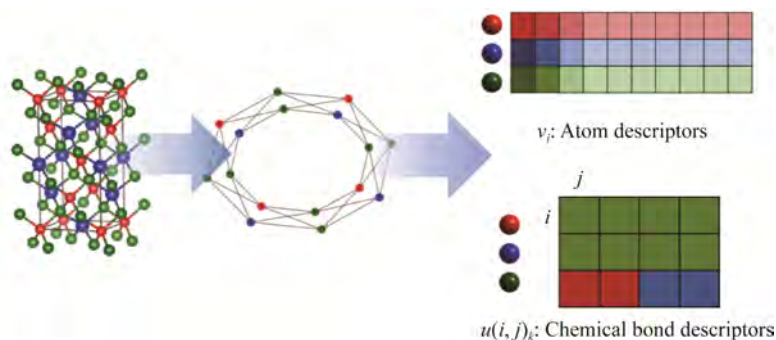


图 4 晶体结构图表示流程

Fig. 4 Graphic representation flow of crystal structure

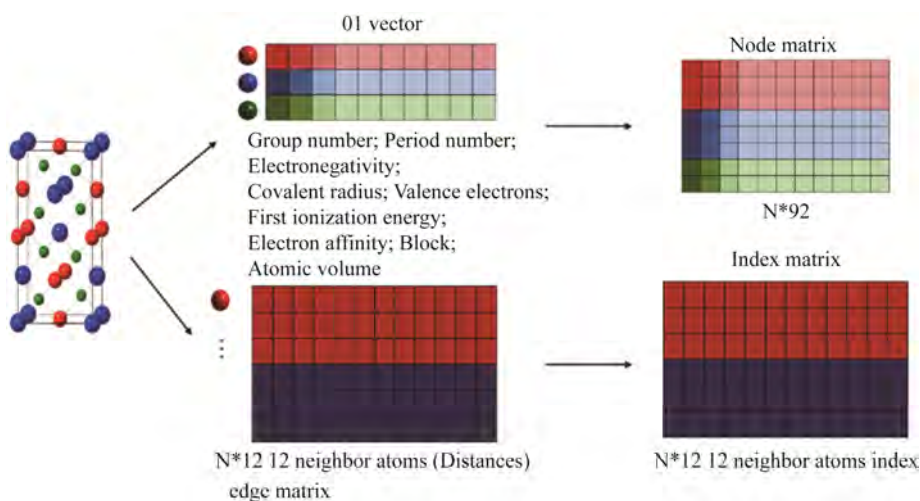


图 5 CGCNN 图表示过程

Fig. 5 Graphic representation of CGCNN

特征中加入了 Voronoi 多面体的相关性质。由于这个网络的应用是确定化合物的稳定性, 因此不做详细的讨论。

2) 材料图网络(MEGNet)^[50]

MEGNet 的图表示(图 6a)除了以原子节点对应的特征向量 $\mathbf{v}_i$ 和以原子间键作为边对应的特征向量 $\mathbf{e}_k$ 外, 还加入了状态的特征向量 $\mathbf{u}$ 来表示晶体所处环境状态, 例如温度等。原子性质组成的节点特征向量仅考虑原子序数 Z , 边特征向量仅涉及原子间距离, 边特征向量也使用高斯展开。

3) 原子线形图神经网络(ALIGNN)^[54]

ALIGNN 图方法(图 6b)的节点特征向量与边特征向量参考了 CGCNN 方法, 节点特征向量涉及的原子性质包括原子的主族数、电负性、共价半径、价电子数、第一电离能、电子亲和能、原子体积、*spdf* 轨道排布。边的特征也是原子间距离。此外, ALIGNN 还加入了化学键角度相关的三元特征。

1.2.2 图卷积 卷积通常应用于图像处理, 输入

是一个形状规则的网格(图像可视为由像素点组成的网格), 而卷积操作则是通过格点及近邻格点的数值加权平均生成新的格点值, 从而使得卷积后的格点包含一个区域的信息(图 7a)。实际上, 图卷积操作与图像卷积操作类似, 也是通过整合节点与近邻节点的节点特征及边特征而生成新节点的特征。不同的是, 图像中格节点有固定的位置关系, 而图卷积中的近邻关系有节点间的边来决定(图 7b)。

1) CGCNN 的图卷积与网络结构

CGCNN 卷积神经网络结构如图 8a 所示, 从输入到输出 CGCNN 网络依次有嵌入层、卷积层、池化层、全连接层和输出层。其中池化函数为归一化的求和, 卷积的公式为:

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \sum_{j,k} \sigma(\mathbf{z}_{(i,j)_k}^{(t)} \mathbf{W}_f^{(t)} + \mathbf{b}_f^{(t)}) \odot g(\mathbf{z}_{(i,j)_k}^{(t)} \mathbf{W}_s^{(t)} + \mathbf{b}_s^{(t)}) \quad (2)$$

其中: $\mathbf{z}_{(i,j)_k}^{(t)} = \mathbf{v}_i^{(t)} \oplus \mathbf{v}_j^{(t)} \oplus \mathbf{u}_{(i,j)_k}$; t 为卷积层标号; $\mathbf{v}_i$, $\mathbf{v}_j$ 分别是 i 原子和 j 原子的节点特征; $\mathbf{u}_{(i,j)_k}$ 是 i

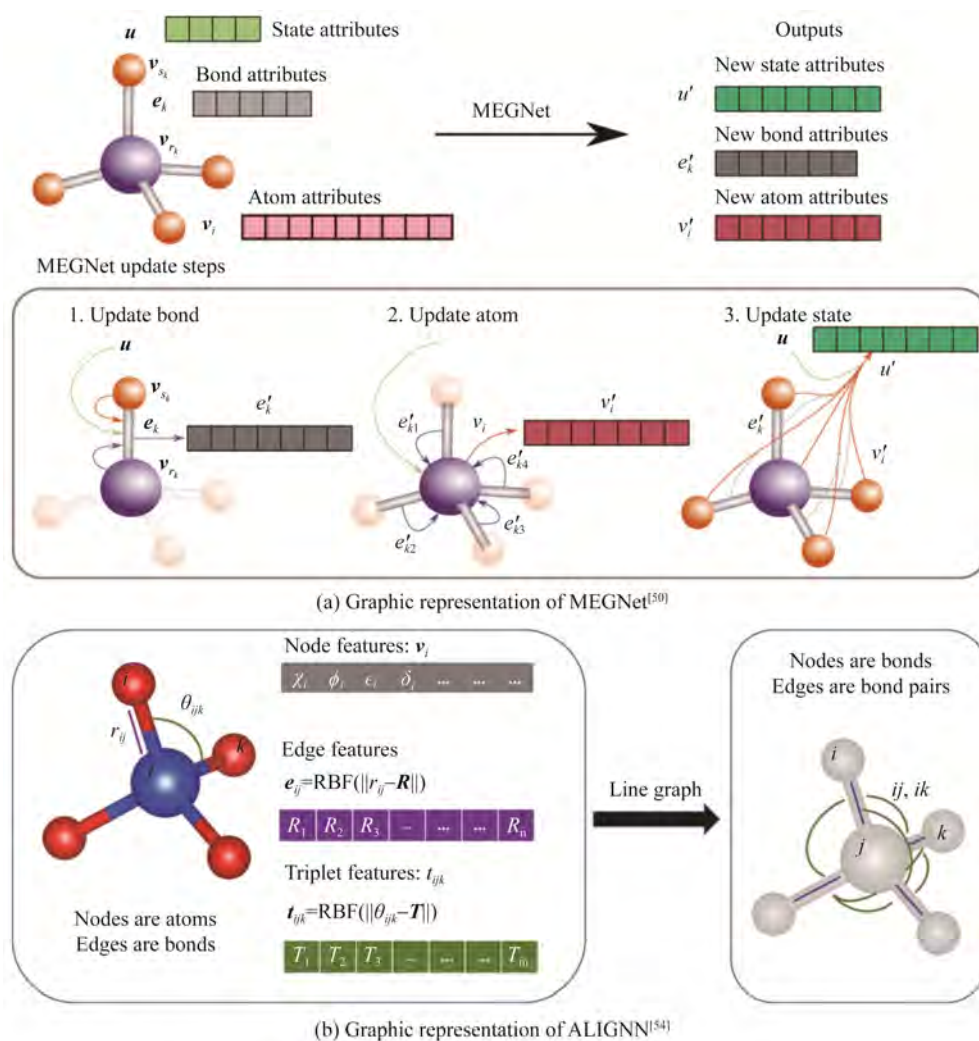


图 6 图表示

Fig. 6 Graphic representation

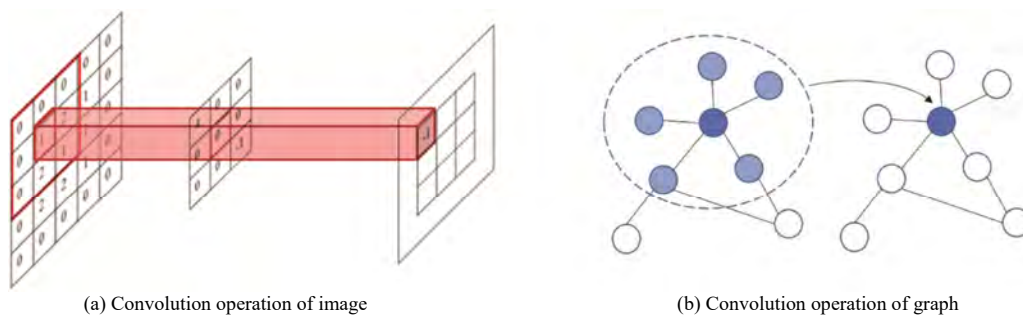


图 7 卷积操作

Fig. 7 Convolution operation

原子与 j 原子间的边的特征: $\sigma(z_{(i,j),k}^{(t)} W_f^{(t)} + b_f^{(t)})$ 为 Sigmoid 函数; $g(z_{(i,j),k}^{(t)} W_s^{(t)} + b_s^{(t)})$ 为激活函数; W_f , W_b 分别为卷积权重矩阵, b_s 和 b_f 分别为偏置。

2) MEGNet 的图卷积与网络结构

MEGNet 的网络结构如图 8b 所示, 其卷积模块不同于 CGCNN 的节点更新, 而是逐一更新边特征 ϕ_e 、节点特征 ϕ_v 和状态特征 ϕ_u 。

3) ALIGNN 的图卷积与网络结构

ALIGNN 图表示与前 2 种方法不同, 除了以原子为节点与原子间距离为边的原子图 g 外, 还有以原子间距离为节点, 三元特征为边的线性图 $L(g)$ 。其网络结构如图 8c 所示, ALIGNN 卷积层首先在原子图 g 卷积, 然后再线性图 $L(g)$ 上卷积, 随后基于更新后的节点特征和边特征进行卷积。

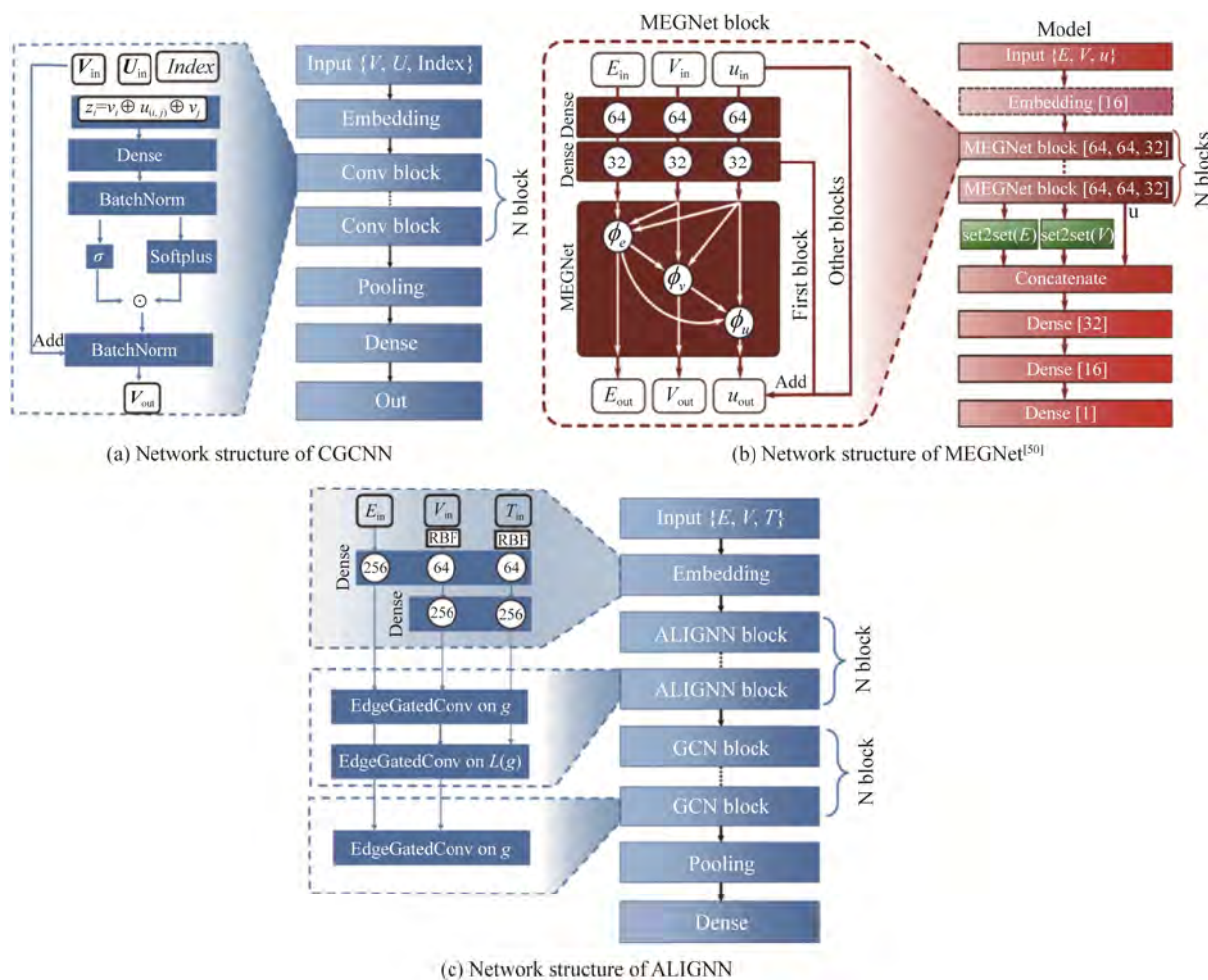


图 8 网络结构

Fig. 8 Network structure

4) Global Attention Graph convolutional neural networks (GATGNN)^[55]的图卷积与网络结构

GATGNN 的卷积操作分为两部分(图 9): 增强图关注层用于提取区域的特征, 全局图关注层用于提取整个晶体图的特征。增强图关注层的公式为:

$$\mathbf{v}_i^k = \sigma \left(\sum_{j \in N(i) \cup j} \alpha_{ij}^{(k)} \mathbf{W}^{(k)} \mathbf{v}_j^{k-1} \right) \quad (3)$$

$$\alpha_{ij}^{(k)} = \text{softmax} \left(g \left(\mathbf{a}_{ij}^{(k)} \right) \right) =$$

$$\text{softmax} \left(g \left(\mathbf{a}^T [\mathbf{W}^{(k)} \mathbf{v}_i^{(k-1)} \parallel \mathbf{W}^{(k)} \mathbf{v}_j^{(k-1)}] \right) \right) \quad (4)$$

其中: k 为卷积层标号; $\mathbf{a}$ 和 $\mathbf{W}$ 分别是权重向量和权重矩阵; $\mathbf{v}_i$ 和 $\mathbf{v}_j$ 为原子 i 和原子 j 的节点特征; $g(z_{(i,j)_k}^{(t)} \mathbf{W}_s^{(t)} + b_s^{(t)})$ 为激活函数, $\parallel$ 表示向量拼接; 其中为进一步加入边特征, 加入了一归一化层 $\mathbf{v}_i^{*(k)} = \mathbf{v}_i^{(k)} \parallel \mathbf{e}_{ij}$, $\mathbf{e}_{ij}$ 为边特征。

GATGNN 中全局图关注层, 基于针对节点 k -means 或 Spectral-clustering 聚类的团簇结果, 集合更新节点特征后解池, 再对所有节点特征求和,

得到全图的特征。最终的特征是全图特征与增强图关注层输出特征的乘积。

1.2.3 池化 池化操作目的在于集合各节点的特征, 输出一个整个图的一维特征, 其需要保持对于节点排序的不变性。最常见的池化方法是各节点归一化求和(CGCNN)或求平均(ALIGNN)的方法, 也有采用对顺序不变的网络作为池化层, 例如 MEGNet 采用了 Set2set 网络实现池化。

1.2.4 图神经网络精度比较 由于各晶体图卷积网络结构的发表时间不同, 使用的数据集也有差异, 表 1 收集了目前研究无机晶体的图神经网络预测 DFT 计算的形成功能和带隙的精度, 括号内显示了所使用的数据集来源与样本数量。

在 MatHub-3D 数据库基础上, 3 个主流图卷积模型在带隙和功率因子的表现(表 2)。对于同一性质带隙, 在不同的数据集里, 3 个模型的趋势不一致。在同一数据集下, 不同的材料性质, 精度最高的模型也不相同, 如 p 型功率因子精度最高的是

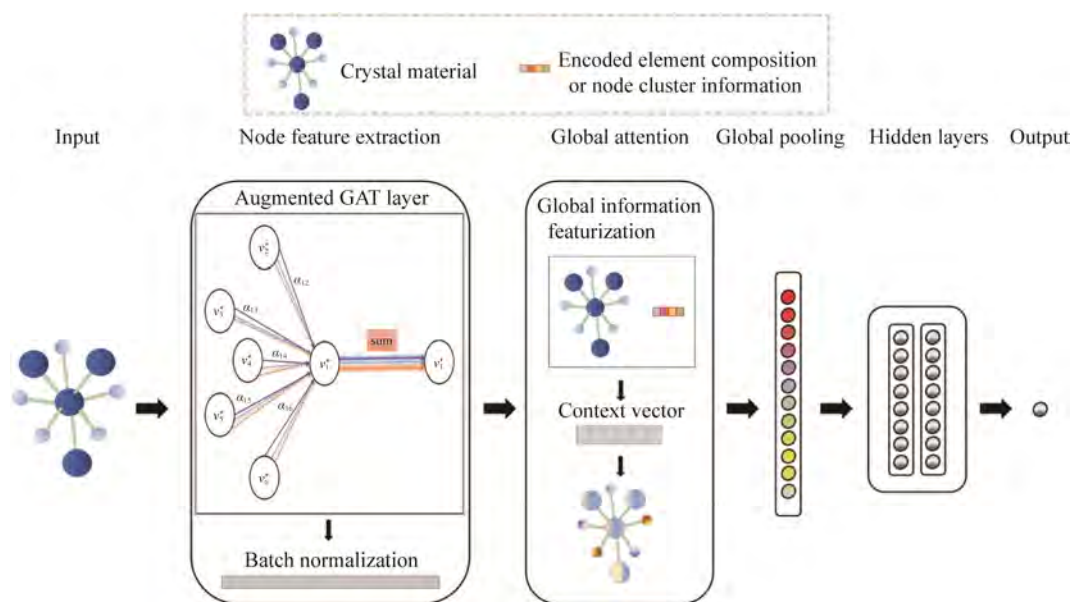
图9 GATGNN 网络结构^[55]Fig. 9 Network structure of GATGNN^[55]

表1 各图卷积神经网络方法精度
Table 1 Accuracies of Graph Neural Networks

Model	Formation energy, E_f (eV·atom ⁻¹)	Bandgap E_g /eV
CGCNN ^[53]	0.039 (MP 28 046) ^[53]	0.39 (MP 16 485) ^[53]
	0.049 (MP 36 839) ^[56]	0.26 (36 837) ^[56]
MEGNet ^[50]	0.028 (MP 60 000) ^[54]	0.33 (MP 36 720) ^[54]
	0.048 (MP 36 839) ^[56]	0.27 (36 837) ^[56]
MPNN ^[49]	0.046 (MP 36 839) ^[56]	0.26 (MP 36 837) ^[56]
SchNet ^[51]	0.035 (MP 60 000) ^[54]	0.28 ((MP 36 837)) ^[56]
	0.050 (MP 36 839) ^[56]	
ALIGNN ^[54]	0.022 (MP 60 000) ^[54]	0.22 (MP 36 720) ^[54]
GATGNN ^[55]	0.039 (MP 28 046) ^[55]	0.32 (MP 16485) ^[55]
	0.048 (MP 60 000) ^[54]	0.31 (MP 36 720) ^[54]
	0.045 (MP 36 839) ^[56]	0.27 (36 837) ^[56]
iCGCNN ^[52]	0.031 (OQMD>200 000)	
DeeperGATGNN ^[56]	0.030 (MP 36 839) ^[56]	0.26 (MP 36 837) ^[56]

表2 CGCNN、MEGNet 和 ALIGNN 在 Mathub-3d 上的平均绝对误差

Table 2 Mean absolute errors of CGCNN, MEGNet and ALIGNN on Mathub-3d

Model	Bandgap E_g /eV	Power factor (p-type)/ (mW·cm ⁻¹ ·K ⁻²)	Power factor (n-type)/ (mW·cm ⁻¹ ·K ⁻²)
CGCNN	0.495	14.853	24.767
MEGNet	0.345	14.843	24.800
ALIGNN	0.378	16.515	21.627

MEGNet 模型, 而 n 型功率因子精度最高的是 ALIGNN 模型。但是整体趋势来看, 对于大数据量和基础的材料性质, 图卷积网络的预测会更准确; 对于影响机理更为复杂的材料性质如功率因子, 其

精度要逊色许多。

2 热电图像机器学习

除了数值数据外, 图像数据也是材料数据中重要的大类, 日益成熟的表征技术提高了实验图像数据获取的效率, 但实验图像分析在传统上需要大量的人工识别操作, 因此实验图像的后处理成为高通量表征分析的一大瓶颈。把基于图像的机器学习方法引入材料研究中, 可以加速和改进基于图像的材料表征技术。由于钢铁的力学性能与其微观结构体积分数、尺寸和形貌有显著的联系, 因此, 实验光学显微图像的机器学习在钢铁领域中被广泛应用, 如 Bulgarevic^[57]等使用机器学习来标记钢中的铁素体/珠光体、铁素体/珠光体/贝氏体和贝氏体/马氏体微观结构。但热电材料的图像机器学习研究工作刚刚起步, 涉及的文献很少。

在 2021 年, Sheng 等^[58]结合了高通量实验和机器学习图像分析方法实现了全流程高通量地寻找新型 Cu–Sn–S 三元热电材料。流程包括 3 个部分: 高通量合成、高通量表征和高通量分析。图 10 为高通量合成和表征的示意图。利用改进的扩散偶高通量合成方法获得成分梯度样品, 该圆柱状样品包含了 9 个不同原材料比(CuS 和 SnS)的片段。随后对每一个片段用扫描电镜进行表征, 每个片段拍摄 11 张背散射电子图像(共 99 张)。

随后, Sheng 等^[58]提出了 2 种不同的机器学习策略来对 99 个背散射电子图像中的不同区域进行

分类, 并进一步识别不同的相。第 1 种策略(图 11a)应用全连接神经网络的主动学习循环, 旨在实现快速和自动的图像分割。主动学习图像分割训练包含注释、人工神经网络训练和模型效果查看 3 个过程, 它们是迭代进行的。第 2 种策略(图 11b)采用无监督

聚类方法, 在没有先验知识的情况下找到可能被忽视的化合物, 提取神经网络最后一个隐藏层的值作为特征进行聚类。通过 2 种机器学习策略分别发现了 2 个新化合物 $\text{Cu}_7\text{Sn}_3\text{S}_{10}$ 和 $\text{Cu}_{1.6}\text{S}$, 且前者的热电优值超过 0.6, 具有良好的热电性能。

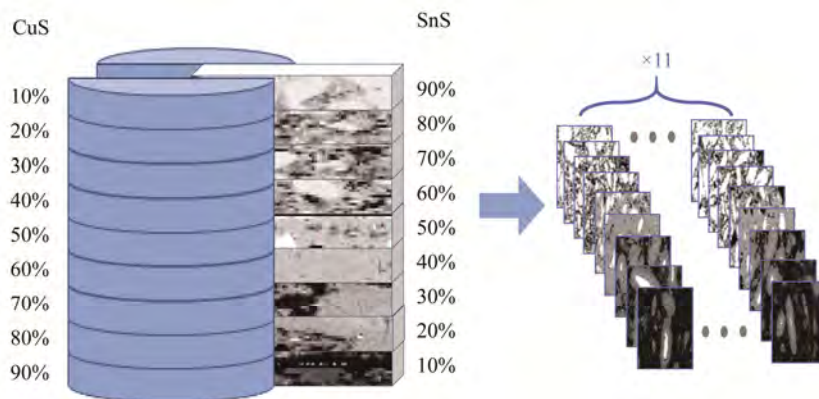


图 10 扩散偶高通量制备的块体样品示意图与其切割后每段的横截面的背散射电子图像^[58]

Fig. 10 Schematic diagram of a bulk sample of a diffusion-couple prepared by a high-throughput method and back-scattered electron image of cross section of each section after cutting^[58]

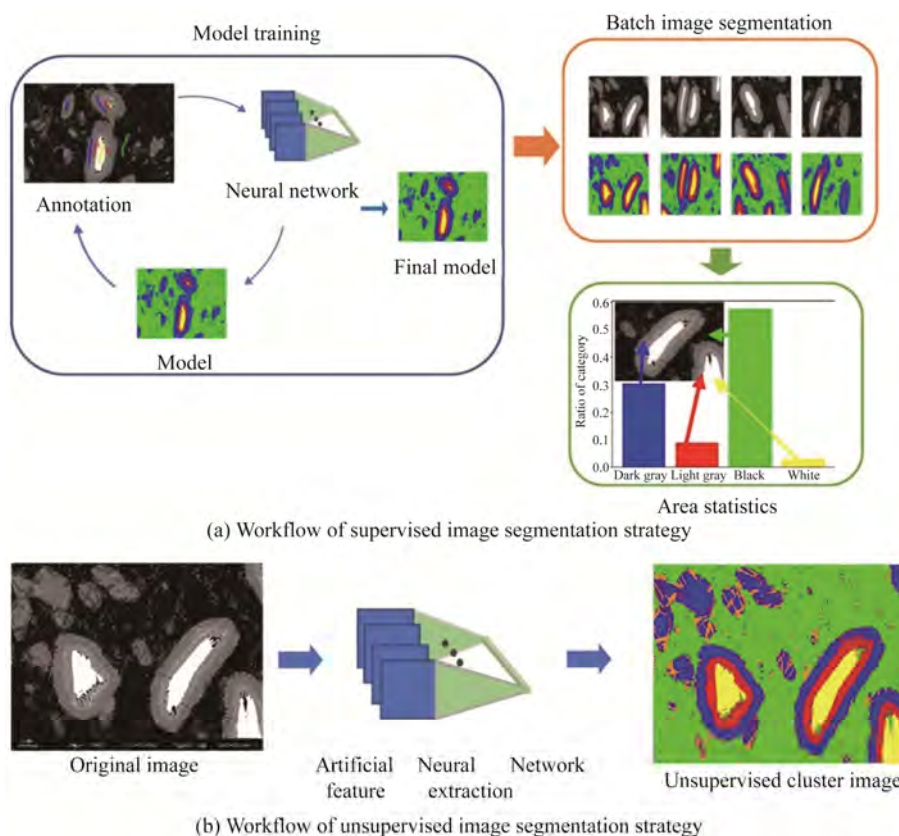


图 11 Cu-Sn-S 背散射图像分割工作流程^[58]

Fig. 11 Workflow of Cu-Sn-S back-scattered electron image segmentation^[58]

3 结论与展望

本文综述了在热电材料中针对不同的数据类型机器学习案例。对于小样本数值数据, 浅层的机器学习方法是一般的常用手段, 主动学习的框架适

用于小样本的机器学习, 特别是需要外推的情况, 可以用尽可能少的样本数据获取全局合理的预测估计。对于大样本数值数据, 早期的由于特征问题, 分析手段局限于统计和可视化。随着深度学习方法 and 图表示方法的引入, 特征问题得以解决, 所有样

本数据可用统一的模型进行学习, 由于深度的神经网络往往有较为复杂的结构, 其拟合精度也有所提高。对于热电材料的图像分析, 这方面的研究较为缺乏。图像的机器学习在热电领域大有可为, 如相分析, 位错定位, 特定微结构识别等。同时, 高通量实验表征技术的快速发展大大提升了需要分析的图片数量, 图像的机器学习正是提升图像分析效率的一个有力手段, 因而图像机器学习在热电领域有着良好的前景。

本文讨论的工作都是正向的机器学习, 也就是用晶体结构/基础元素性质预测化合物相对高阶的性质。人们逐渐不满足于机器学习预测化合物的性质, 而尝试让机器学习产生需要性质的化合物, 也就是材料逆向设计。在逆向设计中主要的机器学习方法有变分自编码器^[59]和生成式对抗网络^[60], 这一领域在热电材料中的研究也刚起步, 但是随着热电数据体量的增加和逆向设计需求的驱动下, 是一个有前景的发展方向。

但是, 机器学习在热电材料中的研究还存在有一些挑战:

1) 对于复杂热电性质从原子和结构级别的机器学习的精度相比于形成能和带隙其精度还需要进一步的提高;

2) 模型的解释性, 相对复杂的模型往往可以有较好的拟合精度但是理解分析困难, 简单模型可解释但精度又有所限制。这种现象在不仅是热电材料在许多材料中都存在^[61–63]。

参考文献:

- [1] BELL L E. Cooling, heating, generating power, and recovering waste heat with thermoelectric systems[J]. *Science*, 2008, 321(5895): 1457–1761.
- [2] SOOTSMAN J R, CHUNG D Y, KANATZIDIS M G. New and old concepts in thermoelectric materials[J]. *Angew Chem Int Ed*, 2009, 48(46): 8616–8639.
- [3] SNYDER G J, TOBERER E S. Complex thermoelectric materials[M]. *Materials for Sustainable Energy*. Co-Published with Macmillan Publishers Ltd, UK, 2010: 101–110.
- [4] FROST R T, CORELLI J C, BALICKI M. Reactor irradiation PbTe, Bi₂Te₃ and ZnSb[J]. *Adv Energy Convers*, 1962, 2: 77–78.
- [5] TANG Y, QIU Y, XI L, et al. Phase diagram of In–Co–Sb system and thermoelectric properties of In-containing skutterudites[J]. *Energy Environ Sci*, 2014, 7(2): 812–819.
- [6] BAI S Q, PEI Y Z, CHEN L D, et al. Enhanced thermoelectric performance of dual-element-filled skutterudites Ba_xCe_yCo₄Sb₁₂[J]. *Acta Mater*, 2009, 57(11): 3135–3139.
- [7] TANG Y, HANUS R, CHEN S-W, et al. Solubility design leading to high figure of merit in low-cost Ce–CoSb₃ skutterudites[J]. *Nat Commun*, 2015, 6(1): 7584.
- [8] NOLAS G, MORELLI D, TRITT T. Skutterudites: A phonon–glass–electron crystal approach to advanced thermoelectric energy conversion applications[J]. *Annual Rev Mater Sci*, 1999, 29: 89–116.
- [9] BOBEV S, SEVOV S C. Clathrates of group 14 with alkali metals: An exploration[J]. *J Solid State Chem*, 2000, 153(1): 92–105.
- [10] HERRMANN R F, TANIGAKI K, KAWAGUCHI T, et al. Electronic structure of Si and Ge gold-doped clathrates[J]. *Phys Rev B*, 1999, 60(19): 13245.
- [11] NOLAS G S, COHN J L, SLACK G A, et al. Semiconducting Ge clathrates: Promising candidates for thermoelectric applications[J]. *Appl Phys Lett*, 1998, 73(2): 178–180.
- [12] SHI X, XI L, FAN J, et al. Cu–Se bond network and thermoelectric compounds with complex diamondlike structure[J]. *Chem Mater*, 2010, 22(22): 6029–6031.
- [13] CULP S R, POON S J, HICKMAN N, et al. Effect of substitutions on the thermoelectric figure of merit of half-Heusler phases at 800 °C[J]. *Appl Phys Lett*, 2006, 88(4): 042106.
- [14] CULP S R, SIMONSON J W, POON S J, et al. (Zr, Hf)Co(Sb, Sn) half-Heusler phases as high-temperature (>700 °C) p-type thermoelectric materials[J]. *Appl Phys Lett*, 2008, 93(2): 022105.
- [15] ZHU T, FU C, XIE H, et al. High efficiency half-Heusler thermoelectric materials for energy harvesting[J]. *Adv Energy Mater*, 2015, 5(19): 1500588.
- [16] CHEN S, REN Z. Recent progress of half-Heusler for moderate temperature thermoelectric applications[J]. *Mater Today*, 2013, 16(10): 387–395.
- [17] XIE W, WEIDENKAF A, TANG X, et al. Recent advances in nanostructured thermoelectric half-Heusler compounds[J]. *Nanomaterials*, 2012, 2(4): 379–412.
- [18] YAN X, JOSHI G, LIU W, et al. Enhanced thermoelectric figure of merit of p-type half-Heuslers[J]. *Nano Lett*, 2011, 11(2): 556–560.
- [19] QIU W, XI L, WEI P, et al. Part-crystalline part-liquid state and rattling-like thermal damping in materials with chemical-bond hierarchy[J]. *Proceed National Academy Sci*, 2014, 111(42): 15031–15035.
- [20] VENKATASUBRAMANIAN R, SHIVOLA E, COLPITTS T, et al. Thin-film thermoelectric devices with high room-temperature figures of merit[J]. *Nature*, 2001, 413(6856): 597–602.
- [21] HSU K F, LOO S, GUO F, et al. Cubic AgPbSbTe₂: Bulk thermoelectric materials with high figure of merit[J]. *Science*, 2004, 303(5659): 818–821.
- [22] PICHANUSAKORN P, BANDARU P. Nanostructured thermoelectrics[J]. *Mater Sci Eng R-Reports*, 2010, 67(2/4): 19–63.
- [23] SHI X, CHEN L, UHER C. Recent advances in high-performance bulk thermoelectric materials[J]. *Int Mater Rev*, 2016, 61(6): 379–415.
- [24] SHI X-L, ZOU J, CHEN Z G. Advanced thermoelectric design: From materials and structures to devices[J]. *Chem Rev*, 2020, 120(15): 7399–7515.
- [25] CHEN W, PÖHL S J H, HAUTIER G, et al. Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment[J]. *J Mater Chem C*, 2016, 4(20): 4414–26.
- [26] RICCI F, CHEN W, AYDEMIR U, et al. An ab initio electronic transport database for inorganic materials[J]. *Sci Data*, 2017, 4: 170085.
- [27] HEDEGAARD E M J, JOHNSEN S, BJERG L, et al. Functionally graded Ge_{1-x}Si_x thermoelectrics by simultaneous band gap and carrier density engineering[J]. *Chem Mater*, 2014, 26(17): 4992–4997.

- [28] HEDEGAARD E M J, MAMAKHEL A A H, REARDON H, et al. Functionally graded $(\text{PbTe})_{1-x}(\text{SnTe})_x$ thermoelectrics[J]. *Chem Mater*, 2017, 30(1): 280–287.
- [29] KOHRI H, NISHIDA I A, SHIOTA I. Improvement of thermoelectric properties for n-type PbTe by adding Ge[J]. *Mater Sci Forum*, 2003, 423–424: 381–4.
- [30] YOU L, LI Z, MA Q, et al. High thermoelectric performance of Cu-doped PbSe-PbS system enabled by high-throughput experimental screening[J]. *Research (Wash D C)*, 2020, 2020: 1736798.
- [31] DENG T, XING T, BROD M K, et al. Discovery of high-performance thermoelectric copper chalcogenide using modified diffusion-couple high-throughput synthesis and automated histogram analysis technique[J]. *Energy Environ Sci*, 2020, 13(9): 3041–53.
- [32] OLIYNYK A O, ADUTWUM L A, RUDYK B W, et al. Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases ABC[J]. *J Am Chem Soc*, 2017, 139(49): 17870–81.
- [33] CHONGHE L, YIHAO T, YINGZHI Z, et al. Prediction of lattice constant in perovskites of GdFeO_3 structure[J]. *J Phys Chem Solids*, 2003, 64(11): 2147–56.
- [34] JAVED S G, KHAN A, MAJID A, et al. Lattice constant prediction of orthorhombic ABO_3 perovskites using support vector machines[J]. *Comput Mater Sci*, 2007, 39(3): 627–34.
- [35] LEE J, SEKO A, SHITARA K, et al. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques[J]. *Phys Rev B*, 2016, 93(11): 115104.
- [36] XU Y, WANG X, LI X, et al. New materials band gap prediction based on the high-throughput calculation and the machine learning[J]. *Scient Sin Technol*, 2018, 49(1): 44–54.
- [37] FURMANCHUK A, SAAL J E, DOAK J W, et al. Prediction of seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach[J]. *J Comput Chem*, 2018, 39(4): 191–202.
- [38] SHENG Y, WU Y, YANG J, et al. Active learning for the power factor prediction in diamond-like thermoelectric materials[J]. *npj Comput Mater*, 2020, 6(1): 171.
- [39] YANG Z, SHENG Y, ZHU C, et al. Accurate and explainable machine learning for the power factors of diamond-like thermoelectric materials[J]. *J Materiomics*, 2022, 8(3): 633–639.
- [40] LI M, CAO G, LUO Y, et al. Predicting the lattice thermal conductivity of alloyed compounds from the perspective of configurational entropy[J]. *npj Comput Mater*, 2022, 8(1): 1–5.
- [41] YANG H, ZHU Y, DONG E, et al. Dual adaptive sampling and machine learning interatomic potentials for modeling materials with chemical bond hierarchy[J]. *Phys Rev B*, 2021, 104(9): 094310.
- [42] JAIN A, ONG S P, HAUTIER G, et al. Commentary: the materials project: A materials genome approach to accelerating materials innovation[J]. *APL Mater*, 2013, 1(1): 011002.
- [43] JAIN A, HAUTIER G, MOORE C J, et al. A high-throughput infrastructure for density functional theory calculations[J]. *Comput Mater Sci*, 2011, 50(8): 2295–2310.
- [44] SETYAWAN W, CURTAROLO S. High-throughput electronic band structure calculations: Challenges and tools[J]. *Comput Mater Sci*, 2010, 49(2): 299–312.
- [45] CURTAROLO S, SETYAWAN W, WANG S, et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations[J]. *Comput Mater Sci*, 2012, 58: 227–235.
- [46] TAYLOR R H, ROSE F, TOHER C, et al. A RESTful API for exchanging materials data in the AFLOWLIB. org consortium[J]. *Comput Mater Sci*, 2014, 93: 178–92.
- [47] SAAL J E, KIRKLIN S, AYKOL M, et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD)[J]. *JOM*, 2013, 65(11): 1501–1509.
- [48] DRAXL C, SCHEFFLER M. NOMAD: The FAIR concept for big data-driven materials science[J]. *MRS Bull*, 2018, 43(9): 676–682.
- [49] GILMER J, SCHOENHOLZ S, RILEY P, et al. Neural Message Passing for Quantum Chemistry[Z]. DOINA P, YEE WHYIE T. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research; PMLR. 2017: 1263–72.
- [50] CHEN C, YE W, ZUO Y, et al. Graph networks as a universal machine learning framework for molecules and crystals[J]. *Chem Mater*, 2019, 31(9): 3564–3572.
- [51] SCHUTT K T, SAUCEDA H E, KINDERMANS P J, et al. SchNet - A deep learning architecture for molecules and materials[J]. *J Chem Phys*, 2018, 148(24): 241722.
- [52] PARK C W, WOLVERTON C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery[J]. *Phys Rev Mater*, 2020, 4(6): 063801.
- [53] XIE T, GROSSMAN J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties[J]. *Phys Rev Lett*, 2018, 120(14): 145301.
- [54] CHOUDHARY K, DECOST B. Atomistic line graph neural network for improved materials property predictions[J]. *NPJ Computational Materials*, 2021, 7(1): 185.
- [55] FUNG V, ZHANG J, JUAREZ E, et al. Benchmarking graph neural networks for materials chemistry[J]. *npj Comput Mater*, 2021, 7(1): 84.
- [56] OMEE S S, LOUIS S Y, FU N, et al. Scalable deeper graph neural networks for high-performance materials property prediction[J]. *Patterns*, 2022, 3(5): 100491.
- [57] BULGAREVICH D S, TSUKAMOTO S, KASUYA T, et al. Pattern recognition with machine learning on optical microscopy images of typical metallurgical microstructures[J]. *Scient Rep*, 2018, 8(1): 2078.
- [58] SHENG Y, DENG T, QIU P, et al. Accelerating the discovery of Cu–Sn–S thermoelectric compounds via high-throughput synthesis, characterization, and machine learning-assisted image analysis[J]. *Chem Mater*, 2021, 33(17): 6918–6924.
- [59] GÓMEZ-BOMBARELLI R, WEI J N, DUVENAUD D, et al. Automatic chemical design using a data-driven continuous representation of molecules[J]. *ACS Central Sci*, 2018, 4(2): 268–276.
- [60] DE CAO N, KIPF T J A P A. MolGAN: An implicit generative model for small molecular graphs[J]. *arXiv preprint*, 2022, 10(4): arXiv:2018.1805.11973.
- [61] LIU Y, GUO B, ZOU X, et al. Machine learning assisted materials design and discovery for rechargeable batteries[J]. *Energy Storage Mater*, 2020, 31: 434–450.
- [62] 施思齐, 涂章伟, 邹欣欣, 等. 数据驱动的机器学习在电化学储能材料研究中的应用[J]. *储能科学与技术*, 2022, 11(3): 739–59.
- SHI Siqu, TU Zhangwei, ZOU Xinxin, et al. *Energy Storage Sci Technol (in Chinese)*, 2022, 11(3): 739–759.
- [63] 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.