

基于机器学习的热导率预测模型研究

牛程程¹ 李少波^{1,2} 但雅波² 曹卓² 李想² 胡建军^{2,3*}

(1.贵州大学现代制造技术教育部重点实验室,贵阳 550025;2.贵州大学机械工程学院,贵阳 550025;
3.美国南卡罗来纳大学计算机科学与工程系,哥伦比亚 29208)

摘要 传统的理论研究、实验研究及计算仿真已无法满足科学家对新材料的探索与设计。数据驱动的机器学习算法对材料的筛选与性能预测有着推动作用。将机器学习算法应用到材料信息学,基于现有材料热导率数据集,建立机器学习热导率预测模型,通过交叉验证来对机器学习回归模型进行评估。利用机器学习算法建立描述符与热导率属性之间的映射模型,可用于大规模的材料筛选,从而指导实验研究。

关键词 数据驱动,机器学习,热导率

Thermal conductivity prediction based on machine learning

Niu Chengcheng¹ Li Shaobo^{1,2} Dan Yabo² Cao Zhuo² Li Xiang² Hu Jianjun^{2,3}

(1.Key Laboratory of Advanced Manufacturing Technology (Ministry of Education),
Guizhou University,Guiyang 550025;2.School of Mechanical Engineering,Guizhou University,
Guiyang 550025;3 Department of Computer Science and Engineering,University of South
Carolina SC,Carolina SC 29208)

Abstract Traditional theoretical research, experimental research and computational simulation have failed to satisfy scientists' exploration and design of new materials. Data-driven machine learning algorithms have a driving role in material screening and performance prediction. The machine learning was applied to material informatics. Based on the existing material thermal conductivity data set, a machine learning thermal conductivity prediction model was developed and evaluated by cross-validation. Machine learning was used to establish a mapping model between descriptors and thermal conductivity properties, which can be used for large-scale material screening to guide experimental research.

Key words data driven, machine learning, thermal conductivity

材料热导率的预测对于设计热障涂层、热电材料到集成电路等领域的材料热性能都是有益的。近年来,研究者通过降低或增加晶格热导率来实现材料导热性能的优化。其中,晶格热导率(LTC) K_L 是导热性能设计中的重要参数。例如,具有闪锌矿或金刚石的高热导率材料在微电子和纳米电子器件中是必不可少的;低热导率材料则构成了新一代热电材料和热障涂层的基础^[1]。

传统的测量方法包括稳态和瞬态方法,在实验过程中,由于热探针和材料之间的热接触仍然存在,导致接触热阻不被消除,从而影响测量结果的准确性;使用热盘热常数分析仪测量热导率的方法不仅需要样品严格处理,而且由于仪器体积大不方便

随时搬运而导致检测不便,难以处理。除了传统的测量方法外,理论计算也是解决材料热导性能的一种方法。Giardino 等^[2]使用密度泛函微扰理论(DFPT)通过获得三阶 IFCs 研究声子线宽,已成功解决玻尔兹曼输运方程(BTE)并预测不同的导热系数^[3-4],由于其中一些参数是从拟合实验中获得^[5-6],这导致降低了计算的预测能力。另一种方法则是基于 Green-Kubo^[7]公式积分微观热流的自相关函数,该公式采用平衡分子动力学模拟计算来获得热导率,但高阶散射过程会导致自相关函数波动大、收敛慢、数值噪声较大,最终迫使热导率计算结果精度较低。还有,Seko 等^[8]使用昂贵的非谐晶格动力学计算来识别低晶格导热材料;Madsen 等^[9]和

基金项目:国家自然科学基金(51741101)

作者简介:牛程程(1993-),女,硕士研究生,主要研究方向为材料信息学。

联系人:胡建军(1973-),男,教授,主要研究方向为智能制造、材料信息学和深度学习。

Toher 等^[10]利用 Debye 模型近似计算包括 K_L 在内的热学属性。

上述材料热导率的获取方法不适用于低成本快速生成和筛选材料属性数据库。最近,机器学习已被用于预测基本材料属性^[11-12],对相关属性进行识别^[13-15]筛选数据并指导实验设计^[16-17]解决材料科学中的问题。在材料科学领域,机器学习加速了用于电介质聚合物^[18-19]、超导材料^[20]、镍基合金^[21]、钙钛矿^[22]、纳米结构^[23]等材料的预测。为了加快材料热导性能的预测能力,本课题组提出了基于机器学习框架模型对材料热导率进行分析预测。

材料的合理设计是现代材料科学与工程的目标。作为实现这一目标的一部分,在材料科学界已经编译了大量的材料特性数据集,以供科学家和工程师对已知材料性质的研究准备。目前,国际材料研究一个非常活跃的领域是使用机器学习从现有材料数据中自动提取预测模型的设计方法。机器学习不是直接依赖于直觉,而是能够通过相关可靠数据来快速、自动地创建精确预测模型。数据驱动的机器学习方法被期望用于建立目标物理特性的预测模型和目标属性类别的分类模型。通过简单计算来精确预测各种材料的热性能属性将有利于缩短实现新材料的时间、减少计算成本。因此,借助机器学习与现有的数据对材料热导性能进行预测,以加快新热导材料的发现。

1 研究方法

通常情况下,建立材料属性机器学习模型的通用方法由 3 部分组成:训练数据、一组描述材料属性的描述符和将特征映射到属性的机器学习算法。

材料学界已经进行了大量的工作来编译材料特性数据集,以提供科学家和研究者的科学实验研究工作。现已证明,这些数据库可以作为创建预测模型和设计新材料的资源。面对数据集过大而导致计算成本过高的事实,笔者从 Materials Project 数据库进行筛选得到一组数据集,其中包括金刚石、闪锌矿、岩盐等结构。但为了避免因小数据集质量差而导致计算局部误差的问题,对该数据进行了数据预处理,减少数据噪音,最终获得一组 232 种材料的热导率数据集(包括具有立方、四方、六方和菱形对称性的材料)。值得注意的是,例如,广泛分布的轻元素 Li、N、O 和 F 对 LTC 的影响通常是可以忽略不计的,数据集中许多由轻元素 Be 和 B 组成的化合物往往显示出高 LTC,而由 Te 和 Bi 组成的 Bi_2Te_3

是一种具有低 LTC 的热电化合物,这些可能因为元素分布而引起热导特性趋势,说明训练数据集在某种程度上是“有偏见”的。由于数据集涵盖着不同的化学成分和晶体结构,因此很自然地会考虑以化学元素和结构的组合表示方法作为描述符。对训练集中的样本进行了数据筛选与特征选择,以特征映射到属性对描述符进行简化。

材料性能预测的一个关键步骤是选择最相关的材料描述符。评估了以下材料特征提取方法:

(1) Magpie 为每种材料计算一组属性,捕获了重要的化学信息,这些信息在决定材料的结构和物理特性方面起着决定性的作用,其中包括化学计量特征(仅取决于元素的比例)、元素属性统计(原子序数、原子半径、熔化温度等)、电子结构属性(s、p、d 和 f 的价电子)和离子化合物特征。(2) Atom2Vec 可以从已知化合物和材料数据库中学习原子的基本属性,且用高维向量对其属性进行表示。材料的输入或特征描述仅取决于预测下的材料特性^[24-26],从而获取原子的某些物理或化学属性。由于缺少材料属性标签,Atom2Vec 自然不会偏向于某个特征,这个数据集以无人监督的方式用于材料学的学习功能。将材料的分子式用原子进行表示,如 Bi_2Se_3 中,2 个 Bi 周围有 3 个 Se,则 Bi 表示为 $(Bi, 2Se_3)$;3 个 Se 周围有 2 个 Bi,则 Se 表示为 $(Se, 3Bi_2)$ 。因此,为捕获原子的属性,将每个原子选作目标生成原子-环境矩阵,学习原子向量以获得材料热导性的相关属性。(3) One-hot 编码为材料属性的每个标签值提供概率数字,将材料属性数据转换成二进制数字,对数据的输入和输出变量保持一致,保证了数据分类的准确性。

本研究中使用 Scikit-learn^[27] 支持向量机(SVR)、随机森林(RF)和核岭回归(KRR)这 3 种回归方法来构建机器学习模型。为了更加准确地预测比较材料热导性,选择一个合适的机器学习算法保证实验结果的稳定性,选择实验结果最佳的 10 折交叉验证来评估几种算法的预测能力。该技术将数据集随机分割 10 份,在其中 9 份上进行训练模型,并尝试预测剩余集合的属性。每份数据在每一次重复过程中作为测试集,且模型预测能力取所有重复过程中的平均值。使用以上几种机器学习算法预测材料的热导率,并将预测结果与自动 GIBBS^[10] 库(AGL)用于高通量计算的实验值进行比较。

最后,通过使用 10 折交叉验证对支持向量机(SVM)、随机森林(RF)和核岭回归(KRR)这 3 种

机器学习算法进行评估,选取最优模型对材料热导性进行预测。为了对所建立模型性能进行量化评价,以决定系数 R^2 和均方根误差 $RMSE$ 这两个统计指标来衡量模型的效果。若 R^2 值越高且 $RMSE$ 值越低,即认定回归模型的精度越高,材料热导性预测更精确。

决定系数 R^2 为相关系数 r 的平方值,表明预测值解释实际值变差的程度,计算公式见式(1)。

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (1)$$

式中, y_i 为样本 i 的测试值; $\bar{y}$ 为 y_i 的平均值; $\hat{y}_i$ 为样本 i 的预测值; n 为样本数。

均方根误差($RMSE$)是检验测试值和实验值的符合精确度,计算公式见式(2)。

$$RMSE = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - y_i')^2} \quad (2)$$

式中, y_i' 为样本 i 的高通量计算的实验值。

2 结果与讨论

为了说明机器学习算法对材料性能预测的优势,结合以下几种材料特征抽取工具对实验结果进行比较。用于信息学和探索的材料不可知论平台^[28-29](Magpie)用于计算验证属性,并执行材料搜索;对离散化后的数据进行独热编码(One-hot Encoding),使得数据分类获得更为准确地预测;由 Zhou 等^[30]提出来的原子向量(Atom2Vec)学习算法从原子-环境矩阵中提取原子知识,并在原子向量中编码学习属性;卡拉威模型(Callaway Model)^[6,31]被认为是在对声子玻尔兹曼输运方程的弛豫时间近似(RTA)的改进,被广泛用于研究晶格热导率。

首先,通过 Magpie 来对相关材料特征属性补充知识源,获得了 132 种特征属性数据集。在分析各特征参数与该属性的相关性基础上,在此环境下对材料属性进行计算与预测。经 10 折验证后,RF 模型下的预测评估值更优。

然而,Atom2Vec 仅考虑材料数据库中化合物的存在,不考虑材料的其他特定属性。Atom2Vec 将每个原子选作目标类型,从原子和环境间的相互关联中提取信息,然后以向量的形式进行表示。输入数据集中的化学分子式得到一个 232×21 特征向量矩阵,通过从原子-环境矩阵中提取相关原子知识对材料的热导性进行预测。虽然预测效果不如

Magpie 预测结果,但也证明了该模型对材料属性有一定的预测。

One-hot 编码作为一种传统的特征选择方法,利用二进制将特征值转化成数字并对特征信息进行提取,在一定程度上扩充了材料信息特征。但由于数据集的局限性而无法得到更好的预测结果。并且此环境下的预测值不如其他特征环境下的预测结果。

尽管原则上可以精确计算晶格热导率,但缺乏晶体的振动光谱和晶体中非谐振力而导致玻尔兹曼方程难以精确克服。因此,要借助机器学习算法对玻尔兹曼方程实验进行改进,从而提高对晶格热导率的快速预测。在该工作中假设所有声子散射过程可以由作为频率和温度函数的弛豫时间表示。通过 Callaway Model 计算公式的分析,借助声速 v_s 、原子个数 n 、Gruneisen 参数 γ 、体积 V 等特征属性来对材料进行热导性能预测,如式(3)所示。其中,声速 v_s 近似于弹性模量 B 和密度 d 比值的算术平方根,如式(4)所示。这样不仅可以节约计算时间和昂贵的费用,同时也提高了材料预测的准确性。由评估预测可知,RF 较优于其余两种算法。

$$k_L = A_1 \frac{M v_s^y}{T \gamma^2 V^2 n^x} + \frac{3k_B}{2} \left(\frac{\pi}{6}\right)^{1/3} \frac{v_s}{V^2} \left(1 - \frac{1}{n^{2/3}}\right) \quad (3)$$

$$v_s \approx \sqrt{\frac{B}{d}} \quad (4)$$

将预处理后的数据集通过 SVM、RF 和 KRR 回归方法建立预测模型,经 10 折交叉验证后,对各实验环境下的输出值进行比较。为了确保实验结果的稳定性,在训练时每组算法均采用了 10 次 10 折交叉验证求平均值的方法,并得出实验结果,见表 1 和表 2。

表 1 各环境下机器学习预测模型的 R^2 值

	SVM	RF	KRR
Magpie	0.6643	0.7129	0.5752
Atom2Vec	0.4535	0.5865	0.4313
One-hot	0.5730	0.6248	0.3327
Callaway Model	0.7224	0.7261	0.7182

表 2 各环境下机器学习预测模型的 $RMSE$ 值

	SVM	RF	KRR
Magpie	54.3008	41.2650	62.0740
Atom2Vec	86.9464	60.9510	86.8005
One-hot	67.9520	52.9804	100.4141
Callaway Model	52.4803	51.7262	52.7120

由表1和表2可见,RF模型下的预测评估值更优。在决定系数 R^2 值的比较中,结合Callaway Model计算公式特征属性的预测评估值 R^2 和RMSE分别为0.7261和51.7262,预测结果详见图1。经验证,4组环境下的实验结果,均显示机器学习与材料热导率之间存在一定的相关性。

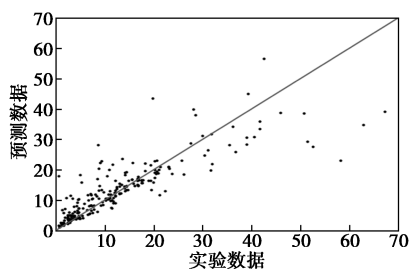


图1 Callaway Model 特征 RF 模型下的预测评估图

3 结论

提出了一种基于机器学习的材料热导率预测方法,并通过实验证明此方法在材料热导率的预测上有较高的准确率。但是,由于数据集规模的局限性,预测性能并没有达到最理想的结果。这促使我们对下一步研究工作进行规划。本课题组认为,在描述符的表示方法和学习模型架构的选择上需要更全面的探索与研究。基于机器学习对材料热导率的研究方法仍有很大的进步空间。

参考文献

- [1] Toberer E S, Zevalkink A, Snyder G J. Phonon engineering through crystal chemistry[J]. *Journal of Materials Chemistry*, 2011, 21(40): 15843-15852.
- [2] Giardino D, Bettio D, Simoni G. Ab initio calculation of the linewidth of various phonon modes in germanium and silicon[J]. *Physical Review B*, 2003, 67(14): 1393-1406.
- [3] Brodido D A, Malorny M, Birner G, et al. Intrinsic lattice thermal conductivity of semiconductors from first principles[J]. *Applied Physics Letters*, 2007, 91(23): 231922.
- [4] Li W, Mingo N, Lindsay L, et al. Thermal conductivity of diamond nanowires from first principles[J]. *Physical Review B*, 2012, 85(19): 195436.
- [5] Allen P B. Zero-point and isotope shifts; relation to thermal shifts[J]. *Philosophical Magazine Part B*, 1994, 70(3): 527-534.
- [6] Callaway J. Model for lattice thermal conductivity at low temperatures[J]. *Physical Review*, 1959, 113(4): 436-441.
- [7] Rowley R L, Painter M M. Diffusion and viscosity equations of state for a Lennard-Jones fluid obtained from molecular dynamics simulations[J]. *International Journal of Thermophysics*, 1997, 18(5): 1109-1121.
- [8] Seko A, Togo A, Hayashi H, et al. Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization[J]. *Physical Review Letters*, 2015, 115(20): 205901.
- [9] Madsen G K H, Katre A, Bera C. Calculating the thermal conductivity of the silicon clathrates using the quasi-harmonic approximation[J]. *Physica Status Solidi*, 2016, 213(3): 802-807.
- [10] Toher C, Plata J J, Levy O, et al. High-throughput computational screening of thermal conductivity, Debye temperature, and Gruneisen parameter using a quasiharmonic Debye model[J]. *Physical Review B*, 2014, 90(17): 174107.
- [11] De J M, Chen W, Notestine R, et al. A statistical learning framework for materials science; application to elastic moduli of k-nary inorganic polycrystalline compounds[J]. *Sci Rep*, 2016, 6: 34256.
- [12] Medasani B, Gamst A, Ding H, et al. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning[J]. *npj Computational Materials*, 2016, 2(1): 1.
- [13] Ma X, Li Z, Achenie L E, et al. Machine-learning-augmented chemisorption model for CO₂ electroreduction catalyst screening[J]. *Journal of Physical Chemistry Letters*, 2015, 6(18): 3528.
- [14] Oliynyk A O, Antono E, Sparks T D, et al. High-throughput machine-learning-driven synthesis of full-heusler compounds[J]. *Chemistry of Materials*, 2016, 28(20): 7324-7331.
- [15] Sendek A D, Yang Q, Cubuk E D, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials[J]. *Energy & Environmental Science*, 2016, 10(1): 306-320.
- [16] Dey P, Bible J, Datta S, et al. Informatics-aided bandgap engineering for solar materials[J]. *Computational Materials Science*, 2014, 83: 185-195.
- [17] Xue D, Balachandran P V, Hogden J, et al. Accelerated search for materials with targeted properties by adaptive design[J]. *Nature Communications*, 2016, 7: 11241.
- [18] Ghanshyam P, Wang C, Jiang X, et al. Accelerating materials property predictions using machine learning[J]. *Scientific Reports*, 2013, 3(9): 2810.
- [19] Mannodikanakkithodi A, Pilania G, Huan T D, et al. Machine learning strategy for accelerated design of polymer dielectrics[J]. *Scientific Reports*, 2016, 6: 20952.
- [20] Stanev V, Oses C, Kusne A G, et al. Machine learning modeling of superconducting critical temperature[J]. *npj Computational Materials*, 2018, 4(1): 29.
- [21] Conduit B D, Jones N G, Stone H J, et al. Design of a nickel-base superalloy using a neural network[J]. *Materials & Design*, 2017, 131: 358-365.
- [22] Schmidt J, Shi J M, Borlido P, et al. Predicting the thermodynamic stability of solids combining Density functional theory and machine learning[J]. *Chemistry of Materials*, 2017, 29(12): 5090-5103.