

机器学习优化 Clausius-Mossotti 介电常数计算方程

秦锦成¹, 马名生¹, 张发强¹, 李永祥², 刘志甫¹

(1. 中国科学院上海硅酸盐研究所, 关键陶瓷材料全国重点实验室, 上海 201899;

2. 皇家墨尔本理工大学工程学院, 澳大利亚 墨尔本 3000)

摘 要: 介电常数的机器学习预测模型准确性虽然较经典 Clausius-Mossotti 方程更高, 但模型解释性仍需提升。本工作提出了领域知识和机器学习双向嵌入的方法, 在 C-M 方程中引入机器学习修正项以提升方程准确性。以机器学习模型所识别出的介电常数关键参量作为修正项的自变量, 结合符号回归和遗传算法生成修正项的数学表达式。基于 Pareto 前沿选出了平衡精准度和复杂度的结果, 并采用偏导数揭示了“低值补偿、高值抑制”的极化纠偏机制。对于采用 Shannon 极化率数据的介电常数计算, 修正后的 C-M 方程 R^2 提升了 42.30%, RMSE 下降了 14.72%。本工作初步探索了“双向嵌入”方法在兼顾模型准确性和解释性方面的潜力, 并尝试为材料设计提供新的思考角度。

关键词: 介电常数; Clausius-Mossotti 方程; 机器学习; 符号回归; 微波介质陶瓷

中图分类号: TB3; TP3 文献标志码: A 文章编号: 0454-5648(2025)09-2643-08

网络出版时间: 2025-08-12



Clausius-Mossotti(C-M)方程作为预测介电常数最经典的理论模型之一, 因其形式简洁、物理内涵清晰且准确性尚可, 尽管已经提出 1 个多世纪之久, 至今仍广泛应用于介质材料研究中。该方程基于 Lorentz 有效场近似和恒定极化假设, 建立了宏观介电常数与微观极化率和晶胞体积之间的关系^[1-3]:

$$\frac{\epsilon_r - 1}{\epsilon_r + 2} = \frac{4\pi\alpha_D^M}{3V_m} \quad (1)$$

式中: V_m 为原胞体积; α_D^M 为单个分子的介电极化率, 可通过求算所组成离子的介电极化率之和获得。然而, 随着国内外研究者在微波介质陶瓷的构效关系研究中不断深入, 发现介电常数的影响因素远不止分子的介电极化率和原胞体积。研究表明, 键长、键能、键价和、对称性、有序度、离子堆积密度等材料属性与介电常数之间存在深刻的关联^[4-10]。这些研究揭示了 C-M 方程的局限性, 并为构建更全面和精确的介电常数预测模型提供了研究依据。

近年来, 机器学习方法在介电性能研究中取得了显著进展, 提出了系列介电常数预测模型, 其准确性往往优于经典 C-M 方程。在构建介电常数模型时, 研究者通常采用领域知识嵌入的机器学习方法, 即将电介质理论知识系统地整合到数据预处理、特征工程、模型选择和模型约束等机器学习过程中。尤其在特征工程阶段, 利用领域知识选择或构建具有电介质物理意义的特征参量, 以提升模型解释性, 发掘影响介电常数的关键因素^[11-13]。基于离散傅里叶变换(DFT)计算数据, Takahashi 等^[14]利用随机森林算法发现了平均原子质量及理论密度是决定电子介电常数(ϵ_{el})的关键特征; 主量子数的标准差和平均近邻距离变化量则显著影响离子介电常数(ϵ_{ion})。Morita 等^[15]通过支持向量机算法也识别出了 ϵ_{el} 值与理论密度呈强正相关, 且与带隙呈负相关。针对 ABO₃ 型钙钛矿材料, Noda 等^[16]通过偏最小二乘回归模型揭示了离子半径、Bader 电荷、径向分布函数和态密度与介电常数存在密切关联。此外,

收稿日期: 2025-01-15。 修订日期: 2025-02-12。

基金项目: 国家自然科学基金联合基金(U24A2052); 国家资助博士后研究人员计划 C 档(GZC20232825); 上海市东方英才计划(QNKJ2024026)。

第一作者: 秦锦成(1995—), 男, 博士。

通信作者: 刘志甫(1975—), 男, 博士, 研究员;

马名生(1986—), 男, 博士, 研究员。

Received date: 2025-01-15. Revised date: 2025-02-12.

First author: QIN Jincheng (1995—), male, Ph.D.

E-mail: qinjincheng@mail.sic.ac.cn

Correspondent author: LIU Zhifu (1975—), male, Ph.D., Professor.

MA Mingsheng (1986—), male, Ph.D., Professor.

E-mail: liuzf@mail.sic.ac.cn; mamingsheng@mail.sic.ac.cn

禁带宽度、理论密度、费米能级、原子半径、电负性、电离能等特征也对 ABO_3 材料的介电常数有重要影响^[17–18]。对于尖晶石结构材料, Liu 等^[19]通过 XGBoost 模型确定了极化率、电负性、原子半径和价电子数等特征量对介电常数有显著影响。本团队在先前工作中, 利用 SVR.RBF 模型发现决定介电常数的 3 个关键特征量: 单位体积分子介电极化率(ppv); 键长平均值(blm); 单位原子的晶胞体积(va)^[2], 其中: 特征量 ppv 定义为分子介电极化率(α_D^M)与原胞体积(V_m)之比, 正是 C–M 方程计算介电常数的唯一参数, 即:

$$\epsilon_r = \frac{1 + \frac{8}{3}\pi \frac{\alpha_D^M}{V_m}}{1 - \frac{4}{3}\pi \frac{\alpha_D^M}{V_m}} = \frac{1 + \frac{8}{3}\pi \times \text{ppv}}{1 - \frac{4}{3}\pi \times \text{ppv}} \quad (2)$$

机器学习模型验证了传统介电理论的正确性, 同时发现了特征量 blm 和 va 对介电常数也存在贡献。

上述介电常数模型研究均属于领域知识嵌入机器学习的范畴, 而将机器学习模型提炼的新知识反向嵌入物理方程的研究鲜见报道。机器学习模型通常为“黑箱模型”, 虽然准确性较高, 但解释性较差; 而物理方程基于理论推导, 解释性较好, 但准确性往往不如机器学习模型, 亟须提出两面兼顾的新思想和方法^[20–23]。物理方程可通过引入机器学习修正项进行优化, 即该修正项为模型所识别并提取出的关键特征参量的函数, 再通过算法构建数学表达式以拟合修正项, 最终实现理论方程的优化。这种新方法旨在平衡理论方程的准确性和解释性, 在材料领域中具有广阔的应用前景, 但目前研究仍然匮乏。符号回归基于遗传算法, 能够发掘数据之间的解析关系, 并自驱动地迭代生成数学公式, 且无须预设函数形式或先验假设。相比传统的线性回归、LASSO、KAN 等方法, 该方法在揭示复杂非线性关系方面更具解释性^[24–27]。

本工作在领域知识嵌入的介电常数机器学习模型所发现的关键特征参量基础上, 进一步借助符号回归对领域知识——经典 C–M 方程进行修正。基于符号回归结合遗传算法的研究框架, 拟合修正项并构建一系列的候选数学表达式。通过 Pareto 前沿分析, 筛选出在精准度和复杂度之间达到平衡的表达式, 不仅显著提升了介电常数计算结果的可靠性, 而且同时保留了方程的物理意义。通过分析 ppv 修正项对 3 个特征量的偏导数, 揭示其变化规律及修正机制。本工作初步展现了领域知识与机器学习“双

向嵌入”的方法在平衡模型准确性与解释性方面的应用潜力, 为材料设计提供新的思考角度。

1 研究方法

1.1 数据集

本工作采用的介电常数数据均来源于已发表的文献, 针对的是单相介质陶瓷, 且均通过 Hakki-Coleman 方法在 5 ~ 18 GHz 频段下进行测试^[28]; 忽略了测试环境、设备型号和样品状态等因素, 因为这些因素对测试结果影响较小。在微波测试频段下, 偶极子转向极化和空间电荷极化通常已经完全退出响应, 因此晶界、杂质等缺陷对介电常数影响不大。陶瓷材料中由于孔隙的存在会让实测介电常数偏低, 因此, 介电常数需经过 Penn 修正公式排除孔隙的影响^[29]:

$$\epsilon_r^{\text{exp}} = \epsilon_r \left(1 - \frac{3P(\epsilon_r - 1)}{2\epsilon_r + 1} \right) \quad (3)$$

式中: ϵ_r^{exp} 为陶瓷介电常数的实验测试值; P 为孔隙率; ϵ_r 为孔隙修正后的介电常数。孔隙率 $P = 1 - \rho_r$, ρ_r 为相对密度。若未给出相对密度, 则采用默认值 95%。对于同种材料, 尽可能收集不同文献的数据。特征量选取的是团队先前领域知识嵌入的机器学习研究中发现的介电常数关键参数, 包括 ppv 和 blm 及 va。特征量的原始数据主要来源于 Materials Project 数据库网站, 并利用 pymatgen 工具实现数据批量自动下载^[30–31]。经过数据清洗排除离群值和重复值, 最终保留了 297 种单相微波介质陶瓷。

本研究目标是修正 C–M 方程, 因此目标量并非直接定为介电常数, 而是 ppv 的残差。因为特征量 ppv, 即 α_D^M 与 V_m 之比是 C–M 方程计算介电常数的唯一参数。其中, 化合物的分子介电极化率可根据加和法则, 通过加和离子介电极化率(α_D)计算得出, 而目前最为广泛接受的 α_D 数据库是由 Shannon 于 1993 年提出的^[3]。本团队在此基础上, 进一步提出了优化的 α_D 数据库^[28]。在本工作中, 前者称为 Shannon 数据库, 后者称为 Optimized 数据库, 因此基于前者和后者计算的 ppv 分别用 ppv_s 和 ppv_o 表示。在明确了特征量 blm 和 va 对介电常数也存在贡献的基础上, 保留了 C–M 方程的形式及物理意义, 并引入修正项(Δppv)拟合 ppv 残差。修正项是与 ppv、blm 和 va 有关的函数, 将通过符号回归方法解析出其表达式。为了便于算法准确捕捉和描述目标量的变化规律, 计算过程中将 ppv 残差放大 1000 倍。待计算完成后, 再将所得结果还原至原始比例。

最终, 经过修正的 C-M 方程表达为:

$$\varepsilon_r = \frac{1 + \frac{8}{3}\pi \times (\text{ppv}_{\text{cal}} + \Delta\text{ppv})}{1 - \frac{4}{3}\pi \times (\text{ppv}_{\text{cal}} + \Delta\text{ppv})} \quad (4)$$

1.2 符号回归

符号回归是一种可解释的机器学习方法, 旨在通过数据驱动的方式自动构建数学表达式^[26-27, 32-33]。该方法通过组合和变换基本函数与运算符拟合目标量并生成表达式, 利用进化算法优化其结构与参数, 以最小化预测误差。在本工作中, 我们选择采用 Python 语言的 `gplearn` 包实现基于遗传算法的符号回归, 基本流程如图 1 所示。算法最初随机生成 5000 个数学表达式, 构建成初代种群。这些表达式的自变量即特征量 `ppv`、`blm` 和 `va`, 而因变量则为 `Δppv`。算法随机从数据集中选取 70% 的数据用于模型训练, 建立数学表达式; 剩余的 30% 用于模型测试, 通过计算平均绝对误差(MAE)评估表达式的准确性。

完成初代模型训练与测试之后, 保留 20% 最准确的表达式, 并在它们之间执行遗传与变异操作。在符号回归中, 数学表达式可通过符号表达式和表达式树的形式表示, 如图 2 所示。针对二叉树结构, 遗传变异操作通常包含 4 种, 即提升变异法、点变异法、子树变异法和交叉法。提升变异法随机选择表达式中的 1 个子树, 并用该子树的某个子树替换原本子树, 简化结构; 点变异法随机替换表达式中的某个节点或常数, 保持结构不变; 子树变异法随机替换表达式中的某个子树, 引入新的结构; 交叉法随机选择 2 个表达式交换部分子树, 生成新的后代。经过遗传变异, 上一代优胜表达式的特征得以保留, 并从中生成 1000 个表达式加入下一代。同时, 随机生成 4000 个新表达式, 以确保新代总数达到 5000 个。上述优化过程不断迭代, 直到满足终止条件(MAE<5 或遗传代数>20)。

符号回归结果的准确性、复杂度, 以及算法的收敛速度受到模型超参数的影响, 因此, 需要通过网格搜索法进行超参数寻优。为确保优势特征的高效延续, 交叉概率 `pc` 设置为较大值, 范围为 0.5~0.9, 步长 0.02, 共有 20 个值; 子树变异概率 `ps` 范围为 $(1-\text{pc})/3 \sim (1.06-\text{pc})/3$, 步长 0.002, 共有 10 个值; 提升变异概率 `ph` 设为等于 `ps`, 在引入新结构同时控制表达式的复杂度; 点变异概率则为 $1-\text{pc}-\text{ps}-\text{ph}$, 确保概率总和为 1。另外, 还引入节俭系数限制表达式过度膨胀。对于误差较大的 Shannon 数据库, 节俭系数设置范围为 0.01~0.04, 步长 0.01, 共有 4

个值; 对于误差较小的 Optimized 数据库, 则设置为 0.001~0.004, 步长 0.001, 共有 4 个值, 以避免表达式过于简化而牺牲准确性。因此, 网格搜索包含 $20 \times 10 \times 4 = 800$ 种超参数组合, 而每种组合有 20 代, 最终产生 16 000 个表达式。运算符包含, 加(add)、减(sub)、乘(mul)、除(div)、平方根(sqrt)和倒数(inv), 而常数项的数值范围采用默认值(-1, 1)。

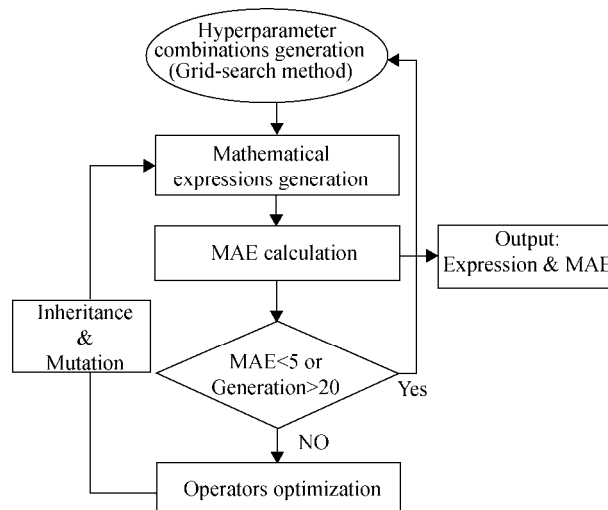


图 1 基于遗传算法的符号回归流程图

Fig. 1 Flowchart of symbolic regression based on genetic algorithm

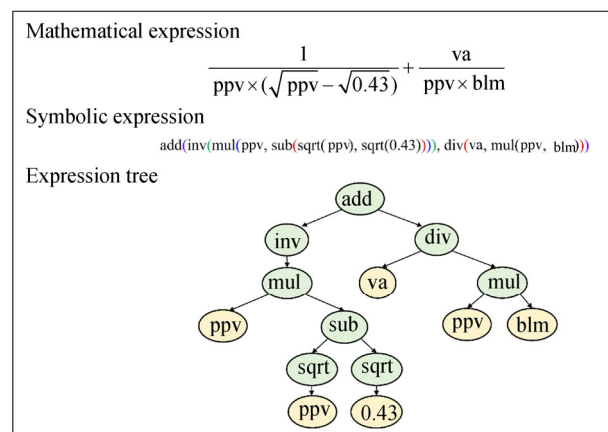


图 2 符号回归表达式及其树结构示意图

Fig. 2 Schematic diagram of symbolic regression expression and its tree structure

2 结果与讨论

采用上述方法生成了 16 000 个修正项 `Δppv` 的数学表达式, 同时输出了其误差值和复杂度。图 3 展示了基于 Shannon 数据库构建的修正 C-M 方程的复杂度和 MAE 之间的权衡, 颜色深浅表示该范围内表达式的个数。位于 Pareto 前沿的数学表达式, 表明其在准确性和复杂度之间达到了最优平衡。这

表明在不增加复杂度的情况下,无法进一步提高其准确性;同时,在不降低准确性的情况下,也无法进一步简化其复杂度。考察了其中 5 个表达式,如表 1 所示。表 1 中显示的复杂度是表达式的原始复杂度,其数值由运算符的数量决定。然而,部分原始表达式中出现了诸如 $\text{sub}(\text{va}, \text{va})$ 的冗余结构,其运算结果恒为 0,对结果准确性无实质影响,但增加了不必要的复杂度。因此,表达式简化后的复杂度与原始复杂度可能不一致。综合考虑准确性、复杂度和解释性,选择了表达式 C 为基于 Shannon 数据库的 Δppv 最优表达式。采用相同的策略,基于 Optimized 数据库建立的 Δppv 最优表达式为

$$\Delta\text{ppv} = [2\text{ppv}_0 \times (\text{blm} - \frac{\text{va}}{0.199 \times \sqrt{\text{ppv}_0}})]^{-1} \quad (5)$$

表 1 基于 Shannon 数据库的 Pareto 前沿数学表达式

Table 1 Mathematical formulas based on Shannon database at the Pareto front

Point	Complexity	MAE	Formula
A	9	7.81	$\frac{(\text{ppv}_s^2 - \text{blm} \times \text{ppv}_s + 1) \times (0.297 \times \text{blm} - \text{va})}{0.297 \times \text{ppv}_s \times \text{va}}$
B	11	7.34	$\frac{\text{va} - \text{ppv}_s - \text{blm} \times \sqrt{\text{ppv}_s} \times (20 \times \sqrt{\text{ppv}_s} + 1)}{\sqrt{\text{ppv}_s}}$
C	13	6.99	$\frac{1}{\text{ppv}_s \times (\sqrt{\text{ppv}_s} - \sqrt{0.43})} + \frac{\text{va}}{\text{ppv}_s \times \text{blm}}$
D	17	6.84	$\frac{\text{va}^{1/2}}{\text{ppv}_s^{3/2}} - \frac{\text{va}^{1/4}}{\text{ppv}_s^{5/4}} + \frac{1}{\text{ppv}_s^{3/2} - \text{ppv}_s \times \sqrt{0.43}}$
E	23	6.80	$\sqrt{\frac{\text{va} \times \text{blm}^{1/2}}{0.522} + \text{ppv}_s^{-1/8}} \times (\sqrt{\frac{\text{va}}{\text{ppv}_s}} + \frac{1}{\sqrt{\text{ppv}_s} - \sqrt{0.43}} - \text{blm} - 0.37)$

图 4 为 C-M 方程修正前后预测能力的对比。对于 Shannon 数据库,引入修正项对 C-M 方程准确性有较明显的提升,其决定系数 R^2 从 0.392 0 提升至

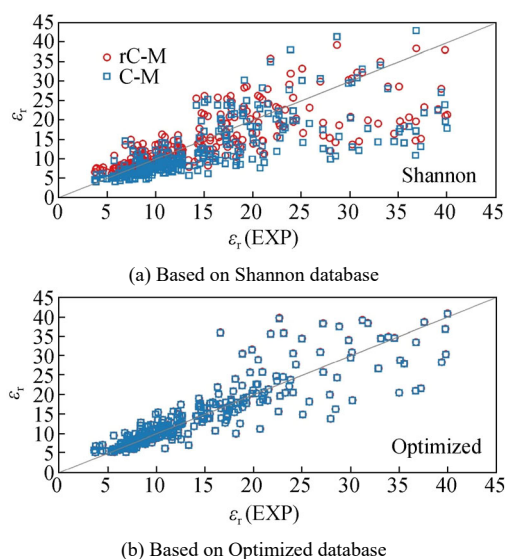


图 4 C-M 方程修正前后预测能力对比

Fig. 4 Predictability of the C-M equation and revised C-M equation

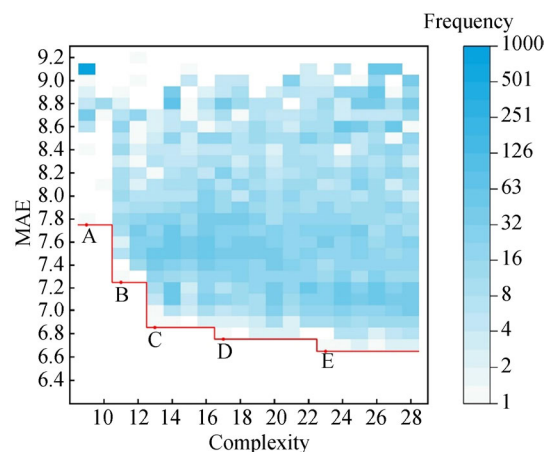


图 3 数学表达式复杂度与 MAE 的 Pareto 前沿

Fig. 3 Pareto frontier between complexity and MAE of mathematical formula

0.557 8, 而均方根误差(RMSE)从 6.519 9 降低至 5.560 3。由于建模数据集的特征量 ppv_s 、 blm 和 va 数值范围分别为 0.121 9~0.222 8、1.63~2.98 Å 和 8.40~20.67 Å³, 因此,在该范围内的预测结果可靠性较高。然而,对于 Optimized 数据库,修正前的 C-M 方程准确性已达到 $R^2 = 0.566 9$, $\text{RMSE} = 5.502 9$, 而修正后在准确性上几乎没有变化($R^2 = 0.568 8$, $\text{RMSE} = 5.491 0$)。这表明该数据库已经较为充分地考虑了晶体结构对极化率的影响,因此难以通过 blm 和 va 对其进行进一步优化。

针对基于 Shannon 数据库建立的 C-M 方程修正项,对 Δppv_s 与 3 个特征量之间的关系进行了可视化,如图 5 所示。计算结果大体上能够反映出理想值的变化趋势,但对于较小值拟合欠佳。从整体上进行分析, Δppv_s 与 ppv_s 、 blm 呈负相关,与 va 表现出较弱的正相关性。随着 ppv_s 增大, Δppv_s 逐渐减小且由正数过渡到负数。其中, Δppv_s 介于 $\pm 1 \times 10^{-3}$ 之间的样本共 19 个,对应的 ppv_s 平均值约为 0.208 7。当 $\text{ppv}_s < 0.208 7$ 时,Shannon 值往往低估

了实际极化率, 需提升该值以修正误差; 反之, 当 $ppv_s > 0.2087$ 时, 高极化率材料可能存在被高估的风险, 则需降低该值实现纠偏。尽管上述分析归纳了修正项“低值补偿、高值抑制”的纠偏机制, 然而对于不同特征量对修正项的贡献程度及特征量之间关系的定量解析, 还需结合偏导数计算进行深入分析。

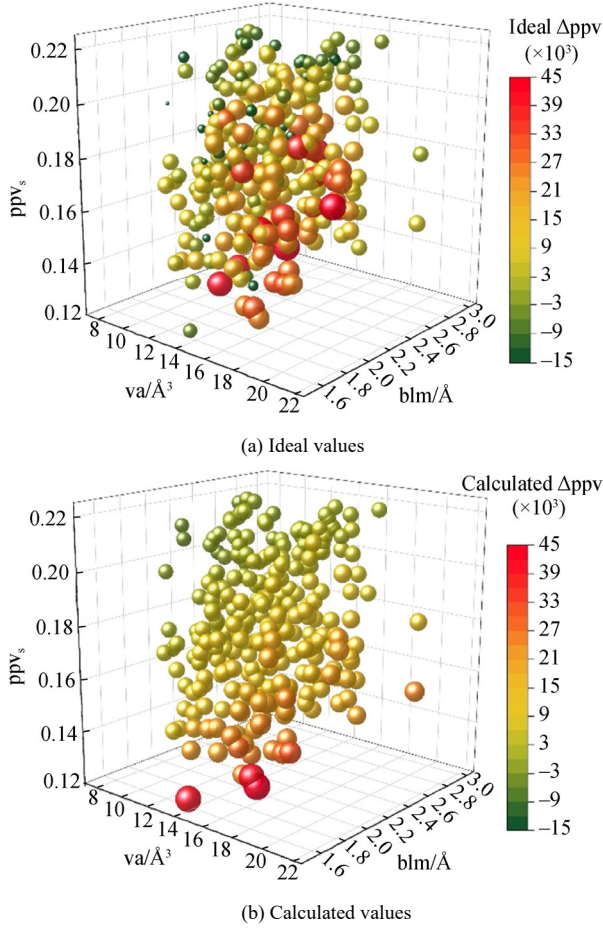


图 5 Δppv_s 与特征量的关系
Fig. 5 Relationships between Δppv_s and features

图 6 为修正项 Δppv_s 对特征量 ppv_s 、 blm 和 va 计算的一阶偏导和二阶偏导。一阶偏导数用于描述多元函数在某一点处沿某一变量方向的变化率。 Δppv_s 分别对特征量计算的一阶偏导为

$$\frac{\partial \Delta ppv_s}{\partial ppv_s} = \frac{0.66 - 1.5\sqrt{ppv_s}}{ppv_s^2(\sqrt{ppv_s} - 0.66)^2} - \frac{va}{blm \cdot ppv_s^2} \quad (6)$$

$$\frac{\partial \Delta ppv_s}{\partial blm} = -\frac{va}{ppv_s \cdot blm^2} \quad (7)$$

$$\frac{\partial \Delta ppv_s}{\partial va} = \frac{1}{ppv_s \cdot blm} \quad (8)$$

由于 3 个特征量均为正数, 因此恒有 $\frac{\partial \Delta ppv_s}{\partial blm} < 0$ 且

$\frac{\partial \Delta ppv_s}{\partial va} > 0$ 。而对于数据集中的材料, 特征量 ppv_s 、 blm 和 va 范围分别为 $0.1219 \sim 0.2228$ 、 $1.63 \sim 2.98 \text{ \AA}$ 和 $8.40 \sim 20.67 \text{ \AA}^3$, 因此 $\frac{\partial \Delta ppv_s}{\partial ppv_s}$ 、 $\frac{\partial \Delta ppv_s}{\partial blm}$ 和 $\frac{\partial \Delta ppv_s}{\partial va}$ 数值范围分别为 $-451.49 \sim -105.71$ 、 $-41.46 \sim -8.42$ 和 $1.55 \sim 5.04$; 平均值分别为 -180.23 、 -15.17 和 2.55 ; 中位数分别为 -166.05 、 -14.76 和 2.49 。由此可见, Δppv_s 对 ppv_s 单调递减特性也反映了“低值补偿、高值抑制”的纠偏机制。同时, 该结果也反映了 ppv_s 对修正项贡献度最大, 远高于 blm , 而 va 的贡献最弱, 符合前述分析。通常, 化学键的键长反映化学键强度及稳定性, 降低键长能有效抑制离子位

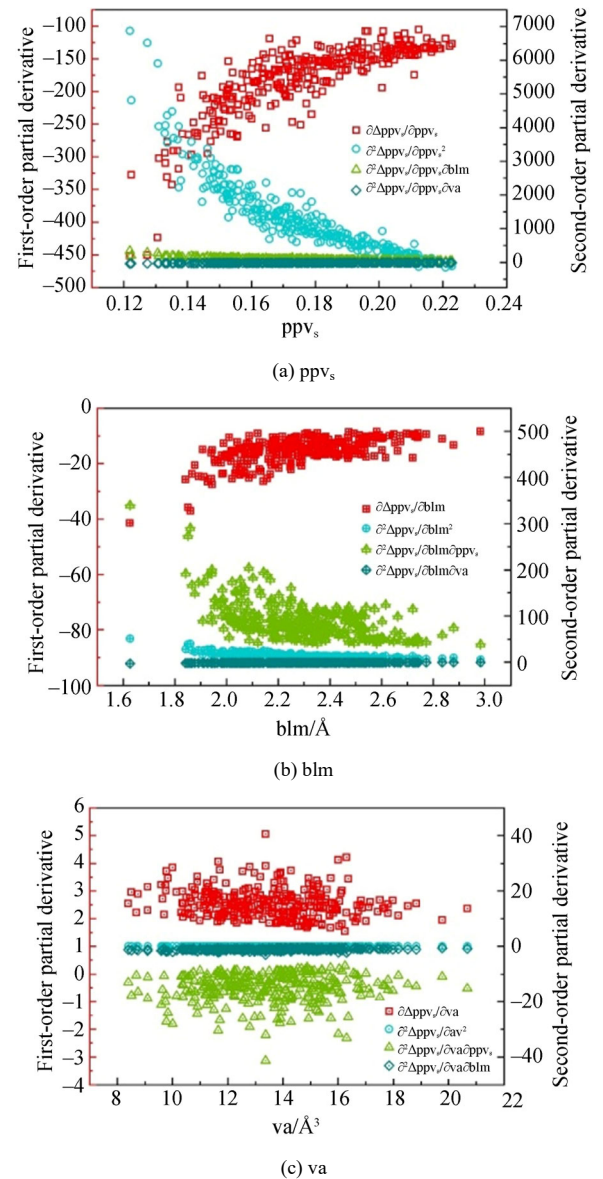


图 6 修正项对特征量 ppv_s 、 blm 和 va 的偏导
Fig. 6 Partial derivatives of correction term to the features of ppv_s 、 blm and va

移极化；自由体积越大对介电极化有稀释作用，也会降低实际极化率。然而， Δppv_s 对 blm 和 va 的一阶偏导结果却与这 2 个规律相反。传统理论通常探讨的是单一因素对介电极化的直接作用，而较少考虑多因素的非线性相互作用。在特征量中， blm 和 va 可能并非直接影响极化率，而是通过更复杂的物理机制，如对晶体结构对称性、离子堆积模式、局域电场及化学键性质等因素进行修正，间接影响介电行为。对表 1 中其余公式进行一阶偏导计算，也发现同样的规律。

二阶偏导数反映了多元函数在某一点处关于某一变量变化率的变化快慢。修正项 Δppv_s 分别对 ppv_s 、 blm 和 va 求二阶偏导，结果为

$$\frac{\partial^2 \Delta ppv_s}{\partial ppv_s^2} = \frac{-5.90 ppv_s + 3.75 ppv_s \sqrt{ppv_s} + 3.12 \sqrt{ppv_s} - 0.56}{ppv_s^3 (\sqrt{ppv_s} - 0.66)^4} + \frac{2va}{blm \cdot ppv_s^3} \quad (9)$$

$$\frac{\partial^2 \Delta ppv_s}{\partial blm^2} = \frac{2va}{ppv_s \cdot blm^3} \quad (10)$$

$$\frac{\partial^2 \Delta ppv_s}{\partial va^2} = 0 \quad (11)$$

在数据集中， $\frac{\partial^2 \Delta ppv_s}{\partial ppv_s^2}$ 和 $\frac{\partial^2 \Delta ppv_s}{\partial blm^2}$ 范围分别为

$-147.95 \sim 6871.20$ 和 $5.64 \sim 50.98$ ，而 $\frac{\partial^2 \Delta ppv_s}{\partial va^2}$ 则恒为

零。这个结果揭示了 Δppv_s 随 ppv_s 和 blm 的变化是非线性的，而随 va 变化是线性的。在数据集中，

$\frac{\partial^2 \Delta ppv_s}{\partial ppv_s^2}$ 几乎均为正数，仅 8 个样本其值为负数，

而该值介于 ± 25 之间的样本共 5 个，对应 ppv_s 平均值约为 0.215 8，与前述 Δppv_s 由正转负的临界点 (0.208 7) 接近。综合分析，在 ppv_s 临界点前，修正项的补偿作用逐渐减弱，而在临界点后极化率降低效应逐渐增强。类似的，结合一阶和二阶偏导可以发现，随着 blm 增大， Δppv_s 下降速度在逐渐减缓；而 va 增大， Δppv_s 则线性增大。

修正项 Δppv_s 在数据集范围内连续可微，二阶混合偏导数的求导顺序对结果无影响，因此有

$$\frac{\partial^2 \Delta ppv_s}{\partial ppv_s \partial blm} = \frac{\partial^2 \Delta ppv_s}{\partial blm \partial ppv_s} = \frac{va}{blm^2 \cdot ppv_s^2} \quad (12)$$

$$\frac{\partial^2 \Delta ppv_s}{\partial ppv_s \partial va} = \frac{\partial^2 \Delta ppv_s}{\partial va \partial ppv_s} = -\frac{1}{blm \cdot ppv_s^2} \quad (13)$$

$$\frac{\partial^2 \Delta ppv_s}{\partial blm \partial va} = \frac{\partial^2 \Delta ppv_s}{\partial va \partial blm} = -\frac{1}{ppv_s \cdot blm^2} \quad (14)$$

在数据集中， $\frac{\partial^2 \Delta ppv_s}{\partial ppv_s \partial blm}$ 和 $\frac{\partial^2 \Delta ppv_s}{\partial ppv_s \partial va}$ 和

$\frac{\partial^2 \Delta ppv_s}{\partial blm \partial va}$ 范围分别为 38.80 ~ 340.09、-41.37 ~ -7.12

和 -3.10 ~ -0.52，进一步印证了修正项的非线性变化规律受 blm 和 va 的影响较小。

3 结论

通过结合领域知识与机器学习的“双向嵌入”方法，在经典 Clausius-Mossotti 介电常数计算方程中引入修正项，在兼顾方程解释性的同时提升其预测精度。修正项的自变量为本团队先前建立的机器学习模型所识别的介电常数 3 个关键参量，即 ppv 、 blm 、 va 。结合符号回归和遗传算法拟合 ppv 的残差，并采用不同超参数组合生成了 16000 个数学表达式。通过 Pareto 前沿分析，最终选择了在精准度与复杂度间达到最佳平衡的修正方程。结果表明，修正后的 C-M 方程在采用 Shannon 极化率数据时，决定系数 R^2 从 0.392 0 提升至 0.557 8，而均方根误差 RMSE 从 6.519 9 降低至 5.560 3，证明了修正结果的可靠性。而对于采用 Optimized 极化数据计算介电常数，C-M 方程修正前后精度接近。结果分析表明，Shannon 值通常低估了低极化率材料 ($ppv_s < 0.208 7$) 的 ppv_s 值，而对于高极化率材料，该值则可能存在被高估的风险。因此， Δppv_s 的修正是通过“低值补偿、高值抑制”的纠偏机制实现的，即对低估的数据施加补偿性提升，对高估的数据施加抑制性降低。一阶偏导数表明， ppv_s 对 Δppv_s 贡献度远高于 blm 和 va ；二阶偏导数则表明，随着 ppv_s 提升，修正项的补偿作用逐渐减弱，随后抑制效应逐渐增强。本工作初步探讨了“双向嵌入”方法的应用潜力，表明能在一定程度上构建兼顾准确性和解释性的预测模型。这一结果不仅为经典物理方程的优化带来了新的思考角度，也为新材料的设计与性能预测探索了新的工具和方法。

参考文献:

- [1] KIRKWOOD J G. On the theory of dielectric polarization[J]. 1936, 4(9): 592–601.
- [2] QIN J C, LIU Z F, MA M S, et al. Machine learning approaches for permittivity prediction and rational design of microwave dielectric ceramics[J]. J Materiomics, 2021, 7(6): 1284–1293.
- [3] SHANNON R D. Dielectric polarizabilities of ions in oxides and fluorides[J]. 1993, 73(1): 348–366.

- [4] SONG X Q, ZOU Z Y, LU W Z, et al. Crystal structure, lattice energy and microwave dielectric properties of melilite-type $\text{Ba}_{1-x}\text{Sr}_x\text{Cu}_2\text{Si}_2\text{O}_7$ solid solutions[J]. *J Alloys Compd*, 2020, 835: 155340.
- [5] KIM E S, CHUN B S, FREER R, et al. Effects of packing fraction and bond valence on microwave dielectric properties of $\text{A}^{2+}\text{B}^{6+}\text{O}_4$ (A^{2+} : Ca, Pb, Ba; B^{6+} : Mo, W) ceramics[J]. *J Eur Ceram Soc*, 2010, 30(7): 1731–1736.
- [6] WILSON G J, CHAN R K, DAVIDSON D W, et al. Dielectric properties of ices II, III, V, and VI[J]. 1965, 43(7): 2384–2391.
- [7] QIN J C, LIU Z F, MA M S, et al. Structure and microwave dielectric properties of gillespite-type $\text{ACuSi}_4\text{O}_{10}$ ($\text{A} = \text{Ca, Sr, Ba}$) ceramics and quantitative prediction of the $Q \times f$ value via machine learning[J]. *ACS Appl Mater Interfaces*, 2021, 13(15): 17817–17826.
- [8] WU S, SONG K X, LIU P, et al. Effect of TiO_2 doping on the structure and microwave dielectric properties of cordierite ceramics[J]. *J Am Ceram Soc*, 2015, 98(6): 1842–1847.
- [9] 唐莹, 相怀成, 李洁, 等. $\text{Ca}_{3-x}\text{Mg}_x\text{Yb}_2\text{Ge}_3\text{O}_{12}$ ($0 \leq x \leq 3$) 石榴石的 A 位 Rattling 效应与微波介电性能[J]. *硅酸盐学报*, 2023, 51(4): 872–881. TANG Ying, XIANG Huaicheng, LI Jie, et al. *J Chin Ceram Soc*, 2023, 51(4): 872–881.
- [10] 杜康, 尹长志, 杨佳庆, 等. Ge^{4+} 取代对 $\text{BaSnSi}_3\text{O}_9$ 陶瓷烧结行为和微波介电性能的优化[J]. *硅酸盐学报*, 2023, 51(4): 866–871. DU Kang, YIN Changzhi, YANG Jiaqing, et al. *J Chin Ceram Soc*, 2023, 51(4): 866–871.
- [11] MURDOCK R J, KAUWE S K, WANG A Y, et al. Is domain knowledge necessary for machine learning materials properties?[J]. *Integr Mater Manuf Innov*, 2020, 9(3): 221–227.
- [12] 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.
- [13] 秦锦成, 刘志甫, 马名生, 等. 数据驱动范式下微波介质陶瓷材料及器件研究进展[J]. *硅酸盐学报*, 2023, 51(4): 921–933. QIN Jincheng, LIU Zhifu, MA Mingsheng, et al. *J Chin Ceram Soc*, 2023, 51(4): 921–933.
- [14] TAKAHASHI A, KUMAGAI Y, MIYAMOTO J, et al. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations[J]. *Phys Rev Mater*, 2020, 4(10): 103801.
- [15] MORITA K, DAVIES D W, BUTLER K T, et al. Modeling the dielectric constants of crystals using machine learning[J]. *J Chem Phys*, 2020, 153(2): 024503.
- [16] NODA Y, OTAKE M, NAKAYAMA M. Descriptors for dielectric constants of perovskite-type oxides by materials informatics with first-principles density functional theory[J]. *Sci Technol Adv Mater*, 2020, 21(1): 92–99.
- [17] LIN X Y, LI C J, HAO H, et al. Accelerated search for ABO_3 -type the electronic contribution of polycrystalline dielectric constants by machine learning[J]. *Comput Mater Sci*, 2021, 193: 110404.
- [18] YAO T Z, CHEN L P, HU N, et al. Discovering ABO_3 -type perovskite with different dielectric constants via intelligent optimization algorithm[J]. *AIP Adv*, 2024, 14(7): 075012.
- [19] LIU X B, SU C, HUANG Q X, et al. Machine learning enhanced prediction of permittivity of spinel microwave dielectric ceramics compared to traditional C–M calculation[J]. *Modelling Simul Mater Sci Eng*, 2024, 32(3): 035002.
- [20] RUDIN C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead[J]. *Nat Mach Intell*, 2019, 1(5): 206–215.
- [21] 孙中体, 李珍珠, 程观剑, 等. 机器学习在材料设计方面的研究进展[J]. *科学通报*, 2019, 64(32): 3270–3275. SUN Zhongti, LI Zhenzhu, CHENG Guanlian, et al. *Chin Sci Bull*, 2019, 64(32): 3270–3275.
- [22] YUAN R H, LIU Z, BALACHANDRAN P V, et al. Accelerated discovery of large electrostrains in BaTiO_3 -based piezoelectrics using active learning[J]. *Adv Mater*, 2018, 30(7): 1702884.
- [23] JIANG L, FU H D, ZHANG H T, et al. Physical mechanism interpretation of polycrystalline metals' yield strength via a data-driven method: A novel Hall–Petch relationship[J]. *Acta Mater*, 2022, 231: 117868.
- [24] BUTLER K T, DAVIES D W, CARTWRIGHT H, et al. Machine learning for molecular and materials science[J]. *Nature*, 2018, 559(7715): 547–555.
- [25] 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.
- [26] GUO W J, WANG L Y, ZHU L A, et al. Exploration of high-temperature oxidation resistance laws in ultra-high temperature boride ceramics through data-driven approaches[J]. *Corros Sci*, 2024, 230: 111943.
- [27] LEE J A, FIGUEIREDO R B, PARK H, et al. Unveiling yield strength of metallic materials using physics-enhanced machine learning under diverse experimental conditions[J]. *Acta Mater*, 2024, 275: 120046.
- [28] QIN J C, LIU Z F, MA M S, et al. Optimizing and extending ion dielectric polarizability database for microwave frequencies using machine learning methods[J]. *NPJ Comput Mater*, 2023, 9: 132.
- [29] PENN S J, ALFORD N M, TEMPLETON A, et al. Effect of porosity and grain size on the microwave dielectric properties of sintered alumina[J]. *J Am Ceram Soc*, 1997, 80(7): 1885–1888.
- [30] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation[J]. 2013, 1(1): 011002.
- [31] ONG S P, RICHARDS W D, JAIN A, et al. Python materials genomics (pymatgen): A robust, open-source Python library for materials analysis[J]. *Comput Mater Sci*, 2013, 68: 314–319.
- [32] KOZA J R. Genetic programming as a means for programming computers by natural selection[J]. *Stat Comput*, 1994, 4(2): 87–112.
- [33] ANGELIS D, SOFOS F, KARAKASIDIS T E. Artificial intelligence in physical sciences: Symbolic regression trends and perspectives[J]. *Arch Comput Methods Eng*, 2023(30): 3845–3865.

Machine Learning-Based Optimization of Clausius-Mossotti Equation for Dielectric Constant Calculation

QIN Jincheng¹, MA Mingsheng¹, ZHANG Faqiang¹, LI Yongxiang², LIU Zhifu¹

(1. State Key Laboratory of High Performance Ceramics and Superfine Microstructure, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 201899, China; 2. School of Engineering, RMIT University, Melbourne 3000, Australia)

Extended Abstract

Introduction The Clausius-Mossotti (C–M) equation is one of the most fundamental theoretical models to calculate the dielectric constant of materials. As was proposed over a century ago, it has been widely applied in dielectric material research due to its relatively simple mathematical form and clear physical interpretation. The equation, derived based on the Lorentz effective field approximation and the assumption of constant polarization, establishes a relationship among the macroscopic dielectric constant and the microscopic polarizability and unit cell volume. However, with the growing understanding of dielectric materials, it becomes clear that dielectric constant is affected by a range of additional factors, such as bond length, packing density and symmetry. These factors are not captured by the conventional C–M equation, indicating the limitations of the equation. Recently, machine learning (ML) methods show a promise in improving the prediction accuracy of dielectric properties. The ML models can outperform classical equations in terms of accuracy, but they often lack interpretability, which hinders their widespread use in material design and optimization. To address these issues, this study proposed a novel bidirectional embedding approach with domain knowledge and machine learning. The ML-based correction term was integrated into the classical C–M equation, enhancing its predictive accuracy, while retaining physical interpretability. This work could explore a potential of combining physical formulas with ML methods to create more robust and explainable models for material design.

Methods The data of single-phase microwave dielectric ceramics were collected from published literatures and the Materials Project database. The dataset included materials whose dielectric constants were measured by the Hakki-Coleman method in a frequencies range from 5 to 18 GHz. A machine learning model identified molecular dielectric polarizability per volume (ppv), average bond length (blm) and unit cell volume (va) as the most significant contributors to the dielectric constant in a previous study. To enhance the accuracy of the C–M equation, correction terms were introduced based on the three key features. Symbolic regression, combined with genetic algorithms, was used to derive mathematical expressions for the correction terms. The symbolic regression technique utilized evolutionary algorithms to generate and evolve potential expressions. The structures and parameters of expressions were optimized to minimize prediction errors. Hyperparameters were optimized by a grid search method to identify the best-performing models. The process was carried out iteratively. A total of 16,000 candidate mathematical expressions were generated and compared. Finally, the Pareto front analysis was used to select the optimal expression that balances accuracy and complexity.

Results and Discussion The revised C–M equation results in significant improvements in its predictive performance. For the Shannon polarizability dataset, the R^2 value of the revised C–M equation is increased by 42.29%, and the RMSE value is decreased by 14.72%. The correction term compensates for the inaccuracies of the original equation, particularly in cases where the molecular dielectric polarizability per volume of Shannon value (ppv_s) is either overestimated or underestimated. For low-polarizability materials with ppv_s values less than 0.2087, the Shannon database tends to underestimate the dielectric constant, and the correction term compensates for this error. Conversely, for high-polarizability materials, the dielectric constant is often overestimated, and the correction term reduces the overestimation. The analysis of the relationship between the correction term (Δppv_s) and features reveals that the correction term is most sensitive to ppv_s, followed by blm, with va having the least impact. This finding indicates that the dielectric polarizability plays a dominant role in determining the accuracy of the revised C–M equation. The influence of the features on the correction term is further explored through partial derivatives. The first-order partial derivatives show that the correction term's contribution from ppv_s is much greater than that from blm or va. The second-order partial derivatives reveal a non-linear relationship between the correction term and ppv_s and blm, while the relationship with va is linear. The results indicate that the compensatory effect of the correction term gradually diminishes, and the suppression effect becomes more pronounced as ppv_s increases. This behavior is consistent with the correction mechanism of “elevating the underrated, reducing the overrated”.

Conclusions This study introduced a bidirectional embedding approach that could integrate machine learning with domain knowledge to enhance both the accuracy and interpretability of the Clausius-Mossotti equation. The revised equation achieved a higher prediction accuracy with a 42.29% increase in R^2 and a 14.72% decrease in RMSE via incorporating the correction term derived from symbolic regression. The revised equation retained the physical meaning of the original one. The research could highlight a potential of bidirectional embedding approach to create models that could balance both accuracy and interpretability, offering a promising perspective on optimizing classical equations and material design.

Keywords dielectric constant; clausius-Mossotti equation; machine learning; symbolic regression; microwave dielectric ceramics