项目编号：_IPP28101_

上海交通大學

“大学生创新实践计划”
项目研究论文

论文题目：___基于机器学习的芯片互联材料性能预测方法__

项目负责人：__曹宇轩__　学院（系）：　___材料科学与工程_

指 导 教 师：___杭弢___　学院（系）：　__材料科学与工程__

参与学生：_____曹宇轩_____

项目执行时间：_2023 年_11_月 至　_2024_年_10_月

摘要

随着摩尔定律失效，芯片内部互连已成为先进半导体制程研究的热点。通过筛选在尺寸微缩条件下具有高电导率优势的新型互连材料，可降低互连延时和损耗，从而支持芯片性能、能效与可靠性提升。为此，本项目应用机器学习材料研发范式，结合使用图神经网络、前馈神经网络，采用 ρ λ 作为衡量材料微缩潜力的指标，以文献数据与第一性原理计算结果作为支撑，并借助特征提取、特征筛选策略，训练机器学习模型以高通量筛选候选互连材料。项目还借助支持向量机（Support Vector Regression, SVR）、岭回归等机器学习模型，综合运用多种特征工程手段，建立了对 MAX 体系芯片互联材料的快速性能预测模型，在抗电迁移性能上具有良好预测表现。本项目为机器学习在材料领域的深入应用提供了大量工作经验。

关键词：机器学习，芯片互联，材料性能预测，图神经网络，MAX 化合物

ABSTRACT

With the obsolescence of Moore's Law, interconnects within chips have become a hot topic in advanced semiconductor process research. By screening new interconnect materials with high electrical conductivity advantages under size reduction conditions, it is possible to reduce interconnect delay and loss, thereby supporting the enhancement of chip performance, energy efficiency, and reliability. To this end, this project applies a machine learning material development paradigm, combining graph neural networks and feedforward neural networks. It uses ρλ as an indicator to measure the potential for material miniaturization, supported by literature data and first-principles calculation results. With feature extraction and feature selection strategies, machine learning models are trained to high-throughput screen candidate interconnect materials. The project also leverages machine learning models such as Support Vector Regression (SVR) and ridge regression, employing a variety of feature engineering techniques to establish a rapid performance prediction model for MAX system chip interconnect materials, showing good predictive performance in electromigration resistance. In addition, this project provides a wealth of work experience for the in-depth application of machine learning in the field of materials.

KEY WORDS： Machine learning, Interconnect, Materials properties prediction, GNN, MAX phase

1 绪论

后道工艺（Back-end-of-line, BEOL），是指在集成电路制造过程中，位于前道工艺（Front-end-of-line, FEOL）之后的工艺序列，主要涉及芯片内部的互连技术。这一阶段的核心任务是通过形成多层金属化线路和通孔（via），实现晶体管层与外部电路的电连接，以及不同晶体管之间或不同金属层之间的互连。随着集成电路特征尺寸的不断缩小和集成度的提高，后道工艺在芯片制造中的重要性日益凸显。当前，芯片互连线的进一步尺寸微缩主要受到了以下制约：互连线横截面的进一步缩小导致的电阻增大问题；随电阻 R 增大导致的信号 RC 延迟问题；线路微缩、比表面积增加，和阻挡层的减薄所带来的电迁移可靠性问题；以及封装工艺、功耗和散热的问题等[1]。

早期的芯片互连主要采用铝（Al）材料，但由于其电阻率较高，随着特征尺寸的减小，互连线路的 RC 延迟问题日益严重。为了解决这一问题，铜（Cu）因其较低的电阻率和良好的抗电迁移性能，在 20 世纪 90 年代末被引入作为主要的互连材料，并伴随着双大马士革（Dual-Damascene）工艺的发展，实现了互连线图案化的突破。然而，随着芯片特征尺寸进一步微缩至纳米和亚纳米尺度，铜互连材料面临越来越多的挑战，包括量子隧穿效应、原子级加工工艺等问题，这些问题已成为制约摩尔定律延续的重要因素[1]。为了应对这些挑战，研究人员开始探索新型互连材料，以期在保持低电阻的同时，提供更好的可靠性和性能。在这些候选材料中，钴（Co）、钌（Ru）和钼（Mo）等金属因其与铜相近或更低的电子平均自由程和更高的内聚能而成为有潜力的候选材料。特别是钴，由于其较小的尺寸效应、更好的抗电迁移性能以及与现有工艺的兼容性，已经在一些先进工艺节点中得到应用。此外，钌因其优异的电阻率和内聚能，展现出在更小尺寸下替代铜的潜力。除了单一金属元素，金属间化合物和 MAX 化合物也成为互连材料研究的新热点。

MAX 化合物是一类非范德华层状化合物。其通式为 Mn+1AXn，其中 n 为正整数，M 代表前过渡金属（元素周期表中第 3 副族至第 7 副族中的所有过渡金属元素，包括镧系和锕系元素，是一些 d 轨道（或 f 轨道）没有填满电子或其轨道能级接近于外层价电子轨道能级因而可以利用 d 轨道（或 f 轨道）成键的一些金属元素。），如钛（Ti）、钒（V）、铬（Cr）、锆（Zr）、铌（Nb）、钼（Mo）、铪（Hf）或钽（Ta）；A 主要代表周期表第 13 或 14 族的元素，如铝（Al）、硅（Si）、磷（P）、镓（Ga）、锗（Ge）、砷（As）、镉（Cd）、铟（In）、锡（Sn）、铊（Tl）和铅（Pb）；X 代表碳（C）或氮（N）。MAX 相的结构特点是由 MX 单元层与 A 原子层交替堆垛而成。211、312 和 413 类 MAX 相的 MX 单元层具有相似性，均是由 M6X 八面体构成，这些八面体通过边相互连接。三种结构的主要的区别在于每个 MX 单元层中的原子层数，符合 Mn+1Xn 的规律，例如 211 结构中仅具有 2 层 M 原子和 1 层 X 原子。在 MX 层，各个小层之间距离不同，不是简单的重复关系。在每个 MX 单元层之间是一层 A 层原子，从垂直原子层的方向看，三种元素各自的原子层均为六角晶格，依次堆垛形成了层状的晶体结构。[2]由层间堆垛方式的不同，MAX 结构具有多种同分异构体。除 211 结构层数较少，不具有同分异构体外，312 结构均由于 A 原子层和 MX 层的堆垛方式不同，形成 α 相和 β 相。两组原子层都是以两层一周期重复，形成了 x-1-y-2 和 x-2-y-1 两种堆叠方式（MX 层用 x、y 表示，A 层分别用 1、2 表示）。对于 413 结构还具有另一种相，主要在 MX 层内部有所不同，X 原子层在垂直层方向上相互重合，标注为 β 相，而另外两种相，即 α 相和 γ 相，和 312 结构的 α 相和 β 相类似。413 结构的 γ 相并不常见[3]，因此在后续讨论和计算中不被涉及。

MAX 相的晶体结构具有高度的各向异性，这势必导致其在不同方向上的物理性能有显著差异。同时，同分异构现象使得不同相的层间堆垛产生很大变化，也影响到材料的各种物理性能，这得到了第一性原理计算的验证。[4]第一性原理计算结果还表明，MAX 化合物具有作为互联材料使用的潜力，共 69 种稳定的化合物有优于 Cu 的 ρ λ 与内聚能性能，其中有 24 种性能甚至优于 Ru[1]。在晶体结构的稳定性方面，MAX 相展现出良好的热稳定性和机械稳定性。通过第一性原理计算和实验研究，科学家们发现许多 MAX 相在高温下仍能保持其晶体结构，这对

于它们在高性能电子器件中的应用至关重要。此外，MAX 相的晶体结构还允许通过元素替换来调整其性能，例如通过替换 A 位元素来合成新的 MAX 相材料。

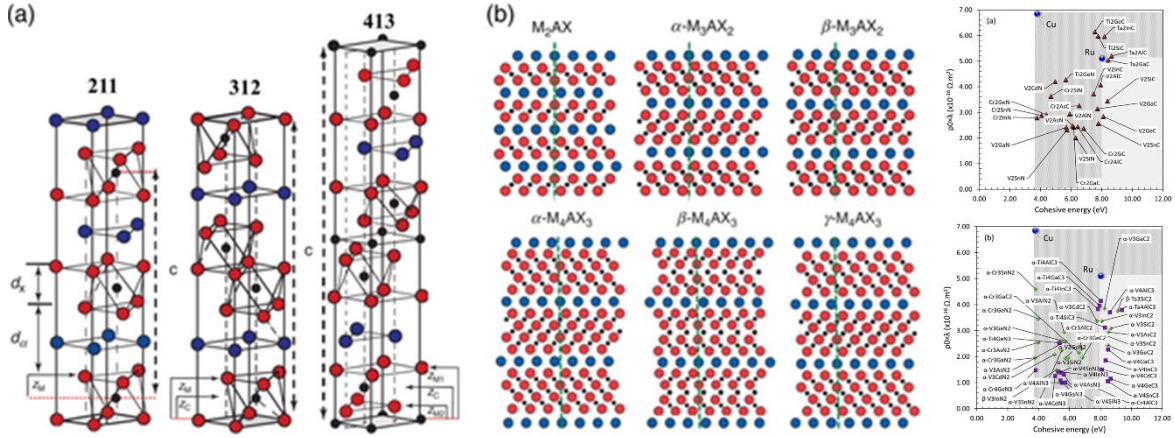

图 1 MAX 化合物的晶体结构[3]（左）和第一性原理得到的互联性能[4]（越往右下性能越好）（右）

MAX 相的研究也对互连材料的研究提出了新思路，即将筛选范围从单质金属扩展到化合物。但由于潜在的化合物数量庞大，绝大部分不具有实验合成的价值，逐个运用第一性原理计算也并不现实。因此，需要高通量广谱筛选方法，运用机器学习模型，仅基于晶体结构猜测晶体的微缩潜力。Xie 和 Grossman[5]在 2018 年的研究中提出了晶体图卷积神经网络（CGCNN）框架，这是一个创新的方法，能够直接从晶体中原子的连接中学习材料性质，提供了一种普适且可解释的晶体材料表示方法。CGCNN 通过构建一个晶体图，其中节点代表原子，边代表化学键，利用图卷积层和池化层自动提取对预测目标性质最优的特征。该框架在多种结构类型和组成的晶体上进行了训练，能够以接近密度泛函理论（DFT）的精度准确预测八种不同的晶体性质。Choudhary 和 DeCost 在 2021 年的研究中[6]进一步发展了 GNN 架构，提出了原子线图神经网络（ALIGNN）。ALIGNN 通过在原子间键图和对应的线图（描述键角）之间进行消息传递，显式且高效地包含了角度信息，从而提高了多个原子性质预测任务的性能。ALIGNN 在预测 52 种固态和分子性质时，表现出比之前报道的 GNN 模型更好的性能，或者在训练速度上具有可比性。Gupta 等人在 2024 年的研究中[7]提出了一个基于结构感知 GNN 的深度迁移学习框架，用于增强多样化材料数据集上的预测分析能力。该框架利用大型源数据集上训练得到的结构信息，通过迁移学习技术显著提高了模型在小型数据集上的预测能力。研究者们在跨属性和跨材料类别的场景中评估了所提出的框架，并发现迁移学习模型在 90%的情况下优于从头开始训练的模型，特别是在数据外推问题上表现更好。这一框架的提出，为材料科学中的材料发现过程提供了加速的可能性，尤其是在数据稀缺的情况下。

2 研究内容及方法

本研究使用机器学习进行互连材料的性能预测，主要分为两个目标标签：微缩潜力因子 $\rho_0\lambda$，以及内聚能 E_{coh}。

首先介绍微缩潜力因子$\rho_0\lambda$。由于在芯片互连材料领域，还没有建立电阻率与微观尺寸之间的物理模型，如何综合考虑材料块体状态下的电阻率，和微缩后材料电阻增大的趋势两大因素，并据此比较材料的微缩潜力，是定义标签工作的关键。本研究参考了 Mayadas-Shatzkes 公式[8]。Mayadas-Shatzkes 公式是 Mayadas 和 Shatzkes 在 1970 年提出的一个用于计算多晶薄膜总电阻率的模型。该模型综合考虑了三种电子散射机制：各向同性的背景散射（由声子和点缺陷

共同作用引起）、由平面势分布（晶界）引起的散射以及由外表面引起的散射。公式具体形式如下：

$$\rho = \rho_0 + \rho_0\lambda\frac{3R}{2D(1-R)} + \rho_0\lambda\frac{3(1-p)}{4d}$$

ρ：总电阻率

ρ_0：体电阻率

λ：与电子散射有关的常数，通常与电子的平均自由程有关

R：晶界反射系数，表示电子在晶界上的镜面反射概率

D：晶界平均间距

p：表面散射参数，表示电子在表面散射时被镜面反射的概率

d：薄膜厚度

随着薄膜或互连线的尺寸微缩，电子的平均自由程会受到导体形状及由此带来的表面散射的影响，从而提高材料的总电阻率。参考公式可以得到，随着薄膜厚度减小，表示表面散射的第三项会增加，且这一项的大小受到$\rho_0\lambda$和 p 两大系数的控制。薄膜在单一维度上达到纳米尺度，而互连线在两个维度上达到纳米尺度，因此我们认为，这一规律对于互连线同样成立。综合考虑数据来源的容易程度，我们选取$\rho_0\lambda$作为表示微缩潜力的因子。

内聚能 E_{coh}（Cohesive Energy）是指独立原子结合成固体时释放的能量，内聚能是表征材料结构稳定性和化学键强度的重要指标。在芯片互联线的工作过程中，由于电子的散射作用，金属原子会发生电迁移（Electronic Migration, EM），改变材料的微观结构，甚至导致互连线脱触而失效，这种现象随着互连线的尺寸微缩而更加显著。结构相同但具有更大内聚能的材料由于原子间的键合更加稳固，能更好抵抗电迁移，避免失效。

2.1 广谱材料性能预测

由 Choudhary 和 DeCost 在 2021 年发布的 ALIGNN 模型[6]，基于晶体的结构文件作为输入，以 JARVIS-DFT 数据作为标签训练，在多个原子性质预测任务上的性能优于先前报道的机器学习模型。故本研究中，我们使用该研究的开源代码进行复现。

模型结构方面，ALIGNN 模型包括 N 层 ALIGNN 和 M 层图卷积网络（GCN），训练中每次迭代均会进行更新。模型使用 Sigmoid Linear Unit（SiLU）激活函数，由于其二阶可微分，且提供了更好的经验性能。模型在经过 N + M 层图卷积更新后，通过网络全局平均池化原子表示，最后通过一个全连接的回归或分类层来预测目标属性。模型的训练通过最小化预测属性和 DFT 计算属性之间的差异来完成，使用了 AdamW 优化器和一周期策略来更新权重。对于模型调优，我们采用试错法调优超参数，尝试不同的图神经网络层数组合来优化模型表现。

由于算力局限，在从头训练模型时我们无法使用大量数据，这势必影响模型的表现。故本研究参考 Gupta 等人 2024 年的研究[7]，将预训练模型视为特征提取器，从中提取原子、键和角度特征。我们使用了论文研究团队提供的 mp_e_form_alignn 预训练模型，该模型基于 Materials Project 数据训练，针对形成能训练，理论上具备描述材料的化学键与电子相关特征的潜力。提取特征后，我们建立了简单的前馈神经网络，并调整模型层数和每层神经元数优化模型表现。

模型复现使用了移动端 GTX1650，基于 Python 3.11.0，CUDA 11.8[9]和 PyTorch 2.2.1[10]。在从头训练中，使用了来自 Materials Project[11]的 3101 个结构文件及对应的$\rho_0\lambda$标签值。

2.2 MAX 体系性能预测

除了基于结构文件的广谱材料筛选方法，本研究还针对 MAX 化合物，探索了对具有特定元素、特定结构的互联材料体系，单独运用特征工程和多种机器学习模型，筛选最优材料的方法。

模型数据的来源是 Sankaran 等人在 2021 年的研究[4]，其中包含了对 107 个 211、312、413 类 MAX 化合物的生成焓、内聚能、$\rho_0\lambda$的第一性原理计算结果，这些结果是机器学习模型训练所使用的标签。采用同一个文献的数据，是由于第一性原理计算中结构弛豫和自洽/非自洽计算中具有大量参数，这样做可以避免不同的计算参数引起数据集本身的系统误差。但同时，这也导致模型缺乏训练数据，无法使用神经网络这类复杂的、需要大量数据来取得较好表现的模型。

为了使模型完全摆脱对第一性原理计算和实验结果的依赖，仅依靠元素的不同来预测 MAX 化合物的性质，本研究充分考虑 MAX 化合物共有的原子化学环境和每种元素各自的属性，组合构建了三个层次的特征。使用了以下十种元素属性（元素属性来源于权威的化学与晶体学手册[12, 13]，电负性采用 Allen 电负性[14]。少数无数据的元素属性均用 0 代替，例如碳和氮元素的功函数（WF）属性）：

Features	Abbreviation
EN	Electronegativity
RAM	Relative atomic mass
WF	Work function
IE	First Ionization Energy
GN	Group number in the periodic table
EA	Electron affinity
MP	Melting point
BP	Boiling point
RW	VDW radii
RC	Covalent radii

表 1 本研究包含的元素属性

使用特征工程和晶体学知识，本研究从上述元素属性构建了三个层次的特征：

原子特征 Average/Deviation atomic features：按照化学计量数加权，得到的元素特性的平均值和方差值，以 A/D 前缀命名。例如 D_EA 是三种元素甲醛的电子亲和能的方差；

位点特征 M/A/X Site-specific features：描述 MAX 三个位点的原子所处的化学环境，分别以 MS/AS/XS 为后缀命名。在 MAX 结构中，每个位点都可看作 12 配位，其中 A 和 X 位点，除了六个同类元素，还有六个 M 位点的元素，故以 M 元素与 A 或 X 元素的属性之差的六倍作为新的特征，前缀标注为 A。M 元素则较为复杂，其中 2/(n+1)的原子,配位的元素是 3A，3X，6M; 另外 1-2/(n+1)配位的元素是 6X，6M，故通过加权得到属性差，标注为 A；对于前一种原子，还可以类比极性的特性，将 A 和 X 原子属性相减，标注为 P。例如 P_EN_MS 表示对应于电负性，M 位点周围的“极性”大小；

结构特征 Structural features：通过分析 MAX 化合物的普遍结构，发现可以用化学计量数中的 n 和相的类型，加上前两种特征对原子特征的描摹，在样本空间可以确定一个相。n 的取

值为 1，2，3，分别对应 211，312 和 413 类 MAX；相的类型使用 One-Hot Coding 的方法，用两个布尔值表现。

根据以上方法，一共构建出 63 种特征，可能需要使用特征筛选的手段降维来提高模型的表现。使用了皮尔森相关系数法，主成分分析法和递归特征筛选，试图提高模型的性能。只保留原子特征或位点特征，评估加入位点特征对模型预测效果的作用。本研究还使用了拆分数据集的方法评估模型泛化性能，具体方法是仅保留其中一部分元素的数据训练模型，用余下的数据作为验证集。

本研究选用线性回归、岭回归、Lasso 回归三种线性模型、随机森林算法和支持向量机（SVR）一共五种模型进行训练和比较。由于上述特征生成方法存在多重共线性的潜在问题，岭回归由于具有 L2 正则化项，在原理上相较其余线性模型具有优势。采用简单模型减短了模型的训练时间，使模型评估和超参数调优更为方便。为保证训练集足够大，采用了 20 折交叉验证（20 Fold Cross-Validation），每折保留 19/20 的数据点用以训练，其余归为测试集。还使用网格搜索（Grid Search）进行自动的模型调优，使各模型达到最好的预测效果。

本研究使用 Python 3.9.10，机器学习和特征筛选调用了 scikit-learn 1.9.0[15]。

3 研究结果和讨论

3.1 广谱材料性能预测结果

3.1.1 ALIGNN 从头训练

在 3101 个结构文件和对应的$\rho_0\lambda$性能上，以 8：1：1 比例拆分出训练集、验证集、测试集，从头训练 ALIGNN 模型，得到的结果如图 3 所示

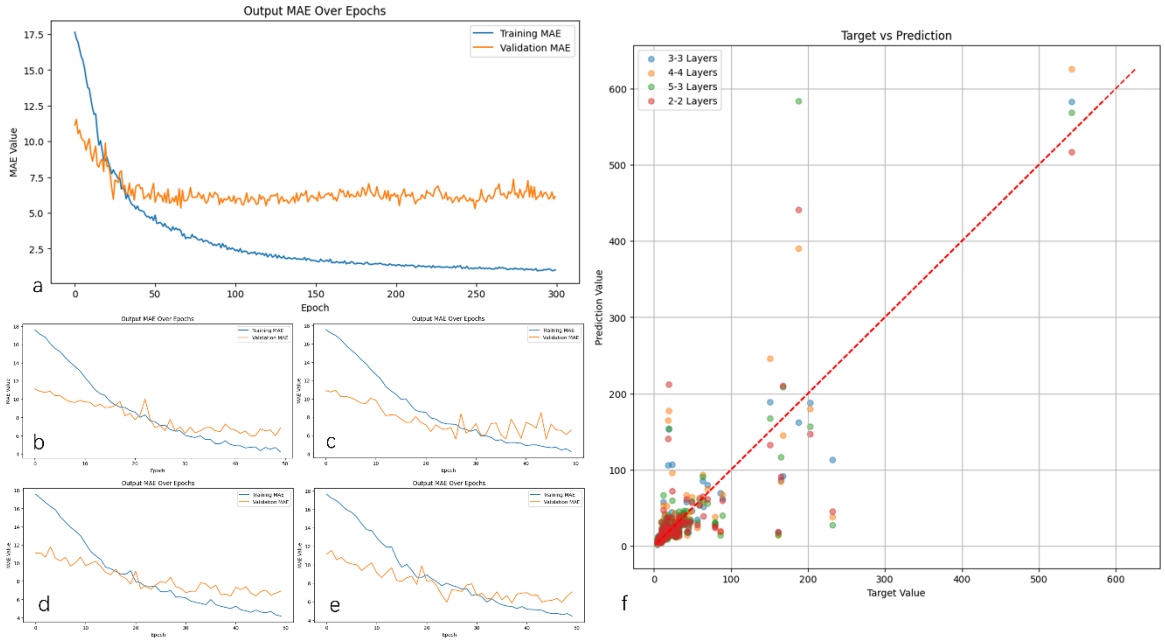

图 3 a~e 训练集和验证集上的模型表现 a：默认配置，alignn-4 & gcn-4，迭代 300 代；b：alignn-3 & gcn-3，迭代 50 代；c：alignn-2 & gcn-2，迭代 50 代；d：alignn-5 & gcn-3，迭代 50 代；e：alignn-4 & gcn-4，迭代 50 代. f 四个超参数配置下模型在测试集中的表现

Hyperparameter (epoch=50)	Test MAE
alignn-4 & gcn-4	8.565
alignn-3 & gcn-3	7.194
alignn-2 & gcn-2	8.100
alignn-5 & gcn-3	8.055

表 2 不同超参数下的模型表现（五次重复训练取中位数）

分析结果可以发现，在训练集上，图神经网络能够很好地表现晶体的特征，模型在这方面具有足够的复杂度，这体现在训练集的损失函数不断下降。但是在 50 代之后，验证集上的损失函数不再下降，并始终高于训练集损失，这意味着模型在 50 代左右已经收敛，并产生了过拟合的现象，泛化能力不足。通过超参数的优化，还发现相较于默认参数，可以取得 16.1%的测试集优化。但是观察和比较模型的预测值与实际值，会发现预测值并没有正确反映实际值中$\rho_0\lambda$的趋势，观察上图的图 f 发现，模型只能定性地预测，尤其在$\rho_0\lambda$较小的位置，还出现了高估的情况，这对筛选工作存在潜在的负面影响。

3.1.2 基于特征提取的迁移学习

我们推测模型数据量是 ALIGNN 无法准确描绘$\rho_0\lambda$和晶体结构间关系的原因，因此，我们进而采用迁移学习的方法，使用 mp_e_form_alignn 预训练模型，从结构文件中提取特征，再针对这些特征，以$\rho_0\lambda$为目标进行训练。经过 ALIGNN，提取出了 768 个特征，由原子特征、键特征和角特征各 256 个组成。经过皮尔森相关系数校验，绝大多数特征之间的相关系数集中在±0.1，几乎不具有共线性关系。为了防止过拟合，本研究仿照论文中的做法采用了前馈神经网络，调整模型层数和每层神经元数优化模型表现，并比较这种方法与从头训练 ALIGNN 的效果。

Models and Hyperparameters	Test MAE
2 layers (384-1)	7.1
3 layers (256-128-1)	7.8
3 layers (384-192-1)	6.1
3 layers (512-256-1)	5.8
4 layers (512-256-128-1)	6.9
4 layers with dropout layers (p=0.5)	7.7
5 layers (512-256-128-64-1)	7.4
5 layers with dropout layers (p=0.5)	7.2
alignn-3 & gcn-3	7.194

表 3 不同超参数下的模型表现（五次重复训练取中位数，去除离域值）

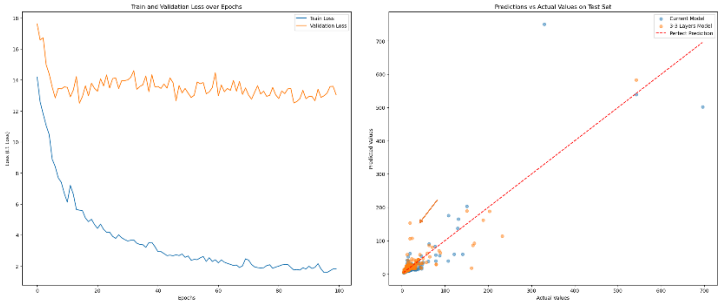

图 4 典型的模型训练过程（左）和比较 3 layers (512-256-1)和 alignn-3 & gcn-3 模型在测试集上的表现（右）

大量的试错法调参表明，3~4 层的简单模型是最优的选择，可以使模型的预测能力超过从头训练的 ALIGNN。观察比较测试集-实际值，模型对于低$\rho_0\lambda$数据点的预测有一定改善，避免了从头训练模型在部分低值上偏高的情况，能够更好描述$\rho_0\lambda$变化的趋势；在此基础上继续提高模型层数（增加模型复杂度）对预测效果则具有反作用，对于较高层数模型采用蒙特卡洛 Dropout 方法没有显著的改善效果。

由于本研究主要关注具有低$\rho_0\lambda$的材料，我们也尝试对数据集设置截止值，只保留$\rho_0\lambda$低于 50 的材料作为训练集和验证集，并与没有设置截止值训练的模型在同一个测试集（测试集始终设置截止值）上作比较。我们发现，采用截断有助于提高模型的预测能力，但过低的截止值会损害模型的泛化能力，导致模型在测试集上表现不佳。

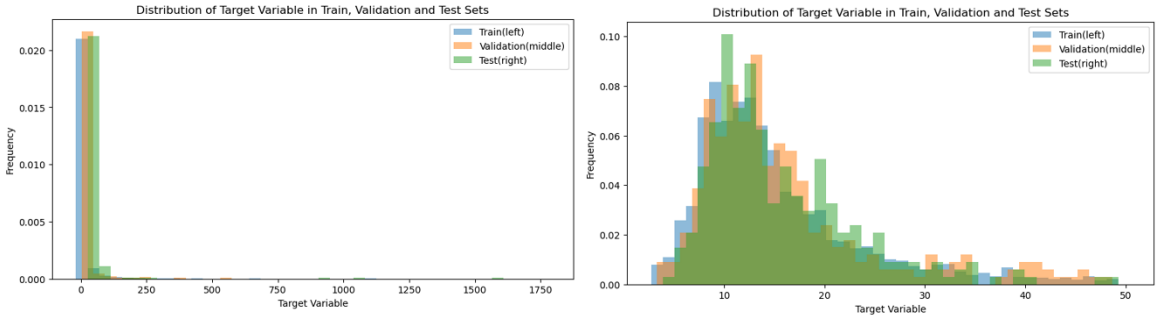

图 5 未设置截断的数据集（左）和设置了截断的数据集（右）

Cutoff	Test MAE
with cutoff	2.89
w/o cutoff	3.69

表 4 比较有无截断值下的模型表现（五次重复训练取中位数）

3.2 MAX 体系性能预测结果

为了提高模型的收敛能力，我们首先对 63 个特征进行了归一化处理，即对特征作线性变换，使其均值为 0、标准差为 1。运用线性回归、岭回归（Ridge Regression）、随机森林算法、支持向量机（Support Vector Regression，SVR）和 Lasso 回归共五个机器学习模型进行训练和网格搜索（Grid Search）调参，训练结果（图 5）显示，模型对内聚能标签的预测性能良好，对生成焓预测性能一般，而对$\rho_0\lambda$（Constant τ）模型预测性能很差。由于数据量非常有限，无法支持神经网络训练，我们放弃对$\rho_0\lambda$的预测，聚焦在对内聚能的预测上，以评估 MAX 化合物抗电迁移能力。以下，我们仅保留岭回归、支持向量机和 Lasso 回归三个模型，通过特

征筛选手段寻找规律（统一使用了 20 折交叉验证），对内聚能预测任务上模型的表现进行解释。

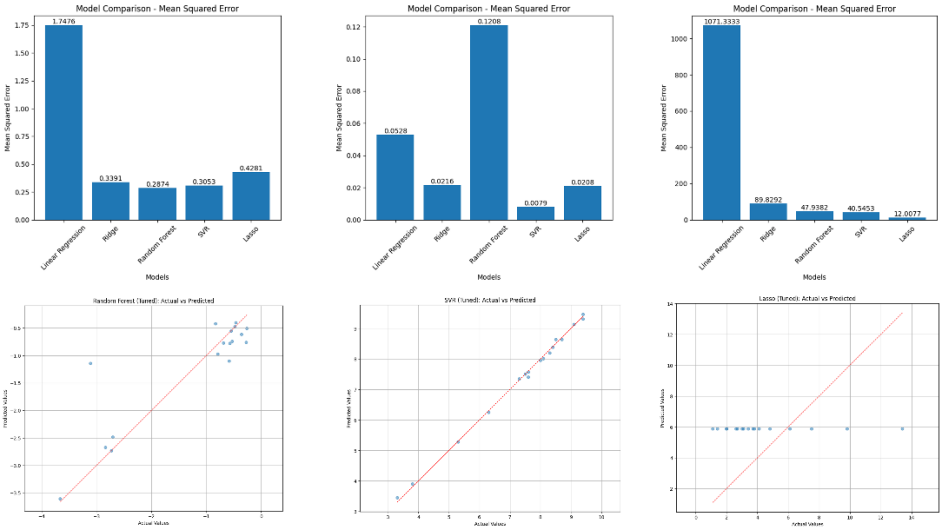

图 5 各模型分别在生成焓、内聚能、$\rho_0\lambda$（Constant τ）标签上的预测性能（Test MSE）柱状图及最优模型的表现

3.2.1 特征类型

如上文所述，本项目在 MAX 化合物的预测任务中，将特征分为原子特征、位点特征和结构特征。为了评估加入位点特征的作用，我们比较了仅原子特征、仅位点特征、原子+位点特征三种特征组合下，训练出的模型的表现。图 6 表明，仅用原子特征或者位点特征已经可以相对准确地预测 MAX 化合物的内聚能，但加入位点特征有助于模型预测精度的提高

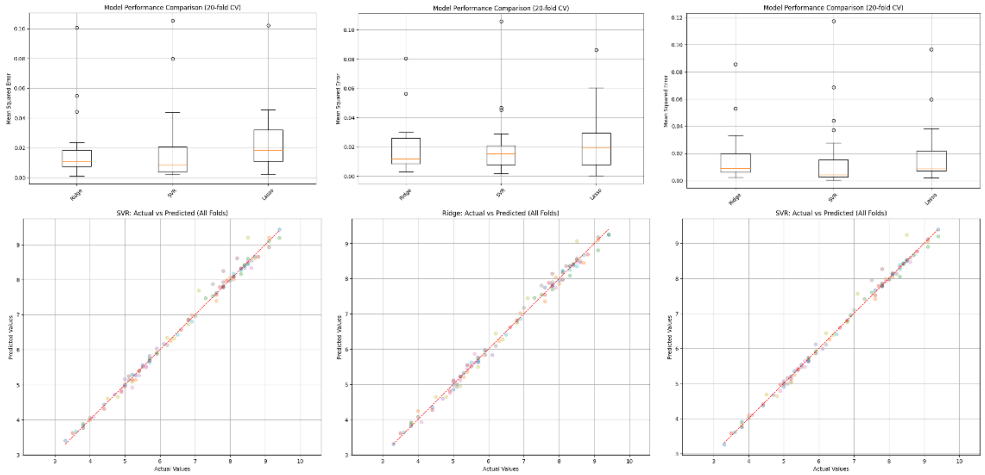

图 6 仅原子特征（左）、仅位点特征（中）、原子+位点特征（右）的预测性能（Test MSE）箱线图及最优模型的表现

3.2.2 特征的信息密度：皮尔森相关系数和主成分分析

皮尔森相关系数是一种衡量两个变量之间线性相关程度的统计量，其值介于-1 和 1 之间，其中 1 表示完全正相关，-1 表示完全负相关，而 0 则表示没有线性相关。该系数通过计算两个变量的协方差与各自标准差的乘积的比值来得出，从而量化它们之间的线性关系强度和方向。

本研究分别使用 0.8 和 0.9 作为筛选的阈值，从图 7 可见，以 $|r| < 0.9$ 为标准，能够缓解模型的离群现象，使模型的预测表现更加稳定。更严格的筛选标准则会反过来劣化模型的表现。

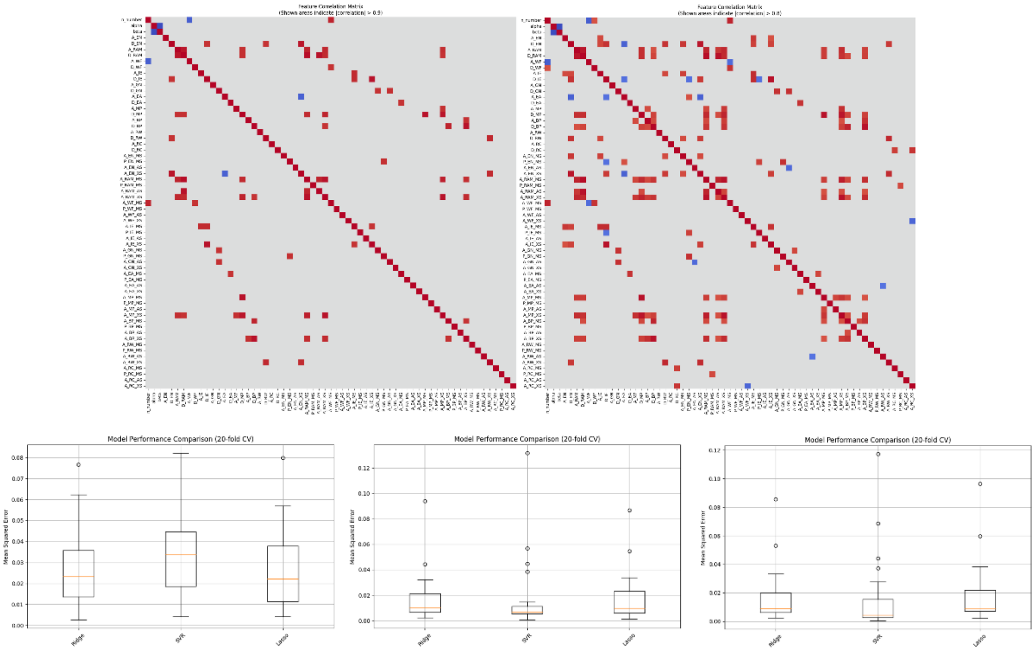

图 7 $|r| < 0.8$ 和$|r| < 0.9$ 判据下的相关性热力图（仅标注被筛选的特征），及从 0.8、0.9 到不筛选（从左到右）的模型表现（Test MSE）箱线图

主成分分析（PCA）是一种降维技术，旨在从原始数据集中提取出最重要的特征子集。在 PCA 中，通过线性变换将数据投影到新的坐标系中，这个新坐标系由数据的主成分构成，即方差最大的方向。成分筛选通常涉及识别并保留那些解释数据大部分变异性的主成分，同时舍弃那些贡献较小的成分，以达到减少数据维度和噪声的目的，同时尽可能保留原始数据集的信息。这种筛选过程有助于简化模型，提高计算效率，并可能增强模型的泛化能力。

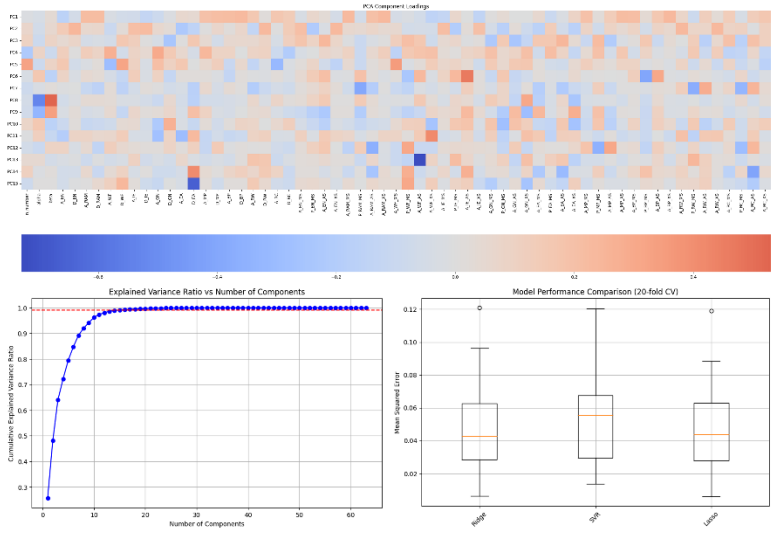

图 8 各特征对主成分的贡献度热力图（上），特征降维过程（左下）和 15 个主成分特征下模型的训练表现

图 8 表明，主成分分析将 61 个变量降维到 15 个，并认为 15 个变量能够解释 99%的变化。对于本研究采用的特征集，由于共线性严重，这种分析有其道理。但使用主成分训练模型，发现模型表现大幅下降了，这说明模型已经足够简化，过拟合的风险很小，且数据本身的噪声不大。

3.2.3 泛化能力评估

进一步，为了评估模型泛化能力，我们筛选出数据集中所有含 Ti 元素的 MAX 化合物，作为测试集，剩余的化合物数据用作训练模型，以此评价模型面对 MAX 体系中新元素组成的化合物的预测能力。结果（图 9）表明，虽然有一定误差，模型仍然能定性描绘包含训练集外元素的 MAX 化合物的内聚能，证明模型具有一定的泛化能力，具有发掘抗电迁移的新型 MAX 化合物的潜力。

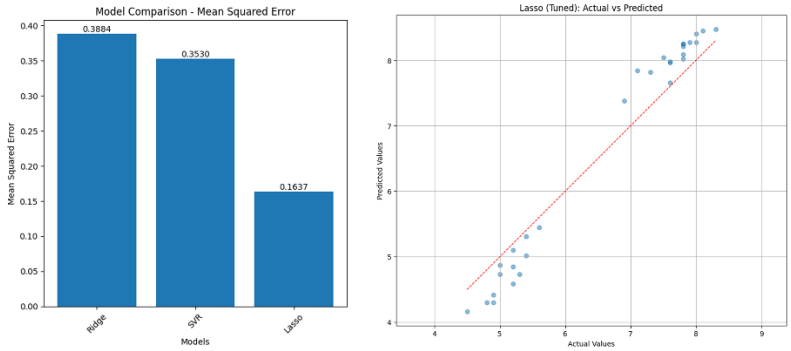

图 9 模型的泛化表现

3.2.4 递归式特征消除

递归式特征消除（Recursive Feature Elimination, RFE）是一种特征选择方法，它通过递归地考虑越来越小的特征集来选择特征。该方法的基本思想是首先使用所有特征训练基模型，并计算每个特征的重要性，然后从当前特征集中删除最不重要的特征，并在修剪后的特征集上重复这个过程，直到达到所需的特征数量或满足早停（Early-stop）条件，即模型表现在容差次数内没有改善。

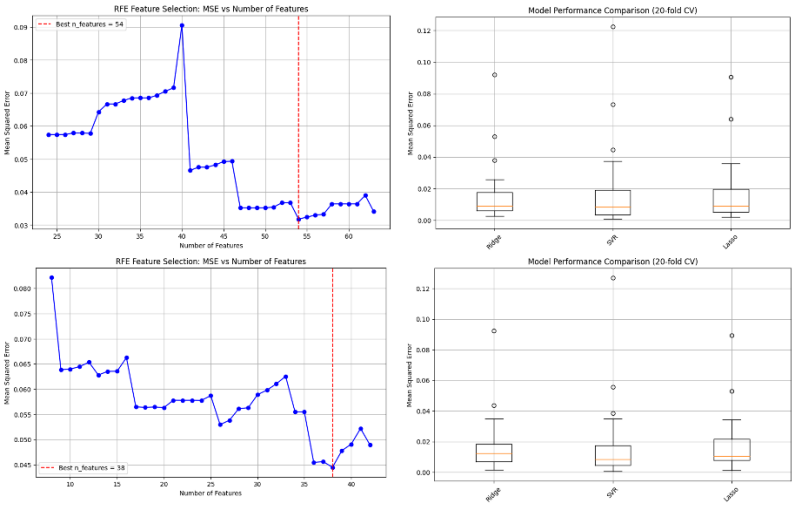

图 10 采用递归式特征消除（上）；联合使用 RFE 和皮尔森相关系数筛选（下）

本研究使用预测能力较高且计算开销小的岭回归模型作为基模型，使用 15 次容差和 20 折交叉验证以避免随机波动影响。RFE 方法筛选出 54 个特征，剔除了以下特征：'A_MP', 'A_RAM_MS', 'A_WF_XS', 'A_MP_MS', 'P_MP_MS', 'A_MP_AS', 'A_BP_MS', 'P_BP_MS', 'A_BP_AS'。训练证明，这种方法可以有效提高模型的预测表现，降低离群值。自然地，我们也尝试了将 RFE 和皮尔森相关系数筛选联合使用，同样起到了降低离群的作用。

4 结论

本研究通过机器学习技术对芯片互联材料的性能进行了预测，特别关注了微缩潜力因子和内聚能 E_{coh} 两个关键指标。通过结合图神经网络和前馈神经网络，本项目成功构建了高通量筛选模型，对候选互连材料进行了有效的性能评估；通过使用多种机器学习模型和特征工程，发现机器学习模型在预测 MAX 体系芯片互联材料的抗电迁移性能上具有较高的准确性，为新型互连材料的性能预测提供了一种高效的筛选方法。

在广谱材料性能预测方面，ALIGNN 模型虽然能够定性预测材料性能，但存在过拟合现象，限制了其泛化能力。通过迁移学习策略，利用预训练模型提取特征，并结合前馈神经网络进行训练，显著提高了模型的预测精度和泛化能力。此外，通过设置数据截断值，模型对低$\rho_0\lambda$材料的预测能力得到了进一步提升。

在 MAX 体系性能预测方面，本研究通过特征工程和多种机器学习模型的结合，建立了对 MAX 化合物内聚能的快速预测模型。特征筛选和降维技术的应用，如皮尔森相关系数法、主成分分析法和递归特征筛选，进一步提高了模型的预测性能和稳定性。模型的泛化能力评估显示，尽管存在一定的误差，但模型能够定性预测包含训练集外元素的 MAX 化合物的内聚能，证明了其在发掘新型抗电迁移 MAX 化合物方面的潜力。

综上所述，本项目不仅为机器学习在材料科学领域的应用提供了实证支持，也为未来芯片互连材料的研究和开发提供了新的思路和工具。随着机器学习技术的不断进步和数据集的日益丰富，预期该领域将取得更多的突破性成果。

5 致谢

感谢杭弢老师对本项目的支持与指导；感谢申炜师兄为我指引学习方向，带领我从零开始探索机器学习和神经网络的运用。特别感谢 Materials Project、PyTorch、Scikit-learn、ALIGNN 等重要工具和平台的开发者和开源社区，您的开源精神，使得这一项目成为可能。

6 补充材料

所有代码在 GitHub 开源可用，请通过以下网址访问：https://github.com/Howard-Cao/ML-Interconnect-Prediction

参考文献

[1] ZHANG S, DENG X, WANG Y, et al. Revolution of next-generation interconnect materials and key processes for advanced chips in post-moore era [J]. SCIENTIA SINICA Chimica, 2023, 53(10): 2027-67.
[2] FU L, XIA W. MAX Phases as Nanolaminate Materials: Chemical Composition, Microstructure, Synthesis, Properties, and Applications [J]. Advanced Engineering Materials, 2021, 23(4): 2001191.
[3] EKLUND P, BECKERS M, JANSSON U, et al. The Mn+1AXn phases: Materials science and thin-film processing [J]. Thin Solid Films, 2010, 518(8): 1851-78.
[4] SANKARAN K, MOORS K, TŐKEI Z, et al. Ab initio screening of metallic MAX ceramics for advanced interconnect applications [J]. Physical Review Materials, 2021, 5(5): 056002.
[5] XIE T, GROSSMAN J C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties [J]. Physical Review Letters, 2018, 120(14): 145301.
[6] CHOUDHARY K, DECOST B. Atomistic Line Graph Neural Network for improved materials property predictions [J]. npj Computational Materials, 2021, 7(1): 185.
[7] GUPTA V, CHOUDHARY K, DECOST B, et al. Structure-aware graph neural network based deep transfer learning framework for enhanced predictive analytics on diverse materials datasets [J]. npj Computational Materials, 2024, 10(1): 1.
[8] MAYADAS A F, SHATZKES M. Electrical-Resistivity Model for Polycrystalline Films: the Case of Arbitrary Reflection at External Surfaces [J]. Physical Review B, 1970, 1(4): 1382-9.
[9] LUEBKE D. CUDA: Scalable parallel programming for high-performance scientific computing; proceedings of the 2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, F 14-17 May 2008, 2008 [C].
[10] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library [J]. Advances in neural information processing systems, 2019, 32.
[11] JAIN A, ONG S P, HAUTIER G, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation [J]. APL Materials, 2013, 1(1).
[12] HAYNES W M. CRC handbook of chemistry and physics [M]. CRC press, 2016.
[13] PAUFLER P. P. Villars, L. D. Calvert. Pearson's handbook of crystallographic data for intermetallic phases. American Society for Metals. Metals Park. Ohio. 1986. Vols. 1–3. 3258 pp, US $ 495.00 ISBN 0-87170-217-7 [J]. Crystal Research and Technology, 1987, 22(11): 1436-.
[14] KAREN P, MCARDLE P, TAKATS J. Comprehensive definition of oxidation state (IUPAC Recommendations 2016) [J]. Pure and Applied Chemistry, 2016, 88(8): 831-9.
[15] PEDREGOSA F, VAROQUAUX G, GRAMFORT A, et al. Scikit-learn: Machine Learning in Python [J]. J Mach Learn Res, 2011, 12(null): 2825–30.