基于机器学习的二维传感材料研究

刘志发，陈浩东，宁上通

【摘要】因现有的气体传感器存在特异性不强、敏感度低的问题，二维材料气敏传感器的研究日益至为重视。然而的二维传感材料的研究中，准确的带隙值、吸附能的计算耗时较长，计算成本高。本文采用机器学习进行初步筛选，从而方便进一步利用第一性原理计算对筛选出的传感材料结构特点，电子特性及其他特性进行了进一步的验证和研究。经验证，该模型具有较好的泛化性能，能在未训练数据集上获得相当好的性能。说明机器学习有望应用于材料掺杂和吸附性能优化设计，本课题的成果可以加速材料基因组的建设，为研究和开发用于传感及其他领域的功能材料和器件提供了有益的参考。

【关键词】机器学习；二维材料；能带；MX2

1 背景

1.1 引言

在我国大气中主要污染物是氨氮化合物，二氧化硫，氮氧化物、VOCs 气体以及重金属颗粒等物质。这些污染物的来源跟我们的日常生活息息相关，比如工业生产，石油矿产的开采，燃煤发电发热，汽车尾气的排放，垃圾的无措施焚烧等等,严重影响了人们的生活。

通过传感器监测大气污染物，可以有效地发现污染源，进而采取相应的措施控制污染物的排放。现有的重金属大气污染物监测方法主要有：原子吸收光谱法，原子荧光光谱法，电感耦合等离子体法，紫外-可见分光光度法，高效液相色谱法，电化学分析法以及生物检测法等。然而，以上的气体检测手段面临着检测仪器体积较大，价格昂贵，检测成本高，检测时间长等问题，无法满足实时快速检测的需求。因此，亟需设计新型的气体污染传感器，以适应低成本、实时测量的检测要求。

1.2 国内外研究现状

现有气体传感器的存在特异性差，敏感度低等问题，二维材料的带隙可调，且不同的二维材料由于晶体结构的特殊性质导致了不同的电学特性或光学特性的各向异性，因此具有很大的发展潜力。因此，基于二维材料的气敏传感器的研究越来越到受人们的关注。2015 年，重庆师范大学冯庆报导了一种 SP3 杂化的气体分子在金红石相二氧化钛（110）的吸附规律，从而为基于该材料的 NH3 的气敏传感器提供了理论基础。2020 年，天津大学吴萍组利用泛函密度理论（DFT）应用双轴应变诱导的绿色磷烯单层膜实现对气体小分子探测，发现绿色磷烯对氮氧化合物具有高的选择性。2020 年，伊朗聚合物和石化研究所石油化工学院的 Mohammad Ghashghae 等通过量子化学计算，发现具有空位缺陷的黑磷烯可用于 NO2 检测，且具有的高灵敏度（19%灵敏度）和可重复使用（4.0 s 恢复时间）。而氧化锌掺杂的黑磷烯是理想的 NO2 吸附剂。2019 年，印度 S.N. Bose national centre for Basic Science 的 Avisek Maity 等以钙钛矿卤化物为传感器材料，制作基于纸质电子元件的廉价固态气体传感器，用于选择性检测 NH3 气体，检测能力优于 1ppm。

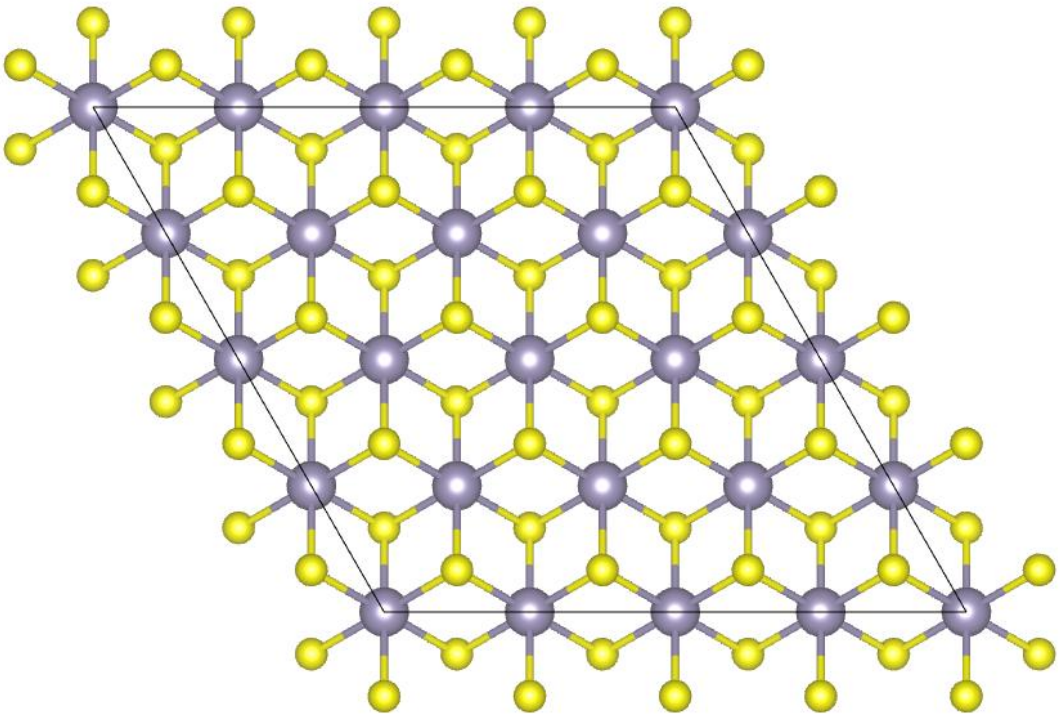
然而，现有的二维传感材料的研究主要是使用基于第一性原理方法的密度泛函理论来计算，计算出的带隙与真实值有偏差，为了获得更精准的带隙，经常使用一些复杂的方法。同时，为了计算吸附能，就通过需要计算吸附位点、吸附能等步骤，计算时间较长，一种传感材料的研究往往需要耗费数十天的时间。而且计算成本较高，需要高性能的计算机，如超算。

机器学习为材料设计提供了新的方向。吉林师范大学计算机学院的董延华等在《基于机器学习的多元化合物带隙预测》论文中采用 2D Materials Encyclopedia 数据库中基于第一性原理和密度泛函理论计算的多元化合物数据集，通过机器学习预测多元化合物的带隙[1]。东南大学物理学院的万新阳等通过机器学习和第一性原理结合的方法，构建了新型 A2BB'O6 型双钙钛矿氧化物材料快速筛选的框架[2]。Tran 等[3]采用主动机器学习模型，筛选出适用于 CO2RR 和 HER 的金属间化合物分别为 10 种和 14 种。然而，目前尚未有使用机器学习对二维传感材料进行研究的报导。

2 主要技术

2.1 带隙及形成能

带隙，也称为能隙或禁带宽度，通常是指绝缘体和半导体中价带顶部和导带底部之间的能量差。带隙是原子上的价电子跃迁到导带所需的能量。位于导带的电子可以在晶格内自由移动，因此作为电荷载体传导电流。带隙的值越大，价带电子就越难以被激发到导带，从而导致本征载流子浓度偏低，进而导致低电导率。带隙是决定固体导电性的一个主要因素。对于传感材料，带隙起到重要作用，可以直接影响到传感器的敏感性和选择性。对于气体或重金属传感器而言，当吸附物到达传感器表面时，传感器表面电子密度会发生变化，进而引起电学特性的变化，从而可以检测气体的存在。



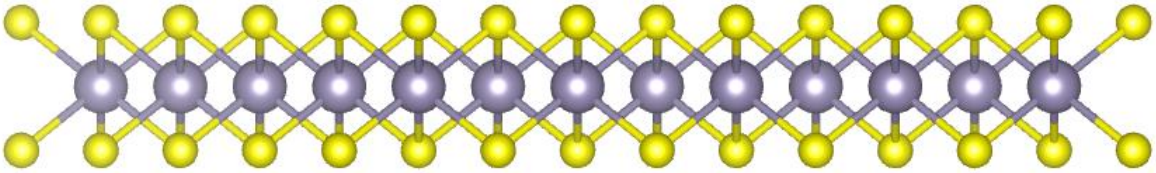


图 1 二维材料 SnS2 上视图及侧视图

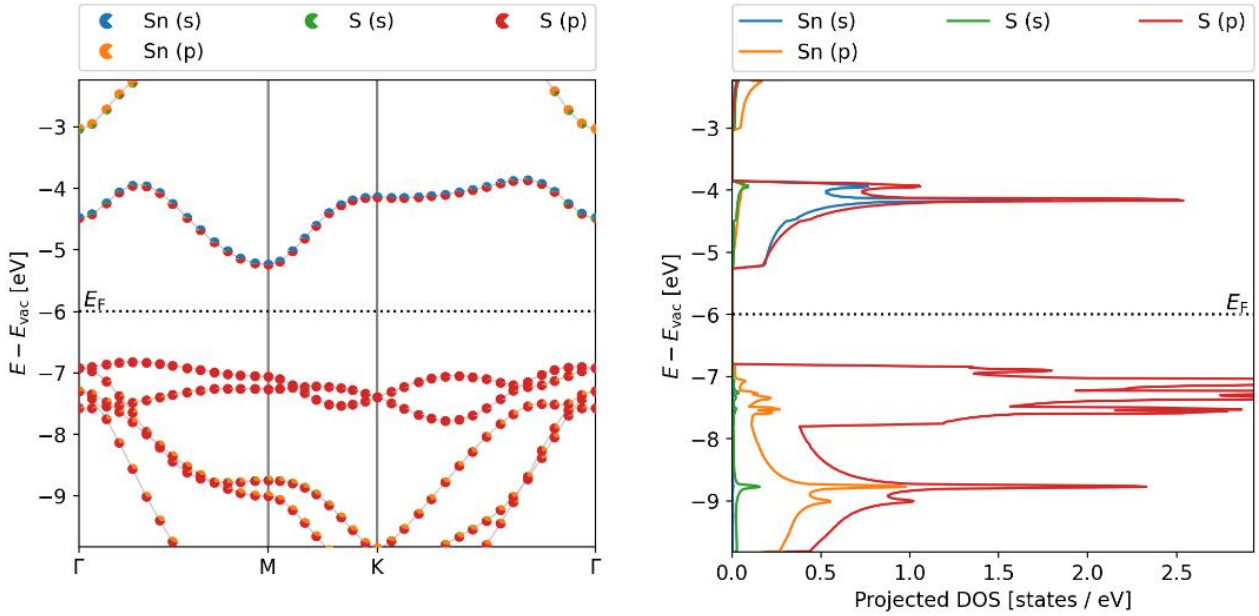


图 2 二维材料 SnS2 的能带及 DOS 图

使用二维材料设计传感器，需要提前评估作为敏感材料的二维材料的带隙。如果敏感带隙过低，则在吸附前后的导电性能对比不明显，影响到了传感器的灵敏度。尤其是带隙接近 0 的二维材料，要设计为电导式传感器，还需要进一步修饰其他原子，以使其具有一定的带隙。同时，当探测对象（例如气体）与本征二维材料之间的吸附作用较弱时，通常需要对二维材料进行修饰或掺杂，以增加其吸附性。这些在材料表面添加的原子或分子可以与探测对象（例如气体）形成更强的相互作用，从而提高二维材料的吸附性。修饰也改变二维材料的电子密度分布，从而改变其对气体的敏感性。

然而，对于不同的基底材料，修饰哪种原子或分子，才能对某种气体具有良好的吸附效果，这正是传感材料设计中遇到的主要问题。在设计阶段通过 CVD 或 ALD 技术大量修饰不同的原子与分子来进行实验，实验成本非常高昂，错误几率非常高，难以找到有效的修饰材料和基底传感器的搭配。

2.2 密度泛函理论

第一性原理是量子力学中最基本的原理，通过研究单个原子中电子和核的相互作用来解释化学和物理问题。而密度泛函理论（DFT）基于第一性原理的框架，是多电子体系电子结构的重要研究方法方法。通过 DFT，可以用来预测原子和分子的结构，能量，离子化能和电子特性，以及密度和成分的可能变化等等。

用密度泛函理论计算半导体材料的电子特性，预测可能的超过实验所能测量到的物理学性质，发现材料的新用途,有助于发现新型材料，设计和合成新型传感材料和传感器。通过建立基底材料修饰不同原子或分子模型，并进行形成能的计算，可以预测基底材料修饰后的稳定性，从而可以排除一些不稳定的基底和修饰物组合。通过计算修饰后的二维材料与吸附原子或分子的吸附能，即可排除吸附效果差，灵敏度低或可复用性差的组合，找到合适的吸附物。

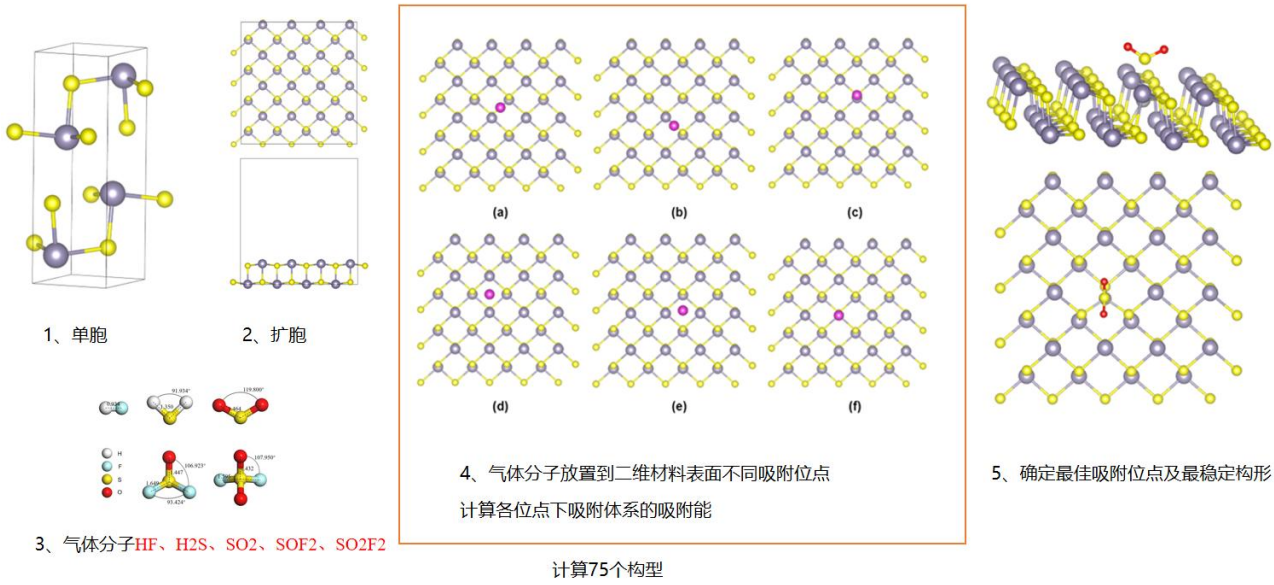


图 3 气体传感器设计步骤图

然而，使用 DFT 方法设计传感器，常消耗大量的计算资源。在计算形成能或吸附能前，需要计算原子或分子吸在基底材料中不同位点的能量，从而找到能量最低点，也就是修饰或吸附时的最稳定点。而这个计算需要确定找到基底材料的不对称点与修饰原子种类、吸附原子或分子不同姿态的组合。这对于结构越复杂的吸附物，组合的可能性越多，计算量越大。由于 DFT 计算往往在周期性势场中进行，这就要求基底材料需要足够大，才能使吸附物之间不相互干扰。而随着基底材料原子数的增加，计算资源需求也快速增长。因此通常情况下，进行传感材料的设计需要使用超算等高性能计算资源，需要耗费大量的计算时间和成本，这也是 DFT 用于传感材料设计的一大难题。

2.3 机器学习

2.3.1 机器学习

机器学习可以快速地从大量数据中提取有价值的信息，生成准确的预测和分类，识别复杂的模式和趋势，而人类很难做到。本文中涉及的机器学习技术包括：

支持向量机（Support Vector Machine，SVM）可用于分类、回归和异常检测等问题。其中，线性回归是指利用线性模型对数据进行回归分析的方法。

随机森林(Random Forest)算法是由多棵决策树组成的模型，每棵决策树都是一个独立的模型，最后将多棵决策树的预测结果集成起来得到最终的预测结果。

DART(Dropouts meet Multiple Additive Regression Trees)算法结合了随机森林和弱学习器的思想，使用多个回归树（Additive Regression Trees）作为基学习器，并通过随机丢弃树（Dropouts）的方法来组合这些回归树，并对剩余的回归树进行训练。

Lightgbm-gbdt 是一种用于机器学习的高效的梯度提升决策树（Gradient Boosting Decision Tree，GBDT）算法。它是针对大数据集进行优化的，可以通过减少模型的训练时间，提高模型的准确性。

2.3.2 深度学习

深度学习是机器学习中的一个分支，是基于人工神经网络（Artificial Neural Networks）的一类技术。深度学习利用了多层次的非线性处理单元（称为神经元）来对数据进行建模和分析，模仿人类大脑的学习方式。深度学习算法将大量的输入数据映射到多个不同的隐藏层上，从而使用许多参数来存储和处理数据。

3 基于机器学习的能带预测

3.1 数据准备

选取了 MX2 型二维材料为研究对象，挑选其中具有一定带隙的化合物进行研究。选取了基于第一性原理和密度泛函理论计算得到的 120 个 MX2 型二维金属化合物数据作为数据集，提取 M, X 位元素的电负性(EN)、第一电离能(I)、原子有效半径(R)、化合物形成热(Hf)、总能量(E)、晶体质量(m)、晶体体积(V)以及基于 GGA-PBE 计算的带隙等性质为特征变量，以 GOWO 带隙为预测目标，组成数据集，进行机器学习，同时也引入了带隙的理论计算值作为模型的参考。

3.2 模型训练

使用 Lightgbm-gbdt(梯度提升决策树)图 4、dart(Dropouts meet Multiple Additive Regression Trees)图 5，随机森林(Random Forest)图 6，SVR(支持向量机线性回归)图 7，机器学习方法对训练集进行训练，可得到如图所示模型，图中横坐标为带宽理论计算值，纵坐标蓝点为对应带宽的预测值，黄点为对应的 PBE 带隙值，可以看到预测值比 PBE 值更接近带宽理论计算值。

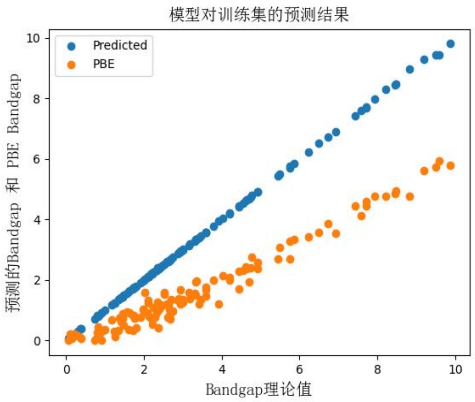


图 4 Lightgbm-gbdt 算法训练图

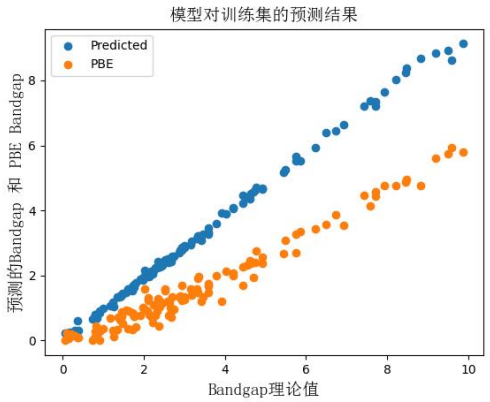


图 5 Lightgbm-dart 算法训练图

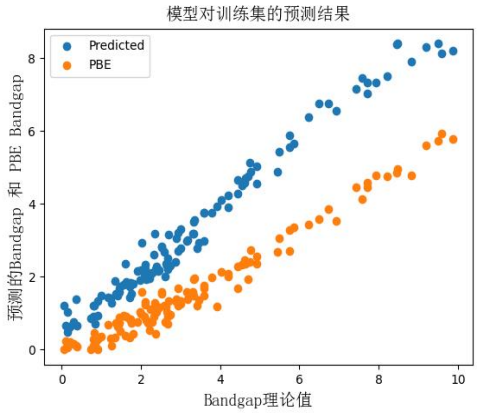


图 6 随机森林算法训练图

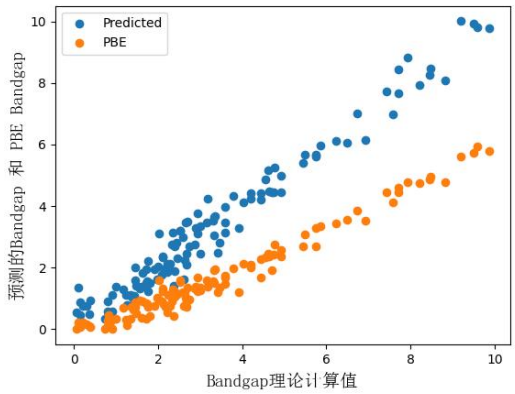


图 7 支持向量机算法训练图

从图 4 到图 7 可以看出，使用 Lightgbm-gbdt 和 Lightgbm-dart 的算法训练出来的模型效果都比较理想，拟合度都很好。而使用随机森林(Random Forest)和 SVR(支持向量机线性回归)训练出来的模型，拟合度一般。四种算法的 10 折交叉验证精确度误差分析如表 1 所示：

	adjusted_r2	mse	rmse	mae	r2	explained_variance_score	runtime (s)
SVR	0.9648	0.2060	0.4539	0.3697	0.9636	0.9638	32.657
Random_Froest	0.9772	0.1042	0.3228	0.4840	0.9794	0.9794	1.271
LightGBM_GBDT	0.9998	0.0006	0.0260	0.1114	0.9998	0.9998	1.706
LightGBM_Dart	0.9915	0.0392	0.1979	0.3748	0.9923	0.9950	3.278

表 1 10 折交叉验证精确度误差分析表

从表 1 可知，以上 4 种算法训练出来的模型效果都比较理想，其中 Light_GBDT 的 adjusted_r2 最高，绝对校正系数(adjusted_r2)能抵消样本数量对 R-Square 的影响，所以 adjuste_r2 越大模型拟合效果越好，模型越好。在样本大小为 120 个的数据集里 LightGBM 算法使用 cpu 训练的时间比传统机器学习算法要快很多，基本上能做到几秒就训练好模型，而传统机器学习算法可能要用到几分钟。从训练时间和模型拟合度来看，LightGBM 算法都比传统机器算法要好。

3.3 能带预测

使用训练好的模型对测试集的样本进行预测，图中横坐标为带宽理论计算值，纵坐标蓝点对应带宽的预测值，黄点为对应的 PBE 带隙值。

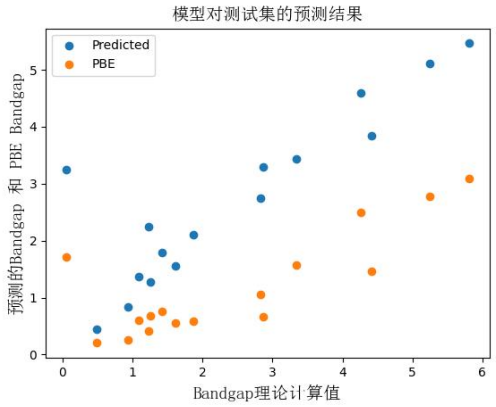


图 8　Lightgbm-gbdt 算法模型预测图

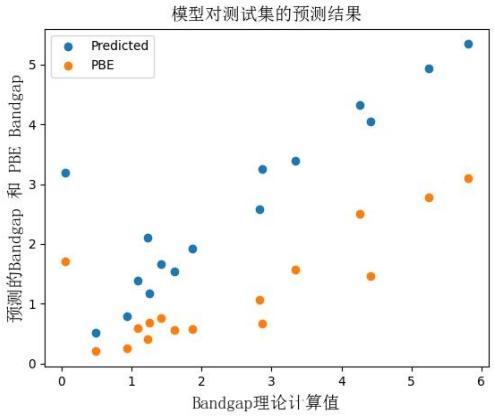


图 9　Lightgbm-dart 算法模型预测图

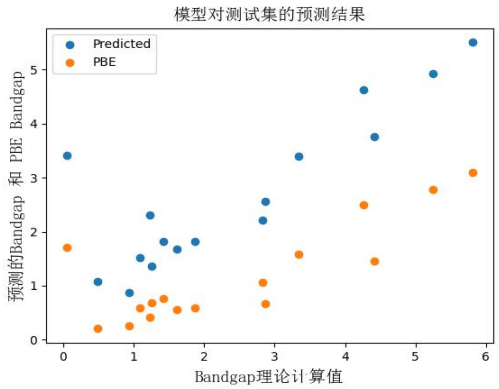


图 10 随机森林算法模型预测图

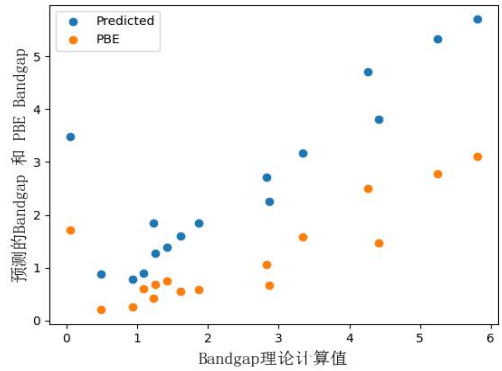


图 11 支持向量机算法模型预测图

从表中数据可以看出模型在二维材料带隙极小(≤ 1eV)的情况下进行预测，有时得到的预测带隙误差与理论计算值相比相差较大，但在合理范围(1eV~6eV)以内，预测值与理论计算值非常接近，平均误差较小。可能是样本之间特征相关性不够强和作为重要特征计算的 PBE 带隙值与带隙理论计算值相差过大导致(如表 2 所示)，从预测结果来看模型对未知数据预测表现良好。

表 2 预测结果对比表

Formula	PBE	GOWO	SVR	RF	Dart	GBDT	SVR 误差	RF 误差	Dart 误差	GBDT 误差
CrO2	0.416	1.23	1.84	2.01	2.16	2.22	0.61	0.78	0.93	0.99
HfS2	1.061	2.83	2.71	2.47	2.58	2.75	-0.12	-0.36	-0.25	-0.08
NiS2	0.583	1.87	1.85	1.88	1.95	2.12	-0.02	0.01	0.08	0.25
PdSe2	1.463	1.62	1.60	1.76	1.55	1.56	-0.02	0.14	-0.07	-0.06
ScO2	0.665	4.41	3.81	4.05	4.07	3.84	-0.60	-0.36	-0.34	-0.57
SnO2	0.259	2.87	2.26	3.07	3.26	3.29	-0.61	0.20	0.39	0.42
ZrTe2	0.687	0.94	0.78	0.77	0.81	0.84	-0.16	-0.17	-0.13	-0.10
FeI2	3.098	1.26	1.28	1.21	1.21	1.27	0.02	-0.05	-0.05	0.01
CdCl2	0.205	5.82	5.70	5.83	5.36	5.46	-0.12	0.01	-0.46	-0.36
FeBr2	1.713	0.49	0.88	0.98	0.55	0.46	0.39	0.49	0.06	-0.03
MnCl2	2.775	0.05	3.49	3.53	3.21	3.25	3.44	3.48	3.16	3.20
PbCl2	1.577	5.25	5.33	4.81	4.95	5.10	0.08	-0.44	-0.30	-0.15

PbI2	1.577	3.34	3.18	3.31	3.40	3.44	-0.16	-0.03	0.06	0.10
SnBr2	2.499	4.27	4.70	4.59	4.34	4.60	0.46	0.32	0.07	0.33
TiBr2	0.758	1.42	1.39	1.62	1.67	1.79	-0.03	0.20	0.25	0.37
TiI2	0.596	1.09	0.90	1.56	1.41	1.36	-0.19	0.47	0.32	0.27
方差							0.8037	0.7808	0.6766	0.6856

上海大学游洋，杜婉，李惟驹，陈竞哲等人通过 SVR 算法、RF、GBR 算法预测的误差方差分别为 0.97,0.89,1.03[4]，远大于本项目组通过 SVR、RF、Lightgbm-GBDT、Lightgbm-Dart 等算法所计算出来的误差方差 0.80,0.78,0.68,0.69。由此可见，本文所预测带宽的精确度高于原有文献。

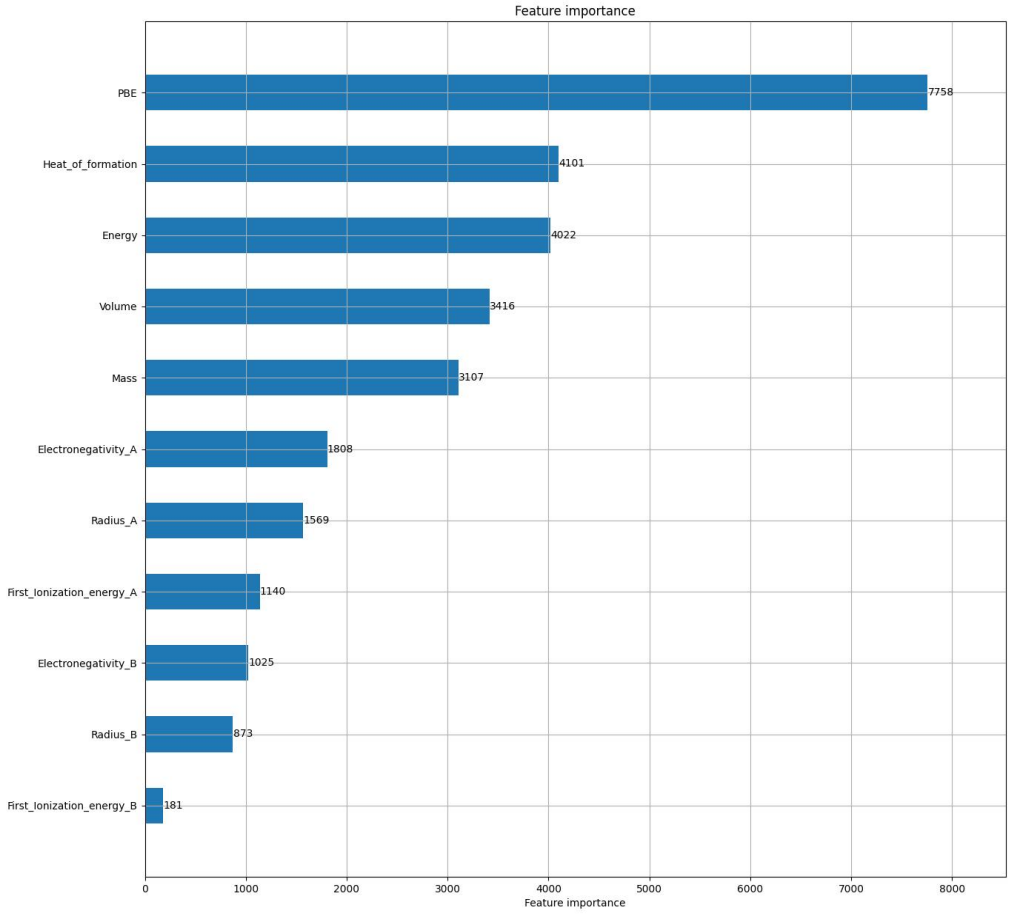


图 12 特征重要性排序

4 基于深度学习的原子掺杂及吸附预测

4.1 数据准备

本工作以 Impurities in 2D Materials Database(IMP2D)和 Computational 2D Materials Database (C2DB)公开材料数据库中选取了 MX2 型二维材料为研究对象，选取了基于第一性原理和密度泛函理论计算得到的 8679 个 MX2 型二维金属化合物数据作为数据集，提取基底材料(host)、掺杂材料(dopant)、掺杂

位置(site)、掺杂方式(defecttype)、融合(converged)、掺杂剂化合势(dopant_chemical_potential)、形成热(eform)、基底空间组(host_space)、能量(host_energy)、超胞(supercell)、基底质量(mass)、基底能量(energy)、基底形成热(hform)、基底 M,X 位元素的电负性(EN)、基底第一电离能(I)、基底原子有效半径(R)、掺杂元素电负性、掺杂原子第一电离能、掺杂原子半径以及基于 GGA-PBE 计算的带隙、HSE 带隙、GOWO 带隙等性质为特征变量，以总能量（en2）为预测目标，组成数据集。数据集包含基底材料 22 种，掺杂原子 79 种，掺杂方式包括吸附（adsorbate）和间隙（interstitial）两种，掺杂位置包括 ads1、ads2、int0、int1、int2、int3、int4。

4.2 数据预处理

因为使用了基底材料、掺杂原子、掺杂方式、掺杂位置作为数据集的特征数据，因此需要把这些数据进行转换，转换为对应的标签数值。同时把数据转化为浮点类型，以便于模型更好的训练。随后对数据进行标准化，把不同规模的数据统一到同一规模，从而使模型能够更有效的收敛。这是因为深度学习会自动调整权重以最大程度地拟合训练数据，但如果数据的规模不一致，则可能导致模型不能找到梯度下降求最优解，导致模型不能很好地拟合数据，从而无法达到期望的训练效果。数据标准化的方式为：把数据变换到均值为 0，标准差为 1 范围内。数据标准化公式见式（1）所示。

$$X' = \frac{x - mean}{\partial} \tag{1}$$

作用于每一列，mean 为平均值， σ 为标准差。

4.3 损失函数

均方误差（MSE）是较为常用的损失函数，通过计算预测值和真实值之间的平方和衡量两者之间的差异。在数据中存在异常值时，会产生较大的 MSE，因此需要对真实值中的异常值进行处理，在数据预测理时需要把缺失值进行处理，MSE 在固定学习率的情况下，可以表现出较好的结果，损失函数的梯度随着损失函数值的增大而增大。MSE 的公式见式（2）所示。

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 \tag{2}$$

式中，m 为样本个数，yi 为预测值，yi^为真实值。

4.4 评估指标

4.4.1 平均绝对误差（MAE）

评价回归算法的方法通常是将预测结果和真实值进行对比，评判差异大小。平均绝对误差（MAE）衡量的是目标值与预测值差值的绝对值，只考虑了平均模长，没有考虑方向，取值范围是 0 到正无穷。MAE 对于异常点有着更好的鲁棒性。MAE 的公式见式（3）所示。

$$\mathrm{MAE}=\frac{1}{M}\sum_{i=1}^{m}|(y_i-\hat{y}i)| \tag{3}$$

式中，m 为样本个数，yi 为预测值，yi^为真实值。

4.4.2 决定系数(R-Square)

决定系数（R-Square）与上述几种评价指标不同，R-Square 拥有上界与下界，在 0 到 1 之间，可以不用考虑量纲的影响，直接反映出模型预测效果的优劣，通常可以理解为待预测的数据能够被回归方程解释的比例，其值越接近 1，说明模型的预测效果越好，越接近 0，说明模型的预测效果越差。R-Square 的公式见式（4）所示。

$$R2=1-\frac{\sum_{i=1}^{m}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{m}(\bar{y}_i-y_i)^2} \tag{4}$$

其中，分子部分表示真实值与预测值的平方差之和，类似于均方差 MSE；分母部分表示真实值与均值的平方差之和，类似于方差 Var。

随着样本数量的增加，r2 必然增加，无法真正定量说明准确程度，只能大概定量。所以需要抵消样本数量的影响，而校正决定系数(Adjusted R-Square)，抵消了样本数量的影响，真正做到了 0~1，越大越好。Adjusted R-Square 的公式见式（5）所示。

$$R2\_\mathrm{adjusted}=1-\frac{(1-R2)(n-1)}{n-p-1} \tag{5}$$

其中，n 是样本数量，p 是特征数量。

4.5 实验步骤

4.5.1 数据划分

数据集按照 8:1:1 的比例划分为训练集、验证集和测试集。训练集用于训练模型，即模型拟合的数据样本集合。验证集用于确定网络结构或者控制模型复杂程度的超参数，是模型训练过程中单独留出的样本集，它可以用于调整模型的超参数和用于对模型的能力进行初步评估。通常用来在模型迭代训练时，用以验证当前模型泛化能力，以决定如何调整超参数。测试集用来评估模型最终模型的性能如何，测试集没有参于训练，主要是测试训练好的模型的准确能力等，只用于评价模型好坏的一个数据集。

4.5.2 模型训练

本文使用的神经网络模型是由四个具有 ReLU 激活的全连接层和一个 regression head 组成。

网络模型图如图 13 所示。

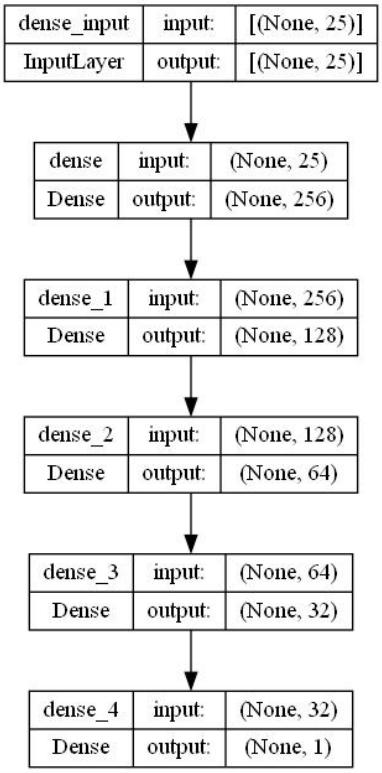


图 13　网络模型图

因为 en2 的取值是连续值，研究的问题是一个回归问题，所以采用均方误差（MSE）作为网络的损失函数，以获得更好的表现。采用 ReduceLROnPlateau 学习率动态调整网络的学习率(lr)方法，初始学习率为 0.001，当验证集 loss 有 120 个 step 的 loss 不变化时，调整学习率，学习率更新为 lr=lr*0.1，使用 EarlyStopping 早停，提前终止训练，当有 200 个 step 的 loss 不变化时，终止模型训练，来获得最好的模型拟合效果。采用 ADAM 优化器，Batchsize 取 128，epoch 取 5000，评估指标为平均绝对误差（MAE）和校正决定系数(Adjusted R-Square)。

在训练集中 5389 个样本的预测值和真实值之间的校正决定系数(Adjusted R-Square)、均方误差(MSE)、平均绝对误差(MAE)，结果如图 14、图 15 所示、图 16 所示。

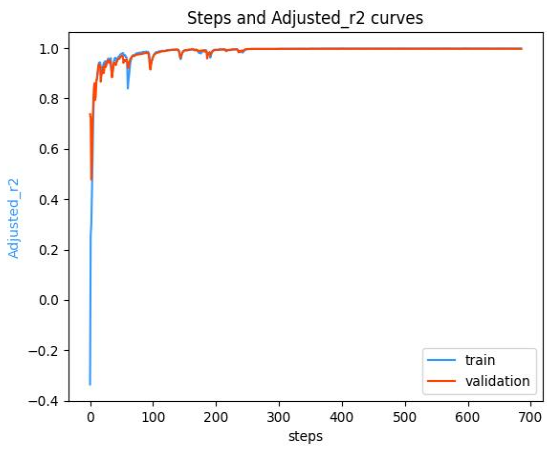


图 14 轮次与校正决定系数曲线

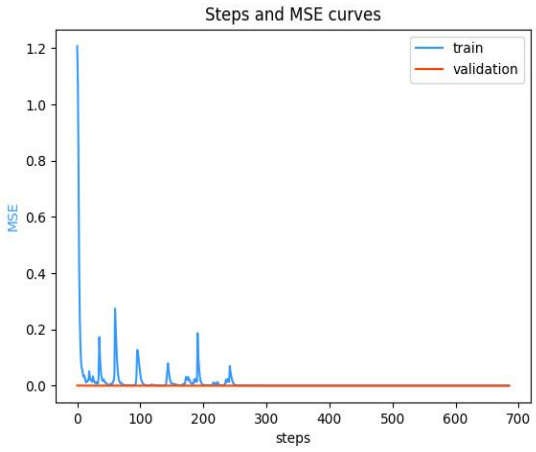


图 15 轮次与 MSE 曲线

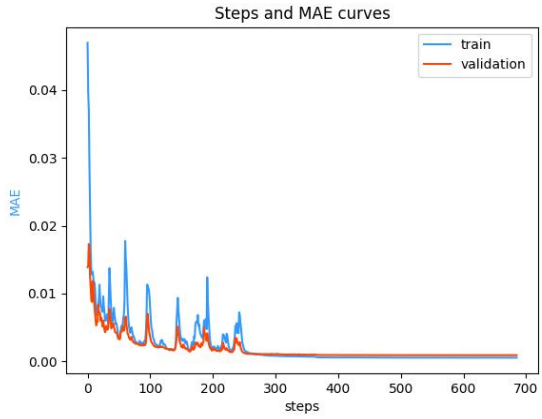


图 16 轮次与 MAE 曲线

在本研究中采用 Adjusted R-Square 和 MSE 作为神经网络模型回归性能的度量。图 14、图 15、图 16 分别为训练过程中训练轮次（epochs）与 Adjusted R-Square、MSE、MAE 的关系。可以看出随着训练轮次逐渐增加，Adjusted R-Square 越来越接近 1，MSE、MAE 越来越接近 0。本文选取训练轮次为 685 时的模型作为最终训练完成的模型。

随后从训练数据集中选取 674 个数据，来得到其预测值。如图 17 所示，横坐标为真实值(true_en2)，纵坐标为预测值(predicted_en2)。图中的蓝线为 Y=X 曲线，可以看出图中的点集中在 Y=X 直线附近，且预测值与真实值的 R2=0.9941，MSE=3.0624e-06，MAE=9.1720e-04 即预测值与真实值非常接近。图 18 为预测值与真实值误差分布直方图。横坐标为预测值与真实值差值的分布区间，其中差值的范围从-14~12，每 5 个偏差值为一个区间的范围。纵坐标为每个区间所占百分比，图中红色的线为正态拟合曲线，可以看出预测值与真实值的差值呈正态分布。综上所述，训练好的神经网络模型可以很精确的在训练集的数据上进行材料 en2 的预测。

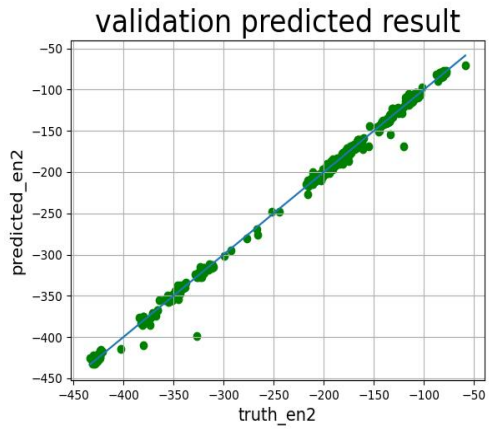


图 17 验证集的预测结果

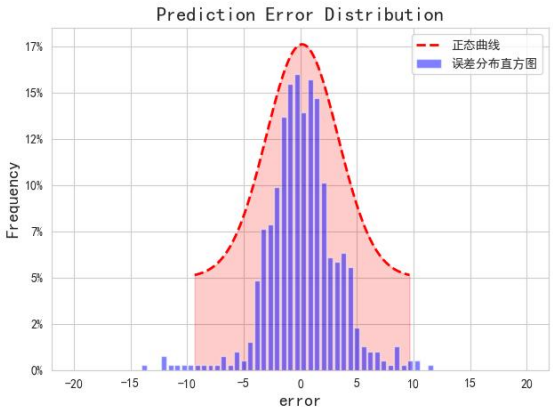


图 18 验证集预测与真实值误差分布图

4.5.3 模型预测

基于已训练完成的神经网络深度学习模型，可以对基底材料中掺杂原子的任意位置、任意方式的 en2 进行预测。本文测试数据集中包含 674 个数据，模型来验证神经网络的泛化性能。图 19 为 674 个测试数

据的预测值与真实值分布的散点图，横坐标为真实值(true_en2)，纵坐标为预测值(predicted_en2)。可以看出图中的点均匀分布在 Y=X 直线附近，且预测值与真实值的 Adjusted_R2=0.990，MSE=1.7862e-05，MAE=0.0011。图 20 为预测值和真实值误差的分布直方图，横坐标为预测值与真实值差值的分布区间，其中差值的范围主要从-10~5 之间,以每 5 个误差值为一个区间的范围。图中红色的线为正态拟合曲线，可以看出预测值与真实值的差值呈正态分布。基于以上结果可以看出训练出来的神经网络具有良好的泛化性能，可以在没有经过训练的数据集上取得相当不错的表现。也说明了所训练出来的深度学习模型有望被用于进行材料掺杂及吸附性能的优化设计。

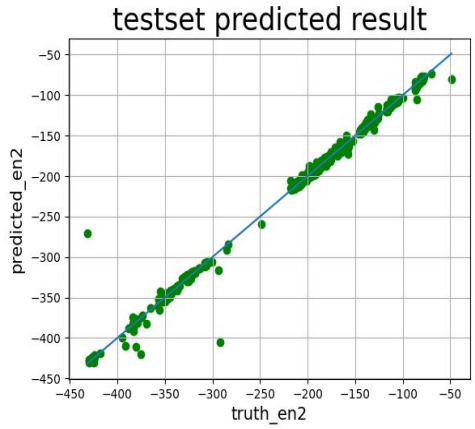


图 19 测试集的预测结果

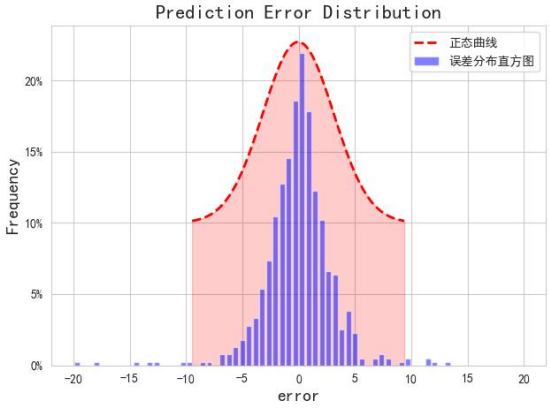


图 20 测试集预测与真实值误差分布图

5 结束语

将机器学习技术与材料科学领域的相关研究进行结合，是今后材料科学发展的大方向。对于传感材料，带隙起到重要作用，可以直接影响到传感器的敏感性和选择性。对于气体或重金属传感器而言，当吸附物到达传感器表面时，它们可能通过吸附在表面上，并影响表面电子结构。如果气体分子可以吸附在表面上，则传感器表面电子密度会发生变化，进而引起电学特性的变化。这种变化可以通过检测传感器的电学特性来检测气体的存在。使用二维材料设计传感器，需要提前评估作为敏感材料的二维材料的带隙。如果敏感带隙过低，则在吸附前后的导电性能对比不明显，影响到了传感器的灵敏度。尤其是带隙接近 0 的二维材料，要设计为电导式传感器，还需要进一步修饰其他原子，以使其具有一定的带隙。同时，当探测对象与本征二维材料之间的吸附作用较弱时，通常需要对二维材料进行修饰，以增加其吸附性。修饰可以通过在二维材料表面添加其他原子或分子来实现。然而，对于不同的基底材料，修饰哪种原子或分子，才能对某种气体具有良好的吸附效果，这正是传感材料设计中遇到的主要问题。在设计阶段通过 CVD 或 ALD 技术大量修饰不同的原子与分子来进行实验，实验成本非常高昂，错误几率非常高，难以找到有效的修饰材料和基底传感器的搭配。传统的计算方法和实验方法耗时、耗力，而且会受到不同环境的影响导致结果不稳定，通过机器学习技术可以大大节省财力物力还可以为材料科学未来的理论和实验指引方向。本文通过材料的性质数据对材料的能带和掺杂吸附总能量进行预测，是一种探索，也是一种尝试，为今后材料科学家们研究材料吸附掺杂指引了方向。本文对材料的性质数据，针对集成学习等方法和深度学习神经网络两种不同

的模型，最终对材料的能带和吸附掺杂的总能量进行了预测，预测效果达到了不错的精度。本文的具体研究内容如下：

（1)通过选用基于第一性原理和密度泛函理论计算得来的材料性质数据,如:M,X 位元素的电负性(EN)、第一电离能(I)、原子有效半径(R)、化合物形成热(Hf)、总能量(E)、晶体质量(m)、晶体体积(V)以及基于 GGA-PBE 计算的带隙作为特征数据对材料的能带进行预测。本文尝试了多次，对需要设置的参数进行了网格搜索的方法，得到了最优参数，使用集成学习算法，如 LightGBM 算法训练出来的模型效果良好，拟合度较好，而传统机器学习算法，如支持向量机 SVM 训练出来的拟合度一般，通过使用集成学习算法的模型误差较小。

（2）针对神经网络模型预测基底掺杂吸附总能量，因为使用了基底材料、掺杂原子、掺杂方式、掺杂位置作为数据集的特征数据，因此需要把这些数据进行转换，转换为对应的标签数值。同时把数据转化为浮点类型，以便于模型更好的训练。随后对数据进行标准化，把不同规模的数据统一到同一规模，模型才能够正常收敛。

（3）基于已训练完成的神经网络深度学习模型，可以对基底材料中掺杂原子的任意位置、任意方式的掺杂总能量进行预测。其预测真实值和误差值呈现正态分布基于以上结果可以看出训练出来的神经网络具有良好的泛化性能，可以在没有经过训练的数据集上取得相当不错的表现。也说明了所训练出来的深度学习模型有望被用于进行材料掺杂及吸附性能的优化设计。

参考文献：

[1] Dong Yanhua, She Anqi, Wang Ming, Liang Jiuxin, Sun Hongyu. Multivariate compound band gap prediction based on machine learning [C].Journal of Jilin Normal University (Natural Science Edition),2022,43(2):120-121

[2] Wan Xinyang,Zhang Yehui,Lu Shuaihua,Wu Yilei,Zhou Shuhua,Wang Jinlan. Machine learning accelerates the search for novel diperovskite oxide photocatalysts[J/OL].Acta Physica Sinica:1-21[2022-08-13].http://kns.cnki.net/kcms/detail/11.1958.04.20220602.1835.004.html

[3] Tran K, Ulisi Z W. Natural Catalysis, 2018, 1, 696.SONG Qinggong，CHANG Binbin，DONG Shanshan，GU Weifeng，KANG Jianhai，WANG Mingchao，LIU Zhifeng. Machine learning and its influence on materials research and development [C].Materials Reports，2022,36(1):3

[4] 游洋，杜婉，李惟驹，陈竞哲，基于机器学习的二维材料带隙预测[N].上海：上海大学学报，2020.