

## 卷积神经网络辅助无机晶体弹性性质预测

刘宇杰<sup>1#</sup> 王振宇<sup>1#</sup> 雷航<sup>1#</sup> 张国宇<sup>1)</sup> 咸家伟<sup>2†</sup>  
高志斌<sup>1‡</sup> 孙军<sup>3)</sup> 宋海峰<sup>2)</sup> 丁向东<sup>3)</sup>

1) (西安交通大学材料科学与工程学院, 金属多孔材料全国重点实验室, 西安 710049)

2) (北京应用物理与计算数学研究所, 计算物理全国重点实验室, 北京 100088)

3) (西安交通大学材料科学与工程学院, 金属材料强度全国重点实验室, 西安 710049)

# 同等贡献作者.

† 通信作者. E-mail: xian\_jiawei@iapcm.ac.cn

‡ 通信作者. E-mail: zhibin.gao@xjtu.edu.cn

### 1. 数据来源

Materials Project (MP)数据库中的晶体结构均基于密度泛函理论(DFT)计算优化获得, 而非直接来源于实验测定。其计算框架采用针对固体优化的 PBEsol 广义梯度近似泛函, 并通过平面波基组(截断能设置为 520 eV)和投影缀加波(PAW)赝势(包含半核电子处理, 如过渡金属 d 电子)完成电子结构计算。 $k$  点网格通过 Monkhorst-Pack 方法自动生成, 确保  $k$  点间距不超过  $0.03 \text{ \AA}^{-1}$  以满足布里渊区积分精度要求。结构优化过程中, 原子位置和晶胞参数均收敛至力小于  $0.01 \text{ eV}/\text{\AA}$ 、应力小于  $0.1 \text{ GPa}$  的严格标准, 从而保证计算结果的物理合理性与热力学稳定性。MP 的结构优化参数经过严格验证, 其晶格常数与实验值的平均偏差小于 1%。然而, 计算结构可能忽略实验中的缺陷、温度效应和杂质影响。若目标值(如弹性常数、带隙)对结构敏感, 需结合实验数据或更高精度方法(如杂化泛函)验证<sup>[S1]</sup>。

Merchant 数据集<sup>[S2]</sup>通过图神经网络发现未在实验或数据库中记录的新结构, 并通过高通量计算优化。结构弛豫主要使用 Perdew-Burke-Ernzerhof 泛函, 部分实验验证结构采用更精确的 rSCAN 泛函。计算采用 PAW 赝势, 与 Materials Project 标准一致, 平面波截断能设为 520 eV。 $k$  点生成遵循 Materials Project 标准, 使用 Gamma-centred 方法, 并通过 pymatgen 的 MPNonSCFSet 自动生成。结构优化分为离子弛豫和静态计算两步, 磁性初始化采用铁磁性假设。在 381000 个稳定结构中, 736 个实验验证结果已添加至 ICSD 数据库, 其余为计算预测。

## 2. CGCNN 模型参数设置

训练迭代步数、卷积层数量和隐藏层数量都属于超参数,无法通过优化算法来优化,为了让模型更好的收敛,我们采用了自适应矩估计(adaptive moment estimation, Adam)优化方法(初始学习率为 0.001)。Adam 结合了动量和均方根反向传播(root mean square propagation, RMSProp)的优点,通过计算梯度的一阶矩(均值)和二阶矩(未中心化的方差)来动态调整学习率,对初始学习率不敏感。

最终我们的模型测试集平均绝对误差(mean absolute error, MAE)达到  $0.0981 \log_{10}(\text{GPa})$  (剪切模量)和  $0.0790 \log_{10}(\text{GPa})$  (体积模量),验证了其预测精度。为了获取更好的模型,在设置 Adam 优化器的基础上,我们手动设置了训练迭代步数、卷积层数量和隐藏层数量,选取各个超参数条件下验证集最好的模型,并计算对应模型的训练集,验证集和测试集的 MAE 和  $R^2$  score 以此来确定参数。

图 S1 展示了随着训练步数增加,训练集(Training set)、验证集(Validation set)和测试集(Testing set)的 MAE 和  $R^2$ ,虽然训练集的 MAE 和  $R^2$  分别随着训练步数(epoch)的增加逐渐下降和上升,但是验证集和测试集的结果并没有在 epoch 增加的时候有明显提升,说明增加训练步数不一定可以使模型变得更好,反而会增加过拟合的风险。所以我们选用 30 步的 epoch,一方面减少过拟合的风险,另一方面节省计算资源。

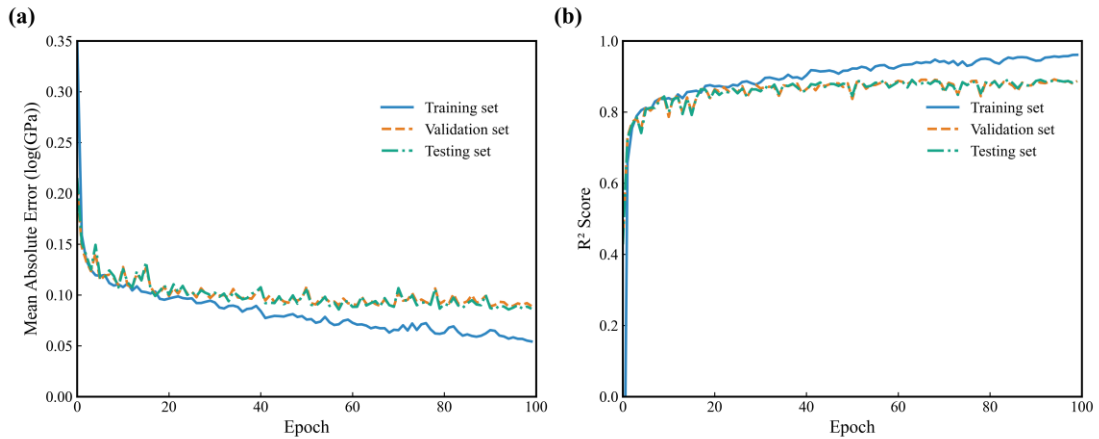


图 S1 卷积层为 3、隐藏层为 2 时, 剪切模量(a)和体模量(b)在不同训练步数下的训练集、验证集和测试集的 MAE 和  $R^2$  变化

Fig. S1. MAE and  $R^2$  variations of shear modulus (a) and bulk modulus (b) for training, validation, and test sets with different training steps when the convolutional layer is 3 and the hidden layer is 2.

为了进一步确认其他超参数,我们手动设置了训练迭代步数、卷积层数量和隐藏层数量,选取各个超参数条件下验证集最好的模型,并计算对应模型的训练

集、验证集和测试集的 MAE 和  $R^2$  score。从图 S2 可以看到，不论是在训练剪切模型和体模量时，不同的超参数条件下在学习率为 0.001 时往往有更好的效果，因此我们将学习率调整为 0.001。不过，当超参数变化时，大部分超参数条件下，模型的验证集和测试集的 MAE 和  $R^2$  的变化都比较接近，说明 CGCNN 模型对于不同参数的设置均能够提取到晶体中原子、元素和键的特征，具有较好的性能。少数几个超参数学习率为 0.01 对应的模型的 MAE 和  $R^2$  很差的原因可能是因为学习率太大，模型没有收敛的原因。

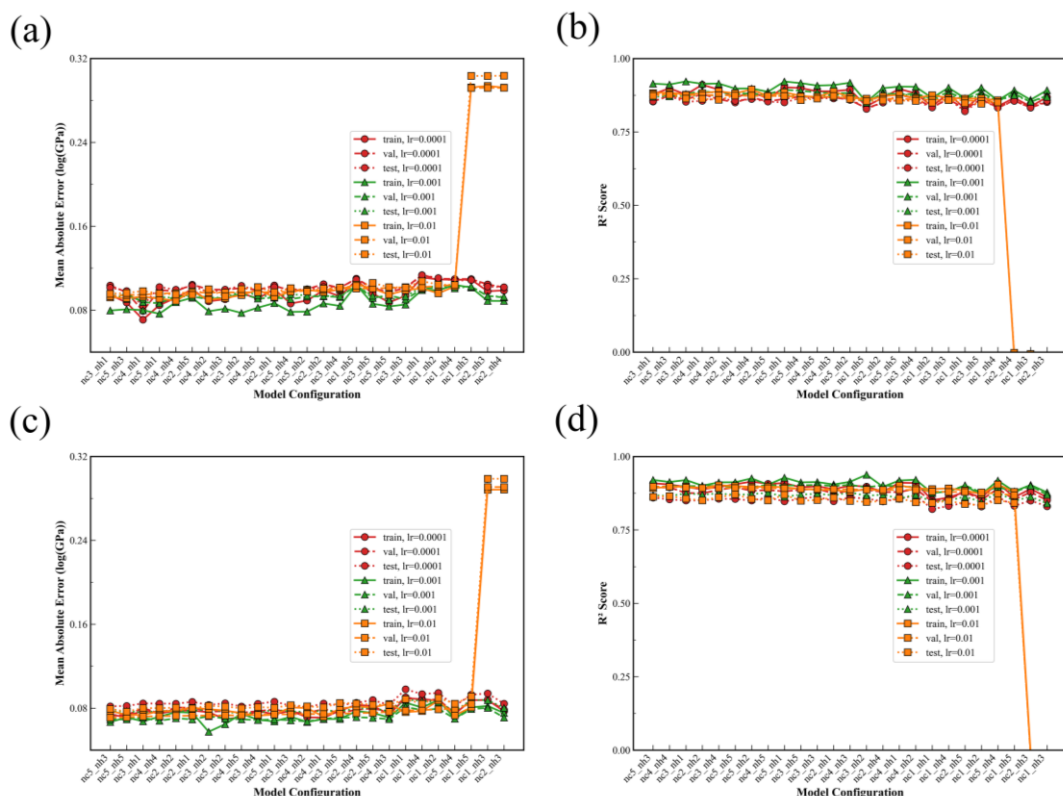


图 S2 不同卷积层数(nc)、隐藏层数(nh)、学习率(lr)，CGCNN 模型剪切模量(a), (b)和体模量(c), (d)的训练集(Train)、验证集(Val)和测试集(Test)的平均绝对误差(MAE)和决定系数( $R^2$ )  
Fig. S2. Mean absolute error (MAE) and coefficient of determination ( $R^2$ ) for the training set (Train), validation set (Val), and test set (Test) of the CGCNN model with different number of convolutional layers (nc), number of hidden layers (nh), and learning rate (lr) for the CGCNN model with shear modulus (a), (b) and bulk modulus (c), (d).

在确定最优学习率为 0.001 后，我们通过热图(图 S3)可视化不同超参数组合的性能对比。该热图以隐藏层维度为横坐标、卷积层深度为纵坐标，色块数值分别代表训练集、验证集和测试集的平均绝对误差(MAE)与决定系数( $R^2$ )的均值。针对剪切模量预测任务，当卷积层为 5、隐藏层维度为 1 时，模型达到最优性能(测试集 MAE = 0.0321,  $R^2 = 0.8969$ )，其  $R^2$  值较原基准模型(卷积层 3、隐藏层 2,  $R^2 = 0.8905$ )仅提升 0.64%。值得注意的是，体积模量预测任务的最优参数为

卷积层 3、隐藏层 2 (测试集 MAE = 0.0287,  $R^2 = 0.9124$ )。然而, 如图 S4 所示, 当两类任务均采用最优模型结构时, 预测统计结果比较混乱, 推测其可能源于模型结构特异性导致的特征提取差异。综合考量跨任务鲁棒性, 最终选定统一参数配置: 卷积层 3、隐藏层 2、学习率 0.001。

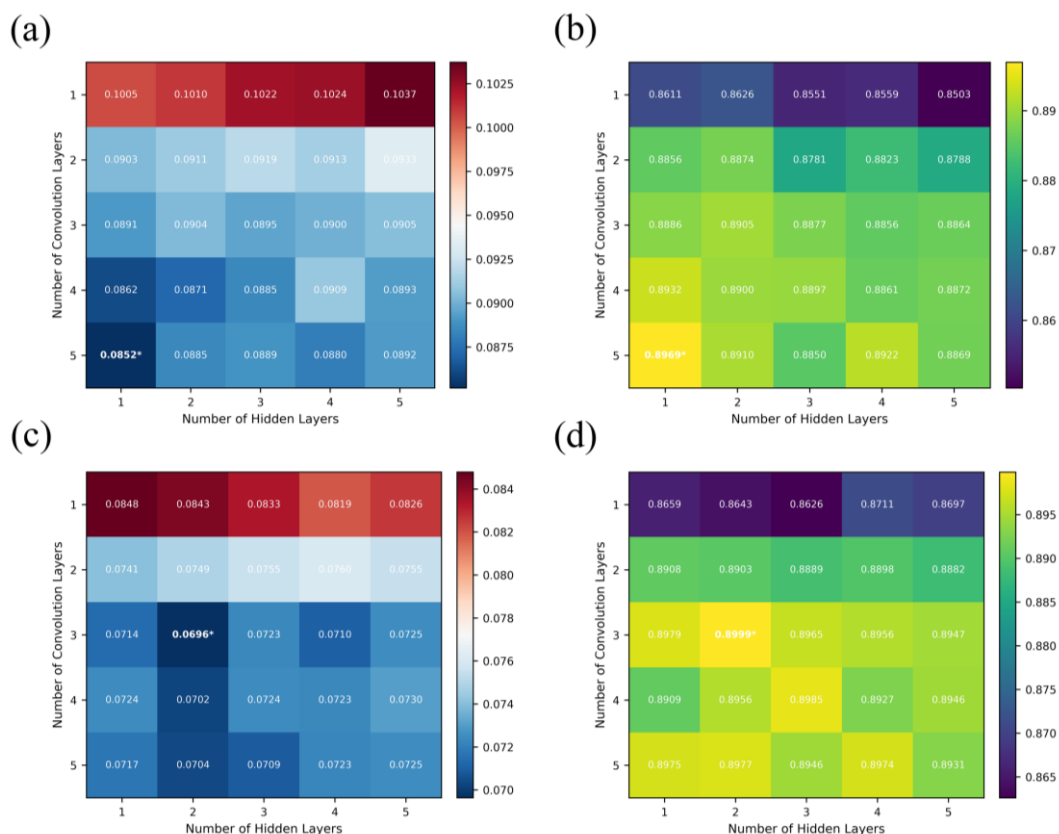


图 S3 学习率为 0.001 时, 不同卷积层数(Number of Convolution Layers)与隐藏层数(Number of Hidden Layers)的 CGCNN 模型架构下, 剪切模量性能评估指标 MAE (a)和  $R^2$  (b)与体模量性能评估指标 MAE (c)和  $R^2$  (d)。其中 MAE 和  $R^2$  均采用训练集(Train)、验证集(Val)和测试集(Test)的均值, 最优参数组合通过\*号标识

Fig. S3. Performance evaluation metrics (a), (c) MAE and (b), (d)  $R^2$  of (a), (b) shear modulus and (c), (d) bulk modulus for CGCNN model with different number of Convolution Layers and Hidden Layers at a learning rate of 0.001. MAE and  $R^2$  both use the mean values of the training set (Train), validation set (Val) and test set (Test), and the optimal parameter combinations are identified by the \* sign

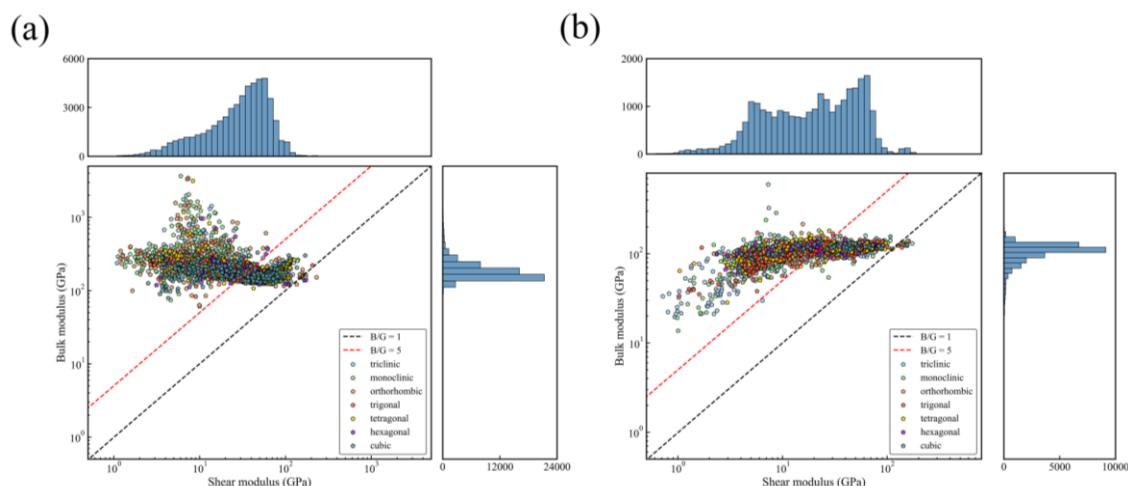


图 S4 MPED 数据集(a)和 NED 数据集(b)剪切模量和体模量分布，用不同颜色对应不同的晶系；条形图表示材料的剪切模量和体模量分布

Fig. S4. Shear and bulk modulus distributions for the MPED dataset (a) and the NED dataset (b). Different colors correspond to different crystal systems. Bars indicate the shear and bulk modulus distributions of the materials.

### 3. 不同机器学习模型与 CGCNN 对比

为了彰显 CGCNN 模型的优势，我们对比了 CGCNN 和其他机器学习模型的性能，例如：随机森林(RandomForest)、极端梯度提升(eXtreme Gradient Boosting, XGBoost)，支持向量回归(Support Vector Regression, SVR)、梯度提升(GradientBoosting)和决策树(DecisionTree)。为了和晶体结构保持相关性，我们选取原胞的平均原子序数、平均原子质量、平均电负性、空间群号、密度和每原子体积为特征构建模型。

图 S5—10 分别为采用相同训练集、验证集和测试集数据时，剪切模量和体模量的晶体图卷积神经网络(CGCNN)、随机森林(RandomForest)、极端梯度提升(XGBoost)、支持向量回归(SVR)、梯度提升(GradientBoosting)和决策树(Decision Tree)模型的训练集、验证集和测试集真实值与预测值结果对比。可以看到随机森林和决策树在训练集上有很好的拟合效果，这固然很好，但存在过拟合的风险，而模型的验证集和测试集的性能可以评价其在训练集之外的泛化能力。表 S1 和表 S2 分别为剪切模量和体模量在不同模型上训练集、验证集和测试集的 MAE 和  $R^2$  指标。从这些对比中可以看到，无论是剪切模量还是体模量，CGCNN 模型都具有更好的性能。

表 S1 相同训练集、验证集和测试集条件下，剪切模量在不同模型上的性能指标(MAE 和  $R^2$ )

Table S1. Performance metrics (MAE and  $R^2$ ) of shear modulus on different models under the same training, validation and test set conditions.

模型	训练集 MAE	训练集 $R^2$	验证集 MAE	验证集 $R^2$	测试集 MAE	测试集 $R^2$
CGCNN	0.0773	0.922	0.0957	0.879	0.0981	0.8704
RandomForest	0.0479	0.9715	0.1309	0.7787	0.1292	0.7845
XGBoost	0.1096	0.847	0.1368	0.7712	0.1352	0.7787
SVR	0.1429	0.7316	0.1464	0.7297	0.1512	0.726
GradientBoosting	0.1431	0.7453	0.1534	0.7222	0.1532	0.7264
DecisionTree	0	1	0.1726	0.6318	0.1786	0.6097

表 S2 相同训练集、验证集和测试集条件下，体模量在不同模型上的性能指标(MAE 和  $R^2$ )

Table S2. Performance metrics (MAE and  $R^2$ ) of the bulk modulus on different models under the same training, validation and test set conditions.

模型	训练集 MAE	训练集 $R^2$	验证集 MAE	验证集 $R^2$	测试集 MAE	测试集 $R^2$
CGCNN	0.0575	0.9379	0.0724	0.8955	0.079	0.8662
RandomForest	0.0382	0.9766	0.1045	0.8269	0.109	0.8133
DecisionTree	0	1	0.1439	0.6878	0.1387	0.6928
XGBoost	0.0838	0.8969	0.1068	0.8263	0.1098	0.8171
SVR	0.1101	0.8089	0.1144	0.8053	0.1188	0.798
GradientBoosting	0.1107	0.8213	0.1189	0.7996	0.122	0.7973

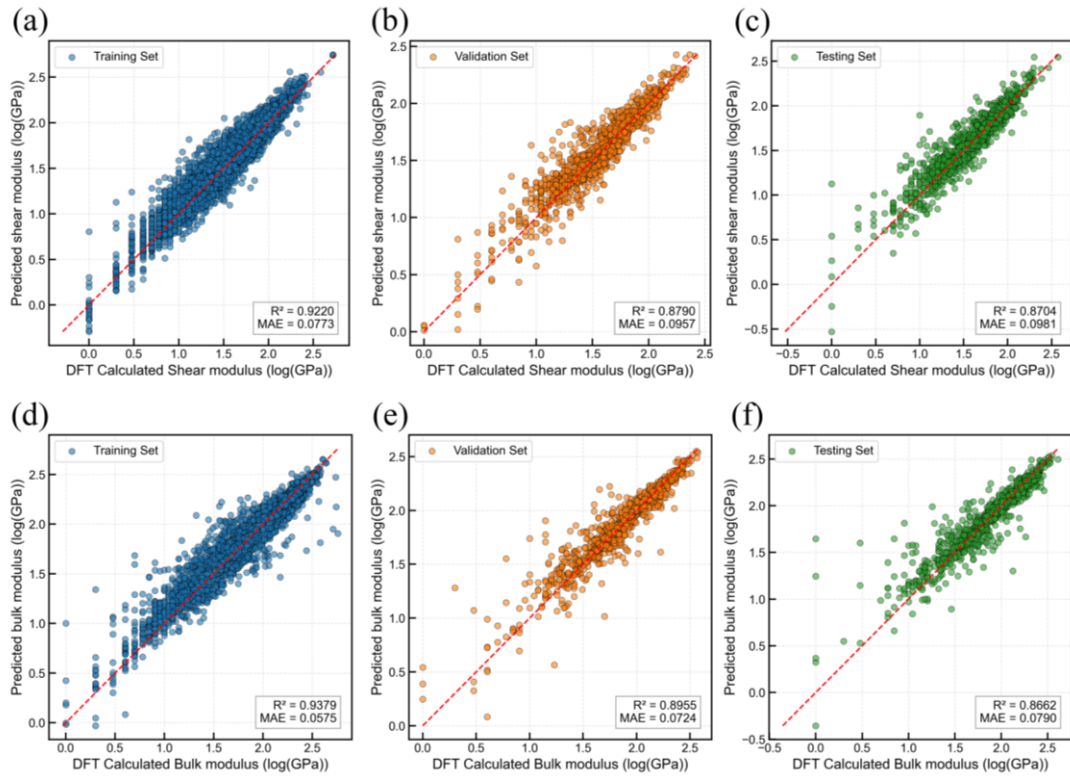


图 S5 CGCNN 剪切模量(a)–(c)和体模量(d)–(f)训练集、验证集和测试集真实值与预测值对比

Fig. S5. Comparison between true values and predicted values for the training, validation and test sets of CGCNN for shear modulus (a)-(c) and body modulus (d)-(f).

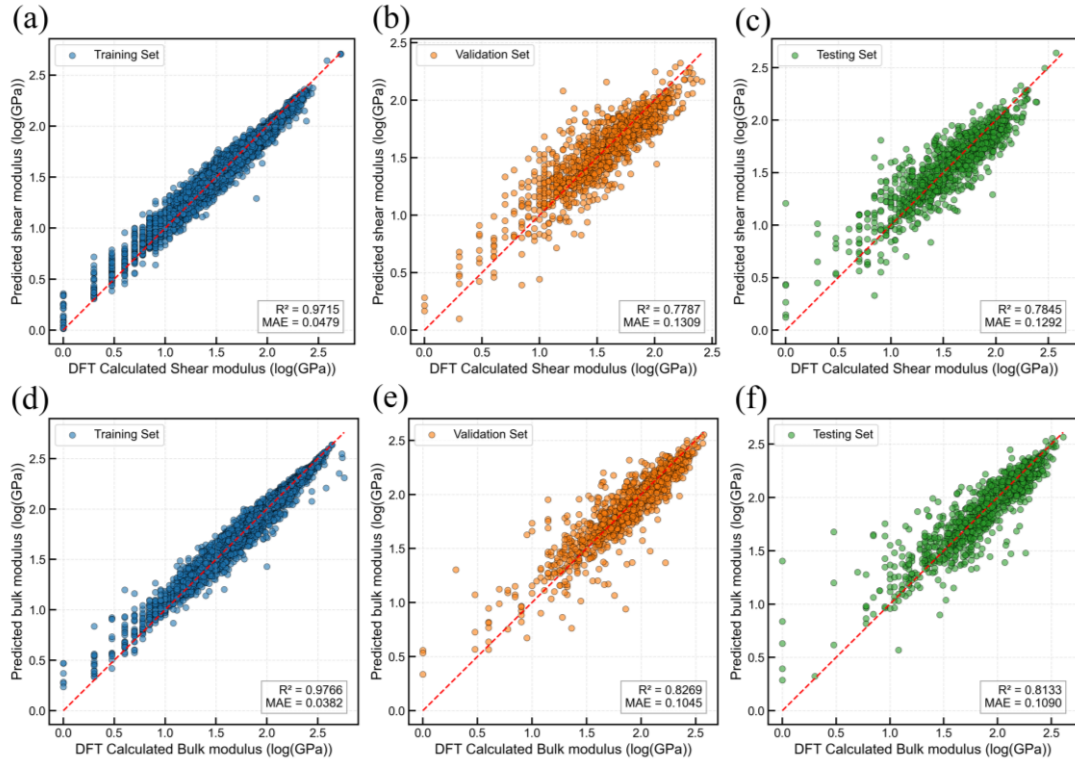


图 S6 随机森林(RandomForest)剪切模量(a)–(c)和体模量(d)–(f)训练集、验证集和测试集真实值与预测值对比

Fig. S6. Comparison between true values and predicted values for the training, validation and test sets of RandomForest for shear modulus (a)-(c) and body modulus (d)-(f).

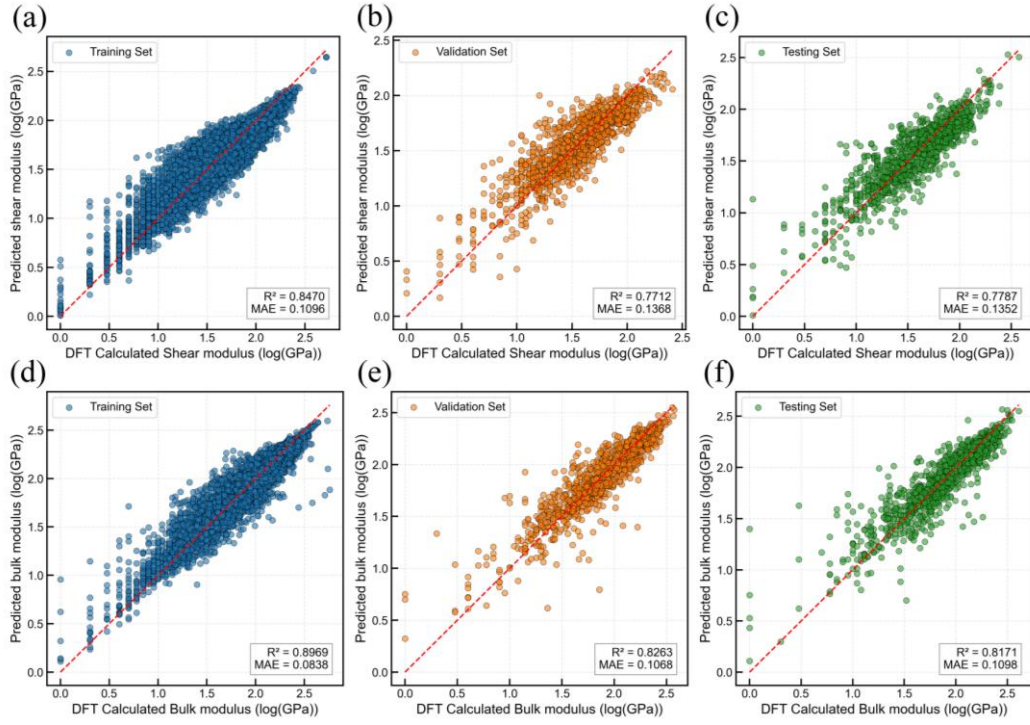


图 S7 极端梯度提升(XGBoost)剪切模量(a)–(c)和体模量(d)–(f)训练集、验证集和测试集真实值与预测值对比

Fig. S7. Comparison between true values and predicted values for the training, validation and test sets of Extreme Gradient Boosting (XGBoost) for shear modulus (a)-(c) and body modulus (d)-(f).

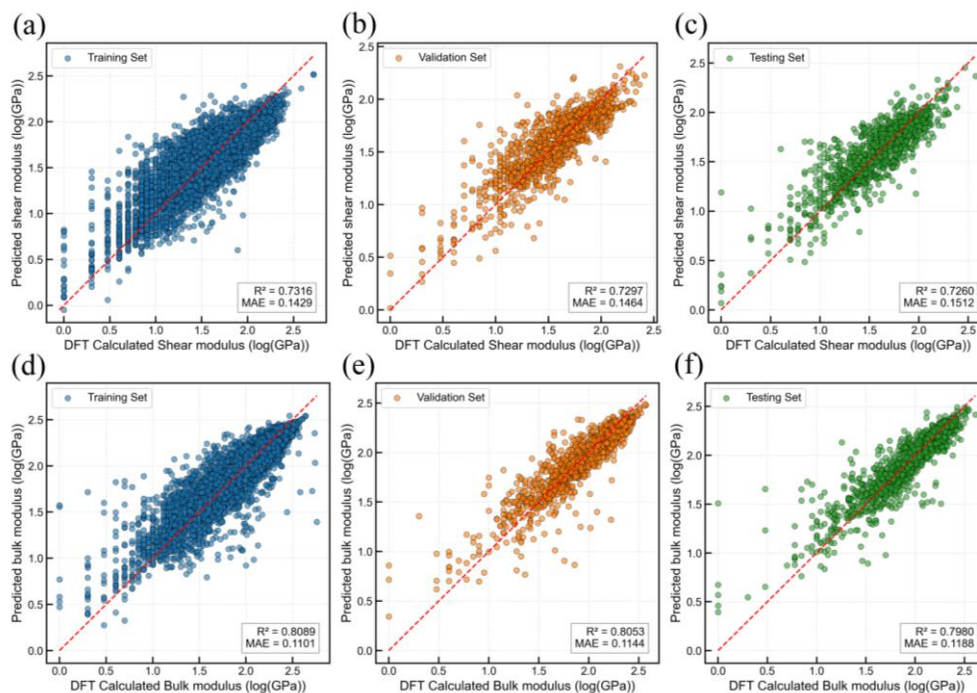


图 S8 支持向量回归(SVR)剪切模量(a)–(c)和体模量(d)–(f)训练集、验证集和测试集真实值与预测值对比

Fig. S8. Comparison between true values and predicted values for the training, validation and test sets of support vector regression (SVR) for shear modulus (a)-(c) and body modulus (d)-(f).

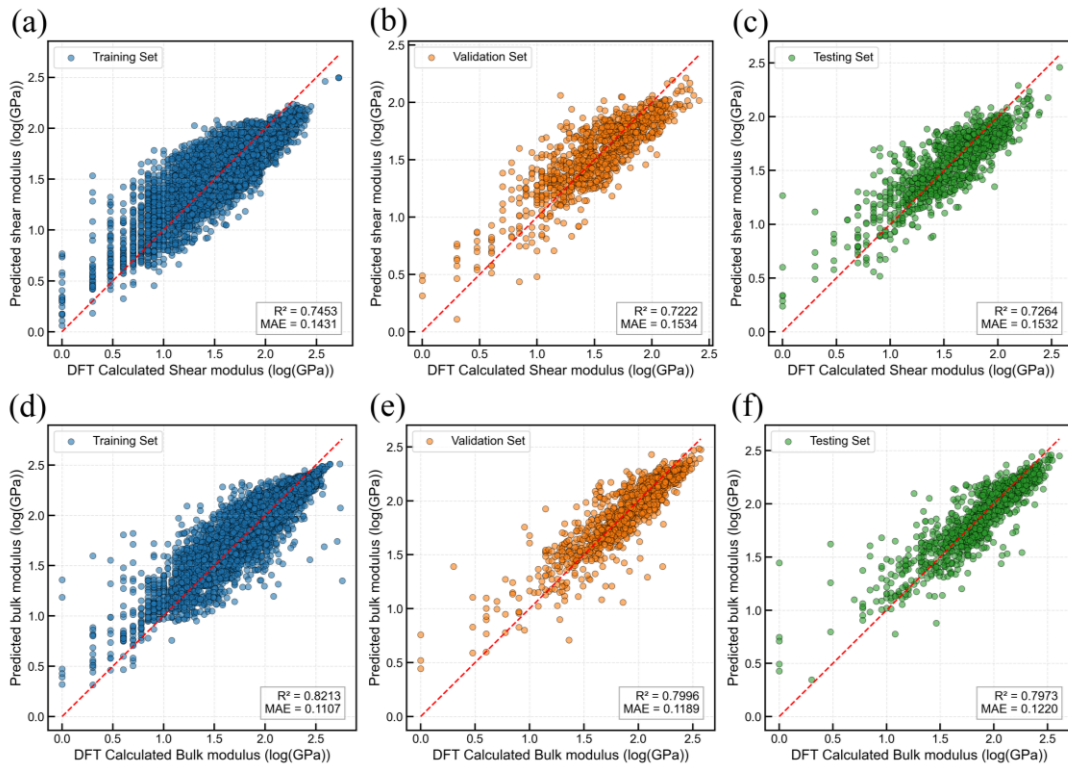


图 S9 梯度提升(GradientBoosting)剪切模量(a)–(c)和体模量(d)–(f)训练集、验证集和测试集真实值与预测值对比

Fig. S9. Comparison between true values and predicted values for the training, validation and test sets of GradientBoosting for shear modulus (a)-(c) and body modulus (d)-(f).

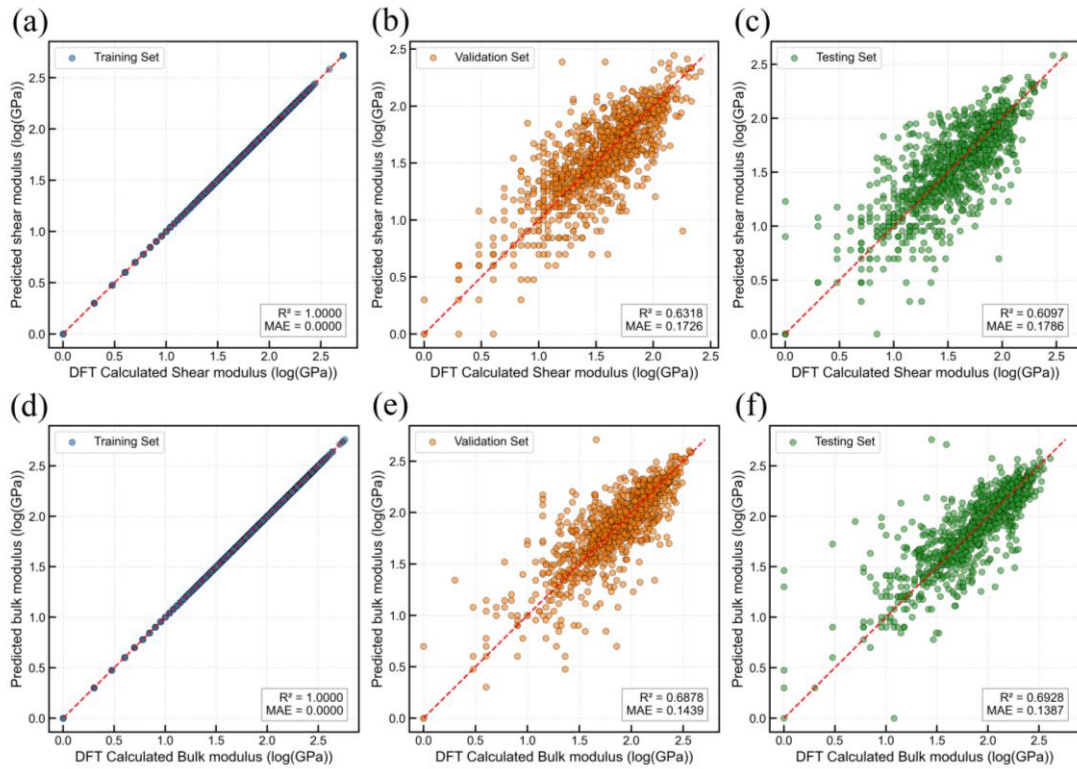


图 S10 决策树(DecisionTree)剪切模量(a)–(c)和体模量(d)–(f)训练集、验证集和测试集真实值与预测值对比

Fig. S10. Comparison between true values and predicted values for the training, validation and test sets of DecisionTree for shear modulus (a)-(c) and body modulus (d)-(f).

### 参考文献

- [S1] Jain A, Ong S P, Hautier G, et al. 2013 *APL Materials* **1** 011002  
[S2] Merchant A, Batzner S, Schoenholz S S, et al. 2023 *Nature* **624** 80