文章编号: 1000-324X(2021)11-1178-07 **DOI:** 10.15541/jim20200748

基于机器学习训练金属离子吸附能预测模型的研究

张瑞鸿[1], 魏 鑫[2], 卢占会[1], 艾玥洁[3]

(华北电力大学 1. 数理学院; 2. 控制与计算机工程学院; 3. 环境科学与工程学院 资源与环境系统优化教育部重点实验室, 北京 102206)

摘 要: 本研究通过密度泛函理论对氧化石墨烯和金属离子的吸附行为进行理论模拟。基于机器学习方法训练预测模型的过程中, 缺失值采用推荐系统中广泛使用的奇异值分解方法处理, 并用梯度提升机解释了影响吸附能的重要因素。结果发现吸附体系中存在九种特征可为吸附能提供 90%的累积重要性, 分别为离子半径、零点振动能量、密立根电荷、沸点、偶极矩、原子量、摩尔定容热容、自旋多重度和键长。定量评估了六种回归方法的预测精度, 包括支持向量回归、岭回归、随机森林、极端随机森林、极端梯度提升和轻梯度提升机。结果表明, 机器学习方法可提供足够的吸附能预测准确性, 其中极端随机森林方法表现出最优的预测性能, 均方误差仅为 0.075。该模型用于香兰素吸附金属离子的测试, 验证了基于机器学习训练金属离子吸附能预测模型的可行性, 但仍需进一步提高其泛化能力。本研究基于机器学习预测吸附能, 简化预测过程、节省计算时间, 可为吸附去除金属离子的理论和实验研究提供参考。

关 键 词: 机器学习; 密度泛函理论; 吸附能; 金属离子; 极端随机森林

中图分类号: TQ174 **文献标志码:** A

Training Model for Predicting Adsorption Energy of Metal Ions Based on Machine Learning

ZHANG Ruihong[1], WEI Xin[2], LU Zhanhui[1], AI Yuejie[3]

(1. College of Mathematics and Physics, North China Electric Power University, Beijing 102206, China; 2. College of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China; 3. MOE Key Laboratory of Resources and Environmental System Optimization, College of Environmental Science and Engineering, North China Electric Power University, Beijing 102206, China)

Abstract: The adsorption behavior of graphene oxide and metal ions was simulated theoretically by density functional theory. In the process of training the prediction model based on the machine learning method, the missing values were processed by matrix completion method, which was widely used in the recommendation systems, and gradient boosting machine (GBM) was trained to explain the importance of factors that affect the adsorption energy. The result showed that nine properties of the adsorption, namely ionic radius, zero-point vibration energy, Mulliken charge, boiling point, dipole moment, atomic weight, molar heat capacity at constant volume (CV), spin multiplicity and bond length, were found to provide 90% importance of the cumulative adsorption energy. Then six regression

收稿日期: 2020-12-31; **收到修改稿日期:** 2021-04-15; **网络出版日期:** 2021-06-01

基金项目: 国家自然科学基金(22076044); 国家重点研发计划(2017YFA0207002); 中央高校基础研究经费(2017YQ001)
National Natural Science Foundation of China (22076044); National Key Research and Development Program of China (2017YFA0207002); Fundamental Research Funds for the Central Universities (2017YQ001)

作者简介: 张瑞鸿(1996–), 女, 硕士研究生. E-mail: zhangruihong@ncepu.edu.cn
ZHANG Ruihong(1996–), female, Master candidate. E-mail: zhangruihong@ncepu.edu.cn

通信作者: 艾玥洁, 副教授. E-mail: aiyuejie@ncepu.edu.cn
AI Yuejie, associate professor. E-mail: aiyuejie@ncepu.edu.cn

methods, including support vector regression, ridge regression, random forest, extremely randomized trees, extreme gradient boosting, and light gradient boosting machine, were used to quantitatively evaluate the prediction accuracy. The results showed that machine learning could provide sufficient accuracy to predict adsorption energy. Among them, extremely randomized trees displayed the best prediction performance, with a mean square error only 0.075. Furthermore, the trained model was tested in a system of vanillin adsorbing metal ions, verifying the feasibility of training the prediction model of adsorption energy based on machine learning. But it is still necessary to be further improved. In general, this research takes the advantage of machine learning on the basis of saving experimental time to provide an instructive reference for theoretical research on metal ion removal.

Key words: machine learning; density functional theory; adsorption energy; metal ions; extremely randomized trees

随着工业发展, 重金属离子以及放射性元素的排放和积累, 对生态环境以及人类健康产生重大威胁。因此, 开发和合成对金属污染物具有更高亲和力、容量和选择性的材料, 成为无机材料领域和环境化学领域的关键问题之一。其中, 氧化石墨烯(Graphene oxide, GO)因其合成简便、成本低廉以及优异的吸附性能, 被广泛应用于金属污染物的吸附去除[1-2]。吸附能是吸附过程中评价吸附性能的关键参数之一, 可通过多种方法估计或计算其数值。GO与金属污染物间的吸附能参数可通过对二者相互作用作理论模拟而得到[3]。目前计算化学领域广为接受的方法是使用密度泛函理论(Density functional theory, DFT)计算吸附反应前后的分子能量差值。然而, 对于复杂的分子结构, 精确的理论计算将消耗大量机时。

近年来, 利用机器学习(Machine learning, ML)建立有针对性或泛化性能良好的模型, 有望解决长期以来寻找优质材料和预测反应性能这两大难题, 并显著降低计算成本。ML 是一系列基于统计学的方法, 利用各种数据挖掘算法从历史数据中获取信息, 预测未知数据。人们现已开发了诸多特定的 ML 算法[4], 广泛应用于高通量材料筛选和材料性能预测等[5-8], 如 Feng 等[9]建立的一种神经网络模型, 可基于原子种类和电荷分布, 对材料的键能作高精度和高准确性预测, 为优化化学反应、加速分子设计和其他重要应用提供了经济有效的工具; Lu 等[10]使用 ML 算法结合 DFT 计算开发了一种靶向驱动方法, 用于发现稳定的无铅钙钛矿, 成为使用 ML 做高通量材料筛选的范例。Tehrani 等[11]利用量子化学计算结合支持向量机(Support vector machine, SVM)提供了一种独特的框架, 以预测无机材料的机械性能。除了进行材料筛选和性能预测外, ML 在为新材料设计提供新思路方面, 与传统的理论计算相比具有更大优势[12]。例如, Brockherde 等[13]利用神经网络预测新的化合物, 节省了大量计算时间, 同时获得了精确结果。Xiao 等[14]利用高通量计算与数据挖掘技术成功预测了新型晶体结构, 展示了由数据驱动的新材料发现研究范式, 在计算材料化学领域的可行性。

虽然 ML 已越来越多地应用于高通量筛选、材料性能预测, 甚至新型材料设计, 但利用其预测材料的吸附性能却鲜见报道。Panapitiya 等[15]使用 ML 中的随机森林算法, 构建了 CO 分子在合金纳米团簇上的吸附能与纳米团簇结构间的构效关系, 进而从结构出发, 较准确地预测了材料对 CO 的吸附能。Pardakhti 等[16]引入化学描述符作吸附结构分析和化学描述符评估, 结合决策树、泊松回归法、支持向量机和随机森林等多种 ML 算法, 对金属有机框架的甲烷吸附性能进行了预测, 大大提高了预测精度。然而, 这些研究大多侧重于从结构或实验数据出发, 并未对 ML 在材料吸附金属离子性能方面的建模和预测作深入研究。

本研究以 GO 和金属离子相互作用为研究对象, 基于密度泛函理论, 结合理论计算结果, 探讨吸附能和金属离子基本状态信息之间的定量关系; 选择材料的 21 个特征参数, 基于 ML 训练并比较了 6 种模型的预测准确性; 依据打分机制作定量评估, 得到性能最优的预测模型。最后, 本研究将该模型应用于多个材料体系的吸附能预测, 验证了模型对不同吸附剂的有效性。

1　实验方法

1.1　基于密度泛函理论的参数计算

本研究选取了一个包含 62 个碳原子的碳环结构以模仿石墨烯表面, 通过在石墨烯结构上添加羰基, 最终得到了氧化石墨烯的模型。所有结构优化及能量计算均使用 Gaussian09 软件包完成, DFT 计算选择

Perdew-Burke-Ernzerhof(PBE)交换相关函数完成[17-19]。对 C、H、O 原子和金属原子分别使用 6-31G(d)基组和 SSD 基组进行几何优化[3], 使用导体极化连续介质模型(Conductor polarized continuum model)对溶剂效应进行模拟[20]。根据下式对吸附能 E_{ad}(eV)进行计算：

$$E_{ad}=E_{(GO,\ M^{n+})}-E_{(GO-M^{n+})} \quad (1)$$

其中, $E_{(GO,\ M^{n+})}$表示分离的 GO 和 M^{n+}体系的总能量(M^{n+}表示不同金属离子), 而 $E_{(GO-M^{n+})}$代表优化后的复合物的总能量, E_{ad} 表示最终吸附能。

1.2 数据的预处理

在后续研究中, 开发了一种 ML 预测方案用于预测金属离子在 GO 上的吸附能, 涉及到 64 种金属离子。为解决建模过程中数据采集不完整和可能存在的模型稳定性波动问题, 本研究采用奇异值分解算法(Singular value decomposition, SVD)对缺失值进行补全[21]。针对样本矩阵所包含的缺失值, 使用 Python 计算中的线性代数库 Scipy(https://docs.scipy.org)求解得到奇异矩阵和奇异值, 并求得近似矩阵填充原矩阵中的缺失值。

通常比较成功的 ML 模型需要大量的参数和数据进行训练, 但实际情况中数据有时并不足够多, 获取新数据又需要大量成本。深度学习常借助数据增强来扩充数据[22], 因此为了提升模型的泛化能力, 本研究为数据添加了高斯噪声。除此以外, 为了消除变量间不同量纲的影响, 在此之前对所有数据进行了标准化。

1.3 基于梯度提升机的特征选择

本研究选择了 11 种金属离子的物理特性作为特征描述符, 这些特性可以从元素周期表、化学手册[23]和公共数据库中获得; 除此以外, 还通过频率分析获取了 10 种其他特征描述符。所有用于训练 ML 模型的特征描述符如表 1 所列。如何在众多特征中选择一组具有代表性的特征用于 ML 预测是本研究的难点。考虑到可能存在过拟合风险, 本研究通过多方面分析, 并引入了梯度提升机(Gradient boosting machine, GBM)[24], 在 21 种特征描述符中筛选出最为重要的若干特征, 重点分析其对吸附能造成的影响。

1.4 基于 ML 方法的吸附能预测

为了实现吸附能的预测, 本计算将数据集划分为“训练集”和“验证集”两个不相交的集合。首先用训练集训练 ML 模型, 然后用该模型对验证集的吸附能进行预测, 最终计算预测值与真实值之间的均方误差(Mean squared error, MSE), 进行可预测性评价。本计算建立了打分机制对 6 种广为使用的 ML 模型进行了预评估。由于传统的线性方法需要假设预测目标和描述符之间存在线性关系且不适用于很多实际情况, 所以本研究聚焦于核方法和树集成方法。此外, 考虑到要提高 ML 模型的预测精度并减少过拟合的风险, 本计算选择了网格搜索方法(GridSearch)进行超参数调优[25]。通过 5 折交叉验证(5-fold cross validation)方法分配训练集和验证集[26]。计算预测值与真实值之间的均方误差 MSE, 并取其平均值作为评价最终模型的性能指标。MSE 的计算方式为：

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(h(x_i)-y_i)^2 \quad (2)$$

其中, $h(x_i)$为模型预测值, y_i 为真实值, 本研究中即为 DFT 计算得到的吸附能。对比 6 种 ML 模型得到一种吸附能最优预测模型, 将该模型延伸用于香兰素吸附金属离子的体系进行测试。为实现 ML 的预测, 本研究使用了 ML 领域已获广泛应用的工具包 Scikit-learn (http://scikit-learn.org)。

2 结果与讨论

2.1 几何构型与能量

GO 与金属离子间的吸附结构优化后, 排除了坍塌和吸附位置不合理的吸附体系, 用 DFT 优化后的

表 1 基于 DFT 计算得到的 21 个特征描述符

Table 1 21 feature descriptors calculated based on DFT

No.	Feature descriptor	No.	Feature descriptor	No.	Feature descriptor
1	Charge	8	Ionic radius	15	CV (Cal/mol-K)
2	Spin	9	Melting point	16	*S*(Cal/mol-K)
3	Atomic radius	10	Boiling point	17	Zero-point vibrational energy/(kCal·mol^{-1})
4	Atomic number	11	First ionization energy	18	Molecular mass
5	Atomic weight	12	Electronegativity	19	Mulliken charges
6	Density/(g·cm^{-3})	13	M–O (bond length)	20	APT charges
7	Atomic volume	14	*E*(Thermal)/(kCal·mol^{-1})	21	Dipole moment/D

Note: 1 Cal=4.184 J; 1 D≈0.020819434 nm

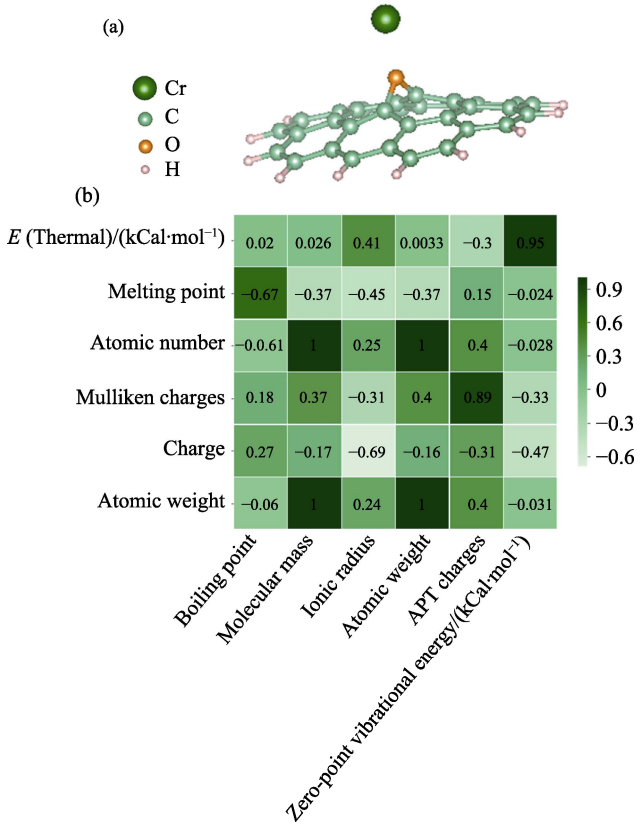

图 1　(a)相关系数>0.6 的特征间相关性热力图和(b)GO 吸附 Cr^{3+}的吸附结构示例

Fig. 1　(a) Thermal map of correlation between features with correlation coefficient>0.6, and (b) example of adsorption structure of GO adsorbing Cr^{3+}

Note: 1 Cal=4.104 J

GO-Cr^{3+}的吸附结构如图 1(a)所示。在溶液环境中, 对每个离子分别用方程(1)计算吸附能, 总结金属离子与 O 原子的键长。GO 对 M^{n+}的吸附能范围主要分布在 0.03~10 eV, 较大的吸附能表明 GO 对 M^{n+}具有更好的吸附能力, 同样从 M^{n+}–O 的键长中也可以得出相同的结论。M^{n+}与 O 原子的距离主要分布在 0.156 ~ 0.400 nm 之间, 键长越短往往吸附能越大, 表明 GO 与各种金属离子的相互作用越强。另一方面, GO 与金属离子的相互作用主要是金属离子与 O 原子的相互作用。此外, GO 对 Hg^{4+}、Tc^{4+}、Co^{3+}、Ba^{2+}等的吸附能小于 0, 本研究认为对于实际应用而言, 这不是有效吸附, 在后续的 ML 模型训练中不考虑其实际意义。

2.2　特征选择及重要性分析

本研究探索了 21 种特征间的相关性, 相关性较强的特征如图 1(b)所示, 颜色越深表示正相关性越强, 而颜色越浅表示变量间负相关性越强, 本研究将强相关特征剔除以得到特征的非冗余子集, 提升模型预测效果。通过 GBM 训练[24,27]得到特征重要性

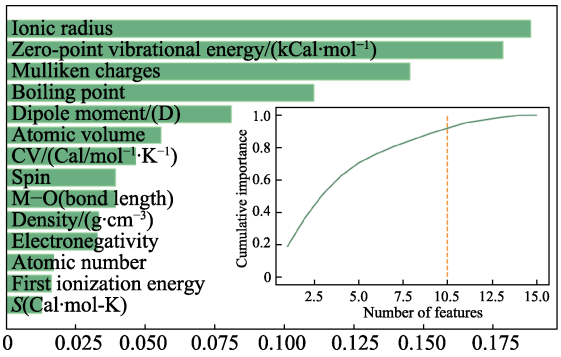

图 2　特征重要性排名

Fig. 2　Feature importance ranking

Note: 1 Cal=4.104 J; 1 D≈0.020819434 nm

排名(图 2), 发现只需前 9 个特征即可为模型提供 90%的累计重要性。在 GO 吸附金属离子的体系中, 金属的离子半径、密里根电荷以及原子体积等固有属性对吸附能的影响十分显著。此外, 另有 6 个特征对 GBM 训练结果的贡献极其微小, 从数据集中去除或者保留这些特征不对吸附能产生影响。

2.3　预测模型的评估

为实现对吸附能的数据驱动预测, 本研究将数据集分为两个不相交的集合: 大小为 75%的“训练集”和大小为 25%的“验证集”。从三个方向对 ML 方法的选择进行了预评估, 包括核方法、随机森林方法和提升方法, 共建立了 6 种回归模型, 分别是支持向量机回归(Support vector regression, SVR)[28]、岭回归(Ridge regression)[29]、随机森林(Random forest, RF)[30]、极端随机森林(Extremely randomized trees, ERT)[31]、极端梯度提升方法(Extreme gradient boosting, XGBoost)、轻梯度提升方法(Light gradient boosting machine, LightGBM)[32]。首先利用训练集建立 ML 模型, 然后利用该模型对验证集的吸附能进行预测, 计算预测值与真实值(DFT caculated)之间的均方误差(MSE), 并通过计算决定系数 R^2 为模型打分[25]。

对于大多数 ML 方法来说, 影响预测性能的关键因素在于适当地设置重要超参数的值。本研究选择了网格搜索方法[33]进行调参, 并对训练集进行 5 次交叉验证以确定最佳超参数, 从而得到最优模型。不同方法的最优参数如表 2 所列。

RF、ERT 和 XGBoost 均由 31 棵决策树组成, 其中 RF 每棵树最大深度(max_depth)为 6, 而 ERT 每棵树最大深度为 7, XGBoost 每棵树最大深度为 2。此外, XGBoost 最小叶子节点样本权重和(min_child_weight)为 13, 学习率设置为 0.32, 学习率表示梯度下降的步长, 其值过小时收敛速度较慢并易导致过拟合。而 LightGBM 则由 17 棵决策树构成, 步长为

表 2 六种机器学习方法的最优超参数

Table 2 Optimal hyperparameters of six machine learning methods

Category	Method	Optimal hyperparameters
Kernel	Support vector regression (SVR)	*C* = 2, kernel="rbf"
	Ridge regression	Alpha = 30
Random forest	Random forest (RF)	*n*_estimators = 31, max_depth = 6, max_features = 2
	Extremely randomized trees (ERT)	*n*_estimators = 31, max_depth = 7, random_state = 1
Boosting	Extreme gradient boosting (XGBoost)	*n*_estimators = 31, max_depth = 2, min_child_weight = 13, learning_rate = .32
	Light gradient boosting machine (LightGBM)	*n*_estimators =17, objective = 'regression', num_leaves = 31, learning_ rate = 0.32

0.32。其余超参数均设置为 Scikit-learn 的缺省值。不同模型在训练集上的预测效果(图 3)显示, SVR 和 Ridge 方法明显不适于预测吸附能, 其 R^2 仅为 0.285 和 0.114, 后续将不再讨论。而树集成方法则表现出较为优异的效果, 特别是 RF 和 ERT, 二者的 R^2 分别达到了 0.816 和 0.899, 表明模型可以很好地解释原数据。从预评估中可以看到, 随机森林方法似比两种提升方法能更有效地预测吸附能, 但评估一个模型好坏最重要的标准在于其对未知数据的预测能力, 即泛化性能。为此本研究深入讨论了模型对验证集的均方误差(MSE)。

本研究中测试的 4 种集成方法可在 0.075 ~ 0.344 的 MSE 范围内较好地预测 GO 去除金属离子的吸附能。总体显示随机森林方法(RF、ERT)和 XGBoost 的预测性能都较完美, 远优于单一模型方法(SVR、Ridge)和 LightGBM。R^2 评分越高的模型往往具有越小的 MSE, 四种集成方法的 MSE 如图 4(a)所示, R^2 更高的 ERT 和 RF 获得了更小的训练误差(分别为 0.075 和 0.137), 以及更小的验证误差(分别为 0.197 和 0.252)。进一步分析回归拟合效果, 绘制了真实值和预测值的相关图, ERT 模型预测值与 DFT 计算值具有较强的线性关系。总体来说 ERT 对未知数据的预测更精确, 控制预测方差的效果更强, 并且 ERT 的训练时间要短于 RF。因此, 可以认为 ERT 模型是最理想的吸附能预测模型。

2.4 预测模型的测试

为了测试 ERT 模型对不同体系吸附能的预测性能, 本研究将训练好的模型用于香兰素单体[34]对金

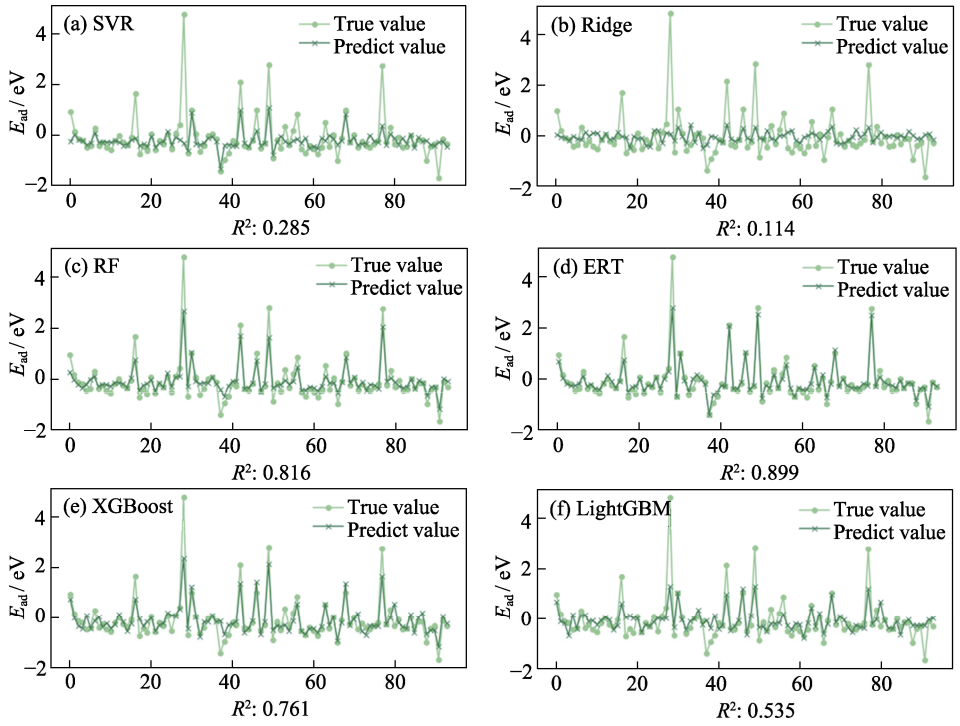

图 3 6 种机器学习方法的模型拟合效果及评分

Fig. 3 Fitting effect diagram and score of six machine learning methods.

(a) Support vector regression (SVR); (b) Ridge regression (Ridge); (c) Random forest (RF); (d) Extremely randomized trees (ERT); (e) Extreme gradient boosting (XGBoost); (f) Light gradient boosting machine (LightGBM)

属离子的吸附预测，涉及到 54 种吸附结构(图 5(a))。基于上述较为重要的 9 个特征进行预测，计算得到 MSE 为 0.596，预测值和真实值的拟合效果(图 5(b))，进一步支持了 ERT 方法是较为稳定而准确的吸附能预测方法的结论。但从真实值和预测值的皮尔逊相关图(图 5(c))可以看出，有很多数据的预测出现了较大的偏差，其原因可能来自于不同的吸附位点。与 GO 的含氧官能团不同的是，香兰素与金属离子的吸附主要归因于 N 原子与金属离子 M^{n+}之间的相互作用，这可能是模型验证效果不理想的主要原因。此外也存在一些其他可能的因素，如数据集较小会增加过拟合风险，降低了模型的泛化能力。

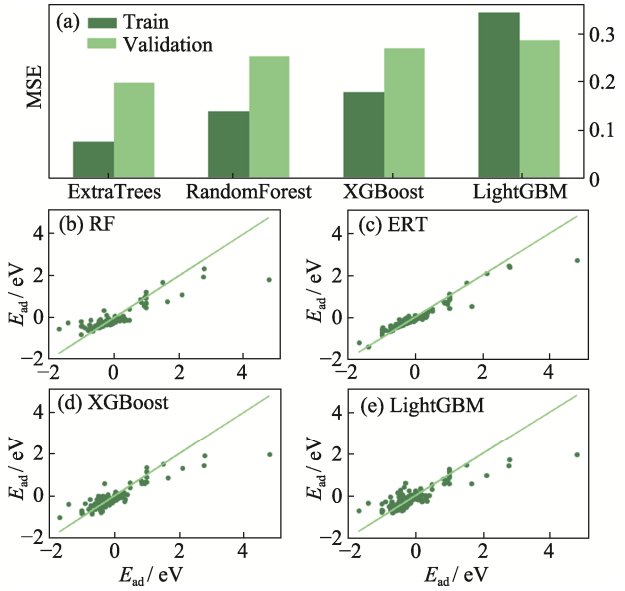

图 4　(a)四种集成方法的 MSE; (b~e)四种集成方法真实值和预测值的相关图

Fig. 4　(a) Mean square error (MSE) of the four ensemble methods, and (b-e) correlation graphs of the true and predicted values of the four ensemble methods

(b) Random forest (RF); (c) Extremely randomized trees (ERT); (d) Extreme gradient boosting (XGBoost); (e) Light gradient boosting machine (LightGBM)

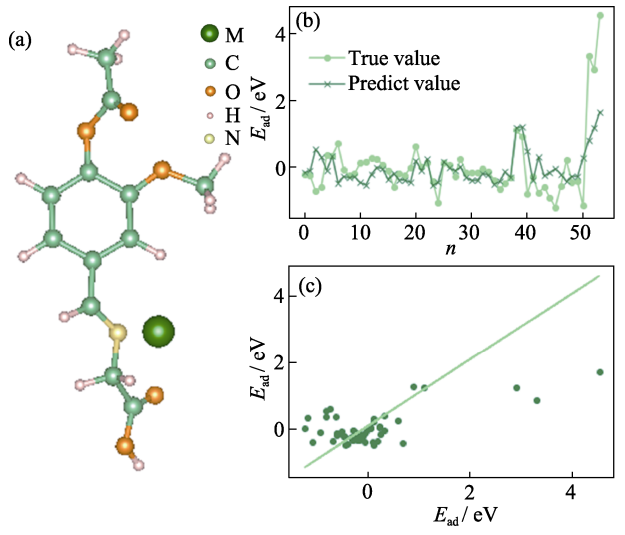

图 5　(a)香兰素单体吸附金属离子的结构示例图; (b)ERT 用于 VMA-M^{n+}吸附能的拟合效果图; (c)ERT 用于 VMA-M^{n+}吸附能的相关图

Fig. 5　(a) Example of the structure of vanillin monomer adsorbing metal ions; (b) Fitting effect graph of Extremely Randomized Trees (ERT) for VMA-M^{n+} adsorption energy; (c) Correlation diagram of ERT for VMA-M^{n+} adsorption energy

3　结论

本研究通过特征选择确定了对吸附能最重要的 9 个特征，对比 6 种不同的回归方法(SVR、Ridge、RF、ERT、XGBoost 和 LightGBM)，确定了 ML 方法对吸附能的预测有较高的准确性，其中极端随机森林方法表现出最优异的预测性能，均方误差仅 0.075，并可节省大量机时。以香兰素吸附金属离子的体系测试所建立的预测模型，发现所探索的机器学习模型可应用于其他金属离子吸附去除材料的预测，但模型的参数选择问题及过拟合问题仍有待进一步研究，因而模型的泛化能力尚待进一步提升。

参考文献:

[1] PENG W J, LI H Q, LIU Y Y, *et al.* A review on heavy metalions adsorption from water by graphene oxide and its composites. *Journal of Molecular Liquids*, 2017, **230:** 496–504.

[2] AHMAD S Z N, SALLEH W N W, ISMAIL A F, *et al.* Adsorptive removal of heavy metal ions using graphene-based nanomaterials: toxicity, roles of functional groups and mechanisms. *Chemosphere*, 2020, **248:** 126008.

[3] LIU Y, ZHAO C F, ZHANG A R, *et al*. Theoretical study on the removal of uranyl by nitrogen, phosphorus and sulfur doped graphene materials. *Scientia Sinica Chimica*, 2019, **49(1):** 91–102.

[4] PENG X J, WANG Y F. Efficient stochastic simulation algorithm for chemically reacting systems based on support vector regression. *Chinese Journal of Chemical Physics*, 2009, **22(5):** 502–510.

[5] CAI C, LI L, DENG X, *et al*. Machine learning and high-throughput computational screening of metal-organic framework for separation of methane/ethane/propane. *Acta Chimica Sinica*, 2020, **78(5):** 427.

[6] ORUPATTUR N V, MUSHRIF S H, PRASAD V. Catalytic materials and chemistry development using a synergistic combination of machine learning and *ab initio* methods. *Computational Materials Science*, 2020, **174:** 109497–16.

[7] LI X, XI L L, YANG J. First principles high-throughput research on thermoelectric materials: a review. *Journal of Inorganic Materials*, 2019, **34(3):** 236–246.

[8] MENG Y, WANG X, YANG J, *et al.* Research on machine learning based model for predicting the impact status of laminated glass. *Journal of Inorganic Materials*, 2021, **36(1):** 61–68.

[9] FENG C, SHARMAN E, YE S, *et al.* A neural network protocol for predicting molecular bond energy. *Sci. China Chem.*, 2019, **62(12):** 1698–1703.

[10] LU S, ZHOU Q, OUYANG Y, *et al.* Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites *via* machine learning. *Nature Communications*, 2018, **9(1):** 3405.

[11] TEHRANI A M, OLIYNYK A O, PARRY M, *et al.* Machine learning directed search for ultraincompressible, superhard materials. *Journal of the American Chemical Society*, 2018, **140(31):** 9844–9853.

[12] KANG Y, LI L, LI B. Recent progress on discovery and properties prediction of energy materials: simple machine learning meets complex quantum chemistry. *Journal of Energy Chemistry*, 2020, **54:** 72–88.

[13] BROCKHERDE F, VOGT L, LI L, *et al.* By-passing the Kohn-Sham equations with machine learning. *Nature Communications*, 2017, **8(1):** 872.

[14] XIAO Y, MIARA L J, WANG Y, *et al.* Computational screening of cathode coatings for solid-state batteries. *Joule*, 2019, **3(5):** 1252–1275.

[15] PANAPITIYA G, AVENDANO-FRANCO G, REN P, *et al.* Machine learning prediction of CO adsorption in thiolated, Ag alloyed Au nanoclusters. *Journal of the American Chemical Society*, 2018, **140(50):** 17508–17514.

[16] PARDAKHTI M, MOHARRERI E, WANIK D, *et al.* Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs). *ACS Combinatorial Science*, 2017, **19(10):** 640–645.

[17] SI Y, SAMULSKIE T. Synthesis of water soluble graphene. *Nano Letters*, 2008, **8(6):** 1679–1682.

[18] CHEN D, FENG H, LI J. Graphene oxide: preparation, functionalization, and electrochemical applications. *Chemical Reviews*, 2012, **112(11):** 6027–6053.

[19] SHENG Z H, SHAO L, CHEN J J, *et al.* Catalyst free synthesis of nitrogen-doped graphene *via* thermal annealing graphite oxide with melamine and its excellent electrocatalysis. *ACS Nano*, 2011, **5(6):** 4350–4358.

[20] MAURIZIO C, VICENZO B, MICHAEL A R. A direct procedure for the evaluation of solvent effects in MC-SCF calculations. *Journal of Chemical Physics*, 1999, **111(12):** 5295–5302.

[21] BRAND M. Incremental Singular Value Decomposition of Uncertain Data with Missing Values. Computer Vision-ECCV 2002, Berlin, Heidelberg, 2002**:** 707–720.

[22] NEELAKANTAN A, VILNIS L, LEQ V, *et al.* Adding gradiant noise improves learning for very deep networks. *arXiv: Machine Learning*, 2015, **1511:** 06807.

[23] 迪安 J A, 魏俊发. 兰氏化学手册, 2 版. 北京: 科学出版社, 2003**:** 1–1579.

[24] FRIEDMAN J H, JAO S. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 2001, **29(5):** 1189– 1232.

[25] LIASHCHYNSKYI P, LIASHCHYNSKYI P. Grid search, random search, genetic algorithm: a big comparison for NAS. *arXiv: Learning*, 2019, **1912:** 06059.

[26] MARTENS H A, DARDENNE P J, CSYSTEMS I L. Validation and verification of regression in small data sets. *ChemomERTics & Intelligent Laboratory Systems*, 1998, **44(1/2):** 99–121.

[27] FRIEDMAN J H. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 2002, **38(4):** 367–378.

[28] SMOLA A J, SCHOLKOPF B. A tutorial on support vector regression. *Statistics and Computing*, 2004, **14(3):** 199–222.

[29] RUPP M. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 2015, **115(16):** 1058–1073.

[30] SVETNIK V. Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Modeling*, 2003, **43(6):** 1947–1958.

[31] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees. *Machine Learning*, 2006, **63(1):** 3–42.

[32] DRUCKER H. Improving regressors using boosting techniques. *Morgan Kaufmann Publishers Inc*, 1997**:** 107–115.

[33] FRANLLIN J. The elements of statistical learning: data mining, inference, and prediction. *The Mathematical Intelligencer*, 2005, **27(2):** 83–85.

[34] SANTOS R I H, REIS D T, PEREIRA D H. A DFT based analysis of adsorption of Cd^{2+}, Cr^{3+}, Cu^{2+}, Hg^{2+}, Pb^{2+}, and Zn^{2+}, on vanillin monomer: a study of the removal of metalions from effluents. *Journal of Molecular Modeling*, 2019, **25(9):** 267.