高储氢密度金属氢化物蓄热性能预测

杨宜坤[1]，吴 震[1]，刘洪皓[1]，张早校[1,2]
（1.西安交通大学 化学工程与技术学院，陕西 西安 710049；2.绿色氢电全国重点实验室，陕西 西安 710049）

摘 要：金属氢化物材料具有储氢/热密度高，工作温度区间广，无污染无腐蚀性的优点，被认为是理想的储氢/热材料。金属氢化物储氢/热材料可以通过掺杂不同元素形成多元合金，以开发具有不同目标性能的材料。这种方法依赖实验合成，十分耗费时间和经济成本。数据驱动的机器学习性能预测模型可以解决这一问题，通过测试对比最小二乘回归、最小绝对收缩和选择操作符回归、岭回归、弹性网络回归、支持向量回归和随机森林回归多种回归算法，成功建立了金属氢化物微观材料性质与宏观形成能之间的关系。测试结果显示随机森林回归具有最好的预测性能，在训练集和测试集上相对误差均较小，仅为 3.078 和 8.201 1，且决定系数较高，具有良好的回归能力和泛化能力。SHAP 分析中表明组成金属氢化物的基态原子体积的均值和最值具有高达 5.56 和 1.26 的 SHAP 值，这 2 个因素很大程度上决定了金属氢化物材料的形成能大小。对 Mg 基，Ca 基，AB、AB_2 及 AB_5 型金属氢化物材料预测结果分析显示预测相对误差均在 9% 以下，证明了模型准确性及普适性，可用于未知数据集的形成能预测。

关键词：太阳能热利用；金属氢化物；储氢储热；机器学习；性能对比
中图分类号：TK91 **文献标志码：**A **文章编号：**1006−6772(2024)12−0134−13

Comparative study of machine learning regression algorithms for predicting thermal energy storage performance of metal hydrides with high hydrogen density

YANG Yikun[1], WU Zhen[1], LIU Honghao[1], ZHANG Zaoxiao[1,2]
(1. *School of Chemical Engineering and Technology*, *Xi'an Jiaotong University*, *Xi'an* 710049, *China*; 2. *State Key Laboratory of Green Hydrogen and Electricity*, *Xi'an* 710049, *China*)

Abstract: Metal hydride thermal/hydrogen energy storage material is considered ideal candidate due to high energy density, wide working temperature range and lack of corrosive pollutants. Multi-component metal hydride alloys can be formed by doping with different elements to obtain various target properties. However, conventional material development method relies on experimental synthesis, having the disadvantages of time-consuming and costly. Data-driven machine learning prediction model is capable of addressing this problem. By comparing varieties of regression algorithms such as least squares regression, least absolute shrinkage and selection operator regression, ridge regression, elastic net regression, supporting vector regression, and random forest regression, the relationship between the microscopic properties of metal hydride materials and their macroscopic formation energy are established. Results show that random forest regression have the best prediction performance, with lowest relative errors on both the training and test sets of 3.078 and 8.201 1, high R-squared values, and great generalization and regression abilities. SHAP analysis reveals extreme and mean value of ground state atom of metal hydride exhibit the greatest SHAP value of 5.56 and 1.26, suggesting their significant influence on the formation energy. Analysis for the prediction value of Mg-base, Ca-base, AB type, AB_2 type, and AB_5 type metal hydrides shows the highest relative error below 9%, further proving the accuracy and universality of the model for all types of metal hydride. This model can be used to predict the formation enthalpy of unknown datasets.

收稿日期：2024−10−30；**策划编辑：**白娅娜；**责任编辑：**李雅楠 **DOI:** 10.13226/j.issn.1006-6772.HH24103101
基金项目：国家自然科学基金面上资助项目(52376208; 52176203)
作者简介：杨宜坤(1997—)，男，河南三门峡人，博士。E-mail: yikun.yang@stu.xjtu.edu.cn
通讯作者：张早校(1963—)，男，山西晋城人，教授，博士生导师。E-mail: zhangzx@mail.xjtu.edu.cn
引用格式：杨宜坤，吴震，刘洪皓，等.高储氢密度金属氢化物蓄热性能预测 [J].洁净煤技术，2024，30(12)：134−146.
YANG Yikun, WU Zhen, LIU Honghao, et al. Comparative study of machine learning regression algorithms for predicting thermal energy storage performance of metal hydrides with high hydrogen density[J].Clean Coal Technology, 2024, 30(12): 134−146.

Key words: thermal application of solar energy; metal hydride; hydrogen and heat storage; machine learning; performance comparison

0 引 言

自 21 世纪以来，全球气候变暖和化石能源危机日益加剧，开发利用清洁的可再生能源技术迫在眉睫[1]。我国幅员辽阔，多种可再生能源都具备极大发展潜力，其中以太阳能资源最为丰富，年辐射量高达 5 000 亿 GJ，有望助力我国达到“双碳”目标[2]。然而自然界中的可再生能源都具有不稳定性，例如太阳能具有时变性高的特点，其辐射量会随着四季和昼夜变化产生较大波动，因此需要结合储能技术存储在二次能源中以实现其稳定高效利用[3]。氢能是优质的二次能源，具有稳定性高、储存周期长和无污染的优点，符合清洁能源发展要求，因此在近些年来被广泛研究。此外，传统化石能源的清洁利用也需要通过氢能实现，比如洁净煤技术旨在通过高效、低污染和低碳排放的方式利用我国丰富的煤炭资源，通过煤气化技术，可以从煤炭中提取氢气，有助于减少煤炭燃烧带来的环境污染，这一过程同样需要结合氢储能技术实现。不同于传统的物理储能，氢储能通过含氢化合物可逆化学反应过程中的热效应实现可再生能源的存储与释放，具有储能稳定的优势，既可用于能量跨区域运输，也可用于集中式热能存储。氢储能/热技术发展的关键基础便是储氢/热材料的选取。在储热领域中，常见的储热材料有金属氢化物、金属氧化物、氢氧化物、碳酸盐和硫酸盐等，现有的储热材料性能见表 1。通过对比可以看出金属氢化物储热材料具有工作温度区间广，储热密度大的特点，且储热过程中不存在硫酸盐和碳酸盐等熔融盐类储热材料常见的腐蚀问题，因此具有更优异的性能而广受关注。然而目前可用于储热的金属氢化物材料种类较少，只有 MgH_2 和 CaH_2 等几种储氢材料可供使用，无法满足低中高温各个工况的应用需求，且金属氢化物储氢/热材料的性能对整个储热系统影响极大[4]，因此储热用高密度金属氢化物储氢材料的设计及选型十分关键。

金属氢化物可以通过元素掺杂进行改性，以获得具有不同吸放氢焓变的金属氢化物材料。在氢储能领域中，金属氢化物储热追求单位质量或物质的量的氢气有更高的吸放氢焓变，而储氢由于低能耗要求则希望单位质量或物质的量的氢气具有更低的吸放氢焓变。因此无论是面对储氢或者储热应用，金属氢化物的热力学性能改性对于其发展应用都极为重要。为了开发高性能的金属氢化物材料，国内外研究人员在此领域进行了大量的试验研究。DANGWAL 等[5] 研究了 TiVNbFe 合金化金属氢化物体系下的材料性能变化，发现 $(TiVNb)_{75}Cr_{16.7}Fe_{8.3}$ 合金的脱氢焓变比 $(TiVNb)_{75}Cr_{25}$ 低 10 ~ 20 kJ/mol H_2。AGAFONOV 等[6] 通过理论计算和实验结合的方法开发出了合金化金属氢化物 $Ti_{0.5}Zr_{1.5}CrMnFeNi$ 和 TiZrCrMnFeNi，实验测得其脱氢焓变分别为 ΔH=−36.5 kJ/mol 和 ΔH=−28.8 kJ/mol。YIN 等[7] 研究了在 $Mg_{85}Zn_5Ni_{10}$ 合金中分别掺杂质量浓度为 4% 和 8% 的 Cr_2O_3，其吸放氢反应焓变由原始的 −82.446 kJ/mol 分别变化至−80.067 和 74.726 kJ/mol。钟海长等[8] 将钇氧化物掺杂进镁锌金属氢化物合金中，发现其吸放氢活化能显著降低，其焓变对比纯 MgH_2 也减少了 3.8 kJ/mol。除了试验测试外，研究人员也通过第一性原理理论计算分析了不同掺杂元素及不同掺杂比例对金属氢化物吸放氢焓变的影响。RKHIS 等[9] 研究了 $ZrNiH_3$ 系列的金属氢化物焓变，分别计算了不同 Zr 和 Ni 元素比例下材料的吸放氢焓变变化，其结果显示 $ZrNi_{0.92}H_{3.04}$ 的吸放氢焓变由原始的−72.14 kJ/mol 变化到 −37.89 kJ/mol，绝对值减少幅度接近 50%。类似的，GU 等[10] 也通过第一性原理计算在 $ZrCoH_3$ 体系中换用不同比例的 Co、Nb 和 Cr 元素，计算结果表明降低 Zr 元素的比例会使材料的吸放氢焓变降低，当材料组分为 $Zr_{0.75}Nb_{0.25}Co_{0.875}Cr_{0.125}H_3$ 时，吸放氢焓变减少了 10.57 kJ/mol。

表 1 各类储热材料性能

Table 1 Properties of varieties of heat storage materials

储热材料类型	工作温度/℃	储能密度
金属氢化物	50 ~ 1 000	1 600 ~ 5 000
金属氧化物	640 ~ 1 200	50 ~ 1 000
氢氧化物	250 ~ 800	600 ~ 1 500
碳酸盐	270 ~ 870	250 ~ 2 000
硫酸盐	900 ~ 1 300	1 600 ~ 3 000

以上研究结果不难看出，通过掺杂多种元素可以改善金属氢化物材料的宏观热力学性能。然而，由于金属氢化物储能/热氢材料改性机理复杂，因此这一过程严重依赖于传统试错法式的试验合成–材料表征–性能测试流程，实验周期长，十分耗费时间和经济成本，且规模有限，无法在同一批次中产出和处理大量材料数据。此外，通过第一性原理计算材料性能也面临计算时间长，计算成本高，无法

大批量处理数据的问题[11]。显然，这 2 种材料开发方法都只能针对单一或几种金属氢化物材料的改性处理，具有一定的局限性。且 2 种方法的改性机理不明确，无法满足金属氢化物储氢/热材料设计的需求，因此构建新的金属氢化物储氢/热材料设计选型方法势在必行。随着数据学科的兴起，机器学习回归作为一种数据驱动的新方法，能够通过数据挖掘和拟合，捕捉数据之间潜在的趋势与联系，并使用特定算法进行自优化和改进对数据趋势进行拟合同时不断降低模型的拟合误差，从而高效智能地建立起大数据输入特征值与输出目标值之间的关系[12]。因其强大的数据处理能力，机器学习模型已经被广泛应用于多种预测任务，包括固体氧化物燃料电池电极性能预测[13]，有机金属框架相似性量化等[14]。在金属氢化物领域，GHEYTANZADEH 等[15] 使用了高斯过程回归研究了 AB_2 型金属氢化物材料的形成焓与多种元素之间的关系，结果显示钒和铬元素的添加对形成焓影响最大。这些研究通常是使用单一的机器学习算法模型来完成回归任务，未考虑不同算法对金属氢化物储热材料设计任务的影响，且使用的描述符较为单一，无法描述材料的微观组分信息。

因此，笔者将首先为多种金属氢化物材料进行特征描述符计算以最大程度抓取其微观组分信息，随后对比最小二乘法回归、最小绝对收缩和选择操作符回归、岭回归、弹性网络回归、支持向量回归、随机森林回归的回归结果，比较多个算法的金属氢化物形成焓预测性能及其适用性，针对多元金属氢化物储氢/热材料设计这一科学问题筛选出最有效的机器学习模型。分析对回归结果影响最大的特征描述符，即从金属氢化物材料的多个微观特征中挖掘出对吸放氢焓变影响最大的微观组分因素，为金属氢化物储氢/热材料设计提供理论基础和技术指导，整体的工作流程如图 1 所示。首先通过金属氢化物数据收集，为模型搭建提供基础；随后对所收集的金属氢化物材料数据进行特征计算生成，最大限度地捕捉其微观信息；然后使用特征数据进行多种机器学习算法模型性能测试，筛选最优算法；最后根据 SHAP 结果分析找出影响金属氢化物吸放氢焓变的关键微观因素，为材料设计开发过程中所关注的材料能耗问题提供解决思路。

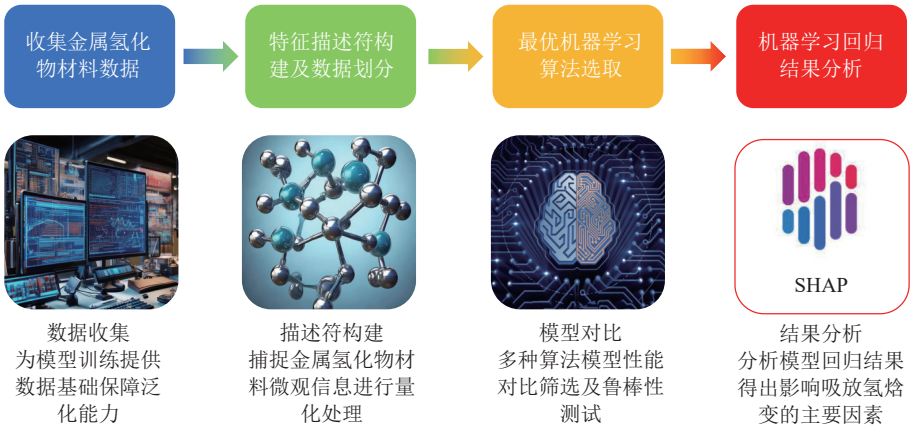


图 1　机器学习预测金属氢化物吸放氢焓变工作流程

Fig. 1　Metal hydride hydrogenation/dehydrogenation enthalpy prediction workflow

1　材料数据与预测方法

1.1　材料数据集构建

数据集是机器学习模型的关键基础，数据集极大程度上决定了机器学习模型性能的优劣。为了开发出能够预测多种类型金属氢化物材料热力学性能的模型，笔者从材料数据库 OQMD[16-17] 和 Materials Project[18] 中收集了来自各个形成能区间共计 398 个金属氢化物材料数据组成数据集。对于该体量数据，为最大程度避免模型的过拟合同时保证模型精度，需要保证训练集具有相当的比例，同时测试集也不能占比过低，因此笔者采用了训练集：测试集比例为 70∶30 的划分方法。数据集的形成能频数分布如图 2 所示，整个数据集的频数分布基本符合正态分布，形成能从 0 到 100 kJ/mol 皆有覆盖，且所使用的训练集和测试集均包含 AB、AB_2、AB_5 和镁基等多种类型的金属氢化物材料数据，因此数据量和数据分布形态可以满足机器学习训练的基本要求。

为了对所有收集的金属氢化物材料微观信息进行准确的量化描述，本研究使用 WARD 等[19] 开发的架构为每个材料进行特征描述符计算。该架构主要针对无机材料的性能预测回归任务，提供一种普适性较好的材料特征描述符计算工具。如图 3 所

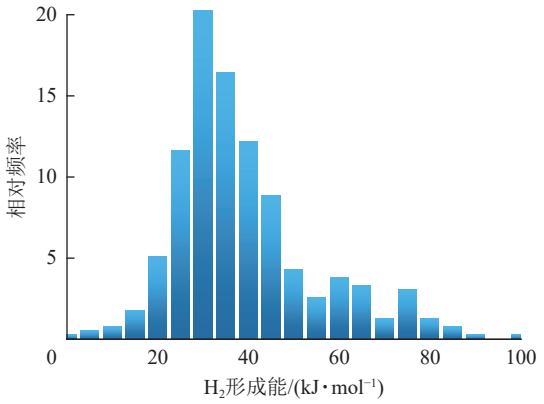


图 2　金属氢化物样本数据集形成能相对频率分布直方图

Fig. 2　Distribution histogram of metal hydride formation energy

示，其所得描述符主要分为 4 类：化学计量属性、元素性质数据、电子结构属性和离子化合物属性。其中化学计量属性主要包括原子的数量和晶体的范数。元素性质数据则包含了化合物内元素的原子序数、元素周期表行数和原子半径等微观信息。电子结构属性则包括了化合物内所有元素的 s、p、d、f 价层电子的平均比例。离子化合物属性侧重于描述化合物内元素电负性相关的属性。特征中包含了极值、众数和均值等类型的数据，通过此类数据能够从不同角度反应材料围观信息特征，因此加入这些特征后能够使模型从多角度捕捉影响金属氢化物吸放氢焓变的微观因素。

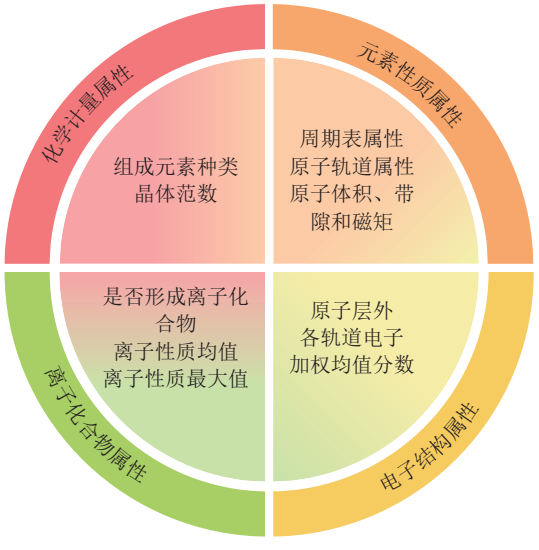


图 3　4 类材料特征描述符

Fig. 3　Four types of feature descriptors

通过该架构计算，结合材料数据库 OQMD 和 Materials Project，对 398 种氢化物材料依次计算生成 145 种描述符，其计算方式如式（1）—式（4）所示。化学计量属性计算公式为

$$||x|| = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \tag{1}$$

式中：x_i 为某一元素的原子分数；p 为范数级数。

元素性质属性：

$$\bar{y} = \sum x_i y_i$$
$$\hat{y} = \sum x_i |y_i - \bar{y}| \tag{2}$$

式中：y、$\bar{y}$和$\hat{y}$为元素性质的原始值、均值和平均偏差。

价电子轨道填充分数：

$$F = \frac{\sum E_i \times O_i}{\sum E_i \times O_f} \tag{3}$$

其中，F 为价电子轨道填充分数；E_i 为第 i 个元素计量比重；O_i 为钙元素的原子轨道实际填充数；O_f 为该元素原子轨道完全填充数。

离子属性：

$$I(X_A, X_B) = 1 - \exp[-0.25(X_A - X_B)^2] \tag{4}$$

其中，X_A 和 X_B 表示元素的电负性。使用这一特征计算方法所得的特征见表 2，其中所有的元素性质类的特征均计算了均值、极差、离差、最大值、最小值和众数作为特征。为了便于机器学习模型处理，本研究采用金属氢化物的形成能作为标签值，吸放氢焓变可通过公式（5）进行转化：

$$\Delta H = \frac{1}{(N/2)}(E_{\text{alloy}} + E_{\text{H}_2} - E_{\text{hydride}}) \tag{5}$$

其中，E_{alloy}、E_{H_2} 和 E_{hydride} 分别为未吸氢金属、氢气和吸氢后的金属氢化物的形成能。通常 E_{H_2} 取−3.345 eV/atom。E_{alloy} 与 E_{hydride} 的计算方法相同，如式（6）所示：

$$E_{\text{化合物}} = E_{\text{kin}} + E_{\text{enuc}} + E_{\text{ee}} + E_{\text{xc}} + E_{\text{ion}} + E_{\text{mix}} \tag{6}$$

其中，E_{kin} 是电子的动能，E_{enuc} 是电子与原子核之间的吸引力能量，E_{ee} 是电子之间的库伦排斥能量，E_{xc} 是交换相关能量，E_{ion} 是离子间的排斥能量，E_{mix} 是合金中由于不同元素的混合而产生的额外能量，也称为混合能。

1.2　机器学习模型构建

机器学习作为数据驱动的方法，通过分析描述符与标签值之间潜在的数据倾向和潜在联系，在输入新的描述符数据时给出相应的目标值预测。笔者测试对比了最小二乘法回归（least squares regression）、LASSO 回归（Least Absolute Shrinkage and Selection Operator）、岭回归（ridge regression）、弹性网络回归（elastic net regression）、支持向量回归（supporting vector regression）、随机森林回归（random forest regression）多种回归算法。

表 2　特征描述符类型及名称

Table 2　Feature descriptor type and name

描述符类型	描述符名称	描述符类型	描述符名称
化学计量属性	元素种类数	元素性质属性	f轨道价电子数
	L2范数		s轨道未填充数
	L3范数		p轨道未填充数
	L5范数		d轨道未填充数
	L7范数		f轨道未填充数
	L10范数		基态原子体积
元素性质属性	原子序数		基态原子带隙
	门捷列夫数		基态原子磁矩
	原子质量		空间群数
	元素熔化温度		价电子总数
	周期表行数	价电子轨道填充分数	s轨道价电子数加权均值分数
	周期表列数		p轨道价电子数加权均值分数
	共价半径		d轨道价电子数加权均值分数
	电负性		f轨道价电子数加权均值分数
	s轨道价电子数	离子属性	是否离子化合物
	p轨道价电子数		离子性质最大值
	d轨道价电子数		离子属性均值

1.2.1　线性回归模型

在线性回归模型中，最小二乘回归是最简单最常见的回归方法，具有直观、效率高和可解释性强的优点，但其对异常值较为敏感，易受其影响。最小二乘回归通过最小化标签值和预测值之间误差的平方和来估计回归系数的方法。它的目标是找到一组回归系数，使得预测值和真实值之间的差距最小。对于给定的输入数据 X，即计算所得的特征值矩阵，其回归方程如式（7）所示：

$$\hat{y}=X\beta \tag{7}$$

其中，β 是算法需要估计和拟合的回归系数，$\hat{y}$是模型的预测值。最小二乘回归的目标是最小化以下目标函数（损失函数）：

$$LOSS=\min_{\beta}\| y-X\beta \|_2^2 \tag{8}$$

LASSO 回归在最小二乘回归的基础上添加了 L_1 范数项，来实现特征选择和回归系数的稀疏化。它倾向于将一些不重要的回归系数缩减为零，从而选择出对预测最有用的特征。LASSO 回归的目标函数如式（9）所示：

$$LOSS=\min_{\beta}(\| y-X\beta \|_2^2+\lambda_1 \| \beta\|_1) \tag{9}$$

其中，$\|\beta\|_1$ 为回归系数的 $L1$ 范数，即绝对值加和；λ_1 是正则化参数，控制模型的复杂度。类似的，岭回归在最小二乘回归的基础上添加了不同于 LASSO 回归的范数项，其目标函数为：

$$LOSS=\min_{\beta}(\| y-X\beta \|_2^2+\lambda_2 \| \beta \|_2^2) \tag{10}$$

其中，$\|\beta\|_2^2$ 是回归系数的 L_2 范数，即平方和；λ_2 是 L_2 正则化参数。弹性网络回归则是兼具了 LASSO 回归和岭回归二者的范数项，结合了 LASSO 和岭回归的优点，同时对回归系数施加 L_1 和 L_2 正则化。这种方法既能实现特征选择，又能处理多重共线性问题防止过拟合，但是超参数调优更为复杂。其目标函数如式（11）所示：

$$LOSS=\min_{\beta}(\| y-X\beta \|_2^2+\lambda_1 \| \beta\|_1+\lambda_2 \| \beta \|_2^2) \tag{11}$$

通过调整 λ_1 和 λ_2，可以在 LASSO 和岭回归之间平衡，达到两者的效果。

1.2.2　支持向量回归模型

支持向量回归（Supporting Vector Regression, SVR）通过在高维特征空间进行线性回归，在处理高维数据时表现良好，泛化能力好，但计算复杂度很高。其核心思想是通过寻找一个超平面来预测目标变量，同时控制预测误差在某个阈值范围内，并尽量减少模型的复杂度。对于给定的 X 特征值矩阵和 Y 标签值矩阵，SVR 的目标是找到函数 $f(X)=\omega^T\varphi(X)+b$，使得大多数数据点的预测误差不超过 1 个预设的阈值 ε。其目标函数为：

$$LOSS=\min_{\bar{\omega},b,\xi_i,\xi_i^*}\frac{1}{2}\| \bar{\omega}\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{12}$$

其中，$\|\omega\|^2$ 是正则化项，用于控制模型的复杂度，C 为惩罚参数，可以控制误差项的权重。ξ_i 和 ξ_i^*允许一些数据点的预测误差超过 ε，但这些超出部分会受到惩罚。在支持向量回归中，数据可以通过非线性映射函数 $\varphi(x)$ 映射到更高维空间。在该空间中，支持向量回归的超平面可以更好地拟合复杂的非线性关系。常用的核函数包括线性核、径向基函数核和多项式核等。

1.2.3　随机森林回归模型

随机森林通过集成多个决策树回归的结果，以投票的形式决定最终的输出结果，可以处理大规模高维非线性数据的回归问题，面对多特征回归问题，算法会随机选择特征和样本构建多个树，避免某一特征主导决策树的问题，能有效避免过拟合，其性能已在其他领域的应用中得到证实[20]。随机森林算法首先进行样本选择，如式（13）所示，从数

据集中随机放回地抽取 m 个样本，生成 B 个样本子集 D_b（其中 b=1,2,⋯,B）。接下来对 D_b，构建回归决策树，在每个节点随机选择 k 个特征（k<p，p 是总特征数），找到最佳分割点，最小化分割后的均方误差。

$$E_{\mathrm{MS,split}}=\frac{|D_L|}{|D|}E_{\mathrm{MS,split}}(D_L)+\frac{|D_R|}{|D|}E_{\mathrm{MS,split}}(D_R) \quad (13)$$

最后综合所有 B 棵决策树的回归结果，对其求均值输出模型的预测结果。为了提高算法的优化效率，本文在研究测试过程中引入了 Optuna 模块实现各个算法的调参优化，能够在应对模型多个参数调节这种复杂任务时以极高的效率提供最优超参数[21]。

1.3　模型评价指标

为了直观地评价和对比各个回归算法对金属氢化物形成能的预测性能，笔者引入了平均绝对误差（mean absolute error, MAE）、均方误差（mean squared error, MSE）、均方根误差（root mean squared error, RMSE）和决定系数 R^2（coefficient of determination）以量化的方式衡量各个机器学习算法在回归任务中的表现，其计算方式如下：

$$E_{\mathrm{MA}}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (14)$$

$$E_{\mathrm{RMS}}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \quad (15)$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (16)$$

其中，y_i 为数据的真实标签值，$\hat{y}_i$ 为模型根据输入值预测的标签值，$\bar{y}$为数据标签值的均值，E_{MA} 为预测值与实际值之间的平均绝对差值，反映了模型预测的标签值与真实标签值之间的平均绝对误差，不易受异常值影响。MSE 和 RMSE 都描述了模型预测的标签值与真实标签值之间的误差平方，对异常值更加敏感。决定系数 R^2 是一个无量纲指标，衡量模型解释数据中变异的比例，反映模型拟合的好坏程度。

1.4　模型评价指标

为了解释和分析机器学习模型的结果，同时揭示所构建的 145 种特征描述符对金属氢化物储热材料形成能的影响程度，笔者引入特征重要性分析，这一缓解通过 SHAP（Shapley Additive exPlanations）值计算分析实现。SHAP 值为每个特征分配一个“贡献值”，反映该特征对预测结果的影响：对于一个预测模型 $f(x)$，输入特征 x 的 SHAP 值定义为

$$\varphi_i=\sum_{S\subseteq N,i\notin S}\frac{|S|!(|N|-|S|-1)!}{|N|!}[f(S\cup i)-f(S)],\hat{y}_i=f(x_i) \quad (17)$$

其中，N 为所有特征的集合。S 为 N 的子集，不包含特征 i。f(S) 为使用特征子集 S 对输入样本 X 进行模型预测的值。通过对种特征在回归过程中的 SHAP 值计算，可以用数值的形式直观对比各个特征的预测结果的影响程度，筛选掉冗余的特征描述符。

2　结果与讨论

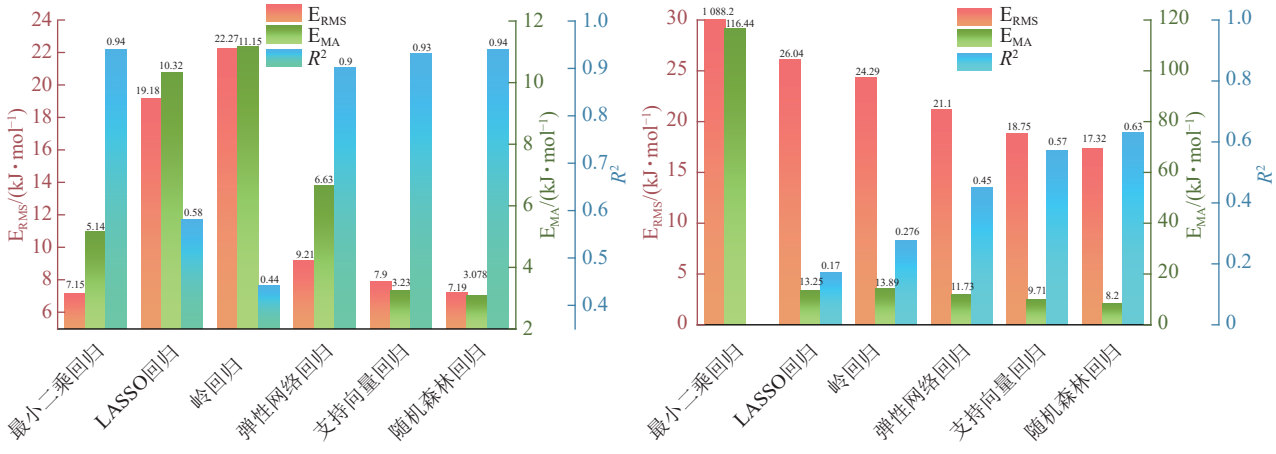
为了筛选出金属氢化物热力学性能预测结果最优的机器学习算法，笔者对上述的 4 种线性模型和 2 种非线性模型的回归结果在相同数据集上进行了回归测试。在运行每种算法时，为了控制变量，均以 7∶3 的比例将数据集划分为训练集和测试集分别进行模型训练和鲁棒性测试。

2.1　线性回归结果

回归算法结果如图 4 和表 3 所示，最小二乘回归中不含有任何范数项，无法进行特征选择或回归系数的稀疏化，因此对模型性能有负面影响的特征描述符，即可能会误导算法回归趋势的描述符无法被有效规避，导致即使在训练集上，最小二乘回归仍出现了图 5c 以中轴对称分布的预测值发散现象，这表明仅包含差值项的优化函数存在对于算法的性能优化存在上限。此外，最小二乘回归出现了极为严重的过拟合现象，决定系数甚至出现了负值。

在优化函数中加入 L_1 正则化项后，如图 6 所示，LASSO 回归的预测值数据点在测试集上向中轴出现了明显的收敛，然而训练集上出现了略微的发散。由图 4 中模型的误差数据可发现，LASSO 回归模型在训练集的 RMSE 和 MAE 分别上升至 19.18 和 10.32，测试集上的 RMSE 和 MAE 则分别降低至 26.04 和 13.25。这一结果表明 L_1 正则化项可以有效地降低模型的过拟合程度，使得模型的泛化能力得到了显著提升，然而测试集上极低的决定系数表明仅加入 L_1 正则化项无法有效提高模型精度和鲁棒性。

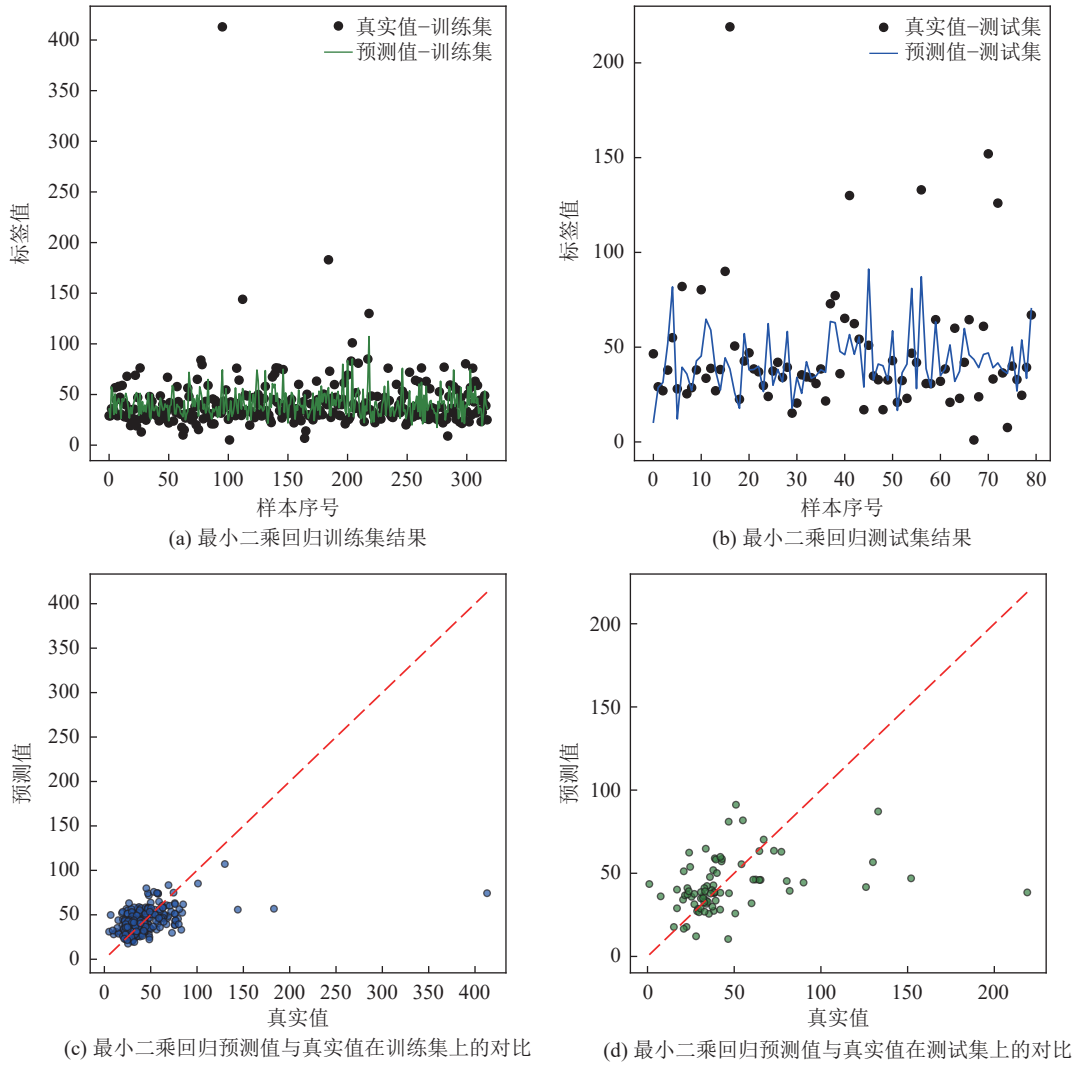
如图 7 所示，搭载 L_2 正则化项的岭回归进一步提升了模型的泛化能力，减小了模型在测试集上的 RMSE 和 MAE。对比 LASSO 回归，岭回归在测试集上的 RMSE 由 26.04 降低至 24.29，同时决定系数也得到了有效提升。L_2 正则化项在模型的自由化过程中对于过拟合现象的惩罚更为严重，因此对比 $L1$ 正则化项有更强的削弱过拟合现象的能力。LASSO 回归和岭回归的结果表明正则化项的加入可以一定程度上缓解部分有毒描述符对算法回归趋势的误导作用，然而 2 个单一正则化项的模型在训练



(a) 训练集结果　　(b) 测试集结果

图 4　回归算法结果对比

Fig. 4　Comparison of regression algorithms



(a) 最小二乘回归训练集结果　　(b) 最小二乘回归测试集结果

(c) 最小二乘回归预测值与真实值在训练集上的对比　　(d) 最小二乘回归预测值与真实值在测试集上的对比

图 5　最小二乘回归算法预测结果对照

Fig. 5　Comparative Prediction results of least square regression

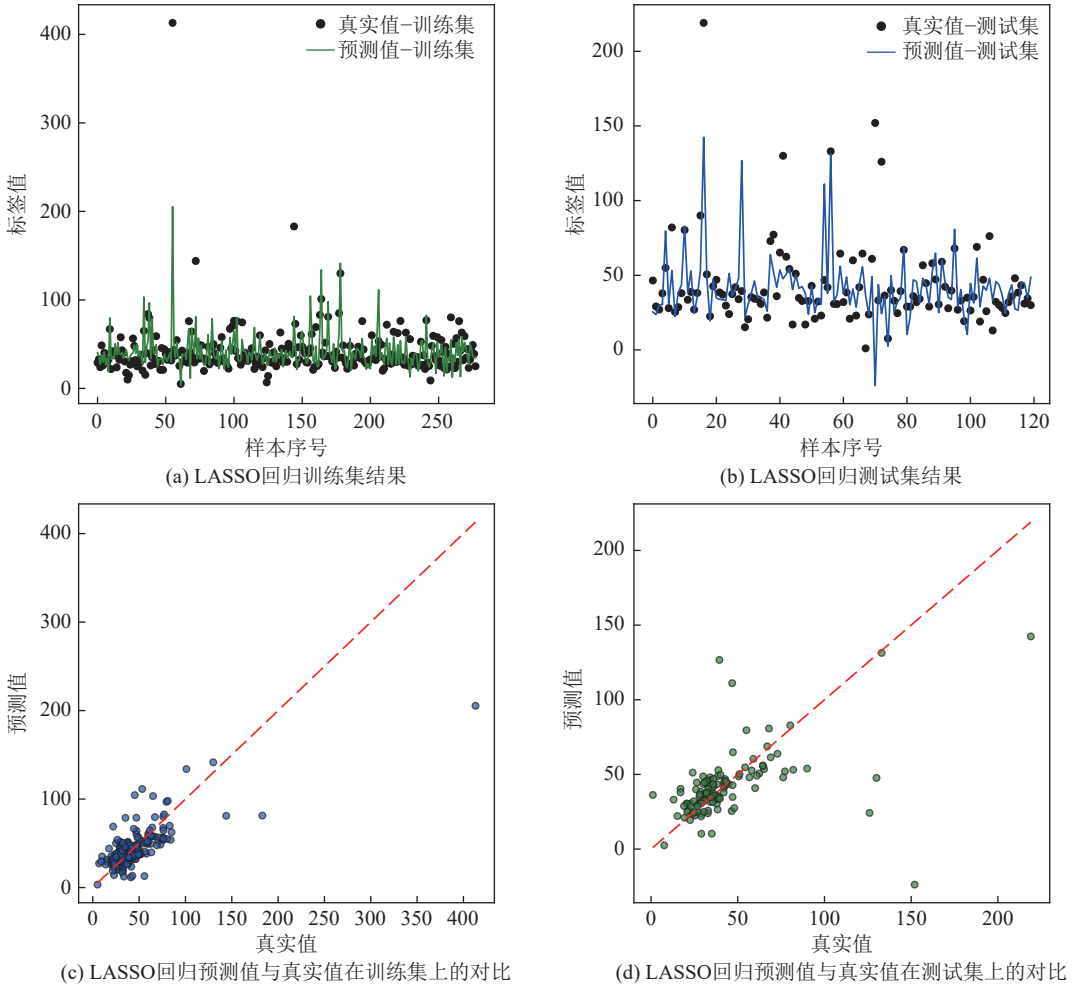
集上的决定系数仍处在较低水平，表明单一正则化项加持下模型的精度和鲁棒性存在对立关系。

为了同时兼顾 L_1 和 L_2 正则化项对线性模型过拟合现象的优化能力，本研究测试了同时搭载 L_1 和 L_2 正则化项的弹性网络回归模型。如图 8 所示，同时加入 L_1 和 L_2 正则项的弹性网络回归在四

表 3　回归算法预测性能对比

Table 3　Performance compassion of regression algorithm

回归算法	训练集性能			测试集性能		
	E_{RMS}	E_{MA}	R^2	E_{RMS}	E_{MA}	R^2
最小二乘回归	7.15	5.14	0.94	1 088.2	116.44	−1 452.4
LASSO回归	19.18	10.32	0.582	26.04	13.25	0.17
岭回归	22.27	11.15	0.44	24.29	13.89	0.276
弹性网络回归	9.21	6.63	0.9	21.1	11.73	0.45
支持向量回归	7.894 9	3.225 47	0.929 1	18.745 66	9.708 6	0.568 7
随机森林回归	7.19	3.078	0.941 2	17.32	8.201 1	0.632



(a) LASSO回归训练集结果　(b) LASSO回归测试集结果

(c) LASSO回归预测值与真实值在训练集上的对比　(d) LASSO回归预测值与真实值在测试集上的对比

图 6　LASSO 回归算法预测结果对照

Fig. 6　Comparative prediction results of LASSO regression

种线性回归算法中表现最好，对于每一个金属氢化物的形成能预测值与真实值差异最小。同时图 8 中弹性网络回归模型的 RMSE 和 MAE 无论是在训练集和测试集都低于其他 3 种线性回归模型，训练集上的决定系数也达到了较高的 0.9，这表明耦合 2 种正则化项能同时对精度和鲁棒性进行提升。然而弹性网络回归在测试集上决定系数仍较低，无法满足精度要求。

通过对比不难发现，线性回归中鲁棒性和精度存在一定冲突，面对大数据回归任务，加入正则化项是必要的，可以有效降低过拟合现象。然而即使是性能最优的线性回归算法依然无法对金属氢化物热力学性能做到较好的预测，这表明线性回归模型的拟合能力无法满足预测需求，同时金属氢化物的

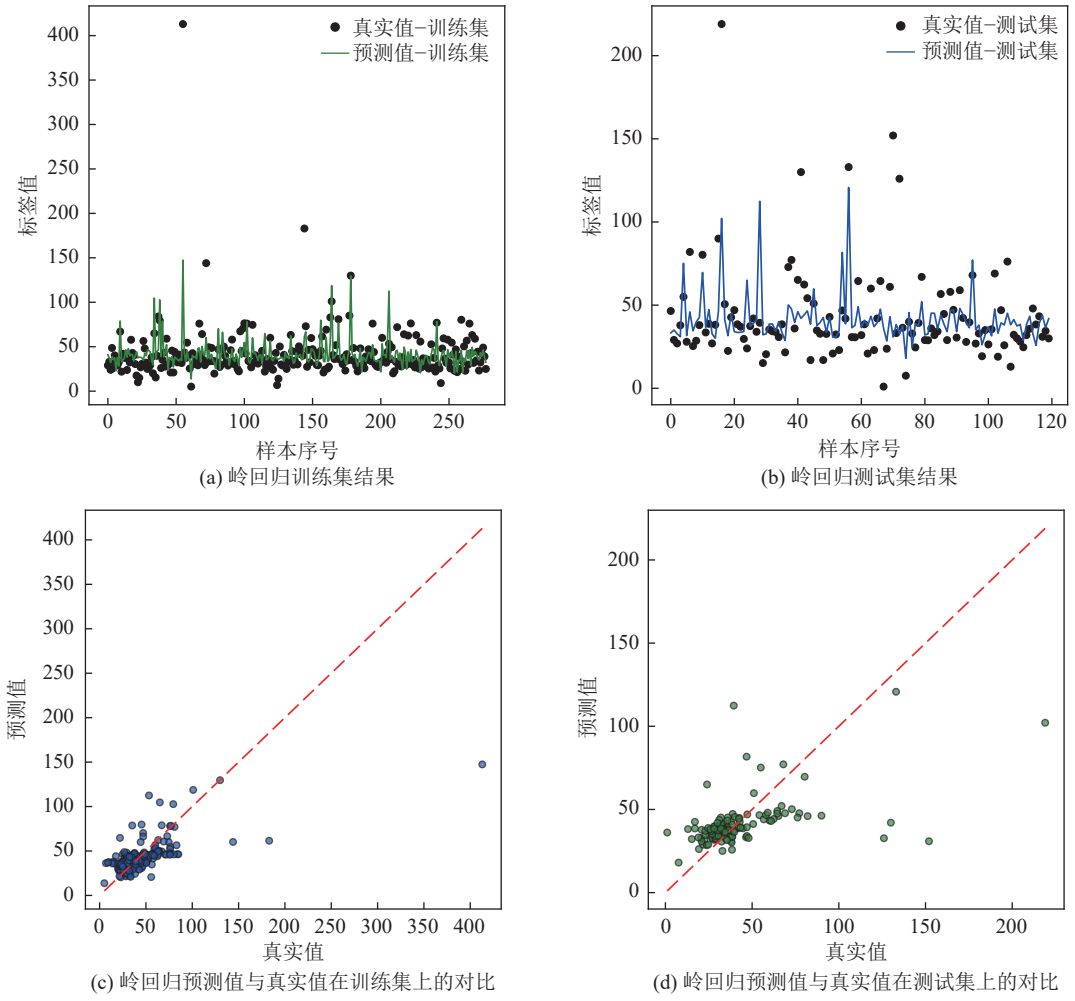


图 7　岭回归算法预测结果对照

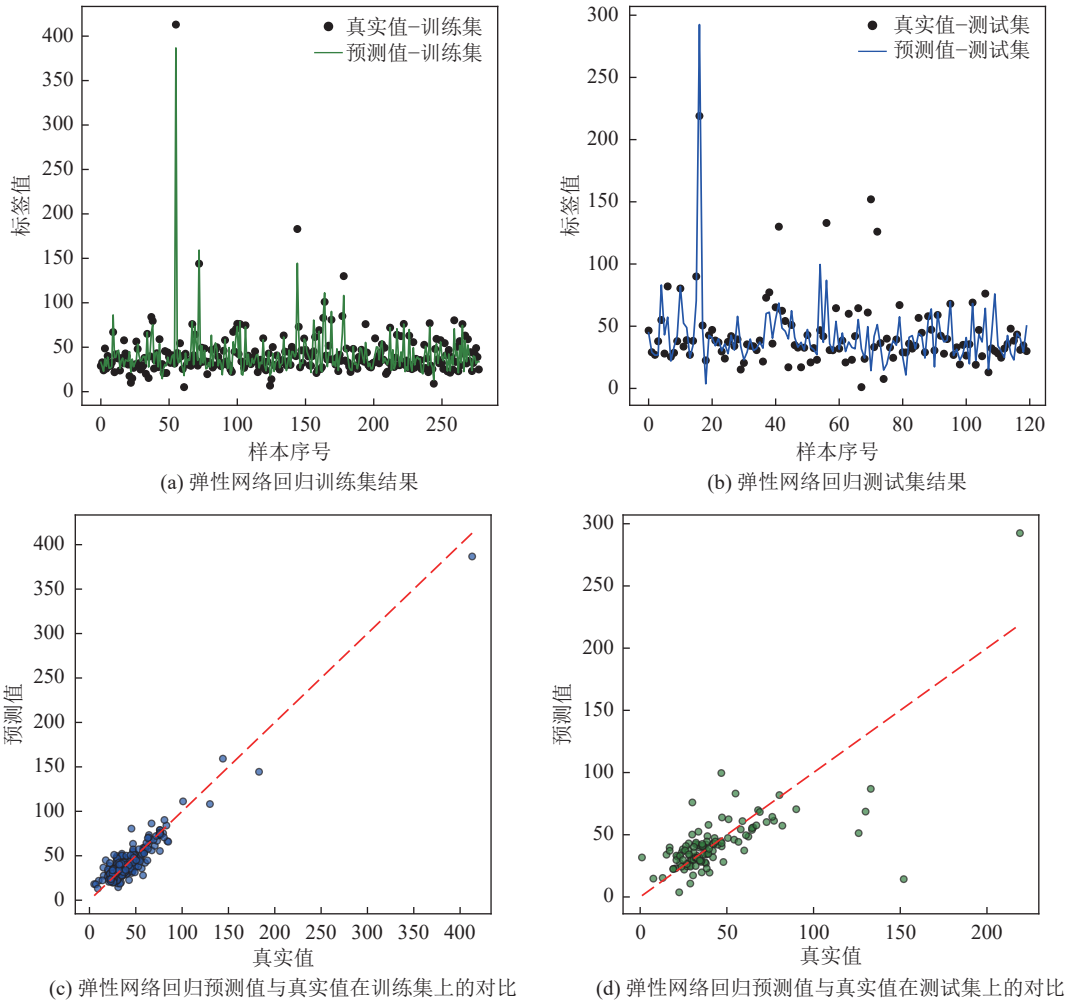
Fig. 7　Comparative prediction results of ridge square regression

形成能与所构建的 145 种描述符并不是简单的线性关系，而是更为复杂的非线性关系，因此需要使用非线性回归算法进行训练和预测。

2.2　非线性模型回归结果

SVR 和随机森林算法都善于处理多维特征变量输入的回归问题，因此其单次模型训练时长均在 10 min 以内，表明二者并未发生高维特征输入引发的模型崩溃现象。为保证模型具有充足的输入特征以捕捉金属氢化物材料微观信息，首先使用非线性模型对 145 种特征进行了回归测试，再通过结果分析进行特征筛选。SVR 的核心思想是通过寻找一个超平面来预测目标变量，同时控制预测误差在某个阈值范围内。超平面的变换使得这一回归算法具有了非线性的性质，因此相较于线性回归算法，SVR 的误差有了进一步的下降，且决定系数到达了接近 0.6 的水平（图 4）。从图 9c 和图 9d 中也可以看出，SVR 在训练集上的收敛程度很好，预测值基本都在中轴线上，在测试集上有部分发散，但仍可直观地看出优于线性回归算法。这些精度和鲁棒性的改善可以归结于 SVR 的超平面非线性变换操作，一定程度上符合了金属氢化物形成能与特征描述符之间的非线性关系。

为了进一步优化模型的精度和鲁棒性，本研究选用了随机森林算法，具有更强的非线性数据处理能力，其非线性主要来自于决策树的组合、树的随机性、特征的非线性组合和投票或平均机制。随机森林是由多个决策树组成的，每个决策树本身就是一个非线性模型。决策树通过一系列的分割规则将数据空间划分成多个子区域，并在每个子区域上预测一个常数输出。这种分割是非线性的，因为它可以创建任意形状的决策边界。在构建随机森林的过程中，每棵树都是在不同的数据样本和特征子集（随机特征选择）上训练的。这种随机性导致每棵树都能捕捉到数据中不同的模式，当这些树组合起来时，它们可以共同形成一个更加复杂和灵活的非线性模型。在决策树中，特征之间的组合可以是高度非线性的。例如，一棵树可能会在一个特定的特征值范围内进行分割，而另一棵树可能会在另一个



(a) 弹性网络回归训练集结果　(b) 弹性网络回归测试集结果

(c) 弹性网络回归预测值与真实值在训练集上的对比　(d) 弹性网络回归预测值与真实值在测试集上的对比

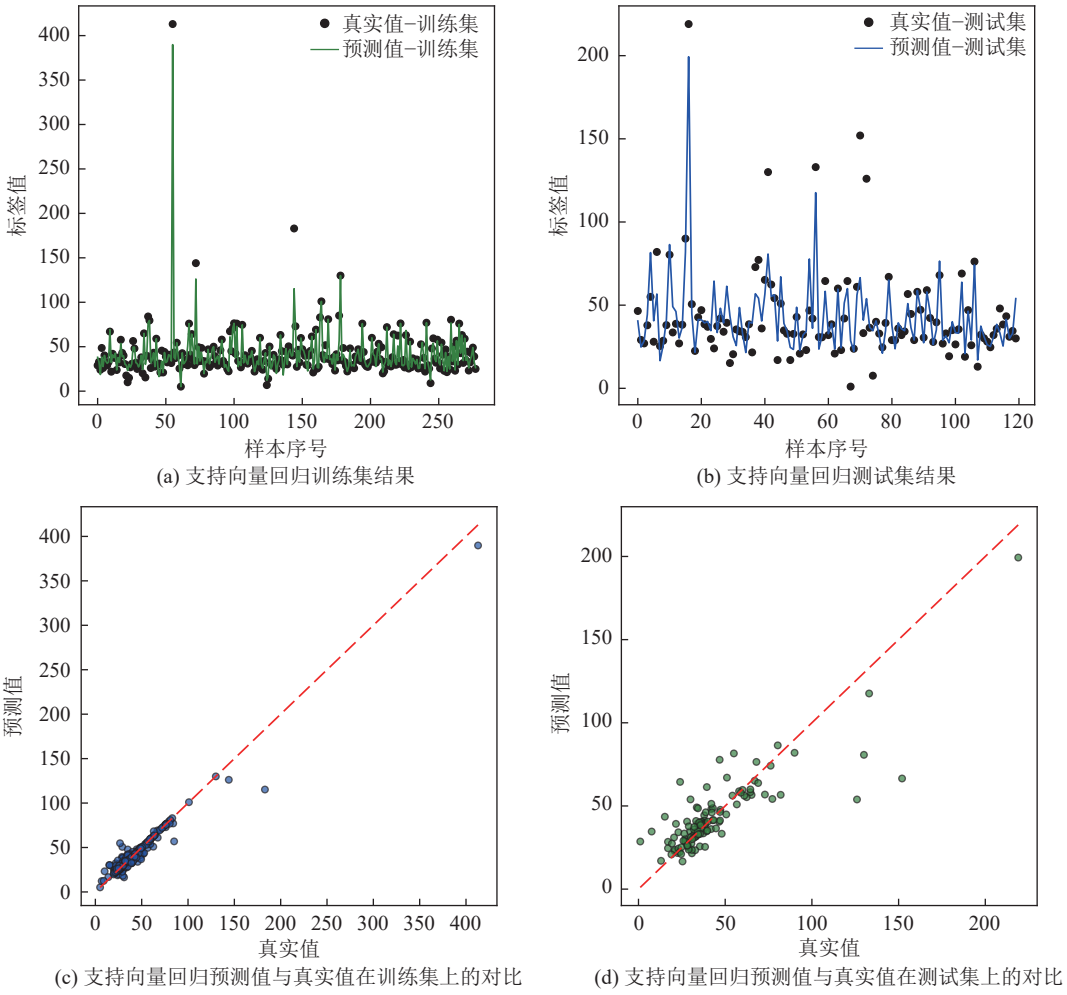
图 8　弹性网络回归算法预测结果对照

Fig. 8　Comparative prediction results of elastic net regression

特征值范围内进行分割。当这些树的结果被聚合时，它们可以表示特征之间复杂的交互作用和非线性关系。最后随机森林通过聚合所有树的预测来得到最终的预测结果。在回归问题中，这通常是通过取所有树预测的平均值来实现的。这种聚合过程可以平滑单个决策树的预测，从而产生一个整体的非线性预测函数。整体预测过程可以有效避免某些特征的过度主导作用，从而避免了过拟合，同时决策树的深度机制保障了算法的拟合能力不至于出现欠拟合，在 111 棵决策树以及每棵树最大深度设置为 16 时随机森林算法的表现最好，误差降到了最低的水平，对比图 9 和图 10 可发现，随机森林回归模型的预测值与真实值相较于支持向量回归模型更为接近。如图 4 和表 3 所示，MAE 在测试集上仅有 8.201 1，决定系数更是超过了 0.6，模型的各项误差及决定系数在训练集和测试集上的差距在正常范围内，没有出现在训练集上表现优异而测试集上表现较差的过拟合现象，这说明随机森林算法的非线性拟合能力能够有效拟合金属氢化物形成能与特征描述符之间的数据关系，可以对未知金属氢化物热力学数据进行预测。

2.3　回归结果分析

在完成模型性能测试及误差分析后，本研究对训练中使用的 145 种材料微观特征描述符进行了 SHAP 分析，计算所有特征的 SHAP 绝对值，以得出不同特征对金属氢化物形成能的影响程度，结果如图 11 所示。SHAP 分析结果表明金属氢化物形成能受其组成元素的平均基态原子体积影响最大，其 SHAP 值高达 5.56。其次是最大基态原子体积，SHAP 值为 1.26。由于金属氢化物的形成过程是氢原子从表面渗透进晶格并与金属原子结合的过程，因此形成能的大小反映了这一过程的难易程度。平均基态原子体积和最大基态原子体积的高 SHAP 值表明这一过程受组成金属氢化物原子的体积因素较大，即组成金属氢化物的基态原子体积的均值和最值很大程度上决定了金属氢化物材料的形成能大小。在微观层面，组成金属氢化物材料的元素原子基态体积越大，则合金倾向于形成的晶胞越大，可



(a) 支持向量回归训练集结果　(b) 支持向量回归测试集结果

(c) 支持向量回归预测值与真实值在训练集上的对比　(d) 支持向量回归预测值与真实值在测试集上的对比

图 9　支持向量回归算法预测结果对照

Fig. 9　Comparative prediction results of SVR

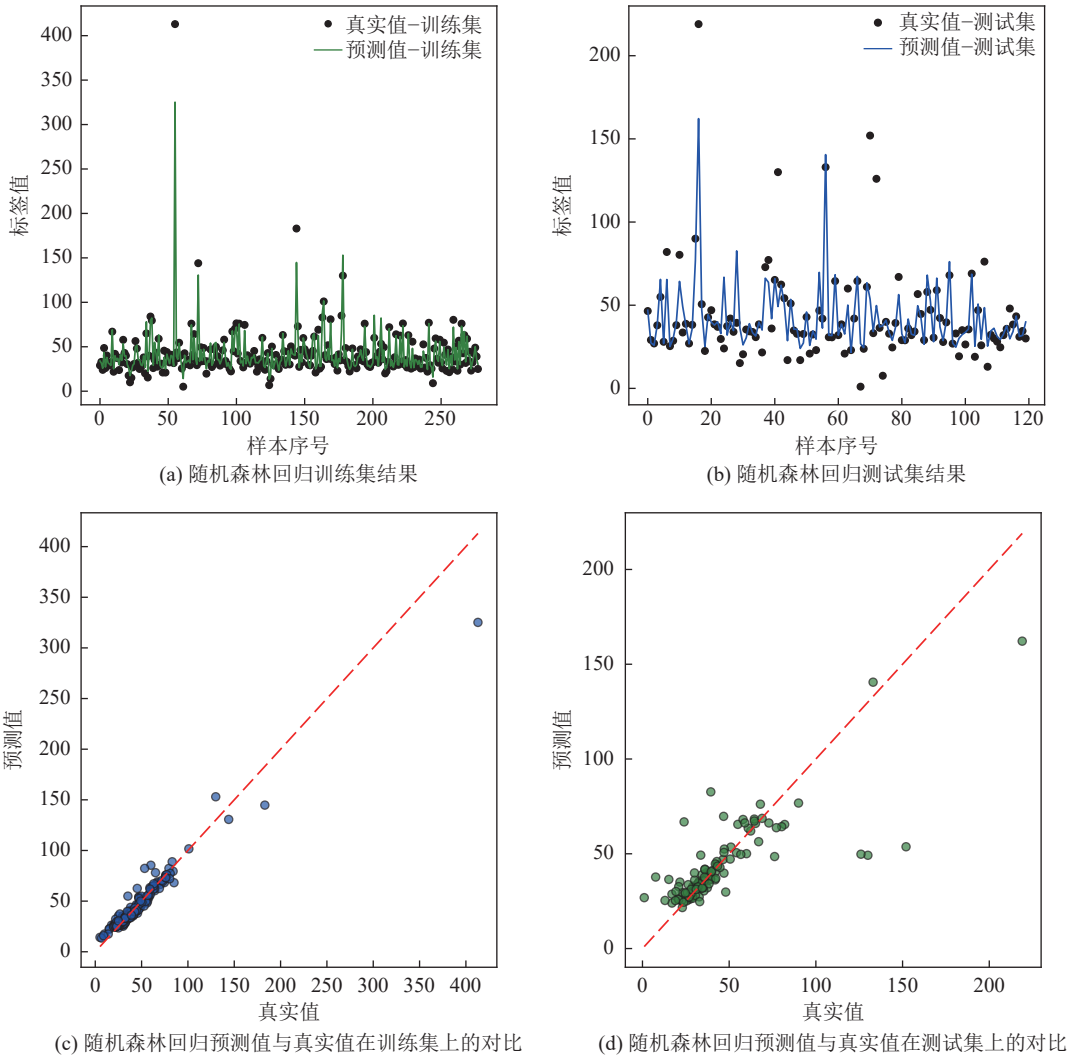
用于氢原子渗透并嵌入的晶胞可用体积就越大，金属氢化物稳定性就越高，越容易形成，对应较大的金属氢化物形成能绝对值，即较大的吸放氢焓变。因而在后续改性研究中应当关注组成元素的基态原子体积方面的因素，当设计开发储热用金属氢化物材料时，应当倾向于使用基态原子体积均值和极值较大的元素作为组分材料以获取更高的储热密度。当开发的金属氢化物材料是用于储氢运氢时，则需要考虑使用基态原子体积均值和极值较小的组成因素，以控制材料的放氢热量消耗。

为了进一步分析模型的结果，本研究对多个不同类别的金属氢化物预测值进行了分析，包含 Mg 基金属氢化物材料，Ca 基金属氢化物材料，AB、AB_2 及 AB_5 型的金属氢化物材料，结果见表 4。通过预测值以及预测相对误差可看出，模型对于多个类别的金属氢化物材料都具有较好的普适性预测能力，而不局限于某一特定类别的金属氢化物材料。模型对几种代表性金属氢化物的预测相对误差均在 9% 以下，由于不同类别的金属氢化物往往在微观特征值上具有一定的相似性，因此模型对几种代表性金属氢化物吸放氢焓变预测值的低误差证明了其对金属氢化物材料吸放氢焓变预测的准确性及普适性。

3　结　　论

金属氢化物储热材料可以通过掺杂不同元素形成多元合金，以获得不同的目标性能。以往材料开发方法严重依赖实验合成，需要以试错的方法合成多个材料样品，并对所合成材料进行性能测试和表征，十分耗费时间和经济成分，且无法处理大批量的材料数据。笔者对比了多种数据驱动的机器学习回归模型，通过测试最小二乘回归、LASSO 回归、岭回归、弹性网络回归、支持向量回归和随机森林回归的性能，筛选出最优算法对金属氢化物材料热力学性能进行预测。

1）分析结果表明线性回归算法均存在过拟合现象，无法根据所构建的 145 种量化微观组分信息对金属氢化物形成能做出合理预测，同时也印证了金



(a) 随机森林回归训练集结果　(b) 随机森林回归测试集结果

(c) 随机森林回归预测值与真实值在训练集上的对比　(d) 随机森林回归预测值与真实值在测试集上的对比

图 10　随机森林回归算法结果示意

Fig. 10　Comparative prediction results of random forest regression

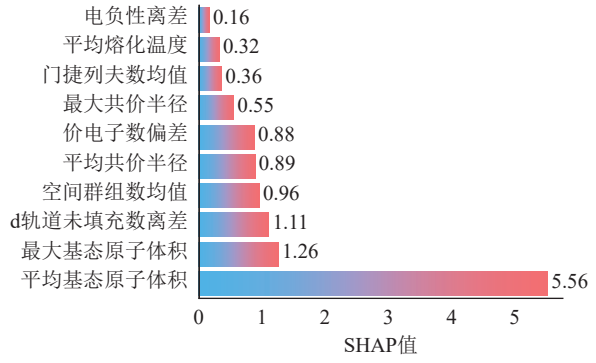


图 11　特征重要性 SHAP 分析

Fig. 11　Feature importance SHAP analysis

属氢化物形成能与这些组分信息之间的数据关系是高度非线性的。随后的测试中，随机森林回归算法展现出最好的预测性能，随机森林回归算法在训练集和测试集上平均绝对误差分别为 3.078 和 8.201 1，决定系数分别为 0.94 和 0.63。这些数据表明随机森林回归算法具有良好的回归能力和泛化能力。

表 4　代表性金属氢化物材料随机森林模型预测值及误差

Table 4　Random forest regression prediction value and error for representative metal hydride species

真实值/（kJ·mol^{-1}）/（kJ·mol^{-1}）	预测值	相对预测误差/%	合金组分	合金类型
64.5	66.62	5.12	Mg_2Ni	A_2B
64.5	67.8	5.12	$Mg_2Co_{0.25}Ni_{0.75}$	A_2B
28.1	30.83	9	TiFe	AB
41.0	40.578 31	1.02	$TiFe_{0.8}Ni_{0.15}V_{0.05}$	AB
49.4	48.06	2	ZrFeCr	AB_2
40.2	40.395 84	4.87	ZrCoCr	AB_2
30.1	29.4	2	$LaNi_5$	AB_5
42.7	40.817	4.41	$LaCu_5$	AB_5
63.2	64.08	1.4	Mg_2Al_3	Mg系
76.0	73.170 27	3.72	Mg_2Co	Mg系
43.5	40.78	7	$Ca_{0.5}La_{0.5}Ni_3$	过渡金属系
64.2	60.918 02	5.11	$LaMg_2Cu_2$	过渡金属系

2）SHAP 分析中金属氢化物材料的平均基态原子体积和最大基态原子体积具有较高的 SHAP 值，即组成金属氢化物的基态原子体积的均值和最值很大程度上决定了金属氢化物材料的形成能大小。当设计开发储热用金属氢化物材料时，应当倾向于使用基态原子体积均值和极值较大的元素作为组分材料以获取更高的储热密度。当开发的金属氢化物材料是用于储氢运氢时，则需要考虑使用基态原子体积均值和极值较小的组成因素，以控制材料的放氢热量消耗。

3）本研究最后对多个不同类别的金属氢化物预测值进行了分析，包含 Mg 基金属氢化物材料，Ca 基金属氢化物材料，AB、AB_2 及 AB_5 型的金属氢化物材料，模型对几种代表性金属氢化物材料的预测相对误差均在 9% 以下，由于不同类别的金属氢化物往往在微观特征值上具有一定的相似性，因此模型对几种代表性金属氢化物吸放氢焓变预测值的低误差证明了其对金属氢化物材料吸放氢焓变预测的准确性及普适性，可用于未知金属氢化物材料的形成能预测。

参考文献（References）：

[1] BELAÏD F, AL-SARIHI A, AL-MESTNEER R. Balancing climate mitigation and energy security goals amid converging global energy crises: The role of green investments[J]. Renewable Energy, 2023, 205: 534–542.

[2] 姚玉璧, 郑绍忠, 杨扬, 等. 中国太阳能资源评估及其利用效率研究进展与展望 [J]. 太阳能学报, 2022, 43(10): 524–535.
YAO Yubi, ZHENG Shaozhong, YANG Yang, et al. Progress and prospects on solar energy resource evaluation and utilization efficiency in China[J]. Acta Energiae Solaris Sinica, 2022, 43(10): 524–535.

[3] BEZDUDNY A V, BLINOV D V, DUNIKOV D O. Single-stage metal hydride-based heat storage system[J]. Journal of Energy Storage, 2023, 68: 107590.

[4] MALLESWARARAO K, DUTTA P, MURTHY S S. Applications of metal hydride based thermal systems: A review[J]. Applied Thermal Engineering, 2022, 215: 118816.

[5] DANGWAL S, IKEDA Y, GRABOWSKI B, et al. Machine learning to explore high-entropy alloys with desired enthalpy for room-temperature hydrogen storage: Prediction of density functional theory and experimental data[J]. Chemical Engineering Journal, 2024, 493: 152606.

[6] AGAFONOV A, PINEDA-ROMERO N, WITMAN M, et al. Destabilizing high-capacity high entropy hydrides *via* earth abundant substitutions: From predictions to experimental validation[J]. Acta Materialia, 2024, 276: 120086.

[7] YIN Y, LI B, YUAN Z M, et al. Microstructure and improved hydrogen storage properties of $Mg_{85}Zn_5Ni_{10}$ alloy catalyzed by Cr_2O_3 nanoparticles[J]. Journal of Physics and Chemistry of Solids, 2019, 134: 295–306.

[8] ZHONG H C, HUANG Y S, DU Z Y, et al. Enhanced Hydrogen Ab/De-sorption of Mg(Zn) solid solution alloy catalyzed by YH2/Y_2O_3 nanocomposite[J]. International Journal of Hydrogen Energy, 2020, 45(51): 27404–27412.

[9] RKHIS M, LAASRI S, TOUHTOUH S, et al. New insights into the electrochemical and thermodynamic properties of AB-type ZrNi hydrogen storage alloys by native defects and H-doping: Computational experiments[J]. International Journal of Hydrogen Energy, 2023, 48(27): 10089–10097.

[10] GU X D, WANG F, CHENG J L, et al. Positive correlation of Nb/Cr doping with dehydrogenation performance of ZrCo-based hydrides[J]. International Journal of Hydrogen Energy, 2023, 48(67): 26276–26287.

[11] DONG S Y, WANG Y Y, LI J Y, et al. Exploration and design of Mg alloys for hydrogen storage with supervised machine learning[J]. International Journal of Hydrogen Energy, 2023, 48(97): 38412–38424.

[12] MOOSAVI S M, JABLONKA K M, SMIT B. The role of machine learning in the understanding and design of materials[J]. Journal of the American Chemical Society, 2020, 142(48): 20273–20287.

[13] ZHAI S, XIE H P, CUI P, et al. A combined ionic Lewis acid descriptor and machine-learning approach to prediction of efficient oxygen reduction electrodes for ceramic fuel cells[J]. Nature Energy, 2022, 7: 866–875.

[14] MOOSAVI S M, NANDY A, JABLONKA K M, et al. Understanding the diversity of the metal-organic framework ecosystem[J]. Nature Communications, 2020, 11(1): 4068.

[15] GHEYTANZADEH M, RAJABHASANI F, BAGHBAN A, et al. Estimating hydrogen absorption energy on different metal hydrides using Gaussian process regression approach[J]. Scientific Reports, 2022, 12(1): 21902.

[16] SAAL J E, KIRKLIN S, AYKOL M, et al. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)[J]. JOM, 2013, 65(11): 1501–1509.

[17] KIRKLIN S, SAAL J E, MEREDIG B, et al. The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies[J]. NPJ Computational Materials, 2015, 1: 15010.

[18] 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): 011002.

[19] WARD L, AGRAWAL A, CHOUDHARY A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials[J]. NPJ Computational Materials, 2016, 2: 16028.

[20] RODRIGUEZ-GALIANO V, SANCHEZ-CASTILLO M, CHICA-OLMO M, et al. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines[J]. Ore Geology Reviews, 2015, 71: 804–818.

[21] AKIBA T, SANO S, YANASE T, et al. Optuna: a next-generation hyperparameter optimization framework[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage AK USA. ACM, 2019: 2623–31.10. 1145/3292500.3330701.