**目 录**

[1 研究概述 1](#_Toc2005)

[1.1研究背景及意义 1](#_Toc27857)

[1.2研究内容 1](#_Toc10915)

[1.2.1非贵/低贵金属基纳米材料协同催化作用机理 1](#_Toc6412)

[1.2.2快速筛选高性能氢能关键催化材料方法 2](#_Toc19365)

[1.2.3高性能氢能催化材料的高通量智能筛选与精准设计 2](#_Toc6670)

[1.3 研究技术路线 2](#_Toc7463)

[1.3.1 建立氢能关键催化材料催化反应性能的理论评估方法 2](#_Toc9740)

[1.3.2 建立氢能关键催化材料的量子化学数据库 3](#_Toc32062)

[1.3.3 氢能关键催化材料的高通量智能筛选设计平台 3](#_Toc26803)

[2 研究进展 4](#_Toc16166)

[2.1氢能关键催化材料催化析氢反应的设计与筛选 4](#_Toc28040)

[2.1.1 TiVC MXenes催化HER反应机理研究 5](#_Toc6304)

[2.1.2 理论设计及高通量筛选 6](#_Toc21076)

[2.2氢能关键材料量子化学数据库 11](#_Toc6960)

[2.2.1 计算数据库介绍 12](#_Toc10720)

[2.2.2 文献库介绍 13](#_Toc8307)

[2.2.3 材料数据库介绍 15](#_Toc31110)

[2.3 机器学习精准设计 18](#_Toc2522)

[2.3.1关键特征选取 20](#_Toc20584)

[2.3.2 机器学习模型训练 22](#_Toc30608)

[2.3.3 机器学习模型预测及DFT验证 23](#_Toc11599)

[3 结论 24](#_Toc30560)

[参考文献 25](#_Toc20598)

**插图清单**

[图 1 表面物种稳定性图 6](#_Toc13854)

[图2 TiVCO](#_Toc17780)[2](#_Toc17780)[结构示意图 7](#_Toc17780)

[图 3 非金属掺杂位点及结合能图 7](#_Toc31100)

[图 4 金属掺杂位点和氢吸附位点及自由能图 8](#_Toc19585)

[图 5 改性TiVCO](#_Toc21726)[2](#_Toc21726)[结构及氢吸附位点图 8](#_Toc21726)

[图 6 过渡金属原子与TM-NM-TiVCO](#_Toc4867)[2](#_Toc4867)[结合能图 9](#_Toc4867)

[图 7 氢吸附自由能（ΔG](#_Toc7402)[H](#_Toc7402)[）和电流密度分布图 10](#_Toc7402)

[图 8 TM-B-TiVCO](#_Toc31209)[2](#_Toc31209)[态密度图 11](#_Toc31209)

[图 9 数据库构建流程图 12](#_Toc23098)

[图 10 部分计算数据库展示 13](#_Toc28603)

[图 11 文献库数据展示 14](#_Toc6576)

[图 12 文献PDF下载流程 14](#_Toc25483)

[图 13 材料数据库关键词分类 16](#_Toc5897)

[图 14 PDF文献提取图片流程 17](#_Toc25640)

[图 15 部分材料数据库展示 17](#_Toc30133)

[图 16 机器学习流程图 19](#_Toc16276)

[图 17 XGBoost代码展示 20](#_Toc31718)

[图 18 相关性及特征重要性排序 22](#_Toc14008)

[图 19 不同算法模型的混淆矩阵及ROC曲线 23](#_Toc1000)

# 1 研究概述

## 1.1研究背景及意义

随着“十四五”电力规划的实施，我国正加速能源清洁化转型进程，脱碳减排需求日益增长，在“双碳”战略指导下，“十四五”期间风电、氢能等可再生能源将迎来爆发式增长，可再生能源将逐步替代传统化石能源占据能源领域主导地位。氢能是一种来源丰富、绿色低碳、应用广泛的二次能源，正逐步成为全球能源转型发展的重要载体之一。

氢能材料催化剂存在活性中心种类复杂多样、不同电极反应中作用机理不明确、材料微结构与性能之间的关系不清楚、能源材料数据库缺失、高效筛选困难等问题，制约着氢能的发展。探索适用不同反应的活性中心结构与催化机制，建立交互式、可视化的能源材料数据库并实现高性能氢能催化材料的智能设计与高效筛选是关键科学问题之一。基于不同电极反应过程对非贵/低贵金属催化剂表界面功能结构和多维分布的不同要求，拟通过密度泛函方法与过渡态理论构建高活性结构模型，认识催化剂构效关系，并使用机器学习辅助实现非贵/低贵金属催化剂的智能设计与筛选。

## 1.2研究内容

### 1.2.1非贵/低贵金属基纳米材料协同催化作用机理

采用实验观察手段很难获知非贵/低贵金属基纳米材料协同催化作用机理，如何通过理论研究手段揭示非贵/低贵金属基纳米材料的催化性能优于其他贵金属基催化剂的机理，是本项目拟解决的第一个关键科学问题。本项目拟采用基于DFT的量子化学计算方法，结合微观动力学分析方法，形成一套完整的评估非贵/低贵金属基纳米材料催化氧还原、氧析出、氢氧化和氢析出反应性能的理论方法，实现计算结果与实验结果相互验证。通过分析非贵/低贵金属基纳米材料的结构对其理论电催化活性的影响，深入理解非贵/低贵金属基纳米材料在协同催化电解水反应中的作用，揭示已有实验中非贵/低贵金属基纳米材料高活性的深层机理。

### 1.2.2快速筛选高性能氢能关键催化材料方法

现阶段尚缺少完整准确的电解水催化反应数据库用于快速筛选高性能氢能关键催化材料，揭示影响非贵/低贵金属基纳米材料催化性能的核心结构特性参数并形成能源材料数据库，这是本项目需要解决的第二个关键科学问题。本项目拟构建针对氧还原、氧析出、氢氧化和氢析出反应的相应表面结构作为计算模型，在模型催化剂上进行计算模拟，通过评估方法计算它们的催化性能并寻找可能对催化活性产生影响的结构特性参数，作为后续构建机器学习算法模型的源数据。通过构建适用于不同反应结构模型的结构描述符，开发基于非贵/低贵金属基纳米催化剂表面结构特征的中间产物吸附能预测模型，揭示非贵/低贵金属基纳米催化剂的结构特征对电解水反应的调控机制。

### 1.2.3高性能氢能催化材料的高通量智能筛选与精准设计

如何根据氢能源材料数据库实现高性能氢能催化材料的高通量智能筛选与精准设计，最终得到用于电解水反应的高活性非贵/低贵金属基纳米催化剂体系，是亟待解决的第三个关键科学问题。本项目拟采用机器学习技术对包含催化剂结构特性参数（电子结构和几何结构）、中间体吸附能、理论过电势的数据库进行数据挖掘，分析反应位点电子结构特性对活性影响的核心因素，建立材料成分、工艺、微结构、性能之间的关联模型，根据机器学习训练模型初步得到影响催化活性的关键参数与结构描述符，完成数据库的交互式、可视化设计，并完成高效率智能设计平台的建立，实现催化性能导向的逆向分析，最终实现氢能关键催化材料的高通量筛选及精准设计。

## 1.3 研究技术路线

### 1.3.1 建立氢能关键催化材料催化反应性能的理论评估方法

采用量化计算，根据氢能电化学相关反应（氧还原、氧析出、氢氧化和氢析出）反应路径，在模型催化剂上进行计算模拟，计算在不同表面模型上的反应物、中间产物、最终产物的吸附能，获取每步基元反应的活化能垒。结合微观动力学分析获取各步基元反应的反应速率。通过比较不同反应路径的速控步的反应速率，找到主（副）反应的最优反应路径。最后结合表面模型上主（副）反应的反应速率和已有实验中催化剂每种反应位点的比例，转化成特定输出电压条件下的电流密度，形成一套完整的非贵/低贵金属基纳米材料催化电极反应性能的理论评估方法。进行计算结果与实验结果的对比，验证理论评估方法的准确性。

### 1.3.2 建立氢能关键催化材料的量子化学数据库

基于已有实验对非贵/低贵金属基纳米催化剂的反应位点的精确表征，构建针对氢能电化学相关反应催化剂的相应表面结构作为计算模型。利用上述建立的理论评估方法计算它们的催化性能，作为后续构建机器学习算法模型的源数据。由于决定氧还原、氧析出、氢氧化和氢析出等相关电化学反应活性的动力学参数之间存在内在关联式，因此可以建立关联中间产物吸附能与相关反应主（副）产物转化速率的“火山型”等高线图。基于不同表面模型上反应位点的电子结构和几何结构，构建适用于不同反应的结构模型的结构描述符，开发基于非贵/低贵金属基纳米催化剂表面结构特征的中间产物吸附能预测模型。对结构描述符包含的各个结构参数进行敏感性测试，获取非贵/低贵金属基纳米催化剂反应位点的不同结构特征对氧还原、氧析出、氢氧化与氢析出等反应催化性能影响的比重，进一步揭示非贵/低贵金属基纳米催化剂的结构特征与氧还原、氧析出、氢氧化与氢析出相关反应的调控机制。

### 1.3.3 氢能关键催化材料的高通量智能筛选设计平台

针对上述包含催化剂结构特性参数、中间体吸附能、理论过电势的数据库，采用机器学习技术进行数据挖掘，针对数据库规模与特性测试监督学习与非监督学习对数据库的适用性，利用多种机器学习算法分析反应位点电子结构特性对活性影响的核心因素，建立材料成分、工艺、微结构、性能之间的关联模型。根据机器学习训练模型初步得到影响催化活性的关键参数与结构描述符，完成数据库的交互式、可视化设计，并完成高效率智能设计平台的建立，实现催化性能导向的逆向分析。并根据课题2的实验结果不断优化数据库与机器学习模型，达到氢能关键催化材料的高通量筛选及精准设计。

# 2 研究进展

## 2.1氢能关键催化材料催化析氢反应的设计与筛选

氢能作为一种绿色、来源广泛的二次能源，在未来能源体系中发挥重要作用，高效的电解水制氢技术是可再生能源制氢的核心，具有绿色环保、生产灵活、产品纯度高等优势。目前性能最佳的催化剂仍然是铂基催化剂，但其高成本、资源短缺以及有限的稳定性制约了电解水产业的大规模发展。

在众多可能取代铂用于HER的催化剂中，类似MXenes的2D材料已经渐露头角，引起人们的广泛关注。MXenes是二维过渡金属碳化物、氮化物或碳氮化物材料，一般化学式为Mn+1XnTx，其中n=1-3，M为早期过渡金属，X = C和/或N1, 2。MXenes材料具有高活性位点、高亲水性、高导电性和高结构稳定性，在电解水制氢方面被寄予厚望3, 4。超过50种类型的MXenes已被证明在理论上是可行的5。

通过文献调研，可以通过表面端基物种修饰，掺杂过渡金属原子（TM）、非金属原子（NM）、调节厚度、拉伸等手段提高MXene材料的HER催化活性。表面端基物种主要包含O、OH、F等基团，甘等人6通过表面相图计算，发现在标况下，Mo2C表面完全被O所覆盖形成Mo2CO2，在计算生成热时，Mohnish等人7发现，在没有表面物种钝化表面的情况下，大多数MXenes是不稳定的。在掺杂非金属和金属原子改性方面，高等人8通过计算过渡金属原子吸附能和电子结构信息发现，单原子吸附或取代能够改善功能化Ti3C2材料的催化活性。李等人9研究了TiCO2负载过渡金属单原子，发现Ti-、Rh-、Ir-和Pt-TiCO2具有较低的氢吸附自由能，TM的存在能够降低氢气脱附反应的自由能和提高HER的效率。

目前研究较多的是Mo2C、Ti2C、Ti3C2等上下表面一致的MXenes材料，而对于上下表面不一致的TiVC材料研究的较少，王等人10通过高通量筛选证明TiVC材料具有较好的HER性能，因此，我们以TiVC研究对象，对其进行改性，进一步提高其HER催化性能。

### 2.1.1 TiVC MXenes催化HER反应机理研究

（1） HER反应机理

电解水制氢时，质子或水合氢离子在阴极得到电子，发生还原反应，生成H2析出，该过程简称为析氢反应（HER）。HER可基于Volmer-Heyrovsky机制或Volmer-Tafel机制发生，碱性和酸性条件的演化步骤相似。

碱性条件下HER的步骤如下：

H2O + e- → OH- + Hads （Volmer）

Hads+ H2O + e- → OH- +H2 （Heyrovsky）

or 2Hads → H2 （Tafel）

酸性条件下HER的步骤如下:

H++e- + \* → Hads （Volmer）

Hads + H++ e- → H2 （Heyrovsky）

or 2Hads → H2 （Tafel）

氢吸附自由能（ΔGHads）已经成为公认的HER反应的能量描述符，通过计算ΔGHads来判断催化剂是否具有较好的催化性能。良好的HER催化剂应具有接近于0的ΔGHads，这往往作为高通量筛选的标准。并且认为氢吸附自由能的绝对值（|ΔGHads|）≤0.2 eV时具有较好的催化活性。

1. 表面端基物种

为了确定TiVC的表面端基物种，绘制相关电位（USHE）和pH值下热力学最稳定的表面Pourbaix图，如图1所示。由于TiVC材料具有Ti和V两个不同的表面，为了确定最稳定的反应面，我们在Ti面和V面分别进行了表面端基物种的稳定性研究。以Ti面为例，图1a为TiVC的表面Pourbaix图。当pH为0时，保护裸TiVC单层免受水氧化所需的USHE低至-2.39 V。随着pH值的增加（斜率为-0.059 V/pH），需要更大的负电位来防止TiVC氧化。一旦电位超过TiVC的最大保护电位，水就开始被氧化，TiVC表面部分被OH\*覆盖。随着电位不断增加，在USHE=-1.98 V时，TiVC表面完全被OH\*终止。随着电位不断增大，OH\*开始被氧化形成O\*。此外，当电位增加到-0.21 V时，TiVC表面的OH\*全部被氧化，形成稳定的O\*端-TiVCO2。因此，在标准条件下，TiVC将被O\*完全终止。

此外，在USHE=0 V时，高效催化剂还应具有较高的稳定性。因此，我们研究了不同末端的TiVC在USHE=0 V时不同pH值下的相对稳定性。在给定的pH下，自由能最低的状态是最稳定的，对应于图1b的底线。从图1b可以看出，在很宽的pH范围内，以O\*完全终止的TiVC是最稳定的状态。V面和Ti面是一样的。因此，我们以TiVCO2为初始结构对其进行改性研究。

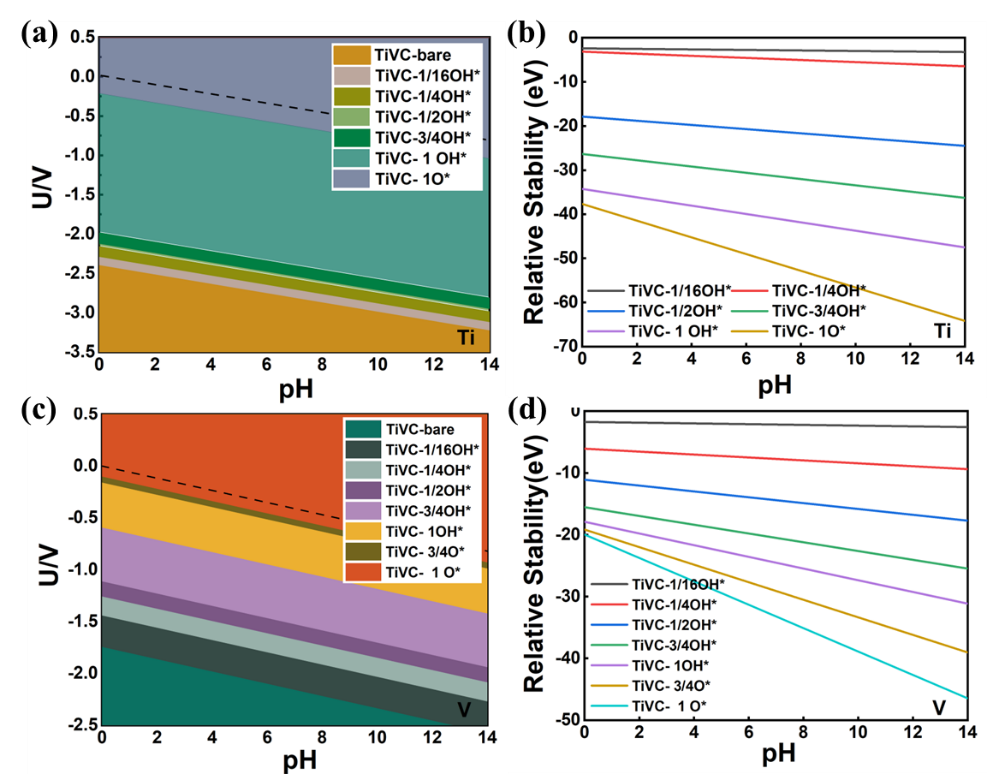


图 1 表面物种稳定性图

### 2.1.2 理论设计及高通量筛选

（1）理论设计

TiVCO2由O-Ti-C-V-O五个原子层组成，其中氧原子构成上下表面的最外层，Ti和V的过渡金属层构成上下两个次级外表面。C层夹在Ti和V过渡金属层之间（图2）。为了确定HER反应是发生在Ti面还是V面，我们分别考虑一个氢原子在Ti面和V面氧原子上的吸附，结果显示钛表面的|ΔGH|较小，所以，后续研究都在Ti表面上进行。

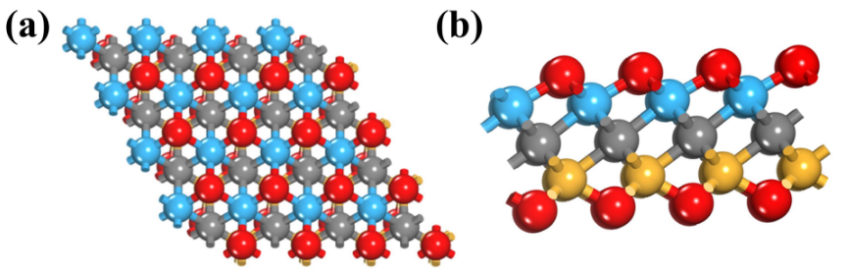


图2 TiVCO2结构示意图

为了提高TiVCO2的HER催化性能，分别引入非金属原子（B、N、P、S）和过渡金属原子（3d、4d、5d），由于在Ti表面进行反应，所以不考虑引入Ti的情况，有毒性、放射性和液体金属除外，共考虑26种过渡金属。首先考虑了非金属原子可能引入的三个位点，如图3a所示，分别计算了四种非金属元素在三种位点处的结合能（图3b），可以发现在LS位点时，四种非金属元素的结合能变化幅度不大，且都能够稳定存在，所以，确定LS为非金属原子引入位点。在此基础上，我们引入金属原子。以引入非金属元素B和金属元素Sc为例，考虑到非金属原子与金属原子之间可能存在相互作用及配位环境的影响，共考虑了四种替代方式和两种掺杂方式，每种方式考虑三种不同的氢吸附位点，共18种结构，如图4a所示，相应的氢吸附自由能如图4b所示。发现hcp掺杂位点的氢吸附位点1具有较低的氢吸附自由能，为-0.415 eV。所以后续均按照该构型进行研究，如图5所示。

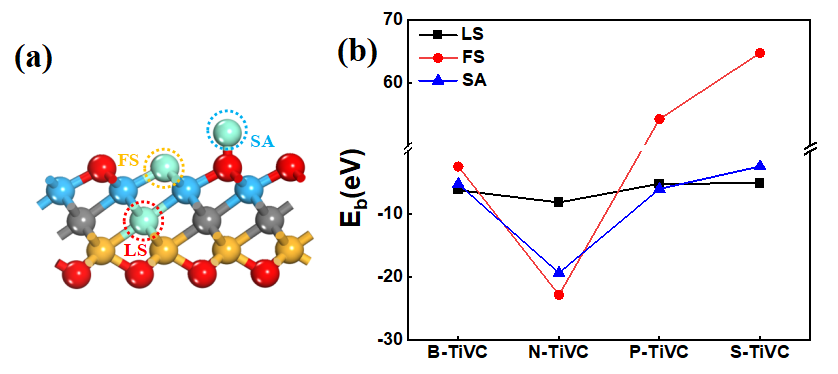


图 3非金属掺杂位点及结合能图

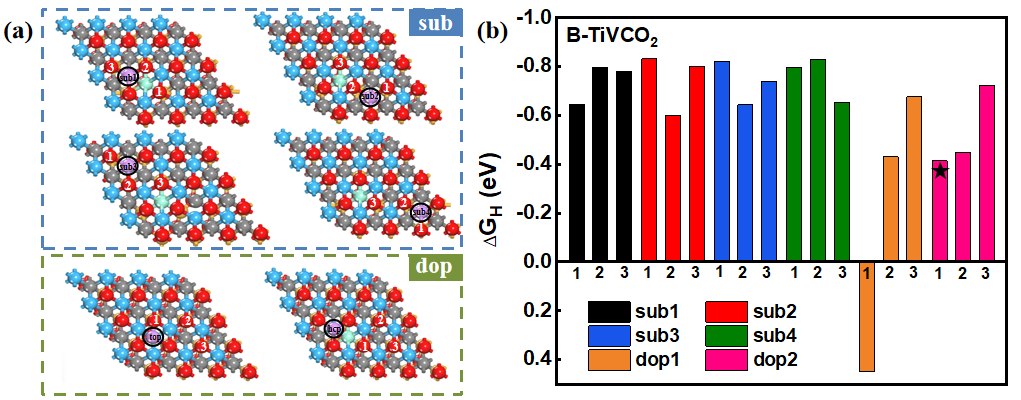


图 4 金属掺杂位点和氢吸附位点及自由能图

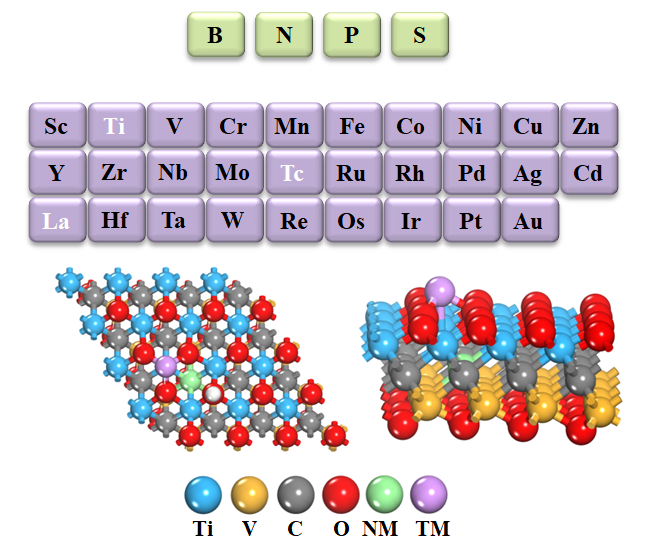


图 5改性TiVCO2结构及氢吸附位点图

1. 高通量筛选

图6展示了B-TiVCO2、N-TiVCO2和P-TiVCO2与26种TM原子的结合能情况。由图可知，所有过渡金属的结合能均为负值，说明能够稳定存在于TiVCO2中，结合能与过渡金属原子之间存在一定的周期性关系。在同一周期，结合能随价电子数的增加而增大。

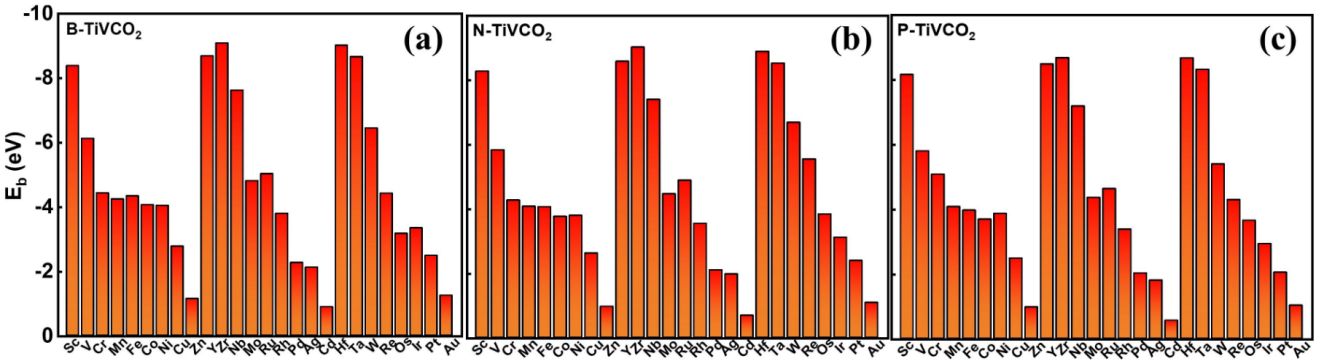
****

图 6 过渡金属原子与TM-NM-TiVCO2结合能图

理想的HER催化剂与H的结合既不能太强也不能太弱，即ΔGH应该接近于0。由于有足够的活性位点，大多数2D材料表现出良好的HER性能，其ΔGH为-0.2 ~ 0.2 eV11, 12。TMs-NM-TiVCO2的氢吸附自由能和电流密度如图7所示。引入TMs后，除Os-B-TiVCO2外，其余催化剂与TiVCO2相比ΔGH的负值显著增加至正值。正的ΔGH（Co-B-TiVCO2和Os-N-TiVCO2）代表与氢原子之间具有较弱的相互作用，有利于HER反应13。交换电流密度位置越靠近火山曲线的峰值，催化剂催化性能越好。具有负的ΔGH和正的ΔGH的催化剂分别位于火山的左右两侧，而ΔGH接近于零的催化剂位于火山曲线的峰顶。从图7（d）、（e）和（f）可以看出，Nb、Ta、Co-B-TiVCO2、Os-N-TiVCO2和Nb-P-TiVCO2的交换电流密度值位于顶点附近，具有优异的HER活性。显然，TMs的掺杂可以有效地提高大多数NM-TiVCO2的HER活性，特别是Nb、Ta、Co-B-TiVCO2、Os-N-TiVCO2和Nb-P-TiVCO2。

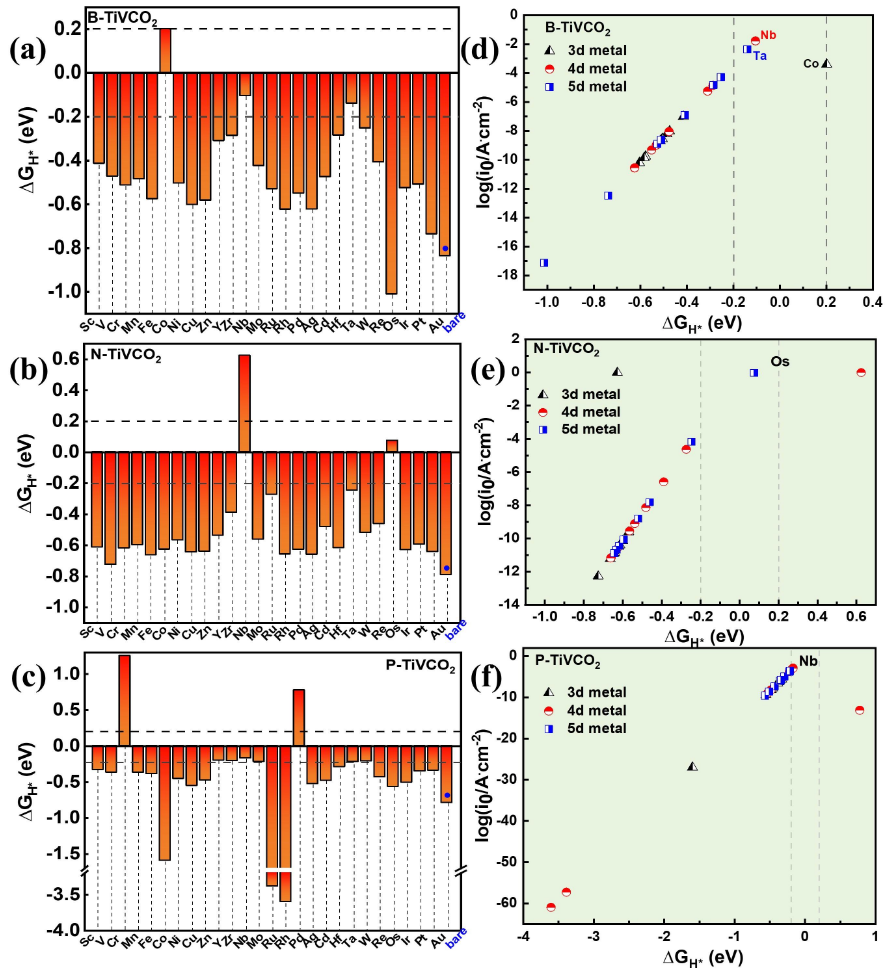


图 7氢吸附自由能（ΔGH）和电流密度分布图

1. 电子结构分析

电子结构分析能够从电子角度了解反应的内在机理，较好的解释和指导催化剂的理论设计，具有重要的研究意义。常见的电子结构分析包括：态密度分析、Bader电荷分析、差分电荷分析和静电势分析等。我们同样对计算体系进行电子结构分析，以TM-B-TiVCO2为例，我们进行了态密度计算及分析，如图8所示。蓝色虚线代表表面氧原子的p带中心，它被定义为2p轨道相对于费米能级的平均能量位置。可以看出，从Nb掺杂到Co掺杂，B-TiVCO2中氧原子的p带中心离费米能级越来越远。由于金属修饰导致p带中心左移，H和O之间的相互作用减弱，与裸B-TiVCO2相比，HER活性提高。这些结果表明，TM掺杂会使氧的p带中心下移，减小成键轨道和反键轨道之间的能隙，降低对H的吸附强度，最终提高催化活性。

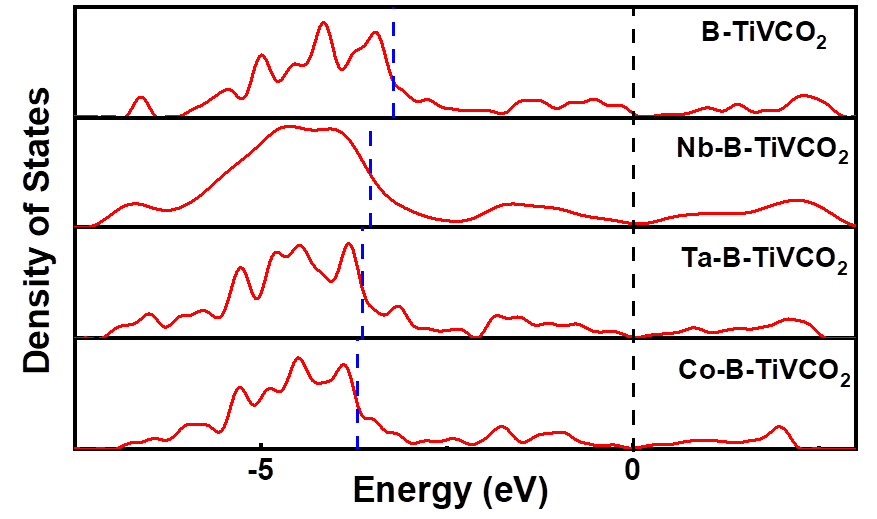


图 8 TM-B-TiVCO2态密度图

综上所述，我们采用能量描述符ΔGH对HER反应进行高通量筛选，ΔGH能够很好的反映催化活性的好坏，避免了计算整个反应路径，大大降低了计算时间和成本。通过以ΔGH为标准的高通量筛选，我们筛选出Nb、Ta、Co-B-TiVCO2、Os-N-TiVCO2和Nb-P-TiVCO2五种性能优异的催化剂，与TiVCO2催化剂相比，催化性能大大提高，并通过电子结构分析，解释了性能提高的内在原因，为后续催化剂的设计提供指导和帮助。

## 2.2氢能关键材料量子化学数据库

本工作提供完整的数据库构建流程，在这个过程中将量子化学数据库分为计算数据库、文献库和材料数据库。对于计算数据库，我们对TiVCO2材料进行改性，研究其对HER反应活性的影响，以氢吸附自由能为标准，计算了78种材料的催化活性，最后得到包含催化剂结构信息、电子结构信息和元素特有信息共23种数据。对于文献库，主要利用Web of Science核心合集数据库和zotero文献管理软件提取论文中作者、DOI、期刊、摘要等信息。对于材料数据库，基于文献库所提供的DOI批量下载PDF，然后利用ChatGPT编写代码，提取PDF中的图片以及材料数据等信息。每个部分的流程将在下文进行介绍（图9）。



图 9 数据库构建流程图

### 2.2.1 计算数据库介绍

上节研究了改性TiVCO2对HER反应活性的影响，我们引入26种过渡金属原子和3种B，N，P非金属原子对其进行改性研究，以氢吸附自由能为标准，计算了78种材料的催化活性。基于对改性TiVCO2催化HER反应的研究，我们建立了相关计算数据库，主要包括催化剂结构信息、电子结构信息和元素特有信息共23种。以TM-B-TiVCO2为例，我们对部分计算数据库进行展示（图10）：催化剂结构信息主要为键长等信息；电子结构信息包括p带和d带中心、Bader电荷转移数、费米能级等信息；原子结构特征信息包括元素价电子数n、第一电离能Ei、电子亲合能EA、电负性Χ、原子半径r、原子序数N等信息。后续随着计算体系的不断扩展，计算数据库包含的种类和催化剂类型将随之增加，根据计算体系不断充实计算数据库。

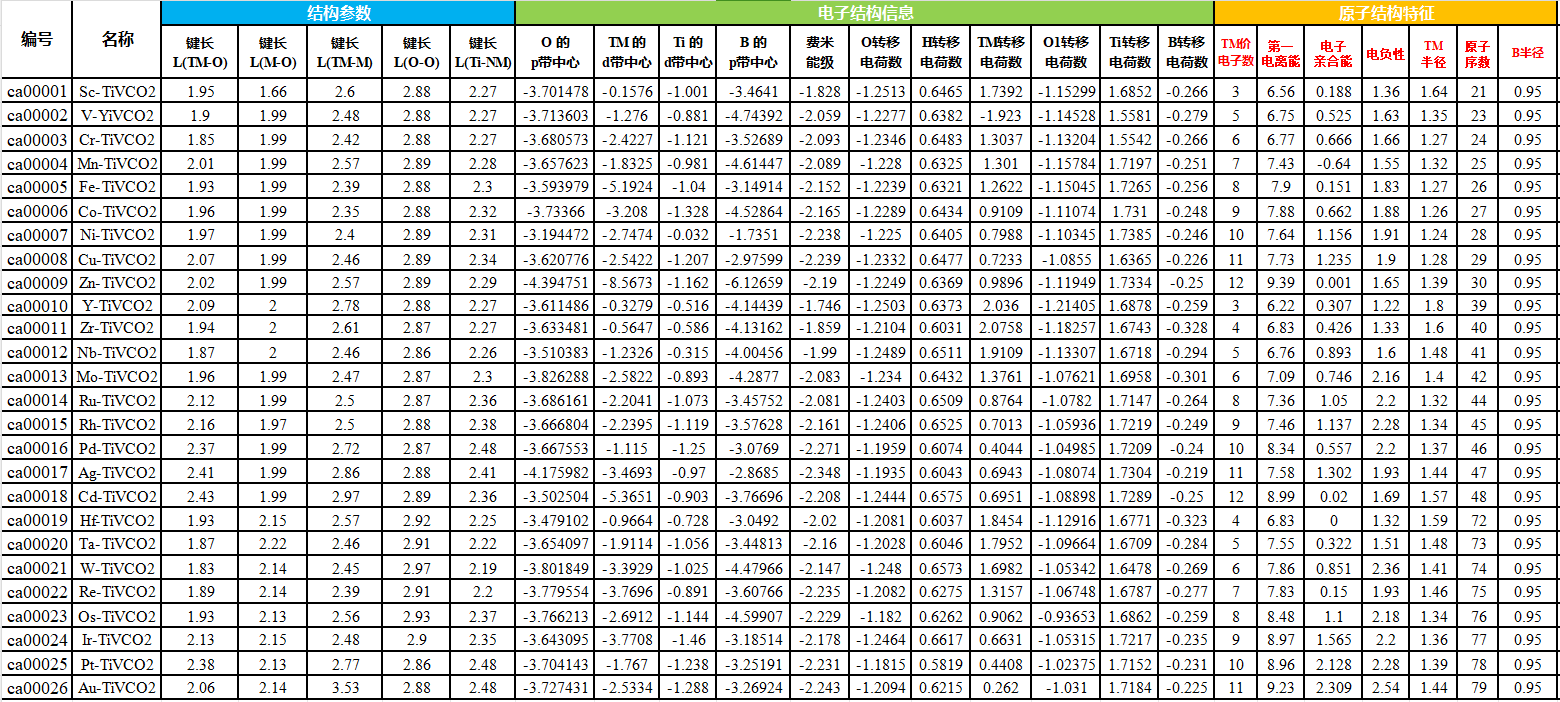


图 10 部分计算数据库展示

### 2.2.2 文献库介绍

（1）文献采集

本工作中所构建的文献库中的文献主要来源于web of science核心合集数据库，根据本项目的研究方向，确定检索主题为：Hydrogen Evolution Reaction、Oxygen Evolution Reaction、Hydrogen Oxidation Reaction、Oxygen Reduction Reaction，关键词之间用or连接，这四种催化反应类型共检索到318441篇，为保证检索文献均为化学材料方向，因此，以Materials Science进行筛选，最后得到112128篇文献。

将得到的112132篇文献以RIS格式导出，得到所有文献（RIS格式）后，全部导入zotero文献管理软件，然后以CSV文献库形式导出，表格中包含文献基本信息，例如题目、期刊、DOI、摘要、作者等。从web of science核心合集数据库导出的112132篇文献中得到共有111099条信息。文献库数据以re000001-re111099进行编码，下图是部分数据展示。（图11）。

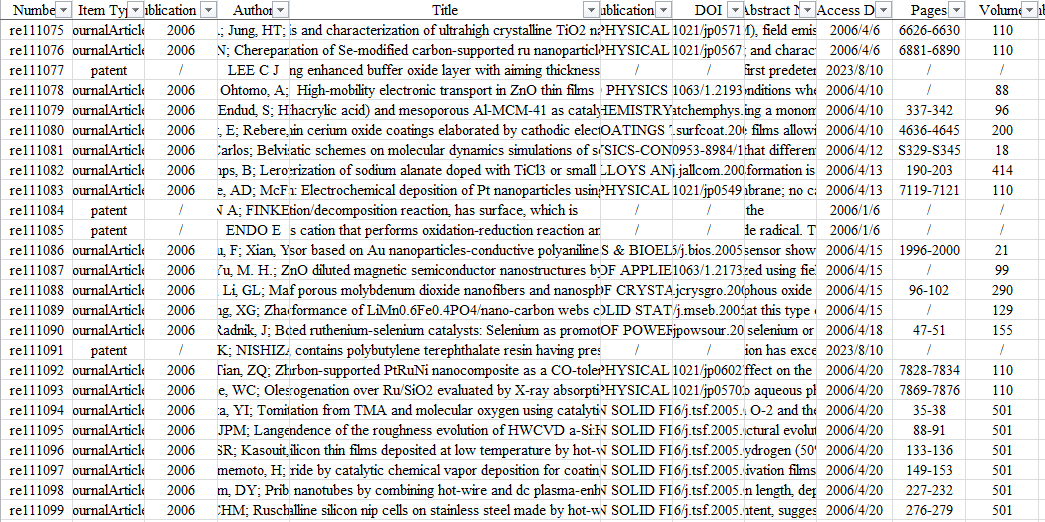


图 11 文献库数据展示

（2）文献处理

该流程的第二步是对文献进行处理，下载所有文献的PDF格式。采用python代码下载，分为以下几步：（a）在导出的CSV表格中提取DOI列到txt文本中；（b）从提供的txt文本中读取DOI列表；（c）逐个DOI访问免费网站下载PDF；（d）将成功下载的PDF保存到一个文件夹；（e）将下载失败的DOI保存到一个名为“failed”的txt文本文件中。对于下载失败的文献，在zotero文献管理软件中下载或者人工去进行下载（图12）。

因文献数量较大，本工作首先将PDF正文下载，后续陆续下载正文所对应的支持信息。



图 12 文献PDF下载流程

### 2.2.3 材料数据库介绍

**（1）数据提取**

一般来说，对材料的描述大多会在摘要中存在。数据库的第一步是将摘要中的材料名字提取。因文献数量较大，将所有的进行编号，而提取的材料名字也和文献编号对应。

流程的第一步是通过关键词在文献中找到所对应的数据。根据对理论、实验文献的阅读了解，对催化剂进行分类，主要分为单原子、双原子、合金、过渡金属二硫化物、过渡金属氧化物、异质结、非金属、Mxene材料等。为了避免数据整理混乱，本工作目前将数据分为量化理论数据、实验数据、量化理论和实验共有数据三类，基于以上分类，实验数据包含：周转频率TOF、金属原子负载量wt%、质量活性A/mgPt、比活性mA/cm2、形貌、电化学活性表面积ECSA cm2、循环次数/稳定性、导电性、表面积、孔体积等；量化理论数据包含：吸附能、热力学（反应能垒eV）、费米能级、带隙、电荷差、电荷转移能等；量化理论和实验共有数据包含：原子性能参数（电离能、电负性、电子亲和能、最外层电子数、原子序数）、结构参数（晶格常数a、晶格常数b、旋转角度、键长、配位数等）、表面性质（活性位点等）、性能参数（电流密度mA/cm2、速率常数、活化能KJ/mol、起始电位（V）、过电势（V）、交换电流密度等（图13）。



图 13 材料数据库关键词分类

**（2）图片提取**

第二步是基于PDF文章，通过python编写的代码进行图片提取。

1. 提取PDF文件中的图像

加载PDF中的每一页，提取页面中的所有图像，并将其保存为PNG文件。保存时，文件名包含了页数和图像在页中的索引。

1. 提取Word文档中的图像

加载Word文档，遍历文档中的所有图像关系。对于每个图像关系，提取图像数据，并将其保存为相应的图像文件。文件扩展名从图像的内容类型中获取。

1. 处理指定目录中的所有文件。

遍历目录中的所有文件，并根据文件扩展名来决定如何处理文件。对于PDF和Word文件，会调用相应的图像提取函数来提取图像。

以上步骤可以在指定的目录中批量处理包含PDF和Word文档的文件，并将其中的图像提取并保存到相应的子文件夹中（图14）。

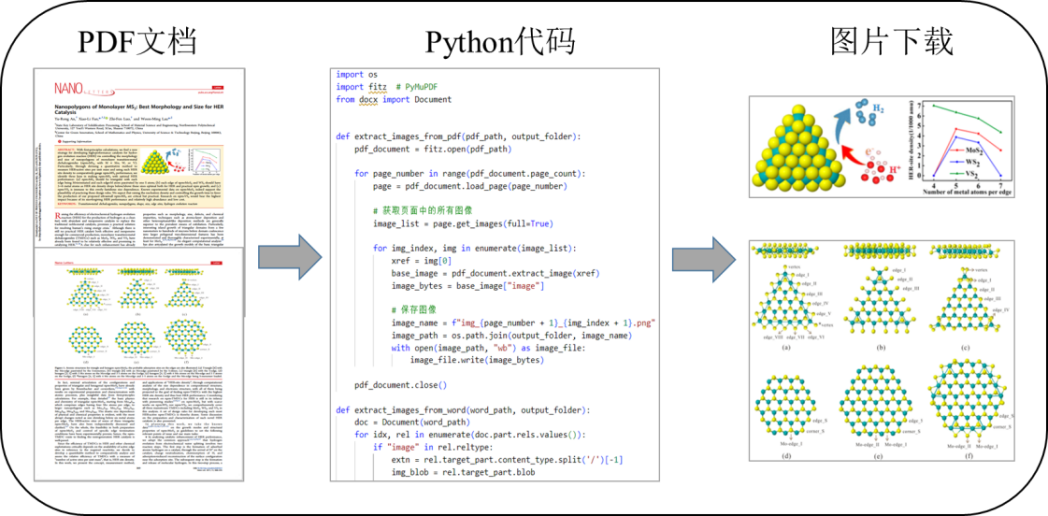


图 14 PDF文献提取图片流程

在量子化学数据库构建方面，我们分为计算数据库、文献库和材料数据库进行构建。在计算数据库中，目前我们搭建78种改性材料，基于对HER反应研究，最后得到23种数据信息。在文献库中，已初步搭建含11万篇左右的文献库，以CSV表格呈现文献中作者、DOI、期刊、摘要等信息，建成11万条左右的文献数据库。在材料数据库中，基于文献库所提供的DOI批量下载PDF，将数据分为实验、量化理论、实验和量化理论共有数据三类，接下来主要通过关键词遍历所有文献，最后利用ChatGPT提取PDF中的图片以及材料数据等信息，最后建成包含10万条左右的材料数据库（图15）。

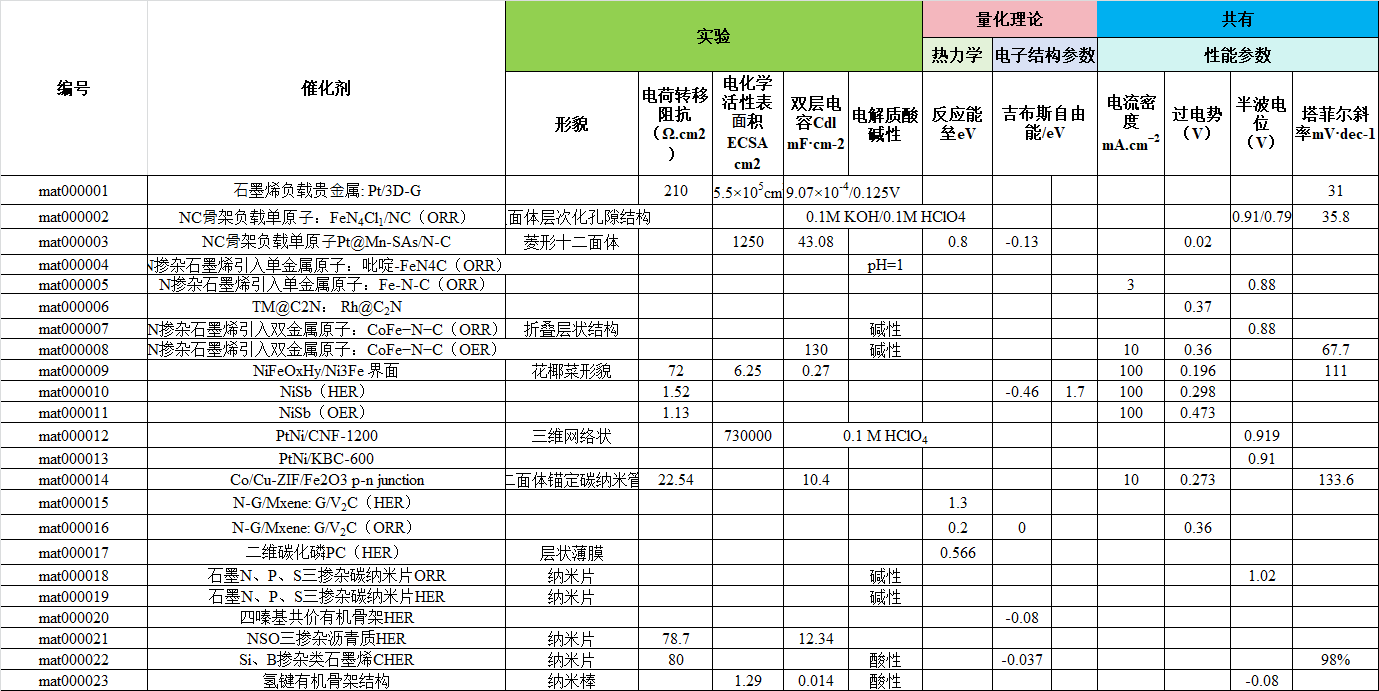


图 15 部分材料数据库展示

2.3 机器学习精准设计

目前寻找新型催化剂的困难之一是候选催化剂的数量太大，传统的试错法效率低下且成本高昂，迫切需要一种更合适的方法来加快这一进程。机器学习（Machine Learning，ML）可以通过各种算法在合理的时间内准确预测团簇形态、自由能、带隙等特性，耦合DFT高通量计算筛选致力于基于计算数据训练机器学习模型用于预测性能优异的催化剂，成为加速新催化剂设计和筛选的有力工具。高通量计算筛选结合机器学习开始在设计过程中发挥关键作用。在TiVC MXenes催化HER反应机理研究课题中，我们使用DFT计算耦合机器学习来加速催化剂的智能设计和筛选。

随着大数据时代的到来，数据成为了驱动机器学习模型训练的关键因素之一。人们普遍认为，拥有更多的数据可以提升机器学习模型的性能和泛化能力。然而，数据量对于模型性能的影响并不总是简单的线性关系。前面建立了材料数据库、计算数据库和文献数据库，虽然目前数量达到了十万条，但预测某种具体材料时还需要进行数据的选取和清洗，以使得数据的质量有所提高，因此利用计算数据库数据特征齐全及计算数据准确的优势，我们将从计算数据库中提取数据作为这次预测新材料性能的初始数据集。

经计算得到了引入非金属原子（B, N, P）和TM原子（3d, 4d, 5d）的一系列的TiVCO2催化材料，积累了一定的数据量，为了加速计算过程，我们利用机器学习来预测引入非金属原子S的TM-S-TiVCO2催化材料的HER性能。图16是机器学习工作流程图，首先计算得到的初始数据集经过特征筛选和降维后代入7种不同的算法进行训练，而采用这么多不同算法的原因就是著名的“没有免费的午餐”定理，即由于对所有可能函数的相互补偿，最优化算法的性能是等价的。因此应用的算法有：K邻近算法（KNN）、二次判别分析（QDA）、决策树（DT）、逻辑回归（LR）、朴素贝叶斯（Naive Bayes）、支持向量机（SVM）和集成算法（XGBoost），最后将测试集代入训练好的模型中再对模型进行评估筛选出最优模型。其中XGBoost算法是一种基于梯度提升树的算法，因此首先要了解梯度提升树的原理。梯度提升树是一种集成学习方法，通过逐步迭代地训练一系列弱学习器（通常是决策树），每一次迭代都尝试纠正前一次迭代的误差，最终将这些弱学习器组合成一个强学习器，由于具有这种强学习器的优势是目前研究工作中使用较多的算法之一，下面是我们导入材料数据库XGBoost模型训练的代码演示（图17）。

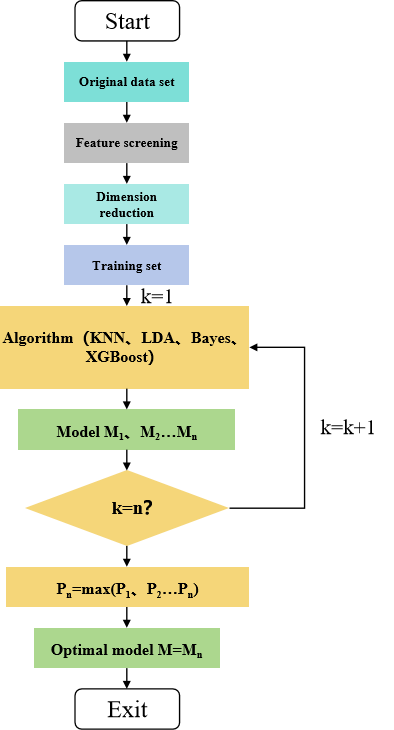


图 16 机器学习流程图



图 17 XGBoost代码展示

### 2.3.1关键特征选取

关键特征的选择和设计是ML模型开发的关键。所选的特性应该是独特的、快速的和易于访问的。基于这些原则，我们选择了23个主要特征，如表1所示。它们被分为两大类，DFT计算特征和元素自有特征。在计算特征中需要经过一次DFT计算的有L（TM-O）、 L（TM-M）、L（OH-M）、 L（O-OH）和L（Ti-NM）这5个特征，而其余的则需要通过第二次DFT计算才能得到。元素特征中的7个特征则是材料对应元素的本征属性不需要计算就可获取。

表1 TM-NM-TiVCO2的DFT计算和元素自有特征

|  |  |  |  |
| --- | --- | --- | --- |
|  | DFT calculated features |  | Elemental features |
| L（TM-O）  L（TM-M） | Average bond length between doping TM and the nearest oxygen atoms/metal atom | nTM | Valence electron of the doping TM atom |
| L（OH-M） | Average bond length between oxygen atom adsorbing hydrogen and the nearest metal atoms | Ei | First ionization energy of the doping TM atom |
| L（O-OH） | Average bond length between oxygen atom bonded to TM and the oxygen atom adsorbing hydrogen | EA | Electron affinity of the doping TM atom |
| L（Ti-NM） | Bond length between Ti atom and B/N/P | χ | Electronegativity of the doping TM atom |
| pO, pB/N/P | The p-band center of the O and B/N/P | rTM rB/N/P | Atomic radius of the doping TM atom and B/N/P |
| dTM | The d-band center of the TM | N | Atomic number of the doping TM atom |
| dTi | The d-band center of the Ti |  |  |
| EF | Fermi level of adsorbed hydrogen structure |  |  |
| qB/N/P, qO, qH, qTM | Bader charge transfer of B/N/P, O, H and TM |  |  |
| qO-TM | The average bader charge transfer of O connected to the TM |  |  |
| qTi | Bader charge transfer of Ti closest to the TM |  |  |

在计算了掺入B和N元素的催化材料后，有56个材料且特征量为23个的数据集，为了加速氢能关键催化材料的筛选过程，去除冗余的特征是必不可少的一步，因此为了了解各个特征之间的相关性作了皮尔逊相关系数（PCC）图，如图18a所示。当|PCC| >0.8时，说明两者存在较强的相关性14，例如：PCCqB/N/P-L（Ti-NM），PCCrB/N/P-L（Ti-NM），PCCrB/N/P-qB/N/P，PCCqB/N/P-L（O-OH），PCCrB/N/P-L（O-OH），PCCnTM-EF，PCCχ-qO-TM和PCCnTM-Ei，代表着在一定情况下这些特征可以相互替代，同样从图18b中看出相关性强的两个特征的数据分布也是相似的。图18c是对这23个特征进行了归一化的重要性排序（XGBoost算法），并且加入了时间因子，其中N为该特征所需DFT计算的次数，并对前十个的特征进行了标注。为了进行特征降维，综合考虑特征的相关性以及排序，rB/N/P和L（Ti-NM）具有很强的正相关性（PCC=0.92），而L（Ti-NM）则需要进行一次DFT计算，为了加速整个筛选过程，仅保留rB/N/P即可。通过这种方式我们确定了在前十的范围内的6个元素自有特征，这些特征分别是nTM、Ei、EA、χ、rTM和N，最终保留了7个特征作为接下来机器学习训练的数据集特征。

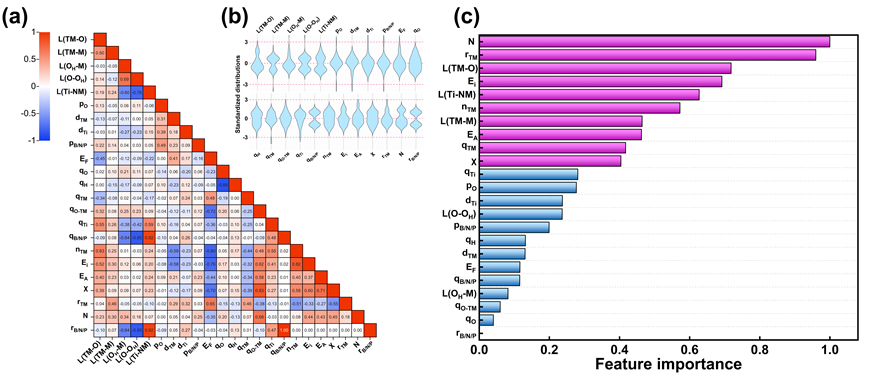


图 18相关性及特征重要性排序

### 2.3.2 机器学习模型训练

对于HER催化材料来说，吸附能是衡量材料催化性能的指标之一。为了尽可能全面地预测具有良好HER性能的TM-S-TiVCO2材料，我们将在-0.3~0.3 eV范围内预测∆GH的材料。我们将DFT计算的78种TM-B-TiVCO2，TM-N-TiVCO2和TM-P-TiVCO2催化材料作为源数据，并将7种元素自有特征作为特征值，共546个数据进行训练。我们采用分类算法，将∆GH在-0.3~0.3 eV范围内分类为0，表示催化剂具有优异的HER性能，在-0.3~0.3 eV范围外分类为1，表示材料催化性能较差。根据图16中的机器学习流程图，我们将其放入不同的模型中进行训练，得到的结果如图19（a~g）所示，对应XGBoost、DT、QDA、LR、Bayes、SVM、KNN等不同算法模型的混淆矩阵。由图可知，XGBoost算法模型的预测准确率最高，为93.6%。另外，图19h可以清楚地看到，曲线AUC达到0.9264的高值。因此，XGBoost算法模型是预测TM-S-TiVCO2材料HER性能的最佳选择。

### 2.3.3 机器学习模型预测及DFT验证

根据TM-S-TiVCO2的7个特征，采用训练好的XGBoost模型对其ΔGH进行预测。筛选出Zr-S-TiVCO2、Nb-S-TiVCO2、Hf-S-TiVCO2、Ta-S-TiVCO2和W-S-TiVCO2五种性能优异的催化剂。为了验证得到的结果，我们计算了这五种材料的吸附能，都在-0.3~0.3 eV的范围内，表明它们具有良好的HER催化性能，XGBoost模型具有较高的准确度（表2）。其中Hf-S-TiVCO2的ΔGH最低，为-0.21eV，具有优异的HER性能。

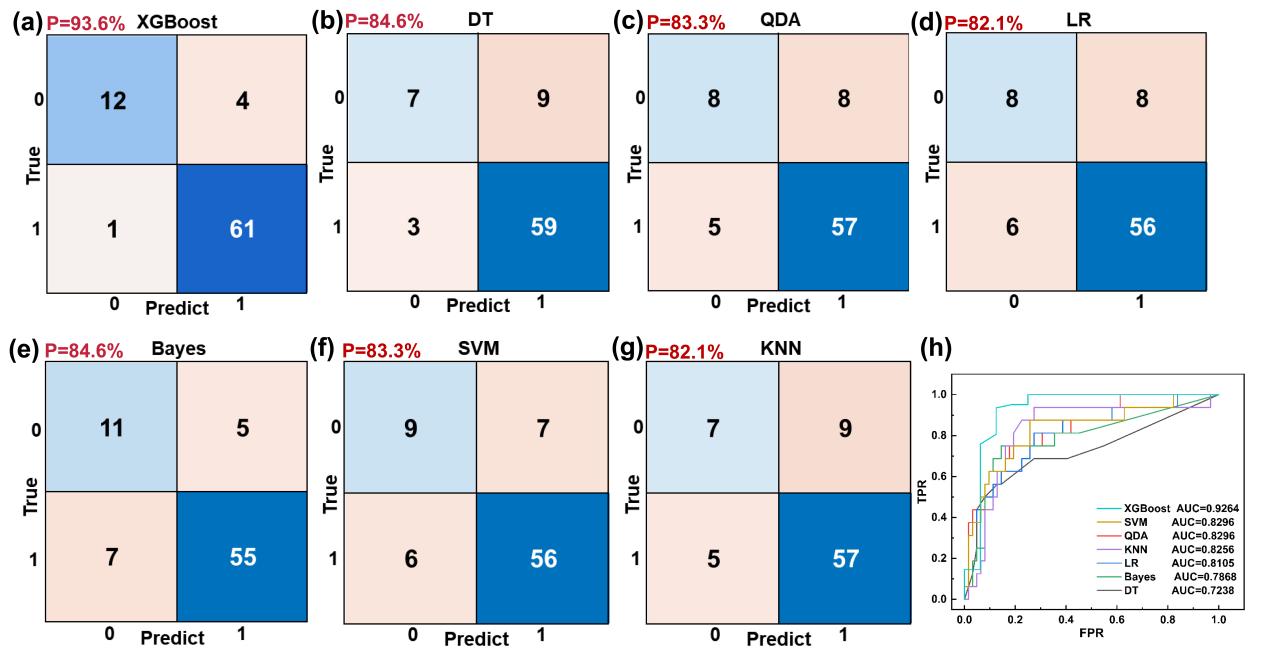


图 19 不同算法模型的混淆矩阵及ROC曲线

表2 XGBoost模型预测和计算验证

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ∆GH （eV） | Material（TM-S-TiVC） | | | | |
| Zr | Nb | Hf | Ta | W |
| prediction | （-0.3~0.3） | （-0.3~0.3） | （-0.3~0.3） | （-0.3~0.3） | （-0.3~0.3） |
| calculation | -0.2493 | -0.2635 | -0.2179 | -0.241 | -0.2556 |

综上所述，我们将DFT计算结合ML模型，对非金属和金属掺杂的TiVCO2催化剂进行HER性能研究，筛选出Nb-B-TiVCO2、Ta-B-TiVCO2、Co-B-TiVCO2、Os-N-TiVCO2、Nb-P-TiVCO2和Hf-S-TiVCO2六种性能优异的催化剂，大大加速了催化剂的筛选过程，并且具有较高的准确度，显著提升能源材料的理论研发效率，为后续催化剂的设计提供新的思路和策略。

# 3 结论

在建立氢能关键催化材料催化反应性能的理论评估方法方面，我们以TiVC-MXenes催化HER反应机理研究为例，采用能量描述符ΔGH对78种改性材料进行HER性能计算，避免计算整个反应路径来实现催化剂的高通量筛选。筛选出Nb-B-TiVCO2、Ta-B-TiVCO2、Co-B-TiVCO2、Os-N-TiVCO2、Nb-P-TiVCO2五种性能优异的材料。并得到催化剂结构特征，电子结构特征和元素特有特征共23种，1196个数据作为机器学习的源数据。

在量子化学数据库构建方面，我们分别搭建了计算数据库、文献库和材料数据库。在计算数据库方面，搭建了约1196个数据的计算数据库，后续会根据计算进度逐步增加；在文献库方面，已初步搭建含111099篇文献的文献库，以CSV表格呈现文献中作者、DOI、期刊、摘要等信息。在材料数据库方面，基于文献库所提供的DOI批量下载PDF，将数据分为实验、量化理论、实验和量化理论共有数据三类，利用ChatGPT提取PDF中的图片以及材料数据等信息，建成含112000个材料的材料数据库。

在机器学习精准设计方面，探索DFT计算结合ML模型已初具成效，后续将继续采用机器学习技术进行数据挖掘，利用多种机器学习算法（分类、决策树、线性回归、神经网络等）分析反应位点结构特性对活性影响的核心因素，建立材料成分、工艺、微结构、性能之间的关联模型，为后续实现数据库的交互式、可视化设计，建立高效率智能设计平台奠定基础。

在这一阶段，十万条数据库已初步搭建完成，文章已完成，正在投稿中，年度考核指标已全部完成。在本年度内，项目成员每半个月开展一次交流讨论会，保证课题进度的正常进行，定期与各课题成员讨论交流，后续将继续保持该工作模式，以确保课题的正常进行。

# 参考文献

[1] VahidMohammadi, A.; Rosen, J.; Gogotsi, Y., The world of two-dimensional carbides and nitrides (MXenes). **2021,** *372* (6547), eabf1581.

[2] Tang, X.; Zhou, D.; Li, P., et al., MXene-Based Dendrite-Free Potassium Metal Batteries. **2020,** *32* (4), 1906739.

[3] Jiang, Y.; Sun, T.; Xie, X., et al., Oxygen-Functionalized Ultrathin Ti(3) C(2) T(x) MXene for Enhanced Electrocatalytic Hydrogen Evolution. *ChemSusChem* **2019,** *12* (7), 1368-1373.

[4] Anand, R.; Nissimagoudar, A. S.; Umer, M., et al., Late Transition Metal Doped MXenes Showing Superb Bifunctional Electrocatalytic Activities for Water Splitting via Distinctive Mechanistic Pathways. *Advanced Energy Materials* **2021,** *11* (48).

[5] Anasori, B.; Lukatskaya, M. R.; Gogotsi, Y., 2D metal carbides and nitrides (MXenes) for energy storage. *Nature Reviews Materials* **2017,** *2* (2), 16098.

[6] Gan, J.; Li, F.; Tang, Y., et al., Theoretical Study of Transition-Metal-Modified Mo(2) CO(2) MXene as a Catalyst for the Hydrogen Evolution Reaction. *ChemSusChem* **2020,** *13* (22), 6005-6015.

[7] Pandey, M.; Thygesen, K. S., Two-Dimensional MXenes as Catalysts for Electrochemical Hydrogen Evolution: A Computational Screening Study. *J. Phys. Chem. C* **2017,** *121* (25), 13593-13598.

[8] Gao, Y.; Cao, Y.; Gu, Y., et al., Functionalization Ti3C2 MXene by the adsorption or substitution of single metal atom. *Applied Surface Science* **2019,** *465*, 911-918.

[9] Gu, Y.; Wei, B.; Legut, D., et al., Single Atom‐Modified Hybrid Transition Metal Carbides as Efficient Hydrogen Evolution Reaction Catalysts. *Adv. Funct. Mater.* **2021,** *31* (43).

[10] Ling, C.; Shi, L.; Ouyang, Y., et al., Searching for Highly Active Catalysts for Hydrogen Evolution Reaction Based on O-Terminated MXenes through a Simple Descriptor. *Chemistry of Materials* **2016,** *28* (24), 9026-9032.

[11] Zhang, X.; Chen, L.; Sun, Q., et al., Modulating oxygen electronic orbital occupancy of Cr-based MXenes via transition metal adsorbing for optimal HER activity. *International Journal of Hydrogen Energy* **2021,** *46* (50), 25457-25467.

[12] Shen, Z.; Fan, X.; Ma, S., et al., 3d transitional-metal single atom catalysis toward hydrogen evolution reaction on MXenes supports. *International Journal of Hydrogen Energy* **2020,** *45* (28), 14396-14406.

[13] Zheng, J.; Sun, X.; Qiu, C., et al., High-Throughput Screening of Hydrogen Evolution Reaction Catalysts in MXene Materials. *J. Phys. Chem. C* **2020,** *124* (25), 13695-13705.

[14] Goltl, F.; Müller, P.; Uchupalanun, P., et al., Developing a descriptor-based approach for CO and NO adsorption strength to transition metal sites in zeolites. **2017,** *29* (15), 6434-6444.