Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Material Science and Engineering—Article

科学中的第五范式——以智能驱动的材料设计为例

冷灿[a,b,c]，唐卓[c,d,*]，周一歌[e]，田泽安[d]，黄维清[f]，刘杰[a,b]，李克勤[d,g]，李肯立[c,d,*]

[a] *Science and Technology on Parallel and Distributed Processing Laboratory, National University of Defense Technology, Changsha 410073, China*
[b] *Laboratory of Software Engineering for Complex Systems, National University of Defense Technology, Changsha 410073, China*
[c] *National Supercomputing Center in Changsha, Changsha 410082, China*
[d] *College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China*
[e] *Institute of Chemical Biology and Nanomedicine, State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China*
[f] *Department of Applied Physics, School of Physics and Electronics, Hunan University, Changsha 410082, China*
[g] *Department of Computer Science, State University of New York, New Paltz, NY 12561, USA*

ARTICLE INFO

Article history:
Received 8 December 2021
Revised 6 June 2022
Accepted 29 June 2022
Available online 14 April 2023

关键词
催化材料
第五范式
智能驱动
机器学习
跨学科专家的协同效应

摘要

材料科学研究正在进入“机器学习+大数据”为标志的数据驱动范式阶段，预示着以机器学习为代表的智能系统融入传统的材料科学计算，具备数据挖掘和知识发现的智能驱动能力。在此，本研究通过在天河一号超级计算机系统上构建的为催化材料专门设计的典型平台案例，生动地阐明了第五范式的本质，旨在促进第五范式在其他领域的发展。第五范式平台主要包括模型自动构建（原始数据提取）、指纹自动构建（神经网络特征选择）以及跨学科知识串联的重复迭代（“火山图”）。与分解一起进行的是对迭代中实现的体系结构的性能评估。通过讨论，第五范式的智能驱动平台可以极大地简化和改进研究中极其繁琐和具有挑战性的工作，并通过补偿机器学习中样本的不足，以及替代一些计算资源不足导致的数值计算，实现数值计算与机器学习的相互反馈，加快探索过程。跨学科专家的协同作用和对实时数据需求的急剧增长仍然是一个挑战。我们相信，对第五范式平台的关注可以为其在其他领域的应用铺平道路。

1. 引言

人类社会发生翻天覆地的变化，离不开对自然的无穷探索。这种变革性的发展已经从自然观察演变为通过各种工具和前沿方法逐渐实现[1–2]，并逐渐形成涵盖各个学科整体和相互关联的不同规范发展范式[3–4]。每一次范式转移都是由统治理论内部的基本假设在一定时代内为了适应后续的要求而发生变化导致的，从而产生新的范式[5]。第五范式现在被描述为智能驱动、以知识为中心的研究范式，紧随数据密集型第四范式的转变，紧随实验、理论和计算机模拟范式从第一范式向第三范式的转变[6–10]。

对第五范式来说，对物理宇宙的探索不仅仅是由智力驱动的密集数据的数学可能领域所投射出来的，而且整个研究过程也涉及人类专业知识的无差别意识过程。基于这些特征，第五范式的应用可以被视为一种认知系统或认知应用[9–10]。以材料科学的发展为例，第五范式的认知系

* Corresponding authors.
E-mail addresses: ztang@hnu.edu.cn (Z. Tang), lkl@hnu.edu.cn (K. Li).

英文原文：Engineering 2023, 24(5): 126–137
引用本文：Can Leng, Zhuo Tang, Yi-Ge Zhou, Zean Tian, Wei-Qing Huang, Jie Liu, Keqin Li, Kenli Li. Fifth Paradigm in Science: A Case Study of an Intelligence-Driven Material Design. *Engineering*, https://doi.org/10.1016/j.eng.2022.06.027

统是通过经典的螺旋进化过程从原始的早期范式演变而来。在牛顿定律和相对论出现之前，金属和陶瓷等材料在古代就已经被发现和使用。然后，相对论和量子力学的出现使得模拟分子的电子结构成为可能[11–13]。近年来，人工智能（AI）和机器学习的迅速兴起促进了数据驱动材料设计的研究[14–18]。因此，通过将相关创新技术加工成越来越大的数据集，可以找到金属和陶瓷等新材料的隐藏特性[19–22]。特别是当下基于智能驱动的认知材料研究，接过数据密集型材料的接力棒形成了一种新的发展趋势，进一步加快材料科学的探索进程。

目前，第五范式正处于萌芽期，还有很长的路要走。材料领域的智能驱动方法伴随着数据密集型科学研究范式的发展而被逐渐应用到材料创新中。随着成熟的数据密集型科学研究范式在多个领域迅速暴发，有关自动驾驶汽车、计算机视觉和大脑建模等工业和科学等领域的应用技术被广泛开发并逐步实现[23–27]。以知识为中心的第五范式仍处于蓬勃发展阶段，与之不同的是，数据智能驱动在认知应用中需要打破传统科学计算和数据密集型研究的界限，通过融合和扩展现有技术形成新的生态系统。一些科学家正逐步针对这一需求开展研究，例如，Malitsky等[10]提出的MPI（Spark-message passing interface）集成平台，可用于推动数据密集型应用向智能驱动方向转变；Zubarev和Pitera [9]研究的认知计算，如自然语言处理、知识表征和自动推理，能够促进认知特征面向智能驱动的实现。在材料研究领域，数据智能驱动需要基于不同领域专家知识的整合，以及来自实验观察和理论模拟的大量数据整合，以此推断面向跨学科应用之间的共同属性，设计出更多研究方案以促进材料创新的发展。因此，尽管第五范式的发展任务艰巨，但其应用前景十分广阔。

从数据密集型科学向复合认知计算应用的第五范式的战略转变是一个长期的过程，有许多未知因素。本文通过解析催化材料（https://github.com/ulissigroup/GASpy）[28]中称为Python广义吸附模拟（GASpy）的框架来解决第五范式平台，旨在将人类智慧和高性能计算中的算法和深度学习方法结合起来，以解决数据驱动应用的新领域。本文的其余部分组织如下：第2章提供了对第五范式平台的简要概述和讨论，第3章进一步阐述了平台的性能评估，第4章总结了本文工作。

2. 第五范式平台

在材料研究过程中，实验数据、理论模型和机器学习的协同作用，依赖于不同领域的专家协同分析和处理数据，需要巨大的人类智慧，这些环节均可以通过智能驱动实现。将各个环节的通用结构相结合，实现智能驱动就显得尤为重要。本章介绍一个在催化材料领域使用的第五范式平台，如图1所示。第五范式平台与第三范式和第四范式平台耦合，后两者包括第一范式和第二范式的过程。其中，原始数据来自第一范式的实验观察和第二范式的理论指导，以及第三范式的数值计算，然后可以通过第四范式的机器学习进行智能驱动。结合实验专家和理论专家的工作集成知识，可以对机器学习选择的材料进行第二次筛选，筛选结果被再次反馈到第三范式的数值模拟中。在第三范式中获得的结果仍然可以由第四范式中的数据驱动。然后，通过实验专家和理论专家的知识整合，对预测结果进行再次过滤，再反馈给第三范式进行数值模拟。这些方法产生了第五范式平台。通过智能控制高通量物理模型的计算，不断为机器学习提供样本，以弥补机器学习样本的不足。此外，利用整合到不同领域的知识，机器学习可以代替部分数值计算，解决由于计算资源不足而导致的大量模型耗时的问题。

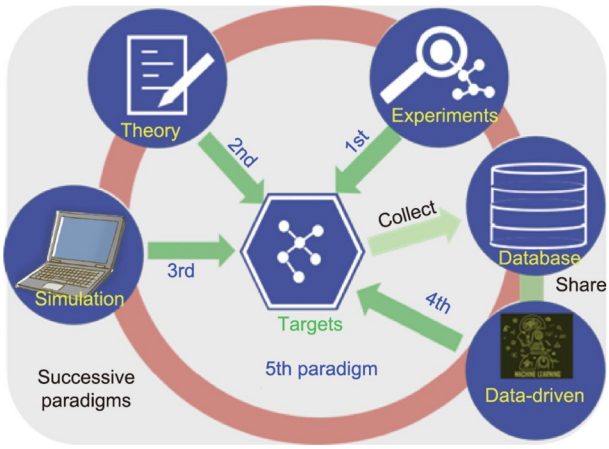

图1. 科学中的范式。科学范式的演变已经从简单的第一范式发展到复杂的第五范式。第五范式的核心是以知识为中心和智能驱动，包括从第一范式、第二范式、第三范式到第四范式，分别以实验、理论、模拟和数据驱动过程为标志。

第五范式平台的综合工作源于Tran和Ulissi [28]为材料科学双金属催化剂研究设计的框架，该框架使用机器学习来加速基于密度泛函理论（DFT）的数值计算，该计算由维也纳大学Hafner小组开发的模拟包（VASP）[29]进行，可以推动高性能电催化剂的发现。该平台可以对双金属晶体各稳定低折射率表面的活性位点进行分类，得到成百上千个可能的活性位点。同时，采用基于人工神经网络的替代模型来预测这些位点的催化活性[30]。发现的高活性位点可以进一步用于未来的DFT计算。

2.1. 自动模型构建和验证

智能驱动原始数据提取的能力体现在模型的自动构建中，由第五范式平台验证。可以自动构造有或没有吸附物的更大结构并通过DFT计算对该结构进行验证。由于表面物质的吸附是多相催化不可缺少的过程，在通过评估吸附能来确定催化活性之前，在实验和DFT计算中构建许多结构可能会耗费大量时间。因此，自动化模型构建和验证对于解决问题至关重要。

如图2所示，整个任务计算包括标准模拟原始数据的准备过程，然后进行数值计算。理论模拟所使用的原始数据全部来自Material Project网站，通过定制模块Generate_Gas/Generate_Bulk实现，并将用户信息、任务位置、计算状态和其他属性通过键值对存储，创建名为“Fireworks”“Atoms”“Catalog”和“Adsorption”的集合。

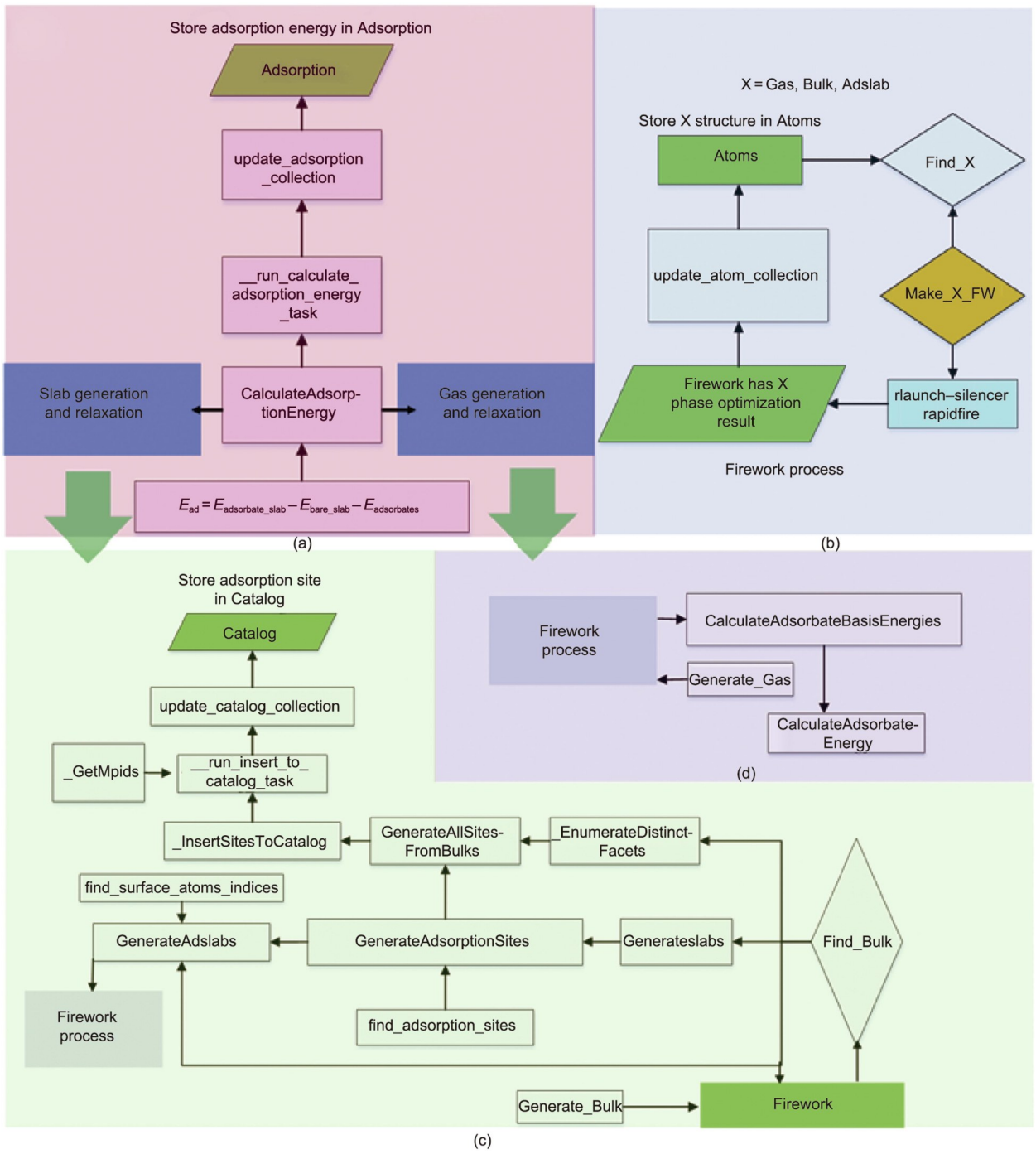

图2. 第五范式框架示例。通过原子运算、生成和计算模块，实现了GASpy框架下原始数据提取的智能驱动。（a）该模块的功能是自动计算第五范式平台中气体和板状相的吸附能。（b）该模块用于通过Firework自动创建高通量任务，用于优化有/无吸附质的Gas、Bulk、Adslab。（c）、（d）模块表示部分（a）中描述的Slab生成（c）和Gas生成以及结构弛豫（d）。

然后，FireWorks工作流管理器可以生成图2（a）中任务的松弛计算，以便在图2（b）提交。FireWorks中的结果属性包含“gasphase optimization”作为气体松弛的列表格式，“gasphase optimization”用于批量优化（bulk_relaxation）。属性“status”是“COMPLETED”“RUNNING”“READY”和其他状态（如“FIZZLED”等）的计算状态，由Find_Bulk/Find_Gas函数判断，以将完成的计算过程存储在Atoms集合中，或者生成等待尚未开始的计算的FireWorks任务工作流。

如果Find_Bulk/Find_Gas确定的状态是“COMPLETED”，则把计算结果存储到数据库中，再从Atoms集合中获取优化后的晶体结构，进行不可约晶面指数枚举（通过EnumerateDistinctFacets函数实现），再根据给定米勒指数，通过扩胞（Atom_operates的函数），枚举晶体表面slab，添加吸附物，从而找到切面的所有吸附位点（由GenerateAdsorptionSites函数实现），如图2（c）和（d）所示。对于所有材料上的指定米勒指数切面吸附位点，可通过由EnumerateDisdinctFacets函数和GenerateAdsorptionSites函数组成的GenerateAllSitesFromBulks函数进行遍历并生成所有吸附位点。所有这些信息都由函数update_catalog_collection写入Catalog集合。

对于每个找到吸附位点的切面，通过GenerateAdslabs函数将吸附物添加到吸附位，生成“slab + adsorbate optimization”计算模型（adslab_relaxation）。还可以通过GenerateAdslabs函数去掉吸附物，生成“bare slab optimization”计算模型（bare_slab_relaxation）。然后可以通过FireWorks工作流管理器提交这些计算模型进行计算。

完成后，所有计算结果将通过函数update_atom_collection存储到Adsorption集合中。Find_Adslab函数将通过查找Atoms集合中是否存在相应的计算结果来确定是否应该再发射对应的DFT任务。对于吸附能量E_{ad}计算，CalculateAdsorptionEnergy函数用于从Atoms集合中提取吸附物能量$E_{adsorbates}$、含吸附物的切面能量$E_{adsorbate_slab}$和不含吸附物的切面能量E_{bare_slab}，利用update_adsorption_collection函数将E_{ad}和相关的初始及最终结构等其他信息添加到Adsorption集合中，构成下一步机器学习提取标记指纹的数据集来源。以上过程为整个计算流程的实现。

2.2. 自动指纹构建

神经网络特征选择的智能驱动性体现在第五范式平台的自动指纹构建中。在该框架中，自动构建的指纹由每个材料吸附模型的所有原子结构转换为卷积神经网络（CNN）数值输入的图形表示[31]。在原子结构信息中，考虑了三种类型的特征，如图3所示，即原子特征（F_{N1}）、邻居特征（F_{N2}）和连接距离（F_{N3}）。原子特征中的基本原子性质有原子序数、电负性、配位数/共价半径、基团、周期、价电子、第一电离能、电子亲和、块积、原子体积。基本邻域特征性质由吸附位点附近相邻原子之间的配位数组成，由Voronoi多面体算法计算得到[32]。连接距离是指从吸附物到所有原子的距离。目标指纹图谱为吸附能（E_{adN}）。

自动指纹构建过程包括通过DFT计算提取最终结构和吸附能的过程、指纹生成过程、机器学习过程以及学习

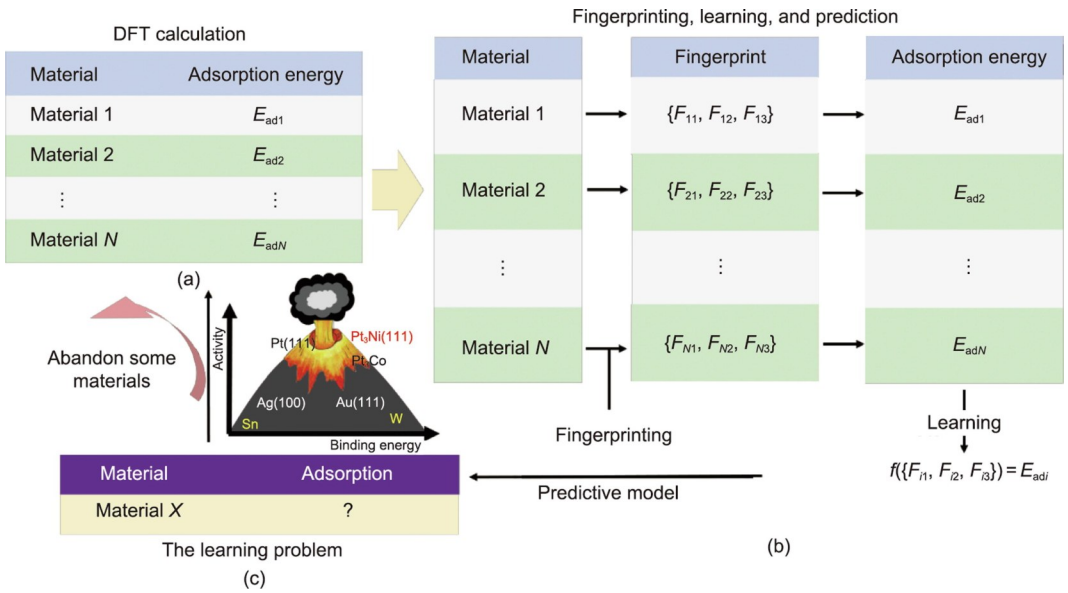

图3. 第五范式平台中神经网络特征选择的智能驱动。它是在GASpy框架下通过自动指纹构建来实现的。（a）DFT计算被示意性地视为示例数据集（N是训练示例的数量）；（b）通过指纹识别和学习步骤过程，利用预测模型实现自动指纹构建；（c）陈述了学习问题，然后通过比例关系从学习结果中放弃一些材料，并进行进一步的DFT计算筛选。

问题的陈述。GASpy中构建的指纹来源于原始模型，没有DFT计算，也没有DFT计算结果。首先，经过DFT计算，得到初始标记指纹E_{adN}作为代理模型的训练集来训练模型，即从原始结构模型中提取指纹$\{F_{N1}, F_{N2}, F_{N3}\}$进行学习和预测，获得训练模型$f$。此处的三个指纹描述符以原子序数、泡利电负性、配位数为代表，整个实现过程如图3（a）和（b）所示。这些特征将被用作机器学习中的验证数据集。接着，函数f将对未标记的指纹进行下一次预测，如图3（c）所示，再通过主动学习过程来选择最有效的目标标注指纹。这个学习问题是由著名的缩放关系[33–34]决定的，如图3所示。标度关系是吸附能-催化活性（又称结合能-催化活性）曲线，像一座先升后降的火山，又称“火山图”（volcano plot）。

火山标定关系中的吸附能和催化活性数据来自理论和实验科学家多次尝试的工作，用来作为特征工程中数据清理的标准。以火山标定关系为依据，图3（c）中描述的预测材料将被进一步利用，意味着火山标定关系不匹配的预测吸附能材料将被丢弃。在下一个周期，被利用的候选者将通过DFT再次计算以增加数据集。随着DFT计算的材料类型的增加，数据集的数量在增加，自动化的探索过程使指纹数量不断得到更新。

2.3. DFT计算和机器学习的理论模型

在第五范式平台中，Kohn-Sham理论与一种集成CNN和高斯过程（GP）的方法[31,35–37]是DFT和机器学习过程的核心理论模型。因此，简要介绍这些理论模型的细节。

2.3.1. DFT计算的理论模型

DFT计算是第一性原理计算的典型代表。通过DFT计算获得吸附能是目前研究材料结构的主要手段之一。吸附能计算过程主要涉及通过不断调整原子和电子结构来达到最终能量稳定的结构状态的每个面结构的优化过程。通过基于量子力学近似求解多体薛定谔方程（many-body Schrödinger equation）求解Kohn-Sham方程的DFT方法是该近似解的主要方法之一。

Kohn-Sham方程为：

$$E[n(r)] = T[n(r)] + \int v(r)n(r)\mathrm{d}^3r + E_{\mathrm{xc}}[n(r)] + \frac{\mathrm{e}^2}{2}\iint \frac{n(r)n(r')}{|r-r'|}\mathrm{d}^3r\mathrm{d}^3r' \quad (1)$$

$$n(r) = \sum_{i=1}^{k}|\psi_i(r)|^2 \quad (2)$$

$$T[n(r)] = \sum\psi_i^*(\boldsymbol{r})(-\frac{\hbar^2}{2m}\nabla^2)\psi_i(\boldsymbol{r})\mathrm{d}^3r \quad (3)$$

$$E_{\mathrm{xc}}[n(\boldsymbol{r})] = \int n(\boldsymbol{r})\varepsilon_{\mathrm{xc}}[n(\boldsymbol{r})]\mathrm{d}^3r \quad (4)$$

给定一个包含K个离子的体系，即三维坐标空间r中的K个已占轨道，其中$\psi_i(r)$为离子i的波函数，其坐标为r，其共轭波函数为ψ^*。$n(r)$是局部电子密度，即在离子i中找到r中的电子的概率。$E[n(r)]$是整个系统的能量。$\hbar$是普朗克常数，m是粒子的质量。$\varepsilon_{\mathrm{xc}}[n(\boldsymbol{r})]$是具有局域电子密度$n(r)$的均质电子气体的交换相关能。$E_{\mathrm{xc}}[n(\boldsymbol{r})]$是交换和相关能。例如，局域电子密度近似是交换相关函数之一，它只以均匀电子气体密度为变量，而广义梯度近似法则以电子密度和密度梯度为变量。$v(r)$为离子i在r位置的势能。因此式（1）中第一项$T\left[n(r)\right]$为动能，第二项$\int v(r)n(r)\mathrm{d}^3r$为外电位。式（1）中最后一项为Hartree能量（电子-电子斥力），其中r'为相对于r的坐标摄动，$\boldsymbol{r}$表示r的矢量。∇为矢量微分算子，∇^2为坐标求导的拉普拉斯算子。

自洽迭代过程描述如下：

给定一个任意$\psi_0(\boldsymbol{r})$的初始电子密度$n(\boldsymbol{r})$，

$$n(r) = \sum_{i=1}^{\mathrm{occ.}}\psi_0^*(r)\psi_0(r) \quad (5)$$

式中，occ.代表已占轨道的数量，于是

$$H[n_0(r)]\psi_1(r) = \varepsilon\psi_1(r) \quad (6)$$

式中，H代表波函数ψ的哈密顿量，其能量用ε表示；然后可以获得一个新的电子密度，

$$n_1(r) = \sum_{i=1}^{\mathrm{occ.}}\psi_1^*(r)\psi_1(r) \quad (7)$$

于是

$$H[n_1(r)]\psi_2(r) = \varepsilon\psi_2(r) \quad (8)$$

…

$$H[n_n(r)]\psi_{n+1}(r) = \varepsilon\psi_{n+1}(r) \quad (9)$$

当$\psi_{n+1}(r) - \psi_n(r)$达到收敛标准后，迭代终止。$E_{\mathrm{ad}}$可以通过$E_{\mathrm{adsorbate_slab}}$与$E_{\mathrm{bare_slab}} + E_{\mathrm{adsorbates}}$之间的能差来计算。

2.3.2. 机器学习的理论模型

卷积馈送的高斯过程（CFGP）[37]是一种使用网络卷积层的池化输出，为高斯过程回归器提供特征的方法[38]。该方法通过将晶体图卷积神经网络（CGCNN）和高斯过程（GP）相结合，进行训练，产生均值，实现关于以吸附能为代表的材料性能预测。Chen等[39]、Xie和Grossman [40]将CNN应用于晶体的图形表示之上，用来预测各种特性，并由Back等[31]进一步使用Voronoi多面体[32]收集邻居信息的改进方法，用来预测非均相催化剂表面的结合能（如吸附能）。在CFGP方法中，首先训练一个完整的CNN来创建最终的固定网络的权重，卷积层的所有池化输出都用作GP中的特征，再通过使用这些特

征来训练GP生成对吸附能的平均和不确定性预测。

在CFGP方法中，晶体结构由晶体图G表示，其中表示晶体中原子之间连接的原子和边由具有原子特征和近邻特征信息的节点编码提供，而CNN构建在无向多重图的顶部[40]。由于晶体图的周期性特征，同一对端节点之间允许存在多条边，每个节点 i 可以用一个特征向量$\boldsymbol{v}_i$表示。类似地，每条边$(i,j)_k$可以用特征向量$\boldsymbol{u}_{(i,k)_k}$表示，对应于连接原子i和原子j的第k个键。考虑每个原子特征与邻居之间相互作用的差异，第一个卷积层通过迭代更新原子特征为：

$$\boldsymbol{z}_{(i,k)_k}=\boldsymbol{v}_i\oplus\boldsymbol{v}_j\oplus\boldsymbol{u}_{(i,k)_k}\ (i,j)_k\in G \tag{10}$$

式中，$\boldsymbol{z}_{(i,k)_k}$为晶体图G中原子i与原子j连接的第k个键的更新原子特征；⊕表示原子和键特征的串联。此时，非线性图卷积函数定义如下：

$$\boldsymbol{v}_i^t=\boldsymbol{v}_i^{t-1}+\sum_{j,k}\sigma\left(\boldsymbol{z}_{(i,j)_k}^{t-1}W_{\mathrm{f}}^{t-1}+b_{\mathrm{f}}^{t-1}\right)\odot g(\boldsymbol{z}_{(i,j)_k}^{t-1}W_{\mathrm{s}}^{t-1}+b_{\mathrm{s}}^{t-1}) \tag{11}$$

式中，⊙代表矩阵对应元素相乘；σ为sigmoid函数；g是一个非线性激活函数（这里指“Leaky ReLu”或“Softplus”）；W和b分别表示神经网络的权重和偏差；$\sigma(\cdot)$函数是相邻之间不同交互的学习权重矩阵；f和s分别代表first和self的缩写。在R层卷积之后，获得的向量通过k个隐藏层完全连接，然后将该向量线性变换为标量值。根据连接距离的收集过滤器，排除离吸附物太远的原子的贡献，然后再使用平均池化层来生成整体特征向量$\boldsymbol{v}_{\mathrm{c}}$，该特征向量可以由池化函数表示：

$$\boldsymbol{v}_{\mathrm{c}}=\mathrm{Pool}(\boldsymbol{v}_0^{(0)},v_1^{(0)},\ldots,\boldsymbol{v}_N^{(0)},\ldots,\boldsymbol{v}_N^{(R)}) \tag{12}$$

通过优化代价函数$J\left(E_{\mathrm{ad}},\hat{E}_{\mathrm{ad}}\right)$进行训练，产生由权重$W$参数化的函数$f$，该函数将晶体$C$的晶体图映射到目标属性$\hat{E}_{\mathrm{ad}}$。通过使用反向传播和随机梯度下降（SGD），使用DFT计算数据迭代更新权重来解决以下优化问题：

$$\min_W J(E_{\mathrm{ad}},f(C;W)) \tag{13}$$

此时，相应地学习权重W代替目标属性E_{ad}，与池化输出的倒数第二层一起被进一步提取作为GP中的特征。节点的描述符为$V=[\boldsymbol{v}_0^0,\boldsymbol{v}_1^0,\ldots,\boldsymbol{v}_N^R]$，使用其相应的能量（$E_{\mathrm{ad}}$）进行训练。预测函数为：

$$f(\boldsymbol{v})\sim\mathrm{GP}[P(\boldsymbol{v}),k(\boldsymbol{v},\boldsymbol{v}')] \tag{14}$$

式中，$P(\boldsymbol{v})$是先验函数的常数均值；$k(\boldsymbol{v},\boldsymbol{v}')$是具有通过最大似然估计方法训练的长度尺度Matern核。这些均可通过GPyTorch实现[41]。

2.4. 机器学习和数值计算之间的迭代

第五范式平台的智能驱动、以知识为中心的本质可以通过机器学习和数值计算之间的迭代被很好地描述，这些迭代由“火山图”的跨学科知识串联起来。这突破了机器学习与数值计算之间人工筛选研究的新材料瓶颈，实现了科学实验与AI的相互促进，如图4（a）所示。实验涉及从Materials Project网站获取原始晶体（或原胞）并存储在数据库中的过程、进行火山标定关系信息比对的过程，以及智能模型的构建以创建大量吸附能计算模型的过程。通过第一性原理计算，将优化后的模型和吸附能数据存储在数据库中，并从中提取指纹以训练合适的机器学习模型。训练后的模型可以使用从尚未经过理论计算的待筛选材料中提取的指纹来预测吸附能，并将结果再次存储在数据库中。通过鲁棒松弛方案，比如对切面最低预测吸附能量的吸附点模型进行智能分析，从中筛选需要进一步开展DFT计算的模型。整个循环如下：①②③④⑤⑥⑦⑧⑨⑩, ④⑤⑥⑦⑧⑨⑩, ..., ④⑤⑥⑦⑧⑨⑩。

当且仅当计算框架中所有材料预测或计算完毕时，机器学习与DFT计算迭代反馈的过程才会停止。机器学习与第一性原理融合的功能在这些步骤中得到了很好的体现。步骤⑤表示数值计算得到的数据集补充了机器学习过程中没有数据集和数据集较少的问题。步骤⑩表明，通过机器学习预测后，再利用火山标定关系可以过滤掉大量数值计算，达到实现材料快速筛选的目的。此外，机器学习的结果可以通过整合实验和理论科学家的知识（跨学科专家的协同作用）的“火山图”进行智能分析，形成以知识为中心的智能驱动的第五范式。

2.5. 信息科学工具

第五范式的框架是通过使用各种Python包构建的，如Python Materials Genomics（pymatgen）、自动模拟环境（ASE）、FireWorks、Luigi和MongoDB [42–45]。pymatgen是Python支持的用于高通量材料计算的功能强大的程序包之一。该程序包标准化了运行高吞吐量计算之前所需的初始化设置，并提供了计算生成的数据的过程分析。ASE旨在建立、引导和分析原子模拟。FireWorks的功能是对运行在高性能计算集群上的高吞吐量计算工作流进行作业管理。Luigi可以用来构建复杂的批处理作业管道，处理依赖关系，并进行工作流管理。MongoDB是用C++语言编写的，用于实时数据存储，可以共同满足JavaScript Object Notation数据交换格式。

如图4（b）所示，在基于Lustre文件系统的天河超级计算机上开展数据密集型DFT计算工作[46]，可以通过运行部署在集群上的安全监控系统及服务实现高通量任务的管理与调度。针对该模型设计两类服务：Luigi服务和

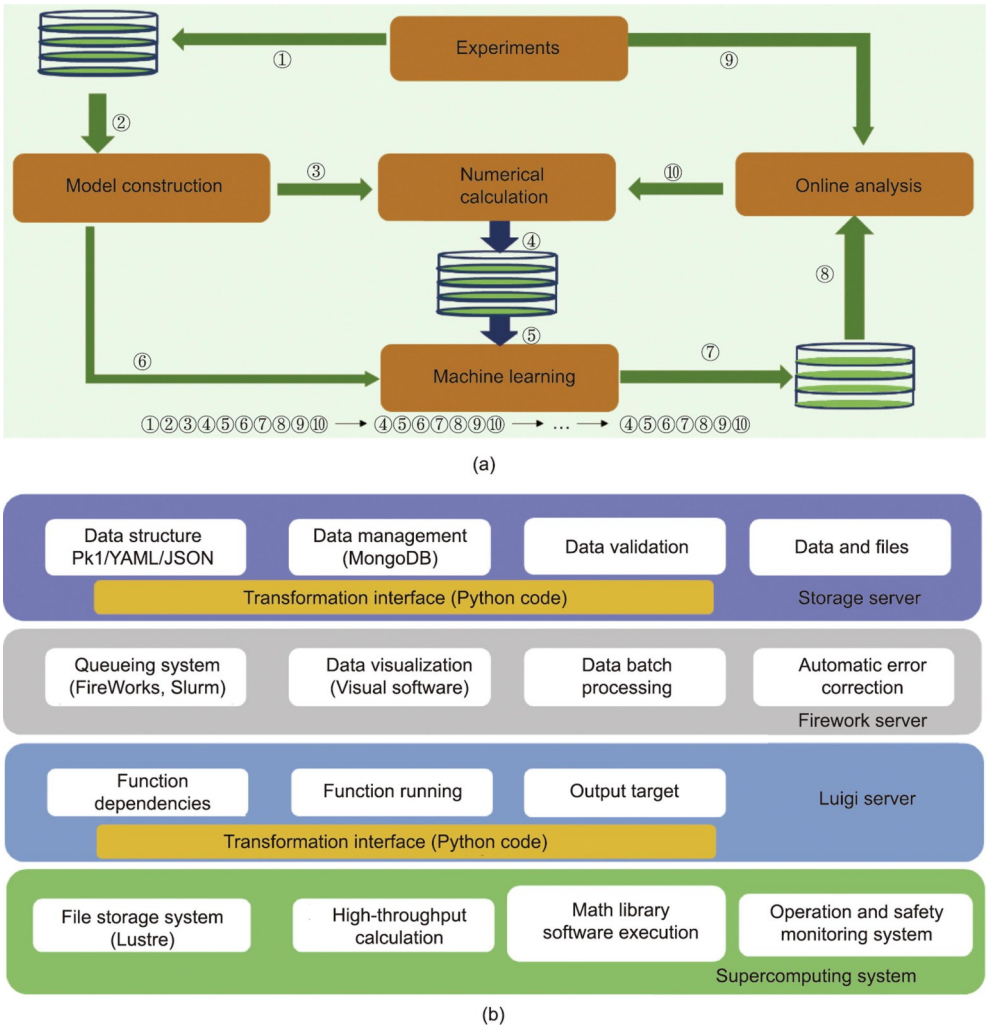

图4. 第五范式的架构。(a) 重复迭代框架包括第五范式平台中的机器学习和数值计算。步骤①和②以及步骤⑦和⑧分别将实验结果和机器学习结果拉入和拉出数据库。步骤③显示了为从头计算准备的构造模型。步骤④是计算结果的存储过程，步骤⑤和⑥分别是从计算结果和实验结果中提取的指纹。步骤⑨是指通过“火山图”对机器学习结果进行在线分析。”步骤⑩显示了在线分析（筛选）后的剩余模型，这些模型需要进一步的数值计算。(b) 服务和功能的实现基于天河一号超级计算机的第五范式平台。GASpy 中专用于服务的类型组件包括存储服务器、Firework 服务器和 Luigi 服务器。超级计算系统的基本环境是在软件层面。

FireWorks 联合 MongoDB 的服务。Luigi 通过管理和解析依赖（函数依赖、运行和输出目标），构建各种物理模型，再通过 FireWorks 对任务管理系统进行配置和计算，通过超级计算机[47]中的 Slurm 资源管理系统对这些任务进行批处理提交。综上，基于天河超级计算机的高通量材料计算环境能够灵活管理和运行单个作业，最终实现高通量任务被持续提交与执行的过程。

3. 性能评估

为了说明第五范式平台在催化材料筛选中的性能，本研究进行了比较测试，以解释机器学习过程如何加速数值计算，以及数值计算过程如何为机器学习迭代提供可训练的样本。在本文中，没有在每个模型的学习周期中使用包含在线 DFT 计算过程的更新数据集，而是使用 DFT 计算数据集提取相应的指纹进行研究。由于目标预测与 DFT 计算的结构没有直接关系，而是与从初始结构中提取的指纹有关，因此没有进行任何模拟处理。本文研究团队认为这不会影响对平台的评价。

准备的测试交叉验证过程的数据集来自 Github（https://github. com/ulissigroup/uncertainty_benchmarking）。由 H、CO、OH、O 和 N 五个吸附物组成，其中主要数据集来自前两个吸附物（21 269 和 18 437）。采用 CFGP 的方法创建模型，超参数采用第二章用到的设置，通过 R^2、平均绝对误差（MAE）、均方根误差（root-mean-square error, RMSE）等性能指标比较不同机器学习模型、数据集总数对催化剂筛选准确率的影响。该方法已与其他几种机器学习方法经过对比，得出效果最佳的结论。数据集的

超参数已经由Back等[31]和Tran等[37]进行了调优，而本文的研究重点是同一方法下的不同模型的性能，因此这些超参数仍然适用。在本文研究工作中，学习问题的表述是通过著名的“火山图”来确定的，用于评估其吸附能的大小和活性水平。在实验中，以H吸附物为例，析氢反应（HER）是一种利用吸附能预测催化性能的方法。最佳吸附能ΔE_H为−0.27 eV [48]，火山标定关系的最优范围定义为[−0.37 eV, −0.17 eV]。如果每个循环的结果都达到最优范围（在接近最优范围内的命中），则选择将该结果作为候选材料在下一个循环开始前继续进行DFT计算的工作。

机器学习和数值计算之间相互反馈实现的是DFT计算提供的可训练样本可以补充机器学习迭代。在该平台中，一旦发生迭代，就确定了包含目标特征的数据集，这意味着确定了相应迭代的机器学习模型。此外，作为第五范式平台的典型案例，每个迭代过程的性能比较都是从相同数据生成条件下的模型比较中得出的。根据表1所示，为评估高通量材料计算在智能驱动过程中反馈迭代的机器学习性能，构建了10组数据集。首先将整个数据集随机打乱并拆分为10个模型，将总数据集的10%作为第一个模型数据集，然后递增直到将总数据集的100%作为第十个模型数据集，形成10个模型对应的数据集。前一个模型的数据集包含在下一个模型的数据集中。对于交叉验证过程，每个模型的训练/验证/测试比为64/16/20，如文献[37]所述将所有单金属添加到训练集中。交叉验证及其结果列于表1和图5中，图5（a）为训练和测试数据集的相关系数（R^2）。图形越细长，表示二者之间的差异越大。如果二者相同，则可以是一条线。model 1、model 2、model 5、model 6、model 9的细长图形表示训练和测试结果差异非常大，为过拟合或欠拟合的指标；model 3、model 4、model 7、model 10次之；model 8表现最好。随着数据集的增加，表1中的MAE和RMSE逐渐减小，而图5中验证和测试过程的R^2趋势逐渐增大，表明数据集不断迭代产生的训练模型比之前的模型更准确。此外，还列出了经DFT计算（N_{DFT}）和机器学习预测（N_{ML}）验证的H吸附的命中数，其趋势也随着数据集的扩大而增加，如图S1所示（见附录A）。将model 1数据集设置为基准。为了寻找机器学习迭代数值计算提供的不断增加的可训练样本的性能，定义公式如下：

$$\eta = \frac{D_n - D_1}{M_n - M_1}\ n \in \mathbf{N},\ 1 < n \le 10 \qquad (15)$$

式中，η表示N_{DFT}与机器学习预测数量N_{ML}相比的增量。D_n和M_n指模型n在接近最优范围（即命中数）的N_{DFT}和N_{ML}。随着数据集的扩大，η的趋势越来越大，当接近1时，说明命中数N_{ML}正在慢慢接近命中数N_{DFT}，表明DFT计算的训练样本越大，机器学习模型的准确率越高。此外，η在图5（b）中拟合得很好，即使有些点不在线性范围内。例如，模型4的η与其他点相比非常小，将其归因于模型5和模型6中较大值的补偿。

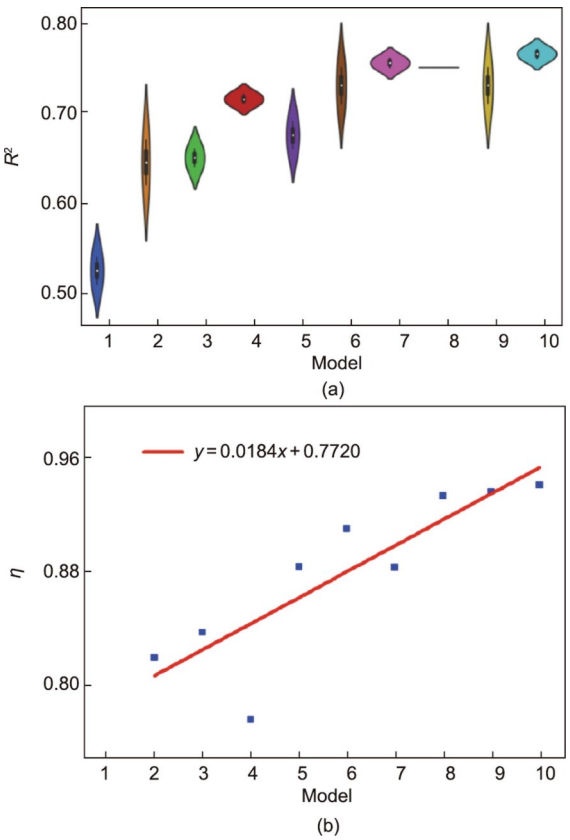

图5. 第五范式平台下学习模式的绩效指标评价。（a）10个模型中验证和测试过程的R^2相关系数。（b）所有模型的η的线性拟合。

数据集指每个模型的数据集总数，评价指标包含MAE和RMSE及η。通过DFT计算和机器学习预测验证火山标定关系中H吸附能接近最优范围的切面数量，分别以N_{DFT}和N_{ML}为代表，η用于评估模型性能变化的趋势。

为了说明机器学习和数值计算之间相互反馈的实现（例如，机器学习解决了数值计算中计算资源不足导致的大量模型耗时问题，而数值计算过程提供了机器学习训练样本），本文准备了三种类型的预测案例来了解上述训练和验证模型的性能。在预测过程中使用的数据集来自Tran和Ulissi [28]的工作，其中包含22 675个H吸附DFT结果。事实上，该数据集已经涵盖了上面提到的21 269 H数据集的大部分。然而，这与重复数据集无关，因为本研究的目标是比较在大小不同的样本上生成的机器学习模型的性能，并找出大小不同的预测样本下机器学习的加速行为。此外，要预测的与数据集对应的材料结构并不取决于是否

进行了模拟计算。因此，该机器学习预测数据集取自DFT计算数据集的决策不会影响对智能驱动过程的整体评估。

表1 机器学习模型性能评估

Model	Dataset	MAE (eV)	RMSE (eV)	H-dataset	Near-optimal activity		η
					N_{DFT}	N_{ML}	
1	4 728	0.30	0.52	2 090	72	60	–
2	9 456	0.26	0.49	4 155	141	144	0.82
3	14 184	0.24	0.43	6 293	208	222	0.84
4	18 912	0.22	0.40	8 384	275	321	0.78
5	23 640	0.23	0.44	10 530	351	375	0.89
6	28 368	0.22	0.41	12 678	417	438	0.91
7	33 096	0.21	0.37	14 756	512	557	0.89
8	37 824	0.21	0.38	16 926	598	622	0.94
9	42 552	0.22	0.41	19 086	658	684	0.94
10	47 290	0.19	0.36	21 269	709	735	0.94

就高通量材料计算在智能驱动过程中的反馈迭代特点而言，每个周期（第一个周期除外）进行的DFT计算均从机器学习结果中获得。表2列出了三种方法：累加异测保守丢弃、累加异测和同值异测方法。累加异测方法是指上面形成的model 1到model 10的机器学习模型对应的总预测数据集的10%到100%增量数据集。此外，整个预测数据集也可以在每个周期中保持相同，正如同值异测方法所定义的那样。对于累加异测保守丢弃方法，意味着机器学习在最优范围内预测的模型在下一次模型预测中被丢弃。过程如下：从model 1开始，在预测的整个22 675个模型中，发现机器学习预测的4960个模型是命中的。当使用模型2进行预测时，机器学习预测的模型将从预测的22 675个模型中剔除4960个模型，只剩下17 715个模型（22 675 − 4960 = 17 715）。然后，模型2发现，机器学习预测的860个模型被命中，并提供另一个简化样本16 855（17 715 − 860 = 168 55）用于模型3的预测。在这10个模型的预测完成之前，这种波动不会结束。请注意，N_{Hits}应该等于N_{ML}，但是样本中的某些材料必须排除在接近最优的活动过程之外。

表2列出了三种方法在最优范围内的结果。在累加异测保守丢弃方法中，由于前一个模型预测的N_{ML}从下一个模型的预测样本中被扣除（model 1除外），所以从model 1到model 10的N_{DFT}、N_{ML}和N_{Hits}也相应减少。在累加异测方法中，随着预测样本的增加，N_{DFT}和N_{ML}逐渐扩展。在同值异测方法中，N_{ML}在4177和4556之间波动，而N_{DFT}保持不变。因此推断，这是由机器学习模型的不同精度引起的，模型中的数据集总体呈现越多，N_{ML}被预测命中得越多。从加速的角度来看，累加异测保守丢弃方法可以保证在上一轮中被预测过的数据集不会在下一轮中被再次预测，而其他两种方法在每轮预测中涉及数据集的重复预测。因此，理想的累加异测保守丢弃方法保证了所有数据集仅被预测一次，因此每轮迭代过程中不重复的数据集能够提供更快的机器学习过程。

为了评估这些方法在加速DFT计算方面的差异，在图6中比较了N_{ML}取代的N_{DFT}的数量，以及N_{ML}/N_{DFT}的值。机器学习替代DFT计算的定义如下：

$$\mathrm{RE} = \begin{cases} T_n - M_n & n = 1 \\ T_n - M_n - M_{n-1} & n \in \mathbf{N},\ 1 < n < 10 \end{cases} \tag{16}$$

式中，RE和T_n为每个模型中被机器学习取代的DFT计算次数和所有预测数据集。如图6（a）所示，累加异测保守丢弃法的所有模型的替换量都在15 000以上。从model 1到model 10的替换量略有减少，但与其他方法相比，累加异测保守丢弃法的N_{DFT}替换量最大。对于累加异测方法，替换次数从1800线性增加到与model 10中的其他方

表2 接近最优范围内的三种预测及所有模型性能

Model	Hit_no_split				No_hit_with_split			No_hits_no_split		
	Dataset	N_{DFT}	N_{ML}	N_{Hits}	Dataset	N_{DFT}	N_{ML}	Dataset	N_{DFT}	N_{ML}
1	22 675	4 027	4 446	4 960	2 268	389	392	22 675	4 027	4 282
2	17 715	1 707	775	860	4 536	774	853	22 675	4 027	4 331
3	16 855	1 484	485	539	6 804	1 178	1 299	22 675	4 027	4 380
4	16 316	1 336	374	419	9 072	1 573	1 722	22 675	4 027	4 293
5	15 897	1 264	309	324	11 340	1 970	2 133	22 675	4 027	4 177
6	15 573	1 194	161	174	13 608	2 417	2 578	22 675	4 027	4 249
7	15 399	1 157	127	141	15 876	2 846	3 058	22 675	4 027	4 413
8	15 258	1 118	193	210	18 144	3 231	3 448	22 675	4 027	4 215
9	15 048	1 058	328	352	20 412	3 650	3 799	22 675	4 027	4 273
10	14 696	968	365	383	22 675	4 027	4 556	22 675	4 027	4 556

The dataset refers to the total number of data sets for each model. The N_{Hits} is the number of machine learning predictions that do not exclude certain materials.

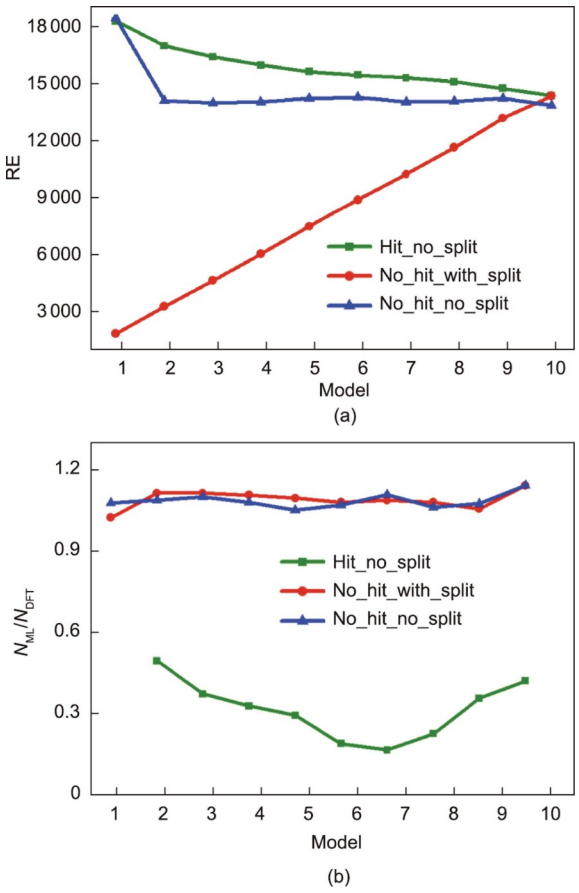

图 6. 在第五范式平台中构建的所有模型的预测性能。上图是 DFT 计算（N_{DFT}）的数量替换为机器学习预测（N_{ML}）的数量。下图是预测过程中不同模型的 N_{ML}/N_{DFT} 在接近最优范围内的变化。在 Hit_no_split 方法中，model 1 由于其对其他模型的基线函数而被放弃。

法相同的次数。对于同值异测方法，除 model 1 外，所有模型的替换次数都在 14 000 次左右，并且有轻微的下降趋势。对于同值异测方法中 model 1 的大量替换以及随后的突然减少，本文研究团队认为这是由欠拟合引起的，因为 model 1 使用少量的数据集来训练模型以预测更大的数据集。在这些方法中，累加异测保守丢弃可以代替最大 N_{DFT}，这正如所期望的那样。

在图 6（b）中，通过比较 N_{ML}/N_{DFT}，可以从另一个角度反映每个模型的性能。理想的 N_{ML}/N_{DFT} 值应该均为 1。在累加异测和同值异测方法中，N_{ML}/N_{DFT} 略微增大到接近 1，表明两种方法的预测行为相似，适合加速 DFT 计算。而在累加异测保守丢弃方法中，model 1 作为参考基准不考虑其性能，N_{ML}/N_{DFT} 值从 model 2 逐渐减小到 model 7，然后在其余模型中逐渐增加，且所有模型均低于 0.5。一方面，这些较小值的浮动是由机器学习模型的精度变化引起的；另一方面，随着预测样本命中数的减少，在下一个模型中可以命中的 N_{ML} 逐渐减少。此外，对于累加异测保守丢弃方法，每轮会去掉上一轮的命中数，在下一个模型中可以命中的 N_{ML} 逐渐减少。该方法由于不涉及重复命中材料在其他迭代中被再次命中的情况，因此虽然该方法在速度方面的优势更加明显，但并不适合利用 N_{ML}/N_{DFT} 指标对精度进行评估。

另外，由于机器学习模型本身在交叉验证小样本的扩展过程中呈现出不良拟合逐渐减少的特点，故在 model 2 到 model 10 的预测过程中会出现一定程度的精度损失。例如，预测的机器学习数据集本应被命中但未被命中，或者数据集不应被命中但被命中，导致命中数据丢失或未命中数据在下一个模型的数据集中增加。预测样本量不够大，导致机器学习模型的欠拟合或过拟合。累加异测保守丢弃方法具有替代更多 DFT 计算的优势，虽然不适合用 N_{ML}/N_{DFT} 指标进行精度的评估，但这并不代表累加异测保守丢弃方法不适用于第五范式平台。当预测模型足够好且数据集足够大时，该方法能够减少数据的重复预测过程，同时保持结果的可靠性，以加速机器学习的优势，实现在高精度下加速材料的筛选效果。

基于这三种方法的结果，利用机器学习预测相对于 DFT 计算的精度损失来评估第五范式平台的性能。精度损失可以定义为：

$$L = \frac{M_n - D_n}{T_n} \quad n \in \mathbf{N},\ 1 \le n \le 10 \qquad (17)$$

式中，L 为精度损失。鉴于累加异测和同值异测方法具有相对合适的预测性能，因此只考虑这两种方法的精度损失。如图 7 所示，在探索未知世界的过程中，科学实验、理论计算和机器学习的相互验证过程代表了材料科学研究范式的准确性。对于累加异测方法，虽然 model 1 的准确率损失最低，但数据集很小导致机器学习取代 DFT 的规模非常小，所以速度不高。model 9 机器学习取代 DFT 的规模非常大，且机器学习与 DFT 计算数量趋向一致，因此精度损失最低。对于同值异测方法，model 5 机器学习取代 DFT 的规模非常大，且机器学习与 DFT 计算数量趋向一致，因此精度损失最低。因此，随着数据集的扩大，机器学习将继续取代 DFT 计算，并且会有不同程度的精度损失。

本研究认为，第五范式的精度损失与机器学习、理论计算和火山标定关系反馈的实验样本量有关，这正是第五范式在精度方面以知识为中心的特征。如图 7 所示，准确的第五范式应该使机器学习、理论计算和科学实验在共同探索未知世界时呈现一致的结果。虽然这个标准要求很高，但始终是人类对未知世界不断探索以期达到的高度。

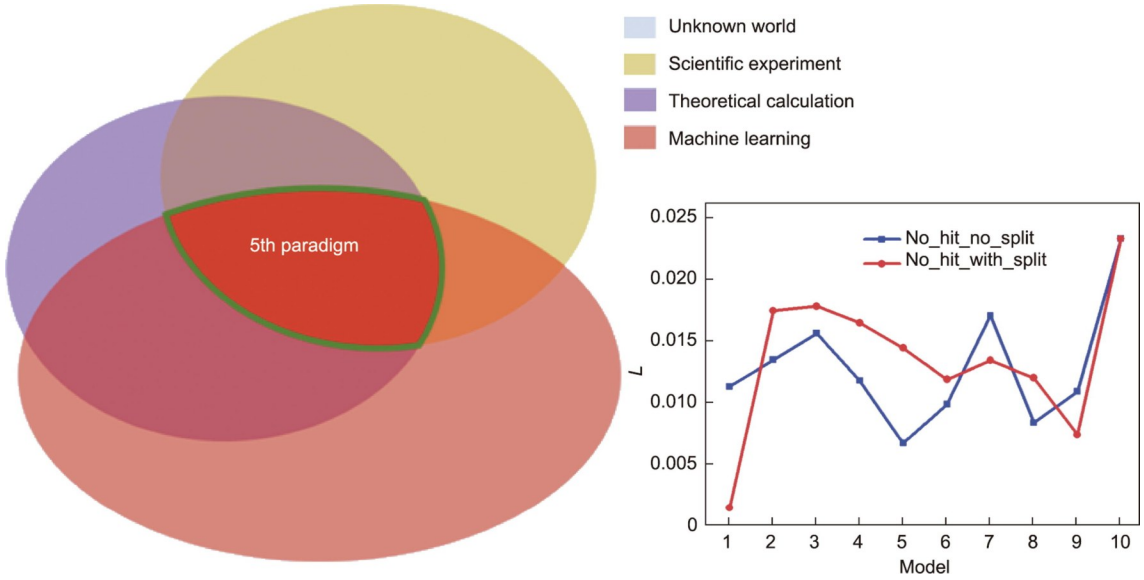

图7. 第五范式的准确性。在探索未知世界的过程中，科学实验、理论计算和机器学习的相互验证过程代表了第五范式的准确性。在第五范式平台中构建了所有模型的机器学习和DFT计算之间的No_hit_No_split和No_hit_with_split方法的精度损失（*L*）。

4. 第五范式平台的讨论

模型的自动构建、指纹的自动提取、密集数据与DFT计算的智能耦合以及“火山图”的机器学习构成了第五范式平台的架构。在智能驱动的框架下，充分利用当前各种信息工具和方法的发展，有效地减少了传统模型构建和计算的工作量，大大简化和改进了材料研究中极其繁琐和具有挑战性的工作。

该框架面临的挑战之一是在第五范式中实现的有限应用领域。这是因为第五范式最典型的特征是智能驱动，需要跨学科专家协同进行深入研究。例如，在本工作介绍的材料科学中，需要智能驱动实验专家和理论专家的高效协同，这可以通过“火山图”过滤机器学习结果来实现。对于一些高通量的跨学科工作，在设计类似的第五范式框架之前，最好首先考虑利用适当的方法来量化不同应用领域的专家之间的协作工作。

此外，机器学习模型的鲁棒性和泛化性又与对应的数据集规模有关。缺乏大的数据集将导致训练模型较差的泛化能力，因此需要积累更多的有效数据集才能实现高精度的机器学习模型预测。有效数据集的来源依赖于更多DFT计算产生的结果，比如，目前Facebook AI Research研究团队和卡内基梅隆大学化学工程系联合研发的Open Catalyst 2020 [49]数据集项目正在持续进行相关数据集更新。因此，智能驱动过程中机器学习对于DFT计算产生的有效数据集规模及相关的准确度对量化精度过程非常重要。最后，利用第五范式实现对未知世界探索的准确性受到机器学习、理论计算和科学实验的影响。高精度的第五范式倾向于在其合理的发现、推导和判断范围内，通过三种合作从未知世界中探索同一客观事物。因此，对第五范式案例的剖析可以极大地促进材料科学第五范式在未来的发展。

5. 结论

在本研究中，讨论了因人工智能带来的繁荣而出现的最新范式的科学解释。然后，以第五范式平台为典型案例进行了详细的讨论。该平台符合一个具体而明确的框架，能够促进材料科学的发展。跨学科的知识和智能驱动的特征是第五范式的关键，这可以在包括自动模型构建和验证、自动指纹构建以及机器学习和理论计算之间的理论模型和重复迭代的工作中解决。还详细讨论了构建框架所需的信息学工具。最后，进行了测试和比较，以展示在第五范式案例的框架下，人工智能和数值计算之间的相互作用如何有意义地相互促进、减少数值计算，并在相互反馈过程中创建更多可训练的样本。数值计算和机器学习模型以及技术的管理使第五范式平台更具可解释性。

随着数据集的扩大，一方面，机器学习取代的DFT计算越多，材料的筛选就越快。另一方面，最终机器学习预测的候选材料数量与DFT计算的候选材料数量越一致，机器学习的预测就越准确。这种最小精度损失的辨别，代表了科学第五范式下材料研究的精准探索前提，即要求在

利用机器学习、理论计算、科学实验共同探索未知世界时，得到一致的结果。

虽然本文为催化材料领域中所代表的第五范式平台提供了科学的解释，但也要承认还有更多的东西需要讨论。跨领域第五范式的整体发展仍然面临着需跨学科专家之间的协同作用和数据驱动学科中数据需求的急剧增长的挑战。尽管面临这些挑战，但可以预见，在各方的共同努力下，人工智能技术与传统学科的结合将不断深化，使每个模拟和计算环节具有更高的智能和自动化特征，最终作为一个平台去运行，从而提高传统科学计算的效率，推动材料研究向更加智能和高精度的方向发展。未来，对第五范式平台的关注可以为第五范式在其他领域的应用铺平道路。

致谢

感谢卡内基梅隆大学的Zachary W. Ulissi教授和Pari Palizahti教授关于第五范式平台提供的建议；感谢国家重点研发计划项目（2021ZD40303）、国家自然科学基金项目（62225205 和 92055213）、湖南省自然科学基金项目（2021JJ10023）和深圳市基础研究项目（自然科学基金项目）（JCYJ20210324140002006）的资助。

Compliance with ethics guidelines

Can Leng, Zhuo Tang, Yi-Ge Zhou, Zean Tian, Wei-Qing Huang, Jie Liu, Keqin Li, and Kenli Li declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2022.06.027.

References

[1] Barber B. Resistance by scientists to scientific discovery. Science 1961; 134(3479):596–602.
[2] Dampier WCD. A history of science, technology and philosophy in the eighteenth century. Nature 1939;143(3613):134–5.
[3] Crombie AC. Scientific change: historical studies in the intellectual, social and technical conditions for scientific discovery and technical invention, from antiquity to the present. London: Heinemann; 1963.
[4] Bidney M, Piekielek N. Towards a new paradigm in map and spatial information librarianship. J Map Geogr Libr 2018;14(2–3):67–74.
[5] Li J, Huang W. Paradigm shift in science with tackling global challenges. Natl Sci Rev 2019;6(6):1091–3.
[6] Tolle KM, Tansley DSW, Hey AJG. The fourth paradigm: data-intensive scientific discovery. Proc IEEE 2011;99(8):1334–7.
[7] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015; 518(7540):529–33.
[8] Bainbridge WS. The scientific research potential of virtual worlds. Science 2007;317(5837):472–6.
[9] Zubarev DY, Pitera JW. Cognitive materials discovery and onset of the 5th discovery paradigm. In: Pyzer-Knapp EO, Laino T, editors. Machine learning in chemistry: data-driven algorithms, learning systems, and predictions. Washington, DC: American Chemical Society; 2019. p. 103–20.
[10] Malitsky N, Castain R, Cowan M. Spark–MPI: approaching the fifth paradigm of cognitive applications. 2018. arXiv:1806.01110.
[11] Woinaroschy A. A paradigm-based evolution of chemical engineering. Chin J Chem Eng 2016;24(5):553–7.
[12] Si Y, Wu HY, Yang K, Lian JC, Huang T, Huang WQ, et al. High-throughput computational design for 2D van der Waals functional heterostructures: fragility of Anderson's rule and beyond. Appl Phys Lett 2021;119(4):043102.
[13] Li B, Peng W, Zhang J, Lian JC, Huang T, Cheng N, et al. High-throughput one-photon excitation pathway in 0D/3D heterojunctions for visible-light driven hydrogen evolution. Adv Funct Mater 2021;31(18):2100816.
[14] Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci 2019;6(21):1900808. Corrected in: Adv Sci 2020;7(2):1903667.
[15] Hardian R, Liang ZW, Zhang XL, Szekely G. Artificial intelligence: the silver bullet for sustainable materials development. Green Chem 2020;22(21):7521–8.
[16] Xu X, Ma WP, Yan B. An electrodeposited nano-porous and neural network-like Ln@HOF film for SO_2 gas quantitative detection via fluorescent sensing and machine learning. J Mater Chem A 2021;9(46):26391–400.
[17] Kumar S, Ignacz G, Szekely G. Synthesis of covalent organic frameworks using sustainable solvents and machine learning. Green Chem 2021; 23(22): 8932–9.
[18] Ding WL, Lu YM, Peng XL, Dong H, Chi WJ, Yuan X, et al. Accelerating evaluation of the mobility of ionic liquid-modulated PEDOT flexible electronics using machine learning. J Mater Chem A 2021;9(45):25547–57.
[19] Vandenberg P. The fourth industrial revolution. J Asia Pac Econ 2020; 25(1): 194–6.
[20] Feng R, Zhang C, Gao MC, Pei Z, Zhang F, Chen Y, et al. High-throughput design of high-performance lightweight high-entropy alloys. Nat Commun 2021;12(1):4329.
[21] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses,opportunities, and threats. Engineering 2021;7(9):1201–11.
[22] Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 2019;5(6):1017–26.
[23] Chen S, Zhang S, Shang J, Chen B, Zheng N. Brain inspired cognitive model with attention for self-driving cars. 2017. arXiv:1702.05596.
[24] Xu Z. Principle analysis of computer vision and its application research. In: Proceedings of the 2018 7th International Conference on Advanced Materials and Computer Science; 2018 Dec 21 – 22; Dalian, China. Ottawa: Clausius Scientific Press; 2018. p. 478–82.
[25] Itaya K, Takahashi K, Nakamura M, Koizumi M, Arakawa N, Tomita M, et al. BriCA: a modular software platform for whole brain architecture. In: Hirose A, Ozawa S, Doya K, Ikeda K, Lee M, Liu D, editors. Neural information processing. Cham: Springer International Publishing; 2016. p. 334–41.
[26] US Department of Energy. Synergistic challenges in data-intensive science and exascale computing. Summary report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee. Washington, DC: US Department of Energy, Office of Science; 2013.
[27] Wang C, Yu F, Liu Y, Li X, Chen J, Thiyagalingam J, et al. Deploying the Big Data Science Center at the Shanghai Synchrotron Radiation Facility: the first superfacility platform in China. Mach Learn Sci Technol 2021;2(3):035003.
[28] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 2018;1(9):696–703.
[29] Kresse G, Furthmüller J. Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci 1996;6(1):15–50.
[30] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, et al. Accelerated

discovery of CO_2 electrocatalysts using active machine learning. Nature 2020; 581 (7807):178–83.
[31] Back S, Yoon J, Tian N, Zhong W, Tran K, Ulissi ZW. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. J Phys Chem Lett 2019;10(15): 4401–8.
[32] Wigner E, Seitz F. On the constitution of metallic sodium. Phys Rev 1933; 43(10):804–10.
[33] Abild-Pedersen F, Greeley J, Studt F, Rossmeisl J, Munter TR, Moses PG, et al. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. Phys Rev Lett 2007;99(1):016105.
[34] Calle-Vallejo F, Martínez JI, García-Lastra JM, Rossmeisl J, Koper MTM. Physical and chemical nature of the scaling relations between adsorption energies of atoms on metal surfaces. Phys Rev Lett 2012;108(11):116103.
[35] Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev 1964;136(3B): B864–71.
[36] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev 1965;140(4A):A1133–8.
[37] Tran K, Neiswanger W, Yoon J, Zhang Q, Xing E, Ulissi ZW. Methods for comparing uncertainty quantifications for material property predictions. Mach Learn Sci Technol 2020;1(2):025006.
[38] Garrido Torres JA, Jennings PC, Hansen MH, Boes JR, Bligaard T. Low-scaling algorithm for nudged elastic band calculations using a surrogate machine learning model. Phys Rev Lett 2019;122(15):156001.
[39] Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater 2019; 31(9):3564–72.
[40] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018; 120(14):145301.
[41] Gardner JR, Pleiss G, Bindel D, Weinberger KQ, Wilson AG. In: GPyTorch: blackbox matrix−matrix Gaussian process inference with GPU acceleration. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, editors. Proceedings of the 32nd International Conference on Neural Information Processing Systems; 2018 Dec 3−8. Montréal, QC, Canada. Red Hook: Curran Associates Inc.; 2018. p.7587–97.
[42] Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. Comput Mater Sci 2013;68:314–9.
[43] Hjorth Larsen A, Jørgen Mortensen J, Blomqvist J, Castelli IE, Christensen R, Dułak M, et al. The atomic simulation environment—a Python library for working with atoms. J Phys Condens Matter 2017;29(27):273002.
[44] Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, et al. FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr Comp Pract E 2015;27(17):5037–59.
[45] Jiao YQ, Li YJ, Li B, Song YG, inventors; Inc.Goertek, assignee. [MongoDB-based test data storage query method and system]. Chinese patent CN 105550333A. 2021 May 4. Chinese.
[46] Wang Y, Lu Y, Qiu C, Gao P, Wang J. Performance evaluation of a infiniband-based lustre parallel file system. Proc Environ Sci 2011;11(Pt A):316–21.
[47] Yoo AB, Jette MA, Grondona M. SLURM: simple Linux utility for resource management. In: Feitelson D, Rudolph L, Schwiegelshohn U, editors. Job scheduling strategies for parallel processing. Berlin: Springer; 2003. p. 44–60.
[48] Nørskov JK, Bligaard T, Logadottir A, Kitchin JR, Chen JG, Pandelov S, et al. Trends in the exchange current for hydrogen evolution. J Electrochem Soc 2005;152(3):J23–6.
[49] Chanussot L, Das A, Goyal S, Lavril T, Shuaibi M, Riviere M, et al. Open catalyst 2020 (oc20) dataset and community challenges. ACS Catal 2021; 11(10):6059–72.