第 44 卷 第 4 期
2025 年 4 月

中国材料进展
MATERIALS CHINA

Vol. 44 No. 4
Apr. 2025

引用格式：程晋荣，何鹏飞，李艺欣，等. 数据与模型驱动的钙钛矿材料智能计算框架[J]. 中国材料进展，2025，44(4)：309-317.
CHENG J R，HE P F，LI Y X，*et al*. Data and Model Driven Intelligent Computing Framework for Perovskite Materials[J]. Materials China，2025，44(4)：309-317.

特约专栏

数据与模型驱动的钙钛矿材料智能计算框架

程晋荣[1]，何鹏飞[1]，李艺欣[2]，雷咏梅[2]

（1. 上海大学材料科学与工程学院，上海 200444）
（2. 上海大学计算机工程与科学学院，上海 200444）

摘　要：钙钛矿材料因其复杂的化学成分、多样的晶体结构和丰富的物理特性，成为现代材料科学研究热点之一。结合模型驱动方法和数据驱动方法，构建特征工程融合主动学习的材料智能计算框架，提高模型精度和系统性能。通过数据布局和动态调度协同优化，提出针对材料特征的确定独立筛选和稀疏算子(SISSO)并行计算方法，缓解 SISSO 算法在建立特征工程模型时面临的精度较低与计算成本较高的问题，降低数据质量对模型的影响。构建面向材料数据的主动学习方法，以处理材料数据标记的复杂性，剔除噪声数据。

关键词：SISSO 算法；智能计算；主动学习；钙钛矿材料
中图分类号：TQ174.1；TP181　　**文献标识码**：A　　**文章编号**：1674-3962(2025)04-0309-09

Data and Model Driven Intelligent Computing Framework for Perovskite Materials

CHENG Jinrong[1], HE Pengfei[1], LI Yixin[2], LEI Yongmei[2]

(1. School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China)
(2. School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: Perovskite materials have become one of the hotspots in modern materials science research due to their complex chemical compositions, diverse crystal structures and rich physical properties. In this paper, by combining the model-driven approach and the data-driven approach, a materials intelligent computing framework integrating feature engineering and active learning is constructed to improve the model accuracy and system performance. Through the collaborative optimization of data layout and dynamic scheduling, a sure independence screening and sparsifying operator (SISSO) parallel computing method for material features is proposed to alleviate the problems of low accuracy and high computational cost faced by the SISSO algorithm when establishing the feature engineering model and reduce the impact of data quality on the model. An active learning method oriented to material data is constructed to deal with the complexity of material data labeling and eliminate noisy data.

Key words: SISSO algorithm; intelligent computing; active learning; perovskite materials

收稿日期：2024-12-04　　**修回日期**：2025-04-01
基金项目：国家自然科学基金资助项目(52472133，91427304)；上海市自然科学基金原创探索项目(22ZR1481100)；水声对抗技术重点实验室开放基金资助项目(JCKY2024207CH12)；中国博士后科学基金资助项目(2024M751931)
第一作者：程晋荣，女，1969 年生，研究员，博士生导师
通讯作者：程晋荣，女，1969 年生，研究员，博士生导师，Email：jrcheng@shu.edu.cn
雷咏梅，女，1965 年生，教授，博士生导师，Email：lei@shu.edu.cn
DOI：10.7502/j.issn.1674-3962.202412002

1　前　言

数据驱动的机器学习模型凭借高效的数据处理能力在材料科学领域展现出巨大的发展潜力[1]，广泛应用于从多源、异构的材料数据中挖掘潜在的有效信息[2]，以及材料知识挖掘与表示[3]、分子动力学势场的开发[4]、新材料体系及其性质的预测[5]、实用型材料发现[6]等问题的研究。相较于传统的材料科学使用的“经验试错型”研究方法，机器学习方法能在节省时间和成本的基础上提高开发效率，对现有的实验数据

实现最大化的利用并构建经验模型用于优化和预测材料系统[7, 8]。

在材料领域的机器学习中，数据主要来源于实验测量数据、科学文献文本及材料性能数据库。然而，这类数据通常具有多源、异构显著、不确定性强、样本稀缺、维度复杂等特点[9-11]，使得数据驱动建模前需要提高材料数据的质与量[10, 12, 13]。本研究通过材料制备—表征的全流程标准化实验设计，建立具有统一环境基准的样本数据集，并结合领域专家知识的人工数据标注机制，有效降低实验误差与环境因素对数据一致性的干扰。

钙钛矿材料的多样性使得通用的特征工程方法难以充分捕捉其关键特征，必须针对其特有的结构和性质设计特征工程模型。欧阳润海等提出的确定独立筛选和稀疏算子(sure independence screening and sparsifying operator, SISSO)[14]是一种结合符号回归[15, 16]和压缩感知技术[17, 18]的数据驱动方法，能够从大量候选特征子集中筛选出最具预测能力的特征组合，并生成精确的数学模型。SISSO 基于给定的数学运算符和特征，在规定复杂度内构造大量高维描述符，并将这些描述符作为“基函数”用于描述目标属性，通过稀疏算子识别出最稀疏的解[19, 20]。Bartel 等[21]利用 SISSO 发现了一个准确且可解释的新容差因子，这些因子在判断化合物是否为钙钛矿结构方面表现出很高的精度。胡红青等[22]基于 SISSO 和机器学习方法对钙钛矿结构的稳定性进行预测，并通过建立新的容许因子加以验证。

模型特征在机器学习的模型训练和推断过程中作为预测的输入起着至关重要的作用，特征的质量和数量直接影响模型的性能和准确性。因此，特征工程作为机器学习中不可或缺的一部分，旨在从原始数据中提取和构造有用的特征，以提高模型的预测能力。为了确保模型的性能和可靠性，需要进一步优化和改进数据处理方法。材料数据的复杂性带来的高昂标记成本是影响算法模型性能的重要因素之一，而生产和测试过程中的差异可能引入噪声，增加数据处理的难度。SISSO 在自动化特征工程研究领域扮演了关键角色，通过其特有的符号回归方法，SISSO 能自动生成具有物理意义的特征，推动了自动化特征工程在材料科学和其他复杂领域的应用。主动学习通过选择对模型训练贡献最大的样本，减少训练集中噪声数据或无关数据，从而提高 SISSO 特征工程模型的精度。

随着材料数据复杂性的不断增加，SISSO 算法在处理这些数据时存在精度低与时间成本过高的问题。构建一种面向材料数据的主动学习方法有利于解决材料数据标记的复杂性问题，并剔除噪声数据。通过并行计算技术加速算法的执行，优化特征生成和选择过程，提高模型的预测精度和计算效率。尽管 SISSO 在高维特征空间处理上已经展现出显著优势，但通过分布式计算提高高维模型计算效率的研究仍然相对不足。本文提出一种分布式计算框架以提高 SISSO 特征选择的精度和效率，提高处理复杂材料数据时建立的高维模型的性能。该框架不仅在特征选择和模型构建过程中展现了灵活性和高效性，还能够有效应对材料科学研究中的多样化数据需求，为智能化材料开发和探索提供了具体实现方式。图 1 展示了数据与模型驱动的材料应用流程。

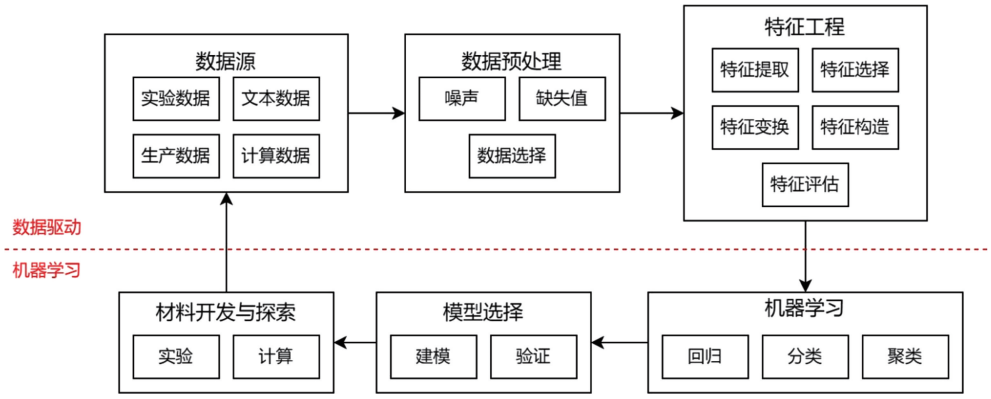

图 1 数据与模型驱动的材料应用流程

Fig. 1 Data and model driven material application process

2 数据与模型双驱动的建模方法

2.1 不同体系差异造成的精度问题

为了获得更全面的信息以及建立泛化性更好的模型，常将多个渠道的数据合并扩充数据集来增加训练样本的数量，从而提升模型的泛化能力。引入更大体系的数据集能够更有力地验证高维特征的普适性，但会引入数据的异质性，即不同来源或体系的数据之间存在差异。由

于数据的异质性，相关系数的计算可能偏离实际的线性关系，影响确定独立筛选(sure independence screening，SIS)阶段筛选出的特征子集。这些特征子集可能包含大量冗余或无关特征，无法很好地捕捉各体系数据之间的关系，从而影响模型的性能和可靠性。以钙钛矿型三元系压电陶瓷为例，不同体系数据可能具有不同的特性、背景噪声或实验条件，导致特征重要性方面存在较大差异，增加 SISSO 特征选择的复杂性。

数据信息帮助模型捕捉不同材料体系间的共性特征和规律，增强模型在不同应用场景中的适应能力。即使存在数据异质性，更大体系的数据集仍然能够帮助模型更好地泛化。在处理不同体系数据子集合并的数据集时，SISSO 算法面临的主要问题是数据异质性造成的模型精度下降，无法很好地适应大部分数据的规律。为了充分利用这些数据，根据 SISSO 算法特征筛选及稀疏建模特点，可通过设计一种能够尽可能适应不同体系数据特点的策略，提高模型对不同体系数据的适应性和泛化能力，从而更准确地预测材料的性质。图 2 给出了适应不同体系数据特点 SISSO 特征工程模型的建立策略。

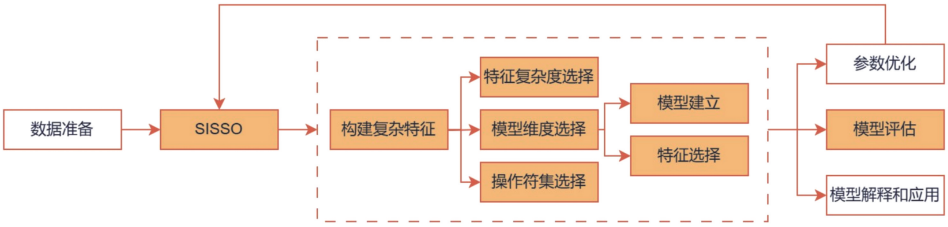

图 2　适应不同体系数据特点 SISSO 特征工程模型的建立策略

Fig. 2　Establishment strategy of SISSO feature engineering model adapted to the characteristics of different system data

2.2　参数优化模型驱动方法

参数优化是改善模型性能和泛化能力的重要环节。在参数优化过程中，可以尝试不同的参数组合，并评估每个组合的效果。通过比较不同参数配置的性能，选择具有最佳性能的参数配置。针对给定任务，模型驱动机器学习方法的基本过程包括构建模型族和设计求解模型族的算法集合。根据任务背景，如目标、物理机制和先验知识，构建的模型族是具有大量未知参数的函数族，相当于机器学习中的假设空间。与模型驱动方法中的精确模型不同，模型族仅提供解空间非常粗略和广泛的定义，具有模型驱动方法的优点，大大降低了精确建模的压力。算法集合是指具有未知参数的算法，所有参数都是可学习的。

本文选取 SISSO 和主动学习作为基本算法集合进行研究。SISSO 适用于处理高维但特征相对稀疏的数据集，通过筛选出与目标变量最相关的少数特征，减少模型的复杂性和过拟合风险，在小样本数据集，尤其是需要高可解释性的场景中表现出色。引入分衡策略与分布式并行的 SISSO 模型，能更好地适应来自不同环境中更大体系的数据集，缓解不同来源或体系数据集特征关系复杂或数据分布不均匀的问题。聚类相关的分组计算模式可以将高度相关的特征进行分组，减少冗余特征的计算，可针对特征之间存在显著相关性的高维数据集快速构建出简洁且高效的特征模型，提升建模效率，减少计算成本。通过结合这些算法的特性，特征工程能够在不同类型和规模的数据集上，构建出性能更优、效率更高的模型，助力科学研究和工程应用中的数据分析和决策过程。

3　材料数据的特征工程融合主动学习方法

3.1　分衡 SISSO 算法设计

基于多体系数据的可分性和 SISSO 算法各阶段的独立性，将 SISSO 算法逻辑划分为子集筛选层和总集建模层，这种层次划分为分布式结构的设计提供了基础。在节点间采用平衡机制提高特征筛选空间中子集特征的信息量，平衡子集与总集的特征选择。针对各节点内不同数据子集特征信息的不同贡献度，引入特征权重动态调整策略，充分利用各子集的特征信息。图 3 是钙钛矿材料分衡 SISSO 算法的调度框架图，其主要思想为“分”和“衡”两个层面。

在分衡 SISSO 算法中，以“分”为基础，将算法逻辑分为子集的特征筛选以及总数据集的模型构建两个部分，基于主从结构的通信模型重新设计了 SISSO 算法。主节点负责将数据集按材料体系划分为若干子集，并将这些子集分配到不同的工作节点，每个工作节点独立处理其分配的子集数据。主节点不仅负责任务分配，还协调和管理整个计算过程，包括结果的收集及最终模型的整合。SISSO 算法充分利用分布式结构，实现高效计算任务的分配与处理。每个工作节点专注于特定子集的特征处理，避免在全局范围内处理大量数据所带来的计算负担和复杂性。这种设计显著提升了算法的计算效率和扩展性，能够处理更大规模的数据集。在子集筛选层，每个工作节点执行特征空间构造和特征选择计算，分布计算以提高特征构造及筛选的效率。在总集建模层，主节点 man-

ager 负责合并子集特征筛选集合，建立子模型，向工作节点发送模型残差。工作节点迭代计算特征筛选集并发送主节点。

“衡”作为分衡 SISSO 算法的重要组成部分，在多体系数据集中实现特征信息的平衡，降低数据异质性对建模精度的影响。从节点内和节点间两个层面对平衡机制进行重新设计，增强模型对不同数据子集的适应性，提升算法处理复杂材料数据的性能。在节点内采用动态权重调整策略(主节点 manager 除外)解决节点内子集数据质量差异导致的特征信息贡献度不一致的问题。节点间采用平衡机制协调各子集与总集的特征信息，实现全局模型特征信息的平衡。计算节点构造不同数据子集特征空间，根据上一轮的残差迭代计算不同体系数据的特征筛选集合，合并扩展总数据的特征筛选空间，平衡总数据集与子集的特征选择，提高模型对不同子集数据的拟合性能和泛化能力。

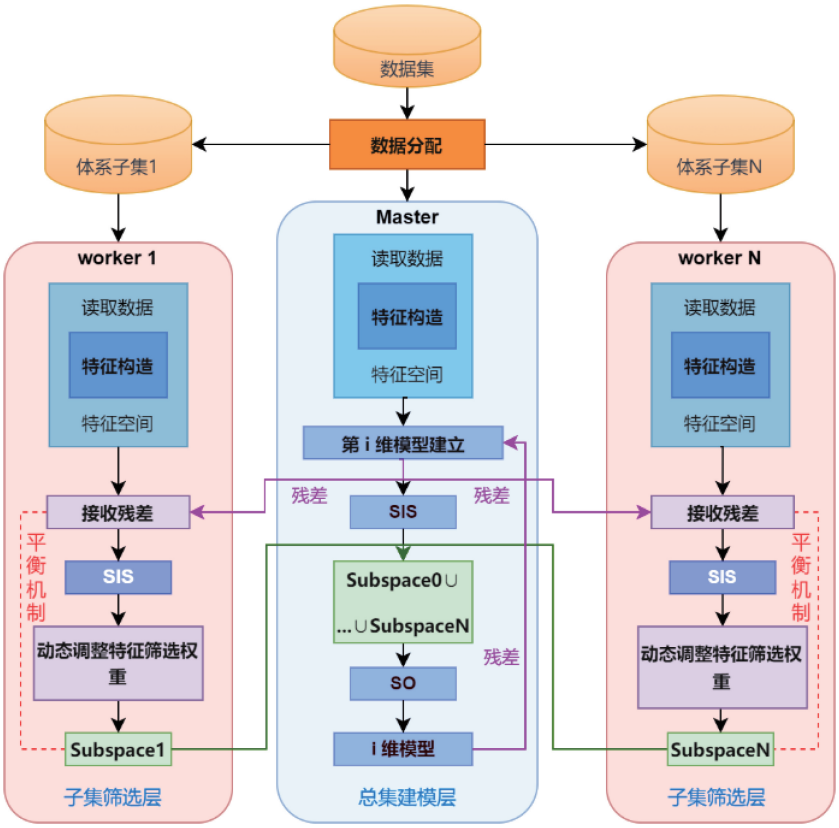

图 3 钙钛矿材料分衡 SISSO 算法框架图

Fig. 3 Perovskite material separation SISSO algorithm frame diagram

3.2 主动学习和特征工程的融合框架

针对材料数据结构复杂以及噪声多的特点，构建一种面向材料数据的主动学习方法，使用基于少数领域知识的多重分层采样模块确保初始阶段样本具有代表性、多样性和覆盖性，通过异常点检测模块剔除可能不符合真实数据分布的异常点，应用贝叶斯优化模块找到未标记数据中具有最大潜力的样本，在有效剔除噪声的同时解决材料数据难以标记的问题。

3.2.1 多重分层采样模块

初始样本选择是主动学习中的一个关键步骤，它对模型的性能和效果有直接影响。合适的初始样本能够为模型提供充足的信息和指导，使得模型能够更快地收敛并提高学习效率。多重分层采样模块根据目标以及自定义组分重要性对数据集进行分层均匀采样，有效覆盖关键目标及组分范围。

设计一个含有 M 个未标记样本，K 个标记数据的数据集的初始采样模块。首先，根据领域知识和目标需求，通过人工选择 N 个可能对目标影响较大的组分，多重分层采样模块中的分层次数为 $N+1$。将所有样本按照目标变量数值进行升序排序，并根据目标范围间隔 S 划分为若干个采样组，S 是每个采样组内目标变量最大值与最小值之间的差距，其中 S 与未标记数据个数 M 及最终标记数据个数 K 的关系如式(1)所示：

$$S = \frac{M}{K} \tag{1}$$

采样时还需考虑目标范围分布情况，根据每个采样组内

的样本数目，设置采样权重，使每个采样组的权重与组内样本的数量成正比。使用随机采样方式对每个组进行采样，可以较好地实现采样平衡，并确保样本覆盖整个目标范围。第 i 层采样总数如式(2)所示：

$$K_i = K(N - i + 2),\ 1 \leqslant i \leqslant N+1 \tag{2}$$

最后，根据选择的重要组分及浓度范围，依次进行多轮分层采样。当迭代次数 $i=N+1$，输出最终的 K 个初始标记数据。通过迭代，每一轮采样都能在前一轮基础上获取新的信息，覆盖数据集不同范围的特征和目标属性，保持样本的代表性和多样性。

3.2.2　异常点检测模块

异常点检测模块旨在剔除不符合真实数据分布的异常点，以减少其对模型训练和预测的负面影响。该模块利用任务模型与有标记数据的信息，构造一个二分类问题来判断数据样本是否为离群点。假设任务模型已经在有标记数据集上进行了训练，那么有标记数据真实值与任务模型预测值之间的误差越大，则越有可能是离群点。构造一个决策函数用于判断新的样本是否为异常点，见下式：

$$\hat{y_i} = \begin{cases} 1 & \text{if } \varphi_i < \text{threshold} \\ -1 & \text{othersise} \end{cases} \tag{3}$$

$$\varphi_i = (y_i - \hat{y_i})^2 \tag{4}$$

如果决策函数输出为 1，则该样本为正常样本；如果输出为-1，则为异常样本。其中，φ_i 表示任务模型的真实值与预测值差值的平方，y_i 是当前有标记数据的真实值，$\hat{y_i}$ 是任务模型对当前有标记数据的预测值。阈值 threshold 由所有标记数据误差的均值 μ 和标准差 σ 决定，见式(5)~式(7)：

$$\text{threshold} = \mu + k\sigma \tag{5}$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N}\varphi_i \tag{6}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\varphi_i - \mu)^2} \tag{7}$$

基于有标签的 N 个样本训练异常点检测分类模型，对所有无标记数据进行异常点预测，根据其 φ_i 是否大于阈值 threshold 来判断样本是否为异常点。其中 k 是一个超参数，用于控制阈值的灵敏度。较大的 k 会提高阈值，减少检测出的异常点数量，适用于数据误差分布集中且噪声较少的情况；较小的 k 会降低阈值，增加检测出的异常点数量，适用于数据误差分布广泛且噪声较多的情况。该模型能够根据不同数据集确定异常点数量，适应不同噪声程度的数据集。

3.2.3　贝叶斯优化模块

面向材料数据的主动学习方法结合了贝叶斯优化算法集合，通过多次迭代，根据先前观测到的样本点更新代理函数，选择下一个最有希望的采样点，以逐步降低目标函数总损失，并使代理函数逼近真实的目标函数。代理函数充当了真实目标函数的近似，根据贝叶斯优化的思想在不断观测到新的样本点后进行更新。引入期望改进函数(expected improvement，EI)作为采集函数，根据代理函数的预测值和不确定性来衡量每个样本点的潜在改进程度。

贝叶斯优化的整体过程包括：首先选择一个任务模型来对有标记数据建模构建一个代理函数。然后，根据观测到的数据，利用贝叶斯规则更新代理模型的后验分布。这个后验分布包含了在观测到数据后对未知函数的新认识，同时也提供了对函数值的预测以及预测的不确定性。接着，使用采集函数 EI，在考虑当前已观测的数据基础上，通过平衡对新区域的探索与已有信息的利用，决定下一个最有希望的采样点。最后，将新采样的数据添加到已观测的数据集中，重复上述步骤，直到达到最大迭代次数，使代理函数接近目标函数。

3.3　主动学习和特征工程的融合框架实现流程

主动学习和特征工程的融合框架通过结合智能样本选择和有效特征提取，实现更有效可靠的模型优化和训练过程。主动学习和特征工程的融合是一个迭代优化的过程，通过不断迭代优化，逐步改善模型的性能。在每一轮迭代中，主动学习算法能够有效选择最具代表性和信息量的样本进行标注，最大限度地提高数据的利用效率，减少需要标注的样本数量。在样本选择过程中，引入异常点检测模块，过滤掉不符合真实数据分布的样本，确保数据质量，进一步提高模型的鲁棒性。通过使用特征工程方法基于标记的样本筛选出最相关的特征，为机器学习模型提供更好的输入。这不仅能提高模型的精度，还能增强模型的可解释性，帮助理解数据中的关键因素。数据流示意图如图 4 所示。

首先，基于多重分层采样模块从未标记数据中采样，选取初始标记数据。将标记好的数据输入特征工程模型中，进行特征选择和特征提取。通过特征工程，获得经过处理和优化的标记数据，为模型提供高质量的输入。使用更新的标记数据训练任务模型，并用训练好的模型对预测集进行预测，评估模型的性能。用基于特征工程更新的标记数据训练异常点检测分类模型，对未标记数据进行异常点检测，预先去除未标记数据中的异常值。最后，利用贝叶斯优化模块从未标记数据中选择下一步的采样点，将选中的采样数据加入标记数据集中进行迭代。

整个过程持续评估模型的性能，根据评估结果进一步优化特征选择和样本选择策略。通过不断循环上述过程，模型的精度和可靠性逐步提升，构建主动学习和特征工程的融合框架，实现样本选择与特征提取的高效结

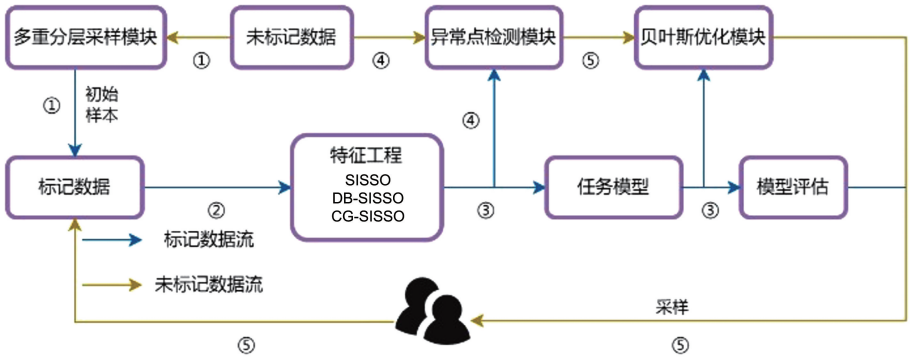

图 4 钙钛矿材料主动学习和特征工程的融合框架数据流示意图

Fig. 4 Data flow diagram of fusion framework of active learning and feature engineering of perovskite materials

合。智能计算框架可以融合 SISSO、分布式 SISSO(DB-SISSO)和聚类分组 SISSO(CG-SISSO)等多种特征工程算法，构建多种特征工程和机器学习模型，适用于从特征提取到模型训练的全流程。

4 针对钙钛矿材料的实验结果及分析

钙钛矿型三元系压电陶瓷数据集[23]由不同体系 $BiFeO_3$-$PbTiO_3$-$BaTiO_3$ (BF-PT-BT)、$BiFeO_3$-$PbTiO_3$-$BaZrO_3$(BF-PT-BZ)和 $BiFeO_3$-$PbTiO_3$-Ba(Zr, Ti)O_3(BF-PT-BZT)子集合并构成，分别用 BT、BZ 和 BZT 指代。相关特征描述符及其物理意义如表 1 所示。

表 1 特征描述符及其物理意义

Table 1 Feature descriptor and its physical meaning

Descriptor	Name	Physical meaning
μ	Reduced mass	Reduced mass of atoms
A/B	Ionic radius	Shannon ion radius
A/B_NU	NUnfilled	The number of unfilled orbitals that have been occupied by electrons
Ba, Pb, Bi	Element content	Element content ratio of metal ions

4.1 分布式 SISSO 特征工程对钙钛矿居里温度的预测

利用 DB-SISSO 特征工程模型，以钙钛矿数据集为研究对象，探索特征和目标(居里温度 T_C)之间的关系，构建钙钛矿材料的基础描述符和高精度预测模型，并通过外部数据集进行验证。

4.1.1 DB-SISSO 特征工程模型构建钙钛矿基础描述符

本实验基于领域知识(融合材料领域知识的数据准确性检测方法)，从钙钛矿数据集中筛选了 22 条 BT、BZ 和 BZT 数据作为训练集，该训练集均匀地覆盖了目标居里温度 T_C 的可调控范围，并且特征分布范围具有代表性。选择 5 组 BT 外部实验数据作为测试集，具体的数据收集于表 2 中。为了构建基础描述符，基于上述训练集建立 DB-SISSO 特征工程模型，获得了新的组合后的描述符，分别拟合出的 i 维预测模型如式(8)所示：

$$y_i = a_i x_1 + b_i x_2 + \cdots + d_i x_i + e_i \tag{8}$$

其中，x_1，…，x_i 为 i 维模型构建的高维描述符，a_i，…，d_i 分别是 i 维高维描述符的系数，e_i 则是 i 维模型的截距，基于 DB-SISSO 建立四维模型，每一维预测模型得到的基础描述符及相关系数如表 3 所示。表 4 展示了不同维度 DB-SISSO 模型的拟合误差。其中，四维模型的均方根误差和最大误差较小，拟合精度较高。

根据表 3 中出现次数最多的描述符初步推断这些描述符与居里温度 T_C 之间存在一定的关联，因为它们在建模过程中出现的频率较高。出现次数最多的描述符有 μ、

表 2 测试数据集详情

Table 2 Test data set details

Sample	$T_C(y)$	$\mu(x_1)$	A/B(x_2)	A/B_NU(x_3)	Ba(x_4)	Pb(x_5)	Bi(x_6)
1	547	0.2821	0.3882	0.5036	0.0750	0.1150	0.1300
2	540	0.2690	0.4606	0.5000	0.0750	0.1250	0.3000
3	502	0.2295	0.6789	0.4879	0.0750	0.1550	0.2700
4	600	0.3404	0.0778	0.5330	0.0500	0.1100	0.3400
5	600	0.3658	−0.0651	0.5625	0.0500	0.0900	0.3600

Notes: 1: 0.62BF-0.23PT-0.15BT, 2: 0.60BF-0.25PT-0.15BT, 3: 0.54BF-0.31PT-0.15BT, 4: 0.68BF-0.22PT-0.10BT, 5: 0.72BF-0.18PT-0.10BT

表 3　DB-SISSO 在不同模型维度下的基础描述符及相关系数

Table 3　The basic descriptors and correlation coefficients of DB-SISSO under different model dimensions

Model dimension	Basic descriptor	Correlation
1D	x_1: [(exp((A/B)/(A/B_NU)) * ((Ba−μ) * (Ba/μ)))]	a_1: 534.933 623 3 e_1: 608.856 365
2D	x_1: [(((Ba−μ) * (Ba/(A/B_NU)))/((μ)^3/ln(μ)))] x_2: [(abs(A/B_NU−abs(μ−A/B))/abs((A/B_NU-Ba)−abs(Bi−A/B)))]	a_2: −37.375 715 60 b_2: 6.105 165 938 e_2: 602.943 005 6
3D	x_1: [(exp(((A/B)/(A/B_NU))) * ((Ba−μ) * (Ba/μ)))] x_2: [(abs((A/B−Ba)−(Pb+A/B_NU))/abs((A/B−Bi)−(A/B_NU−Ba)))] x_3: [abs(((Ba/μ) * (A/B_NU−μ))−abs((Pb−μ)−(μ−A/B_NU)))]	a_3: 530.193 351 2 b_3: 7.097 671 652 c_3: −116.822 686 6 e_3: 604.462 924 9
4D	x_1: [(exp((A/B)/(A/B_NU)) * (Ba−μ) * (Ba/μ))] x_2: [(abs((A/B−Ba)−(Pb+A/B_NU))/abs((A/B−Bi)−(A/B_NU−Ba)))] x_3: [abs(((Ba/μ) * (A/B_NU−μ))−abs((Pb−μ)−(μ−A/B_NU)))] x_4: [(((A/B_NU−A/B)/(Ba)^2)/(((A/B)/μ))−(μ/Pb))]	a_4: 527.193 668 8 b_4: 7.128 221 428 c_4: −102.349 294 7 d_4: −0.024 941 821 e_4: 602.972 244 2

表 4　DB-SISSO 在不同模型维度下的拟合误差

Table 4　The fitting error of DB-SISSO under different model dimensions

Model dimension	RMSE	MaxAE	R^2
1D	2.952 757	9.238 664	0.995 932
2D	1.592 191	4.171 645	0.998 817
3D	0.904 270	2.011 015	0.999 618
4D	0.620 823	1.318 696	0.999 820

A/B_ NU 以及 Ba，分别出现了 20，15 和 16 次。这些高频出现的描述符对居里温度的影响较大，可以作为指导钙钛矿材料设计和发现的重要参考。

4.1.2　钙钛矿多维描述符模型验证

为了对 DB-SISSO 特征工程模型的性能进行评估，并确定它在处理外部数据集时的准确性和可靠性。使用测试集对模型进行验证，评估模型的泛化性能。图 5 显示了训练集以及测试集在 DB-SISSO 不同维度模型上的验证结果。从图 5 可以看出模型在处理外部数据集时具有较高的准确性和可靠性，且整体精度满足了目标要求。表 5 给出了 5 组测试集数据在 DB-SISSO 拟合的四维模型的验证结果，并使用不同指标对其进行评估。平均相对误差 MRE 的定义如式(9)所示：

$$\mathrm{MRE} = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{9}$$

其中，$\hat{y}_i$ 表示预测值，y_i 表示实验值，n 为样本的数量。MRE 表示预测值与真实值之间的相对误差的平均值，用于评估预测模型的准确性。根据式(9)，验证集的平均相对误差计算结果为 0.91%，表明 DB-SISSO 在整体上具备较好的预测性能。

4.2　钙钛矿材料主动学习和特征工程的融合框架

支持向量回归(support vector regression，SVR)在小样本数据场景下表现良好，使用 SVR 作为任务模型，并选用支持向量机(support vector machine，SVM)作为异常点检测模块辅助模型。随机采样(random sampling，RS)是从未标注样本池中随机筛选出一批样本进行标注，广泛用作主动学习中的基准方法。贝叶斯优化是一种用于黑箱函数优化的有效方法，通过优化获得的函数可以在复杂高维空间中高效搜索最优值，在材料研究领域广泛用于发现目标功能材料。面向材料数据的主动学习方法结合贝叶斯优化筛选策略可以提高算法的效率和有效性。

4.2.1　钙钛矿材料实验结果

在钙钛矿材料数据集上进行实验，对面向材料数据的主动学习方法进行性能评估，选择基于随机采样和贝叶斯优化的主动学习方法作为对照组进行对比实验，并分别对结合 SISSO 特征工程的 3 种框架进行对比实验。图 6 展示了面向材料数据的 3 种主动学习方法(active learning and feature engineering for material data)和 3 种主动学习与特征工程融合计算框架(active learning and feature engineering)的对比实验结果。相比根据随机采样获得的初始样本，面向材料数据的主动学习方法在初始采样时就能拥有更好的预测精度。面向材料数据的主动学习方法增加了异常点检测模块，减少了标注异常点的概率。主动学习算法集与 SISSO 特征工程融合框架的预测精度得到显著提升。

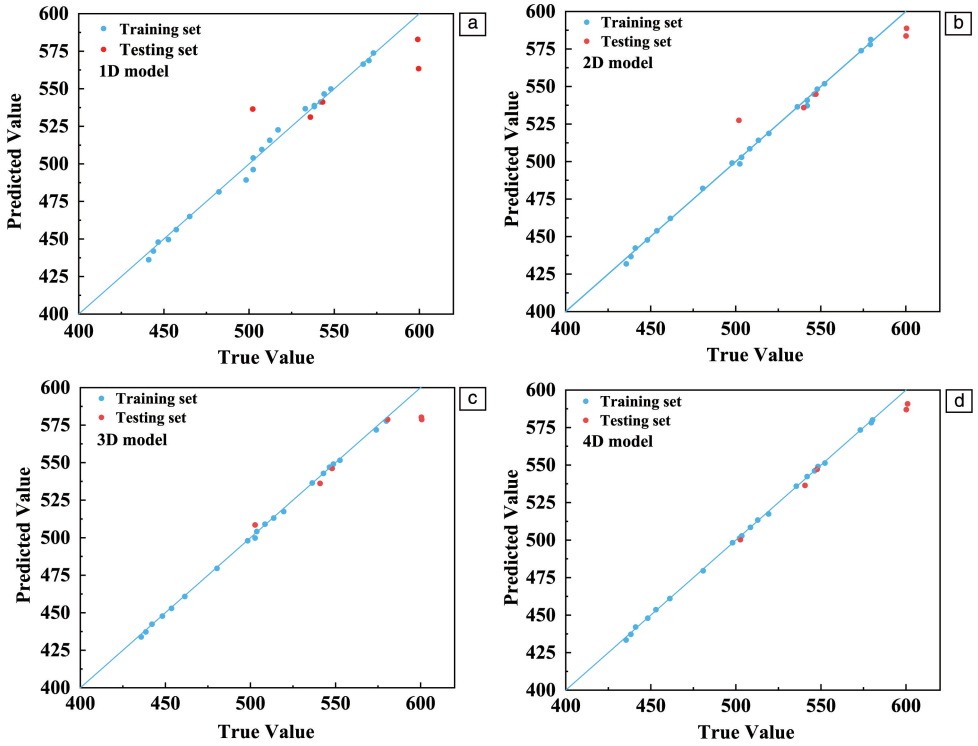

图 5 训练集以及测试集在 DB-SISSO 不同维度模型下的验证结果

Fig. 5 Verification results of the training set and testing set under the DB-SISO different dimension models

表 5 外部验证集预测结果

Table 5 External validation set prediction results

Sample	Experimental value T_C/℃	Predicted value T_C/℃	Absolute error/℃	Relative error/%
1	547	547. 14	0. 14	0. 02
2	540	536. 71	3. 29	0. 61
3	502	500. 81	1. 19	0. 24
4	600	587. 38	12. 62	2. 10
5	600	590. 57	9. 43	1. 57

Notes: 1: 0. 62BF-0. 23PT-0. 15BT, 2: 0. 60BF-0. 25PT-0. 15BT, 3: 0. 54BF-0. 31PT-0. 15BT, 4: 0. 68BF-0. 22PT-0. 10BT, 5: 0. 72BF-0. 18PT-0. 10BT

4. 2. 2 外部材料验证结果

为了验证模型对于外部数据的有效性，选取了 103 条公开的混凝土数据作为验证对象，采取随机选取的方式选择 5 个样本作为测试集，并使用剩余样本建立的模型作为基准线，测试集的验证结果如图 7 所示。与钙钛矿材料数据集的结果类似，随着标注样本的增加，面向材料数据的主动学习方法的预测精度也在不断地上升，进一步融合 SISSO 特征工程后，该主动学习方法的预测精度得到了显著提升。总之，主动学习融合特征工程的智能计算框架对混凝土数据集也具有不错的预测精度，对不同材料数据具有较好的普适性。

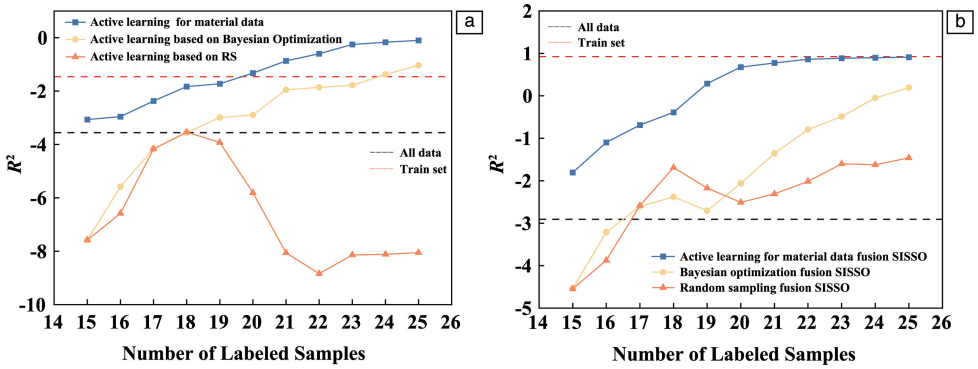

图 6 3 种主动学习(a)和 3 种主动学习算法结合 SISSO 特征工程(b)方法在钙钛矿材料数据集和训练集上的表现

Fig. 6 The performance of three active learning (a) and three active learning algorithms combined with SISSO feature engineering (b) methods on perovskite material data sets and training sets

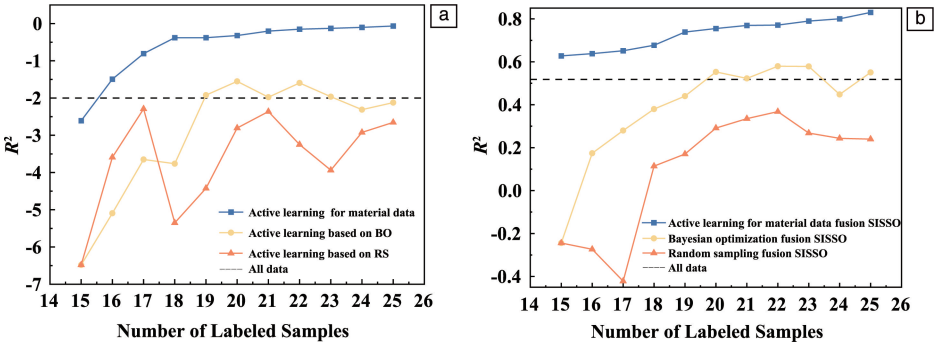

图 7　3 种主动学习(a)和 3 种主动学习算法结合 SISSO 特征工程(b)方法在混凝土数据集和训练集上的表现

Fig. 7　The performance of three active learning (a) and three active learning algorithms combined with SISSO feature engineering (b) methods on concrete material data sets and training sets

5　结　论

本文采用数据和模型双驱动的思路，通过结合特征工程的特征选择能力与主动学习的样本选择策略，提出了主动学习与特征工程的融合计算框架。通过引入分衡策略与分布式并行模型，处理特征关系复杂或数据分布不均匀问题。分布式确定独立筛选和稀疏算子(DB-SISSO)计算模式尤其适合复杂数据情景，在多体系合并钙钛矿整体数据集上获得了反映组分-居里温度关系的高精度预测模型。主动学习方法可以用于解决材料数据噪声及高昂的标记成本问题，在降低数据标记成本的同时提升模型精度。主动学习融合特征工程的钙钛矿材料智能计算框架集成算法集和数据智能标记技术可以提高模型精度、泛化能力和计算效率，有助于加快高性能钙钛矿材料的发现和探索。

参考文献　References

[1] POLLICE R, GOMES G, ALDEGHI M, *et al*. Accounts of Chemical Research[J], 2021, 54(4): 849-860.

[2] SCHLEDER G R, PADILHA A C M, ACOSTA C M, *et al*. Journal of Physics: Materials[J], 2019, 2(3): 032001.

[3] ZHANG J, YU Y, ZHANG L, *et al*. Nano Energy[J], 2023, 114: 108656.

[4] PILANIA G. Computational Materials Science[J], 2021, 193: 110360.

[5] QADIR A, ALI S, DUSZA J, *et al*. Open Ceramics[J], 2024, 19: 100634.

[6] CAI J, CHU X, XU K, *et al*. Nanoscale Advances[J], 2020, 2(8): 3115-3130.

[7] CHAN C H, SUN M, HUANG B. EcoMat[J], 2022, 4(4): e12194.

[8] STERGIOU K, NTAKOLIA C, VARYTIS P, *et al*. Computational Materials Science[J], 2023, 220: 112031.

[9] 刘悦, 邹欣欣, 杨正伟, 等. 硅酸盐学报[J], 2022, 50(3): 863.
LIU Y, ZOU X X, YANG Z W, *et al*. Journal of the Chinese Ceramic Society[J], 2022, 50(3): 863.

[10] 刘悦, 刘大晖, 葛献远, 等. 物理学报[J], 2023, 72(7): 070701.
LIU Y, LIU D H, GE X Y, *et al*. Acta Physica Sinica[J], 2023, 72(7): 070701.

[11] 刘悦, 马舒畅, 杨正伟, 等. 硅酸盐学报[J], 2023, 51(2): 427.
LIU Y, MA S C, YANG Z W, *et al*. Journal of the Chinese Ceramic Society[J], 2023, 51(2): 427.

[12] 施思齐, 孙拾雨, 马舒畅, 等. 无机材料学报[J], 2022, 37(12): 1311.
SHI S Q, SUN S Y, MA S C, *et al*. Journal of Inorganic Materials[J], 2022, 37(12): 1311.

[13] LIU Y, YANG Z, ZOU X, *et al*. National Science Review[J], 2023, 10(7): 125.

[14] OUYANG R H, CURTAROLO S, AHMETCIK E, *et al*. Physical Review Materials[J], 2018, 2(8): 083802.

[15] SMITS G F, KOTANCHEK M. Genetic Programming Theory and Practice II[J], 2005, 8: 283-299.

[16] ZHAO S, ZHANG Y, ZHANG Y, *et al*. Acta Materialia[J], 2022, 228: 117791.

[17] CANDES E J, WAKIN M B. IEEE Signal Processing Magazine[J], 2008, 25(2): 21-30.

[18] NELSON L J, HART G L W, ZHOU F, *et al*. Physical Review B[J], 2013, 87(3): 035125.

[19] OUYANG R, AHMETCIK E, CARBOGNO C, *et al*. Journal of Physics: Materials[J], 2019, 2(2): 024002.

[20] OUYANG R H. Chemistry of Materials[J], 2020, 32(1): 595-604.

[21] BARTEL C J, SUTTON C, GOLDSMITH B R, *et al*. Science Advances[J], 2019, 5(2): eaav0693.

[22] 胡红青, 吴邵刚, 郭治廷, 等. 中国有色金属学报[J], 2020, 30(8): 1887-1894.
HU H Q, WU S G, GUO Z T, *et al*. The Chinese Journal of Nonferrous Metals[J], 2020, 30(8): 1887-1894.

[23] 焦志翔, 贾帆豪, 王永晨, 等. 无机材料学报[J], 2022, 37(12): 70+1322-1328.
JIAO Z X, JIA F H, WANG Y C, *et al*. Journal of Inorganic Materials[J], 2022, 37(12): 70+1322-1328.