

单位代码：11232

分类号：TP391

密级：公开



北京信息科技大学

专业硕士学位论文

面向材料科学的多保真度数据 学习策略研究

学 院：计算机学院

专 业：电子信息

学 号：2022020654

作 者：王滋明

学校指导教师：杨涛 助理研究员

企业指导教师：刘金家 高级工程师

完成日期：二〇二五年三月二十四日

学位论文版权使用授权书

本人完全了解北京信息科技大学关于收集、保存、使用学位论文的规定，按照学校要求提交学位论文的印刷本和电子版本。学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。学校有权适当复制、公布论文的全部或部分内容。学校有权将本人的学位论文加入《中国优秀硕士学位论文全文数据库》和编入《中国知识资源总库》。

学位论文作者签名：王涵明

2025 年 3 月 24 日

☒公开 ☐保密（____年____月）（保密的学位论文在解密后应遵守此协议）

指导教师签名：杨涛

学位论文作者签名：王涵明

2025 年 3 月 24 日

2025 年 3 月 24 日

硕士学位论文原创性声明

本人郑重声明：所呈交的论文题目为《面向材料科学的多保真度数据学习策略研究》学位论文，是本人在导师指导下，进行研究工作所取得的成果。尽我所知，除了文中特别加以标注的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或者没有公开发表的作品的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明并表示了谢意。本学位论文原创性声明的法律责任由本人承担。

作者签字：王溢明

2025 年 3 月 24 日

摘要

多保真度数据是指具有不同精度、数量和获取成本的数据集合，它们广泛存在于各大领域中，合理利用多保真度数据有助于提高数据驱动机器学习模型的预测精度。当前，材料科学领域中有关多保真度数据的研究多集中在模型层面，本研究则从数据修正的层面探讨其对模型精度的影响。所涉及的多保真度数据集来自人为构造、理论计算及实验。

考虑到多保真度数据学习方法强调区别对待不同保真度的数据，因此，本研究首先探索数据保真度的评价方法，采用领域专家知识和模型这两种方法评价数据保真度。实验表明，不同评价方法直接影响数据保真度，合理的评价方法使得预测结果更可信。此外，在使用模型评价方法时，本研究提出了一种基于材料图神经网络的不确定性量化方法，并利用其不确定性初步探索了多保真度数据学习方法。

之后，本研究更加详细地探索了线性修正和非线性修正的多保真度数据学习方法。在含噪程度、类型不同的多保真度数据集下定量评价了数据学习方法对模型精度的提升规律。实验表明，多保真度数据学习方法在不同场景下对模型预测精度的提升规律不同。综上所述，本研究的主要创新之处体现在以下几个方面：

（1）基于材料图神经网络的不确定性量化方法。赋予材料图神经网络预测不确定性的能力，同时在一定程度上提高了预测精度。

（2）多保真度数据学习方法的扩展。提出基于不确定性加权修正与迭代降噪的多保真度数据修正方法，扩展了多保真度数据学习方法的实现策略。

（3）含噪多保真度数据集下数据学习方法对模型提升规律的探索。系统评估了线性修正与非线性迭代降噪方法在提升模型预测精度方面的适用性，揭示了各方法在不同噪声条件下的优劣。

关键词：多保真度数据；数据噪声；不确定性量化；机器学习；属性预测

ABSTRACT

Multi-fidelity data refers to data sets with different precision, quantity and acquisition cost. They are widely used in various fields. Rational use of multi-fidelity data can help improve the prediction accuracy of data driven machine learning model. At present, the research on multi-fidelity data in the field of materials science mostly focuses on the model level. This study discusses its impact on model accuracy from the level of data correction. The multi-fidelity data set involved comes from artificial construction, theoretical calculation and experiments.

Considering that the multi-fidelity data learning method emphasizes to treat data with different fidelity differently, this study first explores the evaluation method of data fidelity, and uses domain expert knowledge and models to evaluate data fidelity. The experiment shows that the reasonable evaluation method makes the prediction results more credible. In addition, when using the model evaluation method, this study proposes an uncertainty quantification approach based on a materials graph neural network and utilizes its uncertainty to preliminarily explore multi-fidelity data learning methods.

Subsequently, this study delves deeper into multi-fidelity learning methods using both linear and nonlinear correction approaches. The effectiveness of these methods in improving model accuracy is quantitatively assessed across multi-fidelity datasets with different noise levels and types. Experiments show that the improvement patterns of model prediction accuracy through multi-fidelity data learning methods vary across different scenarios. In summary, the main innovations of this study include the following:

(1) Uncertainty quantification method based on material graph neural network. This method endows materials graph neural networks with the ability to predict uncertainty while also enhancing prediction accuracy to some extent.

(2) Extension of multi-fidelity data learning methods. We propose a multi-fidelity data correction approach that integrates uncertainty-weighted correction with iterative denoising, thereby broadening the implementation strategies for multi-fidelity data learning.

(3) Investigation of the effectiveness of multi-fidelity data learning methods in noisy datasets. We systematically evaluate the applicability of linear correction and

nonlinear iterative denoising methods in improving model prediction accuracy, revealing the strengths and limitations of each approach under different noise conditions.

KEY WORDS: multi-fidelity data; data noise; uncertainty quantification; machine learning; property prediction

目 录

第 1 章 引 言	1
1.1 研究背景及意义	1
1.2 国内外研究现状	2
1.2.1 来自不同方法的多保真度数据	3
1.2.2 来自同一方法不同参数的多保真度数据	4
1.2.3 多保真度数据在材料领域中的应用	5
1.3 研究内容	6
1.3.1 数据保真度评价方法	6
1.3.2 多保真度数据学习方法	7
1.4 创新点	8
1.5 章节安排	9
第 2 章 关键理论与技术基础	11
2.1 相关性系数和方差膨胀因子	11
2.1.1 皮尔逊相关性系数	11
2.1.2 斯皮尔曼相关性系数	11
2.1.3 方差膨胀因子	12
2.2 高斯过程回归	12
2.3 材料图神经网络	13
2.3.1 MEGNet	13
2.3.2 CGCNN	15
2.4 本章小结	17
第 3 章 数据保真度评价方法	18
3.1 领域专家知识评价数据保真度	18
3.1.1 研究方案	18
3.1.2 领域专家知识评价方法	19
3.1.3 实验流程及结果	20
3.1.4 与传统理解的对比和讨论	25
3.1.5 小结	26
3.2 模型评价数据保真度	26
3.2.1 研究方案	26
3.2.2 基于材料图神经网络的不确定性量化方法	27
3.2.3 基于不确定性的加权修正方法	29

3.2.4 实验结果及分析.....	30
3.2.5 小结.....	32
3.3 本章小结	33
第 4 章 多保真度数据学习方法	34
4.1 线性修正方法	34
4.1.1 研究方案.....	34
4.1.2 线性修正方法.....	35
4.1.3 实验结果及分析.....	36
4.1.4 小结.....	38
4.2 多保真度数据迭代降噪方法	39
4.2.1 研究方案.....	39
4.2.2 训练策略及降噪方法.....	40
4.2.3 单保真度数据集大小及噪声的影响.....	41
4.2.4 不同训练策略的比较.....	42
4.2.5 迭代降噪的效果.....	43
4.2.6 小结.....	44
4.3 本章小结	45
第 5 章 总结与展望	46
5.1 本文工作总结	46
5.2 未来工作展望	49
致 谢.....	50
参考文献.....	51
个人简历、在学期间发表的学术论文及研究成果	56

第1章 引言

1.1 研究背景及意义

机器学习技术在解决现实问题中展现出巨大潜力,尤其在分类和预测问题的处理上,已显示出卓越的性能。然而,随着应用的深入,机器学习对数据的要求也日益增加,数据的质量和数量直接影响模型的效果。对于数据集的获取,采用相对简单、获取成本较低的方式能在单位时间内能够获得更多的数据,但其质量较低,通常被称之为低保真度 (Low-Fidelity, LF) 数据;相反,采用复杂、获取成本较高的方式在单位时间内获得的数据量相对较少,但其质量较高,通常被称之为高保真度 (High-Fidelity, HF) 数据。不同保真度的数据合称为多保真度 (Multi-Fidelity, MF) 数据,其保真度水平并非仅限于高低两种,往往呈现出多种级别^[1,2]。在多保真度数据的生产与消费过程中,存在“精度与计算资源之间的权衡”^[3,4],如图 1.1 所示。在生产阶段,若计算资源固定,采用高精度计算方式通常只能获得少量高质量数据;而采用低精度或近似计算方式,则可以获得更多的低质量数据。数据的消费环节中,使用不同精度与资源消耗的多保真度模型进行训练,也会影响模型的效率和准确度^[5,6]。

多保真度数据的计算资源-精度权衡

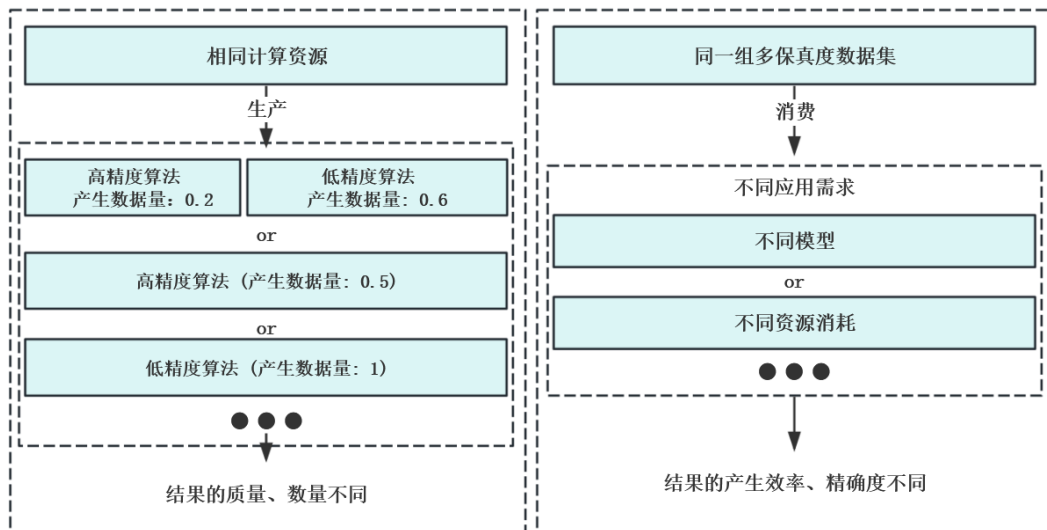


图 1.1 多保真度数据的“精度-计算资源权衡”

多保真度数据广泛存在于各个领域,但目前大多数机器学习模型仅关注单一保真度数据集上的训练效果^[7-9],这些模型要么面临数据不足的问题,要么面临数据质量较差的问题^[10]。采用同一标准笼统地对待多保真度数据集也会遗漏多

保真度数据中的一些信息，导致模型的训练效果不佳。因此，如何充分利用多保真度数据成为当前研究的热点问题，而多保真度技术也因此应运而生。

多保真度技术旨在合理利用多保真度数据中的信息，其强调区别对待多保真度数据来权衡计算资源和精度^[11-13]。从数据层面利用多保真度数据的方法被称为多保真度数据学习方法，其通常以修正低保真度数据的方式来提高模型预测精度。常见的多保真度数据学习方法是通过对代理模型构建高、低保真度数据之间的关系，以此利用多保真度数据信息^[10]。而代理模型，又称为响应曲面模型或者元模型，是对真实模型的一种简化，它无需了解输入变量与输出响应之间的物理意义，可以仅关注数据本身快速建模^[14]。目前，多保真度技术被广泛应用于各大领域，然而其在材料领域上的研究却较为缺乏。

对于材料领域来说，多保真度数据往往来自于理论计算（例如密度泛函理论）和实验（实验工具计算测量），其中实验得到的数据通常被认定为具有较高保真度。目前，从数据角度研究多保真度数据学习方法的相关工作仍较有限，大多数研究仍停留在模型层面，尚未充分挖掘材料数据中的潜在信息。因此，本研究将重点探讨如何从数据角度高效利用多保真度数据学习方法，涵盖数据保真度的评价方法及多保真度数据学习策略，系统分析不同学习方法对模型性能的影响与提升规律。

1.2 国内外研究现状

多保真度技术常出现在材料、航空航天科学、机械工程等领域中，目前关于多保真度数据的获取方式多样，它们普遍存在成本和数据量之间权衡的现象。其中，获取成本较低的数据集单位时间内获取的数据量较大，但是数据质量较低，即所谓的低保真度（Low-Fidelity, LF）数据；获取成本较高的数据集单位时间内获取的数据量较小，但是数据质量较高，即高保真度（High-Fidelity, HF）数据。其他领域中对多保真度技术的系统阐述和验证已有相关综述^[15-17]。然而，材料科学领域仍然缺乏关于多保真度数据学习方法的综述和验证。

因此，本研究首先统计了近年来不同领域下多保真度数据的获取方式。其中，多保真度数据的获取方式通常可以分为两种：不同方法和同一方法不同参数，具体内容如表 1.1 所示。无论在哪个领域，计算同一对象时往往会存在多种不同的计算结果，这些结果可能具有不同的精度和数量，因此产生了多保真度数据。本文将在后续小节中对这些获取方法进行介绍，并详细介绍多保真度数据在材料领域中的应用情况。

表 1.1 不同领域多保真度数据的不同获取方法

材料科学领域	不同方法	泛函和实验 ^[1, 2]
		泛函 ^[18 - 22]
		实验 ^[23]
	同一方法不同参数	物理模型 ^[24]
		泛函和物理模型 ^[25]
		网格大小 ^[26]
航天航空领域	不同方法	时间步长 ^[27, 28]
		不同超参数 ^[29]
		计算流体动力学方法 ^[30 - 37]
	同一方法不同参数	计算流体动力学工具 ^[38]
		计算流体动力学方法和实验 ^[39]
		多旋翼机动力学和实验 ^[40]
机械工程领域	不同方法	网格大小 ^[41 - 43]
		数学模型 ^[44 - 47]
		测量工具 ^[48, 49]
	同一方法不同参数	网格大小 ^[50 - 52]
		收敛条件和模拟对象 ^[53]
		迭代限制 ^[54]
其他领域	不同方法	数学模型 ^[55]
		数学模型和模拟工具 ^[56]
		主要或简化的数值模型 ^[57]
	同一方法不同参数	网格大小 ^[58]
		物理条件 ^[59]

1.2.1 来自不同方法的多保真度数据

在材料科学领域，多保真度数据的存在主要归因于不同方法的计算。这些方法通常涵盖了实验和理论计算。实验方法通常被认为比理论方法更耗时，但它们提供了更高的准确性。此外，即使在理论计算方法中，数据保真度也存在显著差异。与密度泛函理论（Density Functional Theory, DFT）相比，常用的基于经验的势模型一般被认为是低保真度方法，而 DFT 通常被视为高保真方法。但这并不意味着 DFT 计算的结果都会得到高保真度数据，其在计算某些属性时仍然表现乏力。例如，有文献报道，DFT 计算往往低估带隙宽度 30%到 100%^[60]。

此外，DFT 中也存在不同精度的泛函。例如表 1.2 中，PBE 泛函在某些情况下可以被归类为高保真，但在其他情况下可能被认为是低保真。在使用 DFT 计算材料性质时，与真实值相比，通常存在系统误差和随机误差。随机误差通常源于 DFT 计算中使用的初始试探电荷密度，不同的试探电荷密度可能导致不同的随机误差。此外，自洽过程中的不同收敛准则也可能引入随机误差。而系统误差

往往是在 DFT 计算哈密顿量过程中引入的。在计算交换-关联势并求解电子波函数的 Kohn–Sham 方程中，通常采用不同的近似方法，这些方法引入了不同的误差。误差的存在使得 DFT 计算与真实值之间产生偏差。对于相同的计算目标，不同泛函所引起的误差可以视为高保真与低保真之间的差异，而这些差异则可以通过多保真数据学习方法进行修正。

表 1.2 材料领域高/低保真度（HF/LF）数据采集方法的分布

文献	LDA ^[61]	PBE ^[62]	PBE0 ^[63]	HSE ^[64]	势模型 ^[65]
[18,19]		LF		HF	
[20]	LF	LF		HF	
[21]		LF	HF		
[22]		LF		HF	
[25]		HF			LF

此外，在航空航天领域中，不同的计算流体动力学（Computational Fluid Dynamics, CFD）方法或不同的 CFD 工具（如 Q3D 和 MATRICS-V^[38]）也会产生多保真度数据。Reynolds-Averaged Navier-Stokes（RANS）方法得到的结果通常是高保真数据^[35–37]，而势流模型的结果则对应低保真数据^[31,34]。在机械工程领域中，不同的数学模型（例如，不同的有限元模型^[44,47]和代理模型^[16,45,46]）或不同的测量工具（例如，不同的应变传感器^[48]和专用于风力涡轮机的气动伺服弹性计算模拟器^[49]）也会产生多保真度数据。在生物医学和经济领域，多保真数据的产生方式列于表 1.1 中。此外，相对于理论计算获得的 MF 数据，通过人工实验获得的数据更为准确，这些实验数据通常被视为具有最高保真度的数据^[1,39,40]。

1.2.2 来自同一方法不同参数的多保真度数据

对于不同超参数在相同方法下也可能生成多保真度数据，这里的超参数通常是一些影响计算复杂度的参数。如表 1.1 中所示，常见的一种方法是调整网格大小，其中较小的网格通常对应 HF，而较大的网格则对应 LF^[26,41,50]。在材料科学领域，通过使用相同的泛函但指定不同的参数（如平面波截断能量、k 点、原子弛豫等），得到的数据保真度也会有所不同^[29]。此外，多保真度数据也可能来自不同时间步长的模型^[27]或不同收敛条件及模拟对象获得^[53]。不同迭代次数也同样会导致生成多保真度数据，从部分收敛获得的数据^[66]可以视为 LF 数据，而完全收敛获得的数据可以视为 HF 数据。例如，在参考文献^[54]中，大约 600 次迭代和 100 次迭代的结果分别被视为 HF 和 LF。多保真度数据的获取方式决定了数据的保真度。通过不同方式获得的多保真度数据的保真度类型需要通过评估方法

或先验知识进行判断，而数据保真度评价方法的选择或先验知识的使用也会影响数据的保真度。

1.2.3 多保真度数据在材料领域中的应用

在材料科学领域，现有且被广泛使用的数据集包括：Materials Project (MP)^[67]、Open Quantum Materials Database (OQMD)^[68]等其他数据集。这些数据集通常由密度泛函理论 (DFT) 或其他理论方法计算而得，与现实世界中的真实属性值存在不同程度的误差。此外，对于某些材料的属性，甚至存在多种测量方法，导致结果不尽相同^[69]。这些来自不同方法得到的数据集构成了材料领域的多保真度数据集，使得多保真度数据广泛存在于材料领域中。

目前，数据驱动的机器学习研究在材料领域已经存在较多应用，用于预测属性的机器学习模型层出不穷，如本研究所使用到的 CGCNN^[70]和 MEGNet^[71]。MP 更是制作了由十余种 Benchmark 数据集组成的竞赛 MatBench^[72]供研究人员对各种属性预测模型进行公平评测。然而，目前传统的预测算法对待数据都并不精细，没能根据化学数据固有特点最大化利用数据中的信息。随着多保真度数据越来越受到重视，多保真度机器学习方法在材料领域中也有关工作涌现，所涉及的方法包括迭代降噪^[1]、贝叶斯优化^[58,59]、信息融合^[73,74]等。迭代降噪通过对特定排列的多保真度数据进行迭代训练，并在每次迭代后对特定数据进行降噪操作来提高模型的准确性。贝叶斯优化利用高低保真度数据之间的相关性，在迭代寻找最优值时减少了计算成本，提高了模型预测的效率。信息融合算法通过学习高低保真度数据之间的关系来融合数据信息，可以有效利用低保真度数据的信息。此外，2021 年 Chen 等人提出利用多保真度数据来训练属性预测模型^[2]。然而，作为 MEGNet 的作者，Chen 的思路依然是对模型的改进，发展出了 MFGNet (Multi-Fidelity materials Graph Networks)，即多保真度材料图网络。该模型最显著的改进便是为每种保真度的数据类型设置一个整数的全局状态值，然而其并没有针对多保真度数据本身进行修改。

多保真度机器学习技术在多个领域已有成功应用，在材料领域的应用尤为引人关注。然而，遗憾的是，目前针对多保真度数据本身进行研究的工作仍然较少，大多数多保真度技术的研究集中在模型层面。综上所述，尽管属性预测在材料领域已有大量实践，但从多保真度数据角度进行问题分析和研究仍处于初步探索阶段。因此，本研究将重点研究材料领域中多保真度数据，考虑到不同数据保真度评价方法的影响，从数据保真度评价方法和多保真度数据学习方法两大方面进行研究，探索多保真度数据学习方法对于模型的提升规律。

1.3 研究内容

多保真度数据在材料领域中得到了广泛应用，本研究从数据保真度的评价方法和多保真度数据学习方法两个方面展开。鉴于多保真度技术强调针对性地处理不同保真度数据，本研究首先聚焦于数据保真度的评价方法，探讨领域专家知识和模型评价两种方法。其中，本研究在模型评价方法中使用了不确定性加权修正，初步探索了多保真度数据学习方法，为后续进一步研究奠定基础。随后，本研究探讨了线性修正和非线性修正的多保真度数据学习方法，旨在发现不同多保真度数据学习方法在不同场景下对模型的提升规律。整体研究架构如图 1.2 所示，后续小节将对其进行介绍。

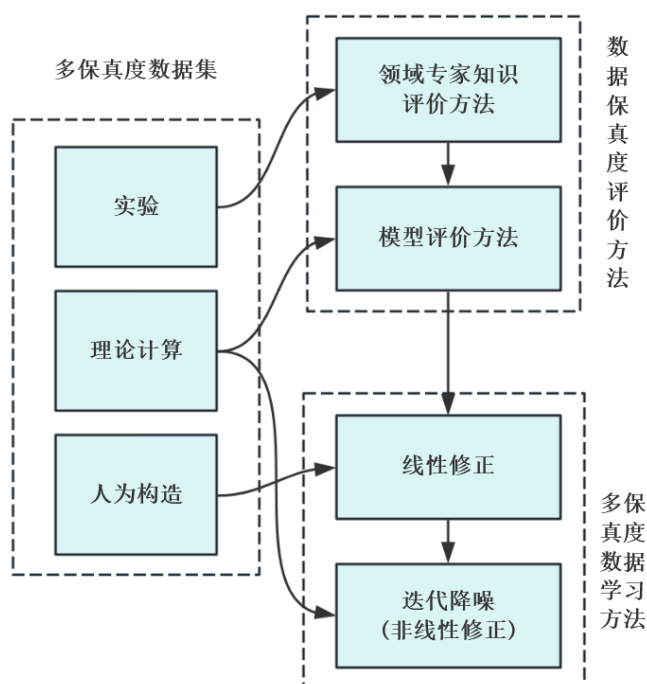


图 1.2 研究内容整体架构

1.3.1 数据保真度评价方法

多保真度技术强调针对性地处理不同保真度的数据，如何区分并选择合适的数据作为高、低保真度数据，直接影响着模型最终的预测效果。因此，本研究首先探讨数据保真度的评价方法，从领域专家知识评价和模型评价两个角度出发，分析材料领域中的数据保真度评估方法。

领域专家知识评价侧重于利用先验经验对数据保真度进行评判。材料领域中的普遍观点是，实验测得的数据通常被认为具有较高的保真度，而理论计算所得的数据则为低保真度数据。在对比不同理论计算方法得到的数据保真度时，可以

通过其内在机理进行判断。然而，实验数据由于实验条件、过程的不可透明性以及人为因素的影响，往往使得其保真度无法直接确认，因此在此情境下，数据保真度的评定更依赖于领域专家的知识 and 经验。本研究通过领域专家知识对材料领域的具体实验得到的多保真度数据集进行评价，筛选出高保真度数据，并在此基础上展开实验验证。具体而言，我们从文献和实验中收集了锂基液态金属电池阴极在不同组分和电流密度条件下的能量密度数据，通过专家知识筛选出高保真度数据集。随后，结合实验数据和理论分析，验证了所采用的数据保真度评价方法的有效性。此外，这一高保真度数据集还被应用于锂基液态金属电池阴极元素推荐，考虑了电池的能量密度和倍率性能，为后续材料设计提供了新的方向。

上文提到领域专家知识适用于不同实验数据的评价，至于不同理论计算的数据保真度来说往往可以通过方法本身内在机理进行判断。然而，同一组多保真度数据集在不同模型中可能会呈现出不同的保真度。因此，如何根据具体模型进行分析，评估多保真度数据在不同模型中的表现，也是本研究关注的重要内容之一。为此，本研究提出了基于模型的数据保真度评价方法，并设计了一种基于材料图神经网络的不确定性量化方法。具体而言，研究通过对比同一组多保真度数据集在不同模型中与实验测试集的误差大小，得出了数据集在不同模型中的保真度排序。此外，得益于不确定性量化的优点，本研究进一步采用了基于不确定性大小的加权修正方法，初步探索了线性修正的多保真度数据学习方法的可行性，为本文后续多保真度数据学习方法的深入研究奠定了基础。

两种数据保真度评价方法各有优势，使用领域专家知识去评价数据保真度相对于不确定性量化来说更节省时间，但受限于个人领域知识水平；使用模型的评价方法能一定程度提高模型最终预测精度提高，但会增加耗时并提高代码复杂度。在本研究中，这两种方法将在实验和理论计算的多保真度数据集上进行实验，所涉及的具体研究内容及其实验结果将在本文后续第3章中介绍。

1.3.2 多保真度数据学习方法

多保真度数据学习方法旨在充分利用不同保真度数据，旨在通过对低保真度数据进行修正、降噪等操作提升模型预测精度。本研究针对多保真度数据的修正策略分为两大类：线性修正和非线性修正。其中，非线性修正方法采用了迭代降噪策略。

线性修正方法通过使用多保真度代理模型（Multi-Fidelity Surrogate Model, MFSM）^[73]去寻找高、低保真度数据之间的关系，并采用线性的方式修正低保真度数据。MFSM 根据参数估计方法可以分为确定性方法^[75]和非确定性方法^[76]，

本研究则以非确定性方法中的高斯过程（Gaussian Process, GP）作为代理模型并用于后续实验，这是因为 GP 在多保真度数据上具有出色表现^[77,78]及其普遍的利用率^[79]。同时，得益于 GP 在建模过程中自然地考虑了噪声，使其能够较好地处理含噪多保真度数据集^[80]。另一方面，有研究表明，与确定性方法相比非确定性方法往往更准确^[81]。基于上述方法，本研究探讨了加性修正、乘性修正及综合修正这三种线性修正方式。尽管这些方法在不同的研究中已有应用，但目前尚缺乏系统评估其在含噪多保真度数据集上的表现。因此，本研究通过人为构造的多保真度数据集，探索三种线性修正方法在加性、乘性及混合噪声下的修正效果。通过这种方式，可以更准确地评估不同线性修正方法在实际场景中的适用性，寻找线性修正方法对模型的提升规律。

除了线性修正方法，本研究还提出了一种迭代降噪的非线性修正方法，以更加全面地探索多保真度数据学习方法。具体而言，迭代降噪方法通过降噪函数修正每一轮迭代后的低保真度数据，最终实现模型精度的提升。考虑到不同的训练策略会对模型结果产生不同的影响，本研究还设计了逐一训练和洋葱训练两种训练策略。其中，逐一训练方法按照数据保真度从低到高依次进行训练；而洋葱训练方法则是在每次训练后剔除当前保真度最低的数据集，再进行下一轮训练。与线性修正的研究方案类似，为了更加全面地评估迭代降噪方法对模型精度的提升规律，本研究在理论数据集上加入了不同量的线性噪声和采样噪声来模拟真实的多保真度数据。

两种多保真度数据学习方法从线性和非线性两个角度进行研究，分别引入不同类型和不同程度的噪声来模拟不同场景下的多保真度数据集，其中非线性修正方法还从训练策略的角度对多保真度数据集进行研究。这两种方法均探索了多保真度数据学习方法在含噪多保真度数据集中对模型精度的提升规律，所涉及的具体研究内容及其实验结果在本文第4章中介绍。

1.4 创新点

结合上述研究内容，本研究的创新点可总结为如下三个：

（1）基于材料图神经网络的不确定性量化方法

不确定性量化在材料领域有助于指导材料设计并降低成本。然而，现有的材料图神经网络属性预测模型并不具备不确定性量化功能。本研究提出了一种基于材料图神经网络的不确定性量化方法，该方法从两种流行的材料图神经网络中提取数据描述子，然后将其作为输入用于高斯过程回归的训练和预测。实验结果表明，与传统模型相比，该方法在一定程度上提高了预测精度，并具备了不确定性

量化能力。

(2) 多保真度数据学习方法的扩展

通过不确定性量化方法,本研究获得了多保真度数据集在具体模型上的不确定性大小。基于此,我们提出了一种不确定性加权修正方法,按照不确定性大小对多保真度数据集的预测值进行加权修正。此外,我们还设计了迭代降噪的多保真度数据学习方法,从非线性修正的角度利用多保真度数据。以上方法从数据修正层面提升了模型的预测精度,为多保真度数据学习方法提供了一种新思路。

(3) 含噪多保真度数据集下数据学习方法对模型提升规律的探索

本研究通过在多保真度数据集中引入不同类型和程度的噪声,探讨了不同数据修正方法在含噪数据集中对模型精度提升规律。所采用的多保真度数据学习方法包括线性修正和非线性的迭代降噪这两种方法。对于线性修正方法,本研究在人为构造多保真度数据集中引入了加性、乘性和混合噪声;对于迭代降噪的非线性方法,则是在理论计算的多保真度数据集中引入了线性、采样噪声。基于以上含噪的多保真度数据集,本研究分别使用不同的多保真度数据学习方法探索其对模型最终预测精度的提升规律。

1.5 章节安排

本研究针对材料科学领域下的多保真度数据集及其机器学习模型来研究多保真度数据学习方法,从数据保真度评价方法和多保真度数据学习方法两大方面进行探索,所涉及的多保真度数据集来自人为构造、理论计算和实验。全文共分为五个章节:

第1章 主要介绍多保真度数据学习方法的研究背景,引入其在材料领域中的重要性,简要叙述本研究中主要的研究内容、创新点以及论文组织结构。

第2章 主要介绍本研究中所使用到的多保真度数据学习方法的相关理论和技术。介绍了相关性系数、方差膨胀因子、高斯过程回归、两种材料图神经网络的理论基础。

第3章 主要研究数据保真度评价方法,针对材料领域中实验和理论计算的多保真度数据集,通过领域专家知识和具体模型评价数据保真度。此外,使用不确定性加权修正方法初步探索多保真度数据学习方法的可行性,为后续的研究奠定基础。

第4章 主要研究多保真度数据学习方法,使用人为构造和理论计算的多保真度数据集进行实验,分析线性和非线性的多保真度数据学习方法在含噪数据集下对模型的提升规律。

第 5 章 对于全文的总结和展望。总结全文内容和实验结果，针对现有研究的不足之处进行讨论并提供未来可能的研究方向。

第2章 关键理论与技术基础

本章主要阐述本研究中所涉及到的关键技术方法的理论介绍,包括皮尔逊和斯皮尔曼相关性系数、方差膨胀因子、高斯过程回归、材料图神经网络 MEGNet 和 CGCNN。

2.1 相关性系数和方差膨胀因子

2.1.1 皮尔逊相关性系数

皮尔逊相关性系数 (Pearson correlation coefficient) 用于衡量两个变量之间线性关系的强度和方向,其值范围为 -1 到 1。计算皮尔逊相关性系数 r 的公式 2.1 所示。

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \cdot \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2.1)$$

在公式 2.1 中, X_i 和 Y_i 分别表示两个变量的单个数据点, $\bar{X}$ 和 $\bar{Y}$ 分别是变量 X 和 Y 的均值。公式的分子部分衡量了两个变量之间的协方差,分母通过两个变量的标准差对协方差进行归一化,从而得出一个无量纲的系数,其范围为 -1 到 1。当系数为 1 时,表示完全正相关,即一个变量增加,另一个变量按比例增加;当系数为 -1 时,表示完全负相关,即一个变量增加,另一个变量按比例减少;当系数为 0 时,表示两个变量之间不存在线性关系。在本研究中,皮尔逊相关性系数将被用于寻找不同元素对电池的能量密度和倍率性能之间的重要性程度。

2.1.2 斯皮尔曼相关性系数

斯皮尔曼相关性系数 (Spearman correlation coefficient) 用于衡量非线性相关性,评估两个变量之间的关系是否可以用单调函数来描述。与衡量线性关系的皮尔逊相关性系数不同,斯皮尔曼相关性系数的取值范围同样是 -1 到 1,但其基于数据的排名(排序后的排名位次)而非原始数据计算。斯皮尔曼相关性系数 ρ 的计算公式为:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad (2.2)$$

在公式 2.2 中, d_i 表示两个变量排序后的位次差值, n 表示观测值的数量。由于斯皮尔曼相关性系数是基于位次差的计算,其对异常值更具鲁棒性,适用于不满

足正态性假设的数据。当数据中存在相同排名时，需采用基于实际排名的皮尔逊相关系数，以避免计算结果的偏差。在本研究中它将与皮尔逊相关性系数共同衡量元素重要性。

2.1.3 方差膨胀因子

为了对相关系数得出稳健的结论，需要评估特征之间的相关性，特征之间存在的线性相关性或共线性的现象被称为多重共线性（Multicollinearity），如果特征间的多重共线性程度较高，那么其相关性系数结果则不太准确。方差膨胀因子（Variance Inflation Factor, VIF）是一种常用的计算多重共线性的方法。在本研究中，使用 Python 库 Statsmodels 中的函数实现了 VIF 的计算。其公式如式 2.3 所示，其中 X_i 是第 i 个预测变量， R_i^2 是 X_i 与所有其他预测变量进行回归时的决定系数。

$$VIF(X_i) = \frac{1}{1 - R_i^2} \quad (2.3)$$

VIF 值表示由与其他预测变量的多重共线性导致回归系数方差膨胀的程度。VIF 值为 1 表示没有多重共线性，值在 1 到 5 之间表示存在适度的多重共线性，值大于 5 表示存在较高的多重共线性，可能会引发问题。特别地，VIF 大于 10 通常被认为是严重多重共线性。当多重共线性较高时，某些变量可能高度相关，那么此时变量之间的相关性程度结果将不再可靠。本研究在后文计算元素与能量密度和倍率性能的相关性系数前，使用了 VIF 去衡量元素间的多重共线性程度。

2.2 高斯过程回归

高斯过程回归（Gaussian Process Regression, GPR）是一种非参数的贝叶斯回归方法，在本研究中该方法将用于不确定性量化以及构建代理模型。高斯过程通过假设函数的分布为一个多元高斯分布，来建模输入与输出之间的关系。其核心思想是利用协方差函数（也称为核函数）来捕捉输入数据之间的相似性，从而对未知数据点进行预测。

高斯过程是一种定义在函数空间上的随机过程，其核心假设是任意有限维输入变量对应的函数值服从多元高斯分布。假设数据集为 (X, y) ，其中 $X \in \mathbb{R}^{N \times d}$ ，为 N 个输入特征长度为 d 向量； $y \in \mathbb{R}^N$ 为对应的观测值。对于一个新的输入点 x_* ，高斯过程假设 (y, y_*) 服从联合多元高斯分布：

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu(X) \\ \mu(x_*) \end{bmatrix}, \begin{bmatrix} K(X, X) + \sigma_n^2 I & K(X, x_*) \\ K(x_*, X) & K(x_*, x_*) \end{bmatrix} \right) \quad (2.4)$$

其中 $\mu(X)$ 和 $\mu(x_*)$ 为输入点的均值函数，常假设为零； $K(X, X)$ 为训练数据间的协方差矩阵； $K(X, x_*)$ 和 $K(x_*, x_*)$ 分别为训练数据与预测点间、预测点自身的协方差矩阵。 σ_n^2 表示建模时加入白噪声的方差，用于表示观测误差； I 表示单位矩阵，确保白噪声仅影响训练数据的协方差矩阵。

协方差矩阵中的每个元素通过协方差函数 $k(x, x')$ 计算得出，常用的核函数为高斯核（即径向基核），如下所示：

$$k(x, x') = \sigma_f^2 \exp \left(-\frac{\|x - x'\|^2}{2\sigma_l^2} \right) \quad (2.5)$$

其中 σ_f 和 σ_l 为超参数，分别控制核函数幅值和特征尺度， $\|x - x'\|$ 表示数据点 x 和 x' 之间的欧氏距离。本研究中所采用的核函数为高斯核函数。通过条件概率推导，预测点 x_* 的输出均值 μ_* 和方差 σ_*^2 如下：

$$\mu_* = K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1}y \quad (2.6)$$

$$\sigma_*^2 = K(x_*, x_*) - K(x_*, X)[K(X, X) + \sigma_n^2 I]^{-1}K(X, x_*) \quad (2.7)$$

其中 μ_* 是预测结果的均值，表示对目标值的最可能估计；方差 σ_*^2 为预测结果的方差，表示不确定性。本研究中则使用高斯过程回归用于训练并预测，特别的，在模型评价数据保真度方法中，高斯过程回归得到的不确定性大小将被用于加权修正预测值。

2.3 材料图神经网络

在材料科学领域，图神经网络（Graph Neural Network, GNN）因其灵活的数据表示能力和高效的特征学习，已成为材料结构和性质预测中的重要工具。GNN能够将复杂的材料结构表示为图，以原子作为节点，原子间的键作为边，直接从数据中学习到物理和化学特性。本研究针对材料领域结构数据集使用了两种代表性的材料图神经网络模型：MEGNet^[71]（MatErials Graph Network）和 CGCNN^[70]（Crystal Graph Convolutional Neural Networks）。

2.3.1 MEGNet

MEGNet 是一种通用的材料图神经网络模型，可以处理分子和晶体结构数据。MEGNet 的核心思想是利用材料的图结构来表示其组成和相互关系，其中节点代表原子，边则代表原子间的化学键。此外，MEGNet 还增加了全局状态用于表达额外信息，例如将平均原子质量或者每个原子的化学键数量作为全局状态属性。

MEGNet 通过信息传递机制，将节点信息逐步传播到邻居节点，迭代更新节点、键和全局状态信息，最终形成整个材料的表示，其更新方式如图 2.1 所示。其中，MEGNet 块包含三个部分：键属性更新、原子属性更新和全局状态更新。对于本研究所使用到的晶体结构数据集来说，原子属性初始为元素编号，键属性为由高斯扩展后的空间距离，全局状态属性默认使用两个零作为占位符。

MEGNet 块首先对键属性进行更新，每个键使用其自身 (e_k)、其连接原子（索引 r_k 和 s_k 对应的原子）和全局状态向量 (u) 的属性进行更新，如公式 2.8 所示：

$$e'_k = \phi_e(v_{s_k} \oplus v_{r_k} \oplus e_k \oplus u) \quad (2.8)$$

其中 ϕ_e 是键更新函数，是带有 2 层隐藏层的多层感知机，默认的激活函数是在 softplus 基础上减去 $\ln(2)$ ； v_{s_k} 和 v_{r_k} 分别表示键 e_k 两端的原子； $\oplus$ 表示拼接操作。接下来对原子属性进行更新，每个原子的属性使用其自身 (v_i)、连接到它的键 ($\bar{v}_i^e$) 和全局状态向量 (u) 的属性进行更新，如下公式所示：

$$\bar{v}_i^e = \frac{1}{N_i^e} \sum_{k=1}^{N_i^e} \{e'_k\}_{r_k=i} \quad (2.9)$$

$$v'_i = \phi_v(\bar{v}_i^e \oplus v_i \oplus u) \quad (2.10)$$

其中 N_i^e 表示与原子 i 相连的键的数量； $\bar{v}_i^e$ 用于局部池化，该操作对连接到原子 i 的键取平均值。 ϕ_v 是原子更新函数，实现方法与键更新函数 ϕ_e 一致。最后，使用自身 (u)、所有原子 ($\bar{u}^v$) 和键 ($\bar{u}^e$) 的信息更新全局状态属性，如下公式所示：

$$\bar{u}^e = \frac{1}{N^e} \sum_{k=1}^{N^e} \{e'_k\} \quad (2.11)$$

$$\bar{u}^v = \frac{1}{N^v} \sum_{i=1}^{N^v} \{v'_i\} \quad (2.12)$$

$$u' = \phi_u(\bar{u}^e \oplus \bar{u}^v \oplus u) \quad (2.13)$$

其中 ϕ_u 是全局状态更新函数，实现方法同 ϕ_e 和 ϕ_v 一致； N^e 和 N^v 分别表示键和原子的总数。

除了 MEGNet 块以外，MEGNet 模型中还包含了其他内容，例如用于将原子和键向量的集合归约为单个向量的 set2set 神经网络^[82]层。同时 MEGNet 模型架构中也包含了与残差神经网络^[83]类似的跳跃连接操作，以实现更深入的模型训练并减少过拟合现象，MEGNet 模型的整体架构如图 2.2 所示。其中，“Add”箭头表示跳跃连接，以实现更深入的模型训练并减少过拟合。在经过多个 MEGNet 块后，使用 set2set 神经网络将原子和键向量的集合归约为单个向量。

MEGNet 对于多保真度数据集的利用也有研究，但其模型 MFGNet^[2]本质上还是在模型层面进行的，MFGNet 为每种保真度进行编号来区分多保真度数据。而本研究则是针对数据层面，不使用 MFGNet。在本研究中 MEGNet 模型将主要被应用于两个方面：训练并预测晶体结构类型的材料数据集以及提取数据描述子。

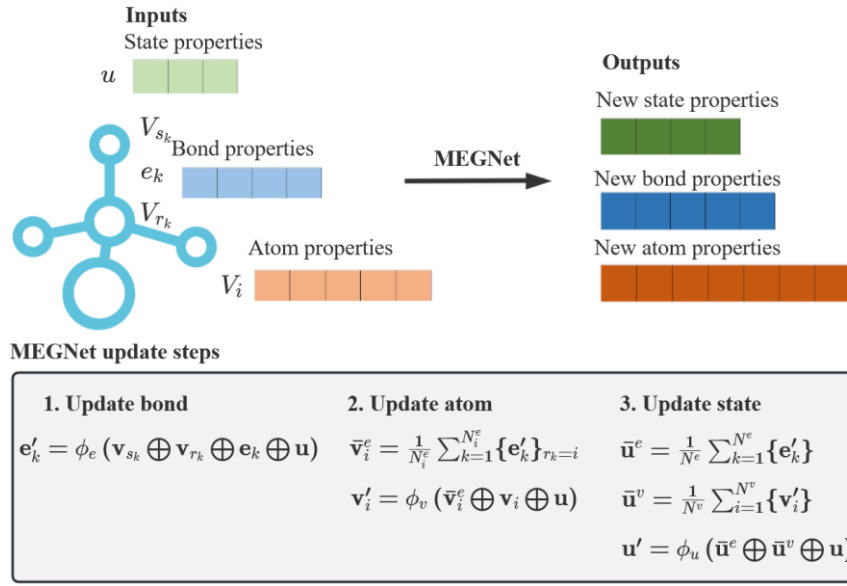


图 2.1 MEGNet 模块。其中 u , v , e 分别表示全局状态、节点和键； v_{s_k} 和 v_{r_k} 分别表示键 e_k 两端的原子； ϕ_e 、 ϕ_v 和 ϕ_u 是更新函数； $\oplus$ 表示拼接操作； N_i^e 表示与原子 i 相连的键的数量， N^e 和 N^v 分别表示键和原子的总数。

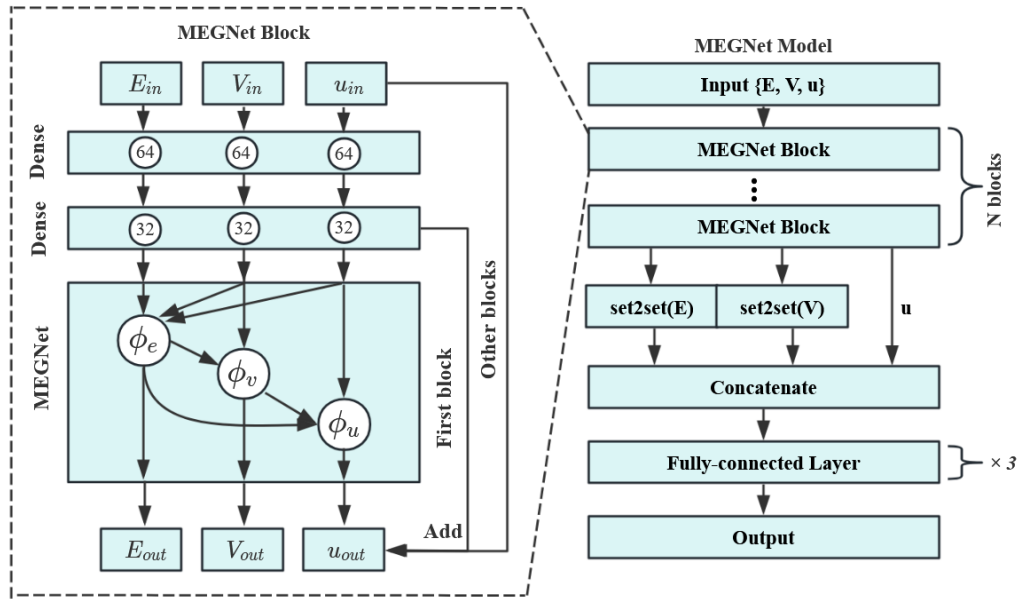


图 2.2 MEGNet 模型架构

2.3.2 CGCNN

CGCNN 是一种经典的材料图神经网络模型，专为晶体结构设计。其核心也

是通过图的形式对晶体结构进行建模,并结合卷积神经网络对图进行特征提取与学习。在将晶体结构转化为图结构时,同样也是用图节点表示原子,边则表示原子间的键。也正如其名字中所描述的那样,CGCNN 中将图结构数据特征进行卷积,从晶体图的节点及其邻居节点中提取特征并迭代更新信息,节点信息的更新如下公式所示:

$$v_i^{(t+1)} = \text{Conv}(v_i^{(t)}, v_j^{(t)}, e_{(i,j)_k}) \quad (2.14)$$

其中 $\text{Conv}(\cdot)$ 表示卷积函数, $v_i^{(t)}$ 表示节点 i 在卷积层第 t 层时的特征向量, $e_{(i,j)_k}$ 表示节点 i 和节点 j 之间第 k 条边的特征。具体来说,节点信息的更新方式如下:

$$z_{(i,j)_k}^{(t)} = v_i^{(t)} \oplus v_j^{(t)} \oplus e_{(i,j)_k} \quad (2.15)$$

$$v_i^{(t+1)} = v_i^{(t)} + \sum_{j,k} \sigma(z_{(i,j)_k}^{(t)} W_f^{(t)} + b_f^{(t)}) \odot g(z_{(i,j)_k}^{(t)} W_s^{(t)} + b_s^{(t)}) \quad (2.16)$$

其中 $\oplus$ 表示拼接操作; $z_{(i,j)_k}^{(t)}$ 表示将相邻节点信息及其边信息都拼接起来用作卷积层输入; $\sigma(\cdot)$ 和 $g(\cdot)$ 分别表示 sigmoid 和 softplus 激活函数; $\odot$ 表示逐元素乘法操作,将两个激活后的特征向量按元素逐一相乘; $W_c^{(t)}$, $W_s^{(t)}$ 和 $b^{(t)}$ 分别表示卷积操作中的权重矩阵,自重矩阵和偏置项。

CGCNN 模型整体的架构图如图 2.3 所示。在经过 R 层卷积操作之后,CGCNN 通过迭代的方式学习周围原子的信息来更新每一个原子的特征向量。之后,进入深度为 $L1$ 的全连接隐藏层,然后进入池化层,池化层用于产生晶体的总体特征向量 v_c ,如下公式所示:

$$v_c = \text{Pool}(v_0^{(0)}, v_1^{(0)}, \dots, v_N^{(0)}, \dots, v_N^{(R)}) \quad (2.17)$$

其中,为了满足原子的置换不变性和晶胞的尺寸不变性,CGCNN 使用了归一化求和的方法作为池化函数 $\text{Pool}(\cdot)$,从而得到表示全局的特征向量。最后,使用该全局特征向量进入深度为 $L2$ 的全连接隐藏层,最后进入一层全连接的输出层来输出结果。

与 MEGNet 仅提供原子序数作为原子属性不同,CGCNN 通过人为经验提供了 9 种属性作为原子属性,包含族数、周期数、电负性、共价半径、价电子数、第一电离能、电子亲和力、分区以及原子体积这 9 种属性,从更多方面涵盖原子信息。键属性为原子间距的高斯扩展特征,通过径向基函数将连续的原子间距映射为离散向量。在本研究中,CGCNN 与 MEGNet 的应用一致,将用于训练并预测晶体结构类型的材料数据集并提取数据描述子。

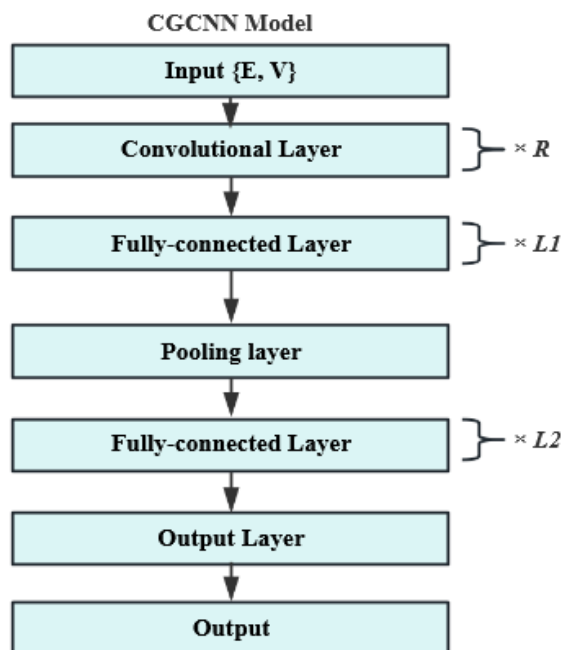


图 2.3 CGCNN 模型架构图

2.4 本章小结

本章节主要介绍了本研究在后续实验中将会使用到的相关理论和技术。其中，第一节介绍的相关性系数和方差膨胀因子将用于领域专家知识评价数据保真度章节当中；第二节介绍了高斯过程回归，在本研究中高斯过程回归将用在不确定性量化和代理模型构建上，用于训练从材料图神经网络中提取的描述子并预测其对应不确定性，此外还将作为线性修正时的代理模型。第三节介绍了本研究所使用到的两种图神经网络：MEGNet 和 CGCNN，它们在后文被用来训练材料领域晶体结构类型的多保真度数据集，此外还将用于提取多保真度数据的描述子。

第3章 数据保真度评价方法

多保真度数据学习方法虽在其他领域早已应用,但有关数据保真度评价方法的研究却少有人关注。在本章节中,本研究分别对领域专家知识和使用具体模型评价这两种数据保真度评价方法进行研究。

对于领域专家知识评价方法,本研究从文献和实验中收集了液态金属电池在不同组分和电流密度条件下能量密度的多保真度数据集。通过领域专家知识进行评价,筛选出高保真度数据集并结合专业知识分析实验结果,验证所提评价方法的有效性。筛选出的高保真度数据集随后应用于最佳元素推荐中,结果仍然与传统理论相符。

对于模型评价方法来说,本研究选择理论计算的带隙数据集作为多保真度数据集,并将实验获得的数据作为最高保真度并用于测试。使用两种材料图神经网络及其不确定性量化扩展评价数据保真度,发现同一组多保真度数据在不同模型上所呈现出的保真度排序有所不同。此外,为了根据具体模型具体利用多保真度数据,本研究利用不确定性大小对预测结果进行加权修正以提高预测精度。该方法初步探索了多保真度数据学习方法,为文章后续进一步研究奠定基础。

3.1 领域专家知识评价数据保真度

3.1.1 研究方案

采用领域专家知识进行数据保真度评价通常较为便捷,但对专家的经验 and 水平要求较高。尤其是在使用未公开具体计算方法或缺乏统一基准评估的多保真度数据集时,数据保真度的评价往往依赖于使用者的经验判断。本节中所使用到的多保真度数据集为材料领域中液态金属电池不同组分和电流密度下能量密度的数据集,其来自文献和实验。每条数据的元素组成和计算方法各异,构成了典型的多保真度数据。本研究则基于该多保真度数据集,从电池的阴阳极质量负载、电流密度与能量密度匹配以及活性成分占比这3方面进行评价,筛选出符合条件的高保真度数据。之后,为了验证评价方法的正确性,本研究在该数据集基础上使用数据驱动的机器学习方法计算各个元素与能量密度和倍率性能的相关性系数。由此得到元素重要性排序并结合传统知识进行对比,用于验证评价方法的正确性。

3.1.2 领域专家知识评价方法

本节所使用的多保真度数据集来自文献和本研究团队的实验数据，实验数据由北京科技大学材料专业相关合作组提供。该数据集涵盖了锂基液态金属电池阴极在不同组分和电流密度条件下的能量密度数据。数据集包括 8 种不同元素的组分含量、电流密度和能量密度，具体元素为：硒（Se）、锑（Sb）、铜（Cu）、铋（Bi）、铅（Pb）、锡（Sn）、碲（Te）、锌（Zn）。这些数据集涉及的元素组合和实验方法各不相同，构成了典型的多保真度数据集。在此基础上，本研究使用领域专家知识对这些数据的保真度进行评价，提出以下三点原则用于筛选出高保真度数据：

（1）阴阳极质量负载不匹配的情况。这种不匹配会导致能量密度偏低，因为部分活性材料未被充分利用，需要剔除阴阳极质量负载不匹配的数据。

（2）能量密度随电流密度增加而升高的情况。较高的充放电电流密度会引起较大的极化电阻，通常应导致放电电压降低、放电容量减少，从而使能量密度下降。

（3）能量密度随活性成分含量增加而降低的情况。活性成分（如 Sb、Bi 和 Te）通常对电池的能量密度有正向贡献。在其他条件不变的情况下，如果这些活性成分的含量上升，那么电池的能量密度也应该有所提高。

通过以上三条原则，本研究从文献和实验中筛选出了 63 条高保真度数据，如表 3.1 所示。随后，本研究基于该数据集使用数据驱动的机器学习方法计算元素重要性程度并结合领域专家知识分析实验结果。通过判断实验结果是否和传统理论经验相符来验证所使用评价方法的正确性。

表 3.1 文献和实验中筛选得到的 61 条高保真度数据

Se (%)	Sb (%)	Cu (%)	Bi (%)	Pb (%)	Sn (%)	Te (%)	Zn (%)	能量密度 (Wh/kg) # 电流密度 (mA/cm^2)
0	30	0	0	70	0	0	0	[84]: 91.15#100; 80.41#250; 64.19#500; 32.12#1000
0	30	0	0	0	0	0	70	[85]: 290.6#100; 284.4#200; 268.8#400; 228.1#750; 208.5#1000
0	0	0	0	0	30	70	0	[86]: 371#100
0	0	0	0	0	14	86	0	[86]: 495.9#100
0	0	0	100	0	0	0	0	[87]: 197.2#200; 189.3#300; 182.9#400; 162.1#750; 134.5#1000; 115.97#1250
0	64	36	0	0	0	0	0	[88]: 353#400

续表 3.1 文献和实验中筛选得到的 61 条高保真度数据

0	50	0	0	0	50	0	0	[89]: 186.16#100; 170.24#200; 145.89#400; 88.41#1000
0	30	0	0	70	0	0	0	[90]: 79.96#275
0	0	0	0	0	100	0	0	[90]: 38.3#100
0	33.3	0	33.3	0	33.4	0	0	[90]: 220.5#100
0	0	0	55.5	44.5	0	0	0	[90]: 80.16#52
0	40	0	60	0	0	0	0	[91]: 217.36#100; 199.34#200; 184.63#300; 173.51#400; 147.26#600; 125.20#800; 116.42#1000
0	0	0	0	0	100	0	0	[92]: 35.3#100; 30.40#200; 20.77#400; 7.37#700; 0.183#1000
0	45	0	45	0	10	0	0	[93]: 261#191; 260#346; 255#462; 202.88#693
0	45	0	45	0	10	0	0	[93]: 262.53#200; 245.29#400; 230.76#800; 226.45#1000; 205.06#1200
0	0	0	55.5	44.5	0	0	0	[94]: 86.58#50; 67.65#100; 54.04#200
0	0	0	70	0	0	0	30	Own data: 213.91#100; 189.59#200; 164.64#400; 129.72#800; 112.25#1000
50	12.5	37.5	0	0	0	0	0	Own data: 312.1#191
50	25	25	0	0	0	0	0	Own data: 320.8#191
50	25	25	0	0	0	0	0	Own data: 237.5#191
50	20	30	0	0	0	0	0	Own data: 405.98#191; 395#200
70	30	0	0	0	0	0	0	Own data: 450.41#100
30	0	0	70	0	0	0	0	Own data: 215.74#382
70	0	30	0	0	0	0	0	Own data: 337.5#44.23
0	0	30	0	0	0	70	0	Own data: 412.3#200

3.1.3 实验流程及结果

本研究通过对该数据集的训练与分析来验证领域专家知识评价方法的正确性。实验流程主要包括：线性回归（Linear Regression, LR）粗略分析、数值法求倍率性能、相关性系数计算、多重共线性分析、实验结果分析及与传统理解对比。

在获得这些高保真度数据集后，本研究首先使用线性回归模型对 8 种元素中电流密度与能量密度之间的关系进行初步探索，得到公式 3.1 的线性关系式：

$$f(\text{Elements}, J) = 1.785Se + 1.007Sb + 0.082Cu - 0.753Bi - 2.795Pb - 2.094Sn + 2.791Te - 0.023Zn - 0.069J + 2.579 \quad (3.1)$$

其中，J 表示电流密度，其系数-0.069 表示能量密度对电流密度的偏导数，即电池

的倍率性能。该负值表明，在其他条件不变时，电流密度升高会导致能量密度下降，与理论预期相符。随后，为了更加准确地判断评价方法的正确性，本研究计算了各元素与电化学属性之间的重要性。与传统仅关注元素与能量密度关系的研究不同，本文同时考察了元素对能量密度和倍率性能的影响，进而综合评估各元素的重要性。

由于倍率性能在文献和实验中未直接给出，本研究首先采用数值方法计算电池的倍率性能。具体而言，对于任意组合下的倍率性能，我们通过计算其左、右两侧偏导数的平均值获得，如图 3.1 所示。其中，左侧虚线和右侧虚线是曲线对应位置的切线， ∂J （例如 $1e-2$ 、 $1e-3$ 等）表示用于计算偏导数的电流密度。

为了更加准确地计算出倍率性能，我们需要尽可能获得不同元素组合、电流密度下的能量密度数据集。因此，本研究使用了多层感知机（Multi-Layer Perceptron, MLP）和高斯过程回归（Gaussian Process Regression, GPR）这两种模型用于额外预测不同组合下的数据。在模型构建阶段，我们为 MLP 和 GPR 设计了多组超参数组合，并采用留一法交叉验证（Leave-One-Out Cross-Validation, LOO-CV）筛选出最佳超参数，其中 LOO-CV 如图 3.2 所示。最终，通过 LOO-CV 确定的最佳模型参数信息如下：

MLP: 本研究使用由 PyTorch 实现的多层感知机模型，共包含 45 个参数。超参数设置为：学习率 0.001，优化器采用 Adam，激活函数为 ELU。模型架构共三层：输入层包含 9 个节点（对应 8 个元素特征和 1 个电流密度特征）、隐藏层包含 4 个节点，输出层包含 1 个节点。训练过程中未采用批量处理，而是在每个 epoch 中对全部 63 个数据点进行整体训练。

GPR: 本研究使用 GPytorch 库实现高斯过程回归模型。该模型采用径向基函数（高斯核）作为核函数，其长度尺度设置为 0.01，以控制函数的平滑程度。高斯似然中的噪声水平设为 1.0，代表观测值的方差。训练时使用 Adam 优化器，学习率为 0.001。

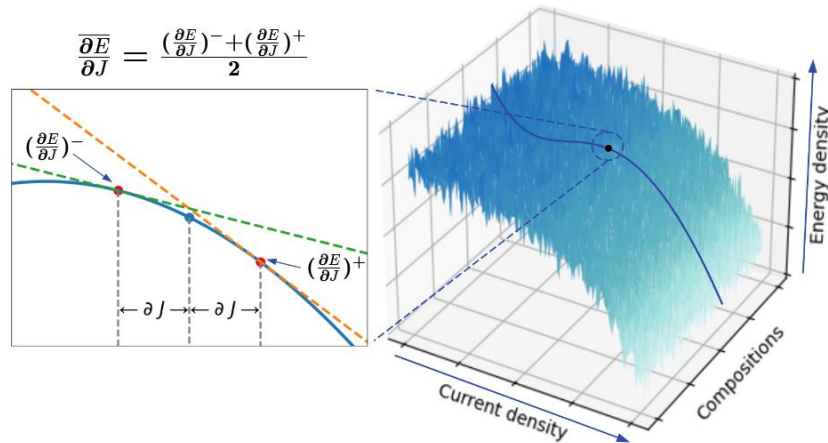


图 3.1 能量密度 (E) 相对于电流密度 (J) 的偏导数示意图

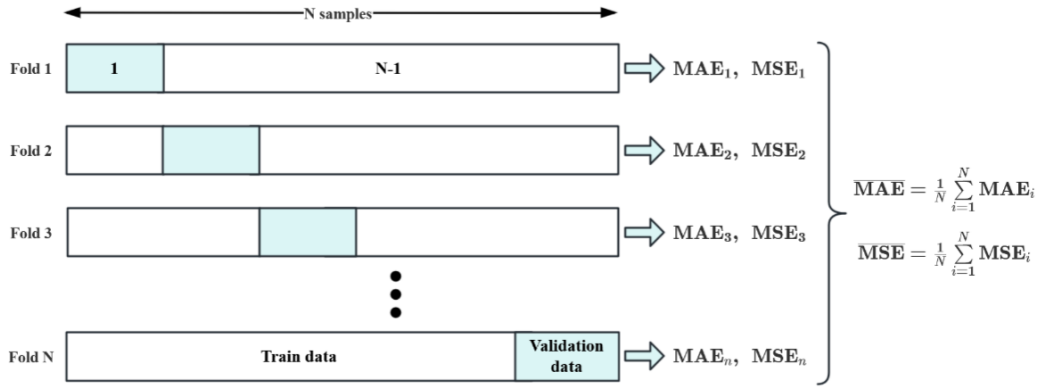


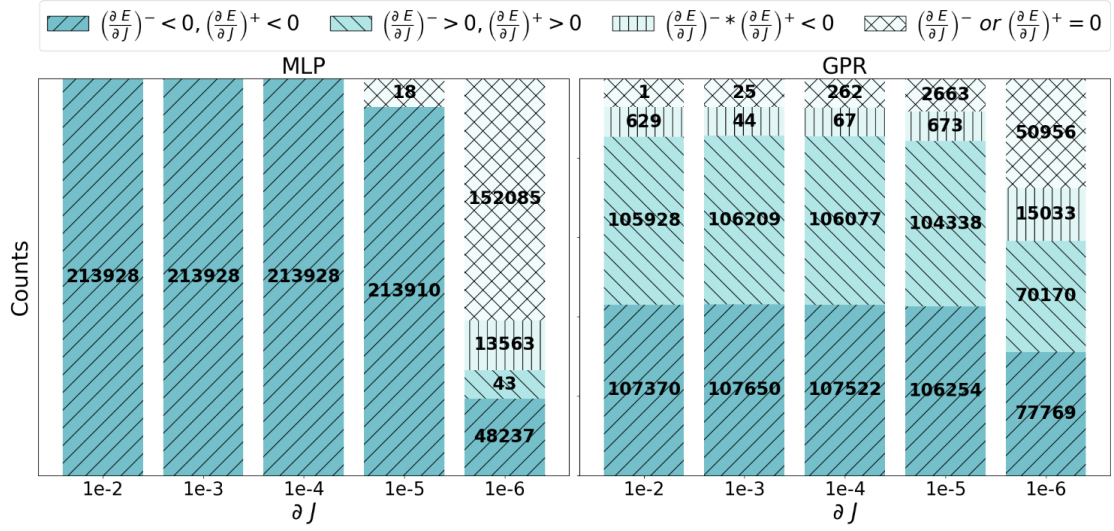
图 3.2 留一法交叉验证示意图

基于上述最佳超参数构建的模型，本研究利用 LOO-CV 计算了各模型的平均 MAE 和 MSE 指标，其中包括之前提到的 LR 模型。表 3.2 显示的结果表明，GPR 和 MLP 模型的性能均明显优于 LR 模型，且 MLP 模型的预测准确性更高。因此，后续实验将主要聚焦于 MLP 模型的计算结果。

表 3.2 各种模型预测能量密度的平均绝对误差（MAE）和均方误差（MSE）

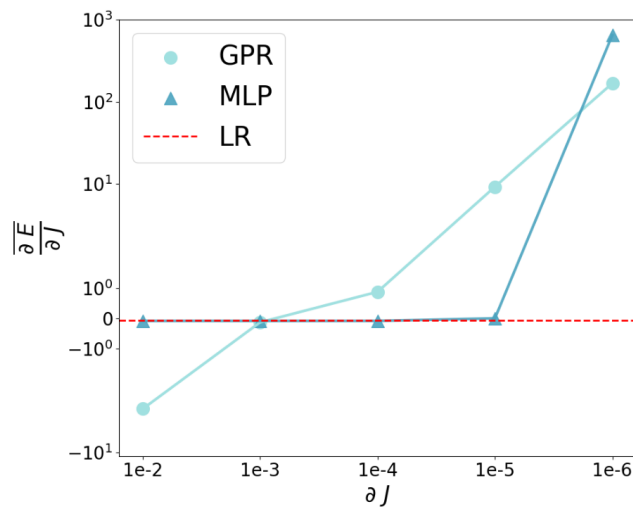
模型	MAE	MSE
LR	30.207	1411.897
GPR	17.318	1099.932
MLP	12.181	390.514

为了尽可能准确地计算能量密度对电流密度的偏导数（即倍率性能），本研究需要获取大量数据点。为此，我们对不同元素和电流密度进行组合，其中 8 种元素的取值均在 0 到 1 之间，步长为 0.1，且要求所有元素的组分之和为 1；电流密度的取值范围为 0 至 1000，步长为 100。基于上述设置，共生成 213,928 ($11 \times C_{17}^7$) 种组合。随后，我们利用之前优化得到的 MLP 和 GPR 模型预测这些组合对应的倍率性能，预测结果如图 3.3 所示。考虑到理论上倍率性能应为负值，即无论是数据点左侧还是右侧的偏导数值都应为负，因此，图 3.3 中所有左斜线部分代表有效数据，其余部分则视为无效数据。


 图 3.3 MLP 和 GPR 的偏导数值 $(\frac{\partial E}{\partial J})$ 分布情况

从图 3.3 中可以明显看出，MLP 模型的计算结果明显优于 GPR 模型。特别是在微分 (∂J) 变化范围从 10^{-2} 到 10^{-4} 之间时，MLP 结果展现出了显著的稳定性。随后，我们针对图 3.3 中所有偏导数均小于 0（即左斜线部分）的数据，分别计算了在五种不同电流密度微分条件下的平均偏导数值。结果显示，不同 ∂J 条件下的平均偏导数存在明显差异。为了更直观地展示这些数据，我们依据公式 (3.2) 对平均偏导数值进行了变换，其中将平均偏导数作为变量 x 参与计算。转化后的结果如图 3.4 所示。

$$f(x) = \begin{cases} x & \text{if } |x| < 3 \\ \text{sign}(x) \cdot \log_{10}(|x| + 1) & \text{if } |x| \geq 3 \end{cases} \quad (3.2)$$


 图 3.4 MLP 和 GPR 在不同电流密度下的平均偏导数 $(\frac{\partial E}{\partial J})$

从图 3.4 可以看出，在 10^{-2} 到 10^{-4} 的范围内，MLP 模型的平均偏导数值明显

比 GPR 更为稳定，这与图 3.3 中的讨论一致。此外，无论是 MLP 还是 GPR，当 $\partial J = 10^{-3}$ 时，其平均偏导数均与线性回归 (LR) 计算得到的常数 (-0.069) 非常接近。因此，后续实验将主要关注 MLP 模型在 $\partial J = 10^{-3}$ 条件下的输出结果，并利用这些数据计算各元素与能量密度及倍率性能之间的相关性来验证评价方法的正确性。

考虑到各元素之间可能存在协同效应，即若某些元素经常同时出现，机器学习模型可能难以准确区分它们对电化学属性的单独贡献。因此，本研究首先计算了各元素的方差膨胀因子 (Variance Inflation Factor, VIF) 以评估多重共线性程度，如图 3.5 所示。结果显示，各元素的 VIF 均位于 1 至 3 之间，其中铜 (Cu) 最高，但其 VIF 仅为 2.139，这表明元素间存在适度的多重共线性，均处于可接受范围 (1~10) 内。但硒 (Se)、锑 (Sb) 和铜 (Cu) 的 VIF 相对较高，表明它们与其他元素之间的共线性较强，可能导致后续相关性系数排序结果的不确定性。基于以上所有因素，本研究接下来使用 MLP 模型在 $\partial J = 10^{-3}$ 时的输出数据来计算相关性系数。

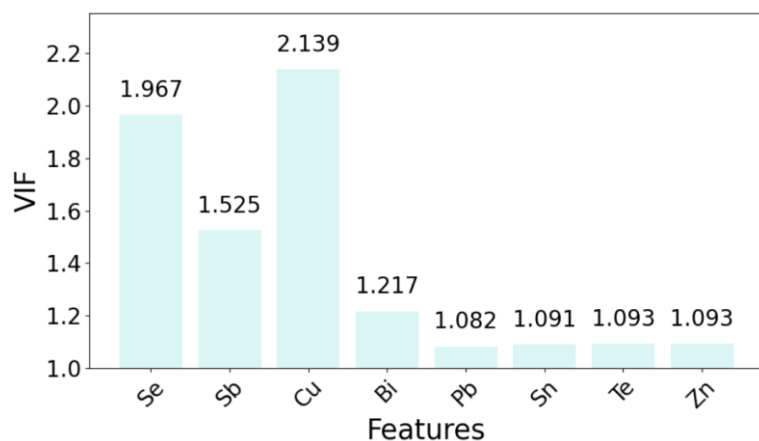


图 3.5 数据集中所有元素对应方差膨胀因子

本研究采用皮尔逊和斯皮尔曼相关性系数衡量各元素与能量密度及倍率性能之间的相关性，其计算结果分别见表 3.3，元素重要性排序见表 3.4。其中，表 3.4 中虚线框住的元素表示其多重共线性程度较高。对于能量密度，我们期望相关性系数越高越好，表明当元素含量上升时，能量密度也随之上升；而对于倍率性能，由于其代表能量密度对电流密度的偏导数（为负值），我们希望下降趋势越平缓，即值越高越好。因此，在表 3.4 中，对倍率性能相关性系数采用降序排序，排名越靠前的元素对能量密度或倍率性能的正向影响越明显。此外，考虑到多重共线性影响，表 3.4 中用虚线框出的元素其方差膨胀因子较高，这可能会影响结果的可信度。综合表 3.3 和表 3.4 的数据可见，Te 和 Se 对于能量密度的积极影响最大，Pb 和 Sn 对于倍率性能的积极影响最大。总体而言，Te 和 Se 在两

项指标中的排序均较高,而其他元素往往难以同时在能量密度和倍率性能上表现出积极作用。

表 3.3 元素与能量密度 (E) 和倍率性能 ($\frac{\partial E}{\partial J}$) 的皮尔逊 (P)、斯皮尔曼 (S) 相关性系数

方法		元素							
		Se	Sb	Cu	Bi	Pb	Sn	Te	Zn
E	P	0.311	0.269	0.210	-0.140	-0.633	-0.451	0.369	0.064
	S	0.283	0.244	0.189	-0.145	-0.580	-0.426	0.336	0.050
$\frac{\partial E}{\partial J}$	P	-0.009	-0.151	-0.068	0.004	0.258	0.086	-0.015	-0.104
	S	-0.020	-0.131	-0.058	0.020	0.162	0.054	-0.031	-0.065

表 3.4 元素与能量密度 (E) 和倍率性能 ($\frac{\partial E}{\partial J}$) 的皮尔逊 (P)、斯皮尔曼 (S) 相关性系数排序

方法	排序
E	P Te>Se>Sb>Cu>Zn>Bi>Sn>Pb
	S Te>Se>Sb>Cu>Zn>Bi>Sn>Pb
$\frac{\partial E}{\partial J}$	P Pb>Sn>Bi>Se>Te>Cu>Zn>Sb
	S Pb>Sn>Bi>Se>Te>Cu>Zn>Sb

3.1.4 与传统理解的对比和讨论

对于锂基液态金属电池来说,由于阳极相同,电池的能量密度主要取决于阴极的性能。基于电化学理论,阴极的高放电电压和大放电容量导致电池的高能量密度。因此,可以理解的是,放电电压高、比容量大的 Se 和 Te 位于能量密度贡献排序的前面,放电电压相对较低但比容量较大的 Sb 和 Bi 位于列表的中间,放电电压低得多且比容量较小的 Zn、Cu、Pb 和 Sn 元素位于列表的底部。然而,Cu 和 Zn 对能量密度的贡献顺序偏离了理论预期。这可以从它们与其他元素合金化的协同效应中理解。Cu 往往显示出较高的电子传导性,但其对锂存储不起作用(通常用作集电器)。此外,在 Se 中引入 Cu 可以明显改善电子传导,从而增强电化学反应,导致阴极的高比容量和高能量密度。Se-Cu 阴极的高能量密度主要来自 Se 的贡献,而 Cu 只是使 Se 的性能得到了较高地利用。由于关于 Cu 基阴极材料的电池相关报道非常有限,可用数据主要与高能量密度 Se 有关,因此对 Cu 对能量密度贡献的评估可能被高估了。与此相一致,Se 的贡献被低估了,这降低了它在能量密度重要性排序中的名次。关于元素 Zn,虽然它本身显示出相对较低的能量密度,但当它与其他活性元素(如 Sb)合金化时,在初始放电过程中会产生三元金属间相,这可能会导致放电电压平台升高,从而导致电池能量密度增加。因此,Zn 对电池能量密度的贡献甚至高于 Bi。对于能量密度与电流

密度的偏导数，即倍率性能，所获得的元素相关关系顺序是可以理解的。唯一的偏差点是 Se 和 Te。它们是电子传导性低得多的非金属和半金属材料。Se-Cu 和 Te-Cu 合金是高能量密度阴极材料的代表性样品。高导电性的 Cu 赋予了 Se-Cu 和 Te-Cu 阴极优异的倍率性能，但这种贡献可能部分归因于计算过程中 Se 和 Te 本身，从而高估了它们对倍率性能的贡献程度。上述分析与 VIF 结果保持一致，也证明了本研究所使用的领域专家知识评价数据保真度的正确性。

3.1.5 小结

本节中，我们采用领域专家知识去评价数据保真度，基于三条原则从多保真度液态金属电池数据集中筛选出高保真度数据。随后，我们使用数据驱动的机器学习方法进行实验并结合专家知识对其进行分析，验证了该评价方法的正确性。由此，我们为研究者们提供了一个由本研究与多篇文献中可靠数据整合而成的高保真度数据集，旨在为该领域后续的研究者提供数据支持。

3.2 模型评价数据保真度

3.2.1 研究方案

传统的数据保真度评价方法通常依赖领域专家知识来确定数据保真度，缺乏针对具体模型的分析。实际上，同一组多保真度数据在不同模型下可能呈现出不同的保真度顺序。例如，某些被认为“低”保真度的数据在某个模型中可能训练效果较差，但在另一个模型中却能取得较好的表现。这表明，多保真度数据在不同模型中所能被学习到的信息存在差异，因此，同一数据集在不同模型下的保真度评价结果可能并不一致。如何根据具体模型合理利用多保真度数据，是本研究探讨的关键问题之一。

为此，本研究使用了两种材料领域较为流行的属性预测模型（MEGNet 和 CGCNN）用于实验。本研究首先计算不同模型在同一组多保真度数据集上所呈现出的数据保真度排序，并对比多保真度数据在不同模型上的排序结果。之后，我们提出了一种结合材料图神经网络与高斯过程回归（GPR）的材料属性预测模型。该模型利用图神经网络提取数据描述子，再将其作为 GPR 的输入进行训练，从而不仅实现了材料属性预测，还扩展了传统图神经网络无法提供的不确定性量化功能，该模型同样也被用于计算数据保真度排序。

在本节研究中所采用的多保真度数据集为材料领域密度泛函理论计算的带隙数据集，总共包含 4 种泛函计算的数据集：Perdew-Burke-Ernzerhof (PBE) [62]，

Heyd-Scuseria-Ernzerhof (HSE) [64,95], Strongly Constrained and Appropriately Normed (SCAN) [96]及 Gritsenko-Leeuwen-Lenthe-Baerends (GLLB) [61]; 使用实验 (Experiment, Exp) [1]获得的 2703 条带隙数据集作为最高保真度数据集并作为测试集, 所有数据集的分布如图 3.6 所示。其中, PBE 包含 52348 个样本, HSE 包含 6030 个样本, GLLB 包含 2290 个样本, SCAN 包含 472 个样本。Exp 来自于实验数据集, 包含 2703 个样本。为了更准确地评价数据保真度, 我们对 SCAN、HSE、PBE 和 GLLB 这四种数据集均采用定量方式重复训练 10 次。具体而言, 对于 HSE、PBE 和 GLLB, 我们随机采样 472 条数据 (与 SCAN 数据量一致) 进行训练, 重复 10 次; 而对于 SCAN 数据集, 则通过设定随机种子同样重复 10 次。最终, 对于每种保真度的数据集, 我们将获得 10 个训练好的模型, 并利用最高保真度的 Exp 数据集作为测试集, 分别计算其 MAE 和 MSE。若所计算的 MAE 和 MSE 较低, 则说明该数据集在当前模型下的预测结果与最高保真度数据集更为接近, 即该数据集的保真度较高。

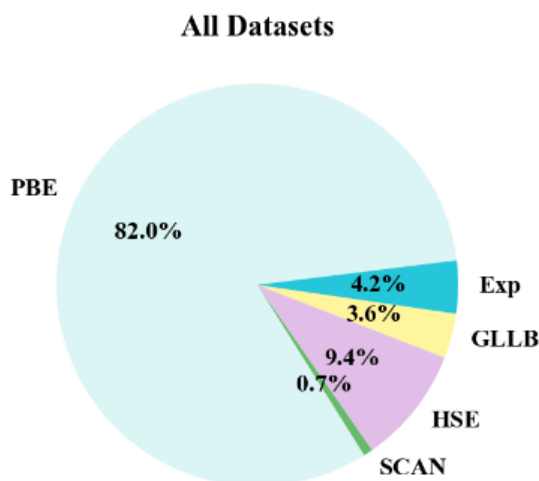


图 3.6. 带隙数据集分布情况

最后, 本研究基于材料图神经网络的不确定性量化结果, 对预测值进行加权修正。该方法从数据修正的角度利用多保真度数据, 初步探索了多保真度数据学习方法对模型最终预测精度的影响, 为后文数据学习方法的进一步研究奠定基础。

3.2.2 基于材料图神经网络的不确定性量化方法

本研究提出了一种通过提取材料图神经网络中数据描述子后经过高斯过程回归进行预测的不确定性量化方法, 使用了较为流行的材料图神经网络 MEGNet 和 CGCNN。有关这两个模型的描述在上文技术说明中已有介绍, 这里不过多阐述。对于 MEGNet 来说, 描述子的选取是在经过 MEGNet Block 和 set2set 层后,

所获取到的描述子是节点、边和全局状态的拼接，如图 3.7(a)所示；对于 CGCNN 来说，描述子的选取则选择在深度为 L2 的全连接隐藏层之后，如图 3.7(b)。

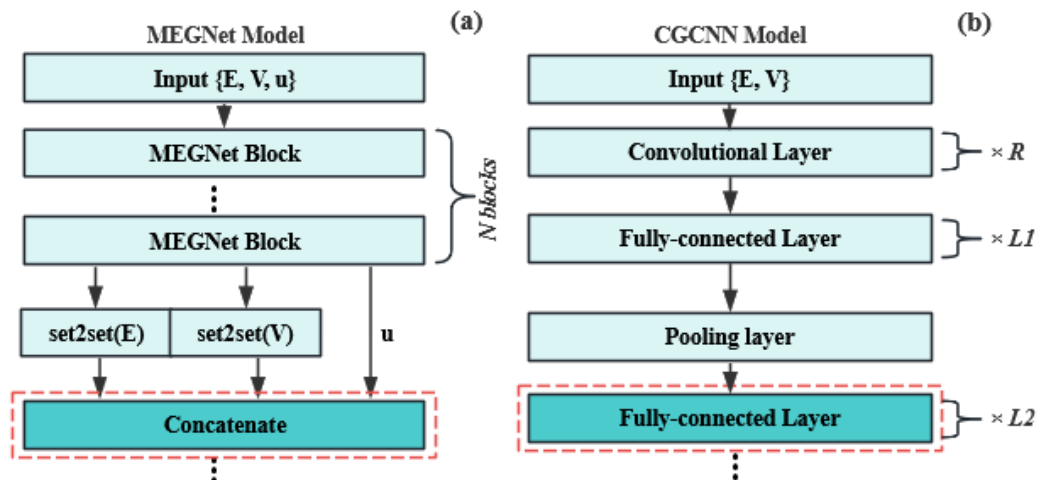


图 3.7 MEGNet 和 CGCNN 描述子提取位置

基于图 3.7 中模型的架构，本研究将每种保真度的数据集分别输入到两种材料图神经网络中，记录对应数据的描述子。其中，MEGNet 和 CGCNN 这两个模型训练时所使用到的超参数参考 Matbench^[72]排行榜中提及的超参数。具体超参数设置如下：

MEGNet: 训练集和验证集的比率为 8:2，batch size 为 32，学习率为 0.001；MEGNet Block 的个数为 3，采用 softplus 激活函数和高斯距离扩展，其中高斯距离扩展的 cutoff 为 5.0，sigma 为 0.4；采用早停策略，其中 patience 为 500。

CGCNN: 训练集和验证集的比率为 8:2，batch size 为 128，学习率为 0.01，卷积层个数为 3；采用 softplus 激活函数；采用早停策略，其中 patience 为 500。

其中未提及的超参数均使用各自模型的默认值。之后，将每种保真度对应的描述子分别作为 GPR 的输入进行训练，所使用到的 GPR 采用开源库 GPytorch，训练集和验证集的比率为 8:2，batch size 为 512，学习率为 0.1，采用径向基函数作为核函数。GPR 同样也采用早停策略，其中 patience 为 50。通过使用提取描述子后经过 GPR 训练的方法，可以获得在多保真度数据集上训练后模型对应的预测值和不确定性。

通过上述训练方式，针对于每种保真度数据集都会采用随机方式重复训练 10 次，对于每种保真度的数据集重复 10 次训练，一共存在 4 种保真度的数据集，因此会存在 40 个训练好的模型，这些模型均是通过材料图神经网络提取描述子后由 GPR 训练得到，具备不确定性量化的能力。

3.2.3 基于不确定性的加权修正方法

针对多保真度数据集，本研究利用 GPR 模型为测试集中每一条数据生成预测值和相应的不确定性。对于每一种保真度数据集，我们重复训练 10 次，共获得 4 种保真度下 40 组预测结果；每组预测结果均包含对应的预测值和不确定性。为从数据角度提升最终预测精度，我们采用标准化的逆比例加权方法，根据不确定性大小对预测值进行加权修正：不确定性越高，赋予的权重越低，且所有权重之和归一化为 1。

具体来说，对于每一条测试数据来说存在 40 组预测结果，这 40 组预测结果对应的权重计算公式如下所示：

$$w_{ij} = \frac{1}{\frac{u_{ij}}{S_i}} \quad (3.3)$$

$$S_i = \sum_{j=1}^{40} \frac{1}{u_{ij}} \quad (3.4)$$

其中 u_{ij} 表示第 i 条测试数据在第 j 组预测结果上对应的不确定性； w_{ij} 表示第 i 条测试数据在第 j 组预测结果上对应的权重； S_i 表示第 i 条测试数据对应的 40 组不确定性倒数的总和。然后，使用这些权重对每一条测试数据对应的 40 组预测值进行加权求和，如公式 3.5 所示：

$$\hat{y}_i^{(w)} = \sum_{j=1}^{40} w_{ij} \hat{y}_{ij} \quad (3.5)$$

其中 $\hat{y}_{ij}$ 表示第 i 条测试数据第 j 组的预测值， $\hat{y}_i^{(w)}$ 表示加权修正后的第 i 条预测值。这种加权方式能够充分利用不确定性量化结果，为每一组预测分配合适权重，从而更合理地整合多保真度数据集的信息，进而修正预测值。

加权修正的整体流程如图 3.8 所示，针对每种保真度数据集重复训练 10 次，得到对测试集 Exp 的 40 组预测结果。通过对这些预测结果的不确定性进行逆比例加权，我们初步探讨了多保真度数据学习方法从数据修正角度对模型预测精度的影响。

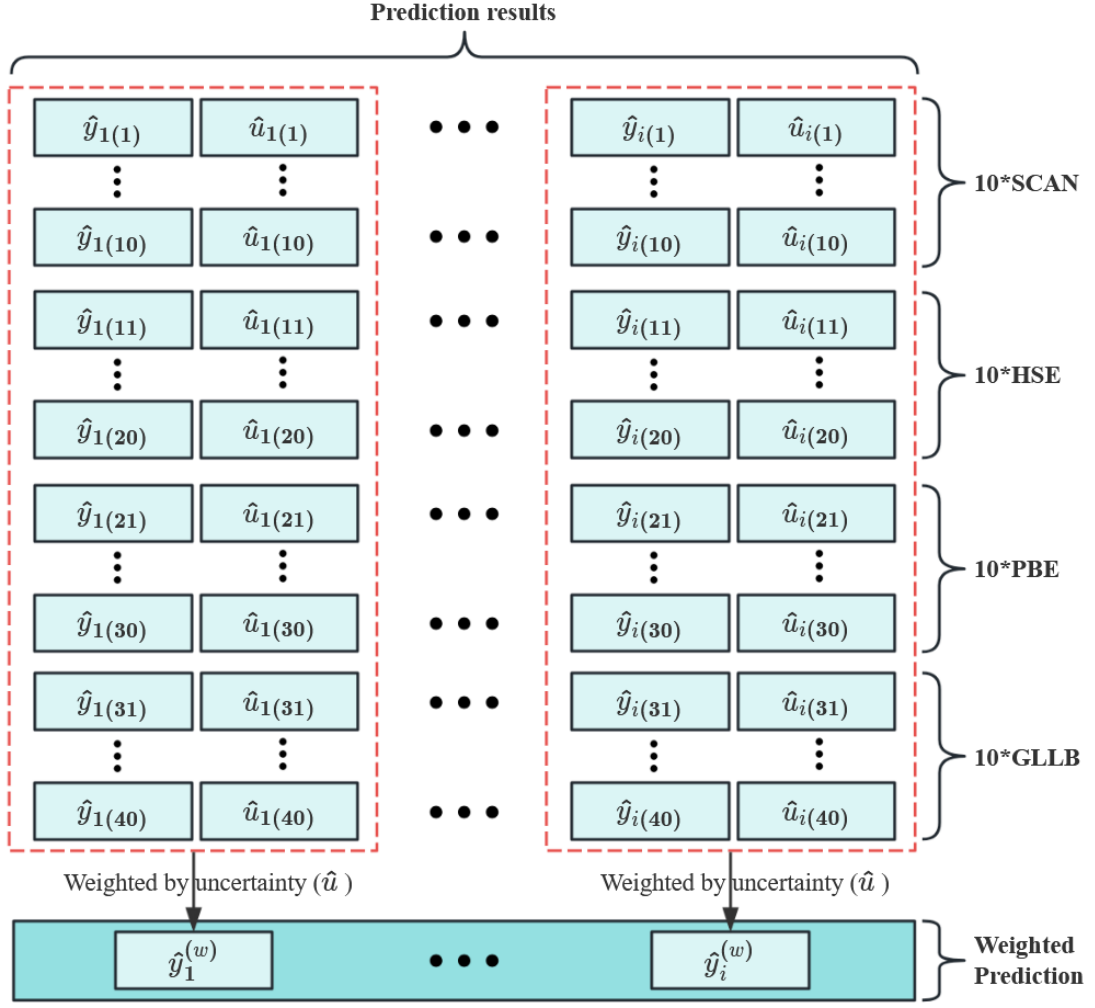


图 3.8 多保真度数据集预测结果加权示意图

3.2.4 实验结果及分析

通过提取描述子并结合高斯过程回归（GPR）训练，我们获得了多保真度数据集对应的 10 组预测结果及其不确定性。以实验数据集（Exp）作为测试集，计算不同模型得到的平均 MAE 和 MSE，结果如表 3.5 所示。其中， $MEGNet_{origin}$ 和 $CGCNN_{origin}$ 分别表示未做任何修改的原始模型，它们不具备不确定性量化功能； $MEGNet_{GPR}$ 和 $CGCNN_{GPR}$ 表示提取描述子后经过 GPR 的不确定性量化方法。其中 SCAN、HSE、PBE、GLLB 对应的 Exp MAE/MSE 表示 10 组测试结果的平均值。对于表 3.5 中的“Exp MAE/MSE”两列，是重复 10 次随机采样后平均的 MAE 和 MSE；“Exp MAE/MSE weighted”两列则表示使用多保真度数据集对应的不确定性对预测值进行加权求和的结果，对应公式 3.3 和 3.4。

表 3.5 不同方法的不确定性量化结果

模型	训练集	Exp MAE	Exp MSE	Exp MAE weighted	Exp MSE weighted
$MEGNet_{origin}$	SCAN	1.714	5.207	None	None
	HSE	1.273	3.635		
	PBE	1.154	3.038		
	GLLB	2.177	7.485		
$MEGNet_{GPR}$	SCAN	1.308	2.755	1.240	2.284
	HSE	1.163	2.649		
	PBE	1.154	2.600		
	GLLB	2.485	8.277		
$CGCNN_{origin}$	SCAN	1.669	4.203	None	None
	HSE	1.489	3.814		
	PBE	1.639	4.404		
	GLLB	1.994	5.518		
$CGCNN_{GPR}$	SCAN	1.024	2.152	0.755	1.365
	HSE	0.855	1.871		
	PBE	0.858	1.845		
	GLLB	1.350	3.690		

从表 3.5 前四列可以看出，原始模型中 $MEGNet_{origin}$ 在 HSE 和 PBE 数据集上的训练结果明显优于 $CGCNN_{origin}$ ，而在 SCAN 和 GLLB 数据集上的表现则相对较差。此外，采用 GPR 方法的模型在一定程度上提高了预测精度，但这种改进在 MEGNet 模型上并不显著。例如，在 PBE 数据集上， $MEGNet_{GPR}$ 的 MAE 几乎和原始模型 $MEGNet_{origin}$ 得到的结果一致，仅仅只是 MSE 结果差距较大。而在 GLLB 数据集中，使用 GPR 的 MEGNet 反而导致预测精度下降，这表明模型对 GLLB 描述子的学习能力较弱，GLLB 数据集在 MEGNet 模型中的保真度较低。相相比之下，采用 GPR 的 CGCNN 模型在所有保真度数据集上均表现出更优异的效果，其中 $CGCNN_{GPR}$ 的预测结果在各模型中最佳，其在基于 HSE 训练的单保真度数据集上取得的最低 MAE (0.855)，较原始 $CGCNN_{origin}$ 模型有显著提升。这表明 CGCNN 能够更有效地捕捉数据信息以构建描述子，可能原因在于其原子描述方式依托于人为经验，而 MEGNet 采用机器学习嵌入方法，在数据量较小的情况下，依赖人为经验获得的描述子可能更具优势。

针对表 3.5 的前四列，我们可以得出各个模型评价得到的数据保真度排序，排序结果如表 3.6 所示。可以发现，相对于不同模型来说，保真度顺序确实有所不同。其中 $CGCNN_{origin}$ 和 $CGCNN_{GPR}$ 所得到的排序结果一致； $MEGNet_{origin}$ 和 $MEGNet_{GPR}$ 所得到的排序结果一致；但 CGCNN 与 MEGNet 的排序不一致。当然，我们同样可以从中发现一些普遍规律：无论是 CGCNN、MEGNet 还是它们对应的不确定性量化模型，SCAN 和 GLLB 数据集相对来说保真度较低，其中

GLLB 保真度最低；而 HSE 和 PBE 数据集的保真度则相对较高。此外，在 CGCNN 模型中，HSE 数据集比在 MEGNet 模型中更易被学习；而在 MEGNet 模型中，PBE 数据集相对来说更容易提供有价值的信息。

表 3.6 不同模型对应的数据保真度排序

模型	保真度排序
$MEGNet_{origin}$	PBE>HSE>SCAN>GLLB
$MEGNet_{GPR}$	PBE>HSE>SCAN>GLLB
$CGCNN_{origin}$	HSE>PBE>SCAN>GLLB
$CGCNN_{GPR}$	HSE>PBE>SCAN>GLLB

以上实验结果证明了同一组多保真度数据在不同模型中呈现的数据保真度水平不同，而如何利用多保真度数据也是本研究的后续关键研究内容之一。因此，表 3.5 最后两列展示了本研究基于不确定性对多保真度数据预测值的加权修正结果。可以发现，相对于 $MEGNet_{GPR}$ 模型来说通过不确定性大小加权修正预测值的方式得到的结果 (MAE: 1.240) 并不如单保真度数据集中的最优结果 (对应 PBE 数据集训练结果, MAE: 1.154)。然而, $CGCNN_{GPR}$ 通过不确定性加权方式得到的结果达到了所有预测结果中的最佳, MAE 为 0.755, 相对于其在单保真度数据集中最佳预测结果的 MAE (0.855) 更低。这可能得益于 CGCNN 更加优秀的数据描述方式。这种通过不确定性大小来对预测值进行加权修正的方式在一定程度上能够对模型预测结果产生积极影响, 证明了多保真度数据学习方法的优势。该方法是本研究对于多保真度数据学习方法的初步探索, 在后文实验中, 本研究将使用更多、更全面的方法来研究多保真度数据学习方法。

3.2.5 小结

本节从模型的角度评价数据保真度水平, 针对具体的模型得到具体的数据保真度排序结果。此外, 本研究还尝试使用一种基于不确定性大小的数据修正方法来利用多保真度数据并提高模型精度。

本节使用了材料领域中四种泛函计算的多保真度带隙数据集 (SCAN、HSE、PBE、GLLB) 用于训练模型, 使用实验数据集 (Exp) 作为测试集评估同一组多保真度数据在不同模型下的保真度水平。此外, 本研究还提出了一种基于材料图神经网络的不确定性量化方法, 该方法将不仅用于评价数据保真度, 其预测结果还被用于融合多保真度数据信息。

实验结果表明, 同一组多保真度数据在不同模型下的保真度排序有所不同, 但是数据保真度的大趋势一致。例如, GLLB 和 SCAN 在 MEGNet 和 CGCNN 中

的保真度相对较低，而 HSE 和 PBE 的保真度则相对较高。结合了不确定性量化方法得到的数据保真度排序与上述结论一致，但在一定程度上比原始模型的预测结果更加优秀。

最后，为了探索多保真度数据学习方法的可行性，本研究利用不确定性大小对多保真度数据预测结果进行标准化的逆比例加权修正。结果发现，在 $CGCNN_{GPR}$ 模型基础上对多保真度数据集采用加权方式得到的结果相较于单保真度数据集最优测试结果来说更加优秀，MAE 最低仅有 0.755。该方法验证了基于数据修正的多保真度数据学习方法的有效性与可行性，为本研究后续开展更丰富、更复杂的多保真度数据学习方法奠定基础。

3.3 本章小结

本章主要介绍了两种数据保真度评价方法：领域专家知识评价方法和基于具体模型的评价方法，并通过实验加以验证。

在领域专家知识评价方法中，我们首先从文献和实验中收集了锂基液态金属电池阴极材料的多保真度数据集，该数据集涵盖了不同特征和计算方法。基于此，本研究提出了三条筛选准则，用以区分高、低保真度数据，并成功筛选出 61 条高保真度数据。随后，我们利用该高保真度数据集，结合线性回归、高斯过程回归和多层感知机三种模型，在同时考虑电池能量密度和倍率性能的基础上，通过多重共线性分析对各元素的重要性进行排序。最后结合传统理论进行对比，验证了领域专家知识评价方法的正确性。

在基于模型的评价方法方面，我们使用了 MEGNet 和 CGCNN 这两种常用的材料图神经网络评价数据保真度。此外，还提出了一种结合材料图神经网络与高斯过程回归的不确定性量化方法。通过将不同泛函得到的多保真度带隙数据集分别输入给模型，我们得到不同模型对应的数据保真度排序。实验结果表明，同一组多保真度数据在不同模型中的排序存在差异。此外，针对高斯过程回归得到的不确定性，本研究利用其对多保真度数据预测结果进行加权修正，为后文多保真度数据学习方法的进一步研究奠定基础。

第4章 多保真度数据学习方法

本章节研究两种多保真度数据学习方法，一种为线性修正方法另一种为非线性修正方法。对于线性修正方法，本研究采用人为构造的多保真数据集，通过引入加性、乘性和混合噪声，系统评估了加性修正、乘性修正以及综合修正三种线性修正方法。非线性修正方法使用了理论计算的带隙数据集，通过引入线性噪声和采样噪声，结合不同的训练策略对多保真度数据集进行迭代降噪，定量地评价了数据学习方法在含噪多保真度数据集上的效果。

4.1 线性修正方法

4.1.1 研究方案

线性修正方法在其他领域的应用已相当普遍，但对其进行系统评估的研究却相对较少。因此，本节首先采用人工构建多保真度数据集的方法，对线性修正方法进行系统性评估，并引入三种不同类型的线性噪声，以探讨具体线性修正方法在不同应用场景下的适用性。线性修正方法的基本思想是构建代理模型来学习高、低保真度数据集之间的信息，然后利用该信息对低保真度数据的结果进行修正。本节中所采用的代理模型为高斯过程回归。

本研究首先构造了一个由 10 个超参数组成的方程作为理想方程，用于计算真实值，如公式 4.1 所示：

$$Y_{true}(x) = p_1 * \sin(p_2 * x^2 + p_3) * (p_4 * x^2 + p_5) + p_6 * (p_7 * x^3 + p_8) + p_9 * x + p_{10} \quad (4.1)$$

其中 $p_1 \sim p_{10}$ 是随机参数，数值是-5 到 5 之间的整数。我们随机选择了 100 组参数来用于实验。用于实验的低保真（LF）数据的输入值为从 0 到 1 等间距分布的 200 个点，记为 X_l 。高保真（HF）数据的输入值包含在低保真数据的输入值中，为从 0 到 1 等间距分布的 20 个点，记为 X_h 。高保真和低保真数据的响应值分别用 Y_h 和 Y_l 表示，对应的计算方法如公式 4.2、4.3 所示：

$$Y_h = E_{h*} * Y_{true} + E_{h+} \quad (4.2)$$

$$Y_l = E_{l*} * Y_{true} + E_{l+} \quad (4.3)$$

其中， E_{h*} 和 E_{l*} 分别表示 HF 和 LF 数据下的乘性噪声； E_{h+} 和 E_{l+} 分别表示高保真和低保真数据下的加性噪声。总的来说，HF 数据中的噪声比 LF 中的较小。

上述公式用于构建混合噪声下的多保真度数据,对于加性或乘性噪声的多保真度数据来说,则需要排除一些噪声添加项。具体而言,对于加性噪声的多保真度数据,我们将参数 E_{h*} 和 E_{l*} 设置为 1,以生成仅包含加性噪声的多保真度数据集;对于乘性噪声的多保真度数据集来说,将参数 E_{h*} 和 E_{l*} 设置为 0,以生成仅包含乘性噪声的多保真度数据集。其中,对于包含加性噪声的数据集,生成范围为 0 到 1 的随机数,并将这些数分别乘以 0.4 和 0.8,以得加性噪声 E_{h+} 和 E_{l+} 。对于乘性噪声,从范围 0.9 到 1.1 中随机选取数值表示 E_{h*} ,从范围 0.8 到 1.3 中随机选取数值表示 E_{l*} 。添加了不同噪声的 HF 和 LF 数据集分别被随机打乱作为训练集,而测试集的输入值是 HF 和 LF 数据集中未包含的从 0 到 1 的额外二十个点。使用均方误差 (Mean Squared Error, MSE) 作为评估指标,并使用不同的参数 ($p1 \sim p10$) 随机重复实验 100 次,计算每种线性修正方法与真值之间的 MSE。仅通过高保真度数据训练得到的模型 (High-Fidelity Data Model, HFDM) 将作为基准来进行对比。

基于以上人为构造的含噪多保真度数据集,本研究系统地评估了三种线性修正方法对模型精度的影响,探索其适用的场景。

4.1.2 线性修正方法

线性修正方法主要分为 3 种,分别是加性修正,乘性修正和综合修正,它们使用不同的修正策略以适用于不同场景。

加性修正 (Additive Correction, AC) ^[97] 通过构建高保真度模型和低保真度模型之间差值的代理模型来修正低保真度模型的响应值,以此作为高保真度模型响应值的近似,用公式可以表示为:

$$\hat{y}_h(x) = y_l(x) + \delta(x) \quad (4.4)$$

其中 $y_l(x)$ 表示低保真度模型在 x 处响应值, $\hat{y}_h(x)$ 是高保真度模型在 x 处响应值的近似, $\delta(x)$ 是一个代理模型,表示高低保真度模型在 x 处的差异,也称为差异函数。将低保真度模型响应值 $y_l(x)$ 与差值函数 $\delta(x)$ 之和作为高保真度模型响应值的近似,称为 AC。

乘性修正 (Multiplicative Correction, MC) ^[98] 通过构造高保真度模型和低保真度模型之间比值的代理模型来修正低保真度模型的响应值,用公式可以表示为:

$$\hat{y}_h(x) = \rho(x)y_l(x) \quad (4.5)$$

其中 $\rho(x)$ 表示高低保真度模型之间比值的代理模型。将低保真度响应值 $y_l(x)$ 与比值函数 $\rho(x)$ 的乘积作为高保真度模型响应值的近似,称为 MC。此外,上文所使用到的不确定性量化的加权修正方法则是利用不确定性大小提供的权重去乘上对应的预测值,其本质上仍然是一种 MC。在本节中则采用公式 4.5 的

乘性修正方法用于实验。

当一个模型无法用简单的 AC 或者 MC 来近似时，可以考虑使用综合修正 (Comprehensive Correction, CC) [99,100]。综合修正方法综合了以上两种方法，用公式可以表示为：

$$\hat{y}_h(x) = \rho(x)y_l(x) + \delta(x) \quad (4.6)$$

其中 $\rho(x)$ 和 $\delta(x)$ 分别是乘法修正代理模型和加法修正代理模型，而 $\rho(x)$ 通常被一个常数 ρ 所替代，如公式 4.7 所示。共同克里金方法便是一种常见的综合修正方法 [101-103]。本研究则使用公式 4.7 作为综合修正方法用于实验。

$$\hat{y}_h(x) = \rho y_l(x) + \delta(x) \quad (4.7)$$

其中， ρ 的值通过最小化公式 4.8 求得，这里使用了最小二乘法来计算。公式 4.8 和 4.9 可以被视为关于 ρ 的二次函数，可以计算它的导数并将其设置为零进行求解。 $\hat{y}_l(\cdot)$ 由代理模型在低保真度数据集上训练获得，并用于预测高保真度数据集 x_h 对应的响应值。

$$\min: (\delta(x_h) - d_h)^2 \quad (4.8)$$

$$d_h = \rho \hat{y}_l(x_h) - y_h \quad (4.9)$$

以上三种线性修正方法中关于代理模型 $\delta(x)$ 或 $\rho(x)$ 的实现本研究均采用了 GPR。具体来说，通过 GPR 训练高、低保真度之间的差值或比值，将训练好的模型作为代理模型代入上述公式修正低保真度数据。这三种线性修正方法都会被应用在人为构造的加性噪声、乘性噪声和混合噪声的多保真度数据集上，以此验证方法的好坏及其适用场景并从中寻找规律。

4.1.3 实验结果及分析

为了较为准确地评估线性修正方法，本研究随机选择 100 组公式 4.1 中的超参数进行测试，并计算不同方法得到的 MSE，取 100 组预测结果的平均值用于最终评判。图 4.1 中，(a) 表示公式 4.1 中选择 100 组随机参数所得到的真值函数。粗体表示公式 4.10 中参数的所对应的真值函数。(b) 表示对应真值函数的放大图。

为了能够较为直观展示各种方法之间的对比，本研究选择 100 组实验结果中的一组用于展示高、低保真度数据分布情况以及各种方法的实现效果。所选择的参数数组为 p_{sample} ，表示公式 4.1 中的超参数 $p1 \sim p10$ 的数组，具体数值如公式 4.10 所示。

$$p_{sample} = [-1, -5, -4, 5, -1, -2, 1, 0, -4, -2] \quad (4.10)$$

其中与参数 p_{sample} 对应的真值函数图形如图 4.1 中加粗曲线所示，我们在该真值函数上展示线性修正结果。

该真值函数对应的 3 种不同噪声类型的多保真度数据如图 4.2 中(a)(c)(e)所示, 而 3 种线性修正方法 (AC, MC 和 CC) 和仅使用高保真度数据训练的模型 (HFDM) 的拟合结果如图 4.2(b)(d)(f)所示。

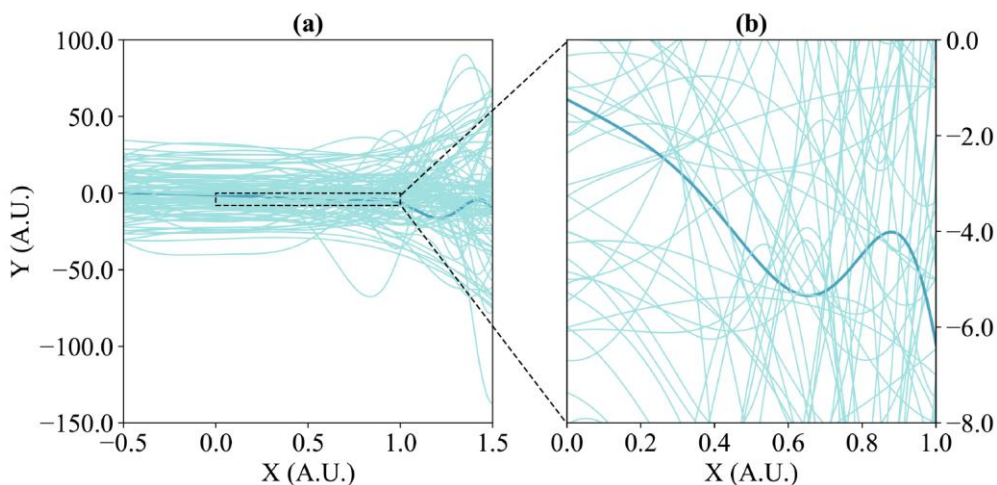


图 4.1 所有的真值函数曲线

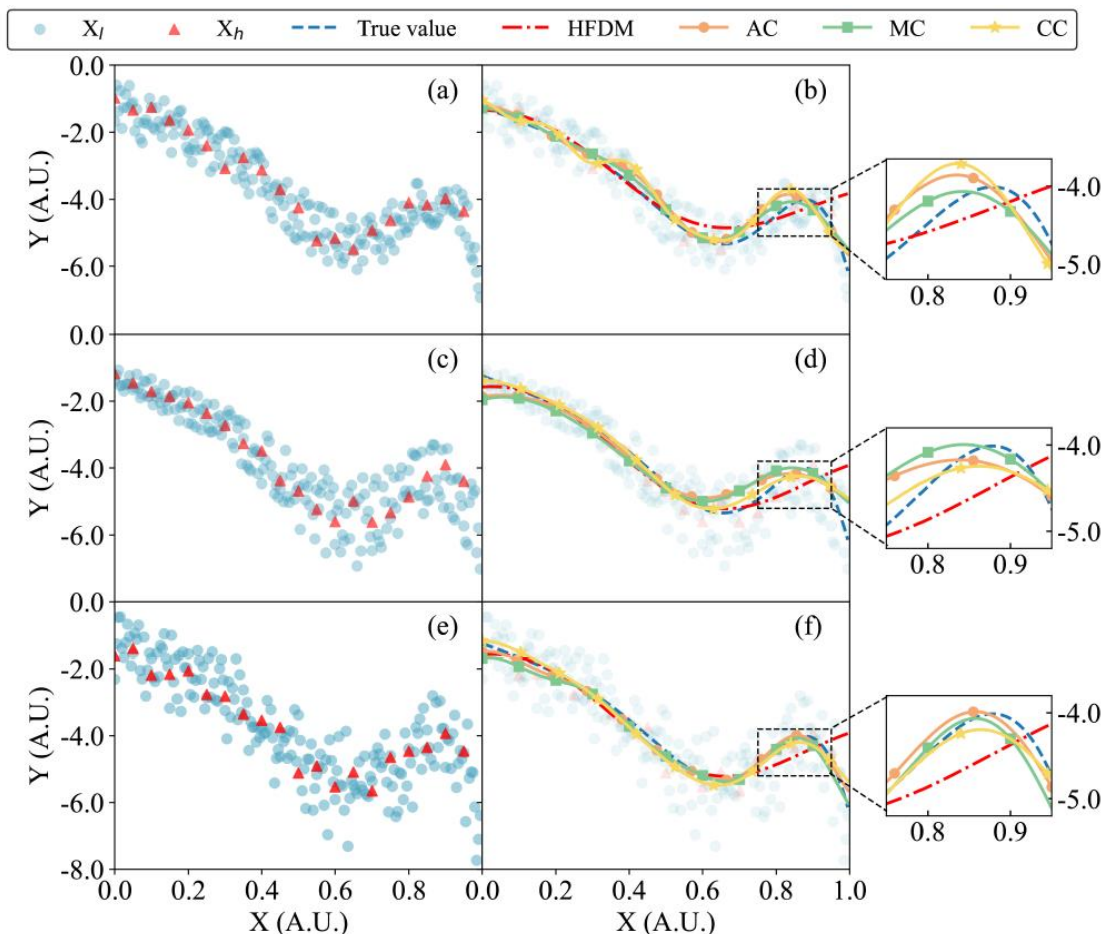


图 4.2 一百次实验结果中的其中一次

从图 4.2 中不难发现, 在 0.75 到 0.95 的范围内线性修正方法获得的结果比高保真数据模型 (HFDM) 获得的结果更接近 “True value”。这是因为高保真度数据的数据量较小, 导致模型拟合的精度较低。相反, 线性修正方法可以同时使

用高保真度和低保真度数据，未知位置的高保真度数据可以用已知的低保真度数据近似，因此拟合结果相对较好。

以上是 100 次实验中的其中一次，对于每次的模型构建中，本研究均使用了 5 折交叉验证方法以确定最佳模型参数，并在此基础上进行预测和测试。随后，我们计算了每种线性修正方法与真实值之间的 MSE，同时也计算了仅使用高保真度数据模型（HFDM）的 MSE，100 次实验的平均结果如表 4.1 所示。结果显示，对于含有任意噪声的多保真度数据集，综合修正（CC）的 MSE 最低，而 HFDM 的 MSE 最高；对于仅存在加性噪声的多保真度数据集，加性修正（AC）的 MSE 低于乘性修正（MC）的 MSE；而对于仅存在乘性噪声的数据集，MC 的 MSE 低于 AC。总体而言，这些线性修正方法的预测结果均优于仅使用高保真度数据训练得到的模型。综上所述，线性修正方法通过对低保真度数据响应值的修正，能够显著提高最终的预测精度，并在对应的线性噪声多保真度数据集下表现出较优的效果，其中综合修正方法适用于任何线性噪声条件下的多保真度数据。

表 4.1 不同方法在不同噪声的多保真度数据集上的平均 MSE

方法	平均 MSE		
	加性噪声	乘性噪声	混合噪声
HFDM	2.545	2.531	2.389
AC	0.065	0.753	0.501
MC	0.608	0.609	1.535
CC	0.057	0.436	0.425

4.1.4 小结

本节旨在探究线性修正方法在含噪多保真度数据集中的应用效果和提升规律。为此，我们首先通过人为构造不同类型的线性噪声（包括加性、乘性和混合噪声）的多保真度数据集，对加性修正、乘性修正和综合修正三种方法进行了系统评估。实验结果表明，线性修正的多保真度数据学习方法相较于仅使用高保真度数据训练的模型能够有效提高预测精度。此外，使用高斯过程回归作为代理模型学习高、低保真度数据集之间的差异是有效的，针对不同场景使用不同修正方法也能够有效提高模型预测精度。

在材料领域数据驱动机器学习日益重要的背景下，从单一保真度向多保真度数据驱动范式转变已成必然，不同场景下采用针对性的多保真度数据学习方法有望进一步提升模型的最终预测性能。

4.2 多保真度数据迭代降噪方法

4.2.1 研究方案

从上文中线性修正方法表现出的效果可以发现,简单的线性修正难以应对含有复杂噪声的多保真度数据集,而材料领域的多保真度数据往往包含不同程度和类型的噪声。因此,本节主要聚焦于非线性修正的多保真度数据学习方法,与上一小节关于线性修正方法的探索形成互补。具体而言,本节基于材料领域中的理论计算数据集,研究了迭代降噪的非线性修正方法。与上一节构造多保真度数据集的方式类似,我们在理论计算得到的带隙数据集中引入不同类型和不同程度的噪声,以定量方式评估该数据学习方法的效果。为此,本研究采用材料图神经网络 MEGNet 对带噪数据集进行迭代降噪,探讨在不同组合训练方式及噪声条件下该方法对模型精度的提升规律。

本节所使用到的数据集同样来自上文提到的 MP 数据库,该数据集中包含来自泛函 PBE, HSE, SCAN 及 GLLB 计算的材料禁带宽度数据。表 4.2 展示了这些多保真度数据集的数据量及禁带宽度的分布情况。

表 4.2 不同禁带宽度 (E_g) 数据集数据分布情况。

泛函	$E_g = 0$		$E_g < 2$		$E_g \geq 2$		Total	
	数量	比例	数量	比例	数量	比例	数量	比例
PBE	19903	38%	15218	29%	17227	33%	52348	100%
HSE	2775	46%	1019	17%	2236	37%	6030	100%
SCAN	16	3%	291	63%	165	35%	472	100%
GLLB	0	0%	531	23%	1759	77%	2290	100%

可以看出, PBE 数据集不仅数据量最大,而且禁带宽度的分布也更为均匀。因此,在本节研究中,我们选取该数据集作为真值数据集,并在此基础上构建多保真度数据。具体来说,我们将 52348 条数据随机划分为 5 个子集,其中训练集(简称 A、B、C、D 集)分别包含 A: 30000, B: 10000, C: 5000, D: 2348 条数据,测试集包含 5000 条数据。鉴于文献报道指出, DFT 计算所得禁带宽度数据通常较真实值低估 30%~100%^[60]。为此,本实验设计 3 个噪声等级,将以上各数据集处理成多保真度数据集,具体噪声系数如表 4.3 所示。

确定噪声等级后,本研究设计了 2 种噪声的添加方式来模拟多保真度数据:

1) 按比例缩小:指遵循公式 $T_{noise} = T_{PBE} * (1 - \theta)$ 对目标值进行处理。此种添加方式通过设计一个系数与数据本身进行相乘,在不同数据集中引入了线性噪声。

2) 采样噪声:指遵循公式 $T_{noise} = T_{PBE} * \max(1 - \sigma)$ 对目标值进行处理。

其中 σ 为采样自正态分布 $\sigma \sim N(\theta, \theta^2)$ 的随机值。此种添加方式在不同数据集中引入了更复杂的非线性噪声。其中 θ 为各数据集在表 4.3 中噪声系数； T_{noise} 为添加噪声后的属性值； T_{PBE} 为 PBE 数据集中原始的属性值。

在此基础上，本研究通过不同组合训练方式以及降噪函数，以迭代的方式来修正上述含噪类型不同、含噪程度不同的多保真度数据。

表 4.3 不同噪声等级应用于不同数据集的噪声系数 (θ)

数据集	噪声系数 $\theta/\%$		
	高噪声	中等噪声	低噪声
A	90	60	30
B	60	40	20
C	30	20	10
D	0	0	0

4.2.2 训练策略及降噪方法

多保真度数据集包含了多个不同保真度及不同大小的子数据集，如何对其加以利用是一个值得推敲的问题。结合本研究团队前期工作中^[1]的训练方式，本研究使用了以下几种数据集的训练策略：

- 1) 逐个训练 (“1by1”)：对应于以上数据集中 $A \rightarrow B \rightarrow C \rightarrow D$ 的训练顺序；
- 2) 洋葱训练 (“Onion”)：对应于以上数据集中 $ABCD \rightarrow BCD \rightarrow CD \rightarrow D$ 的训练顺序，每次训练剔除噪声程度最高、保真度水平最低的数据集；
- 3) 集中训练 (“All together”)：对应于“洋葱训练”方式的第一步 $ABCD$ ，即简单混合所有保真度数据集后进行训练；
- 4) 单独训练 (“Only true”)：对应于以上数据集中 D 数据集（根据表 4.3 中所示，未添加任何噪声），仅训练没有噪声、保真度最高的数据集，这里将其训练结果作为基准。

所测试的机器学习模型选用 MEGNet，在迭代降噪过程中，降噪函数如公式 4.11 所示：

$$T_{denoised} = \begin{cases} P, \text{abs}(P - T_{noise}) > \gamma \\ T_{noise}, \text{abs}(P - T_{noise}) \leq \gamma \end{cases} \quad (4.11)$$

其中 T_{noise} 表示降噪前的数据属性值； P 表示前一个模型的预测值； $T_{denoise}$ 表示降噪后的属性值；超参数降噪阈值 $\gamma = 0.35$ 。这个公式表示，在每一次迭代过程中，如果预测值与原始值之间的误差小于等于降噪阈值，则不修改原始值；如果预测值与原始值之间的误差高于降噪阈值，那么则使用预测值替代原始值。模型在训练过程中同样采用早停的方法防止过拟合，其中 patience 为 10。

4.2.3 单保真度数据集大小及噪声的影响

为了探究噪声及数据集大小对模型训练的影响并为后续实验提供基准，在引入噪声前后的 A、B、C、D 数据集及全部集合 ABCD 上分别进行训练和测试，测试方法使用平均绝对误差(MAE)，结果如图 4.3 所示。其中图 4.3(a)和图 4.3(b)表示 MEGNet 模型在不同类型噪声添加的数据集中的训练结果，图 4.3(a)为对数据集添加线性噪声后训练结果；图 4.3(b)为对数据集添加采样噪声后训练结果。显然，从这 2 幅热力图中可以清楚看出噪声存在对模型预测性能的影响。当无噪声添加时（对应于图 4.3(a)、图 4.3(b)中最下面“None”一行），数据量是影响最终结果的唯一因素，数据越多，最终训练效果越好。同时，MEGNet 模型在本数据集上的性能极限也被限定在 0.4eV 左右。添加噪声后，训练结果显著恶化，不难理解，恶化程度与添加噪声量正相关。在图 4.3 低噪声（噪声等级“Low”对应行）的训练结果中，多保真度数据集协同训练性能显著优于单保真度数据训练结果。这是一个典型的多保真度数据训练特点。然而，低保真度数据并非一无是处，数据量巨大的低保真度数据即使最简单的混合后训练，也可以大幅提高模型性能。

在噪声达到一定程度后，更大噪声的影响完全覆盖数据量的优势。图 4.3(a)和图 4.3(b)的高噪声（“High”行）及中等噪声（“Med”行）添加方案中，大数据、大噪声的数据集 A 单独训练所得结果（图 4.3(a)中 High 对应的 1.32 eV 以及 Med 对应的 0.95 eV，图 4.3(b) High 对应的 1.04 eV 以及 Med 对应的 0.97 eV）已明显逊色于小数据、低噪声数据集 B（图 4.3(a)中 High 对应的 1.01 eV 以及 Med 对应的 0.81 eV，图 4.3(b) High 对应的 1.00 eV 以及 Med 对应的 0.92 eV）和数据集 C 单独训练的结果（图 4.3(a) High 对应的 0.82 eV 以及 Med 对应的 0.83 eV，图 4.3(b) High 对应的 0.81 eV 以及 Med 对应的 0.77 eV）。而 ABCD 全集的训练结果（对应训练策略中的“集中训练”方法结果），随着噪声强度的增加，也逐渐逊色于数据量最小的 D 训练结果。需说明的是，图 4.3(a)、图 4.3(b)最右一列均为相同数据集无噪声添加的训练结果，理论上结果应保持一致。在固定随机数种子后，目前理解实际的细微差异是由 GPU 并行计算过程引入。但这仍不妨碍我们由此给出基线“单独训练”方法的结果：0.76~0.78eV，这将用于评估后续迭代降噪方法的好坏。

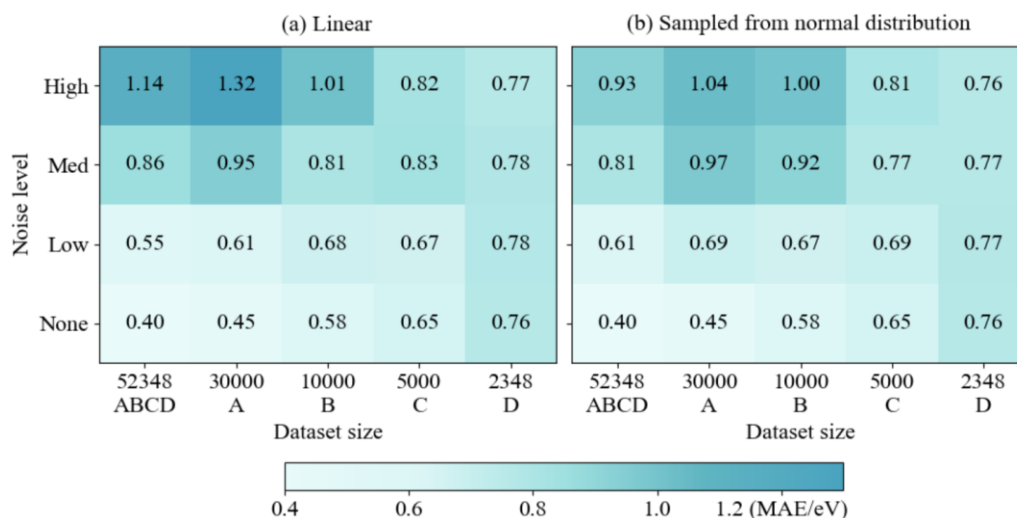


图 4.3 多保真度数据集上训练结果和所添加的噪声类型。(a)表示模型在线性噪声数据集下的训练结果；(b)表示模型在采样噪声数据集下的训练结果。

4.2.4 不同训练策略的比较

在获得各数据集单独训练模型的结果后,本研究进一步将实验推进到对多保真度数据的利用上。图 4.4 中,从下至上的 4 条曲线分别代表了不同噪声添加量(无、低、中、高)的实验结果。为了更清晰比较“逐一训练”和“洋葱训练”方法的性能,对应不同方法采用从两边向中间的方式进行绘图。图 4.4(a)为线性噪声,图 4.4(b)为采样噪声。图中上面的 3 条线为含噪多保真度数据训练结果,可以发现“洋葱训练”(ABCD $\rightarrow$ BCD $\rightarrow$ CD $\rightarrow$ D)与“逐一训练”(A $\rightarrow$ B $\rightarrow$ C $\rightarrow$ D)在不同等级噪声存在下均优于“集中训练”(ABCD, 等效于“洋葱训练”中的第 0 步)。不难理解含噪量越大,最终训练所得模型的平均绝对值误差(MAE)越高。

综合来看,除线性噪声的高噪声情况外,“洋葱训练”训练方式在同样多保真度数据集上均优于“逐一训练”的训练方式。此外,图 4.4 中最下方线条组为无噪声添加时各方法训练效果,可以看出单一保真度数据集训练第一步结果(与数据集大小强相关)便决定了最终性能表现,后续步骤对性能提高影响不大。无论何种方法,在任意数据噪声添加方案上,在最终 Step=3 上的结果均优于图 4.3 中基线“单独训练”给出的 0.76~0.78eV,这再次体现了多保真度数据学习方法的优势。

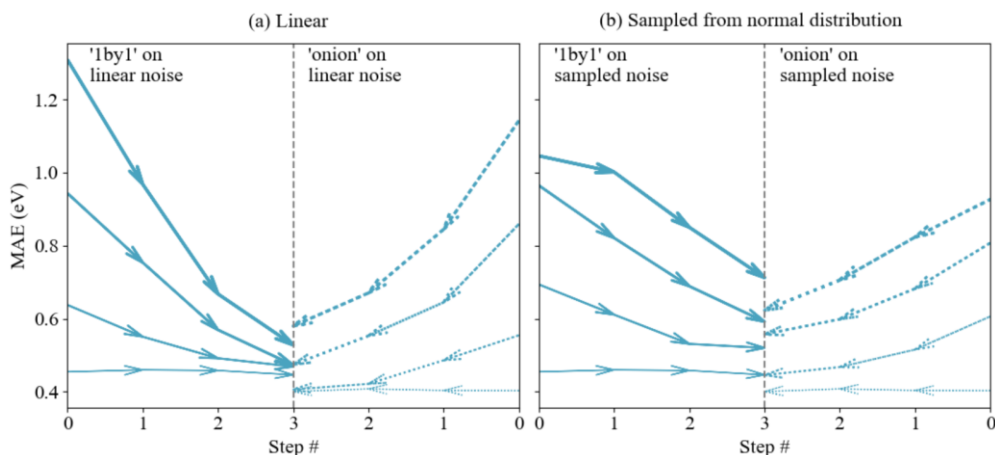


图 4.4 迭代降噪方法在不同噪声等级、不同噪声类型上的比较

4.2.5 迭代降噪的效果

为了较为清晰地展示迭代降噪效果,本节展示了前两次迭代降噪后不同方法对应的误差,如图 4.5 所示。其中,图 4.5(a)(c)对应数据中添加线性噪声的训练结果,而图 4.5(b)(d)所对应的是添加采样噪声的训练结果,图中虚线为不同方式在无噪声数据上的训练结果。

从图 4.5 中可以发现,总体来看线性噪声对模型的影响小于采样噪声,特别是“洋葱训练”得到的结果,其普遍比“逐一训练”方法得到的结果要更加优秀。这可能归因于神经网络强大的线性拟合能力,使得最后一步使用少量(仅 2348 条数据)高保真度的 D 数据集可以轻易地校正模型。然而,线性噪声的多保真度数据集在迭代降噪方面却并不能保证每次迭代降噪过程中都能够降低误差,除图 4.5(c)中“洋葱训练”方法在高噪声数据上的效果外,其余噪声的数据集在第一次迭代降噪后的误差均未出现成功的下降,而这一组所谓的下降,也是由于初次太差导致(与 4.2.4 节相一致)。这进一步说明模型第一轮训练中已经将数据中的信息学习较成功,线性噪声较易被修正。

与之相对应,对于图 4.5(b)、图 4.5(d)中的采样噪声,仅仅使用“逐一训练”或“洋葱训练”的训练方式已很难应对,未经迭代降噪的 2 种多保真度数据学习方法所得结果皆与无噪声数据训练结果有明显差距。这时迭代降噪的优势边能够显示出来,经过两轮迭代降噪可以分别平均降低“逐一训练”及“洋葱训练”这 2 种方法的预测误差 2.1%和 4.1%。综上所述,采样噪声被认为是一种与现实噪声更接近的定量噪声。虽然“洋葱训练”策略得到的结果整体来看更好,但少量的真值数据难以对其训练出的模型进行高效的纠偏,需使用其他降噪算法进一步对整体数据集进行提升。

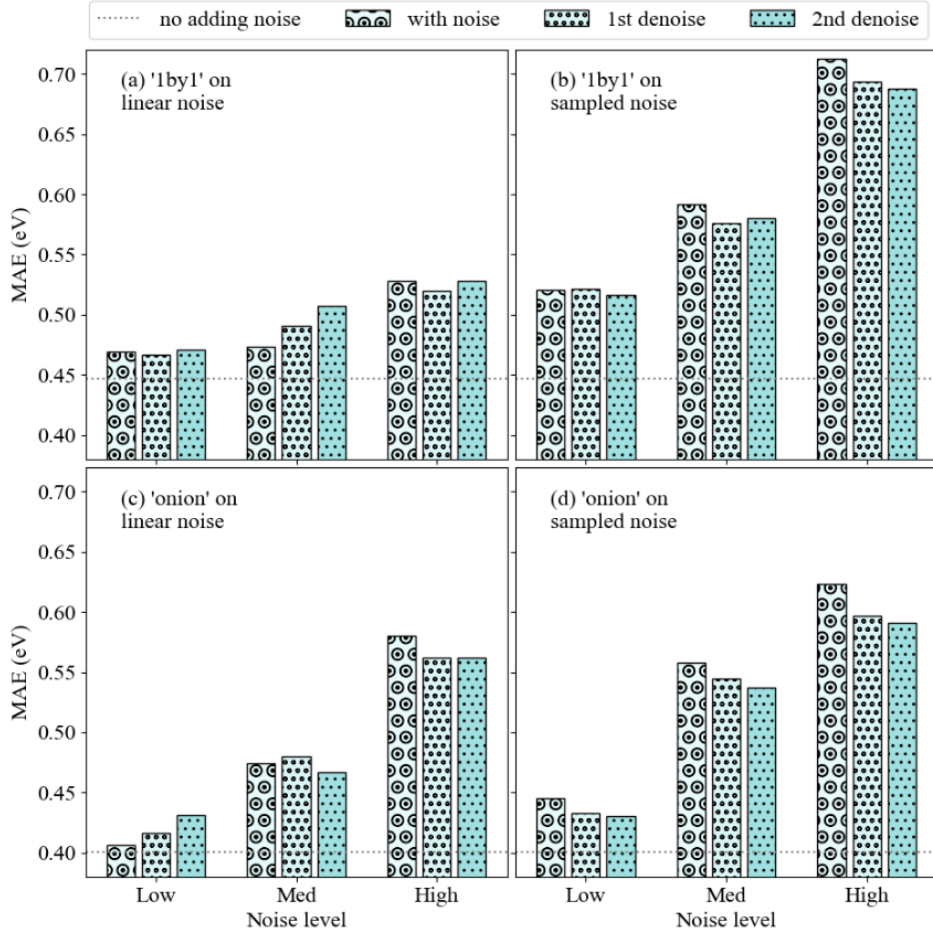


图 4.5 迭代降噪方法在不同噪声上的降噪效果

4.2.6 小结

本节从非线性修正的角度探讨了多保真度数据学习方法,与上一小节线性修正方法形成互补。我们在理论计算的带隙数据集上添加了不同程度和类型的噪声构建多保真度数据集,并研究了迭代降噪的非线性修正方法在不同训练策略下对模型精度的提升规律。实验结果表明,多保真度数据的有效利用至关重要,需要综合考虑数据量和噪声情况。在构造的不同噪声数据集中,得益于数据间的协同效应,逐步剔除低保真度数据的“洋葱训练”方法明显优于按噪声减小顺序的“逐一训练”方法。无论采用何种噪声强度及训练方式,线性噪声对模型的影响较小,训练得到的模型结果与无噪声真值数据的结果更为接近;而对于采样噪声,各个环节均较好地模拟了真实多保真度数据集。定量噪声评价策略有望为未来合理规划数据生产提供指导,从而更好地权衡成本与精度。

本节进一步丰富了多保真度数据学习方法的实现策略,从非线性修正的角度进行研究。同时,考虑了不同的训练策略,探索了不同多保真度数据学习方法对模型精度的提升规律,为未来研究者对于多保真度数据的利用方式提供新的思路。

4.3 本章小结

本章节研究了线性修正和非线性迭代降噪这两种多保真度数据学习方法,通过多保真度数据集中引入不同类型、不同量的噪声来系统评价多保真度数据学习方法,从定量的角度分析了多保真度数据学习方法对模型的提升规律。

对于线性修正方法,本研究采用人为构造多保真度数据集的方式系统评价了加性、乘性和综合修正方法。实验结果表明,线性修正的多保真度数据学习方法相较于仅使用高保真度数据训练的模型能够有效提高预测精度。针对不同噪声场景使用特定修正方法能够有效提高模型预测精度。

与线性修正对应,非线性修正方法是对多保真度数据学习方法的进一步探索。本研究使用降噪函数在每次迭代后筛选低保真度数据,并使用模型预测结果对其进行修正。此外,我们还设计了“逐一训练”和“洋葱训练”两种训练策略,综合探索了不同方法在含噪类型、含噪程度不同的多保真度数据集下的训练效果。实验结果表明,使用了迭代降噪的非线性方法能够在含噪的多保真度数据集中取得较好结果,其中“洋葱训练”策略整体看来更加优秀。

在材料领域数据驱动机器学习日益重要的背景下,从单一保真度向多保真度数据驱动范式转变已成必然,不同场景下采用针对性的多保真度数据学习方法有望进一步提升模型的最终预测性能。

第5章 总结与展望

5.1 本文工作总结

多保真度数据在各领域中广泛存在，材料领域更是如此。然而，目前大多数材料属性预测的机器学习模型仅采用单一保真度数据，未能充分挖掘多保真度数据中蕴含的有价值信息。多保真度数据在材料领域中的应用正在逐渐受到关注，但从数据角度进行的研究少之又少。合理利用多保真度数据有助于提高模型预测精度，多保真度数据学习方法在数据驱动的机器学习模型中具有重要的科研价值和广泛的应用前景。

本研究首先总结了国内外多保真度数据的应用现状及其获取方法。结果显示，多保真度数据在各大领域中广泛存在，但大多数应用均侧重于模型层面的改进，鲜有从数据修正角度开展研究，尤其在材料领域中更为缺乏。针对这一问题，本文从数据保真度评价方法和多保真度数据学习方法两方面入手，力求更全面地探索多保真度数据学习方法，探索不同方法对模型精度的提升规律。其中，之所以需要对数据保真度评价方法进行研究，是因为考虑到多保真度数据学习方法需要区别对待多保真度数据集，合理的数据保真度评价方法直接影响模型最终结果。

针对数据保真度评价方法来说，本研究从领域专家知识和模型评价这两种角度去分析数据保真度。对于领域专家知识评价方法，本研究从文献和实验中收集了锂基液态金属电池不同组分和电流密度下能量密度的多保真度数据集。基于该多保真度数据集，本研究利用领域专家知识从电池的阴阳极质量负载、电流密度与能量密度匹配以及活性成分占比这三方面对其进行评价，从中筛选出高保真度数据。随后，在同时考虑能量密度和倍率性能的前提下，我们利用数据驱动的机器学习模型，并结合多重共线性和相关性系数，详细评估了各元素的重要性。实验结果与传统知识相符，有效验证了所采用的领域专家知识评价方法的正确性。同时，本研究筛选出的高保真度数据集也为后续在本领域的深入探索提供了坚实的数据支持。

与领域专家评价方法相对应，本研究从模型角度探讨了数据保真度。实际上，由于不同模型的学习能力各异，同一组多保真度数据在不同模型中可能呈现出不同的保真度水平。本研究选取了材料领域中通过不同泛函计算得到的带隙数据作为多保真度数据，并采用两种常用的材料图神经网络（MEGNet 和 CGCNN）进行评价。此外，我们基于材料图神经网络设计了一种不确定性量化方法，其通过

提取数据描述子并将作为高斯过程回归的输入，不仅能预测材料属性值，还能获得对应的不确定性。这种方法在一定程度上提高了模型预测精度，也同样被用于评估数据保真度。实验结果表明，同一组多保真度数据在不同模型中的保真度排序存在差异。为更有效地针对具体模型利用多保真度数据，本研究进一步采用不确定性对预测结果进行加权修正，从而提升最终模型的预测性能。这一加权修正策略是本研究对于多保真度数据学习方法的初步探索，为后续更深入的研究奠定了基础。

以上基于不确定性的加权修正方法本质上是多保真度数据学习方法中的乘性修正方法。实际上，多保真度数据学习方法中存在有更多的线性修正方法，例如加性、乘性和综合修正。本研究则是从这三种线性修正方法入手，探索多保真度数据学习方法对于模型精度的提升规律。此外，本研究不仅考虑了线性修正，同样也设计了一种迭代降噪的非线性修正方法。这些方法共同构成了本研究对多保真度数据学习方法的系统探索，旨在寻找不同场景下最合适的修正策略。

考虑到目前有关线性修正方法在含噪多保真度数据集上系统评估较少，我们首先采用人为构造多保真度数据集的方式，探索加性、乘性和综合修正三种方法在不同噪声环境下的适用性和提升效果。具体来说，本研究构造了分别含有加性、乘性和混合噪声的多保真度数据集，并使用高斯过程回归作为代理模型，对各方法的实验结果进行了对比。结果表明，与仅使用高保真度数据训练的模型相比，采用线性修正方法能显著提高预测精度；在对应噪声条件下，使用相应修正策略可进一步提升模型性能。在所有线性噪声多保真度数据集中，综合修正方法的表现最佳。

从综合修正方法表现出的效果可以发现，简单的线性修正难以应对含有复杂噪声的多保真度数据集，而材料领域的多保真度数据往往包含不同程度和类型的噪声。因此，本研究针对材料领域的理论计算数据集设计了非线性修正方法探索其对模型精度的提升规律。具体来说，我们以泛函 PBE 计算的带隙数据集为基础，在数据中分别引入不同程度的线性噪声和采样噪声，以模拟真实多保真度数据。模型训练过程中，我们设计了降噪函数，在每次迭代中筛选并修正低保真度数据，通过模型预测值替换原始数据属性来实现非线性修正。此外，我们提出了逐一训练和洋葱训练两种策略，评估不同训练方式下多保真度数据对模型预测性能的影响。实验结果表明，采用迭代降噪的非线性修正方法能够在含噪多保真度数据集上提高模型预测精度，其中洋葱训练的效果优于逐一训练。无论噪声强度或训练方式如何，线性噪声对模型影响较小，且训练结果更接近无噪声真值；而采样噪声则更真实地反映了多保真度数据的特性。总体而言，迭代降噪的非线性修正方法不仅拓展了多保真度数据学习策略，也为材料领域的多保真度数据利用

提供了新思路。

综合以上内容，本研究的创新点可分为以下三点：

(1) 基于材料图神经网络的不确定性量化方法

在材料领域，不确定性量化有助于指导材料设计并降低成本，但传统的材料图神经网络属性预测模型通常缺乏不确定性量化功能。为此，本研究提出了一种基于材料图神经网络的不确定性量化方法：从主流材料图神经网络中提取数据描述子，将其作为高斯过程回归的输入进行训练和预测。实验结果表明，该方法能在一定程度上提高模型预测精度，同时实现了不确定性量化。

(2) 多保真度数据学习方法的扩展

基于不确定性量化，本研究获取了多保真度数据集在具体模型中的不确定性指标，并据此提出了一种不确定性加权修正方法，对预测值进行权重调整。此外，本研究还设计了迭代降噪的多保真度数据学习方法，从非线性修正的角度利用多保真度数据。

(3) 含噪多保真度数据集下数据学习方法对模型提升规律的探索

本研究通过在多保真度数据集中引入不同类型和程度的噪声，系统探讨了各数据修正方法对模型预测精度的影响。对于线性修正方法，我们在人为构造的多保真度数据集中分别添加了加性、乘性和混合噪声；而在非线性修正方法中，则在理论计算的多保真度数据集中引入了线性和采样噪声。基于这些含噪数据集，我们评估了不同多保真度数据学习方法对最终预测精度的提升效果，并揭示了噪声类型和强度对模型表现的影响规律。

5.2 未来工作展望

针对当前研究工作的不足，本研究提出以下几个未来可能的研究方向：

(1) 材料属性预测模型与不确定性量化方法的更多结合

随着大语言模型和相关技术的发展，未来可探索多种融合策略。将不同类型的材料属性预测模型与不确定性量化方法相结合，以期在指导材料设计和降低研发成本方面取得更佳效果。

(2) 多保真度数据在特征层面的融合

本研究主要从输出数据角度进行修正，未来可尝试对输入特征进行融合，例如采用域对抗训练等方法。从数据预处理层面提升多保真度数据的利用效率，从而进一步提高模型预测性能。

(3) 基于不确定性量化指导的多保真度数据训练策略

利用不确定性量化获得针对具体模型的数据保真度信息，进而设计以不确定性为依据的多保真度数据训练策略，使模型能更精确地利用不同保真度数据，实现预测性能的进一步提升。

(4) 异构多保真度数据集的综合应用

未来可探讨如何整合来自理论计算、实验测量、结构、文本或图像等多种形式的多保真度数据，采用数据对齐等方法实现跨模态融合，为材料及其他领域中的数据驱动模型提供更为丰富和全面的输入信息。

总体来看，多保真度数据学习方法在材料领域仍有很大的发展空间，其他领域中的多保真度技术也有待融入材料领域。

致 谢

感谢刘晓彤老师的指导，感谢杨涛老师在生活上的关心与帮助，感谢北京科技大学材料领域专家赵海雷教授在材料数据集上给予的支持，感谢实验室各位对我的帮助。

参考文献

- [1] Xiaotong Liu, Pierre-Paul De Breuck, Linghui Wang, et al. A simple denoising approach to exploit multi-fidelity data for machine learning materials properties[J]. npj Computational Materials, Nature Publishing Group, 2022, 8(1): 1–13.
- [2] Chen C, Zuo Y, Ye W, et al. Learning properties of ordered and disordered materials from multi-fidelity data[J]. Nature Computational Science, Nature Publishing Group, 2021, 1(1): 46–53.
- [3] Mufti B, Chen M, Perron C, et al. A multi-fidelity approximation of the active subspace method for surrogate models with high-dimensional inputs[C] //AIAA Aviation 2022 Forum. 2022: 3488.
- [4] Hoop M V de, Huang D Z, Qian E, et al. The cost-accuracy trade-off in operator learning with neural networks[J]. arXiv preprint arXiv:2203.13181, 2022.
- [5] Zimmer L, Lindauer M, Hutter F. Auto-pytorch: Multi-fidelity MetaLearning for efficient and robust AutoDL[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(9): 3079–3090.
- [6] Arrieta A. Multi-fidelity digital twins: A means for better cyber-physical systems testing?[J]. arXiv preprint arXiv:2101.05697, 2021.
- [7] Jarin S, Yuan Y, Zhang M, et al. Predicting the crystal structure and lattice parameters of the perovskite materials via different machine learning models based on basic atom properties[J]. Crystals, Multidisciplinary Digital Publishing Institute, 2022, 12(11): 1570.
- [8] Zhao X, Liu D, Yan X. Diameter prediction of silicon ingots in the czochralski process based on a hybrid deep learning model[J]. Crystals, Multidisciplinary Digital Publishing Institute, 2023, 13(1): 36.
- [9] Gao P, Liu Z, Zhang J, et al. A fast, low-cost and simple method for predicting atomic/inter-atomic properties by combining a low dimensional deep learning model with a fragment based graph convolutional network[J]. Crystals, Multidisciplinary Digital Publishing Institute, 2022, 12(12): 1740.
- [10] Fernández-Godino M G. Review of multi-fidelity models[J]. Advances in Computational Science and Engineering, Advances in Computational Science and Engineering, 2023, 1(4): 351–400.
- [11] Wu D, Chinazzi M, Vespignani A, et al. Multi-fidelity hierarchical neural processes[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, 2022: 2029–2038.
- [12] Yoo K, Bacarreza O, Aliabadi M H F. Multi-fidelity robust design optimisation for composite structures based on low-fidelity models using successive high-fidelity corrections[J]. Composite Structures, 2021, 259: 113477.
- [13] Meng X, Babae H, Karniadakis G E. Multi-fidelity bayesian neural networks: Algorithms and applications[J]. Journal of Computational Physics, 2021, 438: 110361.
- [14] Gunst R F. Response surface methodology: Process and product optimization using designed experiments[J]. Technometrics, Taylor & Francis Group, 1996.
- [15] Brevault L, Balesdent M, Hebbal A. Overview of gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems[J]. Aerospace Science and Technology, 2020, 107: 106339.
- [16] Romor F, Tezzele M, Mrosek M, et al. Multi-fidelity data fusion through parameter space reduction with applications to automotive engineering[J]. International Journal for Numerical Methods in Engineering, 2023, 124(23): 5293–5311.
- [17] Absi G N, Mahadevan S. Multi-fidelity approach to dynamics model calibration[J]. Mechanical Systems and Signal Processing, 2016, 68–69: 189–206.
- [18] Pilania G, Gubernatis J E, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids[J]. Computational Materials Science, 2017, 129: 156–163.
- [19] Patra A, Batra R, Chandrasekaran A, et al. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap[J]. Computational Materials Science, 2020, 172:

- 109286.
- [20] Polak M P, Jacobs R, Mannodi-Kanakkithodi A, et al. Machine learning for impurity charge-state transition levels in semiconductors from elemental properties using multi-fidelity datasets[J]. *The Journal of Chemical Physics*, 2022, 156(11): 114110.
- [21] Egorova O, Hafizi R, Woods D C, et al. Multifidelity statistical machine learning for molecular crystal structure prediction[J]. *The Journal of Physical Chemistry A, American Chemical Society*, 2020, 124(39): 8065–8078.
- [22] Mannodi-Kanakkithodi A, Toriyama M Y, Sen F G, et al. Machine-learned impurity level prediction for semiconductors: The example of cd-based chalcogenides[J]. *npj Computational Materials*, Nature Publishing Group, 2020, 6(1): 1–14.
- [23] P. Greenman K, H. Green W, Gómez-Bombarelli R. Multi-fidelity prediction of molecular optical peaks with deep learning[J]. *Chemical Science*, Royal Society of Chemistry, 2022, 13(4): 1152–1162.
- [24] Khatamsaz D, Molkeri A, Couperthwaite R, et al. Adaptive active subspace-based efficient multifidelity materials design[J]. *Materials & Design*, 2021, 209: 110001.
- [25] Tran A, Tranchida J, Wildey T, et al. Multi-fidelity machine-learning with uncertainty quantification and bayesian optimization for materials design: Application to ternary random alloys[J]. *The Journal of Chemical Physics*, 2020, 153(7): 074705.
- [26] Sun G, Li G, Stone M, et al. A two-stage multi-fidelity optimization procedure for honeycomb-type cellular materials[J]. *Computational Materials Science*, 2010, 49(3): 500–511.
- [27] Islam M, Thakur M S H, Mojumder S, et al. Extraction of material properties through multi-fidelity deep learning from molecular dynamics simulation[J]. *Computational Materials Science*, 2021, 188: 110187.
- [28] Razi M, Narayan A, Kirby R M, et al. Fast predictive models based on multi-fidelity sampling of properties in molecular dynamics simulations[J]. *Computational Materials Science*, 2018, 152: 125–133.
- [29] Batra R, Pilania G, Uberuaga B P, et al. Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia[J]. *ACS Applied Materials and Interfaces*, 2019, 11(28).
- [30] Lamberti G, Gorié C. A multi-fidelity machine learning framework to predict wind loads on buildings[J]. *Journal of Wind Engineering and Industrial Aerodynamics*, 2021, 214: 104647.
- [31] Nagawkar J, Leifsson L. Multifidelity aerodynamic flow field prediction using random forest-based machine learning[J]. *Aerospace Science and Technology*, 2022, 123: 107449.
- [32] Kou J, Zhang W. Multi-fidelity modeling framework for nonlinear unsteady aerodynamics of airfoils[J]. *Applied Mathematical Modelling*, 2019, 76: 832–855.
- [33] Thelen A, Leifsson L, Beran P. Aeroelastic flutter prediction using multifidelity modeling of the generalized aerodynamic influence coefficients[J]. *AIAA Journal*, 2020, 58(11): 4764–4780.
- [34] Thelen A S, Leifsson L T, Beran P S. Multifidelity flutter prediction using regression cokriging with adaptive sampling[J]. *Journal of Fluids and Structures*, 2020, 97: 103081.
- [35] Singh D, Antoniadis A F, Tsoutsanis P, et al. A multi-fidelity approach for aerodynamic performance computations of formation flight[J]. *Aerospace, Multidisciplinary Digital Publishing Institute*, 2018, 5(2): 66.
- [36] Ariyarat A, Kanazaki M. Multi-fidelity multi-objective efficient global optimization applied to airfoil design problems[J]. *Applied Sciences, Multidisciplinary Digital Publishing Institute*, 2017, 7(12): 1318.
- [37] Huang L, Gao Z, Zhang D. Research on multi-fidelity aerodynamic optimization methods[J]. *Chinese Journal of Aeronautics*, 2013, 26(2): 279–286.
- [38] Elham A. Adjoint quasi-three-dimensional aerodynamic solver for multi-fidelity wing aerodynamic shape optimization[J]. *Aerospace Science and Technology*, 2015, 41: 241–249.
- [39] Li K, Kou J, Zhang W. Deep learning for multifidelity aerodynamic distribution modeling from experimental and simulation data[J]. *AIAA Journal, American Institute of Aeronautics and Astronautics*, 2022, 60(7): 4413–4427.
- [40] Ryou G, Tal E, Karaman S. Multi-fidelity black-box optimization for time-optimal quadrotor maneuvers[J]. *The International Journal of Robotics Research*, SAGE Publications Ltd STM, 2021, 40(12–14): 1352–1369.
- [41] 刘璟, 边梟, 徐冠峰, 等. 基于多可信度代理模型的尾喷管优化设计[J]. *航空工程进展*, 2022, 13(6): 29–39.

- [42] Brooks C J, Forrester A I J, Keane A J, et al. Multi-fidelity design optimisation of a transonic compressor rotor[C]//Proceedings of the 9th European conference on turbomachinery fluid dynamics and thermodynamics, 2011.
- [43] Shah H, Hosder S, Koziel S, et al. Multi-fidelity robust aerodynamic design optimization under mixed uncertainty[J]. *Aerospace Science and Technology*, 2015, 45: 17–29.
- [44] Lai X, He X, Wang S, et al. Building a lightweight digital twin of a crane boom for structural safety monitoring based on a multifidelity surrogate model[J]. *Journal of Mechanical Design*, 2022, 144(064502).
- [45] Shi R, Liu L, Long T, et al. Multi-fidelity modeling and adaptive co-kriging-based optimization for all-electric geostationary orbit satellite systems[J]. *Journal of Mechanical Design*, 2019, 142(021404).
- [46] Jacobs J P, Koziel S. Cost-effective global surrogate modeling of planar microwave filters using multi-fidelity bayesian support vector regression[J]. *International Journal of RF and Microwave Computer-Aided Engineering*, 2014, 24(1): 11–17.
- [47] Kim H S, Koç M, Ni J. A hybrid multi-fidelity approach to the optimal design of warm forming processes using a knowledge-based artificial neural network[J]. *International Journal of Machine Tools and Manufacture*, 2007, 47(2): 211–222.
- [48] Jin S-S, Kim S T, Park Y-H. Combining point and distributed strain sensor for complementary data-fusion: A multi-fidelity approach[J]. *Mechanical Systems and Signal Processing*, 2021, 157: 107725.
- [49] Abdallah I, Lataniotis C, Sudret B. Hierarchical kriging for multi-fidelity aero-servo-elastic simulators - application to extreme loads on wind turbines[J]. *arXiv preprint arXiv:1709.07637*, 2017.
- [50] Bu H, Yang Y, Song L, et al. Improving the film cooling performance of a turbine endwall with multi-fidelity modeling considering conjugate heat transfer[J]. *Journal of Turbomachinery*, 2021, 144(011011).
- [51] Mell L, Rey V, Schoefs F. Multifidelity adaptive kriging metamodel based on discretization error bounds[J]. *International Journal for Numerical Methods in Engineering*, 2020, 121(20): 4566–4583.
- [52] Koziel S, Pietrenko-Dabrowska A. Accelerated gradient-based optimization of antenna structures using multifidelity simulations and convergence-based model management scheme[J]. *IEEE Transactions on Antennas and Propagation*, 2021, 69(12): 8778–8789.
- [53] Li J, Yang A, Tian C, et al. Multi-fidelity bayesian algorithm for antenna optimization[J]. *Journal of Systems Engineering and Electronics*, 2022, 33(6): 1119–1126.
- [54] Palar P S, Shimoyama K. Multi-fidelity uncertainty analysis in CFD using hierarchical kriging[C]//35th AIAA applied aerodynamics conference. 2017: 3261.
- [55] Parussini L, Venturi D, Perdikaris P, et al. Multi-fidelity gaussian process regression for prediction of random fields[J]. *Journal of Computational Physics*, 2017, 336: 36–50.
- [56] Qiu Y, Song J, Liu Z. A simulation optimisation on the hierarchical health care delivery system patient flow based on multi-fidelity models[J]. *International Journal of Production Research*, Taylor & Francis, 2016, 54(21): 6478–6493.
- [57] Sajjadinia S S, Carpentieri B, Shriram D, et al. Multi-fidelity surrogate modeling through hybrid machine learning for biomechanical and finite element analysis of soft tissues[J]. *Computers in Biology and Medicine*, 2022, 148: 105699.
- [58] Biehler J, Gee M W, Wall W A. Towards efficient uncertainty quantification in complex and large-scale biomechanical problems based on a bayesian multi-fidelity scheme[J]. *Biomechanics and Modeling in Mechanobiology*, 2015, 14(3): 489–513.
- [59] Panda K, King R, Maack J, et al. Visualization of multi-fidelity approximations of stochastic economic dispatch[C]//Proceedings of the Twelfth ACM International Conference on Future Energy Systems. New York, NY, USA: Association for Computing Machinery, 2021: 372–376.
- [60] Morales-García Á, Valero R, Illas F. An empirical, yet practical way to predict the band gap in solids by using density functional band structure calculations[J]. *The Journal of Physical Chemistry C, American Chemical Society*, 2017, 121(34): 18862–18866.
- [61] Kohn W, Sham L J. Self-consistent equations including exchange and correlation effects[J]. *Physical Review*, American Physical Society, 1965, 140(4A): A1133–A1138.
- [62] Perdew J P, Burke K, Ernzerhof M. Generalized gradient approximation made simple[J]. *Physical Review Letters*, American Physical Society, 1996, 77(18): 3865–3868.

- [63] Adamo C, Barone V. Toward chemical accuracy in the computation of NMR shieldings: The PBE0 model[J]. *Chemical Physics Letters*, 1998, 298(1): 113–119.
- [64] Jie J, Weng M, Li S, et al. A new MaterialGo database and its comparison with other high-throughput electronic structure databases for their predicted energy band gaps[J]. *Science China Technological Sciences*, 2019, 62(8): 1423–1430.
- [65] Thompson A P, Swiler L P, Trott C R, et al. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials[J]. *Journal of Computational Physics*, 2015, 285: 316–330.
- [66] Forrester A I J, Bressloff N W, Keane A J. Optimization using surrogate models and partially converged computational fluid dynamics simulations[J]. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Royal Society, 2006, 462(2071): 2177–2204.
- [67] 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.
- [68] 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*, Nature Publishing Group, 2015, 1(1): 1–15.
- [69] Hong J. The band gap problem: The state of the art of first-principles electronic band structure theory[J]. *Progress in Chemistry*, 2012, 24(06): 910–927.
- [70] Xie T, Grossman J C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties[J]. *Physical Review Letters*, American Physical Society, 2018, 120(14): 145301.
- [71] Chen C, Ye W, Zuo Y, et al. Graph networks as a universal machine learning framework for molecules and crystals[J]. *Chemistry of Materials*, American Chemical Society, 2019, 31(9): 3564–3572.
- [72] Dunn A, Wang Q, Ganose A, et al. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm[J]. *npj Computational Materials*, Nature Publishing Group, 2020, 6(1): 1–10.
- [73] Forrester A I J, Sóbester A, Keane A J. Multi-fidelity optimization via surrogate modelling[J]. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Royal Society, 2007, 463(2088): 3251–3269.
- [74] Perdikaris P, Raissi M, Damianou A, et al. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling[J]. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Royal Society, 2017, 473(2198): 20160751.
- [75] Goel T, Haftka R T, Shyy W. Comparing error estimation measures for polynomial and kriging approximation of noise-free functions[J]. *Structural and Multidisciplinary Optimization*, 2009, 38(5): 429–442.
- [76] Le Gratiet L, Cannamela C. Cokriging-based sequential design strategies using fast cross-validation techniques for multi-fidelity computer codes[J]. *Technometrics*, ASA Website, 2015, 57(3): 418–427.
- [77] Takeno S, Tsukada Y, Fukuoka H, et al. Cost-effective search for lower-error region in material parameter space using multifidelity gaussian process modeling[J]. *Physical Review Materials*, American Physical Society, 2020, 4(8): 083802.
- [78] Deringer V L, Bartók A P, Bernstein N, et al. Gaussian process regression for materials and molecules[J]. *Chemical Reviews*, American Chemical Society, 2021, 121(16): 10073–10141.
- [79] Shi M, Lv L, Sun W, et al. A multi-fidelity surrogate model based on support vector regression[J]. *Structural and Multidisciplinary Optimization*, 2020, 61(6): 2363–2375.
- [80] Noack M M, Doerk G S, Li R, et al. Autonomous materials discovery driven by gaussian process regression with inhomogeneous measurement noise and anisotropic kernels[J]. *Scientific Reports*, Nature Publishing Group, 2020, 10(1): 17663.
- [81] Park C, Haftka R T, Kim N H. Remarks on multi-fidelity surrogates[J]. *Structural and Multidisciplinary Optimization*, 2017, 55(3): 1029–1050.
- [82] Vinyals O, Bengio S, Kudlur M. Order matters: Sequence to sequence for sets[J]. *arXiv preprint arXiv:1511.06391*, 2015.
- [83] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 770–778.
- [84] Wang K, Jiang K, Chung B, et al. Lithium–antimony–lead liquid metal battery for grid-level

- energy storage[J]. *Nature*, Nature Publishing Group, 2014, 514(7522): 348–350.
- [85] Xie H, Chu P, Yang M, et al. A novel sb-zn electrode with ingenious discharge mechanism towards high-energy-density and kinetically accelerated liquid metal battery[J]. *Energy Storage Materials*, 2023, 54: 20–29.
- [86] Li H, Wang K, Zhou H, et al. Tellurium-tin based electrodes enabling liquid metal batteries for high specific energy storage applications[J]. *Energy Storage Materials*, 2018, 14: 267–271.
- [87] Ning X, Phadke S, Chung B, et al. Self-healing li–bi liquid metal battery for grid-scale energy storage[J]. *Journal of Power Sources*, 2015, 275: 370–376.
- [88] Chu P, Wang J, Xie H, et al. Sb-cu alloy cathode with a novel lithiation mechanism of ternary intermetallic formation: Enabling high energy density and superior rate capability of liquid metal battery[J]. *Journal of Energy Chemistry*, 2023, 78: 393–400.
- [89] Li H, Wang K, Cheng S, et al. High performance liquid metal battery with environmentally friendly antimony–tin positive electrode[J]. *ACS Applied Materials & Interfaces*, American Chemical Society, 2016, 8(20): 12830–12835.
- [90] Zhou H, Li B, Yu M, et al. Accelerated design of electrodes for liquid metal battery by machine learning[J]. *Energy Storage Materials*, 2023, 56: 205–217.
- [91] Dai T, Zhao Y, Ning X-H, et al. Capacity extended bismuth-antimony cathode for high-performance liquid metal battery[J]. *Journal of Power Sources*, 2018, 381: 38–45.
- [92] Yeo J-S, Lee J-H, Yoo E-J. Electrochemical properties of environment-friendly lithium-tin liquid metal battery[J]. *Electrochimica Acta*, 2018, 290: 228–235.
- [93] Zhao W, Li P, Liu Z, et al. High-performance antimony–bismuth–tin positive electrode for liquid metal battery[J]. *Chemistry of Materials*, American Chemical Society, 2018, 30(24): 8739–8746.
- [94] Kim J, Shin D, Jung Y, et al. LiCl–LiI molten salt electrolyte with bismuth-lead positive electrode for liquid metal battery[J]. *Journal of Power Sources*, 2018, 377: 87–92.
- [95] Heyd J, Scuseria G E, Ernzerhof M. Hybrid functionals based on a screened coulomb potential[J]. *The Journal of Chemical Physics*, 2003, 118(18): 8207–8215.
- [96] Sun J, Ruzsinszky A, Perdew J P. Strongly constrained and appropriately normed semilocal density functional[J]. *Physical Review Letters*, American Physical Society, 2015, 115(3): 036402.
- [97] Lewis R M, Nash S G. Model problems for the multigrid optimization of systems governed by differential equations[J]. *SIAM Journal on Scientific Computing*, Society for Industrial and Applied Mathematics, 2005, 26(6): 1811–1837.
- [98] Alexandrov N M, Dennis J E, Lewis R M, et al. A trust-region framework for managing the use of approximation models in optimization[J]. *Structural optimization*, 1998, 15(1): 16–23.
- [99] Zhou Q, Wu Y, Guo Z, et al. A generalized hierarchical co-Kriging model for multi-fidelity data fusion[J]. *Structural and Multidisciplinary Optimization*, 2020, 62: 1885–1904.
- [100] Liu X, Zhao W, Wan D. Multi-fidelity Co-Kriging surrogate model for ship hull form optimization[J]. *Ocean engineering*, 2022, 243: 110239.
- [101] Toal D J J. Applications of multi-fidelity multi-output Kriging to engineering design optimization[J]. *Structural and Multidisciplinary Optimization*, 2023, 66(6): 125.
- [102] Yi J, Cheng Y, Liu J. A novel fidelity selection strategy-guided multifidelity kriging algorithm for structural reliability analysis[J]. *Reliability Engineering & System Safety*, 2022, 219: 108247.
- [103] Liu J, Shi Y, Ding C, et al. Hybrid uncertainty propagation based on multi-fidelity surrogate model[J]. *Computers & Structures*, 2024, 293: 107267.

个人简历、在学期间发表的学术论文及研究成果

个人简历：王滋明，出生日期：1999.12.10，2022 年在北京信息科技大学获学士学位。

学术论文：

- [1] 第一作者. Using Data-Driven Methods to Analyze the Roles of Different Elements in Liquid Metal Batteries[J]. Journal of Energy Storage, 2025, 106: 114802. (SCI top)
- [2] 第一作者. Exploring Multi-Fidelity Data in Materials Science: Challenges, Applications, and Optimized Learning Strategies[J]. Applied Sciences, 2023, 13(24): 13176. (SCI)
- [3] 第二作者. 多保真度数据学习算法的定量噪声评价[J]. 硅酸盐学报, 2023, 51(02): 405-410. (EI)

其他成果：

- [1] 2024 年 11 月 研究生国家奖学金
- [2] 2024 年 11 月 专利：一种元素重要性排序方法及系统 申请号：2024116812154
- [3] 2024 年 11 月 专利：一种数据保真度评价方法及系统 申请号：2024116805644
- [4] 2024 年 9 月 竞赛：第二届世界科学智能大赛物质科学赛道—催化反应产率预测 初赛 25/575 名 复赛 19/42 名
- [5] 2024 年 7 月 北京高校大学生创新创业校际合作项目 优秀组第二名 学生贡献者第一名
- [6] 2023 年 11 月 研究生三等奖学金
- [7] 2022 年 10 月 软著：QM9 分子数据库处理软件 V1.0 登记号：2023SR0400880