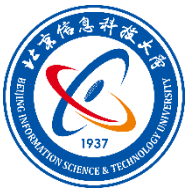
单位代码：11232　　分类号：TP391

密级：公开



北京信息科技大学

工学硕士学位论文

不同嵌入方法对材料结构信息表征能力的研究

学　　院：计算机学院

专　　业：计算机科学与技术

学　　号：2023020604

作　　者：陈昊天

指导教师：刘晓彤 副教授

完成日期：2026 年 3 月 20 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：陈昊天

2026 年 3 月 20 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：刘晓彤　　学位论文作者签名：陈昊天

2026 年 3 月 20 日　　2026 年 3 月 20 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《不同嵌入方法对材料结构信息表征能力的研究》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：陈昊天

2026 年 3 月 20 日

摘 要

随着材料信息学的发展，机器学习已经成为加速材料性能预测与优化的重要工具。在材料图神经网络等模型中，元素嵌入表征能力直接影响模型性能。然而，当前研究往往将元素嵌入作为模型中的次要环节，通过特征拼接或可学习查找表的方式简单实现，且二者被视为互斥选择。事实上，元素作为一种自然界的基本物质单元，其内在化学相似性与周期规律蕴含着丰富的结构信息。因此，系统研究元素嵌入的构建方式及其对模型性能影响，不仅有助于揭示机器学习模型学习机制，也将为后续材料智能设计提供重要基础。基于此，本文从单一表征方法的元素嵌入和混合元素嵌入两方面探索元素表征方法，以期为元素表示提供新思路。

在单一嵌入方法对材料结构信息表征能力的研究方面，将元素间化学可替换性映射到一维空间，并平衡替代关系与元素间距，构建非等间距元素分布。针对高维元素嵌入，本文在两种典型材料图神经网络模型上（CGCNN、MEGNet），分别对基于专家知识的手工设计嵌入与基于数据驱动的可学习嵌入两类元素表示方法进行比较。比较内容涵盖性能预测、嵌入空间结构，并深入探究元素聚类行为。

在混合嵌入方法对材料结构信息表征能力的研究方面，上述两种元素嵌入在材料建模中常被视为互斥选择，而其可融合性往往被忽视。本文在上述两种模型中构建并行双通道元素输入结构，通过调节混合比例参数控制两类嵌入的融合。两类嵌入方法的余弦夹角方差随嵌入维度增加而减小，且融合后的材料属性预测结果大比例优于传统基线，均支持不同嵌入方法在高维嵌入空间中存在互补性的推断。在此基础上，进一步聚合多个模型学习得到的元素关系，构建元素嵌入表 Mat2Vec-*。实验结果表明，相较于领域经典嵌入 Mat2Vec，Mat2Vec-*在 CrabNet 模型的组分任务中综合性能提升 4.3%。

关键词：化学元素嵌入；材料信息表征；材料属性预测；潜空间向量

ABSTRACT

With the development of materials informatics, machine learning has become an important tool for accelerating the prediction and optimization of material properties. In models such as materials graph neural networks, the representational capacity of element embeddings directly affects model performance. However, existing studies often treat element embeddings as a secondary component, implementing them in a simplistic manner through feature concatenation or learnable lookup tables, and these two approaches are typically regarded as mutually exclusive. In fact, as fundamental units of matter in nature, elements possess intrinsic chemical similarities and periodic patterns that contain rich structural information. Therefore, systematically studying the construction of element embeddings and their impact on model performance will not only help reveal the learning mechanisms of machine learning models but also provide an important foundation for subsequent intelligent materials design. Based on this, this work explores element representation methods from two perspectives: single-representation element embeddings and hybrid element embeddings, aiming to provide new insights into element representation.

In the study of the representational capacity of single embedding methods for material structure information, chemical substitutability among elements mapped into a one-dimensional space. By balancing substitution relationships and inter-element distances, a non-equidistant distribution of elements is constructed. For high-dimensional embeddings, this work conducts a systematic comparison of two types of element representation methods, expert knowledge-based handcrafted embeddings and data-driven learnable embeddings, on two representative materials graph neural network models (CGCNN and MEGNet). The comparison covers predictive performance, embedding space structure, and further investigates element clustering behavior.

In the study of the representational capacity of hybrid embedding methods, the aforementioned two types of embeddings are often treated as mutually exclusive in materials modeling, while their potential complementarity is largely overlooked. To address this, a parallel dual-channel element input framework is constructed within the same models, where the fusion of the two embeddings is controlled by a mixing ratio parameter. It is observed that the variance of cosine angles between the two embedding

types decreases as the embedding dimension increases. Moreover, the hybrid embeddings consistently outperform traditional baselines in most material property prediction tasks, supporting the hypothesis that different embedding methods exhibit complementary characteristics in high-dimensional latent spaces. Building upon this, element relationships learned from multiple models are further aggregated to construct a transferable element embedding table, Mat2Vec-*. Experimental results demonstrate that, compared to the domain-standard embedding Mat2Vec, Mat2Vec-* achieves a 4.3% overall performance improvement on composition-based tasks in the CrabNet model.

KEY WORDS: chemical element embedding; materials information characterization; materials property prediction; latent space vector

目 录

第1章 绪 论

1.1 研究背景及意义

随着计算材料学与机器学习技术的发展，基于数据驱动的方法正在逐渐改变传统材料研究模式[1]。传统材料研究依赖经验积累和实验试错，周期长、成本高，机器学习模型通过学习已有材料数据规律，实现材料性质预测与新材料筛选[2,3]，显著提高材料发现效率。

元素周期表长期以来为理解元素性质及其相互关系提供了重要框架。周期表结构体现了元素性质随核电荷数递增而呈现的周期性变化，揭示原本看似独立的元素之间的内在联系，为理解化学反应性、原子结构和成键行为提供了系统化的理论基础。然而，在现代计算材料研究中，元素周期表往往并不能直接满足数据驱动建模和高通量计算的需求[4,5]。

在高通量筛选场景中，元素替换是探索成分空间最有效的策略之一[4,6–8]。由于材料成分组合的数量极其庞大，穷举探索所有可能的成分组合在现实中不可行[9]。因此，在给定的结构或化学环境下，如何快速判断哪些元素之间具有较高的可替换性，成为材料设计中的关键问题，需要一种能够有效刻画元素相似性关系的表示方式。在材料替代设计或成分搜索过程中，若能构建一种反映元素化学相似性的低维表示方法，可以更直观、高效地理解元素之间的关系，为材料筛选与设计提供重要参考。

近年来，在材料科学领域的图神经网络[10]（Graph Neural Network，GNN）凭借其卓越的结构归纳偏置，在材料性质预测任务中展现出卓越性能，成为材料信息学研究的重要工具[11–14]。在材料图神经网络中，元素表示是模型输入的重要组成部分，它决定了模型如何理解不同化学元素之间的关系以及它们在材料结构中的作用。通常每个化学元素会被映射为一个向量表示，即元素嵌入，在模型训练过程中参与消息传递与特征更新。元素嵌入所包含的信息，直接影响模型对化学规律的表达能力，从而影响材料性质预测的准确性。

当前人工智能驱动的材料建模可以分为两大类：一类是基于材料描述符的机器学习方法，另一类是端到端深度学习方法。这两类方法所采用的特征表示方法，可按其生成方式划分为三类：捕捉物理化学和结构特性（描述符和指纹）的表示方法、图论驱动的材料图形表示方法、使用深度学习和自然语言处理算法的数据驱动嵌入方法。当前许多材料机器学习研究[11–14]，在模型设计时将不同元素嵌入方法视为互斥的设计选择，这些嵌入方法的表征能力及其对模型性能的影响，仍

缺乏系统性的比较研究。另外，不同来源的信息是否能在表示空间中形成互补关系，也有待进一步探索。

在实际研究中，不同模型训练得到的元素嵌入通常具有潜在空间未对齐和任务依赖性，这使得不同模型之间学习到的嵌入向量难以直接比较和复用，限制了元素嵌入在跨任务和跨模型应用中的可迁移性。因此，如何从大量模型中提炼稳定的元素关系结构，并构建具有良好泛化能力和可迁移性的元素嵌入表示，是当前材料机器学习研究中的重要问题。

1.2 国内外研究现状

本节将对国内外材料结构结果信息表征的相关研究进行综述，涵盖元素表征与相似性研究、材料结构信息表征的嵌入方法研究以及嵌入空间协同表示与信息融合研究三方面的工作。

1.2.1 元素表征与相似性研究

元素周期表不仅是化学元素的简单排列，更是自然界基本规律的重要体现。其结构反映了周期律，这一规律来源于核电荷的逐渐增加以及由此导致的元素性质周期性变化。通过这种方式，周期表揭示了原本看似独立元素之间的内在联系，为理解化学反应性、原子结构和成键行为提供了统一框架。然而，尽管周期表具有坚实的物理与化学基础，在现代计算材料科学与机器学习应用中，往往并不能满足复杂任务需求。部分研究中，研究人员往往采用更简化或功能性的分类方式，例如将元素划分为金属、非金属、卤素、过渡金属和惰性气体等，在计算模型或数据分析中作为具有化学意义的类别。

随着机器学习在材料发现[15,16]与性质预测[12,14]中的广泛应用，这些人为定义的分类标签也常被用于解释模型学习到的规律或验证模型性能。其中常用的元素分类标签，一种是基于传统化学知识的分类方式，在材料机器学习模型中被广泛采用[12,16–19]。另一种关注元素在具体化学环境中的行为，例如根据离子价态或更细致的化学特性进行分类[5,20–22]。随着数据驱动方法的发展，研究者开始尝试利用无监督学习或表征学习方法自动学习元素之间的相似性关系。例如，SkipAtom 嵌入[23]通过类似自然语言处理的方法学习元素之间的分布式表示；基于变分自编码器的方法[24]从电子构型中学习到元素的潜在坐标系统；PTG（Periodic Table Generator）[25]在无监督条件下重构周期表结构。此外，基于文献语料训练的语言模型嵌入，如 Mat2Vec[5]和 ElementBERT[26]，仅从材料科学文本中也能学习到元

素之间的化学规律与潜在知识，表明分布式表示在捕捉元素相似性方面具有潜力。

在材料发现中，元素之间的相似性尤为关键，因为它直接关系到材料组成空间的探索效率。在高通量筛选或生成式材料设计中，元素替代被认为是一种高效探索新材料的策略[27–29]。通过在已知晶体结构或化合物中替换元素，可以在较低计算成本下生成大量潜在候选材料。由于化学组成空间极其庞大，对所有可能组合进行穷举搜索在实践中不现实，需要依赖合理的元素相似性或分组策略来指导替代过程[9]。实际替代过程不仅需要考虑化学性质，还需要满足结构兼容性。其中，离子半径是影响替代可行性的关键因素之一，半径相近的元素更容易在晶体结构中互相替换而不会导致明显的晶格畸变。一些研究发现传统元素分类与实际化学行为之间存在差异。例如，Pettifor 标度[30]等方法通过一维排序重新组织元素，使相似元素在序列中更接近，后续研究提出基于相似性度量优化元素排序的方法[31]。这些序列聚集化学性质相似的元素同时，发现某些属于同一化学类别的元素在序列中相距较远，元素替代所依赖的化学相似性概念并不完全符合传统周期表分类。随着机器学习的发展，数据驱动模型逐渐能够自动学习元素之间细致关系。从图神经网络模型 MEGNet[12]学习得到的元素嵌入表明，一些镧系元素（如 Eu 和 Yb）在潜在空间中更接近碱土金属，而不是其他镧系元素。这种结果与化学直觉[32]，以及 Pettifor 提出的结构图[33]是一致的。

前期研究通过统计分析大规模材料数据库，发现元素替代的规律并用于新材料预测。例如，Hautier 等人[34]提出基于离子取代的数学模型利用已知晶体结构实验数据库挖掘元素之间的取代可能性，从而指导新化合物的生成。在此基础上，Wang 等人[4]进一步通过在已知晶体结构中用化学相似元素进行替换发现了新的稳定材料，并发现元素替代并非随机过程，而是呈现出明显规律。例如，第一周期元素几乎不能被其他元素替代（即所谓“第一周期异常”）；许多元素只能与周期表同族元素发生替代，这一趋势在碱金属和卤素中尤为明显；存在两个明显的金属替代家族，周期表内部存在某些非连续性，特别是在第 5 族与第 6 族之间。类似的研究还包括利用结构原型或已知化合物进行系统性元素替代以生成新的晶体结构[35,36]，以及通过改进基于周期表相似性的替代规则来提高材料发现效率[37]。近期的一些研究还将这种替代策略与机器学习势能模型结合，在结构预测与材料搜索中进行验证。通用机器学习原子间势能模型的压力测试，模型能够在元素替代驱动的结构预测流程中保持良好性能，证明在材料发现任务中的潜在应用价值[38]。

1.2.2 材料结构信息表征的嵌入方法研究

在材料信息学快速发展的背景下，如何有效表征材料结构信息已成为机器学习模型的关键问题。元素嵌入作为连接化学组成与机器学习模型的重要桥梁，不仅影响模型输入结构，还决定了模型能否有效学习到化学规律。表 1.1 选取了 5 个材料科学领域中经典图神经网络，总结其元素嵌入方式，以及键特征选择。

表 1.1 常见材料图神经网络元素及键嵌入

模型	节点特征选择	键特征选择
SchNet[13]	原子序数	原子距离
CGCNN[11]	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、原子体积	原子距离
MEGNet[12]	原子序数	原子距离
DimeNet++[39]	原子序数	原子距离、键角
ALIGNN[40]	电负性、族数、共价半径、价电子数、第一电离能、电子亲和力、分区、原子体积	原子距离
eqV2 S DeNS[41]	原子序数	原子距离、相对位置矢量

当前元素嵌入方法研究主要分为以下三类：基于文本语料学习的嵌入、基于化学知识的手工特征表示，以及端到端学习的可训练嵌入。第一类方法来源于自然语言处理思想，在这类方法中，材料科学文献被视为语料库，元素符号被视为词元，通过词向量模型学习其分布式表示。例如，Word2Vec 或类似模型通过分析元素在文献中的共现关系来构建嵌入空间[5,42–44]，使化学性质相似的元素在向量空间中彼此接近。随后，一些研究开始将语言模型与材料知识库结合[45,46]，通过预训练的科学语言模型生成更加丰富的元素语义表示，从而在材料性质预测任务中取得良好效果。然而，这类方法的性能往往依赖于语料库规模与覆盖范围，并且学习到的相似性不一定与具体预测任务直接相关。第二类方法是基于化学知识的手工设计元素描述符。这类方法通过组合已知物理和化学性质来构建元素特征，例如原子半径、电负性、第一电离能、价电子数以及周期表位置等[2,11,40,47]。这些特征具有明确的物理意义，并能够为模型提供有用的归纳偏置。在实际应用中，这些元素特征通常通过统计方式（如平均值、加权和或方差）聚合形成化合

物级表示，从而用于传统机器学习模型或神经网络预测材料性质。手工设计描述符具有较强可解释性，能在小数据集上保持稳定性能，但需要人工选择特征，且难以捕捉复杂化学关系。第三类方法是当前模型偏向采用的可学习元素嵌入。这类方法将元素向量作为模型参数的一部分，与下游性质预测任务进行端到端训练。在神经网络模型中，每种元素会被初始化为随机向量，并在训练过程中通过反向传播不断更新[12,14,48]。这种方法自动学习到任务相关的元素相似性结构，例如某些过渡金属可能在嵌入空间中形成聚类，反映其在材料结构中的相似作用。可学习嵌入能够适应复杂的数据分布，但其通常较难解释[49]，且在不同任务或数据集的迁移能力有限[50]。

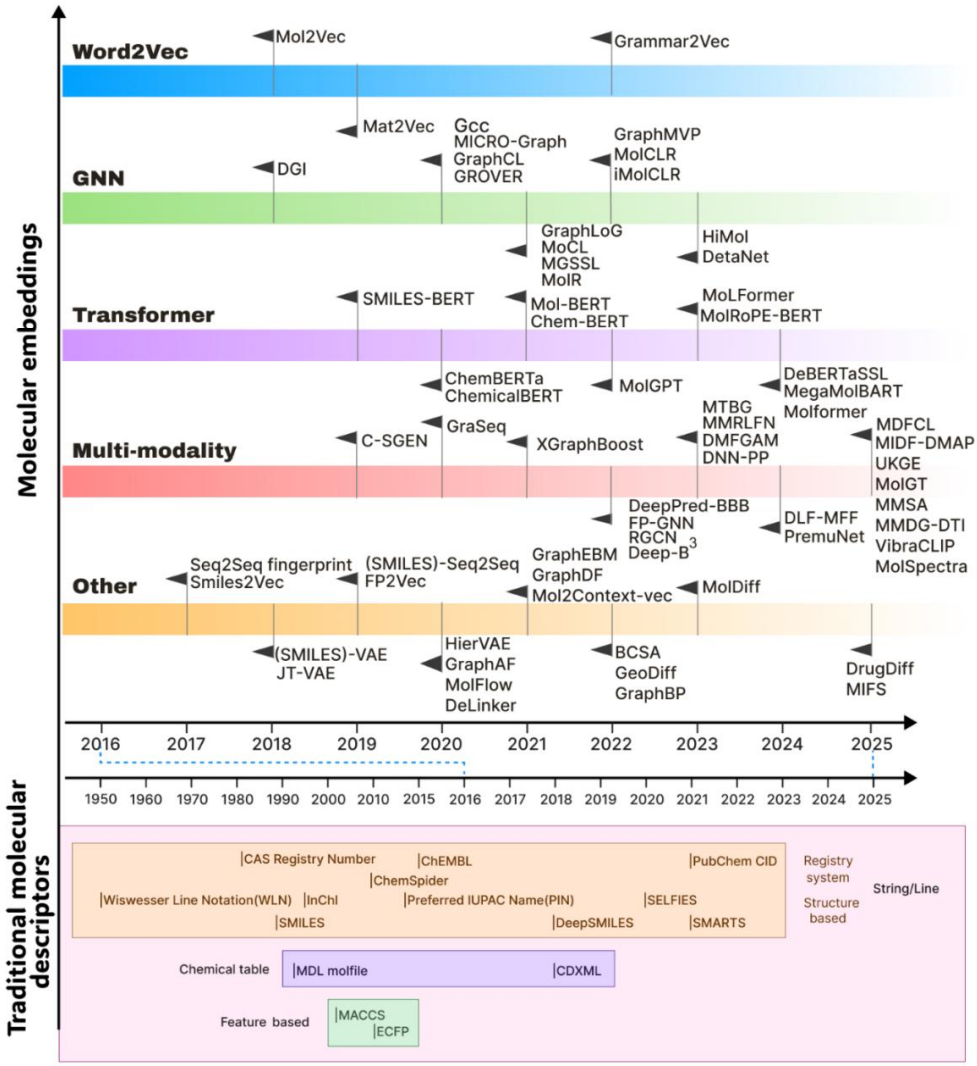


图 1.1 分子材料嵌入研究进展

在材料信息学中，除了元素嵌入外，分子（图 1.1）和晶体（图 1.2）结构的嵌入表示也是结构信息表征的重要研究方向。一些较早或结构较简单的任务，采

用基于特征工程的表示方式。其利用预定义的化学或结构描述符，并通过多层感知机等模型进行性质预测[51,52]。还有一些方法借鉴自然语言处理中的 Word2Vec 模型，从结构片段或材料文本中学习材料嵌入[5,53,54]。两个主流模型架构为基于图神经网络和基于 Transformer 架构。基于图神经网络的表示学习，通过将分子或晶体表示为原子节点与键或邻接关系构成的图结构，并在消息传递过程中学习结构嵌入。包括满足旋转、平移或翻转操作后模型输出不变的不变图神经网络（Invariant Graph Neural Network）[11–13,40,55–59]和操作后输出对应变换的等变图神经网络（Equivariant Graph Neural Networks）[14,60–64]。基于 Transformer 架构的序列建模方法[65–70]，将化学式、结构序列或局部环境编码为序列输入，通过自注意力机制学习材料表示。另外，单一模态往往存在信息局限，实验发现简化分子线性输入规范[71–73]（Simplified molecular input line entry specification，SMILES）、分子指纹[74–76]、分子图[77–79]、分子图像[80–82]、光谱[83,84]以及文本（科研文献）[85–87]等模态往往仅能在特定视角下提供材料信息，融合多种数据来源的多模态嵌入方法可以更加全面地刻画复杂材料体系的多维属性。多模态嵌入通过融合不同来源信息，为模型提供更丰富、稳定的特征表示[81,88–91]。

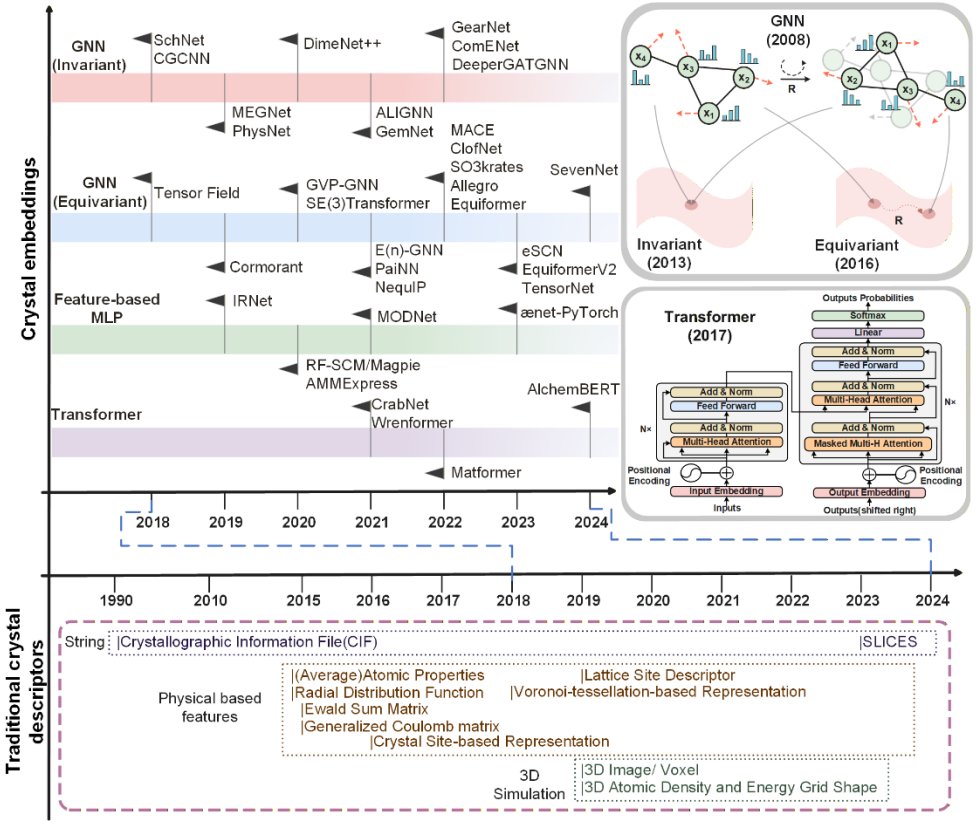


图 1.2 晶体材料嵌入研究进展

1.2.3 嵌入空间协同表示与信息融合研究

在深度学习模型中，将互补特征融合到同一高维表示空间是一种常见且有效的设计策略。典型例子是 Transformer[92]的输入表示，将词元嵌入与位置编码通过逐元素相加，不同语义来源的信息在同一向量空间中共存，而下游网络能够自动学习到如何解耦和利用这些信息。在成分限制注意力网络（Compositionally Restricted Attention-Based network，CrabNet）模型[68]中，仿照 Transformer 位置编码思想，将学习到的元素嵌入和分数嵌入逐元素相加形成输入表示，使同一向量同时编码元素身份和化学计量贡献。近年来，大量研究围绕位置编码与表示融合机制展开，形成了多种方法。最早被广泛使用的是绝对位置编码，在原始 Transformer 中，位置编码采用固定的正弦和余弦函数，将每个序列位置映射到一个确定的向量，并与词元嵌入逐元素相加。其结构简单且不引入额外参数，可以在一定程度上泛化到更长序列。针对位置编码的研究，近年来又出现了多种改进方法，如相对位置编码[93]（Relative Positional Encoding）、旋转位置编码[94]（Rotary Positional Encoding，RoPE）、可学习位置编码[95]、ALiBi[96]（Attention with Linear Biases）等。

1.3 研究内容

围绕材料科学领域属性预测神经网络，研究不同元素嵌入方法对材料结构信息表征能力的影响，本文从元素嵌入的构建、评价与融合三个层面开展系统研究。

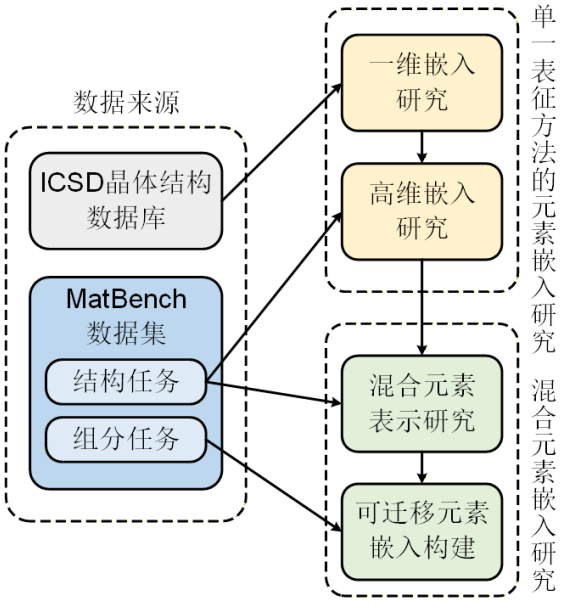


图 1.3 研究内容整体框架

整体研究架构如图 1.3 所示。首先，从实验数据出发构建能够反映元素化学相似性的元素嵌入表示。对不同元素嵌入方法在材料性质预测任务中的表现进行系统比较与结构分析，得到更加符合材料结构合理性的元素分组结果。随后，在此基础上探索不同元素嵌入之间的互补关系，提出混合元素嵌入方法，以融合专家知识与数据驱动信息，并为材料机器学习中的元素表示方法研究提供新的思路和参考。

1.3.1 单一嵌入方法对材料结构信息表征能力的研究

材料机器学习中元素表示方式对模型性能具有重要影响，本文系统研究了单一表征方法的元素嵌入构建方法及潜在空间结构特征。在一维嵌入中，基于化学可替换关系提出一维元素嵌入构建与优化方法。引入平衡替代关系与间距的多目标优化函数，对传统一维元素排序的间距进行优化，在一维表示中增强元素间相似性差异的表达能力。在高维嵌入上，研究材料图神经网络中不同元素嵌入方法。在两种典型材料图神经网络模型上，对基于专家知识的手工描述符和模型学习得到的可训练嵌入两类元素嵌入方法进行分析。从预测性能、嵌入空间结构以及元素聚类行为等角度评估不同嵌入方式的差异，并结合聚类算法与物理信息约束构建数据驱动的元素分组方案。

1.3.2 混合嵌入方法对材料结构信息表征能力的研究

材料机器学习研究中不同嵌入方法通常被视为互斥选择，本文从模型性能和潜在空间结构角度探讨两类信息源在同一潜在空间中的协同表达。基于深度学习中，多源表示能够在高维潜在空间中共存的思想，提出一种混合元素嵌入方法。通过在共享潜在空间中对来自不同信息源的元素表示进行逐元素相加，实现专家知识描述符与可学习嵌入的融合表达。在两种晶体图神经网络模型中构建并行双通道元素输入结构，并通过混合比例参数控制两类表示的融合，在材料结构相关任务上对不同混合策略进行系统实验评估。在绝大多数任务和模型组合中，混合嵌入能够稳定获得优于单一嵌入方法的预测性能。此外，本文从混合模型中提取稳定元素关系，通过聚合多个模型学习得到的元素间余弦距离构建元素嵌入表 Mat2Vec-*。在跨模型与跨数据集验证上，该嵌入均优于传统基线，具有优秀的泛化能力，为材料机器学习研究提供了一个新的可复用元素表示方式。

1.4 章节安排

本研究针对材料科学领域下不同元素表示方式对结构信息表征能力与模型预测性能的影响，围绕元素嵌入的构建方式与融合，从单一嵌入方法和混合嵌入方法对材料结构信息表征能力的研究两大方面进行探索，旨在分析不同嵌入方法对下游任务影响，并探索元素多源信息融合。全文共分为五个章节：

第 1 章 绪论。主要介绍不同元素嵌入方法对材料结构信息表征的研究背景，叙述本研究主要研究内容、创新点以及论文组织结构。

第 2 章 相关理论与技术。主要介绍本研究涉及核心理论方法与技术工具，包括降维方法、聚类算法、多目标优化、相关性分析以及相关神经网络模型等，为后续实验研究提供理论基础。

第 3 章 单一嵌入方法对材料结构信息表征能力的研究。围绕一维元素嵌入构建与优化，以及高维元素嵌入的结构特征与性能表现展开分析，对不同嵌入训练结果采用聚类和物理先验约束，构建符合材料物理规律的元素替代分组。

第 4 章 混合嵌入方法对材料结构信息表征能力的研究。围绕不同嵌入在共享表示空间中的融合机制，首先构建混合元素嵌入框架，分析混合元素嵌入对模型性能的影响。随后在混合模型基础上提炼稳定元素关系，构建可迁移元素嵌入表。

第 5 章 总结与展望。对本文工作进行总结，并对未来研究工作进行展望。

第2章 相关理论与技术

本章主要阐述本研究中所涉及关键技术方法的理论介绍，包括多维尺度变换、K 均值聚类算法、遗传算法、帕累托前沿、皮尔逊和斯皮尔曼相关性系数、CrabNet 嵌入方法及模型、材料图神经网络 CGCNN 和 MEGNet。

2.1 多维尺度变换

多维尺度变换（Multidimensional Scaling，MDS）是一种用于将高维数据映射到低维空间以便可视化和分析的统计方法，算法核心思想是在低维空间中找到一组点尽可能保持样本之间原本的距离关系。给定一个包含n个样本的距离矩阵$D = (d_{ij})$，MDS 的目标是寻找低维空间中的点坐标$X = \{x_1, x_2, \dots, x_n\}$，使得欧氏距离$\|x_i - x_j\|$尽可能接近原始距离$d_{ij}$。常见优化目标是最小化应力函数（Stress）：

$$Stress = \sqrt{\frac{\Sigma_{i<j}\left(d_{ij} - \|x_i - x_j\|\right)^2}{\sum_{i<j} d_{ij}^2}} \tag{2.1}$$

通过迭代优化不断调整点的位置，使应力函数最小，从而得到一个在低维空间中尽量保留原始距离结构的表示。在本研究中，使用 Python 库 sklearn 中的函数实现 MDS 降维。本研究在后文构建一维元素排序（对化学可替换性度量矩阵）和构建 Mat2Vec-*元素嵌入表（对模型元素间余弦距离矩阵）时，使用 MDS 尽可能保持元素之间原本的距离关系。

2.2 K 均值聚类算法

K 均值聚类算法（K-Means clustering algorithm）是一种迭代求解的无监督聚类分析算法，广泛应用于数据挖掘与模式识别领域。其主要作用是将一组没有标签的数据划分为 K 个不同的簇（Cluster），使得同一簇内的数据样本之间尽可能相似，而不同簇之间差异尽可能大。其核心思想是基于距离来衡量样本之间的相似性，通常采用欧氏距离作为衡量标准。给定数据集$X = \{x_1, x_2, \dots, x_n\}$，K-Means 的目标是最小化簇内平方误差（Sum of Squared Errors，SSE），其目标函数为：

$$SSE = min\, \Sigma_{i=1}^{k} \Sigma_{x \in C_i} \|x - \mu_i\|^2 \tag{2.2}$$

其中C_i表示第i个簇，μ_i为该簇的中心（即均值向量）。通过不断优化这一目标函数，算法可以得到较为紧凑且分离度较高的聚类结果。在具体实现上，K-Means 算法通常包括以下步骤：首先随机选择 K 个样本作为初始聚类中心。然后计算每个样本到各个中心的距离，并将其分配到距离最近的簇中。接着根据当前簇内所有样本的均值重新计算每个簇的中心。之后重复分配和更新这两个步骤，直到聚类中心不再发生显著变化或达到最大迭代次数为止。本研究使用 Python 库 sklearn 中的函数 K-Means 对模型学习到的元素嵌入向量进行聚类，分析模型训练后学习到的元素间化学关系。

2.3 多目标优化

2.3.1 遗传算法

遗传算法（Genetic Algorithm，GA）是一种通过模拟自然进化过程搜索最优解的方法。通过模拟“自然选择、优胜劣汰、基因重组与变异”等过程，在解空间中不断进化候选解，从而逼近问题的最优解。与基于梯度的方法不同，GA 不依赖问题的可导性或连续性，通过对种群中多个候选解的并行搜索来提高全局寻优能力。其核心思想是将问题的可行解编码为染色体（由于遗传算法不能直接处理问题空间的参数），通过适应度函数（fitness function）判断群体中个体的优劣程度，适应度函数的设计直接影响遗传算法性能，并在不断迭代中使整体种群适应度提升，从而逐步逼近全局最优解。在实现流程上，遗传算法包括编码、初始化种群、适应度评估、选择、交叉、变异和终止判断等步骤。本研究使用遗传算法搜索能够最大化目标函数的最优候选，从而在多个聚类结果上获取集成后的元素聚类结果。

2.3.2 帕累托前沿

帕累托前沿（Pareto Frontier）是多目标优化问题的一个核心概念，用来描述在多个相互冲突的目标之间能够达到的最优权衡集合。帕累托前沿的作用就是找出一组不可被同时改进的解，对于这些解，如果想让某一个目标变得更好，就必然会使至少另一个目标变差。设决策变量为x，目标函数为$\left(f_1(x), f_2(x), \cdots, f_k(x)\right)$，如果存在两个解$x_a$和$x_b$，使得对所有目标都有$f_i(x_a) \leq f_j(x_b)$，则称解$x_a$帕累托支配解$x_b$。所有不被其他解支配的解构成的集合称为帕累托最优解集，将集合中的解映射到目标空间后形成的边界曲线或曲面，就是帕累托前沿。通过观察这一

前沿，决策者可以直观地理解不同目标之间的权衡关系，从而根据实际需求选择最合适的解。本研究使用帕累托前沿，用于筛选优化传统等间距一维元素排序时目标函数的超参数选择。

2.4 相关性系数

2.4.1 皮尔逊相关性系数

皮尔逊相关系数（Pearson Correlation Coefficient）主要用于衡量两个连续变量之间的线性相关程度，即探究当一个变量发生变化时，另一个变量是否会按固定的比例随之呈直线变化。在计算上，它的核心逻辑是将两个变量的协方差除以它们标准差的乘积，以此来标准化数据并消除量纲的影响，其核心公式为：

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{2.3}$$

其中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。这种方法的特点是其结果的值域严格在-1 到+1 之间，+1 代表完全的正向直线关系，-1 代表完全的负向直线关系，而 0 则表示无线性联系（0 不代表毫无联系，可能存在非线性关系）。皮尔逊相关系数通常假定数据服从正态分布，对极端异常值较敏感，个别离群点可能大幅影响最终的相关性评估结果。在本研究中，皮尔逊相关性系数被用于评估模型性能和元素嵌入向量结构的多个指标间的线性相关程度。

2.4.2 斯皮尔曼相关性系数

斯皮尔曼相关系数（Spearman's Rank Correlation Coefficient）是一种非参数的统计方法，主要用于衡量两个变量之间的单调关系。斯皮尔曼相关系数只关心当一个变量增加时，另一个变量是否总体上也呈现增加（或减少）的趋势，不要求该变化必须是恒定比例的直线。斯皮尔曼相关性系数抛弃了原始数据的具体绝对数值，而是先将所有数据按大小排序转换为名次排名，计算这些名次之间的相关性；在没有重复名次的情况下，通常通过每对数据的秩差d_i利用公式 2.4 计算：

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \tag{2.4}$$

斯皮尔曼相关性系数不要求数据服从特定分布，受异常值的影响较小，能够精准捕捉非线性但严格单调的关系。在本研究中，使用斯皮尔曼相关性系数与皮尔逊

相关性系数共同衡量模型性能与元素嵌入向量结构间的相关性。

2.5 CrabNet 嵌入方法及模型

2.5.1 Mat2Vec

发表在 Nature 的无监督学习嵌入方法 Mat2Vec[5]，将材料的化学信息转化为向量表示，然后通过预训练模型进行各种材料属性的预测。Mat2Vec 是一种在材料科学领域构建文本嵌入的开创性方法，该方法将 Skip-gram[97]变体应用到材料科学文献语料库中，用于预测目标单词（化学元素、材料和相关的物理特性等）附近的上下文单词。上下文的丰富性使得生成的嵌入表示不仅限于简单元素符号或位置关系，而是基于实际材料环境中的多维信息。通过对数百万篇材料科学论文摘要进行大规模训练，将化学元素、化合物、物理性质及科学术语映射到高维连续向量空间（嵌入空间）中。

2.5.2 CrabNet

CrabNet[68]是 Transformer 自注意力机制在材料科学领域的应用，旨在仅提供化学式的情况下，实现与结构无关的材料性质预测。CrabNet 将自注意力机制引入材料性质预测任务，根据元素化学环境动态学习和更新各个元素表征。模型输入将化学成分视为系统，将元素视为系统中的元素。CrabNet 引入了一种特征化方案，该方案在表示和保留各个元素身份的同时，还能在元素之间共享信息。这种表征方式使 CrabNet 能够学习化合物内部元素间的相互作用，并利用这些相互作用生成性质预测。化学成分的输入数据包括组成元素的原子序数和含量分数。CrabNet 默认使用 Mat2Vec 嵌入表示元素，Mat2Vec 嵌入优势在于它已经过预缩放和归一化处理，并且没有缺失元素或元素特征。原子序数用于提取元素表示，根据分数占比获取分数嵌入。通过对元素表示应用全连接网络，生成元素嵌入矩阵，分数嵌入矩阵由分数嵌入生成。然后，将这两个矩阵逐元素相加，形成元素导出矩阵（the Element-Derived Matrix，EDM）。如图 2.1 所示，图中 B 代表批次，d_model 代表元素特征，n_elements 代表元素数量，j 索引表示元素，k 索引为元素嵌入。在计算并叠加元素嵌入与分数嵌入后，将处理完成的元素特征矩阵（EDMs）沿第一维度进行批处理，生成形状为（批次中化合物总数，最大化学成分中元素数量，嵌入维度）的最终输入数据。当化学式中的元素数量少于 EDM 的行数时，多余的数据行用零填充。

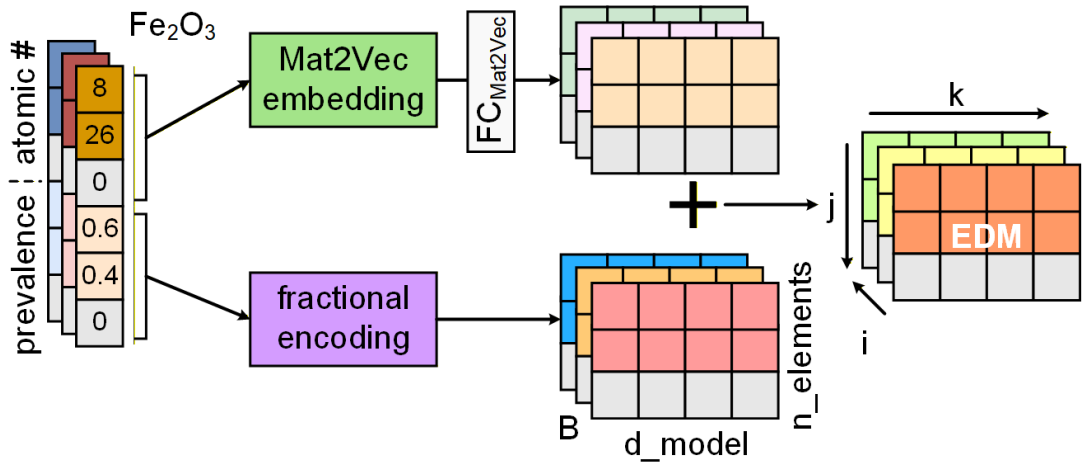


图 2.1 CrabNet 模型 EDM 特征化方案

CrabNet 包含两个主要模块，如图 2.2 所示，第一个模块是 Transformer 编码器，层数默认超参数为 3 层，每层有 4 个注意力头。第二个是模块残差网络，用于将元素向量转换为元素贡献。当 Transformer 编码器完成元素表示更新后，每个 EDM"会通过一个全连接残差网络。该残差网络将 EDMs 的形状转换为大小为（n_elements，n_elements，3）的向量。这三个向量分别为元素原始贡献值、元素不确定性以及元素 logits。通过对元素 logits 进行$Sigmoid(\sigma)$函数运算得到元素缩放因子 s，将元素原始贡献值与其对应的缩放因子 s 相乘得到元素贡献值，元素贡献值的平均值为该化合物属性预测值的最终输出。

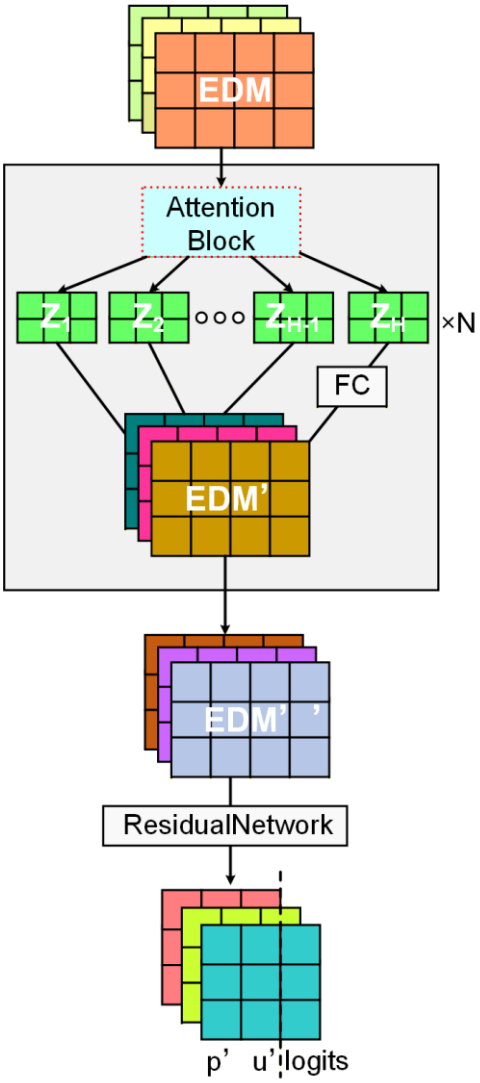


图 2.2 CrabNet 模型架构图

2.6 材料图神经网络

在材料科学领域，图神经网络已成为材料性质预测的重要工具，其流程如图 2.3 所示。图 2.3(a)为图结构的特征标注，(b)为通过消息传递和聚合进行特征更新，(c)为图神经网络中图的迭代更新。材料中原子被定义为节点，化学键被定义为边。本研究将通过两种代表性材料图神经网络 CGCNN[11]（Crystal Graph Convolutional Neural Networks）和 MEGNet[12]（MatErials Graph Network），研究不同元素嵌入策略对模型性能影响。

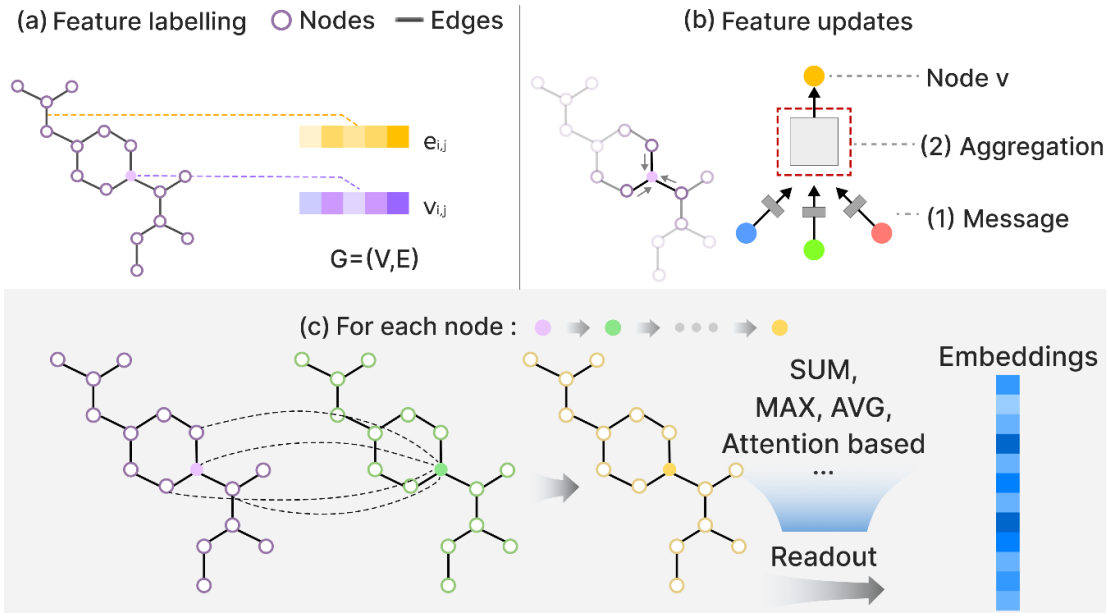


图 2.3 材料图神经网络流程示意图

2.6.1 CGCNN

晶体图神经网络 CGCNN 模型架构如图 2.4 所示。该模型将晶体结构转化为图结构时，节点为原子特征表示，默认采用 9 种原子属性作为原子特征向量（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、分区及原子体积），边为原子间距离。CGCNN 通过卷积操作（默认卷积层为 3 层）进行消息传递：

$$v_i^{(t+1)} = \mathrm{Conv}\left(v_i^{(t)}, v_j^{(t)}, e_{(i,j)_k}\right) \tag{2.5}$$

其中$\mathrm{Conv}(\cdot)$为卷积函数，$v_i^{(t)}$表示节点i在第t层卷积层时的特征向量，$\mathbf{e}_{(i,j)_k}$表示表示连接节点i和节点j的第k条边特征。节点信息更新方式为：

$$z_{(i,j)_k}^{(t)} = v_i^{(t)} \oplus v_j^{(t)} \oplus e_{(i,j)_k} \tag{2.6}$$

$$v_i^{(t+1)} = v_i^{(t)} + \sum_{j,k} \sigma\left(z_{(i,j)_k}^{(t)} W_f^{(t)} + b_f^{(t)}\right) \odot g\left(z_{(i,j)_k}^{(t)} W_s^{(t)} + b_s^{(t)}\right) \tag{2.7}$$

通过将相邻节点信息及边信息拼接（$\oplus$）为$z_{(i,j)_k}^{(t)}$作为卷积层输入。$\sigma(\cdot)$和$g(\cdot)$分别为激活函数 sigmoid 和 softplus，$\odot$表示逐元素乘法，$W_f^{(t)}$、$W_s^{(t)}$、$b^{(t)}$分别表示权重矩阵、自重矩阵和偏置。经过 R 层卷积后，经过 L1 层全连接层后，进入池化层得到晶体全局特征向量。全局特征向量将用于后续晶体属性预测或分类。

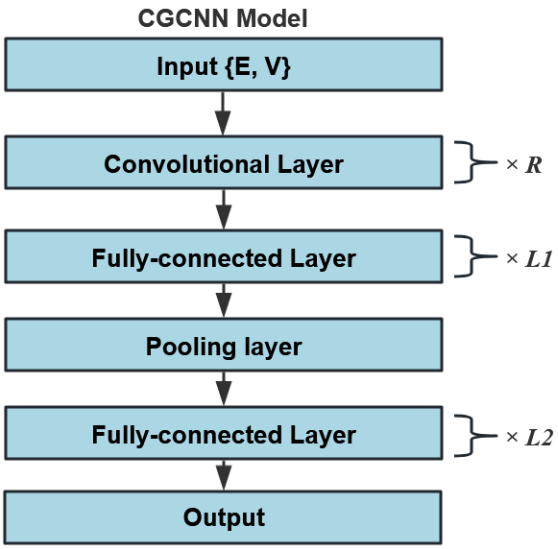


图 2.4 CGCNN 模型架构图

2.6.2 MEGNet

MEGNet 模型架构如图 2.5 所示。该模型可用于分子或晶体结构属性预测，且 MEGNet 用于分子预测时额外增加全局状态（如平均原子质量）表达额外信息，并参与消息传递。MEGNet 块依次对键属性、原子属性（模型初始时为原子序数）和全局状态（本研究中晶体任务默认使用 0 作为全局状态）进行更新。更新过程如公式 2.8-2.13 所示。ϕ_e为键更新函数，$\oplus$为拼接操作，每个键使用其自身键特征e_k、原子索引为r_k和s_k的原子特征和全局状态向量u进行更新。

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

原子特征更新过程中，原子使用其当前特征v_i、连接的键特征$\bar{v}_i^e$和全局状态向量u进行更新，N_i^e为与原子i相连键的数量，$\bar{v}_i^e$为对连接原子i的键特征取均值。ϕ_v为原子特征更新函数，实现方法与ϕ_e一致。

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

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

键更新和原子更新后，使用全局状态向量u、全部原子$\bar{u}^v$和键$\bar{u}^e$特征更新全局状态：

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

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

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

ϕ_u为全局状态更新函数，实现方法同ϕ_e和ϕ_v，N^e和N^v分别为键数量和原子总数。本研究在 MEGNet 模型上，仅调整 MEGNet 块中元素嵌入V_{in}部分。

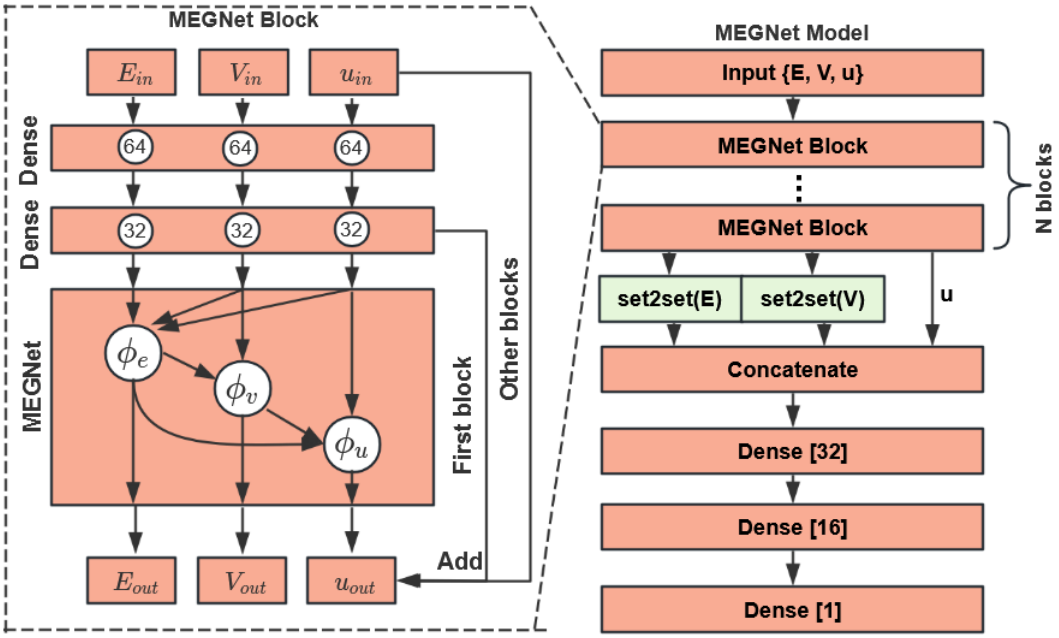


图 2.5 MEGNet 模型架构

2.7 本章小结

本章节主要介绍了本研究在后续实验中使用到的相关理论和技术。其中，多维尺度变换将用于一维排序构建和混合元素嵌入构建。K-Means 聚类算法将用于元素聚类情况分析。遗传算法将用于单一表征方法的元素嵌入研究中，在多个聚类结果上搜索通用元素聚类结果。帕累托前沿将用于筛选优化传统等间距一维元素排序时目标函数的超参数选择。皮尔逊和斯皮尔曼相关性系数用于共同衡量后续不同模型性能与元素嵌入向量结构间的相关性。为了评估不同嵌入策略和研究混合元素嵌入，本研究采用 CGCNN 和 MEGNet 架构作为骨干模型，修改其原始嵌入层。CrabNet 模型和 Mat2Vec 将作为基准，用于比较混合元素嵌入模型获得的元素嵌入表 Mat2Vec-*效果。

第3章 单一嵌入方法对材料结构信息表征能力的研究

元素嵌入作为材料图神经网络中的重要组成部分，嵌入向量所保留的信息量极大程度上决定了模型性能和预测精度，而现有研究大多侧重于模型结构设计与预测精度提升。在本章节中，围绕单一表征方法的元素嵌入从一维嵌入构建和不同高维嵌入方法对材料结构信息表征能力进行研究。

对于一维嵌入构建，本研究从文献中提取了晶体结构实验数据库统计得到的元素间化学可替换性度量，并以此构建元素间的化学可替换性度量矩阵。通过距离转化与降维映射，获得在保持全局距离结构条件下的一维排序。此外，为克服传统等间距排序无法反映相邻元素化学相似性差异问题，本研究引入基于化学可替换性度量的间距优化方法，得到反映元素化学相似程度的一维嵌入表示。

对于高维嵌入研究，本研究基于两种经典材料图神经网络框架。分别采用基于人类知识特征和数据驱动学习的元素嵌入，在多个数据集上进行比较。通过预测误差、嵌入向量结构及元素聚类行为等，评估不同嵌入方式在材料结构属性预测任务中的表现差异，并且过程中发现聚类情况与传统分类有所不同。该研究初步探索了在材料图神经网络中不同元素嵌入方法效果，为后续混合嵌入研究提供对比基础与理论支撑。

3.1 一维元素排序构建与优化研究

3.1.1 研究方案

传统二维周期表结构虽然通过周期与族的排布体现了原子间的化学相似性，但在高通量材料筛选、工业材料替代等许多实际应用场景中，往往需要一种更为简捷的一维排序，使得化学性质相似的元素占据相邻位置。因此，本研究引入晶体结构实验数据库中统计得到的元素间化学可替换性度量[4]，提出了一种基于实验数据驱动的元素一维嵌入构建与优化方案。

本研究对晶体结构实验数据库中元素间化学可替换性度量采用倒数变换，构建初始距离矩阵。考虑到文献中大量元素对缺乏直接可替换性度量，且替换度量受化学环境影响具有方向性。本研究通过中间节点补全非直接关联元素间的拓扑

距离，并对称化处理，构建出元素间距离矩阵。通过最小化高维空间与低维嵌入空间欧氏距离之间的应力函数值，将元素化学可替换性度量映射至一维。针对传统一维等间距元素排序无法区分相邻元素间化学相似性差异的局限，本研究构建包含化学可替换性度量损失与正则化约束的多目标优化函数，对一维排序中的元素间距重新分配。

3.1.2 元素化学可替换性度量与排序构建

本节所使用的元素化学可替换性度量来自 Wang 等人[4]基于实验晶体结构数据库的大规模数据挖掘，元素A被元素B替换的相似度度量S_{AB}如公式 3.1 所示：

$$S_{AB} \coloneqq \frac{1}{N_A} \Sigma_{I,J \neq I} \delta_{AB}^{IJ} \tag{3.1}$$

其中，当材料I和J均存在于实验数据库中，且二者通过化学元素A被B替代而关联时$\delta_{AB}^{IJ} = 1$，否则$\delta_{AB}^{IJ} = 0$。受数据库中材料数量和归一化因子N_A（数据库中包含元素A的材料总数）在不同元素间存在显著差异，所生成的化学可替换性度量矩阵呈现明显的非对称特征。化学可替换性度量数据分布如图 3.1(a)所示，可替换性度量矩阵S中大量空白区域为缺乏直接可替换性度量的元素对。为了将可替换性度量转化为可在流形学习中使用的几何度量，本研究对其进行距离转化处理。首先，采用倒数变换构建初始距离，令$d_{ij} = 1/S_{ij}$，可替换性度量值越高，元素间的化学距离越近。考虑到原始矩阵中存在大量缺失项，利用中间节点（中间元素）寻找任意两个元素间可能的替换路径，如公式 3.2 所示：

$$d_{ij} = min\left(d_{ij}, d_{ik} \times d_{k_j}\right) \tag{3.2}$$

从而尽可能将局部、稀疏的可替换性度量值扩展为更多元素的全局联系。最后，为满足欧式空间的对称性要求，本研究对扩展后的矩阵进行对称化处理，k为中间元素，取d_{ij}和d_{ji}中最小值作为最终元素间的化学距离，构建出如图 3.1(b)所示的对称距离矩阵。

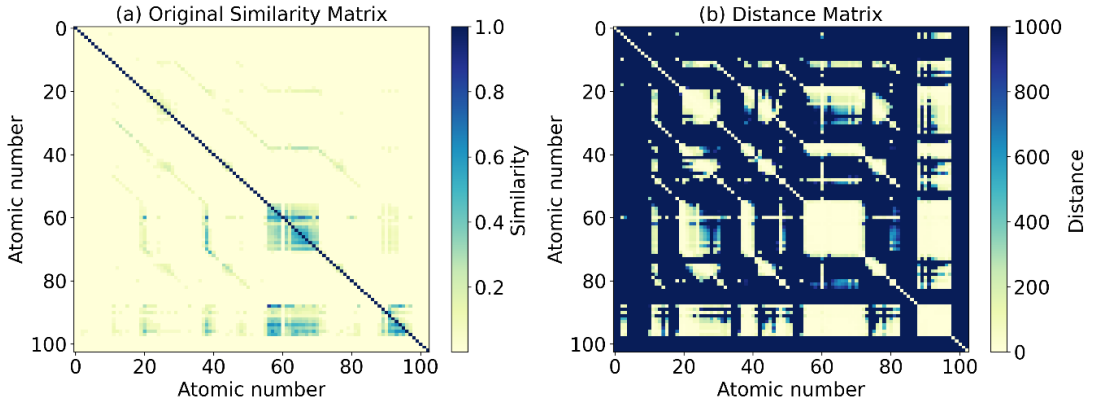


图 3.1 元素间化学可替换性度量矩阵与化学距离矩阵

为确保得到的一维排序能够最大限度保留原始高维空间的距离结构，本研究采用 MDS 算法对元素化学距离矩阵进行降维。为避免算法陷入局部最优解，超参数设置为：迭代初始化次数 20,000 次，最大迭代次数为 500 次，并指定随机种子为 42 以保证结果的可重复性。通过最小化降维后一维欧氏距离与原始拓扑距离的残差，算法自动寻找到一组能够最有效平衡全元素间化学相似关系的排列，最终得到的一维排序结果如图 3.2 所示。

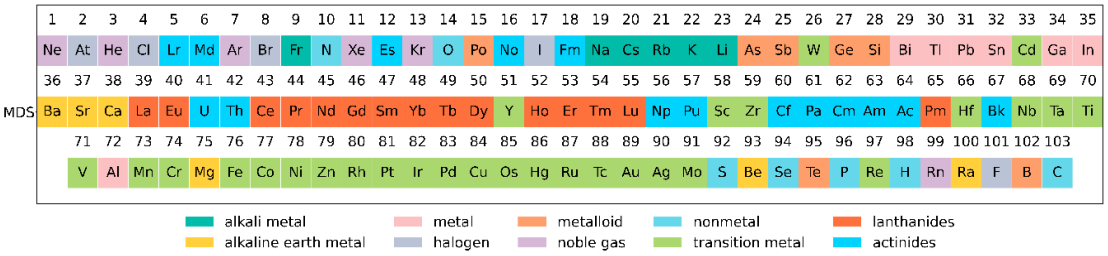


图 3.2 基于 MDS 降维的一维元素排序结果

降维后的一维排序中，碱金属元素（Li、Na、K、Rb、Cs）、过渡金属区域（如 Fe、Co、Ni、Zn、Rh、Pt 等）在序列中紧密相邻，镧系元素（从 La 到 Lu）在序列中形成较长连续片段，且与其具有相似半径和化学性质的钇（Y）、钪（Sc）也分布在附近。然而，本序列也存在一些不同于门捷列夫周期律的非传统排序特征。在序列的两端，惰性气体（He、Ne、Ar、Kr、Xe）与部分放射性元素（如 At、Lr、Md）交替，这种现象可能反映了实验统计数据中，这些元素在晶体结构中的替换数据相对匮乏或无法形成稳定晶体结构。本排序将氢（H）放在序列末端区域，远离碱金属或卤素，符合氢在固态材料中独特的物理化学性质，既可以作为阳离子也可以作为阴离子，且半径极小。这种基于实验数据的一维嵌入，相较于简单的原子序数提供了更丰富的物理化学信息。

3.1.3 基于替代关系的一维排序优化

在神经网络嵌入层初始化时，若能提供贴近化学性质规律的元素间数值距离，使化学性质越相似的元素在嵌入空间中距离越近，将有助于模型更高效地捕捉其规律。针对传统一维等间距元素排序（即任意相邻元素间距均为 1）无法反映相邻元素化学相似性差异的问题，本研究构建了一个多目标优化函数实现非等间距元素分布。该方法可在保持原有一维排序顺序不变的前提下，通过调整元素间的坐标间距$x = [x_1, x_{2,\ldots}, x_{N-1}]$，其中$x_i$表示排序中第$i$个与第$i+1$个元素的距离，使最终嵌入表示能够定量映射出元素间的化学可替换性强度。本研究采用的优化目标函数$L(x, \lambda)$综合考虑了替代关系损失、间距约束以及全局正则化，其数学表达式如下：

$$L(x,\lambda) = \eta\left(\sum_{i=1}^{N-1} x_i - N\right)^2 + \sum_{i=1}^{N-1}\sum_{j=j+1}^{N}\left(S_{ij}\cdot D_{ij} + \frac{\lambda}{D_{ij}}\right) \tag{3.3}$$

其中S_{ij}为 3.1.2 节构建的元素可替换性度量矩阵中的可替换性度量，$D_{ij} = \sum_{k=i}^{j-1} x_k$表示排序中元素$i$与元素$j$之间的累计欧氏距离。$S_{ij}\cdot D_{ij}$旨在压缩高相似度元素对之间的距离，若两元素可替换性度量越大，则较大的距离会产生更高的损失惩罚。λ/D_{ij}为间距平衡项，用于防止所有元素距离坍塌为零，其中λ为控制间距扩张的超参数。$\eta(\sum_{i=1}^{N-1} x_i - N)^2$用于将所有元素间距的总和接近元素总数，$\eta$为正则化系数，本研究取 0.1。

实验过程中，本研究以 Glawe 等人[31]通过遗传算法优化得到的元素排序 GA 作为初始序列，此时初始间距x_0为全 1 向量。采用 scipy.optimize.minize 工具包，对上述目标函数进行最小化求解。为确定最优超参数，对$\lambda \in [1, 2, 4, 6, 8, 12, 16, 32, 64, 100]$的 10 个候选值进行系统筛选。图 3.3 展示了不同超参数下优化所得间距的柱状图分布，元素间距的分布呈现明显波动，而非传统的等间距状态。

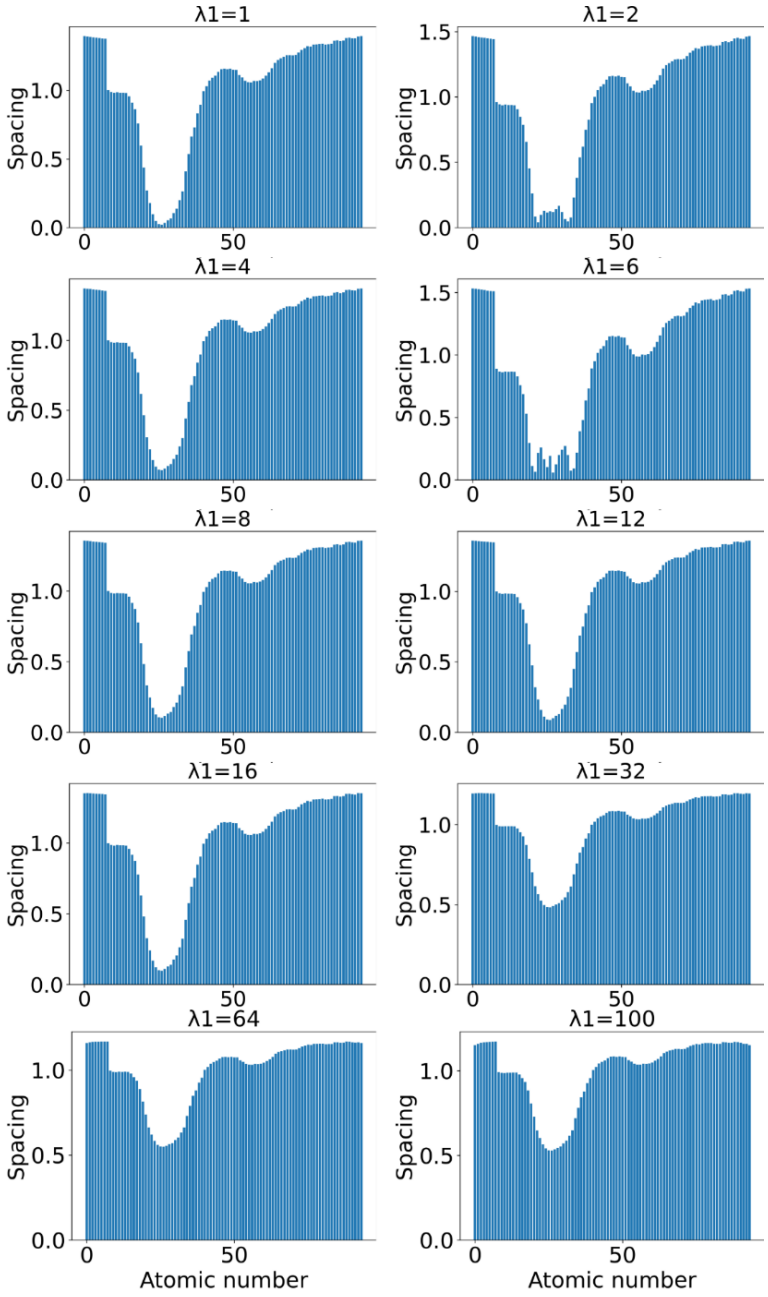


图 3.3 不同超参数λ下元素嵌入间距优化分布图

为科学评估不同λ取值的优劣，本研究采用 Pareto 前沿分析与方差筛选。如图 3.4 左侧所示，通过将不同λ下的替代关系损失（$Loss\ X = \Sigma_{i<j}(d_{ij} \cdot S_{ij})$，$d_{ij}$为元素$i$到元素$j$在一维排序中的距离）与间距平衡损失（$Loss\ Y = \Sigma_{i<j}\frac{1}{d_{ij}}$）映射至双目标坐标系中并构建凸包，处于帕累托前沿的候选点代表了在压缩相似元素距离与保持序列扩张之间达到的权衡最优解。本研究进一步计算了这些候选解在等比例放缩后的间距方差，方差的大小直接反映了优化过程对相似性差异的捕获能力。较大的方差表明，优化后的排序能够让化学可替换性度量差异大的元素尽量远离，化学性质相似的元素尽可能聚拢。选择最大方差结果，可以更明显地体现出元素间关系的非均匀性特征。图 3.4 右侧为优化后排序中元素间距方差，综

合 Pareto 前沿与方差最大化原则，确定超参数$\lambda = 6$。在此参数下，结果既能保持较低的全局损失，又能使间距分布具备表征区分度。

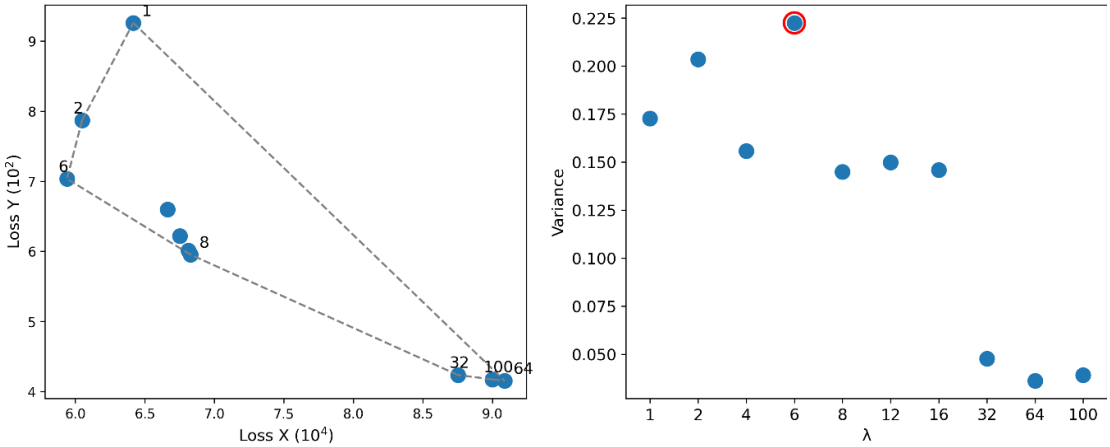


图 3.4 超参数λ的多目标优化筛选

表 3.1 基于化学可替换性度量优化的一维元素嵌入间距结果（$\lambda = 6$）

元素										
He	Ne	Ar	At	Rn	Fr	Kr	Xe	Pm	Cs	Rb
1	2.53	4.06	5.59	7.11	8.63	10.15	11.66	13.17	14.06	14.92
K	Na	Li	Ra	Ba	Sr	Ca	Eu	Yb	Lu	Tm
15.78	16.65	17.52	18.38	19.25	20.08	20.84	21.53	22.05	22.35	22.46
Y	Er	Ho	Dy	Tb	Gd	Sm	Nd	Pr	Ce	La
22.53	22.75	23.01	23.17	23.27	23.47	23.52	23.65	23.85	24.09	24.36
Ac	Am	Pu	Np	U	Th	Pa	Sc	Zr	Hf	Ti
24.56	24.63	24.73	24.94	25.33	25.81	26.45	27.18	28.07	29.02	30.04
Nb	Ta	V	Cr	Mo	W	Re	Tc	Os	Ru	Ir
31.09	32.16	33.27	34.42	35.57	36.72	37.87	39.01	40.15	41.23	42.28
Rh	Pt	Pd	Au	Ag	Cu	Ni	Co	Fe	Mn	Mg
43.28	44.27	45.26	46.26	47.26	48.27	49.3	50.36	51.47	52.63	53.85
Zn	Cd	Hg	Be	Al	Ga	In	Tl	Pb	Sn	Ge
55.1	56.38	57.67	58.98	60.29	61.6	62.92	64.26	65.63	67.03	68.45
Si	B	C	N	P	As	Sb	Bi	Po	Te	Se
69.86	71.3	72.74	74.19	75.63	77.07	78.51	79.96	81.44	82.93	84.4
S	O	I	Br	Cl	F	H				
85.89	87.4	88.92	90.43	91.94	93.47	95				

针对 GA 排序的优化结果如表 3.1 所示，优化过程将原始定性排序转化为具备定量物理意义的一维嵌入表示，可为后续深度学习模型提供包含更多信息的特征输入。最显著的特征体现在镧系元素区域（从 La 到 Lu），镧系元素间的间距从初始的 1.0 被压缩至 0.05-0.3 之间，这些元素在实验数据库中具有极高的可替

换性。碱金属元素（Li、Na、K、Rb、Cs）的间距也被均匀压缩至 0.86-0.89，体现了该主族元素在化学性质上的稳定关系。序列两端的惰性气体（He、Ne、Ar 等）与卤素（F、Cl、Br）区域，其间距普遍扩张至 1.50 左右，最高达到 1.53，这些元素在实验统计中与其他元素的替代关系较弱。

3.1.4 小结

本节旨在解决传统一维元素排序难以描述化学相似性差异的问题，提出并实现了一个元素一维嵌入构建与优化方案。

首先，本研究从晶体结构实验数据库统计得到的元素间化学可替换性出发，构建出能够反映全局化学相似关系的对称距离矩阵。利用 MDS 算法将高维可替换关系映射至一维，获得保持全局结构特征的元素排序。

其次，针对传统等间距元素表示无法区分相邻元素相似度差异，本研究构建包含替代关系损失与间距平衡约束的多目标优化函数。通过对排序中间距的非线性分配，结合 Pareto 前沿分析与方差，找到最优超参数$\lambda = 6$。高相似度元素间的距离最小为 0.05，惰性气体及部分非金属元素间的距离最大至 1.53。

3.2 高维嵌入研究

3.2.1 研究方案

虽然 3.1 节构建的一维嵌入能够提供化学相似性度量，但要对化学元素之间的关系进行最佳描述，需要更高维度视角。当前材料图神经网络中元素嵌入通常采用高维向量以保留更多信息量，参与后续消息传递与下游任务预测。因此，本节针对当前材料图神经网络在元素嵌入选择时常使用的两种策略，结合不同嵌入维度和数据集进行性能评估及嵌入向量结构分析。在此基础上，通过数据驱动与物理约束，提出元素替代分组方案，旨在为材料设计中的元素替代提供参考与初筛依据。

针对不同嵌入方法比较研究，本研究选取经典模型 CGCNN 和 MEGNet，系统对比基于人类知识和模型自主学习两种元素嵌入策略。通过 5 折交叉验证，评估不同嵌入方法和嵌入维度在 MatBench[52]中多个数据集上的预测性能。从信息论与谱分析角度量化不同嵌入方式下元素嵌入向量结构与预测精度的相关性。针对元素聚类情况分析，在不同嵌入方法的模型上分析消息传递过程中嵌入向量聚类情况，以及能否重构传统化学分类。通过整合不同任务、模型下的嵌入结果，通过无监督聚类方法形成统一分类。

3.2.2 不同嵌入方法比较研究

模型对化学元素的表征能力对于理解其预测行为至关重要。在计算化学和材料科学领域，不同 GNN 表示化学元素的两种常见策略如图 3.5 所示，一种将一系列原子性质编码为特征向量作为模型输入，利用已有科学知识指导模型理解化学元素及其相互作用。另一类直接将原子序数输入到嵌入层，采用端到端的数据驱动方法直接从数据中学习元素最佳表示。本节聚焦于材料科学领域两个图神经网络模型 CGCNN 和 MEGNet，二者在众多研究中得到广泛验证且默认元素嵌入分别采用这两类嵌入策略的。

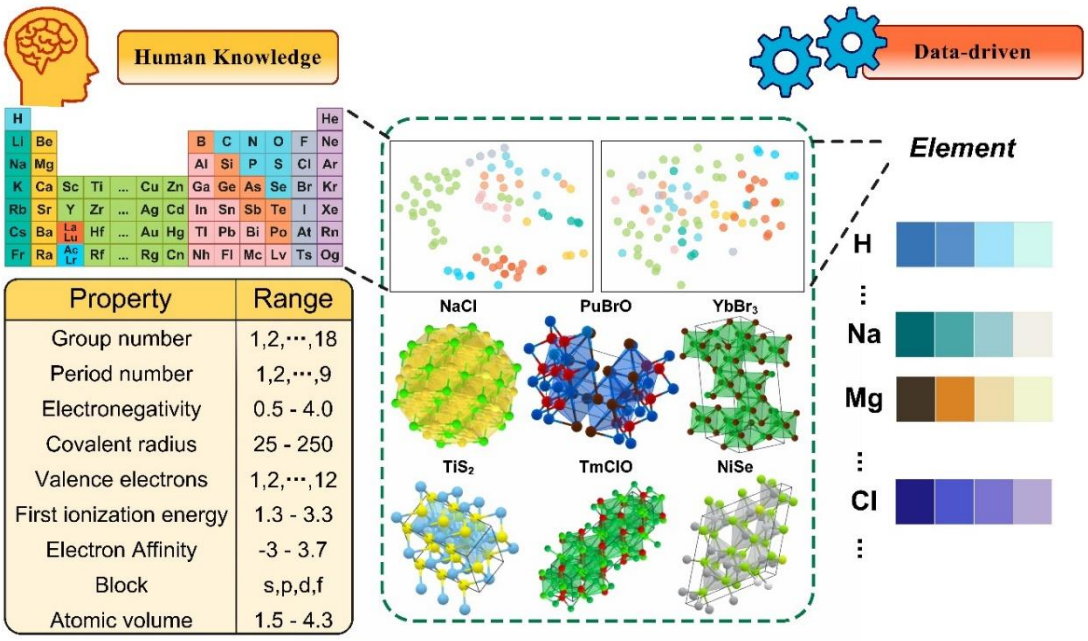


图 3.5 人工设计描述符和数据驱动可学习嵌入策略示意图

本研究使用的数据集来自 MatBench 基准测试中的多个材料结构回归任务，分别为剥离能（meV/原子）、最高频率光学声子模式峰值频率（cm^{-1}）、折射率、平均剪切模量（log10（GPa））、平均体积模量（log10（GPa））、生成热（eV/晶胞）、带隙（eV）、形成能（eV/原子）。这些材料结构属性数据集大小从几百个样本到超过 10 万个样本不等，提供了丰富材料结构和属性，为评估不同嵌入方法奠定基础。

模型架构超参数均参考官方 MatBench 提供参数实现，两模型消息传递层数均设为 3，模型训练批大小设置为 128，基于验证集平均绝对误差，采用 500 个 epoch 的 patience 早停。材料结构中的键特征，两种模型均采用扩展晶体图结构中原子间距离的方式。CGCNN 模型架构采用三个图卷积层和一个全连接层组成。原子嵌入的隐藏特征维度设置为 128，模型使用随机梯度下降法进行训练，学习

率为 0.01，动量为 0.9。应用 MultiStepLR 调度器，里程碑（milestones）位于第 100 个 epoch，衰减因子为 0.1。MEGNet 模型架构采用三个 MEGNet 模块组成，Set2Set 层迭代次数为 3。邻域构建的截止半径设置为 4.0Å。键特征使用具有 25 个中心、标准差为 0.4 的高斯基进行扩展。模型使用 Adam 优化器进行训练，学习率为 0.001。

两种模型各自分别采用两种嵌入策略，嵌入维度设置为{8，16，32，64}，不同嵌入方法的替换仅发生在模型架构的嵌入层。针对手工制作的描述符，选用原子特征包含九种属性，如图 3.5 左侧的属性和范围。离散属性采用类别编码，连续属性被划分为十个区间以捕捉区间变化。手工设计的描述符被线性投影到对应维度的潜在空间中。可学习数据驱动方法使用标准的元素嵌入表实现，在监督训练目标下进行端到端优化为对应维度，直接查找对应元素嵌入向量无需额外变换。CGCNN 在可学习方法的设置中，原始的手工描述符和后续的线性层被移除，替换为实验设置的不同维度的可学习嵌入表。对于最初采用可学习嵌入的 MEGNet，替换为手工嵌入的过程如下：首先使用 CGCNN 手工设计的描述符初始化嵌入表，固定其权重，随后添加线性投影到对应维度。在训练过程中嵌入表被冻结，后续的线性层仍可训练。每个训练任务分两种模型架构、两种嵌入策略、五折交叉验证、四个潜在维度下进行训练，总共训练个 640 模型，所有预处理和归一化步骤限制在当前折内。

实验中使用平均绝对误差（MAE）衡量模型的精度，MAE 取五折交叉验证平均值，评估 CGCNN 和 MEGNet 两种模型在不同材料属性预测任务中的性能。该评估有两个主要目的。一方面，确保学习到的元素嵌入在当前训练条件下具有稳定性能，为后续元素分析奠定基础。另一方面，通过比较不同嵌入方法的泛化能力，探索基于人工设计的描述符和数据驱动的元素嵌入方法之间的差异及其对下游任务的影响。CGCNN 和 MEGNet 两模型实验结果分别如图 3.6 和图 3.7 所示。研究发现，当训练样本量在 10,000 条或更少时，基于手工设计的元素描述符在两种模型中都表现出明显优势。随着训练数据量的增加，当增加到 20,000 条左右时，两种模型在不同嵌入策略下的性能出现差异。在 CGCNN 中，人工设计的描述符仍优于数据驱动方法，表明该模型更依赖于外部先验知识作为有效的归纳偏置来提升性能。MEGNet 借助数据驱动嵌入方法，接近甚至超越了基于手工设计的描述符的性能。此外，两种模型在较高的嵌入维度下，相较于低纬度（如 8 维）通常能取得误差更低的结果，高维潜在空间对元素建模具有积极作用。上述结果表明，基于手工设计的描述符和数据驱动方法学习的元素嵌入在结构属性预测任务的不同情况下具备各自优势，为下一节研究这两种嵌入策略的嵌入向量结构和其所捕捉的元素关系奠定基础。

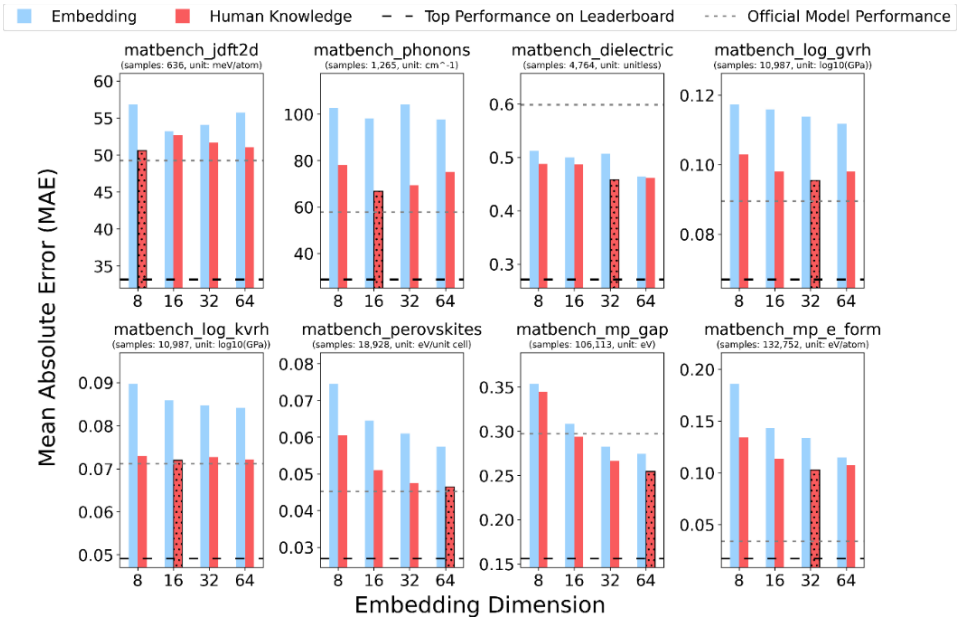


图 3.6 CGCNN 模型两种嵌入策略性能比较

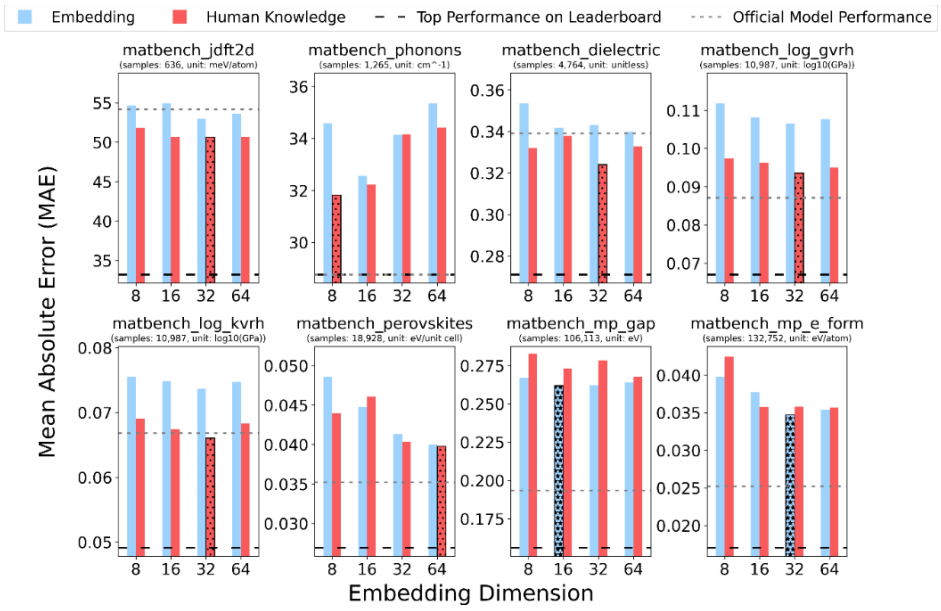


图 3.7 MEGNet 模型两种嵌入策略性能比较

3.2.3 元素嵌入向量分析

在 3.2.2 节中，对两种不同嵌入策略下在不同数据规模下进行性能评估。为探究模型内部学习到的高维向量如何影响材料属性预测精度，仅关注模型预测误差是不够的，须进一步分析嵌入向量的表征结构。在材料科学中，材料性质是晶体结构与化学成分共同作用的结果。对材料图神经网络而言，其学习到的元素嵌入向量既包含元素本身的属性，也包含局部结构环境信息。若某一性质的预测高

度依赖原子的空间排列或拓扑对称性，那么模型学习到的元素嵌入会被结构噪声影响。为减少模型预测的属性由晶体材料结构主导而非原子本身特征带来的预测精度问题，本节首先在 8 种属性预测任务中筛选出 5 种由元素成分主导的任务。如图 3.8 所示，图中以 R^2 反应仅使用元素信息解释目标属性的能力，并将回归结果与平均策略（Dummy）模型和最优模型进行比较，提供相对性能指标。本研究将每个材料样本映射为一个固定长度为 103 维的特征向量，每个维度对应元素周期表中从氢（原子序数 1）到铹（原子序数 103）的一个元素（包含了所有数据集中出现的元素），数值为该元素在化合物中的占比，从而实现对材料化学组成的编码。在不提供任何结构信息的前提下，仅使用元素在化学组成中的成分占比，采用随机森林回归模型对 MatBench 数据集进行预测。在模型训练上，配置 200 棵决策树、不设最大生长深度，并针对每个任务执行 5 折交叉验证。计算过程中，先对目标属性值进行标准化处理以优化拟合效果，在测试集上计算并记录反映模型解释能力的决定系数（R^2）以及还原回原始物理单位后的 MAE。若随机森林模型能够取得较高的R^2或较低的 MAE（图中为随机森林模型达到基准中最优模型性能的程度），表示于成分信息即可近似预测材料属性，元素本身在预测中起核心作用。在这些任务中，CGCNN 和 MEGNet 学习到的元素嵌入向量能够更直接地映射元素的内在规律。基于上述逻辑，本研究选取了 matbench_phonons、matbench_log_gvrh、matbench_log_kvrh、matbench_mp_gap、matbench_mp_e_form 等任务生成的嵌入向量进行后续量化分析。

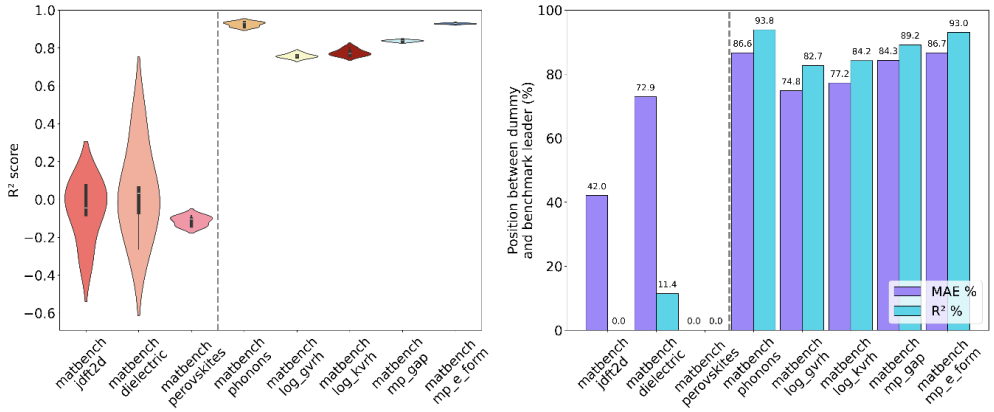


图 3.8 利用元素成分占比评估 MatBench 各项任务的元素相关性

本研究选取 3.2.1 节中这 5 个数据集对应的 400 个模型，并提取模型训练后的元素嵌入向量。实验采用信息熵、谱图理论等指标综合分析，旨在确定嵌入向量是否编码了具有化学意义的结构，并保持了一个良好的潜在空间形态。本实验首先从这 400 个模型中分别提取各元素的嵌入向量，对不同模型分别构建全连接相似性图，其中节点代表化学元素，边反映两侧元素的嵌入向量之间的相似程度。

为使距离值更适用于谱分析，本节采用核函数将欧几里得距离映射为相似度度量。加权邻接矩阵$A \in R^{n\times n}$定义为：

$$\begin{cases} a_{ij} = exp\left(-\|v_i - v_j\|^2\right), i \neq j \\ a_{ii} = 0, i = j \end{cases} \tag{3.4}$$

其中$v_i \in R^d$为元素i的嵌入向量，d为维度，n为元素总数。由于邻接矩阵A为实对称矩阵，其特征值均为实数，设$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda n$为特征值，相应的标准正交特征向量为$u_1, u_{2,\cdots}, u_n$，满足：

$$Au_k = \lambda_k u_k, k = 1,2,\ldots,n \tag{3.5}$$

本研究选取谱半径、谱隙和 Fiedler 值作为谱分析指标，涵盖整体强度、全局结构和连通性。谱半径（λ_1），即最大特征值，衡量了核心元素与其他元素的连接强度。谱隙（$\lambda_1 - \lambda_2$），即最大与次大特征值之差，反应图的整体分离程度或社团结构显著性，谱隙越大的图更易被分割或具有更明显的结构模态。Fiedler 值为图拉普拉斯矩阵$L = D - A$的第二小特征值，反应图的代数连通性。除了基于图的方法，对嵌入层提取的原始矩阵$E \in R^{n\times d}$计算有效秩和嵌入信息熵。通过奇异值分解评估嵌入空间的有效维度，将奇异值σ_i归一化为概率分布$p_i = \sigma_i \,/\, \Sigma\sigma_j$，有效秩定义为该分布的指数熵（$exp(-\sum p_i \, log \, p_i)$），较高的有效秩意味着嵌入向量占据了更大的潜在空间，冗余度更低。嵌入熵分析各维度的方差分布，衡量信息在维度间的分配是否均匀。此外，为了系统分析嵌入空间的几何结构与化学语义之间的一致性，使用 K-Means 聚类算法与元素的真实化学分类进行对比。将每个元素的嵌入向量视为欧氏空间中的样本点。在实践中，研究人员通常使用金属、非金属、卤素、过渡金属和稀有气体等种类的分类方案将元素归类到具有化学意义的类别中。故聚类数设为数据集中实际包含的化学类别数（涵盖碱金属、碱土金属、过渡金属、金属、非金属、准金属、卤素、稀有气体、镧系与锕系等十类典型元素种类），采用调整兰德系数（Adjusted Rand Index，ARI）和轮廓系数。ARI 取值范围为$[-1,1]$，数值接近 1 表示聚类结果和十类典型元素种类分类越吻合，随机聚类 ARI 接近 0，公式为：

$$ARI = \frac{RI - E(RI)}{max(RI) - E(RI)} \tag{3.6}$$

其中 RI（Rand Index）为聚类结果中决策正确的比例：

$$RI = \frac{a + b}{\binom{n}{2}} \tag{3.7}$$

a为同类同簇个数，b为异类异簇个数，$\binom{n}{2}$为n个元素中任选两对的总组合数C_n^2。$E(RI)$为 RI 在随机情况下的数学期望值。轮廓系数：

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.8}$$

其中$a(i)$是样本到同簇其他点的平均距离，$b(i)$是到最近异簇点的平均距离，轮廓系数越接近 1，聚类结果越紧凑且分离度高。

本研究以模型为分析单位，将每一个训练得到的模型视为一个样本点，分析其嵌入结构特征与其归一化误差之间的相关关系，揭示当嵌入空间更具某种结构特征时，模型误差是否呈现系统性变化趋势。针对每种模型框架（CGCNN 和 MEGNet）与嵌入策略组合（手工设计描述符和数据驱动可学习嵌入），在每一个训练任务上分别计算上述各嵌入结构特征与归一化误差之间的皮尔逊相关系数与斯皮尔曼相关系数，并对 5 个任务的相关系数取平均值，得到跨任务的平均相关度量，如图 3.9 所示，相较于 MEGNet，CGCNN 的性能更受嵌入的几何结构影响。

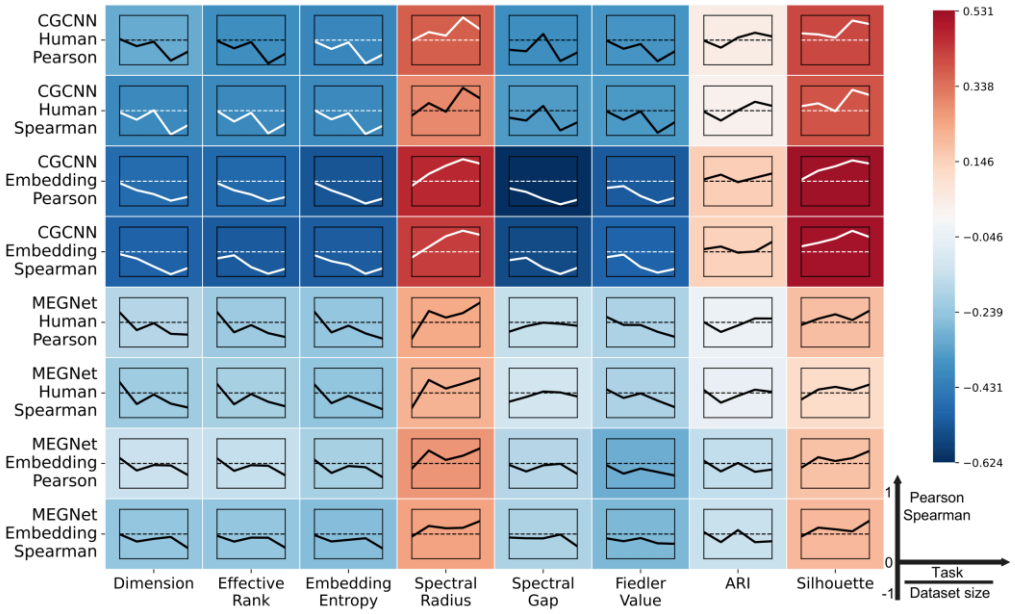


图 3.9 元素嵌入结构属性与模型预测误差在不同任务中的相关性

嵌入向量的维度、信息熵与有效秩三个指标在所有模型设置下均与归一化误差呈一致的负相关趋势，且在 CGCNN 的端到端嵌入策略中相关性最为显著，绝对值在 0.5 以上，在 Pearson 相关系数与 Spearman 相关系数下均保持稳定。这三个指标增大时，嵌入空间不再集中于少数主导方向，这种在多个独立维度上分布更均匀的表示结构与更低的模型误差显著相关。随着数据量增加，模型更容易学习到区分度更高且结构更均衡的嵌入表示，放大负相关趋势。

谱图结构指标中，谱半径与误差呈稳定正相关，谱隙与 Fiedler 值呈负相关，并在 CGCNN 端到端嵌入策略中谱隙相关系数绝对值超过 0.6。谱半径反映邻接矩阵最大特征值规模，谱半径增大表示嵌入集中于少数方向，这种结构往往代表不均衡甚至存在塌缩趋势，与误差升高相对应。相反，谱隙增大通常意味图结构

中主、次特征分离明显，嵌入空间呈现出更清晰的全局结构组织；Fiedler 值增大表示图的代数连通性增强，即整体结构更加稳定且连通紧密。两者与误差的负相关说明，当嵌入空间既具有明确的结构分离趋势，又保持整体连通与稳定时，模型更容易形成具有判别力的表示，从而获得更低误差。结构分离性与整体连通性并存的特征，可能是高性能嵌入的关键几何特征。这种现象在 CGCNN 中显著强于 MEGNet，表明 CGCNN 的预测性能对嵌入空间的谱结构更加敏感，而 MEGNet 整体表现出较弱的结构依赖性，可能与其消息传递机制有关，使其性能不完全由元素嵌入的几何结构决定。

在聚类结构分析中，轮廓系数与预测误差呈现正相关，嵌入形成清晰、分离度高的聚类结构时，模型误差升高，性能更优的嵌入结构并非严格分簇的离散群组。相比之下，ARI 指标整体接近零且波动较小，学习到的嵌入与传统化学类别划分间的一致性与预测误差几乎不存在显著相关性。模型性能的提升并不依赖于重构已有化学分类体系，元素嵌入并未朝着与传统类别严格对齐的方向演化。

3.2.4 元素聚类情况分析

模型训练期间学习到的元素嵌入会产生一个与任务相关的“元素相似性空间”，反映出模型如何在数据和优化约束下将化学关系内化。3.2.3 节中，嵌入空间的几何结构、谱结构指标与模型误差呈现出显著相关关系，但 ARI 指标的结果表明模型学习到的元素表征并未对齐传统化学分类体系。如果高性能模型的嵌入结构并不遵循预定义的化学类别划分，那这些嵌入空间反映了何种不同于传统分组的结构模式？为回答这一问题，本节进一步分析了模型在不同消息传递层中所学习到的元素嵌入表示。

元素嵌入并非静态不变，而是在多层消息传递过程中逐步演化。鉴于初始嵌入相对于传统分组的聚类效果有限，本节进一步追踪了这些嵌入在不同网络层中的变化，判断传统分类是否在学习过程中逐渐重构。图 3.10 定量地描绘了学习到的嵌入与传统化学分组之间一致性的演化，该一致性通过 ARI 指标在不同消息传递层中进行衡量。在某些任务中，元素嵌入和传统分类的 ARI 值略有增加，即使在缺乏明确的先验知识的情况下，模型也能够部分地重构传统的化学分类。但总体 ARI 值仍低于 0.5。事实上，多项文献中结果表明，预定义的元素分组与实际化学行为之间存在不一致之处。即使第三层消息传递层学习到的元素嵌入也无法近似收敛到人类定义的类别系统，强化了关于传统元素分组局限性的假设。

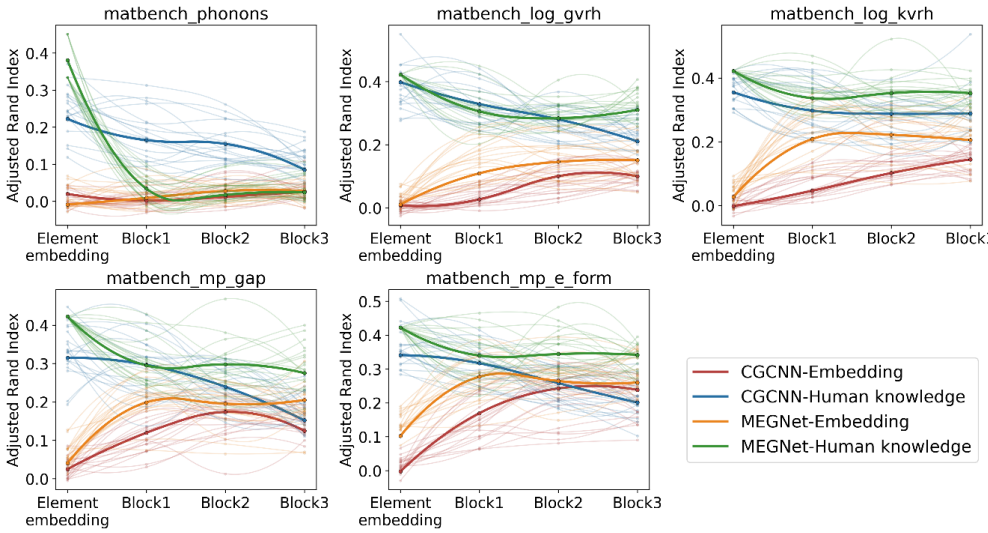


图 3.10 不同消息传递层中元素嵌入与传统元素分类一致性

由于传统的元素分类系统无法反应模型学习到的关系，我们从数据驱动的角度出发，通过无监督聚类方法进一步探索潜在分组。为了从 400 个模型的元素嵌入聚类结果中构建统一且具有代表性的化学元素分类方案，本小节提出了一种基于最大化成对一致性的聚类集成方法。具体而言，对于不同任务、嵌入策略、嵌入维度和交叉验证折数下获得的元素嵌入，执行聚类数为$k \in [2, 10]$的 K-Means 聚类得到多组元素聚类结果。针对每种配置（任务、嵌入策略、维度和交叉验证折数），在上述多组聚类结果中分别选择轮廓系数最高的聚类结果。每个聚类结果都可以看作是一个将一组元素映射到聚类标签的函数。从每个聚类结果中，提取被分配到同一簇的元素对集合 P：

$$P = \{(e_i, e_j) | C(e_i) == C(e_j), i < j\} \tag{3.9}$$

其中e_i和e_j为不同元素，$C(e)$表示分配给元素e的簇标签。同时考虑共聚类对（元素被聚类在一起）和非共聚类对（元素被分开聚类）。P_i表示第i个参考聚类中的共聚类对，P^*表示候选聚类中的共聚类对，$\overline{P_i}$和$\overline{P^*}$分别表示在参考聚类和候选聚类中被分配到不同簇的元素对集合。候选聚类与第i个参考聚类之间的成对一致性得分定义为：

$$Score(i) = \frac{|P_i \cap P^*| + |\overline{P_i} \cap \overline{P^*}|}{\binom{n}{2}} \tag{3.10}$$

其中n表示候选聚类与第i个参考聚类中的元素集的交集数量，$\binom{n}{2}$表示这n个元素中所有可能的配对总数。分子中的第一项统计了在候选聚类和参考聚类结果中都聚为同一组的元素对的数量，分子中的第二项统计了在两个聚类结果中都分离的元素对的数量。该公式反映了正向（同一聚类）和负向（不同聚类）成对分配的一致性。候选聚类的最终得分是通过对所有参考聚类结果，根据公式 3.11 定

义的得分取平均值得到的。得分越高，表明候选聚类越能保留多个来源的聚类和分离模式。本节将该平均成对一致性得分作为目标函数，并应用遗传算法搜索能够最大化该目标函数的最优候选聚类。在遗传算法过程中，聚类数量不设限制，种群规模为 500，迭代次数为 2000。关键参数包括交叉概率 0.5、变异概率 0.2（每个基因的变异率为 0.1）以及规模为 3 的锦标赛选择，这些参数共同平衡探索性和收敛性。基于遗传算法的聚类集成得到的分类结果如图 3.11 所示。该聚类结果将元素分为 17 个类别，部分重构了传统的化学分组，例如类似于元素周期表族群的分组，同时也揭示了传统系统中潜在的元素间关联。

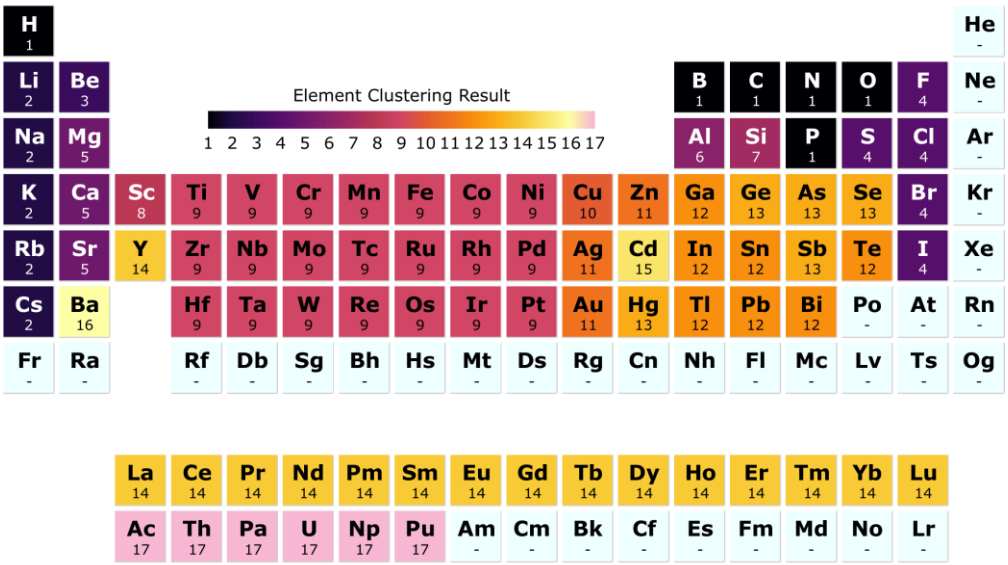


图 3.11 基于成对一致性最大化的元素聚类结果

在诸如高通量筛选或生成式建模等材料发现任务中，元素替代是探索成分空间最有效的策略之一。这些任务的关键在于确定在给定的结构或化学背景下哪些元素可以合理地相互替代，此类取代并非随意进行，用一个原子替换另一个原子需要仔细考虑化学相似性和结构相容性。其中，离子半径起到重要作用，半径相近的元素更有可能进行取代，而不会对主体晶格造成显著畸变。故本研究通过引入离子半径和电荷信息作为附加物理约束进一步优化图 3.11 中数据驱动的元素聚类分组。如图 3.12 所示，分别为标注离子半径值的元素周期表和结合聚类结果和基于离子半径限制的综合分类，背景颜色对应于通过聚类集成方法获得的类别。实验将 17 个元素聚类类别细分为 33 个更精细的类别，确保所提出的取代不仅具有化学意义，而且在晶体环境中也具有结构可行性。基于图 3.12 中元素周期表中离子半径，在每个潜在分组中，按元素的最大离子半径对其进行排序。戈德施密特规则指出原子取代受大小和离子电荷控制，离子半径尺寸差异小于 15%可进行自由替换，尺寸差异在 15%到 30%之间可能发生有限的替代，尺寸差异大

于 30%则几乎不可能进行替代。故在本研究中，当后续元素的半径比前一元素半径大 30%或以上时，引入一个新的子分组。此外，如果一个簇内同时包含阳离子和阴离子时，则将其进一步细分为不同的子组。结果可发现许多元素倾向于只能被同一族元素替换，另外出现了两个主要的金属族：一个为镧系元素，另一个为铁、钴和镍等元素，并且元素周期表内部存在显著的不连续性。通过潜在分组结合基于离子半径的过滤，本节方法将统计相似性与结构兼容性联系起来，为材料发现中的元素替代和新型化合物的设计等应用提供了一个潜在的参考框架。

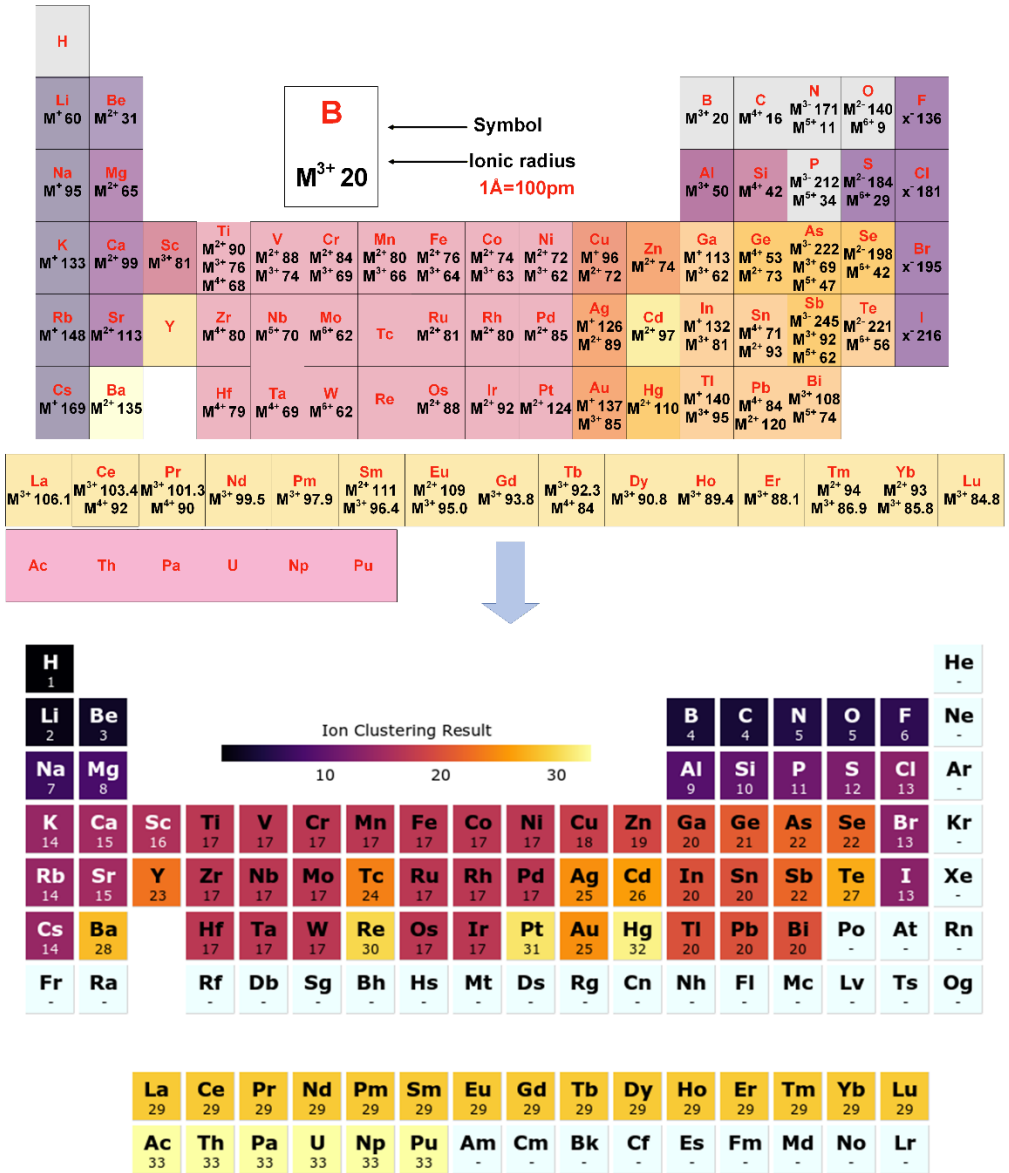


图 3.12 基于聚类和离子半径的元素分类

3.2.5 小结

本节围绕材料图神经网络中的元素高维嵌入展开系统研究，针对模型性能与嵌入结构表征展开分析，并使用数据驱动方式结合物理约束为解释元素分类提供了一个新的视角。

首先，在嵌入策略比较方面，基于 MatBench 中的多个结构属性预测任务，选取 CGCNN 与 MEGNet 两种经典模型，系统对比人工设计描述符与端到端可学习嵌入两类策略在不同嵌入维度与数据规模下的泛化性能。在小样本条件下，基于人工知识的元素描述符能够提供稳定有效的归纳偏置。随着数据规模增加，数据驱动嵌入在 MEGNet 中展现出优势，并且较高嵌入维度通常带来更低误差。

其次，在嵌入向量结构分析方面，从信息论与谱图理论角度量化嵌入空间形态特征，并分析其与模型预测误差之间的统计相关性。结构分离性与整体连通性并存的嵌入几何形态是高性能模型的特征之一。传统化学分类与嵌入结构的一致性与预测误差几乎无显著相关，模型并未复现人类定义的元素分组，而是学习到其他潜在关联。

在元素聚类与分组构建方面，本研究整合 400 个模型的嵌入结果，通过 K-Means 与遗传算法集成优化，提出最大化成对一致性的统一聚类方案，将元素划分为 17 个数据驱动类别。进一步，结合离子半径与电荷信息作为物理约束，将 17 类元素种类划分扩展为 33 个更具结构可行性的子类别。

3.3 本章小结

本章围绕元素表示与嵌入结构在材料建模中的构建与评价问题，通过多个模型的实验与统计分析加以验证。

在元素一维嵌入构建方面，本研究首先从晶体结构实验数据库中统计元素间的化学可替换性关系，构建对称距离矩阵以刻画全局化学相似性。随后利用 MDS 算法将高维替换关系映射至一维空间，获得能够保持整体结构特征的元素排序。针对传统等间距表示无法区分相邻元素相似度差异的问题，本研究进一步构建包含替代关系损失与间距平衡约束的多目标优化函数。优化结果表明，高相似元素间距离可压缩至 0.05，而惰性气体及部分非金属间距离可扩展至 1.53，实现了更具区分度的化学空间表达。

在基于材料图神经网络的高维元素嵌入分析方面，本研究围绕模型性能与嵌入结构之间的关系展开系统研究。基于 MatBench 中的多个结构–属性预测任务，选取 CGCNN 与 MEGNet 两类经典材料图神经网络模型，比较人工设计元素描述符与端到端可学习嵌入在不同数据规模与嵌入维度条件下的泛化表现。并从信息论与谱图理论角度量化嵌入空间的结构特征，分析其与模型误差之间的统计相

关性。结果显示，兼具结构分离性与整体连通性的嵌入几何形态是高性能模型的重要特征之一，而嵌入空间与传统化学分类的一致性与预测误差之间几乎无显著相关性。为进一步研究嵌入学习到的元素间潜在联系，在元素聚类构建方面，本研究整合 400 个模型嵌入结果，采用 K-Means 与遗传算法进行集成优化。并结合离子半径与电荷信息作为物理约束，将元素划分为 33 个具备结构可行性的类别。手工设计的元素描述符与端到端学习的元素嵌入在不同情况下各自展现出优势，模型训练结果为后文开展混合嵌入方法对材料结构信息表征能力的研究奠定基础。

第4章 混合嵌入方法对材料结构信息表征能力的研究

本章节研究一种将专家描述符与可学习嵌入相结合的混合元素表示方法，基于 3.2.2 节不同嵌入方法比较研究的基础，旨在探讨材料属性预测中不同嵌入实现方式是互斥的设计选择还是互补的信息来源。在表示融合和性能评估方面，基于前文研究结果，在 CGCNN 和 MEGNet 两种典型晶体图神经网络中额外加入融合机制。对不同程度专家描述符和可学习嵌入的混合比例，以及潜在空间维度进行系统性的性能评估。在互补机制探索方面，利用余弦距离度量量化两类信息源在共享潜在空间中的几何对齐情况。在此基础上，基于不同混合比例的融合模型生成一组可迁移的元素嵌入表示，并将其应用于下游模型评估其在域外组分数据集上的效果。

4.1 混合元素表示研究

4.1.1 研究方案

元素表示是材料属性预测中常用的计算原语，然而大多数模型将不同的元素嵌入方法视为互斥的设计选择。在实践中，将离散元素映射到连续向量的嵌入层时主要有三种实现方式：（1）通过无监督或自监督学习，在材料语料库上学习得到的基于文本的表示方法；（2）编码了先前的化学和物理知识的手工设计的固定长度元素描述符；（3）通过监督属性预测器进行端到端优化的可学习的元素嵌入。这三种实现方式它们通常被视为互斥的设计选择，而非作为互补的信息来源。但在高维表示空间中，异构信号往往能以极低的干扰程度共存。信息叠加的可行性在深度学习架构中已有先例，CrabNet 模型通过将元素嵌入与化学计量嵌入逐元素相加，同一向量可以同时编码原子身份与组分贡献。这种设计类似于标准的 Transformer 架构的输入构造，将词元嵌入和位置编码相加。尽管词元嵌入表示语义、位置编码表示序列顺序，二者在直觉上属于不同“量纲”，但在高维向量空间中，向量相加并不必然造成语义干扰。简单的加法运算足以让下游层提取并利用各自的信息，特别是在模型容量和训练目标鼓励对维度进行因子化利用的情况下，多个信息通道可以共存于一个共享的嵌入空间内。那么手工设计的描述符与

可学习的元素嵌入是否占据可分离的子空间，从而使得两者的融合能带来系统性的性能提升？

为回答该问题，本研究使用 MatBench 作为评估基准，在结构任务子集上分析可控混合比例的融合效果。评估采用五折交叉验证，排除了其中唯一的分类任务 matbench_mp_is_metal，该任务与 matbench_mp_gap 均源自相同的 Materials Project 原始数据，仅数据标签转变。在 CGCNN 和 MEGNet 两种模型上，本研究修改了元素嵌入层，比较三种元素表示策略，分别为 3.2.2 节不同嵌入方法比较研究中的仅手工特征、仅可学习嵌入，以及在共享潜在空间中融合元素嵌入。通过改变嵌入维度和混合比例，本节旨在探究不同嵌入方法融合能否带来稳定的预测性能提升或在何种条件下能提升预测性能。

4.1.2 不同元素嵌入混合方法

本研究构建了一个混合专家知识与数据驱动的元素嵌入方法。实验采用 CGCNN 和 MEGNet 作为骨干架构，并为两者设计统一的元素输入框架。该框架包含两个并行通道：手工设计通道采用原始 CGCNN 的原子属性描述符作为元素初始表示，通过线性层投影至特定维度的潜在空间（维度$DIM \in \{8,16,32,64\}$）；可学习通道通过标准嵌入查找表实现，进行端到端优化。对原本使用专家设计的原子属性向量的 CGCNN，在保留原始特征的同时，引入可训练的元素嵌入表。同理，对原本使用可学习嵌入表的 MEGNet，利用 CGCNN 未训练的手工描述符初始化一个嵌入表（权重固定），并在该路径后添加一个维度为DIM的可训练线性投影层。两个通道的输出在共享的潜在空间中进行逐元素相加，形成最终元素表示，如图 4.1 所示。开销增长仅发生在元素嵌入层，参数量改变与元素嵌入维度线性相关。相比于整体模型，新增参数和计算量占比较小。在实际训练中，计算瓶颈仍主要来自消息传递。模型超参数选择及训练方式采用和 3.2.2 节一致的设置，仅额外添加一条通道（和原本嵌入方法对应的另一种嵌入方法）。为系统地评估融合效果，研究引入了混合比参数γ：

$$\gamma = log_{10}\frac{Learnable}{Handcrafted} \tag{4.1}$$

其中$Learnable$和$Handcrafted$分别表示可学习通道和手工设计通道权重。通过设定七个离散的γ值，包括纯手工特征H、纯学习嵌入L以及五个中间插值点$\{-4,-2,0,2,4\}$，$\gamma = 0$表示两个通道贡献相等，负值表示手工设计通道贡献更高，正值则表示可学习嵌入通道贡献高。H和L分别对应 3.2.2 节训练所使用的手工设计描述符和可学习嵌入查找表，实验覆盖了从完全依赖先验知识到完全依赖模型自学习的混合嵌入方式。整个实验在 8 个 MatBench 回归任务上展开，结合 5 折

交叉验证、2 种骨干网络、4 种嵌入维度以及 7 种混合比例，总计训练了 2240 个模型，从而能够实现对嵌入融合何时以及如何提高预测准确性进行受控评估。

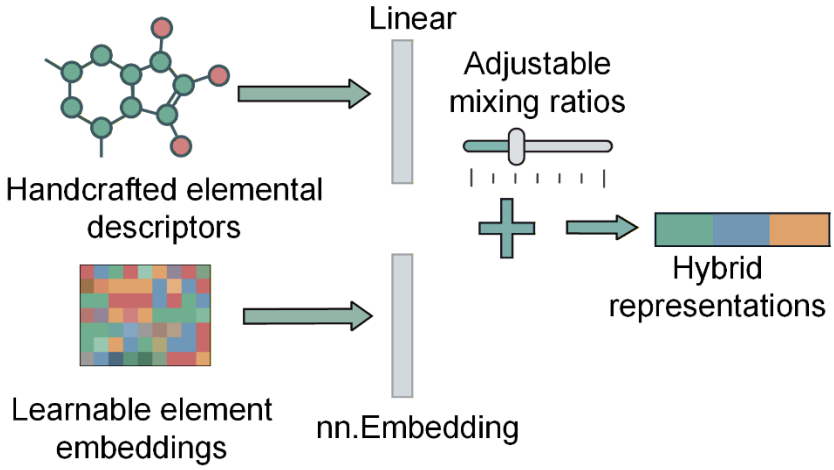


图 4.1 混合嵌入构造方法

4.1.3 实验结果及分析

通过对不同γ值下的 2240 个模型进行性能评估，如图 4.2 所示，图 4.2(a)为 CGCNN 和 MEGNet 在多个 MatBench 任务上的性能，图中折线为同一γ值下，不同嵌入维度、5 折交叉验证下的平均 MAE。阴影区域为不同元素嵌入维度所获得的性能范围，标记点表示最低 MAE 的混合比例。图 4.2(b)为每个骨干网络在每个混合比例下获得最低 MAE 的任务比例。在所有任务和模型系列中，性能最佳的配置很少出现在单一来源端点。相反，对于大多数任务-模型组合，中间混合模型系统性地比纯专家特征或纯可学习嵌入产生更低的 MAE（图 4.2(a)）。值得注意的是，最优混合方案强烈偏向于专家特征一端。对于 CGCNN，每个任务的最低 MAE 均在γ值为$-\infty$/-4/-2 时获得，其中-2 是最常被选择的最优值。而转向以嵌入为主导的比例（γ值为 0、2、4、$+\infty$）通常会降低性能。这种模式表明，即使在已经围绕手工描述符设计的架构中，注入可学习嵌入通道也可能是有益的，但前提是融合仍然以人工设计的专家特征为主导。从相反的范式出发，同样的定性结论也成立。对于标准形式依赖于嵌入查找的 MEGNet，添加专家描述符会使最优解偏离仅基于可学习嵌入的端点：最佳 MAE 集中在混合设置，最常见的γ值为-2 和 0 上，并且对于特定目标（例如 matbench_phonons 和 matbench_dielectric），混合比例 H 更受青睐，而纯可学习嵌入端点 L（$-\infty$）在这些任务的测试比例中从未被选为最佳。汇总所有任务的结果，混合比例在两个模型系列的“优胜者”统计数据中均占据主导地位（图 4.2(b)），模型具体最优性能混合比参数出现频率如图 4.3 所示。实验结果支持了以下观点：两种不同表征可以编码部分互补的信

息，并且可以直接通过简单的逐元素加法融合而不需要专门的融合模块来利用多源信息。

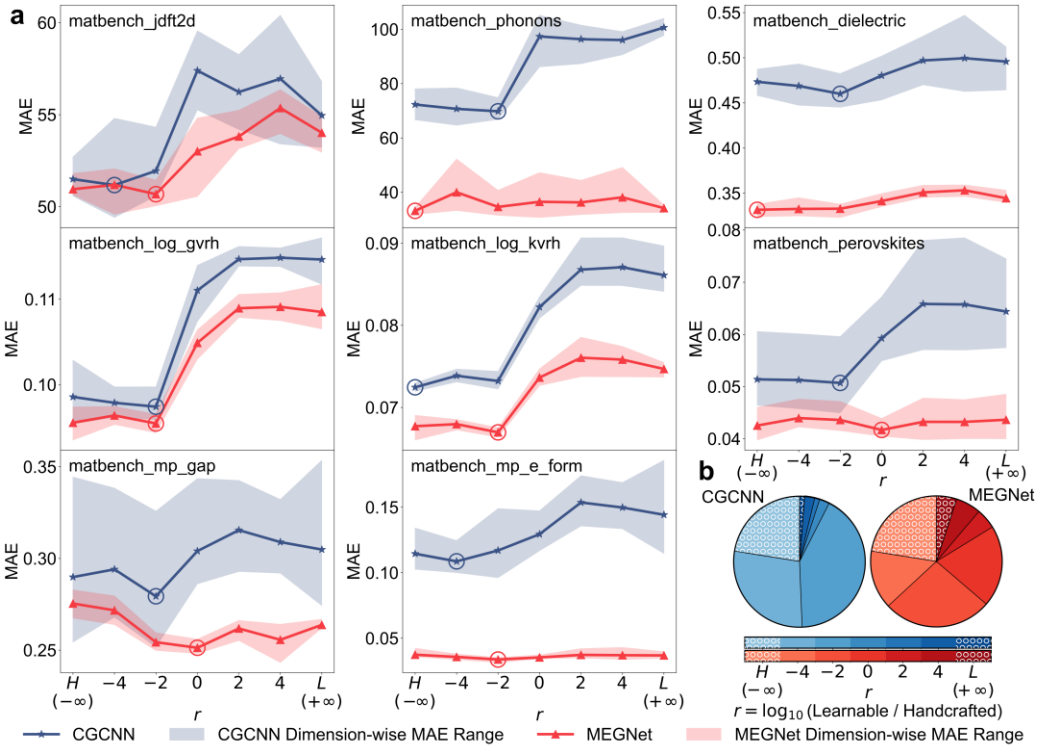


图 4.2 元素嵌入融合对 MatBench 基准测试模型性能的影响

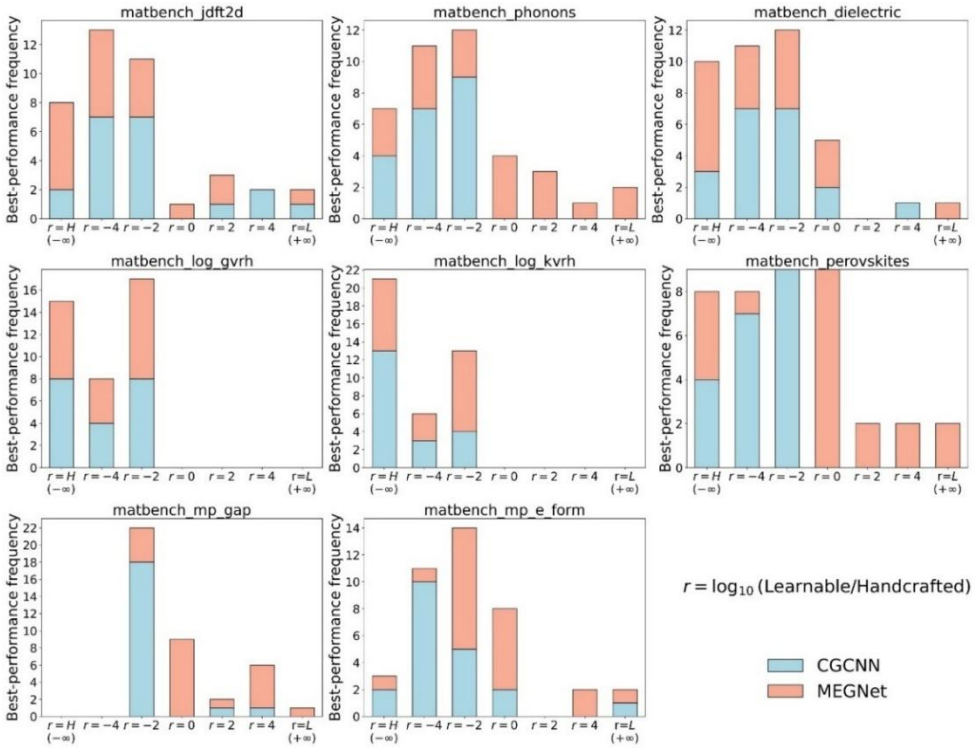


图 4.3 模型最优性能混合比参数出现频率

上述基准测试结果共同证明了混合元素表示法在各种架构中始终具有竞争力，促使本研究进一步分析这种竞争优势效果是否取决于潜在维度以及混合元素嵌入向量伴随哪些几何特征。由于该元素混合表示是通过对两种不同源潜在向量逐元素相加形成的，因此共享潜在空间的维度构成了一个明显的容量瓶颈，空间维度大小决定不同源信息能否互不干扰：它既可能促进异构信号的近乎因子化的共存，也可能在容量不足时导致两种信息互相干扰表达能力下降。为了系统地探究这种效应，本研究对 CGCNN 和 MEGNet 分别在四种潜在维度$DIM = \{8,16,32,64\}$下重复了完整的混合比例扫描，并考察了预测精度和融合行为如何随潜在维度变化。如图 4.4 所示，四个维度的 MAE 均经过归一化处理，其中 1 表示最佳维度，0 表示最差维度，4.4(a)为特征主导模式（H，-4，-2，0），4.4(b)为嵌入主导模式（2，4，L），(c)为任务图例和径向刻度。在所有 MatBench 任务中，维度的增大通常会提高绝对预测准确率，但这种影响的程度很大程度上取决于模型和混合比例。在人工设计特征主导的设置（H、-4、-2）中，性能相对于DIM而言相对稳定，而可学习嵌入主导的设置（2、4、L）则表现出更大的离散性，并且通常在低维度时出现明显的准确率下降。归一化雷达图（图 4.4）直观反映了每个混合比例和骨干网络在不同维度下的任务性能。结果显示出两个一致的模式。首先，极低的容量（$DIM = 8$）很少能在所有任务和混合比例下都具有竞争力，这表明过于狭窄的瓶颈不足以同时编码领域知识描述符和监督关系结构。其次，增大DIM带来的收益递减：从 8 增加到 16，再从 16 增加到 32 通常会带来明显的改进，而进一步扩展到 64 则只会带来较小且取决于任务的变化，有时相对于$DIM = 32$有所提升，有时则几乎没有提升。这种缺乏普遍最优维度的现象表明，有效容量是由目标属性、骨干网络架构和所选混合比例共同决定的。维度本身并不能消除表征选择的影响：有利的混合比例和不利的端点之间的垂直分离通常大于DIM引起的比例内部变化，这表明融合配置是首要驱动因素，而嵌入维度大小则起次要调节作用。上述结果表明，作为构建可迁移元素表示的维度实用默认值为 32，在本研究的其余部分，$DIM = 32$是一个在准确性和效率之间取得良好平衡的选项。同时本研究保留$DIM = 64$作为下游评估和嵌入提炼的更高表达空间。

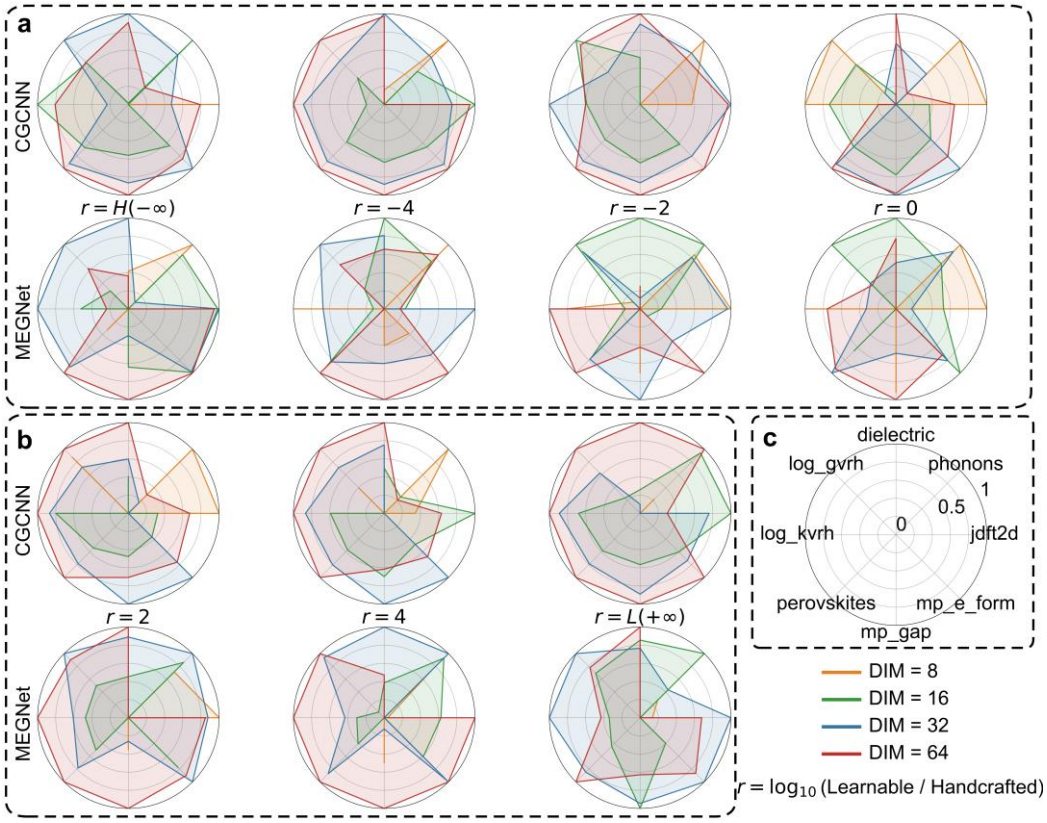


图 4.4 不同任务和混合模式下潜在维度影响

4.1.4 小结

元素表示一直是材料预测中的一个瓶颈，但人们通常将其视为专家设计描述符和可学习嵌入之间的二元选择。本节围绕不同元素表示是否能够在共享潜在空间中实现有效融合这一问题，构建了受控的混合元素嵌入实验框架，并在 MatBench 结构子集数据集上进行系统性评估。本章的研究结果反驳了这种二分法，在各种材料结构性质基准测试和两种广泛使用的晶体图神经网络上，简单的逐元素相加混合模型始终能够达到或超越单一来源模型的性能。这使得专家知识特征向量结合可学习嵌入向量，可以成为一种实用的默认方法，而非相互竞争的替代方案。

通过在 CGCNN 与 MEGNet 两类代表性图神经网络架构中引入并行双通道元素输入（手工设计描述符和可学习嵌入），并以混合比参数对两种信息源进行控制，完成 2240 个模型训练。同时，在实践层面，$DIM = 32$被证明是在性能与计算效率之间取得良好平衡的选择。嵌入混合模型的有效性，为后文提取旋转不变的元素间关系并进行可迁移元素嵌入构建奠定基础。

4.2 可迁移元素嵌入构建

4.2.1 研究方案

在 4.1 节中，本研究从混合嵌入模型性能角度证明，专家设计描述符与监督训练得到的可学习嵌入在共享潜在空间中具有互补性。但单纯的预测性能提升并不足以说明混合表示的科学价值。本研究进一步探讨监督式结构属性任务的学习能够诱导出稳定、可重复的元素间关系几何结构？Mat2Vec 已经证明，无监督文本训练能够恢复一定的周期结构，但其目标函数与材料属性预测并不一致，所编码的元素相似性更多反映文献使用模式，而非直接服务于性质预测。本节探索监督式属性预测是否能够诱导出更贴近材料性质任务的元素关系结构，并将其从模型中提炼出来，构建可迁移的元素嵌入表。

围绕这一目标，本节从三个层面展开研究。在几何层面，通过对混合模型中两条通道的潜在向量进行正交分析，量化专家描述符通道与可学习嵌入通道之间的对齐。通过计算元素级别的余弦角并在任务、折叠与模型维度上进行统计聚合，评估两类表示是否在共享空间中发生方向坍缩，还是保持接近正交的分布特征。

在元素表示构建上，从具体坐标系中提取模型学习到的元素几何关系，转化为旋转不变的元素间关系统计量。由于嵌入空间的坐标系在不同训练过程中可能发生整体旋转或变换，元素坐标无法直接对齐。在高维空间中不同样本对之间的欧氏距离往往集中在一个狭窄的范围内，故本研究采用成对余弦距离作为衡量，对多个性能最佳模型中的元素间距离进行聚合，构建稳健的元素距离矩阵，并在此基础上构建新的元素嵌入表。

在迁移验证层面，测试与构建阶段不同的架构范式和输入设置下评估新元素嵌入表的域外效用。通过在 CrabNet 模型中替换原有嵌入 Mat2Vec，检验其是否具有跨架构、跨数据集的泛化能力。为后续在更广泛模型与任务分布上的推广研究提供理论与技术框架。

4.2.2 元素嵌入正交分析

4.1.3 节中混合方法的优势表明，专家设计的描述符和任务训练的元素嵌入贡献了非冗余信息。为了直接验证这一点，本节量化了共享潜在空间中两个通道之间的几何对齐程度。对于每个元素i，将专家特征投影的潜在向量记为h_i，将可学习嵌入查找表得到的潜在向量记为l_i，并计算它们的余弦角：

$$\theta_i = arccos\left(\frac{h_i \cdot l_i}{\|h_i\|\|l_i\|}\right) \tag{4.2}$$

当两个通道近似正交时，余弦角接近 90°。对各个元素的θ_i进行聚合，以获得每个模型的对齐统计数据。平均夹角和方差如表 4.1 所示，其中“未训练”表示随机初始化的嵌入基线，在同一元素手工通道投影后和可学习嵌入通道之间计算夹角。按任务、混合比参数分类的方差图如图 4.5 所示，图中单元格内为给定设置下逐元素角度分布方差。

表 4.1 不同嵌入维度下同一元素嵌入平均余弦夹角和方差

统计量	模型	嵌入维度			
		8	16	32	64
平均夹角 (°)	未训练	89.5435	90.5969	88.3179	90.9668
	CGCNN	88.6925	90.1415	89.7174	90.0990
	MEGNet	79.3096	80.2907	81.6483	83.4162
方差 (deg^2)	未训练	410.7790	226.3515	73.0965	56.3134
	CGCNN	444.3819	206.9898	102.8234	46.9300
	MEGNet	450.7965	271.6864	163.9447	112.0157

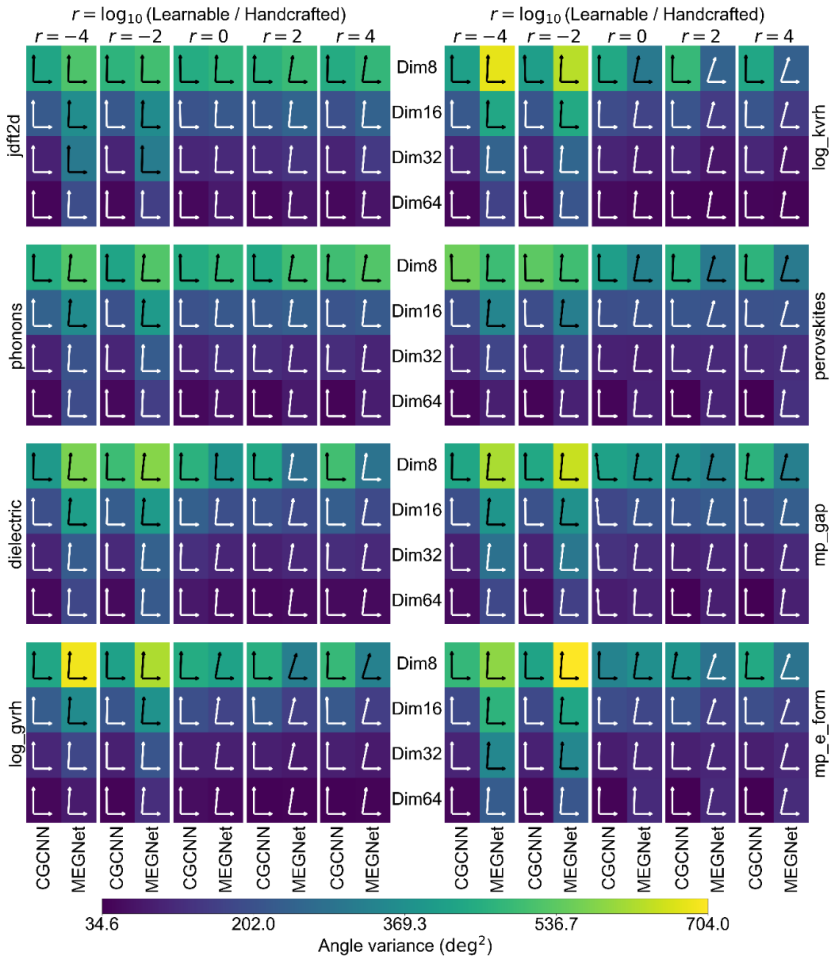


图 4.5 手工嵌入与学习嵌入余弦夹角和方差热力图

如表 4.1 所示，将投影后的手工通道与随机初始化的可学习嵌入层配对时，得到一个中心接近 90°的角度分布。这与先前其他研究者的发现一致，即高维空间中的随机向量预计会接近正交[98–100]。接近 90°的平均夹角本身更多反映的是高维几何基线，而非互补性的直接证据，监督学习改变的是对齐的离散度。此外，方差随维度增加而减小，低维度会导致更宽的对齐离散度。经过监督训练后，该平均角度仍高，离散度显著增加，表明可学习通道不会简单地退化为手工方向。这种更宽泛的分布意味着与手工设计的通道存在元素相关的偏差，与可学习的嵌入能够捕捉其余关系信息而非在嵌入融合下坍缩到人工设计的描述符上相一致。这些结果支持嵌入信息的互补性，4.2.3 节中将进一步将性能最佳的混合模型所编码的稳定元素间关系，提炼成元素距离矩阵与可迁移元素嵌入表。

4.2.3 混合元素嵌入构建与分析

几何分析支持了这两个通道的互补性，如果监督式属性预测能够诱导出稳定且有意义的元素间关系结构，我们能否在多个任务和模型族中提炼出鲁棒的元素表征，并在训练域之外的任务中获得一致的性能提升？因此，除了评估单个任务的性能之外，本研究还对嵌入融合是否能产生稳定的元素间关系进行评估。基于涵盖所有结构输入任务、交叉验证折叠、骨干网络和混合比例的训练模型网格，从学习到的嵌入向量中提炼出旋转不变的元素间关系从而构建了一系列可迁移的监督元素表 Mat2Vec-*。Mat2Vec-*旨在作为语料库衍生 Mat2Vec 的直接替代品，其中包括混合通道嵌入表 Mat2Vec-S（Mat2Vec-Supervised）、单通道嵌入表 Mat2Vec-H（Mat2Vec-Handcrafted）和 Mat2Vec-L（Mat2Vec-Learnable），它们能够恢复显著的元素周期表结构。

Mat2Vec-*构建过程如图 4.6 所示，它是一个监督式的、模型导出的元素嵌入表，通过从大量使用受控混合元素表示训练的属性预测模型中提取元素间关系结构来实现。该过程包含以下两个环节。

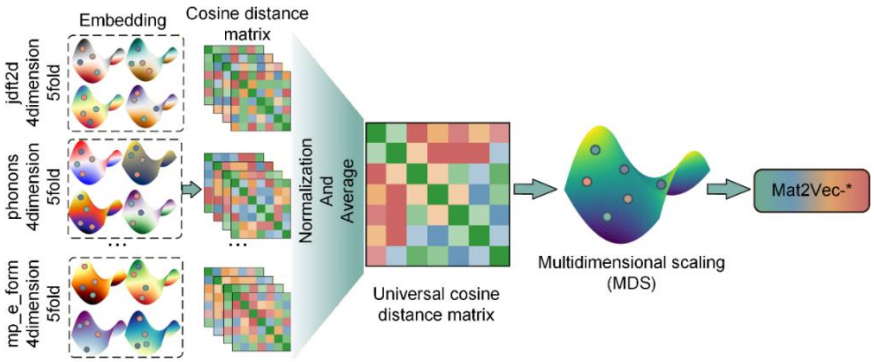


图 4.6 混合元素嵌入表构造流程

第一环节，本研究首先从训练好的模型集合中提炼出稳定的元素关系，元素关系仅考虑训练集中存在元素。在涵盖八个 MatBench 结构任务、五折交叉验证、两个骨干网络、四种嵌入维度和七种混合比例的 2240 个训练模型中，针对每个（任务、折叠、骨干网络、嵌入维度）设置选择性能最佳的混合比例配置，从 2240 个模型网络中筛选出 320 个性能最优模型。从每个选定的模型中，提取其学习到的元素向量，学习到的元素表示矩阵都经过了 L2 归一化。并将其学习到的元素向量转换为旋转不变的元素间距离统计量。对不同训练模型的元素向量直接取平均并不恰当，因为嵌入表示的坐标系不固定，等效解可能因潜在基的任意旋转而有所不同。余弦距离对嵌入空间的全局旋转是不变的，故本实验通过元素间成对余弦距离来提炼关系，而非原始坐标。在高维特征空间中，欧氏距离对向量的大小高度敏感，其值主要取决于跨维度差异的累积。随着维度的增加，不同样本对之间的欧氏距离往往集中在一个狭窄的范围内，导致最近邻和最远邻之间的区别越来越模糊，从而削弱了该度量方法在衡量样本相似性方面的区分能力[101]。相比之下，余弦距离衡量的是向量之间的角度，对向量的整体尺度不敏感，而是关注特征分布模式的相似性。这一特性使得余弦距离能够更有效地区分高维空间中样本之间的相对关系。因此，本研究将元素向量z_i转换为成对元素之间的余弦距离（公式 4.3）量化元素之间的相似性，该距离不受嵌入坐标任意旋转的影响。

$$d_{ij} = 1 - \frac{z_i - z_j}{\|z_i\|\|z_j\|} \tag{4.3}$$

距离值越小表示相似度越高。随后，将性能最优模型集合中的余弦距离矩阵进行聚合。距离矩阵经过全局重缩放，使得非零距离的中位数等于 1。对于每一对元素，最终的提炼距离定义为该元素对出现的所有模型中距离的算术平均值。由于筛选后的 320 个模型已经是在不同任务、折及配置下的最优模型，因此基于性能的加权可能不会显著改变最终聚合结果，同时本研究目标并非优化某一特定预测任务的表现，加权可能会引入对特定任务的偏置，故直接采用算术平均值。聚合后得到一个稳健、具有化学结构的元素距离矩阵，该矩阵概括了由监督式属性预测目标所诱导的关系几何结构，同时消除了单个距离矩阵的特殊性。

第二环节，基于聚合的距离信息重建了一系列嵌入表，记为 Mat2Vec-*，并将其发布以供重复使用。本实验通过重构一组固定维度元素向量，将提炼出的距离几何结构转换为显式嵌入表 Mat2Vec-S（由混合元素嵌入最优模型构建）。通过 MDS 降维后进行高斯分布标准化处理用于生成元素嵌入向量，MDS 诱导的余弦距离能够最好地保留聚合距离矩阵。MDS 参数设置如下，最大迭代次数为 2000，随机重启 8 次，随机种子固定为 42，从而生成维度分别为 32 和 64 的嵌入。另外，本实验使用同样的步骤，还基于相同的任务、折叠和嵌入向量维度，构建了对应于手工描述符（Mat2Vec-H）和可学习嵌入（Mat2Vec-L）的嵌入。

这一步骤生成了一个可直接使用的元素嵌入查找表，其形式类似于语料库衍生的表格（如 Mat2Vec），但完全源自监督模型几何结构。其中，Mat2Vec-S 展现出比原始语料库训练的 Mat2Vec 更宽的元素成对关系范围。

如图 4.7 所示，(a)为基于混合模型导出的距离矩阵重建的 Mat2Vec-S，(b)为经过语料库训练的 Mat2Vec，元素均按原子序数排列。为便于可视化，该图利用距离矩阵对称性，在下三角区域展示所有成对距离，在上三角区域仅展示截断后距离，以突出清晰邻域结构。余弦距离的截断位置在(a)中为 0.7，(b)为 0.4。因为二者距离分布存在差异，故显色范围不同：Mat2Vec-S 的范围在 0.1155 至 1.7340 之间，而 Mat2Vec 的范围在 0.1024 至 0.9022 之间。图中①-⑩分别为：①碱金属和碱土金属（相似的 s 价电子）；②早期过渡金属（低 d 电子数）；③晚期过渡金属（接近填满的 d 轨道）；④同族 p 区元素（共享相同的价电子数）；⑤氧族和卤族元素（高电负性的 p 区元素）；⑥后过渡金属（具有相似成键性质的 p 区金属）；⑦镧系元素（4f 系列的相似性）；⑧镧系-锕系跨系（f 区）结构；⑨内锕系元素块（Ac–Pu；5f 系列的相似性）；⑩第一周（3d）过渡金属（Sc–Zn；紧凑的近主对角线结构）。在图 4.7(a)所有元素成对关系中，Mat2Vec-S 产生的余弦距离范围为 0.1155-1.7340，对应的单元间角度为 27.81°-137.22°。相比之下，Mat2Vec（图 4.7(b)）产生了一个更为集中的几何形状，其余弦距离限制在 0.1024-0.9022 之间，角度范围为 26.15°-84.39°，且完全处于锐角范围内。这种压缩与 Mat2Vec 的文献共现训练一致，其中化学元素仅占整个词元词汇表的极小部分，因此往往占据嵌入空间中相对紧凑的区域。图中，Mat2Vec-S 在距离热力图中展现出对比度更高的条带状和块状结构，与基于语料库的 Mat2Vec 相比，大多数突出显示区域（①-⑨）的组内凝聚力更清晰。一个值得注意的例外是区域⑩（3d 过渡金属），Mat2Vec 显示出更明显的近邻模式。这种差异可能反映出，文献共现捕捉到了化学性质多样的 3d 系列中精细的使用模式，而监督式提炼（以及后续的重建步骤）优先考虑在不同任务和架构中最稳定的关系结构，这可能会使局部邻域变得平滑。因此，本研究将区域⑩视为一个有用的边界案例，有助于区分语料库目标和监督式属性学习分别强调在何种方面的化学相似性。由 Mat2Vec-S 计算得到的余弦距离热力图保留了与集成距离矩阵中观察到的相同的全局周期性结构，表明提炼后的嵌入表示捕捉的是稳定的元素关系，而非特定于任务或架构的结果。我们发布 Mat2Vec-S 以方便重用和复现，接下来将通过将其替换到固定成分模型中来探索其跨领域应用。

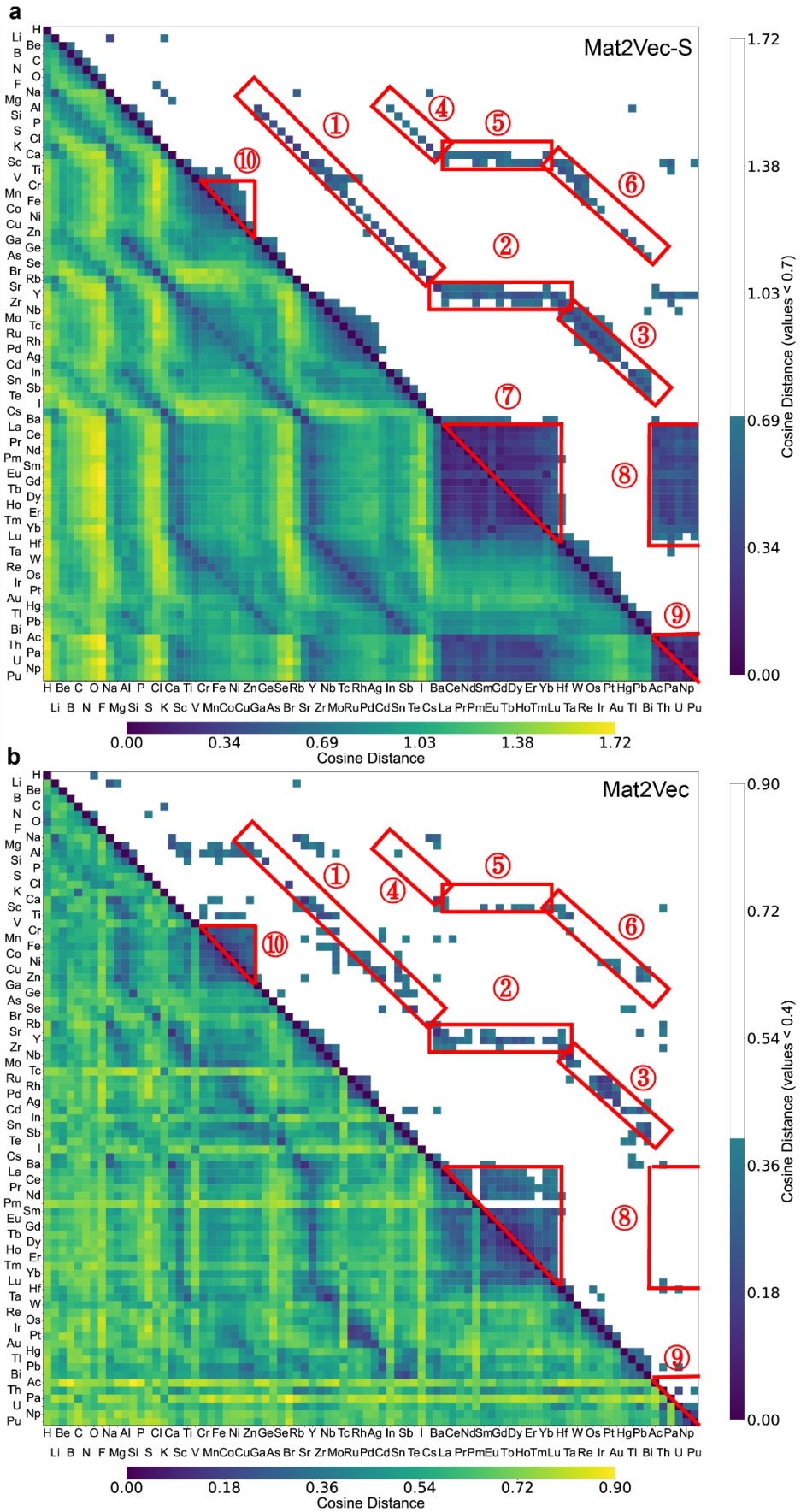


图 4.7 Mat2Vec-S 和 Mat2Vec 嵌入元素间余弦距离

4.2.4 跨模型可迁移结果分析

为了检验提炼出的元素几何结构是否能迁移到用于构建它的结构预测任务之外，本节通过将 Mat2Vec-*替换到一个模型输入仅包含组合元素的下游模型 CrabNet 中评估其域外效用，下游替换过程如图 4.8 所示。实验在 CrabNet 中将 Mat2Vec-*替换语料库衍生的 Mat2Vec 作为模型初始元素嵌入，CrabNet 默认嵌入方法为 Mat2Vec。在 MatBench 所有四个纯组分任务上进行测试评估跨范式迁移，包含两个回归任务（使用 MAE 评估）和两个分类任务（使用 ROC-AUC 评估）。该方法可以使实验结果区分元素嵌入效果优劣，而无需考虑架构或训练方面的修改。测试过程遵循标准的 CrabNet 默认参数设置，元素嵌入表在下游训练期间被冻结，而后续层保持可训练状态。所有模型均训练 500 个 epoch，并应用随机权重平均法（Stochastic Weight Averaging，SWA）形成模型最终权重。使用原始 Mat2Vec 嵌入复现排行榜性能后，使用 Mat2Vec-S、Mat2Vec-H 和 Mat2Vec-L 进行替换，过程中保持模型架构、训练和数据分割设置不变。

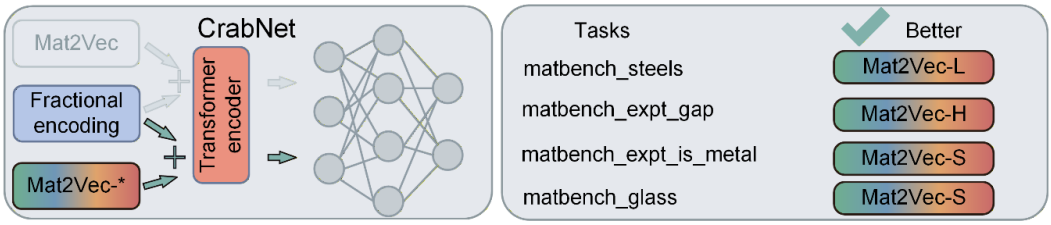


图 4.8 Mat2Vec-*（S/H/L）在 CrabNet 模型替换过程和下游任务效果

实验结果如表 4.2 和表 4.3 所示，DIM=d/512 中 d 表示原始嵌入大小，CrabNet 的输入投影到 512 维。相较于领域经典嵌入 Mat2Vec 和 SkipAtom[102]，Mat2Vec-*在 CrabNet 模型的组分任务中综合性能均提升 4.3%。Mat2Vec-S 和 Mat2Vec-H 在所有四个目标上都优于 Mat2Vec 基线。而 Mat2Vec-L 的迁移效果并不稳定，改进了 matbench_steels 和 matbench_glass，但降低了 matbench_expt_gap 和 matbench_expt_is_metal 的性能。最佳嵌入方式取决于目标，Mat2Vec-H 在两个回归任务上取得最佳的 MAE，而 Mat2Vec-S 在两个分类任务上提供了最高的 ROC-AUC。同时，DIM=32 更适用于 matbench_steels 和 matbench_expt_is_metal，而 DIM=64 更适用于 matbench_expt_gap 和 matbench_glass。另外，如果对每个嵌入方法内的 DIM=32 和 64 在模型预测效果上取平均值，不同任务的优胜者分别分布在所有三种嵌入方法上（L 适用于 matbench_steels，H 适用于 matbench_expt_gap，S 适用于另外两个分类任务）。在严格的跨范式测试下的结果表明，从基于结构的预测器中提取的监督元素关系编码了可迁移的化学先验信

息，即使在不使用晶体结构且下游架构不同的情况下，这些信息仍然有效。

表 4.2 CrabNet 不同嵌入方法性能比较（回归任务）

嵌入方法	数据集	
	matbench_steels (MAE)	matbench_expt_gap (MAE)
Mat2Vec (DIM = 200/512)	106.6691	0.3449
SkipAtom(DIM = 200/512)	104.9032	0.3432
Mat2Vec-S (DIM = 32/512)	**101.9088**	**0.3425**
Mat2Vec-S (DIM = 64/512)	**105.2421**	**0.3338**
Mat2Vec-H (DIM = 32/512)	**96.2327**	**0.3314**
Mat2Vec-H (DIM = 64/512)	**104.3017**	**0.3274**
Mat2Vec-L (DIM = 32/512)	**97.9391**	0.3868
Mat2Vec-L (DIM = 64/512)	**99.2107**	0.3718

表 4.3 CrabNet 不同嵌入方法性能比较（分类任务）

嵌入方法	数据集	
	matbench_expt_is_metal (ROC-AUC)	matbench_glass (ROC-AUC)
Mat2Vec (DIM = 200/512)	0.9641	0.9024
SkipAtom(DIM = 200/512)	0.9659	0.9071
Mat2Vec-S (DIM = 32/512)	**0.9686**	**0.9200**
Mat2Vec-S (DIM = 64/512)	**0.9674**	**0.9210**
Mat2Vec-H (DIM = 32/512)	**0.9675**	**0.9171**
Mat2Vec-H (DIM = 64/512)	**0.9684**	**0.9192**
Mat2Vec-L (DIM = 32/512)	0.9580	**0.9189**
Mat2Vec-L (DIM = 64/512)	0.9606	**0.9192**

4.2.5 小结

本节致力于探索监督式属性预测任务能否诱导出稳定且具科学价值的元素关系结构，并据此构建可迁移的元素嵌入表。首先，通过对混合模型中专家描述符与可学习嵌入两条通道的潜在向量进行几何层面的正交分析，验证了两者在共享空间中提供了互补的非冗余信息。在此基础上，本节提出一种从具体监督模型中提炼旋转不变元素关系的方法，成功构建了一系列包含混合通道及单通道信息的监督式元素嵌入表（Mat2Vec-*）。最后，本节在跨架构与跨数据集的条件下对新构建的嵌入表进行了域外性能验证。将提炼出的 Mat2Vec-*应用于仅包含成分输入的 CrabNet 模型，并在多个 MatBench 纯组分任务上进行测试。结果表明，Mat2Vec-S 和 Mat2Vec-H 在多个回归与分类任务上均稳定超越原始基线，证明其编码了具有高度泛化能力的化学先验信息，即使在脱离晶体结构输入和原有训练

架构的情况下依然有效。

4.3 本章小结

本章系统性地探讨了材料性质预测中元素表示的构建与优化问题，针对专家设计描述符与可学习嵌入的二元选择困境，本章首先构建了受控的混合元素嵌入实验框架，并在两类代表性图神经网络（CGCNN 与 MEGNet）及多种结构-性质基准测试上进行了充分验证。研究结果有力反驳了表征选择的非此即彼论，证明了通过简单的逐元素相加融合双通道输入，混合模型能够稳定达到或超越单一来源模型。

除提升模型性能外，监督式材料结构属性学习能够诱导出稳定的元素间几何关系，并可将其提炼成可复用的先验信息。在证实混合策略有效性的基础上，本章进一步深入模型的潜在空间，探索其元素关系结构。通过几何层面的正交分析，量化了专家描述符通道与可学习嵌入通道间的对齐程度，揭示了两者在共享空间中贡献了互补的非冗余信息。以此为理论支撑，在大量最优模型中聚合旋转不变的成对余弦距离，提炼出包含稳定元素关系几何结构的距离矩阵。基于聚合的元素间关系统计量，本章提供了一系列可迁移的元素嵌入表 Mat2Vec-*（S/H/L）。替代 CrabNet 中基于语料库的 Mat2Vec 嵌入后，Mat2Vec-*能够提升组分任务的预测性能。

综上，本章的模型预测效果和几何结论支持将逐元素相加混合元素表示作为一种强有力的默认元素表示方法，并将监督式关系提炼作为获得可迁移化学先验嵌入的实用途径。未来可以使用同样的方法用于检验等变图神经网络模型和基于 Transformer 的图模型是否也遵循相同的互补性和提炼原理。

第5章 总结与展望

5.1 本文工作总结

随着材料信息学的发展，基于机器学习的材料性质预测方法在高通量材料筛选与新材料发现中发挥着越来越重要的作用。在材料图神经网络等数据驱动模型中，元素表示是模型输入的核心基础，其表征能力直接影响模型预测性能与泛化能力。然而，现有研究往往将不同元素嵌入方式（如专家知识描述符与数据驱动学习嵌入）视为相互独立甚至互斥的设计选择，缺乏对元素嵌入结构与模型性能之间关系的系统性分析。同时，对于不同嵌入的信息互补性及其融合方式仍缺乏系统研究。

针对上述问题，本文围绕元素嵌入表示在材料机器学习模型中的构建与融合问题展开研究，以期为材料机器学习模型中的元素表征提供新的方法与思路。本文提出基于化学可替换性的一维元素嵌入构建与优化方法、基于不同元素嵌入方法的元素替代分组方法，以及混合元素嵌入方法和可迁移元素嵌入构建。现将本文主要工作总结如下：

（1）针对传统一维元素排序难以准确刻画元素化学相似性差异的问题，本文提出了一种基于实验统计，数据驱动的一维元素嵌入构建与优化方法。将晶体结构实验数据库中统计得到元素间化学可替换性关系，构建为元素间距离矩阵。在此基础上利用 MDS 方法将高维替换关系映射至一维空间，获得保持全局结构特征的元素排序。同时，为解决传统等间距表示无法区分相邻元素相似性差异的问题，构建平衡替代关系与间距的多目标优化函数，对元素间距进行优化。实验结果表明，该方法能够在一维空间中较好地保留元素间化学相似结构，高相似元素间距离可压缩至 0.05，而惰性气体及部分非金属元素间距离可扩展至 1.53，从而获得更具区分度的元素表示。

（2）针对材料图神经网络中不同元素嵌入方式缺乏系统比较的问题，本文在 CGCNN 与 MEGNet 两种典型材料图神经网络模型上，系统研究了基于人类知识的手工设计描述符与数据驱动学习嵌入两种元素嵌入方法的表现差异。通过在多个 MatBench 数据集上的五折交叉验证，从预测误差、嵌入空间结构及元素聚类行为等多个角度对嵌入效果进行分析。在此基础上，整合多个模型的嵌入结果，结合 K-Means 聚类与遗传算法构建数据驱动的元素分组方案，并进一步引入离子半径与电荷信息作为物理约束，对元素类别进行细化与扩展。研究结果表明，不同嵌入方式的嵌入空间结构、对模型性能影响存在显著差异，模型学习到

的元素关系并不完全遵循传统化学分类。通过数据驱动聚类与物理信息约束，获得了更符合材料结构可行性的元素分组方案，为理解数据驱动模型中的元素表示结构及材料设计中的元素替代提供新参考。

（3）针对不同元素嵌入方法（如专家知识描述符与可学习嵌入）在材料建模中通常被视为互斥选择的问题，本文从表示空间结构角度出发，探讨两类信息源在高维嵌入空间中协同表达的可能性。基于深度学习中多源表示可以在高维空间中共存的思想，本文提出了一种混合元素嵌入方法，通过在共享潜在空间中对来自不同信息源的元素表示进行逐元素相加，实现专家知识描述符与可学习嵌入的协同表达。实验通过在 CGCNN 和 MEGNet 两种晶体图神经网络模型中构建并行双通道元素输入结构，通过混合比例参数控制两类表示的融合，在 MatBench 结构子集任务上对不同混合策略进行系统评估。实验结果表明，在绝大多数任务和模型组合中，混合嵌入模型能够稳定获得优于单一嵌入的预测性能，专家知识与数据驱动表示在共享潜在空间中能够形成互补信息结构。本文进一步从混合模型中提炼稳定的元素关系，通过聚合多个模型学习到的元素间余弦距离构建元素关系矩阵，并基于 MDS 构建新的元素嵌入表 Mat2Vec-*。实验结果表明，相较于领域经典嵌入 Mat2Vec，Mat2Vec-*在 CrabNet 模型的组分任务中综合性能提升 4.3%。

5.2 未来工作展望

未来在本文研究工作的基础上，可从以下三个方面进一步改进：

（1）本文通过聚类与物理约束获得的元素分组方案，尝试刻画材料空间中元素替代的可行性。未来工作可将该分组结果作为先验知识纳入生成式模型（如扩散模型或变分自编码器）的采样过程中，通过约束元素选择范围测试是否能提升生成结构的化学合理性及可行性。

（2）本文在构建元素关系矩阵时主要基于晶体结构数据库中统计得到的元素可替换关系，该方法能够有效反映元素在材料结构中的替代行为。然而，元素之间的化学相似性不仅体现在是否能够相互替代，还体现在其在不同晶体结构环境中的分布特征。例如，同一元素在不同材料中可能具有不同的配位环境、局域结构类型及键长分布，这些结构环境信息同样能够反映元素的化学行为特征。未来工作可以进一步统计元素在不同结构环境中的分布特征，如配位数分布、结构原型分布或局域几何结构类型，并通过推土机距离（Earth Mover’s Distance，EMD）等分布距离度量方法计算不同元素之间的结构环境相似性。将结构环境分布信息与元素替代统计信息进行结合，构建更加全面的元素关系矩阵，为元素嵌入表示学习提供更加丰富的结构信息约束。

（3）本文在 CGCNN 和 MEGNet 两类典型晶体图神经网络模型上验证了专家知识描述符与可学习嵌入之间的互补关系，并通过混合嵌入策略获得稳定性能提升。然而，近年来材料机器学习领域不断出现新的模型结构，例如等变图神经网络以及基于 Transformer 的图模型等。这类模型在结构信息表达能力和相互作用建模方面具有不同于传统 GNN 的特点，其元素表示方式及信息融合机制也存在差异。研究初步在 Equiformer 模型上，采用小规模数据集进行验证。未来工作可以将本文提出的混合嵌入方法进一步扩展至这些模型以及不同任务中，系统检验不同模型架构下不同嵌入之间是否仍然表现出类似的互补关系，并进一步分析不同模型中元素表示结构的变化规律。

致 谢

感谢我的导师刘晓彤老师，让我在学术与为人处世中受益良多。感谢杨涛老师和刘金家老师，无论学业还是生活上的帮助。感谢同门全体同学、舍友和朋友，一起并肩奋斗的日子。感谢家人养育之恩。感谢周巍老师在我马拉松训练和生活上的指引。

求学之路，我深知自己在科研上造诣不深，天赋一般，也曾无数次陷入迷茫与焦虑。生活如马拉松，迷茫痛苦终会迎来“二次呼吸”。我也是幸运的，所遇之人皆为良善，所拥之家和睦温暖。既然身体这么好，今天就继续笑下去吧。

生活中的挑战，以为遗憾事，已是最好时。正如总书记所说，经风雨、见世面才能壮筋骨、长才干。愿自己信念坚定，不负盛世。

参考文献

[1] Butler K T, Davies D W, Cartwright H, et al. Machine learning for molecular and materials science[J]. Nature, Nature Publishing Group UK London, 2018, 559(7715): 547~555.

[2] Ward L, Agrawal A, Choudhary A, et al. A general-purpose machine learning framework for predicting properties of inorganic materials[J]. npj Computational Materials, Nature Publishing Group, 2016, 2(1): 1~7.

[3] Raccuglia P, Elbert K C, Adler P D, et al. Machine-learning-assisted materials discovery using failed experiments[J]. Nature, Nature Publishing Group UK London, 2016, 533(7601): 73~76.

[4] Wang H-C, Botti S, Marques M A. Predicting stable crystalline compounds using chemical similarity[J]. npj Computational Materials, Nature Publishing Group UK London, 2021, 7(1): 12.

[5] Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature[J]. Nature, Nature Publishing Group UK London, 2019, 571(7763): 95~98.

[6] Goodall R E A, Parackal A S, Faber F A, et al. Rapid discovery of stable materials by coordinate-free coarse graining[J]. Science Advances, 2022, 8(30): eabn4117.

[7] Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery[J]. Nature, Nature Publishing Group UK London, 2023, 624(7990): 80~85.

[8] Liu C, Tamaki H, Yokoyama T, et al. Shotgun crystal structure prediction using machine-learned formation energies[J]. npj Computational Materials, Nature Publishing Group UK London, 2024, 10(1): 298.

[9] Walsh A. The quest for new functionality[J]. Nature chemistry, Nature Publishing Group UK London, 2015, 7(4): 274~275.

[10] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model[J]. IEEE transactions on neural networks, IEEE, 2008, 20(1): 61~80.

[11] Xie T, Grossman J C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties[J]. Physical Review Letters, 2018, 120(14): 145301.

[12] Chen C, Ye W, Zuo Y, et al. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals[J]. Chemistry of Materials, 2019, 31(9): 3564~3572.

[13] Schütt K T, Sauceda H E, Kindermans P-J, et al. Schnet–a deep learning architecture for molecules and materials[J]. The Journal of Chemical Physics, AIP Publishing, 2018, 148(24).

[14] Liao Y-L, Wood B, Das A, et al. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations[J]. arXiv preprint arXiv:2306.12059, 2023.

[15] Zeni C, Pinsler R, Zügner D, et al. A generative model for inorganic materials design[J]. Nature, Nature Publishing Group UK London, 2025, 639(8055): 624~632.

[16] Park H, Onwuli A, Walsh A. Exploration of crystal chemical space using text-guided generative artificial intelligence[J]. Nature Communications, Nature Publishing Group UK London, 2025, 16(1): 4379.

[17] Onwuli A, Hegde A V, Nguyen K V, et al. Element similarity in high-dimensional materials representations[J]. Digital Discovery, Royal Society of Chemistry, 2023, 2(5): 1558~1564.

[18] Cerqueira T F T, Wang H, Botti S, et al. A non-orthogonal representation of the chemical space[J]. arXiv preprint arXiv:2406.19761, 2024.

[19] Teng P, Fu C, Shen S, et al. MatGNet: A graph neural network for crystal property prediction as an alternative to first-principles calculations[J]. Materials Today Communications, Elsevier, 2025, 44: 112021.

[20] Onwuli A, Butler K T, Walsh A. Ionic species representations for materials informatics[J]. APL Machine Learning, AIP Publishing, 2024, 2(3): 036112.

[21] Yadav L. Atoms as words: A novel approach to deciphering material properties using NLP-inspired machine learning on crystallographic information files (CIFs)[J]. AIP Advances, AIP Publishing, 2024, 14(4): 045205.

[22] Mukherjee S, Ghosh M, Basuchowdhuri P. CrysAtom: Distributed Representation of Atoms for Crystal Property Prediction[J]. arXiv preprint arXiv:2409.04737, 2024.

[23] Antunes L M, Grau-Crespo R, Butler K T. Distributed representations of atoms and materials for machine learning[J]. npj Computational Materials, Nature Publishing Group UK London, 2022, 8(1): 44.

[24] Glushkovsky A. AI Discovering a Coordinate System of Chemical Elements: Dual Representation by Variational Autoencoders[J]. arXiv preprint arXiv:2011.12090, 2020.

[25] Kusaba M, Liu C, Koyama Y, et al. Recreation of the periodic table with an unsupervised machine learning algorithm[J]. Scientific reports, Nature Publishing Group UK London, 2021, 11(1): 4780.

[26] Jia Y, Xian Y, Xu Y, et al. Universal Semantic Embeddings of Chemical Elements for Enhanced Materials Inference and Discovery[J]. arXiv preprint arXiv:2502.14912, 2025.

[27] Goodall R E A, Parackal A S, Faber F A, et al. Rapid discovery of stable materials by coordinate-free coarse graining[J]. Science Advances, 2022, 8(30): eabn4117.

[28] Merchant A, Batzner S, Schoenholz S S, et al. Scaling deep learning for materials discovery[J]. Nature, Nature Publishing Group UK London, 2023, 624(7990): 80~85.

[29] Liu C, Tamaki H, Yokoyama T, et al. Shotgun crystal structure prediction using machine-learned formation energies[J]. npj Computational Materials, Nature Publishing Group UK London, 2024, 10(1): 298.

[30] Pettifor D G. A chemical scale for crystal-structure maps[J]. Solid State Communications, Elsevier, 1984, 51(1): 31~34.

[31] Glawe H, Sanna A, Gross E K U, et al. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining[J]. New Journal of Physics, IOP Publishing, 2016, 18(9): 093011.

[32] Gschneidner Jr K A. Physical properties and interrelationships of metallic and semimetallic elements[G]//Solid state physics. Elsevier, 1964, 16: 275~426.

[33] Pettifor D G. Structure maps for. Pseudobinary and ternary phases[J]. Materials Science and Technology, 1988, 4(8): 675~691.

[34] Hautier G, Fischer C, Ehrlacher V, et al. Data Mined Ionic Substitutions for the Discovery of New Compounds[J]. Inorganic Chemistry, 2011, 50(2): 656~663.

[35] Park C W, Wolverton C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery[J]. Physical Review Materials, 2020, 4(6): 063801.

[36] Kusaba M, Liu C, Yoshida R. Crystal structure prediction with machine learning-based element

substitution[J]. Computational Materials Science, Elsevier, 2022, 211: 111496.

[37] Gu L, Wu R. Property-aimed embedding: a machine learning framework for material discovery[J]. arXiv preprint arXiv:1904.08750, 2019.

[38] Jakob K S, Reuter K, Margraf J T. Universally Accurate or Specifically Inadequate? Stress-Testing General Purpose Machine Learning Interatomic Potentials[J]. Advanced Intelligent Discovery, 2025: 202500031.

[39] Gasteiger J, Giri S, Margraf J T, et al. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules[J]. arXiv preprint arXiv:2011.14115, 2020.

[40] Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions[J]. npj Computational Materials, Nature Publishing Group UK London, 2021, 7(1): 185.

[41] Barroso-Luque L, Shuaibi M, Fu X, et al. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models[J]. arXiv preprint arXiv:2410.12771, 2024.

[42] Weston L, Tshitoyan V, Dagdelen J, et al. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature[J]. Journal of Chemical Information and Modeling, 2019, 59(9): 3692~3702.

[43] Jia Y, Xian Y, Xu Y, et al. Universal Semantic Embeddings of Chemical Elements for Enhanced Materials Inference and Discovery[J]. arXiv preprint arXiv:2502.14912, 2025.

[44] Li Y, Lai K, Wang T, et al. Element2Vec: Build Chemical Element Representation from Text for Property Prediction[J]. arXiv preprint arXiv:2510.13916, 2025.

[45] Dagdelen J, Dunn A, Lee S, et al. Structured information extraction from scientific text with large language models[J]. Nature communications, Nature Publishing Group UK London, 2024, 15(1): 1418.

[46] Gupta T, Zaki M, Krishnan N A, et al. MatSciBERT: A materials domain language model for text mining and information extraction[J]. npj Computational Materials, Nature Publishing Group UK London, 2022, 8(1): 102.

[47] Rupp M, Tkatchenko A, Müller K-R, et al. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning[J]. Physical Review Letters, 2012, 108(5): 058301.

[48] Jin L, Du Z, Shu L, et al. Transformer-generated atomic embeddings to enhance prediction accuracy of crystal properties with machine learning[J]. Nature Communications, Nature Publishing Group UK London, 2025, 16(1): 1210.

[49] Wigh D S, Goodman J M, Lapkin A A. A review of molecular representation in the age of machine learning[J]. WIREs Computational Molecular Science, 2022, 12(5): e1603.

[50] Sabando M V, Ponzoni I, Milios E E, et al. Using molecular embeddings in QSAR modeling: does it make a difference?[J]. Briefings in bioinformatics, Oxford University Press, 2022, 23(1): bbab365.

[51] De Breuck P-P, Evans M L, Rignanese G-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet[J]. Journal of Physics: Condensed Matter, IOP Publishing, 2021, 33(40): 404002.

[52] 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 UK London, 2020, 6(1): 138.

[53] Jaeger S, Fulle S, Turk S. Mol2vec: Unsupervised Machine Learning Approach with Chemical

Intuition[J]. Journal of Chemical Information and Modeling, 2018, 58(1): 27~35.

[54] Mann V, Brito K, Gani R, et al. Hybrid, interpretable machine learning for thermodynamic property estimation using grammar2vec for molecular representation[J]. Fluid Phase Equilibria, Elsevier, 2022, 561: 113531.

[55] Liu S, Wang H, Liu W, et al. Pre-training Molecular Graph Representation with 3D Geometry[J]. arXiv preprint arXiv:2110.07728, 2021.

[56] Wang Y, Wang J, Cao Z, et al. Molecular contrastive learning of representations via graph neural networks[J]. Nature Machine Intelligence, Nature Publishing Group UK London, 2022, 4(3): 279~287.

[57] Wang Y, Magar R, Liang C, et al. Improving Molecular Contrastive Learning via Faulty Negative Mitigation and Decomposed Fragment Contrast[J]. Journal of Chemical Information and Modeling, 2022, 62(11): 2713~2725.

[58] Wang L, Liu Y, Lin Y, et al. ComENet: Towards complete and efficient message passing for 3D molecular graphs[J]. Advances in Neural Information Processing Systems, 2022, 35: 650~664.

[59] Omee S S, Louis S-Y, Fu N, et al. Scalable deeper graph neural networks for high-performance materials property prediction[J]. Patterns, Elsevier, 2022, 3(5).

[60] Frank T, Unke O, Müller K-R. So3krates: Equivariant attention for interactions on arbitrary length-scales in molecular systems[J]. Advances in Neural Information Processing Systems, 2022, 35: 29400~29413.

[61] Liao Y-L, Smidt T. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs[J]. arXiv preprint arXiv:2206.11990, 2022.

[62] Passaro S, Zitnick C L. Reducing SO (3) convolutions to SO (2) for efficient equivariant GNNs[C]//International conference on machine learning. PMLR, 2023: 27420~27438.

[63] Simeon G, De Fabritiis G. Tensornet: Cartesian tensor representations for efficient learning of molecular potentials[J]. Advances in Neural Information Processing Systems, 2023, 36: 37334~37353.

[64] Bendaoud A, Hachouf F. SevenNet: rethinking convolutional neural networks with a formula-based architecture[J]. Applied Intelligence, 2026, 56(2): 61.

[65] Bagal V, Aggarwal R, Vinod P K, et al. MolGPT: Molecular Generation Using a Transformer-Decoder Model[J]. Journal of Chemical Information and Modeling, 2022, 62(9): 2064~2076.

[66] Liu Y, Zhang R, Li T, et al. MolRoPE-BERT: An enhanced molecular representation with Rotary Position Embedding for molecular property prediction[J]. Journal of Molecular Graphics and Modelling, Elsevier, 2023, 118: 108344.

[67] Devi K G, Bedadhala R S, Kumar S S, et al. DeBERTaSSL: Enhancing Molecular Property Prediction with Self-Supervised Learning[C]//2023 4th International Conference on Intelligent Technologies (CONIT). IEEE, 2024: 1~7.

[68] Wang A Y-T, Kauwe S K, Murdock R J, et al. Compositionally restricted attention-based network for materials property predictions[J]. Npj Computational Materials, Nature Publishing Group UK London, 2021, 7(1): 77.

[69] Guerrero P, Hašan M, Sunkavalli K, et al. MatFormer: A Generative Model for Procedural Materials[J]. ACM Transactions on Graphics, 2022, 41(4): 1~12.

[70] Wang Y, Liu X, Chen H, et al. Exploring lightweight language models for materials informatics with AlchemBERT[J]. Cell Reports Physical Science, Elsevier, 2025, 6(8): 102724.

[71] Guo Z, Yu W, Zhang C, et al. GraSeq: Graph and Sequence Fusion Learning for Molecular Property Prediction[C]//Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Virtual Event Ireland: ACM, 2020: 435~443.

[72] Liu J, Lei X, Zhang Y, et al. The prediction of molecular toxicity based on BiGRU and GraphSAGE[J]. Computers in biology and medicine, Elsevier, 2023, 153: 106524.

[73] Nguyen D M H, Lukashina N, Nguyen T, et al. Structure-Aware E(3)-Invariant Molecular Conformer Aggregation Networks[J]. arXiv preprint arXiv:2402.01975, 2024.

[74] Wang T, Sun J, Zhao Q. Investigating cardiotoxicity related with hERG channel blockers using molecular fingerprints and graph attention mechanism[J]. Computers in biology and medicine, Elsevier, 2023, 153: 106464.

[75] Zhang H, Wu J, Liu S, et al. A pre-trained multi-representation fusion network for molecular property prediction[J]. Information Fusion, Elsevier, 2024, 103: 102092.

[76] Lu X, Xie L, Xu L, et al. Integrating Chemical Language and Molecular Graph in Multimodal Fused Deep Learning for Drug Property Prediction[J]. Computational and Structural Biotechnology Journal, 2024, 23: 1666~1679.

[77] Deng D, Chen X, Zhang R, et al. XGraphBoost: Extracting Graph Neural Network-Based Features for a Better Prediction of Molecular Properties[J]. Journal of Chemical Information and Modeling, 2021, 61(6): 2697~2705.

[78] Wu J, Su Y, Yang A, et al. An improved multi-modal representation-learning model based on fusion networks for property prediction in drug discovery[J]. Computers in Biology and Medicine, Elsevier, 2023, 165: 107452.

[79] Zheng Z, Wang H, Tan Y, et al. EMPPNet: Enhancing Molecular Property Prediction via Cross-modal Information Flow and Hierarchical Attention[J]. Expert Systems with Applications, Elsevier, 2023, 234: 121016.

[80] Xiang H, Jin S, Xia J, et al. An Image-enhanced Molecular Graph Representation Learning Framework[C]//Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence. Jeju, South Korea: International Joint Conferences on Artificial Intelligence Organization, 2024: 6107~6115.

[81] Ma M, Lei X. A deep learning framework for predicting molecular property based on multi-type features fusion[J]. Computers in biology and medicine, Elsevier, 2024, 169: 107911.

[82] Yin R, Liu R, Hao X, et al. Multi-Modal Molecular Representation Learning via Structure Awareness[J]. IEEE Transactions on Image Processing, IEEE, 2025, 34: 3225-3238.

[83] Yang G, Jiang S, Luo Y, et al. Cross-Modal Prediction of Spectral and Structural Descriptors via a Pretrained Model Enhanced with Chemical Insights[J]. The Journal of Physical Chemistry Letters, 2024, 15(34): 8766~8772.

[84] Chacko E, Sondhi R, Praveen A, et al. Spectro: A multi-modal approach for molecule elucidation using IR and NMR data[J]. ChemRxiv. 05 November 2024. DOI: https://doi.org/10.26434/chemrxiv-2024-37v2j

[85] Hua Y, Feng Z, Song X, et al. MMDG-DTI: Drug–target interaction prediction via multimodal feature fusion and domain generalization[J]. Pattern Recognition, Elsevier, 2025, 157: 110887.

[86] Polat C, Kurban H, Serpedin E, et al. Understanding the Capabilities of Molecular Graph Neural Networks in Materials Science Through Multimodal Learning and Physical Context Encoding[J]. arXiv preprint arXiv:2505.12137, 2025.

[87] Tang X, Tran A, Tan J, et al. MolLM: a unified language model for integrating biomedical text

with 2D and 3D molecular representations[J]. Bioinformatics, Oxford University Press, 2024, 40(Supplement_1): i357~i368.

[88] Wiercioch M, Kirchmair J. DNN-PP: a novel deep neural network approach and its applicability in drug-related property prediction[J]. Expert Systems with Applications, Elsevier, 2023, 213: 119055.

[89] Gong X, Liu M, Liu Q, et al. MDFCL: Multimodal data fusion-based graph contrastive learning framework for molecular property prediction[J]. Pattern Recognition, Elsevier, 2025: 111463.

[90] Chen R, Li C, Wang L, et al. Pretraining graph transformer for molecular representation with fusion of multimodal information[J]. Information Fusion, Elsevier, 2025, 115: 102784.

[91] Liu S, Nie W, Wang C, et al. Multi-modal molecule structure–text model for text-based retrieval and editing[J]. Nature Machine Intelligence, Nature Publishing Group UK London, 2023, 5(12): 1447~1457.

[92] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[93] Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers). 2018: 464~468.

[94] Su J, Ahmed M, Lu Y, et al. Roformer: Enhanced transformer with rotary position embedding[J]. Neurocomputing, Elsevier, 2024, 568: 127063.

[95] Sun F, Liu J, Wu J, et al. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing China: ACM, 2019: 1441~1450.

[96] Press O, Smith N A, Lewis M. Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation[J]. arXiv preprint arXiv:2108.12409, 2021.

[97] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv preprint arXiv:1301.3781, 2013.

[98] Vershynin R. High-dimensional probability: An introduction with applications in data science[M]. Cambridge university press, 2018, 47.

[99] Wang Y-A, Chen Y-N. What do position embeddings learn? an empirical study of pre-trained language model positional encoding[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 6840~6849.

[100] Zuo C, Guerzhoy P, Guerzhoy M. Position information emerges in causal transformers without positional encodings via similarity of nearby embeddings[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 9418~9430.

[101] Durrant R J, Kabán A. When is ‘nearest neighbour’meaningful: A converse theorem and implications[J]. Journal of Complexity, Elsevier, 2009, 25(4): 385~397.

[102] Antunes L M, Grau-Crespo R, Butler K T. Distributed representations of atoms and materials for machine learning[J]. npj Computational Materials, 2022, 8(1): 44.

个人简历、在学期间发表的学术论文及研究成果

一、个人简历

陈昊天，出生日期：2000.11.28，

2018.09-2022.07，北京信息科技大学，计算机科学与技术专业，获工学学士学位。

2023.09-至今，北京信息科技大学，计算机科学与技术专业，攻读硕士学位。

二、学术论文及研究成果

[1] **陈昊天**,杨涛,刘晓彤.化学分子的隐藏空间嵌入方法原理和应用[J].化学进展,2025,37(10):1456-1478.

[2] **Chen H**, Liu X. Hybridizing expert descriptors and learned embeddings yields transferable element representations[J]. Nature Communications (已送审，投稿日期 2026 年 2 月 15 日)

[3] Wang Z, Liu X, **Chen H**, et al. Exploring multi-fidelity data in materials science: Challenges, applications, and optimized learning strategies[J]. Applied Sciences, 2023, 13(24): 13176.

[4] Wang Y, Liu X, **Chen H**, et al. Exploring lightweight language models for materials informatics with AlchemBERT[J]. Cell Reports Physical Science, 2025, 6(8): 102724.

[5] Yang T, Zhang J, **Chen H**, et al. Data-Driven Deciphering Structure–Mössbauer Spectroscopy Relationships in Iron-Based Compounds[J]. The Journal of Physical Chemistry Letters, 2026, 17(7): 2111-2117.

[6] 化学元素序列排序算法及评价软件 V1.0 登记号：2024SR0664976. (软著)

三、所获奖励

[1] 2025 年 研究生学业奖学金 三等奖

[2] 2024 年 研究生学业奖学金 二等奖