分 类 号： TB39　　密　　级：公　开
论文编号：2017022267

贵　州　大　学

2020 届硕士研究生学位论文

基于深度神经网络的无机化合物新材料发现算法研究

学位类别：　工程硕士专业学位
专　　业：　机 械 工 程
校内导师：　胡建军、李少波（教授）
校外导师：　柴旭东（高级工程师）
研 究 生：　但雅波

中国·贵州·贵阳

2020 年 6 月

目　录

摘　要

材料是人类生产生活的物质基础，是直接推动社会发展的动力。但目前的材料发现仍然涉及重大的反复试验，可能需要数十年的研究才能确定适合于技术应用的材料。这一漫长发现过程的主要原因是，潜在材料的数量是巨大的，明智地选择要关注的材料以及进行哪些实验荆棘密布。面对竞争激励的制造业和快速的经济发展，材料科学家和工程师必须缩短新材料研的发周期，以解决 21 世纪的巨大挑战。然而，当前的新材料研发主要依据研究者的科学直觉和大量重复的“尝试法”实验。通过实验或高性能原理模拟的材料数量与预期潜在的多样性相比显得相形见绌。近年来，人工智能（Artificial Intelligence：AI）取得了令人兴奋的进展，其中机器学习（Machine Learning：ML）和深度学习(Deep Learning：DL)技术的应用为人们在各种领域的任务中带来了竞争性的表现，加上越来越多的具有实验和/或计算特性的材料数据库的可用性激发了研究人员采用先进的基于数据驱动的材料发现技术的兴趣，以加速发现具有精选工程的新材料。

本研究在国家自然科学基金项目“基于机器学习与图像处理算法的高通量组合材料实验相图生成与物相辨识方法研究”（项目编号：51741101）支持下，针对巨大的潜在材料的数量而无法明智地选择要关注的材料的问题，采用深度学习方法设计出能高效采样的无机化合物生成模型和能精确预测材料属性的筛选模型，基于这两个模型进行无机化合物新材料的发现，对提高材料的研发效率具有重要的理论和现实意义。主要工作及创新点如下：

提出了一种基于生成对抗网络（Generative Adversarial Networks：GAN）的无机材料生成模型 MatGAN，可以从庞大的无机材料的化学设计空间中进行有效的采样。系统的实验和验证表明，MatGAN 在生成能力方面可以实现高度的唯一性，有效性和多样性。通过扩展无机晶体结构数据库（Inorganic Crystallographic Structure Database：ICDS），材料项目（Materials Project：MP）和开放量子材料数据库（Open Quantum Materials Database：OQMD），本研究设计的 MatGAN 生成模型可用于探索未知的无机材料设计空间。与彻底筛选数十亿个候选对象相比，导出的扩展数据库可用于更高效率的高通量计算筛选。尽管已采用电荷中性和电负性平衡原理来过滤化学上难以置信的成分，以便更有效地搜索新材料，但此类明确的成分规则仍然过于宽松，无法确保在广阔的化学设计空间中对新材料进行有效采样。虽然可以列举出少于 5 种元素的假设材料（对于具有电荷中性和平衡电负性的 4 元素材料而言为 320 亿种材料），但是更多元素的设计空间可能具有挑战性，而 MatGAN 模型可以提供很大帮助。

设计了一种材料的层次式表征方法，基于该表征方法提出了包含全卷积层的卷积神经网络材料属性预测模型。通过设计特殊的卷积算子，该模型能很好的从材料的原始输入矩阵中提取出有用的特征，进行最后的回归任务。利用 ICSD 中的数据进行不同任务的有监督训练，建立了能预测材料带隙的 ICSD-BG 和能预测材料形成能量的 ICSD-FE 的高精度预测模型。建立的模型在测试集上的预测表现和在验证集上的表现基本一致，这保证了后续用预测模型在 MatGAN 生成的假设材料中进行筛选的可靠性。

将 ICSD-FE 施加于 GAN-ICSD 生成的假设材料上进行筛选，筛选出形成能量较低的材料。在筛选出的形成能量较低的材料基础上，用 ICSD-BG 继续筛选带隙在 1.0~2.0eV 之间的材料。将筛选出的材料的带隙在 1.0~2.0eV 以及这些材料用 ICSD-FE 预测的形成能量值放在了 http://github/danyabo/appendix.com 以供研究者们进行后续的 DFT 仿真计算或实验合成。

关键词：深度学习；生成对抗网络；新材料发现；卷积神经网络；带隙；材料的形成能量

Abstract

Materials are the material basis of human production and life, and the driving force that directly promotes social development. However, the current material discovery still involves major repeated experiments, which may take decades of research to determine materials suitable for technical applications. The main reason for this long discovery process is that the number of potential materials is huge, and it is thorny to choose the materials to be concerned about and which experiments to carry out. In the face of competitive manufacturing and rapid economic development, materials scientists and engineers must shorten the research and development cycle of new material discovery in order to solve the huge challenges of the 21st century. However, the current research and development of new materials is mainly based on the scientific instincts of researchers and a large number of repeated "trial method" experiments. The number of materials simulated by experiments or high-performance principles is dwarfed by the expected potential diversity. In recent years, artificial intelligence (AI) has made exciting progress, among which the application of machine learning (ML) and deep learning (DL) technology has brought competitive performance to people in various fields. In addition, the availability of more and more material databases with experimental and/or computational characteristics has inspired researchers to adopt advanced data-driven material discovery technology interest in accelerating the discovery of new materials with selected engineering.

Under the support of the National Natural Science Foundation project "Research on high-throughput composite material experimental phase diagram generation and phase identification method based on machine learning and image processing algorithm" (Project No.: 51741101), aiming at the problem of huge potential material quantity and unable to choose the material to be concerned wisely, this study adopts the deep learning method to design a generation model that can efficiently sample from inorganic compounds and a screening model that can accurately predict the properties of materials. Based on these two models, new materials for inorganic compounds are carried out. The discovery has important theoretical and practical significance for improving the research and development efficiency of materials. The main work and innovations are as follows:

A GAN-based generation model is proposed, which can effectively sample from the huge chemical design space of inorganic materials. System experiments and verifications show

that our GAN model can achieve a high degree of uniqueness, effectiveness and diversity in terms of generating capabilities. By expanding ICDS, Materials Project (MP) and OQMD, our generative model can be used to explore the unknown inorganic material design space. Compared with the thorough screening of billions of candidates, the derived extended database can be used for more efficient high-throughput computational screening. Although the principles of charge neutrality and electronegativity balance have been used to filter chemically incredible components in order to search new materials more efficiently, such clear composition rules are still too lax to ensure that Effective sampling of new materials. Although hypothetical materials with less than 5 elements can be cited (32 billion for 4-element materials with charge neutrality and balanced electronegativity), the design space for more elements can be challenging, and MatGAN model can help a lot.

A hierarchical characterization method for materials is designed. Based on the characterization method, a material property prediction model for convolutional neural networks with fully convolutional layers is proposed. By designing a special convolution operator, the model can extract useful features from the original input matrix of the material and perform the final regression task. Using the data in ICSD for supervised training of different tasks, a high-precision prediction model of ICSD-BG that can predict the band gap of the material and ICSD-FE that can predict the formation energy of the material is established. The prediction performance of the established model on the test set is basically consistent with the performance on the validation set, which ensures the reliability of subsequent screening of hypothetical materials generated by GAN with the prediction model.

ICSD-FE is applied to the hypothetical materials generated by GAN-ICSD for screening, and materials with low formation energy are screened out. On the basis of screening materials with low formation energy, ICSD-BG was used to continue screening materials with a band gap between 1.0 and 2.0 eV. The band gap of the selected materials is 1.0 ~ 2.0eV and the formation energy of these materials predicted by ICSD-FE is placed on http: //github/danyabo/appendix.com for researchers to perform subsequent DFT simulation calculations or Experimental synthesis.

Keywords: Deep learning; generative adversarial network; discovery of new materials; convolutional neural network; band gap; formation energy

第 1 章 绪论

1.1 研究背景及意义

材料是人类生产生活的物质基础，是直接推动社会发展的动力[1]，技术的重大进步很大程度上取决于新材料的发现。从史前的青铜和钢的发现到 20 世纪合成聚合物的发明，新材料的出现不仅推动了人类历史的进步，而且也促进了技术的发展和产业的升级。如今，材料创新也成为我们应对一些最紧迫的社会挑战的关键，例如全球气候变化和未来的能源供应[2,3]。但是，目前的材料发现仍然涉及重大的反复试验，可能需要数十年的研究才能确定适合于技术应用的材料。这一漫长发现过程的主要原因是，潜在材料的数量是巨大的，明智地选择要关注的材料以及进行哪些实验荆棘密布。仅考虑天然存在的元素，9000 个晶体结构原型和化学计量组成[4,5]，大约有3×10^{11}个潜在的四元化合物和10^{13}个五元组合。据估计，理论材料的总数可以达到10^{100}[6]，这给设计具有新的成分和结构的材料带来了巨大的挑战。

面对竞争激励的制造业和快速的经济发展，材料科学家和工程师必须缩短新材料的研发周期，以解决 21 世纪的巨大挑战。然而，当前的新材料研发主要依据研究者的科学直觉和大量重复的“尝试法”实验。通过实验的材料数量与预期潜在的多样性相比显得相形见绌。仅通过实验方法或高性能原子模拟（一种材料可能需要几百个中央处理单元（CPU）几小时才能获得基本性能，而且经常需要一系列操作（在执行性能计算之前，需要计算优化的晶体结构））设计空间不会太大。大多数的研究都采用了资源密集型的实验测量或第一性原理计算，而原子模拟的计算模型，特别是亚微米长的计算方法，由于缺乏精确定义的力场来描述原子间的相互作用而受到影响[7-10]。使用密度泛函理论（Density Functional Theory：DFT）进行计算要求严格的电子结构计算，通常仅限于模拟几百个原子[8-10]。标准材料表征实践，例如计算能带结构，考虑有限尺寸缩放，电荷校正[11]以及超越标准 DFT 时使用 Green 函数方法（如完全自适应的 GW）[12,13]会使计算变得非常昂贵。最终，对这个搜索空间的强力探索，即使是高通量的计算方式[14]，也是不切实际的。

改变计算材料发现的潜在途径是采用深度学习（Deep Learning：DL）的方法。近年来，人工智能（AI）取得了令人兴奋的进展，其中机器学习（Machine Learning：ML）和 DL 技术的应用为人们在各种领域的任务中带来了竞争性的表现，包括图像识别[15-17]，语音识别[18-20]和自然语言理解[21-23]。即使在围棋这个古老的复杂游戏中，人

工智能玩家已经在向人类学习和不学习的情况下令人信服地击败了人类世界冠军[24]。】正是 DL 的这种优越性加上越来越多的具有实验和/或计算特性的材料数据库（MP[25]、OQMD[26]、ICSD[5]等）的可用性激发了研究人员采用先进的基于数据驱动的材料发现技术的兴趣，以加速发现具有精选工程的新材料。

本研究在国家自然科学基金应目“基于机器学习与图像处理算法的高通量组合材料实验相图生成与物相辨识方法研究”（项目编号：51741101）支持下，针对巨大的潜在材料的数量而无法明智地选择要关注的材料的问题，采用深度学习方法设计出能高效采样的无机化合物生成模型和能精确预测材料属性的筛选模型，基于这两个模型进行无机化合物新材料的发现，对提高的研发效率，缩短开发周期，降低研发成本等具有重要的理论和现实意义。

1.2 国内外研究现状

如图 1.1，基于神经网络发现新材料的一般步骤：材料数据的获取；材料数据的表征；采用神经网络建立材料属性的预测模型；用预测模型在材料空间中进行筛选以发现具有某种特性的新材料。如 Jha 等人[27]在 OQMD[26,28]上采用 One-hot [29]材料表征方法，并用全连接神经网络（FNN）[30]算法建立了材料形成能量的预测模型，并在上亿种可能的材料空间（$A_xB_yC_zD_w$，其中 A、B、C 和 D 是元素周期表中的常见元素，x、y、z、w 为整数且$x+y+z+w\leq10$）搜索出了数百种可能存在的新材料。Valen 等人[31]使用 Magpie[32]将每种无机化合表征成 134 维的向量，用随机森林（Random Forest：RF）算法[33]建立了超导体的超导临界温度（Critical Temperature：Tc）预测模型，并用该模型在 ICSD[5]中进行筛选，以查找可能的超导体。Xiang 等人[34]利用具有物理信息的结构描述符并采用迁移学习方法建立了钙钛矿材料的预测模型，然后用该模型从 21316 个假设的钙钛矿结构中筛选出有前途的新型钙钛矿材料。以下分别介绍材料数据集、材料数据表征、预测模型的研究现状。

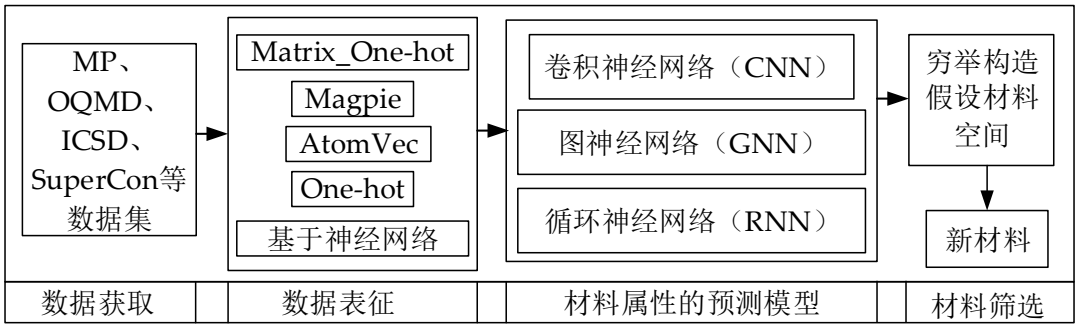


图 1.1 基于神经网络发现新材料的一般步骤

1.2.1 数据研究现状

目前，越来越多的具有实验和/或计算特性的材料数据库（例如，MP[15,25]和 ICSD[5]）的可用性导致了最近出现的用于材料特性预测的机器学习。同时， DFT[35]提供了一种较方便的方法来预测原子级别上的许多材料属性。DFT 的计算结果已存储到大型数据集中，例如 OQMD[26,28]，材料发现数据库（AFLOWLIB）[36]，新材料发现数据库（NoMaD）[37]和超导材料数据库（SuperCond）[38]。它们包含实验观察到的和假设材料的10^4~10^6个 DFT 计算特性。在过去的几十年中，这些材料数据集已应用于新的数据驱动的材料信息学范例[39,40]。这些大型数据资源的可用性激发了研究人员对应用先进的基于数据驱动的 DL 技术的兴趣。在无机材料研究中使用最多的是 OQMD、ICSD、MP 三个数据库，以下对他们进行介绍。

1）材料项目（MP）数据库

材料项目（Materials Project：MP）[15]的绝大多数数据都来自 ICSD 中的化合物。MP 是材料基因组计划的核心程序，该程序利用信息时代的高通量计算能力和最佳实践，为加速材料设计创建一个开放、协作和数据丰富的生态系统以揭示所有已知无机材料的特性。MP 于 2011 年 10 月由麻省理工学院和劳伦斯伯克利国家实验室联合发起，目前已在全球十多家机构建立了合作伙伴关系 MP 结合了高通量计算，基于 Web 的传播和开源分析工具（如图 1.2），为材料科学家提供了解决材料发现问题的新视角，其提供了超过 13 万个无机化合物的结构信息以及性质。可以通过多种渠道访问此开源数据集，以进行交互式探索和数据挖掘。

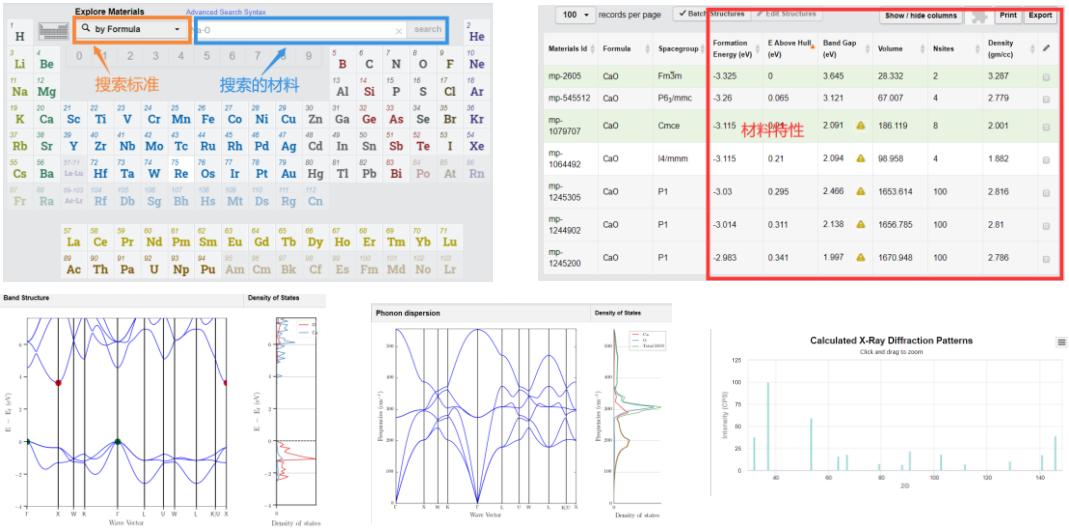


图 1.2 MP 基于 Web 传播和开源分析工具

2）无机晶体结构数据库（ICSD）

无机晶体结构数据库（Inorganic Crystallographic Structure Database：ICDS）[5]，由德国 Fachin-formationszentrum Karlsruhe（FIZ）和美国国家标准与技术研究院（NIST）

合作制作，目前收集了 6 万多个无机材料晶体结构条目。典型的条目包括化学名称、分子式、晶胞、空间群、完整的原子参数（包括原子位移参数）、位置占用因子、和文献引用。ICSD 具有 FIZ-Karlsruhe 和法国格勒诺布尔 Laue-Langevin 研究所（ILL）合作开发的可通过 Web 访问的界面[19]。可以通过简单的图形界面或完全遵循 CRYSTIN 命令语法的专家模式来搜索数据库。

3）开放量子材料数据库（OQMD）

开放量子材料数据库（Open Quantum Materials Database：OQMD）[26]一贯是 DFT 计算出的总能量和松弛晶体结构的集合。目前，使用 Vienna Ab-initio 模拟软件包（VASP），对 ICSD 中没有部分站点占用且原始单元中的原子少于 35 个的条目执行了 32489 次 DFT 计算。OQMD 主要有两个功能：收集了已知结构的大量数据，可从中搜索最佳材料（例如，“高容量转换阳极筛选”，“Li2O 电池筛选”，“HF 清除锂离子电池阴极涂层”等）；对简单和复杂系统的化学势和凸壳的准确描述，可从中轻松进行稳定性测试（例如，“寻找新的增强轻质镁合金中的沉淀物”，“新型三元化合物的数据挖掘”）。 OQMD 主要受 ICSD 实验观察和分类的限制，为了解决这个问题，OQMD 还在数据库中包含了许多一元，二元和三元原型结构的 DFT 计算。在 OQMD 中包含这些原型可为未探索的凸包和可能的未被发现的化合物提供近似值，因为它们会采样未探索的成分和系统。截至 2020 年，OQMD 中的结构总数（包括 ICSD 结构和原型）已超过 600000，并且每天都在增长，可以从[20]免费获得使用。

1.2.2 表征方法研究现状

任何 DL 模型的两个最重要的方面是数据表征和学习算法，材料数据的适当表示对于生成精确模型至关重要。无机材料表征方法主要分为两类，一类是基于分子式的表征方法，一类是基于晶体结构的表征方法。

基于分子式的材料表征方法包括元素属性统计、One-hot 编码、Atom2Vec 编码[34]等。基于分子式的材料表征方法使用最广泛的是元素属性统计，Stanev 等人[41]将其应用到超导材料 *Tc* 的预测模型上，使用 Magpie[32]计算分子式中元素的 22 种统计数据属性，将每种无机化合物描述成 134 维的向量，并用随机森林（RF）算法预测了 *Tc*；后来 Dan 等人[42]根据元素属性统计设计了一种超导体的矩阵表示方法，并用卷积梯度提升决策树（ConvGBDT）算法对 *Tc* 预测模型的精度进行了提升；Zhuo 等人[43]根据化合物组成元素的和，差，最大和最小值，将每个组成属性分为 4 个变量，把材料表征成 136 维的向量，在 3896 个金属化合物的实验带隙值（Band Gap）上用支持向量回归（SVR）算法建立了能预测金属化合物带隙的模型。对于 One-hot 编码，Calfa[29]用 One-hot 编码的方式将材料表示成固定长度的向量并用核岭回归（KRR）算法对 746 个二元金属氧化物的总能量（Total energy）、密度（Density）、带隙等 5 个属性建立了

可靠的预测模型；Jha 等人[27]将其应用在材料发现领域在 OQMD 数据上用 One-hot 编码方式表征材料用 BP 神经网络算法建立了材料形成能量的预测模型，并在上亿种可能的材料空间（$A_xB_yC_zD_w$，其中 A、B、C 和 D 是元素周期表中的常见元素，x、y、z、w 为整数且$x+y+z+w\leq10$）搜索出了数百种可能存在的物质。原子向量(Atom2Vec)是斯坦福大学张首晟等人[44]最近的研究成果，通过矩阵的奇异值分解（SVD）制备的，Li 等人[34]将其用于超导材料的表征上并用 CNN-LSTM 算法建立了超导体 *Tc* 的预测模型。

常用的晶体结构表征方法有利用晶体几何学在每个原子上构建沃罗诺-狄利克雷多面体（VDP）[45,46]表征晶体结构和晶体指纹[47]表征晶体中原子的堆积状况及各个键之间的距离来研究物质的有序性和描述电子的相关性。Isayev 等人[46]用 VDP 表征法，在 AFLOWLIB 数据集上用深度神经网络构件了预测材料属性的通用学习模型，Xie 等人[48]用 VDP 构建晶体图后，首次将图卷积神经网络（GCN）应用到无机材料领域，在 MP 数据集上构建了能预测材料多种属性的晶体图卷积神经网络（CGCNN）模型。对于晶体指纹表征方法，Honrao 等人[49]采用晶体指纹编码晶体结构的相关物理信息，并将两种基于内核的 ML 算法应用到二元 Li-Ge 系统，并表明该方法在整个组成和结构空间上提供了约 20 meV/atom 较小的均方根预测误差；Zhu 等人[50]基于晶体指纹的度量标准，用于测量晶体结构的相似性。

1.2.3 预测模型研究现状

如图 1.1，发现新材料的最后一步是用材料属性的预测模型在材料中间中进行筛选，因此高性能的材料属性筛选模型对新材料发现的准确性至关重要。预测模型的建立主要包括两种，一种是基于机器学习的方法（ML），一种是基于深度学习（DL）的方法。

1）基于机器学习的方法

材料领域常用的机器学习算法包括决策树（DT）[51]、随机森林（RF）[33]、梯度提升决策树（GBDT）[52]、支持向量机（SVM）[53]和核岭回归（KRR）[54]等。DT 是一种树结构算法，其中每个内部节点表示对属性的判断，每个分支表示判断结果的输出，最后每个叶节点表示分类结果。RF 采用的是一种 Bagging 方法，包含多个决策树（DT）[55]模型。梯度提升决策树（GBDT）是 Friedman[56]提出的一种统计学习方法，是集成学习中的一种，它通过学习得到多个弱学习器并对其进行有效的线性组合增强为预测精度较高的强学习器，达到同时减少模型方差和偏差的效果。支持向量机（SVM）是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面，引入核方法后，支持向量机可以用来解决非线性问题。

KRR[54]是岭回归（L2 正则线性回归[57]），使用与支持向量机学习形式相同的核技术，但损失函数不同。

机器学习方法是一种非参数的模型，不需要像神经网络那样需要大量的数据来驱动参数的训练，因此机器学习方法特别适合用在样本少的数据上，很多材料学家将其应用到小数据集上。如，Chen 等人[58]针对晶格热导率（kL）采用理论计算在准确性和计算成本上的限制，使用实验测量的大约 100 种无机材料的 kL 数据，采用高斯过程回归算法，建立了一个预测无机材料的 kL 预测模型。Zhan 等人[59]使用三种不同的机器学习算法来预测热边界电阻，与传统的方法相比，ML 方法显示出更高的准确性。为了精确地预测不同材料的 Seeback 系数，Furmanchuk 等人[60]使用随机森林算法并获得了巨大的成功，预测非常准确，这意味着无需尝试合成或计算材料并获得其 Seeback 系数。Hamidieh 等人[61]使用 Magpie 材料表征方法，使用 GBDT 算法用 SuperCon 的 21263 种超导材料建立了 *Tc* 预测模型。总的来说 RF、GBDT 等基于树和集成学习的机器学习算法在材料特性预测模型的建立上使用的比较广泛。

2）基于深度学习的方法

深度学习算法可模仿人的大脑从经验学习中学习某些输入和输出之间的关系。神经网络的优势在于它的高性能，分层次的抽象模型可以从输入表示中学习，无需考虑输入和输出复杂的变换规律，而是转换为一组可训练的权值，理论上其可以接近任何一种非线性变换。在过去的 10 年中，它变得也来越有吸引力，历了飞速的发展，从最初的全连接网络（FNN），逐步发展出卷积神经网络（CNN）、长短期记忆神经网络（LSTM）、到现在的 Transformer 和图神经网络（GNN），其中卷积神经网络在计算材料领域使用的最为广泛。

卷积神经网络（CNN）通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征，表示后用“简单模型”即可完成复杂的分类或回归等学习任务[62]，其已被用于从材料的微观结构数据建立模型后改进表征方法[63-65]。Dong 等人[66]用卷积神经网络开发了一种能够使结构和带隙相关的材料描述符，通过从头计算的带隙以及相应的结构用作训练数据，在具有任意超级单元配置的硼氮杂化石墨烯材料中实现了带隙预测；Jha 等人[67]针对采用 DFT 数据训练的神经网络继承了实验测量与 DFT 计算差异的问题，证明了使用深度迁移学习，可以将现有的大型 DFT 计算数据集与其他较小的 DFT 计算数据集以及可用的实验观察结合在一起利用 CNN 建立稳健的预测模型；Konno 等人[68]根据超导体分子式中元素 *s*，*p*，*d* 和 *f* 轨道的电子数，提出了一种四通道材料表征方法，并使用 CNN 用 13000 种超导材料构建了临界温度预测模型。

尽管 CNN 在欧氏空间中的数据方面取得了巨大的成功，但在许多实际的应用场景中的数据是从非欧式空间生成的，同样需要进行有效的分析。例如，晶体中原子的连接关系，社会关系中人与人的社交关系。图数据的复杂性对现有的深度学习算法提

出了重大挑战，这是因为图数据是不规则的。每个图都有一个大小可变的无序节点，图中的每个节点都有不同数量的相邻节点，导致一些重要的操作（例如卷积）在图像上很容易计算，但不再适合直接用于图域。此外，现有深度学习算法的一个核心假设是实例彼此独立。然而，对于图数据来说，情况并非如此，图中的每个实例（节点）通过一些复杂的链接信息与其他实例（邻居）相关，这些信息可用于捕获结点之间的相互依赖关系。近年来，人们对深度学习方法在图数据上的扩展越来越感兴趣。在深度学习的成功推动下，研究人员借鉴了卷积网络、循环网络和深度自动编码器的思想，定义和设计了用于处理图数据的神经网络结构，由此图神经网络应运而生[69]。

图神经网络也开始应用到材料领域。Xie 等人[48]首次将图卷积神经网络（GCN）应用到无机材料领域，在 MP 数据集上构建了能预测材料多种属性的晶体图卷积神经网络（CGCNN）模型。之后 Chen 等人[70]对 CGCNN 进行了改进，开发出了通用的 MatErials 图网络模型，用于在分子和晶体中进行准确的属性预测，在约 60000 个晶体上训练的 MEGNet 模型在预测晶体的形成能量，带隙和弹性模量方面明显优于先前的 ML 模型，在更大的范围内，其性能优于密度泛函理论。Goodall 等人[71]开发了一种仅将化学计量作为输入并自动从数据中学习适当且可改进的描述符的神经网络方法，他们的主要见识是将化学计量公式视为元素之间的密集加权图，他们的方法在众多具有挑战性的材料性能上实现了更低的误差。与 CNN 相比，GNN 能从数据的邻接关系中自动的学会数据的表示，因此对材料数据表征的依赖较小。

1.3 研究中的问题与解决方法

由图 1.1，最后需要用建立好的精确预测模型在建立的假设材料空间中进行材料的筛选，假设材料空间的构建有两种方法，一种是通过穷举法，即：$A_xB_yC_zD_w$，其中 A、B、C 和 D 是元素周期表中的常见元素，x、y、z、w 为整数且$x+y+z+w\leq10$。如 Jha 等人[27]就是采用的这种做法，并用 ElemNet 进行了筛选。另一种是在现有的数据上进行筛选。如 Valen 等人[31]在 SuperCond 数据集上建立超导体 Tc 的预测模型后，在 ICSD 上进行超导体的筛选。但是，神经网络在做预测时，必须保证预测数据与训练数据分布的一致性。例如，能对人进行目标检测的网络，很难检查出动物；下围棋的神经网络不能用其来下象棋；能进行英文阅读理解的神经网络不能进行中文阅读理解等。而目前采用的在穷举材料空间或在 A 数据上训练网络在 B 数据上进行筛选的方式都不能保证训练材料与预测材料在数据分布上的一致性，这对预测模型的外插泛化能力提出了很高的要求（如图 1.3），而对最终会导致预测模型在筛选空间的表现不准确。

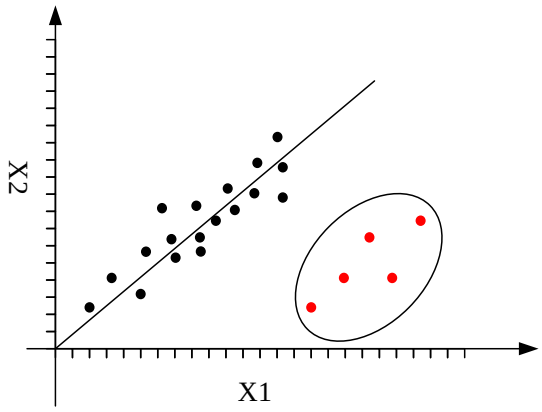


图 1.3 模型的外插泛化能力。模型对黑点的预测表现的非常好，而对分布不一样的红点其预测结果不可靠。

如 1.3.1 所述，在采用预测模型进行材料筛选时必须要保证预测数据与训练数据的一致。给定大量样本，生成对抗网络（GAN）能够学习生成训练数据的复杂隐藏规则，然后将这些学习的规则应用于具有目标属性的新样本。首先训练一个鉴别器，以区分真实样本和伪造样本，然后指导生成器的训练以减小这种差异。交替重复这两个训练过程，他们的军备竞赛将导致生成器和鉴别器的高性能。最终 GAN 的生成器学会了如何生成与训练样本分布规律一致的假样本。因此，我们可以用 GAN 生成数据的能力来构建假设材料空间，保证训练材料与筛选材料在分布规律的一致性。我们改进了现有的材料发现步骤，如图 1.4 所示。先用 GAN 生成与训练材料在原子组合规律一致的假设材料，再用预测模型在 GAN 构建的假设中进行筛选，以发发现有前途的新材料。

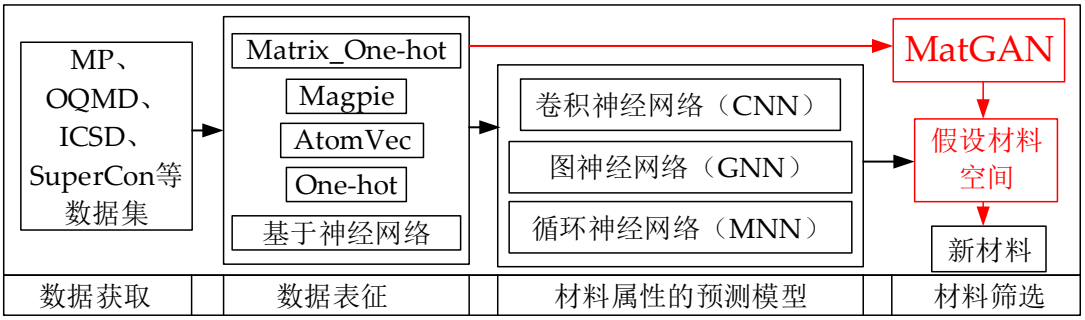


图 1.4 改进的新材料发现步骤

1.4 研究内容与创新性成果

论文以发现新的无机化合物为目标，根据材料表征的研究现状设计出了一种矩阵形式的 One-hot 材料表征方法；为解决新材在料筛选时分布的问题建立了在不施加任何化学规则的基础上能生成与训练材料在原子组合规律一致的假设材料的 GAN 模型；

为了便于层次式的特征提取，设计了一种无机化合物的层次式表征方法，通过设计特殊的卷积算子建立了一种高精度的材料属性卷积神经网络预测模型。

2）创新性成果

1、首次将生产对抗网络（GAN）应用在无机化合物新材料的发现上，解决了新材料在筛选时分布的问题，为无机化合物新材料的发现提供了新的思路。

2、建立的基于生成对抗网络（GAN）的生成深度学习模型（MatGAN），能有效地生成新的假设无机材料。当使用 ICSD 数据库中的材料进行训练，我们的 GAN 模型可以生成训练数据集中不存在的假设性材料，当生成 200 万个样本时，新颖性达到 92.53%。当使用 ICSD 的材料进行训练时，即使 MatGAN 模型中没有明确施加电荷中性和电负性平衡这样的化学规则，生成的假设材料化学有效样本中所占的百分比也达到了 84.5%。MatGAN 可用于加快无机材料的逆向设计或计算筛选。相关成果发表在 SCI 一区 TOP 期刊 npj Computational materials（影响因子 9.8）。

3、为了建立高精度的材料属性预测模型，设计了一种无机化合物的层次式表征方法，通过设计特殊的卷积算子和包含全卷积层的卷积神经网络建立了一种稳键的材料属性预测模型。相关成果发表在 SCI 二区期刊 IEEE Access（影响因子 4.098）

1.5 论文的组织结构

本论文首先系统的介绍了无机化合物的一些表征方法，接着指出了目前在新材料发现过程中无法保证筛选材料数据与训练数据在原子组合规律上的一致性，由此引入了采用生成对抗网络（GAN）构建假设材料空间的方法；其后，为了建立材料的筛选模型，基于卷积神经网络（CNN）训练出了材料属性的预测模型；然后，用预测模型在 GAN 构建的假设材料空间中筛选出了大量满足条件的材料，供研究者们进行后续的 DFT 仿真计算或实验制备；最后，对目前的工作进行了总结和展望，总结了我们工作的意义，指出了以后可能存在的研究方向。论文的组织结构如图 1.5 所示。

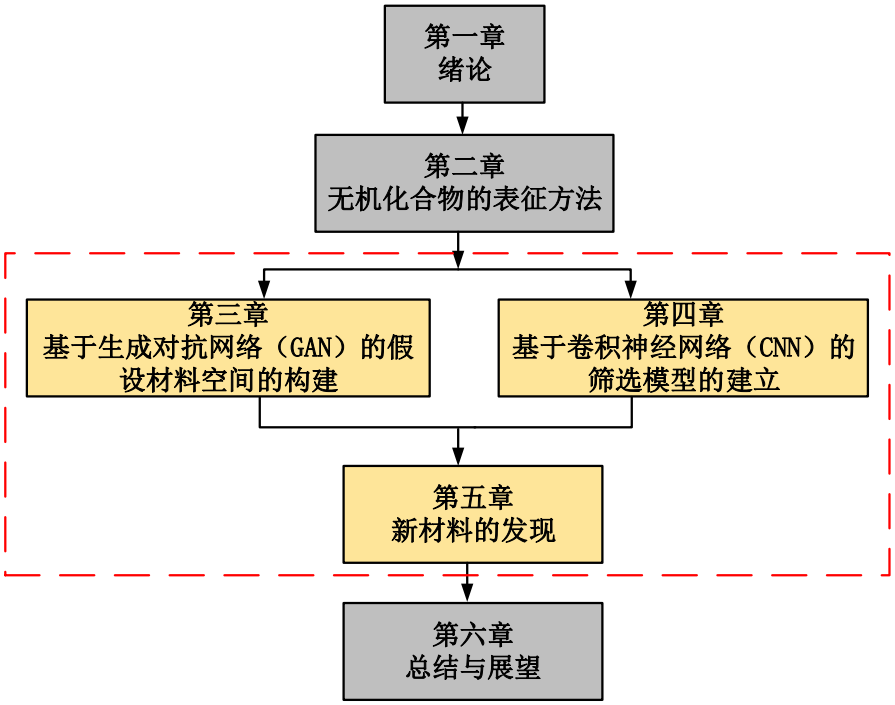


图 1.5 论文的组织结构

根据论文的研究内容及研究思路，论文分为 6 个章节。其余各章具体内容如下：

第二章介绍了无机材料的两类表征方法：基于分子式和基于晶体结构的材料表征方法。并分别介绍了这两类表征方法的优缺点和适用范围。

第三章提出了第一个无机化合物的生成对抗网络模型，通过生成假设的无机材料来有效采样无机材料的设计空间。用来自无机材料数据库（例如 OQMD[26,28]，MP[25]和 ICSD[5]）的材料进行训练后，MatGAN 模型能够从已知材料中学习隐式化学组成规则，从而生成假设但化学上合理的化合物。在没有明确规定化学规则的情况下，我们用 ICSD 子集的所有电荷中性和电负性平衡样本训练的 GAN 生成的假想材料，其具有 84.5%的电荷中性和平衡电负性。分析表明，与详尽的枚举方法相比，MatGAN 在采样无机材料的化学成分空间方面可以实现更高的效率。

第四章设计了一种材料的层次式材料表征方法，基于该表征方法我们提出了包含全卷积层的卷积神经网络材料属性预测模型。通过设计特殊的卷积算子，该模型能很好的从材料的原始输入矩阵中提取出有用的特征，进行最后的回归任务。利用 ICSD 中的数据进行不同任务的有监督训练，建立了能预测材料 Bandgap 的 ICSD-BG，能预测材料形成能量（Formation Energ）的 ICSD-FE 的稳健的预测模型。

第五章将 ICSD-FE 施加于 GAN-ICSD 生成的假设材料上进行筛选，筛选出形成能量较低的材料。在筛选出的形成能量较低的材料基础上，用 ICSD-BG 继续筛选带隙在 1.0~2.0eV 之间的材料。我们将筛选出的所有材料都放在了

http://github/danyabo/appendix.com 以供研究者们进行后续的 DFT 仿真计算或实验合成。

第六章 对论文的总结和展望

第 2 章 无机化合物的表征方法

数据表示的关键是选择用来表征材料的方法，它们是数据挖掘中输入的一部分。材料信息学的一个目的是建立描述符和目标属性之间的映射关系。因此，好的材料表征方法是有效的材料信息学的关键。一旦确定了一系列好的材料表征方法，就可以在数据库中进行内在或外在的最佳材料或性能预测的搜索。无机材料表征方法主要分为两类，一类是基于分子式的表征方法，一类是基于晶体结构的表征方法。基于分子式的表征方法只需以化学成分作为输入，这使得该表征方法能够探索化学组成空间的所有区域。基于结构的表征方法需要借助于我们对材料表示进行特征设计所需的所有领域知识。下面分别对两种表征方法进行介绍。

2.1 基于分子式的表征方法

基于分子式的表征方法只需化学成分作为输入。这使得该表征方法能够探索化学组成空间的所有区域。基于分子式的材料表征方法包括元素属性统计、One-hot 编码、Atom2Vec 编码等。

2.1.1 元素属性统计表征方法

元素属性统计法是指计算材料的一组元素统计属性，如，分子中元素在周期表上的周期数、组数，原子序数，原子半径，熔化温度，所有元素中来自 s、p、d 和 f 轨道的价电子的平均分数等。Magpie[32]是常用来计算元素属性统计表征方法的开源软件，表 1.1 表明了 Magpie 的 22 种属性的统计。

表 2.1 Magpie 的 22 种统计特征

特征类别	Magpie 特征
元素周期表中的位置统计特征	Number, MendeleevNumber, Column, Row
物理属性统计特征	AtomicWeight, MeltingT, CovalentRadius, Electronegativity
价电子特征	NsValence, NpValence, NdValence, NfValence, Nvalence, NsUnfilled, NpUnfilled, NdUnfilled, NfUnfilled, Nunfilled
原子的堆叠特征	Gsvolume_pa, Gsbandgap, Gsmagmom, SpaceGroupNumber

2.1.2 One-hot 编码表征法

One-hot 表征法是以分子式中元素的个数表征材料，表征向量对于化合物中存在的所有元素具有非零值，对于其他元素具有零值。考虑元素周期表中的所有元素[H,He,Li,...]组成一个固定长度的元素向量，化合物中存在的元素具有非零值且该值用元素在化合物中的数量表示，其他地方为零。

表 2.2 给出了 One-hot 编码的一个实例，我们以表格的形式考虑给定的(i = 1, 2, 3, ···, n)数据点的数据集，要预测的属性的数据点（或行）和预测变量。表格的每一行都包含分子式的数据。令 j 属于 J 表示索引的属性集，并且 p 属于 P 表示索引的预测变量集。

表 2.2 One-hot 编码实例

(a)属性($Y_{i,j}$)

i	Formula	Energy(eV)	Formation_Energy_Atom(eV)
1	Li2O	-14.264	-2.071
⋮	⋮	⋮	⋮
11	Na2O2	-8.401	-1.312
⋮	⋮	⋮	⋮
214	TiO2	-26.902	-3.512
⋮	⋮	⋮	⋮

(b)One-hot 表征($X_{i,p}$)

i	H	Li	Na	…	O	…
1		2	0	…	1	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮
11	0	0	2	…	2	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮
214	0	0	0	…	2	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.1.3 Atom2Vec 编码表征法

原子向量(Atom2Vec)是由斯坦福大学张首晟等人[44]首先提出来的。下面简要说明 Atom2Vec 的工作流程。如图 1.2 所示，为了捕获原子和环境之间的关系，作为第一步，为材料数据集中的每种化合物生成原子-环境对。在此之前，需要更明确的环境定义。原子可以方便地用化学符号表示。环境包括两个方面：化合物中目标原子的数量

和残留物中不同原子的数量。例如，我们从图 1 中给出的七个样品的微型数据集中考虑化合物 Bi2Se3。从 Bi2Se3 产生两个原子-环境对：对于原子 Bi，环境表示为“（2）Se3”；对于原子 Se，环境表示为“（3）Bi2。”具体而言，对于第一对，环境中的“（2）”（2）Se3 表示存在两个目标原子（此处为 Bi 的化合物），而“Se3”表示环境中存在三个 Se 原子。

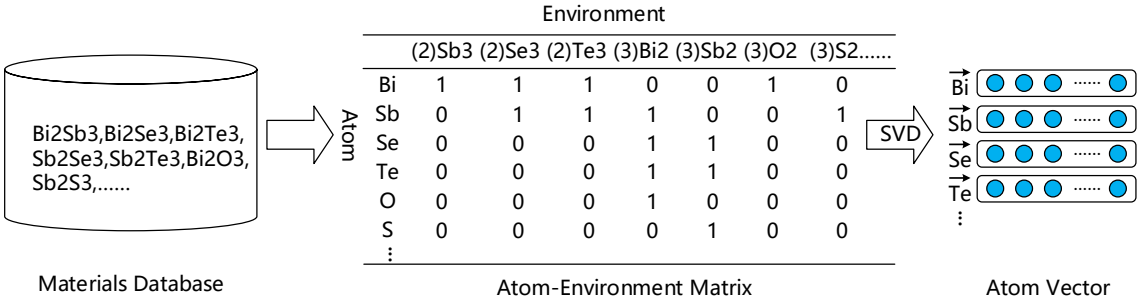


图 2.2 原子向量的生成方法

张首晟等人从 Materials Project 数据库[17]中获取了 73452 多种二元、三元、四元无机化合物，生成了一个稀疏的87×54032的原子环境矩阵，通过矩阵的奇异值分解（SVD）对原子环境矩阵对的列向量进行压缩，并取最大的 20 个奇异值对应的奇异向量，将每个元素描述成 20 维的向量。很显然性质相近的元素在原子环境矩阵中将有相似的行向量，生成相似的原子向量。

这里化合物 AxByCz…的表征，假设元素 A、B、C…原子向量分别为$\vec{A}$、$\vec{B}$ 、$\vec{C}$…可将输入的化合物表征成原子向量的并排，同时为了化合物中原子数量对材料性能的影响，在原子向量的最后加上相应的原子数，则该化合物表征为：V=$\left[\begin{bmatrix}\vec{A}\\x\end{bmatrix},\begin{bmatrix}\vec{B}\\y\end{bmatrix},\begin{bmatrix}\vec{C}\\z\end{bmatrix}...\right]$

其中 x、y、z…分别表示对应元素在超导化合物中的数量。

2.2 基于结构表征方法

基于结构的表征方法是指构建基于矢量的晶体结构数据表示，用来表示材料的晶体结构，常用的晶体结构表征方法有利用晶体几何学在每个原子上构建沃罗诺-狄利克雷多面体（VDP）[45,46]表征晶体结构和利用径向分布函数（RDF）[47]表征晶体中原子的堆积状况及各个键之间的距离来研究物质的有序性和描述电子的相关性。

2.2.1 沃罗诺-狄利克雷多面表征法

原子沃罗诺-狄利克雷多面体（VDP）是一个凸面多面体，其面垂直于连接 VDP 中心原子（VDP 原子）和其他（周围）原子的部分；每个面将相应的片段分成两半，所有原子的 VDP 形成晶体空间的（面对面）Voronoi-Dirichlet 分区[45,46]。

给定晶体结构，第一步是确定其中的原子连通性。通常情况下，原子连通性不是在材料内确定的微不足道的属性。不仅要考虑原子间潜在的键合距离，还要考虑附近原子的拓扑是否允许键合。因此，采用计算几何方法将晶体结构（图 2.3-a）分成原子中心的 Voronoi-Dirichlet 多面体[45]（图 2.3-b）。这种分配方案在金属有机骨架，分子和无机晶体的拓扑分析中是非常重要[72]。通过满足下列两个标准来建立原子之间的连通性：

(1)原子必须共享 Voronoi 面（相邻原子之间的垂直平分线）

(2)原子间距离必须短于 Cordero 共价半径的总和到 0.25 以内 Å 耐受性。

总之，Voronoi 中心共享 Voronoi 面，并且在它们的共价半径的总和内形成定义材料内连通性的三维图形。

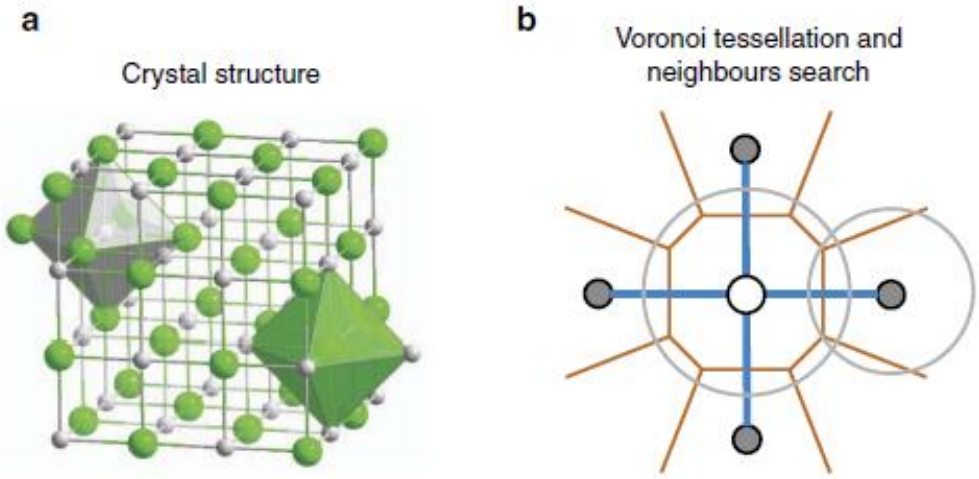


图 2.3 晶体内原子连通性确定步骤

2.2.2 晶体指纹表征法

材料研究人员在过去几年中研究的许多不同数据显示，包括对称函数，原子位置的平滑重叠[73]，径向基函数的傅里叶变换[74]以及其他特殊的数据表示。RDF 满足关于晶胞和晶体对称性的选择的不变性，以及使得两个晶体结构矢量表示x_1和x_2之间的能量差在极限$\|x_1\text{-}x_2\|$中变为零的连续性。

RDF 用$g_{AB}(r)$捕获在 A 和 B 型的原子 i 和 j 之间的平均距离:

$$d_{ij}^{AB}=\left|\vec{r}_i^{A}-\vec{r}_j^{B}\right| \quad 2\text{-}1$$

$$g_{AB}(r)=\frac{1}{N_A}\sum_{i=1}^{N_A}\sum_{j=1}^{\infty}\frac{1}{r^p}\exp\left[-\frac{(r-d_{ij}^{AB})^2}{2\sigma_g^2}\right]\Theta(d_c-d_{ij}^{AB}) \quad 2\text{-}2$$

第一个和是包括晶体内所有 A 型N_A原子，第二个和是 B 型的所有原子截止距离 dc 的和，由 Heaviside 函数（$\Theta(d_c-d_{ij}^{AB})$）强制给出。选择截止距离 dc，使其延伸超出晶体结构的晶胞，以确保数据表示捕获晶体结构的周期性。$1/r^p$项将部分 RDF 重新规范化为距离的函数，使得对于p=2，它在大距离处接近常数，并且对于p>2，它随距离衰减。图 2.4 显示了示例结构 Li4Ge 及其相应的部分 RDF。

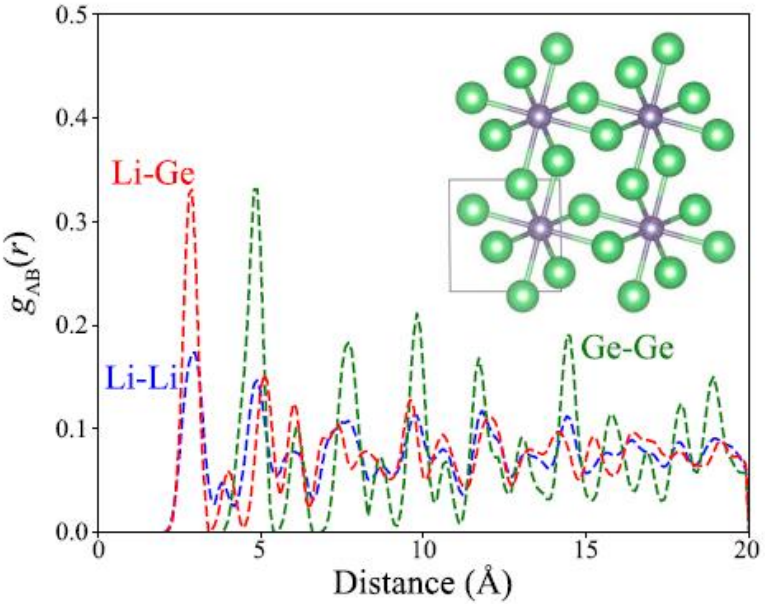


图 2.4 Li4Ge 的结构及其相应的部分 RDF

2.3 本章小结

本章介绍了无机材料的两类表征方法：基于分子式和基于晶体结构的材料表征方法。并分别介绍了这两类表征方法的优缺点和适用范围。在没有材料的晶体结构时只能采用基于分子式的表征方法，因其没有结构信息，因此预测模型往往需要大量的数据来构建。基于晶体结构的材料表征方法需要借助我们对材料表示进行特征设计所需的所有领域知识，其能较全面的表征材料，但在材料研究中有晶体结构的数据相比与只有分子式的数据要少的多。

第 3 章 基于生成对抗网络的假设材料空间的构建

3.1 引言

发现新的无机材料，如锂离子电池的陶瓷电解质材料，对于许多工业应用而言都是至关重要的。尽管近年来在合理的材料设计上已付出了巨大的努力，但由于寻找满足各种技术和经济限制的挑战，进展有限。从计算的角度来看，对于广阔的化学空间模筛选而言，蛮力分子模拟或第一性原理太昂贵。最近为量化多组分无机材料的组成空间的大小所做的努力[75]表明，即使在应用化学过滤器（如电荷中性[76]或电负性平衡[77]）后，四组分/元素材料的空间也超过10^{10}个组合，而五组件/元素空间超过10^{13}个组合。基于机器学习的模型确实已应用于筛选数十亿种假设材料以识别有前途的强离子导体[75]。潜在材料的数量是巨大的，考虑元素混合比例不同的掺杂材料的巨大空间以及许多应用（例如高温超导体），其中有六个到七种成分材料很常见。这种组合爆炸要求需要更有效的采样方法来搜索化学设计空间，该方法采用现有的明确的化学和物理知识以及在已知合成材料中体现的隐式元素组成知识。为了获得更有效的搜索，在计算筛选中已经使用了各种明确的化学规则来评估给定化学计量的可行性以及特定晶体排列的可能性，例如电荷中性（鲍林规则）、电负性平衡、半径比规则[78]，Pettifor匹配[79]等。但是，此类方法仍无法捕获足够的隐式化学规则，无法实现有效的化学设计空间采样。

最近，诸如自编码器（AE）及其变体（VAE，AAE），RNN，生成对抗网络（GAN）等生成式机器学习模型已成功应用于有机材料的逆向设计[80-83]。这些算法主要利用顺序算法或有机材料的图形表示形式，以学习用于生成有效和新颖的假设材料的构件的组成规则。给定大量样本，GAN 能够学习生成训练数据的复杂隐藏规则，然后将这些学习的规则应用于具有目标属性的新样本。应用到逆向设计中时，GAN 证明了其对设计空间进行有效采样的能力[80,84]，比其他采样方法（例如随机采样[85]，蒙特卡洛采样和其他启发式采样（例如遗传算法[86]））更有效。但是目前为止，由于构造块及组成规则的根本差异，这种生成式机器学习模型尚未应用于无机材料的生成。近来，已经提出了变体自编码器[86,87]以产生无机材料的假设晶体结构。但是，这些方法要么限于生成给定材料系统（例如 V-O 系统）的新结构[86]，要么不能生成物理上稳定的分子[87]。

在本章中，我们提出了第一个无机材料的生成对抗网络模型 MatGAN，通过生成假设的无机材料来有效采样无机材料的设计空间。用来自无机材料数据库（例如 OQMD[26,28]，MP[25]和 ICSD[5]）的材料进行训练后，我们的 GAN 模型能够从已知材料中学习隐式化学组成规则，从而生成假设但化学上合理的化合物。在没有明确规定化学规则的情况下，我们用 ICSD 子集的所有电荷中性和电负性平衡样本训练的 GAN 生成的假想材料，其具有 84.5%的电荷中性和平衡电负性。分析表明，与详尽的枚举方法相比，我们生成的 GAN 在采样无机材料的化学成分空间方面可以实现更高的效率。

3.2 MatGAN 的建立

3.2.1 数据与表征

1）实验数据

我们使用沉积在 OQMD[26,28]数据库中的无机材料子集来训练我们的 AE 和 GAN 模型。OQMD 是一种广泛使用的 DFT 数据库，其晶体结构可以通过高通量 DFT 计算得出，也可以从 ICSD[5]数据库获得，目前它有超过 600000 种化合物。我们使用与 Jha 等人类似的筛选标准[27]，选择用于 GAN 训练的 OQMD 子集：对于具有多个报告的形成能的分子式，我们保留最低的一个以选择最稳定的化合物。所有单元素化合物都会与形成能不在 u±5σ 范围内的材料一起去除，其中 u 和 σ 是 OQMD 中所有样品形成能的平均值和标准偏差。最终，数据集具有 291884 种化合物。

为了比较，我们还分别为材料项目（MP）和 ICSD 训练了两个 GAN。 此处的 MP 数据集和 ICSD 都是通过除去所有单原子化合物和晶胞中具有 8 个以上原子的化合物和含有 Kr 和 He 元素的化合物而制备的。使用的最终 MP 数据集包含 63922 种化合物。使用的最终 ICSD 数据集包含 28137 种化合物。使用的 OQMD、ICSD、MP 的数据分布如图 3.1 所示。

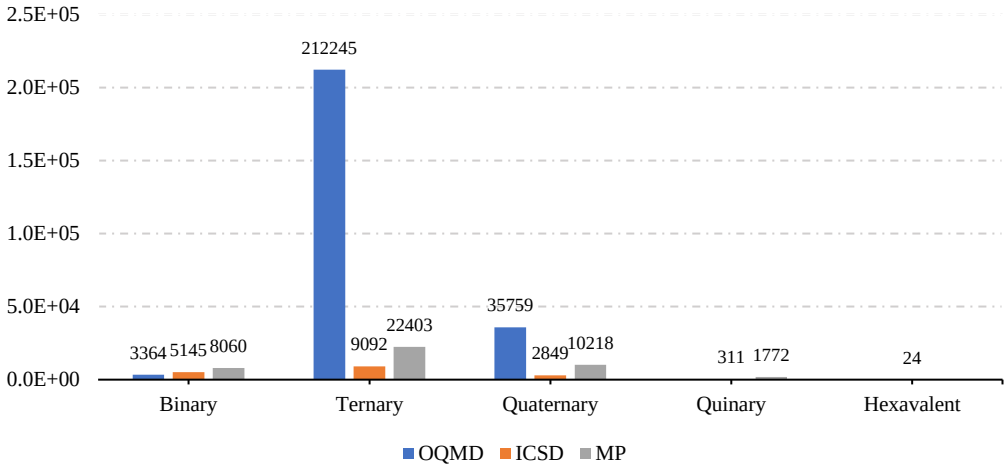


图 3.1 使用的 OQMD、ICSD、MP 的数据分布

2）数据表征

通过对 OQMD 筛选的材料进行简单的统计计算，在元素周期表中的 118 种元素中，找到了 85 种元素，在任何特定的化合物/公式中，每个元素通常小于 8 个原子。然后将每种材料表示为稀疏矩阵$X^r \in R^{n\times m}$，n= 85，m=8。矩阵有 0/1 单元格值，每个列表示 85 个元素中的一个，而列向量则是该特定元素的原子数的一个热编码（如图 3.2）。

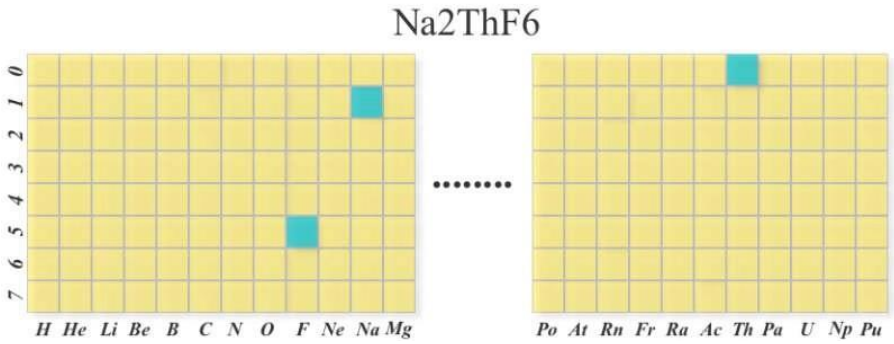


图 3.2 Na2ThF6 的表征示意图

3.2.2 MatGAN 的结构

生成模型可基于多种机器学习算法构建，例如变分自编码器（VAE），生成对抗网络（GAN），强化学习（RL）[88]，递归神经网络（RNN）及其混合[80]。与其他生成模型[89,90]不同，GAN 并不直接使用数据和模型分布的差异来训练生成器。取而代之的是，采用对抗训练方法：首先训练一个鉴别器，以区分真实样本和伪造样本，然后指导生成器的训练以减小这种差异。交替重复这两个训练过程。他们的军备竞赛将导致生成器和鉴别器的高性能。

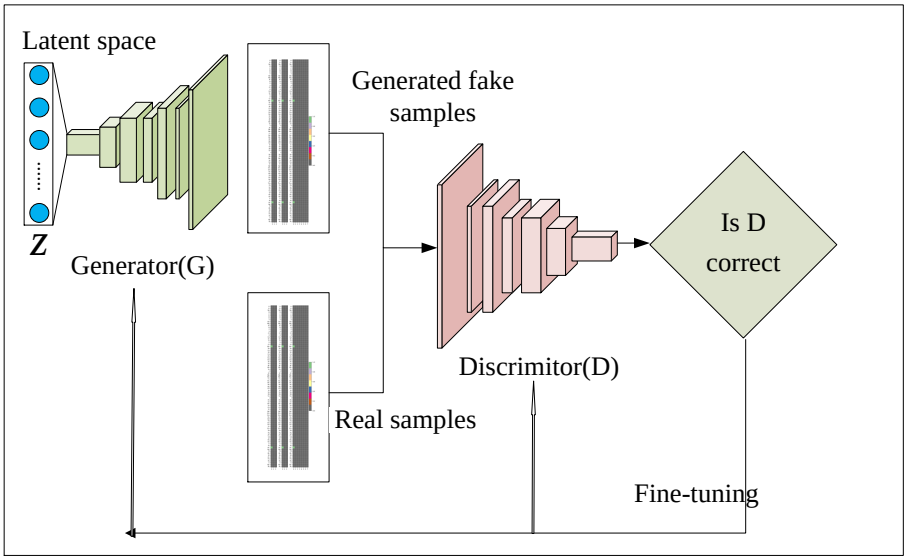


图 3.3 无机材料的 MatGAN 结构

我们的无机材料生成 DL 模型 MatGAN 基于 GAN 方案，如图 3.3 所示。它由将随机向量映射到生成的样本的生成器和试图区分实际材质与生成的假设的鉴别器组成。模型的详细的配置参数在表 3.1 和表 3.2 中列出。我们选择材料样本的 8×85 矩阵表示来构建 GAN 模型。

表 3.1 GAN-OQMD 的配置参数

Model	Layer	Input Shape	Filter	Kernel Size	Stride
Generator	Fc1	[batch, 128]	-	-	-
	Reshape	[batch, 6×4×256]	-	-	-
	DeConv1	[batch, 6, 4, 256]	128	(5, 5, 256)	(2, 2)
	DeConv2	[batch, 11, 8, 128]	128	(3, 3, 128)	(2, 1)
	DeConv3	[batch, 22, 8, 128]	64	(3, 3, 128)	(1, 1)
	DeConv4	[batch, 22, 8, 64]	64	(3, 3, 64)	(2, 1)
	DeConv5	[batch, 43, 8, 64]	32	(3, 3, 64)	(1, 1)
	DeConv6	[batch, 43, 8, 32]	32	(3, 3, 32)	(2, 1)
	DeConv7	[batch, 85, 8, 32]	1	(3, 3, 32)	(1, 1)
Discriminator	Conv1	[batch, 85, 8, 1]	32	(3, 3, 1)	(1, 1)
	Conv2	[batch, 85, 8, 32]	32	(3, 3, 32)	(2, 1)
	Conv3	[batch, 43, 8, 32]	64	(3, 3, 32)	(1, 1)
	Conv4	[batch, 43, 8, 64]	64	(3, 3, 64)	(2, 1)
	Conv5	[batch, 22, 8, 64]	128	(3, 3, 64)	(1, 1)
	Conv6	[batch, 22, 8, 128]	128	(3, 3, 128)	(2, 1)
	Conv7	[batch, 11, 8, 128]	256	(5, 5, 128)	(2, 2)
	Reshape	[batch, 6, 4, 256]	-	-	-
	Fc1	[batch, 6×4×256]	-	-	-

表 3.2 GAN-MP 和 GAN-ICSD 的配置参数

Model	Layer	Input Shape	Filter	Kernel Size	Stride
Generator	Fc1	[batch, 128]	-	-	-
	Reshape	[batch, 6×4×128]	-	-	-
	DeConv1	[batch, 6, 4, 128]	64	(5, 5, 128)	(2, 2)
	DeConv2	[batch, 11, 8, 64]	32	(3, 3, 64)	(2, 1)
	DeConv3	[batch, 22, 8, 32]	16	(3, 3, 32)	(2, 1)
	DeConv4	[batch, 43, 8, 16]	1	(3, 3, 16)	(2, 1)
Discriminator	Conv1	[batch, 85, 8, 1]	16	(3, 3, 1)	(2, 1)
	Conv2	[batch, 43, 8, 16]	32	(3, 3, 16)	(2, 1)
	Conv3	[batch, 22, 8, 32]	64	(3, 3, 32)	(2, 1)
	Conv4	[batch, 11, 8, 64]	128	(5, 5, 64)	(2, 2)
	Reshape	[batch, 6, 4, 128]	-	-	-
	Fc1	[batch, 6×4×128]	-	-	-

我们发现材料的整数表示极大地方便了 GAN 训练。在我们的 GAN 模型中，鉴别器（D）和生成器（G）均被建模为深度神经网络。G 由 1 个完全连接的层和 7 个反卷积层组成。D 由七个卷积层和一个完全连接的层组成。卷积和反卷积层中的每一个都带有批处理规范化层[91]。生成器的输出层将 Sigmoid 函数用作激活函数，而所有其他批处理归一化层将 ReLu[92]用作激活函数。表 3.1 和 3.2 中显示了详细的网络配置。为了避免标准 GAN 的梯度消失问题，我们采用 Wasserstein GAN[85]，它用 Wasserstein 距离代替了 JS 散度。将使用 Wasserstein GAN 方法训练 GAN 模型，方法是将 G 和 D 损耗减到最小，G 和 D 的损失函数定义如下：

$$\mathrm{Loss_G} = -E_{x\sim P_g}[f_w(x)] \tag{3-1}$$

$$\mathrm{Loss_D} = E_{x\sim P_g}[f_w(x)] - E_{x\sim p_r}[f_w(x)] \tag{3-2}$$

其中，P_g是生成的样本和实际样本的分布； $f_w(x)$是判别网络。公式 3-1 和 3-2 用于指导训练过程。$\mathrm{Loss_D}$越小，生成的样本和真实样本之间的 Wasserstein 距离越小，并且 GAN 训练得越好。

3.2.3 MatGAN 的训练

我们通过将学习率从 0.1 设置为10^{-6}（每次减少 10 倍），将批量归一化大小从 32 设置为 1024，并使用了不同的优化器，优化了用于训练 GAN 的超参数。 我们使用来自 OQMD，MP 和 ICSD 的筛选样本以及 ICSD-filter（这是 ICSD 的子集，具有所有电荷中性和电负性平衡材料）数据库训练 GAN。这些 Wasserstein GAN 用 Adam 优化器进行了 1000 个批次的训练，生成器训练的学习率为 0.001，鉴别器训练的学习率

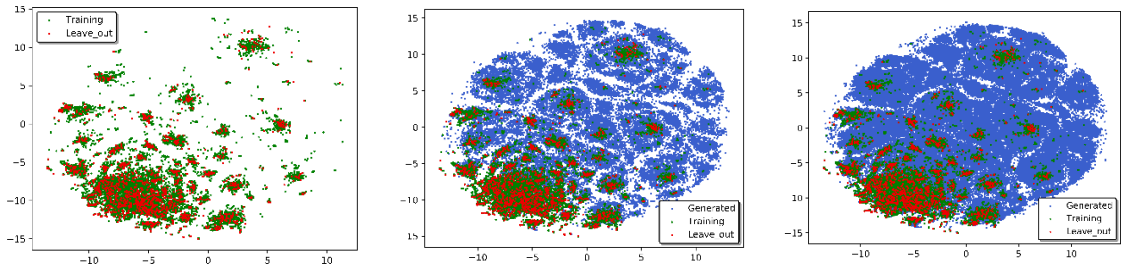
为 0.01。OQMD 上的 GAN 训练的批次大小设置为 512，而其他所有数据集上的 GAN 训练的批次大小设置为 32。使用 Adam 优化算法训练 AE，学习速率为10^{-3}，批处理大小为 1024。

3.3 MatGAN 的性能评估

我们按照 3.2 中详细介绍的程序训练了 GAN。然后，对于所有这些 GAN（包括 GAN-OQMD，GAN-MP，GAN-ICSD），我们使用每个生成器均生成了 200 万个假设材料，并评估了它们的有效性，唯一性和新颖性。

3.3.1 映射无机材料设计空间

在 GAN-ICSD 生成的 200 万个样本中，我们过滤掉不满足电荷中性和平衡电负性的样本，从而生成了 169 万个样本。为了可视化与 ICSD 的训练数据集相比生成的数据分布，我们应用了 T-sne 维度缩减技术[93]来减少生成集、训练集和 Leave-out 验证集样本的矩阵维数。生成的样本相对于训练和验证集的分布如图 3.4 所示。可以观察到，来自 ICSD 的训练样本仅占整个空间的小部分。但是，GAN-ICSD 已经能够生成可能有趣的假设材料，这些材料填补了设计空间，这可能会大大扩展 ICSD 数据库的范围。



(a) (b) (c)

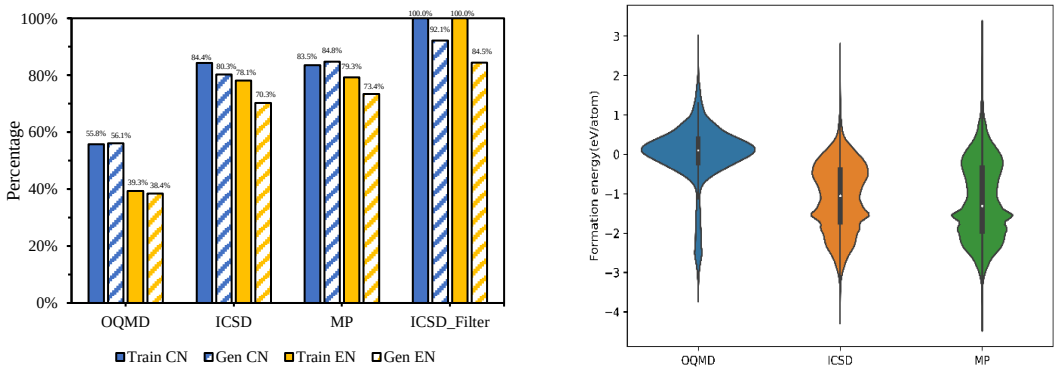
图 3.4 由现有的 ICSD 材料和 GAN-ICSD 生成的假设材料组成的无机材料空间

在基于 T-sne 的尺寸缩减之后，两个轴对应于两个尺寸。 ICSD 材料仅占无机材料化学空间的小部分。3.4-a 是 ICSD 的培训样本（绿点）和验证样本（红点）； 3.4-b 是 50000 个生成的样本（蓝点）以及训练和验证样本；3.4-c 是 200000 个生成的样本以及训练和验证样本。

3.3.2 生成材料的检查

1) 电荷中性和电负性平衡检查

电荷中性和电负性平衡是晶体的两个基本化学规则。因此在没有明确施加这些规则的条件下，有趣的是检查我们的 GAN 模型生成的样本如何满足这些规则。为此，我们采用了文献[75]中提出的电荷中性和电负性检查程序，计算在训练中遵守这些规则的样本百分比，结果示于图 3.4-a。首先，我们发现有效生成的样本的百分比与训练集的百分比非常接近。对于 OQMD，当训练集具有 55.8%的电荷中性样本时，生成集将具有 56.1%。对于 MP 和 ICSD，生成的电荷中性样本的百分比（分别为 84.8%和 80.3%）也接近训练集的百分比（分别为 83.5%和 84.4%）。在电负性检查中发现了类似的观察结果。印象深刻的是，尽管在我们的 GAN 训练模型中没有明确建模或强制执行这些规则，当我们确保 ICSD-filter 中的所有训练样本均达到电荷中性和电负性平衡时，分别生成的样本中分别有 92.1%和 84.5%满足这两个化学规则。为了证明这一高百分比的化学有效候选物的重要性，我们将我们的结果与文献[75]表 1 中的穷举枚举方法进行了比较。穷举计算时，同时满足电荷中性和电负性的所有二元/三元/四元样品的百分比为 0.78%，而我们的为 62.24%，这相当于采样效率的 77 倍。这有力地表明，我们的 GAN 模型已成功学习了用于生成化学上有效的假设材料的隐式化学规则。



(a) (b)

图 3.4 生成材料有效性的评估 （a）生成的样本中的电荷中性（CN）和电负性平衡（EN）样本的百分比与所有四个数据集的训练集的百分比非常接近。 Train / gen CN：满足电荷中性的训练/生成样本的百分比；Train / gen EN：满足平衡电负性的样品百分比。（b）由三个 GAN 生成的含锂化合物的形成能分布。GAN-ICSD 和 GAN-MP 都可以生成很大一部分具有低形成能的假想材料。

2）稳定性检查

评估生成的假设材料质量的另一种方法是检查其稳定性，这可以通过形成能量来衡量[94]。为此，首先，使用 GAN-OQMD，GAN-ICSD 和 GAN-MP 分别生成 200 万个候选材料。然后，我们选择了所有含锂元素的材料，再除去了所有不满足电荷中性和平衡电负性的材料。最后，我们分别从 GAN-OQMD，GAN-ICSD 和 GAN-MP 获得了 15591、137948 和 281320 含锂化合物。接着，我们下载了由 Jha 等人开发的形成

能量预测机器学习模型（ElemNet）[27]，用它来预测所有这些假设材料的形成能。图 3.4-b 表明，这些生成的材料的形成能大多小于 0eV/Atom，特别是对于由 GAN-ICSD 和 GAN-MP 生成的材料，它们的化学性质更强。此外，在图中发现 GAN-OQMD 生成的样本的百分比更高，具有更高的形成能，这是由于 68.48％的 OQMD 训练样本具有大于 0 eV/Atom 的形成能这一事实。

3）唯一性检查

为了检查生成样本的唯一性，我们分别在 OQMD、MP 和 ICSD 数据集上训练的三个 GAN 生成的样本（n）从 1 变为 340000 计算生成的样本中唯一样本的百分比（图 3.5）首先，可以发现随着生成的样本 越来越多，唯一样本的百分比下降，这表明生成新的假设材料更加困难。但是，即使生成了 340000 个样本，我们的 GAN 仍分别对 GAN-OQMD，GAN-MP 和 GAN-ICSD 保持 68.09％，85.90％和 73.06％的唯一性。尽管所有三个曲线都随着生成的样本数量的增加而衰减，但 GAN-MP 保留了更高百分比的独特样本。实际上，GAN-MP 的唯一性曲线主导了 GAN-ICSD，而 GAN-ICSD 又主导了 GAN-OQMD。在仔细检查训练和生成样本的元素数量方面的分布之后，我们发现这主要是由于三个 GAN 的训练集的分布偏差。对于 GAN-OQMD，训练集主要由三元化合物（84.4％）组成，并且倾向于生成三元样本，而 SMACT[75]根据类比和化学理论的半导体材料估计的化学有效材料的总数大约为 200000。因此，它倾向于生成许多重复的三元样本。对于 GAN-ICSD，二元/三元/四元之比约为 2：3：1，这使其可以生成更多样的样本，从而获得更高的唯一性曲线。对于 GAN-MP，二元/三元/四元之比约为 0.8：2：1，这比 GAN-OQMD 和 GAN-ICSD 的均衡得多，并且还具有更多的四元和五元训练样本。这使其可以生成最多样化的样本。

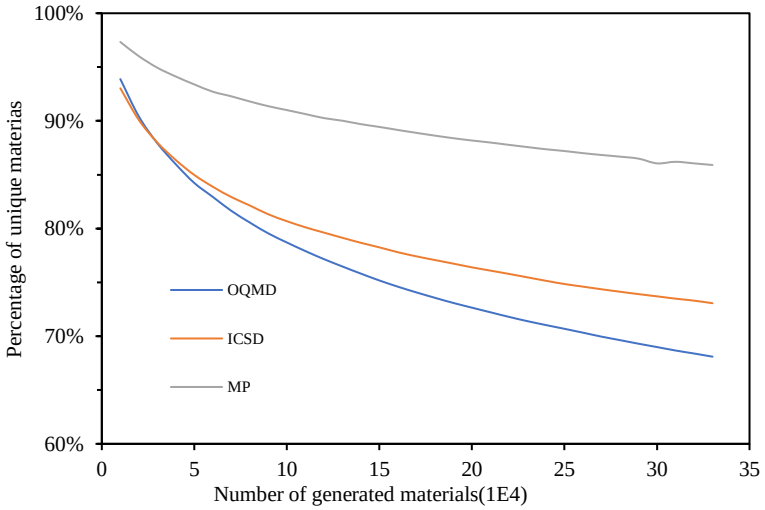


图 3.5 生成材料的唯一性检查。比较三种 GAN 生成的假设材料的唯一性曲线。 GAN-MP 由于其二元/三元/四元训练样本的分布更加均衡而达到了主导曲线。

4）生成材料新颖性检查

为了检查 GAN 生成新颖材料的能力，我们使用了保留验证方法。我们首先将 QMD、MP 和 ICSD 这三个数据集都随机删除 10％的样本，再训练 GAN，并使用它们生成一定数量的样本。然后，我们检查已恢复/重新发现训练样本和保留验证样本的百分比以及已生成了多少新样本。结果显示在表 3.3 中。首先，我们发现，当 GAN 生成一定百分比的训练样本时，还可以覆盖大量的验证样本。例如，当 GAN-MP 恢复其训练集的 47.36％时，还同时生成了约 48.82％的保留样本。这表明我们的 GAN 可以用于发现训练集中不存在的新材料。

表 3.3 GAN 对生成的样本进行的新颖性检查。

	GAN-OQMD	GAN-MP	GAN-ICSD
Training sample	251368	57530	25323
Leave out sample	27929	6392	2813
Generated sample	2000000	2000000	2000000
Recovery % of training samples	60.26%	47.36%	59.54%
Recovery % of validation samples	60.43%	48.82%	60.13%
New samples	1831648	1969633	1983231

为了进一步了解生成性能，我们计算了训练集和遗漏验证集的回收百分比，以及二元，三元和四元样本的新样本的百分比（图 3.6）。首先，GAN-ICSD 通过生成 200 万个样本，生成了 78.1％的训练二元样本，同时还生成/重新发现了 82.7％的遗漏验证二元材料。三元训练和验证样本的回收率分别下降到 30.4％和 31.2％，因为可能的三元样本数量大于二元样本，这也解释了四元训练和验证集的回收率下降到 3.3％和 5.2％。此外，在所有生成的二元/三元/四元样本中，其中 83.15％/ 98.68％/ 99.98％是新颖的假设材料，这强烈表明我们的 GAN 模型具有生成新材料候选物的能力，这些新材料中的大多数候选人符合基本化学规则，如图 3.6 所示。

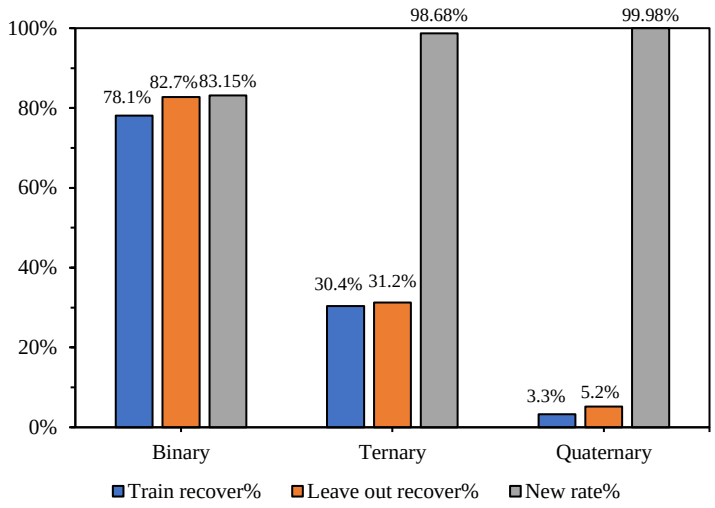


图 3.6 生成材料的新颖性检查。分配训练和验证样品的回收率，以及新产生的假设材料的百分比。

3.3.3 条件生成假设材料

除了生成有效的无机材料外，有趣的是，通过从模型估计的生成分布中进行采样，检查我们的 GAN 模型是否可以生成具有所需属性的新材料[95]。为了验证这一点，我们从 MP 中收集了 30186 个带隙值大于 0 的无机材料。然后，我们使用这些高带隙材料集来训练 GAN-Bandgap 模型，以生成假设的高带隙材料。为了验证所生成样本的带隙值，我们使用具有 Magpie 特征的梯度提升决策树（GBDT）机器学习算法训练了带隙预测模型[32]。学习率设置为 0.06，最大树深度设置为 20，采样率设置为 0.4，估计器的数量设置为 100。

我们使用此模型来预测详尽列举的材料集的带隙值。图 3.7 给出了所产生的材料组与训练组和枚举组的带隙的分布。生成样本的带隙分布与训练集的带隙分布非常相似，这表明我们的 GAN-Bandgap 能够有效地生成假设的高带隙材料。

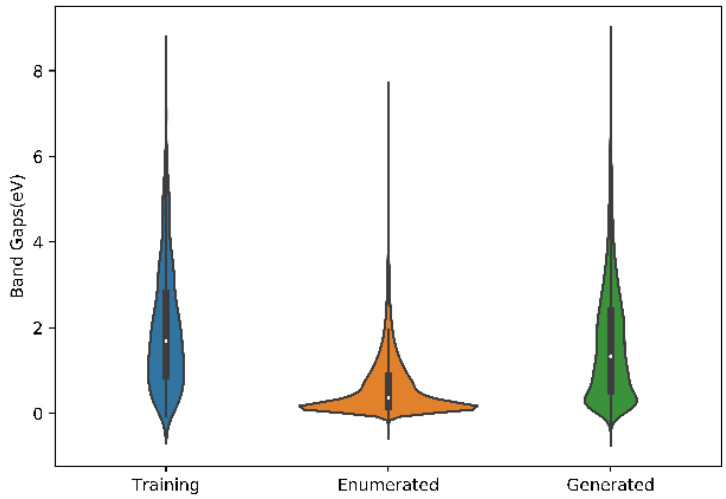


图 3.7 GAN-Bandgap，训练集和枚举集对生成材料的带隙分布的比较

3.3.4 发现潜在的新材料

为了评估我们的 GAN 模型生成已确认新材料的可能性，我们采用了交叉验证评估方法。从本质上讲，对于每个 GAN 模型生成的所有新假设材料，我们都会检查另外两个数据集是否确认/包含了其中的多少。表 3.4 列出了交叉验证确认结果。发现在 GAN-ICSD 生成的 200 万种材料中，MP 数据集确认并包含了 13126 种材料，OQMD 数据集确认了 2349 种新材料。GAN-MP 还具有 ICSD 和 OQMD 分别确认的 6880 和 3601 个生成的样本。

表 3.4 GAN 对生成的新材料进行交叉验证

	ICSD dataset	MP dataset	OQMD dataset
GAN-ICSD	N/A	13126	2349
GAN-MP	6880	N/A	3601
GAN-OQMD	3428	58603	N/A

3.4 MatGAN 的局限性检查

3.4.1 自编码器的建立

在 OQMD 数据集的 GAN 生成实验中，我们发现有时难以生成特定类别的材料。这可能是由于有限的样本学习所需的组成规则所致的。为了研究这个问题，我们建立了一个自编码器（AE）[96]模型，如图 3.8 所示。自编码器由一个编码器和一个解码器组成，该编码器具有七个卷积层，其后是一个全连接层，而解码器是由一个全连接层，

其后是七个反卷积层。在每个卷积和反卷积层之后，都有一个批处理归一化层，用于加速训练并减少初始网络权重的影响[91]。ReLu 用作所有批次归一化层的激活功能。Sigmoid 函数用作解码器输出层的激活函数。详细的配置参数在表 3.5 中列出。

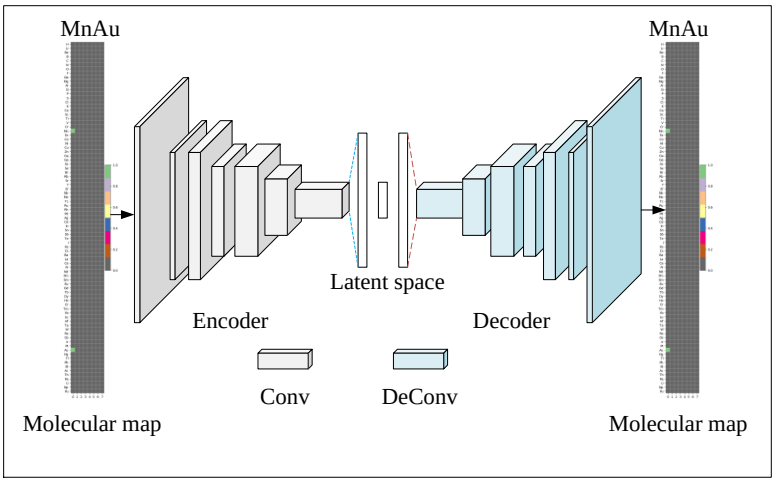


图 3.8 自动编码器的架构

使用从 OQMD 数据库中选择的 291840 种无机材料对自动编码器进行了训练。为了尽可能保证原始输入矩阵X^r和解码器重构的矩阵之间的重叠，我们采用医学图像语义分割中常用的 Dice 系数[97]作为 AE 的损失函数。然后使用反向传播算法训练 AE 模型。损失函数如下式所示：

$$Loss_{AE} = -\,Dice = -\frac{2|A \cap B|}{|A|+|B|} \approx -\frac{2 \times A \bullet B}{Sum(A)+Sum(B)} \qquad 3\text{-}3$$

其中|A∩B|表示矩阵 A 和 B 之间的共同元素，|•|表示矩阵中元素的个数，•是矩阵的点乘，Sum(•)是矩阵中所有元素的和。Dice 系数本质上是衡量两个样本的重叠部分，该指标范围从 1 到 0，其中 1 表示重叠[98]。

AE 模型的解码器模块与 MatGAN 模型中的发生器具有相同的体系结构。 我们的假设是，如果训练有素的 AE 模型无法解码特定材料，则我们的 GAN 模型不太可能生成它。 通过使用 AE 从 OQMD 数据库中筛选出不可解码的物质，我们可以更深入地了解 GAN 模型的局限性。

表 3.5 AE 的参数配置

Model	Layer	Input Shape	Filter	Kernel Size	Stride
Encoder	Conv1	[batch, 85, 8, 1]	32	(3, 3, 1)	(1, 1)
	Conv2	[batch, 85, 8, 32]	32	(3, 3, 32)	(2, 1)
	Conv3	[batch, 43, 8, 32]	64	(3, 3, 32)	(1, 1)
	Conv4	[batch, 43, 8, 64]	64	(3, 3, 64)	(2, 1)
	Conv5	[batch, 22, 8, 64]	128	(3, 3, 64)	(1, 1)
	Conv6	[batch, 22, 8, 128]	128	(3, 3, 128)	(2, 1)
	Conv7	[batch, 11, 8, 128]	256	(5, 5, 128)	(2, 2)
	Reshape	[batch, 5, 4, 256]	-	-	-
	Fc1	[batch, 5×4×256]	-	-	-
Decoder	Fc1	[batch, 128]	-	-	-
	Reshape	[batch, 5×4×256]	-	-	-
	DeConv1	[batch, 5, 4, 256]	128	(5, 5, 256)	(2, 2)
	DeConv2	[batch, 11, 8, 128]	128	(3, 3, 128)	(2, 1)
	DeConv3	[batch, 22, 8, 128]	64	(3, 3, 128)	(1, 1)
	DeConv4	[batch, 22, 8, 64]	64	(3, 3, 64)	(2, 1)
	DeConv5	[batch, 43, 8, 64]	32	(3, 3, 64)	(1, 1)
	DeConv6	[batch, 43, 8, 32]	32	(3, 3, 32)	(2, 1)
	DeConv7	[batch, 85, 8, 32]	1	(3, 3, 32)	(1, 1)

3.4.2 MatGAN 的局限性

在这里，我们旨在检查 AE 不可解码材料的关系以及 GAN 生成它们的难度。 为了训练 AE 模型，我们将 OQMD 数据集随机分成 90％的样本用于 AE 训练和 10％的样本作为测试。学习率设置为10^{-3}，批处理大小为 1024，并使用 Adam 优化器。 最终的 AE 模型被选为在 1000 个训练周期内在测试集中具有最佳性能的模型。我们发现，AE 模型可以从训练集和测试集中解码出 96.31％和 95.50％的样本。这些样品似乎具有一些共同的化学组成规则。

为了显示可解码样本与不可解码样本之间的差异，我们应用了 T-sne 维度缩减技术[93]将所有 OQMD 数据集的矩阵表示维缩减为 2，然后在 2D 上可视化 20％的样本图（图 3.9），其中红点代表不可解码的样本，蓝色点代表可解码的样本。明显不同的分布表明，这两类样品具有不同的组成规则。我们的假设是，可解码样本共享完善的化学成分规则，这使我们的 GAN 生成器可以有效采样相应的化学空间。另一方面，我们的 GAN 模型将难以生成不可降解的样本。为了验证这一点，我们计算了受过训练的 GAN-OQMD 生成的不可降解样品的百分比。可以观察到，即使在生成了 200 万

个样本后，几乎 95%的不可解码的材料超出了生成样本的范围，而 60.26%的可解码训练样本已被重新发现。

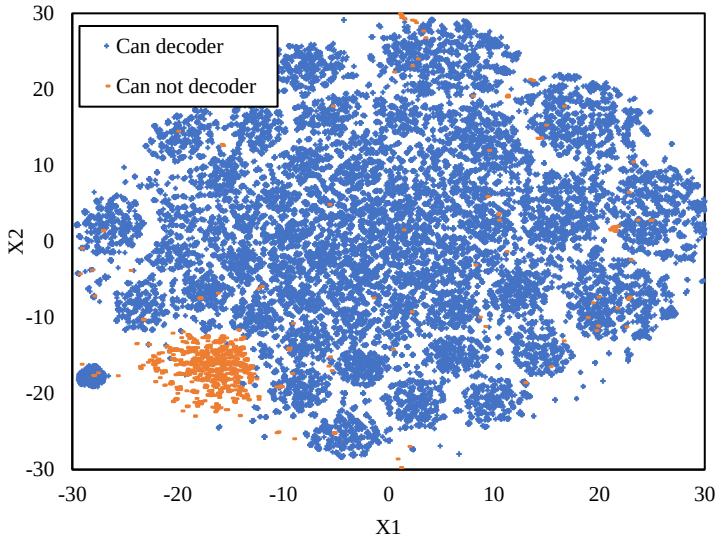


图 3.9 可解码和不可解码材料的分布

这表明我们的 GAN 在生成不可解码材料类型方面存在局限性。也意味着不可分解的材料具有特殊的组成规则，这些规则需要更多的数据或更强大的生成器模型来学习。

3.5 本章小结

新无机材料的构型相空间很大。由周期表的前 103 个元素形成四组分化合物会产生10^{12}种以上的组合。如此巨大的材料设计空间对于高通量实验或第一性原理计算来说是难以企及的。另一方面，当前的无机材料数据库（例如 ICSD 和 MP）都仅占整个无机化学空间的一小部分，需要扩展以进行新材料的计算筛选。

在这里，我们提出了一种基于 GAN 的生成模型，可以有效地采样大量无机材料的化学设计空间。系统的实验和验证表明，我们的 GAN 模型在生成能力方面可以实现高度的唯一性，有效性和多样性。通过扩展 ICDS，材料项目（MP）和 OQMD，我们的生成模型可用于探索未知的无机材料设计空间。与彻底筛选数十亿个候选对象相比，导出的扩展数据库可用于更高效率的高通量计算筛选[99]。尽管已采用电荷中性和电负性平衡原理[75]来过滤化学上难以置信的成分，以便更有效地搜索新材料，但此类明确的成分规则仍然过于宽松，无法确保在广阔的化学设计空间中对新材料进行有效采样。虽然可以列举出少于 5 种元素的假设材料（对于具有电荷中性和平衡电负性的 4 元素材料而言为 320 亿种材料），但是更多元素的设计空间可能具有挑战性，而我们的 GAN 模型可以提供很大帮助。

第 4 章 基于卷积神经网络的筛选模型的建立

4.1 引言

了解材料属性与材料化学和结构之间的关系在理论和实验上都提出了重大挑战。但它的基本物理规律将以数据的形式呈现出来，因此我们可以利用大量现成的相关信息，以最简单的形式计算出每种材料的组成特征。这些收集的变量被开发/训练并用于预测宏观性质。尽管缺乏空间群、电子结构和声子能量等相关性质的信息，但具有这些基本特征的模型仍然具有惊人的准确性。因为成分相似性和属性相似性之间的关系也可以用来建立预测模型，这就是为什么机器学习方法仅利用构图特征成功地应用于材料属性预测问题。卷积神经网络（CNN）通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征，表示后用“简单模型”即可完成复杂的分类或回归等学习任务[35]，其已被用于从材料的微观结构数据建立模型后改进表征方法[63-65]，已证明，卷积神经网络可用于预测晶体结构和分子的性质。

化合物的形成能量（Formation Energy）可以评估材料的稳定性，因此检查某种假设材料存在的可能性的一种方式是计算其形成能量。但是，材料的形成能量只能反映某种材料能够稳定存在的可能，但研究者们往往希望发现的材料拥有某种特性，如半导体材料。半导体材料已成为现代工业中最重要的材料类别之一[100]，显示出多种重要的应用，如二极管，激光器，光电探测器和光伏电池的等[101]。半导体设计中，最值得研究的特性是带隙，因为它是影响光电器件中半导体性能的决定性因素，因此也被认为是最具商业意义的特性。已经证明当合金化合物中不同的金属元素改变时，带隙表现出令人满意的可调谐性[102,103]。

本章我们设计了一种材料的层次式表征方法，基于该表征方法我们提出了包含全卷积层的卷积神经网络材料属性预测模型。通过设计特殊的卷积算子，该模型能很好的从材料的原始输入矩阵中提取出有用的特征，进行最后的回归任务。利用 ICSD 中的数据进行不同的有监督任务，我们建立了能预测材料带隙的 ICSD-BG 和能预测材料形成能量的 ICSD-FE 的高精度预测模型。我们建立的模型在测试集上的预测表现和在验证集上的表现基本一致，这保证了后续我们用预测模型在 MatGAN 生成的假设材料中进行筛选的可靠性。

4.2 卷积神经网络模型的建立

4.2.1 数据与表征

1）实验数据

我们使用沉积在 ICSD 数据库中的无机材料来训练我们的 CNN 模型。我们使用的 ICSD 数据是通过使用除去所有单原子化合物，晶胞中具有 8 个以上原子的化合物和含有 Kr 和 He 元素的化合物而制备的。最终 ICSD 数据集包含 28137 种化合物。

2）数据的表征

本研究采用基于分子式中元素属性的统计方法来表征材料。统计元素属性表示是指计算材料元素性质的所有统计数据，如周期表分子中元素的周期数、族数、原子半径、熔化温度、来自 s、p、d 和 f 轨道的价电子的平均分数等。本文利用 Matminer 包[104]计算了元素的 22 种属性（见表 2.1）。

$$T=\begin{bmatrix} Numb\_min & Numb\_max & Numb\_range & \cdots & Numb\_mode \\ GSmg\_min & GSmg\_max & GSmg\_range & \cdots & GSmg\_mode \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ Spac\_min & Spac\_max & Spac\_range & \cdots & Spac\_mode \end{bmatrix} \in R^{22\times 6} \qquad 4\text{-}1$$

对于每种属性，我们计算其最大值、最小值、范围、平均值、方差和给定材料组成元素的特征，以便将其表征为如上所示的矩阵 $T\in R^{s\times d}(s=22，d=6)$。每种属性有六个统计值，可以看作是材料的局部表示，因此我们可以设计形状为 1×6 的卷积算子对特征矩阵 T 进行连续全行扫描，提取局部特征。

4.2.2 卷积神经网络的结构

CNN 是一种深度学习算法，在计算机视觉[105]等应用领域取得了突破性进展。它的主要优点是能够从高维数据中提取层次特征。CNN 一般由以下六部分组成：输入层、卷积层、激活层、池化层、全连接层和输出层。利用卷积算子提取局部特征；利用池化算子压缩卷积得到的特征映射，简化网络，降低计算复杂度；激活函数是神经网络非线性变换的主要来源；全连接层将学习到的高级特征映射到输出。材料$T\in R^{s\times d}$的特征映射比图像特征映射小得多，为了避免采用池化操作在提取主要特征时信息的丢失，本研究中 CNN 不包含池化操作。我们的 CNN 结构如图 4.1 所示。

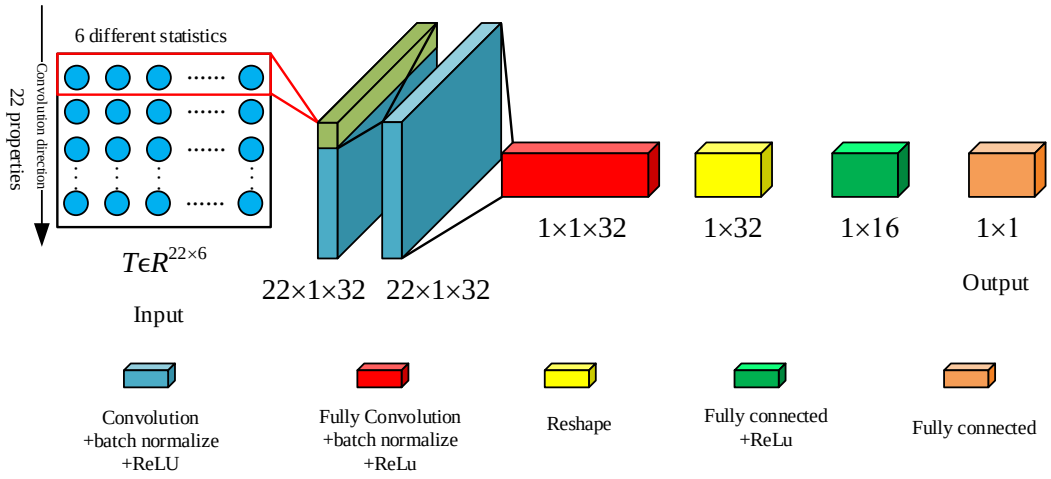


图 4.1.卷积神经网络模型的结构

根据公式4-1，每一行材料表征矩阵T是给定元素性质的6个不同统计量的集合。为便于特征的提取，CNN 模型由两个全行扫描卷积层、一个完全卷积层和两个全连接层组成。CNN 模型各层详细参数见表 4.1

表 4.1 CNN 模型的参数

Layer	Input Shape	Kernel number	Kernel Size	Stride	Output Shape
Conv1	[batch, 22, 6, 1]	32	(1, 6, 1)	(1, 1)	[batch, 22, 1, 32]
Conv2	[batch, 22, 1, 32]	32	(1, 1, 32)	(1, 1)	[batch, 22, 1, 32]
Conv3	[batch, 22, 1, 32]	32	(22, 1, 32)	(1, 1)	[batch, 1, 1, 32]
Reshape	[batch, 1, 1, 32]	-	-	-	[batch, 32]
Fc4	[batch, 32]	-	-	-	[batch, 32]
Fc5	[batch, 32]	-	-	-	[batch, 1]

将第一个卷积层的行扫描卷积核应用于矩阵 T，融合不同的统计值以提取高层特征。卷积核 l 的宽度为 d，与材料特征矩阵 T 的宽度相同，卷积核 h 的高度设为 1。为了提取 22 个元素属性之间的更多关系，我们使用完全卷积层来学习属性间的特征。这种思想最初是在文献[106]中提出的，用于语义分割，卷积核的大小与输入特征映射的大小相同。全卷积层与全连接层的区别在于，全卷积层的特征映射是通过卷积运算生成的，而全连通层的特征映射是通过权值和生成的。在每个卷积层之后，使用批处理规范化层[91]来提高模型的收敛速度，并减小在学习过程中网络权值初始化的影响。除输出层外，神经网络的每一层都使用一个线性整流单元（ReLu）作为激活函数。

4.2.3 数据的处理

1）数据的预处理

为了减少材料属性的差异尺度对模型的影响，我们对表征矩阵进行了归一化预处理。对于每个材质属性，我们将原始值转换为缩放值，如公式 4-2 所示。

$$x' = \frac{x - \min A}{\max A - \min A} \tag{4-2}$$

其中，x 表示原始数据，x'表示规范化数据，$\max A$ 和 $\min A$ 表示统计属性的最大值和最小值。

2）数据的划分

为了训练 CNN 模型，需要将数据集分为训练集、测试集和验证集。训练集用于更新模型的权值，测试集用于调整模型的超参数，验证集用于判断模型的优劣。图 4.2 显示了数据的划分过程。我们首先将 10 倍交叉验证应用于数据，在每一次交叉验证中，数据集被分成一个训练集和一个验证集（占所有样本的 10%）。在每一个折叠中，我们进一步随机地将数据划分为训练集和测试集。训练集和测试集用于训练和调整 CNN 模型的参数，验证集用于评估这模型的质量。在训练 CNN 时，我们保存在 2000 个周期内测试集上R^2表现最好的模型。

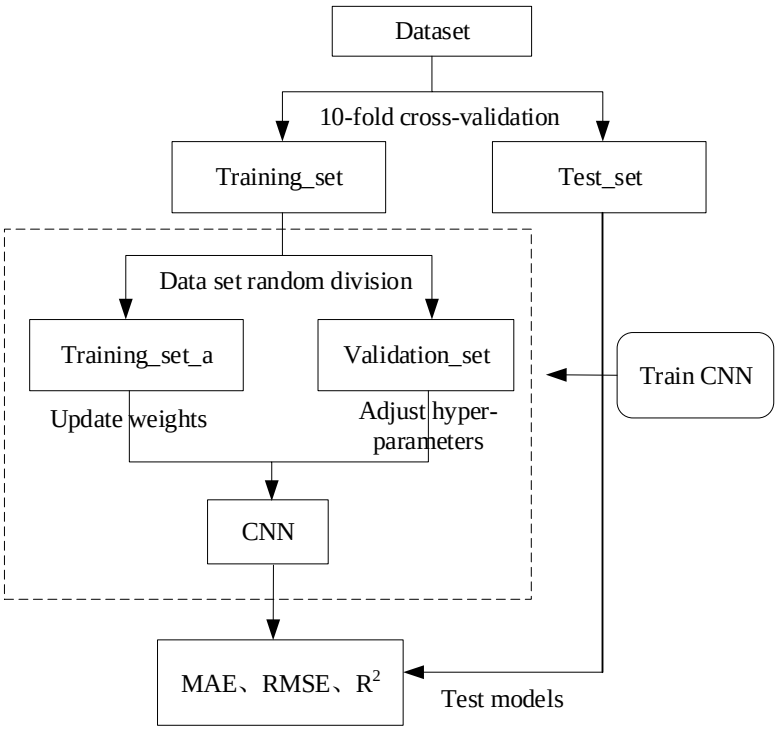


图 4.2. CNN 模型的训练过程及其性能评价

4.2.4 评价指标

为了评价回归模型的性能，我们使用了平均绝对误差（MAE）、均方根误差（*RMSE*）和 R-Squared（R^2）作为评价指标。MAE 用来反映预测值误差的实际情况，RMSE 用来衡量预测值同真值之间的偏差，R^2用来表示预测值和真实值的拟合程度。具体的计算公式如下所示：

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}\left|y_i-\hat{y}_i\right| \tag{4-3}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2} \tag{4-4}$$

$$R^2=1-\frac{\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{m}\left(y_i-\bar{y}\right)^2} \tag{4-5}$$

式中 m 是样本数量，y_i和$\hat{y}_i$分别是第 i 个样本标签的真实值和预测值，$\bar{y}$是m个样本真实标签的平均值。

4.3 实验结果

为了保证计算实验结果的稳定性和可靠性，后续所有的实验都是 10 次 10 折交叉验证计算平均值的结果。整个模型是基于 Python 3.6 开发的，CNN 模型使用 Tensorflow9.0[107]深度学习框架。所有程序均在具有 3.6GHz GPU 和 NVIDIA GPU GTX1080Ti 的 Dell 服务器上运行。

我们在 ICSD 上分别建立了材料的带隙（Band Gap）和形成能量（Formation Energy）的卷积神经网络预测模型，将这两个模型分别命名为 ICSD-BG 和 ICSD-FM。

4.3.1 模型的超参数

在机器学习中，模型可以被认为是具有许多可调旋钮的机器，这些旋钮被称为超参数，调整旋钮可更改模型的表现性能。对于神经网络超参数的搜索空间主要包括动量（Momentum）、学习率（Learning rate）、优化算法（Optimization algorithms）和批处理的数量（Batch size）等。其中学习率是最重要的深度神经网络的超参数之一，我们尝试了从 0.1 到$1e^{-6}$的学习率（每次减小 10 倍）。

我们基于经验直觉为每个超参数设定初始值，然后使用贪婪算法逐步调整每个超参数，而不是执行因为计算成本而不可行的网格搜索。我们根据测试集上的损失值调整超参数，以获得一组最小化测试集上损失值的超参数。最后确定各个模型的超参数

如表 4.2 所示：

表 4.2 预测模型的超参数

Model	Batch size	Learning rate	优化算法
ICSD-BG	256	0.01	Adam
ICSD-FE	512	0.1	Adam

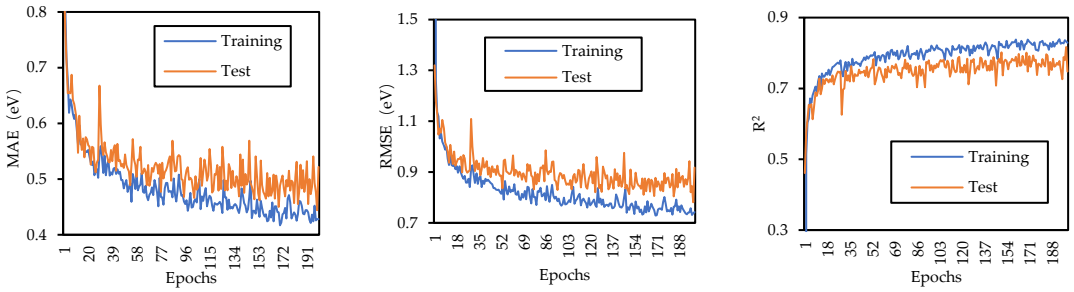
4.3.2 模型训练

图 4.3 给出了 CNN 算法的训练流程图。首先模型根据输入的批次数据完成正向传播，再根据输入数据的真实标签计算出误差，然后根据误差函数进行反向传播计算出梯度，完成神经网络权值的更新。在完成一个训练批次后，我们需要根据模型在测试集上的表现来确定模型训练的好坏，因此需要采用预先划分好的测试数据对进行测试。若模型在测试集上的表现比上一批次训练得到的模型表现的好，则保存当前模型，反之不保存。这样一直循环，最后保存到训练周期内在测试集上表现最好的模型。模型训练完成后，会用预先划分好的验证集对模型进行验证，以观察模型的真实表现。

```
1:  循环开始 每一个训练周期
2:      循环开始 每一个训练批次
3:          1）CNN 模型的正向传播
4:          2）计算损失函数
5:          3）CNN 模型的反向传播
6:          4）神经网络权值更新
7:      内层循环结束
8:      计算 CNN 在测试集上的表现
9:  外层循环结束
10: 计算 CNN 在验证集上的表现
```

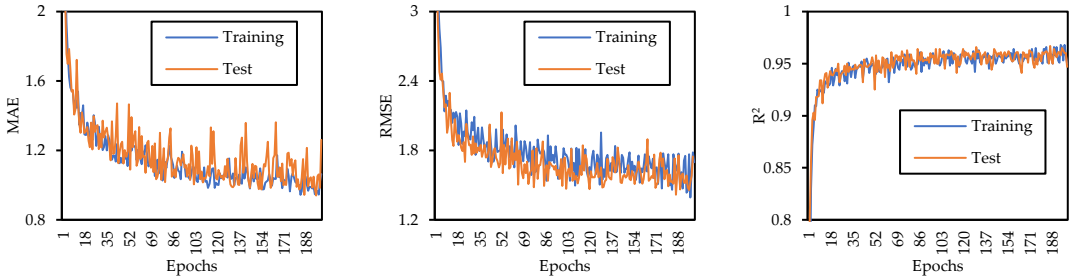
图 4.3 CNN 算法的训练流程图

我们将模型训练 2000 个周期（代），并保存在测试集上表现最好的神经网络权值。图 4.4 到图 4.5 给出了 ICSD-BG、OQMD-FE 在训练过程中训练集和测试集的 MAE、RMSE、R^2的变化过程。



(a) (b) (c)

图 4.4 ICSD-BG 的收敛过程



(a) (b) (c)

图 4.5 ICSD-FE 的收敛过程

随着迭代次数的增加，能反映预测值与真实值之间误差的 MAE 和 RMSE 逐渐减小，能反映两者接近程度的R^2逐渐增加。ICSD-BG 在训练后期，在训练集上的表现略优于在测试集上的表现，出现的一定程度的过拟合，这一方面与有监督模型训练的机制有关，另一方面是因为材料带隙的预测是一个极具挑战的任务。ICSD-FE 在整个训练过程中，在训练集上的表现与在测试集上的表现非常接近，表明 ICSD-FE 能很好的预测 ICSD 上的材料的形成能量。我们提出的卷积神经网络结构在不同任务上都能很好的收敛。

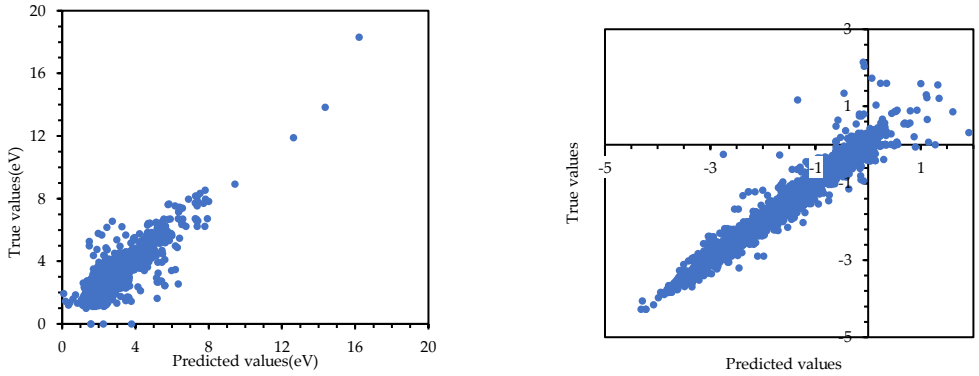
4.3.3 实验结果与分析

为了确保用建立好的预测模型在 MatGAN 生成的假设材料中进行新材料的筛选的准确性，必须要保证模型对未知数据预测的稳健性，我们将未参与训练的验证集输入神经网络，结果如表 4.3。

表 4.3 模型在测试集/验证集上的预测结果

Model	MAE	RMSE	R^2
ICSD-BG	0.406/0.414	0.694/0.754	0.854/0.823
ICSD-FE	0.039/0.046	0.054/0.063	0.963/0.954

ICSD-BG 在验证集上其R^2能达到 0.82，这与 Goodall 等人[71]提出的仅将化学计量作为输入并自动从数据中学习适当且系统可改进的描述符的神经网络方法取得的结果一致。ICSD-FE 在验证集上 MAE 能达到 0.046eV，略好于 Jha 等人[27]在相同的数据上采用全连接网络取得的结果。本章提出的神经网络模型在测试集上与验证集上的结果一致，表明了模型对未知材料数据具有较强的预测的能力，保证了模型在后续筛选材料的可靠性。



(a) (b)

图 4.7 预测值与真实值的比较

为了观察模型预测未知数据的特点，我们比较模型在验证集上预测值和真实值的差距，如图 4.7（当预测值与真实值相等时点将落在 45° 斜线上）。对于带隙的预测（图 4.7-a），ICSD-BG 对带隙高于 6.0eV 和 1eV 到 2eV 的材料预测的较为精确，而对带隙不在此范围的材料预测的偏差较大。如图 4.7-b，ICSD-FE 在整个范围内对化合物的形成能量的预测都表象的比较好。

4.4 本章小结

本章我们设计了一种材料的层次式材料表征方法，基于该表征方法我们提出了包含全卷积层的卷积神经网络材料属性预测模型。通过设计特殊的卷积算子，该模型能很好的从材料的原始输入矩阵中提取出有用的特征，进行最后的回归任务。利用 ICSD 中的数据进行不同任务的有监督训练，我们建立了能预测材料带隙的 ICSD-BG，能预测材料形成能量的 ICSD-FE 的高精度预测模型。我们建立的模型在测试集上的预测表现和在验证集上的表现基本一致，这保证了后续我们用预测模型在 MatGAN 生成的假设材料中进行筛选的可靠性。我们还分析了每个模型对未知数据预测的适用区间，为后续的材料筛选提供了精确的预测模型。

第 5 章 新材料的发现

5.1 引言

第三章建立了能生成在原子组合规律上与真实化合物相同的 MatGAN 模型，利用 ICSD 训练构建的 ICSD-GAN 生成了 200 万中不同的假设材料，建立起了材料的筛选空间。第四章基于卷积神经网络在 ICSD 上建立了能预测材料形成能量的 ICSD-FE 和能预测带隙的 ICSD-BG 模型。接下来，我们将材料属性预测模型应用到 ICSD-GAN 构建的假设材料中，进行新材料的发现。需要说明的是在第三章中通过 OQMD-GAN 和 MP-GAN 也构建了假设材料空间，用建立的预测模型同样能在对应的假设材料上进行筛选，在此我们只以 ICSD 为例进行材料的筛选。同样，本文只建立了两种材料属性的预测模型，根据研究的需要，也可建立其他材料特征的预测模型进行材料的筛选。

本章将 ICSD-FE 施加于 GAN-ICSD 生成的假设材料上进行筛选，筛选出形成能量较低料。在筛选出的形成能量较低的材料基础上，用 ICSD-BG 继续筛选带隙在 1.0~2.0eV 之间的材料。我们将筛选出的材料的带隙在 1.0~2.0eV 以及这些材料用 ICSD-FE 预测的形成能量放在了 http://github/danyabo/papendix.com 以供研究者们进行后续的 DFT 仿真计算或实验合成。

5.2 筛选形成能量较低的材料

化合物的形成能量可以评估材料的稳定性，因此检查某种假设材料存在的可能性的一种方式是计算其形成能量。我们将 ICSD-FE 施加于 GAN-ICSD 生成的假设材料上进行筛选并在表 5.1 中列出了二元、三元和四元组合的前 20 种形成能量最小的材料。

表 5.1 含有二、三、四种原子组合材料的的预测形成能量

2 element	FE(eV/atom)	3 element	FE(eV/atom)	4 element	FE(eV/atom)
TmF3	-4.595	Dy2F2O2	-4.465	SrEuLuF7	-4.460
ErF3	-4.426	DyLuF6	-4.463	MgCeDyF7	-4.393
CeF3	-4.384	LaDyF7	-4.424	LaDy2F6O2	-4.383
ScF3	-4.361	NdLuF6	-4.384	KEuErF7	-4.165
SmF3	-4.345	ScLaF7	-4.380	LaDyPbF7	-4.163
Ho2O3	-4.027	ZrLaF7	-4.374	DyTaAc2O7	-4.103
Dy2O3	-3.935	DyFO	-4.368	LaNdSF5,	-3.886
DyF4	-3.901	CeDyF5	-4.358	DyPuFO3	-3.877
UO2	-3.879	LaYbF5	-4.343	LaLuPbF6	-3.870
Ba2F5	3.783	Sr2LaF6	-4.339	TiLaHoO5	-3.870
YbF2	-3.781	LaEuF6	-4.335	DyHoF6O2	-3.868
Sm2O3	-3.767	ScLaO3	-4.278	NdDyTaO5	-3.862
Ac4O7	-3.762	LaFO	-4.225	DyHoTaO5	-3.853
UF5	-3.746	Dy4F5O4	-4.165	LaEuDy2O6	-3.851
CeF2	-3.727	ZrLaO3	-4.089	CrLaHo2O6	-3.845
ErF5	-3.705	LaDyO3	-4.070	Dy3PuF3O4	-3.844
Sr2F3	-3.651	NdLuO3	-4.058	NdEuDy2O6	-3.836
HoF2	-3.648	EuDy3O6	-4.040	LaDy2ClO4	-3.835
Y2O3	-3.628	Y2LaO5	-4.022	DyLuTaO5	-3.833
Yb2F5	-3.586	Sc2PuO5	-3.950	CrNd2F7O2	-3.830

5.3 筛选半导体材料

材料的形成能量只能反映某种材料能够稳定存在的可能性，但研究者们往往希望发现的材料拥有某种特性，如半导体材料。半导体材料已成为现代工业中最重要的材料类别之一[100]，显示出多种重要的应用，如二极管，激光器，光电探测器和光伏电池等[101]。半导体设计中，最值得研究的特性是带隙，因为它是影响光电器件中半导体性能的决定性因素，因此也被认为是最具商业意义的特性。

带隙是光学和电子应用中表征半导体和绝缘体简单但重要的参数[108]。要从大量化合物中探索具有所需性能的材料，制备满足带隙条件的化合物可能是一个很好的起点。研究者一般将带隙低于 3eV 的材料定义为窄带隙半导体。由于窄带隙半导体材料具有优异的吸光性能，因此窄带隙半导体广泛应用于光电领域。

材料的带隙在很大程度上影响太阳能的转换效率，制作光伏电池的材料一旦选定，其最大可能的转换效率便基本确定，例如单晶硅电池的理论最大转换效率为 24%。在电池的开发和设计中多方面的改进都是为了逼近最大转换效率。根据有关理论研究[109]，针对太阳光谱的最佳带隙为 1.4eV，因此我们可以着重关注带隙在 1.0~2.0eV 之间的材料。

在 5.2 部分筛选出的形成能量低于零的材料基础上，用 ICSD-BG 继续进行筛选，在表 5.2 中列出了 40 种带隙在 1.0~2.0eV 之间的材料。此外我们在附录 C 中列出了筛选出来的 200 多种材料的带隙以及其形成能量。

表 5.2 生成的假设材料的预测形成能量和带隙

Formulas	FE(eV/atom)	BG(eV)	Formulas	FE(eV/atom)	BG(eV)
Br2SeS2	-0.260	1.999	Pb3Se6P	-0.397	1.673
LaGeO6N	-2.173	1.855	Ho2Pb3F6	-3.060	1.667
Sn2NdCl6O2	-2.000	1.855	CsPuO2	-2.620	1.658
K3NbCl7	-2.062	1.819	NdTlS2C	-1.191	1.639
TbPtS3	-1.223	1.805	VBi7O4	-0.271	1.6283
TiBi2AsO6	-1.918	1.796	PbCl4FC	-1.244	1.601
LaSiOC	-1.513	1.795	V2MoO7	-2.348	1.571
ZnGaS4	-0.628	1.784	Sb8F5	-0.887	1.539
TmTaBr6	-1.513	1.784	YCd2O8	-1.549	1.524
K6AsS4	-1.257	1.777	DyUO2	-2.574	1.509
DyThO3	-3.307	1.777	ZnSn3Se7	-0.419	1.475
TmTaSbO6	-2.847	1.767	TeSi2O3	-0.994	1.512
MnCs2F6	-2.967	1.751	ZnSi3	0.433	1.515
BaPbSC	-0.918	1.748	TiSnSO2	-1.940	1.515
RhAs3O8	-1.438	1.732	RuLaF5	-3.106	1.401
MnRuO4	-1.246	1.728	CrPbAs2F6	-2.330	1.364
ScDyTaPuO6	-3.367	1.714	Sn2WCl8	-1.199	1.300
ZnCl6FC	-0.888	1.705	CrPb2Cl5	-1.433	1.214
CeDySbF6	-3.511	1.691	Li2CoCl5	-1.420	1.116
CrZrS4O2	-1.824	1.689	NbRuIO4	-1.892	1.000

5.4 本章小结

本部分我们将 ICSD-FE 施加于 GAN-ICSD 生成的假设材料上进行筛选，筛选出形成能量较低的材料。在筛选出的形成能量较低的材料基础上，用 ICSD-BG 继续筛选带隙在 1.0~2.0eV 之间的材料。我们将筛选出的所有材料都放在了 http://github/danyabo/papendix.com 以供研究者们进行后续的 DFT 仿真计算或实验合成。

第 6 章 总结与展望

6.1 论文研究总结

材料创新成为我们应对一些最紧迫的社会挑战的关键，但目前的材料发现仍然涉及重大的反复试验，可能需要数十年的研究才能确定适合于技术应用的材料。这一漫长发现过程的主要原因是，潜在材料的数量是巨大的，明智地选择要关注的材料以及进行哪些实验荆棘密布。针对巨大的潜在材料的数量而无法明智地选择要关注的材料的问题，采用深度学习方法设计出能高效采样的无机化合物生成模型和能精确预测材料属性的筛选模型，基于这两个模型进行无机化合物新材料的发现具有重要的意义。具体而言，本文的主要研究成果有几个方面：

1）建立的基于生成对抗网络（GAN）的生成深度学习模型（MatGAN），能有效地生成新的假设无机材料。当使用 ICSD 数据库中的材料进行训练，我们的 GAN 模型可以生成训练数据集中不存在的假设性材料，当生成 200 万个样本时，新颖性达到 92.53%。当使用 ICSD 的材料进行训练时，即使我们的 GAN 模型中没有明确施加电荷中性和电负性平衡这样的化学规则，生成的假设材料化学有效样本中所占的百分比也达到了 84.5%。建立的算法可用于加快无机材料的逆向设计或计算筛选。

2）为了建立高精度的材料属性预测模型，设计了一种无机化合物的层次式表征方法，通过设计包含特殊的卷积算子和全卷积层的卷积神经网络并采用 ICSD 训练出了能预测材料带隙的 ICSD-BG 和能预测材料形成能量的 ICSD-FE 的高精度预测模型。

3）将 ICSD-FE 施加于 GAN-ICSD 生成的假设材料上进行筛选，筛选出形成能量较低的材料。在筛选出的形成能量较低的材料基础上，用 ICSD-BG 继续筛选带隙在 1.0~2.0eV 之间的材料。我们将筛选出的材料都放在了 http://github/danyabo/papendix.com 以供研究者们进行后续的 DFT 仿真计算或实验合成。

6.2 未来工作展望

我们的工作可以以多种方式扩展。首先，即使没有将这些规则明确地应用于 GAN 模型中，我们发现 MatGAN 可以隐式地学习化学成分规则。但是，有时需要实施化学规则过滤器以删除化学上无效的候选者，这可以根据我们的材料矩阵表示轻松实现。当前研究的另一个局限性在于，我们在材料表示中仅考虑了化合物中元素的整数比，

而具有分数比的掺杂材料在锂离子电池材料 LiZn0.01Fe0.99PO4 等功能性材料非常常见。我们的研究可以通过在表示矩阵上允许实数来扩展。但是，考虑到掺杂比例的无限可能，我们的 GAN 方法可能需要与其他采样技术（例如遗传算法[110,111]，遗传编程[112]和用于混合参数搜索的主动机器学习[113,114]）一起使用。此外，目前的 GAN 模型无法说明假设材料的晶体结构（晶格常数，空间群，原子坐标等）。但是，有了足够的计算资源，就有可能利用基于 DFT 的计算软件包，例如 USPEX[115]或 CALYPSO[116]来确定给定材料组成及其化学计量的晶体结构。我们的 GAN 模型也可以与材料结构生成器一起使用。

致谢

光阴荏苒，硕士之学将毕，三年之学使我受益匪浅。经大半年之厉，硕士大论文遂成，回首大半年来收集、整理、思、滞、改至终成之。余得良多的关怀与助，今欲示吾至诚之意。

吾欲感硕士三年之恩师胡建军教授，犹忆三年前君并手带余复现第一篇 SCI，君敢创新精神，开吾之学识，请于科研道中少行曲路，大者增吾科研之兴，君亦常戒吾惜日力学，于吾之学供之力助。君与吾亦师亦友，不但教吾统治之科学法，且予吾众人道。又谢李少波教授，谢君在吾科研中供之计利，使吾论实验能遂。

须谢团队之侪伦，盖为汝等，枯者科研生始文，共论时甚喜亦收多。胡杰师兄，师门大师兄亦，靡不耐教，三年而得其之大助。张森师兄，师门二师兄亦，其术甚矣，吾有术难时皆能速决之助。胡甜甜，吾之师妹亦吾之女友，生活上为供至之怀，予每遇烦心也，皆时之教吾。一路长之李琴、权华凤、杨静、张安思等博士同门，及张成龙、姚勇、郑凯、李想、董容智、柘龙炫、赵乐等硕士同门一一谢之。

有感于吾有育之父母，巍巍高山，重若父母之爱。诸公出身贫，其克勤克俭，谓吾学而倾其所。诸公虽学历不高，而皆为义，年来勤身而为余铺就顺学之路，又常嘱吾学之路虽道阻且长，而行则将至，勿惮前崎。正是有诸公一路支与伴，我乃得成业。

但雅波

2020 年 4 月 24 日于湖北

参考文献

[1] 刘石泉. 第三届航天工程和高性能材料需求与应用技术交流会. *导航定位与授时 v.3;No.10*, 3.

[2] Evans, G.W. Projected behavioral impacts of global climate change. *Annual review of psychology* **2019**, *70*, 449-474.

[3] Sampedro, J.; Smith, S.J.; Arto, I.; González-Eguino, M.; Markandya, A.; Mulvaney, K.M.; Pizarro-Irizar, C.; Van Dingenen, R. Health co-benefits and mitigation costs as per the Paris Agreement under different technological pathways for energy supply. *Environment International* **2020**, *136*, 105513.

[4] Bergerhoff, G.; Hundt, R.; Sievers, R.; Brown, I. The inorganic crystal structure data base. *Journal of chemical information and computer sciences* **1983**, *23*, 66-69.

[5] Belsky, A.; Hellenbrandt, M.; Karen, V.L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B: Structural Science* **2002**, *58*, 364-369.

[6] Walsh, A. The quest for new functionality. *Nature chemistry* **2015**, *7*, 274-275.

[7] Yeh, J.W.; Chen, S.K.; Lin, S.J.; Gan, J.Y.; Chin, T.S.; Shun, T.T.; Tsau, C.H.; Chang, S.Y. Nanostructured high‐entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. *Advanced Engineering Materials* **2004**, *6*, 299-303.

[8] Sharma, A.; Singh, P.; Johnson, D.D.; Liaw, P.K.; Balasubramanian, G. Atomistic clustering-ordering and high-strain deformation of an Al 0.1 CrCoFeNi high-entropy alloy. *Scientific reports* **2016**, *6*, 31028.

[9] Sharma, A.; Deshmukh, S.A.; Liaw, P.K.; Balasubramanian, G. Crystallization kinetics in AlxCrCoFeNi (0≤ x≤ 40) high-entropy alloys. *Scripta Materialia* **2017**, *141*, 54-57.

[10] Sharma, A.; Balasubramanian, G. Dislocation dynamics in Al0. 1CoCrFeNi high-entropy alloy under tensile loading. *Intermetallics* **2017**, *91*, 31-34.

[11] Castleton, C.W.; Höglund, A.; Mirbt, S. Managing the supercell approximation for charged defects in semiconductors: Finite-size scaling, charge correction factors, the band-gap problem, and the ab initio dielectric constant. *Physical Review B* **2006**, *73*, 035215.

[12] Lindgren, I. *Relativistic many-body theory: a new field-theoretical approach*; Springer: 2016; Vol. 63.

[13] van Schilfgaarde, M.; Kotani, T.; Faleev, S. Quasiparticle self-consistent g w theory. *Physical review letters* **2006**, *96*, 226402.

[14] Koinuma, H.; Takeuchi, I. Combinatorial solid-state chemistry of inorganic materials. *Nature materials* **2004**, *3*, 429-438.

[15] Choo, K.Y.; Hodge, R.A.; Ramachandran, K.K.; Sivakumar, G. Controlling a video capture device based on cognitive personal action and image identification. Google Patents: 2019.

[16] Berg, M.J.; Robertson, J.C.; Onderdonk, L.A.; Reiser, J.M.; Corby, K.D. Object dispenser having a variable orifice and image identification. Google Patents: 2016.

[17] Yang, J.; Li, S.; Gao, Z.; Wang, Z.; Liu, W. Real-time recognition method for 0.8 cm darning needles and KR22 bearings based on convolution neural networks and data increase. *Applied Sciences* **2018**, *8*, 1857.

[18] Dusan, S.V.; Lindahl, A.M.; Watson, R.D. Automatic speech recognition triggering system. Google Patents: 2019.

[19] Malinowski, L.M.; Majcher, P.J.; Stemmer, G.; Rozen, P.; Hofer, J.; Bauer, J.G. System and method of automatic speech recognition using parallel processing for weighted finite state transducer-based speech decoding. Google Patents: 2019.

[20] Juneja, A. Hybridized client-server speech recognition. Google Patents: 2017.

[21] Clark, K.; Luong, M.-T.; Khandelwal, U.; Manning, C.D.; Le, Q.V. Bam! born-again multi-task networks for natural language understanding. *arXiv preprint arXiv:1907.04829* **2019**.

[22] Thomason, J.; Padmakumar, A.; Sinapov, J.; Walker, N.; Jiang, Y.; Yedidsion, H.; Hart, J.; Stone, P.; Mooney, R.J. Improving grounded natural language understanding through human-robot dialog. In Proceedings of 2019 International Conference on Robotics and Automation (ICRA); pp. 6934-6941.

[23] Marcus, J.N. Initializing a workspace for building a natural language understanding system. Google Patents: 2019.

[24] Wang, F.-Y.; Zhang, J.J.; Zheng, X.; Wang, X.; Yuan, Y.; Dai, X.; Zhang, J.; Yang, L. Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica* **2016**, *3*, 113-120.

[25] Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *Apl Materials* **2013**, *1*, 011002.

[26] Saal, J.E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *Jom* **2013**, *65*, 1501-1509.

[27] Jha, D.; Ward, L.; Paul, A.; Liao, W.-k.; Choudhary, A.; Wolverton, C.; Agrawal, A. Elemnet: Deep learning the chemistry of materials from only elemental composition. *Scientific reports* **2018**, *8*, 1-13.

[28] Kirklin, S.; Saal, J.E.; Meredig, B.; Thompson, A.; Doak, J.W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1*, 1-15.

[29] Calfa, B.A.; Kitchin, J.R. Property prediction of crystalline solids from composition and crystal structure.

AIChE Journal **2016**, *62*, 2605-2613.

[30] Hsu, K.-Y.; Li, H.-Y.; Psaltis, D. Holographic implementation of a fully connected neural network. *Proceedings of the IEEE* **1990**, *78*, 1637-1645.

[31] Stanev, V.; Oses, C.; Kusne, A.G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. Machine learning modeling of superconducting critical temperature. *npj Computational Materials* **2018**, *4*, 29.

[32] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 16028.

[33] Liaw, A.; Wiener, M. Classification and regression by randomForest. *R news* **2002**, *2*, 18-22.

[34] Xie, S.; Stewart, G.; Hamlin, J.; Hirschfeld, P.; Hennig, R. Functional form of the superconducting critical temperature from machine learning. *Physical Review B* **2019**, *100*, 174513.

[35]Parr, R.G. Density functional theory of atoms and molecules. In *Horizons of Quantum Chemistry*, Springer: 1980; pp. 5-15.

[36] Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R.H.; Nelson, L.J.; Hart, G.L.; Sanvito, S.; Buongiorno-Nardelli, M. AFLOWLIB. ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science* **2012**, *58*, 227-235.

[37] Bear, J.E.; Svitkina, T.M.; Krause, M.; Schafer, D.A.; Loureiro, J.J.; Strasser, G.A.; Maly, I.V.; Chaga, O.Y.; Cooper, J.A.; Borisy, G.G. Antagonism between Ena/VASP proteins and actin filament capping regulates fibroblast motility. *Cell* **2002**, *109*, 509-521.

[38] Li, S.; Dan, Y.; Li, X.; Hu, T.; Dong, R.; Cao, Z.; Hu, J. Critical Temperature Prediction of Superconductors Based on Atomic Vectors and Deep Learning. *Symmetry* **2020**, *12*, 262.

[39] Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **2017**, *3*, 1-13.

[40] Ward, L.; Wolverton, C. Atomistic calculations and materials informatics: A review. *Current Opinion in Solid State and Materials Science* **2017**, *21*, 167-176.

[41] Stanev, V.; Oses, C.; Kusne, A.G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. Machine learning modeling of superconducting critical temperature. *npj Computational Materials* **2018**, *4*, 1-14.

[42] Dan, Y.; Dong, R.; Cao, Z.; Li, X.; Niu, C.; Li, S.; Hu, J. Computational Prediction of Critical Temperatures of Superconductors Based on Convolutional Gradient Boosting Decision Trees. *IEEE Access* **2020**, *8*, 57868-57878.

[43] Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *The journal of physical chemistry letters* **2018**, *9*, 1668-1673.

[44] Zhou, Q.; Tang, P.; Liu, S.; Pan, J.; Yan, Q.; Zhang, S.-C. Learning atoms for materials discovery. *Proceedings of the National Academy of Sciences* **2018**, *115*, E6411-E6417.

[45] Lyubartsev, A.P.; Laaksonen, A. Calculation of effective interaction potentials from radial distribution

functions: A reverse Monte Carlo approach. *Physical Review E* **1995**, *52*, 3730.

[46] Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature communications* **2017**, *8*, 1-12.

[47] Zimm, B.H. The scattering of light and the radial distribution function of high polymer solutions. *The Journal of Chemical Physics* **1948**, *16*, 1093-1099.

[48] Xie, T.; Grossman, J.C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.

[49] Honrao, S.; Anthonio, B.E.; Ramanathan, R.; Gabriel, J.J.; Hennig, R.G. Machine learning of ab-initio energy landscapes for crystal structure predictions. *Computational Materials Science* **2019**, *158*, 414-419.

[50] Zhu, L.; Amsler, M.; Fuhrer, T.; Schaefer, B.; Faraji, S.; Rostami, S.; Ghasemi, S.A.; Sadeghi, A.; Grauzinyte, M.; Wolverton, C. A fingerprint based metric for measuring similarities of crystalline structures. *The Journal of chemical physics* **2016**, *144*, 034203.

[51] Quinlan, J.R. Simplifying decision trees. *International journal of man-machine studies* **1987**, *27*, 221-234.

[52] Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In Proceedings of Advances in neural information processing systems; pp. 3146-3154.

[53] Suykens, J.A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural processing letters* **1999**, *9*, 293-300.

[54] An, S.; Liu, W.; Venkatesh, S. Face recognition using kernel ridge regression. In Proceedings of 2007 IEEE Conference on Computer Vision and Pattern Recognition; pp. 1-7.

[55] Safavian, S.R.; Landgrebe, D. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics* **1991**, *21*, 660-674.

[56] Friedman, J.H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, 1189-1232.

[57] Wagner, H.M. Linear programming techniques for regression analysis. *Journal of the American Statistical Association* **1959**, *54*, 206-212.

[58] Chen, L.; Tran, H.; Batra, R.; Kim, C.; Ramprasad, R. Machine learning models for the lattice thermal conductivity prediction of inorganic materials. *Computational Materials Science* **2019**, *170*, 109155.

[59] Zhan, T.; Fang, L.; Xu, Y. Prediction of thermal boundary resistance by the machine learning method. *Scientific reports* **2017**, *7*, 1-9.

[60] Furmanchuk, A.o.; Saal, J.E.; Doak, J.W.; Olson, G.B.; Choudhary, A.; Agrawal, A. Prediction of seebeck coefficient for compounds without restriction to fixed stoichiometry: A machine learning approach. *Journal of computational chemistry* **2018**, *39*, 191-202.

[61] Hamidieh, K. A data-driven statistical model for predicting the critical temperature of a superconductor.

Computational Materials Science **2018**, *154*, 346-354.

[62] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of Advances in neural information processing systems; pp. 1097-1105.

[63] Cecen, A.; Dai, H.; Yabansu, Y.C.; Kalidindi, S.R.; Song, L. Material structure-property linkages using three-dimensional convolutional neural networks. *Acta Materialia* **2018**, *146*, 76-84.

[64] Kondo, R.; Yamakawa, S.; Masuoka, Y.; Tajima, S.; Asahi, R. Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics. *Acta Materialia* **2017**, *141*, 29-38.

[65] Ling, J.; Hutchinson, M.; Antono, E.; DeCost, B.; Holm, E.A.; Meredig, B. Building data-driven models with microstructural images: Generalization and interpretability. *Materials Discovery* **2017**, *10*, 19-28.

[66] Dong, Y.; Wu, C.; Zhang, C.; Liu, Y.; Cheng, J.; Lin, J. Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. *npj Computational Materials* **2019**, *5*, 1-8.

[67] Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nature communications* **2019**, *10*, 1-12.

[68] Konno, T.; Kurokawa, H.; Nabeshima, F.; Ogawa, R.; Iwazume, M.; Hosako, I.; Maeda, A. Deep Learning of Superconductors I: Estimation of Critical Temperature of Superconductors Toward the Search for New Materials. *arXiv preprint arXiv:1812.01995* **2018**.

[69] 王佳. 图神经网络浅析. *现代计算机* **2019**.

[70] Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S.P. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials* **2019**, *31*, 3564-3572.

[71] Goodall, R.E.; Lee, A.A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *arXiv preprint arXiv:1910.00617* **2019**.

[72] Zolotarev, P.N.; Arshad, M.N.; Asiri, A.M.; Al-amshany, Z.M.; Blatov, V.A. A possible route toward expert systems in supramolecular chemistry: 2-periodic H-bond patterns in molecular crystals. *Crystal growth & design* **2014**, *14*, 1938-1949.

[73] Bartók, A.P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

[74] Von Lilienfeld, O.A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *International Journal of Quantum Chemistry* **2015**, *115*, 1084-1093.

[75] Davies, D.W.; Butler, K.T.; Jackson, A.J.; Morris, A.; Frost, J.M.; Skelton, J.M.; Walsh, A. Computational screening of all stoichiometric inorganic materials. *Chem* **2016**, *1*, 617-627.

[76] Dimoulas, A.; Tsipas, P.; Sotiropoulos, A.; Evangelou, E. Fermi-level pinning and charge neutrality level

in germanium. *Applied physics letters* **2006**, *89*, 252110.

[77] Lu, J.; Jin, H.; Dai, Y.; Yang, K.; Huang, B. Effect of electronegativity and charge balance on the visible-light-responsive photocatalytic activity of nonmetal doped anatase TiO2. *International Journal of Photoenergy* **2012**, *2012*.

[78] Jensen, W.B.; Jensen, W.B. The origin of the ionic-radius ratio rules. *Journal of Chemical Education* **2010**, *87*, 587-588.

[79] Ranganathan, S.; Inoue, A. An application of Pettifor structure maps for the identification of pseudo-binary quasicrystalline intermetallics. *Acta materialia* **2006**, *54*, 3647-3656.

[80] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360-365.

[81] Xue, D.; Gong, Y.; Yang, Z.; Chuai, G.; Qu, S.; Shen, A.; Yu, J.; Liu, Q. Advances and challenges in deep generative models for de novo molecule generation. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2019**, *9*, e1395.

[82] Xu, Y.; Lin, K.; Wang, S.; Wang, L.; Cai, C.; Song, C.; Lai, L.; Pei, J. Deep learning for molecular generation. *Future medicinal chemistry* **2019**, *11*, 567-597.

[83] Elton, D.C.; Boukouvalas, Z.; Fuge, M.D.; Chung, P.W. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering* **2019**, *4*, 828-849.

[84] Ferguson, A.L. Machine learning and data science in soft materials engineering. *Journal of Physics: Condensed Matter* **2017**, *30*, 043002.

[85] Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875* **2017**.

[86] Noh, J.; Kim, J.; Stein, H.S.; Sanchez-Lengeling, B.; Gregoire, J.M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370-1384.

[87] Hoffmann, J.; Maestrati, L.; Sawada, Y.; Tang, J.; Sellier, J.M.; Bengio, Y. Data-Driven Approach to Encoding and Decoding 3-D Crystal Structures. *arXiv preprint arXiv:1909.00949* **2019**.

[88] Van Hasselt, H.; Guez, A.; Silver, D. Deep reinforcement learning with double q-learning. In Proceedings of Thirtieth AAAI conference on artificial intelligence.

[89] Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* **2016**.

[90] Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In Proceedings of Advances in neural information processing systems; pp. 2172-2180.

[91] Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* **2015**.

[92] Li, Y.; Yuan, Y. Convergence analysis of two-layer neural networks with relu activation. In Proceedings of Advances in Neural Information Processing Systems 2017; pp. 597-607.

[93] Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **2008**, *9*, 2579-2605.

[94] Ye, W.; Chen, C.; Wang, Z.; Chu, I.-H.; Ong, S.P. Deep neural networks for accurate predictions of crystal stability. *Nature communications* **2018**, *9*, 1-6.

[95] Kang, S.; Cho, K. Conditional molecular design with deep generative models. *Journal of chemical information and modeling* **2018**, *59*, 43-52.

[96] Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; Carin, L. Variational autoencoder for deep learning of images, labels and captions. In Proceedings of Advances in neural information processing systems; pp. 2352-2360.

[97] Shamir, R.R.; Duchin, Y.; Kim, J.; Sapiro, G.; Harel, N. Continuous dice coefficient: a method for evaluating probabilistic segmentations. *arXiv preprint arXiv:1906.11031* **2019**.

[98] Shamir, R.R.; Duchin, Y.; Kim, J.; Sapiro, G.; Harel, N. Continuous Dice Coefficient: a Method for Evaluating Probabilistic Segmentations. *BioRxiv* **2018**, 306977.

[99] Cubuk, E.D.; Sendek, A.D.; Reed, E.J. Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. *The Journal of chemical physics* **2019**, *150*, 214701.

[100]Zakutayev, A. Design of nitride semiconductors for solar energy conversion. *Journal of Materials Chemistry A* **2016**, *4*.

[101] Moustakas, T.D.; Paiella, R. Optoelectronic Device Physics and Technology of Nitride Semiconductors from the UV to the Terahertz: a review. *Reports on Progress in Physics*.

[102] Vurgaftman, I.; Meyer, J.R. Band parameters for nitrogen-containing semiconductors. *Journal of Applied Physics 94*, 3675-3670.

[103] Wu; Junqiao. When group-III nitrides go infrared: New properties and perspectives. *Journal of Applied Physics 106*, 011101.

[104] Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N.E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M. Matminer: An open source toolkit for materials data mining. *Comp Mater Sci* **2018**, *152*, 60-69.

[105] LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **1995**, *3361*, 1995.

[106] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of Proceedings of the IEEE conference on computer vision and pattern recognition; pp. 3431-3440.

[107] Girija, S.S. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. **2016**.

[108] Akimoto, K. Semiconductor device including zinc oxide containing semiconductor film. Google Patents: 2012.

[109]Vermote, E.F.; Tanré, D.; Deuze, J.L.; Herman, M.; Morcette, J.-J. Second simulation of the satellite signal in the solar spectrum, 6S: An overview. *IEEE transactions on geoscience and remote sensing* **1997**, *35*, 675-686.

[110]Atilgan, E.; Hu, J. First-principle-based computational doping of SrTiO $$ _ {} $$ using combinatorial genetic algorithms. *Bulletin of Materials Science* **2018**, *41*, 1.

[111]Atilgan, E.; Hu, J. A Combinatorial Genetic Algorithm for Computational Doping based Material Design. In Proceedings of Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation; pp. 1349-1350.

[112]Atilgan, E. Computational Doping for Fuel Cell Material Design Based on Genetic Algorithms and Genetic Programming. University of South Carolina, Columbia, SC, 2016.

[113]Xue, D.; Balachandran, P.V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. Accelerated search for materials with targeted properties by adaptive design. *Nature communications* **2016**, *7*, 11241.

[114]Lookman, T.; Balachandran, P.V.; Xue, D.Z.; Hogden, J.; Theiler, J. Statistical inference and adaptive design for materials discovery. *Curr Opin Solid St M* **2017**, *21*, 121-128.

[115]Glass, C.W.; Oganov, A.R.; Hansen, N. USPEX—Evolutionary crystal structure prediction. *Computer physics communications* **2006**, *175*, 713-720.

[116]Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications* **2012**, *183*, 2063-2070.

附录

附录 A 攻读硕士学位期间获得的学术成果清单

1. **Yabo Dan(**但雅波**)**, XiangLi, Shaobo Li, Jianjun Hu. Generative adversarial networks (GAN) based efficient sampling of chemical design space of inorganic materials, NPJ computational（SCI 一区 TOP，影响因子 9.8）
2. **Dan Y**(但雅波), Dong R, Cao Z, et al. Computational Prediction of Critical Temperatures of Superconductors Based on Convolutional Gradient Boosting Decision Trees[J]. IEEE Access, 2020, 8: 57868-57878.（SCI 二区，影响因子 4.098）
3. Li S, **Dan Y(**但雅波**)**, Li X, et al. Critical Temperature Prediction of Superconductors Based on Atomic Vectors and Deep Learning[J]. Symmetry, 2020, 12(2): 262. (SCI 四区，影响因子 2.143)
4. Li, X.; **Dan, Y. (**但雅波**)**; Dong, R.; Cao, Z.; Niu, C.; Song, Y.; Li, S.; Hu, J. Computational Screening of New Perovskite Materials Using Transfer Learning and Deep Learning. Appl. Sci. 2019, 9, 5510. (SCI 三区，影响因子 2.217)
5. Zhuo Cao, **Yabo Dan**(但雅波), Zheng Xiong , Chengcheng Niu, Xiang Li, Songrong Qian,Jianjun Hu . Convolutional Neural Networks for Crystal Material Property Prediction Using Hybrid Orbital-Field Matrix and Magpie Descriptors[J]Cystals,2019, 9(4): 191.(SCI 四区，影响因子 2.061)

附录 B 论文中部分核心代码

a MatGAN 模型的训练代码

```
"""
@author: Joker
@file: WGAN.py
"""

import numpy as np
import pandas as pd
import matplotlib
matplotlib.use("Agg")
import shutil
import matplotlib.pyplot as plt
from function.batch_class_gan import Dataset
from function.Evalution import Loss
```

```
from function.Evalution import accuracy
from function.plot_molecule import plot_molecule
from function.Config import Config
import tensorflow as tf
from function.Evalution import evaluation_gan
from function.one_hot_matrix import get_one_hot_matrix
from function.Config import args
from function.Check_Formula import check_formula
import argparse
import time
import os

def load_data(file_path, periodic_table):
    file = open(file_path, "r")
    formula = []
    for ind, form in enumerate(file.readlines()):
        if form == "reduced_cell_formula\n":
            continue
        formula.append(eval(form))

    def statistic(data):
        max_point = 0
        for i, form in enumerate(data):
            if max_point < max(form.values()):
                max_point = max(form.values())
        return max_point

    one_hot_matrix = get_one_hot_matrix(
                                        formula,
                                        periodic_table,
                                        8
                                    )[:,:,:,np.newaxis]

    return one_hot_matrix, formula

def mkdir(path):
    isExists = os.path.exists(path)
    if not isExists:
        os.makedirs(path)
        return True
    else:
        return False

def reduce_molecule(molecular_map, periodic_table, tolerate):
```

```
    periodic_table = np.loadtxt(periodic_table, dtype = str)
    all_for_str, all_for_dic = [], []
    for index, molecular in enumerate(molecular_map):
        formula_str, formula_dic = "", dict()
        for ind, one_vec in enumerate(molecular):
            for j, num in enumerate(one_vec):
                if (num >= (1 - tolerate)) and (num <= 1):
                    formula_str += periodic_table[ind] + str(j + 1)
                    formula_dic[periodic_table[ind]] = j + 1
        all_for_str.append(formula_str)
        all_for_dic.append(formula_dic)
    return np.array(all_for_str), np.array(all_for_dic)

def train_g(sess, model, random):
    g_loss, _ = sess.run(
                    [model.gen_cost, model.opt_g],
                    feed_dict = {
                        model.random : random
                    }
                )
    return g_loss

def train_d(sess, model, config, random, real_data):
    d_lost, _ = sess.run(
                        [model.disc_cost, model.opt_d],
                        feed_dict = {
                            model.real_data : real_data,
                            model.random: random
                        }
                    )
    if config.gan_type == "wgan":
        sess.run(model.wclip)
    return d_lost

def generator_molecule(sess, model, arg, config):
    gen_molecule = []
    for i in range(int(config.gen_num/arg.batch_size)):
        random = np.random.normal(
                            loc = 0,
                            scale = 1,
                            size = [arg.batch_size, 128]
                        )
        generate_sample = sess.run(
                                model.fake_data,
```

```
                                    feed_dict = {
                                        model.random : random
                                    }
                                )
            gen_molecule.append(generate_sample)
        return np.concatenate(gen_molecule, axis = 0)

def main(config, database):
        if database == "OQMD":
            from gan_model.WGANGP_OQMD import GAN
        elif database == "MP":
            from gan_model.WGANGP_MP import GAN
        elif database == "ICSD":
            from gan_model.WGANGP_ICSD import GAN
        elif database == "Bandgap":
            from gan_model.WGANGP_Signal import GAN

        best_rate = 0

        arg = args(config, database)
        gpu_options = tf.GPUOptions(per_process_gpu_memory_fraction = 0.8)

        mo_map, formula = load_data("train_data/" + database + "/train_formula.csv",
config.periodic_table)
        print(mo_map.shape)
        net = GAN(config, database, is_train = True)

        saver = tf.train.Saver(max_to_keep = 1)

        index = Dataset(
                    np.arange(0, mo_map.shape[0]),
                    consider_remain = False
                )

        ev = evaluation_gan()
        model_path = "check_point/iMatGAN/" + database + "/model.ckpt"
        with tf.Session(
                    config = tf.ConfigProto(gpu_options = gpu_options)
                ) as sess:
            saver.restore(sess, model_path)
            # tf.local_variables_initializer().run()
            # tf.global_variables_initializer().run()
            iteration = 0
            while index.epoch != config.times:
```

```
            start_time = time.time()
            for i in range(config.disc_iters):
                ind = index.next_batch(arg.batch_size)
                d_loss = train_d(
                                sess = sess,
                                model = net,
                                random = np.random.normal(
                                            loc = 0,
                                            scale = 1,
                                            size = [arg.batch_size, config.random_len]
                                        ),
                                config = config,
                                real_data = mo_map[ind]
                            )
                ev.updata(d_loss)

            g_loss = train_g(
                            sess = sess,
                            model = net,
                            random = np.random.normal(
                                        loc = 0,
                                        scale = 1,
                                        size = [arg.batch_size, config.random_len]
                                    )

                        )

            if iteration % 5 == 0:
                print(iteration, "g_loss = ", g_loss, "d_loss = ", d_loss)

            if index.check:
                print(ev.total_loss)
                index.reset_check()
                ev.reset()
            iteration += 1

def parse_args():

    parser = argparse.ArgumentParser()
    parser.add_argument("database", choices = ["OQMD", "ICSD", "ICSD_Filter", "Bandgap", "MP"],
                        type = str, help = "choice database to train iMatGAN")
    return parser.parse_args()

if __name__ == "__main__":
```

```
    config = Config()
    main(config, parse_args().database)
```

b CNN 模型的训练代码

```
import numpy as np
from load_data import Dataset, Subset
from evalution import Ev_DL
from model import CNN
import tensorflow as tf
import argparse

def train(sess, model, place, data, args):
    feed_dict =
    _, output = sess.run([model.opt_op, model.outputs], feed_dict = feed_dict)
    return output

def predict(sess, model, place, data):
    feed_dict =
    output = sess.run(model.outputs, feed_dict = feed_dict)
    return output

def main(args):
    data_path = "data/" + args.data_name + ".csv"

    data = Dataset(data_path, "pretty_formula", args.attribute)
    sub_data, val_data = Subset(np.arange(len(data)), 0.9, 520)
    train_data, test_data = Subset(sub_data.idx, 0.9, 520)

    data_shape = data[0][1].shape

    placeholders = {
        'features': tf.placeholder(tf.float32, shape=(None, data_shape[0], data_shape[1], 1)),
        'labels': tf.placeholder(tf.float32, shape=(None, 1)),
        "lr": tf.placeholder(tf.float32)
    }
    name = args.data_name + "_" + args.attribute

    net = CNN(placeholders, True, name = name)
    re = []
    best_re = [0, 0, 0]
    with tf.Session() as sess:
        tf.local_variables_initializer().run()
```

```
            tf.global_variables_initializer().run()
            ev_train = Ev_DL()
            while train_data.epochs != args.epochs:
                train_batch = data[train_data.next_batch(args.batch_size)]
                pre = train(sess, net, placeholders, train_batch, args)
                ev_train.updata(train_batch[-1], pre)
                if train_data.loop_time % 40 == 0:
                    ev_test = Ev_DL()
                    while not test_data.end_one_epoch:
                        test_batch = data[test_data.next_batch(args.batch_size)]
                        pre = predict(sess, net, placeholders, test_batch)
                        ev_test.updata(test_batch[-1], pre)
                    test_data.end_one_epoch = False
                    print(train_data.epochs, ev_train() + ev_test() + best_re)
                    re.append(ev_train() + ev_test() + best_re)

                    ev_train = Ev_DL()
                    if ev_test.R() > best_re[-1]:
                        best_re = ev_test()
                        net.save(sess)
            np.savetxt("result/" + name + ".csv", re, delimiter = ",", fmt = "%.3f")

        compare = []
        with tf.Session() as sess:
            net.load(sess)
            ev_val = Ev_DL()
            while not val_data.end_one_epoch:
                val_batch = data[val_data.next_batch(args.batch_size)]
                pre = predict(sess, net, placeholders, val_batch)
                ev_val.updata(val_batch[-1], pre)
                compare.append(np.concatenate([pre, val_batch[-1]], axis = 1))
            print(ev_val())
            compare = np.concatenate(compare, axis = 0)
            np.savetxt("result/" + name + "_compare.csv", compare, delimiter = ",", fmt = "%.3f")

def parse_args():
    parser = argparse.ArgumentParser()
    parser.add_argument("--data_name", type = str, choices = ["ICSD", "MP", "OQMD"])
    parser.add_argument("--attribute", type = str, choices = ["band_gap", "stability", "delta_e",
"volume_pa", "dosf"])
    parser.add_argument("--lr", type = float, default = 0.001)
    parser.add_argument("--epochs", type = int, default = 1000)
    parser.add_argument("--batch_size", type = int, default = 512)
    return parser.parse_args()
```

```
if __name__ == "__main__":
    main(parse_args())
```

附录 C 筛选出的材料清单

Formulas	FE(eV/atom)	BG(eV)	Formulas	FE(eV/atom)	BG(eV)
SnHoPbSe4O5C	-1.584	2.000	CsLaPb2S2O2	-1.911	1.894
TiNbRuSn2AsO7	-1.308	1.997	ScUPuAs2O6C	-2.031	1.891
Sc2Sn3O6	-2.012	1.994	LiCuPuS2NC	-0.931	1.888
NaK2HoB4H4	0.642	1.991	NbSn3SbO3B4	0.725	1.761
SnPuAs2O3	-1.426	1.985	SnBaPb2S3	-1.192	1.755
Cu2Br6As2O8C3	-0.975	1.982	NbSn2IrBr6O5	-1.549	1.752
Al5Cu2TaO7	-1.919	1.979	Li4RbDyCl4	-1.614	1.749
NbSnSPO4	-2.224	1.976	LiHgC	0.085	1.747
RuAsO6C	-1.549	1.973	La3DyPb2Bi2Br6Cl6	-1.963	1.744
NdPt2Te2AsS4	-0.730	1.970	KScTiCoBa2O6	-2.806	1.741
AlVHoO4C	-2.616	1.967	Sb2Cl2S8C	-0.481	1.738
CrYRuCsAuPbO3	-0.565	1.964	AlLaXe2S2N	-1.091	1.736
Sn3WHg2AsClS3	-0.203	1.961	NbThBr6S2NC	-0.953	1.733
CrDyAs6Cl6NC	-0.810	1.952	Li6AgLaPbS3	-0.879	1.724
CrSn2Pb2As3S2O2	-0.262	1.949	ZnCsO4N	-1.419	1.722
LiCoNbInSnNdO4	-1.714	1.946	CrCs2LaBi2S4O6	-2.072	1.719
Sn2NdAs4B2	0.748	1.943	LaSe2S4C	-0.583	1.716
Pt2Ge2O2	-0.447	1.940	SrTaGe2F3	-2.336	1.713
ZnYbAs3S2O5B2	-1.395	1.937	FeYRuSnAsO4	-1.229	1.710
NdPuAsS2O8	-2.453	1.934	Cu2NbRuSeAs2O6	-0.797	1.708
WPb2N3	0.106	1.928	Sn2CeAuSeGe2O4	-1.457	1.702
NbBa2YbPbTeO4	-2.570	1.926	Sn2DyXe2	-0.478	1.699
MnZnDyTaO5	-2.761	1.923	PuGe2S8	-0.782	1.697
Sn2SO4N3	-0.988	1.920	CuNbTaSe4O4	-1.400	1.694
CoBa2AsO6H2	-1.951	1.917	Sn8NdCl6S2	-0.914	1.691
Mg2MnSb2O6	-1.959	1.914	Co2NbCeTe4SeP2O4	-1.282	1.688
LiNbSnLaTmS2O3	-2.684	1.911	RuPb2O2	-0.274	1.685
LiMn2S6NC	-0.627	1.908	K2CrGa3TaF	0.429	1.683
CuSn2NdS2F	-1.714	1.905	VBaXe2F2O4	-1.860	1.680
NbSn8BiClS4O2	-0.330	1.902	Li4RhSbS2	-1.114	1.677
LaYbPbAs4S2	-0.786	1.899	Cu2YbBi2TeSbCl6S3	-1.223	1.674
LaAs2Cl5	-1.678	1.672	LaAs2Cl5	-1.678	1.672
KNdWPuAs2S2O6	-2.202	1.669	KNdWPuAs2S2O6	-2.202	1.669

LiNbTaAuCl2S2O3	-1.959	1.666	LiNbTaAuCl2S2O3	-1.959	1.666
LiCrCoCs2NdSO3	-1.709	1.660	SeAsCl2S2N	-0.263	1.572
MoTaO4B2	-0.633	1.658	NaSe3S4	0.334	1.570
NbRuYbTaCl7O5	-2.060	1.655	NbPd2Cd2AsS3	-0.588	1.567
AlDy2H6	-0.425	1.652	K2Co3SnTaSbO5C	-1.257	1.564
HoBi2Te2S4B2	-0.572	1.647	SeAsCl2S2N	-0.263	1.572
Cu2Dy3As4F3	-1.852	1.644	NaSe3S4	0.334	1.570
MoEuPuO6	-3.030	1.639	K2Co3SnTaSbO5C	-1.257	1.564
Ni2Pb2SeCl6S2	-0.861	1.633	SeAsCl2S2N	-0.263	1.572
CrNbRuAgS2O8	-1.876	1.631	NaSe3S4	0.334	1.570
Sn2LaBi2Se2As3	-0.274	1.628	NbPd2Cd2AsS3	-0.588	1.567
Li2Br6As2O2	-1.421	1.623	NiAuSe4S2O3C3	-0.360	1.364
Te3As4S3	-0.158	1.620	ZnNbSn2S6	-0.819	1.361
Ba2DyAs2S4	-1.675	1.618	Cu2NbDy3S4O2	-2.257	1.359
AlWAs8S2	0.349	1.615	NiAuAsCl6	-0.842	1.356
PbSe6S4O3	-0.586	1.612	Na2Rh2Ge2S2O5	-1.549	1.353
CoAsS5F	-0.679	1.607	Y2ZrLaPbO6	-3.137	1.348
AgSnWTeS2	-0.434	1.604	CoRuInPbAs2S2N3	-0.310	1.345
Sr2PbGe2S2	-0.876	1.601	Cu2Pb2SbF3	-1.321	1.343
InPbAs2S2N	-0.438	1.596	NiAgSn2Ge2S2F6	-1.812	1.337
SrIrBr6As2	-1.037	1.593	ZnLaCeAs4S2O8N2	-1.748	1.335
NbSePO3	-2.057	1.591	CuTaPuO8	-1.885	1.332
MoSn2LaS6C	-0.999	1.586	AlCuBi2AsS2	-0.011	1.327
Sn3AuF6	-2.335	1.583	Y2LaSe4Cl2S2N	-2.050	1.324
Sr2NbSn2Cl2S2	-1.902	1.580	DyBr6As2Si2H2	-0.702	1.321
Br2Se6C	0.260	1.318	Al3SnDy2Se2O2	-2.148	1.235
CrSn3Ba2O6	-1.993	1.316	NbSe2N6	1.115	1.232
ZrPbCl8S3Si	-1.547	1.313	Sn2Cs2EuThCl2O6	-2.659	1.230
Pd2TaBr2Se2	-0.741	1.310	W6S2O6N	-1.153	1.227
MnZnNbHoAs2O4	-1.547	1.305	K2MnTe2	-0.765	1.222
AgInF6O6N	-1.061	1.302	Na3PuS2O4	-2.181	1.219
Sn2IrTl2O4	-0.839	1.300	SnLaS2ON	-1.802	1.216
Sn2PbBiPuAs2S2C	-0.076	1.297	KCrDy2AuTeSO8	-2.377	1.214
CoGa3MoS2F	-0.736	1.295	Mg2MnDy2O6	-2.624	1.211
VPbXe2O	-0.231	1.292	CrPd2CsSbS4	-0.808	1.208
CrNbAgSn2O2	-0.132	1.289	CrZnBrAsS4	-0.697	1.205
CrAs2S4B6	0.527	1.287	V2ZnSnTaSi2O3N3	-1.433	1.203
Ti3CuAs2S2	-1.057	1.284	NbSn2HoPbBiS3	-0.385	1.200
LaGe2H4	-0.398	1.281	Sr2LaPb2Cl2O4	-2.375	1.197
Ga3PdAgS2O6	-1.546	1.279	Nb3BaIrBr2S2	-1.419	1.195
CeBr6S2O2C	-0.964	1.276	CrZnCs2Ba2O2	-0.185	1.192
Ga2As2S3	-0.479	1.270	MnDy2SbO4	-2.789	1.187

PbPuS6O5	-1.498	1.268	USeAs6F	-0.376	1.184
SnBi2SO2	-0.853	1.265	ZnYNbO4	-2.777	1.181
In3Br6GeO6	-1.552	1.262	AlAgLaPbS2	-0.949	1.179
LiNbRuSnSe2S4	-0.963	1.260	Li2Sn2LaTaPbS4O2	-1.712	1.176
Sn3LaS2O6N7	-1.100	1.257	NbSn2Ba2Dy2GeO4	-2.785	1.173
K2MnBaS2C	-0.931	1.254	AgInSnLaPbS2	-0.618	1.171
MnPbAs2O4H4	-1.141	1.251	Rh2Sn2BrAsCl6	-1.119	1.168
NiAgCs2LaS3N3	-0.971	1.243	CrCu3Dy2S2O4N3	-1.468	1.160
CrZrSn2P2O2	-1.158	1.240	FeWH4	0.110	1.157

附：贵州大学学位论文原创性声明和使用授权声明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的科研成果。对本文的研究在做出重要贡献的个人和集体，均已在文中以明确方式标明。本人在导师指导下所完成的学位论文及相关的职务作品，知识产权归属贵州大学。本人完全意识到本声明的法律责任由本人承担。

论文作者签名：但雅波　日　期：2020 年 6 月 12 日

关于学位论文使用授权的声明

本人完全了解贵州大学有关保留、使用学位论文的规定，同意学校保留或向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅；本人授权贵州大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或其他复制手段保存论文和汇编本学位论文。

本学位论文属于：

保　密（　），在　　年解密后适用授权。

不保密（√）

（请在以上相应方框内打“√”）

论文作者签名：但雅波　导师签名：

日　期：2020 年 6 月 12 日