摘要

目前双碳政策持续发力，材料信息学广泛地利用大数据驱动和机器学习，它们会需要大规模材料特性信息的数据。本文提出了一种基于小数据样本的**数据生成**回归预测模型来预测材料的新方法。在本文的案例中，以**固体胺** CO_2 吸附剂在直接空气捕集的应用，基于表格变分自编码（TVAE）和条件表格生成对抗网络（CTGAN）为数据生成模型框架，通过输入材料的物理、化学特性以及实验条件生成数据，再与原数据训练集整合为新的训练集，这其中与原数据的训练集结合，防止了数据泄露，加强二氧化碳**吸附剂**性能预测泛化的可行性，再运用经典**机器学习**的回归模型（随机森林）进行预测，发现误差相对不使用生成数据模型要下降 9%(TVAE)~12%(CTGAN)。最后进一步采用 **SHAP 分析**量化输入的特征，发现了吸附前后的孔隙体积和吸附剂气体承载量特征性能对吸附容量影响最大。本研究揭示了生成数据模型在数据有限以及小样本材料研究的潜力，能够有效地突破数据瓶颈以及成本有限的条件下提供解决办法，加速新型碳捕集材料的开发。

关键词：数据生成，固体胺，吸附剂，机器学习，SHAP 分析

Abstract

With the current dual-carbon policy continuing to gain momentum, materials informatics makes extensive use of big data-driven and machine learning, and they will require data with large-scale material characterization information. In this paper, we propose a new approach to predict materials based on data-generated regression prediction models for small data samples. In the case of this paper, the application of solid amine CO_2 adsorbent in direct air capture is used as a data generation modeling framework based on tabular variational autocoding (TVAE) and conditional tabular generative adversarial network (CTGAN), which generates the data by inputting the physical and chemical properties of the material as well as the experimental conditions, and then integrates it with the original training set of the original data into the new training set, which which combines with the original training set of the original data, prevents data leakage and enhance the feasibility of generalization of CO_2 adsorbent performance prediction, and then applying the classical machine learning regression model (Random Forest) for prediction, it was found that the error decreased by 9% (TVAE) to 13% (CTGAN) relative to not using the generative data model. Finally, SHAP analysis was further used to quantify the input characteristics, and it was found that the pore volume before and after adsorption and adsorbent gas carrying capacity characterization properties had the greatest impact on the adsorption capacity. This study reveals the potential of generative data modeling in data-limited as well as small-sample material studies, which can effectively break through data bottlenecks as well as provide solutions under cost-limited conditions to accelerate the development of novel carbon capture materials.

Key words： data generation, solid amines, adsorbents, machine learning, SHAP analysis

目录

1 绪论

1.1 课题背景以及意义

根据联合国政府间气候变化专门委员会（IPCC）第六次评估报告综合报告（2023），热浪、强降水、干旱等极端天气发生的频率以及强度自从上世纪以来显著提高。而频发的极端天气归结于气候变暖，这主要是人类的活动加剧了温室气体的排放，使得 2011 年至 2020 年全球平均温度较工业化前即 1850 年至 1900 年升高了$1.1°C$，而陆地升温幅度要明显高于海洋，这深深的破坏了地球的环境，同时人类明确的活动所排放的温室气体导致极端天气的频发又正在反噬了人类的衣食住行，长此以往下去，这样子会对地球带来诸如生态系统崩溃相关的连锁反应以及类似冰川消融的不可逆变化。为了挽回这样的局面，巴黎协定旨在将全球气温升幅控制在工业化前水平的$2°C$，并努力限制在$1.5°C$以内，以显著降低气候变化。不仅如此，巴黎协定还提出了碳中和的愿景，目的在于在本世纪下半叶实现全球温室气体净零排放，推动经济与气候人性深度脱钩。然而当前各国国家自主贡献（NDCs）总和仍然导致$2.5°C$到$2.9°C$升温，这意味着各国依然需要大幅提升行动能力；并且气候资金缺口依然显著，许多国家对于绿色发展的适应需求依然未得到充分的满足。以二氧化碳为主的温室气体是全球气候变化加剧的首要元凶，减少二氧化碳排放已成为了国际社会共同关注的焦点。碳捕集、利用与封存技术是缓解温室效应的关键方法之一，其中高效二氧化碳吸附剂的开发是当中的核心环节。传统的吸附剂开发依托实验的方式以及不断调整吸附剂物理、化学性质以及实验条件进行筛选，这样的长周期开发不仅效率低并且将会导致额外的成本提高。近些年来，机器学习在材料领域展现出强大的能力，它能够将通过实验获得的材料数据挖掘材料当中的构造关系，加速高性能吸附剂的设计，不仅如此当考虑材料的特性较多时候，可以将多种特性转为多维度的矩阵，加速高性能吸附剂的设计。但是现有的研究集中在单个物性的预测，并且基于这样的预测之中，所得到的数据质量以及模型的泛化能力依旧有待提升，并且小样本量的实验数据的物性预测能力也并不良好——实验数据稀缺、分布不均衡、多维特征关联性弱，严重制约着模型的泛化能力以及预测精度。因此基于机器学习构建 CO_2 吸附剂小样本量的物性预测模型，如何通过数据增强技术突破小样本的限制，建立一个高可靠性的预测模型，将会是加速吸附剂设计的关键挑战。

1.2 研究目的与内容

受限于一般情况之下由于成本的原因以及相关的实验操作失误导致的实验数据作废，本研究以吸附剂的物理性质、化学性质和实验条件三组物性关键研究因素，旨在开发一套面向小样本的二氧化碳吸附剂物性预测的框架，通过数据增强技术与机器学习算法，解决吸附剂开发中数据稀缺问题，同时通过所增强得到的模型所测试的物性性能，结合数据的可解释手段工具解析模型决策的逻辑，明确材料的设计规则。将小样本数据增强技术与材料跨尺度特征建模结合，打破了传统机器学习在吸附剂领域的数据依赖，为小样本数据的材料预测场景提供新的范式，加速二氧化碳吸附剂的只能筛选与设计，降低实验成本，为应对气候变化提供坚实的材料基础，具体的目标包括如下：

构建小样本增强的策略：针对吸附剂数据的多维度特性（如孔结构特征、表面化学特性、实验进行条件），设计基于条件表格生成对抗网络（CTGAN）或者基于表格的变分自动编码器（TVAE）的增强模型，生成高保真合成数据，扩充训练集的规模并且有效地提升特征的多样性。

建立物性预测模型：由于数据量的稀缺，在小样本数据的场景之下优先选择复杂度低、抗过拟合能力强的模型，有利于避免了因为数据量不足所导致的模型泛化性差。基于此因素，有以下基线模型将会加入考虑：线性回归、随机森林、梯度提升树等等模型。除此之外模型将会使用网格搜索方式进行超参数的优化，相关的继承学习以及深度学习模型将会限制树的深度以及学习率，对于评估模型的性能指标是均方误差（MSE）、平均绝对误差（MAE）以及决定系数（R2）。

验证技术的有效性：基线模型与通过数据增强的基线模型将会被对比，除此之外通过数据增强所获得的数据会将通过概率分布的是否具有一致性以确保合成数据没有偏离真实数据的流形，另外原始数据与生成数据的相关性也要类似。还有增强数据是否会缓解小样本情况下的过拟合或者欠拟合表现。

指导实验的设计：基于模型反向解析的关键特征，通过模型的可解释分析手段提取关键的特征贡献度，提出基于关键特征贡献度的新型吸附剂优化设计方向，有利于缩短实验的验证周期，减少实验的相关次数，推动高性能材料的快速研究与开发。

1.3 二氧化碳捕集技术的概述

二氧化碳捕集，是指从工业排放源或者大气中分离以及捕获二氧化碳的技术过程，是碳捕集、利用与封存（CCUS）体系的核心环节。当前全球气候危机显著加剧，减少温室气体的排放是目前来说十分紧迫的任务。而国际能源署（IEA）指出，要实现巴黎协定的温控目标，到 2050 年全球通过 CCUS 技术每年捕集 76 吨二氧化碳，占累计减排量的 15%以上。燃烧前碳捕集、燃烧后碳捕集和直接空气捕集方法是捕集二氧化碳的技术。燃烧前捕集是通过气化与水煤气变换反应将碳基材料转化为氢气与二氧化碳，分离二氧化碳之后将氢气用于发电，譬如目前将联合循环使用在燃气轮机上不仅起到吸附二氧化碳同时降废气利用以及高品位的热量重整提高了燃烧前吸附中的联合循环的整体能源和温室气体减排的效率，将在脱碳时间表中提供有建设性的拓展性能[1]，而目前为止，大多数燃烧前捕集项目的设计是捕集 90%的二氧化碳，但在 2021 年美国能源部的点源捕获计划提高了二氧化碳的目标捕获率达到 95%，这为利用二氧化碳选择性聚合物膜的系统是否可以提高吸附率留下了一个疑问，Lie Meng 等人通过建立单级膜气体分离模型，通过模拟揭示使用真实工业合成气实现高性能 CO2 捕集过程所需的最佳条件，并强调了在氧气吹制和空气吹制 IGCC 过程中促进运输膜去除 CO2 的潜力，这为燃烧前吸收效率是否通过使用膜提供了理论基础[2]。燃烧后捕集主要从燃烧延期当中分离大约浓度为 10%到 15%区间的二氧化碳，主流的技术包括胺液化学吸收，该方法利用胺液与二氧化碳发生化学反应来实现吸收；物理吸附通过利用沸石或者活性炭进行吸收；此外，还有膜分离的方式。目前新型胺的开发针对传统胺液在捕集过程中存在的高能耗。燃烧后捕集在一次水泥厂的应用效果中通过将传统与新型方式比较，在传统 MEA 吸收法与新型硅胶-聚乙烯亚胺（SPEI）吸附法均实现了二氧化碳 90%的捕集率，然而进一步研究发现，后者在综合能耗指标上均高于传统的 MEA [3]。物理吸附法则是借助沸石或者活性炭等材料进行吸收，沸石与其他吸附剂相比，因其重复循环的稳定性、较大的表面积和快速的二氧化碳吸附动力学，而活性炭主要优点在于成本效益高、解吸所需的低温、直接的再生过程、低能耗、快速 CO_2 吸附动力学、高热稳定性和化学稳定性、机械强度或出色的导热性。出色的多孔结构、不同的孔隙率和比表面积这些成为了二氧化碳物理吸附的重要材料[4]。直接空气捕集是从大气中直接捕集浓度大约为 420 分压力浓度的二氧化碳，一般采取固态胺以及 MOFs，其目的是在于实现负排放，抵消难减排部门的遗

留排放。在当中固体胺吸附剂比较容易受到温度以及湿度的影响，在低温和相对湿度较高的条件下具有更好的 DAC 性能，同时通过对固体胺吸附剂添加适当的表面活性剂以及纳米花材料会对二氧化碳吸附具有协同效应提升吸附能力[5]，另外有机金属框架（MOFs），为了解决与通常以来有毒溶剂和盐类的传统合成方法相关的环境问题，采用一种利用环保溶剂的绿色合成方法，这一举措使得吸附剂表现出增强的表面积与体积比[6]，进而大幅提高了二氧化碳的吸附容量。

1.4 机器学习（ML）概述

数字化的时代飞速发展，数据正在以前所未有的速度不断累积，而机器学习作为一门让计算机通过数据学习数据的发展模式并做出相关预测的科学，正在成为推动各个领域创新发展的核心技术。

机器学习的定义比较宽泛，它通过算法给予计算机系统一种能力，让计算机无需针对特定任务进行明确的变成，就可以使得其在数据之中自动学习规律并进行预测以及分类。简而言之，通过对海量数据进行深度挖掘与分析，机器学习算法能够有效识别数据中隐含的规律性特征与潜在趋势，并以此为基础构建预测模型，从而实现对未知数据的精准预测与智能决策支持。

在机器学习领域，根据训练范式可划分为监督学习、无监督学习和强化学习三大主要分支。监督学习方法的核心在于利用带有标注的训练数据集，通过建立输入特征与目标输出之间的映射关系，进而实现对未知样本的标签预测。该范式下的典型算法可进一步细分为回归模型和分类模型两大类别，具体涵盖线性回归、决策树、支持向量机以及神经网络等多种经典算法实现。在图像算法领域当中，监督学习可以通过大量已标注的图像数据学习不同物体的特征然后可以基于一些神经网络上的改动与堆叠、结合进行譬如目前在医学上较为广泛应用的有如 B 型超声图像的肿瘤分类[7]。无监督学习则是处理没有标记数据，它的主要目标是发现数据当中的内在结构和模式。聚类算法，像 K-Means 聚类（K 均值聚类，其中 K 为想要通过数据分开的类别数量），能够将相似的数据聚合成一个组别，最常见的是在客户细分、市场调研等方面的广泛应用，如活跃在人人手机上的旅行软件，通过区分旅行者的决策行为，建立模糊无处不在的旅行者聚类和酒店推荐[8]。强化学习则聚焦在智能体如何在环境中采取一系列行动，以最大化累积奖励。智

能体通过与环境不断交互，根据获得的奖励反馈学习到最优的策略，例如 2017 年举世震惊的 AlphaGo 在围棋博弈中，表现出对长期收入的高度敏感性[9]，通过无数次与自己博弈、学习不断探索和优化落子策略，历经海量的训练，最后战胜了人类顶尖棋手。

然而，机器学习的发展也面临诸多挑战。数据质量是关键问题之一，若数据存在噪声、缺失值或偏差，可能导致模型学习到错误的模式，导致数据与学习目标欠拟合，此外当数据的维度过大，也同时会导致模型学习失效。模型的可解释性也是当前研究的热点和难点，复杂的深度学习模型在做出决策时，其内部机制往往难以理解，因为复杂的神经网络模型由大量的神经元相互连接形成复杂的网络结构，当输入数据经过众多在神经网络里面的隐藏层的层层变换得出决策，整个神经网络模型内部就像一个黑箱模型，这在一些对决策可解释性要求较高的场景让人难以确切知晓模型是根据什么特征并计算得出这样的结论，因此成为应用障碍。此外，机器学习模型的训练通常需要大量计算资源，如何提高计算效率、降低能耗也是亟待解决的问题。尽管如此，机器学习依然以前所未有的速度改变热呢生活和工作方式，随着技术的不断完善与创新，它将会在更多领域发挥其本身更大的价值。

1.5 机器学习在吸附剂运用的国内外研究现状

在全球积极应对的气候变化中，致力于二氧化碳排放的大背景环境之下，高效的二氧化碳吸附剂的开发是至关重要的。机器学习在各个领域的深入使用，其强大的数据驱动技术，也慢慢地成为研究二氧化碳吸附剂性能的关键手段，国内外许多科研团队也正在围绕这交叉的课题展开丰富的研究讨论。

而人工智能辅助材料设计的进展在对物理学、化学和材料学的研究领域起着重大的效应，比如通过使用蜂胶提取物油处理对体外气体产生的过程进行建模，以解决最佳的蜂胶处理[10]。通过持续地大数据和算法从广泛的数据集中加速计算学习过程中，机器学习做到了对于减少温室气体的排放。机器学习的方向可能更有助于通过算法以及其余的智能衍生算法揭示数据当中的隐藏关系，目的是将各个吸附剂的描述特性与气体的吸附性能进行相关联。其优势在于在各种规模上可以通过材料的成分配比去发掘最影响因素[11]以及特性进行参数的调整与优化。机器学习可以帮助验证简单的固体吸附剂，像是多孔碳、活性炭等等材料的实验结果的吸附，而分子动力学的引入可以有效性的协同机器学习的作用。在单一的机器学习或者机器学习与分子动力学的组合当中，通常的研究流

程则是 1）吸收二氧化碳 2）优化吸附剂的制造流程、通过筛选不太好的变量以防多重共线性的影响以及二氧化碳在材料当中的吸附量 3）通过单一模型或者混合模型揭示材料在吸附二氧化碳层面的特征强度以及有效性水平 4）通过运行出来的结果筛选以及发现新的优质吸附剂。在固体吸附剂的层面上，它主要是由三种特性决定其在二氧化碳吸附能力性能，由于通过多种的前驱体、热化学转化和化学试剂合成。这些层面的条件，为吸附剂的结构提供了理想的几何、相关原子官能团等等特性。基于于此多个关于机器学习的项目孕育而出，这些特性成为了在电脑程序以及机器学习或者其混合模型研究吸附剂特性的可调参数。建模数据则是由上述材料特性和属性特征作为输入的组要类别，可以通过实验或者数据库进行提取。像 X.Yuan 等人在 2021 年运用多个机器学习模型进行交叉检验和超参数调整以多个物理化学做特征和数据点，解释了特征的影响和气体捕获值的预测[12]。不仅如此 Fathalian[13]等人进一步提出基于多种机器学习算法的智能预测框架，用于预测石墨烯氧化物（GO）基吸附剂的 CO_2吸附能力。研究整合了来自 17 篇文献的 895 条实验数据，使用比表面积、孔体积、温度和压力等输入参数建模，并发现 ANN-MLP 模型预测性能最优，同时研究还可是花了温度以及吸附剂孔结构与吸附性能的关系，指出了面积和孔体积远比化学组成更能解释 CO_2吸附性能。而像 X. Ma[14]等人与 C. Zhang[15]等人分别针对不同的压力之下以及特定的温度之下对吸附剂最大吸附能力进行了很好的预测。而在多个维度之上，采用机器学习模型选择压力、元素和围观结构特征，对之进行 SHAP 分析，去探究各个元素的比例，说明在高维度的情形之下机器学习的适用性也要使用更复杂的模型[16]。混合模型上，基于数据库收集下来的数千个使用遗传算法实现了高吸附的目标[17]，迁移学习通过在大型的数据进行模型的训练再运用回小数据之中，这在一次聚合物以及金属有机框架中也提高了预测的准确性能[18]。现在 AIGC 正如火如荼，在此基础上 Li 等人[19]首次提出将大语言模型应用在材料性能预测中，结合数值类数据以及文本信息，通过 ChatGPT-4o 模型对固态胺二氧化碳吸附剂性能的上下文建模，提出的“三步建模法”整合数据收集、Prompt 设计与模型预测流程，显著提高了预测精度（$R^2 = 0.81$，$MAE = 0.29$），并借助 SHAP 分析确认胺负载量、比表面积及孔容为关键影响因素。这篇研究不仅展示了大语言模型在材料预测中的新范式，也拓展了传统 ML 模型在数据稀缺情况下的适用范围。除此之外，在金属有机框架的领域中，多结构特点[20, 21]会作为机器学习模型的变量，会发现机器学习在评估吸附剂的性能潜力上面以及吸附剂的选择潜力上，机器学习无论是单一模型还是混合的模型对比传

统的模型具有更高的潜力以及更有效的性能。进一步的在材料的研究中，机器学习的应用不仅发掘出结构设计在材料开发的重要性[22]，而拓展使用更多的机器学习模型，如 Amirkhani 等人则将研究拓展至新型多孔液体当中，在他的研究里，他所使用的模型证实出多孔结构中的能力更泛化与强劲[23]。除去 MOFs 和多孔材料，Zhang 等人[24]专注于商业多孔载体负载有机胺的固态吸附剂，采用随机森林与 SHAP 值分析明确了影响 CO_2 吸附性能的主要因素，其中胺负载量和孔体积为主要贡献因子，并实现了对新型载体吸附性能的准确预测。而另一项工作中，作者比较了 TEPA 和 PEI 两种胺在不同孔结构上的吸附行为，发现孔径大小对胺效率有显著影响，且胺类型对吸附剂的循环稳定性至关重要。在基于预测热力学的传统方案上，机器学习已经被证明是经典热力学的一种非常具有建设性的替代方案，通过使用三种模型，Peyman Pakzad[25]等人在关联实验中的数据当中，发现人工神经网络不仅表现良好，在与实验二氧化碳符合数据吻合程度上也要交两种修正的模型来说要优秀。Venkantraman 和 Alsberg 开发一个机器学习模型对二氧化碳在吸附剂溶解度的预测，他们将 100 种分子特性，譬如电荷密度、偶极矩、电子相互作用等被用做新型的胺类碳捕获替代品的分子描述符[26]，其指代将分子结构分解利用一系列数值描述的物理化学、性质，根据这个替代品的描述符，他们通过决策树和随机森林发现影响吸附能力溶解度强度的是二氧化碳在这个给定替代品中的分子相互作用强弱，可见机器模型在挖掘关键因素的重要列[27]，并且机器学习及其混合变形模型正在吸附剂领域越来越展现其独特的能力。

2 基于数据生成的吸附剂物性预测理论

这项工作里面提出了一种基于小规模数据的生成模型预测二氧化碳吸附剂吸附性能的方法。在数据的手机过程中，需要根据文献以及数据库中的实验测量的材料特性以及性能参数，将其分别作为预测模型的输入以及输出特征，结合专家人员的判断构造出一个数据集出来。在数据手机步骤中，从文献报告以及结果获得的数据分别为数字数据以及文本数据。然而数值数据主要是在于结果部分，通常得益于文献通过实验以及仿真的报告的材料特性以及性能得出。文本数据则多半是来自于文献的讨论讨论部分，一般由专家以及相关学者所提出的结论组合而成。

本文所提出的方法预测二氧化碳的直接空气捕集法下固体胺吸附剂的媳妇吸收。将基于小数据规模与生成对抗网络回归类结合使用可以显著地解决在数据规模小以及成本有限的条件之下的性能预测、提高领域预测的准确性。输入特征包括固体胺吸附剂的质构和成分特性以及直接空气捕集的条件下的吸附参数，对应的输出则是二氧化碳的吸附吸收量。通过提供一定量小的数据规模，生成数据模型可以学习数据自己内部的构造关系以及分布关系，使用之后的模型，在该模型生成的数据预测之后可以预测二氧化碳新材料的特性吸附新材料，除此之外，不仅可以解决实验条件有限之下数据不足的预测问题，还可以有效提高相关的吸附预测性能。

2.1 研究思路

随着全球变暖问题日益严重，二氧化碳减排和捕集成为热点研究方向。吸附技术因能效高、成本低、适应性强而在碳捕集中具有重要应用前景。开发新型 CO_2吸附剂材料的关键在于了解其结构与性能之间的关系。但实验测量耗时费力、数据获取困难，导致目前可用的吸附剂物性数据样本量有限，严重制约了传统机器学习模型的准确性和泛化能力。因此，如何在小样本条件下实现有效的物性预测，是本课题需要解决的核心问题。

本研究以机器学习为手段，旨在建立一种能在小样本条件下仍具备良好预测能力的CO_2吸附剂物性预测模型。具体思路包括：构建基于真实实验数据的预测模型，在此基础上引入生成模型对数据进行增强，通过合成数据扩大训练样本，从而提升模型性能。研究将采用两种典型的表格数据生成方法——CTGAN（Conditional Tabular GAN）和TVAE（Tabular Variational Autoencoder），并分别与未增强数据模型进行对比分析，以评

估其效果与可行性。

首先，研究将整理与预处理 CO_2吸附剂的结构及性能数据，包括比表面积、孔径大小、极性参数、化学组成等结构特征，以及 CO_2吸附量作为目标属性。在完成数据标准化预处理及特征工程构建的基础上，本研究选取了随机森林、线性回归、决策树以及梯度提升树等具有代表性的经典机器学习算法进行建模分析。接着利用相关统计指标评价其初始性能表现。

然后，训练 CTGAN 和 TVAE 模型以学习原始样本的特征分布，分别生成多个与原数据相似的新样本。为验证生成数据的质量，将采用概率分布统计绘制出相关的直方图、概率密度分布以及均值和标准差方法对比原始与合成数据的分布差异，确保生成数据在统计意义上具备合理性。接着，将原始数据与生成数据合并，重新训练预测模型，并与原始模型在相同测试集上的表现进行比较，考察合成数据对模型预测能力的提升程度，这样子还可以防止数据泄露导致模型的泛化能力有所弱化

最后，研究还将通过 SHAP 可解释性工具分析重要特征的贡献，确保模型不仅“好用”，而且“可信”。此外，还将探讨不同生成方法在不同特征维度与样本分布下的适用性差异，为后续材料数据增强以及吸附剂性能优化方面提供相关的参考。

2.2 研究方法与技术路线

2.2.1 数据准备与特征工程

数据来源可以选自公开数据库（如 Materials Project、CO_2Adsorbents.org）或文献爬取后以及开源论文提供材料里的吸附剂数据。原始数据应包含吸附剂的结构参数、化学组成、比表面积、孔径大小、表面极性等输入特征，以及目标输出属性如 CO_2吸附量、选择性、热力学稳定性等，除了前面吸附剂的物理性质以及化学性质之外，还有吸附剂在实验条件之下的温度以及相关的湿度和二氧化碳的分压力。

特征工程方面包括：缺失值处理与异常值检测；数值标准化与分类变量编码；高基数类别的数据需要进行对数化

2.2.2 基准模型的构建

在不使用任何合成数据的前提下，采用常规机器学习算法（如随机森林、XGBoost、

SVR、MLP、线性回归等）构建基准模型，作为对比参照。该模型通过交叉验证评估其在原始小样本数据集上的预测能力，获取初始的性能指标（如 RMSE、MAE、R^2等）。

2.2.3 数据增强方法

为缓解样本不足的问题，引入两种代表性的表格数据生成方法，这两个合成数据模型将采用 sdv 库进行相关的运用：

CTGAN（Conditional Tabular GAN）：通过条件生成对不同类别、特征分布进行建模，适用于非平衡、异构类型混合的表格数据生成。

TVAE（Tabular VAE）：基于变分自编码器的结构，对输入数据建模并通过潜在空间采样重建新样本，具备较好的连续性和可控性。

生成步骤包括：

模型训练：在原始样本上训练 CTGAN/TVAE，确保生成数据分布与真实数据相近；

数据合成：生成一定数量的高质量合成数据（可与原始数据进行组合训练）；

数据验证：通过分布比较与比较分布中的平均数和标准差确认生成数据的合理性。

2.2.4 模型训练与对比实验

基于增强后的数据集，重新训练前述的预测模型，并与未增强的数据集模型进行对比。实验组包括：本研究构建了三种不同的训练模型进行对比分析：基于原始数据集训练的基准模型、融合原始数据与 CTGAN 生成数据的增强训练模型，以及结合原始数据与 TVAE 生成数据的混合训练模型，并通过验证集和测试集对上述模型的性能表现进行了系统评估，以评估合成数据对预测能力的提升程度。

2.2.5 结果分析

为增强研究的可靠性与实用性，还需开展模型可解释性分析，利用 SHAP 值分析特征对模型预测结果的贡献，进而找出影响吸附剂性能的最相关特性，方便优化的运用。

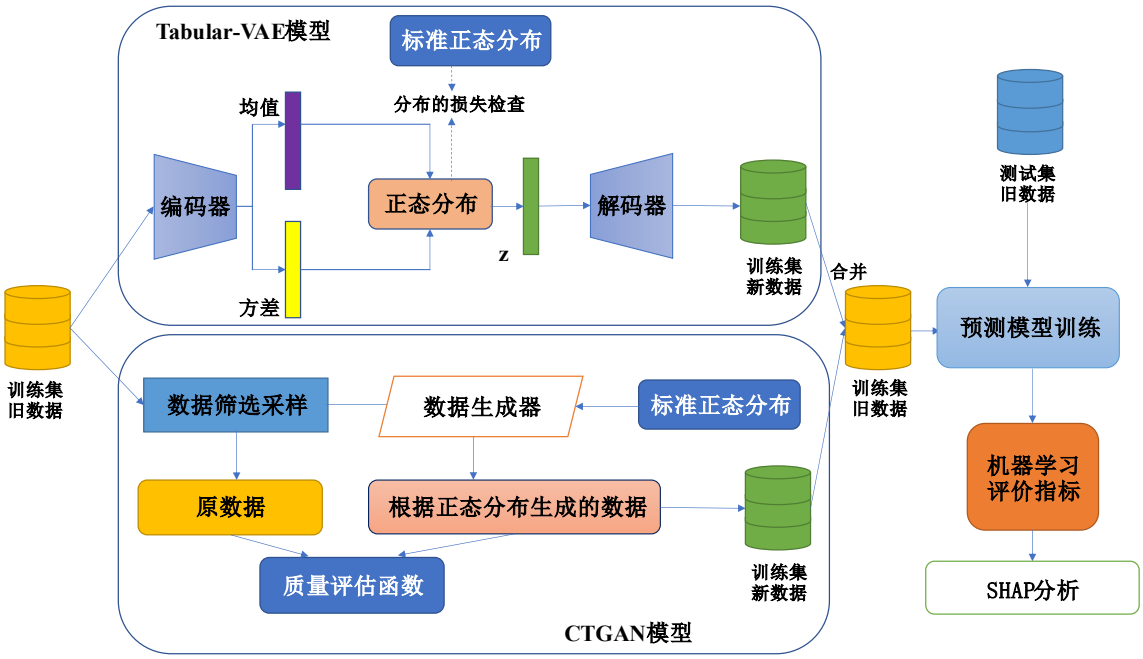

图 2.1　运用生成模型以及预测模型的技术路线图

2.3 数据收集

从各类的实验当中收集了各类综合数据，目的在于构建稳健的生成数据预测模型，对用于直接空气捕集法应用的固体胺吸附剂进行了相关的调查，本文选择了相关的开源数据[19]，从源头上来追述数据的来源均来自于该研究领域的的 22 篇同行评审文献，在此之中利用了 170 个数据点形成数据集[28-49]，这些数据的样本将放在本文的提供资料当中。对于数据的细致分析，所收集到的数据具体来说，收集到的输入特征分别分为三组，第一组是为固体载体的物理性质作为第一组数据，其中包括了吸附剂的比表面积分为吸收前和吸收后的比表面积，还有吸附剂的孔隙体积分为吸收前和吸收后的孔隙体积。第二组的数据则是胺的化学性质，这些化学性质在一定的意义上表示着胺的类型，它们分别是胺的分子量、胺的氮原子含量、不同胺官能团的比例和负载量。第三组数据则是吸附剂在自己本身的一些测试条件，包括吸附剂测试的温度、吸附剂所处环境的二氧化碳分压即二氧化碳浓度以及吸附剂所处的实验条件的相对湿度，在数据集上它们分别表示为胺分散前的比表（SSABD）面积、胺分散后的比表面积（SSAAD）、胺分散前的孔体积（PVAD）、胺分散后的孔体积（PVBD）、胺的分子量（MW）、氮原子的含量（NC）、伯胺比（PA）、仲胺比（SA）、叔胺比（TA）、胺的负载量（LA）、温度（Temp）、二氧

化碳的分压（PPM）和相对湿度（RH）。

在固体胺当中，比表面积以及孔隙体积是材料表面微观结构的直接体现，他主要是取决于吸附剂的颗粒大小、形貌以及吸附剂的孔隙分布决定的。1）具体来说，将胺分散在固体载体表面结合了胺类物质的高亲和力与载体的大表面积和孔体积特性的有点。而胺改性之后可以观察出这两者的减小，这两个参数直接影响材料的物理吸附能力和传质效率，因此作为数据的第一组。2）按类型可以通过分子量、氮原子含量和不同胺官能团的配比进行描述。胺基团与二氧化碳的可逆化学反应，材料在酸性环境的稳定性则是固体胺化学性质的直接表现，胺影响的详细吸附机理与胺官能团密切相关：伯胺和仲胺与二氧化碳反应通过 Zwitterion 机制形成氨基甲酸酯。水分子与二氧化碳相互作用通过亲核以及亲电反应，生成碳酸，随后与胺反应生成碳酸盐和碳酸氢盐[50]，因此作为数据的第二组代表固体胺的化学性质。3）而像温度以及二氧化碳分压力以及实验的湿度将作为实验条件进行研究，从而发掘实验条件过程中除去吸附剂自身的内部物理以及化学特性之外的实验环境特征。

2.4 数据的预处理

本文将采用 python 3.11 去进行数据的分析、机器学习的相关模型搭建以及模型性能的评估，在收集到数据之后首先需要对数据进行相关的预处理，对于数据的处理采用 python 的 pandas 数据处理库进行，其次通过对这 13 则数据的输入标签通过对数据以及输入标签二氧化碳的吸附量进行关于各个标签的正态性检验为以下公式（2.1）为 Shapiro-Wilk 公式[51]：

$$W = \frac{\sum_{i=1}^{n} a_i x_i{}^2}{\sum_{i=1}^{n}(x_i - \overline{x})} \tag{2.1}$$

其中x_i：样本中第 i 个顺序统计量，$a_i = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$由标准正态分布生成，反映正态分布下有序统计量的期望值，m 标准正态分布的顺序统计量期望值向量，V 顺序统计量的协方差矩阵，$\overline{x}$为样本均值，n 时样本量，它是一种检验正态性的方法，W 的取值范围在 0 和 1 之间，当样本数据的正态性检验结果即 W 趋近于 1 时，表明其分布形态更符合正态分布特征，此时应采用 Pearson 相关系数进行变量间线性关系的度量；反之，若正态性假设未被满足，则需选用非参数统计方法中的 Spearman 秩相关系数进行

相关性分析。这是由于需要表明独立性检验未通过代表着变量相关，而采取皮尔逊相关分析还是斯皮尔曼相关分析则取决于数据本身是否满足线性还是正态分布的假设，除此之外当数据不满足皮尔逊假设的时候，尤其是在非线性、非正态、存在异常值等情况之下，斯皮尔曼则是更加安全的选择。

然而，尽管相关性分析可以看出目标变量与输入变量还有各个变量的一些关系，但是变量与变量之间的高相关性并不一定是好事，这是因为回归分析过程中，若解释变量间呈现显著的相关性，则可能引发多重共线性问题，致使模型参数估计值产生较大波动，同时降低统计结果的解释效力，在这里引入方差膨胀因子（VIF：Variance Inflation Factor），由以下公式进行方差膨胀因子的计算如公式（2.2）：

$$VIF_j = \frac{1}{1 - R_j^2} \tag{2.2}$$

R_j^2：第 j 个自变量对其他所有自变量的回归决定系数，相关的判断标准为$VIF < 5$代表着共线性较弱是可以接受的，$5 \leq VIF < 10$代表着存在中等的共线性，$VIF \geq 10$代表着存在严重的共线性，对于这样的自变量的处理方式，则是通过计算相关的 *VIF* 或者相关系数矩阵之后，删除其中一个高相关的变量，保留 *VIF* 较低的变量。

考虑到数据内部可能会存在高基数类别，方差大的数据类别即代表该数据列具有高基数类别特征，对于处理高基数类别，本文通过计算方差然后找出不可接受的方差的数据进行对数转换，对数转换能够稳定数据方差以及解决偏态问题，其详细的转换方法使用以下数学公式（2.3）：

$$y = sign(x) \cdot log_b(x + c) \tag{2.3}$$

b 可以为 e 即自然常数或者为 10，在数据为 0 时，c 要取大于 0 的数，一般为 1，而 sign(x)则为取数据符号运算，只取正负

2.5 基于表格的数据生成模型

2.5.1 Tabular-VAE 简述

VAE 的英文全称为 Variational Autoencoder 是一种生成模型，通过结合深度神经网络以及概率图模型，学习输入数据潜在分布。而 TVAE[52]则是传统 VAE 模型的扩展，其

全称为 Tabular Variational Autoencoder 表格变分自编码，是针对表格数据（结构化数据）设计的变分自编码器的变种。它的核心思想主要是通过编码器-解码器的框架学习原始数据的概率分布，并在隐空间进行采样用来生成合成数据。

2.5.2 Tabular-VAE 模型结构

本研究采用的 Tabular-VAE 模型架构包含两个核心组件：编码器模块负责实现输入数据向潜在空间的非线性映射，而解码器模块则基于潜在变量执行数据重构任务，旨在恢复原始数据的统计特征与分布模式。1) 编码器：编码器的主要功能在于运用神经网络架构实现输入表格数据向隐式特征空间的非线性映射。由于表格数据当中包含不同类别的特征，因此编码器则需要处理不同结构的数据。基于这个情况，编码器通常是由多个分支组成的，其中每个分支对应着一种数据的特征类别，比方说连续型特征的处理、数据类别特征的处理。连续型特征，编码器可以使用全神经网络来学习数据的潜在表示；而对于分类别的特征，那要使用嵌入层进行处理。

在编码器中，输入数据可以通过神经网络变换，然后输出潜在变量的均值和方差，这些潜在的变量会生成重构的数据。为了让模型拥有生成数据的能力，编码器的输出将不只是隐空间的均值，还会提供方差的估算，这样才可以采样潜在的变量。2）解码器：解码器的任务是基于潜在变量重构原始的数据。在 Tabular-VAE 中，解码器一般是由多个全神经网络组合而成，直接从隐空间中生成和输入数据相同维度的重构数据。然而不同类型的数据特征需要不同的解码策略，譬如对于连续型的数据，解码器通常会生成一个高斯分布的均值和方差，用来计算重构得到数据的误差，而对于分类别的数据，则会生成每个数据类别的条件概率。

输出的重构数据在解码器中通过采样生成。加入数据是具有连续特征的，通常需要通过使用均值来回归；而分类特征则需要使用一个 softmax 函数进行数据的分类。

2.5.3 TVAE 损失函数

Tabular-VAE 模型的损失函数由重构损失与 KL 散度损失两项构成：首先，重构损失用于评估生成数据与原始输入数据之间的偏差程度，其中连续型数据采用均方误差作为衡量指标，而分类数据则通过交叉熵损失进行量化；其次，KL 散度作为隐空间分布

与先验分布差异的度量指标，在 VAE 框架中通常假设潜在变量服从正态分布，该损失项通过对潜在空间分布施加正则化约束，有效防止模型在数据生成过程中出现过拟合现象，其具体计算过程可由公式（2.4）予以表征：

$$\mathrm{KL}\big(q(z|x)\big|p(z)\big) = \frac{1}{2}\sum_{i=1}^{d}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2) \tag{2.4}$$

在这则公式当中μ_i和σ_i分别是潜在空间的均值和标准差。

除此之外，变分推断则是 VAE 模型当中的一个计算框架，VAE 采用变分下界优化方法实现对后验分布的近似估计，具体而言，针对表格数据的变分自编码器通过变分推断技术将复杂的后验推断问题转化为可求解的优化问题，该模型通过最大化变分下界的方式，能够高效地从训练数据中学习潜在变量的概率分布特征。从而生成阶段有效采样出潜在变量，由此生成符合数据分布的样本。对比于传统的 VAE，Tabular-VAE 能够处理不同类型的数据，引入 KL 散度避免过拟合的问题因此在处理表格数据上具有很好的优势，能够在数据量不足的情况之下生成新的样本提高模型的泛化能力。

2.5.4 CTGAN 简述

GANs 即生成对抗网络目前因为生成高质量数据，已经在各类数据生成任务中取得了显著的进展，尤其是在图像以及音频的领域当中。而当数据回归到最原始的结构化表格数据当中，尤其是在连续变量和离散变量各标签混合是，经典的生成对抗网络则是面临着巨大的挑战，由此条件生成对抗网络孕育而生。

2.5.5 CTGAN 模型结构

CTGAN[53]是一个改进版本的生成对抗网络，目的是生成满足表格数据当中逼真的数据，然后能够抓捕数据之间的复杂依赖关系，一般的生成对抗网络由生成器和判别器组成：1) 生成器即 Generator：它接受数据噪声向量以及条件输入，生成合成表格的数据样本。2） 判别器即 Discriminator：区分真实以及生辰的数据，同时可以根据特征信息进行条件的判断。

在 CTGAN 中，数据的数值特征不会进行全局的归一化，而是按照数据条件来进行归一化。在这一步的条件之上，使用贝叶斯高斯混合模型来建模数据的数值分布，由此

捕捉数据多模态分布由此来提高数据重新采样的效果。在具体的概率采样公式中，假设 x 为数值变量，条件类别为 c，那么数据的采样概率分布可以表示为以下公式（2.5）：

$$p(x|c) = \sum_{i=1}^{M} \pi_{i\mathcal{N}}(x|\mu_i, \sigma_i^2) \tag{2.5}$$

在这里π_i是高斯混合模型中的混合权重，$\mathcal{N}(x|\mu_i, \sigma_i^2)$是高斯分布，均值为$\mu_i$和方差为$\sigma_i^2$，并且每个类别 c 使用单独的高斯混合模型进行建模。这样的方式能够有效地避免模式坍缩，使得生成器学习到更符合真实数据的特征分布。

2.5.6 CTGAN 损失函数

其中 CTGAN 一个特点是条件训练，生成器基于采样的类别进行条件训练，这样可以保证即使对于较少出现类别，生成器能够平衡训练，能够专注每个条件分布的建模。在损失函数与训练上，CTGAN 遵循生成对抗网络极小极大训练目标

判别器的损失公式（2.6）为：

$$\mathcal{L}_{\mathcal{D}} = -E_{x\sim P_{real}}[\log D\,(x,c)] - E_{\hat{x}\sim G(z,c)}\big[\log\big(1 - D(\hat{x},c)\big)\big] \tag{2.6}$$

生成器的损失公式（2.7）为：

$$\mathcal{L}_{\mathcal{G}} = -E_{\hat{x}\sim G(z,c)}[\log D\,(\hat{x},c)] \tag{2.7}$$

对比两个不同的目标，判别器是基于数据类别条件对真实数据以及生成数据进行区分，而生成器则是生成尽量真实的样本欺骗判别器。同时为了训练稳定，CTGAN 使用小批量方式，将多个样本拼接输入至判别器，避免模式的坍缩问题。

在训练完成之后，将采样一个类别作为条件，选取一个经验分布选择一个值，将条件向量与噪声拼接返回到生成器当中，然后对数据进行逆变换，将生成数据和原始数据相一致结构，反应数据的真实分布。

2.6 模型预测以及模型的评估

在模型的评估上，预测类模型主要的思路是使用回归类的机器学习任务模型，在本研究当中选取四类最具代表性的机器学习模型构建基线：1. 线性回归、2. 决策树、3. 随机森林、4. 梯度提升树、5. 支持向量机。这四类模型将会被对比其在生成数据当中的表

现，全面评估模型在数据上的复杂性以及模型的适用度。

2.6.1 线性回归

在线性回归当中，根据数据的多维层面，建立相关的特征矩阵 X 以及目标变量 y 的线性映射关系，如下公式（2.8）所示：

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i \tag{2.8}$$

而在当中为了找出最佳的参数如β、ϵ则是使用最小二乘法优化参数公式（2.9）的估计：

$$\beta = argmin\beta \parallel y - X\beta \parallel^2 \tag{2.9}$$

但线性回归需要缓解多重共线性在回归任务当中的影响，在本文当中已经采用方差膨胀因子对其进行剔除

2.6.2 决策树模型

在决策树模型中，本研究主要采用 CART 算法进行二叉树的构建，它以基尼不纯度作为数据是否满足条件的分类准则，如下公式（2.10）：

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 \tag{2.10}$$

而树的递归选择则以书当中子节点的纯度以最大的特征进行分裂，如下公式（2.11）：

$$\Delta\, Gini = Gini(D) - \sum_{v=1}^{V} \frac{|D_v|}{|D|} Gini(D_v) \tag{2.11}$$

2.6.3 随机森林模型

它基于树的集合，通过 Bagging 的策略构建多棵决策树，从训练集当中有放回抽样生成多个样本，然后为每一棵树的结点分裂随机选择特征，最终预测则采用分类或者回归的方式。

2.6.4 梯度提升树模型

它是基于树模型的叠加然后足部优化数据预测结果的一个过程，由前一棵树先后一棵数进行梯度提升相关的优化，可以由以下公式得出树递归的相关计算，如公式（2.12）：

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2.12}$$

其中$h_m(x)$为基学习器拟合当前模型的负梯度

2.6.5 支持向量机模型

支持向量机回归是支持向量机的扩展与延申，是为了解决回归问题的。和以往的回归都不同，SVR 是试图在一个可以接受的误差范围内找到一个符合数据分布规律的函数拟合曲线，然后保持模型的泛化能力，通过引入一个不敏感区域，让模型对训练所带来的噪声以及波动不敏感，因此支持向量回归在一些高维度特征中具有良好的鲁棒性，十分适合本次小样本数据的回归预测任务。

为了让模型严格评估模型的性能，数据集根据 8:2 的比例被拆分为训练集和数据集，而这些数据当中的训练集则首先将会被输入到生成数据模型当中进行数据的生成，而剩余的模型则会根据数据增强之后的数据集进行测试以及没有进行数据增强之前的数据集进行测试，在这些预测值与实际实验值进行比较。在比较的过程中，本文将会引入相关的指标量化数据增强前以及数据增强后的预测以及实际值之间的关系。用于量化训练模型性能的指标包括由均方误差（MSE），平均绝对误差（MAE）以及决定系数（R2）。1）MSE 均方误差：用来测量预测误差平方的平均值；2）MAE 平均绝对误差：用来评估预测中误差的平均幅度；3）R2 决定系数：因变量中可从自变量预测的方差比例。

2.7 SHAP 重要性分析

除了预测之外，本研究使用 SHapley Additive exPlanation(SHAP)进行模型的重要性分析，用来预测模型当中每一个特征的相关贡献，它是基于合作博弈论中的 Shapley 值，为每个特征分配预测结果的贡献值。其中正的 SHAP 值表示特征对产生预测结果的增益，而负的 SHAP 则是表示特征对产生预测结果的减少。

SHAP 平均值是为了了解数据当中每个特征的重要性，表示每一个特征对模型的平均影响，为研究数据特征重要性对目标变量的影响提供上帝视角，不仅提高了模型的透明度同时还验证了模型的可靠性以及增强模型的理论重要性。

3 结果与讨论

3.1 数据预处理以及初步分析

在数据预处理的基础上，本文对所得到的开源数据进行预处理，首先对数据进行缺失值寻找发现本文所使用到的开源数据并没有缺失值，通过对数据进行全部标签的独立性检验可以发现所有的数据其所计算得出的独立性参数均是远远小于 0.05，这代表着数据本身具有很明显的显著性，意思是数据本身有可能并不是连续的，具体的显著性信息如下表 3.1 所示。

表 3.1 各变量独立性显著值

吸附剂的相关条件	英文首字母缩写	P 值
吸附前比表面积	SSABD	3.59E-13
吸附后比表面积	SSAAD	9.72E-19
吸附前孔隙体积	PVAD	1.66E-15
吸附后孔隙体积	PVBD	1.48E-09
分子质量	MW	1.99E-24
吸附剂氮原子含量	NC	1.42E-17
伯胺百分比	TA	1.11E-18
仲胺百分比	SA	8.62E-12
叔胺百分比	PA	2.53E-16
气体承载量	LA	9.63E-24
温度	Temp	7.26E-17
CO2 分压	PPM	3.05E-24
相对湿度	RH	3.84E-25
二氧化碳吸附量	CO2 uptake	1.28E-12

尽管如此，将所得到的数据带入到皮尔逊相关性检验当中，如下图 3.1，很明显的发现各个变量之间实际上并没有很高的相关性，最大的相关性的绝对值并没有达到类似 0.8 以及 0.9 的范畴，这意味着其实各个变量之间的数据标签可以得到很好的保留。

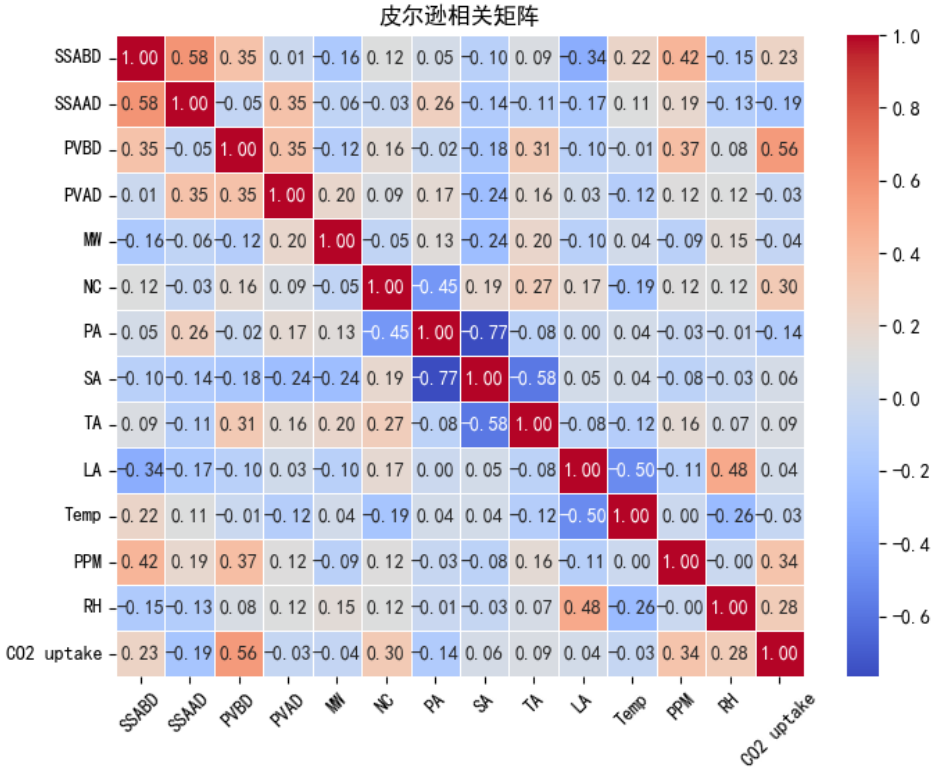

图 3.1　各个变量的皮尔逊相关矩阵热力图

但是考虑到因为数据明显的显著性，这代表着数据在相关性分析当中不可以使用皮尔逊检验，因此采用斯皮尔曼相关性分析是对这则数据最好的选择，如下图 3.2，通过相关性热力图发现，在各个变量中可以发现，各变量当中最高的相关系数为大概 0.7，这同时意味着伯仲叔胺之间的关系虽然高相关，这其实是因为在伯仲叔胺当中它们的比例总和为 1，但是实际上又可以相互独立。

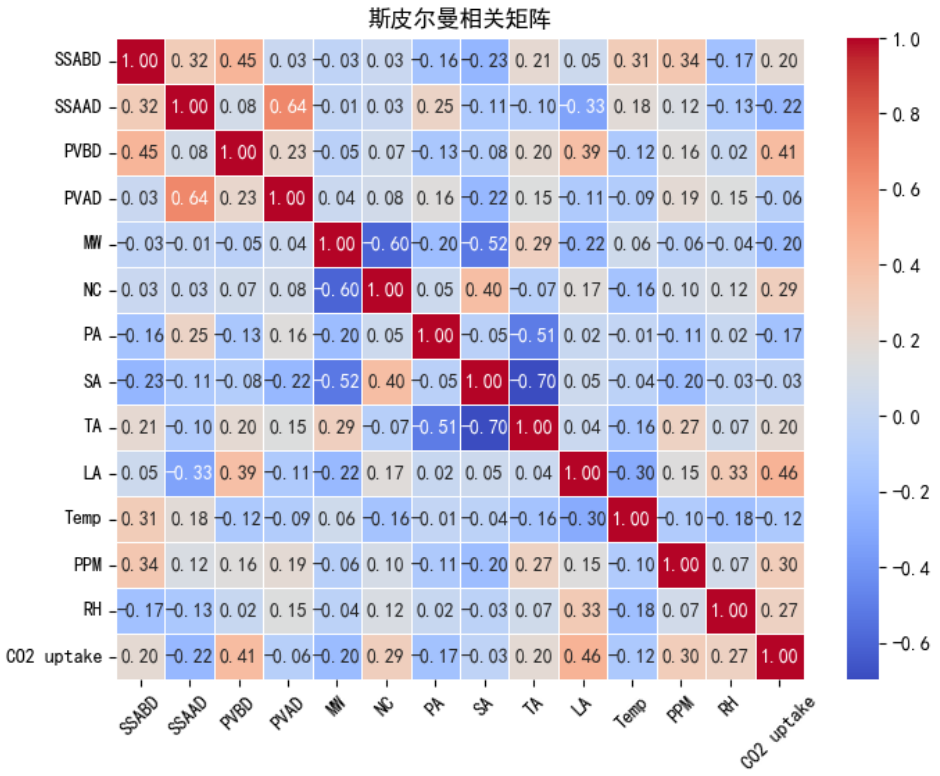

图 3.2　各变量的斯皮尔曼相关系数矩阵热力图

另外第二个具有高相关性的则是吸附剂上的物理性质，这在吸收后的吸附剂表面积和吸收后的孔隙体积之间同时具有大约 0.64 的强正相关性。各变量之间还具备强相关性的则是吸附前的孔隙体积以及吸附剂气体的承载量，它们之间的相关性为 0.39，并且吸附剂气体的承载量与相对湿度的相关性为 0.33，这代表着孔隙体积以及环境的相对湿度对于固体胺本身的气体承载能力具有一定的影响能力。而二氧化碳分压力同时发现与吸收前的比表面积以及吸附剂气体的承载量也有较强的正相关性，这主要是二氧化碳分压同时也会对吸附剂的物理化学能力有所约束。然而根据变量之间的组群特征，像环境来说，实验条件中的相对湿度与吸附剂的承载量以及实验条件当中的气体分压呈正相关，分别为 0.33 和 0.07，而与温度呈负相关(-0.18)，这代表着在进行吸附剂的吸附性能研究时，实验条件的温度以及湿度要选择适宜居中的情况，而二氧化碳分压力与吸附前的比表面积和吸附前的孔隙体积呈正相关分别为 0.34 和 0.16，这代表着分压的条件之下对吸附剂物理化学性能具有激活作用。

表 3.2 各变量的方差膨胀因子

吸附剂的相关条件	英文首字母缩写	VIF
吸附前比表面积	SSABD	3.18E+00
吸附后比表面积	SSAAD	3.12E+00
吸附前孔隙体积	PVAD	2.89E+00
吸附后孔隙体积	PVBD	2.27E+00
分子质量	MW	1.40E+00
吸附剂氮原子含量	NC	1.69E+00
伯胺百分比	TA	1.90E+05
仲胺百分比	SA	4.58E+05
叔胺百分比	PA	3.05E+05
气体承载量	LA	1.98E+00
温度	Temp	1.43E+00
CO2 分压	PPM	1.47E+00
相对湿度	RH	1.56E+00
二氧化碳吸附量	CO2 uptake	2.36E+00

如表 3.2 格代表着各个数据标签的方差膨胀因子，其中可以很明显的发现数据当中在伯胺、仲胺和叔胺这三者当中，它们具有很高的方差膨胀因子，这是因为它们代表着各个官能团在固体胺类的占比同时也反馈着三者的综合实际就是为 1，因此具有很高的方差膨胀因子，然而一般高于 5 的方差膨胀因子需要得到剔除，因此在这次的变量标签当中，所有的特征的方差膨胀因子除了伯仲叔胺的标签都小于 5，但他们三者的各自之间的相关性并没有达到 0.8 的阈值，因此数据当中的所有列都不用剔除。

3.2 生成模型的评估以及评价

本研究主要运用两类生成模型进行数据增强处理：其一是基于表格数据的变分自编码器，其二则是采用表格条件生成对抗网络技术。为了评估原始数据以及生成数据的质量，观察原始数据以及生成数据的概率分布状态以及由概率分布所统筹出来的数据的平均数以及方差是非常重要的，以下是关于 TVAE 数据生成的概率密度分布图以及直方图 3.3 和图 3.4：

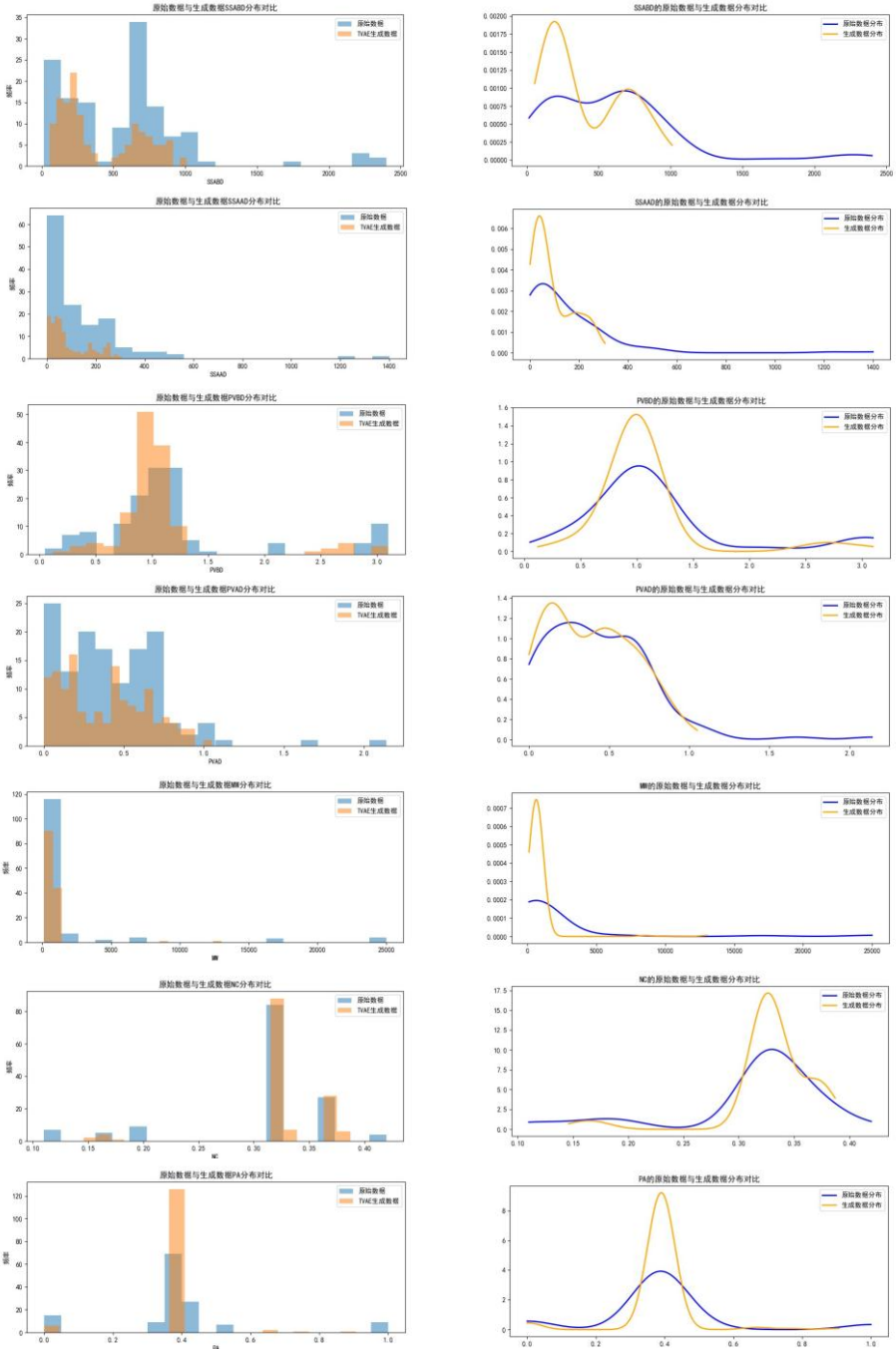

图 3.3　为未进行对数变换时的 TVAE 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为吸附前比表面积、吸附后比表面积、吸附前孔隙体积、吸附后孔隙体积、分子质量、吸附剂氮原子含量、伯胺百分比

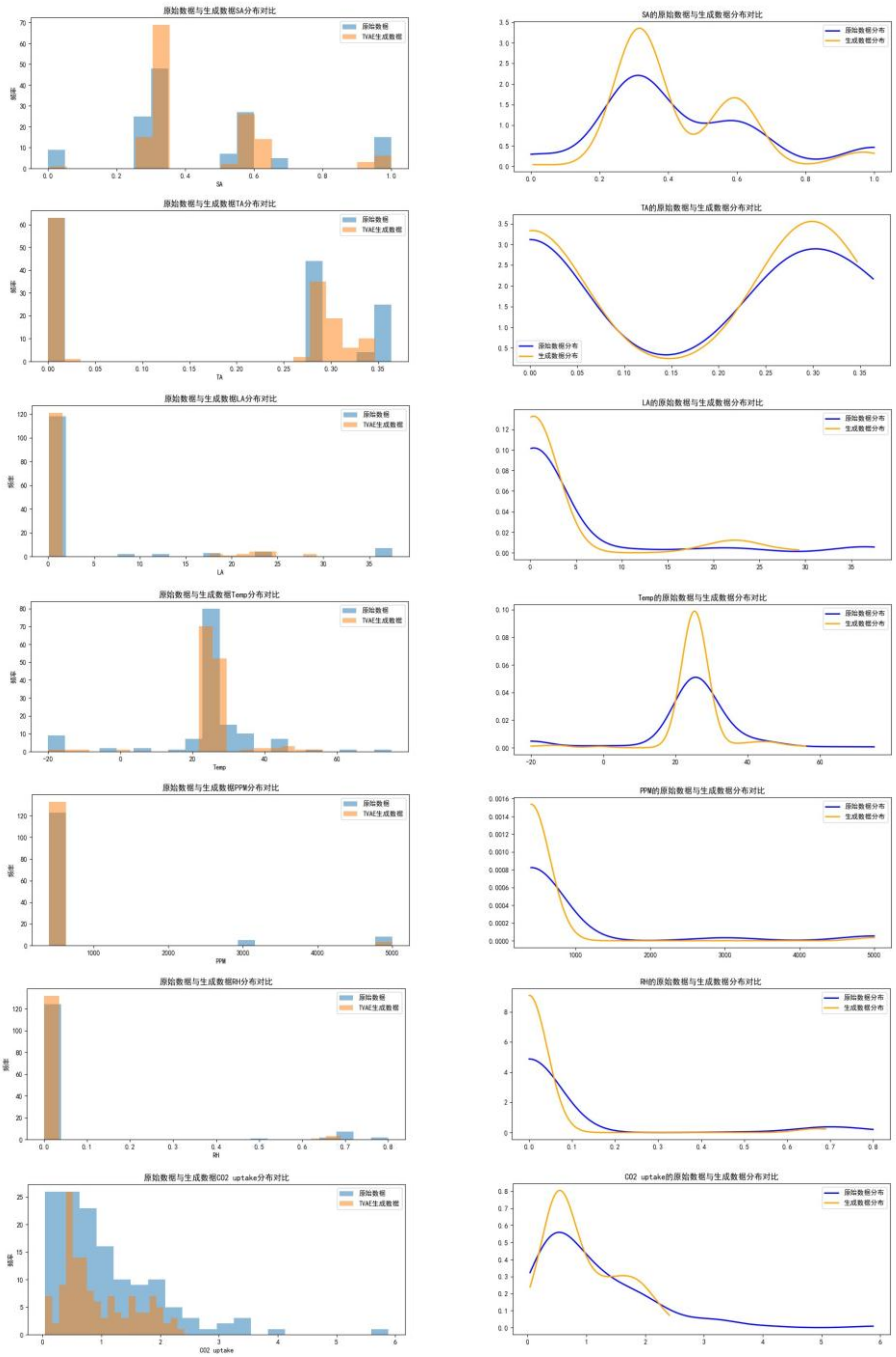

图 3.4　为未进行对数变换时的 TVAE 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为仲胺百分比、叔胺百分比、气体承载量、温度、二氧化碳分压、相对湿度、二氧化碳吸附量

从 TVAE 所生成的数据的概率密度分布函数来看，在直方图上面有大部分数据都是在原始数据之下，其中以吸收前的比表面积、吸附后的比表面积、吸附后的孔隙体积和二氧化碳的吸附量他们生成的数据大部分是在原始数据的覆盖之下，而其余诸如伯仲叔胺的占比以及其余的数据标签均有覆盖以及延伸原始数据，而对比概率密度分布图当中，TVAE 在吸附前后的比表面积当中根据原始数据具有的波形泛化出更高的分布峰值，而在孔隙体积、分子质量、伯仲叔胺等的数据标签当中，在分布的形态上，数据大体都满足原始数据分布所具有的形态以及峰值，但是主要由于数据在这几类标签当中属于是高类别基数，也就代表着这几列数据当中相同组别的基数较多因此在分布当中以及直方图当中，理所应当地在相应的地方上会有相当多的一些数据的积累而导致出现多个波形以及或者说在相应的数据频次上出现的峰值会更大对应着在当前数据生成当中生成该类数据的数量更多。

表 3.3 TVAE 的原始数据和生成数据之比

吸附剂的相关条件	原始数据平均数	生成数据平均数	原始数据标准差	生成数据标准差
吸附前比表面积	575.8621	401.6802	459.7573	276.1604
吸附后比表面积	139.9654	91.214	189.3421	82.7353
吸附前孔隙体积	1.1985	1.0848	0.728	0.5146
吸附后孔隙体积	0.4256	0.3845	0.3295	0.2673
分子质量	1920.1684	763.4265	4816.9141	1277.6598
吸附剂氮原子含量	0.3111	0.329	0.0713	0.0437
伯胺百分比	0.3887	0.3856	0.2067	0.1041
仲胺百分比	0.4448	0.4419	0.2544	0.1927
叔胺百分比	0.1666	0.1603	0.1582	0.1498
气体承载量	3.6142	2.8621	9.0732	7.1307
温度	24.1544	25.6471	14.9512	8.1691
CO2 分压	766.7647	501.4706	1169.9533	678.1222
相对湿度	0.0613	0.0197	0.199	0.1137
二氧化碳吸附量	1.0847	0.9337	0.9231	0.6013

抛开从图上对原始数据和生成数据的评判，数据的平均数以及标准差从一定的程度上可以代表着生成数据以及原始数据是否具有较大的差异如表 3.3。总体的情况而言 TVAE 模型生辰的个大多数数据的大多数特征均值和原始数据是接近的，这代表着 TVAE 在保持数据中心的趋势具有着良好的表现。回看均值在数据之间的对比，偏差是存在的，这一部分大的偏差在于吸附前比表面积、分子质量和二氧化碳分压力在生成数据的均值要远远低于原始数据，分别下降了 30%、60%和 33%，从模型的结构上来看似乎有更大提升的空间，但是回看到数据自己本身，貌似在这一些类别当中，尽管数据是有一定偏态，但由于高类别基数数据都发生在下降明显的数据列当中，因此导致了这些数据的一定偏差。同时相对湿和二氧化碳的均值变化很小，一定程度上反应着模型很好地保留着这些变量的中心趋势，而在标准差上，模型生成数据的标准差通常小于原始数据，说明这模型在生成数据上分布要更为集中。除此之外，很容易发现到，其实均值一

般的数据同时会在方差上也表现的差。

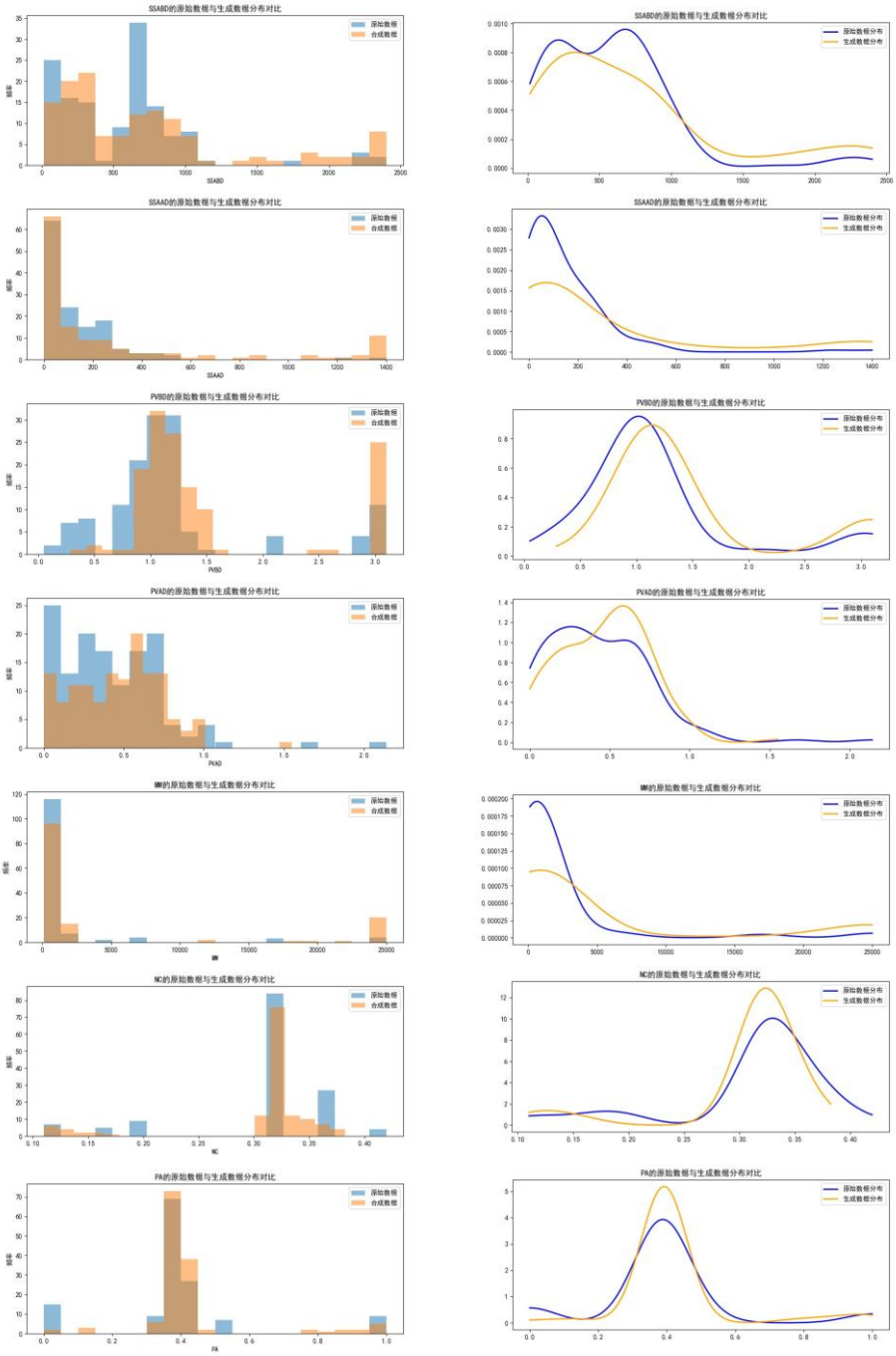

图 3.5　为未进行对数变换时的 CTGAN 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为吸附前比表面积、吸附后比表面积、吸附前孔隙体积、吸附后孔隙体积、分子质量、吸附剂氮原子含量、伯胺百分比

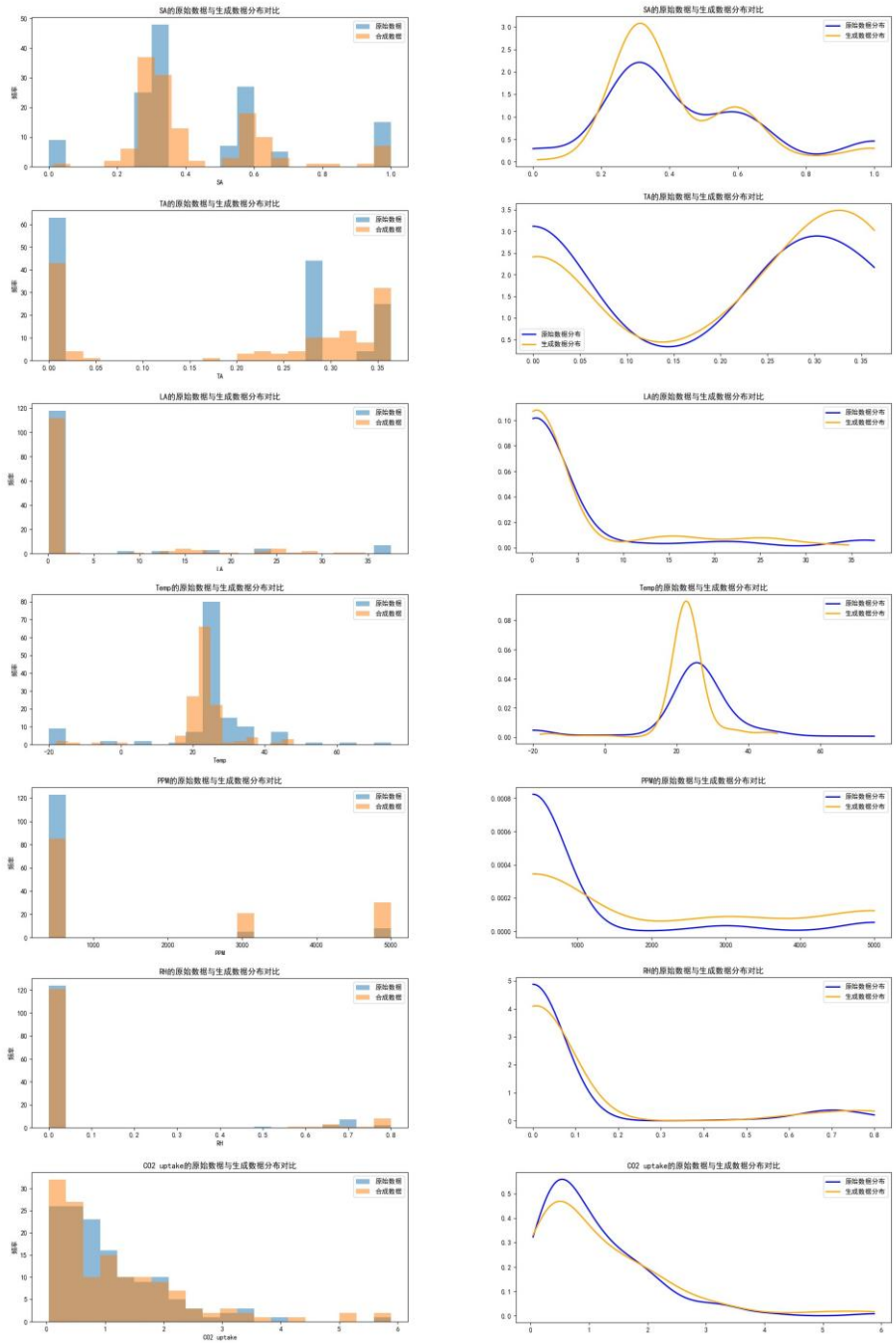

图 3.6　为未进行对数变换时的 CTGAN 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为仲胺百分比、叔胺百分比、气体承载量、温度、二氧化碳分压、相对湿度、二氧化碳吸附量

对比 TVAE 模型所生成的数据，可以从直方图当中很明显的看出大部分由 CTGAN 所生成的数据大部分的分布都和原始的数据相同，尤其在二氧化碳吸附量上除了少部分的原始数据没有被生成数据所覆盖，但在大部分的区域都可以看出生成数据要很好地覆

盖住了大部分的生成数据，而对比 TVAE 生成数据当中的吸附前比表面积、吸附后比表面积以及吸附后的孔隙体积那样子情况的少部分数据均分出现并没有很好地覆盖原始数据来看，CTGAN 在这几个数据标签方面表现都要较 TVAE 要好。当我们看基于直方图频率分布的数据概率密度分布来看，吸收前的比表面积的原始数据呈现的状态是多峰的，而生成数据尽量为了满足多峰的状态又要满足总体的分布呈现的状态却是单峰但还是满足了基本的数据分布，其次还有仲胺的数据分布来看尽管是多峰的，但是在最高的峰处所表现的数据更加集中，侧面反映在原始数据以及现有数据之中这个峰值代表着基于数据类别的频率出现更多。然而部分数据显示的

轮到 CTGAN 的概率密度分布图来看生成的数据具有更高的峰值，侧面反映出数据生成方面更偏向于原始数据出现频次多的进行合成数据的生成。

表 3.4 CTGAN 原始数据和生成数据对比

吸附剂的相关条件	原始数据平均数	生成数据平均数	原始数据标准差	生成数据标准差
吸附前比表面积	575.8621	723.023	459.7573	656.3343
吸附后比表面积	139.9654	291.5815	189.3421	429.8719
吸附前孔隙体积	1.1985	1.5085	0.728	0.7953
吸附后孔隙体积	0.4256	0.4617	0.3295	0.2753
分子质量	1920.1684	4995.6779	4816.9141	8870.1887
吸附剂氮原子含量	0.3111	0.3057	0.0713	0.064
伯胺百分比	0.3887	0.4268	0.2067	0.1669
仲胺百分比	0.4448	0.4293	0.2544	0.1997
叔胺百分比	0.1666	0.2054	0.1582	0.1536
气体承载量	3.6142	3.9386	9.0732	8.1246
温度	24.1544	22.6985	14.9512	8.3766
CO2 分压	766.7647	1818.9706	1169.9553	1930.1404
相对湿度	0.0613	0.0875	0.199	0.2294
二氧化碳吸附量	1.0847	1.2134	0.9231	1.1644

综合分析在表 3.4，CTGAN 原始数据和生成数据在平均数和标准差的差异来看，CTGAN 在吸附前后的比表面积来看，CTGAN 在这两者的生成条件来看是要比 TVAE 所生成的差异要差不多，但在分子的质量上，CTGAN 对比 TVAE 则更加喜欢生成更大的平均以及方差，这在上面的发生了更大的均值偏移，这样的情况同样发生在二氧化碳的分压上，生成出远大于原始数据的平均数以及标准差，使得这一列数据的生成数据出现了较大的均值偏移，这表明 CTGAN 模型能够生成更加极端的样本，这可能是由于 CTGAN 由于其在训练过程中生成器不断尝试通过不同的条件导致数据更加多样但是使得数据更加离散了。

3.3 对数优化模型的提升

考虑到未优化的数据当中，TVAE 模型在生成数据方面均值稳定，适合需要平稳分

布的任务，但是它在波动较大的特征和极端数据生成效果较差。而 CTGAN 生成数据多样性强，能够捕捉较大的波动性和极端样本，但其生成数据可能偏离原始均值较多并且波动性较大。但回归到数据本身的问题，这可能是数据具有大量的高异变量所造成的。在 TVAE 模型当中，吸附前比表面积是其中一个高异变量，在均值中，原始数据为 575.86，而生成数据均值却是 401.68，偏高达约 30%，原始数据的标准差为 456.76，而生成数据的标准差 276.16，具有很大的标准差差异。另外分子质量在 TVAE 当中，原始数据为 1920.17，但生成的数据均值为 763.46，偏差约为 60%，原始数据的标准差为 4816.91，却生成了 1163.95 的标准差，具有很显著的差异，代表着生成的数据具有很严重的离散问题。在 CTGAN 的原始数据和生成数据当中，最明显的的则为依然是分子质量，另一个则是二氧化碳分压，在分子质量上原始数据为 1920.17，生成数据为 4995.68，偏差高达 160%，接近 2 倍的差异，而标准差之间的差值高达 4000，充分反应出分子质量不仅在 TVAE 也在 CTGAN 中具有很大的波动性以及变异性。而二氧化碳分压的均值偏差则到达 150%，这充分说明一些数据的波动很明显需要进行相关的优化。基于此本文后续工作对变异性高的变量进行了对数化处理，下图 3.7 和图 3.8 为优化过后的 TVAE 原始数据和生成数据的概率分布图和概率分布直方图：

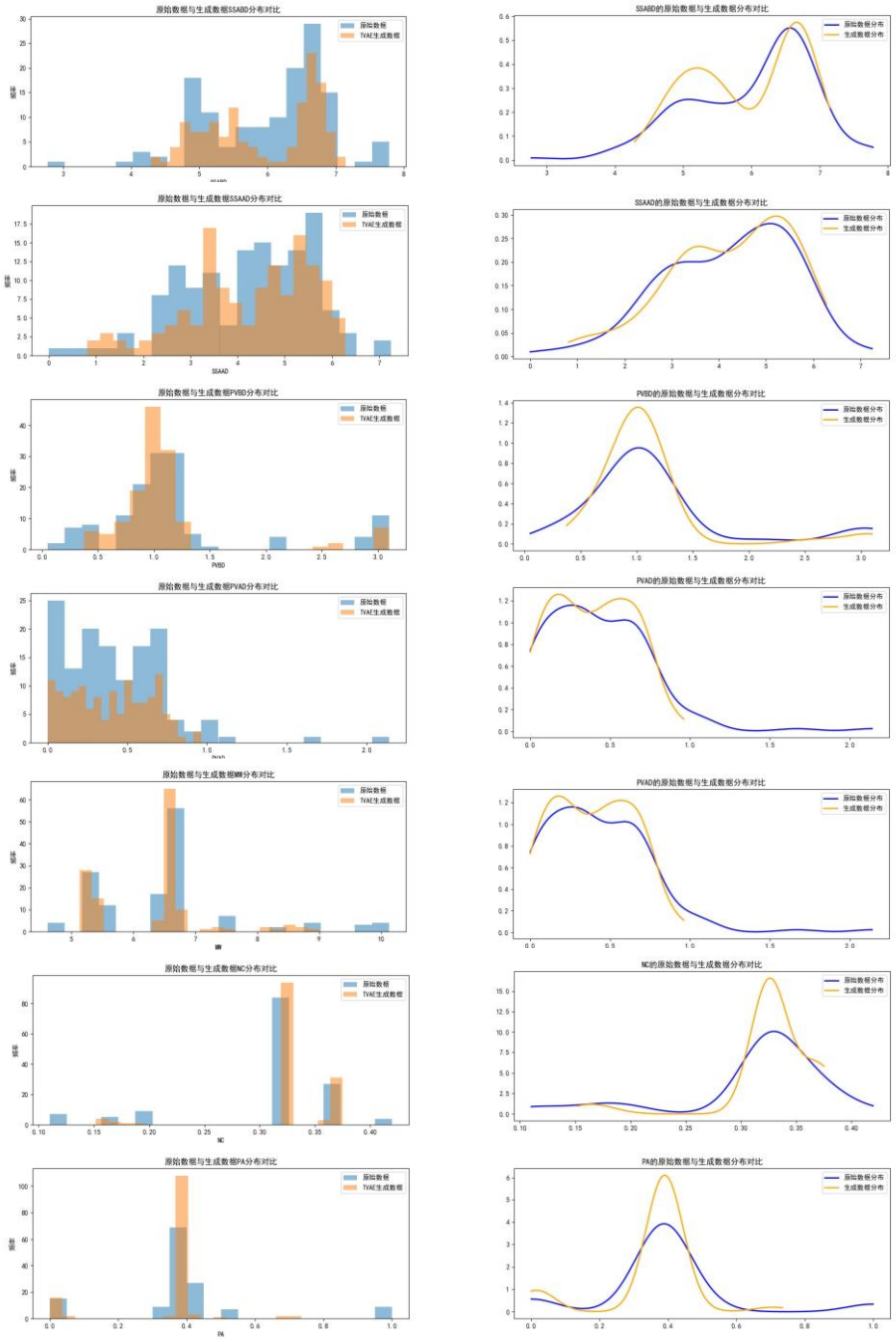

图 3.7　为已经进行对数变换时的 TVAE 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为吸附前比表面积、吸附后比表面积、吸附前孔隙体积、吸附后孔隙体积、分子质量、吸附剂氮原子含量、伯胺百分比

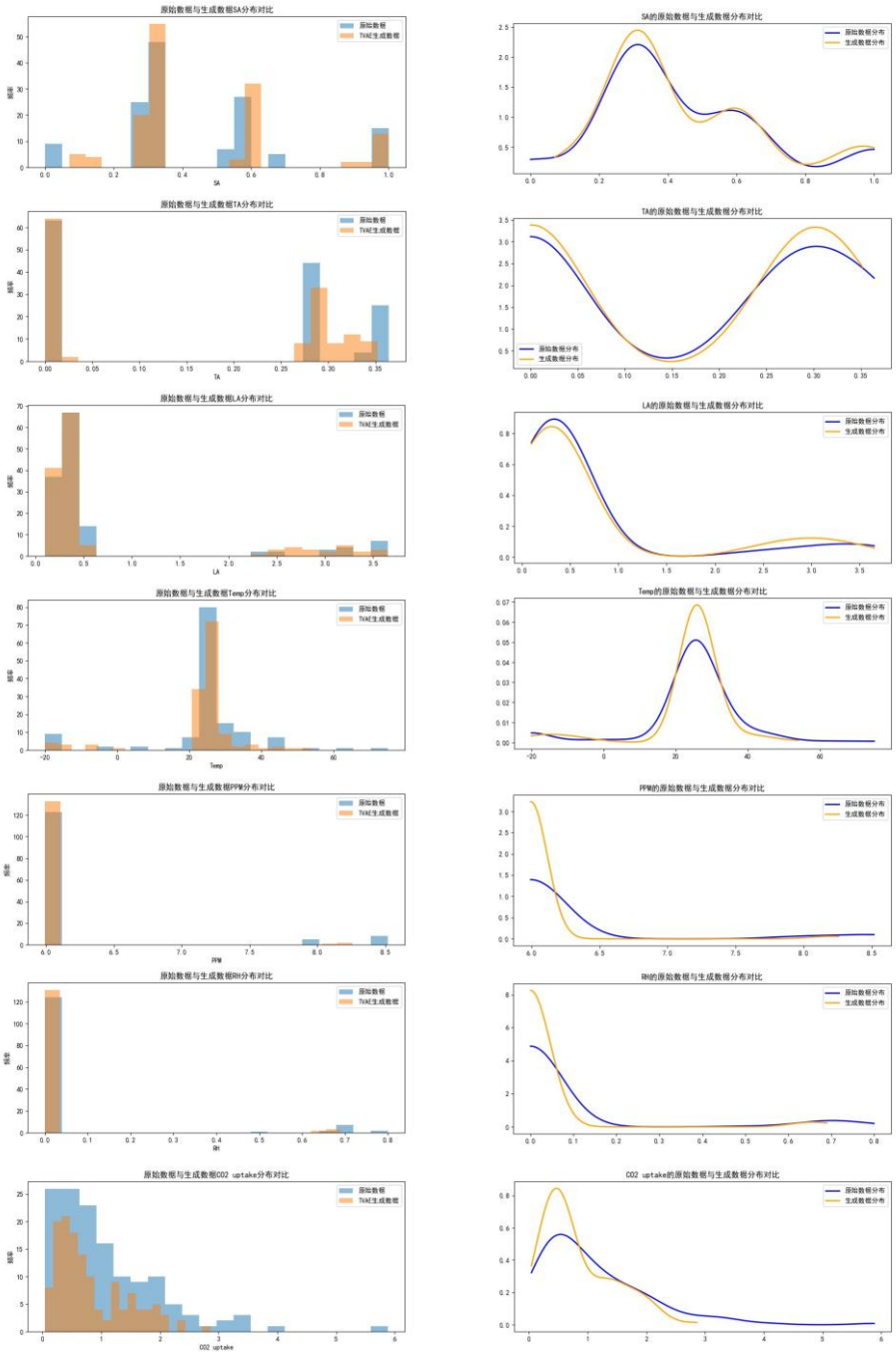

图 3.8　为已经进行对数变换时的 TVAE 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为仲胺百分比、叔胺百分比、气体承载量、温度、二氧化碳分压、相对湿度、二氧化碳吸附量

进行对数变换之后的 TVAE 模型在数据生成方面远远超越了原来没有进行对数变化的 TVAE 模型，具体的表现在处理双峰分布的情况之下时，生成数据对于原始数据具有了更好的拟合方式，不会远远的大于原始数据的频率分布，而在单峰的分布情况之下，

生成数据的频次会更多的分布在原始数据的出现频次多地方，而不会出现在长尾比较严重的区域，这代表着在单峰的数据生成方向中，经历过数据对数化操作之后会出现十分良好的效果，这同时也可以具体的表现在能够代表数据分布的均值和标准差当中，如下表格 3.5：

表 3.5 TVAE 对数变换之后的原始数据和生成数据之比

吸附剂的相关条件	原始数据平均数	生成数据平均数	原始数据标准差	生成数据标准差
吸附前比表面积	6.0201	5.9286	0.9065	0.7955
吸附后比表面积	4.2112	4.2325	1.3599	1.2964
吸附前孔隙体积	1.1985	1.1061	0.728	0.5484
吸附后孔隙体积	0.4256	0.3935	0.3295	0.2472
分子质量	6.499	6.3449	1.1793	0.8632
吸附剂氮原子含量	0.3111	0.3275	0.0713	0.045
伯胺百分比	0.3887	0.3506	0.02067	0.1426
仲胺百分比	0.4448	0.4639	0.2544	0.2404
叔胺百分比	0.1666	0.1569	0.1582	0.1524
气体承载量	0.7063	0.7587	0.9785	1.0213
温度	24.1544	23.8456	14.9512	12.0129
CO2 分压	6.2178	6.0447	0.6906	0.3225
相对湿度	0.0613	0.0242	0.199	0.1244
二氧化碳吸附量	1.0847	0.8253	0.9231	0.6053

综合上表来看，从均值的比较来看，通过数据对数转换之后，大部分的变量生成数据与原始数据的均值差异较小。譬如吸附前比表面积、二氧化碳分压和温度等变量的均值差异较小，这已经说明经历数据变换之后，大部分数据具有很好的集中性。其次在标准差比较来看，大部分数据变量生成的数据和原始数据差异也不大，表示 TVAE 在模拟数据离散程度上能够较好的反应原始数据的分布，但在一些变量上尤其是分子质量，差异十分明显。

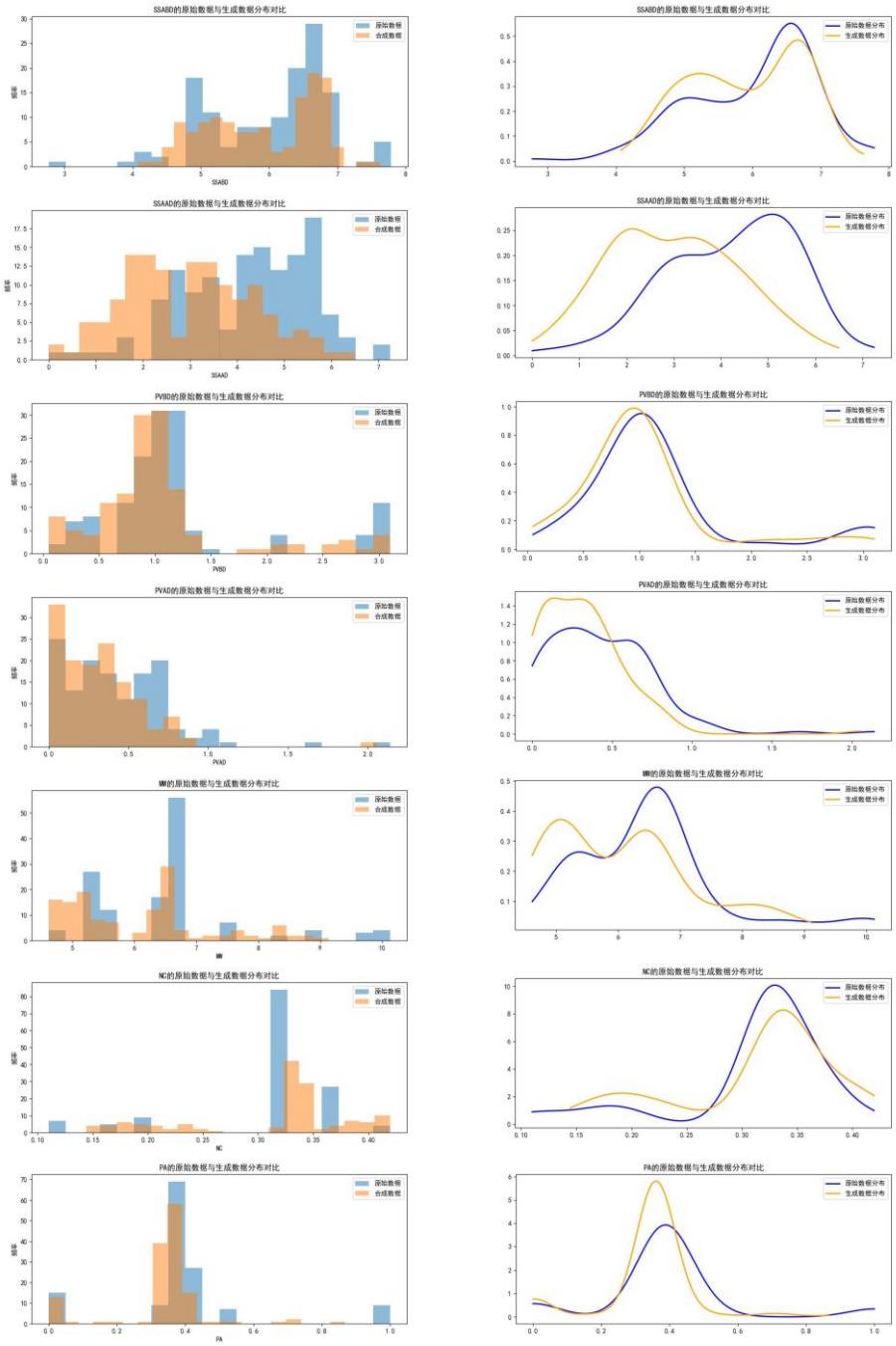

图 3.9　经进行对数变换时的 CTGAN 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为吸附前比表面积、吸附后比表面积、吸附前孔隙体积、吸附后孔隙体积、分子质量、吸附剂氮原子含量、伯胺百分比

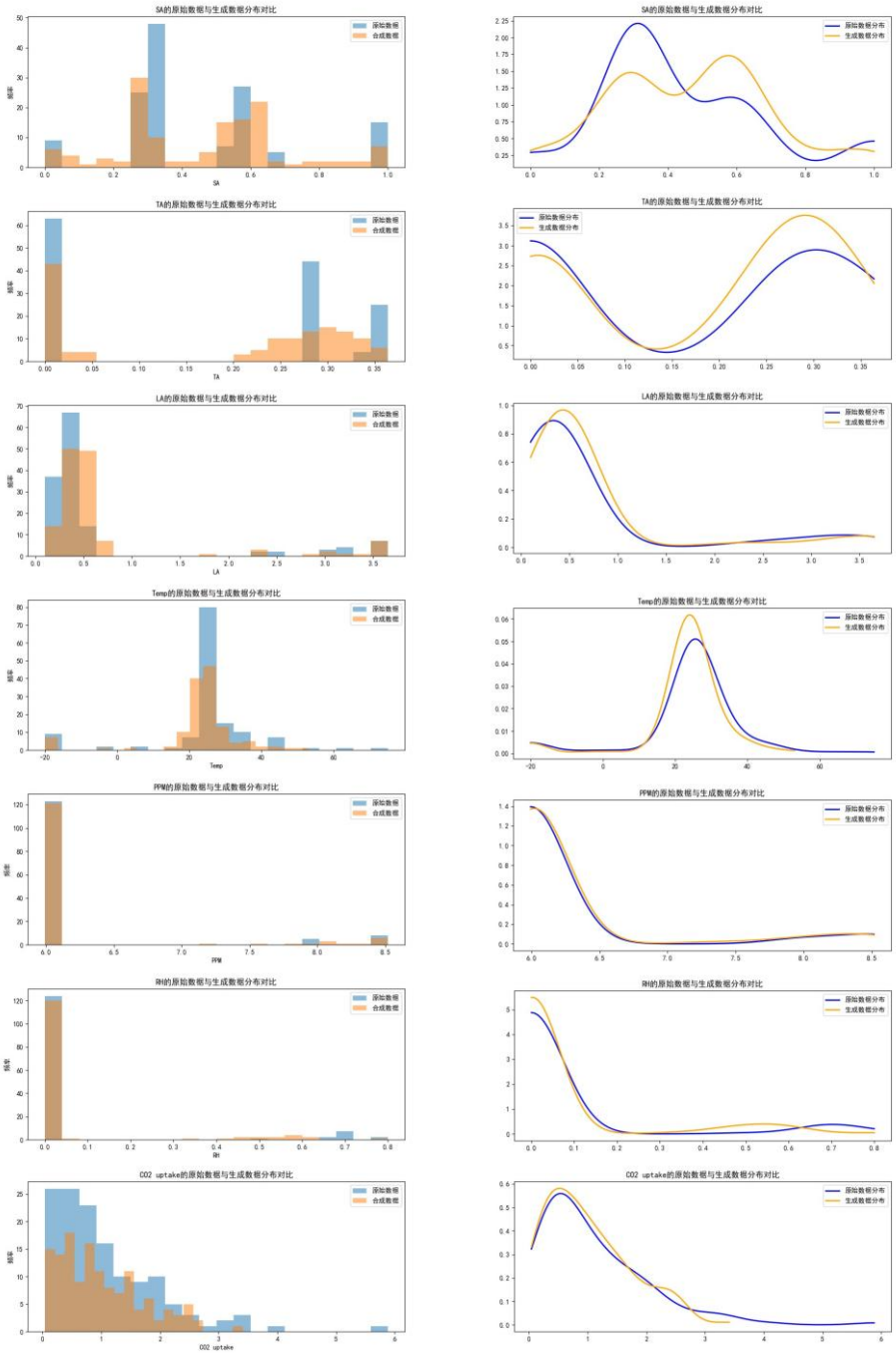

图 3.10　经进行对数变换时的 CTGAN 原始数据和生成数据的频率分布直方图（**左**）以及概率密度分布图（**右**），**左边**和**右边**从上往下的变量分别为仲胺百分比、叔胺百分比、气体承载量、温度、二氧化碳分压、相对湿度、二氧化碳吸附量

从上图 3.9 和图 3.10 对比 TVAE 来看，各个变量的分析中，对比单峰分布的变量，CTGAN 对于单峰的数据分布生成的质量要远高于 TVAE 生成的单峰数据的质量，主要的表现在于，针对所有单峰变量来看，它不会因为数据的集中频次分布而主要将数据生

成在峰度高的数据群里，这在图中主要表示在它不会远远高于原始数据的峰值。但是在多峰的情况之下，生成数据的峰度则是与原始数据的峰度高度是对调的，这主要表现在双峰数据分布中，局部高峰度的原始数据位置会产生最高峰度的生成数据，而最高峰度的原始数据位置会生成局部峰度的数据，反映出面对数据多峰度分布的情况之下，CTGAN 模型较 TVAE 来看，在处理多峰度数据的情况要弱于 TVAE

表 3.6 CTGAN 对数变换之后原始数据和生成数据对比

吸附剂的相关条件	原始数据平均数	生成数据平均数	原始数据标准差	生成数据标准差
吸附前比表面积	6.0201	5.9046	0.9065	0.8187
吸附后比表面积	4.2112	2.9788	1.3599	1.3666
吸附前孔隙体积	1.1985	1.0449	0.728	0.6475
吸附后孔隙体积	0.4256	0.3141	0.3295	0.2695
分子质量	6.499	6.0578	1.1793	1.1234
吸附剂氮原子含量	0.3111	0.3138	0.0713	0.0731
伯胺百分比	0.3887	0.336	0.02067	0.1359
仲胺百分比	0.4448	0.4678	0.2544	0.2402
叔胺百分比	0.1666	0.1846	0.1582	0.1406
气体承载量	0.7063	0.7467	0.9785	0.9021
温度	24.1544	22.9632	14.9512	12.0083
CO2 分压	6.2178	6.2559	0.6906	0.6833
相对湿度	0.0613	0.0614	0.199	0.1717
二氧化碳吸附量	1.0847	0.9905	0.9231	0.7219

从上表 3.6 的反映出的平均数和标准差来看，CTGAN 在吸附后比表面积当中所生成的平均数差距较大，另一个具有较大差值的则是吸附后的孔隙体积以及伯胺的百分比，回看原始数据和生成数据的图来看，这几个变量他们在概率密度分布上来看都是属于多峰分布的数据，因此反映出这几组数据在生成数据上是具有相关的偏置的。在标准差上，很多数据很好地解释了生成数据的离散度与原始数据的差异，由此证明 CTGAN 在数据的多方面上能够很好的抓住原始数据的分布特征不会使得数据具有很强的分散问题。

3.4 对照的模型对比

如图 3.11 所示，在进行性能预测建模的时候，本文选取了五种机器学习模型，通过所使用的数据进行基线模型的搭建，采用了线性回归模型、决策树模型、随机森林模型、梯度提升树模型以及支持向量机回归模型。在这五个模型中，线性模型的 MSE 为 0.6108、MAE 为 0.5514 这两个值相对较高，表示线性回归模型预测误差较大。然而线性回归的 R2 值为 0.4697 接近 0.5，阐述着线性回归模型解释约有 50%的方差。反映着线性回归模型可能只适用于简单的线性回归问题。

而决策树当中的 MSE 和 MAE 都要均大于线性回归模型，相关系数小于线性回归，在模型的测试集上充分说明决策树对于二氧化碳吸附剂数据的拟合效果不佳，同理还有梯度提升树。而支持向量机的表现效果最差，在决定系数上效果连 0.01 都没有到。反观，随机森林具有最好的性能，它的 MSE 为 0.4345，MAE 为 0.4504，其中它的决定系数接近 0.6895，在其中表现为最为优秀的模型。因此随机森林将会在后续数据生成中优先使用。

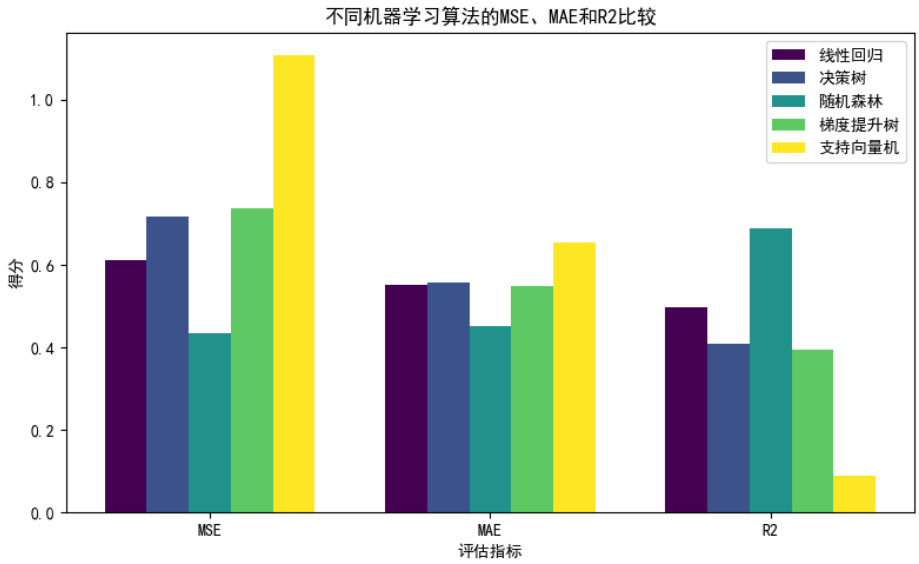

图 3.11　选取的五个机器模型的均方误差、平均绝对误差和决定系数

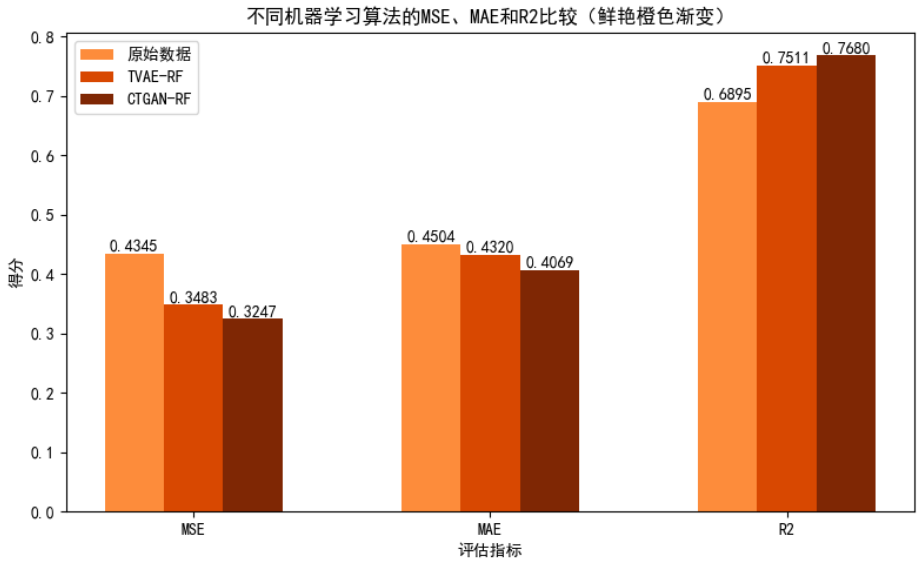

图 3.12　使用数据增强手段和不使用数据增强手段对预测性能的影响

数据生成之后，为了防止数据泄露，生成的数据将会被带入到训练数据当中，隔开测试集进行数据扩展的。如图 3.12，增强之后数据，使用随机森林对二氧化碳吸附剂进行性能预测，在均方误差上为 0.4345，平均绝对误差为 0.4504，决定系数为 0.6895。这代表着在不进行数据增强前随机森林的效果只有 70%的解释能力。本文使用的两个基于表格的数据增强模型中，在 TVAE 中，通过生成数据之后 MSE 要比没有使用的误差要降低到 0.3483，而使用 CTGAN 模型则降低到 0.3247。对于 MAE，使用 TVAE 模型则

降低到 0.4320，而 CTGAN 则表现出更加出色的能力达到了 0.4069。对于决定系数 R2 来说，两个数据生成模型较不使用数据增强手段的来说，提升在 9%以上，前者 MAE 的决定系数为 0.7511，而后者为 0.7680。综合全局表现进行观察，在进行回归处理提升的时候使用 CTGAN 进行数据增强可以使得预测效果具有最低的误差和最高的拟合度，这同时代表着数据增强手段在数据是小样本时通过数据的一些预处理可以使得预测效果更好，充分说明了数据增强手段可以使用在二氧化碳吸附剂性能的预测上。

3.5 SHAP 分析结果

SHAP 分析结果如图 3.13 和图 3.14 所示，基于 SHAP 特征重要性指标对数据特征名称进行降序排列，其中重要性程度最高的特征在图示中位于顶端位置。图 3.13 所展示的 SHAP 值量化了各输入特征对模型预测结果的贡献程度：当特征 SHAP 值为正值时，表明该特征对预测值产生正向促进作用；反之，负值则表征其具有抑制作用。值得注意的是，SHAP 值的绝对值大小与特征对预测结果的影响力呈正相关关系，即数值越大表明该特征对模型输出的影响程度越显著。而图 3.14 则显示每个特征的平均 SHAP 值，代表着每个输入特征对目标输出值的全局影响。这些图的信息要素，除了特征重要性会按要素排序之外，还可以阐述不同输入的特征值对输出变量的影响，反映输入特征对输出目标变量的依赖性。

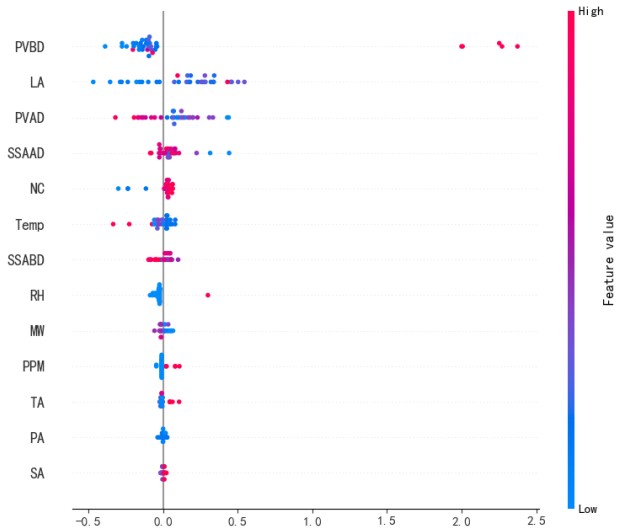

图 3.13　由 SHAP 分析获得的 SHAP 值

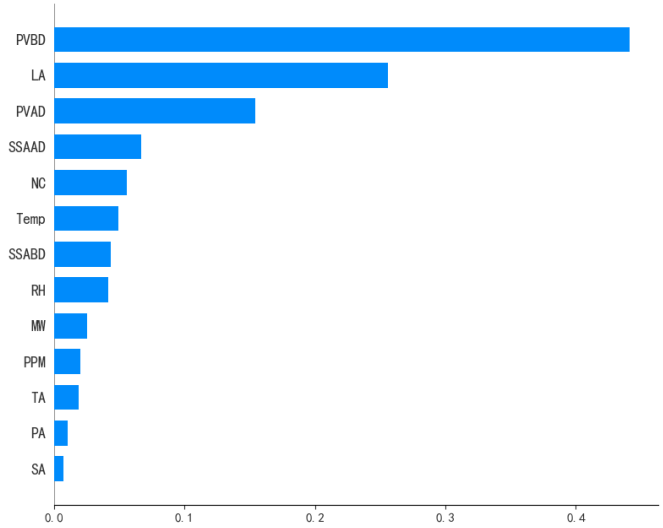

图 3.14　通过 SHAP 分析获得的平均 SHAP 值

结果表明吸附前的孔隙体积、以及气体承载量和吸附后的孔隙体积时影响二氧化碳吸附剂吸附性能的最有影响力的因素。如图 3.14 所示，这代表着虽然物理吸附对吸附的容量的贡献在低浓度二氧化碳当中只是很小的分区，但是固体胺材料的孔体积依旧影响对二氧化碳的亲和能力。此外，胺负载量的调整同时于吸附容量具有很强的相关性。这几个参数的显著影响说明固体胺材料在直接空气捕捉条件下主要涉及在载体上的选择使用一定合适质量的胺，并保证于二氧化碳充分接触即可，具有很好的成效。总的来说固体胺在物理性质上的影响是显著的，考虑到平均的 SHAP 值时，具有相当大的影响。对于这部分的相关优化将有助于实现吸附剂性能更好的发展。

4 结论

本文建立了一种基于小数据的生成数据二氧化碳吸附性能预测方法，基于现有的吸附性能以及机器学习理论基础，已经证实小数据经生成数据模型可以显著提升在数据样本较小情况下的预测性能。

本研究提出的方法主要包含四个关键环节：首先进行原始数据采集，随后基于数据特征相关性实施预处理操作，继而构建生成式模型以扩充数据集规模，最终通过增强后的数据完成回归预测任务。在这个过程中，生成数据模型语序任何形式的数据存在，但是在进行生成数据建模前，需要优先考虑到数据本身的趋势以及态势即数据它本身是高基数类别还是具有相关偏态的问题。然后预处理过后的数据将作为数据增强模型的输入，在本文中主要采用的则是 TVAE 和 CTGAN 两个数据增强模型。将输入特征输入带训练集的数据样本进行数据增强以及泛化。最后合并原始训练数据和基于原始数据生成的新数据将其带入到经典的机器学习预测回归类模型中预测输入和输出的特征的关系。

在本文中，主要考虑了影响吸附剂性能类别的三大性能：物理性能、化学性能和实验条件总共合计 13 个影响关键特征作为本文案例的输入，而材料的二氧化碳吸附吸收量作为输出特征。输入特征以及输出特征将被完全合并以输入到生成模型当中生成数据，然后放入到回归预测模型中训练输入特征与输出特征的关系，通过 SHAP 反映出不同输入对输出的重要性 。

此外，本研究通过对比分析基于数据生成技术构建的预测模型与传统预测模型的性能差异，系统考察了数据生成方法对模型预测效果的影响。实验结果表明在数据样本以及实验样本稀缺的情况之下，通过进行数据生成显著地突破了数据瓶颈，改善了模型泛化能力，使得模型在复杂场景下的性能提升大约 9%~12%，但是这样的方法依旧需要数据支持，在后续工作中可以添加吸附剂的物理约束，使得基于小样本数据模型性能更强、更稳定以及更有理论以及实验材料理论的支撑。

参考文献

[1] Law L C, Gkantonas S, Mengoni A, Mastorakos E. Onboard pre-combustion carbon capture with combined-cycle gas turbine power plant architectures for LNG-fuelled ship propulsion[J]. Applied Thermal Engineering, 2024, 248(B): 123294.

[2] Lie M, Kai T, Nakao S, et al. Modeling of pre-combustion carbon capture with CO_2-selective polymer membranes[J]. *International Journal of Greenhouse Gas Control*, 2023, 123: 103830.

[3] Jaffar M M, Rolfe A, Brandoni C, et al. A technical and environmental comparison of novel silica PEI adsorbent-based and conventional MEA-based CO2 capture technologies in the selected cement plant[J]. *Carbon Capture Science & Technology*, 2024, 10: 100179.

[4] Dziejarski B, Serafin J, Andersson K, et al. CO2 capture materials: a review of current trends and future challenges[J]. *Materials Today Sustainability*, 2023, 24: 100483.

[5] Li W, Fu D. Synergistic effect of surfactant and silica nanoflowers support on CO2 capture from simulated flue gas by solid amine adsorbents[J]. *Chemical Physics Letters*, 2024, 843: 141244.

[6] Kazemi A, Pordsari M A, Tamtaji M, et al. Environmentally friendly synthesis and morphology engineering of mixed-metal MOF for outstanding CO2 capture efficiency[J]. *Chemical Engineering Journal*, 2025, 505: 158951.

[7] Kabir S M M, Bhuiyan M I H. CWC-MP-MC Image-based breast tumor classification using an optimized Vision Transformer (ViT)[J]. *Biomedical Signal Processing and Control*, 2025, 100(Part B): 106941.

[8] Chen T. A fuzzy ubiquitous traveler clustering and hotel recommendation system by differentiating travelers' decision-making behaviors[J]. *Applied Soft Computing*, 2020, 96: 106585.

[9] Zhao P, Wei Y J, Wang S Y. Exploring behavior patterns in human and machine interactions[J]. *Fundamental Research*, 2023.

[10] Vakili A R, Rahimi M, Mashhadimoslem H, et al. Toward modeling the in vitro gas

production process by using propolis extract oil treatment: machine learning and kinetic models[J]. *Industrial & Engineering Chemistry Research*, 2022, 61(40): 14364–14374.

[11] Zhang H, Vo Thanh H, Rahimi M, et al. Improving predictions of shale wettability using advanced machine learning techniques and nature-inspired methods: Implications for carbon capture utilization and storage[J]. *Science of The Total Environment*, 2023, 877: 162944.

[12] Choudhary K, Yildirim T, Siderius D W, et al. Graph neural network predictions of metal–organic framework CO_2 adsorption properties[J]. *Computational Materials Science*, 2022, 210: 111388.

[13] Fathalian F, Aarabi S, Ghaemi A, et al. Intelligent prediction models based on machine learning for CO_2 capture performance by graphene oxide-based adsorbents[J]. *Scientific Reports*, 2022, 12: 21507.

[14] Ma X, Xu W, Su R, et al. Insights into CO_2 capture in porous carbons from machine learning, experiments and molecular simulation[J]. *Separation and Purification Technology*, 2023, 306: 122521.

[15] Zhang C, Li D, Xie Y, et al. Machine learning assisted rediscovery of methane storage and separation in porous carbon from material literature[J]. *Fuel*, 2021, 290: 120080.

[16] Xie C, Xie Y, Zhang C, et al. Explainable machine learning for carbon dioxide adsorption on porous carbon[J]. *Journal of Environmental Chemical Engineering*, 2023, 11: 109053.

[17] Burns T D, Pai K N, Subraveti S G, et al. Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO_2 capture based on integrated molecular simulations, process optimizations, and machine learning models[J]. *Environmental Science & Technology*, 2020, 54(7): 4536–4544.

[18] Guan J, Huang T, Liu W, et al. Design and prediction of metal organic framework-based mixed matrix membranes for CO_2 capture via machine learning[J]. *Cell Reports Physical Science*, 2022, 3(5): 100864.

[19] Li S, Huang Z, Li Y, Deng S, Cao X E. Methodology for predicting material performance by context-based modeling: A case study on solid amine CO_2 adsorbents[J]. *Energy and AI*, 2025, 20: 100477.

[20] Alizamir M, Keshavarz A, Abdollahi F, et al. Accurately predicting the performance of MOF-based mixed matrix membranes for CO_2 removal using a novel optimized extreme learning machine by BAT algorithm[J]. *Separation and Purification Technology*, 2023, 325: 124689.

[21] Daglar H, Gulbalkan H C, Habib N, et al. Integrating molecular simulations with machine learning guides in the design and synthesis of [BMIM][BF4]/MOF composites for CO_2/N_2 separation[J]. *ACS Applied Materials & Interfaces*, 2023, 15: 17421-17431.

[22] Zhu X, Tsang D C.W., Wang L, et al. Machine learning exploration of the critical factors for CO_2 adsorption capacity on porous carbon materials at different pressures[J]. *Journal of Cleaner Production*, 2020, 273: 122915.

[23] Amirkhani H, Anvar A, Dehghani M H. Modeling the solubility of CO_2 in porous liquids using machine learning techniques[J]. *Chemical Engineering Journal*, 2025, 471: 145054.

[24] Zhang S, Chen W, Mu J, et al. Single site Ni(II) anchored tetraethylenepentamine for enhancing CO_2 kinetic adsorption rate and long-term cyclic stability[J]. *Chemical Engineering Journal*, 2022, 436: 135211.

[25] Pakzad P, Mofarahi M, Izadpanah A A, et al. Experimental data, thermodynamic and neural network modeling of CO_2 absorption capacity for 2-amino-2-methyl-1-propanol (AMP) + Methanol (MeOH) + H_2O system[J]. *Journal of Natural Gas Science and Engineering*, 2020, 73: 103060.

[26] Le T, Epa V C, Burden F R, et al. Quantitative structure–property relationship modeling of diverse materials properties[J]. *Chemical Reviews*, 2012, 112(5): 2889–2919.

[27] Venkatraman V, Alsberg B K. Predicting CO_2 capture of ionic liquids using machine learning[J]. *Journal of CO_2 Utilization*, 2017, 21: 162-168.

[28] Al-Absi A A, Benneker A M, Mahinpey N. Amine sorbents for sustainable direct air capture: Long-term stability and extended aging study[J]. *Energy Fuels*, 2024, 38(10): 8938–8950.

[29] Lawson S, Griffin C, Rapp K, et al. Amine-functionalized MIL-101 monoliths for CO_2 removal from enclosed environments[J]. *Energy Fuels*, 2019, 33(3): 2399–2407.

[30] Wang J, Wang M, Li W, et al. Application of polyethylenimine-impregnated solid adsorbents for direct capture of low-concentration CO_2[J]. *AIChE Journal*, 2015, 61(3): 972–980.

[31] Goeppert A, Czaun M, May R B, et al. Carbon dioxide capture from the air using a polyamine based regenerable solid adsorbent[J]. *Journal of the American Chemical Society*, 2011, 133(50): 20164–20167.

[32] Thakkar H, Issa A, Rownaghi A A, et al. CO_2 capture from air using amine-functionalized kaolin-based zeolites[J]. *Chemical Engineering & Technology*, 2017, 40(11): 1999–2007.

[33] Kulkarni V, Panda D, Singh S K. Direct air capture of CO_2 over amine-modified hierarchical silica[J]. *Industrial & Engineering Chemistry Research*, 2023, 62(8): 3800–3811.

[34] Priyadarshini P, Rim G, Rosu C, et al. Direct air capture of CO_2 using amine/alumina sorbents at cold temperature[J]. *ACS Environmental Au*, 2023, 3(5): 295–307.

[35] Zhu X, Ge T, Yang F, et al. Efficient CO_2 capture from ambient air with amine-functionalized Mg–Al mixed metal oxides[J]. *Journal of Materials Chemistry A*, 2020, 8(32): 16421–16428.

[36] Sayari A, Liu Q, Mishra P. Enhanced adsorption efficiency through materials design for direct air capture over supported polyethylenimine[J]. *ChemSusChem*, 2016, 9(19): 2796–2803.

[37] Kuwahara Y, Kang D.-Y, Copeland J R, et al. Enhanced CO_2 adsorption over polymeric amines supported on heteroatom-incorporated SBA-15 silica: Impact of heteroatom type and loading on sorbent structure and adsorption performance[J]. *Chemistry - A European Journal*, 2012, 18(52): 16649–16664.

[38] Panda D, Kulkarni V, Singh S K. Evaluation of amine-based solid adsorbents for direct air capture: A critical review[J]. *Reaction Chemistry & Engineering*, 2023, 8(1): 10–40.

[39] Wang S, Liu Y, Zhang C, et al. High gravity-enhanced direct air capture: A leap forward in CO_2 adsorption technology[J]. *Atmosphere*, 2024, 15(2): 238.

[40] Sanz-Pérez E S, Fernández A, Arencibia A, et al. Hybrid amine-silica materials: Determination of N content by 29Si NMR and application to direct CO_2 capture from

air[J]. *Chemical Engineering Journal*, 2019, 373: 1286–1294.

[41] Chaikittisilp W, Kim H.-J, Jones C W. Mesoporous alumina-supported amines as potential steam-stable adsorbents for capturing CO_2 from simulated flue gas and ambient air[J]. *Energy & Fuels*, 2011, 25(11): 5528–5537.

[42] Zhu X, Lyu M, Ge T, et al. Modified layered double hydroxides for efficient and reversible carbon dioxide capture from air[J]. *Cell Reports Physical Science*, 2021, 2(7): 100484.

[43] Miao Y, He Z, Zhu X. Operating temperatures affect direct air capture of CO_2 in polyamine-loaded mesoporous silica[J]. *Chemical Engineering Journal*, 2021, 426: 131875.

[44] Sakwa-Novak M A, Yoo C.-J, Tan S, et al. Poly(ethylenimine)-functionalized monolithic alumina honeycomb adsorbents for CO_2 capture from air[J]. *ChemSusChem*, 2016, 9(14): 1859–1868.

[45] Sujan A R, Kumar D R, Sakwa-Novak M, et al. Poly(glycidyl amine)-loaded SBA-15 sorbents for CO_2 capture from dilute and ultradilute gas mixtures[J]. *ACS Applied Polymer Materials*, 2019, 1(11): 3137–3147.

[46] Cai H, Bao F, Gao J. Preparation and characterization of novel carbon dioxide adsorbents based on polyethylenimine-modified Halloysite nanotubes[J]. *Environmental Technology*, 2014, 36(10): 1284–1290.

[47] Zhang Y, Zhao S, Zhao S, et al. Revealing the correlation between the performance of silica-based DAC adsorbents and their pore natures[J]. *Gas Science and Engineering*, 2024, 123: 205251.

[48] Park S J, Lee J J, Hoyt C B, et al. Silica supported poly(propylene guanidine) as a CO_2 sorbent in simulated flue gas and direct air capture[J]. *Adsorption*, 2019, 26(1): 89–101.

[49] Wang S, Liu Y, Zhang C, et al. High gravity-enhanced direct air capture: A leap forward in CO_2 adsorption technology[J]. *Chemical Engineering and Processing: Process Intensification*, 2024, 181: 109735.

[50] Kortunov P V, Baugh L S, Siskin M. In situ nuclear magnetic resonance mechanistic studies of carbon dioxide reactions with liquid amines in mixed base systems: The

interplay of Lewis and Brønsted basicities[J]. *Energy & Fuels*, 2015, 29(9): 5967–5989.

[51] Shapiro S S, Wilk M B. An analysis of variance test for normality[J]. *Biometrika*, 1965, 52(3–4): 591–611.

[52] Tan Y, Zhu H, Wu J, Chai H. DPTVAE: Data-driven prior-based tabular variational autoencoder for credit data synthesizing[J]. *Expert Systems with Applications*, 2024, 241: 122071.

[53] Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional GAN[J]. *arXiv*, 2019.