同行专家业内评价意见书编号：20240860020

附件1

浙江工程师学院（浙江大学工程师学院）
同行专家业内评价意见书

姓名： 刘天涛

学号： 22160154

申报工程师职称专业类别（领域）： 生物与医药

浙江工程师学院（浙江大学工程师学院）制

2024年03月27日

一、个人申报

（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】

1.对本专业基础理论知识和专业技术知识掌握情况

申请人刘天涛，就读于浙江大学工程师学院生物与医药专业21级，为专业硕士研究生，在课程学习上，掌握了较为丰富的知识，平均课程成绩为83分，对AI制药领域的前沿应用有较好的认识。

2.工程实践的经历

申请人在浙江大学智能创新药物研究院开展了为期一年的专业实践，在实践期间，针对于AI在化学逆合成、生物合成的应用开展了研究。在中心赵文彬研究员的的专业指导下，基于深度学习技术和数据挖掘技术，构建大规模、标准化的化学反应数据集，采用最新的无监督学习、迁移学习、数据增强、预训练学习等技术，从而发展形成合理、高效、易于解释、对用户友好且完全数据驱动的逆合成路线预测体系，在AI+逆合成项目中，申请人成功地应用了计算机辅助逆合成的方法来合成目标化合物，并取得了以下成果：在AI与生物合成领域的结合应用中，申请人首次使用了迁移学习等策略，提升了生物合成的可用性、广泛性，该成果已在首届AI4science Hackathon中获得银奖，并被浙江大学智能创新药物研究院在公众号和网页报道；在AI与药物合成的结合中，申请通过细粒度聚类、神经机器翻译等技术开发出一个轻量化、可解释性强、准确性好、多样性佳的单步逆合成相关框架SynCluster，相较领域最佳方法的提升超过10%，并在多个上市药物的合成实例中得到了验证，该方法同样在首届AI4science Hackathon中获得银奖，并在ACS金色期刊JACS AU中发表(IF=8.0)，同时，SynCluster已经申请了软件著作权一项；申请人开发的方法兼容度高且易部署，可与各种模型、各种任务结合，作为深度学习与合成化学相结合的前沿交叉学科的具体实例，为合成化学中的智能化与自动化提供新的思路，缓解“创新药物设计”在化合物制取上存在的效率低、成本高等困境，在学位论文的盲审评审中，获得了专家的一致认可，取得了全优秀的成绩。

3.在实际工作中综合运用所学知识解决复杂工程问题的案例

申请人开发了一整套基于AI的合成路线预测技术体系，通过对于开源化学反应数据集的整理以及专利、论文等记录数据的挖掘，构建包含产率、结构、化学标识符、分子SMILES等信息的基准数据集。由此，采用一系列新型AI技术，完全数据驱动地完成对于目标化合物的单步逆合成拆解，并且基于多目标优化，拓展生成合理并且更符合化学直觉的全合成路径。该体系将提升有机合成设计的效率与系统性，加速对于类药空间的探索，缓解实际工作中的“创新药物研发”在化合物制取上存在的效率低、成本高等困境。

基于AI的合成路线预测技术体系的三个主要研究内容如下:

一是基于数据挖掘的化学反应数据库构建。针对于现有反应数据集存在的数据量不足、数据质量不高、开源程度不高等问题，关注于论文、专利等文献所包含的反应数据，尤其是记录详细、连续性强的全合成路径数据。我们将使用开源数据集与数据挖掘并行的方式，采用包括分子图、分子SMILES、反应凝聚图（CGR）等在内的多种特征信息提取方式，构建标准化A、易于部署的基准数据集。

二是基于数据驱动的的单步逆合成模型开发，申请人采用模板抽取、无监督聚类等方法

，重点关注于对于化学空间的二次划分以及对于完整化学反应简化嵌入，并使用大型语言模型，以端到端的方式直接构建逆合成模型。针对于目前部分单步模型存在的领域知识依赖、多样性不高的问题，开发的模型将完全数据驱动，并在生成多样化、高正确率的反应物上具有一定优势。

三是基于目标优化的多步搜索工具开发，该方法关注于如何完成由单步预测向全合成路径生成的推广，以契合现实的化学合成需求。将使用由反应与单个分子共同构建的搜索树，采用化学启发的价值函数，生成更可用的全合成路径。针对于目前该领域方法普遍存在的优化目标单一、评估方法模糊等问题，我们将从多目标优化的角度，并结合专家知识与数据驱动方法并重的评估手段，完成全局性的优化。

基于AI的合成路线预测技术体系的实施方案如下:

一是化学反应信息提取与基准数据集建立，主要基于爬虫与数据挖掘技术，一是完善已有的开源数据集，使用包括指纹、分子图与反应凝聚图在内的多种特征信息抽取技术，以转化为计算机可以读取的向量信息。针对于路径数据缺乏的问题，本方法将采用深度优先搜索（DFAS）等技术，在单步反应数据集中搜索出由可商购分子起始的反应路径。二是从论文、专利中获取全新数据，我们将构建一个用于反应提取的反应提取级联架构，通过文本匹配、角色标记等方法获取真实的反应路径，并对数据进行清洗和规范化。

二是细粒度聚类与端到端生成融合的逆合成预测模型，该方法的出发点源于化学家对于目标分子的拆解，即首先根据结构考虑不同的拆解类型。因此，本方法将基于亚结构指纹推荐目标分子的可拆解类型，并用端到端的方式生成反应物。具体方案如下：（1）
识别出在化学反应中发生变化的原子或者键，以识别到的变化原子为中心，根据不同的半径得到由SMARTS编码所代表的典型反应模板。（2）采用差异指纹将模板映射为向量，考察四种不同指纹的应用（包括Atompairs，ECFP4，FCFP4，以及TopologicalTorsions），由此完成反应嵌入，并基于Butina算法完成反应聚类，得到聚类生成的标签。（3）将聚类标签以词向量的形式嵌入大型语言模型，采用排序与原子约束算法，获得多样性与准确性较高的拆解前体。

三是以真实反应路径为优化目标的多步逆合成预测模型，当前的大部分多步逆合成搜索方法仅以简单的最短路径为优化目标，这忽略了在药物合成工艺设计中考虑其他的其他适当增长路径以获得更高的产率或者更低的成本等目标。因此，本方法将提出一个完全数据驱动，并高度符合现实合成需求的多步逆合成预测模型。具体包括：（1）基于半监督学习对于单步预测模型进行剪枝，以缩短单次推理时间，提高运算效率。（2）使用单步模型生成化学反应副产物，结合化学反应凝聚图与图神经网络技术，获得反应可及性评分。（3）基于所构建的真实反应路径数据集和深度优先搜索形成的反应路径数据集，使用长短期记忆人工神经网络（LSTM）嵌入，并获得对于反应路径的评分。（4）使用基于马尔可夫链（MDP）的决策树搜索，并在优化目标中考虑反应可及性以及路径的评分。

基于AI的合成路线预测技术体系实现的主要优势如下:

基于无监督方法和NMT模型，将细粒度的反应聚类、单步逆合成预测、正向合成预测、试剂预测、合成路线预测等任务相融合，大幅提升了模型的可解释性、合理性、多样性、准确性。此外，该作为一个兼容度高的方法，可以与各种任务、各种模型相适配，并始终提供准

确、多样化、易于解释的结果。在工程领域上，我们开发的方法在材料合成、药物发现、化学工艺优化等领域均有良好的应用前景。例如，在化学工艺优化领域，该方法可以帮助化学家在多种合成方法中选择最优的选项，并可进一步与自动化仪器与机器人平台开发相结合，以“虚拟实验”代替“试错实验”。近年来，“双碳”理论广受关注，工程师还可以使用我们的方法比较不同合成途径的环境影响，选择更绿色、可持续的方法，从而更好的贯彻“绿色化学”理念。

（二）取得的业绩（代表作）【限填3项，须提交证明原件（包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等）供核实，并提供复印件一份】

1.
公开成果代表作【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】

成果名称	成果类别 [含论文、授权专利（含发明专利申请）、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
SynCluster: Reaction Type Clustering and Recommendation Framework for Synthesis Planning.	国际期刊	2023年11月16日	JACS AU	1/8	
基于反应类型推荐的合成预测软件［简称：SynCluster］ V1. 0	计算机软件著作权	2023年03月01日	登记号：2023SR0228555	2/6	导师第一
基于深度学习的合成路线预测研究	学位论文送审专家评阅结果全优	2024年03月04日		1/1	

2.其他代表作【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】

（三）在校期间课程、专业实践训练及学位论文相关情况	
课程成绩情况	按课程学分核算的平均成绩： 83 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年（要求1年及以上） 考核成绩： 86 分（要求80分及以上）
本人承诺	
个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！ **申报人签名：**刘天博	

22160150

二、日常表现考核评价及申报材料审核公示结果

日常表现考核评价	非定向生由德育导师考核评价、定向生由所在工作单位考核评价： ☑优秀　☐良好　☐合格　☐不合格 德育导师/定向生所在工作单位分管领导签字（公章）：柳帅旗　年　月　日 2024.3.28
申报材料审核公示	根据评审条件，工程师学院已对申报人员进行材料审核（学位课程成绩、专业实践训练时间及考核、学位论文、代表作等情况），并将符合要求的申报材料在学院网站公示不少于5个工作日，具体公示结果如下： ☐通过　☐不通过（具体原因：　　） 工程师学院教学管理办公室审核签字（公章）：　　年　月　日

浙 江 大 学 研 究 生 院

攻读硕士学位研究生成绩表

学号：22160154	姓名：刘天涛	性别：男	学院：工程师学院	专业：制药工程	学制：2.5年
毕业时最低应获：24.0学分		已获得：24.0学分		入学年月：2021-09	毕业年月：2024-03
学位证书号：1033532024602175			毕业证书号：103351202402600401		授予学位：生物与医药硕士

学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2021-2022学年秋季学期	人工智能算法与系统		2.0	86	专业学位课	2022-2023学年冬季学期	先进制药技术		2.0	88	专业选修课
2021-2022学年夏季学期	研究生英语基础技能		1.0	免修	公共学位课	2022-2023学年秋冬季学期	工程伦理		2.0	93	公共学位课
2021-2022学年夏季学期	研究生英语		2.0	免修	公共学位课	2022-2023学年春季学期	自然辩证法概论		1.0	81	公共学位课
2022-2023学年冬季学期	新药发现理论与实践		2.0	95	专业学位课	2022-2023学年春季学期	数学建模		2.0	62	专业选修课
2022-2023学年秋冬季学期	科技创新案例探讨与实战		2.0	86	专业选修课	2022-2023学年春季学期	“四史”专题		1.0	82	公共选修课
2022-2023学年冬季学期	药物基因组学		2.0	86	专业选修课	2022-2023学年夏季学期	药品创制工程实例		2.0	89	专业学位课
2022-2023学年秋冬季学期	研究生论文写作指导		1.0	80	专业学位课	2023-2024学年秋季学期	新时代中国特色社会主义理论与实践		2.0	90	公共学位课

说明：1.研究生课程按三种方法计分：百分制，两级制（通过、不通过），五级制（优、良、中、及格、不及格）。

2.备注中“*”表示重修课程。

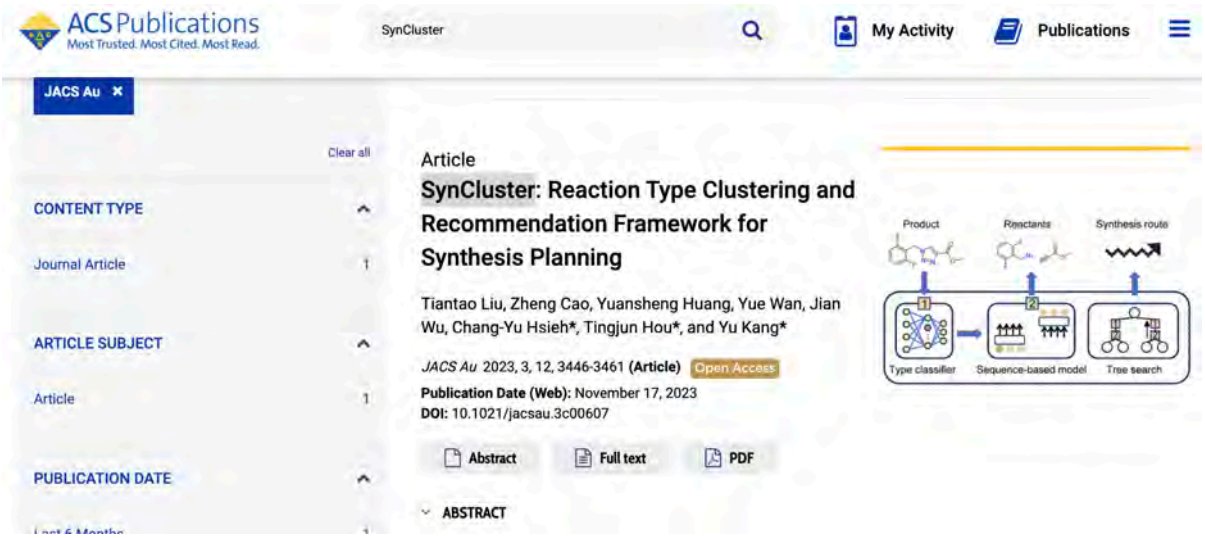
学院成绩校核章：

成绩校核人：张梦依

打印日期：2024-04-02

1. 发表论文证明

论文搜索页：



论文首页： 原文链接（https://pubs.acs.org/doi/10.1021/jacsau.3c00607）

JACS Au

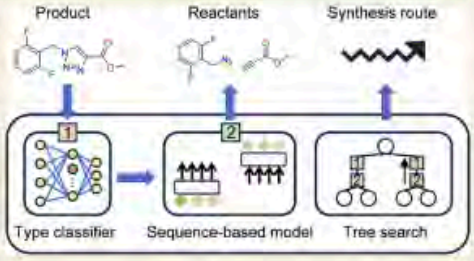
SynCluster: Reaction Type Clustering and Recommendation Framework for Synthesis Planning

Tiantao Liu, Zheng Cao, Yuansheng Huang, Yue Wan, Jian Wu, Chang-Yu Hsieh,* Tingjun Hou,* and Yu Kang*

Cite This: JACS Au 2023, 3, 3446−3461

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



AI-assisted synthesis planning has emerged as a valuable tool in accelerating synthetic chemistry for the discovery of new drugs and materials. The template-free approach, which showcases superior generalization capabilities, is seen as the mainstream direction in this field. However, it remains unclear whether such an end-to-end approach can achieve problem-solving performance on par with experienced chemists without fully revealing insights into the chemical mechanisms involved. Moreover, there is a lack of unified and chemically inspired frameworks for improving multitask reaction predictions in this area. In this study, we have addressed these challenges by investigating the impact of fine-grained reaction-type labels on multiple downstream tasks and propose a novel framework named SynCluster. This framework incorporates unsupervised clustering cues into the baseline models and identifies plausible chemical subspaces which is compatible with multitask extensions and can serve as model-independent indicators to effectively enhance the performance of multiple downstream tasks. In retrosynthesis prediction, SynCluster achieves significant improvements of 4.1 and 11.0% in top-1 and top-10 prediction accuracy, respectively, compared to the baseline Molecular Transformer, and achieves a notable enhancement of 13.9% in top-10 accuracy when combined with Retroformer. By incorporating simplified molecular-input line-entry system augmentation, our framework achieves higher top-10 accuracy compared to state-of-the-art sequence-based retrosynthesis models and improves over the baseline on the diversity and validity of reactants. SynCluster also achieves 94.9% top-10 accuracy in forward synthesis prediction and 51.5% top-10 Maxfrag accuracy in reagent prediction. Overall, SynCluster provides a fresh perspective with chemical interpretability and reinforcement of domain knowledge in the synthesis design. It offers a promising solution for improving the accuracy and efficiency of AI-assisted synthesis planning and bridges the gap between template-free approaches and the problem-solving abilities of experienced chemists.

KEYWORDS: *synthesis planning, fine-grained type, transformer, unsupervised clustering*

INTRODUCTION

Synthesis planning[1] is a critical process in drug discovery and chemical industry that involves fabricating reasonable pathways for synthesizing given compounds. It entails multiple reasoning tasks including forward synthesis prediction, reagent selection, and retrosynthesis. The first two tasks deduce the possible underlying product or reagents by providing building blocks or complete reactions; while retrosynthesis operates in reverse and requires logical disconnection of the desired starting molecules, searching a vast space of possible chemical transformations from a given state.[2] With the continuous expansion of accessible chemical space, the rapid growth in the number of synthetic molecules has made traditional manual treatment of this process inefficient.[3] As a result, computer-aided approaches are now in urgent demand to overcome the limitations of manual methods.

Computer-aided retrosynthetic analysis strategies can be divided into two main categories,[4] template-based and template-free models. The template-based model involves matching generalized reaction rules with target molecules to produce one or more candidate precursors based on defined subgraph patterns of the chemical reaction.[5] Most template-based models[6−9] leverage a classification model for selecting a suitable reaction template, which can be interpreted as a process of searching and matching within established domain knowledge, rather than generating new chemical knowledge

Received: October 7, 2023
Revised: November 7, 2023
Accepted: November 8, 2023
Published: November 17, 2023

https://doi.org/10.1021/jacsau.3c00607
JACS Au 2023, 3, 3446−3461

2. 发表软著证明

该生（刘天涛）在软著基于反应类型推荐的合成预测软件［简称：SynCluster］V1. 排名第二，导师排名第一，且该软著和该生学位论文相关。

导师签字：

软著截图：

中华人民共和国国家版权局

计算机软件著作权登记证书

证书号：软著登字第10815726号

软件名称：基于反应类型推荐的合成预测软件
[简称：SynCluster]
V1.0

著作权人：浙江大学

开发完成日期：2022年11月01日

首次发表日期：2022年11月02日

权利取得方式：原始取得

权利范围：全部权利

登记号：2023SR0228555

根据《计算机软件保护条例》和《计算机软件著作权登记办法》的规定，经中国版权保护中心审核，对以上事项予以登记。

中华人民共和国国家版权局
计算机软件著作权登记专用章

No. 12253715

2023年02月13日

证　明

兹证明软件名称为“基于反应类型推荐的合成预测软件V1.0”，著作权人为：浙江大学，开发人员为：康玉、刘天涛、谢昌谕、侯廷军、潘培辰，流水号为：2022R11L2343988的计算机软件著作权委托杭州天勤知识产权代理有限公司代为办理，目前正在受理中，特此证明。

杭州天勤知识产权代理有限公司

2023-2-1

3. 三优学位论文证明

研究生院系统：

论文目录：

目录

论文摘要：

中文摘要

合成路线设计是从可用的起始材料出发，通过一系列可行的反应步骤，合成目标化学物的过程。合理的合成路线设计可以高效地合成在日化、药学等领域发挥重要作用的分子，因而成为了材料化学、药物化学等学科的基础和核心。然而，人工进行的合成路线设计需要大量的领域知识，同时消耗大量的时间，并可能受到由不同决策者带来的不稳定性的影响。因此，人们开始使用计算机辅助的合成路线预测以提升有机合成设计的效率与系统性。但是，早期的相关工作高度依赖反应规则以及各类启发式方法以拓展和搜索合成网络，这仍然需要大量领域知识，并难以探索未知的化学空间。

深度学习（Deep Learning，DL）通过多个基本处理层组成的计算模型学习到各种抽象的数据表示，相比传统方法在泛化能力、对非专业人员友好度等方面上表现更好。然而，基于 DL 的合成路线预测模型仍存在可解释性不足、多样性不高等问题。因此，亟需完全数据驱动，并高度符合现实合成需求的 DL 模型。基于以上需求，本文综合了神经机器翻译（Neural Machine Translation，NMT）、图神经网络（Graph Neural Network，GNN）、数据增强（Data augmentation）、知识蒸馏（Knowledge Distillation，KD）等 DL 技术，开发了一个合理、高效、易于解释、对用户友好且完全数据驱动的合成路线预测体系。

本文的第二章开发了一个完全数据驱动的单步逆合成相关模型的预测框架，名为 SynCluster。该框架将细粒度的反应聚类信息与单步逆合成、正向合成、试剂预测等任务相融合，完全符合化学直觉并易于理解。为了获取反应聚类信息，我们将识别出反应中发生变化的原子或者键，之后根据不同的半径抽取反应模板，并采用差异指纹将模板映射为向量，最后基于 Butina 算法得到聚类类型；为了与下游任务融合，我们通过 SMILES 和相应的聚类类型来训练下游 NMT 模型（例如 Molecular Transformer（MT）或 Retroformer）。针对于反应聚类，SynCluster 与现行通用专家模型 NameRxn 相比，取得了 0.784 的耦合度（纯度），并对部分不准

确的反应分类实现了修正，展示出其对于化学知识的深刻理解。针对于单步逆合成任务，SynCluster 展现出良好的模型兼容性：在 top-10 预测准确度上，分别使基线模型 MT 和 Retroformer 提升 11.0%和 13.9%，同时大幅改善了现有模型在多样性、合理性、可解释性上的不足。针对于正向合成预测和试剂预测任务，SynCluster 的附加模型在 top-10 准确度和 top-10 最大片段准确度分别达到了 94.9%和 51.5%，展示出其良好的多任务适配。

本文的第三章主要描述了单步模型的场景推广，即基于树搜索的多步合成路线预测。我们首先查看了各个单步模型对于文献路径的复现能力以及推理耗时，并提出了一种全新的 KD 方法以降低计算开销。具体而言，我们设置了具有 4 层解码器和编码器的 Retroformer 与 SynCluster 融合模型作为学生模型，从多个特征学习具有 8 层解码器和编码器的教师网络。之后，各个单步模型将与树搜索方法 Retro*结合，并结合验证机制，以完成多步合成路线预测。结果显示，当 MT 和 SynCluster 的与树搜索的融合模型参与多步合成路线预测时，取得的拆解成功率为 27.9%，路径长度为 3.2，均超过基线模型 AT，反映出 SynCluster 与多步搜索具有有良好适配性。同时，我们提出的 KD 方法，可以降低 43%的单步模型计算开销，21%的多步搜索计算开销。我们的方法减小了基于 NMT 的模型的计算开销，拓展了其使用场景，改善了传统方法效率低、评估方式单一等问题。

综上所述，本文首先通过细粒度聚类、NMT 等技术开发出一个轻量化、可解释性强、准确性好、多样性佳的单步逆合成相关框架 SynCluster；之后结合 KD 与树搜索等技术，开发出以真实的合成需求为导向的多步合成路线预测方法。本文揭示了 DL 模型如何从数据集中提取化学知识，开发的方法兼容度高且易部署，可与各种模型、各种任务结合。由此，作为深度学习与合成化学相结合的前沿交叉学科的具体实例，本文将为合成化学中的智能化与自动化提供新的思路，缓解“创新药物设计”在化合物制取上存在的效率低、成本高等困境。

关键词：合成路线预测、反应聚类、神经机器翻译、树搜索、药物设计

论文展望：

第四章 总结与展望

本文以开发一个合理、高效、易于解释、对用户友好且完全数据驱动的合成路线预测体系为目的，首先提出了一个基于细粒度聚类的单步逆合成相关框架，并探究了其与树搜索结合的能力，最终在多步合成路线预测上完成应用，为合成化学中的智能化与自动化提供了新的思路。

在本文的第二章中，我们首先引入了一种创新性的两阶段框架，名为SynCluster，该框架引入了化学家处理有机合成的思路，将反应预测分为二阶段：首先根据目标产物判断可能的反应类型；其后再进行完整的反应物设计。结果显示，针对于反应聚类，SynCluster 与现行通用专家模型 NameRxn 相比，取得了0.784 的耦合度（纯度），并对部分不准确的反应实现了修正，展示出其对于化学知识的深刻理解。针对于单步逆合成任务，SynCluster 展现出良好的模型兼容性：在 top-10 预测准确度上，分别使基线模型 MT 和 Retroformer 提升 11.0%和 13.9%，同时大幅改善了现有模型在多样性、合理性、可解释性上的不足。针对于正向合成预测和试剂预测任务，SynCluster 的附加模型在 top-10 准确度和 top-10 最大片段准确度分别达到了 94.9%和 51.5%，展示出其良好的多任务适配。相比于常见的基于模板的方法，经由 SynCluster 的输出完全由 NMT 模型给出，因而不受模板库有限拆解方式的限制，拥有较高的创新性与多样性，相比于常见的无模版的方法，经由 SynCluster 的方法通过两阶段的运行方式，缓解了 NMT 模型可解释性弱的缺点；相比于产物一合成子一反应物的工作流，SynCluster 的附加方式更加轻量化，具有更高的适配度。

在第三章，我们使用了树搜索、KD 等技术，将单步逆合成模型发展成快速、有效、合理性改善的多步合成路线预测模型。结果显示出 SynCluster 框架与树搜索框架 Retro*仍具有良好的适配性。当“MT+SynCluster”与树搜索与验证机制结合后，取得的拆解成功率为 27.9%，路径长度为 3.2，均超过基线模型 AT。同时，我们提出的 KD 方式，可以降低 43%的单步模型计算开销，21%的多步搜索计算

开销，而引起的单步模型 top-1 准确度下降以及多步模型成功率下降仅为 1.9%和 1.1%。我们的方法改良了基于 NMT 的模型的计算开销，拓展了其使用场景，并且首次引入了基于文献的评估分数 SLScore，弥补了先前工作仅使用路径长度作为评估依据的不足。

本文的基本贡献是，基于无监督方法和NMT模型，将细粒度的反应聚类、单步逆合成预测、正向合成预测、试剂预测、合成路线预测等任务相融合，大幅提升了模型的可解释性、合理性、多样性、准确性。此外，我们的方法作为一个兼容度高的方法，可以与各种任务、各种模型相适配，并始终提供准确、多样化、易于解释的结果。在工程领域上，我们开发的方法在材料合成、药物发现、化学工艺优化等领域均有良好的应用前景。例如，在化学工艺优化领域，该方法可以帮助化学家在多种合成方法中选择最优的选项，并可进一步与自动化仪器与机器人平台开发相结合，以“虚拟实验”代替“试错实验”。近年来，“双碳”理论广受关注[132]，工程师还可以使用我们的方法比较不同合成途径的环境影响，选择更绿色、可持续的方法，从而更好的贯彻“绿色化学”理念。

本研究中几个问题主要与数据集的局限性有关。例如，数据集中的不对称反应较少，可能会导致模型难以在高优先级输出该类反应。除此之外，由于数据集中副产物记录存在大量缺失的情况，预测离去集团（例如卤素的类型）对于模型较为困难。同时，数据集存在一部分高质量数据和低质量数据混杂的情况，可能也会影响模型预测。因此，我们认为对化学反应数据记录进行标准化、开源化是至关重要的。在今后的进一步研究中，可以结合文本挖掘工具，从文献或者专利中自动提取化学反应数据，以充实数据集。

尽管如此，我们的目标是为化学反应中的 DL 模型提供一种通用技术。我们未来的工作将探索反应聚类的进一步拓展，例如探讨不同反应表示方法在聚类中的应用。同时，我们将探索 SynCluster 框架在其他合成路线预测相关任务上的应用，例如产率预测、副产物预测等，以提高该框架的适配范围。

申请人在 Hackathon AI4science 中获得两项银奖

浙江大学智能创新药物研究院在首届AI4Science Hackathon中喜获银奖

2022.06.09

2022年5月29日，浙江大学智能创新药物研究院侯廷军研究员和吴健研究员项目组成员在首届AI4Science黑客马拉松竞赛，与来自新加坡国立大学、芝加哥大学、清华大学、复旦大学、华中科技大学等高校队伍的激烈角逐中分别取得了两项赛道银牌的优异成绩。

AI4science是加州理工学院的一项倡议，旨在将 AI 研究人员与其他学科的专家聚集在一起，将现代 AI 工具推向科学和工程的各个领域。2022年5月6日至5月29日，DeepVerse深鱼科技联合Bota Bio恩和生物、Chemical.AI智化科技以及Foreseen昱言生物共同举办了全国首届AI4Science黑客马拉松竞赛，包括物理学、化学、蛋白质组学和合成生物学等四项挑战赛。

在研究院侯廷军研究员、吴健研究员和康玉博士的指导下，研究生曹政和刘天涛组成Team TQL团队，参与了化学赛道和合成生物学赛道的项目研发。在化学赛道中，团队提出了一种基于图网络的化学反应副产物预测模型和化学可合成型评估模型；在合成生物学赛道中，团队提出了一种基于迁移学习的酶促反应预测系统和多步预测拓展模型。相关模型已经开源在GitHub上并且取得了各赛道评委的一致好评。

网址（http://zjuaim.com/news_detail/id-60.html）

获奖证书：