

人工智能预测模型在有机化学实验教学中的应用初探

周成卓, 谢召军*

南开大学材料科学与工程学院, 天津 300350

摘要: 有机化学实验是化学及其相关专业的基础必修课程, 合成实验是其中的核心教学内容。本文将人工智能预测模型引入当前有机化学实验中的合成实验, 选取ASKCOS模型进行正向合成预测和逆合成预测。预测结果显示, 现有模型对于课程涉及的合成实验的预测效果良好, 正向合成预测和逆合成预测的命中率最高可达到88.5%。人工智能预测模型在本科实验教学中的应用有助于加深学生对有机化学合成反应特点的理解, 增进人工智能辅助有机合成的理念, 促进学科交叉, 提高教学质量, 有助于培养学生的发散性思维, 为他们今后的科研工作打下基础。

关键词: AI模型; 有机合成; 正向合成预测; 逆合成预测

中图分类号: G642.0; O6-3; O621.3

Exploring the Application of Artificial Intelligence Prediction Models in Organic Chemistry Laboratory Teaching: A Preliminary Study

Chengzhuo Zhou, Zhaojun Xie *

School of Materials Science and Engineering, Nankai University, Tianjin 300350, China.

Abstract: Organic chemistry laboratory is a fundamental, compulsory course in chemistry and related disciplines, with synthesis experiments at their core. This paper introduces artificial intelligence prediction models into the teaching of synthesis experiments in organic chemistry courses. Specifically, the ASKCOS model is applied to perform both forward synthesis prediction and retrosynthesis prediction. The results indicate that the existing model provides reliable predictions for the synthesis experiments covered in the curriculum, with hit rates for both forward and retrosynthesis predictions reaching as high as 88.5%. The integration of AI prediction models into undergraduate laboratory teaching enhances students' understanding of organic chemical synthesis, promotes the concept of AI-assisted organic synthesis, fosters interdisciplinary collaboration, improves teaching quality, and encourages divergent thinking, thereby laying a strong foundation for their future research work.

Key Words: AI Model; Organic synthesis; Forward synthesis prediction; Retrosynthesis prediction

有机化学实验不仅是化学类各专业的必修课, 也是生命科学类、材料类、环境科学与工程类、药学类、医学类、食品科学类等专业的必修课程之一, 与有机化学理论课配套开设并纳入其培养计划。在有机化学实验的课程内容设计中, 通常包含基本操作的练习, 如常压蒸馏、重结晶、分液等, 以及合成实验和天然产物提取和分离实验等。其中, 合成实验是教师在教学中的关注重点, 而且是与理论课关系最为密切的部分。以南开大学开设的有机化学实验课程为例, 对于非化学类专业而言, 有机化学实验课程68学时, 共12个实验项目, 其中合成实验7个, 比例超过50%。化学类专

收稿: 2024-05-10; 录用: 2024-09-05; 网络发表: 2024-09-14

*通讯作者, Email: zjxie@nankai.edu.cn

基金资助: 南开大学实验教学改革项目(23NKSJYC06); 国家自然科学基金重点项目(21933006)

业的有机化学实验共136学时，分两学期开课，第一学期14个实验中的6个是合成实验，而第二学期课程全部都是合成实验。因此，合成实验的教学对于学生理解和掌握有机化学的相关内容，并顺利过渡到今后的科研阶段至关重要。

在有机化学实验课中，学生会亲身体会到有机合成反应的特点，包括反应的复杂性、可逆性、对反应条件的敏感性以及反应过程时间较长等^[1-3]。因此，为了将实验课上学到的知识应用于未来的科研工作，教师需要引导学生进行如下思考：一方面，在给定反应物和反应条件的情况下，能否快速预测可能的产物，完成前端筛选；另一方面，在给定目标化合物的情况下，如何快速选择合适的原料和反应。后者即为逆合成分析^[4]的思路，将复杂的产物通过多步分解，最终过渡到廉价易得的起始原料。但在面对大规模或未知化合物时，单独依靠手工经验判断和手动搜索进行逆合成分析是一项艰巨的工作。

为应对这一挑战，众多研究正致力于探索将人工智能(Artificial Intelligence, AI)技术与有机化学反应的深度融合^[5-7]，尤其是利用各种基于机器学习的人工智能算法建立模型^[8-10]。机器学习是人工智能的一个重要分支，它通过计算机从现有的大量数据中通过自我学习的方式进行预测，从而避免了人工判断对客观数据的主观性影响，这种方法与有机化学反应的复杂性和数量多的特点相契合。随着算法的不断升级和完善，化学反应模型的预测准确度也在不断提高^[11,12]。

同时，计算化学的相关理念也早已融入化学及其相关专业的课程教学中，比如计算机辅助的药物设计^[13]、计算材料学^[14]等，这些方法不需要进行实验，而是通过计算机模拟的方式来预测实验结果，具有前期预先筛选、节省实验成本、减轻实验工作量等诸多优点。最近，已经有研究将计算化学引入到有机化学教学中，利用Gaussian软件进行量子化学计算^[15]，将基于人工智能的预测模型融入教学可以作为计算化学模块的一个良好扩展^[16]。

基于上述讨论，本文拟选取一种有机化学反应的正向预测和逆合成预测的人工智能模型，将其运用于有机化学实验课程中的合成实验，通过预测结果评估该模型的可行性，并进一步在教学中推广应用。

1 实验内容及预测模型的选取

1.1 实验内容的选取

本文选用我校现有教学计划^[17]中已开设多年的有机合成实验——有机化学实验、有机化学实验2-1和有机化学实验2-2，授课对象包括化学类、分子科学与工程、化学生物学、新能源科学与工程、生物科学类、材料类、环境科学与工程类、药学类专业的一年级、二年级本科生。这些实验的反应物、反应试剂、溶剂、产物的化学物质名称及其简化分子线性输入规范(Simplified molecular input line entry system, SMILES)编码如表1所示。

1.2 预测模型的选取

目前，用于预测有机化学反应的人工智能模型分为两类：一类是付费的商业模型，例如默克公司的Chematica^[11]和SRI公司的SynRoute^[18]等，它们在产业中得到了广泛应用，但与高校的应用环境不相符合；另一类是开源免费模型，其源代码通常在GitHub上可公开下载，为科研人员提供了便利，但其在计算机上离线运行基于Python环境，需要学习相关的程序知识，门槛相对较高，对于本科生学习使用有一定的难度。因此，一个可视化程度高、操作便利、结果直观、省时便利的模型更为合适。

针对本科教学的相关需求，具有图形化界面的在线网页系统是一种非常合适的形式。目前，IBM公司的RXN模型^[19]和麻省理工大学开发的基于知识的连续有机合成自动化系统(Automated System for Knowledge-Based Continuous Organic Synthesis, ASKCOS)模型^[20]具备这些优点。RXN模型逆合成的目标反应物命中率最高可达95.3%^[21]，正向预测的准确率最高可达90.4%^[22,23]，ASKCOS模型

表1 实验涉及的反应物、生成物、反应试剂和溶剂的名称及其SMILES号

实验名称	反应物	产物	催化剂或反应试剂	溶剂
乙酸正丁酯的制备	乙酸 <chem>CC(=O)O</chem> 、 正丁醇 <chem>CCCCO</chem>	乙酸正丁酯 <chem>CCCCOC(=O)C</chem>	硫酸 <chem>OS(=O)(=O)O</chem>	
乙酰水杨酸的制备	水杨酸 <chem>c1ccc(c(c1)C(=O)O)O</chem> 、 乙酸酐 <chem>CC(=O)OC(=O)C</chem>	乙酰水杨酸 <chem>CC(OC1=C(C(=O)O)C=CC=C1)=O</chem>	硫酸 <chem>OS(=O)(=O)O</chem>	
二苯甲醇的制备	二苯甲酮 <chem>c1ccc(cc1)C(=O)c2ccccc2</chem>	二苯甲醇 <chem>c1ccc(cc1)C(c2ccccc2)O</chem>	硼氢化钠 <chem>[BH4-].[Na+]</chem>	乙醇 CCO
正丁醚制备	正丁醇 <chem>CCCCO</chem>	正丁醚 <chem>CCCCOCCC</chem>	硫酸 <chem>OS(=O)(=O)O</chem>	
卤代烷的制备	正丁醇 <chem>CCCCO</chem> 、 溴化氢 <chem>Br</chem> 、氯化氢 <chem>Cl</chem>	1-溴丁烷 <chem>CCCCBr</chem> 、 1-氯丁烷 <chem>CCCCCl</chem>	硫酸 <chem>OS(=O)(=O)O</chem>	
环己烯的制备	环己醇 <chem>C1CCC(CC1)O</chem>	环己烯 <chem>C1CCC=CC1</chem>	磷酸 <chem>OP(=O)(O)O</chem>	
环己酮的制备	环己醇 <chem>C1CCC(CC1)O</chem>	环己酮 <chem>C1CCC(=O)CC1</chem>	铬酸 <chem>O=[Cr](=O)(O)O</chem>	水O
肉桂酸的制备	苯甲醛 <chem>c1ccc(cc1)C=O</chem> 、 乙酸酐 <chem>CC(=O)OC(=O)C</chem>	肉桂酸 <chem>c1ccc(cc1)/C=C/C(=O)O</chem>	碳酸钾 <chem>CC(=O)[O-].[K+]</chem>	
固体研磨法合成2-(3,4-二甲氧基苯亚甲基)-2,3-二氢-1-茛菪酮	2,3-二氢-1-茛菪酮 <chem>COc1ccc(cc1OC)C=O</chem> 、 3,4-二甲氧基苯甲醛 <chem>c1ccc2c(c1)CCC2=O</chem>	2-(3,4-二甲氧基苯亚甲基)-2,3-二氢-1-茛菪酮 <chem>COc1ccc(C=C2Cc3ccccc3C2=O)cc1OC</chem>	氢氧化钠 <chem>[OH-].[Na+]</chem>	
1,2,3,4-四氢吡啶的合成	环己酮 <chem>C1CCC(=O)CC1</chem> 、 苯胍 <chem>c1ccc(cc1)NN</chem>	1,2,3,4-四氢吡啶 <chem>c1ccc2c(c1)c3c([nH]2)CCCC3</chem>	乙酸 <chem>CC(=O)O</chem>	
邻氨基苯甲酸的制备	邻苯二甲酸酐 <chem>c1ccc2c(c1)C(=O)OC2=O</chem> 、 氨水 <chem>[NH4+].[OH-]</chem> 、 邻苯二甲酰胺 <chem>c1ccc2c(c1)C(=O)NC2=O</chem> 、 氢氧化钠 <chem>[OH-].[Na+]</chem>	邻苯二甲酰胺 <chem>c1ccc2c(c1)C(=O)NC2=O</chem> 、 邻氨基苯甲酸 <chem>c1ccc(c(c1)C(=O)O)N</chem>		
苄基三乙基氯化铵的制备	氯化苄 <chem>c1ccc(cc1)CCl</chem> 、 三乙胺 <chem>CCN(CC)CC</chem>	苄基三乙基氯化铵 <chem>CC[N+](CC)(CC)Cc1ccccc1.[Cl-]</chem>		
7,7-二氯二环[4.1.0]庚烷的制备	环己烯 <chem>C1CCC=CC1</chem> 、 氯仿 <chem>C(Cl)(Cl)Cl</chem>	7,7-二氯二环[4.1.0]庚烷 <chem>C1CCC2C(C1)C2(Cl)Cl</chem>	氢氧化钠 <chem>[OH-].[Na+]</chem>	
呋喃甲醇和呋喃甲酸的制备	呋喃甲醛 <chem>c1cc(oc1)C=O</chem>	呋喃甲醇 <chem>c1cc(oc1)CO</chem> 、 呋喃甲酸 <chem>c1cc(oc1)C(=O)O</chem>	氢氧化钠 <chem>[OH-].[Na+]</chem>	
微波辐射合成苯甲酸乙酯	甲苯 <chem>Cc1ccccc1</chem> 、 苯甲酸 <chem>c1ccc(cc1)C(=O)O</chem> 、 乙醇 <chem>CCO</chem>	苯甲酸乙酯 <chem>CCOC(=O)c1ccccc1</chem>	高锰酸钾 <chem>[O-][Mn](=O)(=O)O.[K+]</chem> 、 硫酸 <chem>OS(=O)(=O)O</chem>	乙醚 CCOCC
Synthesis of Triphenylmethanol	镁 <chem>[Mg]</chem> 、溴苯 <chem>c1ccc(cc1)Br</chem> 、 苯基溴化镁 <chem>c1ccc(cc1)[Mg]Br</chem> 、 苯甲酸乙酯 <chem>CCOC(=O)c1ccccc1</chem>	苯基溴化镁 <chem>c1ccc(cc1)[Mg]Br</chem> 、 三苯甲醇 <chem>c1ccc(cc1)C(c2ccccc2)(c3ccccc3)O</chem>		乙醚 CCOCC

(待续)

续表1

实验名称	反应物	产物	催化剂或反应试剂	溶剂
Reaction of Triphenylmethanol	三苯甲醇 <chem>c1ccc(cc1)C(c2ccccc2)(c3ccccc3)O</chem>	三苯基甲基溴 <chem>BrC(c1ccccc1)(c1ccccc1)c1ccccc1</chem>		乙酸 <chem>CC(=O)O</chem>
	溴化氢Br 三苯甲醇 <chem>c1ccc(cc1)C(c2ccccc2)(c3ccccc3)O</chem>	三苯基甲烷 <chem>c1ccc(C(c2ccccc2)c2ccccc2)cc1</chem>		乙酸 <chem>CC(=O)O</chem>
	氯化亚锡Cl[Sn]Cl			
薄荷酮的制备	薄荷醇	薄荷酮	铬酸	水O
超声在Knoevenagel 反应中的应用	<chem>C[C@@H]1CC[C@H]([C@@H] (C1)O)C(C)C</chem>	<chem>CC1CCC(C(=O)C1)C(C)C</chem>	<chem>O=[Cr](=O)(O)O</chem>	
	环戊酮 <chem>C1CCC(=O)C1</chem> 、 氰基乙酸乙酯 <chem>CCOC(=O)CC#N</chem>	2-氰基-2-环戊亚基乙酸乙酯 <chem>CCOC(=O)C(C#N)=C1CCCC1</chem>	乙酸 <chem>CC(=O)O</chem> 乙酸铵 <chem>CC(=O)O.N</chem>	
己二酸的制备	环己酮 <chem>C1CCC(=O)CC1</chem>	己二酸 <chem>CCC(=O)O)CC(=O)O</chem>	高锰酸钾 <chem>[O-][Mn](=O)(=O)[K+]</chem>	水O

合成的目标反应物命中率最高可达97.7%^[24,25]，正向预测结果中目标产物在所有化合物中排名第一的准确率最高可达85.6%^[26,27]，预测效果非常理想，但RXN模型的网页使用需要注册账号，响应速度也较慢，而ASKCOS模型的网页在输入SMILES号之后即可直观显示化合物结构，提交10秒之内即可计算出预测结果，因此本文选用ASKCOS模型作为正向预测和逆合成模型。

1.2.1 正向预测模型的选取及参数设置

本文选用的ASKCOS模型的正向反应预测模块(版本2024.01)包含五个部分：条件推荐、反应预测、杂质预测、区域选择性预测和芳香族C-H官能化的选择性预测，其中本文使用的是反应预测模块。

在参数设置中，正向预测模型均是无模板模型，包含wldn5、augmented Transformer (增强变换器)和Graph2SMILES。本文选用wldn5^[26]和augmented Transformer^[27]。wldn5使用的是威斯费勒-莱曼网络(Weisfeiler-Lehman Network, WLN)算法，它是一种无模板的图卷积神经网络算法，基于可能的化学键的变化进行预测，其正向预测模型训练集选取pistachio和uspto_50k。pistachio是一个反应数据集浏览器，提供从美国和欧洲专利局专利中提取的超过900万个化学反应的加载、查询和分析功能；USPTO是Daniel Lowe于2017年开源的美国专利商标局通过文本挖掘收集的数据集，包含了1976年至2016年9月发表在美国国家专利中发表的反应数据。augmented Transformer即增强变换器模型，是基于自然语言处理算法建立，更适合于端到端的预测，其配套的训练集为USPTO_STEREO和pistachio_23Q3。

进入ASKCOS网站^[28]后，用户可以在Reactants输入框中输入实验中投料化合物的SMILES号，若有多个反应物，需在反应物之间加入“.”符号。在Reagents输入框中输入反应发生所需的催化剂或者试剂，例如浓硫酸等，在Solvent输入框中输入反应体系中的溶剂(如有)。然后，在Setting中选择预测模型和训练集。以预测水杨酸和乙酸酐正向合成结果为例，在Reactants输入框中输入水杨酸和乙酸酐，在Reagents输入框中输入硫酸，如图1所示。最后，点击GET RESULTS即可获得预测结果，包含排名(Rank)、产物(Product)、可能性的值(Probability)、最大打分值(Max. Score)等。

1.2.2 逆合成预测模型的选取及参数设置

ASKCOS模型的逆合成预测模块(版本2024.01)包含三个部分：交互式反应设计、单步逆合成和反应树。交互式反应设计和反应树适用于多步合成反应，而本文所涉及的有机合成大部分是单步反应，因此本文使用单步逆合成模块进行预测。

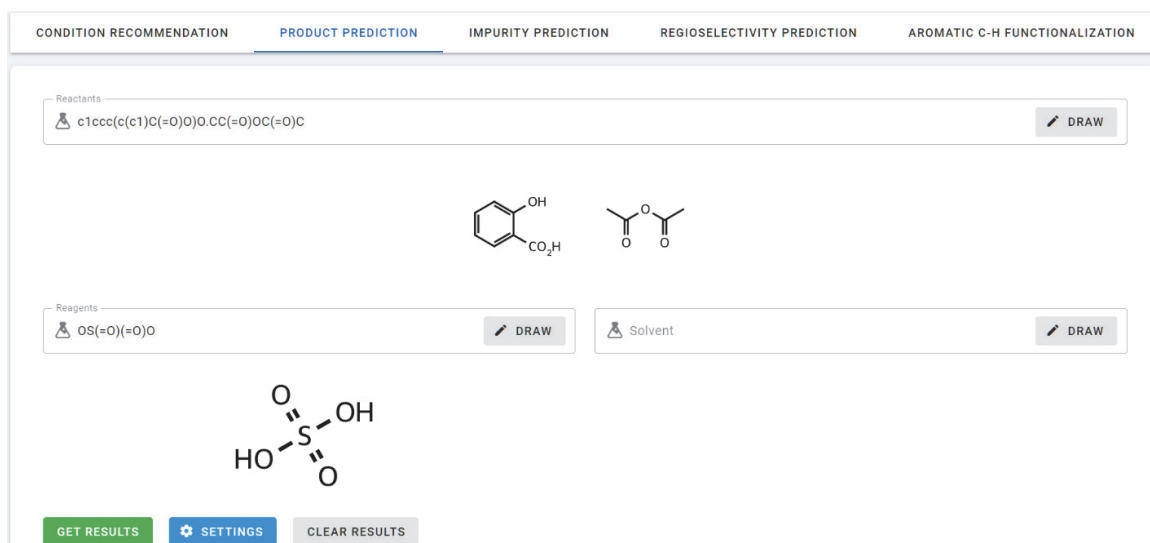


图1 正向预测模型的输入界面

进入ASKCOS网站^[29]后,在Target输入框中输入合成的目标化合物,在Model框中选择模型,包含template_relevance(模板相关)和augment_transformer(增强变换器)两种。对于本文中涉及的反应,使用模板相关模型的预测选用的训练集为pistachio和reaxys,系统中其他的三个训练集bkms_metabolic、pistachio_ringbreaker、reaxys_biocatalysis分别与代谢、开环、生物催化反应相关,与本文中的实验关联不大,因此不选用。reaxys训练集收集了来自105个专利局的新医药和化学专利数据。增强变换器模型包含USPTO_FULL和pistachio两个训练集。以预测乙酰水杨酸逆合成结果的输出界面为例,在Target输入框中输入乙酰水杨酸,如图2所示。最后点击SUBMIT之后得到预测结果,包括前体化合物(Precursor)、排名(Rank)、合理性的值(Plausibility)、SC分数(SCScore)等信息。对于模板相关的预测,系统还会提供模板排名(Template Rank)、模板打分(Template Score)以及相关的模板(Templates)信息。合理性打分范围在0到1之间,值越接近1表示反应发生的可能性越高;预测打分为负值,其绝对值越小代表反应可能性越大;模板打分为正值,分数越大表示模板的可靠性越高。每组前体化合物的合理性值是固定的,与选取的模板和训练集无关,而打分和排名则与选取的模板和训练集相关联。

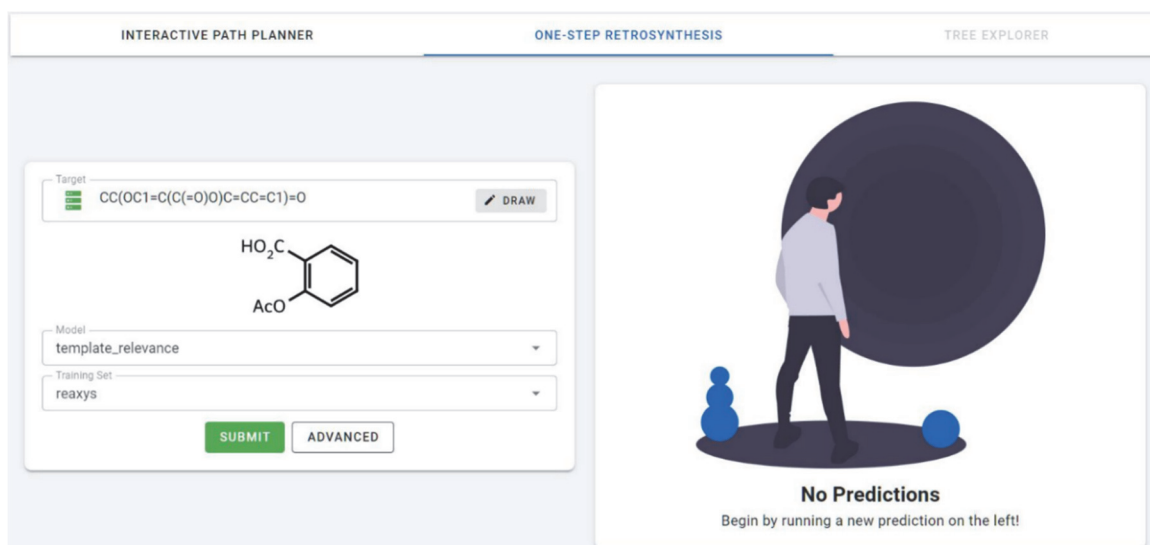


图2 逆合成预测模型的输入界面

2 模型预测结果

2.1 正向预测的结果

以预测水杨酸和乙酸酐正向合成结果为例, 选用了wldn5模型和pistachio训练集, 在图1界面中输入信息后点击GET RESULTS, 在网站中预测出的结果如图3所示。图3中的输出结果表格中, 列出了所有的预测产物(Product), 并按照可能性的值(Probability)从大到小进行排名(Rank), 并给出了预测产物的结构式图片、最大打分值(Max. Score)、分子量(Molecular Weight)等信息。表1中所有反应的正向预测的结果汇总如表2所示, 其中“NA”代表预测结果中不包含实验中的目标产物。

命中率及准确率统计数据如表3所示。其中, “命中率”表示目标产物在相应模型和训练集条件下出现在预测结果中的百分比, “Top-*n*%”代表目标物在预测结果中排名前*n*位的百分比。统计Top-*n*数据时, 将“卤代烷制备”中的三组反应物(正丁醇、溴化氢; 正丁醇、氯化氢; 正丁醇、溴化氢、氯化氢)视为三个不同的反应。对于存在多个目标物的反应(例如呋喃甲醛制备呋喃甲醇和呋喃甲酸), 按照排名最高的目标物进行统计。

根据预测结果显示, ASKCOS模型在大多数情况下都表现出良好的预测能力。在使用两种模型的情况下, 27个反应中的目标产物都被成功预测出来。其中, 采用wldn5模型结合pistachio训练集的组合预测出了24个反应, 命中率达到88.5%。这种组合的可能性值中, 有12个大于0.9, 17个大于0.8, 表明预测发生反应的可能性较高。此外, Top-*n*指标也优于其他预测条件, 在本文选择的预测效果中表现最佳。除了wldn5模型结合pistachio训练集的Top-1指标外, 其他的指标wldn5均优于增强变换器, 这是因为WLN模型是基于可能的键的变化来进行预测, 而本文中实验多为单步反应, 因此该模型应用于这些实验中的预测是比较合适的。然而, 对于一些较为复杂的反应, 例如邻苯二甲酸酐制备邻氨基苯甲酸的反应、镁和溴苯制备格氏试剂的反应, 以及三苯甲醇的制备反应等, 由于机理较为复杂, 需要多步反应并产生较多的中间产物, 模型中键的变化的数据库可能无法涵盖这些反应, 因此预测结果中未能出现产物。然而, 这些反应在增强变换器模型上得到了良好的预测结果, 这是因为该模型针对的是端到端的预测。因此, 在三苯甲醇的制备实验中, 增强变换器模型可以直接预测出最终的产物三苯甲醇并且排名第一, 但wldn5只能预测出中间产物二苯甲醇; 对于邻氨基苯甲酸制备实验的第二步, wldn5未能预测出最终产物, 而增强变换器模型预测出邻氨基苯甲酸排名第一。因此, 不同类型的反应所选择的模型和训练集对预测结果有很大影响。

当然, 有机化学反应的结果受到许多因素的影响, 包括反应时间、反应温度、装置的影响等, 预测结果只能提供一些可能性的排序, 不同的模型适用范围也有所差异, 实验中的产物和产率还需要进行实际的操作实验去验证。

Rank	Product	Probability	Max. Score	Molecular Weight	Predict impurities	Predict regio-selectivities
1		0.8822	-25.943	180.0	→	→
2		0.0403	-29.030	180.0	→	→
3		0.0357	-29.150	222.1	→	→
4		0.0252	-29.497	180.0	→	→
5		0.0108	-30.350	138.0	→	→
6		0.0033	-31.528	180.0	→	→

图3 预测水杨酸和乙酸酐正向合成结果的输出界面

表2 所选实验的正向合成预测结果汇总

反应物	目标物	反应试剂	溶剂	wldn5模型				增强变换器模型			
				Pistachio		uspto_50k		USPTO_STEREO		pistachio_23Q3	
				训练集	训练集	训练集	训练集	训练集	训练集		
排名	可能性	排名	可能性	排名	可能性	排名	可能性				
乙酸、正丁醇	乙酸正丁酯	硫酸		1	0.8773	1	0.95	1	0.8316	1	0.3009
水杨酸、乙酸酐	乙酰水杨酸	硫酸		1	0.8822	1	0.8805	1	0.999	1	0.0903
二苯甲酮	二苯甲醇	硼氢化钠	乙醇	1	0.9874	1	0.9998	1	0.9649	23	0.0145
正丁醇	正丁醚	硫酸		1	0.7612	1	0.7817	10	0.0129	5	0.0411
正丁醇、溴化氢	1-溴丁烷	硫酸		1	0.9955	1	0.5557	1	0.4144	16	0.0083
正丁醇、氯化氢	1-氯丁烷	硫酸		1	0.8728	2	0.2378	NA		NA	
正丁醇、溴化氢、 氯化氢	1-溴丁烷 1-氯丁烷	硫酸		1 2	0.9761 0.019	1 3	0.5378 0.0748	2 NA	0.0358 NA	NA	NA
环己醇	环己烯	磷酸		2	0.1542	2	0.3817	NA		NA	
环己醇	环己酮	铬酸	水	1	0.9969	1	1	1	0.9947	NA	NA
苯甲醛、乙酸酐	肉桂酸制备	碳酸钾		1	0.6229	1	0.6662	NA		NA	NA
2,3-二氢-1-茛酮、 3,4-二甲氧基苯甲醛	2-(3,4-二甲氧基苯 亚甲基)-2,3-二氢- 1-茛酮	氢氧化钠		1	0.9999	1	0.9764	NA		NA	NA
环己酮、苯肼	1,2,3,4-四氢吡唑	乙酸		52	0	NA		1	0.9779	1	0.5444
邻苯二甲酸酐、氨水	邻苯二甲酰胺			5	0.0174	NA		1	0.5914	4	0.0357
邻苯二甲酰胺、 氢氧化钠	邻氨基苯甲酸			NA		NA		NA		1	0.3664
氯化苄、三乙胺	苄基三乙基氯化铵			1	0.8609	1	0.7853	NA		NA	NA
环己烯、氯仿	7,7-二氯二环 [4.1.0]庚烷	氢氧化钠		1	0.9327	5	0.1016	1	0.8718	1	0.4297
呋喃甲醛	呋喃甲醇	氢氧化钠	水	1	0.412	3	0.1432	1	0.5479	NA	NA
	呋喃甲酸			5	0.073	4	0.0394	NA		4	0.0506
甲苯	苯甲酸	高锰酸钾	水	2	0.0544	1	0.997	1	0.9087	1	0.6205
苯甲酸、乙醇	苯甲酸乙酯	硫酸		1	0.9995	1	0.5929	1	0.9996	11	0.0143
镁、溴苯	苯基溴化镁		乙醚	NA		NA		NA		1	0.2007
苯基溴化镁、 苯甲酸乙酯	二苯甲酮 三苯甲醇		乙醚	1	0.9802	1	0.8887	4	0.0003	3	0.0589
二苯甲酮、 苯基溴化镁	三苯甲醇		乙醚	NA		NA		1	0.9947	1	0.8167
三苯甲醇、溴化氢	三苯甲基溴		乙醚	1	0.9979	1	0.9745	1	0.946	1	0.8664
三苯甲醇、氯化亚锡	三苯甲基烷		乙酸	1	0.9932	2	0.1352	1	0.9626	1	0.3621
薄荷醇	薄荷酮	铬酸	水	1	0.889	1	0.9999	1	0.2398	23	0.007
环戊酮、 氰基乙酸乙酯	2-氰基-2-环戊亚基 乙酸乙酯	乙酸、 乙酸铵		1	0.9988	1	1	1	0.6902	1	0.3698
环己酮	己二酸	高锰酸钾	水	1	0.9998	1	0.9914	1	0.4719	22	0.009
				NA		NA		1	0.9046	2	0.1436

表3 正向合成预测命中率及准确率统计

模型	训练集	命中率/%	Top-1/%	Top-2/%	Top-5/%
wldn5	pistachio	88.5	73.1	80.8	84.6
wldn5	uspto_50k	80.8	61.5	73.1	80.8
增强变换器	USPTO_STEREO	73.1	65.4	69.2	69.2
增强变换器	pistachio_23Q3	73.1	38.5	42.3	57.7

2.2 逆合成预测的结果

以预测乙酰水杨酸逆合成的结果为例，选模板相关模型和reaxys训练集进行预测，在图2界面中输入信息后点击SUBMIT，在网站中得到了如图4所示的结果。图4中的输出结果表格中，列出来了所有前体化合物(Precursor)的结构式图片，并按照预测打分(Score)的绝对值从小到大进行排名(Rank)，并给出了前体化合物的合理性(Plausibility)、SC分数(SCScore)等信息。若选用模板相关模型，输出结果中还会显示模板排名(Template Rank)、模板打分(Template Score)以及对应的模板(Templates)等信息。对于表1中的目标化合物的逆合成预测结果进行了汇总，如表4所示。表中，“NA”表示预测结果中未包含实验中的前体化合物。命中率和准确率的统计数据列于表5。若预测结果包含实验中的部分前体化合物(如1-溴丁烷只预测出正丁醇)、或中间产物(如邻氨基苯甲酸预测出邻苯二甲酰胺)、或化合物的其他存在形式(如苯甲酸根)，均被视为预测命中成功。对于预测出多种包含实验中的部分前体化合物的情况(如1-溴丁烷的预测结果中同时包含只有正丁醇和正丁醇、溴化氢两种情况)，使用排名最高的数据进行统计。

图4 预测乙酰水杨酸逆合成结果的输出界面

表4显示，使用两种模型进行预测的26个实验中，有25个成功预测了实验中所用的前体化合物。这些预测结果中的合理性打分平均值为0.89，其中有15个打分在0.95以上。这表明，单步逆合成模型评价本文中的实验前体化合物能够成功合成目标化合物的合理性很高，实验课中的合成方法是合成这些化合物常用的方法。增强变换器模型基于两种训练集的命中率都超过了80%，比模板相关的模型命中率更高，说明增强变换器模型更适用于本文中化合物的预测，而模板相关模型中的模板还存在一定的局限性。不过，将模板相关的模型在两个训练集上的结果综合起来，也成功预测了21个目标化合物，说明预测模型中的模板已经涵盖了大部分反应。在图4中的预测结果界面，学生可以通过

点击具体模板查看相应的反应中化学键的变化，将理论课所学机理和实验课中的实验原理相结合，有助于加深对反应过程的理解。如图4所示，点击红框中的模板，会显示其对应的模板信息(Template Info)、模板数据集(Template set)、总参考文献数量(Total references)以及反应方程式。

表4 所选实验的逆合成预测结果汇总

目标物	预测的前体	合理性	模板相关模型								增强变换器模型			
			reaxys训练集				pistachio训练集				USPTO_FULL训练集		pistachio_23Q3训练集	
			排名	预测 打分	模板 排名	模板 打分	排名	预测 打分	模板 排名	模板 打分	排名	预测 打分	排名	预测 打分
乙酸正丁酯	乙酸、正丁醇	0.97	8	-0.22	3	0.072	6	-1.4	7	0.011	NA	NA	NA	
乙酰水杨酸	水杨酸、乙酸酐	1	1	-0.003	1	0.41	2	-0.71	9	0.0017	1	-0.0023	2	-0.002
二苯甲醇	二苯甲酮	0.92	1	-0.0053	1	0.15	6	-0.31	13	0.0025	1	-0.0016	1	-0.0024
正丁醚	正丁醇	0.49	2	-0.029	3	0.069	NA	NA	NA	4	-3.5	1	-0.0035	
1-溴丁烷	正丁醇	0.94	5	-3.3	1	0.31	NA	NA	NA	3	-0.57	1	-0.0029	
	正丁醇、溴化氢	0.85	NA	NA	NA	4	-0.23	2	0.09	10	-28	NA	NA	
1-氯丁烷	正丁醇	0.95	7	-4	1	0.25	NA	NA	NA	14	-59000	1	-0.0023	
	正丁醇、氯化氢	0.89	22	-150	53	0.00053	NA	NA	NA	9	-110	NA	NA	
环己烯	环己醇	0.12	1	-0.0062	1	0.16	2	-1.9	10	0.00053	3	-0.15	1	-0.0037
环己酮	环己醇	0.98	2	-0.0082	1	0.12	NA	NA	NA	3	-0.032	1	-0.0023	
肉桂酸制备	苯甲醛、乙酸酐	0.73	10	-0.11	19	0.018	1	-0.0036	1	0.56	5	-0.035	6	-0.048
2-(3,4-二甲氧基苯亚甲基)-2,3-二氢-1-茛菪酮	2,3-二氢-1-茛菪酮、3,4-二甲氧基苯甲醛	1	1	-0.0005	1	0.74	1	-0.00074	1	0.74	1	-0.00057	1	-0.00059
1,2,3,4-四氢喹唑啉	环己酮、苯肼	1	1	-0.0052	1	0.38	4	-0.1	7	0.02	1	-0.0037	1	-0.0039
邻苯二甲酰胺	邻苯二甲酸酐	0.91	NA	NA	NA	6	-2.3	36	0.014	14	-1	NA	NA	
邻氨基苯甲酸	邻苯二甲酰胺	0.97	NA	NA	NA	3	-0.089	1	0.86	NA	6	-0.021	NA	
	邻苯二甲酸酐	0.99	NA	NA	NA	NA	NA	NA	NA	17	-3.2	NA	NA	
苄基三乙氯化铵	氯化苄、三乙胺	1	1	-710	900	0.0000042	NA	NA	NA	2	-12	2	-0.03	
7,7-二氯二环[4.1.0]庚烷	环己烯、氯仿	1	2	-0.18	2	0.28	NA	NA	NA	1	-0.12	2	-0.094	
呋喃甲醇	呋喃甲醛	0.89	1	-0.0027	1	0.38	1	-0.0021	1	0.47	2	-0.0097	1	-0.0018
呋喃甲酸	呋喃甲醛	0.75	15	-39	11	0.026	NA	NA	NA	NA	1	-0.0037	NA	
苯甲酸	甲苯	0.91	87	-32	2	0.032	NA	NA	NA	NA	20	-0.7	NA	
苯甲酸乙酯	苯甲酸、乙醇	0.98	15	-0.17	1	0.09	NA	NA	NA	2	-0.038	5	-0.041	
	苯甲酸根、乙醇	0.99	62	-140	315	0.00013	3	-32	7	0.00057	NA	NA	NA	
苯基溴化镁	镁、溴苯	1	NA	NA	NA	6	-10	334	0.00019	2	-0.088	1	-0.0028	
三苯甲醇	二苯甲酮、溴苯	0.99	NA	NA	NA	1	-0.078	3	0.023	1	-0.014	5	-0.045	
三苯甲基溴	三苯甲醇	0.89	NA	NA	NA	NA	NA	NA	NA	NA	1	-0.00059	NA	

(待续)

续表4

目标物	预测的前体	合理性	模板相关模型				增强变换器模型			
			reaxys训练集		pistachio训练集		USPTO_FULLL训练集		pistachio_23Q3训练集	
			排名	预测 得分	模板 排名	模板得分	排名	预测得分	模板 排名	模板得分
三苯基甲烷	三苯甲醇	0.9	NA	NA	NA	NA	1	-0.0012	1	-0.00073
薄荷酮	薄荷醇	1	NA	NA	NA	NA	2	-0.012	1	-0.0041
2-氰基-2-环戊	环戊酮、	1	NA	NA	NA	NA	1	-0.0024	1	-0.0022
亚基乙酸乙酯	氰基乙酸乙酯									
己二酸	环己酮	NA	NA	NA	NA	NA	NA	NA	NA	NA

表5 逆合成预测命中率及准确率统计

模型	训练集	命中率/%	Top-1/%	Top-5/%	Top-10/%
模板相关	reaxys	65.4	26.9	42.3	53.8
模板相关	pistachio	53.8	15.4	38.5	53.8
增强变换器	USPTO_STEREO	80.8	30.8	69.2	73.1
增强变换器	pistachio_23Q3	88.5	57.7	76.9	84.6

对于三苯甲醇的实验，模型只能预测到二苯甲醇和溴苯反应的步骤，无法预测出原始反应物格氏试剂和苯甲酸乙酯。这可能是因为模型更适合于单步反应的逆合成分析，对于反应机理较为复杂、具有较多中间产物的反应的预测能力比较有限。对于己二酸的合成实验，模型均未能成功预测，实验所采用的方法是由环己酮开环氧化得到己二酸，而其逆合成中预测的结果大部分为直链化合物。

从Top-*n*指标的数值来看，几种条件下的预测模型Top-1指标数值较低，这表明除本文中的目标物质外，可能还存在更合适的反应物，本文中选用的反应物是与本科教学实验的培养目标相匹配的，而在实际的科研或生产中，选择其他反应物可能会更好地合成目标物。Top-5和Top-10指标有明显提高，这说明预测模型对本文中所用的反应物的评价相对靠前，合理性较高，其中增强变换器加pistachio_23Q3训练集的组合在命中率和Top-*n*指标上达到最高。当然，反应物的选择还与原料的可得性、价格、毒性以及相对应的反应条件相关，在实际实验中需要进行综合考量。

对目标化合物的逆合成分析，可以使学生了解到除现有实验所用的反应物及实验方法外，还有哪些方法可以合成目标物，并分析不同途径的可行性和效率，有助于学生进行思维发散。在已有的教学中，逆合成的思想已广泛应用于探究和设计化合物的合成路线，例如对氯乙酰苯胺^[30]和肉桂酸^[31,32]，而在本文中使用人工智能模型则可以将逆合成的思想批量拓展到全部的合成实验中。

3 教学实施

使用ASKCOS模型进行实验的正向反应预测和逆合成分析预测取得了良好效果，因此可应用于有机化学实验的课程教学中。教师可在反应原理讲解部分介绍目前人工智能相关模型在有机化学反应预测中的应用，使学生了解相关的概念，激发学生的学习兴趣。进而可以在课后思考题中，增加使用模型进行预测的选做模块，以供感兴趣和有余力的学生利用课后时间使用网页版模型获取相应实验的预测结果，并且结合当次实验的情况进行分析讨论；也可以在学期末进行回顾总结的过程中，将本学期所有的合成实验使用模型进行预测，并对结果进行汇总和分析，提交一份报告作为作业，用于平时成绩评定或考核。

有机化学实验课教学重点是将学生所学的有机化学理论知识初步转化为实际应用，并培养学生未来科研工作所需的基本实验技能。通过人工智能赋能化学教学理念的实施，学生结合动手操作进行实验的全方位研究，能更加深刻地体会到人工智能模型在有机合成中对于节省人力物力和时间成本的重要意义。这种方法不仅能够拓宽传统教学的边界，还可以激发学生的创新思维，有利于实现全面发展应用型人才的培养目标。

4 结语

本文将人工智能模型融入有机化学实验教学中，学生可以利用ASKCOS的网页版程序，通过输入反应物、反应试剂和溶剂，利用正向预测模型来预测可能的生成物；也可通过输入合成的目标化合物，利用逆合成模型预测反应物，将其与实际实验过程进行有机结合。预测结果也显示该模型具有良好的预测效果，将其应用于课程教学中有助于更好地达到课程的教学目标，促进学科交叉，帮助学生加深对有机化学反应特点的理解，培养解决问题和创新思维的能力，为将来迈入科研阶段的工作打下坚实的基础。

尽管人工智能预测模型在有机化学实验教学中的应用具有许多优势，但在教学过程中也需要引导学生认识到这些模型的局限性。例如，学生需要了解到模型预测的不确定性，并认识到实际实验结果可能与模型预测结果不一致的情况，这样有助于培养学生在使用模型时的批判性思维。随着人工智能技术的进一步发展，未来可以寻找界面更友好、功能更丰富、预测准确率更高的模型，并拓展其在有机化学实验课程教学的应用领域，如：(1) 对拟合成的目标化合物，可通过模型预测溶解度、毒性、生物活性等性质评估化合物的应用前景；(2) 对于特定的目标物和产物，可使用模型预测反应体系中的催化剂、溶剂、温度等最佳反应条件；(3) 给定反应物和反应条件，可进行反应活性与某些化合物的对映选择性和官能化选择性的预测等。

参 考 文 献

- [1] 方瑞琴, 谢光波, 冯娟, 邓科君, 杨鹏. 实验科学与技术, **2015**, *13* (5), 83.
- [2] 高惠惠. 广州化工, **2020**, *48* (13), 171.
- [3] 王静, 闫岩, 刘明凯. 化工管理, **2024**, No. 11, 38.
- [4] Corey, E. J.; Wipke, W. T. *Science* **1969**, *166*, 178.
- [5] de Almeida, A. F.; Moreira, R.; Rodrigues, T. *Nat. Rev. Chem.* **2019**, *3*, 589.
- [6] Empel, C.; Koenigs, R. M. *Angew. Chem. Int. Ed.* **2019**, *58*, 17114.
- [7] Gasteiger, J. *ChemPhysChem* **2020**, *21*, 2233.
- [8] Sridharan, B.; Goel, M.; Priyakumar, U. D. *Chem. Commun.* **2022**, *58*, 5316.
- [9] Tu, Z.; Stuyver, T.; Coley, C. W. *Chem. Sci.* **2023**, *14*, 226.
- [10] 刘伊迪, 杨骐, 李遥, 张龙, 罗三中. 有机化学, **2020**, *40* (11), 3812.
- [11] Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. *Angew. Chem. Int. Ed.* **2016**, *55*, 5904.
- [12] 孔祥泰, 张润泽, 张玮, 李叙潼, 郑明月. 中国现代应用药学, **2022**, *39* (21), 2856.
- [13] 梁立, 唐凌凯, 李思瑶, 胡建平. 知识窗(教师版), **2024**, No. 1, 42.
- [14] 崔婷婷, 黄锦涛, 闵永刚. 武汉轻工大学学报, **2024**, *43* (1), 115.
- [15] 李田, 张攀, 边磊, 关玲, 徐炬峰, 张奇涵, 王婕好, 王颖霞. 大学化学, **2021**, *36* (1), 2012048.
- [16] 翟红林, 张晓昀, 曹晶晶. 大学化学, **2024**, *39* (1), 63.
- [17] 邱晓航, 李一峻, 韩杰, 尚贞锋. 基础化学实验. 第2版. 北京: 科学出版社, 2017: 187-270.
- [18] Latendresse, M.; Malerich, J. P.; Herson, J.; Krummenacker, M.; Szeto, J.; Vu, V.-A.; Collins, N.; Madrid, P. B. *J. Chem. Inf. Model.* **2023**, *63*, 5484.

- [19] RXN for Chemistry. [2024-09-12]. <https://rxn.res.ibm.com/>
- [20] Home - ASKCOS. [2024-09-12]. <https://askcos.mit.edu/>
- [21] Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. *Chem. Sci.* **2020**, *11*, 3316.
- [22] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. *Chem. Sci.* **2018**, *9*, 6091.
- [23] Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. *ACS Cent. Sci.* **2019**, *5*, 1572.
- [24] Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. *J. Chem. Inf. Model.* **2020**, *60*, 3398.
- [25] Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. *Chem. Sci.* **2021**, *12*, 1469.
- [26] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. *Chem. Sci.* **2019**, *10*, 370.
- [27] Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G. *Nat. Commun.* **2020**, *11*, 5575.
- [28] Forward Synthesis - ASKCOS. [2024-09-12]. <https://askcos.mit.edu/forward?tab=forward>
- [29] Retrosynthesis - ASKCOS. [2024-09-12]. <https://askcos.mit.edu/network?tab=RP>
- [30] 杜欣, 姜文凤. *实验室科学*, **2022**, *25* (5), 41.
- [31] 郑媛, 兰泉, 查正根. *大学化学*, **2019**, *34* (6), 53.
- [32] 郑媛, 兰泉, 查正根, 李玲玲, 蒋俊, 朱平平. *大学化学*, **2024**, *39* (6), 207.