综述评介

实验规程的过程级语义表示研究综述*

付　芸[1,2]　刘细文[1,2]　朱丽雅[1]　韩　涛[1,2]

[1]（中国科学院文献情报中心 北京 100190）

[2]（中国科学院大学经济与管理学院信息资源管理系 北京 100190）

摘要：**【目的】**揭示实验规程过程级语义表示研究进展，发现尚需解决的关键研究问题，探究发展趋势。**【文献范围】**使用相关主题词在 Web of Science、arXiv、Engineering Village、中国知网、万方、维普中检索筛选出 76 篇文献，并参考知名实验规程专业期刊的提交要求和评审原则文档。**【方法】**在界定实验规程及其过程级语义表示相关概念基础上，从过程级语义表示方法、表示要素抽取方法以及相关表示数据应用三方面进行分析评述。**【结果】**实验规程的过程级语义表示研究整体处于发展初期，表示方法中表示框架尚未统一、表示要素各异，从以自然语言编写为主的实验规程中自动抽取过程级语义表示要素难度较大、效果一般，过程级语义表示的实验规程数据已在部分方向开展应用研究，整体可提升空间较大。**【局限】**未详细阐述面向表示要素自动抽取技术细节及数据应用方法过程。**【结论】**未来应融合各类表示方法的优势以探索构建包含较完整要素的统一表示框架，探索基于先进智能技术的表示要素自动抽取方法研究，探索使用过程级语义表示的实验数据开展广泛应用研究。

关键词：实验规程　过程级语义表示　表示方法　表示要素抽取方法　数据应用

分类号：G35　N19

DOI：10.11925/infotech.2096-3467.2023.0335

引用本文：付芸，刘细文，朱丽雅等 . 实验规程的过程级语义表示研究综述[J]. 数据分析与知识发现，2023，7(8)：1-16.(Fu Yun,Liu Xiwen, Zhu Liya, Han Tao. Review of Semantic Representation of Experimental Protocols at Process-Level[J]. Data Analysis and Knowledge Discovery, 2023, 7(8): 1-16.)

1　引　言

智能科研要求实验规程由语句级向过程级语义表示转变。以 AlphaFold[1] 为代表的“数据驱动知识发现”方法[2]，受所使用数据完整性和实质内容的限制[3]，以材料领域为例，包含材料性质、结构的数据已较为丰富，限制新材料开发速度的步骤已变成合成路径的发现和验证[4]。数据驱动的材料合成路径发现研究，需将语句级实验规程向可计算的过程级语义表示转变[5]。

智能实验平台发展要求实验规程由文本语句向标准操作指令转变。科学实验存在较高的复现危机[6]，以自驱动实验室、实验机器人为代表的智能实验平台是当前解决实验数据质量、多样性及实验复现、效率、安全等问题极具潜力的解决方案[7]。由于

通讯作者（Corresponding author）：刘细文（Liu Xiwen），ORCID：0000-0003-0820-3622，E-mail：liuxw@mail.las.ac.cn。

*本文系国家自然科学基金重点项目（项目编号：72234005）和国家社会科学基金项目（项目编号：22BTQ019）的研究成果之一。

The work is supported by the National Natural Science Foundation of China (Grant No. 72234005), the National Social Science Fund of China (Grant No.22BTQ019).

缺乏规范的数字实验规程表示和交换协议[8]，缺乏标准实验操作指令或通过自然语言处理（Natural Language Processing，NLP）直接访问文献中实验工作流程的能力等[9]，正在限制智能实验平台发展。因此需将文本语句描述的实验规程向标准操作指令转变。

然而，当前尚未发现实验规程的过程级语义表示或标准操作指令的明确定义以及相关研究综述，本文则致力于补足这一研究空白。为最大化获取相关文献，本文使用核心主题词scientific/experiment/experimental protocol/procedure、wet lab protocol，在Web of Science、arXiv、Engineering Village中不限年份检索，经人工判读后筛选出76篇英文文献。基于这些文献内容，提炼总结中文核心主题词“实验规程”“实验程序”或“实验协议”，及限定主题词“数据化”或“数字化”和“表示”“表征”“组织”或“抽取”，在中国知网、万方、维普中不限年份检索，未发现相关文献。同时，参考*Current Protocol*、*Nature Protocol*、*Bio-protocol*等实验规程专业期刊中论文撰写和评审要求文档。

通过梳理76篇文献发现，当前研究主要分为实验规程的过程级语义表示方法、表示要素抽取方法以及该类数据应用三方面。本文在界定实验规程的过程级语义表示相关概念基础上，围绕上述三方面展开综述，并提出对应的关键研究问题及发展趋势。

2　实验规程的过程级语义表示概念界定

当前尚未有研究明确定义什么是实验规程的过程级语义表示。本文在归纳分析前人关于实验规程概念和特征的研究基础上，重新定义实验规程；在总结前人围绕实验规程的过程级语义表示框架和要素研究基础上，明确定义实验规程的过程级语义表示。

2.1　实验规程的内涵解析

实验规程源于科学实验，对应科学实验三个层级[10]（即设计层、模型层（方法）和物理层（结果））中的设计层，是实验设计的书面计划[11]。实验规程具体是指对数据和真实世界对象操作的描述，目的是收集和处理实验数据，和/或构建新对象，其本质上是科学研究活动的显式离散化和组织，通常包含两类任务：结构化实验任务和操作动作（以及附属参数）[3]。在生物领域，实验进一步分为以计算机处理数据为主的干实验和在实验室环境下操作的湿实验，并将湿实验规程（Wet Lab Protocols，WLP）定义为一组用领域专业自然语言编写的用于分布执行生物实验过程的指令集[12]。

根据专业实验规程期刊的刊文要求，可将实验规程文献分为三类：研究型实验规程、丰富型实验规程、精选型实验规程。研究型实验规程对应PLOS ONE中出版的Study Protocol①，强调实验规程的前期研究设计。丰富型实验规程对应*Nature Protocol*中出版的Protocol②以及PLOS ONE中出版的Lab Protocol③，强调实验规程应来自经同行评议后已出版的研究文献。精选型实验规程对应*Current Protocol*中出版的Protocol Article④以及*Bio-protocol*中出版的Protocol⑤，强调由期刊编委精选后邀请已出版研究文献作者，丰富和完善实验规程。

三类实验规程文献均强调实验过程的完整性和复现性特征，即实验规程应包含实验目标、材料、溶剂、配方、设备、数据和软件等要素，且内容编写应：按时间顺序，详细描述实验分步执行说明；按不同的子实验分类、分层描述；通常使用主动语态及现在时，即每步一般以动词开头；除操作动词，每步还应包含详细的实验材料、溶剂、温度、压强等条件；用词一致、词义明确，避免模糊词，如“一些”“大约”等。

基于上述总结可知，前人关于实验规程的定义无法覆盖当前对实验规程内容的新要求和特征。因此，本文将重新定义实验规程：对数据和真实世界的实验活动、行动、操作动作、操作对象、操作顺序、操作条件等的描述，目的是确保实验过程完整且可复现，规范执行实验，收集和处理实验数据，和/或构建

①https://journals.plos.org/plosone/s/submission-guidelines#loc-study-protocols.

②https://www.nature.com/nprot/content-types.

③https://journals.plos.org/plosone/s/submission-guidelines#loc-study-protocols.

④https://currentprotocols.onlinelibrary.wiley.com/hub/authorguidelines.

⑤https://en.bio-protocol.org/en/authors.

新对象。

2.2　实验规程的过程级语义表示概念界定

尽管实验规程专业文献中提出诸多编写要求，但这些使用自然语言编写的文本通常带有术语或口语，很难被计算机或智能实验平台理解执行[13]。围绕将自然语言描述的实验规程转换为机器可读的语义格式，科学家们已开展众多探索研究，并将这种形式的实验规程称为实验动作显性语义[13]、实验动作序列[14]、过程级语义表示[15]、实验操作指令[16]。尽管名称不同，但其本质基本相同：捕获实验过程及其参数之间关系，以确保实验过程完整、可复现及支持机器计算执行。

在兼顾名称的通用性和描述力的基础上，本文将这种机器可读的语义格式统称为实验规程的过程级语义表示。由于当前相关研究成果中仅描述机器可读的实验规程语义表示方法和要素，尚未给出明确定义，本文在理解相关论述的基础上，将实验规程的过程级语义表示定义为：通过定义捕获实验规程复现的全部语义特征，包括实验操作动作、操作对象、操作顺序、操作条件等，在自然语言格式的文献语句与机器可读的低级实验指令之间构建一套语义表示框架。

下面将具体阐述实验规程的过程级语义表示方法、表示要素获取及数据应用等研究进展。

3　实验规程的过程级语义表示方法研究

实验规程的过程级语义表示方法主要有三类：本体、数据模型和结构化图。研究表明，在科学数据的互操作性和知识推理方面，基于本体的方法具有更多优势[17-18]；基于数据模型的方法对于特定用户需求场景更具适应性[19]；结构化图表示方法则是引入图论的思想，将实验规程表示为连贯的动作事件[5]，较为直观。接下来将具体分析各类表示方法的特点和适用情景。

3.1　本　体

当前围绕实验规程的过程级语义表示本体可以分为4类，其典型代表分别为强调完整实验内容的本体EXPO[10]、强调实验规程内容的本体EXACT2[13]、强调实验过程执行计划的本体P-PLAN①、强调实验过程安全的本体OntoSafe[20]。值得一提的是，当前使用最广泛的本体语言是Ontology Web Language（OWL）[21]，最常用的本体编辑工具是Protégé②，通过它可以直观浏览、设置本体语言中定义的类、关系和逻辑公理。考虑到本文关注的实验规程相关本体的差异主要体现在类及关系结构的设计上，因此下文将着重呈现各本体中重要的类及关系结构设计。

（1）强调完整实验内容的本体EXPO

早期实验本体主要面向特定研究方向，如微阵列实验本体（MGED Ontology，MO）③、蛋白质组学实验本体（The Controlled Vocabularies（CV’s）of the Proteomic Standard Initiative（PSI），PSI CV）④等。首个独立于领域的科学实验通用本体EXPO是由Soldatova等[10]于2006年提出的，旨在为有效分析、注释和共享结果提供实验的正式描述，可用于描述计算和物理实验、原子动作和复杂动作、实验装置等内容。如表1所示，EXPO⑤中包含过多非实验规程内容。本文在此仅选取其中与实验规程相关的类及其关系结构，如图1所示。即使EXPO对实验的描述较为丰富，但对实验动作和过程安全等内容的描述力不足。

在EXPO的基础上，融合领域需求研制出众多特色实验本体，主要包括：在摩擦学领域的实验本体tribAIn[22]，基于已发布的tribAIn.owl⑥可知，主要借鉴EXPO中Proposition、Relation、ContentBearingPhysical类别下的部分子类以及hasPart属性下的部分子属性，其他的类及属性则主要依据摩擦学领域特征设置，已用实例证实tribAIn覆盖摩擦学实验的能力；

①http://vocab.linkeddata.es/p-plan/index.html.

②https://webprotege.stanford.edu/#projects/list.

③https://mged.sourceforge.net/ontologies/MGEDontology.php.

④https://www.psidev.info/groups/controlled-vocabularies.

⑤https://sourceforge.net/projects/expo/files/latest/download.

⑥https://github.com/snow0815/tribAIn/blob/master/tribAIn.owl.

表1　4个本体内容概要

Table 1　The Main Content Contained in the Four Ontologies

本体	类	对象属性	标注属性	数据属性	公理	逻辑公理	声明公理	注释断言
EXPO	325个，包括Entity类、Abstract和Physical两个子类，整体类别层级已划分至9层；每个类都明确定义其标签和表示含义、包含的下一子类、对应的公理、不相交类等信息	78个，包括hasAttribute和hasPart两个属性及子属性；每个属性都明确定义其用法、上级属性、对应的类等	2个	0个	2 067条	1 019条	402条	646条
EXACT2	162个，包括experimental action、experimental procedure、alert message等；每个类都明确定义其标签和表示含义、包含的下一子类、对应的公理、不相交类等信息	16个，包括hasAttribute和hasPart两个属性及子属性；每个属性都明确定义其用法、上级属性、对应的类等	7个	1个	858条	371条	168条	319条
P-PLAN	12个，其中Plan和MultiStep分属多个类；每个类都明确定义其标签和表示含义、包含的下一子类、对应的公理等信息	15个，包括correspondsToStep、hasInputVar等；每个属性都明确定义其标签和表示含义、表示语言转化形式、对应的类等	13个	0个	142条	47条	61	319条
OntoSafe	513个，包括Sensor、controller、Hazard等，每个类都明确定义其标签和表示含义、包含的下一子类、对应的公理等信息	80个，包括define、hasElment、isCausedBy等；每个属性都明确定义其用法、表示语言转化形式等	5个	70个	2 496条	1 172条	720	604条

图1　EXPO中与实验规程相关类及关系结构

Fig.1　Experitenmental Protocol Related Classes and Its Relation in EXPO

在生物医药领域，King等基于EXPO开发了系列面向实验机器人使用的定制实验本体——初期面向机器人科学家Adam执行微生物实验的定制版本体LABORS[23]，以及面向机器人科学家Eve执行药物

筛选实验的定制版本体DDI[24]；在计算机领域，机器学习实验本体Exposé[25]旨在促进算法交流，其中关于实验背景的类主要来自EXPO；在材料领域，本体MatOnto[26]旨在表示材料及其结构和特性、材料构成及工程处理步骤等结构化知识，其中处理步骤相关类来源于EXPO。

（2）生物医药实验规程语义本体EXACT2

生物医药实验的通用本体EXACT2[13]是在生物实验的通用本体EXACT[27]的基础上扩建的，最重要的变化是为每个实验动作添加了描述符，如温度、动作持续时间等，确保实验规程可安全复现。EXACT2旨在明确定义实验规程的语义信息，以确保其再现性，支持计算机应用程序读取计算，帮助生物学家准备、维护、提交和共享实验规程。EXACT2中包含的内容要点如表1所示，使用Protégé软件打开后呈现的全部类及其关系结构如图2所示。

EXPO与EXACT2的区别是：EXPO注重科学实验更全面内容的语义表示，实验规程只是其中的一部分；EXACT2更聚焦实验规程的语义表示，最为突出的是规定系列实验动词类和属性描述符、预警信息类和属性信息等，保证实验动作的显式语义可被机器读取计算，实验规程可安全复现，因此更符合当前对实验规程过程级语义表示的需求。

图2　EXACT2类及关系结构①

Fig.2　Classes and Its Relation in EXACT2

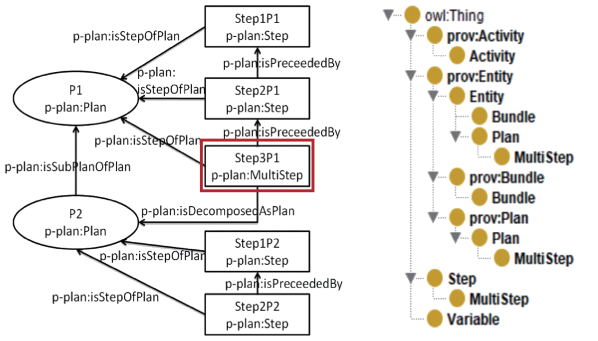
（3）强调实验过程执行计划的本体

实验过程执行的计划本体P-PLAN②，其核心思想是：实验过程执行工作流是一个“计划”，它定义了角色（Entity）、任务（Activity）、要执行的任务（Step）三层结构，即科学工作流通过实施科学方法，实现科学知识收集和组织的总体目标。通过分离工作流步骤，可实现不同工作流对同一任务指令的重用。P-PLAN中包含的内容要点如表1所示，一个P-PLAN示例如图3所示，其中图3（a）为P-PLAN中子计划表示，图3（b）为使用Protégé软件打开后呈现的类及关系结构。P-PLAN已应用于面向FAIR原则的科学工作流数据建设[28]，优化工作流任

①https://bioportal.bioontology.org/ontologies/EXACT.

②http://vocab.linkeddata.es/p-plan/index.html.

务，提高数据复用率。



(a) P-PLAN中子计划表示[①]　　(b) Protégé中的P-PLAN[②]

图 3　P-PLAN 示例

Fig.3　Example of P-PLAN

（4）强调实验过程安全的本体

实验过程安全本体 OntoSafe[20]涵盖化学过程安全基本概念，化学工艺安全系统，工业卫生，安全标准、法规和组织，数学（排放和扩散）模型等与过程安全相关的所有方面，为过程安全社区提供对相关概念及其关系的共同理解，以提高实验过程安全设计、分析和管理的效率，使用和重用社区信息。OntoSafe 中包含的内容要点如表 1 所示，使用 Protégé 软件打开后呈现的类及关系结构如图 4 所示，图中着重展示安全系统、火及暴露物、毒理等特色内容。OntoSafe 已应用于面向实验机器人平台使用的过程控制与模拟框架建设[29]，明确实验过程中的温度边界、时间边界等。

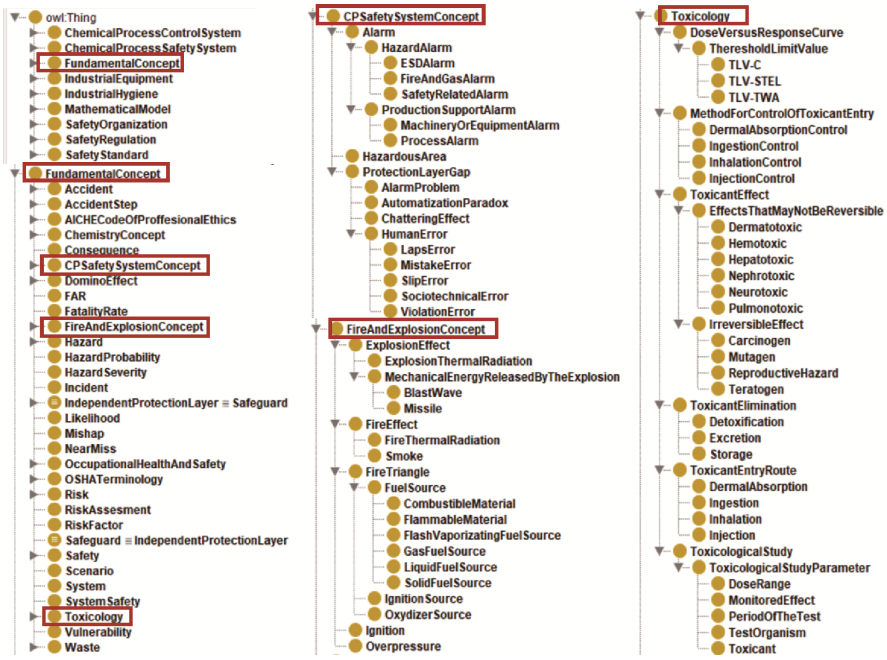


图 4　OntoSafe 示例[③]

Fig.4　Example of OntoSafe

3.2　数据模型

实验规程的过程级语义表示数据模型可以分为三类，典型代表分别是合成实验规程的过程级语义表示数据模型（包括有机[14]、无机[30-33]）、支撑智能实验平台的化学配方文件模型（Chemical Recipe Files, CRF）[34]和化学描述语言模型（Chemical Descriptive Language，χDL）[9]，它们各自的特征及适用场景如下。

①http://vocab.linkeddata.es/p-plan/index.html.

②https://www.opmw.org/p-plan.owl.

③https://webprotege.stanford.edu/#projects/07323336-09a8-4baf-b33c-eff91cb9b249/edit/Classes?selection=Class(owl:Thing).

（1）合成实验规程的过程级语义表示数据模型

有机、无机合成实验规程的过程级语义表示数据模型及其对应的实验动作和表示参数设置如表2所示。其中，有机合成与无机合成的数据模型中实验动作及表示参数差异较大，无机合成的多个数据模型均由Ceder教授团队完成，相应实验动作及表示参数设置差异不大。因此，数据模型表示方法中的关键问题是：选择合适的动作词，并为其预定义合适的属性（表示参数），以完整、高效地表示实验规程，且保证其可复现。

表2　合成实验规程的过程级语义表示数据模型

Table 2　Process-Level Semantic Representation Data Model for Synthesis Experimental Protocols

数据模型名称		实验动作及表示参数
有机合成实验动作序列模型[14]		定义28类实验动作，且为每个动作预定义相关属性： InvalidAction（error）、Add（material，dropwise，temperature，atmosphere，duration）、CollectLayer（layer）、Concentrate（none）、Degas（gas，duration）、DrySolid（duration，temperature，atmosphere）、DrySolution（material）、Extract（solvent，repetitions）、Filter（phase_to_keep）、FollowOtherProcedure（none）、MakeSolution（materials）、Microwave（duration，temperature）、OtherLanguage（none）、Partition（material_1，material_2）、PH（material，ph，dropwise，temperature）、PhaseSeparation（none）、Purify（none）、Quench（material，dropwise，temperature）、Recrystallize（solvent）、Reflux（duration，dean_stark）、SetTemperature（temperature）、Sonicate（duration，temperature）、Stir（duration，temperature）、Triturate（solvent）、Wait（duration，temperature）、Wash（material，repetitions）、Yield（material）、NoAction（none）
无机合成实验规程的过程级语义表示数据模型	固态合成[31]	定义6类实验动作：HeatingOperation、ShapingOperation、DryingOperation 、LiquidGrinding、QuenchingOperation、SolutionMixing 预定义8类通用属性：token、type、conditions、heating_temperature（max_value，min_value，values，units）、heating_time（max_value，min_value，values，units）、heating_atmosphere、mixing_device、mixing_media
	基于液体的无机合成[32]	定义6类实验动作：MixingOperation、PurificationOperation、HeatingOperation、DryingOperation、CoolingOperation、ShapingOperation 预定义8类通用属性：token、type、conditions、temperature（max_value，min_value，values，units）、time（max_value，min_value，values，units）、atmosphere、mixing_device、mixing_media
	金纳米粒子[33]	定义5类实验动作：MIXING、STARTSYNTHESIS、HEATING、COOLING、DRYING 预定义14类通用属性：congtain_recipe、conditions、temperature（value，unit，max_value，min_value）、time（value，unit，max_value，min_value）、string、subject、type、env_toks、op_token、op_type、ref_op、subject、temp_values（max，min，tok_ids，units，values）、time_values（max，min，tok_ids，units，values）
	无机合成动作统一语言模型ULSA[30]	定义8类动作：Starting、Mixing、Purification、Heating、Cooling、Shaping、Reaction、Non-Altering

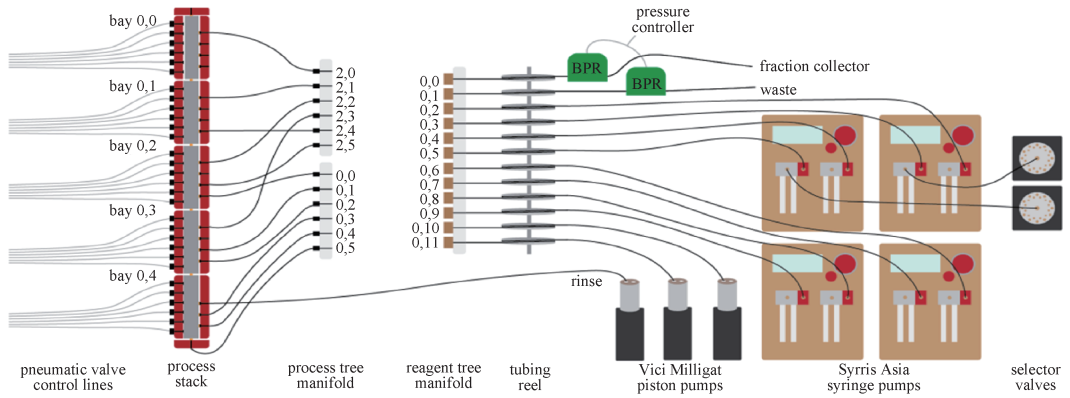
（2）CRF模型

麻省理工学院Coley博士及其团队2019年在*Science*期刊上发布实验机器人平台[34]，使用基于CSV结构的CRF模型，为15个小分子制备实验定制机器可读、可执行的实验规程。为了便于理解CRF模型如何发挥作用，本文以阿司匹林合成为例进行阐述，如图5所示。其中，图5（a）为实验室的统一实验装置；图5（b）为合成阿司匹林的化学方程式；图5（c）为抽象的合成方案，包括反应物、反应条件、实验步骤以及实验结果；图5（d）是为完成阿司匹林实验特别设置的实验装置；图5（e）为基于上述实验方案、实验装置，使用CSV格式描述的阿司匹林合成实验规程部分内容。

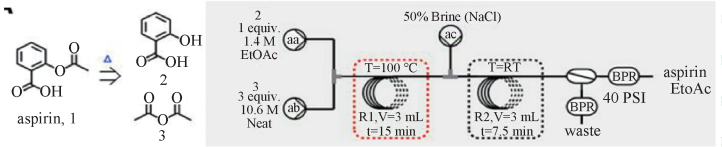
CRF是一种与设备密切关联的表示方法。观察实验规程内容可知，第一列为编号，第二列为自定义的实验动作，第三列为反应器，对应图5（d）中的6、2、3、4号装置，第四列的内容则依据不同的实验动作有所不同，前三列内容充分表明CRF是与实验设备相关的实验规程过程级语义表示方法。

CRF模型的优点：使用CSV结构表示，决定CRF模型简单易操作。缺点：无法表示复杂的操作，每个实验动作后至少连接一个实验设备，后续可连

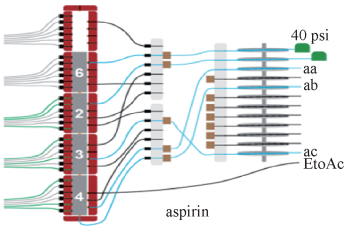
接的动作相关参数有限；灵活性差，实验方案与设备固定，若更换设备，则需重新编制整个实验规程。



(a) 实验装置



(b) 合成方程式　　(c) 合成方案



(d) 合成实验装置设置

```
# Define reactor positions
1  REACTORSETUP  6  0 1
2  REACTORSETUP  2  0 2
3  REACTORSETUP  3  0 3
4  REACTORSETUP  4  0 4
# Define pump connections
5  PUMPSETUP  6  4  0 0
6  PUMPSETUP  2  4  0 1
7  PUMPSETUP  4  1  0 2
8  PUMPSETUP  4  4  0 4
9  PUMPSETUP  3  2  0 11
```

```
# Make connections
10  PUMPCONNECT  6  4
11  PUMPCONNECT  2  4
12  PUMPCONNECT  4  1
13  PUMPCONNECT  4  4
14  PUMPCONNECT  3  2
# Move reactors into place
15  REACTORMOVE  6  TOBAY
16  REACTORMOVE  2  TOBAY
17  REACTORMOVE  3  TOBAY
18  REACTORMOVE  4  TOBAY
```

```
# Open on-board valves
26  PORTOPEN  2  inlet
27  PORTOPEN  2  outlet
28  PORTOPEN  3  inlet
29  PORTOPEN  3  outlet
30  PORTOPEN  4  inlet
31  PORTOPEN  4  outlet
32  PORTOPEN  2  4
33  PORTOPEN  4  1
34  PORTOPEN  3  2
```

```
# Disconnect lines
140  PUMPDISCONNEC  3  2
141  PUMPDISCONNEC  4  4
142  PUMPDISCONNEC  4  1
143  PUMPDISCONNEC  2  4
144  PUMPDISCONNEC  6  4
```

(e) 基于CSV结构的阿司匹林合成实验规程表示示例

图 5　CRF 模型表示的阿司匹林合成示意图[34]

Fig.5　Schematic Diagram of Aspirin Synthesis Represented by CRF Model

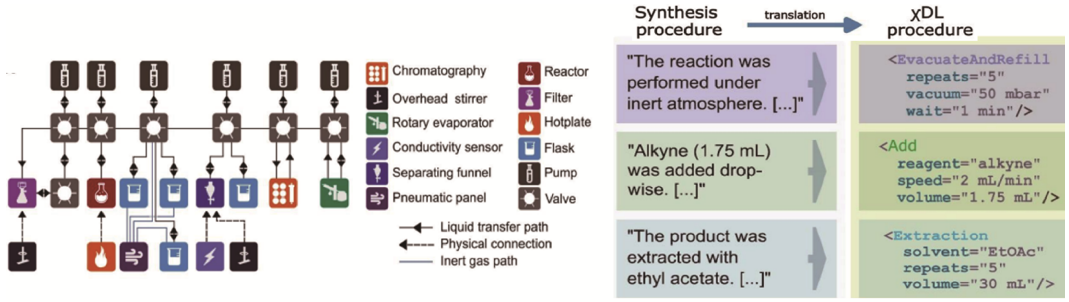
（3）χDL 模型

英国格拉斯哥大学 Cronin 教授团队等[9, 16, 35]在探索化学实验数字化研究中，参照计算机中“程序”的设计理念，提出一套适用于自动化实验平台的通用标准架构（ChemPU），并用 XML 格式编写的高级化学语言 χDL 进行控制，具体示例如图 6 所示。其中，图 6（a）为实验室的统一实验装置；图 6（b）为自然语言格式的实验规程翻译为使用 χDL 语言编码的格式；图 6（c）为抽象的自动实验工作流。从图 6（c）可知，专门设计的 States 中记录了实验应该在哪个设备、如何操作及后续状态，保证 χDL 模型仅关注实验动作表示，无须考虑动作与哪些设备关联。因此，不同设备可重复调用同一动作表示，极大提高编码效率。

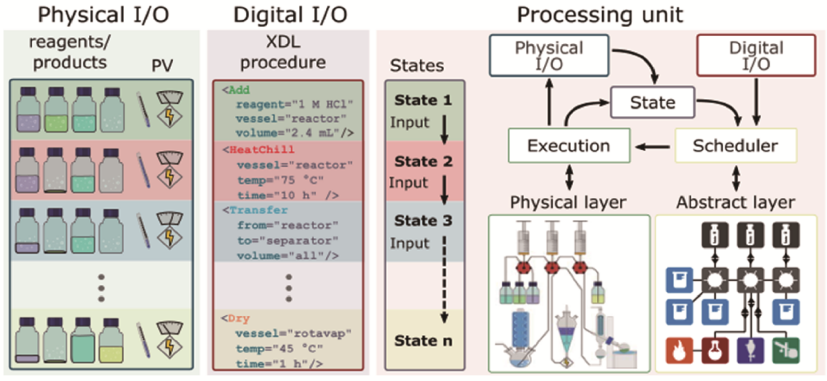
χDL 模型的编码对象是实验动作及其表示参数，与设备无关，一个独立的实验动作及其表示参数也被称为操作单元，χDL 模型可看作不同操作单元的组合。考虑到化学合成的数字化要求以模块化的方式准确捕捉所有相关参数和过程[36]，要实现这一数字化转型，需要考虑不同粒度的化学合成，并尽可能精确和明确[37]。Cronin 教授进一步提出基于操作单元的实验规程模块化表示理念[38-41]，数字化实验信息由一系列具有专用功能的模块组成，对每个模块系统单独考虑进行合成所需的操作单元及其参数。一旦建立这样的功能模块，今后在面对不同反应及场景时，只根据需求调用不同模块组合即可，而无须重新设计每个细节。

χDL 模型的优点是：与设备无关，可重用性和灵活性较高，尤其是模块化表示方法，更进一步提高实

验规程的表示效率;基于XML结构,可表示复杂的实验操作,扩展性较好。缺点是:当前尚未明确表明需要设计多少实验动作及其表示参数,才能表示出所有实验规程。



实验装置　　χDL表示的实验规程



实验执行工作流

图 6　χDL 模型示例[41]

Fig.6　Examples of χDL

3.3　结构化图

结构化图表示方法利用有向无环图(Directed Acyclic Graph, DAG)的思想,节点表示操作、材料、属性等,边表示节点间的关系,每条边都从一个节点指向另一个节点,连边不会形成闭合环。当前主要研究结果包括动作图(Action Graphs)[5]、合成图(Synthesis Graph)[42]、流程图(Flowchart)[43]、过程执行图(Process Execution Graph, PEG)[15],四者基本理念一致,区别在于节点和边标签的定义,具体内容如表3所示。尽管图表示方式较为直观,但是由于操作、材料、属性等均用节点表示,不易于区分和理解计算。

4　实验规程的过程级语义表示要素抽取方法研究

尽管当前已经出现专门出版实验规程的期刊,但其内容依然使用自然语言形式编写。此外,一般科技文献中依然存在数量庞大且丰富的实验规程[44],基于科技文献的实验规程要素自动抽取依然是当前获取大量可用实验规程的主要方式。下面将从科技文献中的实验规程内容特征、文本标注语料构建方法及自动抽取方法三方面梳理总结。

4.1　研究文献中实验规程内容特征分析

综合当前研究结果[3, 14, 33, 44-47]可知,实验规程内容特征表现为以下三方面。

(1)信息缺失或额外信息掺杂,导致实验规程的完整性和清晰度不足。具体表现为:实验目标不够明确,有些是在文本的步骤中逐步分散给出,有些则需要通过实验描述环节推理;实验规程完整性不足,任务或操作以引用形式给出,或表述较为笼统甚至缺失;任务描述通常伴随附加信息,如理论信息或任务说明。

表3　结构化图表示方法

Table 3　The Methods of Structured Graph Representation

图名称	节点标签	边标签
动作图	表示操作和参数，省略/缺失的参数通过添加“隐式参数”节点来处理，灰色节点表示缺少引用边，表示“原始”节点；定义18类：TARGET、MATERIAL、DESCRIPTOR、AMT_UNIT、CND_MISC、CND_UNIT、INTERMED、OPERATION、NUMBER、AMT_MISC、PROP_UNIT、PROP_TYPE、PROP_MISC、SYNTH_APRT、CHAR_APRT、BRAND、META、REF	表示操作与其参数之间的关系，定义两类：association、reference
合成图	表示材料、操作和属性，定义11类：material-start、material-intermedium、material-final、material-solvent、materials-others、operation、property-time、property-temp、property-rot、property-press、property-atmosphere	表示节点间的关系，定义三类：condition、next、coreference
流程图	表示材料、操作和属性，定义19类：operation、material、nonrecipe-material、number、property-misc、property-type、property-unit、amount-unit、amount-misc、condition-misc、condition-type、synthesis-apparatus、apparatus-unit、apparatus-property-type、material-descriptor、apparatus-descriptor、brand、meta、reference	表示节点间的关系，定义三类：Operation-Operation、Operation-Material、Remaining relations
过程执行图	表示操作和参数，其中操作用橙色标记，表示实验操作动词，共定义14类：Transfer、Temperature、Treatment、General、Mix、Spin、Create、Destroy、Remove、Measure、Wash、Time、Seal、Convert 参数用蓝色标记，表示实验物理对象，共定义8类：Reagent、Measurement、Setting、Location、Modifier、Device、Method、Seal	表示操作与其参数之间的关系，定义三类：core-roles、non-core roles、temporal edges

（2）语义模糊或高度依赖上下文，导致实验规程的可理解性或自动抽取存在诸多困难。具体表现为：实体表述不清，如black solid、clear solution等状态表述词，较为定性，无法进行定量判断；句子中多个动作与相同试剂关联，动作嵌套致使很难区分有效操作动作；句子包含高度依赖上下文的动词，如包含followed by的句子，当前的自动抽取模型很难处理。

（3）实验步骤横跨多个句子、段落甚至是章节，连贯性较差。当前研制的基于科技文献自动抽取实验规程模型，通常直接忽略这种跨段情况，将每段看成一个独立完整的实验步骤。部分商业数据库则依靠专家智慧，人工阅读文本后，将这部分信息提取出来。

这些特征为从科技文献中自动抽取实验规程的过程级语义表示要素带来极大困难，理清这些特征可为自动抽取算法设计提供规则依据，提高抽取效率。

4.2　文本标注语料构建方法研究

数据的质量、数量和多样性，对机器学习模型的准确性和通用性施加了上限[48]，文本标注语料建设至关重要。当前实验规程文本标注语料构建一般包括以下三个环节。

（1）标注语料筛选，主要是识别研究内容高度相关的论文、章节、段落或句子。已有的方法包括基于TF-IDF（Term Frequency-Inverse Document Frequency）[33, 49]或BERT（Bidirectional Encoder Representations from Transformers）[50]的论文分类、基于Scikit-learn[51]的段落[4]和句子[49]分类、基于二元分类器[52]的段落分类及基于字典和规则的句子分类[53]、基于递归神经网络的论文分类及基于隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）的段落分类[33]。

（2）语料标注，依据一定的标注框架形成一系列文本标签，由于科技文献中的实验规程编写不够规范，用词不够恰当，当前实验规程文本标注语料多是依靠专家手工标注，在此基础上应用有监督的主动学习[50]、Scikit-learn[51]、LDA与随机森林结合[46, 54]等方法，改善标签分布不均问题。

（3）标注语料检验，保证语料的质量，通常使用多位注释者之间一致性测评指标，两两标注者之间的一致性可用Krippendorff的α指标[55]，面向多个标注者则可用Cohen的Kappa指标[56]等。

4.3　过程级语义表示要素自动抽取方法研究

实验动作序列抽取可转化为NLP中的实体关系抽取任务，相关研究如表4所示。相较于NLP中相关任务已经研发的算法成果，当前应用在实验规

程中的方法依然采用传统的基于实体对预测关系的思想，抽取效果一般。为提升实验规程数据自动抽取方法效果，可考虑引入NLP中对相关任务已经研发的较为先进的方法。

表4　实验动作序列抽取方法

Table 4　The Methods of Extracting Experimental Action Sequence from Literature

模型名称	文献类型	技术方法	抽取内容
ChemDataExtractor 2.0 [57]	科技文献	User Model包含三部分：Quantity Model（包括任何物理量，如时间、密度等）、Compound Model（包括名称、标签和角色）、Base Model（包括用户定义的字段）	基于用户预定义的化学知识本体
ChemicalTagger [58]	专利	基于规则的语法解析树	实验实体、动作及其关系
Synthesis Project [4]	期刊论文	利用ChemDataExtractor[59]和SpaCy解析器，识别合成动词；混合使用神经网络单词标记和遍历依赖分析树提取合成参数；暂无法处理文本中多个分离的合成路线	实验合成操作及参数
动作图抽取[5]	期刊论文	神经网络模型：实体抽取；语法树解析：动作图抽取	实验动作、参数及关系
SynthReader [9]	期刊论文	基于专家定义的模式匹配启发法NLP算法	把实验程序由文本格式转化为χDL格式
专利实验规程抽取[14]	专利	Pistachio中的规则模型：LeadMine + ChemicalTagger，自定义的基于规则的NLP模型	实验动作及相关的化合物、数量和反应条件等
合成图抽取[42]、流程图抽取[43]	期刊论文	基于深度学习的序列标记模型（Mat-ELMo和Bi-LSTM-CRF）：抽取实体；基于简单启发式规则的关系提取器：抽取关系	固态电池制造实验合成图；材料合成实验规程
过程执行图（PEG）抽取[15]	文本标注语料	基于SciBERT和消息传递神经网络的PEG预测管道方法，同时该网络联合学习跨度和关系表示	生物化学实验规程
固体氧化物电池实验规程抽取[60]	期刊论文	BiLSTM-CRF Mat2Vec+Word2Vec	材料、值、装置、实验槽等信息
无机材料合成实验规程抽取[54]	期刊论文	ChemDataExtractor：单词标记及预处理；LDA+RF：实验信息分类；马尔科夫链：动作序列预测	合成方法、实验步骤和详细的加工参数
无机领域合成实验规程抽取[32-33, 46]	期刊论文	MatBERT-BiLSTM-CRF：材料实体识别；RNN+基于规则的句子依赖树解析算法：实验合成动作识别与分类	材料：目标、前体或其他材料，以及实验合成的动作序列
实验规程实体及关系管道抽取[61]	文本标注语料	基于PubmedBERT的神经穷举模型：实验实体识别；神经穷举模型扩展模型：实验关系识别（实验动作序列）	实验实体及实验动作序列抽取效果最好的方法，其他稍弱的方法见文献[62]
实验规程实体及关系管道抽取[63]	文本标注语料	模型分三部分：跨度表示：跨句关系关键特征提取；关系编码器+卷积+解码器：发现长距离输入实体之间的局部关系；多头R-GCN（多头关系图卷积网络）：解决隐式参数	解决隐式论点（动作和实体之间的隐藏关系）以及超越局部的跨句子长范围语义关系

5　实验规程的过程级语义表示数据应用研究

5.1　数据驱动知识发现

基于实验规程的过程级语义表示数据，结合机器学习先进技术，已实现过程条件优化、完整实验规程预测、功能材料发现等。其中，过程条件优化应用最为广泛，完整的实验规程预测及功能材料发现均只出现一项研究，未来可探索空间较大。

在过程条件优化应用中，将实验规程中的实验条件数据与统计分析方法相结合，预测高性能激光融合实验条件设计[64]。基于合成反应和实验条件数据，利用机器学习模型为选择的5个反应类预测合适的溶剂[65]。利用神经网络算法实现预测面向任何反应类的溶剂、溶液、催化剂和温度[66]，并已应用于机器人平台规划合成[34]。更进一步，将反应条件的预测表述为4类交叉耦合反应的多层预测问题，研发的多标签分类模型可为金属类别、配体、碱、溶剂、添加剂、温度、活化剂以及一氧化碳环境的表现和压力选择类别值[67]。

在完整实验规程预测应用中，由于缺乏可直接使用的高质量实验规程数据集和合适的机器学习计算模型，当前仅有Vaucher等[68]发表一篇预测完整实验程序的论文。使用Pistachio数据库中规范化建设的反应、实验程序、分子及相关化合物数据，基于该团队开发的NLP模型（Paragraph2Actions）[14]，将实验程序文本转化为动作序列数据集，训练三个机器学习模型用来预测特定实验的动作序列，结果显示有超过50%的预测结果可在无人工干预下执行。

在功能材料发现应用中，受到可用数据的限制，以往的研究通常仅基于理论或材料功能性质推理预测新功能材料[69-71]，忽略多功能材料设计中特殊属性的权衡问题，如一个需要高温处理的材料绝对不能与纸质基板堆叠，而这些信息通常只会在实验规程中出现，因此结合实验规程数据的多功能材料预测相对更为准确，显著减少试错次数[72]。

5.2 智能实验平台

将实验规程的过程级语义表示数据嵌入智能实验平台，已成功运行并在材料、化学和生命科学等领域实现实验自动化、数字化和标准化，同时助力诸多研究发现。

在材料领域，将定制化的实验规程及优化算法模型同时嵌入模块化机器人平台，在实验执行过程中不断修改模型中薄膜成分和加工条件，自主优化并迭代30次实验后，找到符合目标需求的有机和无机材料[73]。类似地，材料加速操作系统实现按需合成特定功能材料[74]；实时闭环自动化材料发现与优化实验平台发现潜在的实用新材料[75]。2023年，我国也在该方向取得重大突破，发布合成胶体纳米晶材料机器平台，机器人辅助正交实验，单因素、双因素以及三因素实验等，实现纳米晶材料数字智造[76]。

在化学领域，将定制化的实验规程嵌入机器人控制的智能实验平台，结合人工智能驱动的合成路径规划算法，实现15个医学相关小分子的自动化制备[34]。类似地，将实验规程嵌入移动实验机器人，8天时间自主完成668个实验并研发一种全新的催化剂[77]。2022年，我国也在该方向取得突破性进展，发布具有科学思维的全流程人工智能机器化学家AI Chemist，可在14个工作站上自动执行完整的实验过程，包括合成、表征和性能测试，成功执行三类不同的化学任务[78]。

在生物医药领域，将微生物批量生长的实验规程级数据分析软件嵌入智能实验平台Adam，实现特定微生物菌株的培养及其生长曲线监测，从而发现酵母中发生催化生化反应酶的编码基因[23]。将热带病药物筛选、库筛选、命中率确认等实验规程嵌入机器人控制的智能实验平台Eve，实现潜在药物筛选[79]。

6 问题及趋势分析

本文在明确定义实验规程、实验规程的过程级语义表示概念基础上，围绕实验规程的过程级语义表示方法、表示要素抽取方法以及该类数据应用三方面开展综述研究。基于各方向已出现的研究成果，本文认为实验规程的过程级语义表示研究整体处于发展初期，发展前景广阔，具有较高的研究价值。

本文在梳理三方面研究进展中，发现系列尚需解决的关键研究问题，下面将具体论述。

6.1 实验规程的过程级语义表示方法研究

首先，与实验规程相关的本体，从关注实验规程内容，到实验过程执行计划，再到实验规程过程安全，三类本体之间是否能够以及如何更好地关联融合，从而多角度、全面描述实验规程，确保实验规程完整、可重用、可复现。

其次，面向特定应用场景的数据模型，如有机、无机合成，只覆盖部分研究领域，且彼此间实验动作及表示参数设置差异较大，可泛化性、互操作性较差，尤其是对钙钛矿太阳能电池等有机无机杂化合成领域的借鉴性有限。面向智能实验平台的数据模型中，设备无关的实验规程要素表示理念在一定程度上保证实验动作及其参数可灵活组配，但难点在于如何设计合适的实验操作单元。

最后，结构化图表示方法将实验规程的执行过程表示为有向无环图，可视化效果较好，且今后可结合网络的相关方法开展实验动作因果推理研究等，以预测完整的实验规程动作序列。但难点之一是节点同时用来表示实验动作、属性、材料等实体，如何在计算分析时巧妙地区分、避免干扰。

6.2　实验规程的过程级语义表示要素抽取方法研究

一方面，改进现有的实验规程出版模式。尽管已经出现专业的实验规程期刊，并且对实现规程的编写方式等提出较多要求，但依然以自然语言形式编写，尚需使用自然语言处理技术将其转化为计算机可读取计算、智能实验平台可读取执行的格式。因此，可考虑在编写时即以过程级语义表示形式呈现，提供一个丰富的实验操作单元库，在涉及相关动作描述时直接调用，仅修改参数值即可。

另一方面，更好地发挥已有科技文献的价值。当前已出版的科技文献中依然存在大量丰富的实验规程，但普遍存在表述模糊、完整性不足等问题。需要研究利用先进的自然语言处理技术，如计算机领域实体关系联合抽取模型[80]、大语言模型[81]，自动从科技文献中精准、高效地获取实验规程表示要素，更轻松地填充各领域相关数据库，而研究人员不需要手动输入数据。

6.3　实验规程数据应用

当前实验规程数据已在某数据驱动知识发现和智能实验平台中开展部分应用实践，未来伴随可用实验规程的过程级语义表示数据量以及数据质量的提升，将会产生更丰富的应用成果，有待感兴趣的科研人员广泛探索。

参考文献：

[1] Jumper J, Evans R, Pritzel A, et al. Highly Accurate Protein Structure Prediction with AlphaFold[J]. Nature, 2021, 596(7873): 583-589.

[2] De Pablo J J, Jackson N E, Webb M A, et al. New Frontiers for the Materials Genome Initiative[J]. NPJ Computational Materials, 2019, 5: 41.

[3] Girault I, D'Ham C, Ney M, et al. Characterizing the Experimental Procedure in Science Laboratories: A Preliminary Step Towards Students Experimental Design[J]. International Journal of Science Education, 2012, 34(6): 825-854.

[4] Kim E, Huang K, Saunders A, et al. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning[J]. Chemistry of Materials, 2017, 29(21): 9436-9444.

[5] Mysore S, Kim E, Strubell E, et al. Automatically Extracting Action Graphs from Materials Science Synthesis Procedures [OL]. arXiv Preprint, arXiv:1711.06872.

[6] Baker M. 1, 500 Scientists Lift the Lid on Reproducibility[J]. Nature, 2016, 533(7604): 452-454.

[7] Seifrid M, Pollice R, Aguilar-Granda A, et al. Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab[J]. Accounts of Chemical Research, 2022, 55(17): 2454-2466.

[8] Coley C W, Eyke N S, Jensen K F. Autonomous Discovery in the Chemical Sciences Part II: Outlook[J]. Angewandte Chemie, 2020, 59(52): 23414-23436.

[9] Mehr S H M, Craven M, Leonov A I, et al. A Universal System for Digitization and Automatic Execution of the Chemical Synthesis Literature[J]. Science, 2020, 370(6512): 101-108.

[10] Soldatova L N, King R D. An Ontology of Scientific Experiments [J]. Journal of the Royal Society, Interface, 2006, 3(11): 795-803.

[11] Lewis T. Design and Inquiry: Bases for an Accommodation Between Science and Technology Education in the Curriculum? [J]. Journal of Research in Science Teaching, 2006, 43(3): 255-281.

[12] Yang X J, Zhang X L, Zuo J, et al. An Analysis of Relation Extraction Within Sentences from Wet Lab Protocols[C]// Proceedings of the 2021 IEEE International Conference on Big Data. 2021: 562-570.

[13] Soldatova L N, Nadis D, King R D, et al. EXACT2: The Semantics of Biomedical Protocols[J]. BMC Bioinformatics, 2014, 15(14): S5.

[14] Vaucher A C, Zipoli F, Geluykens J, et al. Automated Extraction of Chemical Synthesis Actions from Experimental Procedures[J]. Nature Communications, 2020, 11: 3601.

[15] Tamari R, Bai F, Ritter A, et al. Process-Level Representation of Scientific Protocols with Interactive Annotation[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2021: 2190-2202.

[16] Steiner S, Wolf J, Glatzel S, et al. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language [J]. Science, 2019, 363(6423): eaav2211.

[17] Arch-int N, Arch-int S. Semantic Ontology Mapping for Interoperability of Learning Resource Systems Using a Rule-Based Reasoning Approach[J]. Expert Systems with Applications, 2013, 40(18): 7428-7443.

[18] Daraio C, Lenzerini M, Leporelli C, et al. The Advantages of an Ontology-Based Data Management Approach: Openness, Interoperability and Data Quality[J]. Scientometrics, 2016, 108 (1): 441-455.

[19] Nelson E K, Piehler B, Eckels J, et al. LabKey Server: An Open Source Platform for Scientific Data Integration, Analysis and Collaboration [J]. BMC Bioinformatics, 2011, 12: 71.

[20] Rodríguez M, Laguía J. An Ontology for Process Safety [J]. Chemical Engineering Transactions, 2019, 77: 67-72.

[21] McGuinness D L, Harmelen F V. Web Ontology Language [A]//

Encyclopedia of Social Network Analysis and Mining[M]. New York: Springer, 2014.

[22] Kügler P, Marian M, Schleich B, et al. tribAIn—Towards an Explicit Specification of Shared Tribological Understanding [J]. Applied Sciences, 2020, 10(13): 4421.

[23] King R D, Rowland J, Oliver S G, et al. The Automation of Science [J]. Science, 2009, 324(5923): 85-89.

[24] Qi D, King R D, Hopkins A L, et al. An Ontology for Description of Drug Discovery Investigations [J]. Journal of Integrative Bioinformatics, 2010, 7(3): 126.

[25] Vanschoren J, Soldatova L N. Exposé: An Ontology for Data Mining Experiments[C]//Proceedings of International Workshop on the 3rd Generation Data Mining: Towards Service-Oriented Knowledge Discovery . 2010: 31-46.

[26] Cheung K, Drennan J, Hunter J. Towards an Ontology for Data-Driven Discovery of New Materials[C]//Proceedings of Semantic Scientific Knowledge Integration AAAI/SSS Workshop. 2008: 9-14.

[27] Soldatova L N, Aubrey W, King R D, et al. The EXACT Description of Biomedical Protocols [J]. Bioinformatics, 2008, 24 (13): i295-i303.

[28] Celebi R, Moreira J R, Hassan A A, et al. Towards FAIR Protocols and Workflows: The OpenPREDICT Use Case [J]. PeerJ Computer Science, 2020, 6: e281.

[29] Barrows E, Martin K, Smith T. Markup Language for Chemical Process Control and Simulation [J]. Computers & Chemical Engineering, 2022, 160: 107702.

[30] Wang Z R, Cruse K, Fei Y X, et al. ULSA: Unified Language of Synthesis Actions for the Representation of Inorganic Synthesis Protocols [J]. Digital Discovery, 2022, 1(3): 313-324.

[31] Kononova O, Huo H Y, He T J, et al. Text-Mined Dataset of Inorganic Materials Synthesis Recipes [J]. Scientific Data, 2019, 6: 203.

[32] Wang Z R, Kononova O, Cruse K, et al. Dataset of Solution-Based Inorganic Materials Synthesis Procedures Extracted from the Scientific Literature [J]. Scientific Data, 2022, 9: 231.

[33] Cruse K, Trewartha A, Lee S, et al. Text-Mined Dataset of Gold Nanoparticle Synthesis Procedures, Morphologies, and Size Entities [J]. Scientific Data, 2022, 9: 234.

[34] Coley C W, Thomas III D A, Lummiss J A M, et al. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning [J]. Science, 2019, 365(6453): eaax1566.

[35] Hammer A J S, Leonov A I, Bell N L, et al. Chemputation and the Standardization of Chemical Informatics [J]. JACS Au, 2021, 1 (10): 1572-1587.

[36] Wang Z, Zhao W, Hao G F, et al. Automated Synthesis: Current Platforms and Further Needs [J]. Drug Discovery Today, 2020, 25 (11): 2006-2011.

[37] Collins N, Stout D, Lim J P, et al. Fully Automated Chemical Synthesis: Toward the Universal Synthesizer [J]. Organic Process Research & Development, 2020, 24(10): 2064-2077.

[38] Bubliauskas A, Blair D J, Powell-Davies H, et al. Digitizing Chemical Synthesis in 3D Printed Reactionware [J]. Angewandte Chemie, 2022, 61(24): e202116108.

[39] Angelone D, Hammer A J S, Rohrbach S, et al. Convergence of Multiple Synthetic Paradigms in a Universally Programmable Chemical Synthesis Machine [J]. Nature Chemistry, 2021, 13(1): 63-69.

[40] Wilbraham L, Mehr S H M, Cronin L. Digitizing Chemistry Using the Chemical Processing Unit: From Synthesis to Discovery [J]. Accounts of Chemical Research, 2021, 54(2): 253-262.

[41] Rohrbach S, Šiaučiulis M, Chisholm G, et al. Digitization and Validation of a Chemical Synthesis Literature Database in the ChemPU [J]. Science, 2022, 377(6602): 172-180.

[42] Kuniyoshi F, Makino K, Ozawa J, et al. Annotating and Extracting Synthesis Process of All-Solid-State Batteries from Scientific Literature [OL]. arXiv Preprint, arXiv:2002.07339.

[43] Makino K, Kuniyoshi F, Ozawa J, et al. Extracting and Analyzing Inorganic Material Synthesis Procedures in the Literature [J]. IEEE Access, 2022, 10: 31524-31537.

[44] Guo J, Ibanez-Lopez A S, Gao H Y, et al. Automated Chemical Reaction Extraction from Scientific Literature [J]. Journal of Chemical Information and Modeling, 2022, 62(9): 2035-2045.

[45] Mysore S J Z, Kim E, Huang K, et al. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures [C]// Proceedings of the 13th Linguistic Annotation Workshop (Law Xiii). 2019: 56-64.

[46] Kononova O, Huo H Y, He T J, et al. Author Correction: Text-Mined Dataset of Inorganic Materials Synthesis Recipes [J]. Scientific Data, 2019, 6: 273.

[47] Kim E, Huang K, Kononova O, et al. Distilling a Materials Synthesis Ontology [J]. Matter, 2019, 1(1): 8-12.

[48] Artrith N, Butler K T, Coudert F X, et al. Best Practices in Machine Learning for Chemistry [J]. Nature Chemistry, 2021, 13 (6): 505-508.

[49] Hiszpanski A M, Gallagher B, Chellappan K, et al. Nanomaterial Synthesis Insights from Machine Learning of Scientific Articles by Extracting, Structuring, and Visualizing Knowledge [J]. Journal of Chemical Information and Modeling, 2020, 60(6): 2876-2887.

[50] Zhang Y, Wang C, Soukaseum M, et al. Unleashing the Power of Knowledge Extraction from Scientific Literature in Catalysis [J]. Journal of Chemical Information and Modeling, 2022, 62(14): 3316-3330.

[51] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python [J]. Journal of Machine Learning Research, 2011, 12: 2825-2830.

[52] Kim E, Huang K, Tomala A, et al. Machine-Learned and Codified Synthesis Parameters of Oxide Materials [J]. Scientific Data, 2017, 4: 170127.

[53] Wang W R, Jiang X, Tian S H, et al. Automated Pipeline for Superalloy Data by Text Mining [J]. NPJ Computational Materials, 2022, 8: 9.

[54] Huo H Y, Rong Z Z, Kononova O, et al. Semi-Supervised Machine-Learning Classification of Materials Synthesis Procedures [J]. NPJ Computational Materials, 2019, 5: 62.

[55] Krippendorff K. Content Analysis: An Introduction to Its Methodology[M]. The 4th Edition. Thousand Oaks: SAGE Publications, 2019.

[56] Cohen J. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit [J]. Psychological Bulletin, 1968, 70(4): 213-220.

[57] Mavračić J, Court C J, Isazawa T, et al. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science [J]. Journal of Chemical Information and Modeling, 2021, 61(9): 4280-4289.

[58] Hawizy L, Jessop D M, Adams N, et al. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry [J]. Journal of Cheminformatics, 2011, 3(1): 1-13.

[59] Swain M C, Cole J M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature [J]. Journal of Chemical Information and Modeling, 2016, 56(10): 1894-1904.

[60] Friedrich A, Adel H, Tomazic F, et al. The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1255-1268.

[61] Sohrab M G, Duong Nguyen A K, Miwa M, et al. mgsohrab at WNUT 2020 Shared Task-1: Neural Exhaustive Approach for Entity and Relation Recognition over Wet Lab Protocols[C]// Proceedings of the 6th Workshop on Noisy User-Generated Text. 2020: 290-298.

[62] Tabassum J, Xu W, Ritter A, et al. WNUT-2020 Task 1 Overview: Extracting Entities and Relations from Wet Lab Protocols[C]// Proceedings of the 6th Workshop on Noisy User-Generated Text. 2020: 260-267.

[63] Kulkarni C, Chan J, Fosler-Lussier E, et al. Learning Latent Structures for Cross Action Phrase Relations in Wet Lab Protocols [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 6737-6750.

[64] Gopalaswamy V, Betti R, Knauer J P, et al. Tripled Yield in Direct-Drive Laser Fusion Through Statistical Modelling [J]. Nature, 2019, 565(7741): 581-586.

[65] Walker E, Kammeraad J, Goetz J, et al. Learning to Predict Reaction Conditions: Relationships Between Solvent, Molecular Structure, and Catalyst [J]. Journal of Chemical Information and Modeling, 2019, 59(9): 3645-3654.

[66] Gao H Y, Struble T J, Coley C W, et al. Using Machine Learning to Predict Suitable Conditions for Organic Reactions [J]. ACS Central Science, 2018, 4(11): 1465-1476.

[67] Maser M R, Cui A Y, Ryou S, et al. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions [J]. Journal of Chemical Information and Modeling, 2021, 61(1): 156-166.

[68] Vaucher A C, Schwaller P, Geluykens J, et al. Inferring Experimental Procedures from Text-Based Representations of Chemical Reactions [J]. Nature Communications, 2021, 12: 2573.

[69] Miyao T, Kaneko H, Funatsu K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x) [J]. Journal of Chemical Information and Modeling, 2016, 56(2): 286-299.

[70] Tagade P M, Adiga S P, Pandian S, et al. Attribute Driven Inverse Materials Design Using Deep Learning Bayesian Framework [J]. NPJ Computational Materials, 2019, 5: 127.

[71] Onishi T, Kadohira T, Watanabe I. Relation Extraction with Weakly Supervised Learning Based on Process-Structure-Property-Performance Reciprocity [J]. Science and Technology of Advanced Materials, 2018, 19(1): 649-659.

[72] Fukada K, Seyama M. Designing a Multilayer Film via Machine Learning of Scientific Literature [J]. Scientific Reports, 2022, 12: 930.

[73] MacLeod B P, Parlane F G L, Morrissey T D, et al. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials [J]. Science Advances, 2020, 6(20): eaaz8867.

[74] Li J G, Tu Y X, Liu R L, et al. Toward "On-Demand" Materials Synthesis and Scientific Discovery Through Intelligent Robots [J]. Advanced Science, 2020, 7(7): 1901957.

[75] Kusne A G, Yu H S, Wu C M, et al. On-the-Fly Closed-Loop Materials Discovery via Bayesian Active Learning [J]. Nature Communications, 2020, 11: 5966.

[76] Zhao H T, Chen W, Huang H, et al. A Robotic Platform for the Synthesis of Colloidal Nanocrystals [J]. Nature Synthesis, 2023, 2 (6): 505-514.

[77] Burger B, Maffettone P M, Gusev V V, et al. A Mobile Robotic Chemist [J]. Nature, 2020, 583(7815): 237-241.

[78] Zhu Q, Zhang F, Huang Y, et al. An All-Round AI-Chemist with a Scientific Mind [J]. National Science Review, 2022, 9(10): nwac190.

[79] Williams K, Bilsland E, Sparkes A, et al. Cheaper Faster Drug Development Validated by the Repositioning of Drugs Against

Neglected Tropical Diseases [J]. Journal of the Royal Society, Interface, 2015, 12(104): 20141289.

[80] Wei Z P, Su J L, Wang Y, et al. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 1476-1488.

[81] OpenAI. GPT-4 Technical Report [OL]. arXiv Preprint, arXiv: 2303.08774.

作者贡献声明：

付芸：构思研究内容，设计研究框架，处理数据，撰写及修改论文；
刘细文：提出研究问题，优化研究内容和框架，修改及定稿；
朱丽雅：梳理文献，修改论文；
韩涛：修改研究框架，修改论文。

利益冲突声明：

所有作者声明不存在利益冲突关系。

收稿日期：2023-04-14
收修改稿日期：2023-06-25

Review of Semantic Representation of Experimental Protocols at Process-Level

Fu Yun[1,2] Liu Xiwen[1,2] Zhu Liya[1] Han Tao[1,2]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: [**Objective**] This paper explores the research progress of the process-level semantic representation of experimental protocols. It aims to discover the key issues to be addressed and identify development trends. [**Coverage**] We used related topics to retrieve the relevant literature from Web of Science, arXiv, Engineering Village, CNKI, Wanfang, and VIP. We also examined the requirements of the submission requirements and evaluation principles of renowned journals on experimental protocols. [**Methods**] First, we defined the concepts of experimental protocols and their semantic representation at the process-level. Then, we examined the representation methods, representation element extraction, and application of representative data. [**Results**] The research on process-level semantic representation is in the early development stages. The representation framework was not unified, and the elements were different. The experimental protocols were mainly written in natural language, which were difficult to extract the representation elements automatically. Some studies explored the application of process-level semantic representation data, which leaves more knowledge gaps to be filled. [**Limitations**] This paper does not thoroughly discuss the technical details of extracting representation elements from literature and data application methods. [**Conclusions**] We need to establish a unified representation framework for more complete elements by integrating various representation methods. We should also explore automatic extraction methods based on advanced intelligent technology and application using the process-level semantic representation data.

Keywords: Experimental Protocols Process-Level Semantic Representation Representation Method Representation Element Extraction Method Data Application