

中药材化学成分的晶体结构数据集

李英勇^{1,2*}

ISSN 2096-2223

CN 11-6035/N



文献 DOI:

10.11922/11-6035.csd.2021.0052.zh

数据 DOI:

10.11922/sciencedb.01090

文献分类: 化学

收稿日期: 2021-09-06

开放同评: 2021-09-22

录用日期: 2021-12-08

发表日期: 2021-03-29

1. 中国科学院上海有机化学研究所, 上海 200032

2. 国家基础学科公共科学数据中心, 北京 100190

摘要: 收集提取 1945–2015 年中外科技期刊内的中药材化学成分数据和化合物晶体结构数据, 采用数据录入规范控制和数据加工系统软件检查等多种途径, 对数据进行质量控制和规范化整理, 录入了中药材化学成分数据和化合物晶体结构数据, 并通过化学结构整合两类数据, 最终建成了中药材化学成分的晶体结构数据集。本数据集共收录了 527 种常用中药材、270 种中药化学成分和 345 个化合物的晶体 CIF 数据, 并建立了三项数据间的对应关系, 录入了对应关系的来源文献信息。中药材化学成分的晶体结构数据集的建成, 可以为药物化学家和药学家等科研人员进行中药现代化研究、新药研发等工作提供科研基础数据服务。

关键词: 中药材; 化学成分; 化学结构; 晶体结构

数据集基本信息简介

数据集名称	中药材化学成分的晶体结构数据集
数据通信作者	李英勇 (liyongyong@sioc.ac.cn)
数据作者	李英勇
数据时间范围	1945 – 2015年
地理区域	世界各国
数据量	102 MB
数据格式	*.mdb
数据服务系统网址	https://dx.doi.org/10.11922/sciencedb.01090
基金项目	中国科学院“十四五”网络安全和信息化专项(WX145XQ07-12)
数据集组成	本数据集由1个Microsoft Office Access®文件组成, 包括4个数据表(TcmCmpd表、Compound表、CmpCIF表和CIF表)和2个检索窗体(中药检索窗体和化学成分检索窗体)。数据集共收录了527种常用中药材、270种中药化合物和345个化合物的晶体CIF数据, 主要包括中药材中文名称、拉丁名称、中药化学成分的SRN号、CASRN号、化合物名称、化合物SMILES数据、晶体结构CIF数据和来源文献数据等。

引言

埃博拉病毒、SARS 病毒和新冠病毒等疾病的持续性、大规模爆发, 对人类的生命与健康造成了极大威胁, 埃博拉病毒、SARS 病毒和新冠病毒等疾病的持续性、大规模爆发, 对人类的生命与健康造成了极大威胁, 甚至显著改变了人类的

* 论文通信作者

李英勇: liyongyong@sioc.ac.cn

生产与生活方式。面对疾病的不断威胁, 药物研究受到了国家和政府长期持续的关注。20 世纪初至 80 年代, 化学药物研究获得了飞速发展, 中外科研人员发明出了一系列重要的基本化学药物。但 90 年代以后, 新药研究的风险越来越大, 药物研发逐渐成为一个高投入、高风险、低产出的领域。美国食品和药品监管局(FDA)近期数据显示, 在被研究的众多化合物中, 最终能成为药物的仅占 1/5000, 如此低的研发成功率, 导致仅有约 30% 的化学药物能够最终盈利。在新药研发过程中, 对人体的安全性以及对疾病的有效性问题是导致化学药物临床研究失败的主要原因^[1-2]。而天然产物的化学成分大多具有丰富的结构多样性和较好的生物活性, 相对于纯粹化学合成的化合物, 往往具有毒性相对更小的特点^[1,3]。因此, 天然产物逐渐受到药学研发人员的关注和深入研究。

作为天然产物的一个重要来源, 中药材在东南亚, 特别是在中国有着长期的治病历史, 人们积累了丰富的中药材临床应用经验。因此, 中药材作为寻找新药先导物的研究对象, 逐渐受到新药研发人员的重视, 中医药的现代化研究也取得了飞速发展, 但科研领域仍没有开放的、格式严谨的中药材化学成分数据集可供使用, 相关研究数据散布于中医中药类期刊文献中, 其检索、筛选与使用都极为不便。

针对这一问题, 收集整理中医中药类科研期刊文献, 人工提取出文献中包含的中药材数据和化学成分数据, 建成了数据集。另外, 由于化学物质的晶体结构数据在药物的化学成分分离、结构确认等研究中具有重要价值, 特从化学类期刊文献中收集、录入了化合物的晶体结构数据, 并通过化学结构与中药材的化学成分数据实现了整合, 最终建成了中药材化学成分的晶体结构数据集, 数据集内共包含 527 种中药材、270 种化学成分和 345 个化合物晶体结构数据, 并建立了三种数据间的对应关系。其中, 中药材及化学成分数据来源于收集整理的 131 种中医中药类科技期刊, 化合物晶体结构数据来源于 101 种化学晶体类科技期刊, 数据集中提供了数据的来源文献信息, 便于用户溯源。此数据集的建成, 可为药学家和生物化学家开展中药化学成分与蛋白晶体结构对接、中药活性成分筛选^[4-5]等研究工作, 提供重要的基础数据支撑服务。

1 数据采集和处理方法

1.1 数据来源

中药材数据和化学成分数据来源于 1983–2001 年国内中医中药类科技期刊, 共收集整理了 131 种科技期刊, 来源文献如图 1 所示, 文献中包含中药材数据和化学成分数据。

摘要: 为了研究脱皮马勃 *Lasiosphaera fenzlii* Reich 子实体的化学成分, 通过石油醚索氏提取、硅胶柱色谱、葡聚糖凝胶色谱、重结晶分离纯化, 从脱皮马勃子实体中分离得到了 6 个化合物, 通过波谱数据分析分别鉴定它们为 (22E, 24R) 麦角甾-7, 22-二烯-3 β -醇 (1), 麦角甾-7, 22-二烯-3, 6-二酮 (2), 麦角甾-5 α , 8 α -环二氧-6, 22-二烯-3 β -醇 (3), 麦角甾-5, 7, 22-三烯-3-醇 (4), 麦角甾-7, 22-二烯-3-酮 (5), 硬脂酸 (6), 其中化合物 1~3, 6 是首次从脱皮马勃中分离得到。
关键词: 脱皮马勃; 化学成分; 麦角甾; 硬脂酸

图 1 中药化学成分数据来源文献示例

Figure 1 Example of source literatures for chemical constituents in traditional Chinese medicines

化合物的晶体结构数据来源于 1945–2015 年国外化学类科技期刊, 共收集整理了 101 种科技期刊, 偏重于化学类和化合物晶体结构类科技期刊。晶体数据往往以 CIF^[6]文件形式存在于期刊文献的补充材料中, 晶体学信息文件(Crystallographic Information File, CIF)是一种以“.cif”结尾的计算机文件, 它包含了晶体的详细信息, 如晶胞参数、原子坐标、化合物数据、来源文献等。作为晶体结构

描述、解析、传播和表达时最常使用的文件格式，CIF 文件广泛应用于晶体结构绘图、XRD 精修以及材料理论计算等各个方面，在材料科学和化学结构确认研究中有着重要作用。图 2 是 CIF 文件部分内容示例，完整的 CIF 文件数据示例见第 2 章。

```
loop_
  _atom_site_label
  _atom_site_type_symbol
  _atom_site_fract_x
  _atom_site_fract_y
  _atom_site_fract_z
  _atom_site_U_iso_or_equiv
  _atom_site_adp_type
  _atom_site_occupancy
  _atom_site_symmetry_multiplicity
  _atom_site_calc_flag
  _atom_site_refinement_flags
  _atom_site_disorder_assembly
  _atom_site_disorder_group
O1 O 0.20503(10) -0.19298(9) -0.30573(8) 0.0189(2) Uani 1 1 d . . .
O2 O 0.25946(11) 0.02227(10) -0.58261(9) 0.0286(2) Uani 1 1 d . . .
O3 O 0.07404(11) 0.22642(9) -0.27222(8) 0.0250(2) Uani 1 1 d . . .
O4 O -0.03017(10) -0.00306(10) -0.18992(9) 0.0219(2) Uani 1 1 d . . .
H4O H -0.090(2) -0.0399(19) -0.2605(18) 0.052(6) Uiso 1 1 d . . .
O5 O 0.34076(10) 0.15684(10) -0.05413(9) 0.0230(2) Uani 1 1 d . . .
H5O H 0.388(2) 0.100(2) -0.0824(17) 0.048(6) Uiso 1 1 d . . .
```

图 2 化合物晶体数据来源文件示例

Figure 2 Compound crystal data source file example

1.2 数据采集方法

针对中医中药类期刊文献的内容特点，基于 Microsoft® Office Access® 软件开发了中药材化学成分的晶体数据加工处理系统，加工系统界面如图 3 所示。录入人员通过数据加工系统采集、整理和加工中医中药类期刊文献中的中药材数据和化学成分数据，最后汇集处理录入的数据。



图 3 中药材化学成分信息标引系统界面

Figure 3 Interface of Chemical constituents in traditional Chinese medicines information indexing system

中医中药类文献中往往在正文中表达研究的中药材信息，以及中药的化学成分信息。由于文章

语言的结构松散与表达方式多样，中药材对象与化学成分间对应关系的确定具有一定难度，当一篇文献研究多个中药材对象时，多个中药材对象与多个化学成分间的对应关系就更加复杂。针对来源文献的这些特点，制定了中药材化学成分数据的数据采集与录入规则：

标引对象是来源文献中的中药材信息（包括中药材中文名称、拉丁名称），化合物信息（包括化合物名称、CASRN 号），来源文献信息（包括期刊名称、年、卷、期、起止页数据），以及中药材信息、化合物信息以及来源文献信息间的对应关系。

中药材信息标引规则：与加工系统中中药材名称表记录匹配者是有效的中药材名称。如果遇到名称表未收录的新名称，则由专家审核确认后加入中药材名称表。

化合物信息标引规则：仅收录确定的化合物名称数据。如果化合物名称中包含模糊信息，或者无法根据名称确定对应的化学结构，则不收录。

来源文献信息标引规则：根据期刊信息描述规范，提取来源文献的期刊名称、年、卷、期和起止页数据。

中药材对象与化合物对象的对应关系信息提取规则：根据文献正文描述，如果能确定二者有包含或从属关系，则录入二者间的对应信息。

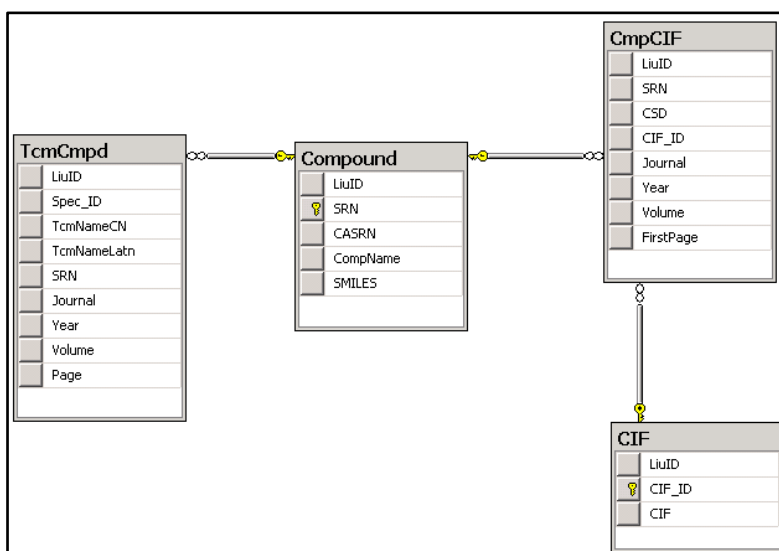
由于化合物的晶体结构数据以 CIF 文件形式存在于来源文献的补充材料中，所以先根据 CIF 文件格式设计开发 CIF 数据分析提取软件，然后制定了化合物晶体结构数据的具体处理流程：首先，采用下载软件批量下载目标文献补充材料中的 CIF 文件。然后，通过 CIF 数据分析提取软件从中提取出来源文献数据和化合物数据。如果化合物数据项为空，则由录入人员从来源文献中人工标引对应的化合物数据。最后，将化合物名称、CIF 文件和来源文献数据导入目标数据集。

1.3 录入数据后处理

录入的中药材化学成分数据和化合物晶体结构数据都包含有化合物，可以通过化合物实现两类数据的整合，具体处理流程是：首先，采用化合物名称生成化学结构 Web 服务程序^[7]，将中药材化学成分名称和 CIF 文件对应的化合物名称分别生成化学结构 SMILES 数据^[8]（简化分子线性输入规范，一种用 ASCII 字符串明确描述分子结构的规范，广泛用于化学结构数据的数据库存储与网络交互传输）；然后，通过化合物登录库 Web 服务系统^[9]得到化学结构的 SRN 号（SIOC Registry Number，化合物登录库根据化学结构建立并发布的化合物标识码，由多位数字组成，可用于识别不同来源、不同格式的化学结构是否属于同一个化合物），并根据 SRN 号归并不同中药材的相同化学成分记录；最后，根据化合物 SRN 号建立起两类数据间的对应关系，从而实现了数据集内中药材化学成分数据和化合物晶体结构数据的整合。

2 数据样本描述

中药材化学成分的晶体结构数据集共包括 4 个数据表：中药材化学成分数据表、化合物信息表、化合物晶体数据表和晶体信息表。中药材化学成分数据表主要存储中药材数据、化学成分数据和来源文献数据；化合物信息表存储了化合物的中文名称、拉丁名称、SRN 号、CASRN 号和化学结构 SMILES 数据；化合物晶体数据表存储了化合物数据、晶体 CIF 文件 ID 号和来源文献数据；晶体信息表主要存储了晶体结构 CIF 数据。4 个数据表的属性名称、属性说明和数据范例分别如表 1-4 所示。图 4 是数据集的实体关系图，展示了 4 个数据表实体间的逻辑关系。



注：TcmCmpd 是中药材化学成分数据表，Compound 是化合物信息表，CIF 表是晶体信息表，CmpCIF 是化合物晶体数据表。Compound 表以主键 SRN 与 TcmCmpd 和 CmpCIF 表连接，CIF 表以主键 CIF_ID 与 CmpCIF 表连接。

图 4 中药材化学成分的晶体结构数据集实体关系图

Figure 4 Entity relationship diagram of the crystal structure dataset of the chemical constituents in traditional Chinese medicines

表 1 中药材化学成分数据示例

Table 1 Example of chemical constituents in traditional Chinese medicines

序号	属性名称	属性说明	范例
1	LiuID	流水号	399569
2	Spec_ID	中药材 ID 号	1465
3	TCMNameCN	中药材中文名称	麦门冬
4	TCMNameLatn	中药材拉丁名	Radix Ophiopogonis seu Liriope
5	SRN	化学成分的 SRN 号	339902
6	Journal	来源期刊名称	河北中医杂志
7	Year	年	1998
8	Volume	卷	20 (3)
9	Page	起止页	175—176

表 2 化学成分数据示例

Table 2 Example of chemical constituents data

序号	属性名称	属性说明	范例
1	LiuID	流水号	1
2	SRN	化学成分的 SRN 号	5339
3	CASRN	化学成分的 CASRN 号	100-09-4
4	CompName	化合物名称	对-甲氧基苯甲酸
5	SMILES	化学结构 SMILES 数据	COc1ccc(cc1)C(O)=O

表 3 化合物晶体信息示例

Table 3 Example of compound crystal information

序号	属性名称	属性说明	范例
1	LiuID	流水号	533
2	SRN	化合物 SRN 号	12977
3	CIF_ID	CIF 数据 ID 号	CDSioc58730824
4	Journal	来源期刊名	Acta Crystallogr.
5	Year	年	1962
6	Volume	卷	15
7	Page	起始页码	1244

表 4 化合物晶体结构数据示例

Table 4 Compound crystal structure data example

序号	属性名称	属性说明	范例
1	LiuID	流水号	289
2	CIF_ID	CIF 数据 ID 号	CDSioc58730824
3	CIF	CIF 数据	<pre> #===== # Crystal Database of SIOC #===== #===== data_CDSioc_58730824 _audit_creation_date '2021-02-01' _audit_creation_method 'by CrystalStructure v3.5.1' _database_code_CDSioc '58730824' _audit_update_record '?' _chemical_formula_sum 'C5 H10 O2' _chemical_formula_moiety ; C5 H10 O2 ; _journal_coeditor_code "IUCr A03686" _journal_date_recd_electronic '?' _journal_date_from_coeditor '?' _journal_date_accepted '?' _journal_coeditor_code '?' _journal_volume 15 _journal_year 1962 _journal_page_first 1244 _journal_name_full 'Acta Crystallogr. ' loop_ </pre>

序号	属性名称	属性说明	范例
			_publ_author_name "R.F.Scheuerman" "R.L.Sass" _chemical_name_systematic ; Valeric acid ; _cell_volume 595.379 _diffrn_ambient_temperature 138.15 _refine_ls_R_factor_gt 0.11 _refine_ls_wR_factor_gt 0.11 _symmetry_cell_setting monoclinic _symmetry_space_group_name_H-M 'P 21/c' _symmetry_Int_Tables_number 14 loop_ _symmetry_equiv_pos_site_id _symmetry_equiv_pos_as_xyz 1 x,y,z 2 -x,1/2+y,1/2-z 3 -x,-y,-z 4 x,-1/2-y,-1/2+z _cell_length_a 5.55 _cell_length_b 9.664 _cell_length_c 11.341 _cell_angle_alpha 90 _cell_angle_beta 101.82 _cell_angle_gamma 90 _cell_formula_units_Z 4 loop_ _atom_type_symbol _atom_type_radius_bond C 0.68 O 0.68 loop_ _atom_site_label _atom_site_type_symbol _atom_site_fract_x _atom_site_fract_y _atom_site_fract_z C1 C -0.77920 0.41120 0.40650 C2 C -0.61380 0.34000 0.33240 C3 C -0.48260 0.44140 0.26280 C4 C -0.33930 0.35820 0.18050

序号	属性名称	属性说明	范例
			C5 C -0.18040 0.45530 0.11360
			O1 O -0.90220 0.32800 0.46890
			O2 O -0.80430 0.54030 0.41460

3 数据质量控制和评估

为保证录入中药材名称数据的正确性,根据《中国药典》^[10]收录的中药材数据建立了中药材名称表,并整合到数据加工系统中。当输入人员录入中药材数据时,加工系统会依据中药材名称表实时检查录入数据的正确性,并及时反馈给输入人员,供后者检查、改正。如果录入员检查后标记数据为生僻中药材名称,则经专家检查、确认正确后,将其入库,并补充加入中药材名称表。通过这种方式,在保证加工系统灵活性的基础上,有效保证了录入中药材数据的正确率。

CASRN (CAS Registry Number) 号码是一种化学物质的唯一数字标识符,由美国化学文摘社(CAS)发布。CASRN 号码广泛应用于化学化工领域,化学文献中往往会提供化合物的 CASRN 号码,作为目标化合物的标识。根据 CAS 发布的 CASRN 号码规则说明^[11],CASRN 号最长是 10 位数字,用连字符“-”分为从左至右的三个部分,第一部分长度可变,有 2–7 位数字,第二部分是 2 位数字,第三部分是 1 位数字,作为整个 CASRN 号的校验码。CAS 发布了校验码的标准计算方法,可以用于校验各种期刊文献中 CASRN 号码的有效性。根据 CASRN 号有效性标准校验方法,数据加工系统中设计开发了程序算法以检查录入 CASRN 数据的有效性。具体步骤是:

1. 录入员输入 CASRN 号码后,数据加工系统先查询数据集,如果数据集中已有此号码,则返回正确标识,结束本流程;如果没有,则以连字符“-”拆分输入的 CASRN 数据;
2. 拆分操作将 CASRN 数据分为三部分:从左起的 2–7 位数字是第一部分,接下来的 2 位数字是第二部分,最后 1 位数字是第三部分。若解析失败则返回错误标识,结束本流程;
3. 将拆分得到的一、二部分字符串合并,作为被校验数,输入 CASRN 号有效性验证公式进行计算,并比较计算结果与第三部分是否相同。如果不同,则返回错误标识;如果相同,则返回正确标识。

通过以上 CASRN 数据校验方法,可有效保证录入 CASRN 号码的正确性。

中药材化学成分的晶体结构数据加工系统在开发完成后,进行了系统录入的正确性测试,具体测试方法是:首先,从收集的中药材期刊文献库中随机选取 100 篇文献,由专家人工标引,并将选取的文献导入数据加工系统中。然后,由录入人员通过数据加工系统标引同一批文献。最后,邀请一名专业领域的专家(高级工程师)检查两种标引结果。检查结果如表 5 所示,表中的错误包括不符合录入规范或者录入数据不正确,比如,中药材名称数据存在录入名称数据不完整、名称前后误录入不相关字词、名称不是拉丁名称等错误。采用数据加工系统标引数据,其各项错误率都显著低于人工标引方式。从表中数据可见,设计开发的中药材化学成分的晶体结构数据加工系统,能够显著降低数据录入操作的错误率。因此,数据加工系统的采用,使数据集的建设质量有了一定的保证。

表 5 中药材化学成分数据录入质量评估结果

Table 5 Quality evaluation results of data entry of chemical constituents in traditional Chinese medicines

数据生产方式	中药材数据 录入错误率	化合物名称数据 录入错误率	CASRN 号码 录入错误率
人工标引	6%	3.0%	0.8%
数据加工系统	0%	0.5%	0%

4 数据使用方法和建议

中药材化学成分的晶体结构数据集共收录了 500 余种常见中药材的化学成分数据，主要包括化合物的中文名称、拉丁名称、化学结构 SMILES 数据和晶体结构 CIF 数据。为了增加数据集的开放性和应用价值，数据集还收录了化合物的 SRN 号和 CASRN 号数据，使用者可以方便地通过 SRN 号或 CASRN 号将此数据集与其他中医药类或化学类数据集进行数据整合，比如化合物毒性数据集、化合物谱图数据集、中药材配方与临床应用数据集等，从而扩大本数据集的应用范围与领域，增大数据集的挖掘价值。

中药材化学成分的晶体结构数据集采用 MDB 文件格式存储，使用者可以通过 Microsoft® Office Access® 软件操作数据集文件，实现对数据集数据的浏览、检索、筛选和聚类分析等操作。数据集还提供了两个检索窗体：中药材检索窗体和化学成分检索窗体。在中药材检索窗体中输入中药材名称，可以检索得到中药材包含的所有化学成分数据，中药材检索如图 5 所示；在化学成分检索窗体中输入化学成分 SRN 号，可以检索得到包含此化合物的所有中药材，化学成分检索如图 6 所示。SRN 号检索功能可识别、筛选中药材的非特征化学成分。使用者还可以按照本数据集的数据表格式，补充完善新的中药材化学成分数据和化合物晶体结构数据，以扩充、完善本数据集的数据量。

中药材化学成分的晶体结构数据集药材名称检索系统

请输入中药材中文名:

流水号	中药材ID号	中药材中文名	中药材拉丁名
149	138	木通	Caulis Akebiae

SRN	CASRN	中药成分来源文献	年	卷	页码
1063371	87-89-8	安徽中医临床杂志	1995	7 (4)	43-44

化合物名称	化学结构SMILES
肌醇	O[C@H]1[C@H](O)[C@H](O)[C@H](O)[C@H](O)[C@H]1O

CIF文件ID号	CIF来源文献	年	卷	起始页码
CDSioc56574346	Acta Crystallogr., Sect. E: Struct.	2007	63	o530

CIF数据

```

_audit_creation_method 'by CrystalStructure v3.5.1'
_database_code_CDSioc '56574346'
_audit_update_record '?'
_database_code_depnum_ccdc_archive 'CCDC 636625'
_chemical_formula_sum 'C6 H12 O6'
_chemical_formula_moiety
:
C6 H12 O6
:
_journal_coeditor_code "IUCr FL2082"
_journal_date_recd_electronic '?'
    
```

记录: 第 1 项 (共 2846) 搜索

图 5 中药材化学成分的晶体结构数据集药物检索示例

Figure 5 Example of drug retrieval from the crystal structure dataset of chemical constituents in traditional Chinese medicines

中药材化学成分的晶体结构数据集化学成分检索系统

请输入化合物SRN号: 71527311 检索

流水号	中药材ID号	中药材中文名	中药材拉丁名
1	16	旋覆花	Inula britannica L.

SRN	CASRN	中药成分来源文献	年	卷	页码
71527311	33620-72-3	安徽中医临床杂志	1998	10 (4)	233

化合物名称: 去乙酰旋覆花次内酯 化学结构SMILES: CC1C[C@H]2OC(=O)C(=C)[C@H]2[C@H](O)C=1[C@H](C)CCCC

CIF文件ID号	CIF来源文献	年	卷	起始页码
CDSioc53326970	Chem.-Eur. J.	2011	17	3332

CIF数据

```
_audit_creation_method 'by CrystalStructure v3.5.1'
_database_code_CDSioc '53326970'
_audit_update_record '?'
_database_code_depnum_ccdc_archive 'CCDC 794896'
_chemical_formula_sum 'C15 H22 O4'
_chemical_formula_moiety
:
C15 H22 O4
:
_journal_date_recd_electronic '?'
_journal_date_from_coeditor '?'
```

记录: 1 第 1 项(共 467 项) 大端清洗器 搜索

图 6 中药材化学成分的晶体结构数据集化学成分检索示例

Figure 6 Example of compound retrieval in the crystal structure dataset of chemical constituents in traditional Chinese medicines

数据作者分工职责

李英勇 (1978—), 男, 河南省南阳市人, 硕士, 高级工程师, 研究方向为化学信息学。主要承担数据库设计和数据库建库工作。

参考文献

- [1] 徐挺军. 植物化学成分数据加工系统的研究与设计[D]. 上海: 上海交通大学, 2016. [XU T J. Research and design of plant Chemical constituents data processing system [D]. Shanghai: Shanghai Jiaotong University, 2016.]
- [2] 王庆利, 胡晓敏, 冯毅, 新药研发需要关注的几个问题[J]. 中国处方药, 2009, 7(2): 46-47. [WANG Q L, HU X M, FENG Y. Some hotspot questions in new drug R&D[J]. Journal of China Prescription Drug, 2009, 7(2): 46-47.]
- [3] 赵丽梅, 谭宁华, 国外天然产物化学成分实物库及数据库建设概况[J]. 中国中药杂志, 2015, 40(1): 29-35. DOI:10.4268/cjcmm20150106. [ZHAO L M, TAN N H. Status of libraries and databases for natural products at abroad[J]. China Journal of Chinese Materia Medica, 2015, 40(1): 29-35. DOI:10.4268/cjcmm20150106.]
- [4] 王恩宇, 王景辉, 丁芮. “黄连-半夏”药对治疗胃癌的网络药理学研究和实验验证[J]. 中国中药杂志, 2020, 45(8): 1779-1788. DOI: 10.19540/j.cnki.cjcmm.20200120.401. [WANG E Y, WANG J

- H, DING R, et al. Network pharmacology research and experimental validation of “Coptidis Rhizoma-Pinelliae Rhizoma” in treating gastric carcinoma[J]. China Journal of Chinese Materia Medica, 2020, 45(8): 1779–1788. DOI:10.19540/j.cnki.cjcmm.20200120.401.]
- [5] Li G, Sun YL, Sng KS, et al. Exploring the mechanism of Shenqisherong pill against cervical spondylotic myelopathy by network pharmacology and molecular docking[J]. Ann Palliat Med, 2021, 10(10): 10253 – 10275. DOI: 10.21037/apm-21-408.
- [6] HALL S R, ALLEN F H, BROWN I D. The crystallographic information file (CIF): a new standard archive file for crystallography[J]. Acta Crystallographica Section A Foundations of Crystallography, 1991, 47(6): 655 – 685. DOI:10.1107/s010876739101067x.
- [7] 中国科学院上海有机化学研究所. 化合物名称到结构的转换[EB/OL]. (2012-05-03) [2021-09-03]. http://www.organchem.csdb.cn/scdb/str/name_change2.asp. [Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences. Conversion from Compound name to structure[EB/OL]. (2012-05-03) [2021-09-03]. http://www.organchem.csdb.cn/scdb/str/name_change2.asp.]
- [8] WEININGER D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules[J]. Journal of Chemical Information and Modeling, 1988, 28(1): 31–36. DOI:10.1021/ci00057a005.
- [9] 中国科学院上海有机化学研究所. 化合物结构数据库[EB/OL]. (2012-06-03) [2021-09-03]. http://www.organchem.csdb.cn/scdb/main/str_introduce.asp. [Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences. Compound structure database. [EB/OL]. (2012-06-03) [2021-09-03]. http://www.organchem.csdb.cn/scdb/main/str_introduce.asp.]
- [10] 国家药典委员会. 中华人民共和国药典[M]. 北京: 中国医药科技出版社, 2015. [State Pharmacopoeia Committee of China. pharmacopoeia of the people's republic of china[M]. BeiJing:China medical science press, 2015.]
- [11] American Chemical Society. Check Digit Verification of CAS Registry Numbers[EB/OL]. [2021-09-03]. <http://www.cas.org/content/chemical-substances/checkdig>.

论文引用格式

李英勇. 中药材化学成分的晶体结构数据集[J/OL]. 中国科学数据, 2022, 7(1). (2021-12-06). DOI: 10.11922/11-6035.csd.2021.0052.zh.

数据引用格式

李英勇. 中药材化学成分的晶体结构数据集[DS/OL]. Science Data Bank, 2021. (2021-08-23). DOI: 10.11922/sciencedb.01090.

A dataset of crystal structure of chemical constituents in traditional Chinese medicines

LI Yingyong^{1,2*}

1. Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, P.R. China

2. National Basic Public Science Data Center, Beijing 100190, P.R. China

* Email: liyingyong@sioc.ac.cn

Abstract: Based on the collection of the chemical composition data and compound crystalline structure data in traditional Chinese medicines from the Chinese and foreign journals of 1945–2015, I adopted a variety of approaches, such as the data entry regulation and real-time data processing system software inspection, for data quality control and standardized sorting of the chemical composition data and compound crystal structure data of traditional Chinese medicine. These two kinds of data were integrated in accordance with different chemical structures. This dataset collects the crystal CIF data of 527 common traditional Chinese materials, 270 Chinese medicinal chemical compounds, and 345 compounds, and presents the corresponding relation and source information among these three kinds of data. This dataset can serve as basic scientific research data for medical chemists, pharmacists and other scientific research personnel in pharmaceutical research, such as the modernization research on traditional Chinese medicine and the research and development of new drugs.

Keywords: traditional Chinese medicine; chemical constituent; chemical structure; crystal structure

Dataset Profile

Title	A dataset of crystal structure of chemical constituents in traditional Chinese medicines
Data corresponding author	LI Yingyong (liyingyong@sioc.ac.cn)
Data author	LI Yingyong
Time range	1945–2015
Geographical scope	Worldwide
Data volume	102 MB
Data format	*.mdb
Data service system	< https://dx.doi.org/10.11922/sciencedb.01090 >
Source of funding	The 14th Five-year Network Security and Informatization Plan of Chinese Academy of Sciences (WX145XQ07-12).
Dataset composition	The dataset consists of one Microsoft Office ACCESS® file, composed of four data tables (TcmCmpd table, Compound table, CmpCIF table, and CIF table) and two retrieval forms (TCM retrieval form and chemical constituent retrieval form). The dataset collects the crystal CIF data of 527 common traditional Chinese materials, 270 Chinese medicinal chemical compounds, and 345 compounds, including TCM's Chinese name, Latin name and chemical constituent, compound's SRN number, CASRN number and SMILES data, crystal structure's CIF data, and source literature data.