Published online 28 November 2016

Nucleic Acids Research, 2017, Vol. 45, Database issue ***D955–D963***
doi: 10.1093/nar/gkw1118

PubChem 生物测定：2017 年更新

Yanli Wang*, Stephen H. Bryant*, Tiejun Cheng, Jiyao Wang, Asta Gindulyte, Benjamin A. Shoemaker, Paul A. Thiessen, Siqian He and Jian Zhang

美国马里兰州贝塞斯达市 20894，美国国立医学图书馆，美国国立卫生研究院，国家生物技术信息中心

收稿日期：2016 年 9 月 29 日；　修订于 2016 年 10 月 26 日；编辑决策于 2016 年 10 月 27 日；接受于 2016 年 11 月 9 日

摘要

自 2004 年以来，PubChem 的生物测定数据库（https://pubchem.ncbi.nlm.nih.gov）一直作为小分子和 RNAi 筛选数据的公共存储库，向公众开放其数据内容。PubChem 接受来自世界各地学术界、工业界和政府机构研究人员的数据提交。PubChem 还与其他化学生物学数据库利益相关者合作进行数据交换。经过十多年的开发努力，它已成为支持药物发现和化学生物学研究的重要信息资源。为了便于数据发现，PubChem 与 NCBI 的所有其他数据库进行了整合。在这项工作中，我们提供了 PubChem 生物测定数据库的最新情况，描述了包括新增的研究数据来源、重新设计的生物测定记录页面、新的生物测定分类浏览器以及上传系统中的新功能（以促进数据共享）在内的几项近期发展。

介绍

PubChem 生物测定数据库（1-4）是由美国国立医学图书馆（NLM）下属的美国国家生物技术信息中心（NCBI）托管的一个开放获取数据库。该数据库始于 2004 年，旨在作为化学基因组学、药物化学和功能基因组学研究中生成的信息的公共存储库。数据库中的所有数据均可免费供公众搜索和下载。近期关于科学界对 PubChem 资源利用情况的综述（5-7）强调，PubChem 生物测定数据库中收集的生物活性和毒性数据极大地支持了药物化学、药物发现、药物基因组学和信息学研究等多个领域的工作。PubChem 生物测定数据库中的小分子数据通过测定中引用的样本与化学结构相互关联。PubChem 生物测定数据库还与 NCBI 托管的其他生物医学和文献数据库（如 PubMed、蛋白质、基因、分类学等）相连接。数据库中的元数据与 NCBI 的搜索引擎 Entrez 集成，使得 PubChem 生物测定数据库可以通过网络界面进行交互式关键词搜索，也可以通过 E-Utilities 进行程序化检索。还可以通过 PubChem 提供的基于网络和程序化的工具来检索和分析测定数据。表 1 提供了用于访问、搜索、下载和分析 PubChem 生物测定数据的服务及其 URL 的更新信息。大多数基于网络的服务也可以在 https://pubchem.ncbi.nlm.nih.gov/assay/ 访问。

在过去的 12 年中，PubChem BioAssay 数据库不断朝着支持开放数据的方向发展，致力于满足社区对信息存档、检索和挖掘日益增长的需求。PubChem BioAssay 通过以下方式保持其作为药物发现相关研究数据领先存储库的地位：（i）采用优化且灵活的数据模型支持广泛的生物活性信息类型；（ii）持续改进数据库基础设施和可扩展性；（iii）利用新技术进行数据存档、查看、索引、搜索和下载；（iv）改进数据上传系统；（v）与其他生物医学资源进行整合。在这项工作中，我们对信息资源的几个方面进行了更新，包括数据内容和数据来源的增长、数据库基础设施的整合、重新设计并模块化的 BioAssay 记录页面、新的 BioAssay 分类浏览器以及为先前提供的网络服务添加的新功能。还介绍了 PubChem 上传系统对实验数据的禁运、发布和搁置数据共享的增强管理。

生物测定数据

PubChem 生物测定数据库目前包含超过 100 万条记录，其中包含由全球 80 多个组织（数据源）提交的 2.3 亿个生物活性结果。2004 - 2013 年和 2014 - 2016 年期间的数据内容见表 2。高通量筛选（HTS）数据由来自学术机构、大学、政府组织以及制药公司的实验室和筛选中心提供。

* To whom correspondence should be addressed. Tel: +1 301 435 7811; Email: ywang@ncbi. nlm. nih. gov Correspondence may also be addressed to Stephen H. Bryant. Tel: +1 301 435 7792; Email: bryant@ncbi.nlm.nih.gov

表 1. PubChem 生物测定服务列表

服务	描述	URL 示例
生物测定记录页面	访问并下载生物测定记录	https://pubchem.ncbi.nlm.nih.gov/bioassay/805
生物测定搜索	使用 Entrez 搜索生物测定数据库	https://www.ncbi.nlm.nih.gov/pcassay/
生物测定搜索，高级页面	一个用于在多个搜索字段中进行搜索的界面 一个用于查看搜索历史并使用布尔运算来优化搜索结果的界面	https://www.ncbi.nlm.nih.gov/pcassay/limits https://www.ncbi.nlm.nih.gov/pcassay/advanced
PubChem 上传	物质与生物测定提交系统	https://pubchem.ncbi.nlm.nih.gov/upload/
生物测定 FTP	所有 PubChem 生物测定记录及相关信息的 FTP 服务	ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/
生物测定数据标准	PubChem 生物测定数据模型的 XML 数据规范	ftp://ftp.ncbi.nlm.nih.gov/pubchem/data规范/
生物测定服务主页	生物测定服务主页	https://pubchem.ncbi.nlm.nih.gov/assay/
生物测定分类	浏览生物测定分类树	https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi? p=classification该链接指向的是美国国家生物技术信息中心（NCBI）的 PubChem 数据库中的一个分类测定页面。
生物活性数据工具	从单个生物测定记录中检索完整的数据表	https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.html?aid=1811
	检索并下载单个物质样本（SID）、化学结构（CID）、蛋白质靶点（GI、UniProt 或 GenBank 登录号）、基因靶点（GeneID）或文献（PMID）的跨实验生物活性数据。	https://pubchem.ncbi.nlm.nih.gov/assay/生物活性.html?sid=103164874 https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.html?cid=2244https://pubchem. ncbi. nlm. nih. gov/assay/bioactivity. html?gi=29725609https://pubchem. ncbi. nlm. nih. gov/assay/bioactivity. html?uniprot=P00533https://pubchem. ncbi. nlm. nih. gov/assay/bioactivity. html?ncbiacc=NP_005219https://pubchem. ncbi. nlm. nih. gov/assay/bioactivity. html?geneid=1956https://pubchem. ncbi. nlm. nih. gov/assay/bioactivity. html?pmid=25728019
生物测定下载工具	灵活的下载界面	https://pubchem.ncbi.nlm.nih.gov/assay/assaydownload.cgi
PubChem哈巴狗/ REST / SOAP	用于数据检索的编程工具和 REST API	https://pubchem.ncbi.nlm.nih.gov/pug rest/ PUG rest .html https: / / pubchem.ncbi.nlm.nih.gov /哈巴狗/ pughelp.html
PubChem 小工具帮助	PubChem 小部件可让您在页面中显示 PubChem 数据。	https://pubchem.ncbi.nlm.nih.gov/widget/docs/widget帮助.html
构效关系分析（SAR）	利用聚类工具和热图样式的展示来分析和可视化构效关系	https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p =热量
剂量 - 反应曲线工具	分析生物测定测试结果并绘制剂量-反应曲线	https://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype = 1
散点图/直方图	用直方图或散点图分析生物测定测试结果	https://pubchem.ncbi.nlm.nih.gov/assay/plot.cgi?plottype = 2
相关生物测定法	通过以下方面总结生物测定关系：相同的测定项目、活性化合物的重叠、活性基因的重叠、目标序列相似性、已存注释、相同的发表文献以及基因相互作用。	https://pubchem. ncbi. nlm. nih. gov/bioassay/1510#section=Same-Project-BioAssays https://pubchem. ncbi. nlm. nih. gov/bioassay/1510#section=Related-BioAssays
生物活性概要 - 以化合物为中心	从化合物的角度对一组记录中的生物活性数据进行总结和分析	https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi?tab=1
生物活性概要 - 以检测方法为中心	从实验检测的角度总结并分析一组记录的生物活性数据。	https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi?tab=2
生物活性概要 - 以靶点为中心	从目标角度对一组记录的生物活性数据进行总结和分析	https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.cgi?tab=3

该数据库还收录了包括参与美国国立卫生研究院分子库计划资助的化学探针开发高通量筛选（HTS）活动的公司在内的多家公司的数据（8,9）。此外，该数据库还包含基于文献的数据，包括由期刊作者提交的出版物驱动的沉积数据以及由 PDBbind（10）、IUPHAR（11）、BindingDB（12）、ChEMBL（13）和其他文献整理项目提供的数据。另外，还收集了针对 PubChem 中的测定数据采用社区认可的词汇和本体（13-15）进行的第三方注释，例如测定格式、测定类型、检测方法和细胞系，并将其与相关数据集关联起来，这些信息显示在 BioAssay 记录页面上，并用于在 Entrez 和数据分析工具中搜索和筛选测定数据。

近三分之一的测定数据源是在过去三年中添加的。这些数据源中的大多数在向支持功能基因组学研究开放获取的期刊提交手稿的同时，将 RNAi 数据存入了 PubChem。《自然·细胞生物学》杂志率先倡导将 RNAi 数据存入公共数据库，最近还开始呼吁共享小分子数据。其他开放获取期刊，如《公共科学图书馆·综合》，也加入了这一行列，推荐通过公共数据库共享数据，并带来了几组小分子数据集存入 PubChem（16,17）。这些来源提供的测定数据与 PubMed 中索引的相应出版物相关联（16-36），这使得 PubChem 用户能够访问文章以获取更多信息，反之亦然，PubMed 用户也能访问支持机器可读格式的 BioAssay 存档中的研究数据。所有 BioAssay 提交者及其相关信息（如所属机构、数据提交摘要）均可在 PubChem 数据源页面查看：https://pubchem.ncbi.nlm.nih.gov/sources/，该页面将数据源按地理位置和其他各类别进行分组。用户可以通过“按类型划分的数据统计”字段中呈现的物质或 BioAssay 记录数量来访问特定提交者的提交内容。

BioAssay 数据库的显著增长需要一个强大且可扩展的数据库系统。为此，建立了一组关系数据库和表，以实现以下功能：（一）存档生物测定提交内容，跟踪更新并提供版本控制；（二）维护数据的保密和发布状态；（三）记录并推导生物测定与其他生物医学信息之间的链接和关系；（四）存储第三方注释并链接到相应的数据集；（五）提供搜索索引；（六）支持用于网页展示、REST API 和数据分析工具的数据检索；（七）为 BioAssay FTP 的日常更新提供便利。为增强数据库基础设施，投入了大量精力。还实施了额外机制来跟踪有关 BioAssay 蛋白质和基因靶点的信息。创建了 NCBI 蛋白质 GI 号、GenBank 登录号和 UniProt ID 之间的映射，以方便数据检索和整合。这些努力旨在促进：（一）生物测定提交；（二）整合已提交的小分子和 RNAi 数据；（三）整合来自公共生物医学数据库的蛋白质和基因的生物注释。（四）利用非 NCBI 序列标识符访问 PubChem 中的生物活性数据；（五）开发新的 PubChem 服务以增强生物测定靶点搜索功能；（六）提高数据库中生物数据的可发现性。

新网络功能和服务

PubChem 生物测定提供了基于网络和程序化的工具，用于数据搜索、访问、分析和下载。这里介绍了几个最近开发的网络服务，通过分类提交的元数据和第三方注释来改进测定数据的搜索和导航。

生物测定记录页面

PubChem 提供对每个已存入的生物测定记录的全面访问权限。PubChem 生物测定记录页面取代了旧版的摘要页面，经过重新设计，旨在优化数据流，支持数据和服务的再利用，并统一 PubChem 资源中的网页呈现形式。借助新的网络技术，该数据驱动型界面针对触摸设备和鼠标设备进行了设计和优化，与近期对 PubChem 化合物摘要页面和物质记录页面的改版采用了类似的方法（37）。该网页采用响应式设计，由多个小部件组成，可自动适应可用的屏幕尺寸，方便用户在台式机、平板电脑和手机上浏览页面内容和查看信息。此外，新设计还提供了将页面的任何部分或子部分嵌入到其他网页作为小部件的功能，无需单独的代码库，从而减轻了第三方的维护负担，这对有意整合 PubChem 生物测定数据的非 PubChem 资源来说大有裨益。有关嵌入 PubChem 小部件的信息和说明，请访问 https://pubchem.ncbi.nlm.nih.gov/widget/docs/widget help.html 。

通过 AID 号（主要访问号）可以访问已提交的生物测定记录。图 1 展示了一个用于鉴定鞘氨醇 1-磷酸受体 4 拮抗剂的数据集示例（AID：1510；https://pubchem. ncbi.nlm.nih.gov/bioassay/1510）。生物测定记录页面提供了对初始提交和所有后续更新的版本控制的主要访问权限。它还提供了第三方注释以及支持数据分析和下载的工具链接。顶部的“下载”按钮允许用户下载提交者提供的元数据和测定结果，以及测试的小分子样品的化学结构。目录可以展开，以便快速导航到每个单独的部分。默认情况下，在数据表部分检索完整数据集。此外，测定数据表可以根据活性结果（例如，活性、无活性或具有微摩尔活性或纳摩尔活性的子集）进行分区，使用户能够快速筛选、选择和下载感兴趣的结果。为了便于命中评估、数据对比和目标识别，小分子样本的结构图像链接到特定的生物活性分析工具，该工具通过 CID（PubChem 化合物唯一化学结构的访问编号）展示化合物的所有跨实验数据。同样，对于 RNAi 实验（例如 https://pubchem.ncbi.nlm.nih.gov/ ），也遵循类似的链接方式。

表 2. PubChem 生物测定统计数据

	总计	化学分析		RNAi 检测	
		2004 至 2013 年	2014 年至今	2004 至 2013 年	2014 年至今
分析记录（AID）	一百二十一万八千六百八十七	七十三万七千九百九十四	四十八万零六百一十六	57	48
物质样本（SID）	三百五十七万六百零六十六	二千七百五十五万零三百二十三	百三十九万六千六百九十三	二一三零三零	二十九万三千四百九十九
化学结构（CID）	二百二十八万三千五百三十三	一百九十五万六千九百九十八	九百八十六万二千三百七十五	-	-
生物活性结果	二百三十一 三百零三 六百零七	222 198 148（由于原文仅包含数字，无法直接翻译为中文数字）		一百零一九百九十三	四百一十九万零八百一十
数据点	一千五百一十四亿二千二百三十	一亿零四百五十一万零三十二	九百四十四万九千九百九十九	七百四十七万三千四百六十五	
物种	3543	2730	1895	6	2
蛋白质靶点	一万零六百三十六	7450	6972	-	-
蛋白质靶点（人类）	4771	3378	3495	-	-
基因靶点	五万五千七百一十四	-	-	三万八千六百九十四	五万二千九百八十六
基因靶点（人类）	二万四千八百八十八	-	-	二十四万四千六百	二万二千六百五十六
基因靶点（表型）	一万五千八百六十六	-	-	一万二千八百一十六	4524

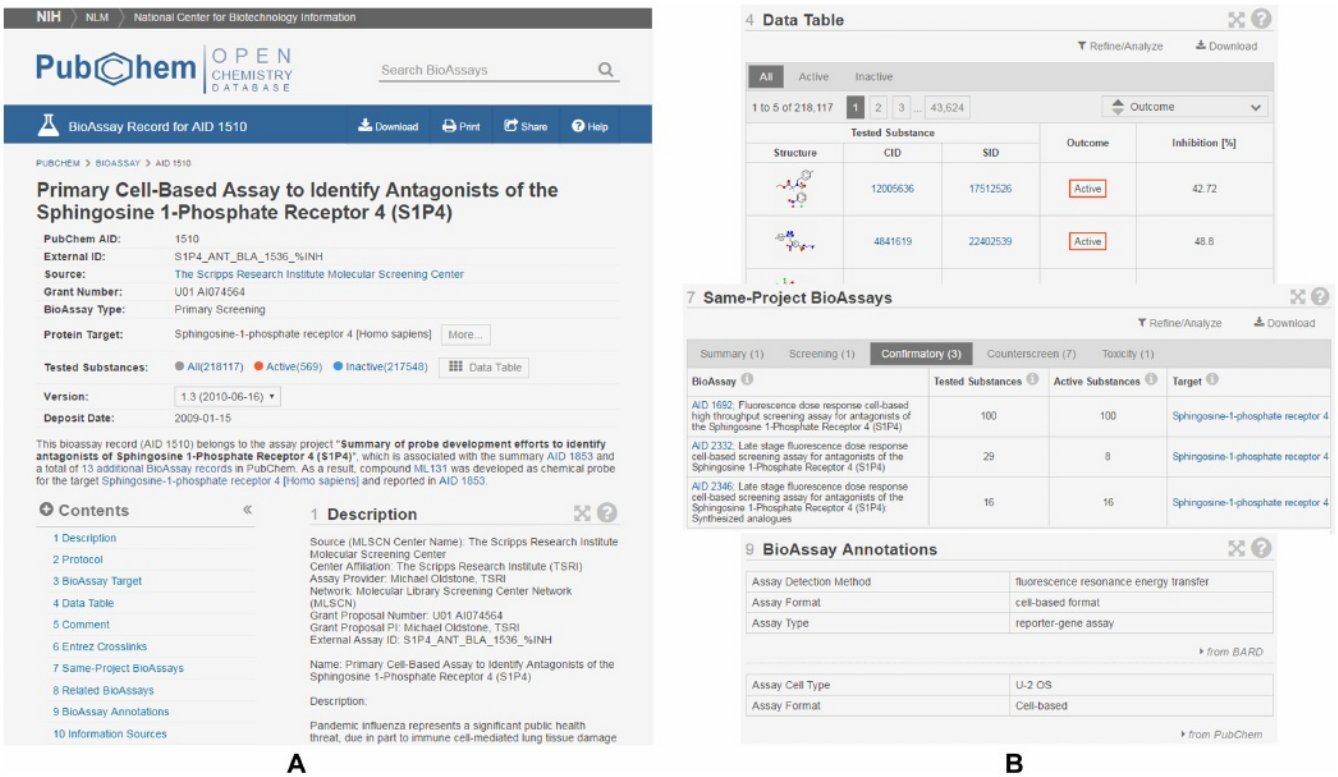

图 1.生物测定记录（AID 1510，https://pubchem.ncbi.nlm.nih.gov/bioassay/1510）。(A) 记录页面概览。目录提供了页面上各部分的快速导航。每个部分都有一个锚点，其 URL 可用于小部件嵌入。(B) 选定的部分：数据表、同项目生物测定和生物测定注释。

（例如：https://www.ncbi.nlm.nih.gov/bioassay/1904），基因目标列中的基因 ID 会链接到一个特定的生物活性分析工具，该工具汇总了所有 RNAi 数据以及针对该基因的小分子，从而有助于发现小分子工具，将结果与其他研究进行比较，并识别其他 RNAi 筛选所提示的生物学功能。在展示提交者提交的信息之后，页面会呈现来自同一项目或多个项目的相关 BioAssay 数据集。提交者提供的相关测定数据集的展示有助于追踪测定项目的进展，并促进数据验证和解释，当与 PubChem 基于靶点和出版物得出的相关测定数据集结合使用时，还能进行跨测定比较。记录页面顶部的摘录会提示此类相关数据的存在，突出了数据整合对于测定数据解释和再利用的重要性。页面的最后一部分展示第三方注释，并标明来源。

生物测定分类工具

基于 PubChem 分类浏览器的软件框架开发了一个分层树形视图，为浏览、搜索和访问生物测定数据提供了另一种途径。该工具（可在 https://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?p=classification 和 https://pubchem.ncbi.nlm.nih.gov/classification/#hid=80 访问）通过整合生物测定元数据，提供了具有共同关注属性的生物测定数据集概览。

第三方注释、检测目标分类、分类学以及相关出版物。它通过层次结构中的节点对生物测定记录进行组织，这些节点展示了各种分类和注释。此工具允许用户浏览感兴趣的层次结构中生物测定数据的分布情况，汇总特定子类的信息，并通过数据源、检测类型等进行特定搜索（图 2）。层次结构显示中节点所显示的计数在点击时会链接到 Entrez 中索引的检测数据条目。从分类浏览器发送的单个检测列表可以使用 Entrez 的查询细化功能进行组合，为用户提供了强大的手段，能够根据脑海中多个搜索条件钻取到所需的数据集。作为使用示例，用户可以钻取到“活性类型/效力”下的 IC50 或 Kd 数据，或者钻取到“检测类型”下的基于细胞或基于生化的方法的数据。用户还可以浏览“高通量筛选项目”下的 7000 多个数据集，包括由美国国立卫生研究院分子库计划资助的全基因组 RNAi 筛查和化学探针开发项目。

“出版物”分支允许用户按相关原始引文的期刊名称和出版年份浏览测定数据集。更重要的是，该工具允许用户根据 PubMed 引文的分类，利用医学主题词表（MeSH）提供的生物医学术语受控词汇表，按研究主题在 BioAssay 中搜索数据。如图 2 所示，用户可以沿着“MeSH 树”下的“化学与药物类别”分支检索有关血管紧张素受体拮抗剂生物活性的测定数据条目，这些条目链接到 PubMed 中相关出版物的摘要。对于工具中的“靶点”分支，已整合了四种蛋白质本体论和分类，用于测定靶点，包括 ChEMBL、GO、IUPHAR 和 KEGG。记录生物分类学 ID（BioAssay 数据模型中的关键元数据字段之一）能够按照生物体对 BioAssay 数据集进行组织，从而基于 NCBI 维护的生物分类学分类提供测定数据集的分层显示，而分类工具中的“生物分类学”树类似于 NCBI 生物分类学树的一个子集，其中包含测定数据的生物分类学节点。

生物活性分析工具的新功能

PubChem BioAssay 提供了多种方式来引用蛋白质和基因靶点。一个测定靶点通常与数百个测定中测试的许多小分子样本相关联，因此将用于识别化学工具的信息与适当的选择性评估相结合非常重要。此前已开发出生物活性分析工具，用于动态检索和汇总针对某一蛋白质或基因靶点的跨测定生物活性数据。随着 BioAssay 提交数量的增长，提供选择功能以对信息进行细分和筛选变得至关重要。此外，对于治疗靶点，必须指出主要通过查询靶点开发用于治疗疾病的药物。而且，强调针对查询靶点测试的所有药物分子也很重要，这有助于识别这些药物的替代治疗效果，从而实现药物再定位。通过整合近期获得的测定注释信息，这些工具新增并升级了若干用于筛选生物活性数据的选择功能，从而能够从特定检测方法中检索测定数据，选择毒性结果或从药物分子中筛选生物活性数据（图 3，https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.html?geneid=1956）。作为另一项功能，新增了“选择性”列，以帮助评估测试化学样品对靶点的选择性，该列统计了化学样品所测试的总独特靶点数量以及其对靶点产生活性的数量。除了接受基因 ID 外，此生物活性分析工具还支持使用 GenBank 登录号或 UniProt ID 指定查询靶点，使用 SID 或 CID 指定测试样品（RNAi 试剂或小分子），使用 PMID 指定包含测定数据的文献。该服务还经过优化，支持通过筛选功能以编程方式下载生物活性数据。

公共访问、搜索、下载

PubChem 提供了多种途径来访问、搜索和下载生物测定数据，包括上述的生物测定记录页面和分类工具。PubChem 生物测定数据在 Entrez（https://www.ncbi.nlm.nih.gov/pcassay/）中按多个字段进行了索引，以支持关键词搜索。它还与多个其他生物医学数据库相互链接，例如通过测定提交时提供的引文与 PubMed 相连，或者通过测定目标规范与 NCBI 基因数据库相连。因此，Entrez 中的基因组信息用户可以检索到与基因靶点相关的 PubChem 生物测试结果，而 PubMed 用户可以前往生物测定数据库检索某篇出版物中讨论的数据。

生物测定数据可通过以下方式下载：（i）在生物测定记录页面使用下载功能，支持 ASN、XML、JSON 和 CSV 格式；（ii）通过 https://pubchem.ncbi.nlm.nih.gov/assay/assaydownload.cgi 提供的基于网络的大批量下载服务，可输入 AID 列表，也可选择性输入 SID（PubChem 物质访问号）列表；（iii）由 NCBI 的 E-Utilities（https://www.ncbi.nlm.nih.gov/books/NBK25497/）、PubChem PUG/SOAP（https://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html） 和 PUG/REST（https://pubchem.ncbi.nlm.nih.gov/pug rest/PUG REST.html）（38）提供的编程工具，这些工具为获取特定 AID 的元数据、测定结果、数据库链接以及化合物或靶点在多个测定中的生物活性数据提供了极大的灵活性；（iv）每日更新的 PubChem 生物测定 FTP 服务器 ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/，主要通过 ASN、XML、JSON 和 CSV 格式提供对所有已存档生物测定记录的开放访问。在生物测定 FTP 的“Extras”目录下新增的信息包括：（i）文件“Aid2Annotation”，提供第三方注释，采用标签/值结构；（ii）包含 AID 以及蛋白质 GI、GenBank 登录号、UniProt ID 和基因 ID 之间映射关系的文件“Aid2GiGeneidAccessionUniprot”；（iii）一个名为“VendorCatalogs”的子文件夹，其中包含多个 RNAi 产品供应商的文件，这些文件中记录了供应商目录 ID 与为 RNAi 样品分配的 PubChem SID 之间的映射关系。

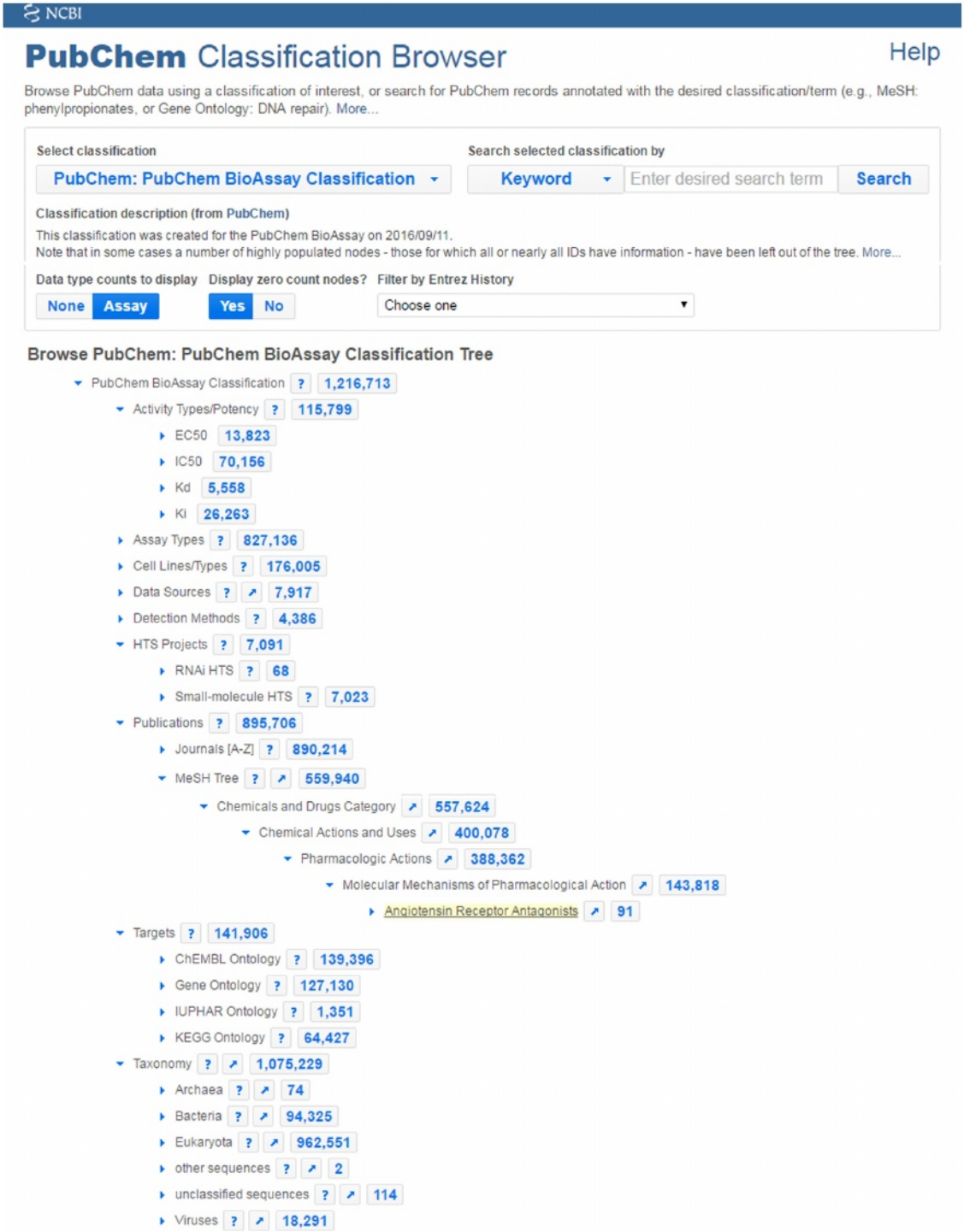

图 2. PubChem 生物测定分类树（https://pubchem.ncbi.nlm.nih.gov/classification/#hid=80）。提供了分层显示，可通过点击三角形图标展开子树进行浏览和探索。点击节点上的数字（显示具有该注释的生物测定记录数量）可进入 Entrez 查看相关测定记录的报告。

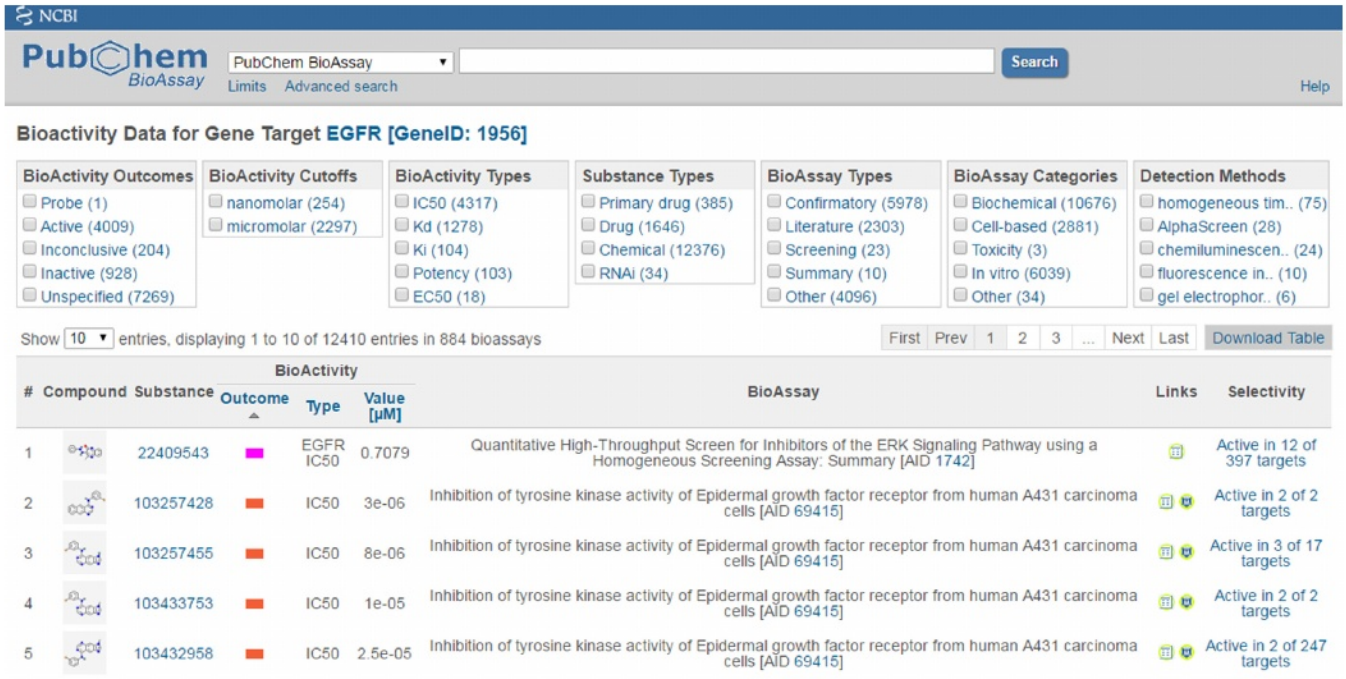

图 3.在 PubChem 生物活性数据库（https://pubchem.ncbi.nlm.nih.gov/assay/bioactivity.html?geneid=1956）中，针对特定基因靶点的集体和跨实验生物活性数据。网页顶部的筛选器可用于缩小到感兴趣的子集。例如，通过点击“物质类型”部分提供的筛选器中的“RNAi（34）”，可以检索到 RNAi 数据。该部分中的“主要药物（385）”筛选器可检索针对查询蛋白/基因专门开发的药物的生物活性数据，而“药物（1646）”筛选器则可检索在实验中测试的任何药物的生物活性数据。后一个筛选器可帮助识别有实验证据（基于 PubChem 生物活性数据）表明与查询蛋白/基因靶点结合或对其产生影响的药物分子，从而进一步探索其药物再定位（针对查询蛋白/基因靶点）的潜力。支持这两个筛选器的药物和靶点信息来自 DrugBank 的注释。

PubChem 用于生物测定提交的上传操作

作为接受复杂研究数据提交的公共存储库，一个强大且用户友好的数据提交系统起着关键作用。PubChem 上传系统（https://pubchem.ncbi.nlm.nih.gov/upload/）的用户界面一直在不断优化。现在新增了若干功能，用于管理数据的保密、发布以及在提交者维护的用户组之间共享待发布数据。PubChem 允许数据保密，以便提交者有足够的时间发表研究成果或完成专利申请。此前，只有上传账户持有者才能访问处于保密状态的生物测定记录。现在新开发的机制允许通过专门请求的 URL 向合作者、期刊审稿人或编辑提供完整访问权限。上传系统现在支持批量发布处于保密状态的生物测定记录，只需提供生物测定记录的访问编号（AID）列表即可。此外，它还简化了生物测定记录及其相关物质记录的发布流程，一旦提交者请求发布 AID 列表，上传系统就会自动查找并发布相关的待发布物质记录。常见问题解答部分（https://pubchem.ncbi.nlm.nih.常见问题解答页面（https://pubchem.ncbi.nlm.nih.gov/gov/upload/docs/upload faq.html）已添加到上传帮助页面（https://pubchem.ncbi.nlm.nih.gov/upload/docs/upload help.html），为常见问题和更新操作提供快速提示。

摘要

PubChem 于 2004 年启动，最初是一个用于存储小分子和 RNAi 筛选所得生物数据的公共库。截至今日，全球已有 80 多家机构和实验室通过 PubChem BioAssay 分享了研究数据。开发和维护公共数据库面临诸多挑战。相关社区已齐心协力，提出关键思考，以制定理想的数据管理指南。最近，有人提出了 FAIR（可查找性、可访问性、互操作性和可重用性）原则，旨在为公共数据管理、数据流通以及分析工具和流程的共享提供指导。这一努力旨在为公共数据的利益相关者带来清晰的方向，鼓励他们与资助机构、研究人员和出版商携手合作，协调研究数据，最大限度地发挥学术数字出版的价值。

PubChem 生物测定库在回顾性审查时，其设计和开发在很大程度上遵循了 FAIR 原则。生物测定数据模型的设计考虑到了机器可读性，数据库中的所有数据都可免费供社区使用。可通过 NCBI Entrez 系统搜索测定数据：https://www.ncbi.nlm.nih.gov/pcassay/。PubChem 还提供了其他工具以支持数据搜索、访问和分析。社区开发了许多附加服务和工具，以扩展和补充 PubChem 资源的功能，并为 PubChem 中的数据内容提供额外注释（5）。两者之间的相互作用

PubChem 与社区的努力是互利的。PubChem 的信息平台在不断发展，以鼓励对 PubChem 中的化学信息学、化学生物学和功能基因组学研究数据的再利用，并通过社区的努力实现和简化整合。随着技术的进步，PubChem 将继续改进服务和工具，与第三方注释和其他公共生物医学数据集成，并与支持研究数据存档和再利用的资助机构和出版商合作。PubChem 欢迎社区共享资源并为该库做出贡献。

致谢

本研究得到了美国国立卫生研究院（NIH）下属的国家医学图书馆（NLM）内部研究计划的支持。作者感谢所有向 PubChem 提交数据的贡献者。

资金

美国国立卫生研究院（NIH）院内研究计划；美国国立医学图书馆（NLM）。开放获取费用资助方：美国国家生物技术信息中心（NCBI）。

利益冲突声明。无申报。

参考文献

1. Wang,Y., Suzek,T., Zhang,J., Wang,J., He,S., Cheng,T., Shoemaker,B.A., Gindulyte,A. and Bryant,S.H. (2014) PubChem BioAssay: 2014 update. *Nucleic Acids Res.*, **42**, D1075–D1082.
2. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J., Zhou,Z., Han,L., Karapetyan,K., Dracheva,S., Shoemaker,B.A. *et al.* (2012) PubChem's BioAssay database. *Nucleic Acids Res.*, **40**, D400–D412.
3. Wang,Y., Bolton,E., Dracheva,S., Karapetyan,K., Shoemaker,B.A., Suzek, T.O., Wang,J., Xiao,J., Zhang,J. and Bryant,S.H. (2010) An overview of the PubChem BioAssay resource. *Nucleic Acids Res.*, **38**, D255–D266.
4. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.*, **37**, W623–W633.
5. Cheng,T., Pan,Y., Hao,M., Wang,Y. and Bryant,S.H. (2014) PubChem applications in drug discovery: a bibliometric analysis. *Drug Discov. Today*, **19**, 1751–1756.
6. Qader,A.A., Urraca,J., Torsetnes,S.B., Tonnesen,F., Reubsaet,L. and Sellergren,B. (2014) Peptide imprinted receptors for the determination of the small cell lung cancer associated biomarker progastrin releasing peptide. *J. Chromatogr. A*, **1370**, 56–62.
7. Zhu,H., Zhang,J., Kim,M.T., Boison,A., Sedykh,A. and Moran,K. (2014) Big data in chemical toxicity research: the use of high-throughput screening assays to identify potential toxicants. *Chem. Res. Toxicol.*, **27**, 1643–1651.
8. Austin,C.P., Brady,L.S., Insel,T.R. and Collins,F.S. (2004) NIH molecular libraries initiative. *Science*, **306**, 1138–1139.
9. Schreiber,S.L., Kotz,J.D., Li,M., Aube,J., Austin,C.P., Reed,J.C., Rosen, H., White,E.L., Sklar,L.A., Lindsley,C.W. *et al.* (2015) Advancing biological understanding and therapeutics discovery with small-molecule probes. *Cell*, **161**, 1252–1265.
10. Liu,Z., Li,Y., Han,L., Li,J., Liu,J., Zhao,Z., Nie,W., Liu,Y. and Wang,R. (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.
11. Southan,C., Sharman,J.L., Benson,H.E., Faccenda,E., Pawson,A.J., Alexander,S.P., Buneman,O.P., Davenport,A.P., McGrath,J.C., Peters,J. A. *et al.* (2016) The IUPHAR/BPS Guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. *Nucleic Acids Res.*, **44**, D1054–D1068.
12. Gilson,M.K., Liu,T., Baitaluk,M., Nicola,G., Hwang,L. and Chong,J. (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, **44**, D1045–D1053.
13. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
14. Visser,U., Abeyruwan,S., Vempati,U., Smith,R.P., Lemmon,V. and Schurer,S.C. (2011) BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, **12**, 257.
15. Howe,E.A., de Souza,A., Lahr,D.L., Chatwin,S., Montgomery,P., Alexander,B.R., Nguyen,D.T., Cruz,Y., Stonich,D.A., Walzer,G. *et al.* (2015) BioAssay Research Database (BARD): chemical biology and probe-development enabled by structured metadata and result types. *Nucleic Acids Res.*, **43**, D1163–D1170.
16. Crowther,G.J., Hillesland,H.K., Keyloun,K.R., Reid,M.C., Lafuente-Monasterio,M.J., Ghidelli-Disse,S., Leonard,S.E., He,P., Jones,J.C., Krahn,M.M. *et al.* (2016) Biochemical screening of five protein kinases from Plasmodium falciparum against 14 000 cell-active compounds. *PloS One*, **11**, e0149996.
17. Wang,Z., Bhattacharya,A. and Ivanov,D.N. (2015) Identification of small-molecule inhibitors of the HuR/RNA interaction using a fluorescence polarization screening assay followed by NMR validation. *PloS One*, **10**, e0138780.
18. Swenson,J.M., Colmenares,S.U., Strom,A.R., Costes,S.V. and Karpen,G. H. (2016) The composition and organization of Drosophila heterochromatin are heterogeneous and dynamic. *eLife*, **5**, e16096.
19. Voter,A.F., Manthei,K.A. and Keck,J.L. (2016) A high-throughput screening strategy to identify protein-protein interaction inhibitors that block the Fanconi Anemia DNA repair pathway. *J. Biomol. Screen.*, **21**, 626–633.
20. Sun,J., Li,N., Oh,K.S., Dutta,B., Vayttaden,S.J., Lin,B., Ebert,T.S., De Nardo,D., Davis,J., Bagirzadeh,R. *et al.* (2016) Comprehensive RNAi-based screening of human and mouse TLR pathways identifies species-specific preferences in signaling protein use. *Science Signal.*, **9**, ra3.
21. Schmidt,C.K., Galanty,Y., Sczaniecka-Clift,M., Coates,J., Jhujh,S., Demir,M., Cornwell,M., Beli,P. and Jackson,S.P. (2015) Systematic E2 screening reveals a UBE2D-RNF138-CtIP axis promoting DNA repair. *Nat. Cell Biol.*, **17**, 1458–1470.
22. Lin,R., Elf,S., Shan,C., Kang,H.B., Ji,Q., Zhou,L., Hitosugi,T., Zhang,L., Zhang,S., Seo,J.H. *et al.* (2015) 6-Phosphogluconate dehydrogenase links oxidative PPP, lipogenesis and tumour growth by inhibiting LKB1-AMPK signalling. *Nat. Cell Biol.*, **17**, 1484–1496.
23. Schmich,F., Szczurek,E., Kreibich,S., Dilling,S., Andritschke,D., Casanova,A., Low,S.H., Eicher,S., Muntwiler,S., Emmenlauer,M. *et al.* (2015) gespeR: a statistical model for deconvoluting off-target-confounded RNA interference screens. *Genome Biol.*, **16**, 220.
24. Lang,L., Ding,H.F., Chen,X., Sun,S.Y., Liu,G. and Yan,C. (2015) Internal ribosome entry site-based bicistronic in situ reporter assays for discovery of transcription-targeted lead compounds. *Chem. Biol.*, **22**, 957–964.
25. Gupte,A., Baker,E.K., Wan,S.S., Stewart,E., Loh,A., Shelat,A.A., Gould, C.M., Chalk,A.M., Taylor,S., Lackovic,K. *et al.* (2015) Systematic screening identifies dual PI3K and mTOR inhibition as a conserved therapeutic vulnerability in osteosarcoma. *Clin. Cancer Res.*, **21**, 3216–3229.
26. Telford,B.J., Chen,A., Beetham,H., Frick,J., Brew,T.P., Gould,C.M., Single,A., Godwin,T., Simpson,K.J. and Guilford,P. (2015) Synthetic lethal screens identify vulnerabilities in GPCR signaling and cytoskeletal organization in E-cadherin-deficient cells. *Mol. Cancer Ther.*, **14**, 1213–1223.
27. Pasetto,M., Antignani,A., Ormanoglu,P., Buehler,E., Guha,R., Pastan,I., Martin,S.E. and FitzGerald,D.J. (2015) Whole-genome RNAi screen highlights components of the endoplasmic reticulum/Golgi as a source of resistance to immunotoxin-mediated cytotoxicity. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E1135–E1142.
28. Pena,I., Pilar Manzano,M., Cantizani,J., Kessler,A., Alonso-Padilla,J., Bardera,A.I., Alvarez,E., Colmenarejo,G., Cotillo,I., Roquero,I. *et al.* (2015) New compound sets identified

from high throughput phenotypic screening against three

kinetoplastid parasites: an open resource. *Sci. Rep.*, **5**, 8771.

29. Nyati,S., Schinske-Sebolt,K., Pitchiaya,S., Chekhovskiy,K., Chator,A., Chaudhry,N., Dosch,J., Van Dort,M.E., Varambally,S., Kumar-Sinha,C. *et al.* (2015) The kinase activity of the Ser/Thr

kinase BUB1 promotes TGF-beta signaling. *Science Signal.*, **8**, ra1.

30. Shan,C., Elf,S., Ji,Q., Kang,H.B., Zhou,L., Hitosugi,T., Jin,L., Lin,R., Zhang, L., Seo,J.H. *et al.* (2014) Lysine acetylation activates 6-phosphogluconate dehydrogenase to promote tumor growth. *Mol. Cell*, **55**, 552–565.

31. Kudryavtsev,D., Makarieva,T., Utkina,N., Santalova,E., Kryukova,E., Methfessel,C., Tsetlin,V., Stonik,V. and Kasheverov,I. (2014) Marine natural products acting on the acetylcholine-binding protein and nicotinic receptors: from computer modeling to binding studies and electrophysiology. *Marine Drugs*, **12**, 1859–1875.

32. Costantino,L., Sotiriou,S.K., Rantala,J.K., Magin,S., Mladenov,E., Helleday, T., Haber,J.E., Iliakis,G., Kallioniemi,O.P. and

Halazonetis,T.D. (2014) Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science*, **343**,

88–91.

33. Crowther,G.J., Booker,M.L., He,M., Li,T., Raverdy,S., Novelli,J.F., He,P., Dale,N.R., Fife,A.M., Barker,R.H. Jr *et al.* (2014)

Cofactor-independent phosphoglycerate mutase from nematodes has limited druggability, as revealed by two high-throughput screens. *PLoS Negl. Trop. Dis.*, **8**, e2628.

34. Falkenberg,K.J., Gould,C.M., Johnstone,R.W. and Simpson,K.J. (2014) Genome-wide functional genomic and transcriptomic analyses for genes regulating sensitivity to vorinostat. *Sci. Data*, **1**, 140017.

35. Hasson,S.A., Kane,L.A., Yamano,K., Huang,C.H., Sliter,D.A., Buehler,E., Wang,C., Heman-Ackah,S.M., Hessa,T., Guha,R. *et al.* (2013) High-content genome-wide RNAi screens identify regulators of parkin upstream of mitophagy. *Nature*, **504**, 291–295.

36. George,A.J., Purdue,B.W., Gould,C.M., Thomas,D.W., Handoko,Y., Qian,H., Quaife-Ryan,G.A., Morgan,K.A., Simpson,K.J., Thomas,W.G. *et al.* (2013) A functional siRNA screen identifies genes modulating angiotensin II-mediated EGFR transactivation. *J. Cell Sci.*, **126**, 5377–5390.

37. Kim,S., Thiessen,P.A., Bolton,E.E., Chen,J., Fu,G., Gindulyte,A., Han,L., He, J., He,S., Shoemaker,B.A. *et al.* (2016) PubChem Substance and Compound databases. *Nucleic Acids Res.*, **44**, D1202–D1213.

38. Kim,S., Thiessen,P.A., Bolton,E.E. and Bryant,S.H. (2015) PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.*, **43**, W605–W611.