“十二五”普通高等教育本科国家级规划教材

普通高等教育“十一五”国家级规划教材

化学信息学

第二版

李梦龙　文志宁　主编

李益洲　郭延芝　战玉华　副主编

HUAXUE XINXIXUE

化学工业出版社

“十二五”普通高等教育本科国家级规划教材

普通高等教育“十一五”国家级规划教材

化学信息学

第二版

李梦龙　文志宁　主编

李益洲　郭延芝　战玉华　副主编

化学工业出版社

·北京·

《化学信息学》（第二版）是“十二五”普通高等教育本科国家级规划教材、普通高等教育“十一五”国家级规划教材。本书在第一版的基础上，着重针对中、外文期刊网络数据库的检索过程进行了介绍，并扩充了专利文献与技术标准数据库的相关内容。全书主要分为四大部分，共12章，其中第1章概述了化学信息学的产生及特点；第2～8章讲述了化学信息的来源，包括纸质版的手册，书籍，搜索引擎，目前广为使用的中、外文期刊文献数据库以及专利文献与技术标准数据库；第9～11章介绍了化学信息的处理工具（即化学软件）、处理方法（相关化学计量学算法）以及定量构效关系（QSAR）的原理及应用，并新增了部分例子；在第12章中，对生物信息学领域的研究进行了概述。

《化学信息学》（第二版）可作为高等院校化学、化工及相关专业本科“化学信息学”课程的入门教材，另外，书中提供了大量与化学信息学相关的网址，亦可作为研究生的参考书籍。

图书在版编目（CIP）数据

化学信息学/李梦龙，文志宁主编. —2版. —北京：化学工业出版社，2018.7

“十二五”普通高等教育本科国家级规划教材　普通高等教育“十一五”国家级规划教材

ISBN 978-7-122-32235-7

Ⅰ. ①化…　Ⅱ. ①李…②文…　Ⅲ. ①计算机应用-化学-信息检索-高等学校-教材　Ⅳ. ①G252.97

中国版本图书馆CIP数据核字（2018）第112648号

责任编辑：杜进祥　　文字编辑：向　东

责任校对：边　涛　　装帧设计：关　飞

出版发行：化学工业出版社（北京市东城区青年湖南街13号　邮政编码100011）

印　　装：三河市双峰印刷装订有限公司

787mm×1092mm　1/16　印张16¾　字数440千字　2018年9月北京第2版第1次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899

网　　址：http://www.cip.com.cn

凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：40.00元

前 言

化学是一门以实验为基础的古老学科，科学家们致力于探索新物质的各类化学属性。随着实验技术的进步，人们获取数据的能力已有了很大的提高。时至今日，面对呈指数级增长的化学数据，化学家们发现在本学科的探索中，最大的瓶颈已不再是如何获取未知物质的性质，而是如何从已有的实验数据中提取更多的化学信息，以总结规律。1995年，美国著名化学家Brown指出，化学家习惯于将99%的精力和资源用于数据的收集上，其余1%用于数据的分析和处理，将数据转化为信息。

庆幸的是，现在越来越多的化学工作者们开始注意到这个问题并投身于化学数据的分析中，化学信息学（cheminformatics）在这种情况下应运而生，它是汇集化学、数学、信息科学等交叉学科知识的研究领域，主要是通过对化学信息的检索、整理、分析以及可视化，最终完成将数据转化为信息的过程。德国Johann Gasteiger出版的《化学信息学教程》一书亦指出，化学信息学的任务就是运用信息学的方法来解决化学的问题。

目前，化学信息学主要涵盖了化学信息的获取、化学信息的表达以及化学信息的处理三个方面的内容。作为一门新的基础课程，如何尽快让化学专业的本科生了解并掌握其中涉及的概念、方法以及网络资源是该学科建设亟待解决的主要问题。

本书第一版是“十二五”普通高等教育本科国家级规划教材、普通高等教育“十一五”国家级规划教材。本次修订，着重对如下内容进行了更新：（1）着重针对中文及外文期刊网络数据库检索过程的介绍进行了扩充，单独在第4章、第5章对目前常用的数据库进行了详细的介绍；（2）在第6章中，对第一版未曾提到的专利文献与技术标准进行了介绍；（3）在生物信息学相关章节中，对目前热门的领域，如精准医疗等进行了介绍；（4）对第一版各章节中的网络资源进行了更新，并新增了一些例子，便于学生们在学习过程中对内容进行理解。

本书由李梦龙、文志宁任主编，李益洲、郭延芝、战玉华任副主编。感谢清华大学图书馆的战玉华老师撰写了本书第4章以及在后期修改过程中提出的宝贵意见。在本书编写过程中，笔者实验室的博士研究生柳媛以及硕士研究生梁羽、谢凡凡、覃柳、吉越、胡文、郝颖异、黄雨瑶、郭佳丽等同学亦参与了资料的收集与整理工作，另外，化学工业出版社在此过程中提供了大力的支持，在此一并表示诚挚的感谢。

由于化学信息学涉及面广，编者的水平和时间有限，书中不足之处在所难免，恳请广大读者批评指正。

编者

2018年3月

目 录

HUAXUE XINXIXUE
化学信息学

第1章 概 述

材料、能源和信息是构成物质世界的三个基本要素。随着社会发展的需要，人们逐渐认识到信息的重要性，并创立了信息论与信息科学。20世纪90年代初，随着美国“信息高速公路”计划的提出，信息科学和信息产业出现了前所未有的飞速增长，成为这一时代的重要标志。同时信息科学加快了向传统科学渗透，化学中的信息学理论基础不断成熟。正是在这一背景下，结合其使用的计算机和因特网（Internet）工具，化学工作者在科研实践中促成了化学信息这一新兴化学分支的出现。化学信息学（chemoinformatics 或 cheminformatics）是化学领域中飞速发展起来的一个新的分支，是建立在多学科基础上的交叉学科，它利用计算机技术和计算机网络技术，对化学信息进行表示、管理、分析、模拟和传播，以实现化学信息的提取、转化与共享，揭示化学信息的实质与内在联系，促进化学学科的知识创新。

1.1 什么是化学信息学

诺贝尔化学奖获得者法国化学家J. M. Lehn在获奖报告中首次提出化学信息的概念，对化学的发展而言具有深远的影响，具有深刻的时代意义。虽然在此之前众多的化学工作者没有对化学信息展开实质性的研究工作，但是传统有机化学、无机化学、生物化学、材料化学以及在受体设计、超分子形成过程的结构化学等方面所积累的大量实验数据，为构建化学信息提供了基础。

化学信息学是个广义的概念，它包含对化学相关信息的设计、创造、组织、存储、处理、恢复、分析、再开发、可视化及应用。另一种关于化学信息学的定义是从药物研发的角度提出的，认为化学信息学是各种信息资源的混合体，目的是将数据转化成信息，再把信息转化成知识，以期更快、更准确地进行药物筛选和设计。

1.2 化学信息学的诞生背景

近十年来，由于计算机及网络技术向智能化、网络化方向发展，应用计算机技术能解决的化学问题也愈来愈多，化学工作者不仅获得了大量物质结构的信息，而且这些信息较从前也更为精确，计算机技术与化学的相互渗透已成为化学和计算机科学工作者的研究热点。由于计算机主要是通过数值计算来解决问题，其特点是能快速进行大量复杂、烦琐的数学运算，而化学是对化学物质进行认识、分析、合成及利用，从而使化学工作者能够对物质化学结构进行解析、表征、模拟与设计；能够处理复杂体系的电子结构、几何结构与其性能关系；完成微观分子工程设计与化学模拟；开展功能材料的研究；进行生物活性分子和药物分

子的相互作用机制及定量构效关系（quantitative structure-activity relationship，QSAR）研究；探讨固体表面结构、固体表面轨道相互作用规律；实现分子以上层次聚集体（超分子体系、界面体系等）结构和性能的模拟等。

然而，纵观早期这一领域的工作，仅仅涉及计算机技术的一些应用层次，要想将计算机技术深入应用到化学中就必须解决化学与计算机的结合问题，从化学工作者的角度应用和设计计算机软、硬件，满足化学工作者处理化学信息的要求。该领域的研究包括计算机与分析仪器的接口、化学类应用软件程序包的开发、化学物质结构数据库的开发和查询。1973 年在荷兰举办了“Computer Representation and Manipulation of Chemical Information”会议，1975 年美国化学会将期刊“Journal of Chemical Documentation”更名为“Journal of Chemical Information and Computer Sciences”，继而又于 2005 年更名为“Journal of chemical information and modeling”。由此可见，化学信息学已逐渐发展壮大，它将给 21 世纪的化学带来全新的面貌。

1.3 信息科学在化学领域的应用

20 世纪中叶，Shannon 在 1948 年发表了关于信息论的著名文章，提出了信息熵计算公式：

$$H(X)=-\sum_{i=1}^{n}\frac{1}{p(x_i)}\log_2[p(x_i)] \tag{1-1}$$

式中，$H(X)$ 为事件 X 的信息熵，它可由该事件当中所有可能出现的情况 x_i 的概率 $p(x_i)$ 计算得到。此后信息理论开始了它的发展，这一理论最早是与通信技术相关联，但在其诞生后十年左右，即从纯粹数学研究渗透到无线电、电视、雷达、心理学、语义学、经济学、生物学等领域。Wiener 认为信息的实质是负熵，并强调信息这种负熵是在调节过程中相互交换而产生的。

化学科学中的分析化学从其诞生起就具有信息科学的特征，Kateman 等从三个方面阐述了分析化学的任务：利用已有的分析方法，提供关于物质化学成分的信息（日常例行分析工作）；研究利用不同学科的原理、方法，取得有关物质系统的相关信息的过程（分析化学的科学研究工作）；研究利用现有分析方法取得关于物质系统的信息的策略（分析实验室的组织工作）。Kowalski 更是明确提出“分析化学作为信息科学”，他认为分析化学不仅在过去是一门信息科学，现在仍然是一门信息科学，在化学的各个分支学科中，分析化学负担的任务与其他分支学科的不同之处在于分析化学的研究对象，它并非某种具体的实物，例如无机或有机材料，而是与这些化学组成或结构相关的信息以及研究获取这些信息的最优方法与策略。

此外，由于化学中熵的概念与 Shannon、Wiener 等提出的信息理论中的熵有着共同的基础，这两门学科之间存在着深刻联系，分析化学以发展分析信息理论作为其基础理论的组成部分，获得了向前发展的动力。随后众多的化学家根据其从事的分析化学工作，发表了多篇用于分析化学的信息理论系列论文，其中捷克学者 Eckschlager 完成了在此领域的第一部专著。

1.4 化学信息采集接口

从图 1-1 中我们可以发现，化学信息和化学实践之间是通过一个信息采集接口相连的，

这与其他化学分支学科明显不同。信息采集接口也是化学信息学科一个极为重要的组成部分，它是现实化学世界通往化学信息的桥梁，也是化学信息的生命源泉。信息采集接口和化学信息的三个层次构成了整个化学信息学科。

图 1-1 信息采集接口的内部结构

如图 1-1 所示，信息采集接口不是一套单一而又具体的软件或硬件，它实际上是一个综合了其他化学分支学科的某些方法以及原始化学信息数字化方法的方法集合。可以用作信息采集接口后端的方法众多、范围非常广泛，一般的各种化学图形处理软件、计算机化学应用软件、文字处理软件、数码照相机、具有 OCR（optical character recognition）光学文字识别功能的扫描仪的录入系统等都可以作为信息采集接口的后端，因为它们都具有一个共同的特点，即能够把现实世界中的化学信息以数字化形式存储起来。对于接口的前端，只要是能从化学实践中获取化学信息的研究方法或仪器设备，如化学分析测量仪器、量子化学计算方法、分析化学及物理化学实验方法等，都可以用作信息采集接口的前端。

1.5 化学信息的结构和特点

仔细分析，我们可以发现化学信息的主体部分实际上是由三个层次构成的，即信息核心层、信息处理层和信息表示层，如图 1-2 所示，化学信息的这种分层结构本质上是计算机技术分层结构的反映。

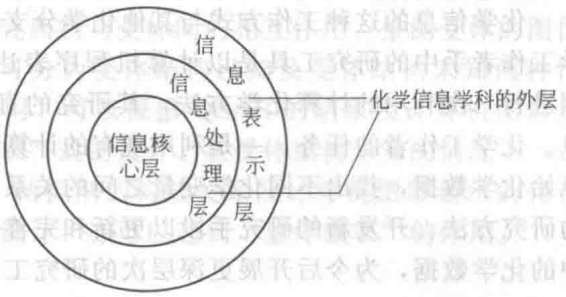

图 1-2 化学信息的结构示意图

在化学实践中产生出来并被计算机处理过的原始化学信息，比如在一个化学实验中发生的各种实验现象、记录的实验数据以及与化学实验相关的外界条件等，它们组成了化学信息的信息核心层。信息处理层由化学计量学方法、药物分子设计方法、QSAR 方法等能够对信息核心层中的数字化化学信息进行二次开发利用的计算方法组成。信息表示层处于化学信息学科的最外层，它根据信息核心层的特定要求在计算机信息科学中寻找适合表达化学信息的技术，从多个角度将化学信息以某种直观的形式，如基于计算机的图形、音频、视频等多媒体表示手段向化学工作者展示出来。信息处理层和信息表示层统称为化学信息学科的外层。

三个层次中最重要的层次为信息核心层，它在化学信息学科中处于基础核心地位，并决定了其他两个层次的构成。信息核心层对外层起着决定性作用，外层对信息核心层也能够产生一定的影响。由于受到计算机信息科学技术和仪器设备本身存在某些特点的影响，要求信息核心层中的化学信息必须要按照一定的要求进行组织和编排。例如，连续吸光度曲线在计算机中只能以离散数据点的形式存储，海量的光谱数据、化学化工测量数据或分子结构参数，唯有按一定的数据结构规范化并形成数据库甚至是专家系统，才能方便日后使用信息处

理层对这些数据进行开发利用。信息处理层在对来自信息核心层的化学信息进行处理之后，所获得的结果一方面将交由信息表示层处理；另一方面，信息处理层将把某些处理结果和原始数据存储在信息核心层，使该层信息量甚至是一些局部结构发生变化。

1.6 化学信息的工作方式

化学实践与化学信息之间的关系是“母与子”的关系。化学信息通过“信息采集接口”从化学实践这一母体中获取原始信息，原始信息经过接口的处理后以数字化形式存储到信息核心层中，并通过外层将其重现出来给化学工作者。利用这些被数字化技术处理过的化学信息，化学工作者可以进行更深层次的化学研究实践，从而生产出新的数字化的原始化学信息，这就是化学信息的工作方式，如图 1-3 所示。

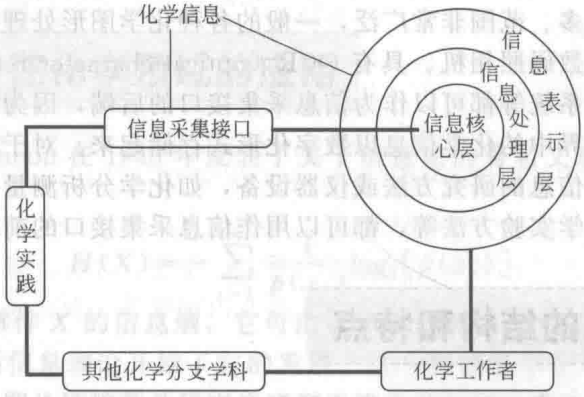

图 1-3 化学信息的工作方式

化学信息的这种工作方式与其他化学分支学科的工作方式基本相同，最大的区别在于化学工作者手中的研究工具是以计算机程序表达的化学计量学、QSAR 等可以对大量化学数据进行二次利用的计算化学方法，其研究的直接对象和得到的产品都是数字化了的化学信息。化学工作者的任务，一是利用现有的计算机软、硬件工具研究大量存储在信息核心层的原始化学数据，找出不同化学变量之间的关系，发现有实际意义的化学规律；二是改进现有的研究方法、开发新的研究手段以更新和完善化学信息外层以及不断丰富和修正信息核心层中的化学数据，为今后开展更深层次的研究工作奠定基础。

1.7 化学信息学的应用

1.7.1 化合物结构绘制

化学信息普遍存在于化学和计算机的结合之中，几乎每一个化学家都是一个绘制结构者，都会使用到 IsisDraw、ChemBioDraw、JchemPaint 等相关软件去绘制一个分子的二维或三维结构。然而，更深入的问题则需要化学信息学的方法来解决，例如如何将化学的结构有效地储存在计算机里；采用何种格式可以用来在不同类型的化学软件中交换数据等。

1.7.2 化学数据库设计与开发

化学数据库的设计、开发、维护和更新是化学信息的重要领域。化学结构和一些相关的

信息主要被存储在化学数据库中，如 Beilstein 数据库。自 1771 年起，化学数据库中存储了超过七百万个有机化合物的信息，可以由 CAS 登录号或分子式查询物质的物理和化学性质，包括光谱数据以及热力学参数。另外在构建这些数据库时，基于物质化学数据信息的数据库查询功能同样至关重要。

1.7.3 化学反应体系模拟

人们已经做了很多的努力去实现用计算机预测化学反应的进行，模拟化学反应的发生，来合成一个设计好的目标产物。目前，德国 Gasteiger 的化学信息研究小组，已经有了这样的一个系统，名叫 EROS (elaboration of reactions for organic synthesis)，能够进行包括两种反应物之间的反应结果的预测，或提供采用何种反应物的建议。

1.7.4 计算机辅助波谱解析

现代分析化学对于仪器的依赖越来越多，谱学手段发展很快，各种光谱、色谱、质谱以及多种仪器联用已成为了分析化学研究的重要工具，因而，计算机辅助谱图解析也成为了研究人员关注的焦点。一般而言，计算机辅助谱图解析方法可以大致分为两类：一是直接对谱图库进行相似性检索 (library search)；二是采取波谱模拟、模式识别及人工智能的手段进行间接识别。由此涌现出了一大批谱图解析专家系统，如由斯坦福大学组建的用于质谱、^{13}C NMR 解析的 DENDRAL 系统；由日本左左木慎一等研制的用于质谱、^{13}C NMR、^{1}H NMR、红外光谱解析的 CHEMICS 系统，等等。

1.7.5 化合物结构与活性关系预测

QSAR (quantitative structure-activity relationship) 定量构效关系方法尝试通过对一系列结构相似的药物分子进行分析，找出分子性质参数和生物活性之间的关系，并以此为依据去预测具有药效的新型分子的结构与性质。目前发展到三维的 3D-QSAR 实际上是 QSAR 与计算机分子图形学相结合的研究方法，是研究药物与受体间的相互作用，推测受体的图像及进行药物设计的有力工具。3D-QSAR 研究可分为受体结构已知及受体结构未知两种情况。受体结构已知（目前仅限于酶作为受体）时，可以根据 QSAR 的结果及计算机图形显示受体的三维结构，并随之进行有如“量体裁衣”式的设计。在受体结构未知的情况下（这是绝大多数情况），则可以根据激动剂或（和）拮抗剂的构效关系及计算机图形显示的化合物优势构象，推测受体的结构，然后进行药物设计，亦可以起到“量体裁衣”的作用。

1.7.6 实验室信息管理系统

实验室信息管理系统 (laboratory information management system，LIMS) 是将以数据库为核心的信息化技术与实验室管理需求相结合的信息化管理工具。该系统的建立是根据 ISO/IEC 17025：2005-5-15《检测和校准实验室能力的通用要求》规范，结合网络化技术，将实验室的业务流程和一切资源以及行政管理等以合理方式进行管理。经过 LIMS 管理，可有效提高实验室样品检测效率、提高分析结果可靠性、提高对复杂分析问题的处理能力、协调实验室各类资源以及实现量化管理。经过多年的发展，国内很多大型实验室均认识到了信息化管理的重要性，引入了 LIMS 系统，并进行了不断的改进和升级。

1.8 展望

随着信息时代的到来，各类化学信息的相关数据不断涌现，这些数据在使我们获得更多