CHEMICAL REVIEWS

Review

pubs.acs.org/CR

Information Retrieval and Text Mining Technologies for Chemistry

Martin Krallinger,[†,○] Obdulia Rabal,[‡,○] Anália Lourenço,[§,∥,⊥] Julen Oyarzabal,[*,‡] and Alfonso Valencia[*,#,▽,■]

[†]Structural Computational Biology Group, Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre, C/Melchor Fernández Almagro 3, Madrid E-28029, Spain

[‡]Small Molecule Discovery Platform, Molecular Therapeutics Program, Center for Applied Medical Research (CIMA), University of Navarra, Avenida Pio XII 55, Pamplona E-31008, Spain

[§]ESEI - Department of Computer Science, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas s/n, Ourense E-32004, Spain

[∥]Centro de Investigaciones Biomédicas (Centro Singular de Investigación de Galicia), Campus Universitario Lagoas-Marcosende, Vigo E-36310, Spain

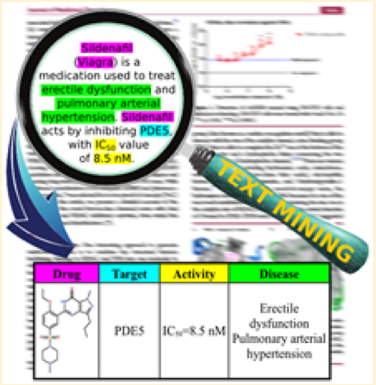
[⊥]CEB-Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga 4710-057, Portugal

[#]Life Science Department, Barcelona Supercomputing Centre (BSC-CNS), C/Jordi Girona, 29-31, Barcelona E-08034, Spain

[▽]Joint BSC-IRB-CRG Program in Computational Biology, Parc Científic de Barcelona, C/ Baldiri Reixac 10, Barcelona E-08028, Spain

[■]Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig de Lluís Companys 23, Barcelona E-08010, Spain

ABSTRACT: Efficient access to chemical information contained in scientific literature, patents, technical reports, or the web is a pressing need shared by researchers and patent attorneys from different chemical disciplines. Retrieval of important chemical information in most cases starts with finding relevant documents for a particular chemical compound or family. Targeted retrieval of chemical documents is closely connected to the automatic recognition of chemical entities in the text, which commonly involves the extraction of the entire list of chemicals mentioned in a document, including any associated information. In this Review, we provide a comprehensive and in-depth description of fundamental concepts, technical implementations, and current technologies for meeting these information demands. A strong focus is placed on community challenges addressing systems performance, more particularly CHEMDNER and CHEMDNER patents tasks of BioCreative IV and V, respectively. Considering the growing interest in the construction of automatically annotated chemical knowledge bases that integrate chemical information and biological data, cheminformatics approaches for mapping the extracted chemical names into chemical structures and their subsequent annotation together with text mining applications for linking chemistry with biological information are also presented. Finally, future trends and current challenges are highlighted as a roadmap proposal for research in this emerging field.



CONTENTS

Received: December 30, 2016
Published: May 5, 2017

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

1. INTRODUCTION

Because of the transition from printed hardcopy scientific papers to digitized electronic publications, the increasing amount of published documents accessible through the Internet and their importance not only for commercial exploitation but also for academic research, data mining technologies, and search engines have impacted deeply most areas of scientific research, communication, and discovery. The Internet has greatly influenced the publication environment,[1] not only when considering the access and distribution of published literature, but also during the actual review and writing process. In this respect, chemistry is a pioneering domain of online information representation as machine-readable encoding systems for chemical compounds date back to the line notation systems of the late 1940s.

Currently, efficient access to chemical information (section 1.1) contained in scientific articles, patents, legacy reports, or the Web is a pressing need shared by researchers and patent attorneys from different chemical disciplines. From a researcher's viewpoint, chemists aim to locate documents that describe particular aspects of a compound of interest (e.g., synthesis, physicochemical properties, biological activity, industrial application, crystalline status, safety, and toxicology) or a particular chemical reaction among the growing collection of published papers and patents. In fact, chemists are among the researchers that spend more time reading articles, and after medical researchers are the ones that overall read more articles per person.[2] For example, there are approximately 10 000 journals publishing "chemistry" articles.[3] Every year, over

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

20 000 new compounds are published in medicinal and biological chemistry journals.[4] Together with journals, patents constitute a valuable information source for chemical compounds and reactions:[5] the Chemical Abstracts Service (CAS) states that 77% of new chemical compounds added to the CAS Registry are disclosed first in patent applications,[6] and the percentage of new compounds added to the CAS REGISTRY database from patents raised from 14% in 1976 to 46% in 2010, with 576 new compounds being added to the CAS REGISTRY after CAS analysis of a standard Patent Cooperation Treaty (PCT) application.[7] By the middle of 2015, the number of published patents was 13 510 for inorganic patents (International Patent Classification, IPC, code C01), 54 075 for organic patents (IPC C07), and 12 524 for metallurgy patent (IPC C22), as consulted in Espacenet.[8] Besides being an important source of information, from a legacy perspective,[9] different patent searches (e.g., state-of-the-art or prior art, patentability, validity, freedom to operate, and due diligence searches) directly influence the work carried out by researchers, patent examiners, and patent attorneys. In summary, these text collections represent a considerable fraction of the overall data generated not only in chemistry but science in general.[10]

Despite the differences in focus and scope of the diverse chemistry branches, final end users have common information demands: from finding papers of relevance for a particular chemical compound, chemical family, or reaction (chemical information retrieval, section 2) to extracting all chemicals or chemical entities mentioned in a document (chemical entity recognition, CER, section 3), including any associated information (e.g., chemical and physical properties, preparatory steps, or toxicological data).

Information retrieval (IR) is defined as finding within a (large) collection of documents the subset of those documents that satisfies a particular information user demand, also known as user information need.[11] A particularity of chemical information retrieval is that these information demands can be expressed as natural language text (text search queries, section 2.4) or can take into account structural information (structure-based search queries, section 5.3) or being a combination of both (hybrid searches, section 2.4). In fact, the main concerns of chemists when searching the literature is using the chemical structure or substructure as query input and/or retrieving the chemical structure as the result of document processing software. Given the multiple representations of chemicals or chemical entities (CE) in documents, with different name nomenclatures and synonyms (section 1.1), as chemical diagrams (images) that require conversion to structures (section 4.3) or with different notations capturing structural information such as line notations (SMILES), connectivity tables, and InChi codes (section 4.4), the quest for the discovery of chemically relevant information in documents is considerably more complex than general-purpose web searches or queries using generic electronic search tools.[12] Thus, there are recent concerns in providing a more formal training to chemists, to instruct them how to construct chemical search queries and acquire the necessary skills for effectively searching chemical information using various search strategies.[12] A survey of the most popular chemical document repositories and search platforms available to the community is presented in section 1.2. Most of them query against unstructured data repositories (without any predefined structure or organized in a predefined manner) and/or access to items of structured data repositories (i.e., defined database fields such as already indexed chemicals) or document metadata (data about the document itself). Metadata attributes of a document (or data in general) do usually provide some descriptive information associated to the document, such as author, publication dates, author names keywords, or journal information, which can be also exploited for retrieval purposes. Metadata attributes are usually much more structured than the actual document content. In the case of chemical documents, chemical compound structure metadata associated to a given document can be regarded as a special type of chemical metadata. For example, to provide a machine-readable version of the key data presented in the articles,[13] in 2014, the *Journal of Medicinal Chemistry* invited authors to submit a spreadsheet with the SMILES and basic information of the compounds presented in the articles. For the inexpert reader, it is important to highlight that IR systems (as well as CER) primarily deal with unstructured or semistructured machine-readable free text (sections 1.2 and 2.1), which require a previous preprocessing (section 2.2) to enable effective indexing and determination of similarities between input query and document contents (section 2.3). This Review focuses on content-based retrieval, that is, the textual content that forms part of the actual documents, rather than only metadata searches, as searching the whole document in addition to indexed fields is crucial to improve the quality of the information obtained. With that purpose, different text search strategies (boolean, subject, keyword, proximity, etc.), common to all IR systems beyond chemistry, are described and exemplified in the context of chemical IR (section 2.4). Here, the impact of CER in the context of IR, for example, to add semantics meaning (knowing that the query is a chemical concept and detecting this concept), is discussed. Also, IR methodologies to enable document classification and clustering, for example, according to a certain topic, are described in section 2.5. Finally, an overview of evaluation metrics for assessment of retrieval efficiency and current state of the art assessments (BioCreative) is discussed in section 2.6.

As commented above, a second goal of chemists when reading a document (e.g., a patent) is extracting all chemical entities mentioned in it, as unstructured data repositories host essential characterizations of chemical compounds obtained through experimental studies that describe their targets, binding partners, metabolism, or, in the case of drugs, the therapeutic use and potential adverse effects.[14,15] Researchers working on diverse chemical topics can benefit from systematic extraction of information on chemicals from document repositories, in particular the scholarly literature, patents, and health agency reports.[16] The term chemical entity recognition (CER) or chemical entity mention recognition refers to the process of automatic recognition of chemical entity mentions in text. Section 3 provides a deep overview of current challenges (section 3.3.3), strategies (sections 3.3−3.5), and available chemical corpora (section 3.8) for CER and ends with assessments on the quality of current methodologies (section 3.9), with an especial focus on BioCreative.

CER does not only constitute a key step within IR systems. Recognized chemical entities can be mapped to their corresponding structural representation (section 4.4) by name to structure conversion software (section 4.1) or by looking up names within the contents of structure chemical entity databases (section 4.2) and then stored, ideally in its canonical form (section 4.5), in chemical knowledge bases supporting

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

structural searches (section 5). These structural databases can be implemented as part of IR systems having chemical intelligent search capability that allows grouping all of the hits relevant to a specific chemical entity, regardless what alias, synonym, or typographical variant is used in the text to refer to the very same chemical object. Alternatively, these chemical knowledge bases can be integrated with, for example, biological data, and accessed by specialized search engines that query specific well-defined database fields. In the field of chemical-biology, the construction of chemical knowledge bases that integrate chemical information and biological data (targets and the associated phenotypic data and toxicological information) extracted from documents is becoming a common task both in academia and industry with a remarkable impact on drug discovery, as discussed in section 6. Although manual information extraction can be very accurate, data mining systems can speed up and facilitate the process, making it more systematic and reliable.[17−19] Automatic information extraction and mining technologies can complement arduous handcrafted annotations and extract chemical entities scattered across multiple data sources. Notwithstanding, automatically finding relations between chemical and biological entities in text can only be achieved efficiently through prior, fine-grained detection of chemical mentions in documents.

The possibilities and limitations of the current resource for decision-making in multifactorial drug discovery will be presented, as well as the main efforts dealing with unprecedented amounts of chemical data generated by the new mining methods. Finally, this Review (covering literature from the mid 1990s to date) is organized to serve as a practical guide to researchers entering in this field but also to help them to envision the next steps in this emerging data science field.

1.1. Chemical Information: What

To fully understand the issues associated with chemical IR and CER, an overview of the multiple representations of chemicals or chemical entities in documents is required. Although obvious, chemical entities can appear in documents either as text (names or chemical structure representations such as line notations) or as chemical diagrams (images) representing its chemicals structure. While this Review is oriented toward text mining (TM), a brief overview on how to extract chemical structures from images is provided in section 4.3.

Chemical entity mentions in text or chemical names can be expressed in many alternative ways, generally classified into systematic (e.g., "propan-2-ol"), semisystematic (e.g., "diacetylmorphine"), trivial/common or generic (e.g., "morphine"), trade/brand names (e.g., "MScontin"), acronyms/abbreviations (e.g., "CPD"), formulas (e.g., "$C_{17}H_{19}NO_3$"), names of groups, names of fragments or plural names (e.g., "diacetylmorphines"), chemical families (e.g., "ketolides"), verbs (e.g., "demethylates"), adjective forms (e.g., "pyrazolic"), chemical database identifiers either from the public domain (e.g., "CAS registry number: 57-27-2"; "MDL number: MFCD00081294"), or as company codes (e.g., "ICI204636"). Most of the chemical names described above do not contain information on the underlying chemical structure (i.e., connectivity between atoms and bonds) and therefore are not directly amenable to IR user demands accounting for structural information. Line notations such as SMILES[20,21] (e.g., "C1CCCCC1") and InChi/InChIKey[22,23] codes (e.g., "InChI = 1S/C6H12/c1-2-4-6-5-3-1/h1-6H2"), addressed in more detail in section 4.4, do capture this structural information and are therefore suitable for that purpose.

With the aim of standardizing the naming of entities in natural sciences, the definition of formal nomenclature and terminological rules to constrain how entities are correctly expressed in natural language has been proposed. Initially, chemicals were named using trivial names. To avoid issues related to the use of trivial names for chemical entities, the International Union of Pure and Applied Chemistry (IUPAC) was formed in 1919 to more systematically consider and review chemical information representation and apply standardization in chemical compound notation. Since 1921, the IUPAC has been organizing committees to deal with chemical nomenclature, aiming to write down rules for systematically naming chemical compounds.[24] Formal attempts to define chemical compound nomenclatures started over a hundred years ago. A milestone in this context was the international Geneva Conference in 1892 on Standardization of Names, which resulted from a series of previous events promoted initially by Friedrich August Kekulé von Stradonitz in 1860.[25] Such systematic names, sometimes also called IUPAC chemical names, are intended to be unambiguous representations of chemical structures. This property is the underlying assumption exploited by name to structure conversion algorithms (see section 4.1).[26] Despite continuous improving, nomenclature rules are not conclusive, with updates being compiled in the Gold Book compendium of technical nomenclature (which interestingly is structure-based searchable).[27] For example, IUPAC names are not sufficient to describe a molecule with complex stereochemistry. In addition to IUPAC names, there are also other efforts to provide some kind of systematic chemical names, like the CAS[28] index names and Beilstein-Institute.[29] Underlying the chemical nomenclature rules is a sort of chemical name grammar, which strengthens regularity in chemical names through the use and combination of building block name segments (e.g., substrings and terminal symbols). Such substrings, like "propyl", "alkyl", or "benzo", are very distinct from regular English words and are thus a useful property for automatic CER.

Manually assigned chemical names provided in publications are sometimes incorrect or misleading in the sense that they contain "mistakes" that make it impossible to generate a structure from the name as published by authors. Such an issue can be, in part, addressed by using computational nomenclature services.[30] Among this kind of chemical mentions are semisystematic names that present some characteristics of systematic names, such systematic chemical substrings, but also include portions of nonsystematic elements often corresponding to common or trivial names of chemicals.

Systematic chemical names, especially in the case of larger molecules, might be lengthy and difficult to read, remember, and construct for nonchemical experts. Therefore, in the scientific literature and, in particular, journal abstracts, more compact chemical entity names are widely used, especially common, trivial, and trade names as well as chemical abbreviations. Health care professionals including pharmacists and prescribers commonly use generic drug names.[31] There are expressly devoted councils, such as the United States Adopted Names (USAN) council[32] and the World Health Organization International Nonproprietary Name (WHO INN) effort,[33,34] to coordinate the official and unique naming of nonproprietary pharmaceutical drugs or active ingredients (i.e., official generic and nonproprietary names). The INN system can be regarded

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

as a standardization of drug nomenclature, providing logical nomenclature criteria to construct and select informative unique names for pharmaceutical substances. The underlying criteria often take into account pharmacological or chemical relationships of the drug. The INN system makes use of specific name stems (mostly suffixes and, in some cases, also prefixes) that group drugs according to certain attributes; for example, the suffix "-caine" is usually used for local anesthetics.

In addition to the previously characterized chemical name types, there are also some more exotic chemical name types, chemical "nicknames", or chemical substances named after their inventor's name, like in the case of "Glauber's salt" (i.e., sodium sulfate), "Jim's juice" (i.e., Cancell), and "Devil's Red" (i.e., Doxorubicin)[35,36] as well as company codes (e.g., ICI204636).

Together with single chemical compounds, chemical reactions and Markush structures are also targets of interest for mining technologies. Markush structures are used to claim a family of related compounds in patents and receive their name after the first successfully granted patent in 1924 by Eugene Markush containing a chemical compound having generic elements. Together with their application in patents, Markush formulas are commonly found in scientific papers describing structure−activity relationships (SAR) of a family of compounds and combinatorial libraries. Markush structures contain a core, depicted as a diagram, having generic notations, normally defined as R-groups, that comprise enumerations of atom lists, bond lists, and homology groups ("heteroaryl", "a bond", "C1−C8 alkyl"). Also, there are variations in the attachment positions for substituents and its frequency of appearance (repeating units).[37]

1.2. Chemical Information: Where

As introduced, with the goal of providing the interested reader a broad vision of the available chemical information sources, this section covers some of the common resources exploited by chemists while seeking information, regardless of whether they are unstructured or indexed repositories, as most complex document retrieval systems (search engines) do in fact aggregate them.

The major types of chemical documents include (i) scientific publications, (ii) patents, (iii) gray literature (conference reports, abstracts, dissertations, and preprints), and (iv) a plethora of regulatory, market, financial, and patent intelligence tools. Each of them has different degrees of format uniformity (Table 1) that greatly influence its preprocessing for TM and IR purposes (document segmentation, section 2.2.3). Patents are highly uniform in structure and consist of a bibliographic section with information on the title, applicant(s), inventor(s), filing date, publication date, patent classification codes, and abstract, which is followed by a description section with background information and exemplary data and ends with a set of claims about the scope of the invention. Scientific journals tend to share a general arrangement (Title, Abstract, Introduction, Materials and Methods, Experiments, Results, Discussion, and Summary and Conclusion sections)[38] although with great variability across publishers and themes. The rest of the document types lack a uniform structure, and each provider, especially commercial ones, arranges the data following its own format.

Table 1. General Characteristics of Main Chemical Information Sources

	scientific journals	patents	conference proceedings reports
uniform format	medium	high	low
content availability	abstracts (most), some full-text	full-text	variable depending on the provider
accessibility in the public domain	low, except for some open initiatives	high (patent offices)	medium
level of data aggregation in search systems	low	high	low

Other important aspects differing between these three major information sources are availability of full text, public accessibility, and level of data aggregation in search systems (Table 1), that is, to what extent a single search system aggregates different sources (e.g., patents from different patent authorities or journals from different publishers), thereby reducing the need of running separate searches.

Figure 1 shows a selected set of the largest and most popular document repositories of each type. As indicated, these repositories can be directly accessed online through specialized web interfaces and be accessed through a huge number of search systems and platforms, which connect to different repositories linking (or not) a variety of entities besides documents, such as chemical structures, chemical reactions, and Markush formula, thereby also enabling structural searches (covered in detail in section 5). In fact, as was recently highlighted by Ellegaard,[39] a main difference between the most popular repositories and search engines is how they index chemical data (if they do). Thus, in an attempt to clarify some common misinterpretations and commonly asked questions regarding highly accessed repositories (e.g., the difference between the user interface PubMed,[40,41] which accesses the repository MEDLINE,[42,43] or the relationship between the search platform SciFinder[44,45] and the CAplus[46] database), we present search engines and platforms (Table 2) connecting different repositories (with several entities or not) separately from document repositories in Figure 1.

1.2.1. Journal Literature and Conference Papers, Reports, Dissertations, Books, and Others. Repositories for journal literature include some of the established, standard references in their field: MEDLINE,[42] Embase,[47] and BIOSIS Previews[48] in biomedicine, TOXLINE[49] in pharmacology/toxicology, Inspec[50] in physics/engineering, as well as broader content databases covering different sources from all areas of chemistry, biochemistry, chemical engineering, and related sciences (CAplus) and nonlife science domains (Scopus[51] and Directory of Open Access Journals, DOAJ[52]). In the biomedical field, MEDLINE (public) and Embase, a product by Elsevier, are the two most prominent resources, with different publications over the last years emphasizing their complementarity,[53−56] while Thomson Reuters includes BIOSIS Previews to complement MEDLINE within their search platform Web of Science[57] (formerly ISI Web of Knowledge). Most of them are bibliographic repositories, containing abstracts (e.g., MEDLINE[42] and TOXLINE[49]), indexed terms extracted from abstracts alone (e.g., Scopus[51]) or from full-texts (e.g., Embase[47] and CAplus[46]), and citations (or are connected to a citations database such as SCI-EXPANDED[58]). In most cases, especially for nonpatent literature, the search services connected to these repositories do not provide the possibility to access documents identified through searches in their full-text version, but have links to the corresponding journal (e.g.,

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

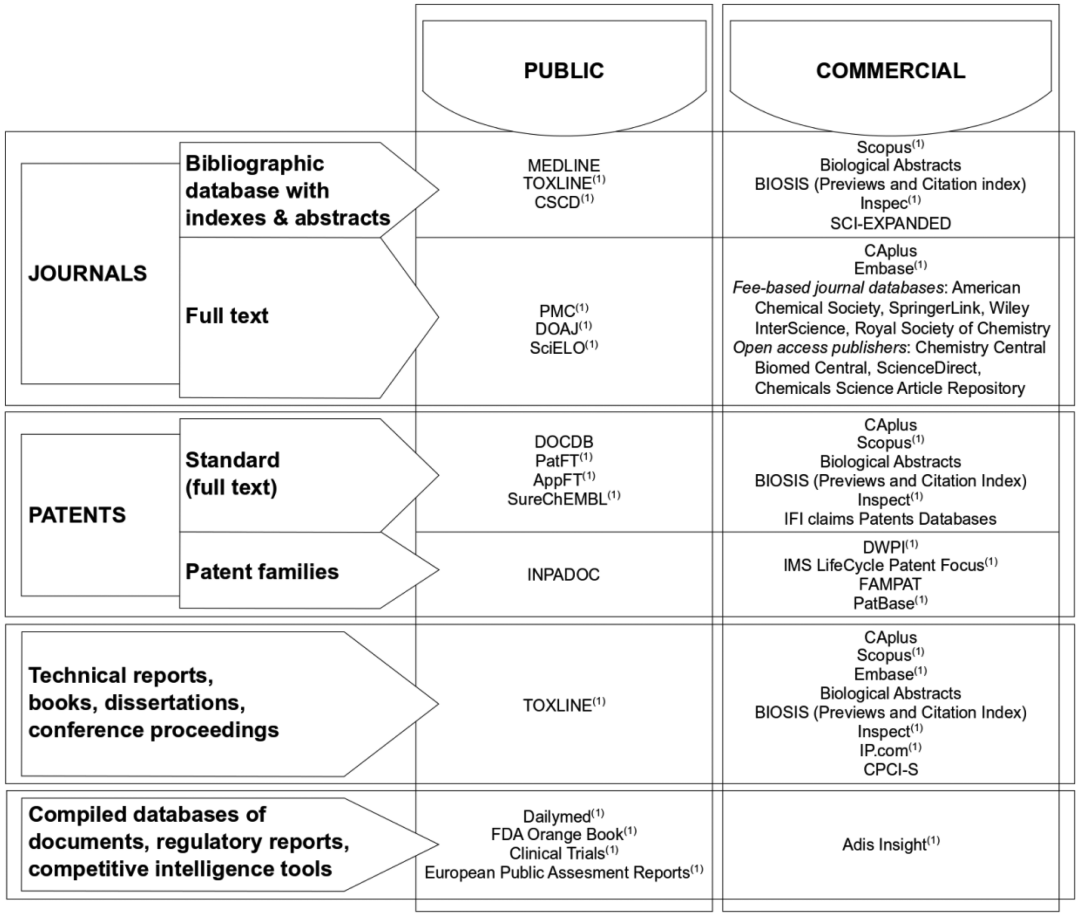


Figure 1. Largest and most popular document repositories. "(1)" indicates that it supports text-based searches, with online access to query the database. Prepared in September 2016.

SciFinder[44,45]). Oppositely, some smaller repositories such as PubMed Central (PMC[59,60]), for the life science domain, and the Directory of Open Access Journals,[52] for different subject areas, provide full-text document access. Apart from these compiled resources, text-based searches can be run over fee-based journal databases or open access publishers. Of interest, some of the public document repositories can be accessed by researchers with TM noncommercial purposes. MEDLINE allows text mining entirely its content, while TOXLINE and PubMed Central restrict full-access to their entire content. ScienceDirect[61] can be accessed via the ScienceDirect API's. Chinese Science Citation Database (CSCD),[62,63] in English and Chinese, and Scientific Electronic Library Online (SciELO),[64] in English, Spanish, and Portuguese, exemplify available resources in other languages besides English.

As seen in Figure 1, the most relevant databases (e.g., CAplus, Embase, and Scopus) do also include other types of documents such as books, conference proceedings, and dissertations, although there exist dedicated resources such as the Conference Proceedings Citation Index-Science (CPCI-S)[65] for the most significant conferences worldwide. Of note, IP.com[66] is a full-text database for companies and individuals to publish and search technical disclosures (defensive publications).

1.2.2. Patents. Patents, including applications and granted patents, are also included and indexed in the largest commercial repositories listed above (i.e., CAplus, Scopus, Biological Abstracts, BIOSIS, and Inspec). In the public domain, national and regional offices track patent applications and granted patents as images of text documents, and either provide Internet access to their internal collections (PatFT,[67] AppFT[68] by the United States Patent and Trademark Office, USPTO) or offer patents in semistructured formats, like XML (DOCDB[69] by the European Patent Office, EPO). This has favored the emergence of many different repositories and search engines (most of them commercial, as listed in Table 2) aggregating patent documents from several offices, which are exclusively specialized in patents, for prior art searching and intellectual property valorization. An interesting feature of patent databases is that they are organized by patent families (e.g., INPADOC,[70] Derwent World Patents Index (DWPI),[71] FAMPAT,[72] and PatBase[73]) rather than single patent records. A patent family is a set/group of published patent documents taken in multiple countries that disclose the same invention and are linked by one or more common priority numbers (the application serial number for the earliest application). Patent family information is useful for determining the scope of international patent protection for a specific patented product or process, identifying the translation of a patent document, and overcoming problems associated with spelling variations and transliteration of inventor names and applicants. Besides manual indexing and data curation (i.e., misspelling corrections, optical character recognition (OCR) errors), some commercial vendors clean up and enhance patent data. For example, the DWPI,[71] by Thomson Reuters, includes enhanced titles and abstracts, and the search engine SciFinder[44,45] proposes a more

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Table 2. Search Systems and Platforms Connecting To Document Repositories (in Figure 1)[a]

service	provider/company	launched	Struct	J	P	URL
Public						
PubMed	U.S. NLM	1996		√		http://www.ncbi.nlm.nih.gov/pubmed
Europe PMC	Europe PMC Consortium	2012[c]		√	√	http://europepmc.org
PMC Canada	CIHR/NRC-CISTI/NLM	2009		√		http://pubmedcentralcanada.ca/pmcc
TOXNET[b]	U.S. NLM	N/A		√		http://toxnet.nlm.nih.gov
Google Scholar	Google	2004		√	√	http://scholar.google.com
Google Patents	Google	2006			√	www.google.com/patents
BASE	Bielefeld University	2004		√		http://www.base-search.net
aRDi	WIPO/publishers	2009		√	√	http://www.wipo.int/ardi/en
PatentScope	WIPO	2003			√	http://www.wipo.int/patentscope/en
Global Patent Search Network	USPTO	N/A			√	http://gpsn.uspto.gov
Espacenet	EPO	1998			√	http://worldwide.espacenet.com
SIPO Patent Search	SIPO	N/A			√	http://211.157.104.77:8080/sipo_EN/search/tabSearch.do?method=init
FreePatentsOnline (FPO)	Patents Online, LLC	2004			√	http://www.freepatentsonline.com
SumoBrain	Patents Online, LLC	2007			√	http://www.sumobrain.com
Patent Lens	Cambia	2001			√	https://www.lens.org/lens
PriorSmart	PriorSmart	2007			√	http://www.priorsmart.com
iScienceSearch	AKos Consulting	2010	√	√	√	http://isciencesearch.com/iss/default.aspx
Commercial						
SciFinder	CAS	1995	√	√	√	http://www.cas.org/products/scifinder
STN	CAS-FIZ-Karlsruhe	1984	√	√	√	https://www.cas.org/products/stn
Reaxys	Elsevier	2009	√	√	√	http://www.elsevier.com/solutions/reaxys
OvidSP	Wolters Kluwer	2007		√		https://ovidsp.ovid.com
Thomson Innovation	Thomson Reuters	2007		√	√	http://info.thomsoninnovation.com
Web of Science	Thomson Reuters	1997	√	√		http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html
ProQuest Dialog (PQD)	ProQuest LLC	2013		√	√	http://www.proquest.com/products-services/ProQuest-Dialog.html
TotalPatent	LexisNexis	2007			√	https://www.lexisnexis.com/totalpatent/signonForm.do
PatBase	Minesoft Ltd.	2003			√	http://www.patbase.com/login.asp
Orbit	QUESTEL	2009			√	http://www.questel.com/index.php/en/product-and-services/prior-art-search
PatSeer	Gridlogics Technologies	2012			√	http://patseer.com
WIPS Global	WIPS Co.	2003			√	http://www.wipsglobal.com/service/mai/main.wips
JP-NET	Japan Patent Data Service	2007			√	http://www.jpds.co.jp/eng
Academic Search	EBSCO publishing	2007		√		https://www.ebscohost.com

[a]Struct = supports structure-based searches; J = journals; P = patents. [b]Fonger, G. C.; Stroup, D.; Thomas, P. L.; Wexler, P. TOXNET:A Computerized Collection of Toxicological and Environmental Health Information. *Toxicol. Ind. Health* **2000**, *16*, 4−6. [c]Previously UKPMC; N/A = not available. Prepared in November 2016.

descriptive patent title. As discussed below, these two major database services have chemical structure indexing, reaction indexing (SciFinder), and Markus indexing, thereby also allowing to perform structural searches. SureChEMBL[74,75] is included in Figure 1 as it is a freely accessible chemically annotated patent document database that enables full-patent searching against both structurally annotated (thereby also enabling structural searches) and unannotated patents. IFI claims[76] provide access to worldwide coverage full-text patents in XML format to different clients, such as SureChEMBL.[75]

1.2.3. Regulatory Reports and Competitive Intelligence Tools. Chemical information can also be found in regulatory documents distributed by government agencies: DailyMed[77] and New Drug Application (NDA) by FDA[78] or the European Public Assessment Reports (EPAR)[79] for marketed drugs in the United States and Europe, respectively. Information on clinical trials can be found at the Clinical Trials Web site[80] in the public domain. Additionally, Adis Insight[81] provides fee-based access to sound profiles of drug programs, clinical trials, safety reports, and company deals.

As mentioned above, Table 2 presents different public and commercial search engines and platforms connecting to the majority of the document repositories in Figure 1. Together with information on the proprietary/institution responsible for the tool and launch year, this table tells whether the platform supports structural searches as a result of the aggregation of the annotated/indexed document with a chemical database, as in the case of SciFinder,[44] Reaxys,[82,83] STN,[84] and Web of Science.[57] Notably, all four major standard tools are commercial, highlighting the need and interest in disposing of

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

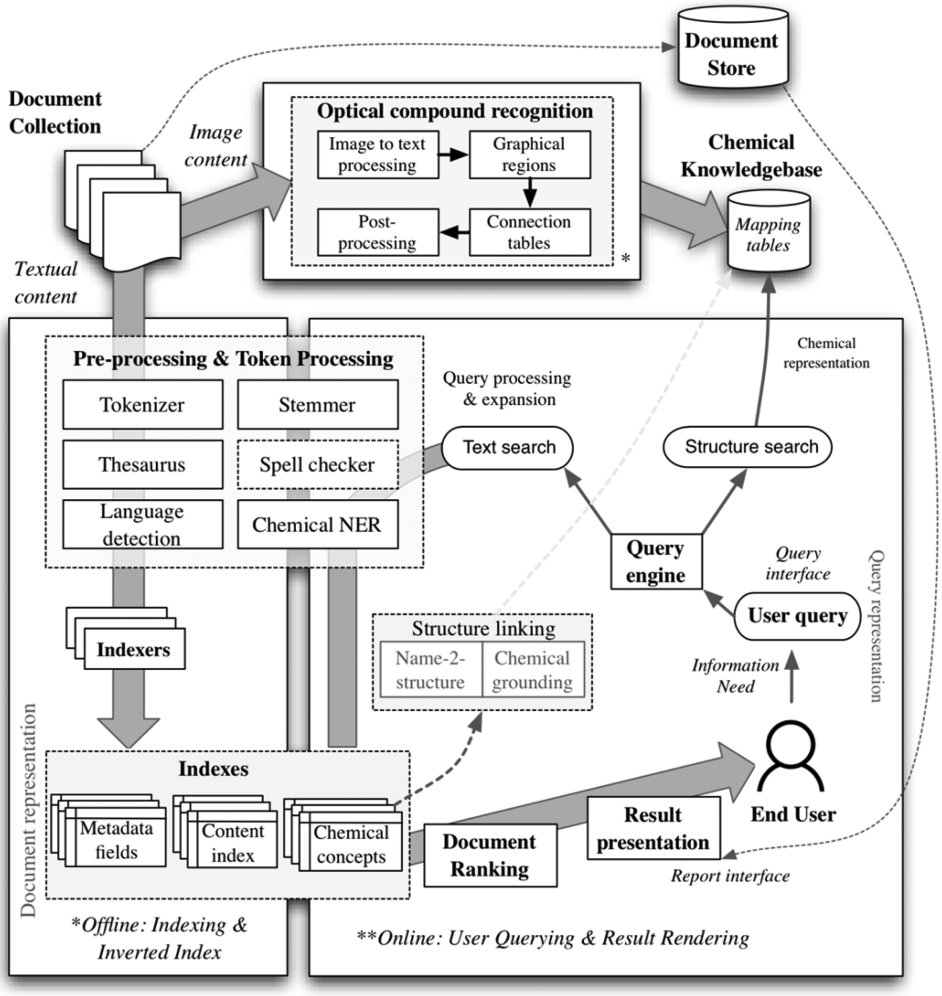


Figure 2. Simplified flowchart of chemical IR systems.

publicly available annotated databases, as was also demonstrated by the recent release of SureChEMBL, which was included in Figure 1 as a patent repository with its own access interface. iScienceSearch[85,86] is an open-access federated search service that allows searching over 86 free chemistry databases, scientific journals, and patents (e.g., using Google Patents[87]), although its retrieval capabilities are not comparable to those of SciFinder. Moreover, Scopus database (Elsevier) and Web of Science (Thomson Reuters) are commonly regarded as equivalent resources,[88] although the latter has the clear advantage of incorporating chemistry and reaction databases, besides being a bibliographic repository.

Commercial systems combine nonpatent literature from different sources and patent collections of one or more patent authorities (Table 2). Because of the vast number of connected databases (e.g., the case of STN[84] with over 180 databases or Web of Science[57] and Thomson Innovation),[89] both distributed by Thomson Reuters until very recently, and since October 2016, by Clarivate Analytics, the interested reader should refer to the link provided in Table 2 to gain deeper insight about their content. In the open-access domain, a few systems merge scientific literature with patents, Europe PubMed Central (Europe PMC)[90,91] and Google Scholar,[92,93] although a commonly criticized aspect of Google Scholar is that it does not provide a clear criteria on the selected journal publications that are indexed.[94] Interestingly, the Access to Research for Development and Innovation (aRDI)[95] program facilitates access (free of charge or with a nominal fee) to scholarly journals from diverse fields of sciences for patent offices and scientific institutions in developing countries. For patents alone, WIPO and national and regional offices (in Table 2 only patent offices with the largest collections WIPO, EPO, USPTO, and SIPO are listed) as well as public free-of-charge patent database producers (e.g., FreePatentsOnline,[96] Patent Lens,[97] and PriorSmart[98]) offer a number of search interfaces. For commercial, private sector databases, only those more commonly used are shown. Recently, closed free-of-charge services, such as Scirus,[99] provided by Elsevier, and Microsoft Academic Search[100] are not tabulated.

Apart from all of these repositories, ontologies and thesauri play a key role in chemical information retrieval, especially in the context of the Semantic Web and query expansion (section 2.4.3) and in CER strategies based on dictionary lookup of chemical names (section 3.4) as they provide standardized explicit descriptions of concepts and entities, providing wealthy vocabularies to index different terms.

2. CHEMICAL INFORMATION RETRIEVAL

The concept information retrieval was first introduced by Mooers in 1950,[101] although it was only in the 1970s when full text analysis and document indexing became common. Before that, most searches could be considered metadata searches,

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

querying for well-defined fields such as title, authors, or keywords. Research related to chemical IR systems had an important initial event in 1951, when IBM did a first presentation on an electronic information-searching machine based on a punched-card equipment for coding and sorting cards that allowed a search rate of 1000 cards per minute. This system was presented back then to the American Chemical Society Committee on Scientific Aids to Literature. Sanderson and Croft provided a short historical description of IR systems, introducing also mechanical and electro-mechanical devices for searching entries in catalogues and early preweb computer-based retrieval systems.[102]

In the following, we provide a comprehensive overview of the core concepts of information retrieval technologies with a focus on those aspects that are specific to chemical search applications, because a deep review of all aspects of information retrieval systems is out of scope.[11]

This section is organized according to the common components of an IR system (Figure 2): (i) unstructured data repositories and characteristics (section 2.1), (ii) document preprocessing to enable effective indexing (section 2.2), (iii) the generation of a new separate representation of the documents optimized for retrieval (section 2.3), (iv) query construction, understood as the actual expression of the user information demand that is provided as an input to the IR system (section 2.4), and (v) result representation, which is the collection of documents returned by the IR system, often presented as a ranked or sorted list of query hits or alternatively as a classification of documents (section 2.5). Figure 2 illustrates a simplified flowchart of chemical IR systems.

2.1. Unstructured Data Repositories and Characteristics

Chemical IR systems can be characterized according to the underlying data sources or document repositories processed.

The hits returned by IR systems are tightly linked to the indexed data and document repositories. As existing online chemical document and literature databases are directly coupled to a user query interface, the practical distinction between databases and search engines at the application level becomes blurry for these complex document retrieval systems. The document collection, referring to the set of documents that are searched/processed by the retrieval system, might be either an in house collection or documents hosted and accessed by some of the existing online retrieval tools. Good understanding of the primary information landscape is crucial to determine which collection of journals and other document types might be relevant (see section 1.2).

For mining in house chemical documents, the combination of existing relational database management systems (e.g., PostgreSQL[103]), modules for handling chemical queries and integrating structural data (e.g., RDKit[104]), and indexing and retrieval components like the popular Lucene library[105] or Lucene-based systems like ElasticSearch[106] can be used in prototypical in house retrieval settings.

The hits returned by chemical search engines might vary due to intrinsic differences in the underlying document databases, and sometimes document collections need to be revised to describe the queries on the basis of documents and their specialized subdomains. For instance, SciFinder, Inspec, Compendex, Web of Science, Scopus, and PubMed index and abstract journal literature, but they show differences in terms of subject coverage for the indexed journals,[12] and therefore, for certain search types, the combination of multiple retrieval tools might yield more exhaustive or relevant results.[107] Typical literature aggregators like PMC Europe address the storage of content from multiple journal sources.[59]

2.2. Preprocessing and Chemical Text Tokenization

Preprocessing of unstructured documents is a key step for both (i) subsequent IR and (ii) in the context of TM technologies, particularly CER.

In practice, chemical documents are available in a range of different input formats, but they are mainly distributed either as electronic text or as image files. With respect to electronic text documents, unfortunately, they are often not directly available as plain text files but do correspond to PDF (portable document format), HTML (HyperText Markup Language), XML (Xtensible Markup Language), or other common file formats.[108] The first step is thus to convert those files into a format that can be better processed by TM software, which is normally plain text. This document transformation step, sometimes also referred to as the document standardization process, may require the conversion of PDF files into plain text, or to select the actual running text content from metadata contained in HTML or XML tags. For both PDF and HTML input, there is a range of open source as well as commercial software tools available, including PDF text parsers like PDFBox,[109] pdftotext, IntraPDF,[110] PDFTron,[111] UTOPIA,[112] and ABBYY PDF Transformer.[113] Since the introduction of the PDF format in 1993, scholarly articles have been increasingly distributed in this format, and it has become nowadays the most commonly used file format for online scientific publications, being a sort of de facto standard format for scientific communication together with HTML. To be able to extract correctly the text content from PDF files in a layout-aware manner, and identify correctly blocks of contiguous text, is still an ongoing field of research.[114] Among the most common errors occurring in PDF transformation are misplaced paragraph separations, sequential layout of tables, and wrong sentence handling of two-column layout articles. Moreover, in the case of scholarly literature, considerable formatting changes may be observed by examining articles published between 1966 and 2007, which represents an additional hurdle for layout-aware PDF transformation.[114]

2.2.1. Document Transformation. The document transformation step is usually carried out as an offline process through analysis of static documents. When text is contained in an image (e.g., derived from scanned papers used to digitize printed documents), optical character recognition (OCR) software, such as Tesseract[115] or CuneiForm, is used to transform images into machine-encoded text. Especially, full text patent documents are often available as images of text documents.[116]

Errors during this step are critical for downstream text processing in general, and particularly in the case of chemical documents. For instance, differences in just a single character between two chemical names result in associated structures that are very different, as in the case of "methylamine" and "menthylamine".[117] Also, the presence or absence of a single space between two chemical word tokens can result potentially in different interpretations of chemical mentions, for example, "methyl ethyl malonate" versus "methyl ethylmalonate".[118] Noisy chemical text documents, corresponding to text generated from scanned image files of patents or historical scientific literature, present additional challenges due to the presence of spelling errors, typographical errors (typos), space

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

errors, truncated words/names, and OCR errors. Letters that resemble numeric characters often result in OCR mistakes, for example, the symbol aluminum Al and its wrong conversion to A1.[108] This type of error is called homoglyphic substitution and occurs when two characters are very similar (e.g., 1 and I, O and 0).[119] In the case of human typos, the distance between characters in a QWERTY keyboard may be a useful strategy to detect potential sources of errors commonly called fat finger syndrome.[119]

OCR software is known to work well for most fonts and is capable to return formatted output that resembles the original page layout for most scenarios. In the case of chemical documents, among frequently encountered errors are wrong line and word detections (word boundary recognition errors) and issues resulting from special characters used in chemical language. This is partially due to the fact that OCR software commonly uses dictionaries of words for the characters segmentation step, and within chemical documents there is a great number of specialized, technical terminologies that do not match dictionary entries. Some attempts have been carried out to detect and correct errors due to OCR failures, or even human spelling errors, encountered in chemical texts.[119]

Even though a considerable fraction of chemical literature and patents is available in English, it is useful, when processing large heterogeneous document sets, to sometimes run language detection tools on the document contents to determine the actual language in which the documents have been written. This can be achieved by applying language detection software, for instance, the widespread Tika tool that can detect 18 different languages.[120]

2.2.2. Character Encoding. A very basic but important aspect that should not be neglected, especially when trying to integrate results from various text processing platforms, or when text is sequentially handled by different modules or systems, is character encoding, that is, the way text characters are represented. The use of character encodings that correspond to internationally accepted standards allows more efficient interchange of text in electronic form. Checking the type of encoding used in text documents and whether the program of choice supports it is one of the first issues that needs to be examined when applying TM strategies. The underlying representation of text by computers is done in the form of binary data.[121] This implies that all of the characters inside text documents have to be represented by numeric codes, or, in other words, they are stored using a particular type of character encoding (generally encoded as bytes). One simple and popular format is ASCII (American Standard Code for Information Interchange, also called US-ASCII), introduced in 1963,[122] which consists of a seven-bit encoding scheme that can be used to encode letters, numerals, and some symbols. ASCII is supported by nearly all text editors and was the most widespread encoding of the World Wide Web until 2007, when it was surpassed by the UTF-8 (8-bit Unicode Transformation Format) encoding,[123] which supports a larger set of characters. UTF-8 is also among the preferred encodings for chemical documents, and it is advisable to make sure that within a given collection all documents use the same encoding. Many literature repositories, such as ScienceDirect, support the UTF-8 character set, making it possible to use search queries that contain UTF-8 characters. Specifically, by using the UTF-8 encoding, it is possible to represent most chemical names, formulas, equations, notations, and expressions, including characters such as Greek letters, subscripts, superscripts, and nonalphanumeric chemically relevant characters (e.g., hyphens, bullets, arrows, daggers, plus/minus signs, and symbols to denote stoichiometric relation, net forward reactions, reactions in both directions, and reactions in an equilibrium).

Another characteristic that should be observed carefully when performing text processing at the level of characters and letters relates to ligatures. Ligatures typically refer to symbols that represent the fusion of more than one character and can thus be regarded as being a sort of character combination or conjoining of letters. Moreover, ligatures are also a common source of OCR errors. Depending on the used representation/encoding model, ligatures can be considered as a single character (composed form), or they can be decomposed into a set of separate characters (decomposed form) resulting in normalized ligatures.

2.2.3. Document Segmentation. For some TM applications, the use of the entire chemical document, regardless of its underlying internal structure, is not practical, requiring document segmentation as a step to handle the documents more efficiently. This is particularly the case when considering lengthy documents such as entire patents or full text scientific papers. Segmenting documents into various sections and identifying chemical entities within those sections enables a more focused contextual search, for instance, by facilitating search constraints that limit hits to figure captions,[116] something supported by tools such as CLIDE.[124,125] Some sort of enhancement of document semantics can be obtained by exploiting document structure and defining structurally relevant units like sections and paragraphs.

Records structured by markup languages like HTML or XML are generally easier to process, as they already often provide explicit tags that can be used to identify section and subsection headers. XML-formatted patent documents from the EPO and USPTO employ explicit delimitation of the major sections and headings of the documents, overall following a patent structure defined by the Common Application Format (CAF).[38] CAF specifies some basic patent anatomy, such as patent abstract, claims, and description sections. Nevertheless, the identification of section and paragraph boundaries within the body of patent documents is not an easy task due to under-specification and the use of implicit document structures that are not directly machine interpretable. Attempts to automatically determine where sections in the patent begin and end have been done, for instance, by matching regular expressions (*regexps*) based on those headings.[38]

Regular expressions are also used for document segmentation purposes of scientific literature, which follows a more general organization (section 1.2). Beyond segmenting documents into main sections, only limited research has been carried out to process chemical literature to detect topically coherent multiparagraph segments,[126] a process called TextTiling.[127] This requires defining a common schema to represent the structure of scientific articles as well as the exploration of the discourse structure of papers.[126]

A very valuable resource for chemical information are tables, which may contain both textual data as well as structure diagrams.[35] Mining tables is still a very preliminary field of research, and despite its importance, there are only few published systems that can handle tables, for instance, the Utopia application.[112] Most of the noncommercial chemical TM software neglects table processing.

2.2.4. Sentence Splitting. Document segmentation can be viewed as a coarse level text processing step. At a more granular

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

level, written chemical documents often need to be divided using text segmentation methods into smaller subunits (tokens) such as sentences and words. Sentence boundary disambiguation (SBD, also known as sentence splitting, sentence tokenization, or sentence segmentation) is a low-level text processing step that consists of separating running written text into individual sentences. Sentences are a foundational unit for most natural language processing (NLP) pipelines, such as assignment of part-of-speech labels to words (POS-tagging), syntactic and semantic processing, or even machine translation. They form logical units of thought. When text is not properly delimited into sentences, the resulting errors propagate upward in the text processing pipeline. In particular, chemical named entity recognition (discussed in section 3) is very sensitive to SBD errors.

Automatic SBD methods are usually rule-based or rely on data-driven machine learning (ML) techniques trained on manually tagged sentence boundaries, although some attempts have been made to also apply unsupervised methods and exploit syntax-based information using POS labels.[128] In English and most other Indo-European languages, using punctuation marks, particularly the full-stop dot character is a reasonable approximation for a very crude SBD strategy. Nevertheless, it is worth noticing that not all written languages contain punctuation characters that could be exploited for approximating sentence limits. One of the most basic rule-based SBD methods, often applied to newswire documents, defines sentence tokenization patterns as periods followed by an upper case letter (and not followed by an abbreviation). Regular expression-based sentence tokenizers have also been tested in the context of patent documents.[129]

Punctuation marks do not always correspond to sentence boundaries; they show ambiguity, for instance, with initials, numbers (decimals, floating points), personal titles, ellipsis, delimiters (e.g., filename extensions, URLs, e-mail addresses), bibliographic references, and especially abbreviations.[130−132] In the case of chemical texts, punctuation marks can also be encountered inside chemical entity mentions (e.g., "$ZnSO_4 \cdot 7H_2O$", "1,3,8-triazaspiro[4.5]decan-4-one", or "2,5-diazabicyclo-[2.2.2]-octane") and mentions of genes/gene products (e.g., "TNF.alpha" or "Kv3.1 channel"), enzyme codes (e.g., "EC 3.4.14.5"), or chromosome locations (LEN.PK113-78). A narrow definition of SBD would consider only the disambiguation of full stop characters as either being a sentence delimiter or not (splitting at a closed set of special characters), while a broader definition of the SBD task would examine every character as a potential sentence delimiter.[128] For more formal written text, such as scientific abstracts and articles, a narrow SBD definition is usually competitive enough, while for spontaneous language, web content and noisy texts like scanned patents, less formal chemical documents, and electronic health records, it is common to encounter missing punctuation marks, and thus a broader SBD definition might sometimes be more appropriate.

Machine learning (ML) approaches have increasingly become the method of choice for many text classification tasks (section 2.5). In the case of SBD, the problem can be viewed as a binary classification task, which requires determining whether a given character in running text does correspond or not to a sentence delimiter. For supervised ML techniques to work well, manually annotated sentence boundaries are required as a training data. From a given set of training examples, a statistical model is then learned and subsequently applied to assign labels to previously unseen data.

For SBD, commonly used annotated corpora in the biomedical domain are the GENIA[133] (16 392 sentences), the PennBioIE[134] (23 277 sentences), and the JULIE[135] (62 400 sentences) corpora. The JULIE corpus also provides more detailed guidelines for the manual annotation of sentence boundary symbols (SBS). Proper guidelines together with corpora for the annotation of chemical document sentence boundaries are not available, and thus most of the existing chemical text processing pipelines use either SBD tools developed using domain-independent data sets or are based on SBD systems tuned for biomedical literature. Overall, SBD algorithms work very well on scientific articles, and even though there are differences in terms of performance depending on the used tool and evaluation corpus, the variability in performance is rather low. Experiments done to evaluate the performance of SBD tools against the GENIA corpus sentence boundary annotations yield a F-score (harmonic mean of precision and recall, described in section 2.6) between 98.3 and 99.6.[128] Among the SBD tools adapted to scientific literature are the LingPipe sentence chunker,[136] the GENIA sentence splitter (GeniaSS),[137,138] and Med-Post.[139] The JULIE sentence boundary detector (JSBD)[135,140] has also been used by both chemical[141] and biomedical text processing pipelines. It yields an F-score between 99.58 and 99.62, depending on the used evaluation corpora. JSBD is based on a ML algorithm very popular for labeling text, called conditional random fields (CRFs),[135] which will be discussed in more detail in section 3.6. Another widely used sentence detector is distributed as part of the openNLP toolkit,[142] and it has been adapted to split sentences derived from chemical patent abstracts.[143] Domain adaptation is worthwhile to obtain a more competitive result for SBD systems, as has been shown for the LingPipe sentence splitter evaluated using the GENIA corpus.[128]

Among common error sources of current SBD tools are the lack of an exhaustive examination of the range of Unicode characters that encode for punctuation marks, the inherent variability of the various text types, and the fact that some SBD tools ignore paragraph boundaries.[128] Applying some simple rule-based pre- or postprocessing steps (e.g., checking balancing of opened and closed parenthesis) can result in performance gain for some SBD systems.

Formal scientific language is characterized by the use of long and complex descriptive phrases, which do represent a challenge for some NLP tasks. Rule-based approaches have been explored for automatic sentence simplification to generate simplified text by exploiting syntactic clues, such as coordinations, relative clauses, and appositions. Sentence simplification has been particularly useful for improving the performance of relation extraction systems (see section 6). Those techniques have been effective for detecting protein−protein interactions from text[144,145] or drug resistance information.[146] bioSimplify[147] and iSimp[148] represent two popular sentence simplification approaches applied to biomedical literature, while the Cafetiere Sentence Splitter has been tested on chemical abstracts.[149,150]

2.2.5. Tokenizers and Chemical Tokenizers. The most critical text preprocessing step is usually tokenization (word segmentation). It consists essentially of the problem of dividing each sentence or string of written language into its constituent tokens (i.e., component words, numbers, punctuations, expression sequences) and therefore requires detecting where

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

word breaks exist. Word tokenization is usually carried out after the SBD step. Tokenization has a deep impact on tasks such as POS tagging (an important task for the recognition of named entities) as well as for text indexing and information retrieval. Splitting of written text into a sequence of word tokens might at first sight seem a trivial undertaking, but tokenizers face several challenges. One of the tokenization challenges is language related, while among other difficulties one can encounter domain-specific language issues. Some scientific domains demand the use of special tokenization approaches, because they contain chemical or mathematical formulas, or have entity names showing internal naming structures like systematic chemical compound names.[151] Although most of the chemical TM work (and as a matter of fact, TM and NLP work in general) has been carried out using English language texts, there is an increasing interest in processing chemical documents in other languages, for instance, the growing number of chemical patents written in Chinese. Unlike Western languages, major East Asian Languages (e.g., Chinese, Japanese, Korean, or Thai) are written without spaces between words. This implies that for these languages it is necessary to run a word boundaries detection program prior to any word-based linguistic processing attempt.[108] In the case of Chinese word segmentation, common tokenization strategies either make use of large vocabulary resources, taking the longest vocabulary match to detect word boundaries, or they apply supervised ML methods trained on manually tagged word boundaries to reach satisfactory results.[152]

Other languages such as German also show particular tokenization intricacies due to the usage of compound nouns without spaces, which can have an effect on the performance of information retrieval systems. To handle compound nouns, a common approach is to apply compound-splitter modules that determine whether a given word can be segmented into multiple subwords that, in turn, appear in a predefined vocabulary list.

Chemically aware text-tokenization approaches have been studied in more detail for documents written in English. For general English texts, tokenization is often done by exploiting punctuation marks and whitespaces or by simply splitting on all nonalphanumeric characters, often also requiring some additional preprocessing to handle apostrophes used for possession and contractions. Efficient tokenization in the case of chemical texts shows considerable differences when compared to general purpose tokenizers.[119] In the case of chemical names and documents describing chemical entities, it is important to take into consideration that chemical names do contain whitespaces, commas, hyphens, brackets, parentheses, digits, and also apostrophes. Therefore, chemical text tokenization is a more demanding process that requires using specially adapted tokenizers able to cope with the peculiarities of chemical expressions, complex naming conventions, and domain-specific terms.

The output of various tokenizers can be greatly different, for instance, depending on how characters such as hyphens are being handled. It has been observed that in biomedical documents, symbols that usually correspond to token boundary symbols (TBS), such as + '/+ %, do not always denote correct boundary elements.[135] Parentheses represent another character type that, while in normal running text do correspond usually to TBS, in the case of chemical texts they are part of the chemical name and require special treatment. Studies have been carried out to make it easier to choose suitable tokenizers by comparing various tokenization algorithms on PubMed abstracts.[153] Especially, hyphens represent a tokenization challenge for chemical texts as they can appear within a chemical name (e.g., "*tert*-butyl peroxide") and in other cases occur between different entities (e.g., "hexane-ethyl acetate"), corresponding to true TBS.[154] Hyphens are also common within chemical expressions and formula ("C−H") and thus require custom chemical tokenization.[155] Word-boundary hyphens were defined by Zamora et al. only as those hyphens that were flanked by alphabetic characters.[156] Corbett and colleagues, in turn, exploited a list of strings that corresponded to nonword boundaries if they were found before a given hyphen (e.g., "*tert*-") with the assumption that they were part of chemical names. They also used certain strings that characterized word boundaries only if they occur after hyphens (e.g., "-induced").[154]

ML-based token boundary detection has also been implemented for biomedical literature.[157] Token boundary detection using the conditional random field algorithm has been used to tokenize PubMed abstracts[135] by training the classifier on semantically motivated word boundary annotations. This resulted in the JULIELab tokenizer module.[135,140]

Currently, the most widely used chemical text tokenizer is the OSCAR4 tokenizer, which employs segmentation rules specifically constructed for chemical texts.[158] For instance, it was used by the ChER chemical mention tagger developed by Batista-Navarro and colleagues.[150]

Another chemical text tokenizer that uses manually defined rules is ChemTok, relying on the examination of the BioCreative CHEMNDER task data set (described in more detail in section 3.8).[159] Dai et al. evaluated both a more fine-grained and a coarse-grained tokenization in the context of chemical entity mention recognition and concluded that a more granular tokenization resulted in better performance for their task.[160] Other chemical text tokenization modules are part of the tmVar module adapted by the tmChem chemical mention tagger[161] and the ChemSpot tokenizer.[162]

2.3. Document Indexing and Term Weighting

Two key initial aspects underlying IR systems are the definition of the document units (i.e., what constitutes a document) and the logical view of how documents are represented internally (i.e., the text representation model). Choosing the appropriate document unit is important for search engines in terms of how granular the returned hits will be. For instance, depending on the underlying end user requests, large documents such as chemical books or thesis could potentially be segmented into mini-documents comprised of individual chapters. Likewise, in the case of separate files, such as a scientific article and its Supporting Information, those files could be merged into one larger document for retrieval purposes.

The entire set of text items, that is, documents on which the search will be performed, is commonly known as document collection (also sometimes referred to as corpus, or body of text). Documents can in principle be represented just as a consecutive stream of plain text characters, which are then searched sequentially through linear scanning of the query text characters against target documents. Command line utilities such as *grep*[163] enable string matching or regular expression based search approaches, including case-sensitive or -insensitive matching and global wildcard pattern matching, usually returning the matching lines. These linear scanning approaches are considered useful when dealing with small or medium sized

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

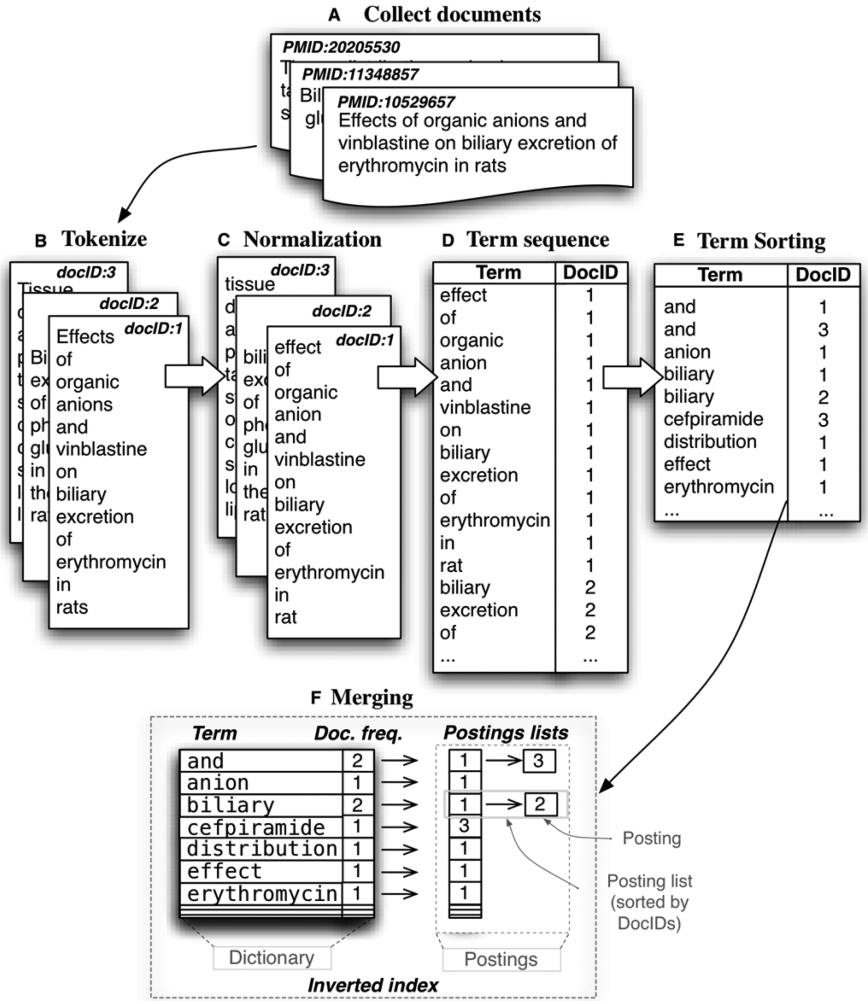


Figure 3. Simplified example process for building an inverted index.

document collections and semistructured or structured database contents. However, linear scanning presents serious efficiency limitations when dealing with large data sets.

Term indexing is considered the best solution to process large document collections, allow more flexible matching operations, and support ranked retrieval. This approach consists of viewing the text of the document as a collection of indexing terms, also sometimes called keywords.[11] IR systems index the documents in advance to generate a term-document incidence matrix, where terms, usually words (the results of the previously described tokenization process), correspond to the indexed units. So, in this representation model, documents are viewed as a set of words. Figure 3 illustrates a simplified example process for building an inverted index. As was already described, tokens correspond to the sequence of characters that are grouped together after the tokenization step and denote the basic semantic unit for document processing. That is, they instantiate sequences of characters in document. All tokens that display the same sequence of characters are grouped into a type (or token type). For instance, let us assume that within a given article there are three mentions of the word "sulfobromophthalein". Each of the individual occurrences would correspond to the "sulfobromophthalein" tokens, while the actual unique string, regardless of the mentions within the documents, would correspond to the token type "sulfobromophthalein". Finally, a term is a (token) type that is incorporated in the dictionary of the IR system. Commonly, IR systems do not use the tokens directly as they appear in the documents. Instead, they carry out a normalization process of the token types to improve retrieval efficiency. In particular, the same tokenization and term normalization process is carried out on both documents and user query words to guarantee that a potential match can be detected. Various normalization procedures will be detailed later in this section.

In the previously introduced term−document incidence matrix, each row corresponds to a particular term, while each column corresponds to a document. This implies that for each term t, there is a vector of document occurrences, and, conversely, for each document d there is a vector representing the terms that are found in this document. In the most basic, binary term−document incidence matrix, term matches in a particular document are recorded as 1; otherwise, a zero is stored. A binary term−document incidence matrix representation model for large document collections does result in a considerably large and very sparse matrix; that is, most of the values would correspond to zero. This phenomenon can be explained in part by a very well-known statistical property of human language known as Zipf's law,[164,165] which holds true for general language documents as well as domain specific collections.[166,167] In essence, Zipf's law states that for a sufficiently large document collection, the frequency of use of a

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

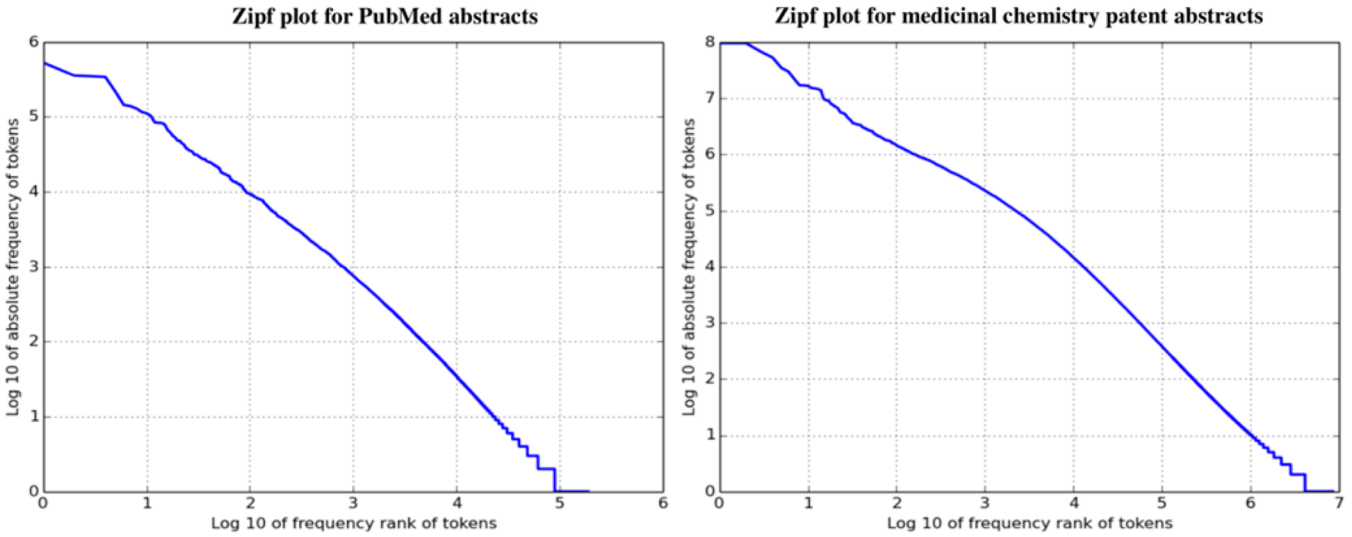


Figure 4. Example Zipf's plot for scientific abstracts and medicinal chemistry patent abstracts.

word n is inversely proportional to its rank r, that is, the position of that word when sorting all words by their absolute frequency. The mathematical representation of this characteristic would correspond to $n \propto 1/r$. This implies that the most frequent word appears approximately twice as often as the second most frequent word, and so forth. A more general formulation of Zipf's law incorporates an exponent α, that is, a parameter with a value usually close to one, resulting in a rank-frequency relation that takes the form of a power law: $n \propto 1/r^{\alpha}$.

For large document collections, most words occur only a few times, and about 40−60% of the words are estimated to occur just once.[168] One common explanation for this language characteristic is based on the principle of least effort, where the speaker (the person or system coding the information) and the listener (the person or system trying to decode the information) try to minimize the effort needed to reach understanding through a tradeoff between the usage of nonspecific/general terms (those that can be easily retrieved but are highly ambiguous) and very specific terms (those that require more effort to be selected, but are less ambiguous and more precise). Figure 4 shows an example Zipf's plot for scientific abstracts and medicinal chemistry patent abstracts.

As a large fraction of words appear at a low frequency, and many often only occur once, a more compact data representation is usually used to capture term−document associations; that is, only the actual matches of terms in documents are stored. This results in a so-called inverted index (also known as inverted file), which comprises for each term a vector corresponding to the set of documents where it occurs. The data structure used by IR systems associated to the resulting collection of terms is commonly called dictionary, while the actual set of terms is usually known as its corresponding vocabulary.

For the sake of efficiency, documents are usually identified in the inverted index by unique serial document identifiers (doc IDs), which are assigned to them during the index construction by simply using successive integers for each new document that is added. The list of documents associated to a given term is called a postings list, while an individual document in this list is named posting and the entire collection of all posting lists that are part of an inverted index are known as postings. In a nonpositional inverted index, postings are just the document identifiers (term−document identifier pairs), while in a positional inverted index, information relative to the position of the term in the document is also stored. The terms comprised in the dictionary are then sorted alphabetically; duplicate terms are merged into a single unique term and, in the posting list, the document identifiers are sorted incrementally. For each dictionary record, the absolute frequency of the term (document frequency corresponding to the length of the postings list) is usually stored. Modern IR systems usually also store within the inverted index a list of all occurrences of the terms (i.e., positions in text) in each document (positional index), which is critical to allow phrase and proximity searches (section 2.4). An exhaustive description of the various indexing and index compression methods is beyond the scope of this Review, and thus only the core concepts relevant to understanding the underlying methodological infrastructure of chemical text search engines are provided.

The tokens, before being added to the term list, are usually normalized; that is, a linguistic preprocessing step is carried out to generate a modified token representing the canonical form of the corresponding term. Typically, this step refers to either stemming or lemmatization. The reduction in size of the term dictionary significantly depends on how rich morphologically is the target language. For instance, Spanish texts are morphologically richer than English, and thus, when applying stemming, the resulting vocabulary reduction is greater than in the case of English texts. The reduction of the dictionary size helps improve the chemical IR systems in terms of processing time and memory, as well as increase recall. However, a detailed analysis of the effect of linguistic preprocessing in chemical IR has not been carried out so far.

For an inflected or derived word, stemming programs, also known as stemming algorithms or stemmers, output its corresponding stem.[169] Word stems often (but not always) correspond to the word base form or morphological root. For example, for the word list "oxidable", "oxidate", "oxidation", "oxidatively", "oxidize", and "oxidizing", the Porter stemmer returns the word stem "oxid". By grouping those word variants under a common stem, the underlying assumption is that those variants in practice should be semantically related (have a similar meaning) and can therefore be used by search engines

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

as synonyms for query expansion (section 2.4.3), a process named conflation. Stemmers usually apply language-specific rules to generate the word stem. Stemming can be based on simple lookup tables of correspondences between inflected and word stems, apply suffix stripping rules, or even explore statistical language analysis techniques. English stemming algorithms have been implemented since the 1960s.[170] The most popular English stemming algorithm is the Porter Stemmer,[171] which has been applied not only to general English texts but also for the processing biomedical, chemical, and scientific literature,[161,172] eventually incorporating domain-specific rules.[173,174] Stemming programs present two common types of errors that can be problematic when processing chemical texts: one is overstemming (i.e., words that have different meanings are grouped together through a common stem) and the other is understemming (i.e., words that are inflectional/derivational variants of the same base form are not linked to the same stem). In the case of chemical named entity recognition strategies (section 3), there is some evidence that applying stemming is detrimental for the performance of systems identifying chemical mentions.[175] One possible explanation for this characteristic is that suffixes can be suitable cues to determine if a word corresponds to a chemical entity. Indeed, suffixes and prefixes are usually examined by chemical-text aware hyphenation systems, which apply heuristic rules for short hyphenated suffixes/prefixes and only carry out token splitting for longer word forms.

A more sophisticated alternative to stemming is to apply lemmatization algorithms. Usually, lemmatizers do not operate just on single words as do stemmers, but they also take into account the sentence context and part-of-speech information to return the linguistic base form (lemma) of an inflected word. The pair formed by the word base form and its corresponding part-of-speech is called lexeme. The biolemmatizer is a domain-specific tool that is able to process biological and biomedical documents mentioning chemicals. It was able to achieve an F-score of 96.37% when evaluated against a gold standard of manually labeled life sciences full text articles.[176]

Other strategies to normalize texts for indexing purposes include case-folding and spelling normalization. Case-folding stands for the process of converting all letters of a token into lowercase letters. This process can generate ambiguous words for proper nouns or person names. In the case of the English language, spelling normalization refers to the conversion of spelling variations, that is, British and American spelling, into one single spelling type.

Reduction of the size of the inverted file index can be achieved by grouping morphological word variants, but also through removal of noninformative words (words with low discrimination power) that do not contribute to the retrieval of relevant documents. Such words, known as stop words, usually correspond to prepositions, articles, or determiners. For most search applications, precompiled lists of stop words are used, although sometimes high frequency words are inspected manually to generate a more tailored set of stop words. However, and because exclusion of stop words from the dictionary might affect phrase searches (see section 2.4), not all IR systems are able to actually eliminate stop words.

Linguistic preprocessing approaches can also be viewed as a sort of lossy compression approach,[177] a concept more commonly used for textual images, and which refers to a reduction of vocabulary that results in a more compact document representation with the cost of losing some marginal information, which in principle does not affect retrieval efficiency noticeably.

2.3.1. Term Weighting. A typical chemical text search engine not only needs to retrieve the documents that mention terms matching the user query, but additionally it should be able to rank or order the returned document hits efficiently, that is, returning the most relevant documents on the top of the result list. To meet this goal, it is fundamental to weight or score the importance of terms on the basis of statistical attributes that model the discriminative power of the terms. A detailed description of term weighting schemes is beyond the reach of this Review, but both *tf* (term frequency) and *tf-idf* (term frequency-inverted document frequency), which are two widespread term weighting approaches, will be introduced here.

The most simple term weighting is called term frequency $tf_{t,d}$, consisting of the number of occurrences of term t in document d, that is, its raw term frequency. Documents are represented by the set of contained words without acknowledging word ordering. This document representation form is known as a bag-of-word (BOW) document model, and it is used for IR as well as for document classification and document clustering purposes (see section 2.5). Raw term frequencies are not sufficient to determine the discriminating power of terms, and therefore some additional weighing factors are being used to scale down the weight of terms using mechanisms that go beyond the level of a single document. The most common document-level statistic is the document frequency df_t, consisting of the number of documents in the collection that contain a particular term t. To return a higher weight for relatively rare terms as opposed to very frequent terms in the document collection N, the inverse document frequency idf_t of a term t is used as follows:

$$idf_t = \log \frac{N}{df_t} \tag{1}$$

The $tf_{t,d}$ and the idf_t are combined into the so-called *tf-idf* weighting scheme to generate a score that down-weights terms that are very frequent in the entire document collection: $tf - idf_{t,d} = tf_{t,d} \cdot idf_t$.

Documents (and also the user query) can be seen as a vector with one component corresponding to each dictionary term together with its corresponding *tf-idf* score. The score for a document d would then be the sum over all query terms q:

$$\text{score}(q, d) = \sum_{t \in q} tf - idf_{t,d} \tag{2}$$

Representing the set of documents in a collection as a set of vectors in a common vector space is called the vector space model, and it is widely used for free text retrieval.[178] Documents are represented as t-dimensional vectors in term space (t corresponds to the vocabulary size), and the query is treated as a short document (i.e., a set of words without specifying any particular query operators between the individual words). The sequential order in which terms appear in the documents or the query is lost when using the vector space representation. Vector operations are then used to compare documents with queries. The standard way to quantify similarity between documents and between a document and a query is to calculate the cosine similarity[179] of their vector representations (defined by distance in vector space using the cosine of the angle between the vectors), as shown in Figure 5. The numerator of the similarity equation is given by the inner

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

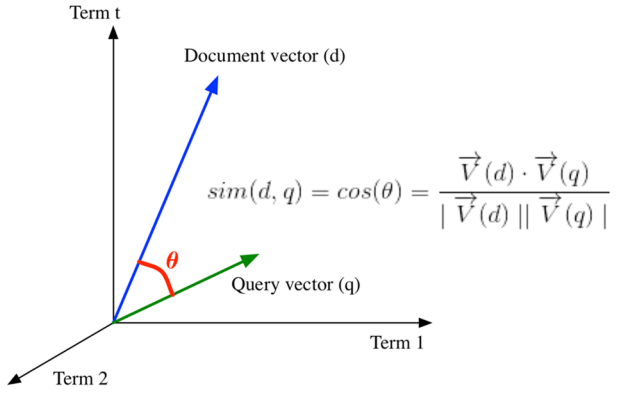


Figure 5. Cosine similarity.

product of the document and query vectors (dot product), while the denominator is essentially the product of their Euclidean distances. After computing the similarity scores between the query−document pairs, the documents are ranked by decreasing similarity scores (cosine scores) with respect to the user query. The popular Apache Lucene indexing and search technology[105] supports this model representation and cosine similarity scoring.

2.4. Information Retrieval of Chemical Data

Figure 6 illustrates the basic IR cycle. There are different ways of characterizing chemical IR systems. They can be examined in

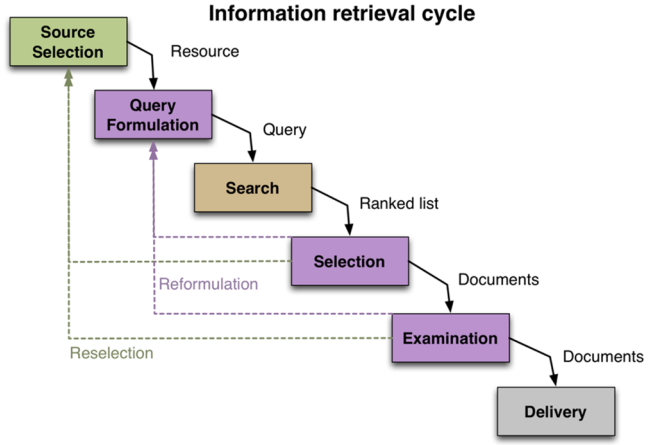


Figure 6. Basic IR cycle.

terms of (i) underlying data sources or document repositories processed, (ii) types of search query specifications and options supported, (iii) the system performance when evaluating retrieval efficiency, and (iv) the document ranking methodologies used. Points (i) and (iii) are addressed in sections 2.1 and 2.6, respectively. This section mainly focuses on the kind of search requests (ii).

Before moving into the kind of search request, a few general aspects of IR systems must be addressed.

When considering the actor launching a query against the IR system, we need to distinguish between machine-focused and human-focused IR infrastructures, with machine-focused retrieval being primarily characterized by the automatic transformation, classification, or processing of collections of unstructured data (typically documents of running text) into some sort of structured or labeled data, for example, chemical entries indexed and organized into a chemical knowledgebase or labeled/grouped document sets. In turn, human-focused information retrieval requires manually defined search queries provided by the information-seeking user.

Retrieval systems can also be divided into those that are characterized in terms of short- versus long-term representation types of user information needs.[180] For somewhat static, long-term information demands, a common approach is to use filtering systems.[181] Basically, the user has a predefined information model about the kind of documents that considers relevant and wants to retrieve new relevant documents as they enter the document repository. Filtering can also be regarded as a classification problem where relevance to a class, with respect to the user information needs, requires assignment of labels to each document.[182] For instance, the ChemXSeer search engine classifies, Web site no longer available, scientific articles into those that correspond to chemical documents and those that do not.[183,184]

Filtering systems often require an iterative search process with the aim of tuning the provided query. Some filtering systems allow periodic query execution, returning the obtained hits through some sort of alert mechanism. For example, the My NCBI system allows registered users to set up e-mail alerts for particular search queries on specified topics at a periodical basis (e.g., monthly, once a week, or daily email schedule).[40] Routing represents another form of long-term IR, similar to filtering, but with the peculiarity that the documents are only ranked according to the relevance to the information need, and without assigning an actual class label.

The most conventional IR system deals with short-term (mostly one-time) user initiated queries consisting of searches executed generally through short queries (usually written in natural language) against large document collections. Often, the document collection is rather static when compared to the typical filtering task. The system then returns the relevant documents as a result to a specified query, and the end user usually inspects the top hits. Ad hoc retrieval tools are concerned with current (short-term) and specific retrieval problems of arbitrary user information demands. The most prevalent examples of ad hoc retrieval are Internet search engines, but there are also search engines that are specialized on chemical document repositories.

Coming back to the kind of search requests (ii), and, as introduced, chemistry-related information needs, that is, the topic of issues in which the user is interested, can be represented through a range of quite heterogeneous query specifications and their combinations: for instance, through examination of complex search syntax, by using controlled vocabulary of thesaurus terms, by means of simple free text natural language query inputs (all of them being text search queries) by attempting structure, substructure, and reaction searches (structure-based chemical searches) or even by combining/integrating the former (hybrid searches).[12] Each kind of query type used, and the way it is conveyed to the retrieval system, implies specific technical development in terms of the chemical information infrastructure. Additionally, search tasks are associated to particular search objectives and often subjected to a number of constraints. Thus, in the case of IR applied to chemical data, it is clear that a single search engine fitting all user needs might not always be the optimal solution, as each type of search task implies a particular tuning and IR setting, and often requires adaptation to attain the most suitable retrieval efficiency. This section focuses on different text search queries consisting of natural language text (i.e., words of

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

combinations of words), while structural searches (regardless of whether the input query is a chemical entity name or its chemical structure) are described in section 5.3. Chemical text search systems rely mostly on standard information retrieval methodologies, described below.

2.4.1. Boolean Search Queries. Boolean search queries are characterized by precise semantics, computational speed, and a retrieval strategy that is based on binary decisions (given a document, the query expression is either satisfied or not). The Boolean search model, one of the first types of search models introduced in IR, represents a classic model of retrieval characterized by an exact match search strategy based on classic set theory. It is also sometimes referred to as set-based searching[185] or terms and connectors search,[186] because document sets (represented as a collection of words) are handled using Boolean operators and the search query connects content-bearing words or terms with content free, logical operators. The Boolean logic expressions (named after the mathematician George Boole) rely on the Boolean operators AND, OR, and NOT. Most chemical and biomedical search engines support Boolean operation searches. Boolean AND search implies the co-occurrence of the search terms in a given document. Boolean OR search is inclusive; that is, it requires the occurrence of, at least, one of the specified query terms in the document. Boolean NOT corresponds to the negation of the query terms, and it is usually included to support more restrictive queries where a specific term should be absent (i.e., setting up filters to narrow the search). It is possible to construct complex Boolean queries through the nested combination of multiple query terms connected by Boolean operators and specifying the order of the constraints. Typically, queries are processed from a given direction (usually, left to right), and the operators are applied according to precedence to avoid ambiguous interpretations of Boolean expressions. Nested Boolean queries are conventionally expressed by enclosing the search terms in parentheses, where terms inside a set of parentheses are handled as a unit.

Complex Boolean search queries are commonly used to perform a very specialized type of professional ad hoc search, called technical survey or prior art searches in the domain of intellectual property and patent search.[187]

For example, the search "troglitazone AND CYP3A4" would retrieve all documents containing both of these terms, and it is obtained by intersecting the postings for "troglitazone" and "CYP3A4", that is, $\mathrm{POSTING}_{\mathrm{troglitazone}} \cap \mathrm{POSTING}_{\mathrm{CYP3A4}}$. The query "rezulin OR troglitazone" would return all of the documents that mention at least one of these terms, either "rezulin" or "troglitazone" or both. It corresponds to the union of the postings for "rezulin" and "troglitazone", that is, $\mathrm{POSTING}_{\mathrm{troglitazone}} \cup \mathrm{POSTING}_{\mathrm{rezulin}}$.

2.4.2. Using Metadata for Searching and Indexed Entries. Documents are primarily described through their actual content, using the BOW representation model (section 2.3) and automatically generated content-based term indices.

As introduced, in most cases documents also have metadata. Document textual metadata show different degrees of organization; that is, it can correspond to a field with a small set of possible finite values or fixed vocabularies (e.g., date/year of publication, page numbers, source name), but it can also contain in principle any arbitrary free text (e.g., chemistry article or patent titles). In the first case, there is usually only a single so-called parametric index for each field. If metadata contains free text, the resulting zone index is often comparable to a regular inverted index. A typical practical use case of publication metadata is citation searching by bibliographic retrieval systems, where bibliographic information such as author names represents a classical user query.

Queries that search against both the textual content and metadata information require the merge of the results obtained when scanning the query terms against the standard inverted index (posting) and parametric indices associated to the metadata fields. A similar principle is applied for hybrid searches using queries consisting of chemical structures combined with textual search terms. Those searches merge results from chemical intelligence software for the structural search component with the hits of text search engine strategies and examine the intersection between the documents satisfying the structural search and those that satisfy the term mention constraint.[188]

Chemical document metadata can be generated manually, automatically, or semiautomatically (i.e., TM assisted). On the other hand, index entries can be content-derived or noncontent based, and they can correspond to controlled vocabulary terms or uncontrolled vocabulary entries. To ensure consistency and quality of manually generated document indexes, the consideration of well-specified indexing rules, that is, how to extract and manually index documents, is crucial. Resources such as the Chemical Abstracts Service (CAS) have carried out manual indexing of chemistry content with keywords and chemical substance entities for many years. In fact, the first issue of CAS was published back in 1907, while the CAS Chemical Registry System to support indexing for Chemical Abstracts dates back to 1965. Since the mid-1990s, CAS has provided access to its content through SciFinder (presented in detail in section 5.3).[44]

The PubMed database indexes scientific citations using the Medical Subject Headings[189−191] (MeSH) terms. MeSH consists of a collection of hierarchically structured vocabulary terms (MeSH Tree Structures) that cover topics relevant for biomedical subject analysis of the literature. PubMed searches can include combinations of multiple MeSH terms. In practice, PubMed translates basic searches into enhanced searches that take into account automatic mapping of query terms to MeSH terms and include in the resulting search more specific MeSH terms (child terms), according to the underlying MeSH hierarchy.

Chemistry-relevant terms for indexing PubMed records include the so-called "Supplementary Concept", which are keywords that correspond to chemical names, protocols, or disease terms. "Butyric acid (PubChem CID: 264)" and "chloroquine (PubChem CID: 2719)" are examples of Supplementary Concepts indexed by PubMed/MeSH. Since 1996, PubMed indexers have also associated drug and chemical MeSH terms to their corresponding pharmacological action (under the pharmacological action MeSH heading). For instance, the pharmacological action terms indexed for "aspirin" include "cyclooxygenase inhibitors" and "antipyretics".

2.4.3. Query Expansion. The use of thesauri, structured vocabularies, and chemical compound registries facilitates the capture of synonyms and aliases of each of the concepts or chemical entities and, as a result, enables the expansion of the original search query, a process often referred to as query expansion. The search term is traditionally looked up in the structured vocabulary list, and its corresponding canonical form or concept identifier is used to select all of the hits for equivalent terms or entities that share the same canonical form

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

or concept identifier. Thus, query expansion exploits some sort of predefined term relationships (explicit equivalence classing), for example, in the form of a query expansion dictionary. When a query term cannot be found in the query expansion dictionary, one common alternative is to normalize the query term (using stemming or case folding) and retrieve equivalent tokens in the inverted index (implicit equivalence classing); another alternative is to carry out a fuzzy/approximate string matching of the query term against the vocabulary list to retrieve the most similar term string.

Several publishers, including the Royal Society of Chemistry (RSC), have tried to complement strategies based on manual indexing of documents with automatic indexing approaches (by using chemical named entity recognition software, discussed in more detail in section 3), aiming to systematically mark/tag publication contents and aid in chemical entity searches. This strategy can be regarded as a way of semantically enriching publications with chemical information and enables content-derived indexing, specifically chemical compound indexing.

2.4.4. Keyword Searching and Subject Searching. The annotation of documents with chemical information, obtained either through automatically recognized chemical entities or by manual indexing, is commonly used to populate databases, which host the chemical metadata information, including chemical names, connection tables, and referring documents. Chemical metadata can be used directly to generate chemistry-aware searchable indices and to support chemical semantic searches or substance searches. Automatic chemical concept-based indexing can thus be regarded as a strategy to yield semantic enrichment of documents, and constitutes a key element of subject searches.

In this context, it is noteworthy to distinguish between two types of search classes, between keyword searching, which does not depend on predefined indexing concepts (uncontrolled vocabulary), and subject searching, where search queries rely on predefined indexing concepts (controlled vocabulary).

Subject searches (also known as topic or thesaurus searches) return only those document records that have the search term in the subject heading field. This implies that subject searches examine only specific subject terms rather than the content of the document itself. Subject searching enables one to retrieve particular categories of information encoded as predefined (controlled) vocabulary terms or subject heading terms assuming that all items concerning the same subject are prearranged and searchable together. A single, unifying term is used for an abstract concept or chemical entity, and a set of related concepts or a class of compounds are mapped to this unifying concept. This empowers search flexibility, characterized by the underlying way in which subject terms are structured and by the use of more specific subheadings that allow focusing the searches. Subject searches may rely on the name of a specific chemical compound or words that stand for classes of compounds, as exemplified earlier for PubMed (MeSH) and SciFinder searches. This implies that the key concept of interest has to be expressed via a subject term, which for complex search types might require the combined use of multiple subject terms. Occasionally, the actual wording used to encode subject terms might not be very intuitive, and therefore may require the identification of the preferred indexing term when the subject query is somewhat different from its corresponding subject term. Strategies to aid users in building a subject search include autocompletion searches, partial matches of query terms and thesaurus entries, and browsing the subdivision lists or hierarchically structured term lists. The thesaurus internally connects alternative expressions or interchangeable terms for a specific concept. The MeSH headings linked to a particular document represent concepts that are a major focus (main topic) of the article. On the other hand, SciFinder employs a CA thesaurus to organize controlled search terms. The CAS vocabulary control system (CA Lexicon on STN thesaurus) is structured as a hierarchy of broader and narrower terms, as well as linked terms, previously used terms, and related terms. Scientific terms are grouped into scientific concept families, and chemical substances for frequently indexed chemicals are organized into compound classes together with their common synonyms. SciFinder automatically processed subject searches account for both singular and plural subject words, spelling variants, and common subject term abbreviations. Concept-based retrieval systems that support synonymy searches also include the Essie search engine, which lets nonmedical experts search using less technical terms (e.g., "heart attack" for the clinical term "myocardial infarction").[192] Subject searches represent a powerful retrieval strategy for topics well covered by indexing subject terms.

When there is little information about a given chemical topic, users want to look for words wherever they may occur in the documents, or when multiple query terms need to be combined in complex ways, keyword searches provide a superior search flexibility. Keyword search queries, also called free text searches, express the query through free text natural language expressions, as opposed to predefined controlled vocabulary terms. They are overall similar to Internet searches carried out by search engines like Google, in the sense that the query terms are directly compared to the words contained in the target documents.

2.4.5. Vector Space Retrieval Model and Extended Boolean Model. The classical Boolean search model is an overall binary classification strategy that assumes equal weight for all query terms. Therefore, it provides unordered result lists. On the other hand, the vector space retrieval model facilitates the calculation of query−document similarity scores as a base for result ranking, but lacks the structure inherit in standard Boolean query formulations. The extended Boolean model, also known as ranked Boolean retrieval or p-norm model, can be regarded as a compromise between the classical Boolean model and the vector space query model.[193] This strategy preserves the properties of Boolean algebra and the underlying query structure while incorporating the use of partial matching and term weights to compute query−document similarities. Additionally, query expansion approaches are commonly integrated together with the extended Boolean query processing for recall improvement.

2.4.6. Proximity Searches. A refined search strategy that allows reducing the number of potentially irrelevant hits is proximity search. Proximity searches have been explored by chemical information retrieval systems to look for words that appear close to each other in a given document.[188] The underlying assumption of this approach is that proximity between search terms implies a relationship between those words, and thus documents where the query terms appear close to each other should have a higher relevance. For instance, particular scientific ideas or topics might be discussed in documents within a single sentence, sentence passage, or into paragraphs. Proximity search allows retrieval of documents where two or more separately matching term occurrences are mentioned within a specified distance. Proximity distance is

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

typically measured as the number of intermediate words (or sometimes characters) between query terms, and thereby constraining returned hits to those documents where specified terms are within the particular maximum proximity distance (proximity-search limit). Proximity searches are powerful to filter results where the query terms are scattered across the entire document, that is, merely co-occur anywhere in the document.

In principle, it is possible to distinguish between implicit/automatic and explicit proximity searches. In the case of implicit proximity searches, the retrieval system automatically applies proximity information to generate the search results, while in the case of explicit proximity searches, the user needs to specify proximity constrains in the query. General Internet search engines normally employ implicit proximity searches. The retrieval results are essentially the same for implicit and explicit searches when only two query terms are used. Nevertheless, when more than two search terms are used, explicit searches enable the definition of subsets or groups of keywords (state the order of term relations) to be considered for the proximity search. In practice, this is particularly important for prior art searches. Another classification of proximity search types is based on word order constraints, that is, whether the order specified in the search query is preserved in the returned hits or not (query terms are in any order in the searched text).

Search engines and chemical information retrieval systems use different query syntax to express proximity searches. Overall, special sets of predefined proximity operators or connectors are used to join together search terms. Among widespread syntax types used to express proximity searches are term1 NEAR/n term2 (used by the Web of Science search system, the Cochrane Library search,[194] or Exalead[195]), term1 W/n term2 (used by the Scopus online retrieval tool), term1 /n term2 (used by Yandex[196]), term1 near:n term2 (used by Bing[197]), term1 *AROUND(n)* term2 (Google[198]), and S term1(nA)term2 (used by CAS STN[199]), where *n* in all cases corresponds to the number of maximum words separating the query terms, and term1 and term2 do correspond to user entered search terms. The use of proximity operators has also been explored as a feature for chemical search engines.[188]

Several retrieval systems allow left-to-right ordered proximity search; that is, query terms are near each other in the order specified by the search query. For instance, in the Scopus online retrieval tool, ordered proximity search is specified by the query syntax term1 PRE/n term2, and in CAS STN searches by using S term1(nW)term2, where *n* is the number of words separating the first term from the second term. Moreover, search engines specify the order or preference of the various operators, usually assigning a higher strength to the order proximity term pairs. CAS STN searches support additional co-occurrence constraints between query terms, such as those where search terms have to co-occur within the same sentence or within the same paragraph.[199]

To support proximity searches, information retrieval systems require indexing of word position information for individual word occurrences found within the documents; that is, they need the index information to capture term position information in the documents. So, the indices are organized in a manner that information relative to whether words appear near each other in a document is captured. Search engines may exploit term offset information to provide word-in-context snippets for online display or for query term mention visual highlighting. Positional indices are also important for KWIC Keyword in Context searches, a term introduced by Hans Peter Luhn who was also one of the pioneers in the development of early chemical compound search engines.[200] KWIC systems require sorting and aligning all occurrences of the matched query word together with the surrounding context on both sides of the word.

Proximity searches are very sensitive with respect to the used tokenization software. Section 2.2 provides a detailed characterization of the various tokenization strategies with emphasis on chemical texts. Hyphenation and quotes pose a challenge for many existing retrieval systems that are unaware of chemical naming characteristics. For instance, the chemical entity name "1,1-diphenyl-2-picrylhydazyl" would not be directly found by search engines depending on how they handle hyphen characters. The PubMed database preserves certain characters, such as hyphens and quotes, during the indexing step to handle chemical names better and improve retrieval of substances.

2.4.7. Wildcard Queries. Partial chemical names searches and tolerant retrieval approaches, based on wildcard and regular expression queries, are an alternative to complete chemical name searches. Wildcard queries can be efficient to account for typographical errors, cases of unclear spelling, and variants of a particular search terms, and to achieve term expansion using wildcard searches. There are an increasing number of retrieval engines that support various types of wildcard searching. Wildcard searches are sometimes referred to as truncation searches, because the most frequent type of wildcard search corresponds to queries using a shortened form of the original search term, commonly its word root, combined with truncation symbols or operators.

IR tools employ different special characters that are interpreted by each system as truncation operators. Wildcard symbols can be regarded as a type of meta-character. The most widespread truncation symbols are *, $, and also !, ?, #, and l. The hash mark (#) traditionally matches one or zero characters at the end of the search term, while the exclamation point (!) usually matches one character at the end or within a term. The * (asterisk) truncation operator (also known as Kleene star) constitutes the most extensively supported truncation symbol and is used to allow searching for alternate word forms. The ? operator usually matches exactly one nonspace character; that is, it matches all terms that have any single character or no character in the position occupied by this symbol (single character replacement or truncation).

Special data structures need to be implemented in retrieval engines to enable wildcard searches, usually exploiting character n-grams for partial matching. A character n-gram can be represented through a sliding character string or window against terms in the inverted index. Search trees and hashing algorithms can also be used for wildcard matching. Trailing wildcard queries are often implemented through normal B-tree searches[201] and leading wildcard queries through reverse B-tree algorithms using a term index dictionary written backward. For example, the wildcard query "succin*" will match any word starting with the string "succin", such as "succinate", "succinic", or "succinates". Searching with a long prefix reduces the number of terms that need to be visited by the candidate term index dictionary matching strategy. In practice, users should define meaningful truncated query terms to avoid unwanted matches. Noteworthy, prefix queries do not usually enable relevance scoring. The SciFinder chemical search engine applies autotruncation of query words, which means that common wildcard symbols like the asterisk are ignored. The PubMed

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

database supports wildcard asterisk searches, but only for search term end-truncation (e.g., "toxicol*"), and it does not support single character truncation; also, in the case of phrase searches, only the final word in the phrase can contain the truncation operator ("ammonium succin*"). PubMed does not recommend using truncation searches because it does not allow automatic term mapping to MeSH terms and may result in a search time out. Moreover, PubMed limits retrieval to only the first 600 variations of the truncated term. Currently, SciFinder does not support truncation in research topic searches, but command-driven searches in STN do.

2.4.8. Autocompletion. Autocompletion, autosuggestion, or automatic phrase completion algorithms aim to retrieve the most suitable completion for a user-provided search prefix consisting of the first few letters of some query term.[202] The autocompletion method suggests automatically (or after some predefined key stroke) appropriate words and phrases to continue the typed sequence of input characters. Autocompletion relies ultimately on computation of string similarities between the user search and a range of candidate corresponding words. Depending on the actual implementation, these autocompletion candidates typically correspond to previously entered text strings/searches, particular controlled vocabulary terms, and inverted index terms. Simple autocompletion methods compute the edit distance of the user query against an index of controlled vocabulary terms, while more advanced autocompletion mechanisms integrate statistical models of frequent user mistakes through weighted edit distances and expected input by a character language model. The PubMed database offers autocompletion search features based on query log analysis. Sections 2.3 and 3.4 provide a more detailed description of text string similarity calculation techniques.

2.4.9. Spelling Corrections. User queries often contain misspelled or alternative spelling variants of words or chemical names. In fact, around 26% of the query terms entered to Web search engines have been estimated to contain misspellings.[203] Efficient chemical information retrieval tools need to be tolerant to spelling mistakes and inconsistent word selection. Chemical entity recognition approaches (section 3) have also to deal with spelling variations and errors when finding chemical mentions.

Spelling correction can be regarded as a practical application of the noisy channel model, where the system receives a user input text and returns a corrected form.[204] This implies that search systems have to carry out spelling correction before looking for matching documents. Alternatively, they might provide means for spelling suggestion by returning a ranked spelling aid list consisting of candidate terms, where the final search is carried out with the user selected suggested spelling correction. Computational techniques for spelling correction have been initially proposed back in 1964.[205] An exhaustive explanation of spelling correction algorithms (spelling checkers or spell checkers) is beyond the scope of this Review. The most frequent strategy underlying spelling correction algorithms is the computation of proximity or similarity measures between the misspelled query and the corrected forms contained in a dictionary of words that are believed to be correct (character n-grams overlap). The words in a dictionary of correct spellings that are most similar to a given misspelling are identified either by maximizing the string-to-string similarity or, alternatively, by minimizing the string-to-string edit distance.[206,207] Therefore, string similarity programs calculate the lexical distance between strings, defined as the minimal number of edit operations (insertions, omissions, substitutions, or transpositions of two adjacent characters) needed to convert one string into another.[208] For instance, the edit distance, also known as Levenshtein distance, between the strings "octadeinol" and "octadienol" is two. In practice, the different types of edit operations are commonly linked to different weight settings reflecting the likelihood of letters substituting each other. This results in the so-called weighted edit distance calculations.

The common steps tackled by spell checkers are error detection followed by error correction. Basic details of English spell checking computational methods are discussed in refs 209 and 210. General-purpose spelling checkers match the user-entered string against computer readable dictionaries. Spelling correction approaches have also been tailored to handle scientific texts,[211] and are critical to improve the quality of the noisy text returned by OCR software (discussed in section 2.2), which generally contains a considerable number of substitution errors as opposed to user-entered keyboard misspellings.

The main types of misspellings are typographical errors and phonetic errors. The former consist of misspelled words with spelling similar to that of correct candidate words, while the latter refer to misspelled words with pronunciation similar to that of the correct candidate words. Phonetic errors typically require the generation of phonetic hashes for each term to be able to group words that sound similar, a process carried out by so-called soundex algorithms.

Before calculating the string-to-string edit distances, spelling checkers usually carry out certain heuristic processing steps, such as converting all letters to the same case, transforming spaces to texts without space, deleting automatically inserted hyphenation, and replacing ligatures. Other widespread heuristics imply the restriction of search terms to those starting with the same letter as the query string or changing certain triple letters to double letters. When building in house retrieval solutions, the Lucene API SpellChecker is a convenient solution.[105] Both the SciFinder retrieval system as well as PubMed[212] support spell checking. In the case of PubMed, instead of using a dictionary of correct spellings, it relies on term frequencies in PubMed to suggest alternative searches (PubMed "Did you mean" function). This function is more suitable for providing alternatives for multiword queries. In turn, SciFinder employs a spelling algorithm to automatically detect misspellings and allows searching for alternative spellings.

Spelling and pronunciation of English chemical names has been a long studied subject[213] and so is the analysis of chemical spelling correction approaches.[214] In the beginning of the 1980s, the SPEEDCOP (Spelling Error Detection/Correction Project) project, carried out at CAS, resulted in the implementation of a n-gram-based approach for the detection of candidate chemical spelling errors and their manual correction by human indexers.[211,215] Kirky et al. proposed a method to assist users to correct erroneous systematic organic chemical names, providing a detailed characterization of error types and the implementation of chemical spellchecker based on simple lexical rules, chemical grammar modification, and soundex check.[216]

Even though chemical names initially have been spelled differently in different parts of the English speaking countries, according to the IUPAC guidelines chemical names should now be written using international standard spellings.[217] Simple

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

dictionary-based chemical spell-checkers have been developed not only for retrieval purposes but also as components to assist users in correct chemical spelling when using word processors, such as Microsoft Word or OpenOffice.[218] Also, chemical name misspelling recognition is part of the preprocessing steps carried out by the CaffeineFix software to recognize chemical entity mentions in patents.[119] Chemical patents contain a considerable amount of typographical misspellings and OCR-associated errors. This chemical spelling correction approach uses chemical text tokenization and fuzzy string matching, allowing a limited number of mismatches between the input string and the dictionary entry. The use of white word lists, to identify potential spelling errors and candidate terms pairs where correction should be avoided (e.g., herein to heroin or cranium and uranium), can reduce errors in automatic chemical spelling correction. ChemSpell is a web-based chemical information system for chemical name spellchecking.[219] ChemIDplus[220] is the computer-readable dictionary used by ChemSpell. It relies on string similarity, lexical distance calculation between strings, and phonetic rules to produce phonetic keys of input words, similar to soundex algorithms. An example phonetic rule is the transformation of "ph" to "f", which would imply that "sulphur" and "sulfur" would share the same phonetic keys. ChemSpell returns for a user-entered misspelling, a list of alternate spelling suggestions. Also, it incorporates a set of heuristics, such as limiting the chemical keys to a maximum of 100 characters, converting all strings to lower case, and ignoring certain characters, such as numeric locants and punctuation characters, Greek letters, and stereo descriptors.

2.4.10. Phrase Searches, Exact Phrase Searching, or Quoted Search. There are cases where the information need or search topic is commonly articulated in documents by means of some frequently used phrases or expressions. Under such circumstances, phrase searches (also known as exact phrase searching or quoted search) can constitute a practical search strategy.[221] In fact, according to some estimates, over 11% of web searches contain phrase queries.[222] These searches essentially consist of fixed phrases or strings of text, usually some particular multiword expressions. This implies that the documents retrieved by the search system have to contain exactly the same search statement provided in the user query, with exactly the same wording and order. The search syntax employed for phrase searches uses quotation marks (usually double quotes, and less frequently single quotes or braces) around a specific query phrase to force phrase searches.

Note that, depending on the underlying retrieval implementation, phrase searches and search terms combined with proximity operators may return different hits, because proximity searches often incorporate word normalization to account for variant word endings. Internally, several indexing strategies (and their combination) can be exploited to support phrase searches more efficiently than to process all documents in the collection sequentially. One approach relies on biword indexes, that is, indexes that consist of two consecutive words treated as a single vocabulary term, and then applying techniques similar to those previously introduced for the inverted index processing (see section 2.3).[11] This method is powerful for phrase searches for pairs of consecutive search terms (phrase query of two words). An extension of biwords, using variable length word sequences, is known as the phrase index approach. Both of these strategies have to deal with a large set of vocabulary. Phrase searches can also be enabled through the construction of an inverted index with position information. This technique retrieves first the intersection of documents, where all of the search terms of the phrase co-occur (merging positional posting lists). It then examines the corresponding positional information of the terms to select those records where search terms are immediately adjacent, in the same order as in the query.

2.5. Supervised Document Categorization

2.5.1. Text Classification Overview. Keyword searching in patents is not fully efficient as patent language can be too legal and text terms are rather general to have a broad scope of protection (e.g., medical uses of compounds). As an initial guide for patent examiners, patent documents are hierarchically classified on the basis of the contents of the patent by using manually assigned classification codes, such as IPC codes. Therefore, most patent-related searches rely on the use of the classification codes assigned by patent offices (e.g., IPC), which are commonly combined with keywords (and additional metadata). Automated patent classification can increase the quality of the information obtained as well as reduce the error rate of the tedious handcrafted patent classification.[223]

Text retrieval is prototypically associated with some momentary information needs, with emphasis on how to best rank a set of documents returned by a search. Filtering is a particular type of retrieval approach, which implies a more long-term retrieval interest. Text filtering can also be regarded as a particular application type of text classification.[224] Text classification is a rather broad term that refers to any kind of assignment of documents into classes. In this context, classes can refer to any particular topic, but also to a certain language, author, or year. Text categorization is a subtype of text classification that requires a predefined classification scheme, in the sense that documents are assigned or sorted by content into predefined categories or labels. Note that, in practice, the term text classification is often used instead of text categorization. Text classification approaches date back to the beginning of the 1960s,[225] but their practical widespread use started during the first half of the 1990s.[224]

The main difference between document ranking and classification is that, in ranking, documents are essentially ordered by some property, while, in classification, a class label is assigned to each document.

Text classification can be defined as, given a set of documents $D = \{d_1, d_2, \ldots d_m\}$ and a fixed set of topics or classes $C = \{c_1, c_2, \ldots c_n\}$, determine the topic of d being $d{:}c(d) \in C$, where $c(x)$ is a classification or categorization function whose domain is D and range is C.

The class labels or categories are associated to some conceptual classification scheme that defines the basic characteristics of membership of a document to a particular class. Although in this subsection we refer to documents as the textual unit or object of classification, in practice, text categorization methods can be applied to any arbitrary textual unit, such as phrases, sentences, abstracts, passages, figure/table legends, full text papers, patient electronic health records, patents, entire article collections, or, in principle, even individual words.

Representative examples of general text categorization cases include assignment of labels to documents according to particular topics,[226] genres,[227] sentiment types,[228] or domain-specific binary categories (e.g., relevant or nonrelevant to a specific subject).[229] In fact, the most frequent type of

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

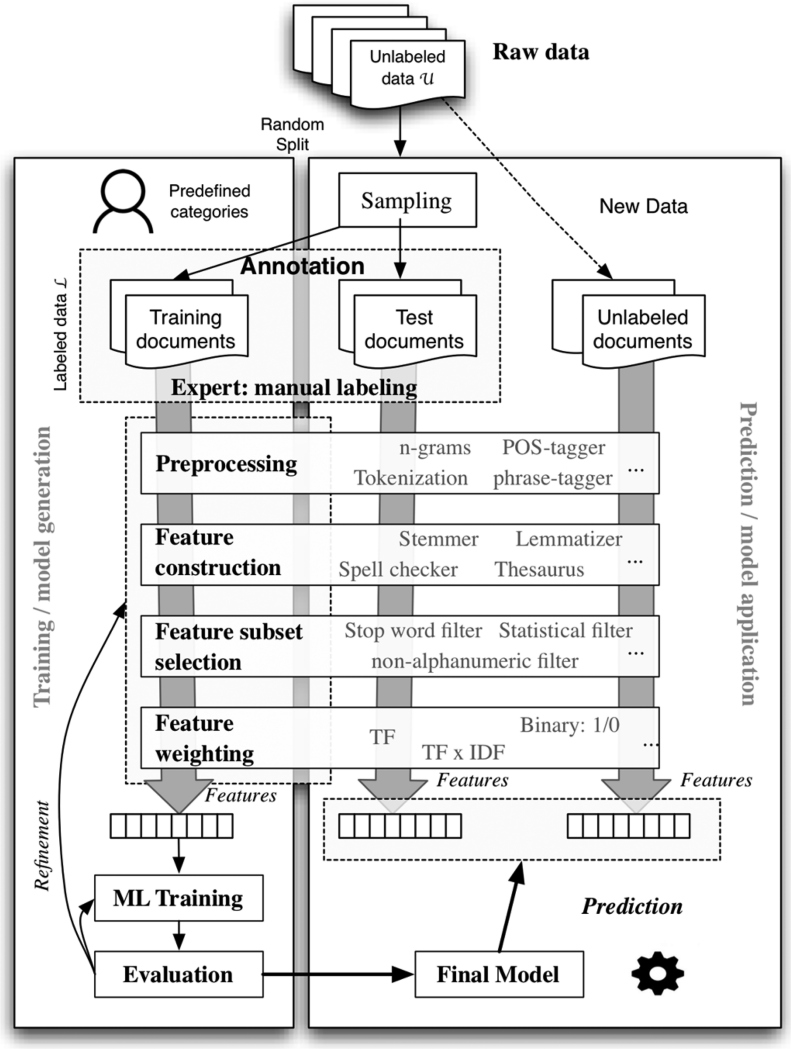


Figure 7. Simplified flow diagram for supervised text categorization.

implementation is the assignment of documents into two mutually exclusive classes, commonly known as binary or two-class text classification, as opposed to multiclass classification, where $c > 2$ classes, often implemented as n binary classifiers.

Text classification can take manual or automatic approaches. Manual classification requires assignment of classes to documents by human experts, usually according to certain classification guidelines or specifications. Manual document labeling by experts is considered to be very accurate and consistent for manageable-sized collections and small annotation teams. However, such labeling is rather time-consuming and difficult to scale to larger document sets. Recently, crowdsourcing has gained popularity as a strategy for labeling text and generating manual annotations, that is, as an approach that outsources labeling tasks to annotators recruited through Internet services.[230,231]

Manual document classification is carried out for the PubMed (expert annotation of MeSH terms) as well as for TOXLINE.[49] Moreover, several text annotation tools[232] and document labeling applications have been developed to speed up this manual process.[233] Currently, manual document labeling is often used in combination with Automatic Text Categorization (ATC) methods, either to review uncertain or difficult cases, or to prepare training and evaluation data sets for ML based text categorization. Also, patent documents are hierarchically classified on the basis of the contents of the patent by using manually assigned classification codes, such as IPC codes. Some efforts for automatic IPC assignation have been pursued.[234]

Automatic text categorization can be achieved by hand-coded rule-based systems, where some expert writes manually a set of rules that, in turn, are applied by machines to automatically classify documents. For example, a rule may assign a particular category to a document if it contains a manually specified Boolean combination of words or terms. Among others, expert rules have been used to assign ICD-9 codes (International Classification of Diseases, 9th revision) to clinical documents.[235] A simplified example expert rule for ICD-9 document coding would be as follows.

Rule: IF "pulmonary collapse" $\in d$ OR "atelectasis" $\in d$ AND "congenital atelectasis" not $\in d$ AND "tuberculous atelectasis " not $\in d$.

ASSIGN label ICD-9-CM 518.0.

Hand-coded rule-based systems usually show a very high accuracy, but require manual expert refinement over time and are difficult to adapt to different document collections. Nowadays, the most widespread approaches for automatic text categorization make use of supervised ML algorithms

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

where the "machine" learns statistical models from training data, that is, example objects (documents) tagged by hand with class labels. The resulting model is then applied to predict class labels for new, unlabeled objects.

Figure 7 illustrates a simplified flow diagram for supervised text categorization.

So, the aim of supervised text classifiers is to build categorization functions (classifiers) that assign a class to an unlabeled document on the basis of some statistical model that, one way or another, measures how "similar" or probable it is that the given document belongs to each possible class. Clearly, there are two distinct phases associated to supervised text categorization, the training phase during which the classifier (categorization function or schema) is generated, and the prediction phase, during which the learned model is applied to new data to predict class labels for each item. Ideally, the document data sets used for training, evaluation, and cross validation should be separate, nonoverlapping sets (randomly) sampled from a common collection.

2.5.2. Text Classification Algorithms. Practically all of the standard classification algorithms have been applied to text categorization tasks. It is beyond the scope of this Review to introduce the ML algorithms used for text categorization. For a general overview on these algorithms, refer to refs 224 and 236 or to ref 237 for bioinformatics use.

Most of the ML algorithms used for text categorization employ some sort of vector space representation model.[178] Among the most popular methods are Naïve Bayes classifiers,[238,239] support vector machines (SVMs),[240,241] Rocchio algorithm,[242] logistic regression,[243] neural nets,[244] boosting algorithms,[245] decision trees,[244] and nearest neighbors.[246]

Typically, text categorization pipelines integrate some basic preprocessing, text tokenization, feature vector generation, feature selection, and model generation modules as well as model validation components. Despite the method, it is necessary to define for the input objects, that is, the documents, a set of features that hold predictive potential to build a good classifier. The basic model of representation of documents for text categorization is some kind of word-based, vector space representation or BOW representation (see details in section 2.3).

Going beyond the BOW representation, it is possible to distinguish different levels of input features. The subword level features are usually composed of text patterns consisting of character n-grams,[247] that is, strings or substrings of n characters like the following 5 g: "cance", "oncog", and "tumor". Word-level features include single words and lexical information (part-of-speech tags). Multiword level features are usually phrases (e.g., noun phrases) and syntactic units, co-occurrence patterns of words, or word n-grams. Semantic level features represent particular units of text associated with certain concepts, entity mentions, or indexing with a fixed vocabulary (e.g., chemical entity mentions or MeSH terms). Pragmatic level features consider the structure and contextual information of the document. Finally, metadata information can also be encoded into feature information.

A key aspect influencing the performance of text categorization methods is the selection and extraction of suitable features from documents that represent only those aspects that are informative to correctly assign labels (eliminate noise). This implies identifying processes for removing, from the large number of features associated to text collections, irrelevant or inappropriate attributes, and thus improving the generalization (performance) of the model, avoiding overfitting, increasing computational efficiency, and reducing classifier training time.

During the feature subset selection step, only a subset of the original attributes is chosen. Typical selection or filtering criteria comprise stop word removal, elimination of non-alphanumeric words or numeric expressions, and use of word character length cut-offs (e.g., retain only tokens with length greater than n characters). Another common feature filtering relies on is statistical feature selection strategies. This can mean applying a simple frequency threshold[246] or sorting all features by their absolute frequency and retaining only the top n most common features. More statistically sound criteria for feature ranking (and selection) are based on hypothesis testing through chi-square measures,[246] information theory-based approaches using mutual information (MI) or information gain (IG),[248] the use of odds ratios,[249] or the choice of attributes on the basis of cross-validation results.[250]

Despite the importance of feature subset selection to reduce the dimensionality and noise of textual data, for categorization purposes, it is advantageous to be able to group equivalent items from the original set of features terms even if they do not share the exact same character string. Building new features through the combination or processing of original features is called feature construction. Classical feature construction approaches include morphological preprocessing, like stemming or lemmatization, conversion into common cases (usually lower case), thesauri matching, term clustering, and spelling correction. Statistical classifiers require that features are linked to some numerical value or feature weights. Widespread feature weighting scores are *tf* (term frequency) and *tf-idf* (detailed previously in section 2.3).

Traditional supervised learning techniques require, at the beginning, the creation of the entire set of manually labeled training and validation data, followed by the construction and training of the classifier. This scenario, where the training data that will be manually labeled were randomly selected, is known as passive learning. An alternative to this setting, with the goal of reducing the needed amount of manually labeled training data, is called active learning.[251] Active learning is usually an iterative process in which first a classifier is trained on a small seed set of manually labeled sample documents and then applied to unlabeled data. From the data are chosen items with predicted labels produced by the classifier potentially informative cases for manual revision and labeling, and subsequent addition to the original training set for model retraining purposes.

2.5.3. Text Classification Challenges. Text categorization strategies face several challenges. From the theoretical point of view, the main obstacle is the accurate definition of classes and how/whether document objects can actually be differentiated according to the proposed class labels. Among the existing practical challenges are the selection of suitable models and algorithms for the classification task, intrinsic class imbalance[252] (e.g., for binary classification irrelevant documents usually significantly outnumber relevant records), the heavy manual workload associated to generating a sufficiently large representative training data set, and the selection of suitable classification features. Other issues that need to be examined carefully while developing automatic text categorization systems are overfitting, parameter optimization for some of the algorithms, and the computational cost associated to dealing

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

with high dimensional feature spaces and its effect on the performance of the classifier.

Text categorization offers a way of selecting documents that are of relevance for the extraction of manual annotations or to run on the natural language processing software, and extract chemical and biomedical entities and associated relations. Spam filtering, word sense disambiguation, and automatic metadata generation[224] are well-known use cases outside the chemical domain. Automatic text categorization has been heavily applied to process scientific data, mainly scholarly articles, paper abstracts, medical records, and patent documents.

Customised classifiers have been implemented for a broad range of biological and medical topics of chemical importance. Text categorization approaches have been used to identify abstracts related to stem cell biology,[253] the cell cycle process,[254] mitotic spindle formation,[255] to detect sentences about transcript diversity,[256,257] and to sort model organism literature according to specific developmental, cellular, and molecular biology topics,[255] among others.

A number of studies used supervised learning methods to detect abstracts describing protein−protein interaction information and thus improve literature curation in domain-specific databases.[229,258] Such effort motivated the implementation of an online application called PIE for scoring PubMed abstracts for protein−protein interaction associations.[259] Text classifier-based literature triage for data curation was used to find articles with peptide information relevant to the PepBank database,[260] and a Naïve Bayes model was applied to discover documents characterizing epitopes for immunology database annotation.[261] Moreover, several classification systems for protein subcellular location prediction have been implemented using ML methods.[262−264]

A particular prolific application area for text categorization is the recognition of medical and disease-related documents or sentences. Text classifiers were tested for screening research literature with respect to genetic association studies,[265] association to melanoma and skin cancer,[266] genome epidemiological studies,[267] or to classify sentences from randomized controlled trial abstracts in PubMed into one of four sentence types (introduction, method, result, or conclusion).[268] SVMs together with Naïve Bayes and boosting algorithms were explored to find articles related to internal medicine in the areas of etiology, prognosis, diagnosis, and treatment.[269] Apart from scholarly documents, electronic clinical records[270] and web-pages[271] are also considered, especially in the oncological domain.[272]

Text classifiers have also been implemented to improve the detection of associations/relations between drugs and chemicals to genes (pharmacogenetic associations)[273] and drug-induced adverse toxic effects.[274] A combination of k-NN-based text categorization together with a chemical dictionary has been used to improve the performance of chemistry-centric search engines.[275]

Using specially trained classifiers has the advantage that these systems usually have competitive performance and do not require the preparation of any extra training data by end users. Nonetheless, in practice, such systems are not necessarily relevant to the topic of interest for a given use case scenario. Several general-purpose classification systems have been published that enabled the classification of PubMed abstracts with user provided training data, like PubFinder[276] or MScanner (validated using example cases from the radiology and AIDS domain).[277] Furthermore, MedlineRanker is still a functional online application for PubMed classification through user-provided training records.[278] This system enables the end user to input lists of PubMed identifiers as training data, or to use surrogate free-text or MeSH term search query output data collections.

A rather novel application type of text classifiers is ranking or prioritizing entities mentioned in the classified documents. Génie is an online application that generates, for a user-provided training set of abstracts (e.g., records discussing some particular biochemical topic), a text classifier model applied, in turn, to classify the entire PubMed database. In this sense, it is rather similar to the previously described MedlineRanker system. In a second step, it ranks all of the genes of a user's defined organism according to the scores and classification of the associated PubMed records.[279]

Alkemio uses a similar strategy for ranking chemicals and drugs in terms of their relevance to a user-defined query topic (document training set). Relevance of PubMed abstracts for the query topic is calculated with a naïve Bayesian text classifier, while chemical/drug ranking is obtained by computing P-values statistics for all chemicals using random simulations.[280] More recently, Papadatos et al. reported a document classifier, freely available, designed for prioritizing papers relevant to the ChEMBL corpus for posterior annotation that are not in journals routinely covered. The BOW document classifier was trained on the titles and abstracts of the ChEMBL corpus using Naïve Bayes and RF approaches. Documents are scored according to the "ChEMBL-likeness" score.[281] With similar purpose, the document relevancy score was created for improving the ranking of literature for the curation of chemical-gene-disease information at the Comparative Toxicogenomics Database (CTD) from PubMed abstracts and titles. Retrieved results were analyzed on the basis of article relevance, novel data content, interaction yield rate, mean average precision, and biological and toxicological interpretability.[282,283]

2.5.4. Documents Clustering. Under some circumstances, for a given document collection, a predefined class definition might be missing, or a training set of annotated documents is unavailable, but nonetheless there is the need to classify the individual items into groups. In this case, the program should identify classes without human feedback, meaning that the machine chooses the classes. This type of strategy based on classification by automatic means, where the machine learns without human feedback, is known as unsupervised learning (UL). The principle behind document clustering is to organize the documents from a collection into groups based on how similar their contents are, defined basically by the words (and their weights) they have in common. Clusters of similar documents (in terms of contents) can be obtained using one of numerous standard clustering methods. Refer to ref 284 for a classical review on document clustering strategies.

Clustering is often used in practice for exploratory text analysis. Typically, it is performed either at the level of documents and based on the contained words (document clustering), or at the level of words/terms based on the documents in which those are mentioned (term clustering). Term clustering attempts to group similar words. Note that document clustering is very sensitive to the characteristics of the document collection, because the implicit grouping of a document depends on the content of the other documents in the collection (interobject similarities).

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

For clustering purposes, a vector-like representation of documents or document parts following the conventional vector space model is commonly applied. To discover the clusters within the document collection, clustering algorithms generally try to maximize the intracluster document similarity and minimize the intercluster similarity scores. A documents cluster is generally represented by a so-called centroid vector, consisting in the simplest case of the normalized sum of the document vectors belonging to that cluster. Clustering strategies commonly try to minimize the average difference of the documents to their cluster centers. The topical terms from each cluster can be selected from words that represent the center of the cluster to generate a more descriptive summarize of the documents in the cluster. Individual documents correspond to the leaf nodes of the cluster tree.

Document similarity scores between individual documents, a particular document and a cluster, or between different clusters can be calculated using the cosine measure introduced in section 2.3. Document similarity was exploited to find similar abstracts within the PubMed database with the aim of detecting potential cases of plagiarism,[285] or to calculate, given some user-provided text passage, the most similar records contained in PubMed.[286,287]

Classical clustering techniques have been applied extensively to automatically group documents, with k-means clustering and hierarchical agglomerative clustering (HAC) among the most widely used techniques. k-Means document clustering is a top-down approach for distance-based flat clustering, which requires that the number of clusters, the parameter k, have been specified in advance. HAC[288] is a bottom-up clustering algorithm that requires the calculation of a pairwise document similarity matrix SIM[*i*,*j*], where every pair of documents is compared and then pairs of documents that are most similar to each other are grouped together. The same principle is followed for grouping documents. The output of HAC is typically a dendrogram of clusters.

There has been rather limited use of document clustering techniques to process chemistry-related document collections. Nevertheless, document clustering was used to analyze the content-based structure of the documents returned by IR systems[289] or to group documents returned in response to a user query.[290] Clustering was used to potentially improve efficiency of ad hoc retrieval under the assumption that when retrieval is restricted to the top clusters (and the documents belonging to them), instead of processing the entire document collection, a reduced number of documents need to be processed. Other common application areas of text clustering include document summarization (clustering similar sentences or paragraphs) and text segmentation of large documents that describe a variety of topics, with the aim of producing smaller semantically coherent portions.

A number of document clustering solutions have been implemented to process records contained in the PubMed database, representing records not necessarily in terms of the words contained in titles or abstracts, but by standard vocabularies linked to the records (e.g., MeSH terms).[291,292] Among these solutions are XPlorMed,[293] GoPubMed (groups abstracts according to the automatically indexed Gene Ontology terms),[294] PubClust,[295] McSyBi,[296] Anne O'Tate[297] (which uses MeSH-term based clustering), PuRed-MCL,[298] and BioTextQuest+.[291] Lin and colleagues carried out a detailed study related to clustering PubMed records where a document-by-word matrix representation was used together with k-means clustering.[299] Each cluster was returned together with a set of summary sentences, informative keywords, and MeSH terms. The BioTextQuest+ system processes both PubMed and the Online Mendelian Inheritance in Man (OMIM) databases. Results of user provided search queries can be clustered using different similarity metrics, including the cosine similarity score, and clustering algorithms including k-means and average linkage hierarchical clustering. Each document cluster is associated to a tag-cloud of informative terms.

2.6. BioCreative Chemical Information Retrieval Assessment

The importance of information technology approaches to help solve the information explosion problem in science and technology has been realized already over 70 years ago.[300] To determine the advantages of using a particular IR or language technology tool requires the use of evaluation strategies that are able to measure the effectiveness of the system, or, in other words, to conclude whether the system helps the user in accomplishing a specific task. Difficulties in evaluating TM and IR results reside intrinsically in the nature of the processed data, where the fraction of relevant items embedded in very large collections of documents is very small, a situation that complicates the evaluation of these systems.

In principle, evaluation of language technologies and retrieval tools can be carried out at two levels.[185] If the entire system, optionally including the interaction with the end user, is evaluated, an overall evaluation is carried out as opposed to the evaluation of the individual components of the system. At a different level, one can differentiate between system and user-oriented evaluation. The former uses standardized tasks to assess the system or part of it, while the latter requires that real users evaluate the system.

There are a number of different aspects and factors that, in practice, influence the performance of these systems, including use case setting, type of information request launched, speed/response time of the system, the content of the searched document collection, or the type of algorithm/processing pipeline being used, just to name a few. A historical overview of IR evaluation is provided by Saracevic.[301]

To evaluate language technologies involved in IR, that is, text classification (section 2.5) or named entity recognition tasks (discussed in section 3), the performance of the system is measured against a test collection, typically consisting of a gold standard (or ground truth) data set generated manually by humans (sometimes experts on a task-specific domain or topic). A typical evaluation stetting for IR systems comprises a test collection and some information needs (queries) or relevance judgments, in the simplest case consisting of a binary classification of relevant and nonrelevant documents. The comparison of annotations/labels performed by multiple humans of an interannotator agreement score is often used to estimate a practical upper boundary for performance of the automated systems.

2.6.1. Evaluation Metrics. Several metrics are extensively used as assessment criteria to measure effectiveness of IR and text classification systems under standard evaluation settings. Understanding how these evaluation metrics are calculated is key for the correct result interpretation. Key concepts playing a role in these evaluation metrics are: false negatives (FN), corresponding to cases missed or incorrectly rejected by the system (type II errors); false positives (FP), corresponding to

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

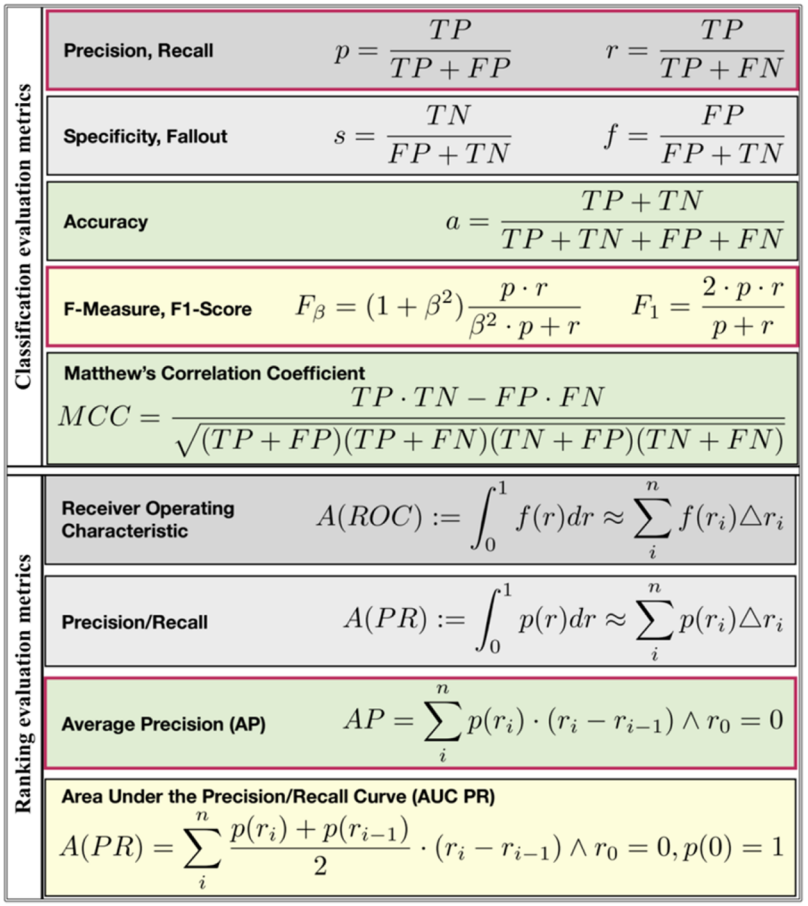


Figure 8. Evaluation metrics for text classification and ranking tasks.

cases wrongly returned by the system (type I errors, incorrect cases); true negatives (TN), being correct negative classification results (correctly rejected predictions); and true positives (TP), corresponding to correct positive classification results (correctly identified cases). Figure 8 provides an overview of widely used evaluation metrics for text classification and ranking tasks.

Kent et al. (1955) were the firsts to propose using precision and recall (originally called relevance) as basic evaluation criteria for IR systems.[302] Precision (p) or positive predictive value is the percentage of correctly labeled positive (relevant) results over all positive labeled results. Recall (also sometimes named coverage, sensitivity, true positive rate, or hit rate) refers to the percentage of correctly labeled positive results over all positive cases. Accuracy, that is, the fraction of correctly labeled results over all results, is less commonly used for IR purposes, but is often used in the context of document classification or named entity recognition tasks. Accuracy is usually not suitable for IR evaluation assessment because IR data are almost always extremely skewed, in the sense that relevant positive documents are a very small fraction as compared to the large collection of nonrelevant (negative) documents.

A widely used metric that provides a trade-of between precision and recall is the F-measure, corresponding to the weighted harmonic mean between precision and recall (see Figure 8). The use of the F-measure for evaluation purposes was introduced by van Rijsbergen (1979) (refer to chapter 7 of ref 303). The F-measure is regarded to be more robust than accuracy when dealing with class imbalance and provides the option to specify a weighting factor β. In the case of the default F-measure, known as F1-score or balanced F measure, $\beta = 1$ and precision and recall have the same weight.

Another balanced measure that is useful for the evaluation of unranked results, such as binary classification scenarios, is Matthew's correlation coefficient (MCC). It is also a more stable score than accuracy for imbalanced class distribution scenarios. The value of the MCC score can range from +1 (perfect classification) to −1 (inverse classification), while 0 corresponds to the average random prediction result. As the F-measure does not consider the TN rate, the MCC score might be more appropriate to assess the performance of binary classifiers, unless the TN rate cannot be established.

Likewise, the mean reciprocal rank (MRR) score corresponds to the average of the reciprocal ranks (of the rank of the first correct answer) of the results generated for a set of queries.

To describe all of the evaluation strategies used for ranked retrieval evaluation goes beyond what can be covered in this Review; therefore, here we will only briefly mention three of the most commonly applied assessment measures: receiver operator characteristic (ROC) curve, mean average precision (MAP), and area under the precision/recall curve (AUC-PR), all of them characterized mathematically in Figure 8.

2.6.2. Community Challenges for IR. Researches, as well as developers of chemical IR systems, face the difficulty of choosing which system or methods yield competitive results for a particular task. Comparing directly different retrieval, or other text extraction systems, is a nontrivial task due to the use of heterogeneous input data, variability of returned results,

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

incompatible evaluation metrics, support of different text formats, and other implicit constraints linked to each approach. In practice, this implies that systems might not support the same data, making it impossible to obtain a clear picture regarding what algorithm or system is more appropriate for a given use case.

Scientific community challenges are being carried out in an effort to introduce independent objective performance assessments and thus standardize evaluation strategies and metrics. Evaluation campaigns are used in computer sciences to judge the performance of different methodologies or algorithms applied on common evaluation data sets and using the same evaluation metrics. This helps to find out the most efficient strategies, features, or parameters. Community challenges are key to determine the state-of-the-art for particular text processing tasks and to better understand the main difficulties associated with them. To address issues that are of relevance for the chemical domain, evaluation efforts have also been made to examine search techniques on chemical documents, such as patents and academic literature.

A typical TM challenge is structured temporally into distinct phases: development/training, test, and evaluation phases. During the first phase, a document collection is released, and, in case of text categorization or entity recognition tasks (or any other task that requires labeled training data), a training/development corpus of manually annotated example cases is distributed to the participating teams. During this phase, teams develop, train, and tune their systems to cope with the posed task. Thereafter, during the test phase, participants receive either some different document set for which they have to produce automatic predictions (e.g., automatic assignment of class labels, ranking of documents, or even extraction of entity mentions; see section 3), or, for classical ad hoc retrieval tasks, a set of queries (information needs) is released. Usually, all of the teams have to return their predictions before some prespecified submission deadline. Finally, during the evaluation phase, the results produced by the systems are evaluated against some manually produced gold standard data set, and obtained performances are returned to the participants.

The Cranfield studies commencing in the late 1950s can be regarded as the first careful formal examination of IR systems (see Cleverdon, 1991 for a detailed retrospective description of the Cranfield experiments).[304] The Cranfield test set consisted of 1398 abstracts from aerodymanics journal articles together with a list of 225 queries and relevance labeling with respect to these queries of the test set documents. Another community assessment with a deep impact on the IR community is the Text REtrieval Conference (TREC). TREC is an annual IR conference and competition organized by the U.S. National Institute of Standards and Technology, with the purpose to promote research and evaluation of IR systems.[305] The TREC evaluation series have been running since 1992[306] and have covered all kinds of IR aspects, including question answering, ad hoc retrieval, and domain-specific retrieval tasks. In the spirit of the TREC evaluation settings, a collection of around 350 thousand MELINE records (known as the OHSUMED test collection) was used to directly compare search strategies based on free text searches using the vector space model versus Boolean query retrieval, considering novice physicians as test users.[307]

The TREC Genomics tracks, running from 2003 to 2007, used PubMed abstracts (2003−2005) and full texts (2006, 2007) as document sources for evaluating different kinds of retrieval-related tasks relevant to gene functional annotation and biomedicine.[308,309]

The main focus of the chemical IR tracks of TREC (TREC-CHEM), which took place in 2009,[310] 2010,[311] and 2011,[312] was the evaluation of two kinds of retrieval strategies carried out on chemical patent documents: prior art (PA) and technical survey (TS) searches. The prior art (PA) track and technical survey (TS) track relied on a document subset consisting of 10% of the MAREC (MAtrixware REsearch Collection) collection, that is, a standardized XML formatted corpus of 19 million patent documents in different languages.[313]

PA searches are important to get the approval of patent claims. They generally refer to information such as prior publications or issued/published patents that are relevant to a patent's claims of originality. For the PA task a collection of over 2.6 million patents and almost 6000 full text articles from the chemical domain were used. System predictions for two subsets were evaluated, one consisting of 1000 patents (full set) and one of 100 patents (short set). Citations in patents and patent families were considered for evaluation purposes. A total of 8 teams provided submissions for the TREC-CHEM 2009,[310] 4 groups for the TREC-CHEM 2010,[311] and only 2 groups for the TREC-CHEM 2011[312] task, respectively. Among the main evaluation metrics used to assess the results of the predictions for the PA task were the MRR and MAP scores.

Top systems of the first TREC-CHEM reached MRR scores of 0.54, while the best MAP scores were of 0.1688 (short set) and 0.1835 (full set). Top systems of the second TREC-CHEM obtained better results, with MRR scores of 0.6452 (short set) and 0.7153 (full set), and MAP scores of 0.3612 (short set) and 0.4121 (full set). The evaluators did not publish a detailed examination of the results obtained by the teams of the third TREC-CHEM competition.

The TS task of the TREC-CHEM competitions was basically a kind of ad hoc retrieval task for a set of topics (queries) that had been provided by patent experts. The evaluation of submissions for this task was done using stratified sampling and manual examination by chemistry graduate students and patent experts. It is worthwhile to notice that there was a rather low agreement between the humans and high variability among the results obtained across the different topics. At the TREC-CHEM 2009, a total of 18 topics were provided to the 8 participating teams, while only 2 teams participated in this task during the 2010 edition (for 6 topics), and 4 teams at the 2011 edition (6 topics).

BioCreative, Critical Assessment of Information Extraction in Biology, is an effort to promote the development and evaluation of IR, extraction, and TM technologies applied to a range of topics relevant to life sciences, biomedicine, and chemistry. During BioCreative II, II.5, and III, specific article classification tasks were held, asking participants to identify and rank articles describing experimentally verified protein−protein interactions (PPIs) data. PubMed abstracts were used in BioCreative II[314] and III,[229] while for BC II.5[315] full-text articles had to processed. An additional task during BioCreative II focused on the retrieval of evidence sentences for the PPIs.

The CHEMNDER (Chemical compound and drug name recognition) tasks posed at BioCreative IV[132] and V[316,317] were the first community calls that challenged the participants with the recognition of chemical entity mentions. Moreover, two CHEMNDER tasks were of particular relevance for chemical IR strategies: the chemical document indexing (CDI) task of BioCreative IV and the chemical passage detection (CPD) text

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

classification task of BioCreative V. For the first CHEMDNER task, PubMed abstracts were exhaustively annotated with chemical entity mentions (see sections 3.8 and 3.9), while, for the second CHEMNDER task, gold standard annotations were done on patent titles and abstracts.

The CDI task addressed the ability of automated systems to detect which compounds are described in a given document. This is relevant for scenarios where a chemist wants to retrieve all of the records that contain a particular chemicals of interest (regardless where exactly they are mentioned within the document). Records used for the CDI task were selected in such a way that all of the main chemical disciplines were covered by the document collection.[318]

For a set of PubMed abstracts, participating systems had to return for each record a ranked and nonredundant list of chemical entities, corresponding to an UTF-8 encoded string of characters found in the text. The CDI task did not attempt to link the chemical entity names to their chemical structures, or some chemical database identifiers. Asking participants for an explicit ranking of the entity lists for each abstract had the objective of empowering systems that are more efficient when combined with manual validation of automatically extracted results, and to facilitate additional flexibility by being able to pick the N top chemicals for each record.

For the CHEMNDER tasks, each participant was allowed to send up to five different runs, having as the only constraint to generate the results totally automatically with no manual adjustment nor correction, and providing submissions in a simple standardized prediction format. All submissions were evaluated against the manually labeled annotations (called the CHEMDNER corpus) using a common, publicly available evaluation script.[315] TP predictions had to match exactly the manually indexed mentions for a particular record, while partial matches or overlaps were not considered correct hits. The evaluation metrics used for the CDI task were recall and precision, and the balanced F-score was the main evaluation criteria. A total of 23 distinct teams participated in this task, returning 91 individual submissions.

Simple baseline performance was determined by using a list of names derived from the training and development sets of the CDI task to index the test set abstracts. This approach yielded a F-score of 53.85% (with a recall of 54.00% and a precision of 53.71%). The best result obtained for this task was an F-score of 88.20%. When examining recall and precision independently, the best precision reached for this task was of 98.66% (with a recall of 16.65%), whereas the best recall was of 92.24% (with a precision of 76.29%). Most of the precision scores obtained by participants were higher than their corresponding recall counterpart. With respect to the ranking strategies of the indexed chemicals, the used criteria could be summarized as: using counts of the number of occurrences of each chemical mentions, applying some manual rules based on the class chemical detected, examining if the chemical name exists in a specific chemical database, scanning if the chemical name was found in the training/development collection, using confidence scores, and marginal probabilities returned by ML models. Most systems participating at the CDI task were an adaptation of their system for the chemical entity recognition task described in section 3.9.

BioCreative V CPD task focused on the classification of patent titles and abstracts with respect to whether they contained chemical compound mentions or not. This means that it consisted of a chemical-aware text classification task. It can also be viewed as an initial selection task of those patents that might describe chemical entities. For this task, a binary classification of patent titles and abstracts was done, (1) does mention chemicals or (0) does not mention chemicals.

The automatically generated categorization labels for patent titles and abstracts were compared to exhaustively annotated manual labels prepared by chemical domain experts. The predictions of each system had to be associated to a rank and a confidence score. A total of 9 teams provided predictions for this task, corresponding to 40 individual runs. The best run with respect to the MCC score was of 0.88, with also the highest sensitivity score of 0.99, and the best accuracy (0.95).

3. CHEMICAL ENTITY RECOGNITION AND EXTRACTION

Generating relevant results for targeted retrieval, as empowered by chemical concept and semantic search strategies mentioned in section 2, requires both understanding the user's intent and the contextual meaning of the chemical search terms as they appear in the documents. This implies adding semantics to the query, for example, knowing that the query corresponds to a chemical concept and detecting this concept in the running text of documents, regardless of the wording used to express it. Together with its utility for IR, as introduced, one of the most interesting applications of CER is the automatic annotation of chemical knowledge bases (section 5).

3.1. Entity Definition

As introduced, the process of automatic recognition of chemical entity mentions in text is usually known as chemical entity recognition (CER) or chemical entity mention recognition. In general, finding (and classifying) mentions of specific predefined types of entities in text is called named entity recognition (NER), although other equivalent terms include named entity recognition and classification (NERC), entity tagging, identification, or extraction.[319] NER is concerned about finding specific information inside documents, rather than working at the level of entire documents as it is usually the case for IR. Information extraction (IE) analyzes documents, extracting structured factual (or conceptual) knowledge from the text for predefined concept types.

Typically, named entities refer to some sort of predefined categories/classes of (real world) objects (names) in text, while, from a linguistic viewpoint, they are often characterized by being referenced in text as a proper name or correspond to a special noun phrase. At the theoretical level, there are philosophical discrepancies regarding the actual formal definition of proper names, which go beyond the more pragmatic interpretation followed by text processing applications. In practice, NER involves locating mentions of such targets/objects (mentions of a particular semantic type) and classifying them into a set of predefined categories of interest. A noun phrase is a phrase or syntactic unit (word/word group acting as a constituent in the syntax of a sentence) that has as its head word a noun or indefinite pronoun.

As was the case for text classification categories (section 2.5), object category definitions of named entities (NE), both in general as well as those related to chemical information, might be intuitively quite clear, but in practice many are associated to vague, or even inconsistent, interpretations. Therefore, NEs require formal definition criteria and explicit text annotation rules, often taking into account conventions that are guided by

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

the underlying practical use case of the resulting technologies (see section 3.8).

At first sight, one may think that labeling entities is a straightforward quest, but systems that attempt to correctly recognize NEs (including chemical compounds) in text have to face two primary written natural language issues: ambiguity and variability. There are various levels of ambiguity intrinsically present in natural language, ambiguity at the lexical, syntactic, morphological, and discourse levels.[320] Lexical ambiguity refers to cases where a word can have multiple alternative senses determined by its contextual properties, that is, shows the same idiosyncratic variation with somewhat unrelated senses (homonymy). For instance, the acronym "CPD" can refer to both "cyclobutane pyrimidine dimer" and "chronic peritoneal dialysis", which leads to lexical ambiguity. An example case of ambiguous grammatical categories for the same word is present in the word "lead", corresponding potentially to either a verb or a noun. Successful NER systems have to cope computationally with lexical ambiguity. The capacity of computationally identifying the correct, context-centered meaning of words is known as word sense disambiguation (WSD). In practice, most WSD and NER systems assume the cocalled one-sense-per-discourse hypothesis that states that, given a particular discourse, all of the individual instances of a specific word tend to have the same meaning.[321] In the example case of "CPD", given a particular document that mentions this acronym, it is highly probable that all mentions of this word will show the same semantics.

Another aspect that influences the strategies used to tag NEs, and specifically chemical mentions, is the existence of naming variations, such as typographical variants (e.g., alternative usage of hyphens, brackets, and spacing), alternative word order, and existence of aliases/synonyms referring to the same entity (e.g., systematic nomenclatures, trade names, and database identifiers).[35] For instance, CER systems have to be able to detect both of the following typographical variants of the same substance: "cyclobutane-pyrimidine-dimer", "cyclobutane pyrimidine dimer".

3.2. Historical Work on Named Entities

The coining of the term named entity (NE) is usually attributed to the sixth Message Understanding conferences (MUC-6, 1995), which introduced a component task requesting participants to mark up all phrases in the provided text that corresponded to named entities of the type person, location, organization, time, or quantity.[322] The MUCs were a very influential series of workshops and community competitions, supported by the Defense Advanced Research Projects Agency (DARPA), with the aim of promoting research on problems related to IE, including entity recognition. After MUC-7 (1998), NER in English newswire text was considered a solved problem as the performance of automatic recognition approaches (balanced F-score of around 93%) was close to an estimated human performance (F-score of 97%).[323]

So, most of the initial NER methods were implemented to recognize predefined types of proper names from the general domain (mainly newswire texts), focusing on three classic entity types: names of organizations, locations, and persons.[319] For general NEs, there is a hierarchy of entity types, with three major classes corresponding to entity names (conventionally labeled as ENAMEX), time (labeled as TIMEX), and numeric expressions (labeled as NUMEX).[323]

Pioneering work on computational processing of chemical names was carried out by Eugene Garfield in the 1960s, who recognized how to convert algorithmically systematic chemical names into molecular formula and line notations[324] (see section 4.1). Several years later, Zamora and colleagues attempted to automatically extract reaction information from journal texts using NLP techniques, including a dictionary of common chemical substances, chemical formula, and chemical word roots and using morphological identification of chemical words.[156,325] Hodge et al. described techniques for recognizing chemical names in text fields to assign CAS registry numbers.[326] Dictionary lookup techniques and word morphology was exploited in a work presented by Blower and Ledwith to identify substances in the experimental section of an organic chemistry journal, while chemical events were extracted using a rule-based approach.[327] More recent activities relevant to CER will be discussed in the following subsections (section 3.9).

NER can be considered a critical building block or key component for other language processing tasks, as it provides, for instance, names for term-based document retrieval and is often a useful feature for document categorization.[229] NER results may be also a prerequisite for event and relation extraction,[319] machine translation,[328] question answering,[329] and automatic summarization.[319]

This section will provide an overview of relevant aspects for CER, including general factors and characteristics of chemical naming in text, different strategies to automatically label chemical entities, existing resources, and evaluation criteria.

3.3. Methods for Chemical Entity Recognition

3.3.1. General Factors Influencing NER and CER. Several factors should be considered when implementing or applying chemical entity taggers. Among the aspects that influence considerably the outcome of such tools are the granularity of the predefined entity categories, the genre of the processed text (i.e., scientific, informal, patents), its domain (e.g., different chemical subdisciplines), the type of document (abstracts vs full texts), and the language in which the documents are written. Some efforts have been made to implement text-type agnostic NER tools, but mostly for classical entities like those used for the MUC challenges.[330] Moreover, some research focused on the exploration of strategies to improve customization of NER tools to new domains.[331]

Nonetheless, in practice, CER tools work as application-specific systems tailored to a particular domain and text type, focusing mainly either on scientific articles (abstracts or full text papers) or on medicinal chemistry patents.[332,333]

3.3.2. Chemical and Drug Names. The type of chemical entity mention (systematic, trivial), see section 1.1, used in text plays a role in the recognition and underlying method used to identify it.[16,132,333]

Several studies indicate that there are considerable differences regarding the preferred use of the various classes of chemical mentions with respect to the type of documents examined.[318,334,335] Systematic chemical names are more frequently used in patent documents and patent abstracts when compared to scientific literature, particularly journal abstracts. On the other hand, abbreviations and acronyms as well as trivial chemical names are more heavily used in scholarly literature when compared to patents. A characteristic of patents is also the heavy use of Markush structures (section 1.1), formulas, and/or systematic substituents to define the scope of

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

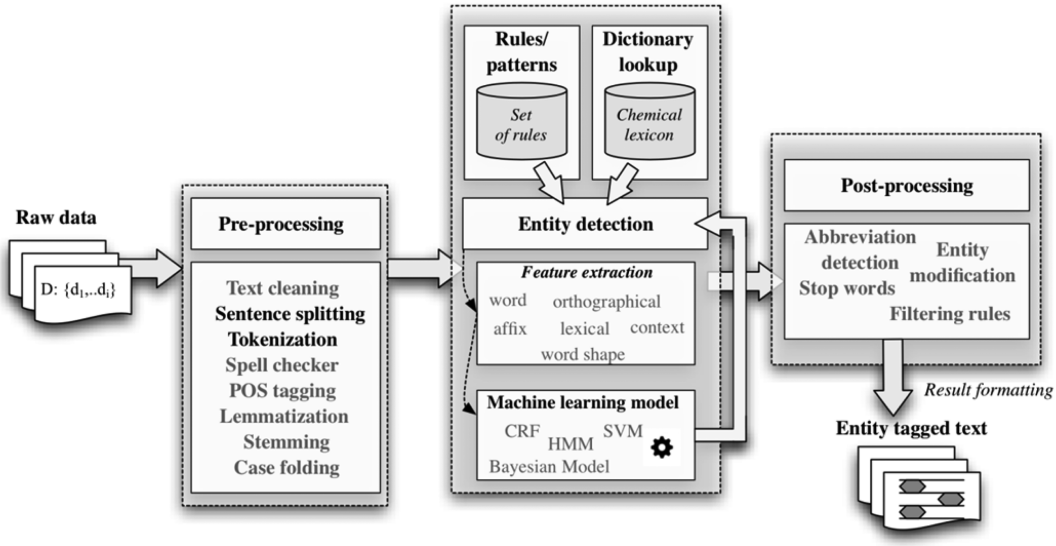


Figure 9. General flowchart illustrating a typical CER pipeline.

a chemical, as well as mentions of novel compounds.[334] While Markush formula in patents appear as structure diagrams (images), the definitions of the variations at the different variations sites are normally found as text, with the additional complication that the chemical mentions to the substituents can be extremely broadly defined, including natural language expressions ("substituted or unsubstituted") and somewhat customized meanings variable from patent to patent whose definition is found in other parts of the patent document ("where the term alkyl refers to"). Moreover, R-group definitions can also be provided as images. Thus, the analysis and interpretation of Markush formula (either manual or automatically) involves the interpretation of the structure diagram of Markush, the interpretation of the text describing R-groups (and images if applicable), and the association of both parts, which can lie at different parts of the document and use a nonhierarchical representation of additional descriptions. The structure diagram can be quite schematic and general and/or present complex drawing features (crossing bonds to indicate variable positions).

Kolarik et al. tried to characterize the relative use of different classes of chemical entity mentions using a small set of 100 manually labeled PubMed abstracts. They concluded that 34.32% of the mentions corresponded to trivial names, 32.42% to IUPAC or IUPAC-like names, 13.55% to abbreviations, 8.21% to chemical families, and the rest to parts of systematic names or other chemical name types.

Recently, a larger study trying to determine the usage of chemical mention types in PubMed abstracts was carried out resulting in the CHEMDNER corpus of 10 000 manually labeled PubMed abstracts[318] (see section 3.8). This study determined that 30.36% of the chemical mentions corresponded to trivial names, 22.69% to systematic names (IUPAC and IUPAC-like), 15.55% to abbreviations, 14.26% to formula, 14.15% to chemical families, 2.16% to chemical identifiers, and the rest to other or multiple noncontinuous mentions of chemicals.

3.3.3. Challenges for CER. Ambiguity is a pervasive problem for practically all NER systems, including the detection of chemical entities. The previously introduced array of different chemical mention classes helps to illustrate some of the difficulties associated to the variable and heterogeneous ways chemicals are named in text. A brief description of difficulties in tagging chemicals is provided by Krallinger et al.[318]

Not all chemical names show distinctive patterns of name segments or chemical word morphology and are therefore missed by some CER strategies.[36]

Chemical NER systems are very sensitive with respect to spelling errors.[336] Applying spelling correction software before chemical entity tagging also can improve the overall recall. A particularly important aspect for CER, and quite distinctive from many other entity types, is their sensitivity with regard to the tokenization strategies used (see section 2.2). Moreover, low-level aspects like punctuation, spacing, and formatting can influence tokenization results. As tokenization errors degrade the performance of CER systems,[162] different tokenizers have been applied to chemical texts, including the exploration of both coarse grained and fine grained tokenization approaches.[160] Tokenization of chemical documents is particularly cumbersome as it is usually loaded with variable use of hyphens, parenthesis, brackets, dashes, dots, and commas. Chemical documents with hyphenated text segments are especially problematic and pose an additional ambiguity for tokenizers, as they have to distinguish between true hyphens (i.e., integral part of complex tokens) and end of line hyphens (i.e., used for typesetting purposes).[337]

Chemical acronyms and abbreviations, as well as some trivial names and a few common English words, such as "gold", "lead", and "iron", are also a source of ambiguity for CER systems. Single and two letter acronyms and abbreviations are particularly challenging for most chemical taggers.[36,161] For instance, one study indicated that, in the case of a particular CER system applied to PubMed abstracts, 18.1% of false positive mentions corresponded to extracted candidate names with a length of one or two characters.[338]

Another common source of difficulties for CER is the correct detection of mention boundaries and associated errors, such as recognizing partially mentions (partial matches) or breaking incorrectly long chemicals into multiple mentions.[339] Incorrect chemical mention boundary recognition is frequently caused by modifiers, for example, the extraction of "aromatic hydro-

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Table 3. Chemical Entity Recognition and Indexing Systems

chemical NER/indexer	description	URL
BANNER-CHEMDNER[343]	CRF-based systematic chemical tagger	https://bitbucket.org/tsendeemts/banner-chemdner
BC4-CHEMDNER Uni. Wuhan CER[344]	CRF-based systematic chemical tagger	https://github.com/zuiwufenghua/biocreative_CHEMDNER
becas-chemicals[338]	online CRF-based chemical/drug tagger	http://bioinformatics.ua.pt/becas-chemicals/
ChemEx[a]	entity tagger integrating ChemicalTagger	http://www3a.biotec.or.th/isl/ChemEx/
chemicalize[b,h]	commercial CER system	https://chemicalize.com
ChemicalTagger[155]	chemical NLP tool	http://chemicaltagger.ch.cam.ac.uk
ChemSpot[162]	hybrid (CRF and dictionary) chemical tagger	https://www.informatik.hu-berlin.de/de/forschung/gebiete/wbi/resources/chemspot/chemspot
chemxseer-tagger[c]	CRF chemical tagger	https://github.com/SeerLabs/chemxseer-tagger
CheNER[345]	CRF-based systematic chemical tagger	http://ubio.bioinfo.cnio.es/biotools/CheNER/
Cocoa[d,h]	hybrid (manual rule/dictionary) chemical tagger	http://relagent.com/Tech.html
iice[346]	online chemical entity and relation tagger	www.lasige.di.fc.ul.pt/webtools/iice/
LeadMine[h,332]	hybrid (manual rule/dictionary) chemical tagger	https://www.nextmovesoftware.com/leadmine.html
MetaMap[347]	tagger for UMLS metathesaurus concepts	https://metamap.nlm.nih.gov
NCBO Annotator (ChEBI ontology)[e]	online tagger of OBO ontologies (incl. ChEBI)	http://bioportal.bioontology.org/annotator
OntoGene[f]	online chemical tagger (incl. lexical lookup)	http://www.ontogene.org/webservices/
Oscar3[g]	naïve Bayesian model-based CER tagger	http://www-pmr.ch.cam.ac.uk/wiki/Oscar3
Oscar4[158]	modular adaptation and update of Oscar3	https://bitbucket.org/wwmm/oscar4
tmChem[161]	CER and normalization using CRF	https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/#tmChem
Whatizit[348]	online tagger or entities (incl. chemicals)	http://www.ebi.ac.uk/webservices/whatizit/info.jsf

[a]Tharatipyakul, A.; Numnark, S.; Wichadakul, D.; Ingsriswang, S. ChemEx: Information Extraction System for Chemical Data Curation. *BMC Bioinf.* **2012**, *13*, S9. [b]Southan, C.; Stracz, A. Extracting and Connecting Chemical Structures from Text Sources Using Chemicalize. Org. *J. Cheminf.* **2013**, *5*, 20. [c]Khabsa, M.; Giles, C. L. Chemical Entity Extraction Using CRF and an Ensemble of Extractors. *J. Cheminf.* **2015**, *7*, S12. [d]Ramanan, S.; Nathan, P. S. Adapting Cocoa, a Multi-Class Entity Detector, for the Chemdner Task of Biocreative IV. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop*; Bethesda, MD, October 7−9, 2013; pp 60−65. [e]Smith, B.; Ashburner, M.; Rosse, C.; Bard, J.; Bug, W.; Ceusters, W.; Goldberg, L. J.; Eilbeck, K.; Ireland, A.; Mungall, C. J.; et al. The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. *Nat. Biotechnol.* **2007**, *25*, 1251−1255. [f]Rinaldi, F.; Clematide, S.; Marques, H.; Ellendorff, T.; Romacker, M.; Rodriguez-Esteban, R. OntoGene Web Services for Biomedical Text Mining. *BMC Bioinf.* **2014**, *15*, S6. [g]*Computational Life Sciences II*; Berthold, M. R., Glen, R., Fischer, I., Eds.; Springer: Berlin, Heidelberg, Germany, 2006. [h]Commercial tool. Prepared in November 2016.

carbons" instead of "polycyclic aromatic hydrocarbons".[161] Finally, just to name some other issues often faced by CER systems are unmatched punctuations of parenthesis and square brackets, false negative mentions due to lack of a sufficiently informative sentence context,[339] and false positives corresponding to larger conventionally not labeled macromolecules.[162]

3.3.4. General Flowchart of CER. Although early general NER taggers typically relied on hand-crafted rules, the current trend increasingly points toward the use of supervised ML techniques for entity recognition, for both domain specific and domain agnostic texts.[319]

Because of the heterogeneous characteristics of the different classes of chemical entities, at the methodological level different complementary strategies have been explored to recognize chemical mentions. Additional reviews on CER are provided by Vazquez et al.[16] and Eltyeb and Salim,[340] whereas an introduction on the detection of systematic names was published by Klinger and colleagues.[141] Two review papers focusing on drug name recognition were published by Liu et al.[341] and Segura-Bedmar and colleagues.[342]

One can distinguish between three general strategies to detect chemical entities in text: (i) chemical dictionary look-up approaches described in section 3.4, (ii) rule/knowledge-based approaches introduced in section 3.5, and (iii) machine learning-based CER covered by section 3.6. Most of the current CER systems are hybrid strategies (section 3.7) that combine ML with lexical features derived from chemical dictionaries.[132] In fact, most of the modern CER combine various approaches at the different stages of the entity tagging process. Figure 9 provides a simplified general flowchart illustrating a typical CER pipeline. Usually, such a pipeline comprises several steps, such as preprocessing, sentence segmentation, word tokenization, entity annotation/detection (e.g., using a dictionary lookup, rule matching, machine learning based token classification, or combinations thereof), postprocessing, and output format conversion.

Because of the difficulty in defining consistent annotation criteria and the considerable workload associated in preparing large manually annotated chemical text corpora, until recently no such annotated resources for chemicals were available. Recent efforts, such as the CHEMDNER tasks posed at the BioCreative challenges, opened the possibility to share training and test data for developing and evaluating CER techniques. This promoted the implementation of a range of new tools and fueled increased interest in the automatic extraction of chemical information from text. Table 3 provides a list of various chemical entity recognition and indexing systems that have been implemented during the past years.

3.4. Dictionary Lookup of Chemical Names

3.4.1. Definition and Background. The most straightforward strategy to detect entity mentions in text relies on a family of techniques commonly known as dictionary lookup or dictionary matching. Classical dictionary lookup requires comparison of terms from a list, dictionary, thesaurus, or

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

catalogue of names to some target text (typically entire documents or sentences). Therefore, dictionary lookup-based CER approaches have two basic components, a dictionary or gazetteer and a matching method.

In the domain of natural language processing, the technical term gazetteer is used to refer to the set of entity name lists, such as chemical compound, drug, or enzyme names. These name lists are exploited to detect occurrences of these names in text for the purpose of named entity recognition. Early attempts to detect chemical entity mentions in text made use of chemical dictionaries.[156,325]

In principle, lookup-based CER does not require any manually annotated text corpus for building a statistical language model, as needed by ML-based CER (section 3.6), and usually solves already partially the entity linking step, that is, associating detected chemical mentions to database records (section 4.2). Short summaries on dictionary-based approaches for the identification of chemical entity mentions are discussed in refs 16, 35, 340, and 341.

Example applications that make use of dictionary matching are Peregrine,[349] ProMiner,[350] and Whatizit.[348]

Dictionary lookup-based CER can be sufficiently competitive for fields that are characterized by using a somewhat limited and comprehensive set of chemical entities, as typically found in medical and clinical documents. Dictionary lookup is thus suitable for scenarios associated to a definite set of well-defined names that are representative of the entity class. Under such circumstances, and assuming that the dictionary is of reasonable size, dictionary lookup can be quite competitive.

Chemical entity classes that are covered sufficiently well by lexical resources include common and generic chemical names, trade names, company codes, and common abbreviations.

Frequent steps used by these sort of CER pipelines are dictionary gathering/construction, dictionary pruning or filtering, dictionary expansion, dictionary matching/lookup, postprocessing, and semantic normalization.

Dictionary construction implies selection of chemical names from thesauri, ontologies, or knowledgebases. Dictionary pruning means removal of problematic names using stop word lists, and filtering rules based on mention statistics, name length, or word type (e.g., POS). Dictionary expansion refers to the process of creating name variations from the original chemical name to cover additional typographical/morphological, spelling, word case, or word order variants. Finally, postprocessing includes techniques to disambiguate or clean up detected mentions, often using some heuristics or context-based disambiguation.

3.4.2. Lexical Resources for Chemicals. A critical component for chemical dictionary lookup is the construction of high-quality and purpose-driven chemical name gazetteers. A description of lexical resources, including thesauri and ontologies containing chemical entity information, is provided by Gurulingappa et al.[35] A list of nonsubscription and open access web resources offering molecular information data is provided in Table 4 of ref 351, while Table 3 of ref 341 provides a collection and description of resources for building drug dictionaries.

In principle, the following types of chemical name gazetteers can be used: manual/hand-crafted name lists, database/derived chemical lexicons, automatic construction of name lists in text (e.g., using results of rule- or machine learning CER tools), and merged/combined gazetteers derived from some of the previously named types of collections. Table 4 shows a selected number of resources for names to construct chemical gazetteers.

Table 4. Resources for Names To Construct Chemical Gazetteers[a]

chemical name resources	URL
BAN (British Approved Name)	https://www.pharmacopoeia.com/
ChEBI	http://www.ebi.ac.uk/chebi/
ChEMBL	https://www.ebi.ac.uk/chembl/
ChemIDplus	https://chem.nlm.nih.gov/chemidplus/
ChemSpider	http://www.chemspider.com/
CTD (Comparative Toxicogenomics Database)	http://ctdbase.org/
DrugBank	http://www.drugbank.ca/
European Pharmacopoeia	http://online.edqm.eu/EN/entry.htm
Hazardous Substances Data Bank	https://sis.nlm.nih.gov/enviro/hsdbchemicalslist.html
HMDB (Human Metabolome Database)	http://www.hmdb.ca/
Jochem	http://www.biosemantics.org/index.php?page=jochem
KEGG COMPOUND	http://www.genome.jp/kegg/compound/
KEGG DRUG	http://www.genome.jp/kegg/drug/
MedlinePlus (drug generic or brand names)	https://medlineplus.gov/druginformation.html
MeSH (MeSH substance record branch)	https://www.nlm.nih.gov/mesh/
NCI Drug Dictionary	https://www.cancer.gov/publications/dictionaries/cancer-drug
NIAID ChemDB	https://chemdb.niaid.nih.gov/
PubChem	https://pubchem.ncbi.nlm.nih.gov/
RxNorm	https://www.nlm.nih.gov/research/umls/rxnorm/
TTD (Therapeutic Target Database)	http://bidd.nus.edu.sg/group/cjttd/
UMLS (Unified Medical Language System)	https://www.nlm.nih.gov/research/umls/
USAN (United States Adopted Name)	https://www.ama-assn.org/about-us/usan-council

[a]Prepared in November 2016.

A chemical dictionary that combines several chemical name resources is Jochem,[336] which integrates names from UMLS, MeSH, ChEBI, DrugBank, KEGG, HMDB, and ChemIDPlus. Rule-based name filtering and manual revision of frequent names was carried out to improve the quality of this dictionary.[336]

Only few attempts have been done so far to generate purely manually constructed chemical gazetteers. Townsend et al. describe a lexicon of common chemical names extracted manually from 295 letter articles from the journal *Nature*.[351] A larger set of 19 805 unique chemical name strings have been produced through manual labeling of chemical names in PubMed abstracts, in the framework of the CHEMDNER corpus,[318] while an automatically generated chemical name gazetteer generated by CER software was a byproduct of the silver standard corpus of the CHEMDNER task.[318]

Dictionary lookup approaches usually need to include postprocessing and manual curation steps (lexicon pruning) to eliminate highly ambiguous names. Strategies to improve the quality of dictionaries commonly rely on manual or rule-based filtering[336] or use of stop word lists.[132,352] Hettne and colleagues present several ways of preprocessing, postprocessing, and disambiguating chemical names for chemical dictionary compilation.[353]

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

3.4.3. Types of Matching Algorithms. Dictionary lookup can be carried out either at the level of characters/strings or at the level of tokens/words. In the first case, two strings, one corresponding to the chemical name and the other to the target text, are compared, usually defining certain additional mention boundary characters or text patterns. In the second case, essential lists of tokens, one resulting from the chemical name and another from the target text, are compared; hence, the used tokenization strategies influence the entity recognition outcome.

There are some discrepancies whether dictionary lookup can be considered a true NER approach, unless some sort of context-based disambiguation step is applied. Dictionary matching techniques can be either exact (exact matching) or approximate (also known as flexible or fuzzy matching). The most common approach is to use exact string or word-level matches, as it shows better precision, but at a recall cost.[354,355] To increase the recall of word-level matches, one option is to consider stemmed versions of the lexicon entries and target articles.

Another choice to boost recall is through approximate string matching methods, which make use of string comparison measures like edit distances (Levenshtein and others) to calculate similarity between character sequences (similar to methods used for spelling correction; see section 2.4). A detailed description of approximate string matching algorithms is beyond the capacity of this Review, and additional details can be found in refs 208 and 356.

The notion of edit distance, introduced by Levenshtein,[206] is the underlying principle of most approximate string matching methods.[207] Fuzzy matching can become computationally infeasible or challenging for very large chemical dictionaries and degrade matching efficiency. Several attempts have been made to apply approximate string matching to detect chemical and drug mentions in text.[357] Approximate string matching was also applied to recognize chemical mentions containing spelling or OCR errors.[119]

Very large dictionaries can pose technical hurdles in terms of lookup efficiency when using sophisticated matching approaches. To overcome the issue of the size of very large chemical gazetteers, one option is to apply heuristics to match first chemical-like tokens or substrings.[16] Alternatively, and exploiting the substrings or terminal symbols building systematic chemical names, matches using character-level n-grams (e.g., 4 g), instead of the entire name, have been tested too.[36]

Recent attempts to combine both dictionary lookup together with chemical grammar rules showed competitive results.[358]

Regular expressions of finite state machines constitute an alternative to fuzzy matching when dealing with ways to capture variations referring to a chemical entity in text. A gazetteer list can be compiled into finite state machines that, in turn, can be used to match text tokens.[330] This essentially implies generating a pattern dictionary from the original lexicon and then scanning each pattern against the target documents.

Finally, instead of using a dictionary as a starting point, an alternative approach is to first extract terms or phrases from text, and then try to check whether it is possible to map them to name entries in a dictionary or lexicon.[359,360]

3.4.4. Problems with Dictionary-Based Methods. Dictionary-lookup-based approaches can usually only cope with a limited number of name variability. Chemical naming variation is considerable; in addition to typographical variants (alternating uses of hyphens, brackets, spacing, etc.), alternative word order can also be encountered.

A particularly problematic class of chemical mentions for dictionary-lookup methods, despite having very distinctive word morphology, are systematic chemical names. These often correspond to long multiword expressions with spelling variability and are not well detected by dictionary lookup.

Moreover, the used chemical gazetteers have to be maintained and constantly updated to include new names added to the used databases or lexical resources. Novel chemical entities are being continuously discovered and characterized in the literature and patents. Such ad hoc chemical names, not yet included in databases or dictionaries, pose a serious bottleneck on dictionary-lookup-based CER systems.

To overcome incompleteness of lexical resources, CER methods based on ML techniques, exploiting chemical name morphology, and handcrafted rules provide complementary solutions. These are described in the following subsections.

3.5. Pattern and Rule-Based Chemical Entity Detection

Rule-based entity recognition represents a suitable strategy to detect entity mentions that show a somewhat fixed structure or do occur in a restricted context of mention. Under such circumstances, entities can be represented through a set of rules. Rule-based NER systems symbolize an attempt to model entity names under the assumption that there are constraints derived from underlying nomenclature or conventional naming aspects in a specific domain. This basically means that rule-based CER is an approach that tries to implement generalized representations of chemical name morphology or context of mention, that is, defining general regularities that characterize chemical entities. In practice, this means that such strategies typically exploit so-called surface clues, that is, how particular chemical compound entities usually look.

Rule-based approaches can be effective when resources such as entity gazetteers (section 3.4) or entity-labeled textual training data are missing (section 3.6). Moreover, such systems are considered to be more amenable for human interpretation and error understanding,[361] and are often regarded as more suitable for closed domain scenarios, where human participation is both feasible and essential. Until the end of the 1990s, rule-based methods were the standard choice together with dictionary-based NER tools. Nowadays, ML-based methods or hybrid systems are widely used, often exploring heuristics and rules mostly at the pre- and postprocessing stages, because purely rule-based methods require input from domain experts and basic linguistic knowledge.

A widespread way to encode individual rules is through the use of regular expressions to be matched against mention instances, each designed to capture and classify a subset of names (or name fragments). In practice, individual elements of these patterns should match specific tokens/strings or classes of tokens/strings with particular features. Regular repressions return all of those strings (sequences of characters) that contain the specified pattern. The representation of regular expressions was already introduced at the end of the 1950s by the mathematician Stephen Kleene[362] and is still a common starting point for many entity extraction tasks, for highly structured entities such as single nucleotide polymorphisms,[363] protein nonsynonymous point mutations,[364] chemical formulas,[365,366] and chemical database identifiers.[365]

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Rule-based systems are generally structured into two basic components. One component consists of a collection of rules, including specifications on how to handle and coordinate relative ordering and matching of multiple rules. The other component is the rule-matching engine and is responsible for the detection of the textual instances of the rule matching patterns.

The first rule-based CER systems relied on human experts to define appropriate rules or regular expressions that characterized chemical entity mentions.[156,367−369] Even today most of the rule-based CER strategies, including those detailed in this subsection, rely mainly on hand-built rules, including competitive commercial systems.[332] A frequent classification of rule-based NER systems distinguishes between corpus-based and heuristic-based rule systems. Humans are very good at creating manual rules or generalizations from a reduced set of illustrative example cases to identify underlying regularities. Corpus-based systems typically require examining a collection of example cases to derive patterns or rules, while heuristic-based approaches usually rely on extensive domain knowledge and/or understanding of existing nomenclature conventions to be able to construct relevant hand-written rules. The creation of manually coded rules by a domain expert is a very labor-intensive and sometimes tedious process, with the danger of resulting in highly customized rules, applicable only to a narrow subdomain that might not necessarily represent the entire space of entity instances.

To overcome these issues, algorithms for learning rules from examples automatically, known as rule learning or rule induction methods, have been devised to detect named entity mentions in running text, including recognition of drug names.[370,371] The typical setting for automatic rule induction requires a small collection of seed examples, that is, hand-crafted patterns and example entity names, and then, based on an iterative process, the collection of extraction patterns is extended by direct induction from unlabeled text.[370] A more detailed description of rule learning for NER can be found in Sarawagi (2008).[372]

Rule-learning techniques are usually classified into bottom-up, top-down, and interactive/hybrid rule learning strategies. In the case of bottom-up rule learning, the idea is to start with seed rules that have a very high precision and low recall, and then to iteratively generalize the rules to increase the associated recall (e.g., by removing tokens or substituting them for more general token representations). Top-down rule learning follows the opposite strategy by starting with very general rules with a high recall and poor precision, and then applying iteratively specialization strategies to increase precision. Interactive or hybrid rule learning basically integrates human experts into the learning process, allowing them to modify or adapt rules, or to add additional seed examples.

Different classes of rules can be used to encode rule-based NER systems. Whole or single token entity rules attempt to model the entire entity mention, neglecting dependencies with other comentioned entities,[330,373] typically in the general form of [Left context] Filler [Right context], while rules to mark entity boundaries (boundary rules)[374] are often applied for detecting mentions of very long entities that are difficult to model entirely. The importance of chemical mention boundary recognition was already realized in an early work by Zamora et al. (1984), through special handling of chemical word segmentation.[156] Start and end mention markers are frequently modeled by independent rules and encoded through mention boundary patterns.

Multiple entity rules[375] try to capture dependencies between entities, for example, to model mentions of noncontinuous, nested, or overlapping entity mentions, or for capturing long name forms and their corresponding abbreviations. Whole entity rules are commonly addressed by using handcrafted rule construction, while boundary rules are also often generated through rule learning techniques.

Rule-based CER extraction systems encode several types of widely used rule token features that try to exploit distinctive properties, like certain character strings, that appear more often in chemical entity mentions in contrast to surrounding text or other entity types. Because of the chemical nomenclature constraints, there is usually some sort of internal evidence (names have an internal structure) encoded in the systematic name that can be exploited by NER rules. Thus, some rules make use of the internal NE formation patterns, including morphological analyzers detecting the presence of specific affixes or morphological characteristics. Generally, morphological recognition of chemical words requires the detection of frequently occurring chemical name fragments like “chlor”, “ethyl”, or “phen”. In addition to orthographic and morpho-syntactic features, string/n-gram properties, grammar/part-of-speech information, and token length are often encoded by rules. NER rules can exploit also features extracted from dictionaries of name constituents or domain knowledge from chemical gazetteers. For instance, the Oscar4 system uses a chemical dictionary as well as syntactic patterns to represent chemical named entities.[158]

Rule can also capture external evidence for NER under situations where chemical names are used in somehow predictive local context. Such context-aware systems have been implemented to exploit contextual clues, like dosage information or treatment duration, as means to detect misspelled drug names and drug names not present in their drug gazetteer in narrative clinical documents (discharge summaries).[376,377]

When multiple rules match a target text, rule preference/priority is used to organize the NER system workflow. In case of an unordered rule list, often ad hoc preference criteria are defined, such as the length of the matching text string, while ordered rules often take into account some statistical measure to sort rules (e.g., sorting by rule f-score or other weighting schema).

In the case of multiword entity names, statistical analysis of name parts extracted from chemical gazetteers can be suitable to select core terms (i.e., meaning bearing elements) and function terms (i.e., function elements or specifiers) that make up entity mentions. Core-terms and function terms can also be defined through manually constructed rules. Narayanaswamy et al. (2003)[369] published a rule-based entity tagger for various entity categories, including chemical names (e.g., “indomethacin”) and chemical parts of names (e.g., “methyl”), which was based on the detection and classification of individual words into chemical core terms (chemical c-terms) and chemical functional terms (chemical f-terms). For recognizing chemical core terms, they used surface feature rules (i.e., morphological features, capital letters, numerals, and special symbols), rules for the detection of chemical root forms/suffixes, and rules derived from IUPAC conventions for naming chemicals. For instance, one rule consisted of the presence of the chemical suffix “-ic” followed by the word “acid” to match chemical c-terms, as in

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

the case of "suberoylanilide hydroxamic acid". Likewise, they used a collection of functional terms corresponding to steroids, drugs, and other chemical term categories. Finally, concatenation, extension, and postprocessing rules were used to refine mention boundaries.[369]

CER systems can be implemented through the use of rules that describe composition patterns or context of chemical entity mentions. Chemical name composition rules try to capture aspects characterizing how names follow predefined conventions or nomenclature rules, as found in the case of systematic IUPAC chemical names and, to some extent, also in INN drug names. Such rules can be expressed using grammars or sets of production rules for transforming strings in a formal language representation. As nomenclature recommendations still allow naming variations (e.g., separating digits in systematic names using dashes or commas), together with the existence of synonyms corresponding to typographical variants (e.g., due to variable use of case, brackets, or whitespaces), rules must capture alternative naming variants.

The LeadMine CER system is a prototypical rule-based system exploiting grammars and heuristic nomenclature specifications to detect chemical entities in text.[332] LeadMine relies on nomenclature and naming convention rules expressed formally through complex expert curated chemical grammars for systematic names (with a total of 486 rules), together with a dictionary-lookup component. An example grammar for systematic names is: Alkane: alkaneStem+"ane", where *alkaneStem*: "meth"|"eth"|"prop". In this system, grammars can inherit also rules from other grammars. Dictionaries can be used as part of rules as they are often practical for detecting semisystematic names, whereas filtering rules help process the initial chemical name gazetteer derived from PubChem and generate a high-quality name list. LeadMine uses several postprocessing rules to modify mention boundaries (i.e., entity extension, trimming, or merging), to detect abbreviations, and to check for correctly nested and balanced brackets.

Some simple postprocessing rules are frequently implemented even by other CER approaches, like supervised learning approaches (section 3.6), in an effort to improve annotation quality. Common applied postprocessing rules are exclusion of mentions, if one of the words is in a predefined stop word list; exclusion of mentions with no alphanumeric characters; or removal of the last character, if it is a dash ("-").

Although less formal and detailed, there are specific nomenclature rules recommended by the World Health Organization (WHO) International Nonproprietary Names (INNs) that can be exploited by rule-based CER approaches. For example, the DrugNer system makes use of naming conventions recommended by WHOINN, and exploits rules that consider drug name stems defined by WHOINN, to recognize drug mentions.[342] By employing such nomenclature rules, it is possible to distinguish between pharmacological substances with regard to pharmacological or chemical families. For instance, to capture names with the affix "-flurane", the regular expression "[A-Za-z0−9]*flurane" is used to match any alphanumeric string ending with such suffix.

Instead of starting the extraction process by examining nomenclature rules, it is also possible to apply a more data-driven extraction process, which examines some selected text corpus and carries out a sublanguage analysis, examining chemical and nonchemical name fragments derived from gazetteers. Heuristics can be used to define how tokenized text is recombined and chemical name boundaries are detected.[367]

Some rule-based CER systems were implemented to process particular types of chemicals or documents from a specific domain. For instance, ChemFrag used rules to recognize organic chemical names,[378] while SERB-CNER (Syntactically Enhanced Rule-Based Chemical NER) uses regular expressions, syntactical rules, heuristics, and recognition dictionary of technical terms and abbreviations to recognize chemical compounds in nanocrystal-development research papers.[379] A later extension of this system combines rule-based and machine learning-based CER.[380] The ChemicalTagger system uses rules and a regular expression tagger to parse experimental synthesis sections of chemistry texts and to mark-up chemistry-related terms.[155]

ChemDataExtractor[365] exploits multiple specialized grammars that merge a list of tags with POS and chemical entity information. Rules are composed of three core elements: the T elements (matching tokens based on its POS or entity label), the W elements (matching the exact text of a token), and the R element (matching text patterns through regular expressions). This tool defines grammars as nested rules that describe how sequences of tagged tokens can be translated into a tree model for entity representation.

A commercial system, the IBM's SIIP (Strategic IP Insight Platform), is an interactive platform for processing patent texts that enables chemical annotators to use a combination of rules and dictionaries to identify chemical names. A set of name grammars were generated by analyzing chemical-related patents manually.[35,381]

Several shortcomings of rule-based systems have promoted the increased use of supervised ML entity taggers. It is well-known that rule construction, domain adaptation, and updating of rule-based strategies imply considerable manual workload, which hinders extension or adaptability to new domains.[382] This makes it challenging to cope with changes of naming conventions over time. Nevertheless, 5 out of 26 teams that participated in the CHEMDNER task used rule-based techniques as part of their CER pipeline,[132] obtaining rather competitive results (notably, the third best system relies exclusively on rules and dictionaries).

Building such rule-based systems required a deep understanding of both the existing chemical nomenclature standards as well as the CHEMDNER annotation guidelines. Surprisingly, two systems relied essentially on the use of lexical resources for chemical names (team 199 and team 222), exploiting a considerable number of different databases and terminologies and obtaining satisfactory results (ranks 11 and 12 in the CEM task).

3.6. Supervised Machine Learning Chemical Recognition

3.6.1. Definition, Types, and Background. When examining chemical documents manually, it becomes clear that systematic chemical names look very different from common English words or surrounding text, largely due to naming conventions imposed by nomenclature recommendations.

An initial exploration of this property was carried out by Wilbur et al. (1999). They applied a segmentation algorithm to split chemical terms into constituent chemical morpheme segments.[36] For instance, the chemical term "triethylaminopropylisothiuronium" was divided into the following segments: TRI-ETHYL-AMINO-PROPYL-ISOTHI-URON-IUM. They

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

applied a Bayesian classifier to classify character strings into chemicals and nonchemicals. Another approach to classify n-grams into chemical and nonchemical substrings using a Naïve Bayes classifier was explored several years later by Vasserman (2004).[383] These initial studies contemplated the CER task as a sort of binary n-gram or string classification problem, categorizing a given substring as being either a chemical or a nonchemical substring. Such strategies have the limitation that they do not specify where exactly the chemical mentions appear in running text, that is, specifying the associated mention boundaries (where mentions start and end within a document). Moreover, these efforts relied on chemical name lists because, at that time, no manually annotated gold standard chemical mention corpus was available.

Since the beginning of 2000, a family of very promising techniques known as machine learning (ML), specifically supervised learning (SL) algorithms, were intensely examined for the purpose of NER tasks. SL algorithms have gradually gained popularity and are substituting other NER approaches when entities have sufficiently large corresponding annotated training data.[319]

ML methods have been used to recognize named entities by posing the underlying problem either as a classification task or as a sequence-labeling problem. Posing named entity recognition as a classification task essentially implies that the underlying problem is to determine the class labels for individual words (or even word characters), instead of labeling entire documents, as was the case of document classification (section 2.5). The used word class labels defined for this classification task are, in the simplest binary scenario, to mark words as being either part or not of an entity mention corresponding to a predefined entity type (e.g., chemical or drug).

For this word or token classification problem, annotated training corpora are examined to automatically induce discriminative features that distinguish them from the surrounding text and learn a statistical model for detecting entity mentions.

When NER is viewed as a sequence-labeling task, sentences are considered to be sequences, whereas words are tokens and entity classes are the labels. For instance, Xu et al.[384] used this approach in CHEMDNER challenge in BioCreative IV, achieving top rank results (88.79% precision, 69.08% recall, and 77.70% balanced F-measure).

The advantage of SL approaches is that they are data-driven techniques in the sense that they make use of information derived directly from documents to model or learn how entity mentions differ from the surrounding text. As compared to rule-based systems that struggle considerably with cases of irregular naming of chemicals (i.e., are not complying explicitly with nomenclature guidelines), ML techniques are flexible enough to distinguish even those mentions that are not following official naming standards. ML-based entity tagging is also more competitive as compared to dictionary-lookup techniques when coping with previously unseen, new, or ad hoc chemical names, as long as they show sufficient morphological or contextual traits that distinguish them from surrounding text. A common characteristic of SL methods is their need of labeled training corpora, typically in the form of exhaustively hand-annotated text corpora, that is, documents with labeled chemical entity mentions. This contrasts with unsupervised learning methods that do not require entity-labeled text, which are currently not a common choice for NER tasks.[385] SL NER taggers are known to have a higher recall, they do not require the development of grammars, and developing such taggers does not require necessarily very deep domain understanding.

3.6.2. Machine Learning Models. To get a general understanding about SL NER algorithms, consider this a word-labeling problem, which can be formalized by assigning to each word w_i one label or $class_i$ and to generate a probabilistic model $P(class_i|w_i)$. For this purpose, the most widespread representation of a sequence of words is a sentence, because a particular named entity mention does not usually span across sentence boundaries (i.e., are constraint to single sentences). Given a sentence, one option is to compute $P(class_i)$ independently for each word. The most probable sequence of tags, using a Viterbi search, then can be computed to establish how to label named entities within the sentence. Given a sentence represented as a sequence of words, considering only the class probability of the current word is usually not very robust. A more competitive representation model requires for the computation of the class probability to take into account, for instance, the preceding, current, and following words $P(class_i|w_{i-1}, w_i, w_{i+1})$, or even to define a local context in terms of a number n of preceding or following words within the sentence.

A range of different SL algorithms have been tested for NER problems,[319] including decision trees (DT)[386] and random forests (RFs),[172] maximum-entropy Markov models (MEMM),[158] hidden Markov models (HMMs),[387−389] support vector machines (SVMs),[390,391] and conditional random fields (CRFs).[141,150,161,175,338,343,344,384,391−393] Providing a detailed overview of the algorithms and mathematical models underlying all of these ML methods goes beyond the scope of this Review.

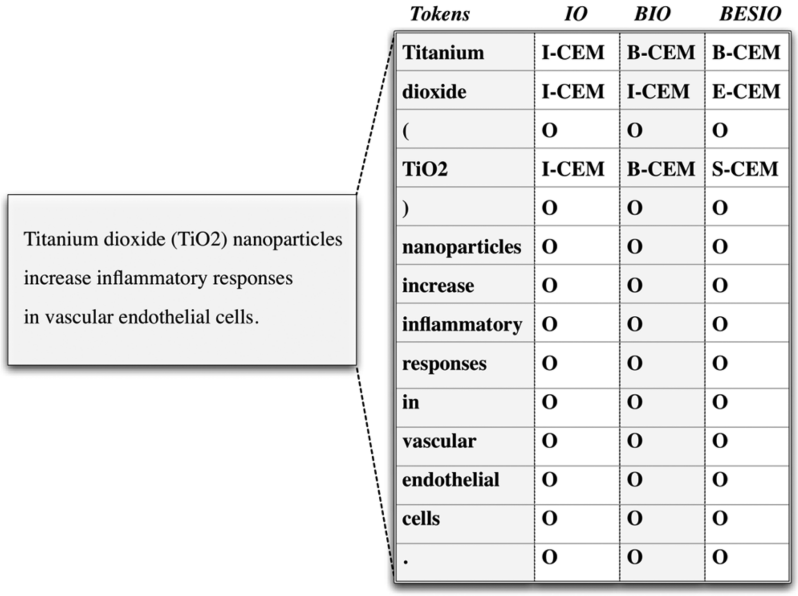
Briefly, maximum entropy models calculate for each used feature (e.g., previous word, current word, etc.) its individual contribution independently. They then combine them multiplicatively under the assumption that each of them contributes separately to the final probability. This can suppose an advantage in situations where two cues occur separately in the used training data. A widely used CER system that employs MEMMs is the OSCAR4 chemical tagger.[158]

Markov chains can be viewed as a kind of stochastic model of processes, represented as a succession of states. The process goes from one state to the next. Each of the corresponding state transitions is associated with a probability. Two frequently used ML techniques based on the Markov principle are HMMs[394] and CRFs.[395]

HMM-based NER taggers generate a distinct statistical model for each name class and also a model for those word sequences that do not correspond to predefined name types. HMMs are a sort of probabilistic automata where a label matches a state while observation symbols represent a word at a state, and state transitions and observation symbols are defined probabilistically. HMM-based entity taggers have been substituted in practice by MEMMs to ease the heavy independence assumptions underlying HMMs,[396] and comparative studies showed that MEMM-based CER strategies yield a better performance when compared to HMM-based chemical taggers.[389]

CRFs can be currently considered as the state-of-the-art method for sequence labeling, including named entity recognition. In fact, 19 out of the 20 teams that used ML methods for CER that participated at the BioCreative CHEMDNER task used this SL approach, including 8 of the top 10 best systems (the other two were rule-based systems).[132]

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761



Tokens	*IO*	*BIO*	*BESIO*
Titanium	I-CEM	B-CEM	B-CEM
dioxide	I-CEM	I-CEM	E-CEM
(	O	O	O
TiO2	I-CEM	B-CEM	S-CEM
)	O	O	O
nanoparticles	O	O	O
increase	O	O	O
inflammatory	O	O	O
responses	O	O	O
in	O	O	O
vascular	O	O	O
endothelial	O	O	O
cells	O	O	O
.	O	O	O

Figure 10. Example case for the most common NER tagging schemas.

CRFs are a type of discriminative undirected probabilistic graphical model trained to maximize the conditional probability of random variables. Publicly available CER taggers like ChemSpot[162] and tmChem[161] are hybrid systems that use CRFs. Seminal work in using CRFs for tagging IUPAC and IUPAC-like chemical names was carried out by Klinger et al. (2008),[141] which also inspired the development of the CRF-based CER tagger CheNER.[345]

3.6.3. Data Representation for Machine Learning. To carry out a sequence-labeling task (and, to a certain extent, word classification), it is necessary to represent text, or more specifically individual sentences, as a sequence of words or tokens. Each token, in turn, is represented by a set of features that are used by the ML algorithm to generate class labels. There are several alternative tagging schemas or ways to encode classes for NER sequence labeling tasks. Figure 10 provides an example case for the most common NER tagging schemas. The IO encoding is the simplest case that only considers whether a token is part of a chemical entity mention. Such an encoding is not particularly suitable to define mention boundaries between two consecutive entity mentions. The most widely followed tagging scheme used by ML-based chemical entity taggers is the BIO, also known IOB, format. This schema labels each token as being either at the beginning of an entity mention (B), inside an entity mention (I), or outside an entity mention (O). This implies that tokens are labeled as "B" if they are the first token of an entity mention, and as "I" if they are part of the following subsequent entity forming words, while all other words that do not form part of the entity mentions are labeled as "O". Other alternative tagging schemas are BEIO, BESIO, and B_{12}EIO,[397] where additional labels are added, a particular tag for tokens at the end of entity mentions (E), single token chemical entity mentions (S), and two labels for the first (B_1) and second (B_2) token of an entity name. Dai et al. examined several entity tagging schemas and arrived to the conclusion that the BESIO schema outperformed other alternative representations.[160]

3.6.4. Feature Types, Representation, and Selection. To build a statistical tagger, each token is represented as a set of features that can be viewed as descriptive properties or characteristic attributes of a particular token (and its flanking tokens) for algorithmic intake. Features are associated both to positive and to negative training examples, that is, tokens that are part of entity mentions and also those that are not part of entities. An overview of feature types commonly used by CER taggers is provided in ref 132. To generate a suitable text representation for building statistical learning models, feature vectors are used. For each token, its corresponding features are encoded as Boolean, numeric, or nominal attributes. Examples of Boolean features could include FirstCap (word is capitalized), AllCap (only contains capital letters) or HasSlash (token contains a slash), HasComa (token contains a coma), HasPrefix1 (token as a specific predefined prefix), HasSuffix1 (token as a specific predefined suffix), IsNoun (has part-of-speech label noun), or HasGreekLetter (token contains a Greek letter). Other aspects that can be encoded as Boolean attributes are POS tags, word classes, or semantic tags. Such attributes have the value "true" if the token does have this characteristic and "false" otherwise. Typical numeric attributes used as chemical entity recognition features are, for instance, token length measured in number of characters, or a counter for the number of times certain characters (e.g., hyphens) or n-grams (e.g., "cyl") appear in a given token. Nominal attributes could, for example, correspond to the token or a lowercased version of a token, or even a pattern that encodes a generalized version of the morphology of the token. For instance, the chemical compound "10-amino-20(S)-camptothecin" can be encoded as its full token shape version "00_aaaaa_00_A_ *aaaaaaaaaaa*", where 0 stands for any number, "_" stands for non-alphanumeric characters, and "a" and "A" stand for lower and upper case letters, respectively.[150] This long word shape representation can be further condensed into its brief shape version or summarized pattern feature "0_a_0_A_a", where consecutive symbol types are conflated into a single symbol.

Another type of nominal attributes are lexicon, dictionary, or gazetteer lookup features where the token is matched against a list of terms, such as entity names or entries in a stop word list that was previously constructed.

It is also common to distinguish between word-level attributes, list lookup attributes, and document or corpus attributes. Word-level attributes include the types of morpho-

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

logical features previously described, while list lookup attributes are often useful to include domain-specific information. Typical document or corpus features include corpus frequency and co-occurrences of other entities in the context.

Prototypical features used by SL-CER methods are derived from morphological/orthographic, lexico-syntactic, and grammatical properties of tokens. Many commonly exploited features examine the presence of specific combinations of orthographic features, including certain patterns of consecutive characters (e.g., character n-grams), token case, presence of digits, special characters (e.g., hyphens, brackets, primes, etc.), and symbols (e.g., Greek letters, @, \$, etc.). Especially useful are features looking at certain affixes at the beginning, within and at the end of tokens as they can help to identify the morphology of systematic chemical names. Context features used by SL CER are usually constrained to the local context of a token, that is, considering a contextual window of previous (w_{i-n}) and next words (w_{i+1}) to the current token (w_i).

3.6.5. Phases: Training and Test. The development of the previously described SL-based CER systems is typically organized into two distinct phases, the training-phase and the test phase. During the training phase, a statistical language model is generated. This requires the selection of suitable training documents that are exhaustively labeled with entity mentions (i.e., all tokens must have a class label), the encoding and examination of suitable features and feature extractors, and training of the statistical sequence model. During the test phase, this predictive model is applied to assign labels for each token in a new text.

3.6.6. Shortcomings with ML-Based CER. Although SL-methods are the general choice for recognizing chemical entities, they also have some limitations, which are mainly tied to the critical need of a significant amount of high-quality manually annotated training data, domain experts (deep domain knowledge), and proper annotation guidelines to carry out such an annotation process. For very specific subtypes of chemical entities or entities of chemical relevance as well as particular target domains or certain document types, labeled training corpora might be missing.

Moreover, feature selection and optimization as well as the detection of term boundaries is still a challenging problem. Finally SL-methods operate generally at the level of individual tokens, and thus they are very sensitive with respect to the used tokenization strategy.

3.7. Hybrid Entity Recognition Workflows

The design of CER systems is driven by the primary goal of obtaining competitive recognition, regardless of the actual methodology used. Therefore, existing chemical mention taggers explored the complementarity of different tagging strategies. In fact, top scoring systems participating in BioCreative challenges did make use of hybrid strategies that combined several of the following techniques: SL-based techniques (based on various CRF models), rules/patterns for certain types of chemical mentions, and dictionary-lookup using chemical gazetteers.[132] Most of these CER systems are primarily SL-based and use dictionary-lookup features (either of entire terms or parts of tokens, such as chemical stems or n-grams) together with rule-based postprocessing techniques.

Chemical dictionary-lookup is integrated into hybrid systems primarily on the basis of either SL-learning[161] or rule-based CER.[332] In the case of SL-learning CER, dictionaries are commonly used to generate list-based features or as part of stop word filtering steps, or alternatively they can be used as a tagging strategy by its own to complement SL results.[398] In the case of rule-based CER, names derived from chemical gazetteers are used as components of specific recognition rules or to generate entity mention patterns. Hybrid CER approaches can also be used to cover different types of chemical names. For instance, rules may be used to detect systematic names and dictionaries may be used to detect trivial names, and in the combined results, the longest mention is retained (in case of overlapping results).[358] Another hybrid system combining both dictionary lookup and rules manually constructed by experts with deep domain understanding is LeadMine.[332]

Rule-based processing steps are integrated into hybrid CER systems to handle different aspects of recognition. They are commonly used to identify highly structured chemical entity mentions, such as chemical formula and sequences of amino acids through pattern matching or regular expressions. At the level of rule-based preprocessing for SL-CER taggers, critical aspects include chemical document-adapted tokenization rules and sentence segmentation. Rule-based techniques are commonly applied to carry out extensive automatic post-processing of results produced by SL-CER taggers. This includes checking whether brackets and parentheses are balanced within detected chemical mentions[161,399] or adjusting chemical mention boundaries. Rule-based methods are also used during postprocessing steps to recognize mentions of chemical abbreviations and acronyms. For instance, tmChem relied on AB3P (Abbreviation Plus P-Precision) for detecting potential abbreviations of chemical names.[161] Rule-based techniques are also used to resolve conflicting results generated by multiple CER techniques. A widely used hybrid CER system applies CRFs mainly for IUPAC chemical names and dictionaries (ChemIDPlus) to derive trivial, brand names, and abbreviations for chemicals is ChemSpot.[162]

Dictionaries or rule-based systems can also be applied to generate surrogate annotations for manually annotated training data sets. In this line of work, rule-based strategies have been explored to produce additional labeled entity mentions that can be used as training data for SL-CER systems.[400]

Another hybrid system is ChemDataExtractor that integrates dictionaries (Jochem), rules, and machine learning (CRF-CER trained on the CHEMDNER corpus) techniques. It uses patterns for the recognition of chemical database identifiers and formulas.[365]

3.8. Annotation Standards and Chemical Corpora

3.8.1. Definition, Types, and Background. The recognition of chemical entities as well as other types of entities of chemical relevance depends heavily on the existence of labeled text collections or corpora to evaluate the reliability and performance of automatically extracted mentions. Moreover, such resources are a critical key resource in case of SL-based CER strategies for the design, training, and evaluation of NER models in the very first place.

Annotation corpora consist typically of (manually) labeled text where all of the mentions of predefined entity types are tagged and, sometimes, also mapped to concept identifiers from ontologies or databases. As the corpus construction process is a highly labor-intensive task, such resources are scarce and constitute extremely valuable resources for chemical TM. Corpora manually annotated by domain experts can be viewed as a gold-standard resource (Gold Standard Corpora) that

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

enables the comparison of various methods, evaluating their relative performance, and the reproduction of obtained results.

A milestone in the construction of text corpora for language processing purposes was the publication of the Brown Corpus by Kučera and Francis, consisting of a corpus of modern American English compiled from different sources with a total of around one million words.[401]

The construction and release of sufficiently large text corpora containing chemical information annotations started relatively recently and will be described in this subsection. It can be said that the release of such annotated resources promoted significantly the research and development of chemical entity recognition systems. In particular, the release of resources such as the CHEMDNER corpus resulted in triplicating the number of published CER taggers.[318] Despite their extensive use, and in contrast to the over 36 corpora constructed for the biomedical domain,[232] only few manually labeled chemical corpora are currently accessible.

The EDGAR corpus was one of the first attempts to annotate chemical entities or, more specifically, drugs, in addition to other entity types such as genes and cells. This corpus was constructed at the beginning of 2000 and comprised annotations for 103 cancer-related PubMed abstracts.[359]

In the case of SL-NER systems, labeled annotation data are commonly divided into two disjoint sets: the training collection from which the model infers its parameters and the test collection used to evaluate the quality of the learned model. This implies that annotated corpora should be sufficiently large to accommodate both of these subsets.

The annotation process itself can be viewed as a sort of semantic enrichment or addition of metadata at the level of adding individual entity tags to the text by following specific annotation criteria or guidelines. To overcome the considerable workload burden associated to manual annotation, corpora with a usually lower annotation quality have also been generated by automatic means, known as silver standard corpora. One example of a silver standard corpus that contains also chemical mentions is the CALBC corpus.[402] Another silver standard corpus was released after the first CHEMDNER challenge consisting of automatic annotations returned by systems that participated in this competition (CHEMDNER silver standard corpus).[318]

Another relevant aspect, both for the consumption of corpora as well as for their construction, are annotation guidelines or rules. Annotation guidelines are essentially a collection of (usually written) rules describing what entities should be labeled and how. This usually implies defining what types of mentions should not be tagged, how to deal with ambiguous cases, and criteria to define the exact boundaries of entity mentions. Manual annotation can only be done systematically through the use of computational text annotation tools. Refer to Neves and Leser for a survey on existing text annotation software.[232] Differences in annotation guidelines are usually tied to the scope of the corpus and the resources that should be implemented from a given corpus. Also, the annotation effort varies considerably depending on the target documents. For instance, patents are considerably longer than scientific articles, with a lower word density.[403]

To measure the consistency and quality of manual annotations, a common strategy is to compare manually labeled mentions. This implies to check whether the offsets of mentions generated by different individuals are the same. To measure quantitatively the annotation consistency, the simplest metric is the interannotator agreement (IAA) or intercoder agreement score, which is based on the percentage agreement of manual annotations between different annotators.

The representation of text annotations and standardization of text annotation formats is still an open research question. In principle, it is possible to distinguish between inline and offline annotation types for defining the labels of entities associated to documents. In the inline annotation strategy, specific tags, labels, or elements that delimit entity mentions are inserted directly into the target text, often represented as XML tags. In the case of stand-off annotations, the original document is not modified; that is, annotations are not embedded into the actual text itself and are stored separate from the document.[404]

For storing, processing, and distribution of corpora, the use of a proper annotation format is also important. Currently, there is no universally accepted, standard chemical text annotation format, but there are several de facto standards or widely used formats. An annotation and representation format proposed for chemical documents is SciXML, which defines inline annotations. It was the original input format required by the CER tool Oscar3.[154] Currently, an alternative way to represent chemical text annotations is through the Chemical Markup Language (CML), whose initial development started back in 1995.[405,406] Another text annotation format, mainly used as a standardized way to represent manual or automatically generated text annotations, is the BioC format.[407]

3.8.2. Biology Corpora with Chemical Entities. The recognition of chemical entities is relevant also for other research disciplines beyond chemistry itself. Therefore, annotations of chemical entities are done in domains such as medicine, pharmacology, and particularly biology. The biological text corpus with the highest influence and importance is the GENIA corpus,[133] consisting of a collection of PubMed abstracts annotated manually with a range of different concept classes, including chemicals. Chemical concepts, as defined in GENIA, can be regarded as a quite broad-spectrum type of chemical substances that can not necessarily be correlated to particular chemical structures.

Another early corpus construction effort was done by Narayanaswamy et al., which prepared a corpus of just 55 abstracts retrieved by keyword searches related to acetylation and that encompassed a number of chemical entity mentions.[369]

The PennBioIE CYP 1.0 corpus is another hand-annotated corpus from the Life Sciences field that labeled chemical entity (substance) mentions. It contains articles related to the inhibition of cytochrome P450 enzymes. Entity mentions annotated as substances in this corpus included also protein names.

Annotating entity names of relevance for metabolic pathways (yeast metabolism) was the main aim of the Metabolites and Enzymes corpus, containing annotations of metabolite mentions in 296 article abstracts.[408]

Several corpora providing annotations for drug mentions in scientific abstracts have been constructed. The ADE corpus hosts annotations for drug-related adverse effects in 3000 abstracts, while the EU-ADR corpus has annotations for drug−target and drug−disease relations extracted from 300 abstracts.[409] Drugs and their relations have been tagged in the DDI (Drug−Drug Interaction) corpus, which contains a total of 700 documents (abstracts and DrugBank[410] database records).[411]

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Only few manually annotated corpora for full text articles have been prepared so far. The CRAFT corpus contains 97 full text biomedical articles annotated with different concept types, including chemical concepts.[412] CRAFT chemical concepts are constrained to those that are covered by the ChEBI ontology,[412] implying that other chemical entity mentions are not annotated. The HANAPIN corpus is another full text scientific article corpus. It consists of 20 articles from the journal Marine Drugs, annotated with several entity types as well with linguistic information. One type of annotated entity corresponded to chemical compounds (280 mentions).[413]

Schlaf and colleagues prepared a patent corpus that focused on chemical entity associations with diseases, extracted from 21 U.S. patents. Those annotations were automatically generated at first and then manually revised.[414]

3.8.3. Chemical Text Corpora. Several manually annotated resources were developed relatively recently that focused primarily on chemical entities and their mentions in text. These chemical corpora are quite heterogeneous according to various properties. At the document type level, one can primarily distinguish between corpora using scientific literature or patents. At the document content level, annotations have been done using only abstracts (and titles) or the entire full text document. Differences are also found regarding the size and format of the different chemical corpora as well as in terms of annotation scope and the underlying manual annotation criteria or guidelines.

Recent work by Habibi et al. examined existing differences between chemical corpora (cross-corpus analysis) to show how adaptable CER systems are when trained on one corpus and tested on another one.[415] This study estimated that the performance CER systems trained on scientific abstracts and evaluated in terms of balanced F-score degraded by about 10% when applied directly to patent abstracts, and decreased by around 18% when run without adaptation on patent full texts. They also proposed that ensemble systems, which combine results provided by different CER systems, can be a strategy for cross-corpus adaptation.[415]

Such variability in performance can be partly explained by differences in the annotation process. A foundational work in defining manual criteria for annotating chemical information in abstracts and chemistry journals was done by Corbett and colleagues,[416] and then refined in the context of the CHEMDNER competition and its corpora.[317,318]

The first efforts to generate comprehensively annotated chemical corpora were the Sciborg (42 full text chemistry papers)[416,417] and the Chemistry PubMed corpus (500 PubMed abstracts).[389,416] The Sciborg corpus was annotated by three chemistry experts and had a total of 4102 manually annotated chemical compound mentions.[417]

These two chemical corpora shared basically the same detailed annotation principles, providing rules for labeling mentions of chemical compounds, reactions, chemical adjectives, enzymes, and chemical prefixes. Although these two corpora were not publicly released, the description of their annotation process inspired the construction of other chemical corpora.

Different classes of chemical names, like systematic names or trivial names, show properties that are of practical importance for its automatic recognition. To take into account this naming characteristic, the developers of the publicly distributed Chem EVAL corpus or SCAI corpus defined more granular chemical mention classes: IUPAC (systematic and semisystematic chemical names), PART (partial IUPAC names), TRIVIAL (trivial names), ABB (abbreviations and acronyms), SUM (sum formula, atoms, molecules, SMILES, and InChI), and FAMILY (chemical family names).[335] This granular chemical entity mention class distinction was of practical importance for the construction of follow-up chemical corpora, even though the original corpus was limited in size (100 abstracts with 1206 chemical mentions) and detailed annotation guidelines were not released together with this corpus.

Another freely accessible chemical corpus is the ChEBI Patent Gold Standard corpus or Chapati corpus,[418] resulting from a collaboration between the European Patent Office and the ChEBI database.[419] It is composed of 40 patents with 18 061 chemical mention annotations, around one-half of which are normalized to ChEBI database records.

Because of the practical importance of chemical patents, another corpus of 200 full text patents with chemical mention annotations was described in 2014,[420] the BioSemantics Patent corpus. Here, a slightly different strategy was used to deal with the lengthy content of patent documents. These patents were automatically preannotated with a CER system. Afterward, during a manual revision phase, automatically detected chemical mentions were examined, and potentially corrected or removed, while missing mentions were added manually.

3.8.4. CHEMDNER Corpus and CHEMDNER Patents Corpus. Two chemical corpora known as the CHEMDNER corpus and the CHEMDNER patents corpus have been publicly released together with detailed annotation guidelines as part of the BioCreative CHEMDNER community challenges.[132,316−318]

These two corpora were large enough to be used as data sets for a community challenge with the aim to promote development and evaluation of CER systems (see section 3.9). This implies that they were used as Gold Standard data sets for the comparison of automatically detected chemical entity mentions against the manually delimited mentions of chemicals annotated in these corpora.

Both of these corpora relied on very similar annotation guidelines with minor adaptations to address differences in scope, or annotation goal of abstracts derived from scientific articles when compared to those from patents. The CHEMNDER corpora annotation guidelines were concerned with ways to define what can be regarded as a chemical compound mentioned in text, focusing on those mentions that, at least at a certain extent, can be associated to structural information. Chemical entity mentions of this kind are known in the context of the CHEMDNER task as Structure Associated Chemical Entity Mentions (SACEMs) and were exhaustively annotated by chemists with experience in literature curation. Example cases of true SACEMs are “nitric oxide”, “resveratrol”, or “malondialdehyde”, while other concepts like “pigment”, “hormone”, “antibiotic”, “metabolite”, or “chelator” do not represent SACEMs and thus were not annotated. Together with the CHEMDNER annotation data and guidelines, additional descriptions of relevant aspects for the interpretation and use of these corpora were published, including the strategies used for corpus document selection/sampling, characteristics and expertise of the human annotators, measurements of corpus annotation consistency, characteristics of the corpus format, and descriptive definitions of annotated chemical name classes. The annotation of chemical name classes was tackled through a granular annotation schema that comprised seven classes of SACEMs: SYSTEMATIC (system-

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

atic or semisystematic names), IDENTIFIERS (chemical database identifiers), FORMULA (chemical formula), TRIVIAL (trivial, common, and trade names of chemicals and drugs), ABBREVIATION (abbreviations or acronyms corresponding to chemicals), FAMILY (chemical families), and MULTIPLE (noncontinuous mentions of chemicals in text). An examination of the CHEMDNER corpus showed that TRIVIAL (30.36%) and SYSTEMATIC (22.69%) names sum up more than one-half of the chemical mentions, while other chemical name classes appear with a lower frequency in text, for example: ABBREVIATION (15.55%), FORMULA (14.26%), FAMILY (14.15%), and IDENTIFIERS (2.16%).

Most chemicals in this corpus occurred only one or two times (over 72% of unique name strings) and were rather long names with a mean chemical mention length of 10.01 characters (median 8). The longest name corresponded to a systematic name with a length of 349 characters.

The consistency of the CHEMDNER corpus was assessed through an interannotator agreement study between human annotators, obtaining a percentage agreement of 91% for exact chemical mentions, regardless of the chemical name class label, and 85.26% when in addition to the chemical mention offset the name class label was also identical.

The CHEMDNER annotation guidelines consisted of documents of roughly 21 pages that contained annotation rules supplemented with example cases and structured into six types of rules. General rules elucidated how to use external knowledge sources and how to deal with unclear mentions. Positive and negative rules described what should and should not be labeled, respectively. Class rules provided criteria for labeling at the level of chemical name classes, while orthography/grammar rules provided instructions for defining entity mention boundaries. Finally, multiword entity rules had to deal with criteria for delimiting multiword chemical entities.

Each of the CHEMDNER corpora contained three subsets, a training set, a development set, and a test set, to clearly define what portions of the corpus should be used to implement, train, and tune the CER systems (training and development data) and which part had to be used for evaluating the performance of the final CER systems (test set).

In the case of the CHEMDNER corpus, the entire data set comprised 10 000 recent PubMed abstracts, while the CHEMDNER patent corpus contained 21 000 medicinal chemistry patent abstracts (and titles). Each of these two corpora contained three randomly sampled subsets for training (3500 PubMed abstracts and 7000 patent abstracts), development (3500 PubMed abstracts and 7000 patent abstracts), and testing (3000 PubMed abstracts and 7000 patent abstracts) the CER systems.

In total, the CHEMDNER corpus contained 84 355 manually labeled chemical entity mentions, while the CHEMDNER patent corpus had 99 634 chemical mentions labeled by hand.

Both CHEMDNER corpora provided details on document selection criteria. For the CHEMDNER corpus, abstracts were selected from journals representing diverse chemistry and chemistry-related disciplines, including organic chemistry, physical chemistry, chemical engineering, medicinal chemistry, biochemistry, pharmacology, and toxicology. In the case of the CHEMDNER patent corpus, patent abstracts from various patent agencies, the World Intellectual Property Organization (WIPO), the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), Canadian Intellectual Property Office (CIPO), the German Patent and Trade Mark Office (DPMA), and the State Intellectual Property Office of the People's Republic of China (SIPO), were selected. A subset of these patent abstracts was chosen taking into account several criteria including their publication date (2005−2014) and associated IPC codes (A61P and A61K31).

The CHEMDNER corpora were distributed using standoff annotation formats, and thus annotation information was kept in a separate file from the original documents. These annotation files contain information on the character offsets of each chemical mention together with its associated chemical name class, being released in simple tab-separated formats as well as in the BioC format.[407]

3.9. BioCreative Chemical Entity Recognition Evaluation

3.9.1. Background. From the point of view of a user of CER systems, or even of chemical text processing tools in general, it is difficult to determine which tool is the most adequate or performs best on a specific problem, document collection, or task. To determine comparative performance of different tools, one option is to perform a benchmark study by evaluating various applications on a common data collection and using the same evaluation metrics. The drawback of benchmark studies is that they are associated with a considerable workload. They typically require some adaptation of the examined software to cope with a common data input and output. They also require designing adequate evaluation strategies that suit all tools and preparing Gold Standard benchmark data sets. Moreover, benchmarking does not actively involve the developers of NER systems in the benchmark process, and thus cannot determine the state-of-the art for a particular task as the various systems are not adapted or optimized for a particular metric and/or evaluation data set.

Community challenges are an alternative evaluation exercise to benchmarking. They actively promote the development of new tools, which might explore different methodologies or algorithms. These systems are, in turn, evaluated and compared on the basis of impartial evaluation metrics and on a common evaluation data set. When community challenges are repeated over time, it is possible to monitor progress over time or to compare results obtained on different document collections.

The advantage for systems that participate in open community challenges is that, traditionally, task organizers are in charge of assembling and providing access to large enough high-quality annotated gold standard data. Organizers of evaluation efforts typically have to collaborate with domain experts to prepare annotation guidelines and data sets. Moreover, they usually also provide software to automatically score the performance obtained by the various participating systems, evaluating their results on a common blind test set.

Already at the former MUC community challenges, tasks were posed to address the recognition of named entity mentions, but outside the scientific domain.[421] The work of Huang et al. provides an overview of the various TM and information extraction community challenges that were carried out in the biomedical field, covering also tasks related to chemical and drug named entity recognition efforts.[422] As NER is a fundamental building block for semantic search, and for higher-level NLP and relation extraction processes, such community challenge NER tasks have attracted considerable attention.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

At the third i2b2 challenge, a task was posed with the aim of asking participating systems to recognize medication mentions (and other information) in deidentified discharge summaries. The task organizers provided overall 1243 summaries, that is, 696 for training and 547 for testing purposes. From the training set, only 17 summaries had conventional manually labeled gold standard annotations. It is noteworthy that the used gold standard test set annotations were collectively produced on the basis of all of the automatic submissions for a subset of 251 summaries. Medications in this challenge comprised names, brand names, generics, and collective names of prescribed substances used to treat patients, for example, including names like "heparin" or "coumandin". Out of the 20 participating teams, most of the top 10 best performing systems relied on some kind of rule-based technique; the best performing system achieved a balanced F-measure of 85.7%.

Another challenge, the DDIExtracion (drug−drug interactions) challenge, also focused on drug-related information, but cannot be considered a true NER task. The 10 participating teams were asked to classify pairs of candidate drugs that co-occurred in a particular sentence whether they were in an interaction relationship or not. The used DrugDDI corpus contained a total of 3160 of these annotations extracted from 579 documents, obtained from the DrugBank database. The best result was obtained by a system that combined ML techniques with case-based reasoning, yielding a balanced F-score was of 65.74%.

These and other efforts[423−425] had as a primary goal the extraction of some particular relationship rather than focusing on the entity recognition task itself. Moreover, annotations were provided by experts from health-related disciplines or biosciences and not by chemistry experts.

3.9.2. CHEMDNER. Since the first BioCreative challenge evaluation, individual tasks focused specifically on the recognition of named entities, under the assumption that solving more complex tasks required modularizing the underlying problem and addressing essential aspects separately.[426] The BioCreative organizers posed the CHEMDNER (chemical compound and drug name recognition) community challenge, with the aim of encouraging the development of original, competitive, and accessible CER systems.

Two comparative assessment tasks, known as the CHEMDNER (BioCreative IV) and the CHEMDNER patents (BioCreative V) tasks, focused primarily on the detection of chemical entity mentions in running text, as part of the BioCreative competitions.[132,316−318] These tasks represent the first effort that systematically tried to assess the performance of CER systems on a common evaluation setting ground for scientific abstracts and patent abstracts.

During the CHEMDNER task, a subtask named CEM (chemical entity recognition) requested participating teams to provide systems able to recognize the exact mention offsets of chemicals in PubMed abstracts. As part of the CHEMDNER patents task, the CEMP (chemical entity mention in patents) subtask requested essentially the same thing, but using as document input medicinal chemistry patent titles and abstracts written in English language.[132,316−318] The used Gold Standard data sets, that is, CHEMDNER corpora, were already introduced in the previous section 3.8.

To assess the performance of automatic recognition of chemical entities, automatically recognized mentions represented by their exact character offsets (start and end character indices, which delimit a specific mention) were compared to the offsets of mentions labeled manually by chemical annotators. This implied that automated CER systems had to return, given a document, the start and end indices of all of the chemical entities mentioned in this document.

Only if the automatically produced mention corresponded exactly to the manually labeled chemical was it regarded as a true positive prediction. Partial chemical mention overlaps or any other mentions that did not match precisely manual annotations were considered to be false positive hits. Those manually labeled chemical mentions that did not have a corresponding automatic detection were scored as false negative predictions.

The evaluation metrics used for the CHEMDNER tasks to assess team predictions were recall, precision, and balanced F-measure (the main evaluation metric).

The CHEMDNER tasks were arranged temporally into several competition phases or periods. During the initial registration and task announcement phase, a small sample collection of illustrative annotations and predictions were distributed among participants to exemplify the type of requested chemical mention annotations. At a later stage, known as the training period, a collection of manually annotated chemical entity mentions (training data) was released. This period was followed by the development phase, when additional annotations similar to the training data were released to allow fine-tuning of the final CER implementations. Finally, during the test period, a blind collection (annotations were held back) of documents (PubMed and patent abstracts) was released. Participating teams had then to return (a predefined short period of time) their automatically generated chemical mention predictions using a common prediction format. Each participating team could send a maximum of five different prediction runs.

A total of 26 systems from commercial and academic research teams (22 academic and 4 commercial groups with a total of 87 researchers) used the CHEMDNER task of BioCreative IV to evaluate their CER tools.[132,318,427] They returned a total of 106 individual system runs for this chemical entity recognition task. In the case of the CHEMDNER patents task, overall 21 teams submitted 93 different runs for the detection of chemicals in medicinal chemistry patent abstracts.[317] For the CHEMDNER patents task, submissions were handled through a new evaluation and visualization platform called Markyt to support online comparative evaluation assessment.[316]

Both CHEMDNER tasks had analogous evaluation settings, and the used training and test data sets relied on highly similar but not totally identical annotation criteria.

Neither of these two tasks evaluated the association of extracted chemical mentions to chemical structures (name-to-structure) or some chemical database identifiers (entity normalization or grounding), a particularly challenging problem,[428] which nonetheless was regarded by the task organizers as a distinct issue independent from the actual NER task. Although the CHEMDNER corpora were annotated at a more granular level, by labeling mentions according to seven chemical name classes (described in section 3.8), classification of chemical mentions according to these classes was not examined.

The best performing CER system at CHEMDNER task (PubMed abstracts) obtained a balanced F-score of 87.39%, a result that is very close to the interannotator agreement (91%). The best precision for this task was of 98.05% (with a recall of

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Table 5. Software for Name to Structure Conversions[a]

	provider (year)	supported chemical classes	features	batch mode	availability	URL
name = struct	CambridgeSoft (1999)	IUPAC, IUBMB, CAS, semisystematic, trade names, trivial	typo errors, warnings about ambiguous English	yes	commercial	http://www.cambridgesoft.com/support/DesktopSupport/Documentation/N2S/
ACD/name	ACD/Laboratories (2000)	systematic, derivatives of 180 natural products, semisystematic, trivial, trade and registered names (INN, USAN, BAN, JAN), CAS Registry, EU numbers, InChi, SMILES	warnings about ambiguous	yes	commercial	http://www.acdlabs.com/products/draw_nom/nom/name/features.php
Lexichem	OpenEye (2005)	IUPAC 79/93 /200x, CAS, traditional, systematic, MDL/Beilstein, AutoNum, OpenEye	multilingual (16)	yes	commercial	http://www.eyesopen.com/lexichem-tk
naming	ChemAxon (2008)	IUPAC (English, Chinese, Japanese), systematic, trivial, drug commercial names, CAS, CAS registry number	multilingual supports user customized dictionaries and larger macromolecules	yes	commercial	https://www.chemaxon.com/products/naming/
OPSIN	Unilever Centre (2008)	systematic organic nomenclature, including support to carbohydrates and peptides	available as standalone JAR warnings about ambiguous typo errors		free	http://opsin.ch.cam.ac.uk/
NamExpert	ChemInnovation Software	IUPAC, 8000 drug names		yes	commercial	http://www.cheminnovation.com/products/nameexpert.asp
IUPAC DrawIt	Bio-Rad Laboratories	systematic IUPAC nomenclature, common names, generic names of drugs		yes	commercial	http://www.bio-rad.com/
name to structure (IC_{N2S})	Infochem	IUPAC, CAS names, semisystematic, abbreviated, trivial, brand names	based on a huge dictionary with over 32 million entries		commercial	http://infochem.de/mining/icn2s.shtml

[a]Prepared in November 2016.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

17.90%), while the highest recall was 92.11% (with a corresponding precision of 76.72). As a general trend, the precision scores of participating systems were noticeably better than the corresponding recall values. Average recall values also varied depending on the chemical mention classes, indicating that overall systematic and trivial names were associated to higher recall values when compared to the other types of mentions. Overall, 99.99% of the manually annotated mentions were detected at least by a single participating system.

The results of the CHEMDNER patents task were very similar to those obtained on scientific abstracts. The top scoring system obtained an F-score of 89.37%. The system with the overall best precision obtained a score of 89.71% (with a recall of 88.22%), while the overall highest recall was of 93.14% (with a precision of 79.67%).

The CER methods explored for the CHEMNDER tasks comprised mostly SL methods. The method of choice for most top performing teams was based on CRFs relying on a large collection of different features, similar to those discussed in section 3.6. Several participants also explored the adaptation of chemical gazetteers and chemical domain-specific rules. Especially problematic cases were single letter chemical entity mentions due to their high degree of ambiguity (e.g., "I", "O", "P", or "H"). Another issue was the detection of trivial chemical mentions corresponding to dyes. In the case of systematic names, finding the correct boundary of very long mentions was challenging for some systems.

4. LINKING DOCUMENTS TO STRUCTURES

As IUPAC guidelines in practice do allow some variability with respect to how names are constructed, together with existing alternative typographical and spelling variants, several chemical entity aliases can be mapped back to a single structure.[324] Moreover, alternative systematic (CAS and Beilstein-Institut), semisystematic, and trivial nomenclatures, as well as the tremendous number of synonyms (e.g., brand names and database identifiers) pose additional difficulties for text-based searches,[429] which are not surpassed by strategies to normalize nearly identical names.[430−432] Still, the normalized chemical names in published literature are not as correct as expectable.[30] Also, chemical trademarks and brand names are commonly annotated in chemical databases, but are not indexed in public patent databases. Thus, identifying the chemical structure associated to a chemical name does not only allow structural searches, but also allows normalization of chemical names and avoids dealing with specific languages.[433] Moreover, disposing of the chemical structures contained in documents opens the door to their chemical registration in molecular databases and a variety of posterior computational analysis, as discussed below. However, it should be noted that chemical structure searches do not replace text-based searches (e.g., many pharmaceutical companies publish reports on compounds in advanced preclinical or clinical phases without disclosing its corresponding structure).

Authors can also present chemical structural information in documents, especially in case of supporting/Supporting Information of scientific articles, in the form of plain text 3D X, Y, Z atom coordinate values. Tools like ChemEngine have been implemented to automatically extract 3D molecular XYZ coordinates and atom information from articles with the aim to directly generate computable molecular structures.[434] This system used pattern recognition and regular expressions to detect molecular coordinates and distinguish it from surrounding nonmolecular free text. After generating the atom coordinate matrix data from the previously detected molecular coordinates, tools like ChemEngine build molecules using the bond matrix and the atom connectivity. Finally, the automatically generated molecules are examined through application of important filtering parameters (e.g., bond length/angles) before returning the final structure in formats like SDF.

Together with optical structure recognition (OSR) or optical chemical structure recognition (OCSR) methods (section 4.2), name-to-structure conversion algorithms (section 4.1) and chemical entity grounding are the main approaches to link documents to structural information.

4.1. Name-to-Structure Conversion

Name-to-structure conversion is the process of generating the chemical structures (chemical diagrams) from chemical names. Earlier works on the direct conversion of names to formulas were developed by Garfield in the 1960s, using a dictionary of name parts or morphemes.[324] CAS then reported internal procedures based on nomenclature rules for the automatic conversion of CAS names and other systematic nomenclatures into chemical structures.[435,436] Later, in 1989, Cooke and colleagues applied grammar-based techniques to the recognition of IUPAC systematic chemical nomenclature and its translation to structure diagrams.[437−439] Soon after, rule-based approaches prompted as an alternative to grammar-based approaches for systematic nomenclature, with the first commercial program, named Name = Struct, being launched in 1999.[117]

In general, systematic nomenclature conversion is based on the parsing of the chemical names and the application of syntax analysis. The procedure starts by dividing input names into name fragments of known type (lexemes) included in internal look-up tables, locants, enclosing marks, and punctuations. This is followed by the syntactic analysis of the chemical name according to the chemical nomenclature grammar. Each fragment name then is assigned its structural meaning, and connections are derived between the different fragments.[117,440] For trivial names and registry numbers, lookup tables are unavoidably required, what makes this approach very sensible to the quality of the internal dictionary (comprehensiveness) and to badly annotated associations between chemical names and structures (e.g., omitted stereochemistry). The difficulty of disambiguating noncommon abbreviations (e.g., different from "DMSO" and "EDTA") depending on their context suggests that the conversion of any trivial name shorter than about 5 or 6 characters is not safe.[440] In Table 5 is tabulated a list of currently available software, with a description of the coverage of chemical mention types and whether the program handles (to some extent) typographical errors generated by OCR or misprints and/or cautions on detected ambiguous names arising from nonstrictly systematic IUPAC nomenclature, or from different word senses (e.g., "oxide" meaning either a ether "dimethyl oxide" or the functional group oxide ("trimethylphosphine oxide")). Formatting issues (e.g., capitalization, font type, or style) and most punctuation marks are ignored by these programs. However, the internal workflows of most of them have been hardly described, with the exception of Name = Struct[117] and OPSIN.[441] Moreover, some prototype systems have been published, such as CHEMorph,[26,442] but to date no actual tool has been released. To benchmark name-to-structure software, some metrics have been developed on the basis of a string comparison between the canonical isomeric SMILES of

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

the starting structure and the final structure (i.e., after name generation and conversion back to canonical isomeric SMILES).[443] Because of the different performance and coverage, a good strategy is to combine different name-to-structure conversion software and intersect comparisons, as currently done by SureChEMBL,[74,75] which uses Name = Struct, ACD/Name, Lexichem, Naming, and OPSIN. Other approaches, such as LeadMine,[332,444] combine the use of OPSIN with dictionaries from sources like ChEMBL. A similar approach combining dictionary-based methods and OPSIN is used by OSCAR.[158] Reviews on the opposite process, that is, conversion of structures into names, have also been published.[445]

4.2. Chemical Entity Grounding

4.2.1. Definition, Types, and Background. Although the detection of chemical entities in text is a fundamental step for chemical TM and retrieval systems, it can only be of practical relevance if systems are capable of linking recognized mentions to particular real-world chemical objects, represented either by a chemical structure or by a unique chemical database identifier. Transforming chemical names directly into structures based on chemical name internal rules is one way to address this issue, but, in practice, when looking at chemical names found in documents, they do not always follow chemical nomenclature recommendations.[318] This downgrades the success rate of strategies that attempt to convert chemical names into structures. Therefore, a number of chemical names cannot be unambiguously converted into structures, and some names cannot be converted at all into structures using traditional name-to-structure conversion algorithms.[428] Revising and associating these chemical mentions to structural information or chemical databases by hand requires a substantial amount of time and workload as well as manual linking criteria. For manually associating chemical names to database identifiers, one possible strategy is conducting chemical name search queries in resources such as PubChem, ChemSpider, Google, or SciFinder.[428]

A complementary strategy is to assign chemical database or concept identifiers to chemical name mentions automatically. This entails grouping together all chemical mentions that refer to the same chemical object, including synonyms and typographical variants, and associate all of them to a common unique identifier.

In reality, this is achieved by returning direct pointers of chemical entity mentions to standard identifiers of entries in chemical databases, or alternatively concept identifiers in chemical ontologies/dictionaries. The process of associating entity mentions to concept or database identifiers is generally known as (named) entity linking, grounding, normalization, or resolution.

A related, but slightly different, historical NLP problem to entity grounding is word sense disambiguation (WSD), referring to the computational identification of the correct meaning or semantic role of a word given its context (e.g., document or text).[446] WSD has been studied in the field of machine translation for a very long time[447] through the analysis of the context where the target word occurs, combining word and word sense statistics with knowledge resources. Knowledge resources, such as dictionaries, ontologies, and thesauri, used for WSD can be viewed as data sources or inventories of word senses.

Chemical entity grounding is similar to WSD. Instead of linking ambiguous words found in a text to a specific sense entry in a dictionary, it links chemical names (given their context of mention) to its corresponding chemical database record/concept identifier. Chemical abbreviations and ambiguous trivial/common chemical names need to be disambiguated (resolving ambiguities) to make sure that they actually correspond to chemicals, and to conclude to which specific chemical they refer to.

Aspects such as quality and completeness of dictionaries and knowledgebase are of critical importance for entity grounding. Selecting a chemical mention from text and linking it to a chemical database identifier is especially challenging in the case of chemicals,[428] as chemical databases or dictionaries are incomplete and only contain a subset of all of the chemicals contained in the literature or patents.

The problem of concept or entity grounding, more at a general level, is being intensively studied under what is known as the Wikification task, which aims to automatically detect concept mentions in text and link them to concept references in a knowledge base, such as Wikipedia.[448,449]

4.2.2. Grounding of Other Entities. Named entity recognition and resolution was intensely studied for years in the life sciences and the medical TM field, focusing primarily on gene mention normalization in case of biology[450,451] and on disease-related concept normalization in case of medicine.[452] Providing a detailed description of disease and gene mention normalization goes beyond the purpose of this subsection. In brief, disease mention normalization systems initially applied dictionary-lookup methods as well as rule-based strategies to deal with name variability related to typographical or word order aspects and map detected mentions to medical thesauri like UMLS. Recent trends in disease name normalization point toward the use of learning-based algorithms to tackle the entity linking problem.[453] This resulted in the release of systems like DNorm, which adapts a pairwise learning to rank algorithm for disease normalization, outperforming classical lexical normalization and dictionary matching techniques.[454]

The gene/protein mention normalization problem was addressed by several community challenges[450,451,455] and differs from the chemical mention normalization problem in that, at least for human and model organisms or other well-studied species, existing databases are rather complete and do contain most of the genes that are described in the literature. Gene mention normalization is nonetheless a difficult task. In addition to issues related to the identification of mentions and the lexical/typographical variability of gene names (e.g., alternative use of case, hyphenation, spaces, and Greek letters), gene symbols (e.g., acronyms) are highly ambiguous at several levels. There is ambiguity of gene names/symbols with other entities or common English words, and there is also intra- and interspecies gene name ambiguity, because different genes from the same organism often share the same symbol, and homologous genes from different species (e.g., mouse and human) do have the same name, but different database identifiers.[456]

Gene mention normalization strategies typically rely on exact or fuzzy matching of gene mentions against database names, by using both regularized gene mentions and dictionary entries (i.e., ignoring case and removing spaces and hyphens). Gene normalization systems, like GNAT, also explore the calculation of similarity between the context of mention and the gene database record for selecting candidate gene database

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

identifiers, together with the detection in the context of potential organism sources to constrain the number of potential database hits.[457]

4.2.3. Grounding of Chemical Entities. Grego et al. addressed the problem of mapping results of a CER system to records of the ChEBI database.[346] They used a collection of 9696 chemical entity mentions that had manually generated mappings to ChEBI identifiers as a Gold Standard data set, consisting of an updated version of the Chapati corpus.[418,419] They evaluated the performance of their CER system in terms of mention detection together with the correct resolution to the corresponding database identifiers. Using an exact matching dictionary-lookup method resulted in a balanced F-measure of 31.93%, while using a SL-based CER system obtained a score of 46.95%. When decoupling entity mention detection from entity grounding, that is, by only evaluating the predicted database mappings of those entities that were correct at the mention level, the results were slightly better, obtaining for the dictionary-based method an F-measure of 38.83% and for the SL CER an F-measure of 57.23%.

They examined some of the frequent errors and discovered that, in the case of the dictionary-lookup method, the automated mapping errors were due to chemical normalizations that had the detected chemical mention as part of their names (partial overlaps). In the case of the SL-based CER results, frequent mismatches were encountered for short chemical terms, like "CN" or "OH" that in the ChEBI database corresponded to "cyano group" (CHEBI: 48819) and "hydroxyl group" (CHEBI:43176), respectively, and which did not have any synonyms similar to the original name found in text.

In the case of the tmChem CER system, it normalizes chemical mentions to two different resources of chemical concept identifiers: one is ChEBI and the other is MeSH.[161] To link automatically recognized names in text to a chemical lexicon generated from ChEBI and MeSH, it converts all names to lowercase, and removes all whitespace and punctuation marks. It then compares the chemical mentions to the lexical entries, and in case it detects a match, the system assigns to the mentions the corresponding concept identifier from either ChEBI or MeSH. In case a mention can be matched to both a ChEBI and a MeSH record, by default, the MeSH identifier is used. In case chemical abbreviation mentions are detected together with its corresponding long form, both mentions are linked to the same chemical identifier.

The online entity recognition tool Whatizit uses a dictionary-lookup pipeline to detect chemical mentions and links directly the recognized names derived from the ChEBI gazetteer to the corresponding chemical identifiers from this database.[348]

A hybrid CER system combining SL-learning and dictionary lookup detects chemical mentions and links the resulting names through two different strategies to chemical database identifiers and structures is ChemSpot.[162] Database identifiers are associated to chemical mentions by matching these mentions to records in the chemical lexicon Jochem[336] (section 3.4), which, in turn, assigns various chemical database identifiers to a given chemical name. Additionally, ChemSpot also integrates the chemical name to structure software OPSIN to associate structural information to chemical mentions (section 4.1).[441] ChemSpot resolves chemical mentions, in addition to ChemIDplus, to several other chemical databases and resources, that is, ChEBI, CAS registry numbers, PubChem compound, PubChem substance, InChI, DrugBank, Human Metabolome Database, KEGG compound, KEGG drug, and MeSH.[336]

TaggerOne represents a recent attempt to couple both the entity recognition and the normalization step for diseases and chemical entities, using learning-based methods and a statistical normalization scoring function to generate normalized entity mentions.[458]

4.3. Optical Compound Recognition

Structural information in patents and scientific articles is often represented by figures and drawings of chemical structures. Optical compound recognition or optical structure recognition (OSR) or optical chemical structure recognition (OCSR) addresses the extraction of structural information from digital raster images.

OSR approaches entail three main steps: image processing into text and graphical regions, analysis of graphical regions and reconstruction of the connection tables, and postprocessing of molecular structures. The algorithms supporting these steps vary among software, in an effort to improve interpretation skills and minimize error, but the principles remain similar. Table 6 tabulates both commercial and free/academic OSR software. Some of the current solutions are based on the combination of consolidate software (e.g., OSRA[459] and CLiDE[124,125]), while original solutions choose to apply automated, advanced learning methods, rule-based logic, artificial intelligence, or supervised learning methods.

Most OSR software is equipped to read contents in a range of common graphical formats, such as .gif, .jpeg, .png, .tiff, .pdf, and .ps. Current software also tries to support formats of well-established chemical drawing/editing software (Figure 11 tabulates meaningful representatives of chemical drawing/editing software). While structure diagrams are typically drawn with black ink on a white background, these diagrams may also contain colors (e.g., for indicating atom types) or the background can be different from white. Hence, the images are usually converted to grayscale and binarized, that is, turn the color-scaled image into a bilevel image by classifying every pixel as an on- or off-pixel (the threshold may be set fixed or adaptive using a data mining algorithm).[460] Next, image segmentation is performed as means to recognize text and graphical regions. The image is scanned on a row basis, and sets of adjacent horizontal on-pixel segments are identified as connected components. Typically, segmentation algorithms are based on a set of criteria, the ratio of black pixels to the total area of the component or the aspect ratio. Further, anisotropic smoothing and thinning processes may be implemented to remove noise, that is, remove small variations in pixel intensities, while preserving global image features, and normalize all detected lines (pixel wide).

The recognition of bonds and atoms may rely on various processing algorithms, which ultimately aim to detect combinations of lines and letters. Vectorization is performed to convert bitmap to vector graphics and be able to detect the positions of atoms and bonds. The Potrace library[461] is the most used software for such processing. Basically, the vectorized form is examined to identify the control points of the Bezier curve, that is, the parametric curve usually used to represent smooth curves. These control points are flagged as atoms, and the connecting vectors are identified as the corresponding bonds. The following processing steps address specific "features" of the recognized atoms and bonds: the recognition of atomic labels and charges (e.g., by considering

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Table 6. Basic Description of Existing OSR Software[a]

software	approach	input	outputs	language	operating system	availability	URL
AsteriX web server[b,462]	reconstruction of 3D ligand coordinates from 2D images	PDF		Java applet and Java web start	n.a.	free for academic use	http://swift.cmbi.ru.nl/bitmapb/
ChemEx[c,463]	extracts compound, organism, and assay information; uses OSRA	reads full-text papers, and recovers SMILES and MOL from structure images	visualization via graphical user interface and exportation to XML	Java and C++	Windows and Linux	free	http://www.biotec.or.th/isl/ChemEx
ChemInfty[464]	mathematical OCR (includes two engines), supporting Markush recognition	image formats	SDF, MOL	Java	Windows, Linux, Mac OS X	commercial	http://www.inftyproject.org/en/ChemInfty
chemOCR (chemical optical character recognition)[d]	expert rules and supervised learning methods	BMP, GIF, PNG, and multipage TIF	SMILES, SDF	Java	Windows and Linux	commercial	http://www.scai.fraunhofer.de/en/business-research-areas/bioinformatics/products/chemocr.html
ChemReader[e]	machine vision approach	image formats	ML, SMILES	C++	Windows	commercial	http://www-personal.umich.edu/kazu/research-areas.html
CLiDE (chemical literature data extraction)[124,125]	image processing and artificial intelligence	documents of the following types: PDF, DOC(X), and HTML; images files: BMP, GIF, JPEG, JPG, JPE, JIF, PBM, PGM, PNG, PNM, PPM, TIFF, TIF, XBM, and XPM	various export options (e.g., can be directly transferred into chemical editors), depends on the version	C++	Windows	commercial	http://www.keymodule.co.uk/products/clide/index.html
D2S	supports CLiDE, OSRA, and Imago	documents in PDF, TXT, HTML, XML, and MS Office formats (e.g., DOC, DOCX, PPT, PPTX, XLS, XLSX), OpenOffice ODT, embedded structure objects (e.g., ChemDraw, SymyxDraw, MarvinSketch), and images in TIFF and BMP formats	MRV (Marvin documents), ML, SMILES, MOL	Java	Windows, Unix/Linux, and server installation	several commercial and free licenses	https://www.chemaxon.com/products/document-to-structure/
IBM OROCS (optical recognition of chemical graphics)[f]	image processing, ability to recognize images containing structure diagrams in documents	image formats	MOL	C	IBM OS/2	commercial	not available
IMAGO OCR	image processing and lexicon-based abbreviation expansion	PNG, JPEG, BMP, DIB, TIFF, PBM, RAS	MOL	C++ (and includes C interface and Java wrapper)	Windows, Linux, Mac OS X	free (GPL-licensed but possible to purchase a commercial license)	http://lifescience.opensource.epam.com/imago/
Kekulé[g]	image processing and rule-based logic, with manual marking of structure diagram regions	ISIS, MOLfile, ROSDAL, and Kekulé's native format	ISIS, MOL, SMILES, ROSDAL, and Kekulé's native format	C++	Windows	no longer commercially available	http://aig.cs.man.ac.uk/research/kekule/
MLOCSR[h,465]	combines a low-level processor with Markov logic (to reason about the low-level entities and relations); images with tables and/or reactions are not supported	image file (jpg, png, jpeg, and TIF supported)	MOL	n.a.	online, but standalone under development	system is available as a web server at http://mlocsr.dinfo.unifi.it; standalone distribution is currently under development	http://mlocsr.dinfo.unifi.it/
OSRA[459]	image processing, character/string recognition, various connection table compilation, and confidence estimation	image file (gif, jpg, png, jpeg, tif, pdf, and ps supported)	SMILES, SD files	C++	Windows and linux	open source	https://cactus.nci.nih.gov/osra/

[a]Prepared in November 2016. [b]Lounnas, V.; Vriend, G. AsteriX: A Web Server to Automatically Extract Ligand Coordinates from Figures in PDF Articles. *J. Chem. Inf. Model.* **2012**, *52*, 568−576. [c]Tharatipyakul, A.; Numnark, S.; Wichadakul, D.; Ingsriswang, S. ChemEx: Information Extraction System for Chemical Data Curation. *BMC Bioinf.* **2012**, *13*, S9. [d]Kral, P. Chemical Structure Recognition

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Table 6. continued

via an Expert System Guided Graph Exploration. Master's Thesis, Ludwig-Maximilians-Universität, 2007. [e]Park, J.; Rosania, G. R.; Shedden, K. A.; Nguyen, M.; Lyu, N.; Saitou, K. Automated Extraction of Chemical Structure Information from Digital Raster Images. *Chem. Cent. J.* **2009**, 3, 4. [f]Casey, R.; Boyer, S.; Healey, P.; Miller, A.; Oudot, B.; Zilles, K. Optical Recognition of Chemical Graphics. *Proceedings of the second International Conference on Document Analysis and Recognition (ICDAR '93)*; IEEE Computer Society Press: Tsukuba City, Japan, October 20–22, 1993; pp 627–631. [g]McDaniel, J. R.; Balmuth, J. R. Kekule: OCR-Optical Chemical (Structure) Recognition. *J. Chem. Inf. Model.* **1992**, 32, 373–378. [h]Frasconi, P.; Gabbrielli, F.; Lippi, M.; Marinai, S. Markov Logic Networks for Optical Chemical Structure Recognition. *J. Chem. Inf. Model.* **2014**, 54, 2380–2390.

the maximum character height and width, the existence of two characters aligned horizontally or vertically, and the presence of the character "−" or "+"), the identification of circle bonds (e.g., the identification of a circle inside a ring may be indicative of an aromatic ring), the evaluation of average bond length and distance within double and triple bond pairs, the identification of dashed and wedge bonds (typically using bond length and positioning criteria), and the disambiguation of bridge bonds.

Connection tables or graphs are constructed on the basis of the previous compiled information, the connections and stereo- and aromaticity flags. Table information and character string data, that is, single atom symbols, atom symbols with charge and mass, or group formulas, then are processed to determine the chemical meaning of the strings. Atom symbols are commonly verified against a lookup list, and strings denoting group or moiety formulas are processed to interpret them into graph format.

At the end, the connection table of the entire chemical structure can be edited to adjust for scanning errors and cleaning up bond angles to standard values. The output formats of choice of most OSR software for the molecular objects are SMILES and Molfiles (also referred to as MOL).

While available systems are already able to handle a good amount of chemical diagram features, there remain several directions for further improvement. These include some graphical ambiguities due to touching and broken characters, or characters touching lines; large macromolecular structures and complicated rings; Markush features, such as substituent replacement in R-groups, link nodes, or repeating units; and recognition of chemical tables or reactions. In part, OCR errors could be minimized by combining the commonly used dictionaries of words (e.g., hash-table-based dictionary lookup for common, trade, and scientific names) with more specialized, technical terminological resources. For instance, CaffeineFix finite state machine encodes a significant fraction of the IUPAC naming rules for organic chemistry, including numerous CAS and Beilstein naming and traditional variants.[119] Likewise, the chemical dictionary of the Structure Clipper encompasses over 46 000 records, which were extracted from subterms found in 25 million IUPAC chemical names.[466]

Currently, the ChemInfty[464] method is one of the few that addresses specifically the recognition of Markush structures from images.

4.4. Chemical Representation

Apart from commonly used structure diagrams and chemical nomenclature, there are multiple representation formats for a chemical compound that have been used throughout the history of chemical information systems. Broadly, they can be classified[467,468] as topological graphs, line notations, connection tables, and combinations of the previous ones (e.g., the IUPAC International Chemical Identifier (InChI)), 3D structure representations, fragment codes, fingerprints, and hash codes, such as CACTVS,[469] which is behind PubChem structure search and the National Cancer Institute's Chemical Structure Lookup.

The range of applications of these formats is very wide: registration in chemical databases, structural and substructural searches, virtual screening, structure−activity relationship (SAR analysis), and crystallography. Here, we will focus on those most commonly used in registration systems, and concepts, such as normalization, will be discussed from this point of view, as registration in database systems does not require the selected

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

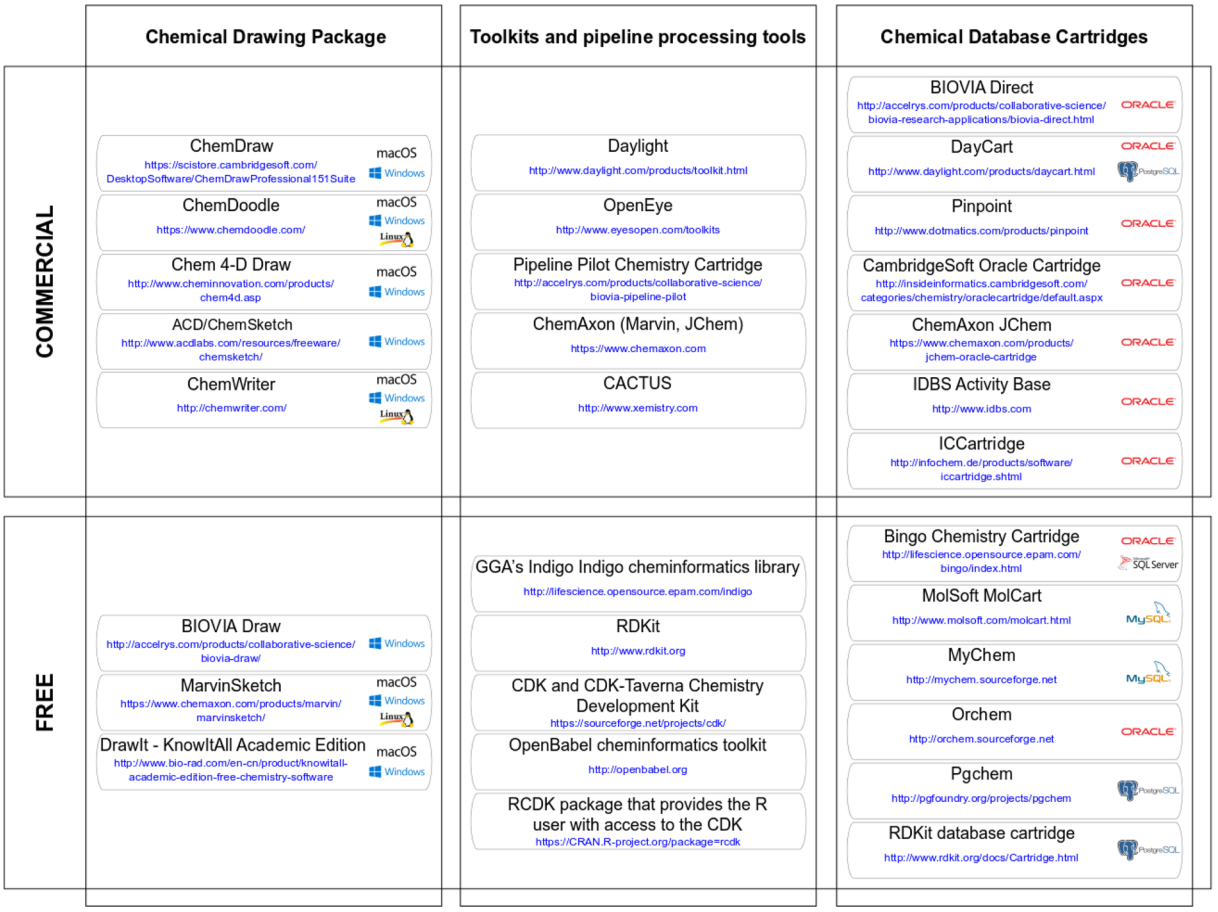


Figure 11. Programs commonly used to generate chemical structure images, toolkits, and pipeline processing tools and chemical database cartridges (prepared in November 2016). Note: Software accessibility is generally described. Many commercial tools have free licenses for academic use, and some of the listed free tools may have restrictions on their use. Moreover, license availability may change with time. Readers are recommended to check license details before use.

representation to be typical of the most physiologically appropriate form.

Two fundamental challenges of structure representations are uniqueness and unambiguity. Uniqueness requires that the structure representation format corresponds to a single chemical entity. For example, InChIKeys have been estimated to have one replicate in 75 billion structures.[23,470] Ambiguity implies that the structure representation format ideally must be unique for a particular chemical entity; that is, there is a unique (or "canonical") encoding of the structure. For example, the SMILES notation contemplates different versions of the same compound (depending on the atom ordering), which enforces the use of the canonical SMILES representation to avoid duplicate redundancies. However, the concept of ambiguity depends on what is considered a unique chemical entity for the purpose of a particular application. For example, to systematically organize compounds in a database, users may opt to differ between different isotopic forms, different tautomers, or remove counterions. All of these issues need to be considered when determining the ambiguity of a particular chemical representation.

In the following, we briefly review the chemical representation formats that are considered most important for chemical and reaction database annotation.

4.4.1. Line Notations. Linear strings of alphanumeric symbols were the first formats developed to surpass systematic nomenclature. In 1965, the Morgan algorithm was introduced to enable unique structure identification when developing the CAS computer system.[471] The Wiswesser line notation (WLN) was also one of the earliest attempts to obtain both unique chemical identifiers and machine interpretation, and supports substructure pattern language.[472] WLN then gave way to the simplified molecular-input line entry system (SMILES),[20,21] a proprietary product of Daylight Chemical Information Systems Inc., which is the predominant line notation nowadays, mostly as it is human readable. Closely related formats, SMARTS (SMILES Arbitrary Target Specification)[473] and SMIRKS (SMIles ReaKtion Specification),[474] are used to match patterns and reaction transformations, respectively. The SYBYL line notation (SLN) was inspired by the SMILES notation with several extensions to enable the specification of full substructure queries, reactions, and some types of Markush structures.[475,476] Finally, ROSDAL (representation of org. structure description arranged linearly) was developed for Dialog patent search platform.[477] Focusing on the highly extended SMILES, the main drawbacks are the existence of different versions of the same compound (i.e., require canonicalization, although to further complicate things, different vendors have their own canonicalization algorithm), the lack of full representation of stereochemistry (e.g., supports absolute stereochemistry but flags such as unknown and relative

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

stereochemistry are lost), and tautomeric dependency (different tautomers have different SMILES).

4.4.2. Connection Tables. These consist of tables of atom-based and bond-based records with columns encoding atom properties (i.e., atomic element and charge) and bond orders and stereochemistry. Among the many variants, the Molfile[478] format, developed by MDL, is the most predominant for exchange of structure representations, especially in the form of structure data file (SDF) for storing multiple molecules and data. The molfile V3000 format, first introduced in the mid-1990s by MDL to overcome issues of molfile V2000 with large structures,[479] currently handles the combination of multiple stereogenic centers into groups of different types (known absolute, known relative, or unknown stereochemistry). As for SMILES, molfile formats (either V2000 or V3000) do not support tautomerism, and interconverting tautomeric forms are not identified as the same structure for these formats. RXNfile (single reaction) and RDfiles (multiple reactions) are variants of the Molfile that contain structural data for the reactants and products of a reaction.[478] These formats are the most popular for data set exchange. Chemical Markup Language (CML) is an XML-based connection table format, proposed in the late 1990s, and, despite its extensive literature and documentation, has not been extensively adopted, especially by commercial vendors.[480] This format has its relevance in the area of semantic web technologies.

4.4.3. InChI and InChIKey. To overcome issues associated with SMILES, the IUPAC International Chemical Identifier (InChI)[22,481] was introduced, and then the InChIKey (a fixed-length hash code representation of the InChI) to improve retrieval of InChI strings by Internet search engines (due to their long number of characters and punctuation symbols). While the InChI to InChIKey hash compression is irreversible, there are a number of InChI resolvers available to look up an InChI giving an InChIKey. InChI has a layered structure (layers of information on connectivity, tautomeric, isotopic, stereochemical, and electronic), which allows one to represent molecular structure with a desired level of detail depending on multitude of options. Because of interoperability concerns, in 2008, the Standard inChI was launched with the aim to always maintain the same level of attention to structure details and the same conventions for drawing perception. The generation of InChIs involves the normalization of the original structure to remove redundant information (e.g., disconnecting metals and protonation), and its canonicalization and serialization. Polymeric molecules are not handled by InChI identifiers, and support for Markush structures and organometallics is incomplete. See the article collection dedicated by the *Journal of Cheminformatics* to the InChi Keys.[482] In contrast to SMILES, InChi code supports mobile hydrogen perception and is able to recognize tautomeric forms. However, it does not distinguish between undefined and explicitly marked unknown sp^3 stereo. A comprehensive review of the many uses of InChi and InChiKeys can be found in ref 483. Reaction InChI (RInChI) is currently being developed.[484]

4.5. Chemical Normalization or Standardization

Normalization or standardization is the process to generate and select a unique accurate representation among all possible variations in which equivalent structures (unique chemical entities) can be represented. This process is key to avoid duplicate structures in chemical registration databases (e.g., to conveniently track different synthetic batches of the same compound) as well as to avoid errors in accurate structure representation (i.e., wrong or misleading structures) that later might translate into computational models[485] and ensure optimal performance of chemical search engines. Moreover, structural normalization is necessary for optimal integration of different data sources, less error-prone, and for mapping recognized chemical names to structural databases. Figure 12

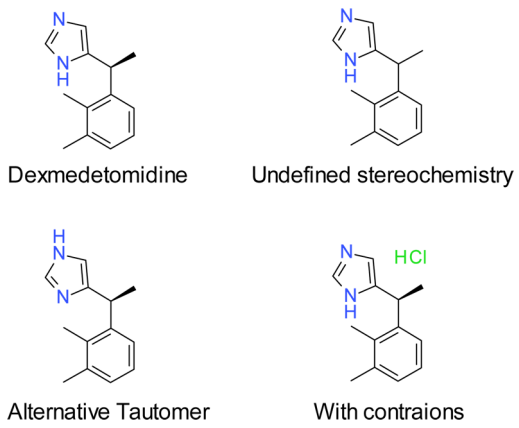


Figure 12. Alternative representations of the chemical structure of the drug dexmedetomidine.

shows four different representations of the chemical structure of the drug dexmedetomidine, with variations in the definition of stereochemistry, tautomeric state, and presence of contraions.

Because of its nature, there is not a single normalization procedure that is able to achieve this goal, although several guidelines, internal workflows, and business rules have been proposed, including those by the FDA and accepted by the Open PHACTS project,[486] the FICTS rules,[487] which by default discard stereochemical marks, and others.[429,488,489] More recently, the freely available Internet-based CVSP platform has been released to assist in this task, with a focus on the registration in ChemSpider.[490] A similar service is offered by PubChem.[491]

A first step when normalizing chemical structures involves structure validation for potential mistakes in atom validity (atomic elements), connectivity, atom valences (e.g., hypervalency), wrong stereochemical assignation (e.g., stereochemical bonds assigned to nonchiral centers), and aromaticity detection (structure aromaticity is detected and validated to be kekulizable). Particularly, aromatic representation is highly dependent on the used software, and users should be aware of its specific requirements. A particular mesomeric representation, alternate forms of functional groups with molecular charge delocalization (such as nitro, azides, or *N*-oxides), is commonly also fixed.[492] For registration in a database system, charges are commonly normalized to a particular ionization state; typically neutral forms for acids and bases are preferred, without zwitterions.[492] As mentioned previously, this normalization procedure might differ when applied to a chemical database for other cheminformatic purposes, such as virtual screening or QSAR modeling, where compounds are typically protonated according to their predicted pK_a at physiological pH.[485,493]

As insinuated in the chemical format sections, the main causes of different chemical representations are the problem of different tautomeric and stereochemical representations (Figure 12). Again, for chemical registration purposes and, in contrast

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

to virtual screening applications,[494] a canonical tautomeric form should be inserted. An alternative approach to surpass the problem with tautomeric-dependent formats (i.e., SMILES and molfile) is generating a standardized canonical tautomer before registration in chemical databases,[495,496] for which many commercial and open source programs and toolkits are available: a comprehensive compilation is provided in ref 496. See the article collection dedicated by the *Journal of Computer-Aided Molecular Design* to the handling of tautomers in cheminformatics.[497] A popular strategy of in-house pharmaceutical systems is to separately track the parent molecule, that is, the molecule responsible for the biological activity of the compound, from the companion salts and solvates, which are chemically stored in a separate table/dictionary and linked through a unique code defining the salt record.[428,498] Together with previous considerations, standardization workflows commonly contemplate chemical drawing rules that disallow some representations: shortcuts and abbreviations such as "Ph", "Bz", and "BOC", certain bonds such as dative bonds and covalent bonds with salts, and certain carbohydrates drawings.[499] For example, the "either" bond (wavy line) used to mark nondefined double bonds is an enhanced feature that, because of tradition, is rarely used by chemists when drawing, and that can be the cause of mismatches in chemical representations if users do not use it to distinguish between the common double bond. Finally, it is also advisable to assign or "clean" the 2D coordinates to have an appropriate layout that ensures that chemical structures are visually interpretative.[500] From the point of view of implementation, SMARTS and SMIRKS enable desired transformations and can be implemented using a variety of cheminformatic toolkits (Figure 11), although other chemical scripting languages (Cheshire,[501] Standardizer,[502] and sdwash,[503] or workflow tools, such as Pipeline Pilot[504] and Knime)[505] are also frequently used.

5. CHEMICAL KNOWLEDGEBASES

Large collections of chemical and biological information have highlighted the need for supporting infrastructures. Since the first computer database and structure retrieval system back in 1957[506] and the release of the CAS database in 1966[507] and chemical and reaction indexing,[508] extensive work in the field of cheminformatics has been focused on the indexing of structures and reactions and the development of efficient algorithms for (sub)structure searching; see ref 509 for a historical perspective. It is important to note that the need to precisely manage and retrieve chemical and biological information is previous and independent, but complementary, of the application of TM techniques in the field of chemistry.

5.1. Management of Chemical Data

Relational database technologies, together with the associated database management systems (DBMS), are valued for their integrity, scalability, and audit trail capabilities. Traditionally, they are built using either a data federation or data warehousing strategies. Data federation technologies (data virtualization) integrate multiple autonomous disparate databases and aggregated in a single conceptual unit. Data warehouses are central repositories from several source systems, which are extracted, transformed, and loaded into the new repository and can be queried from a single schema. Thus, different companies have implemented their own drug discovery informatics platforms on the basis of DBMS: ABCD-Johnson & Johnson (J&J),[510] CCBR-CNIO,[498] ArQiologist-ArQule,[511] Avalon-Novartis,[512] Osiris-Actelion,[513] Chemistry Connect - AstraZeneca,[514] and UCSD-Philip Morris International.[4,428] Interestingly, the Standardised Data Warehouses project of the Pistoia Alliance members aims at proposing a standard harmonized model for the discovery-stage data warehouses, with the goal of increasing data quality and interoperability, while allowing for some local variability.[515]

A drawback of relational databases is that they are comparatively inflexible to changes in the nature of data recorded and require dedicated maintenance. In recent years, the Semantic Web[516] has arisen as an intermediate compromise solution between the old-used uncontrolled data files (e.g., Excel and HTML) and DBMS. Knowledge-based databases and semantics are useful in interpreting the data and derive knowledge.[517,518] Semantic Web solutions require raw data files, a codebook that dictates how the data are entered, and the descriptive metadata to ensure data integrity and curation. The Semantic Web metadata standard is the Resource Description Framework (RDF), a vocabulary for constructing relationships based on triples. An RDF triple consists of three URIs (uniform resource identifier) describing how a subject (one entity, for example a molecular property) relates to an object (one entity, for example a molecular structure) via a predicate attribute (which is a URI with a commonly agreed upon meaning). Thus, RDF schemas express ontologies, as they provide formal explicit descriptions of concepts in a certain domain. Much of the work in this area is based on the seminal work by Murray-Rust and co-workers while developing the Chemical Markup Language (CML),[480] with approaches such as the SemanticEye project[519] initiating the inclusion of metadata within scientific electronic publications (with the chemistry annotated as InChI's). Also, in 2006, Taylor et al. reported the application of semantic web technologies to the storage and access of molecular structures and properties using a hash code directly derived from the InChI.[520] In this line, the Chemistry's Project Prospect of the Royal Society of Chemistry[521] using OSCAR[158] to extract chemical entities applies semantic mark-up to highlight and annotate chemical structures and correlate them with relevant ontological information and additional property data. This enables the use of semantic queries to search for chemical structures present in the papers. Other semantic web projects are the oreChem project,[522] CrystalEye,[523] the ChemXSeer,[183] and the Open PHACTS consortium,[524] and ChemSpider,[525] ChEMBL,[526] and, more recently, PubChem.[527] The amount and diversity of drug discovery data in the -omics and high-throughput-driven paradigms have significantly grown to the point where current relational data models are reaching their performance limits, in terms of both technical and scientific capabilities.[528]

5.2. Chemical Cartridges

Here, we will focus on the implementation issues of traditional relational databases as currently the primary way of substructure search and will not discuss semantic and ontological languages (e.g., XML, OWL, RSS, and RDF) and semantic query languages (e.g., SPARQL) that have been surveyed in a thematic series.[529] The majority of them integrate commercial software from different vendors (e.g., BIOVIA, formerly Accelrys, ChemAxon, OpenEye, Advanced Chemistry Development, CambridgeSoft, Chemical Computing Group, Daylight, ID Business Solutions, and SpotFire) or open access components with some custom developments, especially chemistry editors and chemical cartridges plug-ins to give

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

chemical handling functionalities to the database. There are many available cheminformatics indexing technologies (chemical cartridges available to the public) using different underlying database technologies (Figure 11). Also used are in-house developed cartridges for either search purposes such as the ABCD Chemical Cartridge[530] or for both search and registration as OSIRIS.[513] An important consideration is the wide range of different chemistries that these cartridges must support: small organic molecules, peptides,[531] sugars, polymers, mixtures and formulations (e.g., different ratios of enantiomers), Markush formula, and antibody−drug conjugates. Apart from chemical registration, the use of chemical cartridges for structural searches has the advantage of performing faster than classical cheminformatic tools as they are indexed. With the exception of open source cartridges,[532] little has been published on the inner workings of search engines, and users must refer to the user and development manuals to fully understand their performance. Intermediate solutions that abstract the storing and searching of chemical structures into method calls have been also published[533] or are commercially available (e.g., Compound Registration by ChemAxon[534]).

Besides the normalization of chemical structures, the format of the input data is also key to determine the final annotation. InChI strings can generate different molecules from that was used for InChi generation as they are not intended for backward structure generation. The molfile format is the preferred primary source in many chemical registration modules.[428,489] Remarkably, despite the enhancement of the molfile V3000 format that maintains all stereochemical flags, the V2000 format is still a preferred option for exchange between web-accessible databases. Moreover, the SDF format includes additional fields to be inserted.

In synthetic repositories, the concept of chemical batch is also a clue consideration that determines the decision to prohibit the registration of duplicate chemical structures. However, the detection of duplicate structures depends on the user-selected configuration of the chemical cartridge, regarding the perception of tautomers, isotopes, stereoisomers, and salts as equivalent structures. The most restrictive definition should be selected to avoid the registration of duplicate structures (e.g., different tautomeric forms are detected and assigned to a canonical tautomer representation).

In reaction databases, a database field contains the diagram of a single-step reaction, which consists of one or more reactant molecules and one or more product molecules. Other fields specify additional information, such as the combination of single-step reactions into a multistep reaction, solvents, and catalysts. The first efficient and effective method for the detection of reaction sites was reported by Willett, and it is based on maximum common subgraph isomorphism.[535] This work was key to the development of reaction databases, with the first operational systems (i.e., REACCS, SYNLIB, and ORAC)[536] appearing in the 1980s, and being posteriorly improved and implemented as Oracle cartridges, such as the MDL's reaction database management.[537] To increase the precision and performance of reaction searches, chemical cartridges recommend reaction mapping. Reaction mapping determines the correspondence between the atoms and bonds in the reactants and the atoms and bonds in products: atom−atom map numbers specify the exact correspondence between atoms in reactants and products, and reaction center marks on the bonds define what happens to the bond in the reaction. Chemical editors offer possibilities to automatically or manually map reactions.[538] This technology is behind electronic laboratory notebooks (ELN) for synthetic chemistry.[539]

5.3. Structure-Based Chemical Searches

As compared to chemical text name searching, chemical structure search engines have the advantage of avoiding the problem of synonyms and the disadvantage of disregarding the context of the chemical structure in the text (unless it is dedicatedly annotated, as in the case of SciFinder with the Biological Role, and then the search is complemented with a keyword search). In any case, although the chemical search engines support the simultaneous search of chemistry and keyword search (e.g., "Sildenafil AND PDE5"), the text search does not guarantee the relationship between the drug and the target. On the other hand, text searches are far beyond identifying documents for chemicals that are structurally similar or superstructures for a given chemical of interest. Searching by molecular formula (text-based) has the associated problem of retrieving many hits, especially when searching for fairly common organic substance, as a single molecular formula represents the composition of a variety of substances.

As mentioned, InChI and InChIKeys have been used as a means of indexing the chemical structures mentioned in scientific literature, as its potential effectiveness for chemical IR was early investigated in Google in 2004[540] and the eCrystals/eBank project.[541] However, a recent examination[542] discusses the consistency of Google results and provenance of retrieved links as compared to rigorously maintained major databases. Moreover, InChiKey alone only enables a search engine to identify the exact query structure, not substructure or similarity searches (a simplified similarity option can be run by the insertion of wildcard characters). Alternatively, RSS has recently been considered for chemical structure searching.[543] However, as noted by Frey,[544] structure search has been notably slow to adopt Semantic Web technology.

Together with text-based searches (by chemical name, formula, or database identifiers like CAS Registry Numbers, Derwent DRN and Beilstein BRN, SMILES, and InChi), most chemistry resources include structure searching capabilities as one of the best options of finding a specific substance in the literature. Common structure searches include exact-matching, substructural, and similarity searches. These searches rely on graph theoretical algorithms, where atoms correspond to nodes and bonds to edges joining the nodes. Users draw a substance in the chemistry editor, which is internally transformed into a connection table and normalized and compared for identity (exact matching) with the indexed structures. As for the registration of chemical structures, chemical cartridges offer different configurable options for defining what an exact match is in terms of stereochemistry, tautomers, isotopes, and salts. This can translate into an exact-match search retrieving no hits, even though the structure is present in the database. In these cases, it is advisable to run a similarity search with a high cutoff.[545]

In substructure searches, the connection table of the query structure must be a subset of that of the database substance to be matched. The challenge of these searches is the substructure definition to optimally balance the number of desired retrieve hits and undesired false hits. This implies knowing the flexibility of the search to precisely define bond order, single, double, triple, or dashed bonds; double-bond geometry, stereochemistry, open or specifically defined with wedge and hash bonds; whether the topology of each connection can be fixed as

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

acyclic (chain) or cyclic structure (or contemplates both options); locked ring fusions; the type of substitution permitted at each position, locked atoms, generic groups, or variables (e.g., metals, halogens, alkyl chains, cycles, carbocycles, and atom to hydrogen- or carbon-only), user-defined atom lists, nonatom lists, and R-groups; and variable point of attachment and multiple fragment searches.[221,546,547] Apart from small organic compounds, metal-containing species are especially problematic to search, mostly due to inconsistencies and different drawings within databases[548] as well as for polymers[549] and polymorphic structures.[550] In general, tools for patent priority art searching (e.g., SciFinder, STN, and Reaxys) support more generalizable group definitions to narrow the search, while web-based tools connected to chemical databases have structure editors with less advanced features.[551] Although graphics-based queries are the preferred option for chemistry information retrieval, substructure query languages (text-based) also deserve a mention (e.g., SMARTS, SLN, or Molecular Query Language, MQL,[552] as well as approaches to directly build standard SQL searches[553]).

Chemical cartridges implement similarity-based searches, often termed "fuzzy" matching, that retrieve compounds having a user-defined percentage similarity to the query structure. Two factors determine the search and vary among software packages: the descriptors that characterize the compounds (for database querying, commonly atom, bond, and fragment-based counts expressed as keys or fingerprints) and the metric (commonly Tanimoto), so users should be aware of the characteristics of the underlying approach when analyzing the retrieved hits.[554,555] Recent advances in this field incorporate the use of inverted indexes (as commonly used in text search methods) for similarity-based searches of chemical compounds.[556] Most similarity searching approaches commonly used in virtual screening applications and drug discovery efforts (e.g., 3D) are not implemented with the purpose of searching chemical repositories.

In reaction searches, graphical-based reaction searches can be built by specifying either the reactant or the product (partial) or by defining both reacting species (complete). In general, drawing arrows indicates the role of each molecular species, although specialized reaction searching resources enable the selection of a particular role. Reaction mapping (atom and/or bonds) is very useful to search for specific bonds that are formed or broken during the reaction or that do not change, and helps to reduce the false positive rate and increase precision. Alternatively, tools are available to indicate the reacting bond(s), define nonreacting functional groups, and limit to stereospecific reactions (inverted, retained stereochemistry). See chapter 9 in ref 557 for an excellent comparison between CASREACT[558] searching via SciFinder and Reaxys. Reaction searching engines, such as SciFinder and Reaxys, enable the combination of structure and reaction queries to further refine the returned hits[559] and the capability to combine reaction steps to build and plan the most effective procedure, SciPlanner and Reaxys synthesis planner, respectively. Defining queries in an appropriate way is a clue to reaction searching, as it might translate into long execution times (e.g., multifragment reactants and products should be avoided).

In Markush searches, Markush structures are indexed in fragmentation code systems and topological search systems.[560] The former, available since the 1960s, use structural features (e.g., functional groups) implemented as closed dictionary lists or extractable using fixed rules (examples include the Derwent CPI Fragmentation Codes,[561] IFI Claims Codes,[76] and GREMACS[562]). In the 1980s, topological systems (MARPAT[563,564] and Merged Markush Service − MMS[565]) arose and superseded fragment codes as they fully capture the structural relationships of the patterns. Since then, there have been minor developments.[566] Recent efforts include the development of a search engine that provides a query interface that unifies both structure and keyword conditions include the work by NOVARTIS. The first claimed of these tools is the Canonical Keyword Indexing (ECKI),[188] which converts a chemical entity embedded in a data source into its canonical keyword representation prior to being indexed by text search engines. To perform chemical normalization, all chemical synonyms are aliases of a single entity and their unique canonical keyword representation.

The number of available chemical repositories is vast, and each of them can be used for many different purposes, such as prior art searching engines (e.g., CAS Registry,[567] CASREACT[558]), dictionaries in TM applications (e.g., ChemSpider[568,569]), entity grounding (e.g., PubChem[570,571]), as sources of annotated biological information on different domains (hereafter defined as knowledgebases, such as PubChem, DrugBank,[572,573] ChemIDplus,[220,574] ChEMBL,[575,576] and ChemBank[577,578]), as sources of checking commercial availability and vendors (eMolecules[579] and Zinc[580,581]), and extracted by applying TM techniques (SureChEMBL,[74,75] IBM Watson Patents[582]). Here, we briefly describe those that are relevant for TM purposes (e.g., chemical entity grounding) or that have been derived by TM (only a minority).

PubChem,[570,571] launched in 2004 by the U.S. National Library of Medicine (NLM), contains small molecules with information on their biological activities (BioAssay). Substances (>224 million chemical samples) and unique structures (>92 millions) are tracked separately. It accepts data from multiple repositories, without manual curation, what has raised some critics about quality issues.[583]

ChemSpider,[568,569] by the Royal Society of Chemistry (RSC, 2008), includes 57 million structures with associated chemical information, most of which have been robotically or manually curated and will be by the CVSP platform.[490] The focus on validated chemical name−structure relations has produced a qualified dictionary for TM applications.[584]

CAS Registry,[567] maintained by Chemical Abstracts Service (CAS), is the most authoritative collection of disclosed chemical substance information (over 123 million organic and inorganic substances and 66 million sequences). It is derived manually, and each unique substance is assigned a unique CAS Registry Number and CA Index names (CAS-style systematic nomenclature). It is accessed by SciFinder[44] and STN.[84]

Index Chemicus[585] has over 2.6 million compounds (dated to 1993) covering more than 100 of the world's leading organic journals. Each record contains full graphical summaries (indexed bioactivity), reactions, and complete bibliographic information.

Derwent Chemistry Resource (DCR)[586] is a chemical structure database for searching specific compounds indexed in Derwent World Patents Index bibliographic records. It is accessed by many search services listed in Table 2 (STN, Questel, Dialog, Thomson Innovation).

Beilstein (from 1771)[29] and Gmelin Handbook (from 1700s) databases[587] containing compounds extracted from

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

journals are currently accessible through Reaxys.[82] These "books" extracted chemical and physical property data that were arranged by compound class.

SureChEMBL is a chemical repository of over 17 million unique compounds extracted from patents (US, EP, WO full texts, and JP abstracts), and over 14 million annotated patents as of November 2015[75] using TM techniques (text annotations from 1976 to date and images from 2007 to date). Text-based searches are available to query for patent documents from nonannotated patent authorities. It takes 2−7 days for a published patent to be chemically annotated and searchable.

IBM Watson patent database//IBM BAO strategic IP insight platform (SIIP)[582] is a searchable database of over 2.5 million chemical structures and pharmaceutical data extracted from the patents and scientific literature using SIIP. It was donated to PubChem and the NIH CADD Groups.[588]

Drug Central[589,590] is an open access online drug compendium, containing a total of 4444 Active Pharmaceutical Ingredients (APIs) linked to a list of over 20 617 drug synonyms and research codes. For each API, drug mechanism of action (MoA) target annotations, pharmacological action, bioactivity profiles, drug indications, labels, pharmaceutical formulation, dose, formulation, administration, regulatory approval information, and marketing status are provided. A dedicated effort was put on the annotation of chemical structures (MDL format), especially involving a multistep manual curation process with a hierarchical checking of available public resources (WHO INN, USAN, FDA SRS, CAS, and FDA drug labels) and with an emphasis on stereochemistry annotation.

ChEMBL[576] launched by the EMBL-European Bioinformatics Institute in 2009 is a public, downloadable database of bioactive drug-like small molecules (over 2.0 million records and 1.6 million unique compounds), with calculated properties and abstracted bioactivities from primary literature (over 47 journals) on a regular basis (with releases every 3−4 months), then curated and standardized. It also includes FDA-approved drugs.

DrugBank,[410] publicly available from the University of Alberta since 2006, is a popular knowledgebase that provides a detailed description about the chemical and pharmacological characteristics of over 8200 experimental and approved drugs together with drug target information.

eMolecules[579] and BIOVIA Available Chemicals Directory (ACD)[591] merge vendor's catalogues and are typically consulted to check the commercial accessibility of screening products and building blocks. Zinc, while also being very popular for this purpose, was mainly designed for virtual screening applications. Chapman & Hall/CRC maintains CHEMnetBaSE[592] segregated directories of different types of compounds as well as the Combined Chemical Dictionary (CCD) with access to chemical, physical, and structural data on more than 630 000 compounds.

Together with compound databases, reactions databases are included: CASREACT,[558] Current Chemical Reactions (CCR),[593] SPRESI[594] and its derived ChemReact[595] with unique reaction types, Science of Synthesis,[596] and ChemInform Reaction Library[597] and Selected Organic Reactions Database (SORD).[598] Other more specialized resources containing reaction information include eROS[599] for reagents and catalysts searching, and Comprehensive Heterocyclic Chemistry (CHC)[600] and Synthetic Reaction Updates,[601] a literature updating service with recent developments in synthetic organic chemistry and others reviewed by Zass.[602] CASREACT and the reactions accessed by the search platform Reaxys are the most prominent repositories of chemical reactions, with over 78.4 million single- and multistep reactions and over 42 million reactions, respectively.

For Markush searches, MARPAT and the Merged Markush System (MMS) are the two main databases used for Markush searching, with over 1 million Markush structures.

6. INTEGRATION OF CHEMICAL AND BIOLOGICAL DATA

Retrieval and detection of chemical entities is of key importance on its own, yet there are many practical scenarios where it is important to systematically extract additional information, in particular chemical relationships. Correct chemical relationship extraction depends heavily on the prior detection of the individual entities taking part in the relation. Chemical entity relationships can encompass, for instance, chemical reaction/synthesis relations or relations of chemicals and particular physicochemical attributes. Nevertheless, recognizing associations between chemical entities, more specifically drugs and active pharmaceutical ingredients, with other entities such as proteins/genes or biomedical concepts, like diseases or adverse effects, is likewise significant and has resulted in the publication of a considerable number of chemical−biomedical entity relation extraction approaches. In the field of drug discovery, the identification of relevant chemicals that interact with the target protein opens the door to guided design of analogues sharing similar structural properties, very frequently using computer-aided drug design (CADD) models. Moreover, curate extraction of experimental data associated to this chemical−protein interaction is key for the development of quantitative−structure activity relationships (QSAR) and chemogenomics models. Obviously, this is not restricted to the biochemical target, but also applies to any kind of experimental data delivered by phenotypic, cellular, ADME-Tox, and in vivo assays providing reliable information as to speed up the drug discovery process. Not to forget, associations of chemicals (drugs), protein/gene pathways and clinical outcomes together with systems biology can be a clue for drug repurposing strategies,[603−605] MoA identification,[606] and chemical health risk assessments.[607] Alternatively, fast access to methods for preparing compounds has its interest, not only for the discovery of novel entities, but also for optimizing synthesis (yields, cost of reactants) during scale-up and production of novel or generic drugs. This fully explains the growing number of chemical databases with associated biological data in the postgenomic area, as well as recent efforts in Semantic Web technologies as a way to deliver data integration. Generally, the majority of public chemical and biological data repositories have been implemented by academic, government, or biotechnology institutions,[608,609] although in the past years, as part of their open-innovation collaborations, pharmaceutical companies have started to release selected data sets.[610−614] A desirable feature for these databases is the incorporation of a plethora of source documents, beyond scientific papers, particularly patent literature. Many chemicals, assays, and preparations are exclusively disclosed in patents. For example, the content of only 3−4% of biomedical papers is first published in patent applications.[615] However, patents are frequently disregarded by scientists because of their extension, language, and difficulty in relevant document retrieval, especially when using public repositories from patent agencies

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(as listed in section 2.2). Another important issue are concerns about the quality of the data,[616−618] a problem denominated by Fourches et al. as the "five I's": data may be incompleted, inaccurate, imprecise, incompatible, and/or irreproducible.[619] In this sense, assay standardization approaches, with precise nomenclature, electronic protocols, and formats for bioactive entities[620] as defined in bioassay ontologies (BAO[621] and Catalog of Assay protocols),[622] are indispensable to ensure robust data annotation. TM strategies for linking chemistry and biology are expected to surpass the bottleneck imposed by manual curation in terms of economic costs and time.

This section will primarily focus on the detection of associations of chemical entities and information of biomedical/pharmacological relevance. Integration of chemical and biomedical data is understood in this section as the automatic detection of mentions to relationships between biomedical entities and chemicals in text rather than proper integration of structured chemical data with the contents of biological knowledgebases. Section 5 covers some of the main chemical databases, while section 1.2 introduces existing repositories of documents of chemical relevance. Some of the existing chemical and drug repositories, such as DrugCentral, consider the importance of integrating multiple drug-related databases, using controlled vocabulary concepts and terminologies to annotate and describe key pharmacological aspects, and exploring the importance of TM applications for the retrieval of adverse events from drug labels.[589]

Several databases have been constructed to host collections of chemical reaction information, such as CASREACT[558,623] or Reaxys,[82,83,624] SPRESI,[594,625] SORD,[598] or current chemical reactions.[593] Typically, chemical reaction information is extracted by hand from the chemical literature or patents in a laborious process to feed database records.[626] Several attempts were conducted in the past to extract chemical relations from documents. Also, online chemical search engines, like SciFinder, have been analyzed in terms of different means to execute chemical reaction searches through the combination of various search options.[559]

Early work concerned with the automatic extraction of chemical synthesis reactions was carried out by Reeker and colleagues.[325] The aim of this effort was to assist in the construction of a database on chemical reaction information, by using automated chemical relation extraction to lower manual curation workload. They focused on processing of paragraphs describing the synthesis of organic compounds from the experimental section of the American Chemical Society's *Journal of Organic Chemistry*. To extract the chemical synthesis relations, they focused on the recognition of chemical reactants, products, solvents, and conditions of a chemical reaction. The used relation extraction approach relied on verb arguments (frames). This implied that they defined predicate−argument relationships by using manually generated case frames for each reaction verb, that is, defining a set of possible arguments for each predicate. Reaction conditions examined by this system covered temperature frames and time expressions. An evaluation of this system was carried out on 50 paragraphs, obtaining an accuracy of 78%.[325] A follow-up approach to extract chemical synthesis reactions based on a simple reaction schema (X + Y → Z) was published shortly after.[156] This later approach worked likewise at the level of individual sentences. They first decomposed complex sentences into simpler sentences with a single verb. The authors then exploited the syntactic structure of the sentences together with verb and preposition patterns to determine the role of the chemical substance in the reaction. They also published a more detailed strategy based directly on the work of Reeker and colleagues,[325] where predefined discourse elements of chemical documents were examined in detail, including the synthesis reaction discourse, workup discourse (reaction termination and information on the purification of the product), and the characterization discourse (physical constantans or experimental techniques). This work relied likewise on the use of verb-based frames and tried to map the extracted entities into slots of predefined reaction information form frames.[627] Finally, another paper by Blower and Ledwith, built on these previous attempts, used synthesis frames for the extraction of chemical synthesis reaction information with the goal of generating annotations for CASREACT records from ACS articles.[327] Instead of parting sentences, authors tried to match templates against sentence fragments. They used a list of manually constructed rules to assign specific roles (e.g., reactant, product, reagent, solvent, or catalyst) to participating chemical substances. Frame templates were also used to extract information related to quantity, reaction times, and temperatures. An evaluation of the performance of this system concluded that, in the case of simple synthesis paragraphs, this method could generate usable results in 80−90% of the cases, while for complex paragraphs only 60−70% of the results were acceptable. As a result, it was finally disregarded for CASREACT annotation.

Another frame-based reasoning approach to process analytical chemistry abstracts was presented by Postma et al.[628] These authors provided a formal representation of the input and output of an analytical action related to preparation procedures of chemicals. This formal description focused on the use of sentence structures and physical action verbs.

Jessop and colleagues published a detailed description of a strategy for automatically extracting reaction information from patents as part of a prototype system called PatentEye.[38] They identified that, from a linguistic perspective, chemists archetypically report descriptions of syntheses using past tense and agentless passive voice and that descriptions of syntheses can be classified abstractly into three segments: primary reaction, workup, and characterization. The primary reaction text segments comprise texts describing how the target compound is produced, the workup segment refers to descriptions on how the reaction is quenched/neutralized, that is, the removal of solvents and purification, while the characterization segment provides spectral information and demonstrations of the intended product. Several heuristics and the combination of information extracted from text, NMR and mass spectra, chemical name to structure algorithms, and chemical image to structure conversion results are explored to extract reactions from patents, while structural information is used to support role assignment. Lexical patterns were exploited for the detection of the chemical reaction roles and quantities. For instance, the reactant, which is the substance being consumed during a reaction, is typically reported together with the used quantity expressed by mass and molar amount, while solvents are usually expressed by chemical name together with volume information. This system also used ChemicalTagger for reaction name detection,[155] and OSCAR3 for annotating spectral data and identifying chemical entities referring to the product of the reaction.[154] The PatentEye prototype extracted reactions with a precision of 78% and recall of 64%. Lowe reimplemented a new reaction extraction system directly

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

inspired by the previous PatentEye prototype improving specific aspects, such as the preprocessing steps (tokenization and sentence parsing) and the recognition of chemical concepts.[626] The reaction roles covered by this system included product, reactant, solvent, and catalyst.

All of these works processed various kinds of documents in English language to extract chemical reactions, while only very limited attempts exist so far that address texts written in other languages. A non-English reaction extraction technique has been published using a rule-based approach with multislot frames to extract chemical reaction information from chemistry thesis abstracts written in Thai.[629] The underlying extraction rules/patterns were acquired using a supervised rule-learning approach,[375] which was trained on a manually tagged set of reactions and focused on reaction roles or components such as the reaction name, reaction product, and reactants.

6.1. Biomedical Text Mining

6.1.1. General and Background. One of the most prolific and mature application domains of TM and NLP approaches is the field of biological and biomedical TM.[15,630−634] Biomedical TM research has generated promising outcomes in terms of annotated text resources, biomedical lexica, methodological discoveries, and a considerable number of applications and text extraction components. Many biomedical text processing components have been published, covering a range of fundamental aspects, starting from the analysis and impact of different tokenization approaches,[153] or the implementation of specialized tokenizers to adequately handle the characteristics of biomedical texts.[157] There are also specially tailored linguistic and NLP components for biomedical texts, such as POS taggers,[635] and dependency parsers for syntactic analysis,[636] like Enju/Mogura[637] or GDep,[636] both with specific biomedical domain models to return syntactic dependency relations between words from a sentence. Syntactic dependency parsing essentially consists of, given an input sentence, automatically generating a dependency graph where nodes are words and arcs are dependency relations. Such syntactic relations have been exploited to detect bioentity relationships from text, such as protein−protein interactions,[638] in combination with ML methods.

Providing an exhaustive overview of all of the different types of biomedical TM strategies and application tasks is beyond the coverage of this section. To name a few of the most representative tasks, one can point out efforts to rank or classify articles for topics of relevance,[278] detect a variety of different types of bioentities mentions,[639−641] index or link documents to terms from controlled vocabularies or bio-ontologies,[642,643] and extract binary relationships between bioentities, in particular protein or gene relations like protein−protein interactions,[315,644−647] a topic that has attracted remarkably much attention. Bioentity grounding efforts have mainly focused on the association of gene and protein mentions to open access gene/gene product database identifiers.[451,648−650] Another prevalent biomedical TM research field is the detection of associations between genes and disease concepts[15,651−655] or descriptive functional terms, in particular, Gene Ontology concepts.[656,657]

As fully automatic extraction of bioentity annotations from text has to struggle with performance issues, using TM tools as part of the manual curation pipelines was also explored by several biological databases,[658] with the ultimate goal to scale-up and systematize manual curation of bioentities.[659]

Recent biomedical TM developments have dedicated efforts to prioritize or rank genes on the basis of relevance or association to certain topics or diseases, to detect complex events from text including pathway extraction, and to focus on particular pieces of information that might be directly useful to precision medicine.[660]

An important first step for most biomedical TM tasks is to locate mentions of biological entities of interest. Likewise, research in biomedical sciences is focused on the study of a set of entities, mostly genes, proteins, chemicals, drugs, and diseases. Thus, tools that can enable a more efficient recovery of documents that characterize these entities are greatly appreciated.

Among the bioentities that have been studied in more detail so far are mentions of genes, proteins, DNA, RNA, cell lines, cell types, drugs, chemical compounds, and mutations as well as mentions of species and organisms.[454,661−667]

In line with recent developments of chemical named entity recognition, the availability and use of manually labeled training corpora play an important role in the development of biomedical entity recognition systems.[232,641] Also, from the methodological standpoint, the same basic combination of recognition strategies used for chemicals has been tested to detect bioentities, that is, dictionary-lookup, rule-based, SL-based, and hybrid NER approaches. Nevertheless, the success of these different techniques depends heavily on the entity type, its naming characteristics, the availability of gazetteers, as well as the quality and size of annotated text corpora. For instance, highly structured entity names, such as microRNAs and somewhat mutation, can be detected with a satisfactory performance using rule- or pattern-based methods.[364,668−670] Nevertheless, recent systems like tmVAR explored the use of ML approaches for the detection of gene and protein sequence variants.[663]

Other types of entities have been extracted primarily using dictionary-lookup strategies due to the availability of rather comprehensive lexical resources, such as entity name lists or terminologies. This has been the case of species and organism mention recognition systems like LINNAEUS[667] or SPECIES,[671] and more specific species taggers like TNRS (Taxonomic Name Resolution Service) focusing on the detection of scientific plant names using string matching approaches.[672] There have been also some attempts to build hybrid rule-based/machine learning organism mention recognizers like OrganismTagger.[673] The recognition of species information is of key importance for the correct linking of automatically detected gene and protein mentions to biological database identifiers.[650,674,675]

In Life Sciences, the bioentity types that have attracted most attention are gene and protein mentions.[661] In the case of chemical entities, IUPAC nomenclature guidelines provide naming rules for systematic names, while in the case of genes the HUGO Gene Nomenclature Committee (HGNC) has provided guidelines for naming human gene symbols. The use of systematic names in the case of genes was not particularly successful, and naming standards are not sufficiently used in the literature.[676] When looking at eukaryotic gene names, only in 17.7% of the cases are official gene symbols used.[456] Gene NER systems have to deal with several difficulties for successfully recognizing gene symbols, such as ambiguity (e.g., many genes are referred to using short, highly ambiguous symbols or acronyms), the use of gene aliases, and naming variability due to alternative typographical gene name expressions.[677]

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Gene and protein mentions have been tagged using all main NER strategies and their combinations, that is, using dictionary-based systems,[348,678] rule-based methods,[350,369,679] ML-based approaches,[680−682] and hybrid methods.

For well-characterized species, such as humans and model organisms, online annotation databases constitute rich lexical resources to build gene/protein name gazetteers.[683] Such gene dictionaries have the advantage that they can be used directly for gene mention database grounding purposes, offering mappings between gene aliases and database identifiers, but they also typically require postprocessing or lexicon pruning to remove incorrect and ambiguous names.[451]

The recognition of the gene dictionary names is commonly done using exact string or word-level matching, but approaches based on the construction of name patterns from the original dictionary, and then scanning these patterns against the target or using fuzzy matching methods, have also been tested to enhance recall.[451] Instead of fuzzy matching, an alternative method is to use rules or heuristics to generate typographical gene name variants and then apply string matching using this expanded name lexicon.[684]

SL-NER systems trained on manually labeled text have been implemented to detect biomedical entities, using algorithms like Naïve Bayes,[685] HMMs,[686,687] MEMMs,[688] SVMs,[689,690] and, lately, analogously to CER systems, predominantly using CRFs.[680,682,691] Some taggers, like the GENIA[681] and ABNER[682] taggers, recognize multiple different bioentity mention types, like protein, cell lines, cell types, DNA, and RNA. In turn, GNormPlus also returns database-linked gene mentions.[648] Recognition of anatomical terms is covered by the CRF-based system AnatomyTagger[666] and MetaMap, a popular term matching system using lexical and syntactic term analysis to detect variant candidate terms[347,692] and also disease mentions. Recently, a ML-based disease mention tagger called Dnorm was released.[454]

6.1.2. Evaluations. Similar to the case of CER systems, community challenges that posed biomedical NER tasks and released Gold Standard entity mention training and test data have significantly promoted research and development of biomedical entity taggers. Detecting gene and protein mentions was part of several community efforts, BioCreative I,[426] II,[693] V,[317] and JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and its Applications),[662] which were key to determine the state of the art and cutting edge methodology.

The gene mention tasks of BioCreative I and II evaluated the automatic recognition of gene/protein NER systems using PubMed abstracts, assuming that gene and protein mentions could be regarded as a single entity class. Top scoring teams reached a very competitive performance of balanced F-measure (90%). At the JNLPBA task, a more granular bioentity type distinction was done, differentiating the following entity classes: protein, DNA, RNA, cell line, and cell type. More recently, a task posed at the BioCreative V challenge, the GPRO (gene and protein related object) task, tried to examine the performance of systems recognizing gene and protein mentions from patent abstracts. This task was slightly more difficult as participating system were asked to detect exclusively the subset of gene/protein mentions that can be grounded to biological databases. Top scoring systems reached a balanced F-measure of 81.37%.

6.2. Detection of Chemical and Biomedical Entity Relations

6.2.1. General and Background. Generally, NLP applications interpret relations as associations between entities in text. In principle, one can distinguish also other relation types, including grammatical relations, negation relations, or other linguistic relationships.[694] In practice, and for the purpose of relation mining, the relation extraction task refers to the automatic recovery of semantic relations between two (binary) or more entities.[372] Relation extraction (RE) strategies are very heterogeneous and are normally restricted by the underlying domain and complexity of the relationship categories.[695] RE can be addressed through co-occurrence (comention)-based methods, pattern and/or rule-based approaches, ML-based techniques, methods exploiting syntactic parsing, or hybrid approaches consisting of combinations of multiple strategies.

Early relation extraction efforts relied on a limited linguistic context and the use of word co-occurrences and pattern matching. In this line, relationship extraction between entities was treated as a typical information extraction task, where first predefined entity types are detected and then relations between them are represented as a template or form whose slots are filled with named entities (template filling).[696] Templates would represent particular facts or sometimes also called events (a term often denoting more complex associations), while slots would correspond to particular entity types, such as chemical compounds or drugs, genes/proteins, or diseases and adverse effects. Template filling systems using relation extraction frames encode particular relation types (e.g., "binds_to") and assign relationship roles to the extracted entities. For instance, a simple frame used for protein interaction extraction proposed by Blaschke and Valencia was: "[NOUNS] of (0−3) [PROTEIN] (0−3) by (0−3) [PROTEIN]", where NOUNS would correspond to an event trigger (a noun expressing binding or interaction derived from an interaction noun list), PROTEIN would correspond to a semantic label referring to a protein named entity, while "(0−3)" would correspond to a range of words that are allowed between each of the elements of the frame.[697] Frame-based systems can be a suitable choice for RE when sufficiently large and manually labeled training data are missing, and/or when domain expertise is available and the focus is on high precision, which is typically attained by such approaches. Another example of template filling system was used in the 2009 i2b2 medication extraction challenge, which aimed the extraction of different entity classes including medications, disease or syndrome, therapeutic or preventive procedure, and relationships between those entities.[698] This kind of frame-based systems can be grouped into the class of rule- or knowledge-based relation extraction approaches, which apply extraction rules to encode recurrent ways of expressing certain relations in text. These approaches have the underlying assumption that a considerable amount of relations can be recovered using a somewhat limited number of "typical" relationship text expressions. Relationship extraction rules commonly utilize the presence of event "trigger" terms semantically related to the predefined relation type, in combination with a collection of word patterns.[699] Predefined word patterns might encode POS information, relative position of entities within the sentence, word order, word distances, or sentence length. Manually constructed patterns are often implemented as hand-crafted regular expressions or as a cascade of heuristic rules.

In the biomedical domain, the current trend of relation extraction systems is to use hybrid systems combining different

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

strategies including SL-based techniques.[700,701] Still, other approaches, like rule-based methods, can yield very competitive results.[700]

The simplest approach for RE, a sort of baseline method, relies on comention, cooccurrence, or cocitation of entities within a specific context, defined as individual sentences, abstracts, paragraphs, or whole documents.[702,703] Simple co-occurrence, without any additional constraints, does represent the overall upper boundary in terms of relation extraction recall and lower boundary in terms of precision. Relation extraction using co-occurrence assumes that if entities are mentioned together in a particular unit of text, they should have some sort of association. In this line, Garten and Altman published the Pharmspresso system, which used co-occurrence to cluster comentioned drugs genes, diseases, and genomic variations.[704,705] Li and colleagues used cocitations of drug and disease names (and their synonyms) in PubMed abstracts and GeneRIF sentences to detect entity relationships, and to be able to build entity disease-associated entity networks.[706]

Clear advantages of entity co-occurrence-based applications as compared to other techniques include that they are (i) rather easy to implement, (ii) they can exploit standard statistical association measures to score relations (within a document or across multiple documents), and (iii) simple additional constraints like co-occurrence proximity, that is, how close entities appear in text, can be explored. In the case of the online bioentity co-occurrence search tool PolySearch2,[651,707] comentions between a range of different entity types and concepts, including drugs, metabolites, toxins, diseases, genes/proteins, drug actions, and other concepts, are automatically extracted from documents. PolySearch2 detects entity co-occurrences from various types of documents like PubMed, Wikipedia, U.S. patents, PubMed Central, and others, and also allows carrying out searches, including bioentity or concept synonyms and aliases. It takes into account co-occurrence proximity of comentioned entities by counting the words separating them and scoring higher those co-occurrences where the entities are mentioned closer together. PolySearch2 is primarily concerned about binary co-occurrence relations of user-specified entity types. Another online tool, PubTator, offers searching with user-provided entities as input, that is, chemical, disease, gene, mutations, and species, and then returns all abstracts mentioning the query entity in abstracts, labeled with all other co-occurring entities. A table of all of the co-occurring entity relations is automatically generated for a selected abstract, which can then be manually edited by deleting unwanted pairs and saving or exporting the remaining relations.[652]

To score the strength of entity co-occurrence, the typical measures used are the absolute frequency of co-occurrence, the Pointwise Mutual Information (PMI), and Symmetric Conditional Probability.[708−710] Co-occurrence frequency-based statistics can be used to rank individual relationships.[711] An example of an online application that extracts protein−chemical interactions using multiple data sources, including also text-derived associations, is STITCH (search tool for interacting chemicals). It uses co-occurrence of chemicals and proteins to retrieve relationships detected in PubMed abstracts, OMIM database records, and full-text articles.[712] A recognizable drawback of co-occurrence-based relation extraction is that such strategies do not provide any semantic evidence with respect to the kind or role of relationship existing between the entities. For some binary entity co-occurrences, for example, for comention of a chemical and a protein, a vast number of heterogeneous types of semantic association might be found in text. A partial solution to this issue can be found in the trioccurrence strategy, where additionally to the entities, the mention within the same sentence of certain relation trigger keywords or verbs is examined.[713] Co-occurrence-based relation extraction and trioccurrence, among other strategies, were implemented in the SNPshot system to detect 12 different binary relation types, including gene−drug, drug−disease, drug−adverse effect, drug−population (ethnicities, regions, countries, and inhabitants), drug−allele, and drug−mutation relationships.[714] On the basis of a manual evaluation, using sentences from pharmacogenomics-related PubMed abstracts that had at least one predicted relationship, the precision of this system was estimated, showing relatively good results for sentence co-occurrence, and particularly trioccurrence, in the case of gene−drug (82.7%) and drug−disease relationships (78.5%).

The method presented by Li and Lu also addressed the extraction of gene−drug−disease relationships by means of entity co-occurrences.[715] The used entity gazetteers were derived from the pharmacogenomics database PharmGKB,[716] but instead of focusing on the scientific literature, they examined entity co-occurrences in clinical trial records, trial record metadata, and descriptions from ClinicalTrials.gov records.[80] They discovered not only that the detected relationships intersect considerably with manually annotated relations obtained from the literature, but also that most of the relations detected in clinical trials appear on average five years before they are described in the literature.

SL methods are becoming an increasingly used option to extract entity relationships. These methods treat the underlying task as a classification problem viewed as, given a pair (or more) of entities and their context of mention (e.g., a sentence), and the method classifies them in terms of whether they are in a particular relationship class or not.[314,717−720] In the simplest case, it is treated as a binary sentence classification problem using a typical BOW representation of the sentences, where the entity pairs co-occur, and then applying one of the existing text classification algorithms, for instance, kernel methods such as SVMs. SL-based relation extraction is habitually more competitive in terms of recall when compared to rule-based approaches. Nonetheless, it requires a sufficiently large, representative, and consistent training data set, typically involving a considerable manual annotation workload. Entity relation corpora construction requires not only labeling of participating entities but also marking relationship types and negated relationship assertions. Some relation types require capturing the directionality of the relation event (e.g., substrate−product of an enzymatic reaction). In the case of pattern or rule-based approaches, syntactic relations between words returned by syntactic or dependency parsing software can be a suitable strategy to discover useful patterns, for instance, by examining the syntactic path connecting two entities in the parse tree (minimum path between them).[721] Syntactic parsing refers to the conversion of a sentence, that is, an ordered list of words, into a tree or graph structure where word−word connections encode their syntactic relationships. The word sequence connecting two entities, either viewed as a flat list of words, POS tags, or alternatively using the parse tree path (syntactic relations between words within sentences[722]), can be used as features by SL-based relation extraction methods. Figure 13 illustrates an example of a parse tree

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

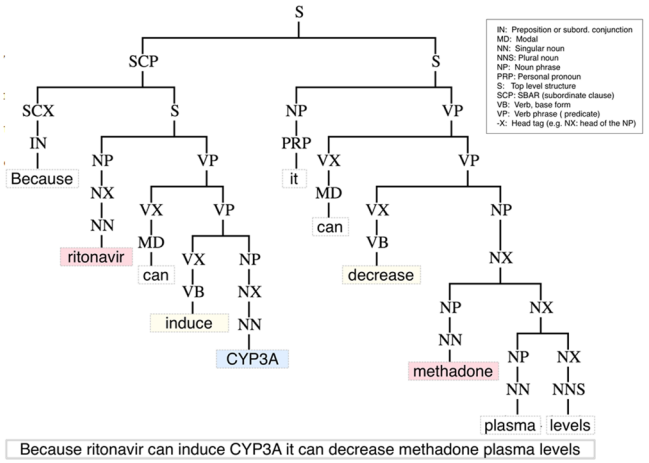


Figure 13. Example parse tree generated for a sentence using the Enju parser;[637] chemical entities are highlighted in red, protein entities are in blue, while relation trigger words are labeled in yellow.

generated for a sentence using the Enju parser.[637] Chemical entities are highlighted in red and protein entities in blue, while relation trigger words are labeled in yellow.

An alternative to generating automatically a full syntactic relation graph between all words is to define relations only between certain groups or "chunks" of words (e.g., noun phrases or verb phrases) without specifying the internal structure within those chunks.[723] This approach is known as shallow parsing or chunking, and it has also been exploited for relation extraction purposes.[718,724,725]

Syntactic word relations in principle do not directly convey semantic information to entity relationships. This can be addressed by examining the presence, within the syntactic path connecting the entities, of event trigger words, that is, words or terms often used to express certain relationships.[725−727] Relation events can then be represented through trigger terms and its arguments (entities) derived from the syntactic path. The resulting syntactic-semantic trees can be used as features for SL-based relation detection.

The scientific literature is a key information resource not only for drug safety but also for drug discovery, as relevant information to find potentially new drug targets often first appears in the academic literature.[728] The two types of relations involving chemical entities that have attracted by far most of the attention are chemical−protein and particularly chemical−phenotype (e.g., disease-related term) associations.

6.2.2. Chemical−Protein Entity Relation Extraction. Craven and Kumlien published an early strategy for the extraction of protein localization relations and stated their intention to extend this method to detect interactions between drugs or pharmacologic agents and protein targets, but they finally did not pursue this idea.[729]

Rindflesch et al. published a system called EDGAR (Extraction of Drugs, Genes And Relations) that detected relations between drugs, genes, and cell lines related to cancer therapy.[359] This early procedure relied mainly on concepts from the UMLS Metathesaurus to detect entity mentions, using the semantic types "pharmacological substance" for finding drugs and "gene or genome", in addition to other resources, for detecting gene mentions. They exploited syntactic information as a strategy to retrieve relationships, relying on relational vocabulary and predications that express interactions of the identified arguments (entities). Relation types extracted by EDGAR included drug−gene relations (drugs affecting gene expression) and gene−drug relations (gene/protein affecting drug activity).

Resources like SuperTarget host drug−target interactions acquired from multiple sources including TM.[730,731] The original SuperTarget database used the former EBIMed application to recover text passages describing candidate drug−target associations followed by manual curation to validate candidate interactions.[702]

Several TM methods have been published with the goal of detecting automatically drug relationships in the context of pharmacogenomics (PGx) research.[732−735] The field of pharmacogenomics studies how individual genomic variants might influence drug-response phenotypes. Ahlers et al. published a rule-based NLP pipeline to detect pharmacogenomics information, called SemRep.[736] It extracted semantic predications whose arguments were essentially certain entity types, such as drugs, genes, diseases, or pharmacological effects, and covered various types of pharmacological relations, including drug−pharmacological effect, drug−drug, and drug−disease relations. For instance, in the case of a drug−disease relation, the example relation type detected by SemRep would correspond to "<DRUG> *CAUSES* <DISEASE>", where DRUG and DISEASE refer to drug and disease mentions and *CAUSES* refers to a predicate or expression expressing a causative association. Other types of relations extracted by this system were pharmacogenomic relations (drug−gene, drug−genome). An evaluation of this system on a small set of relations showed that it had a recall and precision of 50% and 73%, respectively, for substance relations (relation types: INTERACTS_WITH, INHIBITS, STIMULATES) and of 41% and 68% for pharmacological effects (relation types: AFFECTS, DISRUPTS, AUGMENTS).

Tari and colleagues utilized the syntactic dependencies between words resulting from a parse tree to detect bioentity relationships including gene−drug relations. They illustrated gene−drug relation extraction by means of extraction of drug−enzyme metabolic metabolism and inhibition relations.[737]

Coulet et al. published yet another attempt to detect binary pharmacogenomic relationships between genes, drugs, and phenotypes (diseases and adverse effects).[733] They examined recurrently used syntactic structures underlying pharmacogenomic statements. Syntactic parsing uses grammatical rules to detect subjects, objects, and relation types. Pharmacogenomic relationships are represented as subject−object relations, while the use of an entity lexicon is exploited for the detection of participating pharmacogenomic entities. The type of detected relations is defined by certain relation trigger words like "inhibits", "transports", or "treats". A manual evaluation of a sample of automatically detected relations revealed that this system reached a precision of 70.0−87.7%, depending on the considered relation type.

Recently, members of the same research group described a systematic approach to extract drug−target relationships from PubMed abstracts.[738] By using simple dictionary-lookup to detect mentions of drugs and genes in a 2013 version of PubMed, they were able to recover a total of roughly 184 000 sentences containing both drug and gene name mentions, corresponding to around 236 000 unique drug−gene−sentence triplets. These sentences where then further processed using a syntactic dependency parser to recover the dependency path connecting drugs and gene mentions. Although they pointed out that a large portion of these dependency paths are

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

infrequent, implying a high degree of variability in drug−gene descriptions, they were able to apply clustering methods and a small seed set of drug−gene interactions to learn frequent structures of drug−gene relationship assertions.

Buyko et al. implemented a system for the detection of generic (coarse association) relations between genes−diseases, genes−drugs, and drugs−diseases reported in scientific literature.[732] To do this, they used annotations from a manually curated database called PharmGKB[716] to construct a training data set. PharmGKB was used to collect relevant references and to generate, together with additional lexical resources, entity gazetteers. Entity mentions were essentially detected using dictionary-lookup to retrieve sentences with co-occurrences. They adapted a SL-based system, JReX (Jena Relation eXtractor), to predict whether pairs of putative arguments (named entities) comentioned in sentences do have a relation association. Among the used ML features were lexical features, chunking features (i.e., the head words of phrases between the entity mentions), and dependency tree parse features.

Xu and Wang examined how prior knowledge in terms of known drug−gene pairs annotated in manually curated pharmacogenomics resources, like PharmGKB, can influence the automatic detection of PGx-specific drug−gene relationship, referred to by them as conditional relationship extraction approach.[734] Using a co-occurrence method to extract drug−gene pairs, they classified extracted entity pairs according to whether they had been previously known (annotated). According to the study by Xu and Wang, using such conditional relationships can improve drug−gene relationship extraction by 20.10% in terms of balanced F-measure.

Because of the key importance of cytochromes P-450 (CYPs) for the xenobiotic metabolism of drugs and chemical biotransformation reactions, relation extraction systems tailored to this class of enzymes have been implemented. The efforts to obtain systematically information concerning CYPs from the literature focused on particular interaction types,[739,740] CYPs polymorphisms, and metabolic characteristics.[741] For example, Feng at el.[742] tried to detect CYP3A4-compound interactions by applying pattern matching, keywords, and rules, while Yamashita and colleagues used interaction patterns that exploited a trigger verb list to retrieve CYPs-compound associations (i.e., substrate, inhibitors, and inducer relationships).[739]

Chemical relation extraction approaches, concentrating either on particular subsets of chemical entities or on certain types of proteins, have been proposed. In the case of the Herb Ingredients' Targets (HIT) database, manual curation was performed on relations between chemical compounds (corresponding to herbal active ingredients) and their protein targets, detected through a cocitation rule-based TM method.[743]

Chemical relation extraction was also implemented to cover a group of receptor proteins of key relevance for drug discovery, the predominant drug target membrane proteins known as G protein-coupled receptors (GPCRs). Chan et al. implemented a TM pipeline for the detection of GPCR−ligand interactions that included several steps, starting with the detection of GPCR and chemical entity mentions, and their co-occurrences within biomedical literature sentences.[744] This pipeline then picks those sentences mentioning as well certain binding trigger keywords. Binding triggers are words that express GPCR-ligand associations and consist essentially of verbs like "bind", "activate", "antagonize", and nouns and adjectives that refer to ligand attributes (e.g., "agonist", "antagonist"). Those triggers are detected in target sentences using regular expressions. GPCR−ligand interactions are finally detected using a SL-learning-based relation extraction approach, which examines, within the dependency tree, the shorted path connecting GPCR, ligand, and trigger word mentions.

Enzymes and the detection of enzymatic reactions represent another class of proteins for which chemical relation extraction efforts were performed. In fact, one of the first published chemical−bioentity relation extraction systems was EMPathIE (Enzyme and Metabolic Pathways Information Extraction), a template-based approach to extract relations between the following template elements: enzymes, organisms, and compounds.[745] It relied essentially on MUC-style information extraction templates and tried to assign product or activator roles to compounds participating in enzymatic reactions. This system obtained a performance of 23% recall and 43% precision on a corpus of seven journal articles from the journals *Biochimica et Biophysica Acta* and *FEMS Microbiology Letters*.

More recently, a strategy to extract metabolic reaction information was published by Czarnecki et al.[746] They used a pattern-matching and rule-based method that integrated the detection of stemmed metabolic reaction (e.g., "convert" or "hydrolys") and production (e.g., "produc" or "synthesi") keyword lists and variants of the verb "catalyze". Rules were applied to assign a weight depending on several factors, like the number of separating word tokens with respect to relation keyword occurrences or relative location of keywords in sentences labeled with candidate entities, that is, substrate, product, and enzyme mentions. The used rule scoring criteria were derived from an analysis of a small training corpus. This system was evaluated on three different metabolic pathways, resulting in a precision that ranged between 40% and 88% and a recall of 20−82% for substrate−product interactions, while the precision and recall for substrate−enzyme relations was 62−80% and 37−64%, respectively, and for product−enzyme was 58−67% and 36−70%.

Nobata and colleagues did not directly extract metabolic reactions, but focused on the prior recognition of metabolite mentions in the literature.[408] They implemented a NER system specifically for the automatic identification of yeast metabolite mentions in the scientific literature, and constructed therefore a manually annotated corpus of 296 PubMed abstracts labeled with metabolite expressions. The resulting metabolite NER tool consisted of a hybrid strategy using dictionary-lookup and CRF-based tagger that, according to their evaluation, was able to recognize metabolite mentions with a balanced F-measure of 78.49% and precision of 83.02%. Similarly, another hybrid metabolite NER system combining mainly dictionary-lookup with CRFs was recently presented by Kongburan et al.[747]

A TM framework for the detection of metabolic interactions, that is, enzyme−metabolite interactions, was recently developed by Patumcharoenpol and colleagues.[748] They differentiated between four classes of metabolic relation types or events, metabolic production, metabolic consumption, metabolic reaction, and positive regulation relationships. The used framework integrated existing NER systems to detect genes/proteins and metabolite compounds from the CheEBI dictionary. Metabolite events were extracted by using syntactic parsing, metabolic event trigger words, and a publicly available event classification system that uses a range of features, including syntax information and word features. An evaluation of this framework using an in-house corpus showed that it could recover metabolic production relations with a F-measure

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

of 59.15%, metabolic consumption relations with a F-measure of 48.59%, and metabolic reactions and positive regulation relations with F-measures of 28.32% and 36.69%, respectively.

6.2.3. Chemical Entity−Disease Relation Extraction. The largest number of the published chemical relation extraction strategies can be grouped under chemical entity−disease relation extraction systems. By disease, in this context, we mean a broad range of disease-related concepts, adverse effects/events, and side effects, while chemical entities typically are constrained to drugs, chemical substances in therapeutic use, or environmental chemical entities.

We provide a synopsis of some of the most prominent attempts to recover automatically chemical−disease relations from diverse documents, in addition to previously described efforts that extracted multiple relation types, including chemical/drug−disease relations. Typically, the recognition of chemical−disease relations requires first identifying mentions of chemical entities or drugs together with the detection of mentions of disease-related concepts in text. A range of different methods and resources have been developed for the recognition of both general disease terms as well as particular subtypes of disease-related concepts such as adverse effects, processing diverse types of documents including the scientific literature, medical records, social media, drug prospects, free text comments from databases, or patents.[414,420,453,454,680,749−766] The focus in this section will be on the scientific literature, while illustrative cases and references will be covering other document types.

Although most of the chemical−disease relations extraction systems have focused on the identification of adverse effects/events or chemically induced pathological conditions, some approaches have attempted to discover other classes of relationships like drug−disease indication/treatment associations. Drug−disease indication/treatment relations can be of particular importance for evidence-based medicine, clinical decision support, patient safety, and drug repurposing, with the goal to systematically determine and compare drug treatments used for a particular disease. The BioText corpus[767] represents an early effort to construct a small hand annotated resource for disorder−treatment relationships ("TREATS" and "PREVENTS" relations) using scientific abstracts, where treatments include also drug treatments.

Fiszman et al. implemented a system to recognize drug interventions for 53 diseases using the British Medical Journal (BMJ) Clinical Evidence journal[768] together with additional resources, that is, the Physicians' Desk Reference (PDR),[769] hosting information on FDA-approved drug interventions. This work relied on the previously introduced SemRep system[736] to retrieve arguments of treatment relation types to find drug therapies, with the goal to use the recognized relations within a medical text summarization system. Recently, SemRep was exploited together with a module for UMLS Metathesaurus concepts look-up and the TextRank algorithm[770] for ranking sentences describing treatment alternatives for Alzheimer's disease.[771]

The authors of the BeFree system, which detects drug−disease, drug−target, and gene−disease relations, additionally examined the capacity of SemRep to identify the same associations, focusing on different association subtypes including treatment relationships. BeFree, in turn, uses SL-techniques trained on the EU-ADR corpus[409] and exploits the combination of different association features and morpho-syntactic information to recognize drug relations.[772]

Névéol and Lu also adapted the SemRep system to detect "TREATS" predications (relationships) between diseases and drugs.[753] They used certain UMLS semantic types to define disease and drug concepts and applied the relation extraction pipeline to automatically detect drug indications from multiple resources, including DailyMed[77] and MeSH scope notes[773] and DrugBank[410] and PubMed records.

Some drug−disease treatment extraction approaches explored the use of methods based on manually or semimanually constructed rules or patterns.[774,775] Lee and colleagues explored pattern-based approaches to recover treatment associations from abstracts related to colon cancer. First, they explored the use of frequently occurring text patterns to automatically detect treatment associations, but due to limited precision they finally used manually defined linguistic patterns.[774] Abacha and Zweigenbaum proposed another approach relying on hand-crafted linguistic patterns and domain knowledge for detecting four types of drug−disease treatment relations ("causes", "diagnoses", "treats", and "prevents" relationships).[775]

In the work of Embarek and Ferret,[776] automatically constructed patterns were used for detecting four indication relation types, "Detect", "Treat", "Sign", and "Cure" associations, between five types of medical entities, diseases, exams, treatments, drugs, and symptoms. The linguistic patterns were generated through a sentence alignment algorithm using edit distances, wildcard operators, and mappings between sentence parts through a multilevel representation of words defined as their inflected form, POS label, and lemmatized form. SL-methods based on the CRF algorithm were applied by Bundschus et al. to detect diseases−treatment relationships[719] trained on a manually annotated corpus of PubMed sentences.

With the aim of detecting relationships between medications and indications (i.e., drug-condition data), Li et al. implemented a strategy for automatically mining potential reasons of medication prescriptions in clinical outpatient notes.[777] They linked the extracted relation pairs to a knowledgebase of indications for selected drugs assembled from multiple resources.

The automatic detection of treatment-specific relations between approved drugs and diseases can be a valuable knowledge source for drug repurposing, that is, using known drugs to treat novel diseases. This requires the capacity of detecting off-label, new, use of prescribed drugs. Xu and Wang induced potential treatment-specific textual patterns using co-occurrences of known drug−disease pairs in Clinicaltrials.gov sentences.[778] The resulting patterns were ranked on the basis of the associated number of known drug−disease pairs, and, by manually examining the top patterns, those specific to drug treatment relations were selected. These drug treatment specific patterns were then used to systematically obtain drug−disease pairs from PubMed abstracts, resulting in the detection of 34 305 unique drug−disease treatment pairs.

Jung and colleagues applied SVM-based machine learning classifiers to automatically retrieve potential off-label drug treatments obtained from clinical notes and also using database-derived features.[779] To make sure that they actually detected drug−indication usage pairs corresponding to drugs for unapproved indications, they applied filtering strategies to account for adverse effects and comorbidities associated to the approved drug use. Overall, this strategy yielded 403 well-supported novel off-label drug uses.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

ML-techniques were also explored for the purpose of categorizing drugs with respect to anatomical therapeutic chemical class labels, a relevant information resource for drug repurposing.[780] The features used by this classifier were previously obtained by an IE pipeline that detected terms relevant to several drug related characteristics, including therapeutic and pharmacological properties.

Instead of detecting drug indications automatically, Khare et al. explored a hybrid strategy that detects automatically drug and disease mention in drug product labels and then uses a crowdsourcing platform to ask humans to label those binary drug−disease co-occurrences that correspond to treatment relationships.[781]

Recently, a community challenge task, the BioCreative V chemical−disease relation (CDR) task, specifically addressed the issue of disease named entity recognition (DNER subtask) as well as the detection of chemical-induced disease (CID) relations[425,782] in PubMed abstracts and titles. This task can be considered the first systematic and independent evaluation setting of diverse disease entity recognition and chemical-induced disease relation extraction systems applied to scientific literature. From the 16 teams participating in the DNER task, the top scoring system could reach an F-measure of 86.46%, while out of the 18 participating teams of the CID task the best team obtained a F-measure of 57.03%. For the CID task, systems had to return a ranked list of chemical−disease pairs together with normalized concept identifiers. For this task, the organizers provided manually labeled entity and relation mentions. In the case of chemical entities, the CID task annotators followed closely the annotation criteria used for the CHEMDNER corpus whenever possible.[318]

Two review articles by Harpaz et al.[761,783] and one by Karimi and colleagues[784] provide a general overview of selected recent efforts using TM techniques with the goal of detecting pharmacovigilance (PhV, drug safety surveillance) relevant information. In particular, these reviews focus on adverse drug events (ADEs), extracted from heterogeneous document types like scientific literature, medical records, drug product labels, content from social media, and Web search logs. They estimated that roughly 340 thousand ADE relevant articles are contained in the PubMed database and 13 thousand new ADE specific records are indexed each year.[761] It is thus not surprising that the scientific literature, and especially PubMed records, have been used as a source for the computational detection of ADE.

Instead of mining directly the free text content of PubMed records, structured metadata annotations consisting of MeSH terms can be explored to retrieve potential candidate ADE information.[785,786] Following this idea, Avillach et al. developed a system that relied on the combination of particular MeSH descriptors, supplementary concepts, and subheadings, together with a threshold based on the number of publications, to select ADE candidates related to chemically induced adverse effects and pharmacological actions.[785]

MeSH index terms were used by a method published by Shetty et al. to detect ADE associations.[787] In this work, MeSH term-related features were used by a statistical document classifier to eliminate ADE-irrelevant records. Furthermore, they applied filtering steps to remove drug−disease associations corresponding to drug indications by analyzing product label information. According to this study, 54% of the examined ADEs could be identified before the drug warnings were published.

PubMed articles were also the document types processed by Wang et al. to recognize ADEs.[788] The used pipeline consisted of three main steps, a PubMed search, a document classification step, and a drug-ADE classification step. The initial retrieval step relied on a search query comprising keywords referring to drugs and adverse events. The resulting hits were then automatically classified as being ADE-relevant or not by relying on a logistic regression algorithm that used two types of features: (1) 21 ontological features including MeSH headings and chemical compound entities and (2) 14 textual features. The classifier was trained and tested using a set of 400 hand-labeled records. The last step examined the fraction of positively classified articles to decide whether the ADE is correct. This pipeline was tested for two relevant classes of adverse events corresponding to neutropenia and myocardial infarction.[788]

SL-algorithms for adverse reaction relation extraction have been applied to a particular type of PubMed records, case reports, and described in a 2012 publication by Gurulingappa et al.[789] This system was trained on a corpus known as ADE-SCAI corpus, consisting of 2972 PubMed case reports tagged with mentions of conditions (5776 mentions, covering also diseases), drugs (5063 mentions), and sentence-level drug−adverse event condition relationships (6821 mentions).[790] The detection of drugs and conditions was essentially handled through a dictionary-lookup approach, exploiting as lexical resources DrugBank and the medical dictionary for regulatory activities (MedDRA),[791] while the relation detection framework Java Simple Relation Extraction (JSRE),[792] relying on SVM classifiers, was trained to detect sentence-level drug−adverse event associations.

In a previous effort, Gurulingappa and colleagues implemented an ADE sentence classifier with the aim of improving the retrieval of ADE related sentences. They also used the ADE-SCAI corpus for training the sentence classifier, considering those sentences that had at least a single annotated ADE as positively (ADE-relevant) labeled sentences. An analysis of different SL-algorithms was carried out using different feature sets, such as words, lemmatized words, drug-matches, condition-matches, bigrams, trigrams, and token-dependencies.[793] Finally, they used a maximum entropy-based classifier, which according to their evaluation obtained the most competitive performance for their data set.

Yang et al. posed the assumption that letters to the editor of medical journals might be a useful resource for early signals of drug-related adverse effects.[794] They selected this kind of document and implemented a binary SVM-based classifier to score candidate drug−ADE pairs, given the text contents of the letters. This classifier used features derived from MetaMap results and some kind of text patterns relying on n-grams.

PubMed abstracts have been processed to detect very specific types of adverse events of particular relevance for toxicology studies, drug-induced liver injuries (DILI). Fourches et al. applied a commercial text processing pipeline, the BioWisdom's Sofia platform, to extract statements related to drug-induced hepatotoxicity in the form of concept−relationship−concept triplets, comprising mentions of compounds and terms related to hepatobiliary anatomy/pathology.[795]

Kang and colleagues, instead of using ML approaches, implemented a knowledge-based relation extraction system for the detection of ADEs from the literature.[756] The advantage of such a strategy is that it does not require a large training corpus of ADE annotations but still can yield competitive results.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Drugs and adverse events/disorders detected within this work were mapped to UMLS Metathesaurus concepts, and relation paths between concepts encoded in the UMLS structure were exploited by the knowledge-based system for relation extraction purposes.

A rather casual category of text that is becoming extensively mined for adverse drug reaction information are patient-reported or health consumer contributed data as well as social Internet communications, like blogs or twitter messages.[754,755,796−806] This kind of textual data is characterized by the use of somewhat informal descriptions of health-related concepts, including adverse reaction descriptions, differing considerably from how health care professionals or scientists account the same observations.[800] Karimi et al. as well as Harpaz and colleagues[761] provide a synopsis of TM strategies to detect adverse drug events by processing social media data.[784] An illustrative case of social media mining for adverse reactions focusing on four drugs was carried out by Leaman et al. who exploited as input data source user comments of the DailyStrength[807] social network.[796] In this work, authors generated a lexicon of adverse event concepts from multiple lexical resources, including UMLS, and added manually selected colloquial phrases describing adverse reactions. For the detection of the lexicon terms, these authors had to deal with spelling errors through the use of string similarity algorithms for term recognition.

Another valuable source of textual descriptions of adverse drug reaction information can be found in particular sections of drug product labels or package inserts of drugs. For instance, the extensively used resource SIDER (side effect resource) hosts side effect information extracted through TM methods from sections describing the indication areas and side effects of drug labels provided by the FDA.[808] For the extraction process, the authors of SIDER used a dictionary of side effect terms, which was obtained using COSTART (Coding Symbols for a Thesaurus of Adverse Reaction Terms)[809] as seed lexicon, and expanded by selecting synonyms and equivalent terms from the UMLS Metathesaurus.

Duke and colleagues published a system called SPLICER (Structured Product Label Information Coder and ExtractoR) using a rule-based text processing technique, which relies on regular expressions and patterns to retrieve adverse events from product labels, and also associates the detected adverse event terms to the MedDRA (Medical Dictionary of Regulatory Activities) concepts.[809]

Smith et al. published an effort to extract and integrate ADEs derived from multiple publicly available document sources, including human drug labels.[810] First, they identified mentions of drugs, drug ingredients, and brand names and mapped them to unique UMLS concept identifiers. They then detected mentions of diseases, or, more precisely, what they call clinical manifestations within certain sections of drug labels and filtered negations of those terms. The terms detected in the "Adverse Reactions" section were considered to be ADEs, while the terms recovered from the "Indications and Usage" section were tagged as drug indications.

In the work of Bisgin and colleagues, UL methods (topic modeling) were used to process FDA drug labels obtained from DailyMed to detect topics that could be used to group drugs that show similar treatment characteristics and drugs that show similar adverse events. They processed particular sections of FDA labels related to adverse reactions to label each drug with standardized vocabulary in the form of MedDRA terms.[811] During the topic-modeling step, the conditional probability of every topic given a drug was calculated, and, afterward, for each drug, the single topic corresponding to the highest topic conditional probability was determined. During the topic assessment step, the authors resolved which topics were related to adverse effects and which ones were associated to therapeutic or treatment effects.

Another document source that has been exploited to extract mentions of drugs and their associations to adverse effects and diseases are electronic health records (EHRs).[812−821] The use of TM techniques to recover potential adverse drug events from clinical documents is an intense research topic due to its practical importance for postmarketing drug safety and pharmacovigilance. A review paper by Warrer et al. describes several TM and NLP approaches that extract adverse drug reactions from electronic patient records.[816]

Although some EHRs might show a certain degree of structure, under typical circumstances, they comprise written free text portions in clinical narrative format. Clinical documents do show a series of particularities and challenging properties.[817,822] They are characterized by a heavy use of abbreviations, some of them created ad hoc; it is a rather noisy type of text with a high density of typographical and spelling errors, the use of ungrammatical sentences, the lack of punctuation marks, and a frequent use of negated assertions (e.g., "DRUG does not cause ADVERSE EFFECT"). Moreover, in the case of clinical documents, text processing systems have to go beyond processing documents written only in English.

An early attempt to detect ADE from electronic records was done by Honigman and colleagues.[813] They relied on International Classification of Diseases (ICD-9) codes that were related to adverse drug events and used manual revision by researches to determine whether an ADE occurred. A similar attempt was carried out by Field et al. to identify drug-related incidents occurring in the ambulatory clinical setting using multiple sources, including discharge summaries and emergency department notes.[823] They used a computer-based free-text search strategy to identify potential drug-related incidents and relied on pharmacist investigators to review drug-related adverse incidents.

Another early strategy to recover adverse medical events, including adverse drug reactions, was described briefly in a publication by Murff et al., where keyword queries were used to capture trigger words related to adverse events found in medical discharge summaries.[824]

Quantitative association statistics using chi-square statistical tests were used by Wang et al. to detect relations between co-occurring drug mentions and adverse effects extracted from discharge summaries of inpatients.[812] Therefore, they applied entity recognition approaches to detect mentions of medications (drugs) and adverse drug events and used filtering techniques to remove confounding factors (e.g., diseases and symptoms that appeared prior to the administration of the drug) and to exclude cases of negated adverse event assertions.

Hazlehurst et al. used a knowledge-based NLP system called MediClass, for the detection of vaccine-related adverse events (VAEs) from electronic medical records, with particular emphasis on gastrointestinal-related VAEs.[825]

A more recent work with the aim of detecting drug-related adverse effects in free text EHRs was carried out by LePendu.[818] The patient-feature matrix constructed from clinical notes captured associations between patients, drugs,

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

diseases (adverse events), devices, and procedures. The recognition of drugs and diseases was done using a lexicon that was carefully constructed using multiple lexical and ontological resources and filtering steps to remove non-informative terms. Drug prescriptions were mapped into active ingredients using a controlled vocabulary.

Sohn et al. describe the adaptation of an existing clinical NLP system to detect drug side effects in electronic clinical notes from psychiatry/psychology hospital departments.[826] The recognition of potential adverse effect terms (e.g., signs/symptoms or disorders/diseases) and drugs was handled through the clinical concept recognition module provided by the cTAKES framework. A rule-based approach using regular expressions and handcrafted patterns was used to detect associations between drugs and potential adverse effects. Sentences detected by this rule-based approach were, in turn, used as training data to build a ML-based adverse effect sentence classifier.

Although most of the published efforts to use TM strategies for the identification of ADE in clinical documents was carried out in English, some attempts have also been made to recover this kind of information from electronic health records written in other languages. Aramaki et al. published a short description of a system that detects ADE from Japanese discharge summaries.[827] They used a SL-based NER system based on CRFs to label mentions of drugs and symptom expressions. Co-occurring drug and symptom pairs were then classified as being ADEs by combining a pattern-based technique and a SVM-based classifier.

A manually annotated corpus and automatic TM system related to adverse drug reactions extracted from discharge reports in Spanish was presented by Oronoz and colleagues in 2015.[820] The entities covered by this work comprise drugs, procedures, and diseases, while in the case of the entity relationships the focus was set on adverse drug reactions. Using the manually annotated corpus as seed set an automatic annotation system called FreeLing-Med was generated.

The detection of single drug adverse events does only capture partially the space of adverse reactions associated to the administration of drugs. A very common clinical scenario is polypharmacy, referring to the usage of several concomitant drugs for treating medical conditions. Moreover, especially in elderly patients, several medical conditions need to be treated simultaneously, implying the intake of more than one drug at the same time. The administration of multiple drugs can potentially result in drug−drug interactions (DDIs) leading to adverse effects.[828] Several relation mining systems have been developed to detect DDIs in free text,[701,718,828−831] and two community challenges have addressed the task of extracting DDI from text.[701,718,832]

In a work published by Tari et al., TM methods to detect biological facts from Medline abstracts relevant for DDIs were integrated with different information sources derived from structured databases capturing biological domain knowledge important for the understanding of DDIs.[828] For recognizing and disambiguating enzymes, BANNER[680] and GNAT[650] were, respectively, used, while drug mentions were tagged by MetaMap.[692] To produce structured sentence representations for the relation extraction step, syntactic parse trees were generated. Two types of relations were recovered. Explicit interactions characterizing how one drug might influence another one were detected using trigger words like “induce”, “induction”, “inhibit”, and “inhibition” together with word proximity constraints. In the case of implicit drug interactions, relations between drugs are inferred by capturing drug metabolism relations in the form of [protein, metabolizes, drug] triplets.

Most of the current approaches used for the detection of DDIs make use of SL-learning methods exploiting lexical, syntactic, and semantic features[701] and represent this relation extraction task as a classification problem of candidate DDI pairs. The DDI extraction challenges, which were carried out in 2011[718] and 2013,[832] played an influential role in promoting the development of this kind of approaches through the release of manually labeled training data sets. One system that was evaluated on both DDI challenge test sets was published by Bui et al.[830] This extraction pipeline generated syntactically structured sentence representations containing DDI candidate pair mentions and used a SVM classifier, including lexical, phrase, verb, syntactic, and additional features related to drug names, to classify DDI candidate pairs. This system was able to reach a F-score of 71.1% against the DDI challenge 2011 test set and of 83.5% using the DDI challenge 2013 test data set.

7. FUTURE TRENDS

TM and NLP technologies are playing an increasingly important role in many areas of science, technology, and medicine. Indeed, the integration of TM technology with web search engines and related AI products is one of the most active current technological research topics, which is not surprising because the capacities to decode text and interpret human language are essential for the development of systems able to interact with humans.

The implementation of such technologies and their adaption to handle textual data from the domains of biology and biomedicine is also increasingly important, as they provide key methods for medical support systems, scientific assessment tools, and strategies for the interpretation of large scale biological systems. The same is true for the application of TM methodologies in chemistry, which is already starting to follow a similar path as can be seen by the increasing integration of chemical TM and NLP components into search engines and support engines embedded in AI systems. Still, as we have seen in this Review, there are numerous chemical domain-specific challenges that have to be faced more efficiently. In this section, we stress two of them: (i) the annotation of chemical-specific information and (ii) the integration of information extracted from text with the data stored in curated databases.

Documents baring information of chemical importance are very heterogeneous at various levels. Scientific literature and patent documents differ substantially in their length, structure, language and used vocabulary, writing style, addressed topics, and information content, all of which are aspects that influence significantly the use and performance of chemical TM systems. Moreover, beyond technical and methodological challenges, accessibility issues and/or legal constraints faced by chemical information retrieval and TM tools limited research and implementation of systems that process full-text patents and scientific articles, and consequently most efforts were restricted to article abstracts. Considering that patents are in fact a unique source of information (e.g., it is estimated that only 6% of bioactive compounds described in patents are later referred to in scientific papers),[833] chemical patent mining is increasingly becoming a critical research field,[834] as emphasized by the number of participants at the CHEMDNER patents task of BioCreative V,[317] and will probably continue to be so in the

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

near future. Recent progress in patent mining resulted in the extraction, annotation, classification, and release of 1.15 million unique chemical reaction schemes retrieved from pharmaceutical U.S. patents from the period between 1976 and 2015.[5] The multilingual nature of patents constitutes another challenging aspect for chemical TM systems, particularly as compared to the prevalence of English language publications in the case of the scientific literature, and will undoubtedly require the development new tailored multilingual text processing approaches. For instance, lately a substantial growth of novel compounds was published in patents from the Asian Pacific region.[835]

Another chemical patent mining task that will require future developments concerns the intrinsic hierarchical recursive structure of patents claims sections. Notably, there is the need to resolve broad chemical terms (R-group definitions) and their dependency and association to generic Markush structure diagrams, which involves OSR processing. Mining of frequently used structure−activity relationship (SAR) tables found both in medicinal chemistry papers and in patents represents another challenging task that requires the development of new text processing solutions.

Together with scientific publications and patents, the World Wide Web is becoming an increasingly important primary source of unstructured text for chemical TM. Today, social media and mobile applications are routinely and massively used, and are generating a vast amount of unstructured data about collective human behavior. This is of key interest to the construction of population-level observation tools. Within the biomedical domains, some works have already been exploring the potential of nonconventional data sources, for instance, the discovery of adverse drug reactions and drug−drug interactions based on Twitter and Instagram data.[836,837]

Image mining is also of relevance to TM applications and, most notably, to chemical TM. OSR is yet at its infancy, and there is still much room for improvement, particularly dealing with graphical ambiguities (e.g., touching and broken characters, or characters touching lines), the recognition of large macromolecular structures and complicated rings, the correct interpretation of Markush features (e.g., substituent replacement in R-groups, link nodes, or repeating units), and the recognition of chemical tables or reactions requires substantial further research.

A second challenge in chemical TM is the integration of data extracted from different data and information sources. As compared to standardized gene and gene products nomenclatures, the existence of multiple representations for chemicals entails significant field-specific obstacles. Effective data mining requires accurate structural representations and chemical naming, while public databases struggle to provide unambiguous names and synonyms and do not fully solve the problems of uniqueness and unambiguity of structural representations. Even if there is a clear consensus in the need of incorporating chemical structural information into matching identifiers (e.g., as with InChi codes), as the only effective solution to this problem, current results mapping automatically extracted chemical names to identifiers in chemical repositories show the many difficulties of this task and the need of substantial investments in this area.

Beyond existing issues related to entity normalization, in an increasing number of applications, chemical information has to be viewed in context and integrated with other data types such as pharmacological, toxicological, and biological attributes. This is the prototypical scenario of many genomic applications, in which new semantic classes and ontologies are being applied to support semantic web developments and which are increasingly adopted by core biological databases.[838] It is fairly noticeable that we are going to meet new and more powerful developments, based on semantic web technology, and linking information types, such as entities from chemistry and biology, which will make the access and dissemination of information easier and facilitate integrated chemical information. On the other hand, data integration faces an additional critical challenge regarding access to fee-based databases that severely jeopardize information availability, and thus the potential usage of multidisciplinary integrated data. The axis formed by data-information-knowledge is critical for scientific progress; then, promoting data sharing is a "must": from funding agencies to professional associations and research journals (e.g., as condition of publication or receiving grants).

Another key aspect of TM is integration into adequate analysis environments, especially the visualization of semantically enhanced chemical documents, navigation across terms, and annotation types, and the ability to produce semantic summaries and to cross-link annotations to consolidated resources (e.g., databases and ontologies). At a first glance, TM outputs are sometimes seen as simple frequency statistics of annotated terms, where wordclouds are a commonly used way to display how many times each term is occurring in the text or document collection. In turn, graphs are a powerful means of characterization of term representativeness as well term co-occurrence associations, or some other form of correlation. For instance, graph nodes (i.e., annotations) can be adjusted in size according to their annotation frequency, and two chemicals that co-occur very frequently should be connected using a thicker edge.

Visualization technologies also provide support to human experts when creating gold standards and controlled vocabularies, that is, in the manual annotation/labeling of texts according to domain-specific guidelines. So, document rendering abilities have to team up with annotation (highlighting and marking) and navigation abilities to provide an integrated environment for the curation of document contents. The use of Web-based programming, JavaScript libraries, is the leading trend in underlying the construction of easy-to-use and highly flexible curation tools. Efforts concentrate on providing lightweight multiuser environments that enable manual text annotation, calculate interannotator agreement, and provide, to some extent, semiautomated means to revise/validate annotations. Browser compatibility and devising ways of making highlights/annotations noninvasive are two open challenges.

7.1. Technological and Methodological Challenges

TM is experiencing rapid and significant evolution. TM is taking advantage of recent computing technologies to boost and optimize existing software as well as looking for emerging technologies that may help address open challenges. At the technical level, scalability is a key demand in many fields and also in chemical TM. The emergence of flexible cloud-based virtualization techniques promises good/reasonable solutions to this problem. For instance, cloud computing has been applied to the accumulation of concept co-occurrence data in MEDLINE annotation[839] and virtual screening[840] searches.

Server virtualization ensures effective software development and testing but demands extensive system level knowledge. Docker technology has become very popular over the last years,

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

because docker containers employ the kernel of the host machine; that is, they do not require or include the whole operating system, in contrast to virtual machines, which emulate virtual hardware. The myCHEMBL platform is an example of a virtual machine distribution that has recently incorporated Docker technology.[841]

Aside from scalability issues, the lack of interoperability among biomedical TM tools is a major bottleneck in creating more complex applications. The number of methods and techniques for common TM tasks, IR and NER, keeps increasing, but combining different tools requires substantial effort and time due to heterogeneity and variety not only of technologies but also of data formats. Some formats, such as BioC, aim at harmonizing the presentation of text documents and annotations, but their use is far from being widespread. Currently, JavaScript Object Notation (JSON) is the most popular data interchange format, and BioC and other biomedical formats are already migrating. Also noteworthy is that the Chemical JSON (http://wiki.openchemistry.org/Chemical_JSON), which represents chemical molecules, may be useful to TM, to present chemically relevant outputs.

At the methodological level, traditionally, TM has been focused on supervised predictive learning. In the particular case of chemical TM, attention was set on the development of chemical entity taggers. While this area of work is still important, because not all relevant chemical types have been dully addressed so far, new areas are already attracting some attention. For instance, sentiment analysis or opinion mining is becoming of interest due to the increasing interest in social media data. However, document summarization, which is a long-term TM task, is still an open challenge, although the ability to produce domain-focused and semantically enriched summaries is ever more important.

Along with the typical IR methods, many other techniques have been developed to empower search engines and enhance knowledge extraction workflows. IR has incorporated techniques related to term weighting, natural language querying, ranking retrieval results, and query-by-example. More recently, attention has been driven to the uncertainty inherent in the IR task and to add intelligence to IR systems. A key component in terms of intelligent behavior is flexibility, intended as the capability of learning a context and adapting to it. IR systems should be tolerant to uncertainty and imprecision in user–system interaction (i.e., allow a more natural expression of user needs) and, at the same time, be able to learn the user's/domain's notion of relevance (i.e., elucidating information preferences). So, the so-called intelligent IR methods aim to intervene in areas/tasks such as personalized indexing, relevance feedback, text categorization, TM, and cross-lingual IR, among others.

Probabilistic models are being explored as a strategy to deal with uncertainty. For instance, contextual models have been explored in chemical patent search,[842] compound similarity search,[843] and drug connectivity mapping.[844] The *pmra* is another probabilistic topic-based model for PubMed content similarity that computes document relatedness on the basis of term frequencies, that is, the probability that a user would want to examine a particular document given known interest in another.[845]

Typically, query representation is based on keywords, and the retrieval mechanism performs a lexical match of words. One of the main problems with this approach is vocabulary mismatch; that is, users employ different words than those that are found in relevant documents. Moreover, distinct users are likely to employ distinct words to describe the same documents. So, a new area of IR research is tackling the definition of methods for vocabulary expansion. Vocabulary expansion can be achieved by transforming the document and query representations, for example, by using Latent semantic indexing (LSI), or it can be done as a form of a dictionary automatically built by corpus analysis. LSI is basically a sort of algebraic document retrieval representation model that relies on singular value decomposition of the vector space of index terms, and maps each document and query into a lower dimensional space representing their most important features. For instance, LSI has been applied to the search and retrieval of documents with textual, chemical, and/or text- and chemistry-based queries in PubMed.[846]

Inspired by search engine approaches, new methods of evaluating word similarity take into consideration word co-occurrence or other measures of word relatedness. For instance, these methods have been applied to the discovery of protein–protein interactions[847] as well as eliciting antiobesity/diabetes effects of chemical compounds.[848] Semantic relations have also been explored to identify genes, chemicals, diseases, and action term mentions in the Comparative Toxicogenomic Database.[849]

Deep learning has been a key reference for big data in the past few years, in particular for temporally and spatially structured data, such as images and videos. It has also been explored in TM, but with less success. Texts, usually treated as a sequence of words, are not suited for direct use of deep learning systems (there are too many words in a language). However, some works have already explored this avenue. For instance, a system participating in CHEMDNER competition integrated mixed conditional random fields with word clustering (including a Skip-gram model based on deep learning) for chemical compound and drug name recognition.[344] Deep semantic information was also applied to large-scale MeSH indexing.[850]

Last, it is worth mentioning that the use of dimensionality reduction techniques is increasingly demanded. Despite the availability of big data platforms and parallel data processing algorithms, the use of large data volumes may sometimes compromise the performance of the systems. Current techniques include correlation filtering, random forests/ensemble trees, principal component analysis, backward feature elimination, and forward feature construction. Some works have explored these techniques for biomedical applications; for instance, PubMiner uses a feature dimension reduction filter while mining useful biological information, such as protein–protein interaction, from PubMed,[851] and the SparkText Big Data infrastructure for TM uses feature extraction and dimension reduction to provide relevant features and thus condense the feature space for text classification methods.[852]

In summary, the first wave of applications of TM in chemistry is addressing the more demanding issues in the field and is also helping to clarify the many future challenges, including document heterogeneity, complexity of the chemical structures, integration with other data sources, and data representation. The progress in the field will come at the hand of the application of new technologies, including NLP, ML, and semantic encoding together with the implementation in robust distributed computational systems. With no doubt, the future will depend on the collaboration between the experts in these technologies with domain experts.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673–7761

AUTHOR INFORMATION

Corresponding Authors

*Tel.: +34 948 19 47 00, ext 2044. E-mail: julenoyarzabal@unav.es.
*Tel.: +34 93 41 34 084. E-mail: alfonso.valencia@bsc.es.

ORCID

Martin Krallinger: 0000-0002-2646-8782
Obdulia Rabal: 0000-0002-3224-0987
Anália Lourenço: 0000-0001-8401-5362
Julen Oyarzabal: 0000-0003-1941-7255
Alfonso Valencia: 0000-0002-8937-6789

Author Contributions

○M.K. and O.R. contributed equally to this work.

Notes

The authors declare no competing financial interest.

Biographies

Dr. Martin Krallinger has a M.Sc. degree obtained from the University of Salzburg (Austria) and a doctorate degree from the Universidad Autónoma de Madrid (Spain) with special distinction. Dr. Martin Krallinger has been an honorary visiting professor at the Pompeu Fabra University. He is an expert in the field of biomedical text data mining. Martin Krallinger has been working in this and related research topics for more than 10 years, which resulted in over 60 publications, book chapters, and review papers. He has organized text mining tutorials and lectures at various international events, as well as several Spanish hospitals and universities. Among the specific research questions that he has addressed during his scientific career are biomedical named entity recognition and annotation standards (chemical compounds and mutations), information extraction systems (adverse reactions, protein interactions), semantic annotation of documents with biomedical ontology concepts, text-mining assisted biocuration workflows, interoperability standards and formats for text annotations (BioC), and text-bound annotation infrastructures (MyMiner).

Dr. Obdulia Rabal has a M.Sc. degree in Chemistry (Ramon Llull University, Barcelona) and in Biochemistry (University of Barcelona). In 2006 she obtained a Ph.D. degree in Computational Chemistry from the Ramon Llull University with special distinction. In 2007, she joined the Spanish National Cancer Research Centre (CNIO) where she was involved in different drug discovery projects at the Computational Medicinal Chemistry and Chemoinformatics Section, initially as postdoc and as staff scientist from 2009. In March 2011, she left CNIO and joined the Center for Applied Medical Research (CIMA), at the University of Navarra, as a member of the small molecule discovery platform. Her main research interests are knowledge management, molecular design, and medicinal chemistry.

Dr. Anália Lourenço obtained a Ph.D. degree in Computer Science from the University of Minho (Portugal) in 2007. From 2007 to 2012 she worked as Post-Doc researcher and Research Fellow in the Centre of Biological Engineering at the University of Minho (Portugal), engaging in new lines of research in the fields of Bioinformatics and Systems Biology. In 2012, she joined the faculty of the Department of Computer Science of the University of Vigo (Spain). Her main research interests include biomedical text mining, biological model reconstruction and analysis, biosimulation, and development of large-scale biodata analysis applications.

Dr. Julen Oyarzabal obtained his Ph.D. in Pharmaceutical and Organic Chemistry from Universidad del País Vasco/Euskal Herriko Unibertsitatea in 1998. After his Ph.D., he moved to the University of California San Francisco (CA); and later, he joined the University of Southampton (UK) where he worked in computational chemistry. In November 2001 he joined Johnson & Johnson Pharmaceutical R&D where he led several projects, from molecular design and chemo-informatics perspective, in the CNS therapeutic area. In October 2006, after leaving J&J, he joined the Spanish National Cancer Research Centre (CNIO) where he set up and led the Computational Medicinal Chemistry and Chemoinformatics Section. After 4 years, in September 2010 he left CNIO and joined the Center for Applied Medical Research (CIMA), at the University of Navarra, to set up the small molecule discovery platform and led the Molecular Therapeutics Program. With research interests in chemoinformatics and competitive intelligence tools, the group is running a comprehensive program, from drug discovery informatics to medicinal chemistry and screening, focused on derisking drug discovery. Dr. Oyarzabal is Director of Translational Sciences of the CIMA, on the Board of Directors for the Academic Drug Discovery Consortium (ADDC), and a member of the European Medicines Agency (EMA) experts panel.

Prof. Valencia is Director of the Life Science Department of the Barcelona Supercomputing Centre and Director of Spanish Bio-informatics Institute (INB-ISCIII), the Spanish node of the European Bioinformatics Infrastructure ELIXIR. Valencia is a Computational Biologist interested in the analysis of a large collection of genomic information with particular emphasis on the study of protein interaction networks applied to (epi)genomics, cancer biology, and precision medicine. His group is particularly interested in the application of text mining technology in the area of biology and biomedicine. Alfonso Valencia has published more than 300 articles with a Google Scholar profile h-index of 81, i10 of 252. His group participates in various international consortiums including GEN-CODE/ENCODE, BLUEPRINT/IHEC (epigenomics), RD-Connect/IRDiRC (rare diseases), and CLL/ICGC (cancer genomics). Prof. Valencia is President of the International Society for Computational Biology (ISCB), and an elected member of the European Molecular Biology Organization (EMBO). Prof. Valencia is Executive Editor of the main journal in the field since 2006 ("Bioinformatics" OUP) and Professor *Honoris Causa* of the Danish Technical University - DTU. Alfonso Valencia is a founder and current member of the steering committee of the BioCreative text mining challenge, where he has emphasized particularly the importance of text mining in the connection between molecular biology and chemistry.

ACKNOWLEDGMENTS

A.V. and M.K. acknowledge funding from the European Community's Horizon 2020 Program (project reference: 654021 - OpenMinted). M.K. additionally acknowledges the Encomienda MINETAD-CNIO as part of the Plan for the Advancement of Language Technology. O.R. and J.O. thank the Foundation for Applied Medical Research (FIMA), University of Navarra (Pamplona, Spain). This work was partially funded by Consellería de Cultura, Educación e Ordenación Universitaria (Xunta de Galicia), and FEDER (European Union), and the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 (POCI-01-0145-FEDER-006684). We thank Iñigo García-Yoldi for useful feedback and discussions during the preparation of the manuscript.

ABBREVIATIONS

AB3P	Abbreviation Plus P-Precision

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

ACD	Available Chemicals Directory
ADE	adverse drug event
ADME	absorption, distribution, metabolism, and excretion
ADME-Tox	ADME and toxicology
AI	artificial intelligence
AppFT	USPTO Patent Application Database
aRDI	Access to Research for Development and Innovation
ASCII	American Standard Code for Information Interchange
API	active pharmaceutical ingredient
ATC	automatic text categorization
BASE	Bielefeld Academic Search Engine
BAN	British approved name
BioCreative	Critical Assessment of Information Extraction in Biology
BMJ	British Medical Journal
BOW	bag-of-word
CADD	computer aided drug design
CAF	common application format
CAS	Chemical Abstracts Service
CCD	combined chemical dictionary
CCR	current chemical reactions
CDI	chemical document indexing
CDR	chemical disease relation
CE	chemical entity
CER	chemical entity recognition
CHC	comprehensive heterocyclic chemistry
ChEBI	chemical entities of biological interest
CHEMINF	chemical information ontology
CHEMNDER	chemical compound and drug name recognition
CID	chemical-induced disease
CIHR	Canadian Institutes of Health Research
CIPO	Canadian Intellectual Property Office
CML	chemical markup language
COSTART	coding symbols for a thesaurus of adverse reaction terms
CPCI-S	Conference Proceedings Citation Index-Science
CPD	chemical passage detection
CRF	conditional random field
CSCD	Chinese Science Citation Database
CTD	Comparative Toxicogenomics Database
DARPA	Defense Advanced Research Projects Agency
DBMS	database management systems
DCR	Derwent Chemistry Resource
DDI	drug−drug interaction
DNER	disease named entity recognition
DOAJ	Directory of Open Access Journals
DOCDB	Content of the bibliographic database
DPMA	German Patent and Trade Mark Office
DT	decision tree
DWPI	Derwent World Patents Index
ECKI	canonical keyword indexing
EDGAR	extraction of drugs, genes, and relations
EHR	Electronic Health Record
ELN	Electronic Laboratory Notebooks
EMPathIE	Enzyme and Metabolic Pathways Information Extraction
EPAR	European Public Assessment Reports
EPO	European Patent Office
Europe PMC	Europe PubMed Central
FDA	Food and Drug Administration
FN	false negatives
FP	false positives
FPO	FreePatentsOnline
GPCR	G protein-coupled receptor
GPRO	gene and protein related object
GPSN	Global Patent Search Network
HAC	hierarchical agglomerative clustering
HGNC	HUGO Gene Nomenclature Committee
HMDB	Human Metabolome Database
HMM	hidden Markov model
HTML	HyperText Markup Language
IAA	inter-annotator agreement
ICD-9	International Classification of Diseases, 9th revision
IE	information extraction
InChI	international chemical identifier
INN	international nonproprietary name
INPADOC	international patent documentation
IPC	international patent classification
IR	information retrieval
IUPAC	International Union of Pure and Applied Chemistry
JNLPBA	Joint Workshop on Natural Language Processing in Biomedicine and its Applications
JReX	Jena Relation eXtractor
JSBD	JULIE sentence boundary detector
JSON	JavaScript Object Notation
JSRE	Java Simple Relation Extraction
KEGG	Kyoto Encyclopedia of Genes and Genomes
HIT	herb ingredients' targets
LSI	latent semantic indexing
MAP	mean average precision
MAREC	MAtrixware REsearch Collection
MCC	Matthew's correlation coefficient
MedDRA	Medical Dictionary of Regulatory Activities
MEMM	maximum-entropy Markov model
MeSH	medical subject headings
ML	machine learning
MMS	merged Markush system
MoA	mechanism of action
MQL	molecular query language
MRR	mean reciprocal rank
MUC	message understanding conferences
NCI	National Cancer Institute
NDA	new drug application
NE	named entities
NER	named entity recognition
NERC	named entity recognition and classification
NIAID	National Institute of Allergy and Infectious Diseases
NLM	National Library of Medicine
NLP	natural language processing
NRC-CISTI	National Research Council - Canada Institute for Scientific and Technical Information
OCR	optical character recognition
OCSR	optical chemical structure recognition
OMIM	online Mendelian inheritance in man
OSR	optical structure recognition

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

OWL	web ontology language
PA	prior art
PatFT	USPTO Patent Full-Text and Image Database
PCT	Patent Cooperation Treaty
PDF	portable document format
PDR	physicians' desk reference
PGx	pharmacogenomics
PhV	pharmacovigilance
PMC	PubMed Central
PMI	pointwise mutual information
POS	part-of-speech
PPIs	protein−protein interactions
PQD	ProQuest Dialog
QSAR	quantitative structure activity relationship
RDF	resource description framework
RE	relation extraction
RF	random forest
ROC	receiver operator characteristic
ROSDAL	representation of organic structure description arranged linearly
RSC	Royal Society of Chemistry
RSS	really simple syndication
SACEM	structure associated chemical entity mentions
SAR	structure−activity relationships
SBD	sentence boundary disambiguation
SBS	sentence boundary symbols
SciELO	Scientific Electronic Library Online
SCI-EXPANDED	Science Citation Index Expanded
SDF	structure data file
SERB-CNER	syntactically enhanced rule-based chemical NER
SIDER	side effect resource
SIIP	strategic IP insight platform
SIPO	State Intellectual Property Office of the People's Republic of China
SL	supervised learning
SLN	SYBYL line notation
SMARTS	SMILES arbitrary target specification
SMILES	simplified molecular-input line entry system
SMIRKS	SMIles ReaKtion Specification
SMO	small molecule ontology
SORD	Selected Organic Reactions Database
SPEEDCOP	Spelling Error Detection/Correction Project
SPLICER	Structured Product Label Information Coder and ExtractoR
STITCH	Search Tool for Interacting Chemicals
SVM	support vector machine
TBS	token boundary symbols
TM	text mining
TN	true negatives
TNRS	Taxonomic Name Resolution Service
TP	true positives
TREC	Text REtrieval Conference
TS	technical survey
TTD	Therapeutic Target Database
UL	unsupervised learning
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
USAN	United States Adopted Names
USPTO	United States Patent and Trademark Office
UTF-8	8-bit Unicode Transformation Format
VAE	vaccine-related adverse event
WHO	World Health Organization
WHOINN	World Health Organization International Nonproprietary Name
WIPO	World Intellectual Property Organization
WLN	Wiswesser line notation
WSD	word sense disambiguation
XML	Xtensible Markup Language

REFERENCES

(1) McEwen, L. Taking a Long View: Traverses of 21st Century Chemical Information Stewardship. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; pp 1−18.

(2) Tenopir, C.; King, D. W. Reading Behaviour and Electronic Journals. *Learn. Publ.* **2002**, *15* (4), 259−265.

(3) Roth, D. L. Chapter 2 Non-Patent Primary Literature: Journals, Conference Papers, Reports, Abstracts and Preprints. In *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2014; pp 29−52.

(4) Martin, E.; Monge, A.; Peitsch, M. C.; Pospisil, P. Chapter 4. Building a Corporate Chemical Database towards Systems Biology. In *Data Mining in Drug Discovery*; Hoffmann, R. D., Gohier, A., Pospisil, P., Eds.; Wiley-VCH Verlag: Weinheim, Germany, 2014; pp 75−98.

(5) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *J. Med. Chem.* **2016**, *59* (9), 4385−4402.

(6) Pope, Y. CAS RegistrySM: The Quality of Comprehensiveness Is Not Strained. Presented at the EMBL-EBI Industry Programme Workshop; Chemical Structure Resources: Hinxton, Cambridge, December 1, 2010.

(7) Schenck, R.; Zabilski, J. *CAS Registry: Maintaining the Gold Standard for Chemical Substance Information*. Presented at the 241st National Meeting & Exposition of the American Chemical Society, Anaheim, CA, March 27−31, 2011; paper CINF-56.

(8) Espacenet Patent search. http://worldwide.espacenet.com (accessed Oct 20, 2016).

(9) Alberts, D.; Yang, C. B.; Fobare-DePonio, D.; Koubek, K.; Robins, S.; Rodgers, M.; Simmons, E.; DeMarco, D. Introduction to Patent Searching. In *Current Challenges in Patent Information Retrieval*; Lupu, M., Mayer, K., Tait, J., Trippe, A. J., Eds.; Springer-Verlag: Berlin Heidelberg, 2011; pp 3−43.

(10) Southan, C. Expanding Opportunities for Mining Bioactive Chemistry from Patents. *Drug Discovery Today: Technol.* **2015**, *14*, 3−9.

(11) Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, 2008.

(12) Currano, J. N. Teaching Chemical Information for the Future: The More Things Change, the More They Stay the Same. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; pp 169−196.

(13) Gilson, M. K.; Georg, G.; Wang, S. Digital Chemistry in the Journal of Medicinal Chemistry. *J. Med. Chem.* **2014**, *57* (4), 1137.

(14) Hoffmann, R.; Krallinger, M.; Andres, E.; Tamames, J.; Blaschke, C.; Valencia, A. Text Mining for Metabolic Pathways, Signaling Cascades, and Protein Networks. *Sci. Signaling* **2005**, *2005* (283), pe21.

(15) Krallinger, M.; Leitner, F.; Valencia, A. Analysis of Biological Processes and Diseases Using Text Mining Approaches. *Methods Mol. Biol.* **2010**, *593*, 341−382.

(16) Vazquez, M.; Krallinger, M.; Leitner, F.; Valencia, A. Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. *Mol. Inf.* **2011**, *30* (6−7), 506−519.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(17) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to Improve R&D Productivity: The Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203−214.
(18) Garnier, J.-P. Rebuilding the R&D Engine in Big Pharma. *Harv. Bus. Rev.* **2008**, *86*, 68−70 , 72−76, 128.
(19) Williams, M. Productivity Shortfalls in Drug Discovery: Contributions from the Preclinical Sciences? *J. Pharmacol. Exp. Ther.* **2011**, *336* (1), 3−8.
(20) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31−36.
(21) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Model.* **1989**, *29* (2), 97−101.
(22) McNaught, A. The Iupac International Chemical Identifier. *Chem. Int.* **2006**, *28*, 12−14.
(23) IUPAC. The IUPAC International Chemical Identifier. http://old.iupac.org/inchi/release102.html (accessed Oct 20, 2016).
(24) Fennell, R. W. *History of IUPAC, 1919−1987*; Blackwell Science: Oxford, UK, 1994.
(25) Weisgerber, D. W. Chemical Abstracts Service Chemical Registry System: History, Scope, and Impacts. *J. Am. Soc. Inf. Sci.* **1997**, *48* (4), 349−360.
(26) Kremer, G.; Anstein, S.; Reyle, U. Analysing and Classifying Names of Chemical Compounds with Chemorph. *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006)*; Jena, Germany, April 9−12, **2006**; pp 37−43.
(27) IUPAC Compendium of Chemical Terminology - the Gold Book. http://goldbook.iupac.org (accessed Oct 20, 2016).
(28) American Chemical Society. Chemical Abstracts Service. *Naming and Indexing of Chemical Substances for Chemical Abstracts: A Reprint of Appendix IV (Chemical Substance Index Names) from the Chemical Abstracts 1997 Index Guide*; American Chemical Society: Columbus, OH, 1997.
(29) Luckenbach, R. The Beilstein Handbook of Organic Chemistry: The First Hundred Years. *J. Chem. Inf. Model.* **1981**, *21* (2), 82−83.
(30) Eller, G. A. Improving the Quality of Published Chemical Names with Nomenclature Software. *Molecules* **2006**, *11* (11), 915−928.
(31) Workman, M. L.; LaCharity, L. A. *Understanding Pharmacology: Essentials for Medication Safety*, 2nd ed.; Elsevier Health Sciences: St. Louis, MO, 2016.
(32) United States Adopted Names Council (USAN). https://www.ama-assn.org/about/united-states-adopted-names-council (accessed Oct 30, 2016).
(33) WHO. Guidance on INN. http://www.whoint/medicines/services/inn/innquidance/en (accessed Oct 30, 2016).
(34) Gopakumar, K. M.; Syam, N. International Nonproprietary Names and Trademarks: A Public Health Perspective. *J. World Intellect. Prop.* **2008**, *11* (2), 63−104.
(35) Gurulingappa, H.; Mudi, A.; Toldo, L.; Hofmann-Apitius, M.; Bhate, J. Challenges in Mining the Literature for Chemical Information. *RSC Adv.* **2013**, *3*, 16194−16211.
(36) Wilbur, W. J.; Hazard, G. F., Jr; Divita, G.; Mork, J. G.; Aronson, A. R.; Browne, A. C. Analysis of Biomedical Text for Chemical Names: A Comparison of Three Methods. *Proc. AMIA Symp.* **1999**, 176−180.
(37) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Model.* **1991**, *31* (2), 233−253.
(38) Jessop, D. M.; Adams, S. E.; Murray-Rust, P. Mining Chemical Information from Open Patents. *J. Cheminf.* **2011**, *3* (1), 40.
(39) Ellegaard, O.; Wallin, J. A. Identification of Environmentally Relevant Chemicals in Bibliographic Databases: A Comparative Analysis. *SpringerPlus* **2013**, *2* (1), 255.
(40) PubMed. http://www.ncbi.nlm.nih.gov/pubmed (accessed Oct 20, 2016).
(41) Agarwala, R.; Barrett, T.; Beck, J.; Benson, D. A.; Bollin, C.; Bolton, E.; Bourexis, D.; Brister, J.; Bryant, S. H.; Canese, K.; et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2015**, *43*, D6−17 (Database issue).
(42) MEDLINE. https://www.nlm.nih.gov/pubs/factsheets/medline.html (accessed Oct 20, 2016).
(43) MEDLINE, PubMed, and PMC (PubMed Central): How are they different? http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html (accessed Oct 20, 2016).
(44) SciFinder. http://www.cas.org/products/scifinder (accessed Oct 20, 2016).
(45) Wagner, A. Ben. SciFinder Scholar 2006: An Empirical Analysis of Research Topic Query Processing. *J. Chem. Inf. Model.* **2006**, *46* (2), 767−774.
(46) CAPLUS Contents. http://www.cas.org/content/references (accessed Oct 20, 2016).
(47) Embase. http://www.elsevier.com/solutions/embase (accessed Oct 20, 2016).
(48) BIOSIS Previews. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/biosis-previews.html (accessed Oct 20, 2016).
(49) TOXLINE. https://toxnet.nlm.nih.gov/newtoxnet/toxline.htm (accessed Oct 20, 2016).
(50) INSPEC. http://www.theiet.org/resources/inspec (accessed Oct 20, 2016).
(51) Scopus. http://www.elsevier.com/solutions/scopus (accessed Oct 20, 2016).
(52) Directory of Open Access Journals (DOAJ). https://doaj.org (accessed Oct 20, 2016).
(53) Woods, D.; Trewheellar, K. Medline and Embase Complement Each Other in Literature Searches. *Br. Med. J.* **1998**, *316* (7138), 1166−1166.
(54) Dunikowski, L. G. EMBASE and MEDLINE Searches. *Can. Fam. Physician* **2005**, *51*, 1191−1191.
(55) Kelly, L.; St Pierre-Hansen, N. So Many Databases, Such Little Clarity: Searching the Literature for the Topic Aboriginal. *Can. Fam. Physician* **2008**, *54* (11), 1572−1573.
(56) Mesgarpour, B.; Müller, M.; Herkner, H. Search Strategies to Identify Reports On "off-Label" drug Use in EMBASE. *BMC Med. Res. Methodol.* **2012**, *12*, 190.
(57) Web of Science. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/web-of-science.html (accessed Oct 20, 2016).
(58) Science Citation Index Expanded. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/science-citation-index-expanded.html (accessed Oct 20, 2016).
(59) PMC Pubmed Central. http://www.ncbi.nlm.nih.gov/pmc (accessed Oct 20, 2016).
(60) Sequeira, E. PubMed Central-Three Years Old and Growing Stronger. *ARL* **2003**, *228*, 5−9.
(61) Science Direct. http://www.sciencedirect.com (accessed Oct 20, 2016).
(62) Chinese Science Citation Database. http://english.las.cas.cn/ (accessed Oct 20, 2016).
(63) Jin, B.; Wang, B. Chinese Science Citation Database: Its Construction and Application. *Scientometrics* **1999**, *45* (2), 325−332.
(64) Scientific Electronic Library Online (SciELO). http://www.scielo.org (accessed Oct 20, 2016).
(65) Conference Proceedings Citation Index-Science. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/conference-proceedings-citation-index.html (accessed Oct 20, 2016).
(66) IP.com Prior Art Database. https://priorart.ip.com (accessed Oct 20, 2016).
(67) USPTO Patent Full-Text and Image Database. http://patft.uspto.gov (accessed Oct 20, 2016).
(68) USPTO Patent Application Full-Text and Image Database. http://appft.uspto.gov (accessed Oct 20, 2016).

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(69) Content of the bibliographic database (DOCDB). http://www.epo.org/searching/subscription/raw/product-14-7.html (accessed Oct 20, 2016).
(70) INPADOC. http://www.epo.org/searching/subscription/raw/product-14-11.html (accessed Oct 20, 2016).
(71) Derwent World Patents Index. http://thomsonreuters.com/en/products-services/intellectual-property/patent-research-and-analysis/derwent-world-patents-index.html (accessed Oct 20, 2016).
(72) FamPat Database Coverage and Update. https://questel.com/index.php/en/2012-11-20-10-09-15/fampat (accessed Oct 20, 2016).
(73) PatBase. http://www.patbase.com (accessed Oct 20, 2016).
(74) SureChEMBL Open Patent Data. https://www.surechembl.org (accessed Oct 20, 2016).
(75) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; et al. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44* (D1), D1220−D1228.
(76) IFI Claims Patent Services. http://www.ificlaims.com (accessed Oct 20, 2016).
(77) DAILYMED. http://dailymed.nlm.nih.gov/dailymed (accessed Oct 20, 2016).
(78) New Drug Application (NDA). http://www.fda.gov/Drugs/DevelopmentApprovalProcess/HowDrugsareDevelopedandApproved/ApprovalApplications/NewDrugApplicationNDA (accessed Oct 26, 2010).
(79) European public assessment reports. http://www.ema.europa.eu/ema/index.jsp?curl=pages/medicines/landing/epar_search.jsp&mid=WC0b01ac058001d125 (accessed Oct 20, 2016).
(80) Clinical Trials.gov. https://clinicaltrials.gov (accessed Oct 20, 2016).
(81) Adis Insight. http://adisinsight.springer.com (accessed Oct 20, 2016).
(82) Reaxys: The Quickest Path from Q to A. http://www.elsevier.com/solutions/reaxys (accessed Oct 20, 2016).
(83) Lawson, A. J.; Swienty-Busch, J.; Geoui, T.; Evans, D. The Making of Reaxys-Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; pp 127−148.
(84) STN - The choice of patent experts. https://www.cas.org/products/stn (accessed Oct 20, 2016).
(85) iScienceSearch. http://isciencesearch.com/iss/default.aspx (accessed Oct 20, 2016).
(86) Kos, A.; Himmler, H.-J. CWM Global Search—The Internet Search Engine for Chemists and Biologists. *Future Internet* **2010**, *2* (4), 635−644.
(87) Google Patents. https://patents.google.com (accessed Oct 20, 2016).
(88) Scopus vs Web of Science. http://hlwiki.slais.ubc.ca/index.php/Scopus_vs._Web_of_Science (accessed Oct 20, 2016).
(89) Thomson Innovation. http://info.thomsoninnovation.com (accessed Oct 20, 2016).
(90) Europe PMC. http://europepmc.org (accessed Oct 20, 2016).
(91) Gou, Y.; Graff, F.; Kilian, O.; Kafkas, S.; Katuri, J.; Kim, J. H.; Marinos, N.; McEntyre, J.; Morrison, A.; Pi, X.; et al. Europe PMC: A Full-Text Literature Database for the Life Sciences and Platform for Innovation. *Nucleic Acids Res.* **2015**, *43*, D1042−1048.
(92) Google Scholar. http://scholar.google.com (accessed Oct 20, 2016).
(93) Butler, D. Science Searches Shift up a Gear as Google Starts Scholar Engine. *Nature* **2004**, *432* (7016), 423.
(94) Mayr, P.; Walter, A.-K. Studying Journal Coverage in Google Scholar. *J. Libr. Adm.* **2008**, *47* (1−2), 81−99.
(95) Access to Research for Development and Innovation ARDI. http://www.wipo.int/ardi/en (accessed Oct 20, 2016).
(96) Free Patents Online FPO. http://www.freepatentsonline.com (accessed Oct 20, 2016).
(97) Lens. https://www.lens.org/lens (accessed Oct 20, 2016).
(98) Prior Smart. http://www.priorsmart.com (accessed Oct 20, 2016).
(99) Elsevier Scirus. http://www.sciencedirect.com/scirus (accessed Oct 20, 2016).
(100) The decline and fall of Microsoft Academic Search. http://blogs.nature.com/news/2014/05/the-decline-and-fall-of-microsoft-academic-search.html (accessed Oct 20, 2016).
(101) Mooers, C. E. Coding, Information Retrieval, and the Rapid Selector. *Am. Doc.* **1950**, *1* (4), 225−229.
(102) Sanderson, M.; Croft, W. B. The History of Information Retrieval Research. *Proc. IEEE* **2012**, *100* (13), 1444−1451.
(103) PostgreSQL. http://www.postgresql.org.es (accessed Oct 20, 2016).
(104) The RDKit database cartridge. http://www.rdkit.org/docs/Cartridge.html (accessed Oct 20, 2016).
(105) Apache Lucene. http://lucene.apache.org (accessed Oct 20, 2016).
(106) Elasticsearch. https://www.elastic.co (accessed Oct 20, 2016).
(107) Baykoucheva, S. Selecting a Database for Drug Literature Retrieval: A Comparison of MEDLINE, Scopus, and Web of Science. *Sci. Technol. Libr.* **2010**, *29* (4), 276−288.
(108) Gurulingappa, H.; Toldo, L.; Rajput, A. M.; Kors, J. A.; Taweel, A.; Tayrouz, Y. Automatic Detection of Adverse Events to Predict Drug Label Changes Using Text and Data Mining Techniques. *Pharmacoepidemiol. Drug Saf.* **2013**, *22* (11), 1189−1194.
(109) PDFBox. https://pdfbox.apache.org (accessed Oct 20, 2016).
(110) IntraPDF. http://www.intrapdf.com (accessed Oct 20, 2016).
(111) PDFTron. https://www.pdftron.com (accessed Oct 20, 2016).
(112) Attwood, T. K.; Kell, D. B.; McDermott, P.; Marsh, J.; Pettifer, S. R.; Thorne, D. Utopia Documents: Linking Scholarly Literature with Research Data. *Bioinformatics* **2010**, *26* (18), i568−74.
(113) ABBYY PDF Transformer. https://www.abbyy.com/pdf-transformer (accessed Oct 20, 2016).
(114) Ramakrishnan, C.; Patnia, A.; Hovy, E.; Burns, G. A. Layout-Aware Text Extraction from Full-Text PDF of Scientific Articles. *Source Code Biol. Med.* **2012**, *7* (1), 7.
(115) Tesseract. https://github.com/tesseract-ocr (accessed Oct 20, 2016).
(116) Banville, D. L. Mining Chemical Structural Information from the Drug Literature. *Drug Discovery Today* **2006**, *11* (1−2), 35−42.
(117) Brecher, J. Name = struct: A Practical Approach to the Sorry State of Real-Life Chemical Nomenclature. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 943−950.
(118) Shidha, M. V; Mahalekshmi, T. Chem Text Mining - An Outline. *Int. J. Emerg. Technol. Adv. Eng.* **2014**, *4*, 377−386.
(119) Sayle, R.; Xie, P. H.; Muresan, S. Improved Chemical Text Mining of Patents with Infinite Dictionaries and Automatic Spelling Correction. *J. Chem. Inf. Model.* **2012**, *52* (1), 51−62.
(120) Mattmann, C.; Zitting, J. *Tika in Action*; Manning Publications: Greenwich, CT, 2011.
(121) Rusch, P. F. Introduction to Chemical Information Storage and Retrieval. *J. Chem. Educ.* **1981**, *58* (4), 337−342.
(122) Mackenzie, C. E. *Coded-Character Sets: History and Development*; Addison-Wesley Pub: Boston, MA, 1980.
(123) Davis, M. Unicode nearing 50% of the web. https://googleblog.blogspot.com.es/2010/01/unicode-nearing-50-of-web.html (accessed Oct 26, 2010).
(124) Ibison, P.; Jacquot, M.; Kam, F.; Neville, A. G.; Simpson, R. W.; Tonnelier, C.; Venczel, T.; Johnson, A. P. Chemical Literature Data Extraction: The CLiDE Project. *J. Chem. Inf. Model.* **1993**, *33* (3), 338−344.
(125) Valko, A. T.; Johnson, A. P. CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition. *J. Chem. Inf. Model.* **2009**, *49* (4), 780−787.
(126) Rupp, C. J.; Copestake, A. A.; Corbett, P. T.; Murray-Rust, P.; Siddharthan, A.; Teufel, S.; Waldron, B. Language Resources and Chemical Informatics. *Proceedings of the Sixth International Conference*

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

on Language Resources and Evaluation (LREC 2008); Marrakech, Morocco, May 28−30, 2008; pp 2196−2200.
(127) Teufel, S.; Moens, M. Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Comput. Ling.* **2002**, *28* (4), 409−445.
(128) Read, J.; Dridan, R.; Oepen, S.; Solberg, L. J. Sentence Boundary Detection: A Long Solved Problem? *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*; Mumbai, India, December 8−15, 2012; pp 985−994.
(129) Täger, W. The Sentence-Aligned European Patent Corpus. *Proceedings of the 15th Conference of the European Association for Machine Translation (EAMT 2011)*; Leuven, Belgium, May 30−31, 2011; pp 177−184.
(130) Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. Automatic Extraction of Rules for Sentence Boundary Disambiguation. *Proceedings of the Workshop on Machine Learning in Human Language Technology, Advanced Course on Artificial Intelligence (ACAI'99)*; Chania, Greece, July 5−16, 1999; pp 88−92.
(131) Zhou, W.; Torvik, V. I.; Smalheiser, N. R. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics* **2006**, *22* (22), 2813−2818.
(132) Krallinger, M.; Leitner, F.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Valencia, A. CHEMDNER: The Drugs and Chemical Names Extraction Challenge. *J. Cheminf.* **2015**, *7*, S1 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(133) Kim, J.-D.; Ohta, T.; Tateisi, Y.; Tsujii, J. GENIA Corpus−Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics* **2003**, *19* (Suppl 1), i180−2.
(134) Bies, A.; Kulick, S.; Mandel, M. Parallel Entity and Treebank Annotation. *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*; Association for Computational Linguistics: Ann Arbor, MI, June 29, 2005; pp 21−28.
(135) Tomanek, K.; Wermter, J.; Hahn, U. A Reappraisal of Sentence and Token Splitting for Life Sciences Documents. *Stud. Health Technol. Inform.* **2007**, *129*, 524−528.
(136) Carpenter, B. Phrasal Queries with LingPipe and Lucene: Ad Hoc Genomics Text Retrieval. *Proceedings of the 13th Text Retrieval Conference (TREC 2004)*; Gaithersburg, MD, November 16−19, 2004.
(137) GENIA Sentence Splitter. http://www.nactem.ac.uk/y-matsu/geniass (accessed Oct 26, 2016).
(138) Sætre, R.; Yoshida, K.; Yakushiji, A.; Miyao, Y.; Matsubayashi, Y.; Ohta, T. AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS Subtask. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*; Madrid, Spain, April 23−25, 2007; pp 209−212.
(139) Smith, L.; Rindflesch, T.; Wilbur, W. J. MedPost: A Part-of-Speech Tagger for bioMedical Text. *Bioinformatics* **2004**, *20* (14), 2320−2321.
(140) Tomanek, K.; Wermter, J.; Hahn, U. Sentence and Token Splitting Based on Conditional Random Fields. *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*; Melbourne, Australia, September 19−21, 2007; pp 49−57.
(141) Klinger, R.; Kolárik, C.; Fluck, J.; Hofmann-Apitius, M.; Friedrich, C. M. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics* **2008**, *24* (13), i268−76.
(142) OpenNLP. http://opennlp.apache.org/ (accessed Oct 26, 2016).
(143) Lai, H.; Xu, S.; Zhu, L. Chemical and Biological Entity Recognition System from Patent Documents. *Proceedings of the Second International Workshop on Patent Mining and Its Applications (IPaMin 2015)*; Beijing, China, May 27−28, 2015.
(144) Jonnalagadda, S.; Gonzalez, G. Sentence Simplification Aids Protein-Protein Interaction Extraction. *Proceedings of the 3rd International Symposium on Languages in Biology and Medicine (LBM 2009)*; Jeju Island, South Korea, November 8−10, 2009; pp 109−114.
(145) Miwa, M.; Sætre, R.; Miyao, Y.; Tsujii, J. Entity-Focused Sentence Simplification for Relation Extraction. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*; Beijing, China, August 23−27, 2010; pp 788−796.
(146) Bui, Q.-C.; Nualláin, B. O.; Boucher, C. A.; Sloot, P. M. A. Extracting Causal Relations on HIV Drug Resistance from Literature. *BMC Bioinf.* **2010**, *11*, 101.
(147) Jonnalagadda, S.; Gonzalez, G. BioSimplify: An Open Source Sentence Simplification Engine to Improve Recall in Automatic Biomedical Information Extraction. *AMIA Annu. Symp. Proc.* **2010**, *2010*, 351−355.
(148) Peng, Y.; Tudor, C. O.; Torii, M.; Wu, C. H.; Vijay-Shanker, K. iSimp in BioC Standard Format: Enhancing the Interoperability of a Sentence Simplification System. *Database* **2014**, *2014*, bau038.
(149) Cafetiere English Sentence Detector. http://metashare.metanet4u.eu/repository/browse/u-compare-cafetiere-english-sentence-detector/aff1ddc0bc8911e1a404080027e73ea259aeca28412944ea97f7b2580a41caec (accessed Oct 26, 2016).
(150) Batista-Navarro, R.; Rak, R.; Ananiadou, S. Optimising Chemical Named Entity Recognition with Pre-Processing Analytics, Knowledge-Rich Features and Heuristics. *J. Cheminf.* **2015**, *7*, S6.
(151) Appelt, D. E.; Hobbs, J. R.; Bear, J.; Israel, D.; Kameyama, M.; Martin, D.; Myers, K.; Tyson, M. SRI International FASTUS System: MUC-6 Test Results and Analysis. *Proceedings of the 6th Conference on Message Understanding (MUC-6)*; Columbia, MD, November, 1995; pp 237−248.
(152) Xue, N. Chinese Word Segmentation as Character Tagging. *Comput. Linguist. Chinese Lang. Process.* **2003**, *8*, 29−48.
(153) He, Y.; Kayaalp, M. *A Comparison of 13 Tokenizers on MEDLINE*; Technical Report LHNCBC-TR-2006-003; The Lister Hill National Center for Biomedical Communications: Bethesda, MD, December 2006.
(154) Corbett, P.; Murray-Rust, P. High-Throughput Identification of Chemistry in Life Science Texts. In *Computational Life Sciences II. Volume 4216 of the Series Lecture Notes in Computer Science*; Berthold, M. R., Glen, R. C., Fischer, I., Eds.; Springer Berlin Heidelberg: Cambridge, UK, 2006; pp 107−118.
(155) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A Tool for Semantic Text-Mining in Chemistry. *J. Cheminf.* **2011**, *3* (1), 17.
(156) Zamora, E. M.; Blower, P. E., Jr Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 1. Lexical and Syntactic Phases. *J. Chem. Inf. Model.* **1984**, *24* (3), 176−181.
(157) Barrett, N.; Weber-Jahnke, J. Building a Biomedical Tokenizer Using the Token Lattice Design Pattern and the Adapted Viterbi Algorithm. *BMC Bioinf.* **2011**, *12* (Suppl 3), S1.
(158) Jessop, D. M.; Adams, S. E.; Willighagen, E. L.; Hawizy, L.; Murray-Rust, P. OSCAR4: A Flexible Architecture for Chemical Text-Mining. *J. Cheminf.* **2011**, *3* (1), 41.
(159) Akkasi, A.; Varoğlu, E.; Dimililer, N. ChemTok: A New Rule Based Tokenizer for Chemical Named Entity Recognition. *BioMed Res. Int.* **2016**, *2016*, 4248026.
(160) Dai, H. J.; Lai, P. T.; Chang, Y. C.; Tsai, R. T. Enhancing of Chemical Compound and Drug Name Recognition Using Representative Tag Scheme and Fine-Grained Tokenization. *J. Cheminf.* **2015**, *7*, S14 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(161) Leaman, R.; Wei, C.-H.; Lu, Z. tmChem: A High Performance Approach for Chemical Named Entity Recognition and Normalization. *J. Cheminf.* **2015**, *7*, S3 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(162) Rocktäschel, T.; Weidlich, M.; Leser, U. ChemSpot: A Hybrid System for Chemical Named Entity Recognition. *Bioinformatics* **2012**, *28* (12), 1633−1640.
(163) Bambenek, J.; Klus, A. *Grep Pocket Reference*; O'Reilly Media: Sebastopol, CA, 2009.
(164) Zipf, G. K. *Selected Studies of the Principle of Relative Frequency in Language*; Harvard University Press: Cambridge, MA, 1932.
(165) Zipf, G. K. *Human Behavior and the Principle of Least Effort*; Addison-Wesley Press: Cambridge, MA, 1949.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(166) Kalankesh, L.; New, J. P.; Baker, P. G.; Brass, A. The Languages of Health in General Practice Electronic Patient Records: A Zipf's Law Analysis. *J. Biomed. Semant.* **2014**, *5* (1), 2.
(167) Rebholz-Schuhmann, D.; Kirsch, H.; Couto, F. Facts from Text–Is Text Mining Ready to Deliver? *PLoS Biol.* **2005**, *3* (2), e65.
(168) Kornai, A. *Mathematical Linguistics*; Springer-Verlag: Berlin, Germany, 2008.
(169) Kraaij, W.; Pohlmann, R. Viewing Stemming as Recall Enhancement. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'96)*; Zurich, Switzerland, August 18–22, 1996; pp 40–48.
(170) Lovins, J. B. Development of a Stemming Algorithm. *Mech. Transl. Comput. Linguist.* **1968**, *11*, 22–31.
(171) Porter, M. F. An Algorithm for Suffix Stripping. *Program* **1980**, *14* (3), 211–218.
(172) Lamurias, A.; Ferreira, J. D.; Couto, F. M. Improving Chemical Entity Recognition through H-Index Based Semantic Similarity. *J. Cheminf.* **2015**, *7*, S13 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(173) Andrade, M. A.; Valencia, A. Automatic Extraction of Keywords from Scientific Text: Application to the Knowledge Domain of Protein Families. *Bioinformatics* **1998**, *14* (7), 600–607.
(174) GENIA tagger: Part-of-speech tagging, shallow parsing, and named entity recognition for biomedical text. http://www.nactem.ac.uk/GENIA/tagger/ (accessed Oct 31, 2016).
(175) Huber, T.; Rocktäschel, T.; Weidlich, M.; Thomas, P.; Leser, U. Extended Feature Set for Chemical Named Entity Recognition and Indexing. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop Vol. 2*; Bethesda, MD, October 7–9, 2013; pp 88–91.
(176) Liu, H.; Christiansen, T.; Baumgartner, W. A.; Verspoor, K. BioLemmatizer: A Lemmatization Tool for Morphological Processing of Biomedical Text. *J. Biomed. Semant.* **2012**, *3*, 3.
(177) Witten, I. H.; Moffat, A.; Bell, T. C. In *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd ed.; Fox, E., Ed.; Morgan Kaufmann Publishers: San Diego, CA, 1999.
(178) Salton, G.; Wong, A.; Yang, C.-S. A Vector Space Model for Automatic Indexing. *Commun. ACM* **1975**, *18* (11), 613–620.
(179) Salton, G. In *The SMART Retrieval System--Experiments in Automatic Document Processing*; Salton, G., Ed.; Prentice-Hall: Upper Saddle River, NJ, 1971.
(180) Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval: The Concepts and Technology behind Search*; Addison-Wesley Longman Publishing: Boston, MA, 1999.
(181) Singhal, A. Modern Information Retrieval: A Brief Overview. *Bull. IEEE Comput. Soc. Technical Committee Data Eng.* **2001**, *24*, 35–42.
(182) Schapire, R. E.; Singer, Y.; Singhal, A. Boosting and Rocchio Applied to Text Filtering. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*; Melbourne, Australia, August 24–28, 1998; pp 215–223.
(183) ChemXSeer Official web site. http://chemxseer.ist.psu.edu (accessed Jan 24, 2016).
(184) Li, N.; Zhu, L.; Mitra, P.; Mueller, K.; Poweleit, E.; Giles, C. L. oreChem ChemXSeer: A Semantic Digital Library for Chemistry. *Proceedings of the 10th annual joint conference on Digital libraries (JCDL 2010)*; Gold Coast, Australia, June 21–25, 2010; pp 245–254.
(185) Hersh, W. *Information Retrieval: A Health and Biomedical Perspective*, 3rd ed.; Springer: New York, 2009.
(186) Jackson, P.; Moulinier, I. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*, 2nd ed.; John Benjamins Publishing: Philadelphia, PA, 2007.
(187) Kim, Y.; Seo, J.; Croft, W. B. Automatic Boolean Query Suggestion for Professional Search. *Proceedings of the 34th international ACM SIGIR conference on Research and development in information retrieval (SIGIR'11)*; Beijing, China, July 24–28, 2011; pp 825–834.
(188) Zhou, Y.; Zhou, B.; Jiang, S.; King, F. J. Chemical-Text Hybrid Search Engines. *J. Chem. Inf. Model.* **2010**, *50* (1), 47–54.
(189) Medical Subject Headings. https://www.nlm.nih.gov/mesh (accessed Oct 20, 2016).
(190) Sewell, W. Medical Subject Headings in MEDLARS. *Bull. Med. Libr. Assoc.* **1964**, *52*, 164–170.
(191) NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2013**, *41*, D8–D20 (Database issue).
(192) Ide, N. C.; Loane, R. F.; Demner-Fushman, D. Essie: A Concept-Based Search Engine for Structured Biomedical Text. *J. Am. Med. Inf. Assoc.* **2007**, *14* (3), 253–263.
(193) Salton, G.; Fox, E. A.; Wu, H. Extended Boolean Information Retrieval. *Commun. ACM* **1983**, *26* (11), 1022–1036.
(194) Cochrane Library. http://www.cochranelibrary.com (accessed Oct 20, 2016).
(195) Exalead. https://www.exalead.com/search (accessed Oct 20, 2016).
(196) Yandex. https://www.yandex.com (accessed Oct 20, 2016).
(197) Bing. https://www.bing.com (accessed Oct 20, 2016).
(198) Google. https://www.google.com (accessed Oct 20, 2016).
(199) STN Pocket Guide. http://www.cas.org/training/stn/stn-pocket-guide (accessed Oct 20, 2016).
(200) Luhn, H. P. Key Word-in-Context Index for Technical Literature (Kwic Index). *Am. Doc.* **1960**, *11* (4), 288–295.
(201) Comer, D. Ubiquitous B-Tree. *ACM Comput. Surv.* **1979**, *11* (2), 121–137.
(202) Bast, H.; Mortensen, C. W.; Weber, I. Output-Sensitive Autocompletion Search. *Proceedings of the 13th international conference on String Processing and Information Retrieval (SPIRE'06)*; Glasgow, UK, October 11–13, 2006; pp 150–162.
(203) Wang, P.; Berry, M. W.; Yang, Y. Mining Longitudinal Web Queries: Trends and Patterns. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54* (8), 743–758.
(204) Brill, E.; Moore, R. C. An Improved Error Model for Noisy Channel Spelling Correction. *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00)*; Hong Kong, October 1–8, 2000; pp 286–293.
(205) Damerau, F. J. A Technique for Computer Detection and Correction of Spelling Errors. *Commun. ACM* **1964**, *7* (3), 171–176.
(206) Levenshtein, V. I. Binary Codes with Correction for Deletions and Insertions of the Symbol 1. *Probl. Peredachi Inf.* **1965**, *1*, 12–25.
(207) Gusfield, D. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*, 1st ed.; Cambridge University Press: New York, 1997.
(208) Hall, P. A. V; Dowling, G. R. Approximate String Matching. *ACM Comput. Surv.* **1980**, *12* (4), 381–402.
(209) Mitton, R. *English Spelling and the Computer*; Longman Group: London, 1996.
(210) Kukich, K. Techniques for Automatically Correcting Words in Text. *ACM Comput. Surv.* **1992**, *24* (4), 377–439.
(211) Pollock, J. J.; Zamora, A. Automatic Spelling Correction in Scientific and Scholarly Text. *Commun. ACM* **1984**, *27* (4), 358–368.
(212) New PubMed Spell Checking Feature. https://www.nlm.nih.gov/pubs/techbull/nd04/nd04_spell.html (accessed Oct 20, 2016).
(213) Norton, T. H. The Spelling and Pronunciation of Chemical Terms. *Science* **1892**, *20* (510), 272–274.
(214) Davis, C. H.; Rush, J. E. *Information Retrieval and Documentation in Chemistry*; Greenwood Press: Westport, CT, 1974.
(215) Pollock, J. J.; Zamora, A. Collection and Characterization of Spelling Errors in Scientific and Scholarly Text. *J. Am. Soc. Inf. Sci.* **1983**, *34* (1), 51–58.
(216) Kirby, G. H.; Lord, M. R.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 6. (Semi)-automatic Name Correction. *J. Chem. Inf. Model.* **1991**, *31* (1), 153–160.
(217) *A Guide to Iupac Nomenclature of Organic Compounds Recommendations 1993 (International Union of Pure and Applied Chemistry Organic Chemistry Division)*, 2nd ed.; Panico, R., Powell, W. H., Eds.; Blackwell Scientific Publications: Oxford, UK, 1994.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673–7761

(218) Azman, A. M. A Chemistry Spell-Check Dictionary for Word Processors. *J. Chem. Educ.* **2012**, *89* (3), 412−413.
(219) ChemSpell. http://chemspell.nlm.nih.gov/spell (accessed Oct 20, 2016).
(220) ChemIDplus. http://www.chem.sis.nlm.nih.gov/chemidplus (accessed Oct 20, 2016).
(221) *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2014.
(222) Johnson, D.; Malhotra, V.; Vamplew, P. More Effective Web Search Using Bigrams and Trigrams. *Webology* **2006**, *3*, 35.
(223) Eisinger, D.; Tsatsaronis, G.; Bundschus, M.; Wieneke, U.; Schroeder, M. Automated Patent Categorization and Guided Patent Search Using IPC as Inspired by MeSH and PubMed. *J. Biomed. Semant.* **2013**, *4* (Suppl 1), S3.
(224) Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34* (1), 1−47.
(225) Maron, M. E. Automatic Indexing: An Experimental Inquiry. *J. Assoc. Comput. Mach.* **1961**, *8* (3), 404−417.
(226) Kazawa, H.; Izumitani, T.; Taira, H.; Maeda, E. Maximal Margin Labeling for Multi-Topic Text Categorization. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*; Saul, L. K., Weiss, Y., Bottou, L., Eds.; MIT Press: Cambridge, MA, 2005; pp 649−656.
(227) Stamatatos, E.; Fakotakis, N.; Kokkinakis, G. Automatic Text Categorization in Terms of Genre and Author. *Comput. Ling.* **2000**, *26* (4), 471−495.
(228) Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up?: Sentiment Classification Using Machine Learning Techniques. *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10 (EMNLP'02)*; Philadelphia, PA, July 6−7, 2002; pp 79−86.
(229) Krallinger, M.; Vazquez, M.; Leitner, F.; Salgado, D.; Chatr-Aryamontri, A.; Winter, A.; Perfetto, L.; Briganti, L.; Licata, L.; Iannuccelli, M.; et al. The Protein-Protein Interaction Tasks of BioCreative III: Classification/ranking of Articles and Linking Bio-Ontology Concepts to Full Text. *BMC Bioinf.* **2011**, *12* (Suppl8), S3.
(230) Hsueh, P.-Y.; Melville, P.; Sindhwani, V. Data Quality from Crowdsourcing: A Study of Annotation Selection Criteria. *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing (ALNLP-09)*; Boulder, CO, June 5, 2009; pp 27−35.
(231) Hirschman, L.; Fort, K.; Boué, S.; Kyrpides, N.; Islamaj Doğan, R.; Cohen, K. B. Crowdsourcing and Curation: Perspectives from Biology and Natural Language Processing. *Database* **2016**, *2016*, baw115.
(232) Neves, M.; Leser, U. A Survey on Annotation Tools for the Biomedical Literature. *Briefings Bioinf.* **2014**, *15* (2), 327−340.
(233) Salgado, D.; Krallinger, M.; Depaule, M.; Drula, E.; Tendulkar, A. V; Leitner, F.; Valencia, A.; Marcelle, C. MyMiner: A Web Application for Computer-Assisted Biocuration and Text Annotation. *Bioinformatics* **2012**, *28* (17), 2285−2287.
(234) Fall, C. J.; Benzineb, K. Literature Survey: Issues to Be Considered in the Automatic Classification of Patents. *World Intellect. Prop. Org.* **2002**, *29*, 1−68.
(235) Farkas, R.; Szarvas, G. Automatic Construction of Rule-Based ICD-9-CM Coding Systems. *BMC Bioinf.* **2008**, *9* (Suppl 3), S10.
(236) Pawar, P. Y.; Gawande, S. H. A Comparative Study on Different Types of Approaches to Text Categorization. *IJMLC* **2012**, *2* (4), 423−426.
(237) Larrañaga, P.; Calvo, B.; Santana, R.; Bielza, C.; Galdiano, J.; Inza, I.; Lozano, J. A.; Armañanzas, R.; Santafé, G.; Pérez, A.; et al. Machine Learning in Bioinformatics. *Briefings Bioinf.* **2006**, *7* (1), 86−112.
(238) Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the 10th European Conference on Machine Learning (ECML'98)*; Chemnitz, Germany, April 21−24, 1998; pp 137−142.
(239) Robertson, S. E.; Jones, K. S. Relevance Weighting of Search Terms. *J. Am. Soc. Inf. Sci.* **1976**, *27* (3), 129−146.
(240) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20* (3), 273−297.
(241) Vapnik, V. N. *Statistical Learning Theory*, 1st ed.; Wiley: New York, 1998.
(242) Rocchio, J. J. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System--Experiments in Automatic Document Processing*; Salton, G., Ed.; Prentice-Hall: Upper Saddle River, NJ, 1971.
(243) *Machine Learning, Neural and Statistical Classification*; Michie, D., Spiegelhalter, D. J., Taylor, C. C., Campbell, J., Eds.; Ellis Horwood: Upper Saddle River, NJ, 1994.
(244) *Machine Learning: An Artificial Intelligence Approach*, 1st ed.; Michalski, S. R., Carbonell, G. J., Mitchell, M. T., Eds.; Morgan Kaufmann Publishers: San Francisco, CA, 1986.
(245) Freund, Y.; Schapire, R. E. Experiments with a New Boosting Algorithm. *Proceedings of the 13th International Conference on Machine Learning (ICML'96)*; Morgan Kaufmann Publishers: Bari, Italy, July 3−6, 1996; pp 148−156.
(246) Yang, Y.; Pedersen, J. O. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of the 14th International Conference on Machine Learning (ICML'97)*; Morgan Kaufmann Publishers: Nashville, TN, July 8−12, 1997; pp 412−420.
(247) Han, B.; Obradovic, Z.; Hu, Z.-Z.; Wu, C. H.; Vucetic, S. Substring Selection for Biomedical Document Classification. *Bioinformatics* **2006**, *22* (17), 2136−2142.
(248) Cover, T. M.; Thomas, J. A. *Elements of Information*; Wiley-Interscience: New York, 1991.
(249) Van Rijsbergen, C. J.; Harper, D. J.; Porter, M. F. The Selection of Good Search Terms. *Inf. Process. Manage.* **1981**, *17* (2), 77−91.
(250) Caruana, R.; Freitag, D. Greedy Attribute Selection. *Proceedings of the 11th International Conference on Machine Learning (ICML'94)*; Morgan Kaufmann Publishers: New Brunswick, NJ, July 10−13, 1994; pp 28−36.
(251) Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648; University of Wisconsin-Madison: Madison, WI, January, 2010.
(252) Chawla, N. V; Japkowicz, N.; Kotcz, A. Editorial: Special Issue on Learning from Imbalanced Data Sets. *SIGKDD Explor.* **2004**, *6* (1), 1−6.
(253) Suomela, B. P.; Andrade, M. A. Ranking the Whole MEDLINE Database according to a Large Training Set Using Text Indexing. *BMC Bioinf.* **2005**, *6*, 75.
(254) Krallinger, M.; Leitner, F.; Valencia, A. Retrieval and Discovery of Cell Cycle Literature and Proteins by Means of Machine Learning, Text Mining and Network Analysis. In *8th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2014)*; Saez-Rodriguez, J., Rocha, M. P., Fdez-Riverola, F., De Paz Santana, J. F., Eds.; Springer: Switzerland, 2014; Vol. *294*: Series Advances in Intelligent Systems and Computing, pp 285−292.
(255) Krallinger, M.; Rojas, A. M.; Valencia, A. Creating Reference Datasets for Systems Biology Applications Using Text Mining. *Ann. N. Y. Acad. Sci.* **2009**, *1158*, 14−28.
(256) Shah, P. K.; Jensen, L. J.; Boué, S.; Bork, P. Extraction of Transcript Diversity from Scientific Literature. *PLoS Comput. Biol.* **2005**, *1* (1), e10.
(257) Shah, P. K.; Bork, P. LSAT: Learning about Alternative Transcripts in MEDLINE. *Bioinformatics* **2006**, *22* (7), 857−865.
(258) Donaldson, I.; Martin, J.; de Bruijn, B.; Wolting, C.; Lay, V.; Tuekam, B.; Zhang, S.; Baskin, B.; Bader, G. D.; Michalickova, K.; et al. PreBIND and Textomy−Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine. *BMC Bioinf.* **2003**, *4*, 11.
(259) Kim, S.; Kwon, D.; Shin, S.-Y.; Wilbur, W. J. PIE the Search: Searching PubMed Literature for Protein Interaction Information. *Bioinformatics* **2012**, *28* (4), 597−598.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(260) Shtatland, T.; Guettler, D.; Kossodo, M.; Pivovarov, M.; Weissleder, R. PepBank–a Database of Peptides Based on Sequence Text Mining and Public Peptide Data Sources. *BMC Bioinf.* **2007**, *8*, 280.
(261) Wang, P.; Morgan, A. A.; Zhang, Q.; Sette, A.; Peters, B. Automating Document Classification for the Immune Epitope Database. *BMC Bioinf.* **2007**, *8*, 269.
(262) Shatkay, H.; Höglund, A.; Brady, S.; Blum, T.; Dönnes, P.; Kohlbacher, O. SherLoc: High-Accuracy Prediction of Protein Subcellular Localization by Integrating Text and Protein Sequence Data. *Bioinformatics* **2007**, *23* (11), 1410−1417.
(263) Brady, S.; Shatkay, H. EpiLoc: A (Working) Text-Based System for Predicting Protein Subcellular Location. *Pac. Symp. Biocomput.* **2008**, 604−615.
(264) Stapley, B. J.; Kelley, L. A.; Sternberg, M. J. E. Predicting the Sub-Cellular Location of Proteins from Text Using Support Vector Machines. *Pac. Symp. Biocomput.* **2002**, 374−385.
(265) Yu, W.; Clyne, M.; Dolan, S. M.; Yesupriya, A.; Wulf, A.; Liu, T.; Khoury, M. J.; Gwinn, M. GAPscreener: An Automatic Tool for Screening Human Genetic Association Literature in PubMed Using the Support Vector Machine Technique. *BMC Bioinf.* **2008**, *9*, 205.
(266) MelanomaMine. http://melanomamine.bioinfo.cnio.es (accessed Oct 20, 2016).
(267) Polavarapu, N.; Navathe, S. B.; Ramnarayanan, R.; ul Haque, A.; Sahay, S.; Liu, Y. Investigation into Biomedical Literature Classification Using Support Vector Machines. *Proc. IEEE Comput. Syst. Bioinform. Conf.* **2005**, 366−374.
(268) McKnight, L.; Srinivasan, P. Categorization of Sentence Types in Medical Abstracts. *AMIA Annu. Symp. Proc.* **2003**, 440−444.
(269) Aphinyanaphongs, Y.; Tsamardinos, I.; Statnikov, A.; Hardin, D.; Aliferis, C. F. Text Categorization Models for High-Quality Article Retrieval in Internal Medicine. *J. Am. Med. Inf. Assoc.* **2005**, *12* (2), 207−216.
(270) Nguyen, A.; Moore, D.; McCowan, I.; Courage, M.-J. Multi-Class Classification of Cancer Stages from Free-Text Histology Reports Using Support Vector Machines. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2007**, *2007*, 5140−5143.
(271) Aphinyanaphongs, Y.; Aliferis, C. Text Categorization Models for Identifying Unproven Cancer Treatments on the Web. *Stud. Health Technol. Inform.* **2007**, *129*, 968−972.
(272) Spasić, I.; Livsey, J.; Keane, J. A.; Nenadić, G. Text Mining of Cancer-Related Information: Review of Current Status and Future Directions. *Int. J. Med. Inform.* **2014**, *83* (9), 605−623.
(273) Rubin, D. L.; Thorn, C. F.; Klein, T. E.; Altman, R. B. A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge. *J. Am. Med. Inf. Assoc.* **2005**, *12* (2), 121−129.
(274) LimTox. http://limtox.bioinfo.cnio.es (accessed Oct 20, 2016).
(275) Liang, C.-Y.; Guo, L.; Xia, Z.-J.; Nie, F.-G.; Li, X.-X.; Su, L.; Yang, Z.-Y. Dictionary-Based Text Categorization of Chemical Web Pages. *Inf. Process. Manage.* **2006**, *42* (4), 1017−1029.
(276) Goetz, T.; von der Lieth, C.-W. PubFinder: A Tool for Improving Retrieval Rate of Relevant PubMed Abstracts. *Nucleic Acids Res.* **2005**, *33*, W774−8 (Web Server issue).
(277) Poulter, G. L.; Rubin, D. L.; Altman, R. B.; Seoighe, C. MScanner: A Classifier for Retrieving Medline Citations. *BMC Bioinf.* **2008**, *9*, 108.
(278) Fontaine, J.-F.; Barbosa-Silva, A.; Schaefer, M.; Huska, M. R.; Muro, E. M.; Andrade-Navarro, M. A. MedlineRanker: Flexible Ranking of Biomedical Literature. *Nucleic Acids Res.* **2009**, *37*, W141−6 (Web Server issue).
(279) Fontaine, J. F.; Priller, F.; Barbosa-Silva, A.; Andrade-Navarro, M. A. Génie: Literature-Based Gene Prioritization at Multi Genomic Scale. *Nucleic Acids Res.* **2011**, *39*, W455−461 (Web Server issue).
(280) Gijon-Correas, J. A.; Andrade-Navarro, M. A.; Fontaine, J. F. Alkemio: Association of Chemicals with Biomedical Topics by Text and Data Mining. *Nucleic Acids Res.* **2014**, *42*, W422−429 (Web Server issue).
(281) Papadatos, G.; van Westen, G. J.; Croset, S.; Santos, R.; Trubian, S.; Overington, J. P. A Document Classifier for Medicinal Chemistry Publications Trained on the ChEMBL Corpus. *J. Cheminf.* **2014**, *6* (1), 40.
(282) Davis, A. P.; Wiegers, T. C.; Johnson, R. J.; Lay, J. M.; Lennon-Hopkins, K.; Saraceni-Richards, C.; Sciaky, D.; Murphy, C. G.; Mattingly, C. J. Text Mining Effectively Scores and Ranks the Literature for Improving Chemical-Gene-Disease Curation at the Comparative Toxicogenomics Database. *PLoS One* **2013**, *8* (4), e58201.
(283) Vishnyakova, D.; Pasche, E.; Ruch, P. Using Binary Classification to Prioritize and Curate Articles for the Comparative Toxicogenomics Database. *Database* **2012**, *2012*, bas050.
(284) Willett, P. Recent Trends in Hierarchic Document Clustering: A Critical Review. *Inf. Process. Manage.* **1988**, *24* (5), 577−597.
(285) Errami, M.; Sun, Z.; Long, T. C.; George, A. C.; Garner, H. R. Deja vu: A Database of Highly Similar Citations in the Scientific Literature. *Nucleic Acids Res.* **2009**, *37*, D921−924 (Database issue).
(286) Lewis, J.; Ossowski, S.; Hicks, J.; Errami, M.; Garner, H. R. Text Similarity: An Alternative Way to Search MEDLINE. *Bioinformatics* **2006**, *22* (18), 2298−2304.
(287) Errami, M.; Wren, J. D.; Hicks, J. M.; Garner, H. R. eTBLAST: A Web Server to Identify Expert Reviewers, Appropriate Journals and Similar Publications. *Nucleic Acids Res.* **2007**, *35*, W12−15 (Web Server issue).
(288) Zhao, Y.; Karypis, G.; Fayyad, U. Hierarchical Clustering Algorithms for Document Datasets. *Data Min. Knowl. Discov.* **2005**, *10* (2), 141−168.
(289) Cutting, D. R.; Karger, D. R.; Pedersen, J. O.; Tukey, J. W. Scatter/gather: A Cluster-Based Approach to Browsing Large Document Collections. *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92)*; Copenhagen, Denmark, June 21−24, 1992; pp 318−329.
(290) Hearst, M. A.; Pedersen, J. O. Reexamining the Cluster Hypothesis: Scatter/gather on Retrieval Results. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*; Zurich, Switzerland, August 18−22, 1996; pp 76−84.
(291) Papanikolaou, N.; Pavlopoulos, G. A.; Pafilis, E.; Theodosiou, T.; Schneider, R.; Satagopam, V. P.; Ouzounis, C. A.; Eliopoulos, A. G.; Promponas, V. J.; Iliopoulos, I. BioTextQuest+: A Knowledge Integration Platform for Literature Mining and Concept Discovery. *Bioinformatics* **2015**, *31* (6), 979−979.
(292) David, M. R.; Samuel, S. Clustering of PubMed Abstracts Using Nearer Terms of the Domain. *Bioinformation* **2012**, *8* (1), 20−25.
(293) Perez-Iratxeta, C.; Keer, H. S.; Bork, P.; Andrade, M. A. Computing Fuzzy Associations for the Analysis of Biological Literature. *Biotechniques* **2002**, *32*, 1380−1382 1384−1385..
(294) Doms, A.; Schroeder, M. GoPubMed: Exploring PubMed with the Gene Ontology. *Nucleic Acids Res.* **2005**, *33*, W783−6 (Web Server issue).
(295) Fattore, M.; Arrigo, P. Knowledge Discovery and System Biology in Molecular Medicine: An Application on Neurodegenerative Diseases. *In Silico Biol.* **2005**, *5*, 199−208.
(296) Yamamoto, Y.; Takagi, T. Biomedical Knowledge Navigation by Literature Clustering. *J. Biomed. Inf.* **2007**, *40* (2), 114−130.
(297) Smalheiser, N. R.; Zhou, W.; Torvik, V. I. Anne O'Tate: A Tool to Support User-Driven Summarization, Drill-down and Browsing of PubMed Search Results. *J. Biomed. Discovery Collab.* **2008**, *3*, 2.
(298) Theodosiou, T.; Darzentas, N.; Angelis, L.; Ouzounis, C. A. PuReD-MCL: A Graph-Based PubMed Document Clustering Methodology. *Bioinformatics* **2008**, *24* (17), 1935−1941.
(299) Lin, Y.; Li, W.; Chen, K.; Liu, Y. A Document Clustering and Ranking System for Exploring MEDLINE Citations. *J. Am. Med. Inf. Assoc.* **2007**, *14* (5), 651−661.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(300) Bush, V. As We May Think. *The atlantic monthly* **1945**, *1*, 101−108.
(301) Saracevic, T. Evaluation of Evaluation in Information Retrieval. *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'95)*; Seattle, WA, July 09−13, 1995; pp 138−146.
(302) Kent, A.; Berry, M. M.; Luehrs, F. U.; Perry, J. W. Machine Literature Searching VIII. Operational Criteria for Designing Information Retrieval Systems. *Am. Doc.* **1955**, *6* (2), 93−101.
(303) van Rijsbergen, C. J. *Information Retrieval*, 2nd ed.; Butterworth-Heinemann: Newton, MA, 1979.
(304) Cleverdon, C. W. The Significance of the Cranfield Tests on Index Languages. *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'91)*; Chicago, IL, October 13−16, 1991; pp 3−12.
(305) Voorhees, E. M.; Harman, D. K. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*; MIT Press: Cambridge, MA, 2005.
(306) Harman, D. K. The First Text Retrieval Conference (TREC-1) Rockville, MD, U.S.A., 4−6 November, 1992. *Inf. Process. Manage.* **1993**, *29* (4), 411−414.
(307) Hersh, W.; Buckley, C.; Leone, T. J.; Hickam, D. OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research. *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*; Dublin, Ireland, July 03−06, 1994; pp 192−201.
(308) Hersh, W. R.; Bhupatiraju, R. T. TREC GENOMICS Track Overview. *Proceedings of the 12th Text Retrieval Conference (TREC 2003)*; Gaithersburg, MD, November 18−21, 2003; pp 14−23.
(309) Hersh, W.; Voorhees, E. TREC Genomics Special Issue Overview. *Inf. Retr.* **2009**, *12* (1), 1−15.
(310) Lupu, M.; Huang, J.; Zhu, J.; Tait, J. TREC-CHEM: Large Scale Chemical Information Retrieval Evaluation at TREC. *ACM SIGIR Forum* **2009**, *43* (2), 63−70.
(311) Lupu, M.; Tait, J.; Huang, J.; Zhu, J. TREC-CHEM 2010: Notebook Report. *Proceedings of the 19th Text Retrieval Conference (TREC 2010)*; Gaithersburg, MD, November 16−19, 2010.
(312) Lupu, M.; Huang, J.; Zhu, J.; Tait, J. TREC Chemical Information Retrieval–An Initial Evaluation Effort for Chemical IR Systems. *World Pat. Inf.* **2011**, *33* (3), 248−256.
(313) Lupu, M.; Piroi, F.; Huang, X.; Zhu, J.; Tait, J. Overview of the TREC 2009 Chemical IR Track. *Proceedings of the 18th Text Retrieval Conference (TREC 2009)*; Gaithersburg, MD, November 17−20, 2009.
(314) Krallinger, M.; Leitner, F.; Rodriguez-Penagos, C.; Valencia, A. Overview of the Protein-Protein Interaction Annotation Extraction Task of BioCreative II. *Genome Biol.* **2008**, *9* (Suppl 2), S4.
(315) Leitner, F.; Mardis, S. A.; Krallinger, M.; Cesareni, G.; Hirschman, L. A.; Valencia, A. An Overview of BioCreative II.5. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2010**, *7* (3), 385−399.
(316) Pérez-Pérez, M.; Pérez-Rodríguez, G.; Rabal, O.; Vazquez, M.; Oyarzabal, J.; Fdez-Riverola, F.; Valencia, A.; Krallinger, M.; Lourenço, A. The Markyt Visualisation, Prediction and Benchmark Platform for Chemical and Gene Entity Recognition at BioCreative/CHEMDNER Challenge. *Database* **2016**, *2016*, baw120.
(317) Krallinger, M.; Rabal, O.; Lourenço, A.; Perez, M. P.; Rodriguez, G. P.; Vazquez, M.; Oyarzabal, F. L. J.; Valencia, A. Overview of the CHEMDNER Patents Task. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*; Sevilla, Spain, September 9−11, 2015; pp 63−75.
(318) Krallinger, M.; Rabal, O.; Leitner, F.; Vazquez, M.; Salgado, D.; Lu, Z.; Leaman, R.; Lu, Y.; Ji, D.; Lowe, D. M.; et al. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *J. Cheminf.* **2015**, *7*, S2 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(319) Nadeau, D.; Sekine, S. A Survey of Named Entity Recognition and Classification. *Lingvist. Invest.* **2007**, *30* (1), 3−26.
(320) MacDonald, M. C.; Pearlmutter, N. J.; Seidenberg, M. S. The Lexical Nature of Syntactic Ambiguity Resolution. *Psychol. Rev.* **1994**, *101* (4), 676−703.
(321) Gale, W. A.; Church, K. W.; Yarowsky, D. One Sense per Discourse. *Proceedings of the Workshop on Speech and Natural Language (HLT '91)*; Harriman, NY, February 23−26, 1992; pp 233−237.
(322) Grishman, R.; Sundheim, B. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th conference on Computational linguistics - Volume 1 (COLING'96)*; Copenhagen, Denmark, August, 1996; pp 466−471.
(323) Chinchor, N.; Robinson, P. MUC-7 Named Entity Task Definition. *Proceedings of the 7th Conference on Message Understanding (MUC-7)*; Fairfax, VA, 1998.
(324) Garfield, E. An Algorithm for Translating Chemical Names to Molecular Formulas. *J. Chem. Doc.* **1962**, *2* (3), 177−179.
(325) Reeker, L. H.; Zamora, E. M.; Blower, P. E. Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions. *Proceedings of the first conference on Applied natural language processing (ANLC '83)*; Santa Monica, CA, February 1−3, 1983; pp 109−116.
(326) Hodge, G. M.; Nelson, T. W.; Vleduts-Stokolov, N. *Automatic Recognition of Chemical Names in Natural-Language Texts*; Presented at the 197th National Meeting of the American Chemical Society, Dallas, TX, April 7−9, 1989; paper CINF-17.
(327) Ai, C. S.; Blower, P. E., Jr; Ledwith, R. H. Extraction of Chemical Reaction Information from Primary Journal Text. *J. Chem. Inf. Model.* **1990**, *30* (2), 163−169.
(328) Babych, B.; Hartley, A. Improving Machine Translation Quality with Automatic Named Entity Recognition. *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools: Resources and Tools for Building MT (EAMT '03)*; Budapest, Hungary, April 9−16, 2003; pp 1−8.
(329) Mollá, D.; Van Zaanen, M.; Smith, D. Named Entity Recognition for Question Answering. *Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006)*; Sydney, Australia, November 30−December 1, 2006; pp 51−58.
(330) Maynard, D.; Tablan, V.; Ursu, C.; Cunningham, H.; Wilks, Y. Named Entity Recognition from Diverse Text Types. *Recent Advances in Natural Language Processing 2007 Conference (RANLP - 2007)*; Tzigov Chark, Bulgaria, September 5−7, 2001; pp 257−274.
(331) Karkaletsis, V.; Spyropoulos, C. D.; Petasis, G. Named Entity Recognition from Greek Texts: The GIE Project. In *Advances in Intelligent Systems: Concepts, Tools and Applications*; Tzafestas, S. G., Ed.; Springer: Dordrecht, Netherlands, 1999; pp 131−142.
(332) Lowe, D. M.; Sayle, R. A. LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition. *J. Cheminf.* **2015**, *7*, S5 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(333) Leaman, R.; Wei, C.-H.; Zou, C.; Lu, Z. Mining Chemical Patents with an Ensemble of Open Systems. *Database* **2016**, *2016*, pii: baw065.
(334) Lowe, D. M.; Sayle, R. A. Recognition of Chemical Entities in Patents Using LeadMine. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*; Sevilla, Spain, September 9−11, 2015; pp 129−134.
(335) Kolárik, C.; Klinger, R.; Friedrich, C. M.; Hofmann-Apitius, M.; Fluck, J. Chemical Names: Terminological Resources and Corpora Annotation. *Workshop on Building and Evaluating Resources for Biomedical Text Mining (Sixth International Conference on Language Resources and Evaluation - LREC 2008 Workshop)*; Marrakech, Morocco, May 26, 2008; pp 51−58.
(336) Hettne, K. M.; Stierum, R. H.; Schuemie, M. J.; Hendriksen, P. J.; Schijvenaars, B. J.; Mulligen, E. M.; Kleinjans, J.; Kors, J. A. A Dictionary to Identify Small Molecules and Drugs in Free Text. *Bioinformatics* **2009**, *25* (22), 2983−2991.
(337) Grefenstette, G.; Tapanainen, P. What Is a Word, What Is a Sentence?: Problems of Tokenisation. *3rd International Conference on Computational Lexicography (COMPLEX'94)*; Budapest, Hungary, 1994; pp 79−87.
(338) Campos, D.; Matos, S.; Oliveira, J. L. A Document Processing Pipeline for Annotating Chemical Entities in Scientific Documents. *J.*

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Cheminf. **2015**, *7*, S7 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(339) Zhang, Y.; Xu, J.; Chen, H.; Wang, J.; Wu, Y.; Prakasam, M.; Xu, H. Chemical Named Entity Recognition in Patents by Domain Knowledge and Unsupervised Feature Learning. *Database* **2016**, *2016*, pii: baw049.baw04910.1093/database/baw049
(340) Eltyeb, S.; Salim, N. Chemical Named Entities Recognition: A Review on Approaches and Applications. *J. Cheminf.* **2014**, *6*, 17.
(341) Liu, S.; Tang, B.; Chen, Q.; Wang, X. Drug Name Recognition: Approaches and Resources. *Information* **2015**, *6* (4), 790−810.
(342) Segura-Bedmar, I.; Martínez, P.; Segura-Bedmar, M. Drug Name Recognition and Classification in Biomedical Texts: A Case Study Outlining Approaches Underpinning Automated Systems. *Drug Discovery Today* **2008**, *13* (17−18), 816−823.
(343) Munkhdalai, T.; Li, M.; Batsuren, K.; Park, H. A.; Choi, N. H.; Ryu, K. H. Incorporating Domain Knowledge in Chemical and Biomedical Named Entity Recognition with Word Representations. *J. Cheminf.* **2015**, *7*, S9 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(344) Lu, Y.; Ji, D.; Yao, X.; Wei, X.; Liang, X. CHEMDNER System with Mixed Conditional Random Fields and Multi-Scale Word Clustering. *J. Cheminf.* **2015**, *7*, S4 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(345) Usié, A.; Cruz, J.; Comas, J.; Solsona, F.; Alves, R. CheNER: A Tool for the Identification of Chemical Entities and Their Classes in Biomedical Literature. *J. Cheminf.* **2015**, *7*, S15 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(346) Grego, T.; Pesquita, C.; Bastos, H. P.; Couto, F. M. Chemical Entity Recognition and Resolution to ChEBI. *ISRN Bioinf.* **2012**, *2012*, 619427.
(347) Aronson, A. R.; Lang, F.-M. An Overview of MetaMap: Historical Perspective and Recent Advances. *J. Am. Med. Inf. Assoc.* **2010**, *17* (3), 229−236.
(348) Rebholz-Schuhmann, D.; Arregui, M.; Gaudan, S.; Kirsch, H.; Jimeno, A. Text Processing through Web Services: Calling Whatizit. *Bioinformatics* **2008**, *24* (2), 296−298.
(349) Schuemie, M. J.; Jelier, R.; Kors, J. A. Peregrine: Lightweight Gene Name Normalization by Dictionary Lookup. *Proceedings of the Second BioCreative Challenge Evaluation Workshop*; Madrid, Spain, April 23−25, 2007; pp 131−133.
(350) Hanisch, D.; Fundel, K.; Mevissen, H.-T.; Zimmer, R.; Fluck, J. ProMiner: Rule-Based Protein and Gene Entity Recognition. *BMC Bioinf.* **2005**, *6* (Suppl1), S14.
(351) Townsend, J. A.; Adams, S. E.; Waudby, C. A.; de Souza, V. K.; Goodman, J. M.; Murray-Rust, P. Chemical Documents: Machine Understanding and Automated Information Extraction. *Org. Biomol. Chem.* **2004**, *2* (22), 3294−3300.
(352) Xu, H.; Stenner, S. P.; Doan, S.; Johnson, K. B.; Waitman, L. R.; Denny, J. C. MedEx: A Medication Information Extraction System for Clinical Narratives. *J. Am. Med. Inf. Assoc.* **2010**, *17* (1), 19−24.
(353) Hettne, K. M.; Williams, A. J.; van Mulligen, E. M.; Kleinjans, J.; Tkachenko, V.; Kors, J. A. Automatic vs. Manual Curation of a Multi-Source Chemical Dictionary: The Impact on Text Mining. *J. Cheminf.* **2010**, *2* (1), 4.
(354) Kolárik, C.; Hofmann-Apitius, M.; Zimmermann, M.; Fluck, J. Identification of New Drug Classification Terms in Textual Resources. *Bioinformatics* **2007**, *23* (13), i264−i272.
(355) Chhieng, D.; Day, T.; Gordon, G.; Hicks, J. Use of Natural Language Programming to Extract Medication from Unstructured Electronic Medical Records. *AMIA Annu. Symp. Proc.* **2007**, 908.
(356) Wagner, R. A.; Fischer, M. J. The String-to-String Correction Problem. *J. Assoc. Comput. Mach.* **1974**, *21* (1), 168−173.
(357) Levin, M. A.; Krol, M.; Doshi, A.; Reich, D. L. Extraction and Mapping of Drug Names from Free Text to a Standardized Nomenclature. *AMIA Annu. Symp. Proc.* **2007**, 438−442.
(358) Akhondi, S. A.; Hettne, K. M.; van der Horst, E.; van Mulligen, E. M.; Kors, J. A. Recognition of Chemical Entities: Combining Dictionary-Based and Grammar-Based Approaches. *J. Cheminf.* **2015**, *7*, S10 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(359) Rindflesch, T. C.; Tanabe, L.; Weinstein, J. N.; Hunter, L. EDGAR: Extraction of Drugs, Genes and Relations from the Biomedical Literature. *Pac. Symp. Biocomput.* **2000**, 517−528.
(360) Sanchez-Cisneros, D.; Martinez, P.; Segura-Bedmar, I. Combining Dictionaries and Ontologies for Drug Name Recognition in Biomedical Texts. *Proceedings of the 7th International Workshop on Data and Text Mining in Biomedical Informatics (DTMBIO'13)*; San Francisco, CA, November 1, 2013; pp 27−30.
(361) Jurafsky, D.; Martin, J. H. *Speech and Language Processing*, 2nd ed.; Russell, S., Norvig, P., Eds.; Prentice-Hall: Upper Saddle River, NJ, 2014.
(362) Kleene, S. C. *Representation of Events in Nerve Nets and Finite Automata*; Research Memorandum. U.S. Air Force PROJECT RAND. DTIC Document; Santa Monica, CA, December 1951.
(363) Thomas, P.; Rocktaschel, T.; Hakenberg, J.; Lichtblau, Y.; Leser, U. SETH Detects and Normalizes Genetic Variants in Text. *Bioinformatics* **2016**, *32* (18), 2883−2885.
(364) Caporaso, J. G.; Baumgartner, W. A.; Randolph, D. A.; Cohen, K. B.; Hunter, L. MutationFinder: A High-Performance System for Extracting Point Mutation Mentions from Text. *Bioinformatics* **2007**, *23* (14), 1862−1865.
(365) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894.
(366) Batista-Navarro, R.; Ananiadou, S. Adapting ChER for the Recognition of Chemical Mentions in Patents. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*; Sevilla, Spain, September 9−11, 2015; pp 149−153.
(367) Kemp, N.; Lynch, M. Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (4), 544−551.
(368) Chowdhury, G. G.; Lynch, M. F. Automatic Interpretation of the Texts of Chemical Patent Abstracts. 1. Lexical Analysis and Categorization. *J. Chem. Inf. Model.* **1992**, *32* (5), 463−467.
(369) Narayanaswamy, M.; Ravikumar, K. E.; Vijay-Shanker, K. A Biological Named Entity Recognizer. *Pac. Symp. Biocomput.* **2003**, 427−438.
(370) Xu, R.; Morgan, A.; Das, A. K.; Garber, A. Investigation of Unsupervised Pattern Learning Techniques for Bootstrap Construction of a Medical Treatment Lexicon. *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP 2009)*; Boulder, CO, June 4−5, 2009; pp 63−70.
(371) Coden, A.; Gruhl, D.; Lewis, N.; Tanenblatt, M.; Terdiman, J. SPOT the Drug! An Unsupervised Pattern Matching Method to Extract Drug Names from Very Large Clinical Corpora. *Proceedings of the 2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB'12)*; La Jolla, CA, September 27−28, 2012; pp 33−39.
(372) Sarawagi, S. Information Extraction. *Found. trends databases* **2008**, *1* (3), 261−377.
(373) Califf, M. E.; Mooney, R. Relational Learning of Pattern-Match Rules for Information Extraction. *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*; Orlando, FL, July 18−22, 1999; pp 9−15.
(374) Ciravegna, F. Adaptive Information Extraction from Text by Rule Induction and Generalisation. *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2 (IJCAI'01)*; Seattle, Washington, August 4−10, 2001; pp 1251−1256.
(375) Soderland, S. Learning Information Extraction Rules for Semi-Structured and Free Text. *Mach. Learn.* **1999**, *34* (1−3), 233−272.
(376) Gold, S.; Elhadad, N.; Zhu, X.; Cimino, J. J.; Hripcsak, G. Extracting Structured Medication Event Information from Discharge Summaries. *AMIA Annu. Symp. Proc.* **2008**, 237−241.
(377) Hamon, T.; Grabar, N. Linguistic Approach for Identification of Medication Names and Related Information in Clinical Narratives. *J. Am. Med. Inf. Assoc.* **2010**, *17* (5), 549−554.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(378) Mack, R.; Mukherjea, S.; Soffer, A.; Uramoto, N.; Brown, E.; Coden, A.; Cooper, J.; Inokuchi, A.; Iyer, B.; Mass, Y.; et al. Text Analytics for Life Science Using the Unstructured Information Management Architecture. *IBM Syst. J.* **2004**, *43* (3), 490−515.
(379) Dieb, T. M.; Yoshioka, M.; Hara, S. Automatic Information Extraction of Experiments from Nanodevices Development Papers. *Advanced Applied Informatics (IIAIAAI), 2012 IIAI International Conference on*; Fukuoka, Japan, September 20−22, 2012; pp 42−47.
(380) Dieb, T. M.; Yoshioka, M. Extraction of Chemical and Drug Named Entities by Ensemble Learning Using Chemical NER Tools Based on Different Extraction Guidelines. *Trans. Mach. Learn. Data Min.* **2015**, *8*, 61−76.
(381) Yan, S.; Spangler, W. S.; Chen, Y. Learning to Extract Chemical Names Based on Random Text Generation and Incomplete Dictionary. *Proceedings of the 11th International Workshop on Data Mining in Bioinformatics (BIOKDD '12)*; Beijing, China, August 12, 2012; pp 21−25.
(382) Chiticariu, L.; Krishnamurthy, R.; Li, Y.; Reiss, F.; Vaithyanathan, S. Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*; Cambridge, MA, October 9−11, 2010; pp 1002−1012.
(383) Vasserman, A. Identifying Chemical Names in Biomedical Text: An Investigation of the Substring Co-Occurrence Based Approaches. *Proceedings of the Student Research Workshop at HLT-NAACL 2004 (HLT-SRWS'04)*; Boston, MA, May 2, 2004; pp 7−12.
(384) Xu, S.; An, X.; Zhu, L.; Zhang, Y.; Zhang, H. A CRF-Based System for Recognizing Chemical Entity Mentions (CEMs) in Biomedical Literature. *J. Cheminf.* **2015**, *7*, S11 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(385) Mansouri, A.; Affendey, L. S.; Mamat, A. Named Entity Recognition Approaches. *IJCSNS* **2008**, *8*, 339−344.
(386) Sekine, S. NYU: Description of the Japanese NE System Used for MET-2. *Proceedings of the 7th Conference on Message Understanding (MUC-7)*; Fairfax, VA, 1998.
(387) Bikel, D. M.; Miller, S.; Schwartz, R.; Weischedel, R. Nymble: A High-Performance Learning Name-Finder. *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLC '97)*; Washington, DC, March 31−April 3, 1997; pp 194−201.
(388) Ponomareva, N.; Rosso, P.; Pla, F.; Molina, A. Conditional Random Fields vs Hidden Markov Models in a Biomedical Named Entity Recognition Task. *Recent Advances in Natural Language Processing 2007 Conference (RANLP - 2007)*; Borovets, Bulgaria, September 27−29, 2007; pp 479−483.
(389) Corbett, P.; Copestake, A. Cascaded Classifiers for Confidence-Based Chemical Named Entity Recognition. *BMC Bioinf.* **2008**, *9* (Suppl 11), S4.
(390) Asahara, M.; Matsumoto, Y. Japanese Named Entity Extraction with Redundant Morphological Analysis. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1 (NAACL'03)*; Edmonton, Canada, May 31- June 1, 2003; pp 8−15.
(391) Tang, B.; Feng, Y.; Wang, X.; Wu, Y.; Zhang, Y.; Jiang, M.; Wang, J.; Xu, H. A Comparison of Conditional Random Fields and Structured Support Vector Machines for Chemical Entity Recognition in Biomedical Literature. *J. Cheminf.* **2015**, *7*, S8 (Suppl 1 Text mining for chemistry and the CHEMDNER track).
(392) McCallum, A.; Li, W. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction and Web-Enhanced Lexicons. *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (CONLL'03)*; Edmonton, Canada, May 31−June 1, 2003; pp 188−191.
(393) Zitnik, S.; Bajec, M. Token-and Constituent-Based Linear-Chain Crf with Svm for Named Entity Recognition. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop Vol. 2*; Bethesda, MD, October 7−9, 2013; pp 144−151.
(394) Rabiner, L.; Juang, B. An Introduction to Hidden Markov Models. *IEEE ASSP Mag.* **1986**, *3* (1), 4−16.
(395) Sutton, C.; McCallum, A. An Introduction to Conditional Random Fields for Relational Learning. In *Introduction to statistical relational learning*; Getoor, L., Taskar, B., Eds.; MIT Press: Cambridge, MA, 2007; pp 93−127.
(396) Ratnaparkhi, A. A Maximum Entropy Model for Part-of-Speech Tagging. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'96)*; Philadelphia, PA, May 17−18, 1996; pp 133−142.
(397) Zhao, H.; Huang, C.-N.; Li, M.; Lu, B.-L. A Unified Character-Based Tagging Framework for Chinese Word Segmentation. *TALIP* **2010**, *9* (2), 5.
(398) He, L.; Yang, Z.; Lin, H.; Li, Y. Drug Name Recognition in Biomedical Texts: A Machine-Learning-Based Method. *Drug Discovery Today* **2014**, *19* (5), 610−617.
(399) Akhondi, S. A.; Pons, E.; Afzal, Z.; van Haagen, H.; Becker, B. F. H.; Hettne, K. M.; van Mulligen, E. M.; Kors, J. A. Chemical Entity Recognition in Patents by Combining Dictionary-Based and Statistical Approaches. *Database* **2016**, *2016*, pii:baw061.
(400) Tikk, D.; Solt, I. Improving Textual Medication Extraction Using Combined Conditional Random Fields and Rule-Based Systems. *J. Am. Med. Inf. Assoc.* **2010**, *17* (5), 540−544.
(401) Kucera, H.; Francis, W. N. *Computational Analysis of Present-Day American English*; Brown University Press: Providence, RI, 1967.
(402) Rebholz-Schuhmann, D.; Yepes, A. J. J.; Van Mulligen, E. M.; Kang, N.; Kors, J.; Milward, D.; Corbett, P.; Buyko, E.; Beisswanger, E.; Hahn, U. CALBC Silver Standard Corpus. *J. Bioinf. Comput. Biol.* **2010**, *8* (1), 163−179.
(403) Müller, B.; Klinger, R.; Gurulingappa, H.; Mevissen, H.-T.; Hofmann-Apitius, M.; Fluck, J.; Friedrich, C. M. Abstracts versus Full Texts and Patents: A Quantitative Analysis of Biomedical Entities. *Proceedings of the First international Information Retrieval Facility conference on Advances in Multidisciplinary Retrieval (IRFC'10)*; Vienna, Austria, May 31, 2010; pp 152−165.
(404) Rebholz-Schuhmann, D.; Kirsch, H.; Nenadic, G.; Rebholz-Schuhmann, D.; Kirsch, H.; Nenadic, G. IeXML: Towards an Annotation Framework for Biomedical Semantic Types Enabling Interoperability of Text Processing Modules. *Proceedings of the Joint BioLINK and Bio-Ontologies SIG Meeting*; Fortaleza, Brazil, August 4−5, 2006.
(405) Townsend, J. A.; Murray-Rust, P. CMLLite: A Design Philosophy for CML. *J. Cheminf.* **2011**, *3* (1), 39.
(406) Chemical Markup Language. hhttp://www.xml-cml.org (accessed Oct 20, 2016).
(407) Comeau, D. C.; Islamaj Doğan, R.; Ciccarese, P.; Cohen, K. B.; Krallinger, M.; Leitner, F.; Lu, Z.; Peng, Y.; Rinaldi, F.; Torii, M.; et al. BioC: A Minimalist Approach to Interoperability for Biomedical Text Processing. *Database* **2013**, *2013*, bat064.
(408) Nobata, C.; Dobson, P. D.; Iqbal, S. A.; Mendes, P.; Tsujii, J.; Kell, D. B.; Ananiadou, S. Mining Metabolites: Extracting the Yeast Metabolome from the Literature. *Metabolomics* **2011**, *7* (1), 94−101.
(409) Van Mulligen, E. M.; Fourrier-Reglat, A.; Gurwitz, D.; Molokhia, M.; Nieto, A.; Trifiro, G.; Kors, J. A.; Furlong, L. I. The EU-ADR Corpus: Annotated Drugs, Diseases, Targets, and Their Relationships. *J. Biomed. Inf.* **2012**, *45* (5), 879−884.
(410) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672 (Database issue).
(411) Herrero-Zazo, M.; Segura-Bedmar, I.; Martínez, P.; Declerck, T. The DDI Corpus: An Annotated Corpus with Pharmacological Substances and Drug−Drug Interactions. *J. Biomed. Inf.* **2013**, *46* (5), 914−920.
(412) Bada, M.; Eckert, M.; Evans, D.; Garcia, K.; Shipley, K.; Sitnikov, D.; Baumgartner, W. A.; Cohen, K. B.; Verspoor, K.; Blake, J. A.; et al. Concept Annotation in the CRAFT Corpus. *BMC Bioinf.* **2012**, *13* (1), 161.
(413) Batista-Navarro, R. T.; Ananiadou, S. Building a Coreference-Annotated Corpus from the Domain of Biochemistry. *Proceedings of*

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

BioNLP 2011 Workshop (BioNLP'11); Portland, OR, June 24, 2011; pp 83−91.
(414) Schlaf, A.; Bobach, C.; Irmer, M. Creating a Gold Standard Corpus for the Extraction of Chemistry-Disease Relations from Patent Texts. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*; Reykjavik, Iceland, May 16−31, 2014; pp 2057−2061.
(415) Habibi, M.; Wiegandt, D. L.; Schmedding, F.; Leser, U. Recognizing Chemicals in Patents: A Comparative Analysis. *J. Cheminf.* **2016**, *8* (1), 59.
(416) Corbett, P.; Batchelor, C.; Teufel, S. Annotation of Chemical Named Entities. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing (BioNLP'07)*; Prague, Czech Republic, June 29, 2007; pp 57−64.
(417) Rupp, C. J.; Copestake, A.; Teufel, S.; Waldron, B. Flexible Interfaces in the Application of Language Technology to an Escience Corpus. *Proceedings of the UK e-Science All Hands Meeting*; Nottingham, UK, September 18−21, 2006; pp 622−629.
(418) ChEBI Chapati corpus. http://chebi.cvs.sourceforge.net/viewvc/chebi/chapati/ (accessed Oct 20, 2016).
(419) Degtyarenko, K.; de Matos, P.; Ennis, M.; Hastings, J.; Zbinden, M.; McNaught, A.; Alcántara, R.; Darsow, M.; Guedj, M.; Ashburner, M. ChEBI: A Database and Ontology for Chemical Entities of Biological Interest. *Nucleic Acids Res.* **2008**, *36*, D344−D350 (Database issue).
(420) Akhondi, S. A.; Klenner, A. G.; Tyrchan, C.; Manchala, A. K.; Boppana, K.; Lowe, D.; Zimmermann, M.; Jagarlapudi, S. A. R. P.; Sayle, R.; Kors, J. A.; et al. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. *PLoS One* **2014**, *9* (9), e107477.
(421) Hirschman, L. The Evolution of Evaluation: Lessons from the Message Understanding Conferences. *Comput. Speech Lang.* **1998**, *12* (4), 281−305.
(422) Huang, C.-C.; Lu, Z. Community Challenges in Biomedical Text Mining over 10 Years: Success, Failure and the Future. *Briefings Bioinf.* **2016**, *17* (1), 132−144.
(423) Arighi, C. N.; Wu, C. H.; Cohen, K. B.; Hirschman, L.; Krallinger, M.; Valencia, A.; Lu, Z.; Wilbur, J. W.; Wiegers, T. C. BioCreative-IV Virtual Issue. *Database* **2014**, *2014*, bau039.
(424) Li, J.; Sun, Y.; Johnson, R.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; Lu, Z. Annotating Chemicals, Diseases, and Their Interactions in Biomedical Literature. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*; Sevilla, Spain, September 9−11, 2015; pp 173−182.
(425) Wei, C.-H.; Peng, Y.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Li, J.; Wiegers, T. C.; Lu, Z. Assessing the State of the Art in Biomedical Relation Extraction: Overview of the BioCreative V Chemical-Disease Relation (CDR) Task. *Database* **2016**, *2016*, pii: baw032.
(426) Yeh, A.; Morgan, A.; Colosimo, M.; Hirschman, L. BioCreAtIvE Task 1A: Gene Mention Finding Evaluation. *BMC Bioinf.* **2005**, *6* (Suppl 1), S2.
(427) Krallinger, M.; Leitner, F.; Rabal, O. Overview of the Chemical Compound and Drug Name Recognition (CHEMDNER) Task. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2*; Bethesda, MD, October 7−9, 2013; pp 2−33.
(428) Martin, E.; Monge, A.; Duret, J.-A.; Gualandi, F.; Peitsch, M. C.; Pospisil, P. Building an R&D Chemical Registration System. *J. Cheminf.* **2012**, *4* (1), 11.
(429) Williams, A. J.; Ekins, S.; Tkachenko, V. Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation. *Drug Discovery Today* **2012**, *17* (13−14), 685−701.
(430) Brecher, J. S. The ChemFinder WebServer: Indexing Chemical Data on the Internet. *Chim. Int. J. Chem.* **1998**, *52*, 658−663.
(431) Krebs, H. A.; Jordis, U. How to Add Chemical Abstracts Service Registry Numbers and Structures to Databases via Chemical Names Comparison. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (2), 293−294.
(432) ChemHits. Chemical compound name normalization and matching. http://sabio.h-its.org/chemHits (accessed Oct 20, 2016).
(433) Sayle, R. Foreign Language Translation of Chemical Nomenclature by Computer. *J. Chem. Inf. Model.* **2009**, *49* (3), 519−530.
(434) Karthikeyan, M.; Vyas, R. ChemEngine: Harvesting 3D Chemical Structures of Supplementary Data from PDF Files. *J. Cheminf.* **2016**, *8* (1), 73.
(435) Vander Stouw, G. G.; Naznitsky, I.; Rush, J. E. Procedures for Converting Systematic Names of Organic Compounds into Atom-Bond Connection Tables. *J. Chem. Doc.* **1967**, *7* (3), 165−169.
(436) Stouw, G. G. Vander; Elliott, P. M.; Isenberg, A. C. Automated Conversion of Chemical Substance Names to Atom-Bond Connection Tables. *J. Chem. Doc.* **1974**, *14* (4), 185−193.
(437) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Introduction and Background to a Grammar-Based Approach. *J. Chem. Inf. Model.* **1989**, *29* (2), 101−105.
(438) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 2. Development of a Formal Grammar. *J. Chem. Inf. Model.* **1989**, *29* (2), 106−112.
(439) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 3. Syntax Analysis and Semantic Processing. *J. Chem. Inf. Model.* **1989**, *29* (2), 112−118.
(440) Williams, A. J.; Yerin, A. Automated Identification and Conversion of Chemical Names to Structure-Searchable Information. In *Chemical Information Mining: Facilitating Literature-Based Discovery*; Banville, D. L., Ed.; CRC Press: Boca Raton, FL, 2008; pp 21−44.
(441) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *J. Chem. Inf. Model.* **2011**, *51* (3), 739−753.
(442) Reyle, U. Understanding Chemical Terminology. *Terminology* **2006**, *12* (1), 111−136.
(443) Cannon, E. O. New Benchmark for Chemical Nomenclature Software. *J. Chem. Inf. Model.* **2012**, *52* (5), 1124−1131.
(444) LeadMine. https://www.nextmovesoftware.com/leadmine.html (accessed Oct 20, 2016).
(445) Williams, A. J.; Yerin, A. Automated Systematic Nomenclature Generation for Organic Compounds. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3* (2), 150−160.
(446) Navigli, R. Word Sense Disambiguation: A Survey. *ACM Comput. Surv.* **2009**, *41* (2), 10−79.
(447) Locke, W. N.; Booth, A. D. *Machine Translation of Languages: Fourteen Essays*; Technology Press: Cambridge, MA, 1955.
(448) Bunescu, R.; Pasca, M. Using Encyclopedic Knowledge for Named Entity Disambiguation. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*; Trento, Italy, April 3−7, 2006; pp 9−16.
(449) Ratinov, L.; Roth, D.; Downey, D.; Anderson, M. Local and Global Algorithms for Disambiguation to Wikipedia. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*; Portland, OR, June 19−24, 2011; pp 1375−1384.
(450) Hirschman, L.; Colosimo, M.; Morgan, A.; Yeh, A. Overview of BioCreAtIvE Task 1B: Normalized Gene Lists. *BMC Bioinf.* **2005**, *6* (Suppl1), S11.
(451) Morgan, A. A.; Lu, Z.; Wang, X.; Cohen, A. M.; Fluck, J.; Ruch, P.; Divoli, A.; Fundel, K.; Leaman, R.; Hakenberg, J.; et al. Overview of BioCreative II Gene Normalization. *Genome Biol.* **2008**, *9* (Suppl 2), S3.
(452) Névéol, A.; Li, J.; Lu, Z. Linking Multiple Disease-Related Resources through UMLS. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI'12)*; Miami, FL, January 28−30, 2012; pp 767−772.
(453) Doğan, R. I.; Leaman, R.; Lu, Z. NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *J. Biomed. Inf.* **2014**, *47*, 1−10.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(454) Leaman, R.; Islamaj Dogan, R.; Lu, Z. DNorm: Disease Name Normalization with Pairwise Learning to Rank. *Bioinformatics* **2013**, *29* (22), 2909−2917.
(455) Lu, Z.; Kao, H.-Y.; Wei, C.-H.; Huang, M.; Liu, J.; Kuo, C.-J.; Hsu, C.-N.; Tsai, R. T.-H.; Dai, H.-J.; Okazaki, N.; et al. The Gene Normalization Task in BioCreative III. *BMC Bioinf.* **2011**, *12* (Suppl 8), S2.
(456) Chen, L.; Liu, H.; Friedman, C. Gene Name Ambiguity of Eukaryotic Nomenclatures. *Bioinformatics* **2005**, *21* (2), 248−256.
(457) Hakenberg, J.; Plake, C.; Leaman, R.; Schroeder, M.; Gonzalez, G. Inter-Species Normalization of Gene Mentions with GNAT. *Bioinformatics* **2008**, *24* (16), i126−132.
(458) Leaman, R.; Lu, Z. TaggerOne: Joint Named Entity Recognition and Normalization with Semi-Markov Models. *Bioinformatics* **2016**, *32* (18), 2839−2846.
(459) Filippov, I. V; Nicklaus, M. C. Optical Structure Recognition Software to Recover Chemical Information: OSRA, an Open Source Solution. *J. Chem. Inf. Model.* **2009**, *49* (3), 740−743.
(460) Uchida, S. Image Processing and Recognition for Biological Images. *Dev., Growth Differ.* **2013**, *55* (4), 523−549.
(461) Selinger, P. Potrace. http://potrace.sourceforge.net (accessed Oct 27, 2016).
(462) Lounnas, V.; Vriend, G. AsteriX: A Web Server to Automatically Extract Ligand Coordinates from Figures in PDF Articles. *J. Chem. Inf. Model.* **2012**, *52* (2), 568−576.
(463) Tharatipyakul, A.; Numnark, S.; Wichadakul, D.; Ingsriswang, S. ChemEx: Information Extraction System for Chemical Data Curation. *BMC Bioinf.* **2012**, *13*, S9.
(464) Fujiyoshi, A.; Nakagawa, K.; Suzuki, M. Robust Method of Segmentation and Recognition of Chemical Structure Images in Cheminfty. *Pre-Proceedings of the 9th IAPR International Workshop on Graphics Recognition (GREC2011)*; Seoul, Korea, September 15−16, 2011.
(465) Frasconi, P.; Gabbrielli, F.; Lippi, M.; Marinai, S. Markov Logic Networks for Optical Chemical Structure Recognition. *J. Chem. Inf. Model.* **2014**, *54* (8), 2380−2390.
(466) Kibbey, C. E.; Klug-McLeod, J. L. Structure Clipper−an Interactive Tool for Extracting Chemical Structures from Patents. Presented at the 248th National Meeting & Exposition of the American Chemical Society, San Francisco, CA, August 10−14, 2014; CINF-56.
(467) Barnard, J. M.; Kenny, P. W.; Wallace, P. N. CHAPTER 6 Representing Chemical Structures in Databases for Drug Design. In *Drug Design Strategies: Quantitative Approaches*; Livingstone, D. J., Davis, A. M., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2012; pp 164−191.
(468) Warr, W. A. Representation of Chemical Structures. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (4), 557−579.
(469) Ihlenfeldt, W. D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular Structure Elements. *J. Comput. Chem.* **1994**, *15* (8), 793−813.
(470) Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S. InChIKey Collision Resistance: An Experimental Testing. *J. Cheminf.* **2012**, *4* (1), 39.
(471) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5* (2), 107−113.
(472) Smith, E. G. *The Wiswesser Line-Formula Chemical Notation*; McGraw-Hill Book: New York, 1968.
(473) SMARTS - A Language for Describing Molecular Patterns. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Oct 20, 2016).
(474) Leach, A. R.; Bradshaw, J.; Green, D. V; Hann, M. M.; Delany, J. J. Implementation of a System for Reagent Selection and Library Enumeration, Profiling, and Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1161−1172.
(475) Ash, S.; Cline, M. A.; Homer, R. W.; Hurst, T.; Smith, G. B. SYBYL Line Notation (SLN): A Versatile Language for Chemical Structure Representation. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 71−79.
(476) Homer, R. W.; Swanson, J.; Jilek, R. J.; Hurst, T.; Clark, R. D. SYBYL Line Notation (SLN): A Single Notation to Represent Chemical Structures, Queries, Reactions, and Virtual Libraries. *J. Chem. Inf. Model.* **2008**, *48* (12), 2294−2307.
(477) Rohbeck, H.-G. Representation of Structure Description Arranged Linearly. In *Software Development in Chemistry 5*; Gmehling, J., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, Germany, 1991; pp 49−58.
(478) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, *32* (3), 244−255.
(479) V3000 molfile format. http://www.ccl.net/chemistry/resources/messages/2002/12/05.005-dir/index.html (accessed Oct 20, 2016).
(480) Casher, O.; Chandramohan, G. K.; Hargreaves, M. J.; Leach, C.; Murray-Rust, P.; Rzepa, H. S.; Sayle, R.; Whitaker, B. J. Hyperactive Molecules and the World-Wide-Web Information System. *J. Chem. Soc., Perkin Trans. 2* **1995**, No. No. 1, 7−11.
(481) The IUPAC International Chemical Identifier (InChI). http://www.iupac.org/home/publications/e-resources/inchi.htm (accessed Oct 20, 2016).
(482) The IUPAC International Chemical Identifier (InChI) and its influence on the domain of chemical information. http://www.jcheminf.com/series/InChI (accessed Oct 20, 2016).
(483) Warr, W. A. Many InChIs and Quite Some Feat. *J. Comput.-Aided Mol. Des.* **2015**, *29* (8), 681−694.
(484) Grethe, G.; Goodman, J. M.; Allen, C. H. International Chemical Identifier for Reactions (RInChI). *J. Cheminf.* **2013**, *5* (1), 45.
(485) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50* (7), 1189−1204.
(486) Substance Registration System - Unique Ingredient Identifier (UNII). http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm (accessed Oct 20, 2016).
(487) Muresan, S.; Sitzmann, M.; Southan, C. Mapping between Databases of Compounds and Protein Targets. *Methods Mol. Biol.* **2012**, *910*, 145−164.
(488) Kenny, P. W.; Sadowski, J. Chapter 11. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery (Methods & Principles in Medicinal Chemistry 23)*; Oprea, T. I., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2006; pp 271−285.
(489) Gobbi, A.; Lee, M.-L. Handling of Tautomerism and Stereochemistry in Compound Registration. *J. Chem. Inf. Model.* **2012**, *52* (2), 285−292.
(490) Karapetyan, K.; Batchelor, C.; Sharpe, D.; Tkachenko, V.; Williams, A. J. The Chemical Validation and Standardization Platform (CVSP): Large-Scale Automated Validation of Chemical Structure Datasets. *J. Cheminf.* **2015**, *7*, 30.
(491) PubChem Standardization Service. https://pubchem.ncbi.nlm.nih.gov/standardize/standardize.cgi (accessed Oct 20, 2016).
(492) Sayle, R. A. So You Think You Understand Tautomerism? *J. Comput.-Aided Mol. Des.* **2010**, *24* (6−7), 485−496.
(493) Cummings, M. D.; Arnoult, É.; Buyck, C.; Tresadern, G.; Vos, A. M.; Wegner, J. K. Preparing and Filtering Compound Databases for Virtual and Experimental Screening. In *Virtual Screening: Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2011; pp 35−59.
(494) Oellien, F.; Cramer, J.; Beyer, C.; Ihlenfeldt, W.-D.; Selzer, P. M. The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2342−2354.
(495) Sitzmann, M.; Ihlenfeldt, W.-D.; Nicklaus, M. C. Tautomerism in Large Databases. *J. Comput.-Aided Mol. Des.* **2010**, *24* (6−7), 521−551.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(496) Warr, W. A. Tautomerism in Chemical Information Management Systems. *J. Comput.-Aided Mol. Des.* **2010**, *24* (6−7), 497−520.
(497) Martin, Y. C. Overview of the Perspectives Devoted to Tautomerism in Molecular Design. *J. Comput.-Aided Mol. Des.* **2010**, *24* (6−7), 473−474.
(498) Urbano-Cuadrado, M.; Rabal, O.; Oyarzabal, J. Centralizing Discovery Information: From Logistics to Knowledge at a Public Organization. *Comb. Chem. High Throughput Screening* **2011**, *14* (6), 429−449.
(499) Batchelor, C.; Karapetyan, K.; Sharpe, D.; Tkachenko, V.; Williams, A. Carbohydrate Structure Representation and Public Chemistry Databases. Presented at the 245th National Meeting & Exposition of the American Chemical Society, New Orleans, LA, April 7−11, 2013; Paper CARB-110.
(500) Brecher, J. Graphical Representation Standards for Chemical Structure Diagrams (IUPAC Recommendations 2008). *Pure Appl. Chem.* **2008**, *80* (2), 277−410.
(501) BIOVIA. Cheshire. http://accelrys.com/products/pdf/cheshire.pdf (accessed Oct 20, 2016).
(502) ChemAxon. Standardizer. https://www.chemaxon.com/products/standardizer (accessed Oct 20, 2016).
(503) SD File Processing with MOE Pipeline Tools. https://www.chemcomp.com/journal/sdtools.htm (accessed Oct 20, 2016).
(504) BIOVIA Pipeline Pilot Overview. http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot (accessed Oct 20, 2016).
(505) Knime. https://www.knime.org (accessed Oct 20, 2016).
(506) Ray, L. C.; Kirsch, R. A. Finding Chemical Records by Digital Computers. *Science* **1957**, *126* (3278), 814−819.
(507) CAS History. https://www.cas.org/about-cas/cas-history (accessed Oct 20, 2016).
(508) Lynch, M. F.; Willett, P. The Automatic Detection of Chemical Reaction Sites. *J. Chem. Inf. Model.* **1978**, *18* (3), 154−159.
(509) Willett, P. From Chemical Documentation to Chemoinformatics: 50 Years of Chemical Information Science. *J. Inf. Sci.* **2008**, *34* (4), 477−499.
(510) Agrafiotis, D. K.; Alex, S.; Dai, H.; Derkinderen, A.; Farnum, M.; Gates, P.; Izrailev, S.; Jaeger, E. P.; Konstant, P.; Leung, A.; et al. Advanced Biological and Chemical Discovery (ABCD): Centralizing Discovery Knowledge in an Inherently Decentralized World. *J. Chem. Inf. Model.* **2007**, *47* (6), 1999−2014.
(511) Rojnuckarin, A.; Gschwend, D. A.; Rotstein, S. H.; Hartsough, D. S. ArQiologist: An Integrated Decision Support Tool for Lead Optimization. *J. Chem. Inf. Model.* **2005**, *45* (1), 2−9.
(512) The Novartis Avalon Datawarehouse Project. http://www.daylight.com/meetings/emug00/Rohde (accessed Oct 20, 2016).
(513) Sander, T.; Freyss, J.; von Korff, M.; Reich, J. R.; Rufener, C. OSIRIS, an Entirely in-House Developed Drug Discovery Informatics System. *J. Chem. Inf. Model.* **2009**, *49* (2), 232−246.
(514) Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H. Making Every SAR Point Count: The Development of Chemistry Connect for the Large-Scale Integration of Structure and Bioactivity Data. *Drug Discovery Today* **2011**, *16* (23−24), 1019−1030.
(515) Pistoia Alliance. Standardised data warehouses. https://main.qmarkets.org/live/pistoia/node/1355 (accessed Oct 20, 2016).
(516) Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, *284* (5), 34−43.
(517) Slater, T.; Bouton, C.; Huang, E. S. Beyond Data Integration. *Drug Discovery Today* **2008**, *13* (13−14), 584−589.
(518) Frey, J. G. The Value of the Semantic Web in the Laboratory. *Drug Discovery Today* **2009**, *14* (11−12), 552−561.
(519) Casher, O.; Rzepa, H. S. SemanticEye: A Semantic Web Application to Rationalize and Enhance Chemical Electronic Publishing. *J. Chem. Inf. Model.* **2006**, *46* (6), 2396−2411.
(520) Taylor, K. R.; Gledhill, R. J.; Essex, J. W.; Frey, J. G.; Harris, S. W.; De Roure, D. C. Bringing Chemical Data onto the Semantic Web. *J. Chem. Inf. Model.* **2006**, *46* (3), 939−952.
(521) RSC Semantic publishing. http://www.rsc.org (accessed Oct 20, 2016).
(522) Microsoft Research.oreChem Project. http://research.microsoft.com/en-us/projects/orechem (accessed Oct 20, 2016).
(523) Crystallography Open Database. http://www.crystallography.net (accessed Oct 20, 2016).
(524) Open PHACTS. https://www.openphacts.org (accessed Oct 20, 2016).
(525) Tkachenko, V.; Williams, A. J.; Pshenichnov, A.; Karapetyan, K.; Batchelor, C.; Steele, J.; Day, A.; Sharpe, D. ChemSpider Compound Database as One of the Pillars of a Semantic Web for Chemistry. Presented at the 244th National Meeting & Exposition of the American Chemical Society, Philadelphia, PA, August 19−23, 2012; paper CINF-106.
(526) Willighagen, E. L.; Waagmeester, A.; Spjuth, O.; Ansell, P.; Williams, A. J.; Tkachenko, V.; Hastings, J.; Chen, B.; Wild, D. J. The ChEMBL Database as Linked Open Data. *J. Cheminf.* **2013**, *5* (1), 23.
(527) Fu, G.; Batchelor, C.; Dumontier, M.; Hastings, J.; Willighagen, E.; Bolton, E. PubChemRDF: Towards the Semantic Annotation of PubChem Compound and Substance Databases. *J. Cheminf.* **2015**, *7*, 34.
(528) Mons, B.; van Haagen, H.; Chichester, C.; Hoen, P.-B. 't; den Dunnen, J. T.; van Ommen, G.; van Mulligen, E.; Singh, B.; Hooft, R.; Roos, M.; et al. The Value of Data. *Nat. Genet.* **2011**, *43* (4), 281−283.
(529) Journal of Cheminformatics, Thematic series: RDF technologies in chemistry. http://www.jcheminf.com/series/acsrdf2010 (accessed Oct 20, 2016).
(530) Agrafiotis, D. K.; Lobanov, V. S.; Shemanarev, M.; Rassokhin, D. N.; Izrailev, S.; Jaeger, E. P.; Alex, S.; Farnum, M. Efficient Substructure Searching of Large Chemical Libraries: The ABCD Chemical Cartridge. *J. Chem. Inf. Model.* **2011**, *51* (12), 3113−3130.
(531) Jensen, J. H.; Hoeg-Jensen, T.; Padkjaer, S. B. Building a BioChemformatics Database. *J. Chem. Inf. Model.* **2008**, *48* (12), 2404−2413.
(532) Rijnbeek, M.; Steinbeck, C. OrChem - An Open Source Chemistry Search Engine for Oracle(R). *J. Cheminf.* **2009**, *1* (1), 17.
(533) Kiener, J. Molecule Database Framework: A Framework for Creating Database Applications with Chemical Structure Search Capability. *J. Cheminf.* **2013**, *5* (1), 48.
(534) Compound Registration - ChemAxon. https://www.chemaxon.com/products/compound-registration (accessed Oct 20, 2016).
(535) McGregor, J. J.; Willett, P. Use of a Maximum Common Subgraph Algorithm in the Automatic Identification of Ostensible Bond Changes Occurring in Chemical Reactions. *J. Chem. Inf. Model.* **1981**, *21* (3), 137−140.
(536) *Modern Approaches to Chemical Reaction Searching: Proceedings*; Willett, P., Ed.; Gower Publishing Co.: Aldershot, UK, 1986.
(537) Chen, L.; Nourse, J. G.; Christie, B. D.; Leland, B. A.; Grier, D. L. Over 20 Years of Reaction Access Systems from MDL: A Novel Reaction Substructure Search Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1296−1310.
(538) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic Reaction Mapping and Reaction Center Detection. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3* (6), 560−593.
(539) Taylor, K. T. The Status of Electronic Laboratory Notebooks for Chemistry and Biology. *Curr. Opin. Drug Discovery Dev.* **2006**, *9*, 348−353.
(540) Googling for INChIs; A remarkable method of chemical searching. http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019 (accessed Oct 20, 2016).
(541) Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web through the Use of InChI Identifiers. *Org. Biomol. Chem.* **2005**, *3* (10), 1832−1834.
(542) Southan, C. InChI in the Wild: An Assessment of InChIKey Searching in Google. *J. Cheminf.* **2013**, *5* (1), 10.
(543) Murray-Rust, P.; Rzepa, H. S. Towards the Chemical Semantic Web. An Introduction to RSS. *Internet J. Chem.* **2003**, *6*, 4.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(544) Frey, J. G.; Bird, C. L. Cheminformatics and the Semantic Web: Adding Value with Linked Data and Enhanced Provenance. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3* (5), 465−481.
(545) Miller, M. A. Chemical Database Techniques in Drug Discovery. *Nat. Rev. Drug Discovery* **2002**, *1* (3), 220−227.
(546) Willett, P. Chemoinformatics: A History. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (1), 46−56.
(547) Willett, P. Matching of Chemical and Biological Structures Using Subgraph and Maximal Common Subgraph Isomorphism Algorithms. *Rational Drug Design*; Springer: New York, 1999; pp 11−38.
(548) Currano, J. N. Chapter 5 Searching by Structure and Substructure. In *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2014; pp 109−145.
(549) Wrublewski, D. T. Chapter 8 Searching For Polymers. In *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2014; pp 206−223.
(550) Faber, J.; Needham, F. The New Organic Powder Diffraction File: Applications for Polymorph and Search-Indexing. *Am. Pharm. Rev.* **2002**, *5*, 70−75.
(551) Ertl, P. Molecular Structure Input on the Web. *J. Cheminf.* **2010**, *2* (1), 1.
(552) Proschak, E.; Wegner, J. K.; Schüller, A.; Schneider, G.; Fechner, U. Molecular Query Language (MQL)−a Context-Free Grammar for Substructure Matching. *J. Chem. Inf. Model.* **2007**, *47* (2), 295−301.
(553) Golovin, A.; Henrick, K. Chemical Substructure Search in SQL. *J. Chem. Inf. Model.* **2009**, *49* (1), 22−27.
(554) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons: Hoboken, NJ, 1996; Vol. 7, pp 1−66.
(555) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (6), 983−996.
(556) Nasr, R.; Vernica, R.; Li, C.; Baldi, P. Speeding up Chemical Searches Using the Inverted Index: The Convergence of Chemoinformatics and Text Search Methods. *J. Chem. Inf. Model.* **2012**, *52* (4), 891−900.
(557) Currano, J. N. Chapter 9 Reaction Searching. In *Chemical Information for Chemists: A Primer*; Currano, J. N., Roth, D. L., Eds.; The Royal Society of Chemistry: Cambridge, UK, 2014; pp 224−254.
(558) Reactions - CASREACT - Answers to your chemical reaction questions. https://www.cas.org/content/reactions (accessed Oct 20, 2016).
(559) Ridley. Strategies for Chemical Reaction Searching in SciFinder. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (5), 1077−1084.
(560) Downs, G. M.; Barnard, J. M. Chemical Patent Information Systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (5), 727−741.
(561) Simmons, E. S. Central Patents Index Chemical Code: A User's Viewpoint. *J. Chem. Inf. Model.* **1984**, *24* (1), 10−15.
(562) Rössler, S.; Kolb, A. The GREMAS System, an Integral Part of the IDC System for Chemical Documentation. *J. Chem. Doc.* **1970**, *10* (2), 128−134.
(563) Markush - MARPAT - Database containing the keys to substances in patents. https://www.cas.org/content/markush (accessed Oct 20, 2016).
(564) Fisanick, W. The Chemical Abstract's Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Model.* **1990**, *30* (2), 145−154.
(565) Benichou, P.; Klimczak, C.; Borne, P. Handling Genericity in Chemical Structures Using the Markush DARC Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37* (1), 43−53.
(566) Barnard, J. M.; Downs, G. M. Recent and Current Developments in Handling Markush Structures from Chemical Patents. *J. Cheminf.* **2012**, *4* (Suppl 1), O18.
(567) CAS REGISTRY - The gold standard for chemical substance information. https://www.cas.org/content/chemical-substances (accessed Oct 20, 2016).
(568) ChemSpider. http://www.chemspider.com (accessed Oct 20, 2016).
(569) Pence, H. E.; Williams, A.; ChemSpider. An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87* (11), 1123−1124.
(570) PubChem. http://pubchem.ncbi.nlm.nih.gov (accessed Oct 20, 2016).
(571) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. Chapter 12 − PubChem: Integrated Platform of Small Molecules and Biological Activities. *Annu. Rep. Comput. Chem.* **2008**, *4*, 217−241.
(572) DrugBank Drug & Drug Target Database. http://www.drugbank.ca (accessed Oct 20, 2016).
(573) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−7 (Database issue).
(574) Tomasulo, P. ChemIDplus-Super Source for Chemical and Drug Information. *Med. Ref. Serv. Q.* **2002**, *21* (1), 53−59.
(575) ChEMBL. https://www.ebi.ac.uk/chembl (accessed Oct 20, 2016).
(576) 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. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−1090 (Database issue).
(577) ChemBank. http://chembank.broadinstitute.org (accessed Oct 20, 2016).
(578) Seiler, K. P.; George, G. A.; Happ, M. P.; Bodycombe, N. E.; Carrinski, H. A.; Norton, S.; Brudz, S.; Sullivan, J. P.; Muhlich, J.; Serrano, M.; et al. ChemBank: A Small-Molecule Screening and Cheminformatics Resource Database. *Nucleic Acids Res.* **2008**, *36*, D351−9 (Database issue).
(579) eMolecules. https://www.emolecules.com (accessed Oct 20, 2016).
(580) ZINC. http://zinc.docking.org (accessed Oct 20, 2016).
(581) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52* (7), 1757−1768.
(582) IBM BAO strategic IP insight platform (SIIP). http://www-935.ibm.com/services/us/gbs/bao/siip (accessed Oct 20, 2016).
(583) All that glisters is not gold: Quality of Public Domain Chemistry Databases. http://blogs.scientificamerican.com/guest-blog/all-that-glisters-is-not-gold-quality-of-public-domain-chemistry-databases (accessed Oct 20, 2016).
(584) Hettne, K. M.; Williams, A. J.; van Mulligen, E. M.; Kleinjans, J.; Tkachenko, V.; Kors, J. A. Automatic vs. Manual Curation of a Multi-Source Chemical Dictionary: The Impact on Text Mining. *J. Cheminf.* **2010**, *2* (1), 3.
(585) Index Chemicus. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/index-chemicus.html (accessed Oct 20, 2016).
(586) Derwent World Patents Index. Contents. http://thomsonreuters.com/content/dam/openweb/documents/pdf/intellectual-property/fact-sheet/derwent-world-patents-index.pdf (accessed Oct 20, 2016).
(587) Nebel, A.; Olbrich, G.; Deplanque, R. The Gmelin Information System the Connection between Handbook and Database. In *Software Development in Chemistry 4: Proceedings of the 4th Workshop "Computers in Chemistry" Hochfilzen, Tyrol, November 22−24, 1989*; Gasteiger, J., Ed.; Springer Berlin Heidelberg: Berlin, Heidelberg, 1990; pp 51−56.
(588) IBM strategic IP insight platform and the National Institutes of Health. http://www-935.ibm.com/services/us/gbs/bao/siip/nih (accessed Oct 20, 2016).
(589) Ursu, O.; Holmes, J.; Knockel, J.; Bologa, C. G.; Yang, J. J.; Mathias, S. L.; Nelson, S. J.; Oprea, T. I. DrugCentral: Online Drug Compendium. *Nucleic Acids Res.* **2017**, *45*, D932.
(590) DrugCentral. http://drugcentral.org/ (accessed Oct 20, 2016).

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(591) BIOVIA Available Chemicals Directory (ACD). http://accelrys.com/products/collaborative-science/databases/sourcing-databases/biovia-available-chemicals-directory.html (accessed Oct 20, 2016).
(592) CHEMnetBASE. http://www.chemnetbase.com (accessed Oct 20, 2016).
(593) Current Chemical Reactions. http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/current-chemical-reactions.html (accessed Oct 20, 2016).
(594) SPRESI. http://www.spresi.com/ (accessed Oct 20, 2016).
(595) ChemReact. http://infochem.de/products/databases/chemreact41.shtml (accessed Oct 20, 2016).
(596) Science of Synthesis. https://www.thieme.de/en/thieme-chemistry/science-of-synthesis-54780.htm (accessed Oct 20, 2016).
(597) ChemInform Reaction Library. http://www.cheminform.com (accessed Oct 20, 2016).
(598) Selected Organic Reactions Database (SORD). http://www.sord.nl (accessed Oct 20, 2016).
(599) e-EROS Encyclopedia of Reagents for Organic Synthesis. http://onlinelibrary.wiley.com/book/10.1002/047084289X (accessed Oct 20, 2016).
(600) Comprehensive Heterocyclic Chemistry. http://www.sciencedirect.com/science/referenceworks/9780080965192#ancv0020 (accessed Oct 20, 2016).
(601) Synthetic Reaction Updates. http://pubs.rsc.org/lus/synthetic-reaction-updates (accessed Oct 20, 2016).
(602) Zass, E. Databases of Chemical Reactions. In *Handbook of Chemoinformatics: From Data to Knowledge in 4 Volumes*; Gasteiger, J., Ed.; Wiley-VCH Verlag: Weinheim, Germany, 2003; pp 667−699.
(603) Oprea, T. I.; Nielsen, S. K.; Ursu, O.; Yang, J. J.; Taboureau, O.; Mathias, S. L.; Kouskoumvekaki, L.; Sklar, L. A.; Bologa, C. G. Associating Drugs, Targets and Clinical Outcomes into an Integrated Network Affords a New Platform for Computer-Aided Drug Repurposing. *Mol. Inf.* **2011**, *30* (2−3), 100−111.
(604) Kringelum, J.; Kjaerulff, S. K.; Brunak, S.; Lund, O.; Oprea, T. I.; Taboureau, O. ChemProt-3.0: A Global Chemical Biology Diseases Mapping. *Database* **2016**, *2016*, bav123.
(605) Tari, L. B.; Patel, J. H. Systematic Drug Repurposing through Text Mining. *Methods Mol. Biol.* **2014**, *1159*, 253−267.
(606) Ali, I.; Guo, Y.; Silins, I.; Högberg, J.; Stenius, U.; Korhonen, A. Grouping Chemicals for Health Risk Assessment: A Text Mining-Based Case Study of Polychlorinated Biphenyls (PCBs). *Toxicol. Lett.* **2016**, *241*, 32−37.
(607) Korhonen, A.; Séaghdha, D. O.; Silins, I.; Sun, L.; Högberg, J.; Stenius, U. Text Mining for Literature Review and Knowledge Discovery in Cancer Risk Assessment and Research. *PLoS One* **2012**, *7* (4), e33427.
(608) Williams, A. J. A Perspective of Publicly Accessible/open-Access Chemistry Databases. *Drug Discovery Today* **2008**, *13* (11−12), 495−501.
(609) Nicola, G.; Liu, T.; Gilson, M. K. Public Domain Databases for Medicinal Chemistry. *J. Med. Chem.* **2012**, *55* (16), 6987−7002.
(610) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; et al. A Quantitative Analysis of Kinase Inhibitor Selectivity. *Nat. Biotechnol.* **2008**, *26* (1), 127−132.
(611) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; et al. Thousands of Chemical Starting Points for Antimalarial Lead Identification. *Nature* **2010**, *465* (7296), 305−310.
(612) Ang, K. K. H.; Ratnam, J.; Gut, J.; Legac, J.; Hansell, E.; Mackey, Z. B.; Skrzypczynska, K. M.; Debnath, A.; Engel, J. C.; Rosenthal, P. J.; et al. Mining a Cathepsin Inhibitor Library for New Antiparasitic Drug Leads. *PLoS Neglected Trop. Dis.* **2011**, *5* (5), e1023.
(613) Lee, J. A.; Chu, S.; Willard, F. S.; Cox, K. L.; Sells Galvin, R. J.; Peery, R. B.; Oliver, S. E.; Oler, J.; Meredith, T. D.; Heidler, S. A.; et al. Open Innovation for Phenotypic Drug Discovery: The PD2 Assay Panel. *J. Biomol. Screening* **2011**, *16* (6), 588−602.
(614) Ballell, L.; Bates, R. H.; Young, R. J.; Alvarez-Gomez, D.; Alvarez-Ruiz, E.; Barroso, V.; Blanco, D.; Crespo, B.; Escribano, J.; González, R.; et al. Fueling Open-Source Drug Discovery: 177 Small-Molecule Leads against Tuberculosis. *ChemMedChem* **2013**, *8* (2), 313−321.
(615) Agarwal, P.; Searls, D. B. Can Literature Analysis Identify Innovation Drivers in Drug Discovery? *Nat. Rev. Drug Discovery* **2009**, *8* (11), 865−878.
(616) Tiikkainen, P.; Bellis, L.; Light, Y.; Franke, L. Estimating Error Rates in Bioactivity Databases. *J. Chem. Inf. Model.* **2013**, *53* (10), 2499−2505.
(617) Ruusmann, V.; Maran, U. From Data Point Timelines to a Well Curated Data Set, Data Mining of Experimental Data and Chemical Structure Data from Scientific Articles, Problems and Possible Solutions. *J. Comput.-Aided Mol. Des.* **2013**, *27* (7), 583−603.
(618) Williams, A. J.; Ekins, S. A Quality Alert and Call for Improved Curation of Public Chemistry Databases. *Drug Discovery Today* **2011**, *16* (17−18), 747−750.
(619) Fourches, D.; Muratov, E.; Tropsha, A. Curation of Chemogenomics Data. *Nat. Chem. Biol.* **2015**, *11* (8), 535−535.
(620) Orchard, S.; Al-Lazikani, B.; Bryant, S.; Clark, D.; Calder, E.; Dix, I.; Engkvist, O.; Forster, M.; Gaulton, A.; Gilson, M.; et al. Minimum Information about a Bioactive Entity (MIABE). *Nat. Rev. Drug Discovery* **2011**, *10* (9), 661−669.
(621) Visser, U.; Abeyruwan, S.; Vempati, U.; Smith, R. P.; Lemmon, V.; Schürer, S. C. BioAssay Ontology (BAO): A Semantic Description of Bioassays and High-Throughput Screening Results. *BMC Bioinf.* **2011**, *12* (1), 257.
(622) de Souza, A.; Bittker, J. A.; Lahr, D. L.; Brudz, S.; Chatwin, S.; Oprea, T. I.; Waller, A.; Yang, J. J.; Southall, N.; Guha, R.; et al. An Overview of the Challenges in Designing, Integrating, and Delivering BARD: A Public Chemical-Biology Resource and Query Portal for Multiple Organizations, Locations, and Disciplines. *J. Biomol. Screening* **2014**, *19* (5), 614−627.
(623) Blake, J. E.; Dana, R. C. CASREACT: More than a Million Reactions. *J. Chem. Inf. Model.* **1990**, *30* (4), 394−399.
(624) Goodman, J. Computer Software Review: Reaxys. *J. Chem. Inf. Model.* **2009**, *49* (12), 2897−2898.
(625) Roth, D. L. SPRESIweb 2.1, a Selective Chemical Synthesis and Reaction Database. *J. Chem. Inf. Model.* **2005**, *45* (5), 1470−1473.
(626) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Ph.D. Thesis, University of Cambridge, June 2012.
(627) Zamora, E. M.; Blower, P. E., Jr Extraction of Chemical Reaction Information from Primary Journal Text Using Computational Linguistics Techniques. 2. Semantic Phase. *J. Chem. Inf. Model.* **1984**, *24* (3), 181−188.
(628) Postma, G. J.; Kateman, G. A Systematic Representation of Analytical Chemical Actions. *J. Chem. Inf. Model.* **1993**, *33* (3), 350−368.
(629) Intarapaiboon, P.; Nantajeewarawat, E.; Theeramunkong, T. Extracting Chemical Reactions from Thai Text for Semantics-Based Information Retrieval. *IEICE Trans. Inf. Syst.* **2011**, *94* (3), 479−486.
(630) Jensen, L. J.; Saric, J.; Bork, P. Literature Mining for the Biologist: From Information Retrieval to Biological Discovery. *Nat. Rev. Genet.* **2006**, *7* (2), 119−129.
(631) Zweigenbaum, P.; Demner-Fushman, D.; Yu, H.; Cohen, K. B. Frontiers of Biomedical Text Mining: Current Progress. *Briefings Bioinf.* **2007**, *8* (5), 358−375.
(632) Cohen, A. M.; Hersh, W. R. A Survey of Current Work in Biomedical Text Mining. *Briefings Bioinf.* **2005**, *6* (1), 57−71.
(633) Ananiadou, S.; Kell, D. B.; Tsujii, J. Text Mining and Its Potential Applications in Systems Biology. *Trends Biotechnol.* **2006**, *24* (12), 571−579.
(634) Rodriguez-Esteban, R. Biomedical Text Mining and Its Applications. *PLoS Comput. Biol.* **2009**, *5* (12), e1000597.
(635) Tsuruoka, Y.; Tateishi, Y.; Kim, J.-D.; Ohta, T.; McNaught, J.; Ananiadou, S.; Tsujii, J. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics*; Bozanis, P., Houstis, E.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

N., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2005; Vol. *3746*, pp 382−392.
(636) Sagae, K.; Tsujii, J. Dependency Parsing and Domain Adaptation with LR Models and Parser Ensembles. *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*; Prague, Czech Republic, June, 2007; pp 1044−1050.
(637) Miyao, Y.; Tsujii, J. Feature Forest Models for Probabilistic HPSG Parsing. *Comput. Ling.* **2008**, *34* (1), 35−80.
(638) Miwa, M.; Sætre, R.; Miyao, Y.; Tsujii, J. Protein-Protein Interaction Extraction by Leveraging Multiple Kernels and Parsers. *Int. J. Med. Inform.* **2009**, *78* (12), e39−e46.
(639) Abacha, A. B.; Zweigenbaum, P. Medical Entity Recognition: A Comparison of Semantic and Statistical Methods. *Proceedings of BioNLP 2011 Workshop (BioNLP'11)*; Portland, OR, June 24, 2011; pp 56−64.
(640) Goulart, R. R. V.; de Lima, V. L. S.; Xavier, C. C. A Systematic Review of Named Entity Recognition in Biomedical Texts. *J. Brazilian Comput. Soc.* **2011**, *17* (2), 103−116.
(641) Neves, M. An Analysis on the Entity Annotations in Biological Corpora. *F1000Research* **2014**, *3*, 96.
(642) Spasic, I.; Ananiadou, S.; McNaught, J.; Kumar, A. Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. *Briefings Bioinf.* **2005**, *6* (3), 239−251.
(643) Krallinger, M.; Valencia, A.; Hirschman, L. Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology. *Genome Biol.* **2008**, *9* (Suppl 2), S8.
(644) Papanikolaou, N.; Pavlopoulos, G. A.; Theodosiou, T.; Iliopoulos, I. Protein-Protein Interaction Predictions Using Text Mining Methods. *Methods* **2015**, *74*, 47−53.
(645) Zhou, D.; He, Y. Extracting Interactions between Proteins from the Literature. *J. Biomed. Inf.* **2008**, *41* (2), 393−407.
(646) Hoffmann, R.; Valencia, A. A Gene Network for Navigating the Literature. *Nat. Genet.* **2004**, *36* (7), 664.
(647) Blaschke, C.; Valencia, A. The Potential Use of SUISEKI as a Protein Interaction Discovery Tool. *Genome Inform.* **2001**, *12*, 123−134.
(648) Wei, C.-H.; Kao, H.-Y.; Lu, Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. *BioMed Res. Int.* **2015**, *2015*, 1−7.
(649) Huang, M.; Liu, J.; Zhu, X. GeneTUKit: A Software for Document-Level Gene Normalization. *Bioinformatics* **2011**, *27* (7), 1032−1033.
(650) Hakenberg, J.; Gerner, M.; Haeussler, M.; Solt, I.; Plake, C.; Schroeder, M.; Gonzalez, G.; Nenadic, G.; Bergman, C. M. The GNAT Library for Local and Remote Gene Mention Normalization. *Bioinformatics* **2011**, *27* (19), 2769−2771.
(651) Cheng, D.; Knox, C.; Young, N.; Stothard, P.; Damaraju, S.; Wishart, D. S. PolySearch: A Web-Based Text Mining System for Extracting Relationships between Human Diseases, Genes, Mutations, Drugs and Metabolites. *Nucleic Acids Res.* **2008**, *36*, W399−W405 (Web Server Issue).
(652) Wei, C.-H.; Kao, H.-Y.; Lu, Z. PubTator: A Web-Based Text Mining Tool for Assisting Biocuration. *Nucleic Acids Res.* **2013**, *41* (Web Server Issue), W518−W522.
(653) Al-Mubaid, H.; Singh, R. K. A Text-Mining Technique for Extracting Gene-Disease Associations from the Biomedical Literature. *Int. J. Bioinf. Res. Appl.* **2010**, *6* (3), 270−286.
(654) Gonzalez, G.; Uribe, J. C.; Tari, L.; Brophy, C.; Baral, C. Mining Gene-Disease Relationships from Biomedical Literature: Weighting Protein-Protein Interactions and Connectivity Measures. *Pac. Symp. Biocomput.* **2007**, 28−39.
(655) Pinero, J.; Queralt-Rosinach, N.; Bravo, A.; Deu-Pons, J.; Bauer-Mehren, A.; Baron, M.; Sanz, F.; Furlong, L. I. DisGeNET: A Discovery Platform for the Dynamical Exploration of Human Diseases and Their Genes. *Database* **2015**, *2015*, bav028.
(656) Mao, Y.; Van Auken, K.; Li, D.; Arighi, C. N.; McQuilton, P.; Hayman, G. T.; Tweedie, S.; Schaeffer, M. L.; Laulederkind, S. J. F.; Wang, S.-J. Overview of the Gene Ontology Task at BioCreative IV. *Database* **2014**, *2014*, bau086.
(657) Blaschke, C.; Leon, E. A.; Krallinger, M.; Valencia, A. Evaluation of BioCreAtIvE Assessment of Task 2. *BMC Bioinf.* **2005**, *6* (Suppl 1), S16.
(658) Hirschman, L.; Burns, G. A. P. C.; Krallinger, M.; Arighi, C.; Cohen, K. B.; Valencia, A.; Wu, C. H.; Chatr-Aryamontri, A.; Dowell, K. G.; Huala, E.; et al. Text Mining for the Biocuration Workflow. *Database* **2012**, *2012*, bas020.
(659) Winnenburg, R.; Wächter, T.; Plake, C.; Doms, A.; Schroeder, M. Facts from Text: Can Text Mining Help to Scale-up High-Quality Manual Curation of Gene Products with Ontologies? *Briefings Bioinf.* **2008**, *9* (6), 466−478.
(660) Gonzalez, G. H.; Tahsin, T.; Goodale, B. C.; Greene, A. C.; Greene, C. S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings Bioinf.* **2016**, *17* (1), 33−42.
(661) Leser, U.; Hakenberg, J. What Makes a Gene Name? Named Entity Recognition in the Biomedical Literature. *Briefings Bioinf.* **2005**, *6* (4), 357−369.
(662) Kim, J.-D.; Ohta, T.; Tsuruoka, Y.; Tateisi, Y.; Collier, N. Introduction to the Bio-Entity Recognition Task at JNLPBA. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*; Geneva, Switzerland, August 28−29, 2004; pp 70−75.
(663) Wei, C.-H.; Harris, B. R.; Kao, H.-Y.; Lu, Z. tmVar: A Text Mining Approach for Extracting Sequence Variants in Biomedical Literature. *Bioinformatics* **2013**, *29* (11), 1433−1439.
(664) Settles, B. Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA '04)*; Geneva, Switzerland, August 28−29, 2004; pp 104−107.
(665) Kaewphan, S.; Van Landeghem, S.; Ohta, T.; Van de Peer, Y.; Ginter, F.; Pyysalo, S. Cell Line Name Recognition in Support of the Identification of Synthetic Lethality in Cancer from Text. *Bioinformatics* **2016**, *32*, 276−282.
(666) Pyysalo, S.; Ananiadou, S. Anatomical Entity Mention Recognition at Literature Scale. *Bioinformatics* **2014**, *30* (6), 868−875.
(667) Gerner, M.; Nenadic, G.; Bergman, C. M. LINNAEUS: A Species Name Identification System for Biomedical Literature. *BMC Bioinf.* **2010**, *11*, 85.
(668) Bagewadi, S.; Bobić, T.; Hofmann-Apitius, M.; Fluck, J.; Klinger, R. Detecting miRNA Mentions and Relations in Biomedical Literature. *F1000Research* **2014**, *3*, 205.
(669) Li, G.; Ross, K. E.; Arighi, C. N.; Peng, Y.; Wu, C. H.; Vijay-Shanker, K. miRTex: A Text Mining System for miRNA-Gene Relation Extraction. *PLoS Comput. Biol.* **2015**, *11* (9), e1004391.
(670) Doughty, E.; Kertesz-Farkas, A.; Bodenreider, O.; Thompson, G.; Adadey, A.; Peterson, T.; Kann, M. G.; Kann, M. G. Toward an Automatic Method for Extracting Cancer- and Other Disease-Related Point Mutations from the Biomedical Literature. *Bioinformatics* **2011**, *27* (3), 408−415.
(671) Pafilis, E.; Frankild, S. P.; Fanini, L.; Faulwetter, S.; Pavloudi, C.; Vasileiadou, A.; Arvanitidis, C.; Jensen, L. J. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. *PLoS One* **2013**, *8* (6), e65390.
(672) Boyle, B.; Hopkins, N.; Lu, Z.; Raygoza Garay, J. A.; Mozzherin, D.; Rees, T.; Matasci, N.; Narro, M. L.; Piel, W. H.; Mckay, S. J.; et al. The Taxonomic Name Resolution Service: An Online Tool for Automated Standardization of Plant Names. *BMC Bioinf.* **2013**, *14* (1), 16.
(673) Naderi, N.; Kappler, T.; Baker, C. J. O.; Witte, R. OrganismTagger: Detection, Normalization and Grounding of Organism Entities in Biomedical Documents. *Bioinformatics* **2011**, *27* (19), 2721−2729.
(674) Wei, C.-H.; Kao, H.-Y.; Lu, Z. SR4GN: A Species Recognition Software Tool for Gene Normalization. *PLoS One* **2012**, *7* (6), e38460.
(675) Wei, C.-H.; Kao, H.-Y. Cross-Species Gene Normalization by Species Inference. *BMC Bioinf.* **2011**, *12* (Suppl 8), S5.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(676) Tamames, J.; Valencia, A. The Success (or Not) of HUGO Nomenclature. *Genome Biol.* **2006**, *7* (5), 402.
(677) Hirschman, L.; Morgan, A. A.; Yeh, A. S. Rutabaga by Any Other Name: Extracting Biological Names. *J. Biomed. Inf.* **2002**, *35* (4), 247−259.
(678) Pafilis, E.; O'Donoghue, S. I.; Jensen, L. J.; Horn, H.; Kuhn, M.; Brown, N. P.; Schneider, R. Reflect: Augmented Browsing for the Life Scientist. *Nat. Biotechnol.* **2009**, *27* (6), 508−510.
(679) Fukuda, K.; Tamura, A.; Tsunoda, T.; Takagi, T. Toward Information Extraction: Identifying Protein Names from Biological Papers. *Pac. Symp. Biocomput.* **1998**, 707−718.
(680) Leaman, R.; Gonzalez, G. BANNER: An Executable Survey of Advances in Biomedical Named Entity Recognition. *Pac. Symp. Biocomput.* **2008**, 652−663.
(681) Tsuruoka, Y.; Tsujii, J. Bidirectional Inference with the Easiest-First Strategy for Tagging Sequence Data. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*; Vancouver, Canada, October 6−8, 2005; pp 467−474.
(682) Settles, B. ABNER: An Open Source Tool for Automatically Tagging Genes, Proteins and Other Entity Names in Text. *Bioinformatics* **2005**, *21* (14), 3191−3192.
(683) Sasaki, Y.; Montemagni, S.; Pezik, P.; Rebholz-Schuhmann, D.; McNaught, J.; Ananiadou, S. Biolexicon: A Lexical Resource for the Biology Domain. *Proceedings of the Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*; Turku, Finland, September 1−3, 2008; pp 109−116.
(684) Krallinger, M.; Rodriguez-Penagos, C.; Tendulkar, A.; Valencia, A. PLAN2L: A Web Tool for Integrated Text Mining and Literature-Based Bioentity Relation Extraction. *Nucleic Acids Res.* **2009**, *37*, W160−W165 (Web Server Issue).
(685) Hatzivassiloglou, V.; Duboué, P. A.; Rzhetsky, A. Disambiguating Proteins, Genes, and RNA in Text: A Machine Learning Approach. *Bioinformatics* **2001**, *17* (Suppl 1), S97−106.
(686) Collier, N.; Nobata, C.; Tsujii, J. Extracting the Names of Genes and Gene Products with a Hidden Markov Model. *Proceedings of the 18th conference on Computational linguistics - Volume 1 (COLING'00)*; Saarbrücken, Germany, July 31-August 4, 2000; pp 201−207.
(687) Shen, D.; Zhang, J.; Zhou, G.; Su, J.; Tan, C.-L. Effective Adaptation of a Hidden Markov Model-Based Named Entity Recognizer for Biomedical Domain. *Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine - Volume 13 (BioMed'03)*; Sapporo, Japan, July 11, 2003; pp 49−56.
(688) Torii, M.; Hu, Z.; Wu, C. H.; Liu, H. BioTagger-GM: A Gene/protein Name Recognition System. *J. Am. Med. Inf. Assoc.* **2009**, *16* (2), 247−255.
(689) Kazama, J.; Makino, T.; Ohta, Y.; Tsujii, J. Tuning Support Vector Machines for Biomedical Named Entity Recognition. *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical domain - Volume 3 (BioMed '02)*; Philadelphia, PA, July 11, 2002; pp 1−8.
(690) Mika, S.; Rost, B. Protein Names Precisely Peeled off Free Text. *Bioinformatics* **2004**, *20* (Suppl 1), i241−i247.
(691) McDonald, R.; Pereira, F. Identifying Gene and Protein Mentions in Text Using Conditional Random Fields. *BMC Bioinf.* **2005**, *6* (Suppl 1), S6.
(692) Aronson, A. R. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *AMIA Annu. Symp. Proc.* **2001**, 17−21.
(693) Smith, L.; Tanabe, L. K.; Ando, R. J. nee; Kuo, C.-J.; Chung, I.-F.; Hsu, C.-N.; Lin, Y.-S.; Klinger, R.; Friedrich, C. M.; Ganchev, K.; et al. Overview of BioCreative II Gene Mention Recognition. *Genome Biol.* **2008**, *9* (Suppl 2), S2.
(694) Zhou, D.; Zhong, D.; He, Y. Biomedical Relation Extraction: From Binary to Complex. *Comput. Math. Method Med.* **2014**, *2014*, 298473.
(695) Cohen, K. B.; Hunter, L. Getting Started in Text Mining. *PLoS Comput. Biol.* **2008**, *4* (1), e20.
(696) Hobbs, J. R. The Generic Information Extraction System. *Proceedings of the 5th Conference on Message Understanding (MUC5'93)*; Baltimore, MD, August 25−27, 1993; pp 87−91.
(697) Blaschke, C.; Valencia, A. The Frame-Based Module of the SUISEKI Information Extraction System. *IEEE Intell. Syst.* **2002**, *17* (2), 14−20.
(698) Spasić, I.; Sarafraz, F.; Keane, J. A.; Nenadić, G. Medication Information Extraction with Linguistic Pattern Matching and Semantic Rules. *J. Am. Med. Inf. Assoc.* **2010**, *17* (5), 532−535.
(699) Hu, Z.-Z.; Narayanaswamy, M.; Ravikumar, K. E.; Vijay-Shanker, K.; Wu, C. H. Literature Mining and Database Annotation of Protein Phosphorylation Using a Rule-Based System. *Bioinformatics* **2005**, *21* (11), 2759−2765.
(700) Wei, C.-H.; Peng, Y.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Li, J.; Wiegers, T. C.; Lu, Z. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*; Sevilla, Spain, September 9−11, 2015; pp 154−166.
(701) Segura-Bedmar, I.; Martínez, P.; Herrero-Zazo, M. Lessons Learnt from the DDIExtraction-2013 Shared Task. *J. Biomed. Inf.* **2014**, *51*, 152−164.
(702) Rebholz-Schuhmann, D.; Kirsch, H.; Arregui, M.; Gaudan, S.; Riethoven, M.; Stoehr, P. EBIMed–Text Crunching to Gather Facts for Proteins from Medline. *Bioinformatics* **2007**, *23* (2), e237−44.
(703) Crasto, C.; Luo, D.; Yu, F.; Forero, A.; Chen, D. GenDrux: A Biomedical Literature Search System to Identify Gene Expression-Based Drug Sensitivity in Breast Cancer. *BMC Med. Inf. Decis. Making* **2011**, *11*, 28.
(704) Garten, Y.; Altman, R. B. Pharmspresso: A Text Mining Tool for Extraction of Pharmacogenomic Concepts and Relationships from Full Text. *BMC Bioinf.* **2009**, *10* (Suppl 2), S6.
(705) Garten, Y.; Tatonetti, N. P.; Altman, R. B. Improving the Prediction of Pharmacogenes Using Text-Derived Drug-Gene Relationships. *Pac. Symp. Biocomput.* **2010**, 305−314.
(706) Li, J.; Zhu, X.; Chen, J. Y. Building Disease-Specific Drug-Protein Connectivity Maps from Molecular Interaction Networks and PubMed Abstracts. *PLoS Comput. Biol.* **2009**, *5* (7), e1000450.
(707) Liu, Y.; Liang, Y.; Wishart, D. PolySearch2: A Significantly Improved Text-Mining System for Discovering Associations between Human Diseases, Genes, Drugs, Metabolites, Toxins and More. *Nucleic Acids Res.* **2015**, *43* (W1), W535−42.
(708) Holzinger, A.; Yildirim, P.; Geier, M.; Simonic, K.-M. Quality-Based Knowledge Discovery from Medical Text on the Web. In *Quality Issues in the Management of Web Information*; Pasi, G., Bordogna, G., Jain, L. C., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, Germany, 2013; pp 145−158.
(709) Tsuruoka, Y.; Tsujii, J.; Ananiadou, S. FACTA: A Text Search Engine for Finding Associated Biomedical Concepts. *Bioinformatics* **2008**, *24* (21), 2559−2560.
(710) Wermter, J.; Hahn, U. You Can't Beat Frequency (Unless You Use Linguistic Knowledge): A Qualitative Evaluation of Association Measures for Collocation and Term Extraction. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44)*; Sydney, Australia, July 17−18, 2006; pp 785−792.
(711) Chen, E. S.; Hripcsak, G.; Xu, H.; Markatou, M.; Friedman, C. Automated Acquisition of Disease–Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *J. Am. Med. Inf. Assoc.* **2008**, *15* (1), 87−98.
(712) Kuhn, M.; Szklarczyk, D.; Pletscher-Frankild, S.; Blicher, T. H.; Von Mering, C.; Jensen, L. J.; Bork, P. STITCH 4: Integration of Protein-Chemical Interactions with User Data. *Nucleic Acids Res.* **2014**, *42* (Database Issue), D401−D407.
(713) Li, C.; Jimeno-Yepes, A.; Arregui, M.; Kirsch, H.; Rebholz-Schuhmann, D. PCorral--Interactive Mining of Protein Interactions from MEDLINE. *Database* **2013**, *2013*, bat030.10.1093/database/bat030
(714) Hakenberg, J.; Voronov, D.; Nguyên, V. H.; Liang, S.; Anwar, S.; Lumpkin, B.; Leaman, R.; Tari, L.; Baral, C. A SNPshot of PubMed

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

to Associate Genetic Variants with Drugs, Diseases, and Adverse Reactions. *J. Biomed. Inf.* **2012**, *45* (5), 842−850.
(715) Li, J.; Lu, Z. Systematic Identification of Pharmacogenomics Information from Clinical Trials. *J. Biomed. Inf.* **2012**, *45* (5), 870−878.
(716) Whirl-Carrillo, M.; McDonagh, E. M.; Hebert, J. M.; Gong, L.; Sangkuhl, K.; Thorn, C. F.; Altman, R. B.; Klein, T. E. Pharmacogenomics Knowledge for Personalized Medicine. *Clin. Pharmacol. Ther.* **2012**, *92* (4), 414−417.
(717) Xu, R.; Wang, Q. Comparing a Knowledge-Driven Approach to a Supervised Machine Learning Approach in Large-Scale Extraction of Drug-Side Effect Relationships from Free-Text Biomedical Literature. *BMC Bioinf.* **2015**, *16* (Suppl 5), S6.
(718) Segura-Bedmar, I.; Martínez, P.; de Pablo-Sánchez, C. Using a Shallow Linguistic Kernel for Drug−Drug Interaction Extraction. *J. Biomed. Inf.* **2011**, *44* (5), 789−804.
(719) Bundschus, M.; Dejori, M.; Stetter, M.; Tresp, V.; Kriegel, H.-P. Extraction of Semantic Biomedical Relations from Text Using Conditional Random Fields. *BMC Bioinf.* **2008**, *9*, 207.
(720) Chowdhury, M. F. M.; Lavelli, A. Drug-Drug Interaction Extraction Using Composite Kernels. *Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011)*; Huelva, Spain, September, 2011; pp 27−33.
(721) Bunescu, R. C.; Mooney, R. J. A Shortest Path Dependency Kernel for Relation Extraction. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*; Vancouver, Canada, October 6−8, 2005; pp 724−731.
(722) Cohen, R.; Elhadad, M. Syntactic Dependency Parsers for Biomedical-NLP. *AMIA Annu. Symp. Proc.* **2012**, *2012*, 121−128.
(723) Thessen, A. E.; Cui, H.; Mozzherin, D. Applications of Natural Language Processing in Biodiversity Science. *Adv. Bioinf.* **2012**, *2012*, 391574.
(724) Segura-Bedmar, I.; Martínez, P.; de Pablo-Sánchez, C. Extracting Drug-Drug Interactions from Biomedical Texts. *BMC Bioinf.* **2010**, *11* (Suppl5), P9.
(725) Vanegas, J. A.; Matos, S.; González, F.; Oliveira, J. L. An Overview of Biomolecular Event Extraction from Scientific Documents. *Comput. Math. Method Med.* **2015**, *2015*, 571381.
(726) Zhang, Y.; Lin, H.; Yang, Z.; Wang, J.; Li, Y. Biomolecular Event Trigger Detection Using Neighborhood Hash Features. *J. Theor. Biol.* **2013**, *318*, 22−28.
(727) Campos, D.; Bui, Q.-C.; Matos, S.; Oliveira, J. L. TrigNER: Automatically Optimized Biomedical Event Trigger Recognition on Scientific Documents. *Source Code Biol. Med.* **2014**, *9* (1), 1.
(728) Agarwal, P.; Searls, D. B. Literature Mining in Support of Drug Discovery. *Briefings Bioinf.* **2008**, *9* (6), 479−492.
(729) Craven, M.; Kumlien, J. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB-99)*; Heidelberg, Germany, August 6−10, 1999; pp 77−86.
(730) Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E. G.; Gewiess, A.; Jensen, L. J.; et al. SuperTarget and Matador: Resources for Exploring Drug-Target Relationships. *Nucleic Acids Res.* **2008**, *36*, D919−22 (Database issue).
(731) Hecker, N.; Ahmed, J.; von Eichborn, J.; Dunkel, M.; Macha, K.; Eckert, A.; Gilson, M. K.; Bourne, P. E.; Preissner, R. SuperTarget Goes Quantitative: Update on Drug-Target Interactions. *Nucleic Acids Res.* **2012**, *40*, D1113−7 (Database issue).
(732) Buyko, E.; Beisswanger, E.; Hahn, U. The Extraction of Pharmacogenetic and Pharmacogenomic Relations–a Case Study Using PharmGKB. *Pac. Symp. Biocomput.* **2012**, 376−387.
(733) Coulet, A.; Shah, N. H.; Garten, Y.; Musen, M.; Altman, R. B. Using Text to Build Semantic Networks for Pharmacogenomics. *J. Biomed. Inf.* **2010**, *43* (6), 1009−1019.
(734) Xu, R.; Wang, Q. A Knowledge-Driven Conditional Approach to Extract Pharmacogenomics Specific Drug-Gene Relationships from Free Text. *J. Biomed. Inf.* **2012**, *45* (5), 827−834.
(735) Pakhomov, S.; McInnes, B. T.; Lamba, J.; Liu, Y.; Melton, G. B.; Ghodke, Y.; Bhise, N.; Lamba, V.; Birnbaum, A. K. Using PharmGKB to Train Text Mining Approaches for Identifying Potential Gene Targets for Pharmacogenomic Studies. *J. Biomed. Inf.* **2012**, *45* (5), 862−869.
(736) Ahlers, C. B.; Fiszman, M.; Demner-Fushman, D.; Lang, F.-M.; Rindflesch, T. C. Extracting Semantic Predications from Medline Citations for Pharmacogenomics. *Pac. Symp. Biocomput.* **2007**, 209−220.
(737) Tari, L.; Hakenberg, J.; Gonzalez, G.; Baral, C. Querying Parse Tree Database of Medline Text to Synthesize User-Specific Biomolecular Networks. *Pac. Symp. Biocomput.* **2009**, 87−98.
(738) Percha, B.; Altman, R. B. Learning the Structure of Biomedical Relationships from Unstructured Text. *PLoS Comput. Biol.* **2015**, *11* (7), e1004216.
(739) Yamashita, F.; Feng, C.; Yoshida, S.; Itoh, T.; Hashida, M. Automated Information Extraction and Structure-Activity Relationship Analysis of Cytochrome P450 Substrates. *J. Chem. Inf. Model.* **2011**, *51* (2), 378−385.
(740) Jiao, D.; Wild, D. J. Extraction of CYP Chemical Interactions from Biomedical Literature Using Natural Language Processing Methods. *J. Chem. Inf. Model.* **2009**, *49* (2), 263−269.
(741) Preissner, S.; Kroll, K.; Dunkel, M.; Senger, C.; Goldsobel, G.; Kuzman, D.; Guenther, S.; Winnenburg, R.; Schroeder, M.; Preissner, R. SuperCYP: A Comprehensive Database on Cytochrome P450 Enzymes Including a Tool for Analysis of CYP-Drug Interactions. *Nucleic Acids Res.* **2010**, *38*, D237−43 (Database issue).
(742) Feng, C.; Yamashita, F.; Hashida, M. Automated Extraction of Information from the Literature on Chemical-CYP3A4 Interactions. *J. Chem. Inf. Model.* **2007**, *47* (6), 2449−2455.
(743) Ye, H.; Ye, L.; Kang, H.; Zhang, D.; Tao, L.; Tang, K.; Liu, X.; Zhu, R.; Liu, Q.; Chen, Y. Z.; et al. HIT: Linking Herbal Active Ingredients to Targets. *Nucleic Acids Res.* **2011**, *39*, D1055−D1059 (Database issue).
(744) Chan, W. K. B.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Özgür, A.; Zhang, Y. GLASS: A Comprehensive Database for Experimentally Validated GPCR-Ligand Associations. *Bioinformatics* **2015**, *31* (18), 3035−3042.
(745) Humphreys, K.; Demetriou, G.; Gaizauskas, R. Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures. *Pac. Symp. Biocomput.* **2000**, 505−516.
(746) Czarnecki, J.; Nobeli, I.; Smith, A. M.; Shepherd, A. J. A Text-Mining System for Extracting Metabolic Reactions from Full-Text Articles. *BMC Bioinf.* **2012**, *13*, 172.
(747) Kongburan, W.; Padungweang, P.; Krathu, W.; Chan, J. H. Metabolite Named Entity Recognition: A Hybrid Approach. *Proceedings of the 23rd International Conference on Neural Information Processing (ICONIP 2016)*; Kyoto, Japan, October 16−21, 2016; pp 451−460.
(748) Patumcharoenpol, P.; Doungpan, N.; Meechai, A.; Shen, B.; Chan, J. H.; Vongsangnak, W. An Integrated Text Mining Framework for Metabolic Interaction Network Reconstruction. *PeerJ* **2016**, *4*, e1811.
(749) Chowdhury, M. F. M.; Lavelli, A. Disease Mention Recognition with Specific Features. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP '10)*; Uppsala, Sweden, July 15, 2010; pp 83−90.
(750) Dogan, R. I.; Lu, Z. An Improved Corpus of Disease Mentions in PubMed Citations. *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP'12)*; Montreal, Canada, June 8, 2012; pp 91−99.
(751) Kaewphan, S.; Hakaka, K.; Ginter, F. UTU: Disease Mention Recognition and Normalization with CRFs and Vector Space Representations. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*; Dublin, Ireland, August 23−24, 2014; pp 807−811.
(752) Jin, Y.; McDonald, R. T.; Lerman, K.; Mandel, M. A.; Carroll, S.; Liberman, M. Y.; Pereira, F. C.; Winters, R. S.; White, P. S.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

Automated Recognition of Malignancy Mentions in Biomedical Literature. *BMC Bioinf.* **2006**, *7*, 492.
(753) Névéol, A.; Lu, Z. Automatic Integration of Drug Indications from Multiple Health Resources. *Proceedings of the 1st ACM International Health Informatics Symposium (IHI'10)*; Arlington, VA, November 11−12, 2010; pp 666−673.
(754) Karimi, S.; Metke-Jimenez, A.; Kemp, M.; Wang, C. Cadec: A Corpus of Adverse Drug Event Annotations. *J. Biomed. Inf.* **2015**, *55*, 73−81.
(755) Segura-Bedmar, I.; Revert, R.; Martínez, P. Detecting Drugs and Adverse Events from Spanish Health Social Media Streams. *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014*; Gothenburg, Sweden, April 26−30, 2014; pp 106−115.
(756) Kang, N.; Singh, B.; Bui, C.; Afzal, Z.; van Mulligen, E. M.; Kors, J. A. Knowledge-Based Extraction of Adverse Drug Events from Biomedical Text. *BMC Bioinf.* **2014**, *15*, 64.
(757) Jimeno, A.; Jimenez-Ruiz, E.; Lee, V.; Gaudan, S.; Berlanga, R.; Rebholz-Schuhmann, D. Assessment of Disease Named Entity Recognition on a Corpus of Annotated Sentences. *BMC Bioinf.* **2008**, *9* (Suppl3), S3.
(758) Wang, Y. Annotating and Recognising Named Entities in Clinical Notes. *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*; Suntec, Singapore, August 4, 2009; pp 18−26.
(759) Gurulingappa, H.; Klinger, R.; Hofmann-Apitius, M.; Fluck, J. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. *2nd Workshop on Building and evaluating resources for biomedical text mining (Seventh edition of the Language Resources and Evaluation Conference - LREC 2010 Workshop)*; Valetta, Malta, May 17−23, 2010; pp 15−22.
(760) Pradhan, S.; Elhadad, N.; South, B. R.; Martinez, D.; Christensen, L.; Vogel, A.; Suominen, H.; Chapman, W. W.; Savova, G. Evaluating the State of the Art in Disorder Recognition and Normalization of the Clinical Narrative. *J. Am. Med. Inf. Assoc.* **2015**, *22* (1), 143−154.
(761) Harpaz, R.; Callahan, A.; Tamang, S.; Low, Y.; Odgers, D.; Finlayson, S.; Jung, K.; LePendu, P.; Shah, N. H. Text Mining for Adverse Drug Events: The Promise, Challenges, and State of the Art. *Drug Saf.* **2014**, *37* (10), 777−790.
(762) Leaman, R.; Khare, R.; Lu, Z. Challenges in Clinical Natural Language Processing for Automated Disorder Normalization. *J. Biomed. Inf.* **2015**, *57*, 28−37.
(763) Kang, N.; Singh, B.; Afzal, Z.; van Mulligen, E. M.; Kors, J. A. Using Rule-Based Natural Language Processing to Improve Disease Normalization in Biomedical Text. *J. Am. Med. Inf. Assoc.* **2013**, *20* (5), 876−881.
(764) Huang, Z.; Hu, X. Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. *IJMLC* **2013**, *3* (6), 494−498.
(765) Lee, H.-C.; Hsu, Y.-Y.; Kao, H.-Y. AuDis: An Automatic CRF-Enhanced Disease Normalization in Biomedical Text. *Database* **2016**, *2016*, baw091.
(766) Chen, Y.; Pedersen, L. H.; Chu, W. W.; Olsen, J. Drug Exposure Side Effects from Mining Pregnancy Data. *SIGKDD Explor.* **2007**, *9* (1), 22−29.
(767) Rosario, B.; Hearst, M. A. Classifying Semantic Relations in Bioscience Texts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04)*; Barcelona, Spain, July 21−26, 2004; p 430.
(768) Fiszman, M.; Demner-Fushman, D.; Kilicoglu, H.; Rindflesch, T. C. Automatic Summarization of MEDLINE Citations for Evidence-Based Medical Treatment: A Topic-Oriented Evaluation. *J. Biomed. Inf.* **2009**, *42* (5), 801−813.
(769) Physicians' Desk Reference. http://www.pdr.net (accessed Oct 29, 2016).
(770) Mihalcea, R.; Tarau, P. TextRank: Bringing Order into Texts. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004)*; Barcelona, Spain, July 25−26, 2004; pp 404−411.
(771) Jonnalagadda, S. R.; Del Fiol, G.; Medlin, R.; Weir, C.; Fiszman, M.; Mostafa, J.; Liu, H. Automatically Extracting Sentences from Medline Citations to Support Clinicians' Information Needs. *J. Am. Med. Inf. Assoc.* **2013**, *20* (5), 995−1000.
(772) Bravo, À.; Piñero, J.; Queralt-Rosinach, N.; Rautschka, M.; Furlong, L. I. Extraction of Relations between Genes and Diseases from Text and Large-Scale Data Analysis: Implications for Translational Research. *BMC Bioinf.* **2015**, *16* (1), 55.
(773) Lowe, H. J.; Barnett, G. O. Understanding and Using the Medical Subject Headings (MeSH) Vocabulary to Perform Literature Searches. *JAMA, J. Am. Med. Assoc.* **1994**, *271* (14), 1103−1108.
(774) Lee, C.-H.; Khoo, C.; Na, J.-C. Automatic Identification of Treatment Relations for Medical Ontology Learning: An Exploratory Study. *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference*; London, England, July 13−16, 2004; pp 245−250.
(775) Abacha, A. Ben; Zweigenbaum, P. Automatic Extraction of Semantic Relations between Medical Entities: A Rule Based Approach. *J. Biomed. Semant.* **2011**, *2* (Suppl 5), S4.
(776) Embarek, M.; Ferret, O. Learning Patterns for Building Resources about Semantic Relations in the Medical Domain. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*; Marrakech, Morocco, May 28−30, 2008.
(777) Li, Y.; Salmasian, H.; Harpaz, R.; Chase, H.; Friedman, C. Determining the Reasons for Medication Prescriptions in the EHR Using Knowledge and Natural Language Processing. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 768−776.
(778) Xu, R.; Wang, Q. Large-Scale Extraction of Accurate Drug-Disease Treatment Pairs from Biomedical Literature for Drug Repurposing. *BMC Bioinf.* **2013**, *14*, 181.
(779) Jung, K.; LePendu, P.; Chen, W. S.; Iyer, S. V; Readhead, B.; Dudley, J. T.; Shah, N. H. Automated Detection of off-Label Drug Use. *PLoS One* **2014**, *9* (2), e89324.
(780) Gurulingappa, H.; Kolárik, C.; Hofmann-Apitius, M.; Fluck, J. Concept-Based Semi-Automatic Classification of Drugs. *J. Chem. Inf. Model.* **2009**, *49* (8), 1986−1992.
(781) Khare, R.; Burger, J. D.; Aberdeen, J. S.; Tresner-Kirsch, D. W.; Corrales, T. J.; Hirchman, L.; Lu, Z. Scaling Drug Indication Curation through Crowdsourcing. *Database* **2015**, *2015*, bav016.
(782) Li, J.; Sun, Y.; Johnson, R. J.; Sciaky, D.; Wei, C.-H.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Wiegers, T. C.; Lu, Z. BioCreative V CDR Task Corpus: A Resource for Chemical Disease Relation Extraction. *Database* **2016**, *2016*, 1.
(783) Harpaz, R.; DuMouchel, W.; Shah, N. H.; Madigan, D.; Ryan, P.; Friedman, C. Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis. *Clin. Pharmacol. Ther.* **2012**, *91* (6), 1010−1021.
(784) Karimi, S.; Wang, C.; Metke-Jimenez, A.; Gaire, R.; Paris, C. Text and Data Mining Techniques in Adverse Drug Reaction Detection. *ACM Comput. Surv.* **2015**, *47* (4), 56.
(785) Avillach, P.; Dufour, J.-C.; Diallo, G.; Salvo, F.; Joubert, M.; Thiessard, F.; Mougin, F.; Trifirò, G.; Fourrier-Réglat, A.; Pariente, A.; et al. Design and Validation of an Automated Method to Detect Known Adverse Drug Reactions in MEDLINE: A Contribution from the EU-ADR Project. *J. Am. Med. Inf. Assoc.* **2013**, *20* (3), 446−452.
(786) Garcelon, N.; Mougin, F.; Bousquet, C.; Burgun, A. Evidence in Pharmacovigilance: Extracting Adverse Drug Reactions Articles from MEDLINE to Link Them to Case Databases. *Stud. Health Technol. Inform.* **2006**, *124*, 528−533.
(787) Shetty, K. D.; Dalal, S. R. Using Information Mining of the Medical Literature to Improve Drug Safety. *J. Am. Med. Inf. Assoc.* **2011**, *18* (5), 668−674.
(788) Wang, W.; Haerian, K.; Salmasian, H.; Harpaz, R.; Chase, H.; Friedman, C. A Drug-Adverse Event Extraction Algorithm to Support Pharmacovigilance Knowledge Mining from PubMed Citations. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 1464−1470.
(789) Gurulingappa, H.; Mateen-Rajpu, A.; Toldo, L. Extraction of Potential Adverse Drug Events from Medical Case Reports. *J. Biomed. Semant.* **2012**, *3* (1), 15.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(790) Gurulingappa, H.; Rajput, A. M.; Roberts, A.; Fluck, J.; Hofmann-Apitius, M.; Toldo, L. Development of a Benchmark Corpus to Support the Automatic Extraction of Drug-Related Adverse Effects from Medical Case Reports. *J. Biomed. Inf.* **2012**, *45* (5), 885−892.
(791) Brown, E. G.; Wood, L.; Wood, S. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Saf.* **1999**, *20* (2), 109−117.
(792) Giuliano, C.; Lavelli, A.; Pighin, D.; Romano, L. FBK-IRST: Kernel Methods for Semantic Relation Extraction. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*; Prague, Czech Republic, June 23−24, 2007; pp 141−144.
(793) Gurulingappa, H.; Fluck, J.; Hofmann-Apitius, M.; Toldo, L. Identification of Adverse Drug Event Assertive Sentences in Medical Case Reports. *Proceedings of the First International Workshop on Knowledge Discovery and Health Care Management (KD-HCM), European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*; Athens, Greece, September 9, 2011; pp 16−27.
(794) Yang, C.; Srinivasan, P.; Polgreen, P. Automatic Adverse Drug Events Detection Using Letters to the Editor. *AMIA Annu. Symp. Proc.* **2012**, *2012*, 1030−1039.
(795) Fourches, D.; Barnes, J. C.; Day, N. C.; Bradley, P.; Reed, J. Z.; Tropsha, A. Cheminformatics Analysis of Assertions Mined from Literature That Describe Drug-Induced Liver Injury in Different Species. *Chem. Res. Toxicol.* **2010**, *23* (1), 171−183.
(796) Leaman, R.; Wojtulewicz, L.; Sullivan, R.; Skariah, A.; Yang, J.; Gonzalez, G. Towards Internet-Age Pharmacovigilance: Extracting Adverse Drug Reactions from User Posts to Health-Related Social Networks. *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing (BioNLP'10)*; Uppsala, Sweden, July 15, 2010; pp 117−125.
(797) Chee, B. W.; Berlin, R.; Schatz, B. Predicting Adverse Drug Events from Personal Health Messages. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 217−226.
(798) Nikfarjam, A.; Gonzalez, G. H. Pattern Mining for Extraction of Mentions of Adverse Drug Reactions from User Comments. *AMIA Annu. Symp. Proc.* **2011**, *2011*, 1019−1026.
(799) Liu, X.; Chen, H. AZDrugMiner: An Information Extraction System for Mining Patient-Reported Adverse Drug Events in Online Patient Forums. *International Conference on Smart Health (ICSH 2013)*; Beijing, China, August 3−4, 2013; pp 134−150.
(800) Yang, C. C.; Jiang, L.; Yang, H.; Tang, X. Detecting Signals of Adverse Drug Reactions from Health Consumer Contributed Content in Social Media. In *Proceedings of ACM SIGKDD Workshop on Health Informatics (HI-KDD'12)*; Beijing, China, August 12−16, 2012.
(801) Sampathkumar, H.; Chen, X.; Luo, B. Mining Adverse Drug Reactions from Online Healthcare Forums Using Hidden Markov Model. *BMC Med. Inf. Decis. Making* **2014**, *14* (1), 91.
(802) Lardon, J.; Abdellaoui, R.; Bellet, F.; Asfari, H.; Souvignet, J.; Texier, N.; Jaulent, M.-C.; Beyens, M.-N.; Burgun, A.; Bousquet, C. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. *J. Med. Internet Res.* **2015**, *17* (7), e171.
(803) Katragadda, S.; Karnati, H.; Pusala, M.; Raghavan, V.; Benton, R. Detecting Adverse Drug Effects Using Link Classification on Twitter Data. *Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2015)*; Washington DC, November 9−12, 2015; pp 675−679.
(804) Bian, J.; Topaloglu, U.; Yu, F. Towards Large-Scale Twitter Mining for Drug-Related Adverse Events. *Proceedings of the 2012 International Workshop on Smart Health and Wellbeing (SHB'12)*; Maui, HI, October 29 - November 02, 2012; pp 25−32.
(805) Ginn, R.; Pimpalkhute, P.; Nikfarjam, A.; Patki, A.; O'Connor, K.; Sarker, A.; Smith, K.; Gonzalez, G. Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark. *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014)*; Reykjavík, Iceland, May 31, 2014.
(806) Sarker, A.; Ginn, R.; Nikfarjam, A.; O'Connor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; Gonzalez, G. Utilizing Social Media Data for Pharmacovigilance: A Review. *J. Biomed. Inf.* **2015**, *54*, 202−212.
(807) DailyStrength https://www.dailystrength.org (accessed Oct 20, 2016).
(808) Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L. J.; Bork, P. A Side Effect Resource to Capture Phenotypic Effects of Drugs. *Mol. Syst. Biol.* **2010**, *6* (1), 343.
(809) Wood, K. L. The Medical Dictionary for Drug Regulatory Affairs (MEDDRA) Project. *Pharmacoepidemiol. Drug Saf.* **1994**, *3* (1), 7−13.
(810) Smith, J. C.; Denny, J. C.; Chen, Q.; Nian, H.; Spickard, A.; Rosenbloom, S. T.; Miller, R. A. Lessons Learned from Developing a Drug Evidence Base to Support Pharmacovigilance. *Appl. Clin. Inform.* **2013**, *4* (4), 596−617.
(811) Bisgin, H.; Liu, Z.; Fang, H.; Xu, X.; Tong, W. Mining FDA Drug Labels Using an Unsupervised Learning Technique-Topic Modeling. *BMC Bioinf.* **2011**, *12* (Suppl 10), S11.
(812) Wang, X.; Hripcsak, G.; Markatou, M.; Friedman, C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. *J. Am. Med. Inf. Assoc.* **2009**, *16* (3), 328−337.
(813) Honigman, B.; Lee, J.; Rothschild, J.; Light, P.; Pulling, R. M.; Yu, T.; Bates, D. W. Using Computerized Data to Identify Adverse Drug Events in Outpatients. *J. Am. Med. Inf. Assoc.* **2001**, *8* (3), 254−266.
(814) Hazlehurst, B.; Sittig, D. F.; Stevens, V. J.; Smith, K. S.; Hollis, J. F.; Vogt, T. M.; Winickoff, J. P.; Glasgow, R.; Palen, T. E.; Rigotti, N. A. Natural Language Processing in the Electronic Medical Record: Assessing Clinician Adherence to Tobacco Treatment Guidelines. *Am. J. Prev. Med.* **2005**, *29* (5), 434−439.
(815) Bates, D. W.; Evans, R. S.; Murff, H.; Stetson, P. D.; Pizziferri, L.; Hripcsak, G. Detecting Adverse Events Using Information Technology. *J. Am. Med. Inf. Assoc.* **2003**, *10* (2), 115−128.
(816) Warrer, P.; Hansen, E. H.; Juhl-Jensen, L.; Aagaard, L. Using Text-Mining Techniques in Electronic Patient Records to Identify ADRs from Medicine Use. *Br. J. Clin. Pharmacol.* **2012**, *73* (5), 674−684.
(817) Jensen, P. B.; Jensen, L. J.; Brunak, S. Mining Electronic Health Records: Towards Better Research Applications and Clinical Care. *Nat. Rev. Genet.* **2012**, *13* (6), 395−405.
(818) LePendu, P.; Iyer, S. V; Bauer-Mehren, A.; Harpaz, R.; Mortensen, J. M.; Podchiyska, T.; Ferris, T. A.; Shah, N. H. Pharmacovigilance Using Clinical Notes. *Clin. Pharmacol. Ther.* **2013**, *93* (6), 547−555.
(819) Iyer, S. V; LePendu, P.; Harpaz, R.; Bauer-Mehren, A.; Shah, N. H. Learning Signals of Adverse Drug-Drug Interactions from the Unstructured Text of Electronic Health Records. *AMIA Jt. Summits Transl. Sci. Proc.* **2013**, *2013*, 98.
(820) Oronoz, M.; Gojenola, K.; Pérez, A.; de Ilarraza, A. D.; Casillas, A. On the Creation of a Clinical Gold Standard Corpus in Spanish: Mining Adverse Drug Reactions. *J. Biomed. Inf.* **2015**, *56*, 318−332.
(821) Visweswaran, S.; Hanbury, P.; Saul, M. I.; Cooper, G. F. Detecting Adverse Drug Events in Discharge Summaries Using Variations on the Simple Bayes Model. *AMIA Annu. Symp. Proc.* **2003**, 689−693.
(822) Meystre, S. M.; Savova, G. K.; Kipper-Schuler, K. C.; Hurdle, J. F. Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *Yearb. Med. Inform.* **2008**, *35*, 128−144.
(823) Field, T. S.; Gurwitz, J. H.; Harrold, L. R.; Rothschild, J. M.; Debellis, K.; Seger, A. C.; Fish, L. S.; Garber, L.; Kelleher, M.; Bates, D. W. Strategies for Detecting Adverse Drug Events among Older Persons in the Ambulatory Setting. *J. Am. Med. Inf. Assoc.* **2004**, *11* (6), 492−498.
(824) Murff, H. J.; Forster, A. J.; Peterson, J. F.; Fiskio, J. M.; Heiman, H. L.; Bates, D. W. Electronically Screening Discharge Summaries for Adverse Medical Events. *J. Am. Med. Inf. Assoc.* **2003**, *10* (4), 339−350.
(825) Hazlehurst, B.; Naleway, A.; Mullooly, J. Detecting Possible Vaccine Adverse Events in Clinical Notes of the Electronic Medical Record. *Vaccine* **2009**, *27* (14), 2077−2083.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761

(826) Sohn, S.; Kocher, J.-P. A.; Chute, C. G.; Savova, G. K. Drug Side Effect Extraction from Clinical Narratives of Psychiatry and Psychology Patients. *J. Am. Med. Inf. Assoc.* **2011**, *18* (Suppl 1), i144−9.
(827) Aramaki, E.; Miura, Y.; Tonoike, M.; Ohkuma, T.; Masuichi, H.; Waki, K.; Ohe, K. Extraction of Adverse Drug Effects from Clinical Records. *Stud. Health Technol. Inform.* **2010**, *160*, 739−743.
(828) Tari, L.; Anwar, S.; Liang, S.; Cai, J.; Baral, C. Discovering Drug-Drug Interactions: A Text-Mining and Reasoning Approach Based on Properties of Drug Metabolism. *Bioinformatics* **2010**, *26* (18), i547−53.
(829) Percha, B.; Garten, Y.; Altman, R. B. Discovery and Explanation of Drug-Drug Interactions via Text Mining. *Pac. Symp. Biocomput.* **2012**, 410−421.
(830) Bui, Q.-C.; Sloot, P. M. A.; van Mulligen, E. M.; Kors, J. A. A Novel Feature-Based Approach to Extract Drug–Drug Interactions from Biomedical Text. *Bioinformatics* **2014**, *30* (23), 3365−3371.
(831) Kim, S.; Liu, H.; Yeganova, L.; Wilbur, W. J. Extracting Drug-Drug Interactions from Literature Using a Rich Feature-Based Linear Kernel Approach. *J. Biomed. Inf.* **2015**, *55*, 23−30.
(832) Segura Bedmar, I.; Martínez, P.; Herrero Zazo, M. Semeval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (Ddiextraction 2013). *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*; Atlanta, Georgia, June 14−15, 2013; pp 341−350.
(833) Southan, C.; Várkonyi, P.; Muresan, S. Quantitative Assessment of the Expanding Complementarity between Public and Commercial Databases of Bioactive Compounds. *J. Cheminf.* **2009**, *1* (1), 10.
(834) Walter, D. Patent Analytics: Current Tools and Emerging Trends. *Pharm. Pat. Anal.* **2014**, *3* (3), 227−233.
(835) Schenck, R. J.; Zapiecki, K. R. Back to the Future: CAS and the Shape of Chemical Information To Come. In *The Future of the History of Chemical Information*; McEwen, L. R., Buntrock, R. E., Eds.; ACS Symposium Series 1164; American Chemical Society: Washington, DC, 2014; pp 149−158.
(836) Correia, R. B.; Li, L.; Rocha, L. M. Monitoring Potential Drug Interactions and Reactions via Network Analysis of Instagram User Timelines. *Pac. Symp. Biocomput.* **2016**, *21*, 492−503.
(837) Hamed, A. A.; Wu, X.; Erickson, R.; Fandy, T. Twitter K-H Networks in Action: Advancing Biomedical Literature for Drug Search. *J. Biomed. Inf.* **2015**, *56*, 157−168.
(838) Omberg; NiklasWilliams, A. J.; Harland, L.; Groth, P.; Pettifer, S.; Chichester, C.; Willighagen, E. L.; Evelo, C. T.; Blomberg, N.; Ecker, G.; Goble, C.; et al. Open PHACTS: Semantic Interoperability for Drug Discovery. *Drug Discovery Today* **2012**, *17* (21−22), 1188−1198.
(839) Kreuzthaler, M.; Miñarro-Giménez, J. A.; Schulz, S. MapReduce in the Cloud: A Use Case Study for Efficient Co-Occurrence Processing of MEDLINE Annotations with MeSH. *Stud. Health Technol. Inform.* **2016**, *228*, 582−586.
(840) Karthikeyan, M.; Pandit, Y.; Pandit, D.; Vyas, R. MegaMiner: A Tool for Lead Identification Through Text Mining Using Chemo-informatics Tools and Cloud Computing Environment. *Comb. Chem. High Throughput Screening* **2015**, *18* (6), 591−603.
(841) Ochoa, R.; Davies, M.; Papadatos, G.; Atkinson, F.; Overington, J. P. myChEMBL: A Virtual Machine Implementation of Open Data and Cheminformatics Tools. *Bioinformatics* **2014**, *30* (2), 298−300.
(842) Urbain, J.; Frieder, O. Exploring Contextual Models in Chemical Patent Search. *Proceedings of the First International Information Retrieval Facility Conference on Adbances in Multidisciplinary Retrieval (IRFC'10)*; Vienna, Austria, May 31, 2010; pp 60−69.
(843) Salim, N.; Abd. Wahid, M. T.; Alwee, R.; Dollah. *The Study for Probability Model for Compound Similarity Searching*; Final Project Report UTM Research Management Centre Project Vote − 75207; Malaysia, February 18, 2008.
(844) Parkkinen, J. A.; Kaski, S. Probabilistic Drug Connectivity Mapping. *BMC Bioinf.* **2014**, *15*, 113.
(845) Lin, J.; Wilbur, W. J. PubMed Related Articles: A Probabilistic Topic-Based Model for Content Similarity. *BMC Bioinf.* **2007**, *8*, 423.
(846) Singh, S. B.; Hull, R. D.; Fluder, E. M. Text Influenced Molecular Indexing (TIMI): A Literature Database Mining Approach That Handles Text and Chemistry. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 743−752.
(847) Chiang, J.-H.; Ju, J.-H. Discovering Novel Protein-Protein Interactions by Measuring the Protein Semantic Similarity from the Biomedical Literature. *J. Bioinf. Comput. Biol.* **2014**, *12* (6), 1442008.
(848) Dura, E.; Muresan, S.; Engkvist, O.; Blomberg, N.; Chen, H. Mining Molecular Pharmacological Effects from Biomedical Text: A Case Study for Eliciting Anti-Obesity/Diabetes Effects of Chemical Compounds. *Mol. Inf.* **2014**, *33* (5), 332−342.
(849) Hsu, Y.-Y.; Kao, H.-Y. Curatable Named-Entity Recognition Using Semantic Relations. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2015**, *12* (4), 785−792.
(850) Peng, S.; You, R.; Wang, H.; Zhai, C.; Mamitsuka, H.; Zhu, S. DeepMeSH: Deep Semantic Representation for Improving Large-Scale MeSH Indexing. *Bioinformatics* **2016**, *32* (12), i70−79.
(851) Eom, J.-H.; Zhang, B.-T. PubMiner: Machine Learning-Based Text Mining for Biomedical Information Analysis. *Genomics Inform.* **2004**, *2*, 99−106.
(852) Ye, Z.; Tafti, A. P.; He, K. Y.; Wang, K.; He, M. M. SparkText: Biomedical Text Mining on Big Data Framework. *PLoS One* **2016**, *11* (9), e0162721.

DOI: 10.1021/acs.chemrev.6b00851
Chem. Rev. 2017, 117, 7673−7761