Text Mining for Chemical Compounds

Saber Ahmad Akhondi

Financial support for this thesis was provided by:

AstraZeneca, Mölndal, Sweden

ISBN:	978-94-6332-397-0
Layout:	Saber Ahmad Akhondi
Illustration and cover design:	Ghazaleh Beyk
Printed by:	GVO drukkers & vormgevers, Ede, The Netherlands

Text Mining for Chemical Compounds

Tekstmining naar chemische stoffen

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. R.C.M.E. Engels

en volgens besluit van het College voor Promoties
De openbare verdediging zal plaatsvinden

dinsdag 2 oktober 2018 om 13.30 uur
door

Saber Ahmad Akhondi
geboren te Tehran, Iran

Erasmus University Rotterdam

PROMOTIECOMMISSIE

Promotor:

Prof.dr. J. van der Lei

Overige leden:

Prof.dr. B. Mons

Dr. D. Rebholz-Schuhmann

Prof.dr. P.J. van der Spek

Copromotor:

Dr.ir. J.A. Kors

To My Family

TABLE OF CONTENTS

Chapter 1

Introduction

This thesis concerns the exploration of chemical space in chemical-related literature using text-mining. We begin this chapter by introducing chemical information extraction as a discipline. We continue by defining chemical naming conventions. Following we describe chemical information sources. The sources are categorized into chemical databases and chemical-related publications. The section continues by introducing methodologies to assess the quality of chemical databases. We continue by introducing text-mining as a means to automate the extraction of information from chemical-related publications. Furthermore, we present the benefits and challenges to only extract relevant information from chemical-related publications. This chapter concludes by providing the aim and outline of this thesis.

The chemistry domain

The introduction of the internet has resulted into a migration from hardcopy scientific literature to digital electronic publications. This migration has dramatically affected the research in both scientific and commercial environments [1]. The availability of machine-readable encoding systems in the chemistry field (since the 1940s) enabled a faster migration within the chemistry field [1]. Similarly, the number of patents published in the chemistry domain has quintupled annually since the early 1990s (from 1000 per year to around 5000 per year) [2].

Interestingly, researchers in the chemistry field read more scientific publications per person than researchers in other domains, except in the life sciences [3]. Due to the complexity and variety of chemistry-related literature, they spend the most amount of time on reading scientific literature as compared to other researchers [1]. Among the most retrieved information in chemistry is the identification of compounds of interest in chemical documents based on the structure of the compound [1]. Such information can be used for chemical predictive modelling [4] or Quantitative Structure Activity Relationships (QSAR) modelling that can be used in early stages of medicinal chemistry activities [2, 5]. The structure of chemical compounds is essential for chemical research and in most cases chemists focus on chemical structure or substructure for exploring the chemical domain [1].

A publication in the chemistry domain (be it a journal article or patent) can contain chemical-related information in a variety of ways. The information can be stored in the textual part of the document using different chemical identifiers (naming conventions). Additionally, the information can also be stored in chemical diagrams (chemical scaffolds or images) or tables. In some cases, this information can only be extracted by combining information from all of the above (such as for Markush compounds in patents) [1].

The ever-swelling volume of chemical-related documents in the form of scientific articles and patents makes it increasingly hard to manually find and extract relevant information from such texts [6]. These sources contain a large set of unstructured information which is cumbersome to process manually [1]. In order to overcome this obstacle different approaches can be taken into consideration. These approaches include mining currently available commercial or public chemical databases, and using techniques such as chemical text-mining to extract information from the textual part of the documents. The different representations of chemical names in text make these approaches extremely challenging [7].

Chemical entities extracted through text-mining can be valuable for information retrieval systems as they can point to documents mentioning the compound. They can also be used along with additional relevant information (e.g., biological activities extracted from text) to assess specialized search engines with specific well-defined queries [1]. The same information can be used to extend or curate available databases [8]. These systems become even more valuable if they can identify the relevant compounds within a document from the wide range of extracted compounds [9, 10]. Using patent analysis, the information can be used to understand compound prior art, or perform novelty checking, and finally identify new starting points for chemical exploration [9].

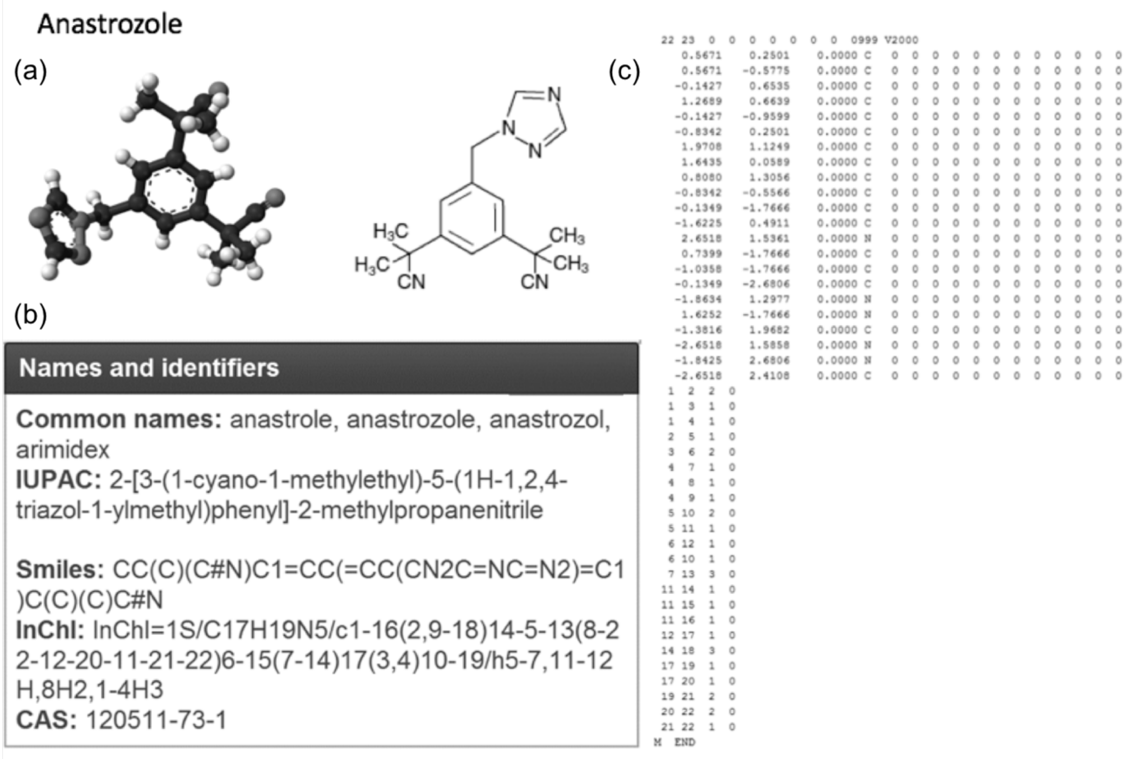
Naming conventions of chemical compounds

A chemical compound consists of two or more atoms of at least two elements which are connected via a chemical bond [11]. In chemistry, the compounds are represented in chemical diagrams and can be digitally stored in MOL files [12]. In short, a MOL file format digitally stores three-dimensional information for a compound based on the orientation of its atoms, bonds and additional chemical properties [12]. A MOL file consists of a table with coordinates of the elements and may contain additional fields regarding the properties of the compound.

Due to the presence of isotopes, charges, tautomers, stereochemistry or fragments for a compound, a chemical structure can be drawn in different ways. Based on the chemical field of study (e.g., organic chemistry vs in-organic chemistry) some of this information can be disregarded and the compound can be standardized [13, 14]. Such standardization approach maps two similar compounds with differing characteristics (e.g., one has stereochemistry, the other not) to one compound.

Chemical compound identifiers are used to refer to a chemical compound in text. Chemical identifiers can be distinguished in two major groups based on how they are generated.

The first group consists of systematic identifiers. These identifiers are generated algorithmically and correspond to the structure of the compound [1]. A set of rules are used for generating these identifiers. SMILES notations [15], InChI strings [16], and IUPAC names [17] are examples of systematic identifiers. A name-to-structure toolkit can be used to convert chemical compound structures to systematic identifiers and vice versa [1]. Systematic identifiers should have a one-to-one correspondence to the compounds. Despite constant improvements to the naming conventions, this is not always the case. For example, IUPAC names suffer from issues for converting stereochemistry information [18]. It is important to note that the standardization of a compound may affect the systematic naming of the compound.



Names and identifiers

Common names: anastrole, anastrozole, anastrozol, arimidex
IUPAC: 2-[3-(1-cyano-1-methylethyl)-5-(1H-1,2,4-triazol-1-ylmethyl)phenyl]-2-methylpropanenitrile

Smiles: CC(C)(C#N)C1=CC(=CC(CN2C=NC=N2)=C1)C(C)(C)C#N
InChI: InChI=1S/C17H19N5/c1-16(2,9-18)14-5-13(8-22-12-20-11-21-22)6-15(7-14)17(3,4)10-19/h5-7,11-12H,8H2,1-4H3
CAS: 120511-73-1

```
22 23  0  0  0  0  0  0  0  0999 V2000
   0.5671    0.2501    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.5671   -0.5775    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.1427    0.6535    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.2689    0.6639    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.1427   -0.9599    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.8342    0.2501    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.9708    1.1249    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   1.6435    0.0589    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   0.8080    1.3056    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.8342   -0.5566    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.1349   -1.7666    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.6225    0.4911    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   2.6518    1.5361    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   0.7399   -1.7666    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.0358   -1.7666    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -0.1349   -2.6806    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -1.8634    1.2977    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
   1.6252   -1.7666    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
  -1.3816    1.9682    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  -2.6518    1.5858    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
  -1.8425    2.6806    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
  -2.6518    2.4108    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0
  1  3  1  0
  1  4  1  0
  2  5  1  0
  3  6  2  0
  4  7  1  0
  4  8  1  0
  4  9  1  0
  5 10  2  0
  5 11  1  0
  6 12  1  0
  6 10  1  0
  7 13  3  0
 11 14  1  0
 11 15  1  0
 11 16  1  0
 12 17  1  0
 14 18  3  0
 17 19  1  0
 17 20  1  0
 19 21  2  0
 20 22  2  0
 21 22  1  0
M  END
```

Figure 1: Different representations of Anastrozole as a chemical compound. (a) 3D and 2D structure of “Anastrozole”. (b) compound naming in non-systematic (common names, CAS) and systematic names (IUPAC, SMILES, InChI). (c) part of MOL file representing Anastrozole.

The second group consists of non-systematic chemical identifiers. These identifiers are generated at the point of registration within the source. Brand names, generic names, research codes, chemical abstracts service (CAS) registry numbers, and database identifiers are examples of such non-systematic identifiers [12]. The only approach to identify the structure of a non-systematic identifier is to look it up in a database. Figure 1 illustrates the different representations of a chemical compound.

Chemical information sources

Chemical-related information is available through structured and unstructured resources. Structured sources include public and commercial chemical databases. Unstructured sources include scientific publications and patents [1]. These sources have different characteristics and extraction of information from them (manual or automatic) has its own challenges.

In the last decade, we have observed a major increase in the number of public and commercial chemical databases [19]. Chemical databases are structured data sources that provide a variety of chemical information on chemical compounds (e.g., SAR data) [13]. These data can be obtained from different means including data obtained from other databases. Chemical databases are built on chemical compound records. Ideally each chemical compound record is dedicated to a unique chemical compound based on its structural representation. In chemistry, the most important information retrieved from these sources is about compound structures [13]. This structural information is then used for different purposes, such as predictive modelling [4]. To retrieve information from databases, researchers mostly query by drawing the structure of the compound of interest (not available through all sources), or using compound systematic and non-systematic identifiers. Examples of such databases are PubChem [20], ChEBI [21], DrugBank [22], and Reaxys [23, 24]. Quality is a major aspect when dealing with databases. Scholars have shown errors within these sources and errors that proliferate from one database to another database through download and reuse of the content [25].

Unstructured data sources include scientific publications and patents. Scientific publications are available through different repositories such as MEDLINE [26]. Journal publications in the chemistry domain usually also contain a section with supplementary information. This section also contains a wide range of information valuable for chemistry research.

Initial public disclosure of new chemical compounds is usually done through patent applications in commercial research and development projects [27]. This makes patents extremely interesting for knowledge discovery. Analyzing patents is crucial in chemistry research [2, 27, 28]. Patent analysis enables the understanding of compound prior art, and provides the means for novelty checking and validation. It can also indicate new starting points for chemical research [9, 29–31]. Chemical patents are complex legal documents (not scientific). They can contain up to hundreds of pages. Patents have uniform structures and consist of title, abstract, claims and description. The European Patent Office (EPO) [32], the United States Patent and Trademark Office (USPTO) [33], and the World Intellectual Property Organization (WIPO) [34] are the biggest patent providers. These sources provide patent full text free of charge. Some patent offices only provide the patent through optical character recognition (OCR) format. OCR processing introduces spelling errors into the patent documents. As mentioned patents are legal documents, which tend to hide interesting chemical information. This results in additional difficulties in extracting relevant chemical information from patents both manually and automatically.

A patent document can contain thousands of mentions of different chemical compounds while defining experiments, claims and description. This is to ensure that the patent protects the chemical compound of interest (key compound). Key compounds are usually well-hidden within the context for commercial purposes [9, 10]. The presence of a large number of compounds in patents makes it difficult to manually or automatically identify the key compound.

Quality of chemical databases

The correctness of a structure that is extracted from chemical databases has great impact on the predictive ability of computational modeling [35]. While this correctness is crucial, qualitative studies have indicated that errors exist within chemical databases [25, 35]. Errors can be in the form of wrong structure associations or ambiguity within databases [19, 25, 35, 36]. Ambiguity is present in cases where an identifier is associated to more than one structure. Presence of such errors in one or multiple databases can also reduce the quality of other databases because databases tend to integrate data from one another [35]. Text-mining methods that use these databases for identification of chemical compounds or for association of the compound to a structure, are also affected by these types of error [1].

Identification of structure correctness of chemical compounds mentioned in databases depends on the chemical identifier mentioned in the databases. Systematic identifiers (generated algorithmically) can be evaluated using name-to-structure toolkits. The correctness of non-systematic identifiers can only be assessed in a manual manner because no algorithmic relationship between non-systematic identifiers and their structures exists [19]. To our knowledge, there has been no quantitative assessment of the consistency of systematic chemical identifiers and the ambiguity of non-systematic identifiers within and across chemical databases.

Text-mining on chemical literature

Exploring the chemical domain in chemical-related publications such as journal articles and patents is a challenging task. Text-mining can apply algorithmic, statistical and data management methodologies on a large set of chemical-related literature and unstructured free text to extract relevant information. In this way text-mining shifts the information overload problem from human to computers [37]. The complexity of textual content can influence the performance and complexity of a text-mining system. To obtain high performance, text-mining engines usually focus on domains (or sub-domains). For example, journal publications and patents have different characteristics (e.g., short vs long, scientific vs legal document, digital vs OCR) that need to be considered by a text-mining system [1, 37, 38].

Different text-mining steps can be taken into account depending on the use case. The performance of a text-mining tool relies on the performance of each of the components used in these steps [38]. Figure 2 illustrates the steps involved in text-mining. These steps are described in more detail below.

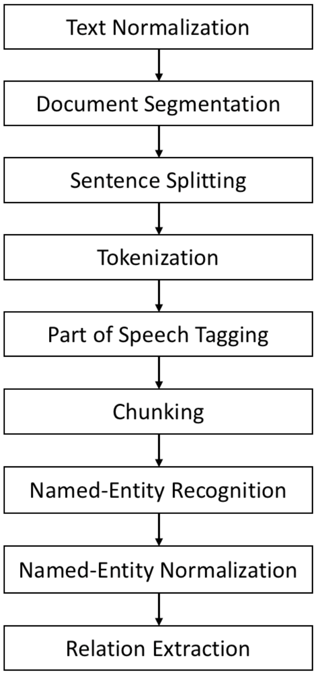


Figure 2: The main steps involved in text-mining.

Text normalization

The first component in text-mining approaches normalizes the input text. Chemical documents are available in a wide range of different formats. This can include PDF (portable document format), HTML (Hypertext Markup Language), XML (Xtensible Markup Language), or other common file formats [39]. The normalization component attempts to convert the data format into a suitable format for text-mining (e.g., plain text) [1]. This step is considerably more difficult when the input data have been generated with the use of OCR. Any errors made in this step can directly influence future steps. The normalization step also takes into account possibly different character encodings within the input data. Different character encoding standards can result in different digital representations for the same character and result in different interpretation of the same character. The use of internationally accepted standard character encodings can prevent possible errors. UTF-8 (8-bit Unicode Transformation Format) encoding supports a wide range of characters and can represent most chemical names and formulas [40]. This encoding is currently widely used for text-mining.

Document segmentation

There can be different segments within a journal or patent document (e.g., title, abstract, methods, results, claims, references). Document segmentation detects and delineates these segments based on the document structure. The extraction techniques of a text-mining tool can differ depending on the segment that is analyzed (e.g., chemical text-mining tools should not look for chemicals in references) [1].

Sentence splitting

This step splits the text into sentences. Sentences form the logical units of thought in human language. Punctuations are good indicators to define a sentence boundary [38]. Usually rule-based approaches are used for sentence detection (e.g., a sentence ends if there is a period, exclamation mark or question mark) [41]. Automatic identification of sentences in a chemical-related publication can be challenging. Systematic chemical identifiers such as IUPAC names can contain punctuation marks and therefore complicate the sentence splitting [1].

Tokenization

The tokenization step is the process of splitting each sentence into words, or tokens [38]. Chemical identifier naming conventions can complicate the tokenization step. Use of punctuation and symbols greatly influence the tokenization of chemical names. For example, in common English, parentheses are token separators. In chemistry, the parentheses can be part of the token (e.g., "(CH3)2CHCH2CH(CH3)2").

Part-of-speech tagging

Part-of-speech (POS) tagging is the process of identifying the part-of-speech information for each word (token) based on its meaning and its context (i.e., the relationship of the word to adjacent words) [38, 42]. For example, a word can be a verb, a noun, or an article.

Chunking

Chunking or shallow parsing is a technique that enables the machine to identify constituent parts of a sentence and link them to units with discrete grammatical meaning. Chunking provides the machine with an understanding of the sentence structure [38, 43]. This step combines tokens into grammatical units such as noun phrases, verb phrases, or prepositional phrases. In chemical identifier recognition, we can use chunks such as noun phrases to validate that a term is a chemical compound [1].

Named-entity recognition and normalization

Named-entity recognition (NER) is the process of identifying and classifying specific entities within a text [38]. An example of chemical NER is the identification of chemical compounds or their subclasses such as formulas, CAS numbers and IUPAC names [1]. Named-entity normalization is the identification of a relevant database identifier for the recognized named entity. This step correlates the extracted named entity to a named entity existing within a database.

Relation Extraction

Extraction of knowledge or facts is performed in the last phase of text-mining. Relation extraction is the process of identifying relations between pairs of identified entities. Examples of relation extraction include the identification of relations between genes and proteins, or between drugs and diseases [38, 43].

Named-entity recognition approaches in chemistry

Three text-mining approaches are used for extracting chemical named-entities from text. These approaches are dictionary-based, morphology-based (or grammar-based), and statistical-based [37].

Dictionary-based approaches use dictionaries as a basis to identify matches of the dictionary terms in the text [37]. The performance of these methods greatly relies on the quality of the used dictionary. These dictionaries are usually produced from chemical identifiers that are contained in well-known chemical databases. This approach is limited to the terms located within the dictionary. Dictionary-based approaches are valuable to extract non-systematic chemical identifiers (non-systematic chemical identifiers are stored in databases) but are less fit to extract systematic identifiers because it is nearly impossible to include all systematic chemical identifiers in a dictionary (systematic identifiers are algorithmically generated). Dictionary-based approaches cannot identify novel chemical compounds (they are not available in the databases upon which the dictionaries are based). Its noteworthy to mention that dictionary-based approaches can utilize the chemical database that was used to generate the dictionary, to identify the structure of the compound [6].

Grammar-based approaches capture systematic chemical identifiers by exploiting the rules that are used to produce them. Therefore, grammar-based approaches can recognize systematic identifiers that are missing from the dictionaries. This also includes new systematic chemical identifiers [1, 6, 37]. Through a set of rules a systematic name can be translated into a chemical structure. Grammar-based approaches utilize the same rules to provide chemical structures for recognized compounds. Building grammar-based systems requires a deep understanding of the naming conventions and the domain. These systems also need to be changed based on the changes of naming conventions over time. Grammar-based approaches are generally limited in identifying non-systematic chemical identifiers, although some of these identifiers may be found with regular expressions [1].

Statistical-based approaches use manually created resources (a training set of documents with annotated chemical identifiers) to automatically train a classifier that can recognize chemical identifiers within text [1]. These approaches can identify both systematic and non-systematic identifiers. The drawback of statistical approaches is that they need a large annotated corpus to train the system. Statistical approaches have no direct means to provide structures for extracted chemical entities.

As mentioned, each of the approaches has its benefits and limitations. An ensemble system that combines multiple approaches can help resolve some of the limitations. It is noteworthy that until recently the focus of text-mining systems has mostly been on the biomedical domain, and relatively limited research in chemical text-mining has been done [44, 45].

Community competitions and tasks for text-mining

A common approach to improve, enhance, and compare the performance of text-mining systems is the introduction of community challenges that address a specific text-mining task [1]. These challenges are performed in the form of conferences or workshops (e.g., BioCreative [46]). Participants (academia and industry) are challenged to develop systems for the task and provide results in a predefined time frame. The outcome of the challenge is a set of systems and methodologies that help progress in the task domain. Comparative performance results are usually published in scientific literature.

Chemical gold standard corpora for NER

The availability of manually annotated corpora is essential for building named-entity recognition systems and validating their performance [1, 6, 37]. The annotations in a corpus are regarded as the ground truth and should have high quality. To obtain a high-quality corpus, the manual annotators must use well-defined annotation guidelines. Preferably, annotations are provided by multiple annotators to reduce the influence of an individual annotator's perspective. The annotations of multiple annotators can be harmonized using methods such as voting [1, 6, 37].

Producing an annotated corpus is laborious and expensive. Currently only a few non-commercial corpora exist for chemical NER [47–49]. These are mostly limited to titles and abstracts from scientific publications. A few corpora are available for patents [50, 51] but they are limited in size and do not contain all patent sections. Extending the current corpora to cover full-text journals and full patents is essential for building text-mining systems that can analyze the full text.

Performance evaluation

The availability a gold-standard corpus enables performance evaluation of text-mining systems. Typically, three performance measures are used: precision, recall, and F-score.

Precision and recall were first introduced in the 1950s for the evaluation of information retrieval systems [52]. The same measures are also used for text mining. Precision or positive predictive value is the percentage of correct system annotations over all annotations made by the system. Recall is the percentage of correct system annotations over all gold-standard annotations [52]. Later F-score was introduced as an aggregate performance measure [53]. F-score is the harmonic-mean of precision and recall.

In order to calculate precision, recall, and F-score three key measurements need to be determined based on the manual annotations and the annotations made by the system. These measurements are the number of true positives (TP, the number of manual annotations correctly identified by the system), the number of false positives (FP, the number of wrong annotations by the system), and number of false negatives (FN, the number of manual annotations that are missed by the system).

Precision, recall and F-score are then calculated as follows:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F - score = 2 * \frac{Precison * Recall}{Precision + Recall}$$

Aim and outline of the thesis

Most chemical research utilizes the structure representation of chemical compounds. The naming conventions that enable the translation of chemical identifiers to chemical structures and vice versa are unique to the chemical field. The characteristics of these identifiers in chemical-related text such as journals and patents have made text-mining challenging in the chemical field. To enhance text-mining in the chemical field, the quality of chemical-related databases needs to be investigated based on their representation of chemical compound structures. The availability of high quality association between compounds and their structures provides the means to build text-mining solutions that can extract chemical identifiers and their associated structures from journals and patents. Analyzing these identifiers based on their relevancy to the field of study can provide understanding of compound prior art, novelty checking, validation of biological assays, and identification of new starting points for chemical exploration. The aim of this study was to use text mining for the identification of chemical identifiers in journal and patent documents. For this:

First, we investigate the quality of chemical-related databases based on their representation of chemical compound structures. In **Chapter 2**, we investigate the consistency of systematic identifiers within and between small molecular databases. In **Chapter 3**, we expand our research and focus on the ambiguity of non-systematic chemical identifiers within and between chemical databases.

Second, we develop new resources that can be utilized to further enhance text-mining systems in the chemical domain. In particular, we develop an annotated chemical patent corpus based on full-text patent documents in **Chapter 4**.

Third, we investigate the development of systems for extracting chemical identifiers from journal articles and patents. To build efficient text-mining engines for journals and patents we investigate a variety of chemical text-mining approaches. In **Chapter 5**, we focus on mining chemical identifiers from journal publications using dictionary-based and grammar-based approaches. In **Chapter 6**, we focus on extraction of chemical entities from patents using dictionary-based and machine-learning approaches.

Finally, we use the methods and techniques studied in previous chapters to identify relevant compounds in patents. In **Chapter 7**, we develop a patent corpus containing relevant compounds and use it along with a high-quality chemical database to train and evaluate our text-mining system.

References

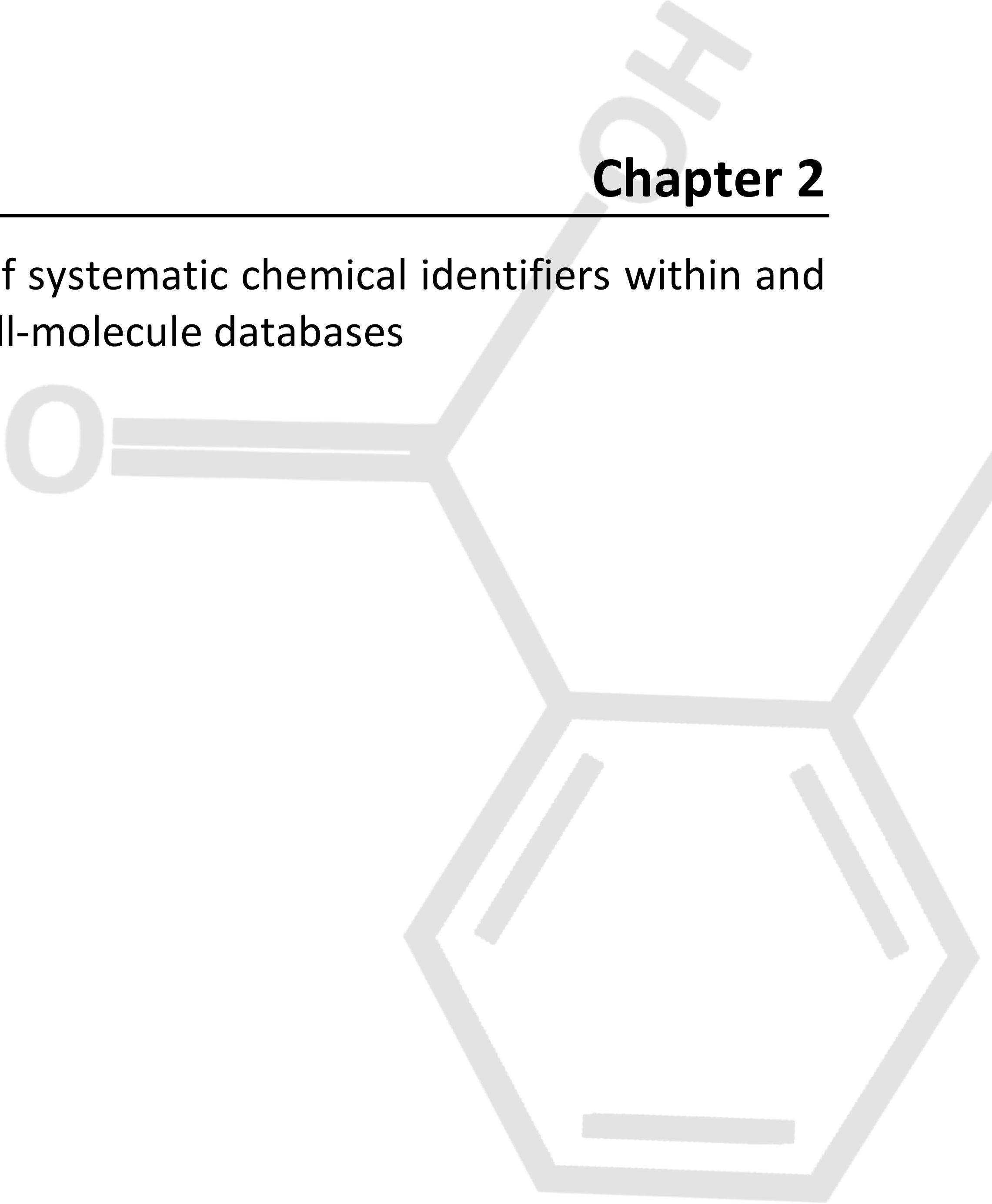
1. Krallinger M, Rabal O, Lourenço A, et al: **Information Retrieval and Text Mining Technologies for Chemistry.** *Chem Rev* 2017, **117 (12):**7673–7761.
2. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, Varkonyi P, Xie PH: **Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data.** *Drug Discov Today* 2011, **16:**1019–1030.
3. Tenopir C, King DW: **Reading behaviour and electronic journals.** *Learn Publ* 2002, **15**:259–265.
4. Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H: **Chemical predictive modelling to improve compound quality.** *Nat Rev Drug Discov* 2013, **12:**948-962.
5. Liaw A, Svetnik V: **QSAR modeling: prediction of biological activity from chemical structure.** *Statistical Methods for Evaluating Safety in Medical Product Development* 2015**:**66-83.
6. Eltyeb S, Salim N: **Chemical named entities recognition: a review on approaches and applications.** *J Cheminform* 2014, **6:**1-12.
7. Currano JN: **Teaching Chemical Information for the Future: The More Things Change, the More They Stay the Same.** *The Future of the History of Chemical Information* 2014, **Chapter 11:**169–196.
8. Williams AJ, Ekins S: **A quality alert and call for improved curation of public chemistry databases.** *Drug Discov Today* 2011, **16**:747–750.
9. Tyrchan C, Boström J, Giordanetto F, Winter J, Muresan S: **Exploiting Structural Information in Patent Specifications for Key Compound Prediction.** *J Chem Inf Model* 2012, **52**: 1480-1489.
10. Hattori K, Wakabayashi H, Tamaki K: **Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone.** *J Chem Inf Model* 2008, **48:**135–142.
11. Pauling L: **General chemistry.** *Dover Publications* 2008.
12. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: **Description of several chemical structure file formats used by computer programs developed at molecular design limited.** *J Chem Inf Comput Sci* 1992, **32:** 244-255.
13. Muresan S, Sitzmann M, Southan C: **Mapping between databases of compounds and protein targets.** *Methods Mol Biol* 2012, **910:**145–164.
14. Sitzmann M, Filippov IV, Nicklaus MC: **Internet resources integrating many small-molecule databases.** *SAR QSAR Environ Res* 2008, **19:**1-9.
15. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28:**31-36.
16. *InChI Trust - developing the InChI chemical structure standard.* http://www.inchi-trust.org/.
17. *IUPAC | International Union of Pure and Applied Chemistry Nomenclature.* https://iupac.org/.
18. Wilkinson A, McNaught A: **IUPAC Compendium of Chemical Terminology.** *Int. Union Pure Appl. Chem.* 1997.
19. Williams AJ: **Public chemical compound databases.** *Curr Opin Drug Discov Devel* 2008, **11:**393–404.
20. Kim S, Thiessen PA, Bolton EE et al: **PubChem Substance and Compound databases.** *Nucleic Acids Res.* 2016, **44:**D1202–D1213.
21. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, et al.: **ChEBI: a database and ontology for chemical entities of biological interest.** ***Nucleic Acids Res*** 2008, **36:** D344-D350.

22. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, 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:**D1091-1097.
23. *Reaxys.* https://www.reaxys.com.
24. Lawson AJ, Swienty-Busch J, Géoui T, Evans D: ***The Making of Reaxys-Towards Unobstructed Access to Relevant Chemistry Information.*** The Future of the History of Chemical Information 2014**:**127–148.
25. Williams AJ, Ekins S: **A quality alert and call for improved curation of public chemistry databases.** *Drug Discov Today* 2011, **16:**747–750.
26. Sayers EW, Barrett T, Benson DA, et al: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2009, D1 **41:**D8-D21.
27. Senger S, Bartek L, Papadatos G, Gaulton A: **Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents.** *J Cheminform* 2015:**7:**49.
28. Asche G: **"80% of technical information found only in patents" – Is there proof of this ?** *World Pat Inf* 2017, **48:**16–28.
29. Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA: **Annotated Chemical Patent Corpus: A Gold Standard for Text Mining.** *PloS one* 2014, **9:**e107477.
30. Papadatos G, Davies M, Dedman N et al: **SureChEMBL: a large-scale, chemically annotated patent document database.** *Nucleic Acids Res.* 2016, **44:**D1220–D1228.
31. Benson CL, Magee CL: **Quantitative determination of technological improvement from patent data.** *PLoS One* 2015, **10(4):**e0121635.
32. *European Patent Office.* https://www.epo.org/index.html.
33. *United States Patent and Trademark Office.* https://www.uspto.gov/.
34. *Word Intellectual Property Organization.* https://patentscope.wipo.int.
35. Young D, Martin T, Venkatapathy R, Harten P: **Are the chemical structures in your QSAR correct?** *QSAR Comb Sci* 2008, **27:**1337–1345.
36. Opera TI, Olah M, Ostopovici L, Rad R, Mracec M: **On the propagation of errors in the QSAR literature**. In *EuroQSAR 2002 designing drugs and crop protectants: processes, problems and solutions*. 2003rd edition. Edited by Ford M, Livingstone D, Dearden J, Van de Waterbeemd H. New York: Blackwell Publishing; **2003:**314–315.
37. Vazquez M, Krallinger M, Leitner F, Valencia A: **Text mining for drugs and chemical compounds: methods, tools and applications.** *Molecular Informatics* 2011, **30:**506–519.
38. Kang N: **Using natural language processing to improve biomedical concept normalization and relation mining.** *J Am Med Inform Assoc* 2013, **20(5):**876-81.
39. Park J, Rosania GR, Shedden K a, et al: **Automated extraction of chemical structure information from digital raster images.** *Chem Cent J* 2009, **3:**4.
40. Davis M: **Unicode nearing 50% of the web. Off**. *Google Blog* 2010*.*
41. Stamatatos E, Fakotakis N, Kokkinakis G: **Automatic extraction of rules for sentence boundary disambiguation.** *In: Proc. Work. Mach. Learn. Hum. Lang. Technol.* 1999 **P:**88–92.
42. Tsuruoka Y, Tateishi Y, Kim J-D, et al: **Developing a Robust Part-of-Speech Tagger for Biomedical Text.** In: *BMC Bioinform* 2013**, 141:**382–392.

43. Berwick R: **Principle-based parsing.** *Computation and Psycholinguistics* 1987.
44. Hettne K, Boorsma A, van Dartel A, Goeman J, de Jong E, Piersma A, Stierum R, Kleinjans J, Kors J: **Next-generation text-mining mediated generation of chemical response-specific gene sets for interpretation of gene expression data.** *BMC Medical Genomics* 2013, **6:**2.
45. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, Mulligen EM, Kleinjans J, Kors JA**: A dictionary to identify small molecules and drugs in free text**. *Bioinformatics* 2009, **25:**2983-2991.
46. *BioCreative.* http://www.biocreative.org/.
47. Kim J-D, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus-a semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19:**i180–i182.
48. Kulick S, Bies A, Liberman M, Mandel M, McDonald R, et al. **Integrated annotation for biomedical information extraction;** 2004. *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* pp. 61-68.
49. Kolárik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J. **Chemical names: terminological resources and corpora annotation**; 2008. *Workshop on Building and evaluating resources for biomedical text mining*.
50. Kiss M, Nagy Á, Vincze V, Almási A, Alexin Z, et al.: **A Manually Annotated Corpus of Pharmaceutical Patents. Text, Speech and Dialogue.** *Springer Berlin Heidelberg* 2012, pp. 135–142.
51. Tiago G, Catia P, Bastos Hugo P: **Chemical entity recognition and resolution to ChEBI.** *ISRN Bioinformatics* 2012.
52. Kent A, Berry M, Luehrs F: **Machine literature searching VIII. Operational criteria for designing information retrieval systems.** *J. Assoc.* 1955. **6:**93-101.
53. van Rijsbergen C: **Information retrieval**. *Butterworth-Heinemann Newton* 1979.

Chapter 2

Consistency of systematic chemical identifiers within and between small-molecule databases



Published: ***Akhondi SA****, Kors JA, Muresan S*

Journal of Cheminformatics 2012, 4: 35

Abstract

Background

Correctness of structures and associated metadata within public and commercial chemical databases greatly impacts drug discovery research activities such as quantitative structure–property relationships modelling and compound novelty checking. MOL files, SMILES notations, IUPAC names, and InChI strings are ubiquitous file formats and systematic identifiers for chemical structures. While interchangeable for many cheminformatics purposes there have been no studies on the inconsistency of these structure identifiers due to various approaches for data integration, including the use of different software and different rules for structure standardisation. We have investigated the consistency of systematic identifiers of small molecules within and between some of the commonly used chemical resources, with and without structure standardization.

Results

The consistency between systematic chemical identifiers and their corresponding MOL representation varies greatly between data sources (37.2%-98.5%). We observed the lowest overall consistency for MOL-IUPAC names. Disregarding stereochemistry increases the consistency (84.8% to 99.9%). A wide variation in consistency also exists between MOL representations of compounds linked via cross-references (25.8% to 93.7%). Removing stereochemistry improved the consistency (47.6% to 95.6%).

Conclusions

We have shown that considerable inconsistency exists in structural representation and systematic chemical identifiers within and between databases. This can have a great influence especially when merging data and if systematic identifiers are used as a key index for structure integration or cross-querying several databases. Regenerating systematic identifiers starting from their MOL representation and applying well-defined and documented chemistry standardisation rules to all compounds prior to creating them can dramatically increase internal consistency.

Background

The past decade has seen a major increase in the availability of public and commercial chemical databases [1]. Resources such as PubChem (released in 2004) [2] and ChEMBL (released in 2009) [3], with their corresponding web services have gained the trust of many researchers in the fields of cheminformatics, bioinformatics, systems biology, and translational medicine. Because large numbers of compounds and associated structure-activity relationships (SAR) data are published in journals and patents every year, many new data sources have become available, each covering different aspects of the connectivity between the SAR-related entities [4]. With the increasing usage of these resources by scientists from both academia and the pharmaceutical industry, quality control of chemical structures and associated metadata is becoming a necessity [5].

Correctness of a structure extracted from databases has a great impact on predictive ability of computational models for quantitative structure-activity relationships (QSAR) [6]. A recent study by Williams and Ekins [7] on a subset of a chemistry database showed more than 70% errors in the absolute structural integrity, a striking difference to the 5-10% level the authors had anticipated. In another study of database quality, Oprea et al. [8] have illustrated how errors within a database are transferred to other databases following data integration (also mentioned by Williams et al. [9]). Quality issues have also been observed in the relationship between chemical structures and the corresponding identifiers, such as chemical names referring to structures with different stereochemistry or CAS numbers incorrectly associated with a particular salt or mixture [9]. Although these problems are known to exist, there have been no studies that quantify the consistency between structures and their identifiers.

Chemical identifiers can be distinguished in two major classes based on how they are generated. The first consists of systematic identifiers, which are generated algorithmically and should have a one-to-one correspondence with the structure (however, different software could generate different flavours, as is the case for SMILES notations [10,11]). The second class comprises non-systematic chemical identifiers. These are source dependent and usually generated at the point of registration within a particular source (e.g. CAS numbers, PubChem compound identifiers (CIDs) and substance identifiers (SIDs), generic or drug brand names).

Structure depictions are the natural language for chemists. In order to convert the images to a form usable by computers, several file formats and chemical identifiers have been introduced. The MOL file format [12], SMILES notations [10], InChI strings [13], and IUPAC names [14] are arguably the most widely used. In the context of this work we will refer to IUPAC names, SMILES notations, and InChI strings as systematic identifiers.

Most chemical databases are built starting from the MOL file representations of chemical structures, which are linked to systematic and non-systematic identifiers. It is thus crucial that different chemical identifier types represent the same compound. Inconsistencies between systematic identifiers and registered chemical structures can occur for several reasons. For example, systematic identifiers can be generated with different structure-to-identifier conversion tools, with different levels of structure standardisation, or structures and systematic identifiers can be integrated without harmonisation from different sources.

In this study, we investigate the consistency of systematic identifiers of well-defined structures within and between some of the commonly used chemical resources. We also examine the effect of standardisation on this consistency.

Methods

Databases

For this study, we selected a set of well-known publicly available small-molecule databases to cover a wide range of bioactive compounds: DrugBank [15], Chemical Entities of Biological Interest (ChEBI) [16], the Human Metabolome Database (HMDB) [17], PubChem [2], and the NCGC Pharmaceutical Collection (NPC) [18]. Table 1 shows the number of structures and corresponding systematic identifiers in each database. All data were downloaded on March 14, 2012. In this study, only compounds that had MOL files were used. Whenever available, we collected SMILES notations, InChIs strings and IUPAC names. If several SMILES notations were available for a single compound, we selected the isomeric SMILES.

Table 1: Number of structures (MOLs) and systematic identifier counts for databases in this study.

Database	MOL	InChI	SMILES	IUPAC
DrugBank	6506	6391	6504	6489
ChEBI	21367	19076	19725	18798
HMDB	8534	8534	8534	7727
PubChem	5069294	5069293	5069294	4769031
NPC	8024	0	8018	0

In addition to systematic identifiers, cross-references linking records between databases were also downloaded.

The following data were extracted from the resources:

DrugBank [15]. The set of compounds consisted of approved drugs, experimental drugs, nutraceutical drugs, illicit drugs, and withdrawn drugs. Cross-references to other databases were extracted from the DrugCards in DrugBank.

ChEBI [16]. All manually checked and annotated (3 stars) structures with their corresponding systematic identifiers were downloaded. For some of these, ChEBI provides several IUPAC names. In these cases, we only used the first IUPAC name in the ChEBI record for our analyses. Cross-references were obtained from the ChEBI ontology file.

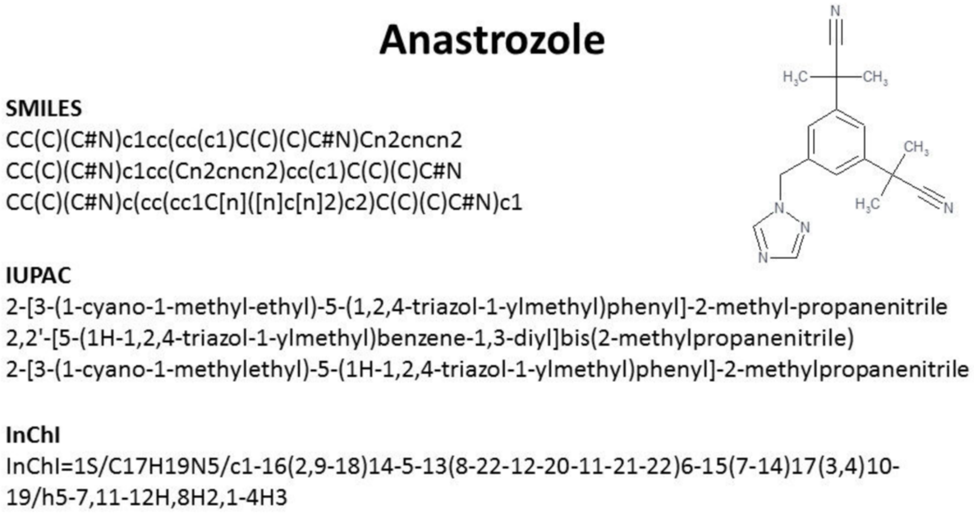
HMDB [17]. All small-molecule metabolites with their corresponding structures were downloaded. Cross-references were extracted from the HMDB MetaboCard files.

PubChem [2]. Based on criteria described previously [4], a set of compounds likely to have SAR and/or other bio-annotations were downloaded from PubChem Compound. PubChem cross-references are only provided on the substance level, not on the compound level, and therefore no PubChem cross-references were used in this study.

NPC [18]. NPC contains the clinical approved drugs from the USA, Europe, Canada and Japan. Compounds and cross-references were downloaded through the NPC Browser 1.1.0 [18]. The export option of the NPC Browser was used to extract data in MOL and SMILES formats. NPC does not provide InChIs strings and IUPAC names.

Consistency of systematic identifiers within a database

To analyse the structural representation consistency of systematic identifiers within a database, we took the MOL representation of a compound as the reference point. Ideally all associated systematic identifiers should represent the same MOL file. In this work, we have used InChI strings for comparisons. InChI (International Chemical Identifier) is a structure-derived tag for a chemical compound. It is an algorithmically produced string of characters, which acts as the unique digital signature of the compound [19]. InChI software developed by IUPAC and InChI Trust, is open-source software and the de facto standard for generating InChI strings [20]. This is not the case for SMILES or IUPAC names (Figure 1). Various flavours of SMILES or IUPAC names are generated by different software to represent the same molecular structure [11,21,22]. Therefore, MOL files and all systematic identifiers were converted into Standard InChIs, using InChI version 1.03, which were then used to perform all comparisons (Figure 2).



Anastrozole

SMILES

CC(C)(C#N)c1cc(cc(c1)C(C)(C)C#N)Cn2cncn2
CC(C)(C#N)c1cc(Cn2cncn2)cc(c1)C(C)(C)C#N
CC(C)(C#N)c(cc(cc1C[n]([n]c[n]2)c2)C(C)(C)C#N)c1

IUPAC

2-[3-(1-cyano-1-methyl-ethyl)-5-(1,2,4-triazol-1-ylmethyl)phenyl]-2-methyl-propanenitrile
2,2'-[5-(1H-1,2,4-triazol-1-ylmethyl)benzene-1,3-diyl]bis(2-methylpropanenitrile)
2-[3-(1-cyano-1-methylethyl)-5-(1H-1,2,4-triazol-1-ylmethyl)phenyl]-2-methylpropanenitrile

InChI

InChI=1S/C17H19N5/c1-16(2,9-18)14-5-13(8-22-12-20-11-21-22)6-15(7-14)17(3,4)10-19/h5-7,11-12H,8H2,1-4H3

Figure 1: Chemical representation of Anastrozole.

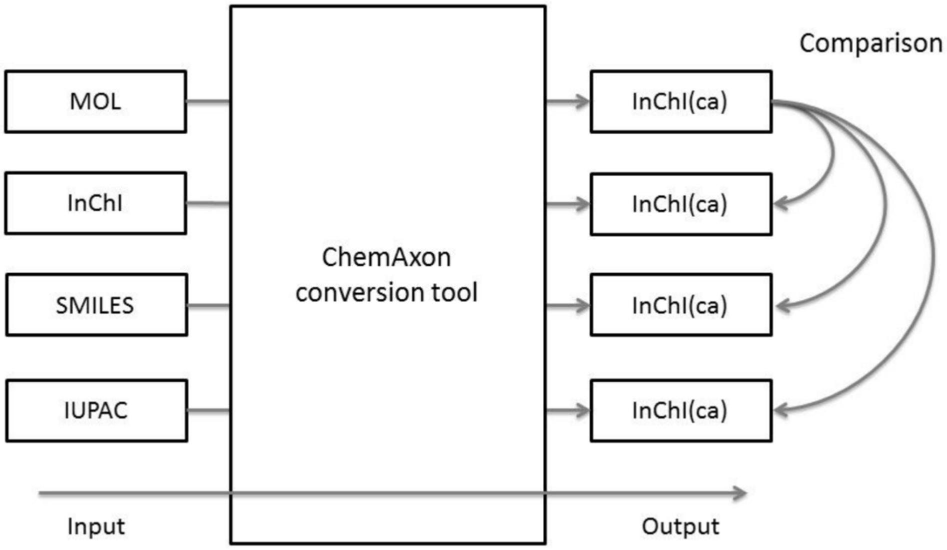


Figure 2: Comparison of MOL representation with systematic identifiers.

Several public and commercial cheminformatics toolkits are currently available for structure manipulation and molecular editing [23]. We used ChemAxon's MolConverter 5.9.1 [24], which has the necessary functionality and is freely available for academic research. For clarity, we refer to Standard InChI strings generated by ChemAxon's MolConverter as InChI(ca).

Consistency of systematic identifiers between databases

To analyse the consistency of systematic identifiers between databases, the cross-reference linkage of compounds was examined. Within the constraints of different chemistry business rules, the chemical entities linked together via the cross-references should represent the same structure based on their MOL representation. We compared the structures using the InChI(ca) generated from the MOLs. We did not consider cross-references where conversion to InChI(ca) failed for one or both of the MOL files. If a compound had multiple cross-references to a single database, each cross-reference was investigated independently. For cross-references to PubChem, we only considered compounds within our subset of the PubChem database.

Standardisation

Inconsistency between systematic identifiers and their MOL representation may partly relate to the different levels of sensitivity in identifier calculation. Currently, different structure normalisation rules can be used to define compound uniqueness [25]. Unfortunately, a unified and agreed set of rules is still lacking [9]. To assess the effect of structure standardisation on the consistency of systematic identifiers within and between databases, we applied a set of rules developed by the Computer-Aided Drug Design group of the National Cancer Institute (NCI/CADD) known as FICTS rules [26,27]. These were applied to each structure and its corresponding systematic identifier.

The FICTS rules include removing small organic fragment (F), ignoring isotopic labels (I), neutralizing charges (C), generating canonical tautomers (T), or ignoring stereochemistry information (S) for a compound. If any of these rules are applied the corresponding upper-case letter is replaced with a “u” (standing for “un-sensitive” [26]). We implemented the FICTS rules using ChemAxon’s Standardizer [28]. To make the results comparable with our other analyses the rules are applied to the InChI(ca) strings.

Results

Conversion of systematic identifiers

Table 2 shows the percentage of successful conversion of the systematic identifiers into InChI(ca) strings by ChemAxon’s MolConverter. This is high for MOLs, SMILES notations and InChI strings in all databases. The lower (90%) MOL conversion for ChEBI was due to the presence of query atom features such as “R” (R-groups) or “*” (= any atom). The main reason for failure in conversion of IUPAC names to Standard InChI strings was challenges for the conversion tool to handle certain structural classes such as steroids, porphyrins, and carbohydrates. The lowest value of IUPAC to InChI(ca) conversion was for HMDB.

Table 2: Successful conversion (in %) of MOL files and systematic identifiers to InChI(ca).

Database	MOL	InChI	SMILES	IUPAC
DrugBank	98.9	100	99.1	93.6
ChEBI	90.6	100	96.8	69.8
HMDB	100	99.9	100	38.1
PubChem	100	100	100	92.6
NPC	99.7	-	100	-

To investigate whether this could be improved, the same procedure was applied with another structure-to-identifier tool, the NCI Chemical Identifier Resolver [29]. This increased successful conversions slightly by 8% but still left the majority of IUPAC names in HMDB unconverted.

Consistency of systematic identifiers within databases

For each compound in a database, we compared the InChI(ca) derived from the MOL file with the InChI(ca) strings from the corresponding systematic identifiers (Figure 2).

Table 3 shows, for each database, the consistency between the MOL representation and the corresponding systematic identifiers, expressed as percentage agreement of matching InChI(ca) strings. If the InChI(ca) could not be generated for a MOL file or a systematic identifier, no comparison was done.

Table 3: Consistency of MOLs and systematic identifiers (in % agreement) within databases.

Database	**MOL–InChI**	**MOL–SMILES**	**MOL–IUPAC**
DrugBank	98.2	98.5	90.0
ChEBI	96.5	96.5	75.3
HMDB	89.3	37.2	55.7
PubChem	97.7	97.8	87.2
NPC	-	93.4	-

In DrugBank there is more than 98% agreement between MOLs and their corresponding InChI strings and SMILES, while the consistency drops to around 90% for IUPAC names. PubChem and ChEBI have slightly lower agreement than DrugBank for InChI strings and SMILES notations, but the IUPAC names in ChEBI show a substantially lower agreement of 75%. The figures are lowest in HMDB with agreements of 37% for MOL-SMILES and 56% for MOL-IUPAC names. NPC only stores SMILES, which have a 93% agreement with their MOL representations.

Standardisation

FICTS rules were applied to the InChI(ca) strings derived from the MOL files and systematic identifiers and all comparisons were redone. Table 4 show the results. Stereochemistry has the most significant impact. For example, the consistency for MOL-SMILES notations and MOL-IUPAC names in HMDB increased with 61 and 29 percentage points. ChEBI and PubChem also show a considerable increase in agreement between IUPAC names and MOL files. In addition to stereochemistry, the changes made by standardising tautomers also improved the consistency, with the largest effect on HMDB. Charges, fragments and isotopic labels had a small or no effect on the consistency.

Table 4: Effect of different standardisation rules on the consistency between MOL files and systematic identifiers (in % agreement).

Database	**Comparison**	**FICTS**	**uICTS**	**FuCTS**	**FIuTS**	**FICuS**	**FICTu**
DrugBank	MOL–InChI	98.2	99.0	99.0	99.0	99.4	99.8
	MOL–SMILES	98.5	98.6	98.6	98.6	99.5	99.7
	MOL–IUPAC	90.0	90.1	90.0	90.1	93.5	96.2
ChEBI	MOL–InChI	96.5	98.9	98.5	98.4	99.2	99.6
	MOL–SMILES	96.5	96.6	96.6	96.6	99.6	99.8
	MOL–IUPAC	75.3	75.6	75.4	77.1	79.7	91.9
HMDB	MOL–InChI	89.3	89.8	89.7	90.3	89.9	98.5
	MOL–SMILES	37.2	37.3	37.2	38.0	43.1	98.3
	MOL–IUPAC	55.7	55.8	55.8	57.5	58.8	84.8
PubChem	MOL–InChI	97.7	97.9	97.9	97.9	99.3	99.9
	MOL–SMILES	97.8	97.9	97.9	97.8	99.2	99.9
	MOL–IUPAC	87.2	87.7	87.5	87.2	93.7	97.2
NPC	MOL–SMILES	93.4	93.5	93.4	93.4	98.0	99.8

Consistency of systematic identifiers between databases

Table 5 shows the agreement between the MOL files for compounds with inter-database cross-references. This varies from 25.8% to 93.7%, but for most cases is around 60-75%. The low value for cross-references from NPC to PubChem can be attributed to 1527 compounds in NPC that have more than one (average 5.7, median 3) cross-reference to PubChem CIDs. The agreement for the 2475 compounds in NPC that have just one cross-reference to PubChem is 79.3%. Note that the agreement for the cross-references in DrugBank or HMDB to ChEBI is about 20% higher than the other way around.

Table 5: Agreement between MOL files of compounds that have a cross-reference in one database (row) to another database (column). The number of cross-references is given in parentheses.

	DrugBank	**ChEBI**	**HMDB**	**PubChem**	**NPC**
DrugBank	-	72.1% (1666)	-	93.7% (4723)	-
ChEBI	54.3% (1288)	-	45.6% (114)	-	-
HMDB	-	64.0% (1433)	-	76.0% (2217)	-
PubChem	-	-	-	-	-
NPC	76.7% (1320)	-	-	25.8% (9557)	-

Since our results indicate that stereochemistry standardisation may substantially improve the consistency of systematic identifiers within databases (Table 4), we also assessed the consistency between databases after applying the FICTu rule (Table 6).

Table 6: Agreement between MOL files of compounds that have a cross-references in one database (row) to another database (column) after stereochemistry standardisation.

	DrugBank	**ChEBI**	**HMDB**	**PubChem**	**NPC**
DrugBank	-	91.4%	-	95.6%	-
ChEBI	68.6%	-	93.0%	-	-
HMDB	-	82.0%	-	89.8%	-
PubChem	-	-	-	-	-
NPC	93.4%	-	-	47.6%	-

Stereochemistry annotation increases the agreement for most databases by around 15-20%. The largest increase (47.4%) is seen for cross-references linking ChEBI to HMDB.

The agreement between NPC and PubChem also increases but more than half of the cross-references still link MOL files that do not match. For compounds that have just one cross-reference the agreement increased from 79.3% to 91.0%.

Discussion

While the importance of data quality control in chemical resources has been discussed previously [5-7,9], to our knowledge this is the first study to assess the consistency of structural representations of systematic identifiers within and between small-molecule databases. The assumption was that systematic identifiers should correspond with the registered MOL file. Standard InChI strings were used as a basis for this comparison because of the unique algorithm available, unlike for SMILES notations and IUPAC names where multiple strings can represent the same compound.

To provide comparable results and remove the influence of different structure-to-identifier software, only ChemAxon's MolConverter [24] was used for all name conversions. Compounds where MOL files or systematic identifiers did not convert to InChI strings were disregarded. To quantify the potential influence of different structure-to-identifier software we compared the Standard InChI strings generated from the MOL files using ChemAxon's MolConverter [24] with those of Xemistry's CACTVS chemoinformatics toolkit [30,31]. The comparison showed 98.9% agreement for HMDB, 98.3% for PubChem, 97.6% for DrugBank, 96.4% for ChEBI, and 94.2% for NPC in cases were both tools managed to convert MOL files to InChI strings. The differences are small and likely to be caused by the way the tools handle the MOL files. We consider it unlikely that our results would essentially have changed by using another conversion tool.

The consistency of systematic identifiers with their corresponding MOL representations varies widely (Table 3). The highest agreement was obtained for DrugBank and PubChem, the lowest for HMDB. The higher consistency values for PubChem may be explained by their procedure for generating systematic identifiers [32]: starting from the MOL files, InChI strings are calculated based on the IUPAC Standard InChI software and SMILES notations and IUPAC names are generated by OpenEye software [33]. Unfortunately, because other databases do not clearly describe their procedures it remains unclear how possible differences may have affected consistency.

Application of the FICTS sensitivity rules [26] gave us further insight. We found that disregarding stereochemistry and, to a lesser extent, tautomers boosted the consistency, in particular of MOL-IUPAC names (Table 4). The other sensitivity levels had a much lower or no effect. Thus, differences in stereochemistry between MOL files and systematic identifiers appear the single most important cause of inconsistencies. For ChEBI and HMDB, the agreement between MOLs and IUPAC names remained low even with stereochemistry insensitive matching.

The consistency of systematic identifiers between databases, as measured by the agreement of MOL files in different databases linked by cross-references, ranged from 26% to 94% (Table 5). The value of cross-references lies in the consistency of the structural representation of the data and our study shows these have many errors. Disregarding stereochemistry on the registered MOL files increased the agreement, but a considerable percentage of the cross-references remained inconsistent.

Integration of different chemical databases should consider these problems. Merging databases using different structure identifiers as indexes for integration can reduce quality. Instead a unique representation such as MOL files can be used as the basis of integration. Other systematic identifiers can be generated later on the validated structure within the database.

Inconsistencies within databases may steer curation efforts, and by combining the information on inconsistencies for a specific compound may even suggest which of the names or representations are wrong.

In a recent article by Williams et al. [9] several solutions have been proposed to reduce errors in databases. In addition to improved curation the use of structure validation filters for incorrect valance, atom labels, aromatic bonds, charges, stereochemistry and duplication was suggested. In another recent study, O'Boyle [11] proposed a standard method to generate canonical SMILES based on InChI strings, in order to create the same canonical SMILES using different toolkits. Our results quantify the issues raised in these studies. We have shown that a set of well-defined standardisation rules is essential while constructing systematic identifiers (can gain up to 50% increase in consistency), and that stereochemistry has an important contribution to this inconsistency.

Our approach of testing the consistency of systematic identifiers is general and can be applied to other databases and may prove valuable in data curation and integration efforts. Using a similar approach, we also plan to investigate the consistency of non-systematic identifiers in chemical resources.

Conclusions

The degree of consistency within systematic chemical identifiers varies between data sources. When building a new database, de novo recalculation is superior to recycling and creating systematic identifiers starting from the same primary structural representation (e.g. MOL) will improve the quality of the final product. Extra consideration should be taken into account if systematic identifiers are going to be used as a key index for merging databases. Well-defined and documented chemistry standardisation rules applied to all compounds can greatly decrease the number of errors and expedite integration.

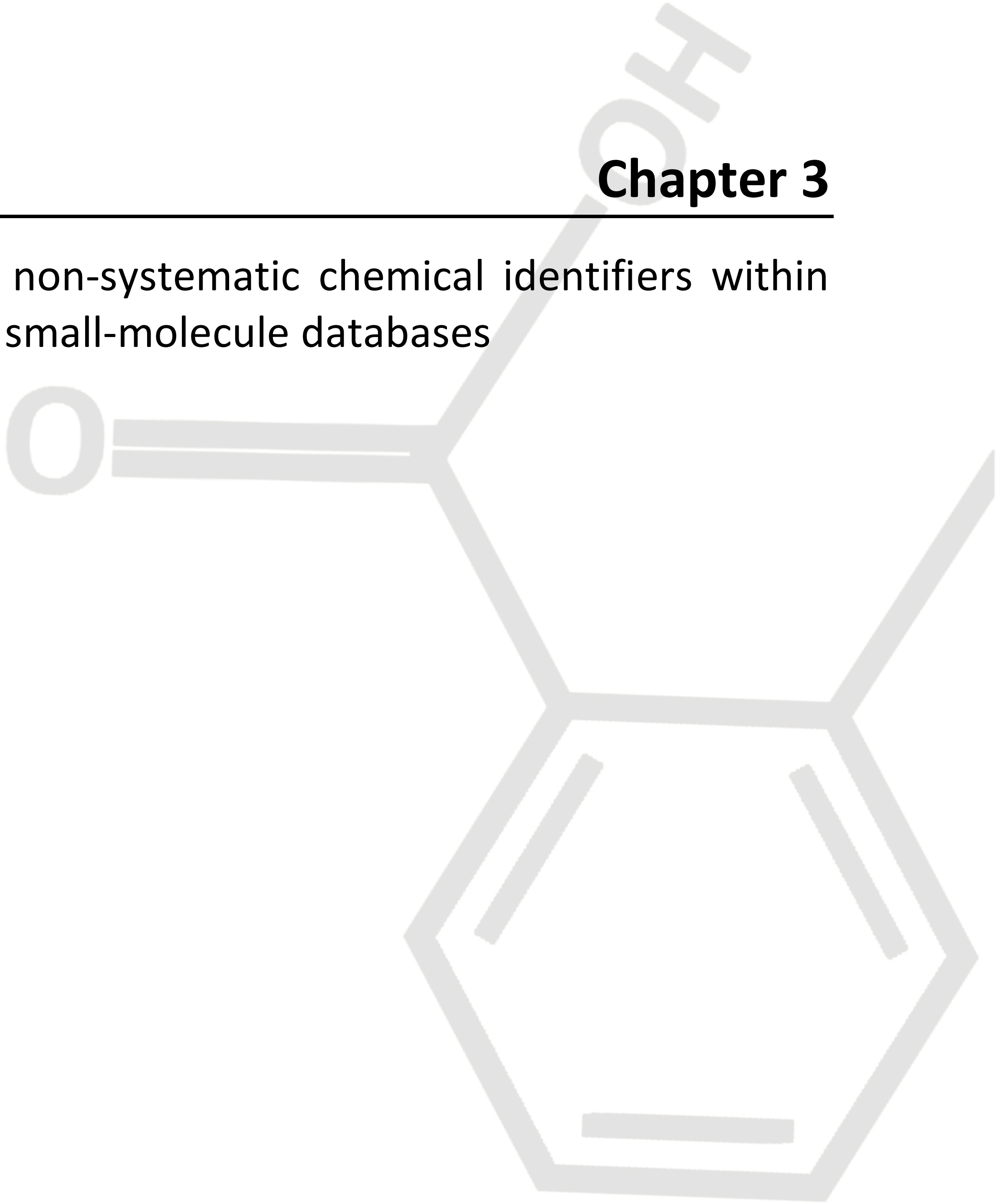
Finally, we have shown that inconsistency exists between the structural representations of compounds that are linked via cross-references within databases. Inconsistency here can have deleterious effects when merging data from or cross-querying multiple databases.

References

1. Williams AJ: **Public chemical compound databases.** *Curr Opin Drug Discov Devel* 2008, **11:**393–404.
2. Bolton E, Wang Y, Thiessen P, Bryant S: **PubChem: integrated platform of small molecules and biological activities.** *Annual reports in computational chemistry*. 12th edition. Washington, DC: American Chemical Society; 2008.
3. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res* 2012, **40:**D1100–D1107.
4. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, Varkonyi P, Xie PH: **Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data.** *Drug Discov Today* 2011, **16:**1019–1030.
5. 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:**1189–1204.
6. Young D, Martin T, Venkatapathy R, Harten P: **Are the chemical structures in your QSAR correct?** *QSAR Comb Sci* 2008, **27:**1337–1345.
7. Williams AJ, Ekins S: **A quality alert and call for improved curation of public chemistry databases.** *Drug Discov Today* 2011, **16:**747–750.
8. Opera TI, Olah M, Ostopovici L, Rad R, Mracec M: **On the propagation of errors in the QSAR literature**. In *EuroQSAR 2002 designing drugs and crop protectants: processes, problems and solutions*. 2003rd edition. Edited by Ford M, Livingstone D, Dearden J, Van de Waterbeemd H. New York: Blackwell Publishing; 2003:314–315.
9. Williams AJ, Ekins S, Tkachenko V: **Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation.** *Drug Discov Today* 2012, **17:**685–701.
10. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28:**31–36.
11. O'Boyle NM: **Towards a universal SMILES representation - a standard method to generate canonical SMILES based on the InChI.** *J Cheminf* 2012, **4:**22.
12. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J: **Description of several chemical structure file formats used by computer programs developed at molecular design limited.** *J Chem Inf Comput Sci* 1992, **32:** 244-255.
13. *History of InChI*. http://www.inchi-trust.org/inchi/.
14. *About IUPAC*. http://www.iupac.org/home/about.html.
15. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39:**D1035–D1041.
16. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical entities of biological interest: an update.** *Nucleic Acids Res* 2010, **38:**D249–D254.
17. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De Souza A, Zuniga A,

Dawe M, *et al*: **HMDB: a knowledgebase for the human metabolome.** *Nucleic Acids Res* 2009, **37:**D603–D610.

18. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP: **The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics.** *Sci Transl Med* 2011, **3:**80ps16.
19. *InChI FAQ*. http://www.inchi-trust.org/fileadmin/user_upload/html/inchifaq/inchi-faq.html.
20. *InChI trust*. http://www.inchi-trust.org/home/.
21. Garfield E: *An algorithm for translating chemical names to molecular formulas*. Philadelphia: Institute for Scientific Information; 1961.
22. Vazquez M, Krallinger M, Leitner F, Valencia A: **Text mining for drugs and chemical compounds: methods, tools and applications.** *Molecular Informatics* 2011, **30:**506–519.
23. Lowe DM, Corbett PT, Murray-Rust P, Glen RC: **Chemical name to structure: OPSIN, an open source solution.** *J Chem Inf Model* 2011, **51:**739–753.
24. *ChemAxon – naming*. http://www.chemaxon.com/products/name-to-structure/.
25. Martin E, Monge A, Duret JA, Gualandi F, Peitsch MC, Pospisil P: **Building an R&D chemical registration system.** *J Cheminf* 2012, **4:**11.
26. Sitzmann M, Filippov IV, Nicklaus MC: **Internet resources integrating many small-molecule databases.** *SAR QSAR Environ Res* 2008, **19:**1–9.
27. Muresan S, Sitzmann M, Southan C: **Mapping between databases of compounds and protein targets.** *Methods Mol Biol* 2012, **910:**145–164.
28. *Standardize - structure canonicalization and more*. http://www.chemaxon.com/products/standardizer/.
29. *Chemical identifier resolver beta 4*. http://cactus.nci.nih.gov/chemical/structure.
30. Ihlenfeldt WD, Takahashi Y, Abe H, Sasaki S: **Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility.** *J Chem Inf Comp Sci* 1994, **34:**109–116.
31. *Xemistry chemoinformatics*. http://www.xemistry.com.
32. *PubChem SD file formatted data, V2.0.1*. ftp://ftp.ncbi.nlm.nih.gov/pubchem/data_spec/pubchem_sdtags.pdf.
33. Wlodek S, Skillman AG, Nicholls A: **Automated ligand placement and refinement with a combined force field and shape potential.** *Acta Crystallogr D: Biol Crystallogr* 2006, **62:**741–749.



Chapter 3

Ambiguity of non-systematic chemical identifiers within and between small-molecule databases

*Published: **Akhondi SA**, Muresan S, Williams AJ, Kors JA*

Journal of Cheminformatics 2015, 7:1–10

Abstract

Background

A wide range of chemical compound databases are currently available for pharmaceutical research. To retrieve compound information, including structures, researchers can query these chemical databases using non-systematic identifiers. These are source-dependent identifiers (e.g., brand names, generic names), which are usually assigned to the compound at the point of registration. The correctness of non-systematic identifiers (i.e., whether an identifier matches the associated structure) can only be assessed manually, which is cumbersome, but it is possible to automatically check their ambiguity (i.e., whether an identifier matches more than one structure). In this study we have quantified the ambiguity of non-systematic identifiers within and between eight widely used chemical databases. We also studied the effect of chemical structure standardization on reducing the ambiguity of non-systematic identifiers.

Results

The ambiguity of non-systematic identifiers within databases varied from 0.1 to 15.2% (median 2.5%). Standardization reduced the ambiguity only to a small extent for most databases. A wide range of ambiguity existed for non-systematic identifiers that are shared between databases (17.7-60.2%, median of 40.3%). Removing stereochemistry information provided the largest reduction in ambiguity across databases (median reduction 13.7 percentage points).

Conclusions

Ambiguity of non-systematic identifiers within chemical databases is generally low, but ambiguity of non-systematic identifiers that are shared between databases, is high. Chemical structure standardization reduces the ambiguity to a limited extent. Our findings can help to improve database integration, curation, and maintenance.

Background

A wide range of chemical compound databases are currently available for pharmaceutical research [1]. They provide a variety of chemical information [2], most importantly compound structures, which can be used for different purposes, such as chemical predictive modelling [3] or quantitative structure-activity relationships modelling [4]. To retrieve information about a compound, researchers can query these chemical databases using one of many available compound identifiers. Information retrieval based on automatic extraction of chemical identifiers from scientific literature or patents, is becoming increasingly important as the large amount of such unstructured texts makes manual extraction and analysis cumbersome [5-7]. Text mining methods that extract compound-target or drug-disease relationships from text, can provide valuable new insights [8] or support database curation [9, 10]. The correctness of the chemical identifiers that link to the chemical structures in the databases can greatly affect the results of cheminformatics analyses [11, 12].

Chemical identifiers fall into two main classes. The first class consists of systematic identifiers, which are algorithmically defined based on the chemical structure of the compound [13]. Among the systematic identifiers are IUPAC names [14], SMILES [15], and International Chemical Identifiers (InChIs) [16, 17]. We have previously investigated the correctness or consistency of systematic identifiers (i.e., whether an identifier matches the associated structure) within and across small-molecule databases, and found many inconsistencies [13]. We also checked whether the inconsistencies could be reduced by different chemical structure standardizations (e.g., removal of fragments, or ignoring isotopes), but this was only the case to a limited extent [13].

The second class of chemical identifiers consists of non-systematic identifiers. These are source-dependent identifiers which are usually assigned to the compound at the point of registration in a chemical database [13]. Brand names, generic names, research codes, chemical abstracts service (CAS) registry numbers, and database identifiers are examples of such non-systematic identifiers. Since there is no algorithmic relationship between non-systematic identifiers and structures, the correctness of these identifiers can only be assessed manually, which has proven cumbersome [1]. However, it is possible to automatically check the ambiguity of non-systematic identifiers (i.e., whether an identifier matches more than one structure). The extent of this ambiguity problem is unknown and not yet quantified.

Here, we investigate the ambiguity of non-systematic identifiers within and between small-molecular databases, before and after chemical structure standardisation.

Methods

Databases

We selected eight well-known chemical databases covering a wide range of bioactive compounds: Chemical Entities of Biological Interest (ChEBI) [18], ChEMBL [19], ChemSpider [20], DrugBank [21], the Human Metabolome Database (HMDB) [9, 22], the NCGC Pharmaceutical

Collection (NPC) [23], PubChem [24], and the Therapeutic Target Database (TTD) [25, 26]. We focused on compound records that had associated chemical structures in the form of MOL files [27]. For each record, we extracted the structure file and gathered all chemical identifiers (available from possibly different record fields), except for identifiers explicitly tagged as IUPAC names, SMILES, or InChIs. For example, identifiers for the antibiotic “ampicillin” included “ampicilina”, “ampicillin acid”, “AMP”, “AP”, “ABPC”, “ay-6108”, “DB00415”, “penbritin”, “totacillin”, “PEN A/N”, “Prestwick3_000114”, “Ampi-bol”, “Aminobenzylpenicillin” and, “brl 1341”. Note that extracted identifiers may include database identifiers (such as “DB00415”) that appear in the name fields of the chemical records. Typically, for a given chemical database, database identifiers in its name fields come from other databases, and local database identifiers are only used as record identifiers (and not extracted). All data were downloaded in February 2013. The identifiers extracted from all databases, except ChemSpider which is a commercial database, are made available through www.biosemantics.org. In the following, we briefly describe the databases, indicating the version that was used (if versioning was available) and the fields from which identifiers were extracted.

ChEBI is a database of molecular entities, focusing on small chemical compounds [18]. ChEBI provides an ontological classification with parent and child relationships. We extracted data for all three-star (i.e., manually annotated) compounds from ChEBI SD files. This included synonyms, ChEBI names, brand names, and International Non-proprietary Names (INN).

ChEMBL is a large-scale bioactivity database containing information for drug-like bioactive compounds [19]. In addition to literature-derived data ChEMBL also contains Food and Drug Administration (FDA) approved drugs. The data available through ChEMBL have been manually extracted and standardized [19]. We used a local installation of ChEMBL version 14. Extracted fields include preferred name, synonyms, FDA alternative names, trade names, INN, United States Adopted Names (USAN), and United States Pharmacopoeia names (USP).

ChemSpider is a chemical database containing information of compounds gathered from over 500 different data sources [20]. ChemSpider structures and their corresponding identifiers were made available from the Royal Society of Chemistry (RSC) [28]. We focused on compounds that have structure-activity relationships or other biological annotations. Similar selection criteria as defined by Muresan et al. [29] were provided to the ChemSpider team to extract the ChemSpider data. Subsets of chemicals such as “make on demand” chemicals from screening library vendors without names other than computationally generated systematic names were excluded, as were the datasets that have been deprecated from ChemSpider during curation. We also considered a subset of the ChemSpider data that only contained information that was validated with the use of crowdsourcing, including curation work performed by members of the ChemSpider technical support team (ChemSpider-V) [20, 30]. For each compound, we were provided with all preferred terms and synonyms.

DrugBank provides information regarding drugs, including chemical, pharmacological and pharmaceutical drugs and their targets [21]. DrugBank data are curated by a curation team based on primary literature sources. During production and maintenance all synonyms and brand names within DrugBank are extensively reviewed and only the most common synonyms are kept

[31]. We used DrugBank version 3.0, and extracted generic names, synonyms, CAS numbers, and brand names from the DrugBank SD files and DrugCards.

HMDB contains small-molecule metabolites found in the human body. The database links chemical, clinical, molecular-biology, and biochemistry data. HMDB is both automatically and manually curated [9, 22]. We used HMDB version 3. All generic names, CAS numbers, and synonyms were extracted from HMDB SD files and MetaboCards.

NPC provides clinically-approved drugs from USA, Europe, Canada, and Japan for high-throughput screening [23]. In addition NPC provides chemical-related information gathered from different sources, such as the KEGG database. Using NPC browser 1.1.0, we extracted preferred names and synonyms.

PubChem is a database that provides information on the biological activities of small molecules [24]. PubChem consists of three different databases: a compound database (with currently about 61 million entries), a substance database (about 157 million entries), and a bioassay database (more than 1 million entries). The compound database was used to extract structures for a subset of compounds that had structure-activity relationships or other biological annotations. This subset of compounds was introduced by Muresan et al. [29] and is the same subset of PubChem compounds that we used in our previous study on the consistency of systematic identifiers [13]. The PubChem compound database does not contain non-systematic identifiers. This information is available through the PubChem substance database. The relations between PubChem substance identifiers (SIDs) and compound identifiers (CIDs), which have been created by PubChem through in-house chemical structure standardization [24], are specified in the "PubChem_CID_associations" tag available in the downloadable SD files [32]. We used the relations between SIDs and CIDs to extract the non-systematic identifiers (synonyms and identifiers) from the substance database and assign them to the corresponding compounds [24].

TTD provides therapeutic protein and nucleic acid targets and drug information including targeted disease and pathway [25, 26]. We used TTD version 4.3.02. All synonyms, trade names, and drug names were extracted.

Filtering

The fields with non-systematic identifiers that were extracted from the databases may also contain systematic identifiers (e.g., a field with synonyms may not distinguish between the two types of identifiers). Systematic identifiers were automatically filtered out from the extracted identifiers with the use of two name-to-structure converters, ChemAxon's MolConverter [33] and the open source tool OPSIN (Open Parser for Systematic IUPAC Nomenclature) [34]. Both tools are freely available for academic research. We used two different name converters since the algorithms that they implement to recognize systematic identifiers may differ slightly (mostly when considering IUPAC names). Each extracted identifier was fed into the converters and only considered non-systematic if neither tool recognized it as systematic. For example, the term "(2S,5R,6R)-6-{[(2R)-2-amino-2-phenylacetyl]amino}-3,3-dimethyl-7-oxo-4-thia-1-azabicyclo[3.2.0]heptane-2-carboxylic acid" was not labelled as a IUPAC name in DrugBank "DB00415" but it was filtered out through this step.

Ambiguity within and across databases

A non-systematic identifier was considered ambiguous within a database if it appeared in multiple records in the database, i.e., if multiple structures were provided for the same identifier. Ambiguity was measured as the percentage of unique identifiers within a database that are ambiguous.

An identifier was considered ambiguous across two databases if the structures (as defined by their MOL files) of the compounds associated with the identifier in the two databases were different. If an identifier was ambiguous in one or both of the databases (i.e., the identifier was associated with multiple compounds within the database(s)), the identifier was also considered ambiguous across databases. Ambiguity was calculated as the percentage of unique shared identifiers between databases that are ambiguous.

To compare two MOL files, we used the same approach as in our previous study [13]. Briefly, each MOL file was converted into a Standard InChI with ChemAxon's MolConverter [33], providing a unique textual representation of the MOL file. The two InChI strings were then compared to determine whether the corresponding structures were the same. No comparison was made if an InChI could not be generated.

Standardization

In the process of creating MOL files for compounds, databases can apply different sensitivity settings [2]. These settings pertain to including or ignoring fragments, isotopic labels, charges, canonical tautomers, or stereochemical information. Different sensitivity settings can result in different Standard InChI strings for the same compound, and thus are a potential source of ambiguity. Standardization of the MOL files can help to reduce such ambiguities.

The Computer-Aided Drug Design group of the National Cancer Institute defined a set of rules called FICTS to standardize the structural representation of compounds [2, 35]. FICTS rules correspond to five standardisation levels that affect structural information. The rules remove small fragments (F), disregard isotopes (I) and charges (C), generate canonical tautomers (T), or ignore stereochemical information (S). Any combination of the five rules can be applied and is expressed by converting the corresponding upper-case letter of the term "FICTS" into a "u" (for "un-sensitive"). ChemAxon's Standardizer [36] was used to execute these standardization rules.

Results

Databases

For each database, Table 1 shows the number of compounds with at least one non-systematic identifier, and the total number of non-systematic identifiers (not unique). The databases vary greatly in size and in the average number of non-systematic identifiers per compound, ranging from 1.3 for ChemSpider-V and ChEMBL to 35.4 for TTD. The large average for TTD can be attributed to the presence of a large number of database identifiers for many of the compounds.

Table 1: Number of compounds and non-systematic identifiers in different chemical databases.

Database	Compounds	Identifiers	Identifiers / compound
PubChem	4,232,875	15,211,133	3.6
ChemSpider	6,646,902	10,063,709	1.5
ChemSpider-V	654,052	850,601	1.3
HMDB	37,761	308,733	8.2
NPC	14,814	131,290	8.9
TTD	2,977	105,407	35.4
ChEBI	15,633	41,956	2.7
ChEMBL	21,398	28,011	1.3
DrugBank	3,769	26,780	7.1

Ambiguity of non-systematic identifiers within databases

Table 2 shows the ambiguity of non-systematic identifiers and the average number of compounds per ambiguous identifier within the databases. HMDB has 15.2% ambiguity, much larger than for any of the other databases. On average, an ambiguous identifier in HMDB is associated with 6.1 compounds, but the distribution is highly skewed. For example, the two most ambiguous identifiers in HMDB, “Triglyceride” and “Triacylglycerol”, are each associated with about 14,000 compounds. Moreover, HMDB contains 176 non-systematic identifiers with more than 100 structures (100 being an arbitrary number chosen for the purpose of comparison). The only other databases that contain identifiers that are associated with more than 100 structures, are ChemSpider (39 identifiers) and PubChem (16 identifiers). Some of these identifiers are unspecific, e.g., “ester” is linked to 228 structures in ChemSpider.

TTD is the database with the second-largest ambiguity (4.6%), but none of the ambiguous identifiers in TTD are associated with more than three compounds. This is also reflected in the low average number of compounds per ambiguous identifier (2.1), close to the minimum of 2 that would be reached if all ambiguous identifiers were associated with exactly two compounds. The ambiguity of ChemSpider-V (0.6%) is much lower than the ambiguity of ChemSpider (2.5%), suggesting a positive effect of curation. However, when we recalculated the ambiguity of the ChemSpider-V records prior to curation, we found an ambiguity of 0.7%. Therefore, the curation effort only slightly reduced ambiguity within ChemSpider-V, possibly because it focused more on establishing the correctness of compound structures. DrugBank has the lowest ambiguity of non-systematic identifiers (0.1%).

Table 2: Ambiguity of non-systematic identifiers and the average number of compounds per ambiguous identifier, within databases.

Database	Unique identifiers	Ambiguous identifiers	Ambiguity (%)	Compounds / ambiguous identifier
HMDB	173,455	26,430	15.2	6.1
TTD	100,570	4,607	4.6	2.1
ChEMBL	26,910	1,050	3.9	2.1
NPC	112,717	3,455	3.1	2.1
ChemSpider	9,691,277	245,541	2.5	2.5
ChEBI	41,023	827	2.0	2.1
PubChem	14,937,728	201,621	1.3	2.4
ChemSpider-V	842,128	5,401	0.6	2.3
DrugBank	26,759	20	0.1	2.1

Ambiguity of non-systematic identifiers between databases

Table 3 presents for each pair of databases the number of unique non-systematic identifiers that are shared between the databases. The first figure in the parentheses indicates the ambiguity of these shared identifiers, i.e., the percentage of shared identifiers for which the corresponding structures in the two databases are different. For example, the identifier “floxuridine” occurs in ChEBI and in ChEBML, but the corresponding structures in these two databases do not match, and thus the identifier is ambiguous. The second figure in the parentheses shows the percentage of the shared identifiers that are ambiguous within one or both of the databases, and thus are ambiguous across databases by definition. For example, “ofloxacin” is shared between ChEMBL and HMDB, but is also ambiguous within HMDB because it is associated with two different structures (in records HMDB01929 and HMDB15296). Therefore, the identifier is considered ambiguous, even though one of the structures in HMDB (HMDB15296) matches the one in ChEMBL.

Table 3: Number of shared non-systematic identifiers between databases, ambiguity of the shared identifiers (first figure in parentheses, in italic), and the percentage of shared identifiers that are ambiguous within at least one of the databases (second figure in parentheses).

Database	**ChEBI**	**ChEMBL**	**ChemSpider**	**ChemSpider-V**	**DrugBank**	**HMDB**	**NPC**	**PubChem**
ChEMBL	1,886 (*39.5*/18.2)							
ChemSpider	28,281 (*30.9*/24.1)	23,584 (*29.9*/22.3)						
ChemSpider-V	5,081 (*39.9*/9.8)	4,303 (*43.6*/16.6)						
DrugBank	2,981 (*28.7*/3.4)	4,108 (*39.6*/14.7)	19,222 (*50.7*/32.0)	6,985 (*45.2*/6.7)				
HMDB	4,529 (*49.6*/10.6)	2,325 (*48.4*/17.7)	27,608 (*57.3*/29.5)	11,774 (*43.9*/8.8)	5,515 (*30.7*/5.2)			
NPC	5,516 (*40.7*/6.1)	6,858 (*46.4*/15.1)	62,527 (*60.2*/26.8)	18,709 (*48.6*/7.4)	22,377 (*21.9*/2.0)	6,815 (*44.4*/7.4)		
PubChem	24,331 (*36.9*/26.1)	25,607 (*33.1*/28.9)	2,275,338 (*17.7*/8.5)	99,334 (*41.6*/19.2)	24,929 (*46.8*/39.4)	35,905 (*43.3*/28.3)	68,280 (*49.8*/29.6)	
TTD	4,854 (*27.7*/7.6)	5,019 (*36.9*/16.9)	50,182 (*32.3*/18.4)	8,305 (*40.3*/10.4)	17,232 (*18.2*/6.8)	6,256 (*43.0*/11.2)	23,669 (*22.4*/6.9)	98,853 (*25.4*/23.0)

Ambiguity between two databases varies widely, from 17.7% (for DrugBank and TTD) to 60.2% (for NPC and ChemSpider). Overall, the lowest ambiguity values between a given database and the other databases are seen for TTD (median ambiguity over all databases 30.0%), while highest values occur for NPC (median 45.4%), and HMDB (median 44.2%).

The percentage of shared identifiers that are ambiguous within either or both of the databases (i.e., are ambiguous across databases by definition) also varies greatly. For instance, 39.4% of the shared identifiers between DrugBank and PubChem are also ambiguous within the databases, largely accounting for the overall ambiguity of 46.8%. (This means that only 7.4% of the shared identifiers are ambiguous across but not within the databases.) Similar values are seen for ChEMBL and PubChem (33.1% overall ambiguity and 28.9% ambiguity due to identifiers that are ambiguous within the databases) and PubChem and TTD (25.4% and 23.0%, respectively). On the other hand, for DrugBank and NPC only 2.0% ambiguity is due to ambiguous identifiers within the databases (overall ambiguity 21.9%), and for DrugBank and ChEBI only 3.4% (overall 28.7%).

Effect of standardisation

Table 4 shows the effect of different types of standardization on reducing the ambiguity of non-systematic identifiers within databases. For most databases, standardization has little effect on ambiguity (median change for each setting less than 0.5 percentage point). The largest changes are seen for TTD and ChEMBL, in particular for removing fragments (uICTS). Overall, removing fragments and disregarding stereochemistry (FICTu) gives the largest changes, while disregarding isotopes (FuCTS) has the lowest effect. Notably, standardization does not affect HMDB, the most ambiguous database.

Table 4: Effect of standardization on the ambiguity of non-systematic identifiers (in %) within databases.

Database	**FICTS**	**uICTS**	**FuCTS**	**FIuTS**	**FICuS**	**FICTu**
HMDB	15.2	15.2	15.2	15.2	15.2	15.2
TTD	4.6	1.8	2.1	2.0	2.1	2.1
ChEMBL	3.9	2.0	3.8	3.9	3.9	3.4
NPC	3.1	2.7	2.7	2.7	2.7	2.7
ChemSpider	2.5	2.3	2.5	2.5	2.2	1.9
ChEBI	2.0	1.8	1.9	1.4	1.8	1.6
PubChem	1.4	1.2	1.3	1.3	0.6	0.6
ChemSpider-V	0.6	0.6	0.6	0.6	0.6	0.3
DrugBank	0.1	0.1	0.1	0.1	0.1	0.1

We also computed the effect of different standardization settings on the ambiguity of non-systematic identifiers across databases. Table 5 shows the results for removing fragments (uICTS) and disregarding stereochemistry (FICTu), which gave the largest reductions in ambiguity. Results for the other standardization settings (FuCTS, FIuTS, and FICuS) are available as Additional file 1.

Overall, ignoring stereochemistry information gave the largest ambiguity reduction (median decrease of 13.7 percentage points), but the remaining ambiguity between databases was still considerable (median 25.4%). The largest improvements were seen for HMDB and NPC (23.2 percentage points) and for HMDB and ChemSpider (21.9 percentage points). Removal of small fragments resulted in a median reduction in ambiguity of 4.9 percentage points. The highest reduction was obtained for ChEBI and ChEMBL (17.5 percentage points).

Table 5: Effect of standardization on the ambiguity of non-systematic identifiers (in %) across databases.

Database	Standardization	ChEBI	ChEMBL	ChemSpider	ChemSpider-V	DrugBank	HMDB	NPC	PubChem
ChEMBL	FICTS	39.5							
	uICTS	22.0							
	FICTu	32.6							
ChemSpider	FICTS	30.9	29.9						
	uICTS	28.4	25.0						
	FICTu	19.5	17.8						
ChemSpider-V	FICTS	39.9	43.6						
	uICTS	36.5	34.1						
	FICTu	26.1	27.3						
DrugBank	FICTS	28.7	39.6	50.7	45.2				
	uICTS	15.5	22.6	41.4	37.2				
	FICTu	23.3	32.6	35.9	33.4				
HMDB	FICTS	49.6	48.4	57.3	43.9	30.7			
	uICTS	47.4	36.1	54.4	42.4	30.4			
	FICTu	32.3	33.0	34.4	23.3	16.1			
NPC	FICTS	40.7	46.4	60.2	48.6	21.9	44.4		
	uICTS	31.2	31.1	45.9	37.3	21.3	43.5		
	FICTu	26.8	36.2	45.1	31.6	13.5	21.2		
PubChem	FICTS	36.9	33.1	17.7	41.6	46.8	43.3	49.8	
	uICTS	32.9	25.2	16.1	37.1	37.6	40.9	38.4	
	FICTu	24.1	24.0	9.0	25.4	34.6	26.7	35.1	
TTD	FICTS	27.7	36.9	32.3	40.3	18.2	43.0	22.4	25.4
	uICTS	20.9	24.6	27.8	32.7	16.8	41.1	20.6	21.6
	FICTu	15.2	26.0	17.8	23.0	10.1	22.0	9.2	13.8

Discussion

We quantified the ambiguity of non-systematic identifiers within and between eight widely used chemical databases. Our results show an ambiguity between 0.1% and 15.2% (median 2.5%) within databases, whereas ambiguity between databases ranged from 17.7% to 60.2% (median 40.3%). Standardization reduced the ambiguity to some extent. Removal of small fragments gave the largest reduction (median 1.8 percentage point) in ambiguity within databases, while removing stereochemistry information provided the best improvement in reducing ambiguity (median 13.7 percentage point) across databases. Possibly, the addition of three-dimensional information to structures either by hand or through automated processes introduces an extra complexity that is responsible for the ambiguity. These results complement our findings in a previous study, where we investigated the consistency of systematic identifiers (i.e., whether a systematic identifier was consistent with the associated MOL file) and showed that this consistency varied greatly within and across databases [13].

Ambiguity of non-systematic identifiers within databases is generally low, with on average few compounds associated with an ambiguous identifier. HMDB was an outlier with 15.2% ambiguity and an average of 6.1 compounds per ambiguous identifier. Among the most common ambiguous identifiers in HMDB are different classes of Triglyceride (TG, triacylglycerol, TAG, tracylglycerol), which is an ester derived from glycerol and three fatty acids, and Phosphatidylcholine (PC), a class of phospholipids. The IUPAC-IUB Commission on Biochemical Nomenclature discourages the use of “triglyceride” as the ambiguity of this identifier will result in inconsistencies [37]. Chemical compound records representing drugs, metabolites, and biochemicals of other types are usually records with a higher number of non-systematic identifiers, which might lead to a higher ambiguity. However, our results suggest that there is no clear association between number of non-systematic identifiers per compound and ambiguity within the different databases. Drugbank, for example, has a fairly large average number of identifiers per compound (7.1) but showed lowest ambiguity (0.1%), whereas ChEMBL has a low number of identifiers per compound (1.3) but relatively high ambiguity (3.9%).

Another reason for ambiguity is that many databases massively integrate information from other databases, but may use different standardization procedures. This can result in different compound structures that have the same, but now ambiguous, non-systematic identifiers.

The ambiguity within databases is much lower than the ambiguity across databases, which varies between 17.7% (for PubChem and ChemSpider) and 60.2% (ChemSpider and NPC). Factors that may affect the ambiguity between databases are the ambiguity within the separate databases, the level of (manual) database curation, and standardization procedures. The ambiguity between databases that could be attributed to identifiers that are already ambiguous within one or both of the databases, varied between 2.0% (DrugBank and NPC) and 39.4% (DrugBank and PubChem), but generally was considerably lower than the overall ambiguity between databases. This suggests that reducing the ambiguity within databases will only partly resolve the ambiguity across databases. It should also be noted that the ambiguity between two databases is based on the number of identifiers that the databases share, which may be much lower than the number of identifiers in either database. This explains why the ambiguity between databases for identifiers that are already ambiguous in one of the databases can be much higher than the ambiguity within databases. For example, the ambiguity between DrugBank and PubChem is 39.8%, whereas it is only 0.1% within DrugBank and 1.4% within PubChem. This shows that identifiers that are ambiguous within these databases are relatively frequently shared between the databases.

Database curation does not appear to affect the level of ambiguity of shared non-systematic identifiers between databases. For instance, DrugBank and ChemSpider-V, which are both considered highly curated databases [20, 38], show that 45.2% of the shared identifiers are ambiguous (while only 6.7% of the ambiguity between these databases could be attributed to identifiers that were already ambiguous in the separate databases). This ambiguity ranks among the highest ambiguities between databases.

The effect of chemical structure standardization on reducing the ambiguity of non-systematic identifiers is limited. The largest reductions were seen for disregarding stereochemistry and small fragments (median ambiguity reduction of 13.7 and 4.9 percentage points, respectively), but the remaining ambiguity was still considerable. The other standardization settings that we tested hardly reduced the ambiguity.

Our study may have several implications for database curation and integration efforts. First, our findings indicate that some non-systematic identifiers are very ambiguous within databases (e.g., TG, triacylglycerol, ester). These identifiers are more likely to represent classes of chemicals than individual compounds, and may be considered for removal from the databases.

Second, our study suggests that efforts to disambiguate non-systematic identifiers should not only pay attention to ambiguity within databases, which is generally low, but also consider identifiers that are ambiguous across databases. This will reveal many ambiguous and potential problematic identifiers that will not be apparent if only single databases are considered. Our method to detect these ambiguous identifiers can provide helpful information to database curators to direct their disambiguation efforts. Crowdsourcing approaches that involve the chemical community to improve database quality [20, 29, 39], may also benefit from this information to resolve ambiguity issues. All ambiguous identifiers in this study, within and between databases, are available through www.biosemantics.org.

Third, our findings are relevant for database integration and maintenance. Many chemical databases are increasing their coverage by regularly integrating data from other sources [40], or existing databases are merged and made available as a new resource [41]. As mentioned in our previous study [13], integration of databases should focus on a unique representation of compounds (e.g., MOL files) as their base of integration. InChI strings derived from the MOL files have been shown to facilitate the process as they are unique and can encode multiple types of information [42], although limitations also exist [43]. Ambiguity of systematic identifiers can be reduced by regenerating them from the structures [13], but such an approach is not possible for non-systematic identifiers, which are generated at the point of registration. Our results show that there is a large ambiguity of non-systematic identifiers across databases, and suggest that the integration of these identifiers from different databases without proper manual curation can greatly increase their ambiguity. It has previously been proposed to use a voting approach to disambiguate non-systematic identifiers when integrating multiple databases, assigning the identifier to the compound to which it was most frequently associated in the databases [29], but this approach may be biased by error propagation when one database includes an erroneous identifier from another database.

Our study has several limitations. First, although we included a variety of commonly used chemical databases, the number of databases is not very large and our results may not apply to databases that were not considered. Moreover, as the content of the databases evolves over time, the ambiguity within and between databases is likely to have changed since we downloaded the data. For example, recently an effort has been made to reduce ambiguity within the ChemSpider database by using a subset of records with non-systematic identifiers that had manually been validated, and automatically removing these identifiers from any record that had not been validated. A second limitation is that we quantified the ambiguity of non-systematic identifiers within and across databases, but did not determine which of the associations between non-systematic identifiers and compounds were correct, and thus could not rank the databases on their performance in this respect. A reference set of correctly assigned non-systematic identifiers would allow such an analysis, but may be cumbersome to establish. Finally, to assess whether two structures were the same, we used one tool to convert MOL files into InChI strings. Other tools might occasionally produce different conversions, because of differences in MOL file processing, but in our previous study [13] such differences were negligible and did not significantly influence the results.

Conclusions

Ambiguity of non-systematic identifiers within chemical databases is generally low. A much higher ambiguity was observed for non-systematic identifiers that are shared across databases. Chemical structure standardization reduces the ambiguity to a limited extent. The largest reductions are obtained when disregarding stereochemistry information or removing small fragments. The results of our study can help to improve database integration, curation and maintenance.

References

1. Williams AJ: **Public chemical compound databases.** *Curr Opin Drug Discov Devel* 2008, **11:**393-404.
2. Muresan S, Sitzmann M, Southan C: **Mapping between databases of compounds and protein targets.** *Methods Mol Biol* 2012, **910:**145-164.
3. Cumming JG, Davis AM, Muresan S, Haeberlein M, Chen H: **Chemical predictive modelling to improve compound quality.** *Nat Rev Drug Discov* 2013, **12:**948-962.
4. Liaw A, Svetnik V: **QSAR modeling: prediction of biological activity from chemical structure.** *Statistical Methods for Evaluating Safety in Medical Product Development* 2015:66-83.
5. Eltyeb S, Salim N: **Chemical named entities recognition: a review on approaches and applications.** *J Cheminform* 2014, **6:**1-12.
6. Vazquez M, Krallinger M, Leitner F, Valencia A: **Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications.** *Molecular Informatics* 2011, **30:**506-519.
7. Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA: **Annotated Chemical Patent Corpus: A Gold Standard for Text Mining.** *PloS one* 2014, **9:**e107477.
8. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, Valencia A: **CHEMDNER: The drugs and chemical names extraction challenge.** *J Cheminform* 2015:S1.
9. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, et al: **HMDB: the Human Metabolome Database.** *Nucleic Acids Res* 2007, **35:**D521-D526.
10. Alex B, Grover C, Haddow B, Kabadjor M, Klein E, Matthews M, Roebuck S, Tobin R, Wang X: **Assisted Curation: Does Text Mining Really Help?** In *Pacific Symposium on Biocomputing*. Citeseer; 2008, **13:**556-567.
11. 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:**1189-1204.
12. Southan C, Sitzmann M, Muresan S: **Comparing the chemical structure and protein content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database.** *Molecular Informatics* 2013, **32:**881-897.
13. Akhondi SA, Kors JA, Muresan S: **Consistency of systematic chemical identifiers within and between small-molecule databases.** *J Cheminform* 2012, **4:**35.
14. *About IUPAC.* http://www.iupac.org/home/about.html.
15. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28:**31–36.
16. *History of InChI.* http://www.inchi-trust.org/index.php?q=node/2.
17. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I: **InChI - the worldwide chemical structure identifier standard.** *J Cheminform* 2013, **5:**7.
18. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical Entities of Biological Interest: an update.** *Nucleic Acids Res* 2010, **38:**D249-D254.
19. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res* 2012, **40:**D1100-D1107.

20. Pence HE, Williams AJ: **ChemSpider: An Online Chemical Information Resource.** *J Chem Educ* 2010, **87:**1123-1124.

21. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39:**D1035-D1041.

22. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, et al: **HMDB: a knowledgebase for the human metabolome.** *Nucleic Acids Res* 2009, **37:**D603-D610.

23. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP: **The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics.** *Sci Transl Med* 2011, **3:**80ps16.

24. Bolton EE, Wang Y, Thiessen PA, Bryant SH: **PubChem: integrated platform of small molecules and biological activities.** *Annual reports in computational chemistry* 2008, **4:**217-241.

25. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C, Chen Y: **Update of TTD: Therapeutic Target Database.** *Nucleic Acids Res* 2010, **38:**D787-D791.

26. Chen X, Ji ZL, Chen YZ: **TTD: therapeutic target database.** *Nucleic Acids Res* 2002, **30:**412-415.

27. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, J L: **Description of several chemical structure file formats used by computer programs developed at molecular design limited.** *J Chem Inf Comput Sci* 1992**:**244-255.

28. *Royal Society of CHEMISTRY.* http://www.rsc.org/.

29. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, Varkonyi P, Xie PH: **Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data.** *Drug Discov Today* 2011, **16:**1019-1030.

30. Williams AJ, Ekins S, Tkachenko V: **Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation.** *Drug Discov Today* 2012, **17:**685-701.

31. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, 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-D1097.

32. *PubChem SD file formatted data, V2.0.1.* ftp://ftp.ncbi.nlm.nih.gov/pubchem/.
    a. data_spec/pubchem_sdtags.pdf.

33. *ChemAxon, Naming.* http://www.chemaxon.com/products/naming/.

34. Lowe DM, Corbett PT, Murray-Rust P, Glen RC: **Chemical name to structure: OPSIN, an open source solution.** *J Chem Inf Model* 2011, **51:**739-753.

35. Sitzmann M, Filippov IV, Nicklaus MC: **Internet resources integrating many small-molecule databases.** *SAR QSAR Environ Res* 2008, **19:**1-9.

36. *Standardizer - Structure canonicalization and more.* http://www.chemaxon.com/products/standardizer/.

37. *Nomenclature of Lipids, IUPAC-IUB Commission on Biochemical Nomenclature (CBN).* http://www.chem.qmul.ac.uk/iupac/lipid/.

38. Wishart DS, Knox C, Guo AC, 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.

39. Williams AJ: **Facilitating scientific discovery through crowdsourcing and distributed participation.** *EMBnet journal* 2013, **19:**12.

40. Williams AJ: **A perspective of publicly accessible/open-access chemistry databases.** *Drug Discov Today* 2008, **13:**495-501.

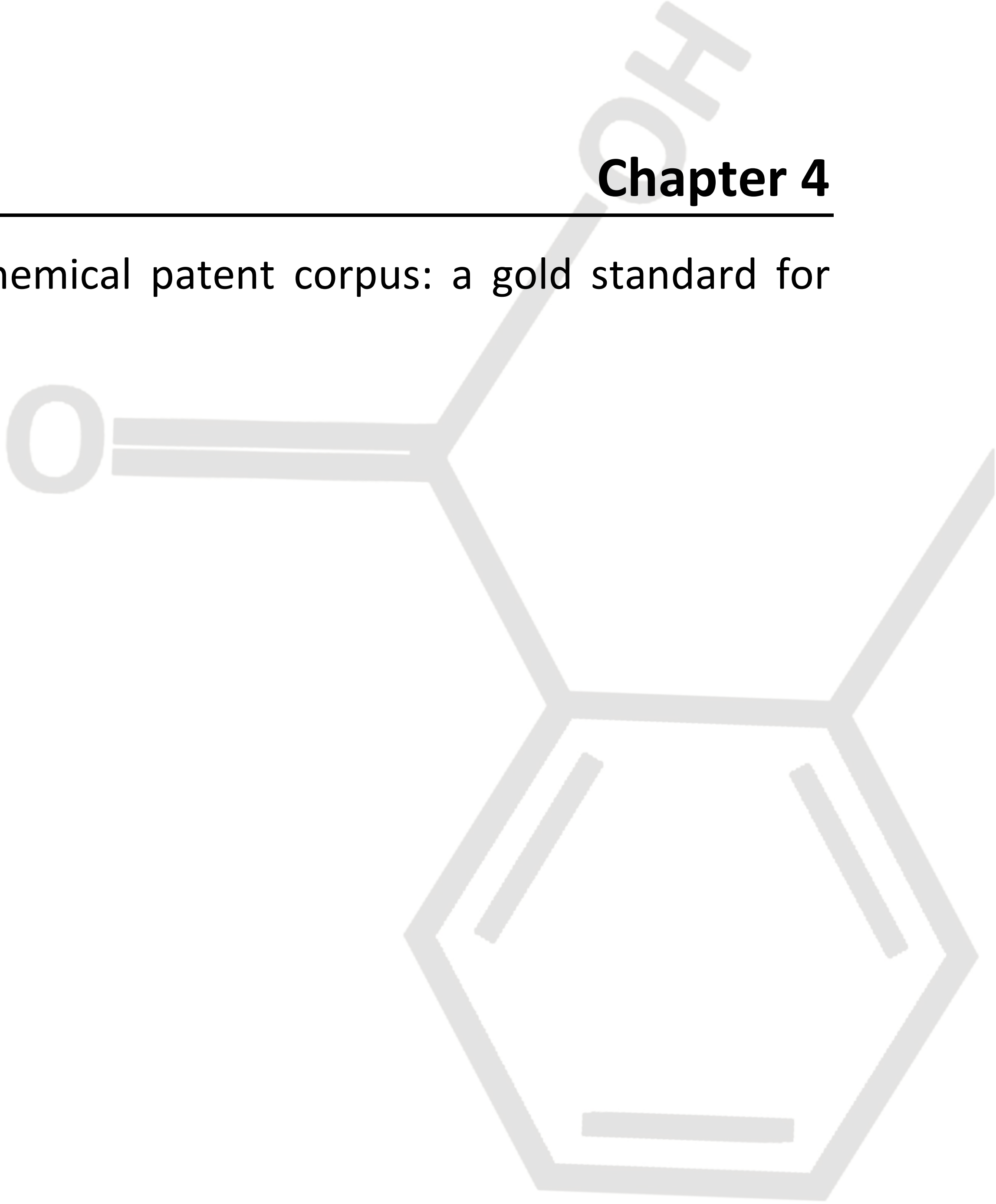
41. Samwald M, Jentzsch A, Bouton C, Kallesoe CS, Willighagen E, Hajagos J, Marshall MS, Prud'hommeaux E, Hassenzadeh O, Pichler E, Stephens S: **Linked open drug data for pharmaceutical research and development.** *J Cheminform* 2011, **3:**19.

42. Haraldsdottir HS, Thiele I, Fleming RM: **Comparative evaluation of open source software for mapping between metabolite identifiers in metabolic network reconstructions: application to Recon 2.** *J Cheminform* 2014, **6:**2.

43. Galgonek J, Vondra Ek JI: **On InChI and evaluating the quality of cross-reference links.** *J Cheminform* 2014, **6:**15.

Chapter 4

Annotated chemical patent corpus: a gold standard for text-mining



*Published: **Akhondi SA**, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA, Muresan S*

PloS one 2014, 9:e107477

Abstract

Exploring the chemical and biological space covered by patent applications is crucial in early-stage medicinal chemistry activities. Patent analysis can provide understanding of compound prior art, novelty checking, validation of biological assays, and identification of new starting points for chemical exploration.

Extracting chemical and biological entities from patents through manual extraction by expert curators can take substantial amount of time and resources. Text mining methods can help to ease this process. To validate the performance of such methods, a manually annotated patent corpus is essential.

In this study we have produced a large gold standard chemical patent corpus. We developed annotation guidelines and selected 200 full patents from the World Intellectual Property Organization, United States Patent and Trademark Office, and European Patent Office. The patents were pre-annotated automatically and made available to four independent annotator groups each consisting of two to ten annotators. The annotators marked chemicals in different subclasses, diseases, targets, and modes of action. Spelling mistakes and spurious line break due to optical character recognition errors were also annotated. A subset of 47 patents was annotated by at least three annotator groups, from which harmonized annotations and inter-annotator agreement scores were derived. One group annotated the full set.

The patent corpus includes 400,125 annotations for the full set and 36,537 annotations for the harmonized set. All patents and annotated entities are publicly available at www.biosemantics.org.

Background

A substantial number of patent applications are filed every year by the pharmaceutical sector [1]. Exploring the chemical and biological space covered by these patents is crucial in early-stage medicinal chemistry activities [1,2]. Patent specifications are one of many information sources needed to progress drug discovery projects. Patent analysis can provide understanding of compound prior art, novelty checking, validation of biological assays, and identification of new starting points for chemical exploration [3].

Extracting chemical and biological entities from patents is a complex task [4,5]. Different approaches are currently used including manual extraction by expert curators, text mining supported by chemical and biological named entity recognition, or combinations thereof [6]. Chemical patents are complex legal documents that can contain up to hundreds of pages. The European Patent Office (EPO) [7], the pharmaceutically relevant patents within the United States Patent and Trademark Office (USPTO) [8], and the World Intellectual Property Organization (WIPO) [9] can be accessed and queried on-line via their websites. The patents are freely available from the patent offices, usually as XML, HTML or image PDFs, although EPO limits the number of downloads per week for non-paying users. Using optical character recognition (OCR), the image PDFs can be prepared for text mining. In fact, the available HTML and XML documents are mainly the OCR output prepared and published by the patent offices.

However, the text mining itself is a rather challenging task [10,11]. Methods and their output can suffer dramatically from the large number of complex chemical names, term ambiguities, complex syntactic structures and OCR errors [12].

To validate the performance of named entity recognition techniques, the availability of a manually annotated patent corpus is essential [13]. Producing such annotated text is laborious and expensive. Most of the prior focus on corpora development has been on genes and proteins and less effort has been put into creating corpora for chemical terms [14]. Among the latter efforts, Kim et al. [15] in 2003 developed the GENIA corpus consisting of several classes of chemicals. The BioIE corpus by Kulick et al. [16] was made available in 2004 and included annotations of chemicals and proteins. In 2008, Kolárik et al. [17] released a small corpus of scientific abstracts annotated with chemical compounds. Recently, the CHEMDNER corpus, annotated with different classes of chemicals, was made available as part of the BioCreative challenge [18]. All these corpora consist of scientific abstracts from Medline. In a collaborative project between the EPO and the Chemical Entities of Biological Interest (ChEBI) in 2009 [19] a chemical patent corpus containing annotations of chemical entities and, if possible, their mapping to ChEBI chemical compounds [20] was developed. In a later study [21], the updated version of ChEBI [22] was used to increase the number of mappings. A larger patent corpus was developed in 2012 by Kiss et al. [12] which included name entity recognition of generic chemical compounds.

To our knowledge, the development of a gold standard patent corpus has not been systematically tackled before. Among the obvious reasons for this are the length and complexity of the patent text. In previous attempts only limited number of chemicals have been annotated and subclasses have not been defined. Other biological entities such as diseases or modes of actions have not been included and errors due to misspellings or OCR procedures have not been considered. Most previous studies on annotated corpora did not provide insights into inter-annotator agreement. This information would be valuable in assessing and comparing the performance of text mining applications.

Here we present a gold standard annotated corpus of 200 full patents for benchmarking text mining performance. The patent corpus includes annotation of chemicals with subclasses, diseases, targets and modes of action. Also spelling mistakes and spurious line break due to OCR errors are annotated within this corpus. The full-text patents and annotated entities are publicly available at www.biosemantics.org.

Methods and Materials

Corpus development strategy

The development of the gold standard patent corpus consisted of several phases. First, annotation guidelines were developed and a set of 200 diverse patents was chosen. The patents were pre-annotated automatically and made available to four independent annotator groups. The annotator groups could choose to consider or disregard the pre-annotations. Two patents were used to refine the annotation guidelines. The remaining patents were distributed between multiple annotator groups in a way that a subset of 47 patents was annotated by at least three groups, from which harmonized annotations were derived. Inter-annotator agreement scores between the annotator groups and against the harmonized set were computed. One annotator group annotated the complete set of patents.

Patent corpus selection

The GVK BIO target class database [23] was used as a starting point for patent corpus selection. Patents from the EPO [7], USPTO [8], and WIPO [9] are available through this database, which includes relationships between documents, assays, chemical structures, assignees and protein targets, manually abstracted by expert curators [1]. Within the database, patents are binned based on different classes of protein families such as kinases or GPCRs [23].

All English language patents containing between 10 and 200 exemplified compounds, with a named primary target, were selected from the GVK BIO database. We made sure that all compounds had a molecular weight below 1000 to bias towards small-molecule patents. We did not specify limits on the time of the application. Overall 28,695 patents fulfilled the above criteria.

Chemical patents are known to include long sentences with complex syntactic structure [12]. Individual companies may have different ways of writing patents and we wanted to include diversity over assignees in the corpora. Therefore, if assignees had written multiple patents for one primary target, only one was randomly kept and the rest was disregarded.

Based on these selection criteria we were left with 8,016 patents grouped in 11 target classes. To make sure that a collection of well-known patents are included in the corpus, 50 drug patents from Sayle et al. [24] were added. Subsequently patents were randomly picked from each target group with a minimum of 10 patents per group. The diversity of the final selection is shown in Table 1. The final set consists of 121 USPTO, 66 WIPO, and 13 EPO patents, and contains over 11,500 pages and 4.2 million words.

Table 1: Target class distribution of the 8,066 patents from which the final set was drawn.

Target class	**Number of patents**	**Final selection**
GPCR	3,569	20
Protease	1,093	17
Kinase	1,046	12
Ion-Channel	433	14
Oxidoreductase	404	17
Hydrolase	364	15
NHR	349	15
Transporters	323	18
Other	218	11
Transferase	152	12
Phosphatase	65	17
Drugs from Sayle et al. [24]	50	32
Total	8,066	200

The patents were downloaded from the sources (EPO, USPTO, and WIPO) in XML format. Whenever multiple consecutive line breaks were encountered, they were replaced with a single line break. Images were also removed for all patents.

Annotated entities

We annotated all compounds, diseases, protein targets, and modes of actions (MOA) mentioned in the patents. Compounds were assigned to a number of subclasses based on how they are generated: systematic identifiers and non-systematic identifiers [25]. The following systematic identifiers were annotated: IUPAC names [26], such as "ammonium phosphate" or "2-[2-(4-{2-[ethyl(2-fluorobenzyl)amino]-2-oxoethoxy}phenyl)ethoxy]benzoic acid"; SMILES notations [27], such as "n1c[nH]cc1"; and InChI strings [28,29], such as "InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3". We also annotated the following non-systematic identifiers: trademarks, such as "Aspirin", "Mesupron", and "Arimidex"; abbreviations, such as "DCM", "TBTU" and "DMAP"; CAS numbers [30,31], such as "7732-18-5"; formulas, such as "MgSO4"; registry numbers, such as "ly256548"; and generic names, such as "iodotamoxifen", "cycloalkylamines" and "racemate". Any mention of diseases, such as "diabetes", protein targets, such as "trypsin", and MOAs, such as "antagonist", were also annotated. OCR errors were also annotated in terms of spelling mistakes and spurious line breaks.

Annotation guidelines

Initial annotation guidelines were developed based on previous work [14,16-18]. Two patents (US5023269 [32] and US4659716 [33]) were randomly chosen from the patent corpus for training the annotators and fine-tuning the annotation guidelines. The following rules were defined:

1. When an entity is nested or has an overlap with another entity, annotate the entity that is more specific and informative. For example "5-HT1D" should not be annotated as target when it is embedded within the target annotation of "5-HT1D Serotonin Receptors".

2. Annotate simple IUPAC names such as "water, "ammonia", and "ethanol".

3. Prefixes should be included within annotations, for example "1,4-" in "1,4-butanediol".

4. Simple formulas such as "NaOH" and "(NH4)2SO4", should be annotated as Formulas.

5. Counterions, such as "acetate", "oxalate", "propionate", should be annotated as IUPAC names.

6. Generic structures such as "4-halo-phenol" or "xylene", should be annotated as Generic names.

7. Polymers, e.g., "Polystyrene", should be annotated as Generic names.

8. Trivial names, e.g., "Sildenafil", should be annotated as IUPAC names.

9. Enumerations, like "hydrochloric" and "hydrobromic" in "include inorganic acids such as hydrochloric, hydrobromic", should be annotated as IUPAC names.

10. Elements like "N", "O", and "C" should not be annotated.

11. Misspelled terms should be annotated as spelling mistakes (e.g., "hydrobroml:c").

12. Annotations spanning over multiple lines because of spurious line breaks should be annotated as one term and be tagged with spurious line breaks.

13. Extra white space should be annotated as spelling mistakes (e.g., "hydro bromic").

14. Do not annotate a term if it is splitted due to reasons other than OCR errors.

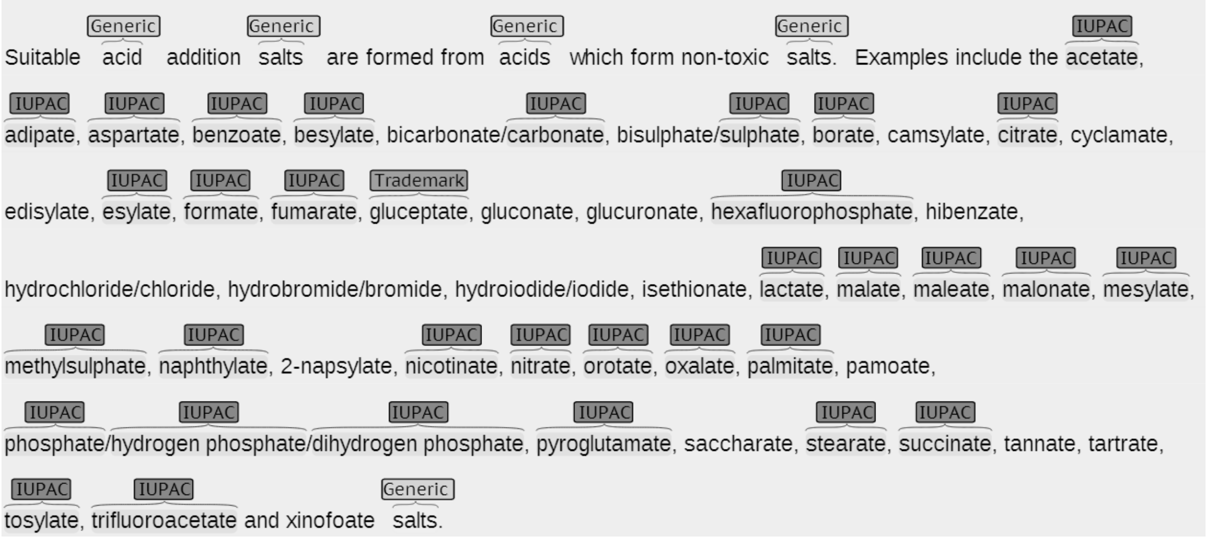
15. All symbols such as comma, charge symbol or brackets, should be included in the annotation (e.g., “n1c[nH]cc1”).

Annotation process

Each patent was automatically pre-annotated using LeadMine (NextMove Software, UK) [34]. LeadMine can identify chemicals, protein targets, genes, species, company names, and also has the ability to recognise terms with spelling mistakes and suggest corrections. This increases the likelihood of detecting terms with OCR errors by the human annotator.

A pre-annotation consists of the span of text corresponding with the entity and its location within the text file. The following entity types were pre-annotated by LeadMine: IUPAC names, trivial names, CAS numbers, registry numbers, generic names, formulas, and targets. We did not pre-annotate SMILES and InChIs, as they are rarely present in patents. Diseases, and MOAs were also not included as this was not possible through our version of LeadMine.

For the annotation process the Brat rapid annotation tool (version 1.3) was used [35]. Brat allows online annotation of text using pre-defined entity types. It can display the pre-annotations and annotators can add new annotations and modify or delete the pre-annotated entities. To reduce mistakes and increase readability each entity type was marked by a specific color. For performance reasons we split the patents into pages with 50 paragraphs for display in Brat. Figure 1 shows a screenshot of Brat with pre-annotations.


Generic Generic Generic Generic IUPAC
IUPAC IUPAC IUPAC IUPAC IUPAC IUPAC IUPAC IUPAC
IUPAC IUPAC IUPAC Trademark IUPAC
IUPAC IUPAC IUPAC IUPAC IUPAC
IUPAC IUPAC IUPAC IUPAC IUPAC IUPAC IUPAC
IUPAC IUPAC IUPAC IUPAC IUPAC IUPAC
IUPAC IUPAC Generic

Suitable acid addition salts are formed from acids which form non-toxic salts. Examples include the acetate, adipate, aspartate, benzoate, besylate, bicarbonate/carbonate, bisulphate/sulphate, borate, camsylate, citrate, cyclamate, edisylate, esylate, formate, fumarate, gluceptate, gluconate, glucuronate, hexafluorophosphate, hibenzate, hydrochloride/chloride, hydrobromide/bromide, hydroiodide/iodide, isethionate, lactate, malate, maleate, malonate, mesylate, methylsulphate, naphthylate, 2-napsylate, nicotinate, nitrate, orotate, oxalate, palmitate, pamoate, phosphate/hydrogen phosphate/dihydrogen phosphate, pyroglutamate, saccharate, stearate, succinate, tannate, tartrate, tosylate, trifluoroacetate and xinofoate salts.

Figure 1: Example patent text with pre-annotations as shown by the Brat annotation tool.

Patents were annotated by annotators from four groups: AstraZeneca, Fraunhofer, GVK BIO, and NextMove. The GVK BIO annotation group consisted of ten annotators, while the other annotator groups had two annotators. One annotator group (Fraunhofer) chose to disregard the pre-annotations made by LeadMine. The patents were distributed between annotators within a group, such that each patent was annotated by only one annotator in a group. In the context of this paper, annotator group will refer to any individual annotator within the group.

Annotators had to correct any misidentified pre-annotation and had to add annotations that were missed in the pre-annotation step. Entities containing misspellings or spurious line breaks were separately annotated.

Resolving misannotation of ambiguous terms

After the completion of the annotations by all groups, a group of annotators reviewed the results to reduce the number of ambiguous terms within the corpus. A term is defined as ambiguous if different groups annotated it with different entity types throughout the corpus.

A list of ambiguously annotated terms was compiled and annotators were asked to review the list only based on the different entity types assigned to each ambiguous term (i.e., the context of the terms was not provided). The annotators had to classify each term in one of three groups:

1. None of the entity types assigned to the term is applicable. All annotations of the term were removed from the corpus. For example, “nitrogen” was annotated as both IUPAC and Generic multiple times throughout the corpus. However, either entity type is incorrect since the term is an element. Therefore annotations of nitrogen are removed from the corpus.

2. One entity type is applicable. All occurrences of the term within the corpus were assigned to this entity type. For example, the term “DMF” was assigned 43 times as Trademark, 289 times as Abbreviation, and once as Formula. Regardless of the context of the text, DMF is an abbreviation and therefore the entity type of the term was changed to Abbreviation throughout the corpus.

3. More than one entity type is applicable. Only term annotations with an entity type that is not applicable, were removed throughout the corpus. For example, the term “5-ht” has been annotated 17 times as Abbreviation, 25 times as Generic, and 23 times as Target. Depending on the context of the text, the term can be either Target or Abbreviation but not Generic. Therefore all annotations of the term as Generic were removed from the corpus.

Harmonization

To develop the gold standard corpus, the annotations of the 47 patents annotated by more than three groups were merged into a harmonized set. The centroid algorithm described by Lewin et al. [36] was used for this purpose.

Briefly, the algorithm tokenizes the annotations of different annotators at the character level and counts the number of agreeing annotators over pairs of adjacent annotation-internal characters [36]. Calculating votes over annotation-internal character pairs and not individual characters, guarantees that boundaries (starting and ending positions of an annotated entity) are considered in situations where two terms are annotated directly adjacent to each other [36]. The harmonized annotation consists of the characters pairs that have a vote equal to or larger than a specified threshold. In this work, we used a voting threshold of two, i.e., at least two annotators had to agree on the annotation.

The centroid algorithm was executed separately for each entity type. Therefore votes were only calculated if at least two annotators annotated a term with the same entity type.

Inter-annotator agreement

Similar to Corbett et al. [14] and Kolárik et al. [17], we used the F-score (harmonic mean of recall and precision) to calculate the inter-annotator agreement between the annotator groups and between each annotator group and the harmonized set. For the comparison of two sets of annotations, one set was arbitrarily chosen as the gold standard (this choice does not affect the F-score). An annotation in the other set was counted as true positive if it was identical to the gold standard annotation, i.e., if both annotations had the same entity type and the same start and end location. If a gold standard annotation was not given, or not rendered exactly in the other set (i.e., non-matching boundaries or a different entity type), it was counted as false negative; if an annotation found in the other set did not exactly match the gold standard, it was counted as false positive.

Results

Patent distribution among groups

The number of annotated patents varied between annotation groups. Apart from the two patents used for training, 27 patents were annotated by NextMove, 36 by Fraunhofer, 49 by AstraZeneca, and 198 by GVK BIO. A total of 47 patents were annotated by at least three of the groups (three patents were annotated by all four groups).

Initial harmonized set

The initial harmonized set, prior to disambiguation, was generated over the 47 common patents, yielding a total of 35,337 annotations (Table 2). The results show that IUPAC names and generic names have been annotated significantly more than any other chemical type, as has also been shown previously [13]. On the other hand, InChIs, CAS registry numbers and SMILES are rarely seen in these chemical patents. Also, a considerable number of diseases, targets, and MOAs have been annotated.

Table 2: Number of annotated terms and unique terms within the harmonized set prior to disambiguation.

Entity type	Annotated terms	Unique terms
IUPAC	14,423	5,365
Generic	7,959	880
Disease	3,777	1,257
Target	3,227	705
Trademark	2,273	987
Abbreviation	1,460	153
Formula	1,069	171
MOA	1,014	211
Registry Number	108	90
SMILES	21	21
CAS	6	5
InChI	0	0
Total	35,337	9,845

Inter-annotator agreement prior to disambiguation

Table 3 shows the inter-annotator agreement between the groups and the harmonized set prior to disambiguation. There is generally a higher inter-annotator agreement between individual annotator groups and the harmonized set than between pairs of groups. The best agreement was 0.78. The agreement between groups ranged between 0.39 and 0.69. Investigation of the reasons for some low agreements suggested that adding a disambiguation step could resolve some of these disagreements.

Table 3: Inter-annotator agreement (F-score) without ambiguity resolution.

	AstraZeneca	**Fraunhofer**	**GVK BIO**	**NextMove**
Fraunhofer	0.42			
GVK BIO	0.60	0.39		
NextMove	0.50	0.69	0.52	
Harmonized	0.78	0.64	0.74	0.72

Disambiguation

A set of 2,135 unique ambiguous terms, corresponding to 47,044 annotations, were provided to annotators for disambiguation as described above. The annotators were able to make a decision for 333 unique ambiguous terms, affecting 9,005 annotations. The results in Table 4 show that most difficulties within the annotations were encountered between IUPAC names, Generic names and Trademarks. Also 23 elements were found that had been annotated 2,499 times with different entity types throughout the corpus. Since elements should not be annotated according to the guidelines, these terms were removed from the corpus.

Table 4: The effect of the disambiguation process on the annotations.

Rules	Type	Affected Terms	Affected Annotations
Add annotation	IUPAC	52	2,275
	Abbreviation	29	1,631
	Generic	67	976
	Trademark	71	442
	Disease	4	387
	MOA	2	203
	Formula	25	177
	Registry Number	28	111
	Target	19	32
Remove annotation	Elements	23	2,499
	IUPAC	7	103
	Trademark	3	101
	Generic	2	67
	Target	1	1
Total		333	9,005

Inter-annotator agreement after disambiguation

After resolving the ambiguous terms, the harmonized set was recalculated. This resulted in an increase of inter-annotator agreement scores by 0.01 to 0.09 points (Table 5).

Table 5: Inter-annotator agreement after ambiguity resolution. The lower left triangle presents the inter-annotator agreement scores (F-score). The upper right triangle shows the improvement gained through disambiguation.

	AstraZeneca	**Fraunhofer**	**GVK BIO**	**NextMove**	**Harmonized**
AstraZeneca		+ 0.04	+ 0.09	+ 0.08	+ 0.06
Fraunhofer	0.46		+ 0.05	+ 0.03	+ 0.01
GVK BIO	0.69	0.44		+ 0.06	+ 0.05
NextMove	0.58	0.72	0.58		+ 0.03
Harmonized	0.84	0.65	0.79	0.75	

Recalculating the inter-annotator agreement by only considering text boundaries and disregarding the entity types, further increases the agreement with up to 0.04 points. To analyze the reasons behind some of the low agreements, inter-annotator agreement scores were calculated for the main entity types (Table 6). The major difficulty in the annotation was encountered for non-systematic identifiers and MOAs, while identification of targets, diseases, and systematic identifiers were made with higher agreements.

Table 6: Inter-annotator agreement (F-score) between the harmonized set and the annotator groups for the main entity types.

	AstraZeneca	**Fraunhofer**	**GVK BIO**	**NextMove**
	Harmonized	**Harmonized**	**Harmonized**	**Harmonized**
Overall	0.84	0.65	0.79	0.75
Chemicals	0.89	0.65	0.78	0.75
Systematic	0.94	0.81	0.91	0.93
Non-systematic	0.85	0.38	0.68	0.56
Disease	0.47	0.82	0.87	0.86
Targets	0.76	0.57	0.81	0.86
MOA	0.65	0.29	0.67	0.17

The inter-annotator agreement between the groups and overall, chemicals and systematic names were between 0.65 and 0.94. The inter-annotator agreement for non-systematic terms between Fraunhofer and the harmonized set was only 0.38. To investigate the reasons behind this low agreement, we recalculated the inter-annotator agreement between Fraunhofer and the harmonized set by considering cases where one annotation was embedded within the other annotation as an agreement. This only increased the inter-annotator score to 0.46. Further analysis showed that counting annotations that overlap as an agreement increased the score to 0.62. The main reason for the remaining differences was that annotators at Fraunhofer did not annotate formulas and had low agreements with others within the generic terms.

Table 6 shows that apart from AstraZeneca, all groups managed to gain a high inter-annotator agreement (0.82 to 0.86) between diseases and the harmonized set. Further analysis showed that the low inter-annotator agreement between AstraZeneca and the harmonized set on diseases is due to annotation differences in the boundaries. Calculating inter-annotator agreement on diseases by also accepting embedded terms increased the agreement to 0.70.

The inter-annotator agreement between Fraunhofer and the harmonized set for targets was only 0.57. Additional investigation showed that accepting embedded terms increased the agreement to 0.64.

The annotations of MOA for Fraunhofer and NextMove were also greatly affected by how the boundaries were chosen. An example is the term “mixed agonist” for which one group annotated the whole term as MOA and the other only annotated “agonist” as MOA. Accepting such cases as an agreement increases the agreement between NextMove and the harmonized set from 0.17 to 0.72, and between Fraunhofer and the harmonized set from 0.29 to 0.62.

The gold standard patent corpus

The gold standard patent corpus consists of two sets: the harmonized corpus and the full corpus. The harmonized corpus consists of 47 patents with a total of 36,537 annotations for 9,813 unique terms (Table 7). In addition, 1,239 OCR errors have been annotated, of which 1,189 are spelling mistakes. The full patent corpus of 198 patents contains only the GVK BIO annotations with 400,125 annotations for 80,977 unique terms. The set includes 5,096 OCR error annotations, of which 4,403 are spelling mistakes.

Table 7: Number of annotated terms and unique terms in the harmonized set and in the full patent set of the gold standard corpus after disambiguation.

	Harmonized set (47 Patents)		**Full set (198 Patents)**	
	Unique terms	**Annotated terms**	**Unique terms**	**Annotated terms**
IUPAC	5,325	14,377	50,893	135,603
Generic	881	8,384	14,305	169,133
Disease	1,256	3,776	4,503	20,229
Target	703	3,235	3,514	14,398
Trademark	994	2,366	3,365	9,574
Abbreviation	153	2,088	778	21,087
Formula	169	1,127	3,108	25,716
MOA	210	1,017	110	3,837
Registry Number	96	140	188	329
SMILES	21	21	166	166
CAS	5	6	47	53
InChI	0	0	0	0
Total	9,813	36,537	80,977	400,125

Discussion and conclusion

We have produced a gold standard chemical patent corpus consisting of 198 full patents of which 47 patents have been annotated by at least three annotators. The patent corpus contains a selection of patents from WIPO, USPTO and EPO with annotation of compounds, diseases, targets, and MOAs. We have also annotated spelling errors for the mentioned entity types.

We have released the inter-annotator agreements along with the gold standard corpus. Making inter-annotator agreement scores available will hopefully prove to be useful for performance assessment of automatic annotations of the patent corpus.

To our knowledge this is the first patent gold standard corpus containing full patents with different entity types (chemicals and their sub entities, diseases, MOAs, and targets). Patents are one of the richest knowledge sources with high information content and detailed description of chemistry and technology. Our annotation process showed the complexity of the annotation task. The OCR process added a significant level of noise to the text. A high inter-annotator agreement was seen on the annotation of entities such as systematic names. In contrast, we observed lower inter-annotator agreements for non-systematic names and MOAs. This emphasizes the challenges in identifying named entities from patent text. Annotation of OCR errors may also be helpful to improve patent informatics systems by facilitating the development of algorithms to correct such errors.

The annotated gold standard corpus should prove a valuable resource for developing and evaluating patent text analytics approaches.

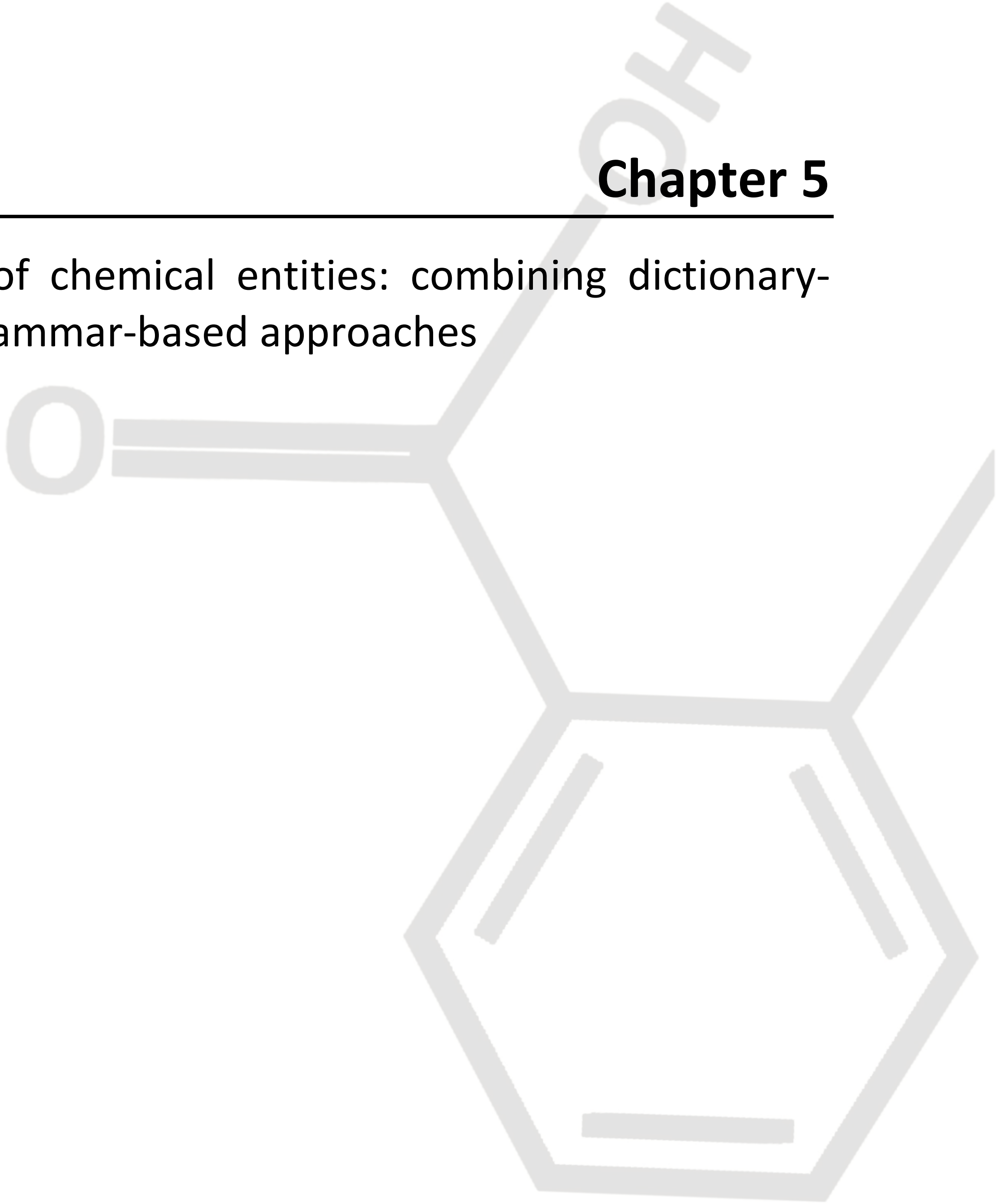
References

1. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, et al.: **Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data.** *Drug Discov Today* 2011, **16:** 1019-1030.
2. Southan C, Boppana K, Jagarlapudi SA, Muresan S: **Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds.** *J Cheminform* 2011, **3:** 14.
3. Tyrchan C, Boström J, Giordanetto F, Winter J, Muresan S: **Exploiting Structural Information in Patent Specifications for Key Compound Prediction.** *J Chem Inf Model* 2012, **52**: 1480-1489.
4. Kolarik C, Hofmann-Apitius M, Zimmermann M, Fluck J: **Identification of new drug classification terms in textual resources.** *Bioinformatics* 2007, **23:** i264-272.
5. Klinger R, Kolarik C, Fluck J, Hofmann-Apitius M, Friedrich CM: **Detection of IUPAC and IUPAC-like chemical names.** *Bioinformatics* 2008, **24:** i268-276.
6. Zimmermann M, Fluck J, Thi LT, Kolarik C, Kumpf K, et al.: **Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology.** *Curr Top Med Chem* 2005, **5:** 785-796.
7. *European Patent Office.* http://www.epo.org/.
8. *United States Patent and Trademark Office.* http://www.uspto.gov/.
9. *World Intellectual Property Organization.* http://www.wipo.int/.
10. Tseng Y-H, Lin C-J, Lin Y-I: **Text mining techniques for patent analysis.** *Inf Process Manag* 2007, **43:** 1216-1247.
11. Jessop DM, Adams SE, Murray-Rust P: **Mining chemical information from open patents.** *J Cheminform* 2011, **3:** 40.
12. Kiss M, Nagy Á, Vincze V, Almási A, Alexin Z, et al.: **A Manually Annotated Corpus of Pharmaceutical Patents. Text, Speech and Dialogue.** *Springer Berlin Heidelberg* 2012, pp. 135–142.
13. Vazquez M, Krallinger M, Leitner F, Valencia A: **Text mining for drugs and chemical compounds: methods, tools and applications.** *Mol Inform* 2011, **30:** 506-519.
14. 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. Morristown, NJ, USA*: Association for Computational Linguistics 2007. pp. 57-64.
15. Kim JD, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus--semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19 Suppl 1:** i180-182.
16. Kulick S, Bies A, Liberman M, Mandel M, McDonald R, et al. **Integrated annotation for biomedical information extraction;** 2004. *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* pp. 61-68.
17. Kolárik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J. **Chemical names: terminological resources and corpora annotation**; 2008. *Workshop on Building and evaluating resources for biomedical text mining*.
18. Krallinger M, Leitner F, Rabal O, Vazquez M, Oyarzabal J, et al. **Overview of the chemical compound and drug name recognition (CHEMDNER) task**; 2013. *BioCreative Challenge Evaluation Workshop*. vol. 2. pp. 2.

19. Grego T, Pęzik P, Couto FM, Rebholz-Schuhmann D: **Identification of chemical entities in patent documents. Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living**: *Springer* 2009, pp. 942-949.
20. Degtyarenko K, De Matos P, Ennis M, Hastings J, Zbinden M, et al.: **ChEBI: a database and ontology for chemical entities of biological interest.** ***Nucleic Acids Res*** 2008, **36**: D344-D350.
21. Tiago G, Catia P, Bastos Hugo P: **Chemical entity recognition and resolution to ChEBI.** *ISRN Bioinformatics* 2012.
22. De Matos P, Alcántara R, Dekker A, Ennis M, Hastings J, et al.: **Chemical entities of biological interest: an update.** ***Nucleic Acids Res*** 2010, **38:** D249-D254.
23. *GVK BIO Target Class Based Compound Database.* http://www.gvkbio.com/products-services/informatics-analytics/products/standalone-databases/.
24. Sayle R, Xie PH, Muresan S: **Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction.** *J Chem Inf Model* 2012, **52:** 51-62.
25. Akhondi SA, Kors JA, Muresan S: **Consistency of systematic chemical identifiers within and between small-molecule databases.** *J Cheminf* 2012, **4:** 35.
26. *About IUPAC.* http://www.iupac.org/home/about.html/.
27. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28:** 31–36.
28. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I: **InChI - the worldwide chemical structure identifier standard.** *J Cheminform* 2013, **5:** 7.
29. *About the InChI Standard.* http://www.inchi-trust.org/about-the-inchi-standard/.
30. Undefined **CAS Registry System.** *J Chem Inf Model* 1978, 18: 58-58.
31. *CAS REGISTRY - The gold standard for chemical substance information.* http://www.cas.org/content/chemical-substances/.
32. Krushinski JH, Robertson DW, Wong DT (1991) **3-aryloxy-3-substituted propanamines.** USPTO US5023269 A.
33. Villani FJ, Wong JK (1987) **Antihistaminic 8-(halo)-substituted 6,11-dihydro-11-(4-piperidylidene)-5H-benzo[5,6]cyclohepta[1,2-b]pyridines.** USPTO US4659716 A.
34. Lowe DM, Sayle RA**. LeadMine: A grammar and dictionary driven approach to chemical entity recognition;** 2013. ***BioCreative Challenge Evaluation Workshop.*** vol. 2 pp. 47.
35. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, et al. **BRAT: a web-based tool for NLP-assisted text annotation**; 2012. ***Association for Computational Linguistics.*** pp. 102-107.
36. Lewin I, Kafkas S, Rebholz-Schuhmann D. **Centroids: Gold standards with distributional variation**; 2012. ***LREC.*** pp. 3894-3900.

Chapter 5

Recognition of chemical entities: combining dictionary-based and grammar-based approaches



*Published: **Akhondi SA,** Hettne KM, van der Horst E, van Mulligen EM, Kors JA*

Journal of Cheminformatics 2015, 7:S10

Abstract

Background

The past decade has seen an upsurge in the number of publications in chemistry. The ever-swelling volume of available documents makes it increasingly hard to extract relevant new information from such unstructured texts. The BioCreative CHEMDNER challenge invites the development of systems for the automatic recognition of chemicals in text (CEM task) and for ranking the recognized compounds at the document level (CDI task). We investigated an ensemble approach where dictionary-based named entity recognition is used along with grammar-based recognizers to extract compounds from text. We assessed the performance of ten different commercial and publicly available lexical resources using an open source indexing system (Peregrine), in combination with three different chemical compound recognizers and a set of regular expressions to recognize chemical database identifiers. The effect of different stop-word lists, case-sensitivity matching, and use of chunking information was also investigated. We focused on lexical resources that provide chemical structure information. To rank the different compounds found in a text, we used a term confidence score based on the normalized ratio of the term frequencies in chemical and non-chemical journals.

Results

The use of stop-word lists greatly improved the performance of the dictionary-based recognition, but there was no additional benefit from using chunking information. A combination of ChEBI and HMDB as lexical resources, the LeadMine tool for grammar-based recognition, and the regular expressions, outperformed any of the individual systems. On the test set, the F-scores were 77.8% (recall 71.2%, precision 85.8%) for the CEM task and 77.6% (recall 71.7%, precision 84.6%) for the CDI task. Missed terms were mainly due to tokenization issues, poor recognition of formulas, and term conjunctions.

Conclusions

We developed an ensemble system that combines dictionary-based and grammar-based approaches for chemical named entity recognition, outperforming any of the individual systems that we considered. The system is able to provide structure information for most of the compounds that are found. Improved tokenization and better recognition of specific entity types is likely to further improve system performance.

Background

The past decade has seen a massive increase in the number of chemical publications in the scientific literature. The ever-swelling volume of available documents makes it increasingly hard to manually find and extract relevant information from such texts [1,2]. Automatic indexing of individual publications by the chemical entities mentioned in them, can make it easier to find new information. Ranking these chemical entities by recognition confidence can be helpful in judging the relevance of the publication. Also, knowing the location of every mention of chemical compounds in these publications is of use to establish relationships with other entities or concepts [3].

Different text-mining approaches can be taken to extract chemical named entities from text. The various approaches have been categorized as dictionary-based, morphology-based (or grammar-based), and context-based [3]. In dictionary-based approaches, different matching methods can be used to detect matches of the dictionary terms in the text [3]. This requires good-quality dictionaries. The dictionaries are usually produced from well-known chemical databases. This approach may well capture non-systematic chemical identifiers, such as brand or generic drug names, which are source dependent and are generated at the point of registration. The drawback of a dictionary approach is that it is nearly impossible to also include all systematic chemical identifiers, such as IUPAC names [4] or SMILES [5], which are algorithmically generated based on the structure of the chemical compound and follow a specific grammar [6]. These predefined grammars are sets of rules or guidelines developed to refer to a compound with a unique textual representation (systematic term or identifier). These terms should have a one-to-one correspondence with the structure of the compound. Grammar-based approaches expand their extractions through the capture of systematic terms by utilizing these sets of rules, for example by means of finite state machines [7]. Therefore grammar-based approaches can extract systematic terms that are missing from the dictionaries. Both dictionary-based and grammar-based approaches may suffer from tokenization problems [3]. Following the third approach, context-aware systems use machine learning techniques and natural language processing (NLP) to capture chemical entities. Machine learning techniques utilize the manually annotated chemical terms in a training set of documents to automatically learn and define patterns to extract terms from text [3]. The drawback of machine learning approaches is the need for a sufficiently large annotated corpus for training the system.

Extraction of chemical entities from text has shown to be difficult. Among the main reasons are the large number of terms and synonyms within the chemical domain, the failure to follow guidelines when creating systematic terms by authors, the use of characters such as hyphens and commas within chemical terms, and the ambiguity and inconsistency within and across chemical databases [2,6,8]. Studies have tackled these difficulties using the approaches previously mentioned. Hettne et al. [9] extracted chemical terms from text using a dictionary-based approach (through a system called Peregrine [10]). Funk et al. [11] evaluated the performance of three different dictionary-based systems (MetaMap [12], NCBO Annotator [13], and

ConceptMapper [14]) by examining different parameters over multiple ontologies. Lowe et al. developed Opsin, which uses a grammar to transfer chemical nomenclature into structures [15].

In a later study Lowe et al. [16] further improved dictionary-based approaches by introducing 485 grammar-based rules to identify systematic terms. Others (e.g., Leaman et al. [17]) have investigated machine-learning approaches with a focus on conditional random fields (CRFs) [18], hidden mark models (HMMs), and maximum entropy markov models (MEMMs) [19] to extract chemical terms from text. In a recent study, Campos et al. [20] developed Neji, an open source package that integrates dictionary-based and machine-learning approaches to extract biomedical terms from text.

The BioCreative CHEMDNER challenge [8] intends to encourage the development of systems that can index chemical entities (especially the ones that are associated with a chemical structure) in scientific journals. Challenge participants were invited to submit results for two different tasks. The chemical document indexing (CDI) subtask pursues the creation of a list of the chemical entities in a document, ranked according to their confidence of recognition [8]. The chemical entity mention recognition (CEM) subtask aims at establishing the location of every mentioned chemical entity within a document [8]. The CHEMDNER organizers provided the participants with a manually annotated gold standard corpus [21] for training their systems. Overall 65 groups registered for the challenge and 27 groups (both academic and commercial) submitted results [8].

We investigated an ensemble approach where dictionary-based named entity recognition is used along with grammar-based recognizers and chemical toolkits to extract compounds from text. We analyzed the performance of ten different commercial and publicly available lexical resources using Peregrine, an open source indexing system [10,22], along with three different chemical compound recognizers. Different combinations of resources and recognizers were explored to find the best combination to extract the compounds.

Methods

Our approach was to extract non-systematic chemical identifiers using dictionary-based methods and systematic identifiers using grammar-based methods. We extracted compound family names using a defined ChEBI family dictionary, and database identifiers using a set of manually defined regular expressions. We merged the extractions of these systems. We first concentrated on the CEM subtask where we carried out chemical entity mention recognition. For the CDI subtask we determined confidence scores for all recognized terms and used these to rank the mentions.

Corpus

The CHEMDNER corpus [21] was used for the development and the evaluation of our system. The corpus consists of 10,000 manually annotated Medline abstracts divided in a training set and a development set (3,500 abstracts each), and a test set (3,000 abstracts). An additional sample dataset with 30 abstracts was also made available through the corpus. The abstracts in the test set were provided as part of a blinded set of 20,000 abstracts (participants did not know which of these abstracts were part of the test set), which the teams had to process in the evaluation phase of the challenge. The corpus has been annotated with the following entity types: abbreviation (e.g., "DMSO"), family (e.g., "Iodopyridazines"), formula (e.g., "(CH3)2SO"), identifier (e.g., "CHEBI:28262"), multiple (e.g., "thieno2,3-d and thieno3,2-d fused oxazin-4-ones"), systematic (e.g., "2-Acetoxybenzoic acid"), trivial (e.g., "Aspirin"), and undefined (e.g., "C4-C-N-PEG9"), concentrating on mentions with practical relevance as to potential target applications (focusing on chemical entities with structures) [21]. Therefore general compounds not associated with chemical structures were not annotated throughout the corpus. The combination of sample set, training set, and development set, collectively called the training material further on, was used to develop the ensemble system.

Lexical resources

We extracted all the terms (a term denoting a compound and consisting of one or more words) from the databases described below, including brand names, synonyms, trade names, generic names, research codes, Chemical Abstracts Service (CAS) numbers, and any other compound-relevant information. Since we wanted to focus on compounds with structures, only records with MOL file representations of chemical structures [23] were extracted.

ChEBI [24] Chemical Entities of Biological Interest (ChEBI) is a freely accessible dictionary of small molecular entities. Manually checked and annotated (three star) compounds and their associated MOL file representations of chemical structures were extracted, including all synonyms, brand names, ChEBI names, and International Nonproprietary Names (INNs).

ChEMBL [25] ChEMBL is a freely accessible database of bioactive molecules with drug-like properties. Chemical records are manually curated and standardized. Relevant information was extracted from ChEMBL records with associated MOL files.

ChemSpider [26] The ChemSpider database is a freely accessible chemical structure database, owned by the Royal Society of Chemistry [27]. It contains structures, properties and associated information for compounds gathered from more than 470 data sources. The information in the database is validated automatically by robot software, and manually by annotators and crowdsourcing [26,28,29]. We only used the subset of compounds that were manually validated.

DrugBank [30] DrugBank is a freely accessible database containing information on drugs and drug targets. Most of the data in DrugBank is expertly curated from primary literature sources [31]. All synonyms, brand names, CAS numbers, INNs, and generic names were extracted from DrugBank records with MOL files.

HMDB [32] The Human Metabolome Database (HMDB) contains human body-related small molecule metabolites information. The database links chemical, clinical and biological data. All compounds within HMDB are manually annotated by at least two annotators [33].

NPC [34] NIH Chemical Genomics Center Pharmaceutical Collection (NPC) contains clinical approved drugs from the USA, Europe, Canada and Japan. The data are automatically screened for curation [34]. The NPC browser 1.1.0 was used to extract synonyms, CAS numbers, and structure names for compounds with structures.

TTD [35] Therapeutic Target Database (TTD) contains known and explored therapeutic targets and their corresponding drugs. Targets are only included in TTD if they have been described in the literature [36]. All synonyms and drug names were extracted.

PubChem [37] PubChem is a database that provides information regarding biological activities of small molecules. PubChem stores molecular structures and bioassay data from different contributors [37]. A subset of compounds likely to have structure-activity relationships and/or other biological annotations [38] with all of their corresponding synonyms derived from PubChem substances were downloaded.

In addition to the databases above, which all contain information on compound structure, we also explored two large lexical resources that do not provide structure information.

Jochem [9] The joined lexical resource Jochem is a dictionary of small molecules and drugs, containing information from multiple sources. The dictionary is designed for text mining and all integrated data have been filtered, curated and disambiguated automatically [9]. All compounds and their corresponding information were extracted from Jochem.

UMLS [39] The Unified Medical Language System (UMLS) is a collection of biomedical concepts from different lexical resources grouped by 135 different semantic types [39]. UMLS provides a mapping among these lexical resources. Automatic auditing tools are used to discover and resolve possible errors [40,41]. Concepts belonging to a subset of 21 chemical-related semantic types were selected and extracted from UMLS.

To capture family names, we also created a dictionary from the ChEBI ontology where we only took parent compounds that did not appear in the ChEBI three-star database, assuming that these terms have a high likelihood of being a family name. We call this dictionary **ChEBI family**.

Table 1 shows the number of compounds and the number of terms for each of the resources. The total number of unique (case-sensitive) terms was 25,795,580.

Table 1: Number of records and number of terms in the terminological resources.

Resource	Number of compounds	Number of terms	Structure
ChEBI	23,240	85,036	Yes
ChEMBL	22,245	29,488	Yes
ChemSpider	2,957,105	5,235,393	Yes
DrugBank	6516	31,991	Yes
HMDB	40,200	364,541	Yes
NPC	14,666	131,795	Yes
TTD	3,196	127,568	Yes
PubChem	4,235,189	19,420,462	Yes
Jochem	362,928	2,062,333	No
UMLS	329,464	743,791	No
ChEBI Family	22,635	90,166	No

Stop words

In a recent study, Funk et al. [11] described the effect of different parameters such as use of stop words on automatic extraction of biomedical concepts from text. In this study we investigate the influence of stop words on automatic extraction of chemical terms from text. Several stop-word lists were analyzed for their ability to improve system performance, viz. English basic words (100 words) [42], the PubMed stop-word list (133 words) [43], the Jochem stop-word list (258 words) [9], and stop-words derived from the CHEMDNER annotation guidelines (116 words) [21]. Terms found by dictionary-based or grammar-based matching were disregarded if they were part of the stop-word lists. The basic English stop-word list and the PubMed stop-word list contain common English words, with 51 shared terms like "about", "all", "most", and "make". The Jochem stop-word list and the CHEMDNER derived stop-word list focused on more specific ambiguous terms, such as "crystal" or "acid" for the Jochem set, and "insulin" or "lead" for the CHEMDNER set. These two sets only shared five words.

Dictionary-based recognition

We employed the Peregrine tagger [10,22] to analyze the performance of the individual terminological resources. Tokenization of text that contains chemical terms can be complicated as compound names may include punctuation, such as commas or brackets. We used Peregrine with the tokenizer previously developed by Hettne et al. [9]. All the terms from the terminological resources were used to index the training material with different settings for case sensitivity and noun-phrase (NP) chunking.

Case sensitivity To study the effect of case sensitivity of characters within chemical names on the performance of the system, we indexed the text in separate runs with different matching settings: case insensitive, case sensitive, and partial case sensitive (only case sensitive for abbreviations, defined as terms where the majority of characters consists of capitals and digits, e.g. "BaTiO3").

NP chunking Assuming that chemical compounds will mostly be present in the noun phrases of a sentence, the experiments were also repeated by only feeding noun phrases extracted with the OpenNLP chunker [44] to Peregrine. The OpenNLP chunker has previously been shown to score best in performance and usability on NP recognition in biomedical text [45].

Grammar-based recognition

A number of public and commercial software packages that can find chemical entities in text were used for the grammar-based recognition approach. ChemAxon's Document-to-Structure toolkit (D2S) [46], NextMove's LeadMine [47], and OSCAR 4 [48] were used for this purpose. These tools have also implemented grammar-based recognition of systematic chemical identifiers. D2S uses grammars along with dictionaries to extract chemicals from text. D2S can also extract information from optical character recognition text and has the ability to recognize chemical structures from text (image extraction) [46]. NextMove's LeadMine uses a filtered dictionary along with 485 rules (grammars defined for chemical nomenclatures naming) to find and extract systematic names. The tool provides automatic spelling correction which allows the tool to extract misspelled terms from documents. The tool also supports multiple languages [47]. Oscar is an open-source software package for extracting named entities from chemical publications. The tool uses different types of models (such as a Bayesian model, pattern recognition, and a Maximum Entropy Markov Model) to extract terms from documents [48]. All the tools were used with their default settings, without further training, adjustment or tuning.

Regular expressions

Database identifiers of compounds are one of the entity types annotated in the CHEMDNER corpus [21], e.g., LY541850 or AMN082. This subset was used to define a set of regular expressions that served to index the abstracts for chemical database identifiers. As an example, "LY[\ ]{0,1}[1-9][0-9]{5,6}" captures the letters "LY" followed by a space (optional) and six or seven digits (the first of which is not 0).

Ensemble system

The stop-word lists were employed for both dictionary-based and grammar-based recognition. The dictionary-based recognition was applied using different settings for case sensitivity and NP chunking. We used the BioCreative evaluation script [49] to calculate precision, recall, and F-score (using exact matching of entity boundaries without considering entity type). The scores for the grammar-based recognizers and the regular expressions were also calculated in the same manner. We then heuristically selected different combinations of terminological resources, grammar-based recognizers and regular expressions, and assessed the performance of each ensemble. Our strategy was to have at least one system from each approach. The ensemble system merged the outputs of the various systems. All combinations of up to three lexical resources, the grammar-based recognizers, and the regular expressions were assessed, and the ensemble system with the highest F-score was determined. For comparison, we also investigated a simple voting scheme, where a term is accepted if the number of resources and systems by which the term is found, is at least equal to a voting threshold.

In the final setup we tried to improve our system by extending our dictionary with all gold-standard annotations from the training material that our system initially missed. Further improvement was reached by singling out indexed terms that overlapped. In these cases, the longest term (greater number of characters) was kept. If the terms had the same number of characters, they were ranked based on the subsystems that extracted them: regular expressions, grammar-based, dictionary-based (decreasing priority). If any or both of the overlapping terms were captured by more than one system, the term with highest priority was chosen. In rare cases where the overlapping terms had the same size and the same priority, one term was randomly chosen.

Ranking

To perform the CDI subtask, we needed a sorted list of unique mentions of the chemical terms in each document. The terms should be ranked according to an estimated confidence of recognition. We therefore determined a "confidence score" for each chemical term as follows. Abstracts from the whole of Medline were divided into two subgroups based on subject categories from the ISI Web of Knowledge [50] (Table 2). The first group consisted of 1,979,485 abstracts from chemical journals, employing the same subject categories as described in the CHEMDNER guidelines [21]. The second group contained 73,603 abstracts from non-chemical journals (e.g., journals in the subject category "Agricultural economics & policy") carefully chosen through the ISI Web of Knowledge classification. All abstracts were indexed by Peregrine with all lexical resources. We assumed that chemical terms would be present more frequently in chemical abstracts than in non-chemical abstracts. For each term, the ratio of the tf*idf (term frequency times inverse document frequency) scores for both abstract sets was computed and transformed into a confidence score between zero and one: if ratio < 1 then score = ratio * 0.5 else score = 1 - 0.5/ratio. A term with high confidence is found more frequently in chemical abstracts than in non-chemical abstracts and therefore is likely to be a chemical term. Vice versa, a term with low confidence is likely to be non-chemical, or highly ambiguous. For example, the drug "Indomethacin" (with DrugBank id DB00328) was found 15,421 times in the chemical abstracts and only once in the non-chemical abstracts, resulting in a high confidence score of 0.99. The ambiguous term "Merit" (synonym of "Imidacloprid" with HMDB id HMDB40292) was found 779 times in the chemical and 101 times in the non-chemical abstracts and obtained the low score of 0.14 after normalization.

Table 2: Subject categories in the ISI Web of Knowledge that contain chemical or non-chemical related journals.

Chemical related	**Non-chemical related**
Biochemistry & molecular	Agricultural economics & policy
Biology	Automation & control systems
Chemistry, applied	Computer science, information systems
Chemistry, medicinal	Computer science, software engineering
Chemistry, multidisciplinary	Computer science, theory & methods
Chemistry, organic	Education, scientific disciplines
Chemistry, physical	Instruments & instrumentation
Endocrinology & metabolism	Mathematics
Engineering, chemical	Mechanics
Polymer science	Physics, mathematical
Pharmacology & pharmacy	Robotics
Toxicology	Telecommunications

The confidence score was taken to rank the term. If it was not available (due to time constraints for the challenge we did not compute scores for terms only captured by regular expressions or grammar-based recognition, which took much more processing time than dictionary-based recognition), the term was ranked according to the precision of the system that indexed the term. In cases where multiple systems indexed the term the highest score was applied.

Results

Individual systems

Table 3 shows the baseline performance of the dictionary-based and grammar-based named entity recognition with and without stop-word removal on the 7030 abstracts in the training material. The dictionary-based named entity recognition was performed with case sensitive matching.

Table 3: Performance (in %) of individual systems on the training material, before and after stop-word removal. The highest score in each column is bolded.

	Baseline			**Baseline + stop-word removal**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
Dictionary-Based						
ChEBI	28.3	40.6	33.4	77.7	39.7	52.6
ChEMBL	**87.9**	18.7	30.8	**88.8**	18.7	30.9
ChemSpider	65.4	39.0	48.9	80.4	38.4	51.9
DrugBank	63.0	17.2	27.0	78.1	17.1	28.1
HMDB	53.2	34.5	41.8	81.3	33.9	47.9
NPC	46.8	26.7	34.0	59.7	26.4	36.6
TTD	43.9	14.7	22.1	82.9	14.4	24.6
PubChem	17.4	59.0	26.9	61.1	57.9	**59.5**
Jochem	64.2	52.5	**57.8**	67.1	52.5	58.9
UMLS	37.7	51.1	43.4	45.4	50.8	47.9
ChEBI Family	10.4	16.6	12.8	29.4	16.3	21.0
Grammar-based						
Oscar	25.1	**63.2**	35.9	28.4	**62.4**	39.0
LeadMine	61.3	47.4	53.4	71.1	47.1	56.7
ChemAxon	80.9	41.8	55.1	82.5	41.7	55.4

The baseline F-scores without stop-word removal fluctuate between 12.8% and 57.8%, with Jochem, ChemAxon and LeadMine performing the best. ChEMBL obtained a high precision of 87.9% but with a poor recall of 18.7%. Oscar, PubChem and Jochem had the highest recalls, but with moderate to poor precisions. ChEBI Family gained the lowest F-score, which can be explained by the fact that its scope was limited to chemical family names. Further analysis revealed that 40.3% of the annotated family names were captured by ChEBI Family. The low precision of ChEBI Family is mainly due to the presence of terms such as "role", "proteins", "inhibitors", "metabolites", which are not blocked as they are not present in the stop-word list. The use of the stop-word lists greatly improved the precision and F-score of the majority of resources. The performance of ChEMBL and ChemAxon remained nearly constant showing that these systems extract few of the stop words in our lists. Use of the stop-word lists hardly affects recall, with a largest decrease of only 1.1% for PubChem.

Table 4 gives a further breakdown of the performance improvement for the individual stop-word lists that were used. Clearly, the largest improvements are seen for the Basic English terms (up to 23.7 percentage points with an average of 4.1) and the PubMed stop-word list (up to 22.3 percentage points with an average of 3.6). Among the terms that had a large effect on precision were basic English terms such as "In" (extracted 32367 times of which only 5 are annotated in the corpus as Formula) and "As" (extracted 7087 times of which 33 cases are annotated as Formula). Many more general terms were also extracted mostly as false positives, such as "protein", "DNA", "insulin", and "water".

Table 4: Effect of individual stop-word lists on F-score.

	Baseline			Basic English			PubMed stop words			Jochem stop words			CHEMDNER guidelines		
Resource	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
ChEBI	28.3	40.6	33.4	69.0	40.3	50.9	63.0	40.3	49.2	28.5	40.1	33.3	29.3	40.6	34.0
ChEMBL	87.9	18.7	30.8	87.9	18.7	30.8	87.9	18.7	30.8	88.8	18.7	30.9	87.9	18.7	30.8
ChemSpider	65.4	39.0	48.9	74.4	38.8	51.0	65.3	38.8	48.7	69.3	38.6	49.6	67.9	39.0	49.5
DrugBank	63.0	17.2	27.0	63.0	17.2	27.0	63.0	17.2	27.0	78.1	17.1	28.1	63.1	17.2	27.0
HMDB	53.2	34.5	41.8	74.6	34.4	47.1	72.6	34.4	46.7	55.7	34.0	42.2	55.2	34.5	42.5
NPC	46.8	26.7	34.0	47.6	26.5	34.0	46.6	26.5	33.7	52.2	26.7	35.3	53.3	26.7	35.6
TTD	43.9	14.7	22.1	64.9	14.5	23.7	66.0	14.5	23.8	50.6	14.7	22.8	43.9	14.7	22.1
PubChem	17.4	59.0	26.9	44.5	58.7	50.6	42.4	58.6	49.2	18.9	58.3	28.5	17.9	59.0	27.4
Jochem	64.2	52.5	57.8	65.2	52.5	58.2	64.2	52.5	57.8	64.1	52.5	57.7	67.1	52.5	58.9
UMLS	37.7	51.1	43.4	45.4	50.8	43.6	38.0	51.1	43.6	40.0	50.8	44.9	42.4	51.1	46.4
ChEBI	10.4	16.6	12.8	21.0	16.6	18.5	21.0	16.6	18.5	10.8	16.4	13.1	11.6	16.6	13.7
Family Oscar	25.1	63.2	35.9	25.4	63.0	36.2	25.3	62.9	36.1	25.7	62.7	36.4	27.7	63.2	38.5
LeadMine	64.9	47.4	54.8	66.4	47.4	55.3	64.9	47.4	54.8	68.0	47.1	55.7	72.8	47.4	57.4
ChemAxon	80.9	41.8	55.1	80.9	41.8	55.1	80.9	41.8	55.1	81.1	41.7	55.1	83.3	41.8	55.5

Case sensitivity

To study the influence of case sensitivity on the dictionary-based approach, we indexed the training data using case insensitive, case sensitive, and partial case sensitive matching for all terminological resources (Table 5). The results did not show a large difference in most of the cases although (partial) case sensitive matching improved the F-score of ChEBI by 7.1 percentage points and reduced the score of TTD by 2.7 percentage points.

Table 5: F-score of terminological resources for different case sensitivity matching.

	Insensitive			Sensitive			Partial sensitive		
Resource	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
ChEBI	71.2	33.5	45.6	77.7	39.7	52.6	76.7	40.2	52.7
ChEMBL	91.6	18.9	31.3	88.8	18.7	30.9	88.5	18.8	31.1
ChemSpider	78.4	40.5	53.4	80.4	38.4	51.9	80.3	39.6	53.0
DrugBank	76.0	17.5	28.4	78.1	17.1	28.1	78.4	17.5	28.6
HMDB	79.3	35.1	48.6	81.3	33.9	47.9	81.5	35.1	49.1
NPC	58.5	26.8	36.8	59.7	26.4	36.6	59.9	27.1	37.4
TTD	78.3	16.8	27.6	82.9	14.4	24.6	81.1	14.7	24.9
PubChem	56.4	57.2	56.8	61.1	57.9	59.5	60.4	58.6	59.5
Jochem	67.1	52.5	58.9	67.1	52.5	58.9	66.4	53.5	59.3
UMLS	44.7	51.6	47.9	45.4	50.8	47.9	45.3	51.3	48.1
ChEBI Family	29.4	16.3	21.0	29.4	16.3	21.0	29.4	16.4	21.1

NP chunking

To study the possible gain through NP chunking on dictionary-based approaches, we applied the OpenNLP chunker to extract noun phrases from the training material. The noun phrases were then indexed with Peregrine using all terminological resources. Table 6 shows higher precision and F-scores for most of the systems as compared to the baseline values (cf. Table 3), in particular for PubChem and ChEBI. As expected, recall drops, but only by 0.3 to 1.9 percentage points.

Table 6: Performance (in %) of individual systems in combination with NP chunking, before and after stop-word removal.

	Baseline + NP chunking			Baseline + NP chunking + stop-words		
ChEBI	56.3	39.4	46.4	77.5	38.5	51.5
ChEMBL	87.8	18.2	30.1	88.6	18.2	30.1
ChemSpider	70.1	37.9	49.2	81.5	37.3	51.2
DrugBank	62.9	16.8	26.5	76.6	16.7	27.5
HMDB	73.5	33.7	46.2	82.0	33.1	47.2
NPC	46.8	26.0	33.5	59.1	25.7	35.9
TTD	66.6	14.4	23.6	83.0	14.0	24.0
PubChem	32.7	57.0	41.6	61.5	55.9	58.6
Jochem	64.3	50.6	56.7	67.4	50.6	57.8
UMLS	36.6	49.2	42.0	44.3	48.9	46.5
ChEBI Family	18.4	15.9	17.1	28.8	15.6	20.3
ChEBI	56.3	39.4	46.4	77.5	38.5	51.5

The removal of stop-words in combination with the NP chunking system gives a further improvement of performance, but to a much smaller extent than for the baseline system. This is largely because most of the stop-words are not part of the noun phrases and disregarding them has no effect. Based on a comparison between the performances in Table 3 and Table 6 we decided to dispense with NP chunking as there was no gain.

Regular expressions

The regular expressions detected 44.4% of the chemical database identifiers, with a precision of 90.4%. Further analysis of the false-positive and false-negative detections showed many partial extractions, e.g., "LY2090314" was extracted as an identifier while a prefix had also been annotated as part of the identifier ("[(14)C]LY2090314").

Ensemble system

We evaluated different combinations of terminological resources (applying different case-sensitivity settings), grammar-based recognizers, and regular expressions on the training data. The ensemble system with the best F-score consisted of the combination of ChEBI, HMDB, LeadMine, and the regular expressions, yielding an F-score of 66.6% (Table 7).

Table 7: Performance of the ensemble system on the training material.

Ensemble system	**CDI task**			**CEM task**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
ChEBI, HMDB, LeadMine, and RegEx	70.1	63.7	66.7	70.9	62.8	66.6
+ Missed terms added to dictionary	73.4	91.0	81.3	73.8	89.4	80.9
+ False-positive terms added to stop-word list	87.6	89.4	88.5	86.4	87.6	87.0
+ Removal of overlapping terms	91.8	89.1	90.9	91.8	87.4	89.5

The dictionaries performed best with case-sensitive matching but the differences with partial case-sensitive and with case-insensitive matching were marginal. Further addition of terminological resources to the ensemble system improved recall but decreased precision to a larger extent. For example, the addition of PubChem provided the largest increase in recall (about 7 percentage points), but decreased precision with about 8.9 percentage points, resulting in a drop in F-scores of 2.1 percentage points. Also note that the ensemble system had a better F-score than any of the individual systems (cf. Table 3). When we applied a voting approach, using all our sources and resources and varying the voting threshold between 1 and 15, the best F-score was 65.3% (precision 76.6%, recall 56.9%) for a threshold of 4.

We further analyzed the number of unique true positives (TPs) per entity type found by each of the systems within the ensemble system (Table 8). From a total of 37469 TPs captured by the ensemble system, 4139 cases were unique to ChEBI (mostly formula and abbreviation), 1878 were unique to HMDB (mostly trivial and abbreviation), 9480 cases were unique to LeadMine (mostly systematic terms) and 280 cases were unique to Regular expressions.

Table 8: Number of unique true positives found by each system in the ensemble system.

Entity type	**Regex**	**LeadMine**	**CHEBI**	**HMDB**
Trivial	8	1655	888	**711**
Systematic	0	**3945**	198	136
Family	0	2643	79	325
Formula	0	613	**1866**	110
Abbreviation	39	515	1093	596
Multiple	0	11	2	0
Identifier	**229**	98	13	0
Total	280	9480	4139	1878

We tried to further improve our system by expanding our dictionary with the gold-standard annotations from the training material that were missed by our system. This greatly improved the recall and F-score values (Table 7), although these estimates are optimistically biased since we evaluated the performance on the same dataset from which the newly added terms were derived. We also added all false-positive terms, i.e., terms indexed by our system but not annotated within the corpus (e.g., "peptide" and "carcinogen"), to our stop-word list, which further improved performance. Furthermore, we removed the shorter of two overlapping terms, which added 2.5 percentage points to the F-score, to reach 90.9% for the CDI task and 89.5% for the CEM task.

We submitted various runs to evaluate the system performance on the test set for both the CDI task and the CEM task (Table 9). The F-score of the baseline ensemble system improved by 9 percentage points after adding the false-negative terms of the training material to the dictionary and the false-positive terms to the stop-word list. A small further improvement was seen after the removal of overlapping terms, corroborating our findings on the training material. The best ensemble system obtained F-scores of 77.6% and 77.8% for the CDI and CEM tasks, respectively. Additional runs with a more recall-oriented system that included PubChem improved recall only slightly (about 3 percentage points) but greatly reduced precision (about 16 percentage points). We also tested whether removal of dictionary terms with low confidence scores would further improve the results, but this was not the case.

Table 9: Performance of the ensemble system on the test set.

Ensemble system	CDI task			CEM task		
	Precision	Recall	F-score	Precision	Recall	F-score
ChEBI, HMDB, LeadMine, and RegEx	70.5	64.8	68.0	73.1	64.6	68.6
+ Missed terms added & extended stop-word list	81.0	72.1	76.3	82.5	71.6	76.7
+ Removal of overlapping terms	84.6	71.7	77.6	85.8	71.2	77.8

Discussion

Extracting chemical terms from unstructured text has proven to be a difficult task [3]. Here we present an ensemble approach that combines a grammar-based approach to capture systematic chemical identifiers with a dictionary-based approach and regular expressions to capture non-systematic names. The ensemble system performed better than any individual system. Stop-word removal was shown to greatly improve system performance, as did the addition of false-negative and false-positive terms from the training material to the dictionary and stop-word list, respectively. The effect of different types of case-sensitive matching, use of NP chunking, and removal of dictionary terms that were likely to be highly ambiguous or non-chemical, did not essentially change the performance.

Our initial assumption about the beneficial effect of NP chunking on compound recognition was only partially met, in that the use of NP chunking alone improved performance but there was no additional value in combination with stop-word removal (cf. table 6). In a previous study by Kang et al. [51] dictionary-based recognition of diseases in scientific abstracts was improved by employing NLP techniques, including NP chunking. However, in that study only a small stop-word list was used. Also, chunk recognition in disease-related abstracts may be easier than in chemical abstracts, which can contain complex chemical names with multiple punctuation marks (e.g., hyphens, brackets).

On the test set, our best ensemble system achieved F-scores of 78% for both challenge tasks. The results of our ensemble system on the training material are much better than on the test set (cf. Tables 7 and 9), but clearly this is due to the fact that we used the training data to improve the system. However, if we compare the baseline ensemble system, for which no training was needed, the F-scores on the training and test sets were almost similar for the CDI and CEM tasks.

From the 27 teams that participated in the BioCreative CHEMDNER challenge, 20 teams used machine-learning methods to extract chemical terms from text. The most frequently used method was CRF [8]. The best scoring system for the CDI subtask [52] managed to gain a precision of 87%, a recall of 89%, and an F-score of 88%. This system uses CRF along with word clustering to extract terms. The state of the art system for the CEM subtask [17] obtained 89% precision, 86% recall, and 87% F-score. This system also uses CRF along with several pre-processing steps to extract chemical terms from text. With an F-score that was about 10 percentage points lower than the best systems, our ensemble system ranked eighth for the CDI task and seventh for the CEM task. Tuning of the grammar-based systems that we considered, could have resulted in a higher F-score. For example, LeadMine also participated in the challenge as a separate software system [16]. After tuning, LeadMine achieved an F-score that was nine percentage points higher than our ensemble system, and 32 percentage points higher than the baseline LeadMine system that we used. Also ChemAxon participated in the challenge and obtained an F-score of 77% (an increase of 22 percentage points compared to the version we used). Among the teams who used lexical resources, ChEBI, PubChem and DrugBank were most often used; 13 teams also used a stop-world list. Irmer et al. [53] used a dictionary-based approach along with modules to recognize formulas or handle specific scenarios (such as abbreviation or acronym expansion) and obtained an F-score of 77%. They introduced a set of words in a so-called grey list. Terms in this list were only annotated in specific circumstances. Some systems (e.g. [54]) also tried to create an ensemble system by combining machine learning, dictionary-based approaches and regular expressions, but obtained lower F-scores than our ensemble system. Finally, in our approach the ensemble system merges the outputs of a selected set of individual systems. Our results indicate that this approach produced a better result than a simple voting scheme. However, we did not explore more sophisticated approaches, such as weighted voting or integration into a learning framework [55]. Application of these techniques may further improve the performance of an ensemble system.

Our approach has several advantages. First, use of the terminological resources and grammar-based recognizers did not have to be trained. This is an advantage over machine-learning approaches that require a large training set, which is laborious and expensive to create. On the other hand, our results also indicate that a substantial performance improvement can be gained by using the training data to expand the dictionary and the stop-word list. Thus, if training data are available, they can straightforwardly be used to improve system performance for both dictionary-based and grammar-based approaches.

A second advantage is that our system can provide structures for most of the found terms. Although the supply of information about structures was not required for the CHEMDNER tasks, chemists are generally interested in the chemical structure of a chemical identifier recognized in text. The terminological resources in the ensemble system (ChEBI and HMDB) contained MOL files, and also the grammar-based method (LeadMine) can provide structures for the extracted terms. Only the terms extracted with the regular expressions and terms that were added based on the training data, are not linked to structure information.

There are also several limitations. While the precision of our best ensemble system was an acceptable 86%, the recall was a more modest 71%. Including other dictionaries in the ensemble improved recall, but deteriorated precision to a much larger extent. Also, we noticed that many of the missed chemical terms were due to tokenization issues, e.g., the formulas "WC" and "Na" were missed in the context of "(nano-WC)" and "(I(Na))", respectively (PMID 22954532). Improvement of our tokenizer will further be investigated.

Another limitation of the current ensemble system is that some of the entity types were poorly recognized, in particular the entity types Multiple and Formulas. Terms of these types are not well covered in our dictionary. Better recognition may be possible by the use of regular expressions specifically developed for these types.

Finally, it should be noted that we used the grammar-based recognition tools with their default parameter settings, and did not try to tune them to the tasks at hand. Further improvements may be possible if such tuning were done.

Conclusion

We developed an ensemble system that combines dictionary-based and grammar-based approaches to chemical named entity recognition, and obtained F-scores of 78% on the two CHEMDNER challenge tasks. The baseline version of the system did not require training, but we were readily able to improve performance by making use of the available training data. The system is capable of providing structure information for most of the compounds that are found. Improved tokenization and better recognition of specific entity types will likely further increase system performance.

References

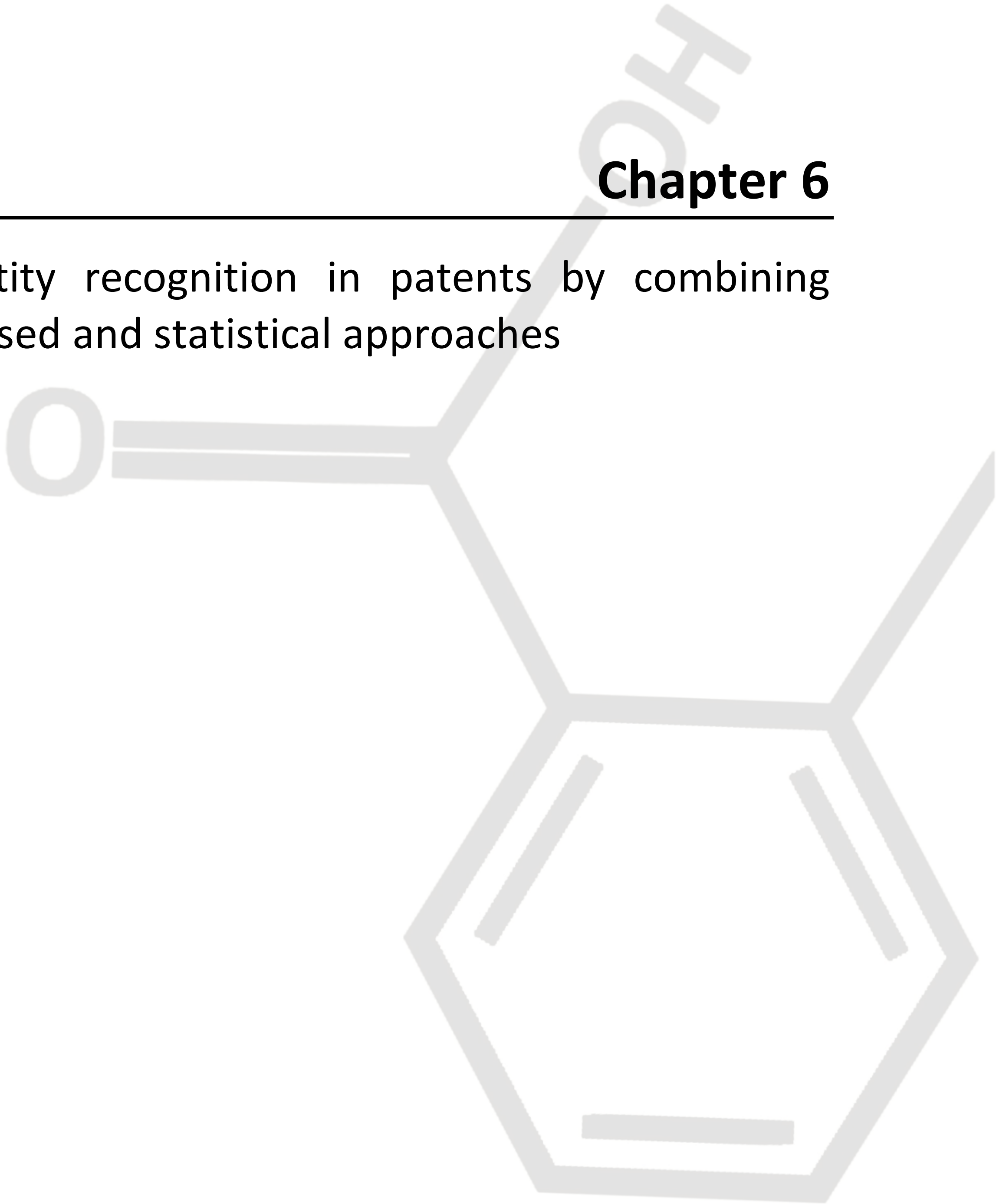
1. Yeh A, Morgan A, Colosimo M, Hirschman L**: BioCreAtIvE task 1A: gene mention finding evaluation.** *BMC Bioinformatics* 2005, **6(Suppl 1):**S2.
2. Eltyeb S, Salim N: **Chemical named entities recognition: a review on approaches and applications.** *J Cheminf* 2014, **6:**1-12.
3. Vazquez Miguel, Krallinger Martin, Leitner Florian, Valencia A: **Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications.** *Mol Inform* 2011, **30(6-7):**506-519.
4. *About IUPAC.* http://www.iupac.org/home/about.html.
5. Weininger D: **SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules.** *J Chem Inf Comput Sci* 1988, **28:**31-36.
6. Akhondi SA, Kors JA, Muresan S: **Consistency of systematic chemical identifiers within and between small-molecule databases.** *J Cheminf* 2012, **4:**35.
7. Sayle R, Xie PH, Muresan S: **Improved chemical text mining of patents with infinite dictionaries and automatic spelling correction.** *J Chem Inf Model* 2012, **52:**51-62.
8. Krallinger M, Leitner F, Rabal O, Vazquez M, Oryazabal J, Valencia A: **CHEMDNER: The drugs and chemical names extraction challenge**. *J Cheminform* 2015, **7(Suppl 1):**S1.
9. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, **Mulligen EM, Kleinjans J, Kors JA: A dictionary to identify small molecules and drugs in free text**. *Bioinformatics* 2009, **25:**2983-2991.
10. Schuemie MJ, Jelier R, Kors JA: **Peregrine: Lightweight gene name normalization by dictionary lookup.** *Proceedings of the Biocreative 2 workshop* 2007, 131-140.
11. Funk C, Baumgartner W Jr, Garcia B, Roeder C, Bada M, Cohen KB, Hunter LE, Verspoor K: **Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters.** *BMC Bioinformatics* 2014, **15:**59.
12. Aronson AR: **Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program.** *Proceedings of the AMIA Symposium* American Medical Informatics Association; 2001, 17.
13. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA: **Comparison of concept recognizers for building the Open Biomedical Annotator.** *BMC Bioinformatics* 2009, **10:**S14.
14. Tanenblatt MA, Coden A, Sominsky IL: **The ConceptMapper Approach to Named Entity Recognition.** *Proceedings of Seventh International Conference on Language Resources and Evaluation (LREC'10)* 2010.
15. Lowe DM, Corbett PT, Murray-Rust P, Glen RC: **Chemical name to structure: OPSIN, an open source solution.** *J Chem Inf Model* 2011, **51:**739-753.
16. Lowe DM, Sayle RA: **LeadMine: A grammar and dictionary driven approach to chemical entity recognition.** *J Cheminform* 2015, **7(Suppl 1):**S5.
17. Leaman R, Wei C-H, Lu Z: **NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem.** *J Cheminform* **2015, 7(Suppl 1):**S3.
18. Wallach HM: **Conditional random fields: An introduction.** *Technical report, Dept. of CIS, Univ. of Pennsylvania* 2004.
19. McCallum A, Freitag D, Pereira FC: **Maximum Entropy Markov Models for Information Extraction and Segmentation.** *ICML* 2000, 591-598.

20. Campos D, Matos S, Oliveira JL: **A modular framework for biomedical concept recognition**. *BMC Bioinformatics* 2013, **14:**281.
21. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe DM, Sayle RA, Batista-Navarro RT, Rak R, Huber T, Rocktaschel T, Matos S, Campos D, Tang B, Xu H, Munkhdalai T, Ryu KH, Ramanan SV, Nathan S, Zitnik S, Bajec M, Weber L, Irmer M, Akhondi SA, Kors JA, Xu S, An X, Sikdar UK, Ekbal A, Yoshioka M, Dieb TM, Choi M, Verspoor K, Khabsa M, Giles CL, Liu H, Ravikumar KE, Lamurias A, Couto FM, Dai H, Tsai RT, Ata C, Can T, Usie A, Alves R, Segura-Bedmar I, Martinez P, Oryzabal J, Valencia A: **The CHEMDNER corpus of chemicals and drugs and its annotation principles.** *J Cheminform* 2015, **7(Suppl 1):**S2.
22. *Peregrine.* https://trac.nbic.nl/data-mining.
23. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, J L: **Description of several chemical structure file formats used by computer programs developed at molecular design limited.** *J Chem Inf Comput Sci* 1992, 244-255.
24. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical Entities of Biological Interest: an update.** *Nucleic Acids Res* 2010, **38:**D249-254.
25. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res* 2012, **40**:D1100-1107.
26. Pence HE, Williams A: **ChemSpider: An Online Chemical Information Resource.** *J Chem Educ* 2010, **87:**1123-1124.
27. *Royal Society of CHEMISTRY.* http://www.rsc.org/.
28. *What is ChemSpider?.* http://www.chemspider.com/About.aspx?.
29. Hettne KM, Williams AJ, van Mulligen EM, Kleinjans J, Tkachenko V, Kors JA: **Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining.** *J Cheminf* 2010, **2:**3.
30. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, et al: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res* 2011, **39:**D1035-1041.
31. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, 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:**D1091-1097.
32. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, et al: **HMDB: a knowledgebase for the human metabolome.** *Nucleic Acids Res* 2009, **37:**D603-610.
33. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, et al: **HMDB: the Human Metabolome Database.** *Nucleic Acids Res* 2007, **35:**D521-526.
34. Huang R, Southall N, Wang Y, Yasgar A, Shinn P, Jadhav A, Nguyen DT, Austin CP: **The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics.** *Sci Transl Med* 2011, **3:**80ps16.
35. Zhu F, Han B, Kumar P, Liu X, Ma X, Wei X, Huang L, Guo Y, Han L, Zheng C, Chen Y: **Update of TTD: Therapeutic Target Database.** *Nucleic Acids Res* 2010, **38:**D787-791.
36. Chen X, Ji ZL, Chen YZ: **TTD: therapeutic target database.** *Nucleic Acids Res* 2002, **30:**412-415.
37. Bolton EE, Wang Y, Thiessen PA, Bryant SH: **PubChem: integrated platform of small molecules and biological activities.** *Annual reports in computational chemistry* 2008, **4:**217-241.

38. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, Varkonyi P, Xie PH: **Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data.** *Drug Discov Today* 2011, **16:**1019-1030.
39. Bodenreider O: **The Unified Medical Language System (UMLS): integrating biomedical terminology.** *Nucleic Acids Res* 2004, **32:**D267-270.
40. Morrey CP, Geller J, Halper M, Perl Y: **The Neighborhood Auditing Tool: a hybrid interface for auditing the UMLS.** *J Biomed Inform* 2009, **42:**468-489.
41. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ: **A review of auditing methods applied to the content of controlled biomedical terminologies.** *J Biomed Inform* 2009, **42:**413-425.
42. *100 English basic words.* http://en.wiktionary.org/wiki/Category:100_English_basic_words.
43. *PubMed Stopwords list.* http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/020_170.html.
44. *Apache OpenNLP library.* http://opennlp.apache.org/.
45. Kang N, van Mulligen EM, Kors JA: **Comparing and combining chunkers of biomedical text.** *J Biomed Inform* 2011, **44:**354-360.
46. *ChemAxon-Document to Structure.* http://www.chemaxon.com/products/document-to-structure/.
47. *NextMove Software-LeadMine.* http://www.nextmovesoftware.com/products/LeadMine.html.
48. Jessop DM, Adams SE, Willighagen EL, Hawizy L, Murray-Rust P: **OSCAR4: a flexible architecture for chemical text-mining.** *J Cheminf* 2011, **3:**41.
49. *BioCreative evaluation library scripts.* http://www.biocreative.org/resources/biocreative-ii5/evaluation-library/.
50. *Web of Knowledge.* http://webofknowledge.com.
51. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA: **Using rule-based natural language processing to improve disease normalization in biomedical text.** *J Am Med Inform* 2013, **20:**876.
52. Lu Y, Yao X, Wei X, Ji D, Liang X: **CHEMDNER system with mixed conditional random fields and multi-scale word clustering.** *J Cheminform* 2015, **7(Suppl 1):**S4.
53. Irmer M, Bobach C, Böhme T, Laube U, Püschel A, Weber L: **Chemical Named Entity Recognition with OCMiner.** *BioCreative Challenge Evaluation Workshop* 2013, **2:**92.
54. Usié A, Cruz J, Comas J, Solson F, Alves R: **CheNER: a tool for the identification of chemical entities and their classes in biomedical literature.** *J Cheminform* 2015, **7(Suppl 1):**S15.
55. Khabsa M, Giles CL: **Chemical entity extraction using CRF and an ensemble of extractors.** *J Cheminform* 2015, **7(Suppl 1):**S12.

Chapter 6

Chemical entity recognition in patents by combining dictionary-based and statistical approaches



Published: ***Akhondi SA****, Pons E, Afzal Z, van Haagen H, Becker BF, Hettne KM, van Mulligen EM, Kors JA*

Database 2016, 2016: baw061

Abstract

We describe the development of a chemical entity recognition system and its application in the CHEMDNER-patent track of BioCreative 2015. This community challenge includes a Chemical Entity Mention in Patents (CEMP) recognition task and a Chemical Passage Detection (CPD) classification task. We addressed both tasks by an ensemble system that combines a dictionary-based approach with a statistical one. For this purpose the performance of several lexical resources was assessed using Peregrine, our open-source indexing engine. We combined our dictionary-based results on the patent corpus with the results of tmChem, a chemical recognizer using a conditional random field classifier. To improve the performance of tmChem, we utilized three additional features, viz. part-of-speech tags, lemmas and word-vector clusters. When evaluated on the training data, our final system obtained an F-score of 85.21% for the CEMP task, and an accuracy of 91.53% for the CPD task. On the test set, the best system ranked sixth among 21 teams for CEMP with an F-score of 86.82%, and second among nine teams for CPD with an accuracy of 94.23%. The differences in performance between the best ensemble system and the statistical system separately were small.

Introduction

Exploration of the chemical and biological space covered by patents is essential in the early stages of activities in the field of medicinal chemistry [1]. Analyzing patents can help to understand compound prior art and to pinpoint alternative starting points for chemical research [2]. Important tasks in patent analysis are the recognition of chemical names, the identification of chemical structure images, and the conversion of the extracted names and images into a structure-searchable form [3]. Other types of entities in medicinal chemistry patents, such as genes and proteins, diseases, or particular numerical values, may also be relevant to extract and to relate to chemical entities [4]. The extracted information is often compiled in structured databases that are easy to query and facilitate computational analysis.

Usually, patent information is manually extracted [5]. This process is laborious and expensive due to the length of chemical patent texts, which may take hundreds of pages, and their complexity (mixture of scientific, technical and legal language, typographical errors, optical character recognition errors, etc.). These problems are aggravated by the sheer number of medicinal chemistry patents [1, 6]. Automatic methods to recognize chemicals in patents can help to ease this process, but have proven to be elaborate and demanding [7, 8]. One of the impediments is that very few large annotated gold-standard corpora for algorithm training and testing are available [9].

The automatic extraction of chemical and biological data from medicinal chemistry patents was addressed in the CHEMDNER-patents track of BioCreative V [10]. The track was organized as a community challenge to stimulate the development and comparative assessment of chemical and biological entity recognizers, and consisted of three tasks: (i) Chemical Entity Mention in Patents (CEMP), focusing on chemical entity recognition in patents; (ii) Chemical Passage Detection (CPD), focusing on the classification of patent titles and abstracts according to whether they contain chemical entities; and (iii) Gene and Protein Related Object (GPRO), focusing on the recognition of gene and protein mentions in patents. Our team participated in the CEMP and CPD tasks.

Previous text-mining research mostly concentrated on chemical name recognition in scientific literature [4, 11]. Recently, a large-scale patent resource, SureChEMBL [12], has become available, which contains compounds extracted from the full-text, images and attachments of patents, and provides comprehensive search capabilities. Chemical entity recognition is the first step in the SureChEMBL data extraction pipeline, but performance figures have not been presented as yet [12]. A variety of systems to extract chemicals from Medline abstracts were developed and evaluated as part of the previous BioCreative IV CHEMDNER task [11]. The top-ranking systems in that challenge used machine-learning techniques based on conditional random fields (CRFs) [11]. However, some systems that combined dictionary-based and rule-based approaches also achieved competitive results [13, 14]. For the current challenge, we combined a dictionary-based approach with a statistical, CRF-based approach, and investigated the performance of the ensemble system for the CEMP and CPD tasks on the CHEMDNER-patents data.

Materials and methods

Data

The CHEMDNER-patent corpus [10] was used for the development and evaluation of our system. The corpus comprises a training corpus of 14 000 manually annotated patent records (each record consisting of a title and an abstract), divided into a training set and a development set of 7000 records each, and a test set of 40 000 patent records, of which only 7000 were manually annotated. The annotation process and guidelines were largely similar to the ones used for the BioCreative IV CHEMDNER corpus, and have been described extensively [10, 15]. Table 1 summarizes the number of annotated chemicals and chemical-related titles and abstracts. Only the annotations of the training and development sets were made available to the participants in the challenge. For evaluating the performance of their system on the test set, teams could submit up to five runs. To produce the evaluation results, we used the BioCreative evaluation software (www.biocreative.org/resources/biocreative-ii5/evaluation-library/) and focused on micro-averaged recall, precision and F-score to assess system performance for the CEMP task, and on sensitivity (=recall), specificity and accuracy for the CPD task. Given the number of true-positive (TP), false-positive (FP), false-negative (FN) and true-negative (TN) detections, these metrics were computed as follows:

recall = TP/(TP + FN), precision = TP/(TP + FP), F-score = 2*precision*recall/(precision + recall), specificity = TN/(TN + FP) and accuracy = (TP + TN)/(TP + FN + FP + TN).

We also used the Markyt prediction analysis toolkit (www.markyt.org/biocreative/analysis) to visualize the results.

Table 1: Characteristics of the CHEMDNER patent corpus.

	Training	Development	Test	Total
Patent records	7,000	7,000	7,000	21,000
Manual chemical annotations	33,543	32,142	33,949	99,634
Unique chemical annotations	11,977	11,386	11,433	34,796
Chemical-related titles and abstracts	9,152	8,937	9,270	27,359

Dictionary-based approach

We used Peregrine, our open-source indexer [16], to analyze the performance of the different chemical dictionaries. Tokenization was done with a tokenizer previously developed by Hettne et al. [17]. Term matching was carried out by partial case-sensitive matching: case-sensitive for abbreviations (defined as terms of which the majority of characters consists of capitals and digits), case-insensitive for all other terms.

Dictionaries

To construct our dictionaries, we selected seven well-known, publicly available chemical databases covering a wide range of compounds, namely: Chemical Entities of Biological Interest (ChEBI) [18], ChEMBL [19], DrugBank [20], the Human Metabolome Database (HMDB) [21], the NCGC Pharmaceutical Collection (NPC) [22], PubChem [23] and the Therapeutic Target Database (TTD) [24]. For each database record, we gathered all chemical terms (available from possibly different record fields). Chemical terms were only extracted from records that had associated chemical structures in the form of MOL files [25]. In the following, we briefly describe the databases and the fields from which identifiers were extracted.

ChEBI is concerned with molecular entities, focusing on small chemical compounds [18]. It provides an ontological classification with parent and child relationships. We extracted data for all three-star (i.e. manually annotated) compounds from ChEBI SD files. This included synonyms, ChEBI names, brand names, International Nonproprietary Names (INNs) and International Union of Pure and Applied Chemistry (IUPAC) names.

ChEMBL contains information on drug-like bioactive compounds [19]. In addition to literature-derived data, ChEMBL also contains Food and Drug Administration (FDA) approved drugs. The data available through ChEMBL have been manually extracted and standardized [26]. Extracted fields include preferred names, synonyms, FDA alternative names, INNs, United States Adopted Names (USANs) and United States Pharmacopoeia (USP) names.

DrugBank provides information regarding drugs, including chemical, pharmacological and pharmaceutical data, and their targets [27]. DrugBank data are curated by a curation team, which relies on primary literature sources. During production and maintenance, all synonyms and brand names within DrugBank are extensively reviewed and only the most common synonyms are kept [20]. We extracted brand names, generic names, synonyms, Chemical Abstracts Service (CAS) numbers, and IUPAC names from the DrugBank SD files and DrugCards.

HMDB lists small-molecule metabolites found in the human body [21]. The database links chemical, clinical, molecular-biology and biochemistry data. HMDB is both automatically and manually curated [21]. All generic names, synonyms, CAS numbers and IUPAC names were extracted from the HMDB SD files and MetaboCards.

NPC provides information on clinically approved drugs from USA, Europe, Canada and Japan for high-throughput screening [22]. We extracted preferred names and synonyms using the NPC browser 1.1.0.

PubChem provides information on the biological activity of small molecules [23]. It consists of three different databases: a compound database, a substance database and a bioassay database. We extracted structures and all corresponding IUPAC identifiers and synonyms for a subset of compounds that had structure–activity relationships or other biological annotations. This subset of compounds was introduced by Muresan et al. [1] and is the same subset of PubChem compounds that we used in our previous study on chemical entity recognition [13]. The PubChem compound database does not contain synonyms. This information is available through the PubChem substance database. The relations between PubChem substance identifiers (SIDs) and compound identifiers (CIDs), which have been created by PubChem through in-house chemical structure standardization [23], are specified in the ‘PubChem_CID_associations’ tag available in the downloadable structure data files. We used the relations between SIDs and CIDs to extract the synonyms from the substance database and assign them to the corresponding compounds.

TTD contains information about therapeutic protein and nucleic acid targets of drugs, corresponding pathways and targeted diseases [24]. All trade names, drug names, CAS numbers and synonyms were extracted.

Dictionary construction and combination

For each database, a dictionary consisting of the extracted chemical terms was constructed. Each term was linked to one or possibly more (in case of ambiguity) compounds, represented by their MOL files. Dictionaries were combined by merging the identifiers of all compounds in the dictionaries. To determine which compounds in different dictionaries were the same, we used the same approach as in previous studies [28, 29]. Briefly, we compared MOL files by converting them into InChI strings, which provide unique textual representations of the MOL files. Compounds with identical InChI strings were considered the same, and the corresponding identifiers were merged.

Term exclusion

To improve the precision of the dictionary-based approach, we applied an exclusion list of terms as previously described [13]. Briefly, the list contains common English words, like ‘about’, ‘all’ and ‘make’, and ambiguous terms, such as ‘acid’, ‘crystal’ and ‘lead’. We expanded this list with exclusion terms mentioned in the annotation guidelines for the CEMP task.

We also removed terms that were false-positive detections in the training data, but only if the ratio of true-positive to false-positive detections was lower than 0.3. This threshold was heuristically set based on the training data in order to prevent erroneous removal of overall correctly recognized terms because of an occasional false-positive detection. When testing on the development set, exclusion ratios were calculated for all false-positive terms in the training set; when evaluating on the test set, ratios were computed for all false-positive terms in the combined training and development sets.

Term inclusion

We identified all missed terms (false negatives) in the training set and re-indexed the texts for these terms. Only those terms that, after re-indexing, did not result in false-positive detections in the training set or had an exclusion ratio larger than 0.5 were added to the dictionary. When evaluating on the test set, the combined training and development sets were used to collect the false negatives and to determine whether they should be included in the dictionary.

Machine-learning approach

We used the tmChem chemical recognizer system [30], one of the best performing systems in the previous BioCreative CHEMDNER challenge [11]. The tmChem system is an ensemble system that combines the output of two CRF-based systems. The first system is a modified version of the BANNER system [31], the second is based on the tmVar system [32], which employs CRF ++ libraries (https://taku910.github.io/crfpp/). Previous results of tmChem showed that the second system outperformed the first as well as the ensemble system [30]. We therefore only used the second system.

Pre-processing

The tmChem system transliterates non-ASCII Unicode characters to a similar ASCII equivalent. As some non-ASCII Unicode characters were not handled (causing a system crash when encountered in text), we expanded the transliteration capacities as necessary. We also replaced a vertical bar enclosed by parentheses or brackets (e.g. [|]), because these combinations caused tmChem to crash as well.

Features

Our initial feature set consisted of all features extracted by tmChem, including stemmed words, prefixes and suffixes, character counts (digit, uppercase, lowercase), semantic affixes (such as trivial rings) and chemical elements [30].

Three additional types of features were determined and used to train tmChem: part-of-speech (POS) tags, lemmas and word-vector clusters. We used the BioC natural language processing pipeline [33] to generate POS tags with MaxentTagger [34] and lemmas with BioLemmatizer [35]. Recent studies have shown that features based on clusters of word vectors can improve classification performance [36, 37]. We used the word2vec tool (https://code.google.com/p/word2vec/) to generate clusters of word vectors. Word2vec employs K-means clustering. The number of the cluster to which a word belonged was taken as a feature.

We generated separate word clusters during the development phase and the test phase of the challenge. During development, the clusters were generated from the 14 000 titles and abstracts in the training and development sets. These data were extended with 200 full-text chemical patents that had been used in a previous study [9]. We experimented with different numbers of clusters (K = 300, 500, 1000). For testing our final system, clusters were generated using all 54 000 records in the corpus plus the 200 full-text patents, with K = 1000.

Post-processing

For the machine-learning approach, the tmChem post-processing steps were applied [30]. These include enforcing tagging consistency (for each term that was found by the CRF at least twice within an abstract, any term mention in the abstract that the CRF had not identified was also tagged), abbreviation resolution (tagging corresponding abbreviations and long forms), boundary revision (adding or removing unbalanced brackets or parentheses) and finding chemical database identifiers (through regular expressions).

We experimented with different sets of dictionaries for the dictionary-based approach and different sets of features for the machine-learning approach. All terms recognized by the dictionary-based system or the statistical system were taken as the output of the final ensemble system.

Text classification

For the CPD task (classification of patent titles and abstracts as chemical-related or not), we used a straightforward approach based on the output of the CEMP task. If our system recognized any chemical term in a text (title or abstract), the text was categorized as a chemical-related. Note that the title and abstract of each record were classified separately.

Results

Table 2 shows the number of compounds and the number of unique identifiers in the chemical databases. Clearly, PubChem is by far the largest database.

Table 2: Number of compounds and unique identifiers in chemical databases.

Database	No. of compounds	No. of identifiers
ChEBI	23,240	82,612
ChEMBL	22,245	28,411
DrugBank	6,516	31,948
HMDB	40,199	228,907
NPC	14,666	128,153
PubChem	4,235,189	19,049,175
TTD	3,196	121,744

The number of identifiers that are shared between pairs of databases is shown in Table 3. Although PubChem contains >90% of the identifiers in ChEMBL, DrugBank and TTD, the other databases are much less well covered by PubChem. The majority of identifiers in DrugBank is covered by NPC and TTD, but the overlap between all other pairs of databases is relatively low.

Table 3: Number of unique identifiers that overlap between pairs of chemical databases. The percentage coverage of the identifiers in the smallest sized database of each pair is given in parentheses.

Database	**ChEBI**	**ChEMBL**	**DrugBank**	**HMDB**	**NPC**	**PubChem**
ChEMBL	1,209 (4.3)					
DrugBank	2,444 (7.6)	3,931 (13.8)				
HMDB	4,885 (5.9)	2,293 (8.1)	5,946 (18.6)			
NPC	3,406 (4.1)	6,508 (22.9)	23,865 (74.7)	7,444 (5.8)		
PubChem	45,021 (54.5)	26,251 (92.4)	28,943 (90.6)	52,533 (22.9)	69,873 (54.5)	
TTD	4,481 (5.4)	4,507 (15.9)	18,028 (56.4)	6,503 (5.3)	23,901 (19.6)	119,819 (98.4)

Table 4 shows the performance of the dictionary-based approach on the development set, with and without use of the list of exclusion terms. Use of the exclusion list gives a substantial precision improvement for most dictionaries. The PubChem dictionary demonstrates the highest recall among the individual dictionaries, which may be explained by the large size of the PubChem dictionary and the fact that it contains the majority of terms from the other dictionaries. The dictionaries from ChEMBL and DrugBank had the highest precision, which is likely due to the fact that these databases are highly curated. The low recall of the dictionaries can be explained by their low coverage of systematic names and chemical family names. Of the 9194 systematic names that were annotated in the development corpus, recognition rates ranged from 7.5% for TTD to 53.8% for PubChem (median 31.0%). For family names, which form the largest annotation group (n = 11 710), recognition rate varied between 3.3% and 20.4% (median 9.1%).

Table 4: Performance of different dictionaries and dictionary combinations with and without removal of exclusion terms.

	Without exclusion			**With exclusion**		
Dictionary	**Precision**	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
ChEBI	56.51	29.47	38.74	78.87	28.42	41.79
ChEMBL	84.53	20.46	32.94	85.11	19.87	32.22
DrugBank	68.20	17.28	27.58	85.15	16.89	28.19
HMDB	66.11	29.38	40.68	79.59	28.19	41.63
NPC	30.90	44.85	36.59	55.23	30.61	39.39
TTD	66.89	14.07	23.24	80.90	13.89	23.71
PubChem	34.30	47.11	39.69	67.03	45.64	54.30
All combined	30.85	50.32	38.25	53.66	48.59	51.00
ChEBI-HMDB	55.46	36.98	44.37	78.12	35.45	48.77
ChEMBL-DrugBank	70.51	23.94	35.74	83.02	23.16	36.21

Table 4 also shows the performance of several combinations of dictionaries. As to be expected, the combination of all dictionaries after term exclusion has the highest recall (49%), but the lowest precision (54%). The combination of dictionaries from ChEBI and HMDB, which we used in the previous BioCreative CHEMDNER task (13), gave a recall of 35% and a precision of 78%. The combination of ChEMBL and DrugBank resulted in the highest precision (83%).

Table 5 shows the incremental performance of the ensemble system trained on the training corpus and evaluated on the development corpus, when different feature sets and term-processing steps were added. We only present dictionary-based results for the combination of ChEMBL and DrugBank as this combination produced the highest F-score on the training data when combined with the CRF. For the CEMP task, all incremental steps improved the F-score, except when terms that were missed in the training set were included in the dictionary. The best ensemble system attained an F-score of 85.21% with a precision of 84.88% and a recall of 85.55%. For the CPD task, the system that comprised all processing steps, including the addition of missed terms, achieved the best performance with an accuracy of 91.84% (sensitivity 97.00%, specificity 82.74%).

When we only used the CRF-based system (trained on all features) to process the development set, we obtained an F-score of 84.78% (precision 86.14%, recall 83.47%) on the CEMP task, and an accuracy of 90.96% (sensitivity 94.23%, specificity 85.19%) on the CPD task.

Table 6 shows the performance for both tasks on the test set. We submitted runs of the ensemble systems with and without the addition of missed terms. For comparison, we also submitted a run for the statistical system alone (including all features and post-processing).

For the CEMP task, the statistical system performed best (F-score 86.82%), slightly better than the ensemble system without the addition of missed terms (F-score 86.55%). For CPD, the ensemble system with missed terms reached the best performance (accuracy 94.23%), slightly better again than the system without missed terms (93.93%). Our best systems ranked sixth among 21 participating teams for the CEMP task, and second among nine teams for the CPD task.

Table 5: Performance of the ensemble system trained on the training set and tested on the development set.

	CEMP task			CPD task		
System	**Precision**	**Recall**	**F-score**	**Sensitivity**	**Specificity**	**Accuracy**
Dictionary-based (ChEMBL-DrugBank)	70.51	23.94	35.74	50.63	88.41	64.29
+ Exclusion list	83.02	23.16	36.21	44.29	94.37	62.40
+ Term removal (exclusion ratio 0.3)	88.85	23.09	36.65	42.14	97.12	62.02
+ CRF original features	84.96	83.83	84.39	95.11	85.33	91.57
+ Post-processing (CRF)	84.50	84.91	84.70	95.39	85.01	91.64
+ POS and lemmatization features	84.72	85.09	84.90	95.40	85.25	91.73
+ Word-vector cluster features	84.88	85.55	**85.21**	95.31	84.87	91.54
+ Missed terms (exclusion ratio 0.5)	75.88	88.63	81.76	97.00	82.74	**91.84**

Table 6: Performance of different systems on the test set.

	CEMP task			CPD task		
System	**Precision**	**Recall**	**F-score**	**Sensitivity**	**Specificity**	**Accuracy**
Statistical	86.83	86.81	**86.82**	96.13	88.67	93.61
Statistical + dictionary without missed terms	84.92	88.25	86.55	97.00	87.91	93.93
Statistical + dictionary with missed terms	77.76	90.84	83.79	98.03	86.79	**94.23**

Discussion

We investigated the combination of dictionary-based and statistical approaches for chemical entity recognition in patents. Our results show that the recall of the chemical dictionaries on the CEMP task is low, and even a combination of all dictionaries gives a recall and precision of only around 50%. The low recall can be explained by the fact that many systematic chemical terms and chemical family names were lacking in our lexical resources. Meanwhile, the machine-learning approach yielded a much higher precision and recall (86% and 83%, respectively). In order to maintain the high precision of the ensemble system, we used the dictionary combination with the highest precision (ChEMBL and DrugBank). For the CEMP task, this supplied us with a system that slightly improved machine-learning performance on the development set, but not on the test set. Thus, there was no performance gain for this task by the use of a combined dictionary-based and statistical approach over a statistical approach alone. For the CPD task, the ensemble system performed better than the statistical system alone, both on the development set and on the test set. This may be explained by the 1.9 percentage point higher sensitivity of the ensemble system, in combination with a similar decrease in specificity. As the majority of titles and abstracts in the development and test sets are chemical-related (see Table 2), sensitivity weighs more heavily than specificity in the accuracy. For both tasks, our results on the test set were better than those on the development set, indicating that overtraining did not occur.

Contrary to our expectation, the inclusion of false-negative terms in the dictionary decreased the performance for the CEMP task, both on the development set and on the test set. This may partly be explained by tokenization issues that split chemical terms in multiple parts. Some of these parts were then erroneously matched with the newly added dictionary terms, resulting in a drop in precision. For the CPD task, the increase in sensitivity more than compensated for the decrease in specificity, yielding a slightly improved accuracy of the ensemble system using the missed terms.

Although furnishing structure information about the recognized chemicals was not part of the challenge, this information is often important in practical applications. We are able to readily associate dictionary terms with structures because we only extracted terms from chemical records with structure information. Of the chemical terms in the development set, 23% is found by the dictionary-based approach and can be linked to structures. For the machine-learning approach, the mapping of recognized terms to structures is less straightforward, but part of these terms will consist of systematic chemical identifiers. These can also be converted into chemical structures using chemical naming conversion software (28, 29).

Considering that annotated patent corpora are scarce, the CHEMDNER corpus of annotated patent titles and abstracts is a highly valuable and important resource for further development and comparative assessment of algorithms. Recently, we have reported on the creation of another corpus of 200 annotated full-text patents, which is publicly available (9). We plan to use this corpus to evaluate and possibly improve the performance of our systems on full-text patents.

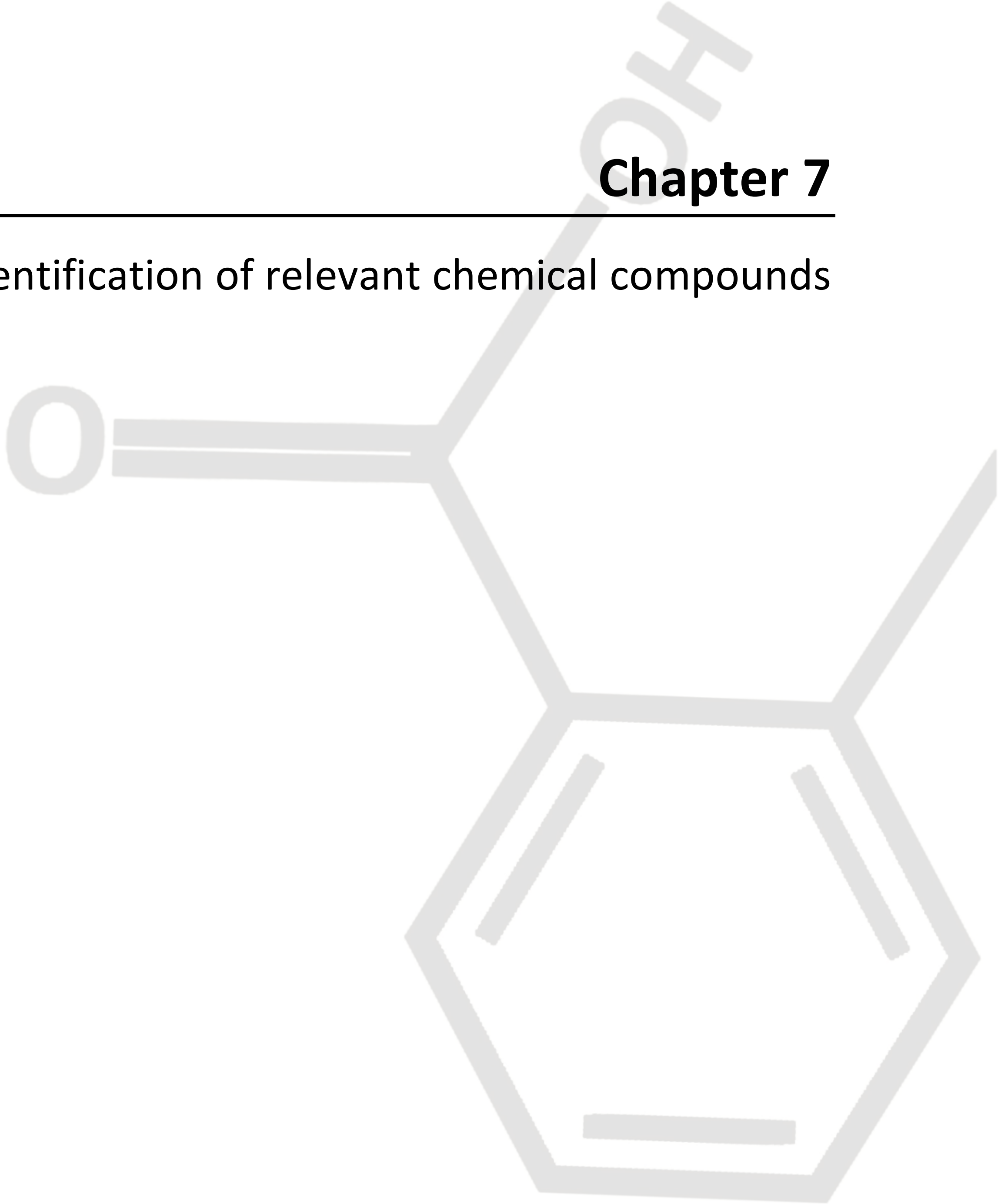
References

1. Muresan S, Petrov P, Southan C, et al: **Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data.** *Drug Discov. Today* 2011, **16:**1019–1030.
2. Tyrchan C, Bostrom J, Giordanetto F, et al: **Exploiting structural information in patent specifications for key compound prediction.** *J. Chem. Inf. Model.* 2012, **52:**1480–1489.
3. Banville DL: **Mining chemical structural information from the drug literature.** *Drug Discov. Today 2006*, **11:**35–42.
4. Vazquez M, Krallinger M, Leitner F, et al: **Text mining for drugs and chemical compounds: methods, tools and applications.** *Mol. Inform.* 2011, **30:**506–519.
5. Zimmermann M, Fluck J, Thi Le TB, et al: **Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology.** *Curr. Top. Med. Chem.* 2005, **5:**785–796.
6. Klinger R, Kolarik C, Fluck J et al: **Detection of IUPAC and IUPAC-like chemical names.** *Bioinformatics* 2008, **24:**i268–i276.
7. Tseng YH, Lin CJ, and Lin YI: **Text mining techniques for patent analysis.** *Inf. Process. Manag.* 2007, **43:**1216–1247.
8. Jessop DM, Adams SE, and Murray-Rust P: **Mining chemical information from open patents.** *J. Cheminform.* 2011, **3:**40.
9. Akhondi SA, Klenner AG, Tyrchan C et al: **Annotated chemical patent corpus: a gold standard for text mining.** *PLoS One* 2014, **9:**e107477.
10. Krallinger M, Rabal O, Lourenc OA. et al: **Overview of the CHEMDNER patents task.** In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* 2015. pp. 63–75.
11. Krallinger M, Leitner F, Rabal O et al: **CHEMDNER: the drugs and chemical names extraction challenge**. *J. Cheminform.* 2015, **7:**S1.
12. Papadatos G, Davies M, Dedman N et al: **SureChEMBL: a large-scale, chemically annotated patent document database.** *Nucleic Acids Res.* 2016, **44:**D1220–D1228.
13. Akhondi SA, Hettne KM, van der Horst E, van Mulligen EM, Kors JA: **Recognition of chemical entities: combining dictionary-based and grammar-based approaches.** *J. Cheminform.* 2015, **7:**S10.
14. Lowe DM and Sayle RA: **LeadMine: a grammar and dictionary driven approach to entity recognition.** *J. Cheminform.* 2015, **7:**S5.
15. Krallinger M, Rabal O, Leitner F et al: **The CHEMDNER corpus of chemicals and drugs and its annotation principles.** *J. Cheminform.* 2015, **7:**S2.
16. Schuemie,M J, Jelier R, and Kors JA: **Peregrine: lightweight gene name normalization by dictionary lookup.** In: *Hirschman L , Krallinger M and Valencia A (eds.). Proceedings of the Second BioCreative Challenge Evaluation Workshop. Centro Nacional de Investigaciones Oncologicas, Madrid, Spain* 2007. pp. 131–133.
17. Hettne KM, Stierum RH, Schuemie MJ et al: **A dictionary to identify small molecules and drugs in free text.** *Bioinformatics* 2009, **25:**2983–2991.
18. Hastings J, de Matos P, Dekker A et al: **The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013.** *Nucleic Acids Res.* 2013, **41:**D456–D463.

19. Bento AP, Gaulton A, Hersey A et al: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014, **42:**D1083–D1090.
20. Law V, Knox C, Djoumbou Y et al: **DrugBank 4.0: shedding new light on drug metabolism.** *Nucleic Acids Res.* 2014, **42:**D1091–D1097.
21. Wishart DS, Jewison T, Guo AC et al: **HMDB 3.0 - The Human Metabolome Database in 2013**. *Nucleic Acids Res.* 2013, **41:**D801–D807.
22. Huang R, Southall N, Wang Y et al: **The NCGC pharmaceutical collection: a comprehensive resource of clinically approved drugs enabling repurposing and chemical genomics.** *Sci. Transl. Med.* 2011, **3:**80ps16.
23. Kim S, Thiessen PA, Bolton EE et al: **PubChem Substance and Compound databases.** *Nucleic Acids Res.* 2016, **44:**D1202–D1213.
24. Zhu F, Shi Z, Qin C et al: **Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery.** *Nucleic Acids Res.* 2012, **40:**D1128–D1136.
25. Dalby A, Nourse JG, Hounshell WD et al: **Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited.** *J. Chem. Inf. Comput. Sci* 1992, **32:**244–255.
26. Gaulton A, Bellis LJ, Bento AP et al: **ChEMBL: a large-scale bioactivity database for drug discovery.** *Nucleic Acids Res.* 2012, **40:**D1100–D1107.
27. Knox C, Law V, Jewison T et al: **DrugBank 3.0: a comprehensive resource for 'omics' research on drugs.** *Nucleic Acids Res.* 2011, **39:**D1035–D1041.
28. Akhondi SA, Kors JA, and Muresan S: **Consistency of systematic chemical identifiers within and between small-molecule databases.** *J. Cheminform.* 2012, **4:**35.
29. Akhondi SA, Muresan S, Williams AJ et al: **Ambiguity of non-systematic chemical identifiers within and between small-molecule databases.** *J. Cheminform.* 2015, **7:**1–10.
30. Leaman R, Wei CH, and Lu Z: **tmChem: a high performance approach for chemical named entity recognition and normalization.** *J. Cheminform*. 2015, **7:**S3.
31. Leaman R and Gonzalez G: **BANNER: an executable survey of advances in biomedical named entity recognition.** *Pac. Symp. Biocomput* 2008, **13:**652–663.
32. Wei CH, Harris BR, Kao HY et al: **tmVar: a text mining approach for extracting sequence variants in biomedical literature.** *Bioinformatics* 2013, **29:**1433–1439.
33. Comeau DC, Islamaj Dogan R, Ciccarese P et al: **BioC: a minimalist approach to interoperability for biomedical text processing.** *Database* 2013, bat064.
34. Toutanova K, Klein D, Manning CD et al: **Feature-rich part-of-speech tagging with a cyclic dependency network. In: Heart,M. and Ostendorf,M.** (eds.). *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics, Stroudsburg (PA), USA* 2003. pp. 252–259.
35. Liu H, Christiansen T, Baumgartner WA et al: **BioLemmatizer: a lemmatization tool for morphological processing of biomedical text.** *J. Biomed. Semant.* 2012, **3:**3.
36. Deng L and Yu D: **Deep learning: methods and applications.** *Found. Trends Signal Process.* 2014, **7:**197–387.
37. Nikfarjam A, Sarker A, O'connor K et al: **Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features.** *J. Am. Med. Inform. Assoc.* 2015, **22:**671–681.

Chapter 7

Automatic identification of relevant chemical compounds from patents



*Submitted: **Akhondi SA,** Rey H, Schwörer M, Maier M, Toomey J, Nau H, Ilchmann G, Sheehan M, Irmer M, Bobach C, Doornenbal M, Gregory M, Kors JA*

Journal of ChemInformatics, 2018

Abstract

Background

In commercial research and development projects, public disclosure of new chemical compounds often takes place in patents. Only a small proportion of these compounds are published in journals, usually a few years after the patent. Patent authorities make available the patents but do not provide systematic continuous chemical annotations. Content databases such as Elsevier's Reaxys provide such services mostly based on manual excerptions, which are time-consuming and costly. Automatic text-mining approaches help overcome some of the limitations of the manual process. Different text-mining approaches exist to extract chemical entities from patents. Majority of them have been developed using sub-sections of patent documents and focus on mentions of compounds. Less attention has been given to relevancy of a compound in a patent. Relevancy of a compound to a patent is based on the patent's context. A relevant compound plays a major role within a patent. Identification of relevant compounds reduces the size of the extracted data and improves the usefulness of patent resources (e.g., supports identifying the main compounds). Annotators of databases like Reaxys only annotate relevant compounds. In this study, we design an automated system that extracts chemical entities from patents and classifies their relevance. To develop and evaluate the system, a patent corpus with annotations for chemical entities and their relevance was constructed.

Results

The gold-standard set contained 18,789 chemical entity annotations. Of these, 10% were relevant compounds, 88% were irrelevant, and 2% were equivocal. The performance (F-score) of the system on compound recognition was 84% on the development set and 86% on the test set. The relevancy classification system had an F-score of 86% on the development set and 82% on the test set.

Conclusions

Our system can extract chemical compounds from patents and classify their relevance with high performance. This enables the extension of the Reaxys database by means of automation.

Background

The number of chemistry-related publications has massively increased in the past decade [1]. These publications are mainly in the form of patent applications and scientific journal articles. A crucial step in early stages of medicinal chemistry activities is the exploration of the chemical space covered by these sources [1–4]. In commercial research and development projects, initial public disclosure of new chemical compounds often takes place in patent applications [4, 5]. On average, it takes an additional one to three years for a small fraction of these chemical compounds to appear in journal publications [5]. Therefore, a large selection of these chemical compounds are only available through patent documents [6]. Additionally, chemical patent documents contain unique information such as reactions, experimental conditions, mode of action [7], bioactivity data, and catalysts [1, 3]. Analysing such information becomes crucial [1, 4, 5, 8] as it allows the understanding of compound prior art, it provides a means for novelty checking and validation, and it points to starting points for chemical research in academia and industry [3, 7, 9, 10].

Patent data is freely available through different patent offices. Major patent authorities include the European Patent Office (EPO) [11], the United States Patent and Trademark Office (USPTO) [12], and the World Intellectual Property Organization (WIPO) [13]. Depending on the patent authority, the data are made available in the form of XML, HTML, text PDF, Optical Character Recognition (OCR) PDF, or image PDF. Patent documents usually follow a systematic structure consisting of title, bibliographic information (such as patent number, dates, inventors, assignees, International Patent Classification (IPC) classes), abstract, description, and claims. Most of the chemical data are available in the experimental section of the description, while chemical compounds that are claimed (i.e., will become protected by the patent) are available in the claim section [4]. Drawings, sequences, or other additional information will normally be found at the very end of the patent.

While patent authorities make available the patent documents, they do not provide systematic continuous chemical annotations and full-text searching capabilities [3], so manual or automatic excerption processes have been considered [1, 5, 7, 14]. Manual excerption processes result in high-quality content but are costly and time-consuming, and are therefore limited to commercial content providers [5]. Examples of content databases are Elsevier Reaxys [15, 16], CAS SciFinder [17], and Thomson Reuters Pharma [18]. These commercial resources provide high-quality content, such as compounds and their associated structures, facts associated to compounds, and reactions. Automatic approaches to extract information from patents have recently come into existence to overcome some of the aforementioned cost and time limitations. Examples of such resources include SureChEMBL [3], SCRIPDB [19], ChEBI database [20], IBM database [21], NextMove Software's reaction database [22], and databases that combine data from different sources (e.g., PubChem [23]). SureChEMBL provides continuous, up-to-date chemical annotations with structures derived from USPTO, EPO, WIPO, and the Japanese Patent Office (JPO) [24]. The information is extracted from full-text patents (except JPO), images, and attachment files [3]. This information is mostly derived by text mining and image mining. SCRIPDB is a chemical structure database from compounds and reactions. This information is built based on the digital chemical structure files provided by USPTO for a subset of its patents (grant patents, from 2001 until 2011) [19]. ChEBI database provides chemistry compounds and structures extracted from a subset of patent documents from the EPO office [20]. IBM database provides chemical compounds and structures derived from a subset of EPO, WIPO, and USPTO patents [21]. This information is derived by text-mining approaches. The reaction database of NextMove Software is also automatically generated by text mining the relevant experimental sections of patents covering the period 1976 - 2013 [22]. It proves difficult to maintain public databases and many of the above have become outdated.

Some of the automatic resources mentioned above incorporate the textual data content supplied by the content providers to build their database (such as SCRIPDB). Others use image mining and text mining approaches to extract data from the patent full-text document (e.g., SureChEMBL and IBM). Image-mining approaches convert images attached to patents into structures using image-to-structure tools (e.g., CLiDE Pro [25] in SureChEMBL) [4]. These tools have limitations in the interpretation of individual drawing features (such as chemical bonds) found in the structure diagrams of some images [25], and will not further be considered in this study. Text-mining approaches focus on the recognition of chemical compounds in patents [4]. Each recognised small compound should also be associated with a chemical structure. Different text-mining approaches exist to extract chemical entities from patents. The approaches can be categorized as dictionary-based, morphology-based (or grammar-based), or statistical [26–29]. Dictionary-based approaches use matching methods to identify compounds mentioned in a dictionary (e.g., generic drug names) within patents. This approach is limited by the compounds contained in the dictionary. Addition of all systematic compound identifiers to a dictionary is almost impossible as they are algorithmically generated based on the structure of a compound and a set of rules [30]. Grammar-based approaches use these rules to overcome this limitation and provide functionality to recognize systematic identifiers [26]. Statistical approaches use machine-learning techniques to recognize chemical compounds. These statistical-based recognizers are trained on manually annotated chemical terms [7]. Among the three approaches, statistical approaches have shown to perform the best [4, 31, 32] but they require a large annotated corpus for training [26, 33] and cannot associate compounds with structures. Correctness of the associated chemical structure to a recognized compound is essential in the field of chemistry [34, 35]. Often a combination of the methods above in the form of an ensemble system is used for chemical compound recognition [31, 36]. All systems require a gold-standard corpus for training, developing, and testing performance [30]. Producing such a corpus is laborious and expensive [7]. It involves development of well-defined annotation guidelines, selection and training of domain experts for annotation, selection of the data, annotation of the data by multiple annotators, and finally harmonization of the annotations [7].

Extracting information from patents automatically is fast but has limitations [7, 29, 37]. The majority of patent text-mining systems have been developed, trained, and tested using the title and abstract of the patent documents. Therefore their usage is not evaluated on full-text documents [31, 36]. More importantly, automatic extraction is mostly focused on extraction of all chemical compounds mentioned. In manually excerpted databases, the focus is on relevant compounds [5, 38]. A compound is relevant to a patent when it plays a major role within the patent application (e.g., starting material or a product in a reaction specified in the claim section).

Relevant compounds are a small fraction of all the compounds mentioned within the patent document [9, 39]. Automatic identification of the relevant compounds would greatly reduce the amount of extracted data from patents and can improve the usefulness of patent resources. Furthermore, these compounds can be used in predictive analyses to identify the key compounds within the patent (key compounds are the main compounds protected by the patent application and are usually well-hidden within the context) [9, 39]. To our knowledge, automatic identification of relevant compounds within patents has not yet been investigated.

The objective of this study is to identify relevant chemical compounds in patents using an automatic approach. To develop and evaluate our approach, a patent corpus with named-entity and relevancy annotations was built.

Methods

Figure 1 shows the relevancy classification workflow. The chemical patents are pulled through patent offices. The patent source documents are first normalized into a unified format. They are then fed into the chemical entity recognition system that consists of two different named-entity extraction systems, Chemical Entity Recognizer (CER) (Elsevier, Frankfurt, Germany) [40] and OCMiner (OntoChem, Halle, Germany) [41]. CER extracts chemical entities and tags them in the normalized input document. OCMiner further enriches the output of CER by extracting additional chemical entities and assigning confidence scores to all extracted entities of both systems. The associated structures of chemical compounds extracted by CER or OCMiner are generated, validated, and standardized using the Reaxys Name Service [42]. The chemical annotations in the patent corpus are used to train and test the chemical entity recognition system. The relevancy annotations in the corpus are used to train and test the relevancy classifier, which labels the chemical entities extracted by the chemical entity recognition system as relevant or irrelevant. Below we describe each of the components in more detail.

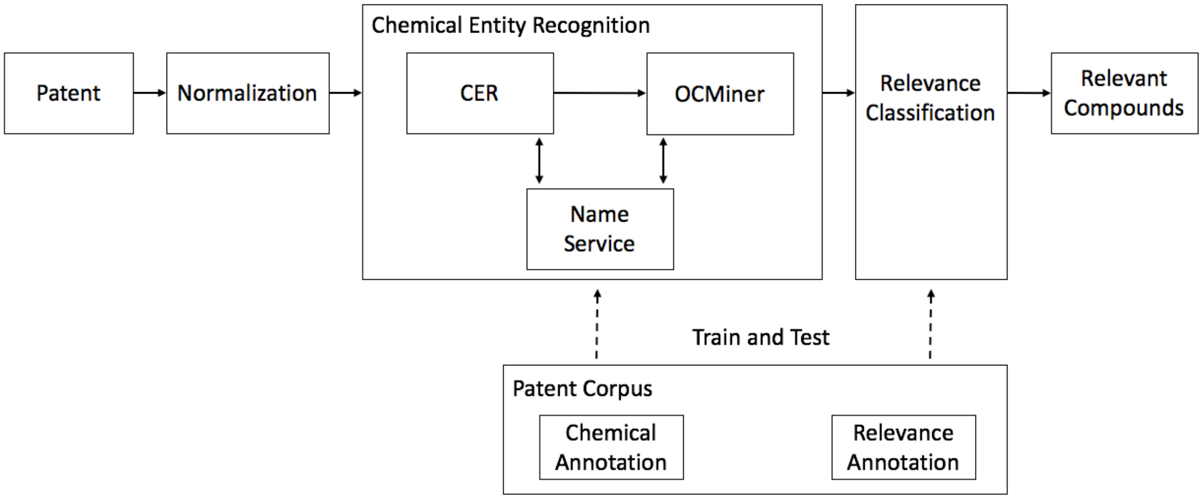


Figure 1: Workflow of the relevancy classification.

Normalization

The variety of input sources and file types need to be normalized into a unified text representation [4]. The normalization step is performed by converting all input files (e.g., XML, HTML, pdf) into a unified XML representation format. Predefined XML tags corresponding to heuristic information such as document sections (title, abstract, claims, description, and metadata) are used within this unified representation. The normalization also converts all character encodings into UTF-8 (8-bit Unicode Transformation Format).

During normalization, we store a one-to-one mapping between each character in the original text and the corresponding character in the normalized document. This provides us with the possibility to go back to the original document from the normalized text and vice versa. It also minimizes the efforts to update the annotations in the patent corpus in case of changes in normalization methodology (note that the documents in the corpus have also been normalized).

Patent corpus development

The development of the chemical patent corpus with chemical entity and relevancy annotations was done in two phases. Figure 2 illustrates the corpus creation process. The first phase focuses on building a corpus with chemical entity annotations. In phase two the corpus obtained from phase one is used to assign relevancy annotations to the entities annotated in phase one. In this phase, annotators also flagged any compounds with spelling mistakes. For each phase, a set of well-defined guidelines was developed that helped achieve annotation consistency.

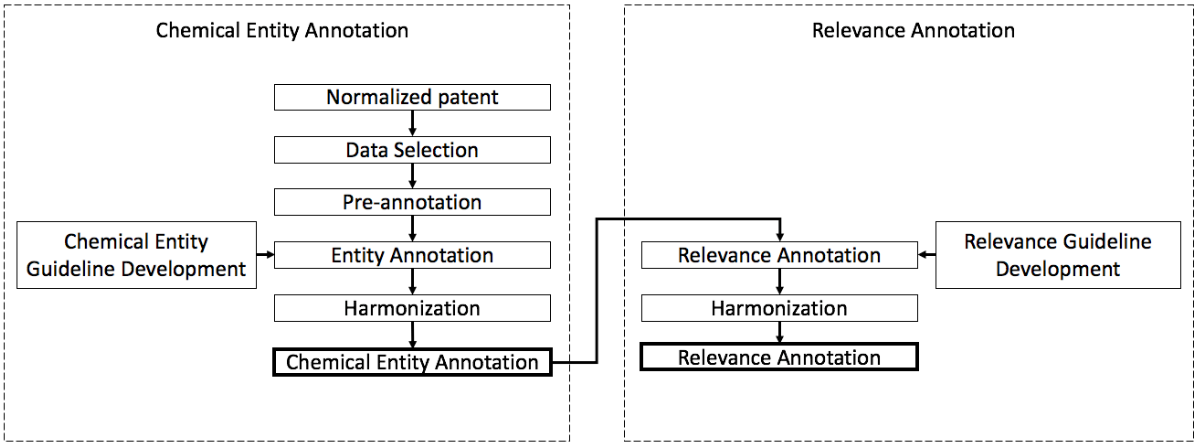


Figure 2: Patent corpus development.

Chemical entity annotation guideline

The chemical entity annotation guideline was developed based on our previous patent corpus development guideline [7], previous work by other scholars [32, 43–46], and the help of subject matter experts in Elsevier. The guidelines define the entities to be annotated. For each entity, positive and negative examples were provided. Additionally, any exception was defined and illustrated through examples. The guideline also defined how the annotation should be performed within the brat rapid annotation tool [47, 48]. Brat allows online annotation of text using pre-defined entity types. Annotators were asked to annotate chemical compounds (e.g., tetrahydrofuran), chemical classes (e.g., zirconium alkoxide), and suffixes or prefixes of these compounds (e.g., “stabilized” as prefix in “stabilized zirconia”, “nanoparticles” as suffix in “silver nanoparticles”).

Chemical compounds could be annotated in three categories: mono-component-compound (pure chemical compounds, e.g., systematic identifiers, trivial names, elements, chemical formulas); compound-mixture-part (e.g., “Magnesiaflux”, which scientifically is a mixture of 30% MgF2 and 70% MgO); or prophetic-compound (specific compounds that are uncharacterized within the text and are mentioned in claims or descriptions only for intellectual property protection).

Compound classes could be annotated in six categories: chemical-class (natural products or substructure names, e.g., heterocycle); biomolecules (e.g., insulin); polymers (e.g., polyethylene); mixture-classes (e.g., opium); mixture-part-classes (e.g., quinupristin), or Markush (textual description of a Markush formula, e.g., HaXbC-C-H).

Relevancy annotation guideline

For the relevancy annotation, a new set of guidelines were developed, which defined how relevant compounds should be identified. The relevancy annotation did not include suffixes and prefixes of compounds. In brief, relevancy is assigned as follows:

1. Prophetic compounds and Markush classes are relevant.

2. Compound-mixture-parts, mixture-part-classes, mixture-classes, polymers, and biomolecules are irrelevant.

3. Mono-component-compounds and chemical-classes are assigned relevance based on the context of the full patent text. They are considered relevant to the patent if: (a) the entity is present in the title or abstract section of the patent; (b) the entity is part of a reaction context (e.g., product, intermediate product, catalyst or starting material used in synthetic procedures); or (c) the entity or its measured property belongs to the invention in the claim section and is connected to the core invention of the patent. The mono-component-compounds and chemical-classes are irrelevant if: (a) the entity is only introduced for further explanation and is described beyond the invention; (b) the entity is described for reference or comparison; or (c) the entity is involved in a chemical reaction but not a starting material, product or catalyst.

Data selection

Patent documents are long and extensive. Annotation of full-text documents is time-consuming and expensive. Complexity was reduced by selecting snippets of patent text from a large set of patent documents that represented the diversity of the data. We downloaded all EPO patents with IPC class A or C (corresponding to chemistry) from a three-month period in 2016 [15, 16]. This yielded 19,274 patents, which were divided into snippets as follows. First, each patent was divided into six snippets containing title, abstract, claims, description, metadata, and non-English section of the patent. Second, since the performance of the brat toolkit drops on long files [7], snippets of more than 50 paragraphs were further divided into multiple snippets. From this set of snippets, a small set was selected for annotation. We performed random stratified sampling based on the sub-classes of IPC A and C (list available at http://web2.wipo.int/classifications/ipc/ipcpub). In addition, the following conditions were satisfied: 10% of the snippets were from titles, 10% from abstracts, 40% from claims, and 40% from descriptions, and all snippets were from different patents.

We selected a total of 131 snippets, which constitute our patent corpus. The IPC sub-classes that occurred most frequently were A61K, A61B, C07D, A61F, A61M, C12N.

Chemical entity annotation process

We selected ten chemistry graduates as annotators. The annotators were located in different European countries. To train the annotators, 11 of the 131 patent snippets were distributed among the annotators using the brat annotation tool [47, 48]. The snippets were pre-annotated with an untuned version of the chemical entity recognition software that is used in this study (see next section for the description of this software). The pre-annotations were displayed in brat and annotators were asked to modify incorrect pre-annotated entities (wrong boundary or entity type) and add missing entities according to the guideline (see Figure 3).

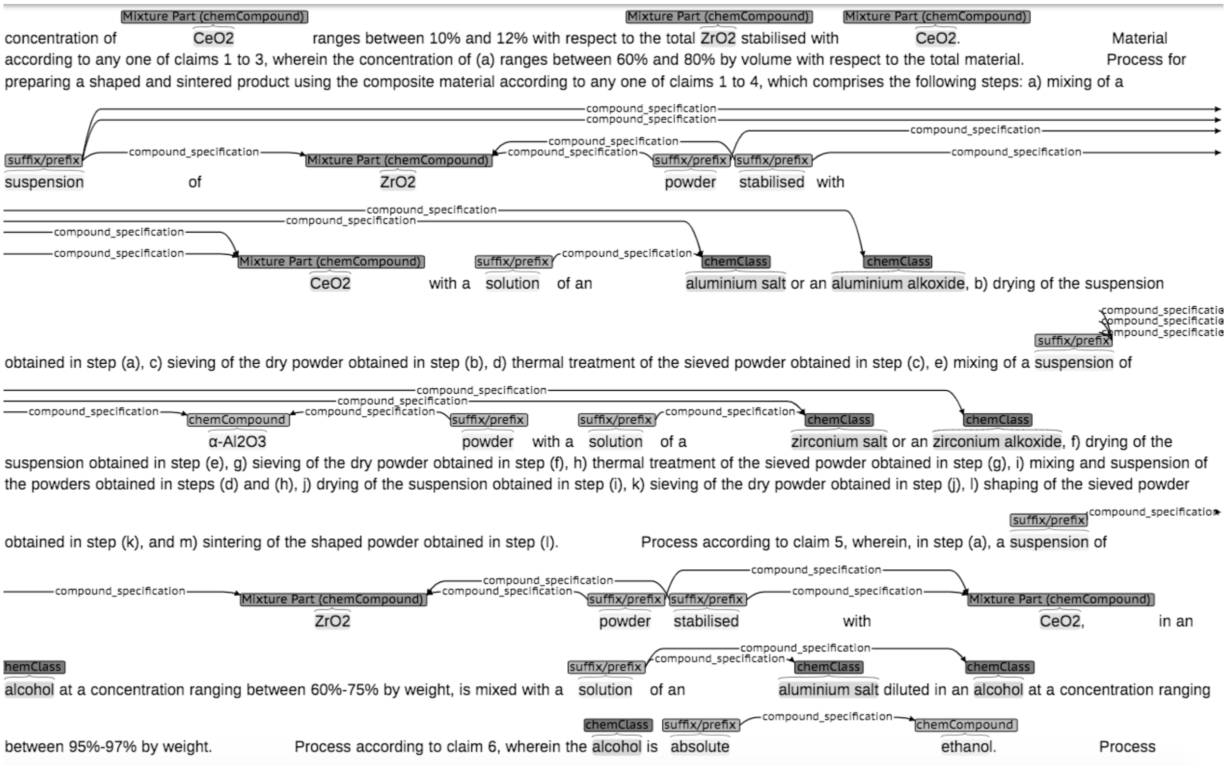

concentration of CeO2 ranges between 10% and 12% with respect to the total ZrO2 stabilised with CeO2. Material
according to any one of claims 1 to 3, wherein the concentration of (a) ranges between 60% and 80% by volume with respect to the total material. Process for
preparing a shaped and sintered product using the composite material according to any one of claims 1 to 4, which comprises the following steps: a) mixing of a
suspension of ZrO2 powder stabilised with
CeO2 with a solution of an aluminium salt or an aluminium alkoxide, b) drying of the suspension
obtained in step (a), c) sieving of the dry powder obtained in step (b), d) thermal treatment of the sieved powder obtained in step (c), e) mixing of a suspension of
α-Al2O3 powder with a solution of a zirconium salt or an zirconium alkoxide, f) drying of the
suspension obtained in step (e), g) sieving of the dry powder obtained in step (f), h) thermal treatment of the sieved powder obtained in step (g), i) mixing and suspension of
the powders obtained in steps (d) and (h), j) drying of the suspension obtained in step (i), k) sieving of the dry powder obtained in step (j), l) shaping of the sieved powder
obtained in step (k), and m) sintering of the shaped powder obtained in step (l). Process according to claim 5, wherein, in step (a), a suspension of
ZrO2 powder stabilised with CeO2, in an
alcohol at a concentration ranging between 60%-75% by weight, is mixed with a solution of an aluminium salt diluted in an alcohol at a concentration ranging
between 95%-97% by weight. Process according to claim 6, wherein the alcohol is absolute ethanol. Process

Figure 3: Annotations in a patent snippet with the brat annotation tool.

The 11 snippets were also annotated by two Elsevier Subject Matter Experts (SMEs) who defined the guidelines. The SMEs had PhDs in chemistry and around 15 years of professional experience in the field. Any discrepancies between the annotations of the two SMEs were resolved in consensus discussions. The resulting annotations were used as a reference and compared to the annotations of each of the other annotators by inter-annotator agreement (IAA) scores. We used the F-score (harmonic mean of recall and precision) as a measure of IAA, similar to other studies [7, 43, 46]. Several review sessions were held to compare annotations and resolve inconsistencies, and the annotation guideline was updated for clarity if needed. For each annotator, training continued until the IAA between the annotator and the SMEs was at least 85%.

After successful completion of the training, the remaining 120 snippets of the corpus were distributed between the annotators. Each snippet was annotated by three annotators, after which the annotations were harmonized. The harmonization was done for each entity as follows: if at least two annotators agreed on the entity boundaries and the entity type, that annotation was added to the gold-standard set, otherwise an SME adjudicated the disagreement.

Relevancy annotation process

The same training set of 11 snippets was also annotated for relevant compounds by the annotators and the SMEs. They were provided with the reference annotations of the chemical entities, and had to indicate whether the annotations were relevant or not. For every snippet, we also delivered the corresponding full patent text to the annotators and the SMEs. This allowed them to determine relevance based on the complete document which included title, abstract, description, and claims. The relevancy annotations of the annotators and SMEs were compared, and questions were resolved.

After training, the 120 snippets of the chemical entity corpus created in the previous step were distributed between the annotators. Each snippet was annotated by five annotators. If more than three annotators annotated the chemical entity as relevant it was considered relevant. If three annotators annotated the chemical entity as relevant it was considered equivocal. If less than three annotators annotated the chemical entity as relevant it was considered irrelevant. The equivocal category was introduced since relevance determination is sometimes complex and judged differently by different experts (as relevance is decided based on the full text). To capture this complexity, we did not try to resolve ambiguity by enforcing a decision by the SMEs. As per the guidelines, relevance is document-based. As a result, if a compound is considered relevant at one occurrence in the snippet, it is marked automatically relevant at any other occurrence. Finally, the annotators were also asked to annotate any spelling errors. This annotation can be helpful for improvement of chemical entity recognition systems. As spelling errors can be hard to detect, we decided to accept each spelling-error annotation, irrespective of the number of annotators that made that annotation. The corpus was divided into a development and test set consisting of 50 and 70 snippets, respectively.

Chemical Entity Recognition

We focused on non-statistical approaches for chemical entity recognition as we wanted to associate a chemical structure to extracted chemical compounds. A dictionary-based approach was used in combination with a morphology-based approach to identify chemical entities. The structures of these compounds were produced, validated, and standardized using Reaxys Name Service [42]. Since the gold-standard annotations showed that only a small set of relevant entities are from compound class categories (see results), we decided to reduce our chemical entity recognition scope to the identification and classification of chemical compounds.

Name Service

The Reaxys system uses a name-to-structure toolkit (Reaxys Name Service [42]) and a set of standardization rules (e.g., eliminate hydrogen bounds when constructing structures) when new compounds are inserted into the database. In this study, the Name Service was used to convert names to structures and standardize those structures as well as the structures in different dictionaries based on the Reaxys standardization rules, and to validate the structures assigned to chemical compounds.

Chemical Entity Recognizers

An ensemble system was used for chemical entity recognition. First, we used Elsevier's Chemical Entity Recognition (CER) software [40]. CER identifies and tags chemical compounds and their physical properties (e.g., colour, melting point and boiling point) within a text document and converts extracted compounds into a chemical structure (using Name Service). In addition, CER also identifies chemical reactions and chemical properties within the patent. The software uses a combination of dictionary-based and morphology-based approaches to extract chemical compounds from patents. CER was loaded with a dictionary derived from the manually curated compounds in the Reaxys database. Similar to previous studies [27, 28], an exclusion list was used to filter out any noise (e.g., frequent compounds such as oxygen) from the extracted compounds. The morphology-based approach in CER identifies different elements within a compound and combines them to create the final compound only if it can validate the compound based on its structural chemistry (e.g., can two elements bind with each other in this manner). This validation is done on the structural level and through a set of pre-defined rules processed by the Name Service. CER cannot assign the extracted compounds to the different compound groups that are defined in the guidelines.

Second, we used and improved OCMiner [41] to identify chemical entities. OCMiner also uses a dictionary-based approach along with a morphology-based approach to extract chemical compounds. The dictionary used for OCMiner was generated from a compound database built from various publicly available sources such as PubChem [23], DrugBank [49], ChEMBL [50], among others [41]. The Name Service was used to standardize the compounds within these dictionaries based on the same standardization rules applied by CER and Reaxys. In comparison to CER, OCMiner has additional functionality, such as abbreviation expansion and spelling-error correction [41]. The software also has post-dictionary modules to identify systematic names. In a separate module built for this study, OCMiner cleans up the chemical entities identified by both CER and OCMiner (e.g., overlapping annotations and combination of simple annotations to complex entities) and assigns compounds to the different compound groups. Finally, OCMiner generates a confidence score for all recognized chemical entities extracted by CER or OCMiner.

Relevancy Classification

Relevance of a chemical compound is defined based on the context of the full patent. To identify the relevance of a unique entity in a snippet, the complete patent should be analysed for that entity. We therefore gathered statistical information for each unique entity (recognised in the snippet) from the whole patent text and used that information to classify the extracted entity. Relevancy classification was based on a continuous relevance score that can vary between zero (irrelevant) and one (relevant). We divided the corpus into a training and a test set to find the best threshold for relevancy classification. The training set was used along with the relevance score to define the best cut-off point for the relevancy classification. The results were then tested on the test set.

Relevance score

Several features derived from the full text are used to calculate the relevance score. These features are combined to calculate the relevance score. Based on the feature type, features are normalized between zero and one. These features include:

***A -** Compound frequency:* Frequency of the compound within the document. Usually compounds that occur frequently in a patent document are less relevant (due to the nature of patents), unless the compound is unique to the patent.

***B -** Compound section:* Occurrence of the compound within specific sections of a patent (e.g., title, claim). A compound in a claim section is more relevant than a compound in a description section of a patent. If a compound appears in multiple sections we prioritize it in the following order: Title, Abstract, Claim, and Description.

***C -** Compound length:* Length of the extracted term. We have noticed that longer names are more likely to be IUPAC names and hence have a higher chance of being relevant.

***D -** Surrounding characters:* Occurrence of the compound within special characters (e.g., "[" , "("). Examples are usually mentioned between special characters and they will be less relevant.

***E -** Compound section uniqueness:* Compound single occurrence within a section of the patent. If a compound is mentioned once in the claims and a few times in the description it has higher probability to be relevant than the other way around.

***F –** Compound with solvent:* If the compound contains solvents or laboratory chemicals, there is a higher chance of the compound being relevant.

***G –** Compound wide usage:* Presence of the compound in one of a number of predefined groups representing the frequency of compounds in a large set of chemistry patents. To create the groups, all chemical entities from a large set of patent documents (selection of chemical patents in 2015, excluding patents from the patent corpus) were extracted using OCMiner and ranked according to their frequency of occurrence. The resultant compound list was divided in 16 equally-sized groups. Note here that we are extending our calculation to data derived from a larger set of patents. If a compound is frequently mentioned in other patents, then there is a lower probability of it being relevant.

Performance evaluation

The performance of the system against the gold-standard annotations was evaluated using recall, precision and F-score, given the number of true positives (TP), false positives (FP), and false negatives (FN). For the entity recognition task, TP represents the total number of correctly identified chemical entities by the system (based on starting and ending position of the entity in text), FP the number of entities wrongly identified by the system, and FN the number of entities that are missed by the system. Recall, precision and F-score metrics are calculated as follows: recall = TP / (TP + FN), precision = TP / (TP + FP), F-score = 2*precision*recall / (precision + recall).

For the relevancy classification task, TP, FP, and FN are determined at the document level and only take into account the unique entities identified in each of the documents. TP represents the number of compounds correctly classified as relevant, FP the number of compounds wrongly classified as relevant by the system, and FN the number of relevant compounds missed by the system. The compounds in the corpus that were annotated as equivocal, were disregarded from relevancy calculation. This is done since true judgment cannot be made on the relevance of these compounds.

Results

Chemical entity annotation

The average IAA between the annotators on the 11 training documents initially was 72%, and reached 92% after two rounds of training. On the gold-standard set of 120 snippets, the average IAA between the annotators and the harmonized annotations was 87%. Table 1 provides the frequency of entities within the corpus. Overall, 18,789 chemical entities were annotated, of which 15,199 chemical compounds and 3,590 chemical classes. The majority of the annotations consisted of mono-component compounds (13,564). In addition, the corpus contains 1,848 relationships from chemical compound or classes to 628 suffix or prefixes annotations (a suffix or prefix can have a relationship with one or more chemical compounds or classes).

Table 1: Number of annotation in the gold-standard set.

Annotation type	Annotation Subtype	Annotation	Relevant	Equivocal	Irrelevant
Compound	Mono Component	13,564	883	362	12,319
	Mixture Part	1,010	0	0	1,010
	Prophetic	625	625	0	0
Classes	Chemical Class	1,848	249	30	1,569
	Biomolecule	1,039	0	0	1,039
	Markush	17	17	0	0
	Mixture	286	0	0	286
	Mixture Part	174	0	0	174
	Polymer	226	0	0	226
Total chemical entities		18,789	1,774	392	16,623
Additional annotation	Suffix & Prefix	628	-	-	-
	Relations	1,848	-	-	-

Relevancy Annotation

All 18,789 chemical entities were annotated for relevance (Table 1). Of the 15,199 compounds, 1,508 (9.9%) were considered relevant and 362 (2.4%) equivocal. Of the 3,590 chemical classes, 266 (7.4%) were relevant, while 30 (0.8%) were equivocal. Thus, the majority of entities were considered irrelevant (87.7% of the compounds and 91.8% of the classes).

Chemical Entity Recognition

The performance of the ensemble system on compound recognition is shown in Table 2 for different thresholds of the confidence score. On the development set, a threshold of 0.2 yielded the best F-score of 83.7% (precision 89.1%, recall 78.9%). For this threshold, the best result was also obtained on the test set (F-score 86.2%, precision 90.1%, recall 82.3%). Error analysis of the results indicated that the performance of the system may further be improved by better recognizing prophetic compounds, reactants, and products of synthesis procedures.

Table 2: Performance of the ensemble system on compound recognition for different confidence score thresholds.

Confidence score	**Development**			**Test**		
	Precision	**Recall**	**F-score**	**Precision**	**Recall**	**F-score**
0.0	88.5	79.3	83.6	86.5	82.3	84.3
0.1	88.6	79.1	83.6	89.1	82.3	85.6
0.2	89.1	78.9	**83.7**	90.1	82.3	**86.2**
0.3	89.1	78.6	83.5	90.1	81.6	85.7
0.4	89.1	78.4	83.4	90.1	81.5	85.6
0.5	89.1	78.4	83.4	90.1	81.5	85.6
0.6	89.1	78.4	83.4	90.1	81.3	85.5
0.7	87.2	60.6	71.5	90.7	69.4	78.6
0.8	82.0	36.2	50.3	96.2	39.8	56.3
0.9	100.0	0.1	0.2	96.4	0.8	1.7
1.0	100.0	0.1	0.2	97.2	0.8	1.7

Relevancy Classification

Figure 4 shows the performance of the relevance system for different relevance score thresholds on the training set. The best performance (in terms of F-score) was obtained for a relevance score threshold of 0.53, with a precision of 85%, a recall of 87%, and an F-score of 86%. For the same threshold, the performance on the test set was slightly lower with 81% precision and 82% recall, resulting in an F-score of 82%. Further investigation into the compounds that the system classified as relevant, showed that 97% of these compounds were annotated as chemical compounds in the chemical entity corpus. Therefore, only 3% of the compounds classified by the system as relevant were not chemical entities.

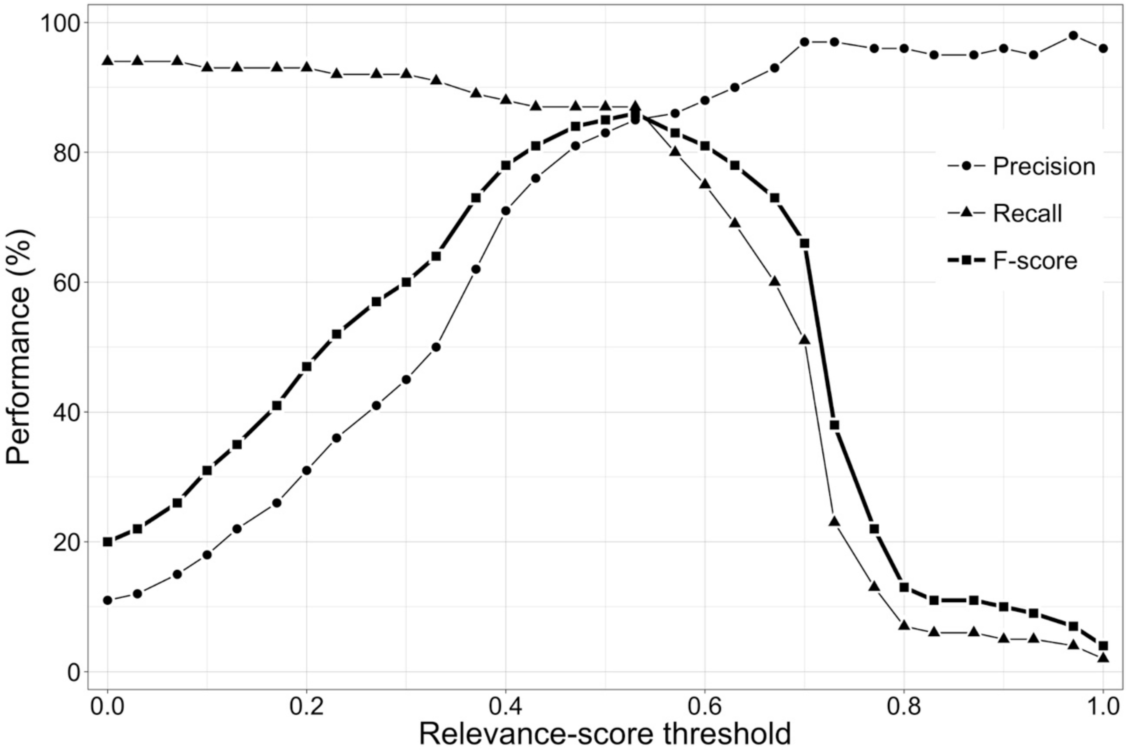


Figure 4: The performance of the relevance system based on precision, recall and F-score.

The relevancy classification is dependent on the performance of the chemical entity recognition system in two ways. First, only compounds that are found by the chemical entity recognizer can be classified as relevant. Second, the relevance-score features for a given chemical entity are based on the full patent text. The recognizer needs to correctly identify all occurrences of that entity in the full text. To assess the effect of the first dependency on the performance of the relevance system, we fed the gold-standard chemical entities as input to the relevance system (simulating a scenario where the chemical entity recognition system has a precision and recall of 100%). Apart from the patent snippet, all other parts of the full patent document were analyzed with the original system because gold-standard annotations were not available. When evaluated on our test set, the relevance classification system obtained 93% precision, 88% recall, and 91% F-score. Further investigation into these scores indicated that the system could have performed better if we could also eliminate the second dependency.

We also investigated the contribution of individual relevancy features to the performance of the relevancy classification system. For this we removed each feature in turn from the relevance score and adjusted the relevance-score threshold for optimal performance. Table 3 shows that the length of the compound is a major indicator of the relevance of the compound (10 percentage points added value). Additionally, the patent section in which the compound was found and compound wide usage in other publications are also good indicators of the relevance of the compound (around 5 percentage points added value respectively). The other features contribute between 1 to 2 percentage points to the relevancy classification performance.

Table 3: The added value of individual features based on “leave-one-out” methodology.

Setting	Threshold	Precision	Recall	F-Score	Added value
All features	0.53	84.8	86.8	85.8	-
A- Compound frequency	0.47	82.8	86.2	84.5	1.3
B - Compound section	0.40	95.5	70.0	80.8	5.0
C - Compound length	0.40	75.9	75.5	75.7	10.1
D - Surrounding characters	0.53	85.1	82.9	84.0	1.8
E – Compound section uniqueness	0.53	84.8	82.9	83.9	1.9
F – Compound with solvent	0.53	85.1	82.9	84.0	1.8
G – Compound wide usage	0.63	83.9	76.4	80.0	5.8

As can be seen from Table 3, leaving out a feature can affect the optimal value of the relevance-score threshold. Figure 5 shows the performance of the relevancy classification system as a function of the threshold value when a feature is left out.

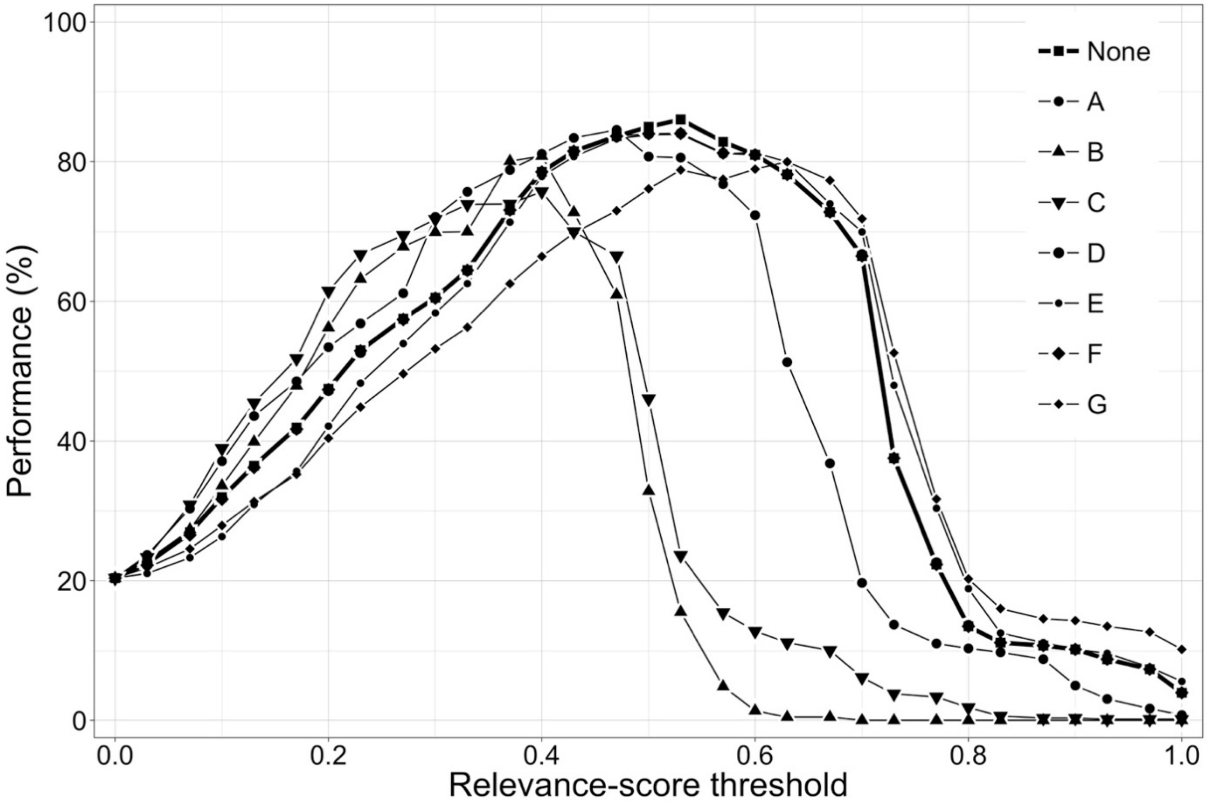


Figure 5: Performance of the relevancy classification system as a function of the relevance-score threshold when one of relevancy features A-G is removed (see Table 3 for feature legend).

Discussion and Conclusion

Extraction of chemical compounds from chemical-related patents has recently been studied, focussing on patent titles and abstracts [28, 31, 51] or full texts [3, 20, 21, 27]. The majority of these studies concentrated on identifying chemical compounds in text while disregarding the structures of the extracted compounds [31, 51]. Some have also looked at associating structures to extracted compounds (e.g., [3, 20, 21]), and have resulted in products and databases of chemical compounds in patents [3, 20, 21]. To our knowledge, ours is the first attempt to narrow down the focus to relevant compounds and their structures within a chemical patent. Relevance of a chemical compound is based on the context of the full patent document. Generally, a relevant compound is a compound that plays a major role in the patent (e.g., a product of a reaction that is mentioned in the Claim section of a patent). We have shown that these compounds are a small subset (less than 10 percent) of all compounds mentioned in the textual part of a patent.

We have presented a two-step approach to identify relevant compounds in patent documents: compound identification (first step) followed by compound classification (second step). This approach allows the use of the output of the first step for additional purposes (such as indexing chemical compounds mentioned in patents) but at the same time introduces dependencies. Obtaining high precision and recall values in the first step is essential for the success of the second step. Based on the findings of our previous studies [27, 28], we used an ensemble approach combining dictionary-based and morphology-based approaches to obtain high precision and recall. These approaches require a small annotated corpus [26, 33] and can provide a structural representation of the extracted compounds. Associating correct chemical structures to compounds is essential when extracting chemical compounds. To reduce the possibility of associating a compound with the wrong structure [34, 35] we regenerated the structures of compounds in different databases with our name to structure toolkit (Name Service) and standardized the structures based on standardization rules used for Reaxys [15].

The structures of non-systematic identifiers associated with a compound within Reaxys are manually drawn by excerpters and are later validated and standardized using Name Service. Adding such structures to the Name Service database allowed us to generate structures for non-systematic identifiers. We used the same toolkit with the same standardization functionalities to validate compounds extracted using the grammar-based approach. This ensures high quality and consistency of the extracted compounds.

To build the chemical entity recognition and relevancy classifier system, a patent corpus annotated with chemical entities and their relevance was needed. To our knowledge, such a corpus did not exist [7]. Currently available patent corpora are either limited to subsections of the patents, mostly title and abstract (e.g., the BioCreative corpus [36]), or had other limitations that prevented their use, such as different guideline definitions (focus on different entity types), harmonization approaches (manual using SMEs vs automation), low or unidentified IAA scores, and limited scope of coverage (only one chemical IPC class or one section of a document) [7]. We developed the corpus in two steps. First, we constructed a chemical entity corpus using random stratified sampling for content selection and manual harmonization to ensure high quality. Later we extended this corpus with relevancy annotation. We took into account the inherent difficulty of classifying relevance of some compounds by introducing “equivocal” as a classification in the corpus. Chemical compounds identified as equivocal can be classified as both relevant and irrelevant. The system can assign relevant or irrelevant for compounds extracted in this area. Any compound identified as equivocal was disregarded from our evaluation. Using five annotators for relevancy annotation, we showed that the equivocal label is only limited to about 2% of the compounds.

Normalized patent documents were used to develop the corpus and the system. Any change in the normalization approach will lead to changes to the corpus and might result in a need for retraining the system. We reduced this dependency by finalizing the normalization before developing the corpus and the software. We also introduced a one-to-one mapping between the original patent document and the normalized patent document to allow possible changes to the corpus with limited efforts. The relevancy classification system has lower dependency to the normalization step as its performance is calculated on unique mentions of compounds within a patent. The dependency to the normalization step relies on the quality of the patent source file. Digital patents (e.g., from EPO [11] or USPTO [12]) have a higher quality than OCR patents (e.g., from WIPO [13]). Therefore, the system is more dependable on the normalization when dealing with OCR patents.

The chemical entity recognition software showed a precision of 90.1% and a recall of 82.3% for compound recognition on EPO patents. State-of-the-art statistical systems (tested on patent title and abstract) have obtained higher recall (precision of 87.5% and recall of 91.3%) [31]. These systems do not generate structures for the identified chemical compounds. Error analysis of our system indicated that the loss in recall in our system is mainly due to the fact that reactants and products of synthesis procedures are not recognized, and prophetic compounds are missed. Identification of prophetic compounds may be improved by taking into account trigger phrases (e.g., “The compound of claim is:”, “A compound selected from”), or negative triggers for these compounds (e.g. “catalysts”).

Our current process only investigates the identification of relevant compounds in the textual part of non-OCR patents. Expanding this approach to chemical classes (such as Markush) can further improve the software. A large proportion of relevant compound information is only available through scaffolds, pictures, and tables. Successful identification of these compounds can result in a higher coverage. Since 2001, some patent offices including the USPTO [12] are requesting applicants to submit chemical structures and reactions (as MDL Molfiles or ChemDraw CDX files [30]) when submitting their patent applications. Note that in many cases these are not drawn by authors or chemists, and are presented usually with defects in the connection table of chemical structure). This can be a good starting point for future research.

We have successfully managed to identify relevant compounds in chemical-related patents. The resulting relevant compounds can be used to predict key compounds within a patent [9, 39, 52, 53]. In future research, we want to extend this work to chemical classes, increase the coverage by dealing with OCR patents (that contain many spelling errors), and utilize data from tables, scaffolds, and images.

References

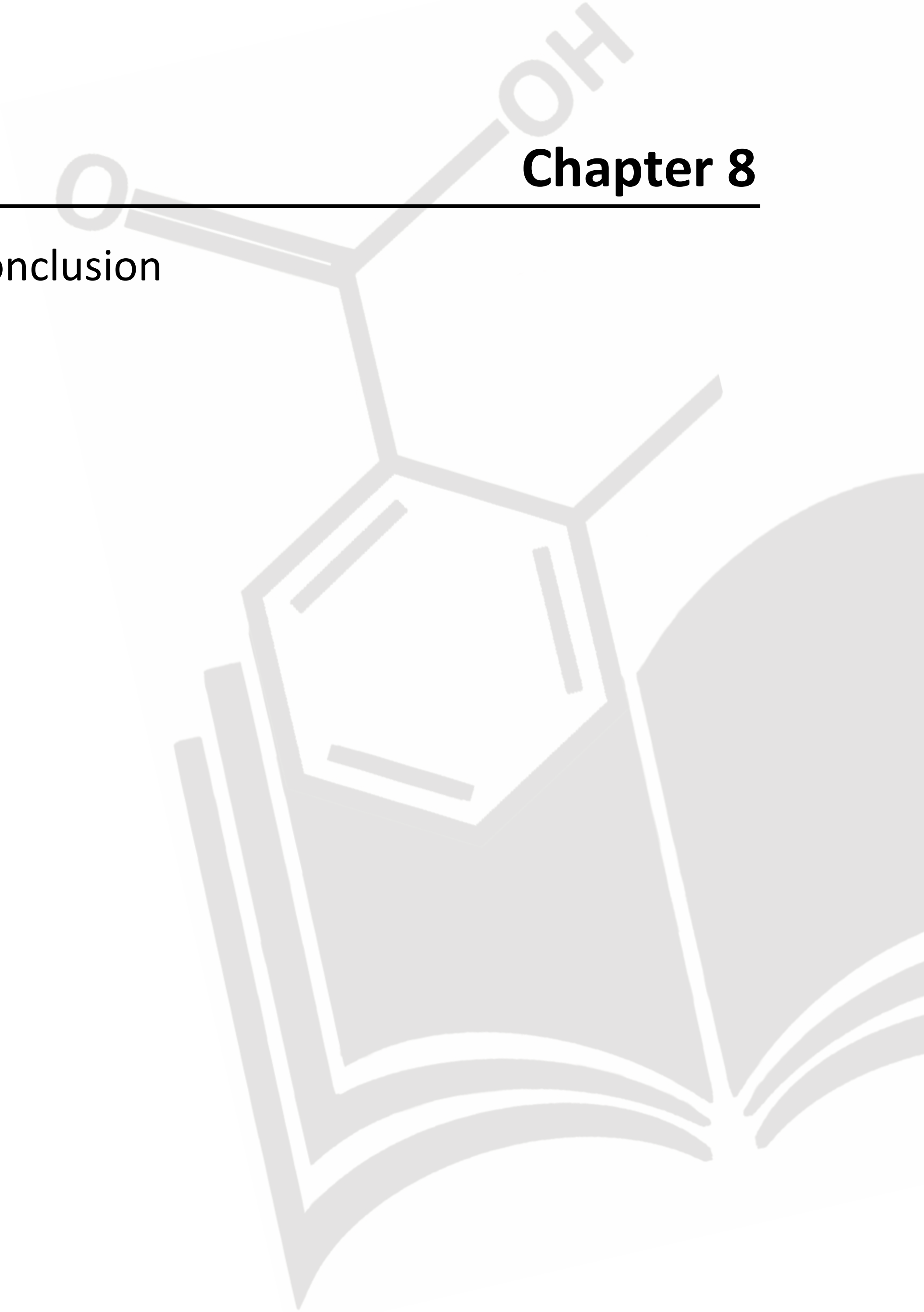
1. Muresan S, Petrov P, Southan C, Kjellberg MJ, Kogej T, Tyrchan C, Varkonyi P, Xie PH: **Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data.** *Drug Discov Today* 2011, **16:**1019–1030.
2. Southan C, Boppana K, Jagarlapudi SA, Muresan S: **Analysis of in vitro bioactivity data extracted from drug discovery literature and patents: Ranking 1654 human protein targets by assayed compounds and molecular scaffolds.** *J Cheminform* 2011, **3:** 14.
3. Papadatos G, Davies M, Dedman N et al: **SureChEMBL: a large-scale, chemically annotated patent document database.** *Nucleic Acids Res.* 2016, **44:**D1220–D1228.
4. Krallinger M, Rabal O, Lourenço A, et al: **Information Retrieval and Text Mining Technologies for Chemistry.** *Chem Rev* 2017, **117 (12):**7673–7761.
5. Senger S, Bartek L, Papadatos G, Gaulton A: **Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents.** *J Cheminform* 2015, **7:**49.
6. Bregonje M: **Patents: A unique source for scientific technical information in chemistry related industry?** *World Pat Inf* 2015, **27:**309–315.
7. Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA: **Annotated Chemical Patent Corpus: A Gold Standard for Text Mining.** *PloS one* 2014, **9:**e107477.
8. Asche G: **"80% of technical information found only in patents" – Is there proof of this ?** *World Pat Inf* 2017, **48:**16–28.
9. Tyrchan C, Boström J, Giordanetto F, Winter J, Muresan S: **Exploiting Structural Information in Patent Specifications for Key Compound Prediction.** *J Chem Inf Model* 2012, **52:**1480-1489.
10. Benson CL, Magee CL: **Quantitative determination of technological improvement from patent data.** *PLoS One* 2015, **10(4):**e0121635.
11. *European Patent Office.* http://www.epo.org.
12. *United States Patent and Trademark Office.* http://www.uspto.gov/.
13. *World Intellectual Property Organization.* http://www.wipo.int/.
14. Zimmermann M, Fluck J, Thi Le TB, et al: **Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology.** *Curr. Top. Med. Chem.* 2005, **5:**785–796.
15. *Reaxys.* https://www.reaxys.com.
16. Lawson AJ, Swienty-Busch J, Géoui T, Evans D: ***The Making of Reaxys-Towards Unobstructed Access to Relevant Chemistry Information.*** *The Future of the History of Chemical Information* 2014, 127–148
17. *SciFinder*. https://scifinder.cas.org/scifinder/.
18. *Thomson Reuters Pharma.* http://lifesciences.thomsonreuters.com/.
19. Heifets A, Jurisica I: **SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents.** *Nucleic Acids Res* 2012, **40:**D428-33.
20. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: **Chemical entities of biological interest: an update.** *Nucleic Acids Res* 2010, **38:**D249–D254.
21. *IBM press release: IBM Contributes Data to the National Institutes of Health to Speed Drug Discovery and Cancer Research Innovation*. http://www-03.ibm.com/press/us/en/pressrelease/36180.wss.

22. *Unleashing over a million reactions into the wild – NextMove Software.* https://nextmovesoftware.com/blog/2014/02/27/unleashing-over-a-million-reactions-into-the-wild/.
23. Kim S, Thiessen PA, Bolton EE et al: **PubChem Substance and Compound databases.** *Nucleic Acids Res.* 2016, **44:**D1202–D1213.
24. *Japan Patent Office.* https://www.jpo.go.jp/.
25. Valko AT, Johnson AP: **CLiDE Pro: The Latest Generation of CLiDE, a Tool for Optical Chemical Structure Recognition.** *J Chem Inf Model* 2009, **49:**780–787.
26. Vazquez M, Krallinger M, Leitner F, Valencia A: **Text mining for drugs and chemical compounds: methods, tools and applications.** *Molecular Informatics* 2011, **30:**506–519.
27. Akhondi SA, Hettne KM, van der Horst E et al: **Recognition of chemical entities: combining dictionary-based and grammar-based approaches.** *J. Cheminform.* 2015, **7:**S10.
28. Akhondi SA, Pons E, Afzal Z, van Haagen H, Becker BF, Hettne KM, van Mulligen EM, Kors JA: **Chemical entity recognition in patents by combining dictionary-based and statistical approaches.** *Database* 2016, **2016:**baw061.
29. Tseng Y-H, Lin C-J, Lin Y-I: **Text mining techniques for patent analysis.** *Inf Process Manag* 2007, **43**:1216-1247.
30. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, J L: **Description of several chemical structure file formats used by computer programs developed at molecular design limited.** *J Chem Inf Comput Sci* 1992, 244-255.
31. Krallinger M, Rabal O, Lourenc OA. et al: **Overview of the CHEMDNER patents task.** In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* 2015. pp. 63–75.
32. Krallinger M, Leitner F, Rabal O et al: **CHEMDNER: the drugs and chemical names extraction challenge**. *J. Cheminform.* 2015, **7:**S1.
33. Eltyeb S, Salim N: **Chemical named entities recognition: a review on approaches and applications.** *J Cheminform* 2014, **6:**1-12.
34. Akhondi SA, Kors JA, Muresan S: **Consistency of systematic chemical identifiers within and between small-molecule databases.** *J Cheminform* 2012, **4:**35.
35. Akhondi SA, Muresan S, Williams AJ et al: **Ambiguity of non-systematic chemical identifiers within and between small-molecule databases.** *J. Cheminform.* 2015, **7:**1–10.
36. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe DM, Sayle RA, Batista-Navarro RT, Rak R, Huber T, Rocktaschel T, Matos S, Campos D, Tang B, Xu H, Munkhdalai T, Ryu KH, Ramanan SV, Nathan S, Zitnik S, Bajec M, Weber L, Irmer M, Akhondi SA, Kors JA, Xu S, An X, Sikdar UK, Ekbal A, Yoshioka M, Dieb TM, Choi M, Verspoor K, Khabsa M, Giles CL, Liu H, Ravikumar KE, Lamurias A, Couto FM, Dai H, Tsai RT, Ata C, Can T, Usie A, Alves R, Segura-Bedmar I, Martinez P, Oryzabal J, Valencia A: **The CHEMDNER corpus of chemicals and drugs and its annotation principles.** *J Cheminform* 2015, **7(Suppl 1):**S2.
37. Jessop DM, Adams SE, and Murray-Rust P: **Mining chemical information from open patents.** *J. Cheminform.* 2011, **3:**40.
38. Ede M, Endacott J, Harper M, Rees D: **Indexing chemical structures: Exemplified compound indexing in patents by the vendors Thomson Reuters, Chemical Abstracts and Elsevier – A comparative study by the Patent Documentation Group (PDG).** *World Pat Inf* 2016, **44:**48–52.
39. Hattori K, Wakabayashi H, Tamaki K: **Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone.** *J Chem Inf Model* 2008, **48:**135–142.

40. Lawson A, Roller S, Grotz H, Wisniewski J: **Method and software for extracting chemical data.** *US7933763*
41. Irmer M, Weber L, Böhme T, Puschel A, Bobach C, Laube U: **OCMiner for Patents. Extracting Chemical Information from Patent Texts.** In: *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop* 2015. pp.119-123
42. Roller S: **Using Reaxys for Searching Chemistry in Patents.** https://new.reaxys.com/
43. Kolárik C, Klinger R, Friedrich CM, Hofmann-Apitius M, Fluck J. **Chemical names: terminological resources and corpora annotation**; 2008. *Workshop on Building and evaluating resources for biomedical text mining*.
44. Kulick S, Bies A, Liberman M, Mandel M, McDonald R, et al. **Integrated annotation for biomedical information extraction;** 2004. *Proc. of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* pp. 61-68.
45. Kim J-D, Ohta T, Tateisi Y, Tsujii J: **GENIA corpus-a semantically annotated corpus for bio-textmining.** *Bioinformatics* 2003, **19:**i180–i182.
46. 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. Morristown, NJ, USA*: Association for Computational Linguistics 2007. pp. 57-64.
47. *Brat rapid annotation tool.* http://brat.nlplab.org/.
48. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, et al. **BRAT: a web-based tool for NLP-assisted text annotation**; 2012. ***Association for Computational Linguistics.*** pp. 102-107.
49. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, 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:**D1091-1097.
50. Bento AP, Gaulton A, Hersey A et al: **The ChEMBL bioactivity database: an update.** *Nucleic Acids Res.* 2014, **42:**D1083–D1090.
51. Pérez-Pérez M, Pérez-Rodríguez G, Rabal O, Vazquez M,Oyarzabal J, Fdez-Riverola F,Valencia A, Krallinger M, Anália Lourenço A: **The Markyt visualisation, prediction and benchmark platform for chemical and gene entity recognition at at BioCreative/CHEMDNER challenge.** *Database* 2016, **2016:**baw120.
52. Lepp Z, Huang C, Okada T: **Finding Key Members in Compound Libraries by Analyzing Networks of Molecules Assembled by Structural Similarity.** *J Chem Inf Model* 2009, **49:**2429–2443.
53. Kettle JG, Ward RA, Griffen E: **Data-mining patent literature for novel chemical reagents for use in medicinal chemistry design**. *Medchemcomm* 2010, **1:**331.

Chapter 8

Discussion and Conclusion



The chemical domain has seen a massive increase in the number of databases and scientific literature in the past decade. Analyzing these data can provide understanding of compound prior art, novelty checking, validation of biological assays, and identification of new starting points for chemical exploration. Applying natural language processing (NLP) techniques such as text-mining can significantly simplify these analyses. In this thesis, we investigate text-mining for chemical identifiers in scientific articles and patents. This chapter discusses our findings. The chapter is divided into the following sections: quality of chemical databases, corpora development for chemistry, chemical text mining, and detection of relevant compounds in patents. The chapter ends by providing concluding remarks and discussing possible future work.

Quality of Chemical Databases

In Chapter 1, we noted that the quality of chemical databases has been debated based on qualitative studies [1, 2]. High-quality chemical databases are essential for chemical research and text-mining. We assessed the quality of chemical databases using a quantitative approach (Chapter 2 and Chapter 3). For this we evaluated consistency and ambiguity of chemical identifiers within and across chemical databases. We used MOL files (a digital representation of a chemical structure) as the foundation of a chemical record in a chemical database and analyzed if all systematic identifiers assigned to this record represent the same structure as defined in the MOL file (consistency within database). We also investigated the ambiguity of non-systematic identifiers within and across databases based on their structural representation (available within the MOL file).

Our results indicated that considerable inconsistency exists for systematic identifiers within and across chemical databases. We showed that non-systematic identifiers are very ambiguous across chemical databases. We also showed that standardization of chemical compounds prior to their inclusion in databases can improve the consistency and reduce the ambiguity.

To improve the quality of chemical databases we can regenerate systematic identifiers based on the MOL files available within the chemical databases (using well-defined standardization rules). Integration of databases also improves if the same set of standardization rules are applied to the MOL files prior to integration. The importance of chemical structures makes it crucial not to integrate data based on the textual similarities of the identifiers.

Improving the quality of non-systematic identifiers is more challenging in chemical databases. This includes considerable manual curation efforts. Nevertheless, identifying ambiguous identifiers within and across databases (using the structural representation of a compound) can help pinpoint problematic identifiers and guide curation efforts. Manual curation steps can be further simplified by using approaches such as crowd sourcing (an approach applied by ChemSpider [3]). Applying rule-based approaches such as voting schemes can help predict the correct structures based on how the non-systematic identifier is linked to a structure in different databases. It is also essential to apply a well-defined standardization method prior to registering a chemical compound structure of a non-systematic identifier within a database.

The standardization of chemical compound structures depends on the use case. For example, the importance of the presence of stereochemistry information varies between the fields of chemistry research. Therefore, we find it crucial for chemical databases to clearly describe their procedures of name-to-structure conversion and standardization.

Finally, it is important to note that chemical name-to-structure rules have limitations. Some of these rules are ambiguous and may result in similar systematic identifiers for different compounds. For this reason, updates are made to these rules through time. These updates may result in different identifiers that need to be updated within a chemical database. This adds an extra complexity also for linking data points between databases (if focus is on the textual representation). Name-to-structure toolkits are designed based on these rules but the implementation and methodology may vary. These variations and possible updates should be considered while using the toolkits. To overcome these limitations standardized structures (represented in MOL files) should be used for integration of databases. Any systematic identifier can be recreated based on the MOL file at any point of time.

Annotated Chemical Corpora

The availability of an annotated corpus is essential for training and evaluating text-mining systems. The performance of text-mining systems is sensitive to the corpus used for training them. A study by Habibi et al. [4] reported that applying a chemical text-mining system trained on scientific articles directly to patent abstracts resulted on average in about 10 percentage points lower performance. When Habibi et al. applied the same study on patent full text using the corpus developed in Chapter 4 they noticed that the performance of the systems trained on journal titles and abstracts dropped about 18 percentage points on patent full text. This study and many others [5] pinpoint the importance of availability of high-quality representative corpora.

In our study in Chapter 4, we annotated a full-text chemical patent corpus. The corpus consists of 200 patent documents. Multiple annotators were used to annotate the corpus. These annotators were in various locations (across Europe and India). The availability of a web-based annotation tool (brat [6]) greatly eased this step.

Developing a well-defined annotation guideline is a crucial step in creating a corpus. These guidelines should be strictly followed by the annotators. Considerable size of the patent document and the corresponding guidelines can make the annotation process challenging for annotators. Automatically generated pre-annotations were used to help speed up the annotation process. Annotators had to accept, modify or add additional annotations to the pre-annotations. Use of pre-annotations eases the annotation process only if the system that provides the pre-annotations has a good performance. If the performance of the system is low more time is spent on modifying the pre-annotations.

In our study in Chapter 4, we noticed that annotators often disagreed on the boundaries of annotated terms. In the development of the chemical relevant full-text patent corpus described in Chapter 7, we tried to mitigate this limitation by creating a small training corpus with the help of chemical experts. We then used this corpus to train the annotators using an extensive guideline. We only moved to corpus development after we made sure that our annotators had a high inter-annotator agreement (IAA). If needed the guidelines were also updated to resolve ambiguity. We also re-evaluated the annotators' IAA during the project. This was done to make sure that the quality of the corpus did not drop due to the extensiveness of the task. In case of a drop in IAA, annotators were notified and if needed guidelines were updated. In Chapter 4, the IAA varied between 64% to 95% F-score depending on the annotated entities and annotators. The training step applied in Chapter 7 allowed us to reach a consistent IAA of about 87% F-score.

Some patent offices such as the World Intellectual Property Organization only provide their documents in OCR format. In the patent corpus described in Chapter 4, errors that are introduced by OCR, were also annotated. Text-mining systems can use this information to further improve their performance. For instance, Lowe et al. [7] used the corpus developed in Chapter 4 to improve the identification of chemical entities in patents by resolving spelling errors.

Chemical Text Mining

The vast amount of chemical-related literature makes it essential to use text-mining approaches to extract information from text. In Chapter 5 and Chapter 6, we applied a variety of chemical named-entity methodologies (dictionary-based, morphology-based, and statistical-based) to identify chemical compounds in journal articles (titles and abstracts) and patents (abstracts). For this we took advantage of the BioCreative community challenges [8]. The outcome of these studies was used to develop a text-mining system applied on patent full text (described in Chapter 7).

Our studies showed several challenges for text-mining in the chemical domain. First, tokenization of the text was particularly cumbersome as chemical identifiers may contain hyphens, parentheses, brackets, dashes, dots, and commas. Second, non-systematic identifiers appearing in the text contained a range of ambiguous terms (due to chemical acronyms, abbreviations, and trivial names). Third, the correct identification of chemical boundaries was difficult in the chemical documents. This is mostly due to the presence of systematic identifiers in the text (containing punctuations within the identifier). This can result in recognizing partial mentions or breaking long chemicals into multiple mentions. Finally, there were spelling mistakes in chemical identifiers. This can be due to OCR errors or mistakes made by the author when writing long systematic names (in some cases in patents deliberate mistakes are made to further hide the chemical compound in the text). Addressing spelling mistakes is most challenging in systematic identifiers. These terms are long and result in a chemical structure. In most cases spelling mistakes may be assumed present if a term fails to translate to a structure using name-to-structure toolkits. Modification of the identifier (to solve the spelling error) may appear to have solved the problem (the name can be translated after modification) but may result in a wrong structure. Analyzing the surrounding context such as reactions or compound characteristics or properties can help evaluate the modified compound based on the chemical structure characteristics.

The results of our study and of the BioCreative challenges illustrate that among the three text-mining approaches for named entity recognition, statistical-based approaches (mostly using conditional random fields (CRFs)) perform better than dictionary-based or grammar-based approaches.

We applied dictionary-based approaches on journal articles and patents. The dictionaries were obtained from well-known chemical databases. Our study illustrates a substantial difference between the coverage of these databases. The best combination of these databases was used for the dictionary-based approach. Our study shows that, in journal articles, dictionaries yield a maximum recall of 60% of the chemical compounds. In patents this figure stands at 50%. The lower coverage in patents is due to the large presence of systematic names in patents. Case sensitive or insensitive matching of dictionary terms did not influence the results. We also found that our dictionaries contain common English terms, such as “Result” (also the brand name for a drug). Removing these terms using exclusion lists improved the performance (F-score) up to five percentage points.

We also tested grammar-based systems on journal and patents. These systems have a lower recall compared to dictionary-based approaches (around 10 percentage points less). The main advantage of grammar-based approaches is the ability to identify systematic identifiers. We have shown that the addition of this approach to dictionary-based or statistical-based approaches improves the final performance of the systems.

Although statistical-based approaches generally perform better than the other approaches (also shown in the BioCreative challenges), these systems need large training sets and are sensitive to the type of document they have been trained on. Therefore, a statistical-based approach trained on journals can perform significantly worse when applied on patents. Feature engineering is crucial in statistical-based system development. We have shown that in chemistry a set of features such as n-grams, prefixes and suffixes, word length, semantic affixes (such as trivial rings) and heuristic features can help identify chemical compounds (see Chapter 6).

Each of the approaches mentioned above have their advantages and drawbacks. Combining these approaches in an ensemble system might yield to a system with better performance. In our studies in Chapter 5 and Chapter 6, we have developed ensemble systems that perform better than the individual systems.

Identifying the structural representation of the recognized chemical compounds is essential in the chemistry domain. Among the three approaches, statistical-based approaches cannot provide structural representation for the identified compound identifiers. This is a big limitation for these approaches. For this reason, in Chapter 7, we disregarded the statistical-based approach when applying text-mining on patent full text.

Pre-processing of documents was necessary for all of our studies. In Chapter 6 and Chapter 7, we showed that fixing character encodings improves the chemical text-mining solution. In Chapter 7, we also normalized the patent documents based on the document structure.

Detection of Relevant Compounds in Patents

Patents are unique sources of information in the chemistry field. For example, a study showed that only 6% of bioactive compounds mentioned in patents are later defined in journal publications [9]. Patents are written as legal documents to protect one or a few compounds (the key compounds). These compounds are hidden within the context of the patent [10, 11]. In some cases, the key compound is not directly mentioned in the patent document [12]. Identification of key compounds in a patent is usually performed by medicinal chemists through the analyzes of available information (e.g. biological data, scale of reaction) [11]. To identify the key compounds, researchers study other compounds (mostly compounds relevant to the patent document) mentioned in the patent. Relevancy of a compound to a patent is based on the patent's context. A compound is relevant to a patent when it plays a major role within the patent application (e.g., starting material or a product in a reaction specified in the claim section).

In Chapter 7, we developed a system that can identify relevant compounds in patents. For this we built a patent corpus with relevancy classification of compounds in snippets selected from patent full text. We showed that relevant compounds (mentioned in text) constitute less than ten percent of the total number of compounds.

To develop this system, we used findings from all previous chapters. We used a well-defined manually created chemical database to ensure the quality of systematic and non-systematic identifiers (based on findings of Chapter 2 and Chapter 3). We developed a corpus with relevancy annotations (based on the process we investigated in chapter 4). We used an ensemble system with a dictionary-based and grammar-based approach to identify relevant compounds and provide structure information for the identified compounds (based on findings of Chapter 5 and Chapter 6).

Relevancy determination is difficult and sometimes judged differently by annotators. In our study, we introduced the compound relevancy class "equivocal" (meaning that for this compound annotators cannot agree on relevancy annotation). We found that only two percent of the compounds fall in this category.

A comparison of our study results with manually excerpted compounds from patents (in the Reaxys database) showed that relevant compounds also exist in chemical diagrams (images), tables, or are embedded in R-groups of Markush structures. Expanding research to extract this type of information can further ease the patent analysis process.

Special Attention Areas and Concluding Remarks

In this thesis, we provided a methodology to evaluate the quality of chemical databases based on the structural representation of chemical identifiers. We showed that considerable inconsistencies and ambiguities exist within and between chemical databases. Additionally, our methodology can be used to systematically integrate chemical identifiers from databases or from mined text into chemical databases.

We also developed corpora for chemical text-mining. While these corpora are essential for text-mining it is notably important to extend these corpora with structural representation of compounds. Our developed systems provide structural information for identified compounds, but these structures need to be validated using a corpus with structural representations. We strongly suggest that future community challenges also focus on this aspect of chemical research.

Furthermore, as chemical documents are very heterogeneous we also suggest the development of more focused corpora (e.g., focus on patents about organic or inorganic compounds).

We have studied the identification and classification of chemical identifiers within chemical-related text (i.e., journal documents and patents). While these systems can extract information with acceptable performance it is important to note that they should be further improved on OCR text. For this the detection of spelling errors needs considerable attention. Future studies could use surrounding information in the document to validate the spelling corrections. The annotated corpus developed in this thesis can be a good starting point for these investigations.

As we have shown in this thesis, an ensemble system performs better than individual text-mining approaches. To create an ensemble system, previously developed systems by academia or commercial venders can be used. Interoperability and scalability of available systems developed in the domain can ease the process of building ensemble systems. Community challenges such as BioCreative can play a major role in defining standards to cover these technical aspects. In recent challenges, BioCreative has been requiring teams to provide a web service for their system. While this is a good starting point, additional measures to improve scalability and interoperability are needed.

In this thesis, we have shown that relevant compounds can be identified in textual parts of patents. While this is a good starting point, future research should investigate the detection of relevant compounds in chemical diagrams (scaffolds and images) and tables within patents. Having the ability to perform analysis on images and tables provides us with the means to investigate Markush structures, identify R-groups within the structure (using the image), and find corresponding groups based on textual information from tables or corresponding text. Future research should also investigate approaches to find the key compounds among relevant compounds. Identification of key compounds has a significant value in fields such as drug discovery.

Finally, integrating chemical text-mining with other data types such as pharmacological, toxicological, and biological attributes extracted through means of text-mining can significantly expand the research domain. For example, within the Open PHACTS project, databases of compounds, targets, pathways, diseases, and tissues have been integrated to allow complex queries that enhance drug discovery [13]. It is essential for any development in integrating chemistry with other domains to pay considerable attention to the structural representation of chemical compounds.

References

1. Young D, Martin T, Venkatapathy R, Harten P: **Are the chemical structures in your QSAR correct?** *QSAR Comb Sci* 2008, **27:**1337–1345.
2. Williams AJ, Ekins S: **A quality alert and call for improved curation of public chemistry databases.** *Drug Discov Today* 2011, **16:**747–750.
3. Pence HE, Williams AJ: **ChemSpider: An Online Chemical Information Resource.** *J Chem Educ* 2010, **87**:1123-1124.
4. Habibi M, Wiegandt DL, Schmedding F, Leser U: **Recognizing chemicals in patents: a comparative analysis.** *J Cheminform* 2016, **8:**59.
5. Krallinger M, Rabal O, Lourenço A, et al: **Information Retrieval and Text Mining Technologies for Chemistry.** *Chem Rev* 2017, **117 (12):**7673–7761.
6. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, et al. **BRAT: a web-based tool for NLP-assisted text annotation**; 2012. ***Association for Computational Linguistics.*** pp. 102-107.
7. Lowe DM and Sayle RA: **LeadMine: a grammar and dictionary driven approach to entity recognition.** *J. Cheminform.* 2015, **7:**S5.
8. *BioCreative.* http://www.biocreative.org/.
9. Southan C, Várkonyi P: **Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds**. *J Cheminform.* 2009, **1(1):**10.
10. Tyrchan C, Boström J, Giordanetto F, Winter J, Muresan S: **Exploiting Structural Information in Patent Specifications for Key Compound Prediction.** *J Chem Inf Model* 2012, **52:**1480-1489.
11. Hattori K, Wakabayashi H, Tamaki K: **Predicting Key Example Compounds in Competitors' Patent Applications Using Structural Information Alone.** *J Chem Inf Model* 2008, **48**:135–142.
12. *gaming the system, loses Viagra patent in Canada.* https://arstechnica.com/tech-policy/2012/11/pfizer-caught-gaming-the-system-loses-viagra-patent-in-canada/.
13. Digles D, Zdrazil B, Neefs J-M, et al: **Open PHACTS computational protocols for in silico target validation of cellular phenotypic screens: knowing the knowns.** *Med Chem Commun* 2016, **7:**1237–1244.

Summary

Samenvatting

Acknowledgements

PhD Portfolio

List of Publications

About the Author

Summary of the main findings

The aim of this study was to use text mining for the identification of chemical identifiers in journal and patent documents. For this we addressed the lack of quality measurements for assessing the correctness of structural representation within and across chemical databases; lack of resources to build text-mining systems; lack of high performance systems to extract chemical compounds from journals and patents; and lack of automated systems to identify relevant compounds in patents.

The consistency and ambiguity of chemical identifiers was analyzed within and between small-molecule databases in **Chapter 2** and **Chapter 3**. In **Chapter 4** and **Chapter 7** we developed resources to enable the construction of chemical text-mining systems. In **Chapter 5** and **Chapter 6**, we used community challenges (BioCreative V and BioCreative VI) and their corresponding resources to identify mentions of chemical compounds in journal abstracts and patents. In **Chapter 7** we used our findings in previous chapters to extract chemical named entities from patent full text and to classify the relevancy of chemical compounds.

A summary of the main findings discussed in each chapter can be found below:

In **Chapter 2**, we analyzed the consistency of systematic identifiers within and between databases. The consistency within a database was analyzed by comparing the structural representation of a compound (based on the MOL file) and the compound structure derived automatically (through a set of predefined rules) from its assigned systematic identifier. The consistency across databases was calculated based on the cross-reference linkage of compounds available via databases (through MOL files). Our results show a considerable inconsistency in structural representation of systematic identifiers within (37.2%-98.5%) and between (25.8%-93.7%) widely known chemical databases (DrugBank, ChEBI, HMDB, PubChem, NPC). Standardizing the chemical compound improved this consistency to 84.8%-99.9% within and 47.6%-95.6% between databases. To improve this consistency, we proposed that databases should regenerate systematic identifiers starting from their MOL representation and apply well-defined and documented chemistry standardization rules to all compounds prior to integration.

In **Chapter 3**, we analyzed the ambiguity of non-systematic chemical identifiers within and between small-molecule databases (ChEBI, ChEMBL, ChemSpider, DrugBank, HMDB, NPC, PubChem). A non-systematic identifier is ambiguous if it has been assigned to multiple structural representations within or between databases. Our results show that the ambiguity of non-systematic identifiers within chemical databases is generally low (0.1%-15.2 %), but the ambiguity of non-systematic identifiers that are shared between databases, is high (17.7%–60.2%). Standardizing the chemical structures hardly reduced the ambiguity (average reduction of less than 0.5 percentage point) within databases. The effect of standardization was higher across databases (average reduction of 13.7 percentage points).

In **Chapter 4**, we developed an annotated chemical patent corpus for text-mining. This corpus was developed in collaboration with four commercial partners and sixteen annotators. To develop the corpus, we selected 200 full-text patents from the World Intellectual Property Organization, the United States Patent and Trademark Office, and the European Patent Office.

The patents were pre-annotated automatically. During a manual revision phase by the annotators, the pre-annotations were examined and potentially corrected or removed, while missing mentions were added. The annotations consist of chemicals in different subclasses (IUPAC, Generic, Trademark, Abbreviation, Formula, Registry Number, SMILES, CAS, InChI), diseases, targets, and modes of action. We also annotated spelling mistakes and spurious line breaks due to optical character recognition errors. From the 200 full-text patents, 47 were annotated by three annotators and later harmonized. The patent corpus contains 400,125 annotations for the full set and 36,537 annotations for the harmonized set.

In **Chapter 5**, we investigated named-entity recognition in titles and abstracts of scientific journals. For this we developed an ensemble system that combines dictionary-based and grammar-based approaches. The dictionary-based approach was used along with dictionaries derived from well-known chemical databases. Our analyses showed that our dictionaries only contain 60% of the compounds mentioned in journals. The ensemble system outperformed the individual systems that were considered. Application of this approach allowed us to provide structure representations of compounds for the recognized mentions. Our system could also rank the recognized compounds at the document level. The system provides a recall of 71% and a precision of 86%. Correct tokenization and identification of chemical formulas were the most challenging aspects in chemical named-entity recognition in journals.

In **Chapter 6**, we investigated the named-entity recognition in patent abstracts. For this we developed an ensemble system that combines dictionary-based and statistical-based approaches. The dictionary-based approach was used along with dictionaries derived from well-known chemical databases. The system had a recall of 86% and a precision of 85%. The difference in performance between the ensemble system and the statistical-based system was small. Our analyzes showed that dictionaries only contain 50% of the compounds mentioned in patents. Correct tokenization was one of the most challenging aspects in this approach. The limitation of the statistical-based approach was that chemical structures could not directly be defined for compounds recognized only by this approach (77% of the named entities).

Finally, in **Chapter 7**, we investigated the identification of relevant chemical compounds in patents. For this we used dictionary-based and grammar-based approaches to extract named entities from patent full text. This decision was made so that we could directly provide chemical structures for all recognized chemical compounds. We derived the dictionaries from a manually created high-quality database (Reaxys database). We also developed a chemical named-entity corpus with relevancy annotations for training and testing the system. We used the corpus to identify different features that can be used to classify the relevant compounds, which comprised only 10% of the total number of compounds within a patent (on average 2000 compounds in a patent). The system obtained a recall of 82% with precision of 90% for chemical named-entity recognition and a recall of 82% and a precision of 81% for relevancy classification.

Samenvatting van de belangrijkste bevindingen

Het doel van deze studie was het herkennen van chemische namen in wetenschappelijke artikelen en patenten door gebruik van tekstmining. Hiervoor hebben we de volgende problemen onderzocht: het ontbreken van kwaliteitsmaten om de correctheid van structurele representaties van chemische stoffen binnen en tussen databases te bepalen; het ontbreken van geannoteerde corpora voor de ontwikkeling van tekstminingsystemen; het ontbreken van nauwkeurige systemen om chemische stoffen in wetenschappelijke artikelen en patenten te herkennen; het ontbreken van geautomatiseerde systemen waarmee de meest relevante chemische stoffen in patenten bepaald kunnen worden.

In **hoofdstuk 2** en **hoofdstuk 3** zijn de consistentie en ambiguïteit van chemische namen binnen en tussen databases voor kleine moleculen onderzocht. In **hoofdstukken 4 en 7** zijn geannoteerde corpora ontwikkeld voor de constructie van chemische tekstminingsystemen. **Hoofdstukken 5 en 6** beschrijven onze ontwikkeling van tekstminingsystemen om chemische stoffen in samenvattingen van artikelen en patenten te herkennen in het kader van publieke competities (BioCreative V en BioCreative VI). In **hoofdstuk 7** worden de bevindingen uit voorgaande hoofdstukken gebruikt om chemische stoffen te herkennen in complete patenten, en om de relevantie van chemische stoffen te classificeren.

Hieronder worden de belangrijkste bevindingen uit elk hoofdstuk samengevat.

In **hoofdstuk 2** analyseerden we de consistentie van systematische chemische namen binnen en tussen databases. De consistentie binnen databases werd geanalyseerd door de structurele representatie van een chemische stof (gebaseerd op zijn MOL file) te vergelijken met de structuur die automatisch kan worden afgeleid van de toegekende systematische naam. De consistentie tussen databases werd bepaald op basis van kruisverwijzingen voor een chemische stof. Onze resultaten laten een aanzienlijke variatie zien in de consistentie van systematische namen binnen (37,2%-98,5%) en tussen (25,8%-93,7%) een aantal bekende chemische databases (DrugBank, ChEBI, HMDB, PubChem, NPC). Standaardisatie van de systematische namen verbeterde de consistentie tot 84,8%-99,0% binnen databases en tot 47,6%-95,6% tussen databases. Om de consistentie verder te verbeteren, stellen we voor om systematische namen op basis van MOL-representaties te genereren en goed-gedocumenteerde standaardisatieregels toe te passen voordat ze in een database worden opgenomen.

In **hoofdstuk 3** analyseerden we de ambiguïteit van niet-systematische chemische namen zowel binnen als tussen chemische databases (ChEBI, ChEMBL, ChemSpider, DrugBank, HMDB, NPC, PubChem). Een niet-systematische naam is ambigu als hij is toegewezen aan verschillende structurele representaties binnen of tussen databases. Onze resultaten tonen aan dat de ambiguïteit van niet-systematische namen binnen chemische databases over het algemeen laag is (0,1%-15,2%), maar tussen databases hoog (17,7%-60,2%). Standaardisatie van de chemische structuren verminderde de ambiguïteit binnen databases nauwelijks (gemiddeld minder dan 0,5 procentpunt). Het effect van de standaardisatie was hoger tussen databases (gemiddelde vermindering 13,7 procentpunt).

In **hoofdstuk 4** ontwikkelden we een geannoteerd corpus van chemische patenten voor tekstmining. Dit corpus werd ontwikkeld in samenwerking met vier commerciële partners en zestien annotatoren. De volledige teksten van 200 patenten van de World Intellectual Property Organization, het United States Patent and Trademark Office, en het European Patent Office werden geselecteerd. Eerst werden de patenten automatisch geannoteerd. Daarna werden deze annotaties bekeken door de annotatoren en waar nodig verbeterd of verwijderd, terwijl ontbrekende annotaties werden toegevoegd. De annotaties bestonden uit chemische stoffen in verschillende subklassen (IUPAC, Generic, Trademark, Abbreviation, Formula, Registry Number, SMILES, CAS, InChI), ziekten, aangrijpingspunten, en werkingsmechanismen. Ook spelfouten en problemen door foutieve optische karakterherkenning werden geannoteerd. Van de 200 volledige patenten werden er 47 door drie annotatoren onafhankelijk van elkaar geannoteerd en later geharmoniseerd. Het patent-corpus bevat 400.125 annotaties in de volledige set, en 36.537 annotaties in de geharmoniseerde set.

In **hoofdstuk 5** onderzochten we de herkenning van chemische namen in titels en samenvattingen van artikelen in wetenschappelijke tijdschriften. Hiervoor ontwikkelden we een ensemblesysteem dat methodes combineert die op chemische vocabulaires en op grammaticaregels gebaseerd zijn. De vocabulaires werden afgeleid van bekende chemische databases. Onze analyses laten zien dat de vocabulaires slechts 60% van de in tijdschriften vermelde chemische stoffen bevatten. Het ensemblesysteem presteerde beter dan de individuele systemen afzonderlijk. Toepassing van deze methode maakt het mogelijk structurele representaties voor de herkende chemische namen te verschaffen. Het systeem behaalde een sensitiviteit van 71% en een precisie van 86%. Correct tokeniseren en identificeren van chemische formules behoorden tot de meest uitdagende aspecten bij het herkennen van chemische termen in wetenschappelijke tijdschriften.

In **hoofdstuk 6** onderzochten we het herkennen van chemische namen in patentsamenvattingen. Hiervoor ontwikkelden we een ensemblesysteem dat een vocabulaire-gebaseerde methode combineert met een statistische methode. De vocabulaire was gebaseerd op chemische namen in bekende chemische databases. Het systeem herkende 86% van de chemische namen in de patenten, met een precisie van 85%. Het prestatieverschil tussen het ensemblesysteem en het statistische systeem was klein. Onze analyses lieten zien dat chemische vocabulaires slechts 50% van de chemische namen in patenten bevatten. Het correct tokeniseren was een van de meest uitdagende problemen. Een beperking van de statistische methode was dat structurele representaties niet direct konden worden bepaald voor de chemische stoffen die alleen door deze methode werden herkend (77% van de namen).

Tot slot hebben we in **hoofdstuk 7** de identificatie van de relevante chemische stoffen in patenten onderzocht. Hiervoor combineerden we vocabulaire- en grammatica-gebaseerde methodes om chemische namen in de volledige tekst van patenten te herkennen. De reden voor deze keuze van methodes was dat daarmee direct structurele representaties konden worden aangeleverd voor de herkende chemische stoffen. De vocabulaire was gebaseerd op een handmatig samengestelde database van hoge kwaliteit (Reaxys). Ook ontwikkelden we een corpus waarin de relevantie van chemische stoffen was geannoteerd om het systeem te trainen en te testen. We gebruikten dit corpus om verschillende variabelen te identificeren waarmee de relevantie van chemische stoffen in een patent geclassificeerd kan worden. Van het totale aantal chemische stoffen die in de patenten vermeld werden (gemiddeld 2000 stoffen in een patent), waren slechts 10% relevant. Het systeem had 82% sensitiviteit en 90% precisie voor het herkennen van chemische namen, en 82% sensitiviteit met 81% precisie voor de classificatie van relevantie.

ACKNOWLEDGEMENTS

It all started with a tap on the back, Sorel asking me, “do you want to do a PhD in the Netherlands? If so, get ready for an interview - and I need an answer soon.” I was surprised and went home to tell my wife and call my dad, asking them for advice.

I was sitting at my office in AstraZeneca in Mondal, Sweden when all of this happened. At the time, I was a bioinformatician at AstraZeneca. During my work, I got to know Sorel and through him understand a little bit of the Chemical world and all of its unique challenges. Now I was selected to do a collaboration PhD between AstraZeneca and Erasmus MC. The title was exploring chemical and biological name entity recognition in drug discovery. After some initial telephone interviews, Sorel and I flew together to Rotterdam where I made a presentation of my previous work and was formally accepted into the PhD.

It was the end of November 2011 when me and my wife put all we had in a rented truck and drove it down to the Netherlands. The drive was an amazing journey with a surprise storm hitting us at one point, but this was nothing compared to what was ahead of me - a new phase in my life had started. Through my PhD, I met a lot of nice people. The journey was a long one, took four years to finish, and an additional two to finalize. It had its ups and downs, along with some surprises, satisfactions and sometimes frustrations. All this could have not been possible without the help, support, and love of my family and the guidance and support of my colleagues and friends. Thank you all.

First of all, to my co-promotor and daily supervisor Dr. Jan A. Kors. Jan, this section has to start with you. Thank you for trusting me and providing me such a great experience. Your friendliness, your energy, your experience, your dedication, your attention to every detail, your amazing feedback, your trust, and your deep understanding of the subject is just amazing. I have learned a lot from you, most of all how to be critical on my own work, how to look at every result of an experiment and make sure it makes sense from different angles. Thanks for all the time you spent going through my work, always making it better and better. Jan, I always remember the car trip we had to Sweden to visit AstraZaneca. We spoke on the car and on the boat for hours and hours about everything, from weather, to study, to work, to politics. Funny thing is that we still had things to talk about. I would also like to thank you and your wonderful family for inviting us many times to your home and making sure we are not feeling alone in the Netherlands.

To my promoter, Prof. Dr. Johan van der Lei, many thanks for all your trust and support. Johan, it has always been a great joy to come and discuss my work, the progress and the next phases with you. I will always appreciate the opportunity I had to learn from your knowledge and experience. Your guidance has always directed us to the right path. Thank you for taking care of us in the Medical informatics department.

To my supervisor from AstraZeneca, Dr. Sorel Muresan. Sorel, thank you for defining this collaboration and then selecting and trusting me for it. Thank you for introducing me to Jan and for supporting us in different phases of the project. Your insight about the chemistry field is just amazing. You taught a non-chemist on how to think like a chemist. I still remember the day that you called me and told me you are moving from AstraZeneca but will try to stay in touch. That was probably one of the hardest days of my project.

I would also like to thank my inner doctoral committee, Prof. Dr. P.J. van der Sepk, Prof. Dr. B. Mons, and Dr. Rebholz-Schuhmann for their valuable time and encouraging feedbacks. I would also like to thank my other committee members Prof. Dr. W. Kraaij, Prof. Dr. G.W. Jenster, Dr. S. Muresan and Dr. M. Gregory for reviewing my work and providing feedbacks.

My colleagues in the BioSemantics group, Erik, Rein, Zubair, Ewoud, Bharat, Kang, Benus, Chinh, Herman, Wytze, Kristina, Solene, thank you for making work a pleasure. Erik, you were always there for us trying to guide and supervise us throughout the work. I remember the time you invited us to your house and showed us Nesselande lake. That kind of made me so want to move to that area (maybe that's why I did). Rein, you always had a smile on your face. Zubair, honestly, not only you are a great colleague, you are a great friend, an awesome companion, a cycling buddy, and the list can go on and on. I really enjoy and appreciate the endless discussions we have, while travel to work, at work or on the way back. I am happy that I have managed to convince you that you are not always right!!! Thank you for such a nice friendship. Ewoud, I remember the day you entered the department and the joy and laughter you brought to it ever since. Thanks for reminding me which server was named after what whisky, for filing my tax returns in Dutch and coming with a great idea on how to name Lagavulin. Bharat, I enjoyed working with you. Kang, to me you were by far the most hard-working guy I have ever seen, at the same time you always made the workplace fun, giving us hints regarding photography and various toolkits for cameras. Benus, I don't know how many times I ran into your room asking you to "get up" and go for coffee. It was always fun working with you. "Doctor Chinh", you're a great friend and a great company, especially over a few drinks. Herman, I have not seen a drummer doing NLP better than you. Wytze, you're so fun to talk to, I have always enjoyed your company. I especially want to appreciate your Dutch skills on translating the summary of this book - thanks man. Kristina, your research through the field paved the way for my research, thank you for all the collaboration work we did together. Solene, we did not get so much time to spend together but I always enjoyed your company and the discussions we had. David and Tiago, it was fun having you in our group for some time. We should definitely spend some more long nights in Seville. Osomeke, thank you for the amazing friendship.

My special thanks to all the staff: Desiree, Petra, Tineke, Carmen, Sander, Andreas, Solane, and Mees. Thank you for your help and support throughout my time at Erasmus MC. Desiree, you are by far the most organized person I ever known. I appreciate all the dedication you put into my work. Petra, what a great friend you have been to me and my family. Tineke, thanks for giving me support whenever needed. Sander, talking to you brings a smile to anyone's face. Andreas and Solane, thank you for supporting me on my endless requests.

I would also like to thank Ghazaleh Beyk for the amount of effort she put for the illustration and design of this dissertation. Ghazaleh, when I asked you if you would agree to do the design of my book, I was sure that you are going to do a great job, but I did not imagine that I had to choose between multiple great designs. Thank you for the dedication, time and effort you put into this work making sure that its delivered on time.

I would also like to thank my friend Shanmukha Sampath for the moral support, continues feedback and proofreading of my work. Sampath, it's always fun spending time with you. I'm sorry that I made you camp in Spain. You should move back to Europe. We can do it again.

I would also like to thank my coauthors, and those with whom I collaborated over the years. Thank you for your contribution.

Dr. Michelle Gregory, Dr. Marius Doornenbal, and Mark Sheehan, thank you for support and understanding during the final phase of my thesis. Without your push and support this work could have not finalize. Michelle, Marius, and Mark - you guys are just awesome, it's so much fun working with you. Michelle, I really appreciate your positive influence on my work, life, and family. Marius, you can officially call me a doctor now! Mark, appreciate your help in proofreading my work, I hope I finally got you the best reason to cycle to Rotterdam!

This research, as with any other achievement in my life, would not have been possible without the endless love, support and prayers of my family. My parents have always been a source of great inspiration for me. To my mother Zohreh Behjati, you raised me in a way where excellence is anticipated, and versatility encouraged. You always pushed me to go forward, asked me to aim for higher, and showed me the value of hard work. I hope I have made you proud. To my father Prof. Dr. Mehdi Akhondi, you always inspired me on what a human can achieve. Since my childhood, I have seen nothing but respect for you and your achievements. I always love to be fearless in life like you. Thanks for guiding me into the great field of Bioinformatics, pushing me to discover uncharted territories. Thanks for your constant guidance in my life, studies, and career. I hope I have also made you proud. To my brothers Saleh and Sadegh, being away from you guys was the hardest thing I have done in my life. I miss you all the time. Saleh, thanks for being always just a phone call away. Although we were far apart but your support has influenced many details of my life. Thank you. Sadegh, I love the way you customize your jokes and make me laugh. Every time I meet you or you call me, I am pumped up with energy. I am so blessed to have a brother like you.

Last but not least, I would like to express my deepest gratitude to my wife Bahareh Beyk. Bahar, your love and support has always been unconditional. I could have not achieved any of this work without you. You have stood by me shoulder to shoulder. The moment I told you if I should take the opportunity, you were the one saying I will be stupid not to, although this would result in you missing great opportunities in Sweden. You have made many sacrifices for me, and always pushed me and encouraged me to succeed. I cannot count the nights you stood up till morning so that I am not working alone. Thank you for being always there and bearing with me.

Finally, to my new joy in life, Lillian my beautiful daughter. You have made life so colorful for me and your mother since you arrived only a month ago. You are my biggest bundle of joy. I am so happy that you are here for my defense. I love you forever.

There are simply too many other people I would like to acknowledge. Please forgive me if I failed to mention your name. Thanks to you all.

Saber Akhondi, Rotterdam, 2018

PhD Portfolio

Name:	Saber Ahmad Akhondi
Promotor:	Prof.dr. J. van der Lei
Copromotor:	Dr.ir. J.A. Kors
Affiliation:	Erasmus University Medical Center
Department:	Medical Informatics

PhD Training

Biomedical and Scientific English Writing and Communication, Erasmus University Medical Center, Rotterdam

Scientific Presentations, Erasmus University Medical Center, Rotterdam

Integrity in Science, Erasmus University Medical Center, Rotterdam

Oral Presentations

CHEMPROT chemical protein relation extraction task: text mining of metabolic, gene regulation and drug-target interactions.

- BioCreative VI Challenge and Workshop, Bethesda, Maryland, United States 2017

Patent mining: combining dictionary-based and machine-learning approaches.

- BioCreative V Challenge and Workshop, Sevilla, Spain 2015

A dictionary-and grammar-based chemical named entity recognizer.

- BioCreative IV Challenge and Workshop, Bethesda, Maryland, United States, Spain 2013

Poster Presentations

Text mining of metabolic, gene regulation and drug-target interactions.

- BioCreative VI Challenge and Workshop, Bethesda, Maryland, United States 2017

Chemical entity recognition in patents by combining dictionary-based and statistical approaches.

- BioCreative V Challenge and Workshop, Sevilla, Spain 2015

Chemical-disease relation extraction using prior knowledge and textual information.

- BioCreative V Challenge and Workshop, Sevilla, Spain 2015

Recognition of chemical entities: combining dictionary-based and grammar-based approaches.

- BioCreative IV Challenge and Workshop, Bethesda, Maryland, United States, Spain 2013

Awards

- **First place –** eHealth Evaluation Lab – Clinical Named Entity Recognition Conference and Labs of the Evaluation Forum (CLEF), Toulouse, France 2015
- **First place –** eHealth Evaluation Lab – Clinical Named Entity Recognition Conference and Labs of the Evaluation Forum (CLEF), Évora, Portugal 2016
- **First place –** eHealth Evaluation Lab – ICD10 Coding of French Death Certificates Conference and Labs of the Evaluation Forum (CLEF), Évora, Portugal 2016
- **Second place –** CHEMDNER-Patents – Chemical passage detection and text classification (CDI), BioCreative V Challenge and Workshop, Sevilla, Spain 2015
- **Second place –** Chemical induced disease relation extraction (CID), BioCreative V Challenge and Workshop, Sevilla, Spain 2015

Scientific Contributions

BioCreative CHEMPROT Task organizer

- BioCreative VI Challenge and Workshop, Bethesda, Maryland, United States 2017

Scientific Advisory Board, IberEval BARR track

- IberEval Challenge and Workshop, Murcia, Spain 2017

LIST OF PUBLICATIONS

List of journal publications:

1. **Akhondi SA,** Rey H, Schwörer M, Maier M, Toomey J, Nau H, Ilchmann G, Sheehan M, Irmer M, Bobach C, Doornenbal M, Gregory M, Kors JA: *Automatic identification of relevant chemical compounds from patents.* J ChemInf. 2018 (submitted).

2. Krallinger M, Rabal O, **Akhondi SA,** Pérez MP, Santamaría J, Rodríguez GP, Tsatsaronis G, Intxaurrondo A, López JA, Nandal U, van Buel E, Chandrasekhar A, Rodenburg M, Laegreid A, Doornenbal M, Oyarzabal J, Lourenço A, Valencia A: *Evaluation of the BioCreative VI CHEMPROT chemical protein relation extraction task: text mining of metabolic, gene regulation and drug-target interactions.* Database, 2018 (submited).

3. **Akhondi SA,** Pons E, Afzal Z, van Haagen H, Becker BF, Hettne KM, van Mulligen EM, Kors JA: *Chemical entity recognition in patents by combining dictionary-based and statistical approaches.* Database 2016, 2016:baw061.

4. Pons E, Becker BFH, **Akhondi SA,** Afzal Z,van Mulligen EM Kors JA: *Extraction of chemical-induced diseases using prior knowledge and textual information.* Database 2016, 2016:baw046.

5. **Akhondi SA,** Muresan S, Williams AJ, Kors JA: *Ambiguity of non-systematic chemical identifiers within and between small-molecule databases.* J Cheminform 2015, 7:54.

6. Krallinger M, Rabal O, Leitner F, Vazquez M, Salgado D, Lu Z, Leaman R, Lu Y, Ji D, Lowe DM, Sayle RA, Batista-Navarro RT, Rak R, Huber T, Rocktaschel T, Matos S, Campos D, Tang B, Xu H, Munkhdalai T, Ryu KH, Ramanan SV, Nathan S, Zitnik S, Bajec M, Weber L, Irmer M, **Akhondi SA,** Kors JA, Xu S, An X, Sikdar UK, Ekbal A, Yoshioka M, Dieb TM, Choi M, Verspoor K, Khabsa M, Giles CL, Liu H, Ravikumar KE, Lamurias A, Couto FM, Dai H, Tsai RT, Ata C, Can T, Usie A, Alves R, Segura-Bedmar I, Martinez P, Oryzabal J, Valencia A: *The CHEMDNER corpus of chemicals and drugs and its annotation principles*. J Cheminform 2015, 7(Suppl 1):S2.

7. **Akhondi SA,** Hettne KM, van der Horst E, van Mulligen EM, Kors JA: *Recognition of chemical entities: combining dictionary-based and grammar-based approaches.* J Cheminform. 2015, 7:S10.

8. Kors JA, Clematide S, **Akhondi SA,** Rebholz-Schuhmann D: *A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC.* J Am Med Informatics Assoc 22:948-956.

9. **Akhondi SA,** Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA: *Annotated chemical patent corpus: a gold standard for text mining.* PloS one 2014, 9:e107477.

10. **Akhondi SA,** Kors JA, Muresan S: *Consistency of systematic chemical identifiers within and between small-molecule databases.* J Cheminform 2012, 4:35.

List of publications in conference proceedings:

1. Intxaurrondo A, Perez-Perez MP, Perez-Rodrıguez G, Lopez-Mart AJ, Santamarıa J, de la Pena S, Villegas M, **Akhondi SA,** Valencia A, Lourenco A, Krallinger M: *The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts.* In: BioCreative VI Workshop 2017.

2. Krallinger M, Rabal O, **Akhondi SA,** Martín P, Santamaría J, Rodríguez GP, Tsatsaronis G, Intxaurrondo A, López JA, Nandal U, van Buel E, Chandrasekhar A, Rodenburg M, Laegreid A, Doornenbal M, Oyarzabal J, Lourenço A, Valencia A: *Overview of the BioCreative VI chemical-protein interaction track.* In: BioCreative VI Work. 2017, pp 140-147.

3. Zhu Z, **Akhondi SA,** Nandal U, Doornenbal M, Gregory M: *Selecting Documents Relevant for Chemistry as a Classification Problem.* In: Eur. Knowl. Acquis. Work. 2016, pp 198-201.

4. Afzal Z, **Akhondi SA,** van Haagen HH, van Mulligen EM, Kors JA: *Concept Recognition in French Biomedical Text Using Automatic Translation.* In: Int. Conf. Cross-Language Eval. Forum Eur. 2016, pp 162-173.

5. van Mulligen EM, Afzal Z, **Akhondi SA,** Vo D, Kors JA: *Erasmus MC at CLEF eHealth 2016: Concept Recognition and Coding in French Texts.* In: CLEF (Working Notes). 2016, pp 171-178.

6. **Akhondi SA,** Pons E, Afzal Z, van Haagen H, Becker B, Hettne KM, van Mulligen EM, Kors JA: *Patent mining: combining dictionary-based and machine-learning approaches. In: Proc.* Fifth BioCreative Chall. Eval. Work. 2015, pp 102-109.

7. Pons E, Becker B, **Akhondi SA,** Afzal Z, van Mulligen EM, Kors JA: *RELigator: Chemical-disease relation extraction using prior knowledge and textual information.* In: Proc. Fifth BioCreative Chall. Eval. Work. 2015, pp 247-253.

8. Afzal Z, **Akhondi SA,** van Haagen H, van Mulligen EM, Kors JA: *Biomedical Concept Recognition in French Text Using Automatic Translation of English Terms.* In: CLEF 2015.

9. **Akhondi SA**, Singh B, van der Host E, van Mulligen EM, Hettne KM, Kors JA: *A dictionary- and grammar-based chemical named entity recognizer.* In: BioCreative Chall. Eval. Work. 2013, vol. p 113.

Patent application:

1. **Akhondi SA,** Rey H, Schwörer M, Maier M, Toomey J, Nau H, Ilchmann G, Sheehan M, Irmer M, Bobach C, Doornenbal M, Gregory M: *Automatic identification of relevant chemical compounds from patents.* Unites States Patent Office, Provisional Patent application (submitted).

ABOUT THE AUTHOR

Saber A. Akhondi was born in Iran, September 1982. He obtained his MSc degree in Bioinformatics and Systems Biology from Chalmers University of Technology, Sweden, in 2011. He did his thesis at AstraZeneca in 2010 with the title of "Integrating heterogeneous ranked sources with different reliabilities: A case study of Gene-Disease associations".

In 2011 he started working as a Scientist and a bioinformatic developer at AstraZeneca R&D Molndal, Sweden, where he was mainly involved in two projects called iSIM and Pharmaconnect.

In December 2011 he started as a PhD student within the biosemantics group. His research aimed at Exploring Chemical and Biological Named Entity Recognition (Chemical NER, BioNER) in drug discovery. The project was a collaboration between AstraZeneca and the department of Medical Informatics at Erasmus MC.

As of December 2015, he is working at Elsevier Amsterdam as a Senior NLP Scientist where he is applying state-of-the-art NLP and machine learning techniques to extract information useful for large commercial and research communities.