Cheminformatics: An Introductory Review

Seifu Chonde and Soundar Kumara
Department of Industrial and Manufacturing Engineering
The Pennsylvania State University, University Park, Pennsylvania 16802-6817

Abstract

Cheminformatics, a sub-field of chemometrics, utilizes "Big Data" to answer questions related to chemical synthesis. In particular, it can help to answer the question of how to discover the next (possibly newly synthesized) useful molecule that is easy to produce, and subsequently, evolve that molecule from a theoretical point of view to the production scale. Cheminformatics offers the application of optimization techniques in an interdisciplinary field to answer these questions. It requires insight from computer science, information science, chemistry, operations research, and statistics. Based on a review of the scholarly literature, this work will introduce the landscape of cheminformatic research and its overlap with Industrial Engineering. A short history of the field is provided and current cheminformatic areas are aligned with Industrial Engineering skill sets and interests. The authors discuss three stages of cheminformatic research: 1) capturing data; 2) storing data; and, 3) mining data. Summary tables of research areas, journals, and overlap between cheminformatics and Industrial Engineering are provided in an analysis of published literature to show what challenges are being heavily researched. Finally, a synergy of Industrial Engineering and cheminformatic research is considered for future directions.

Keywords
Cheminformatics, review, industrial engineering

1. Introduction

Many questions have plagued scientists in general over the ages, but a particularly challenging one that encompasses a major goal of chemistry as a science is as follows: How does one discover a new, useful molecule that is easy to produce and, subsequently, evolve that molecule from a set of raw materials? From synthesis studies to quantum mechanics, the study of chemistry and its subfields has yielded theories and tools to aide in answering such a challenging question.

In general, this question is approached from two different ways. The first approach, often called the forward model, seeks to find the useful synthesized products when a set of reactants are combined. This framing of the forward model is at the compound level and focuses on molecular synthesis. The second approach, often called the backward or inverse model, first defines a set of desirable physical and/or reactive properties then seeks out candidate compounds that display these properties. This approach is much more challenging because the inverse model can be setup as an equation whose unknowns are integers (Diophantine equation) and has a large search space.

Cheminformatics, a recent subfield that has gained considerable momentum (spelled chemoinformatics in Europe), provides algorithms and methodologies to break this governing research question down into manageable problems and, subsequently, present elegant solutions. This article first introduces cheminformatics and discusses why Industrial Engineers are well positioned to get involved in research in this area; Section 2 provides a look at the overlap between cheminformatics and Industrial Engineering as rationale why IEs should enter this domain; Section 3 provides a history of cheminformatics research; Section 4 gives some example data for Aspirin to show what information databases store; and, Section 5 concludes by summarizing how industrial engineers can contribute to this new field and provides the authors’ take on the future of cheminformatics research.

1.1 Definitions of Cheminformatics

In the book Chemometrics and Chemoinformatics, editor Barry Levine describes cheminformatics as a subfield of chemometrics [1]. He describes chemometrics as an approach to analytical chemistry based on the idea of indirect
Barry Levine’s definition of cheminformatics is one of many in literature. The reason for disparate definitions is that cheminformatics is a nascent field. Cheminformatic’s seminal definition comes from Dr. Brown in 1998 as the practice of using “all the information resources that a scientist needs to optimize the properties of a ligand to become a drug” [2]. Yet, this definition, with its strong focus on drug discovery, is deprecated. The field of cheminformatics has yet to choose a single definition. Some definitions include:

- The mixing of information resources to transform data into information, and information into knowledge, for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization.
- A subfield of chemometrics that encompasses analysis, visualization, and use of chemical structural information as a surrogate variable for other data or information.
- A conglomeration of molecular modeling, chemical information, and computational chemistry for the development of experimental techniques to analyze unprecedented volumes of data.

These three definitions, where we have underlined particular portions to identify how various individuals capture the main components of cheminformatics. In the book Chemoinformatics, by Johann Gasteiger a pioneer in this field, the following note is made in the Foreword:

“Chemoinformatics is a generic term that encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization and use of chemical information [Paris 2000]” [3]

The definition above shows cheminformatics as a multifaceted research area. In his review, Chen [4] lists four traditional areas of cheminformatic research (explained in the next subsection below), while Bajorath [5] introduced a pragmatic spectrum of cheminformatics to show various research tasks shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1 Spectrum of Cheminformatics adopted from Bajorath [5]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical data collection, analysis, and management</td>
</tr>
<tr>
<td>Data representation and communication</td>
</tr>
<tr>
<td>Database design and organization</td>
</tr>
<tr>
<td>Chemical structure and property prediction (including likeness)</td>
</tr>
<tr>
<td>Molecular similarity and diversity analysis</td>
</tr>
<tr>
<td>Compound or library design and optimization</td>
</tr>
<tr>
<td>Database mining</td>
</tr>
<tr>
<td>Compound classification and selection</td>
</tr>
<tr>
<td>Qualitative and quantitative structure-activity or –property relationships</td>
</tr>
<tr>
<td>Information theory applied to chemical problems</td>
</tr>
<tr>
<td>Statistical models and descriptors in chemistry</td>
</tr>
<tr>
<td>Prediction of in vivo compound characteristics</td>
</tr>
</tbody>
</table>

This spectrum definition is useful because it includes all concepts and methods designed to interface theoretical and experimental programs involving small molecules. However, all of the definitions and approaches listed above show that cheminformatics has a large landscape of research problems. In the next subsection we explicitly explore how to view cheminformatics from a practical standpoint: that of a task-oriented researcher.

1.2 Main Areas of Cheminformatics
For the purposes of his review, Chen [4] lists four areas as:

- Structure Elucidation: development of chemical structure of uncharacterized substances usually done through analyzing spectral data.
- Synthesis Design: determination of optimal synthesis steps for a given target
Database Systems: the processes and logic that are put in place to actively store, manage, and retrieve chemical data

Structure Builders: computer assisted development of 2D and 3D structures.

This fit the Chen’s goal of identifying the role of computers in cheminformatics throughout history but foregoes discussion on other important topics such as QSAR [6] and highthroughput screening [7].

In this paper we adopt a practical approach to explaining the cheminformatic research landscape. We prefer to discuss cheminformatic research according to stages of research. Specifically, we consider capturing data, storing data, and mining data to be the three stages of research in a cheminformatics. These stages naturally progress as the discovery process continues. Figure 1 shows some research streams or generalized problems in each of the three stages.

**The Three Stages of Cheminformatic Research Areas**

- **Capturing Data**
  - Chemical descriptors and fingerprint collection
  - Storing electronic lab notebooks
  - Chemical 3D and 2D data encoding in various file formats (e.g., Chemical Mark Language, SD File, SMILES/SMARTS)
  - Reaction representation

- **Storing Data**
  - Data management
  - Data representation and communication
  - Compound or library design and database optimization
  - Statistical models and descriptors
  - Laboratory Information Management Systems (LIMS)

- **Minining Data**
  - Structure searching and mining
  - Chemical structure and property prediction
  - Molecular similarity and diversity analysis
  - Prediction of in vivo characteristics
  - Compound classification and selection for syntheses

Figure 1: Stages of cheminformatic research areas and research streams.

Capturing data includes all issues related to collecting and encoding the data for storage. Such tasks include taking chemical information and transforming it into common chemical lexical terms. Examples of this are given in Section 4 below. This process includes not only storage of chemical descriptors for a molecule but also storing laboratory data and chemical reaction data. Storing data includes the management of data warehouses and libraries. Additionally, there has been a large push since the 1990s for automated laboratory information management systems (LIMS) and electronic notebooks (ELN). Finally, mining data seeks to observe and predict interesting patterns in the data especially with regards to the relationship between compound structure and compound properties or activity.

The next step in understanding cheminformatics research is where to look for information. Cheminformatics research is carried out in many journals throughout the world. Table 2 has a sample of the most relevant journals and listed them in order of publication date in. Table 2 shows that many of the journals (5 of 13) are published by the American Chemistry Society (ACS). Many cheminformatics are formally trained in chemistry and have honed additional information science and statistics skills. Also, roughly 40% of the journals have a focus on biological research and drug design. This reinforces that, historically, academic cheminformatics research has been conducted using drug design as a testbed. However, there is a large amount of research into similar problems in industry that goes largely unpublished because of intellectual property concerns. The next section of this report compares cheminformatics and Industrial Engineering.
Table 2: Main Journals in Cheminformatics Ordered by First Publication

<table>
<thead>
<tr>
<th>Journal Name</th>
<th>Publisher</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acta Crystallographica (6 Journals)</td>
<td>Wiley-Blackwell</td>
<td>1948-</td>
</tr>
<tr>
<td>Journal of Chemical Information and Modeling (formerly Journal of Chemical Documentation)</td>
<td>ACS</td>
<td>1961-</td>
</tr>
<tr>
<td>Journal of Medicinal Chemistry</td>
<td>ACS</td>
<td>1963-</td>
</tr>
<tr>
<td>Journal of Computational Chemistry</td>
<td>John Wiley &amp; Sons</td>
<td>1980-</td>
</tr>
<tr>
<td>Journal of Molecular Graphics and Modeling</td>
<td>Elsevier</td>
<td>1983-</td>
</tr>
<tr>
<td>Journal of Computer-Aided Molecular Design</td>
<td>Springer</td>
<td>1987-</td>
</tr>
<tr>
<td>Drug Discovery Today (4 Journals)</td>
<td>Elsevier</td>
<td>1996-</td>
</tr>
<tr>
<td>ACS Combinatorial Science (formerly Journal of Combinatorial Chemistry)</td>
<td>ACS</td>
<td>1999-</td>
</tr>
<tr>
<td>Journal of Proteome Research</td>
<td>ACS</td>
<td>2002-</td>
</tr>
<tr>
<td>Molecular and Cellular Proteomics</td>
<td>ASBMB</td>
<td>2002-</td>
</tr>
<tr>
<td>Journal of Chemical Theory and Computation</td>
<td>ACS</td>
<td>2005-</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Wiley-VCH</td>
<td>2006-</td>
</tr>
<tr>
<td>Journal of Cheminformatics</td>
<td>Chemistry Central</td>
<td>2009-</td>
</tr>
</tbody>
</table>

2. Similarities between Cheminformatics and Industrial Engineering

Before proceeding it is important to note that chemistry knowledge is a necessary component of cheminformatics. And while the challenging chemistry part is somewhat complete—public databases with millions of records, chemical syntaxes, and many data fusion and search techniques have already been created—knowledge of basic chemistry principles is a prerequisite for cheminformatic research. The validation of results may be a confusing and tiresome task without a basic, undergraduate-level understanding of chemistry. In such instances, it is recommended that a researcher with suitable chemistry knowledge participates in the work.

The areas of Industrial Engineering and Cheminformatics have considerable overlap. Industrial Engineers focus on the following: Statistics, Optimization, Logistics, Simulation, Machine Learning, Systems Modeling, Networks, and Big Data. Cheminformatics focuses on the following: {Chemical nomenclature and database creation, graph and text mining, structure similarity & compound clustering, chemical reaction simulation, patent/regulatory information obfuscation DNA computing, mathematical modeling, ELN and LIMS}. Relating these focus areas provides some insight on how Industrial Engineers are aligned to handle cheminformatic questions.

The overlap between these two fields is summarized in Figure 2 below. Little of the cheminformatic skills have to do with the manufacturing and logistics skills developed in Industrial Engineering. This is to say that cheminformatics does not deal with using large-scale machinery such as extruders or with supply chain logistics in the same way as Industrial Engineering does. To fully explore the similarities between these two areas each heading of cheminformatics is considered separately in a subsection below.

![Figure 2: Relationship between Cheminformatic and Industrial Engineering skillsets.](image-url)
2.1 Chemical Nomenclature
Chemical nomenclature refers to the representation of a molecule in a common, ideally canonical, lexicon that is shared among scientists. Examples of chemical nomenclature include the IUPAC International Chemical Identifier (InChI) and InChI Key; linear notations such as SMILES, SMARTS, and WLN; systematic name; and common name. Chemical nomenclature can be extended to include a complete representation of the compounds chemical makeup that is useful for computer storage. Chemical file formats such as the SDFile have been created to capture and store this information. Examples of these are discussed in Section 4. The skills needed to handle problems in chemical nomenclature are data storage and databases, networks and graph theory, and Big Data.

2.2 Graph and Text Mining
Once data has been appropriately translated into a machine readable format and stored it is possible to mine the data based on queries. A large area of research in cheminformatics historically has been into the (sub)graph isomorphism problem. In this problem one seeks to detect an equivalence relation between two graphs or subgraphs. It is used to determine functional group or moiety similarity between molecules. A functional group is a specific set of bonded atoms that produce a characteristic defining behavior. For example, say you have a large database and would like to consider only molecules that have an ester functional group. A database will return many such molecules with Benzyl acetate being an example of one and the ester is highlighted in Figure 3.

Figure 3: Benzyl acetate with ester functional group circled.

There are many freely available databases and search interfaces such as the Royal Society of Chemistry’s ChemSpider (http://www.chemspider.com). Some of these databases rely on SMILES string-based notations of compounds. This is where text mining becomes a necessary skill. These freely available databases approach the size appropriate for Big Data considerations with some having more than 30 million compounds’ data accessible. Additionally, network analysis and systems modeling is important in this area as well as a working knowledge of graph theory.

2.3 Structure Similarity & Compound Clustering
Graph and text mining lead to the realization that compounds can be clustered given their functional groups. This clustering realization, coupled with the Similar Property Principle [8], powered cheminformaticists to develop measures of similarity between compounds. An ongoing issue in cheminformatics remains to determine the effectiveness of such analyses [9]. Structure similarity and compound clustering has been the driving force in research for many years and has sprouted an interest in what is known as Quantitative Structure-Activity Relationship (QSAR) or Quantitative Structure-Property Relationship (QSPR) research. The QSAR problem focuses on finding functional solutions to the model for a particular chemical structure (G) a property (P) or activity (A):

\[ A = f(G) + \varepsilon \]  
and \[ P = f(G) + \varepsilon \]

(1)

Structure similarity and compound clustering research requires many Industrial Engineering skills. Examples of statistics [10]–[12], optimization [13], [14], machine learning [15]–[17], networks [18], and Big Data [19] provide a clear indication that Industrial Engineers are well suited to support research tasks in QSAR.

2.4 Chemical Reaction Simulation
Chemical reaction simulation has long been an interesting area of research for chemists and engineers. Understanding how reactions proceed is vital for high yield, increased safety, and process monitoring. Cheminformatics is useful for chemical reaction simulation at the quantum level up to the macro level depending on the research scale of interest. There remains a need for chemical reaction simulation even as improvements in experimental procedures, such as high throughput screening (HTS), approach Edisonian discovery.

In chemical reaction simulation a forward model is one that predicts how a set of reactants will synthesize various products given a set of operating conditions. The backwards model, on the other hand, forecasts the success of various reactants to synthesize a product of interest. The optimization of a backward model is much more taxing and
chemical reaction simulation has advanced knowledge in this area. Research in this area uses a mix of statistics, optimization, and simulation.

2.5 Patent/Regulatory Information Obfuscation
An important concern for chemical manufacturing companies is patent and regulatory information obfuscation. There are practices employed by companies to ensure that patent information can’t be obfuscated by competitors; however, these measures may not be enough. As technology continues to advance there will be algorithms and tools developed that will be able to determine formulations from multiple patents. Ultimately, this will provide companies with a better view of competitor formulations. Combining such knowledge with somewhat publically available pricing information will allow companies that use such technologies a large market advantage.

Much of the work done in decoding obfuscated patents in the chemical realm has been conducted by IBM. The research in this area is often concerned with code obfuscation in computer science and software development. Some examples include [20]–[22]. Systems modeling and machine learning stand out as necessary skills in this area.

2.6 DNA Computing
DNA computing is a process that utilizes biological processes in a massively parallel manner to perform calculations faster than the world’s fastest silicon-based computers. In the seminal work, Leonard Adleman argued that his DNA and enzyme based computer could roughly make \(10^{14}\) operations per second. And, a computer that made \(10^{20}\) operations per second could feasibly be made by him would be 1000 times faster than the world’s super computer [23]. DNA computing is a useful technique to solve cheminformatic research questions surrounding combinatorial chemistry. It requires the skills of machine learning and simulation to perform research using DNA computing.

2.8 Electronic Laboratory Notebook (ELN) and Laboratory Information Management Systems (LIMS)
With the rise in the amount of data generated during chemical processes there has been a steady rise in Laboratory Information Management Systems (LIMS). LIMSSs are software systems that support capturing, tracking, sorting, and querying data streams in a dynamic, time-based fashion. Many LIMS distribution companies exist, each offering slightly different key features. These features are targeted to specific groups of users based on the common processes they use. Current LIMS software incorporates electronic laboratory notebook (ELN) data “to create, store, retrieve, and share electronic records in ways that meet all legal, regulatory, technical, and scientific requirements” [24].

Analysis of ELN and LIMS datasources requires high computing power and many data fusion and integration algorithms. Particular software suites boast about their integration systems; however, when it comes to selected an ELN or LIMS: one size does not fit all. There are many commercial off the shelf options and some companies choose to develop their own. In all packages, the ability to incorporate and merge sensor and computer-generated data is paramount to the platform’s success. For this reason the development of these packages, while mostly done by industry professionals, requires knowledge of systems modeling, Big Data, and machine learning algorithms.

3. History of Cheminformatics
Cheminformatics is a newly titled field but its roots extend back to the early 19th century. Before there was cheminformatics, there was a battle for the name of an emerging discipline of computer applications in chemistry between chemometrics and computer chemistry. Now all three are used but cheminformatics has broken away from in some ways. The history provided here is a quick glimpse at the most notable discoveries and papers from much before Dr. Brown first used the term chemoinformatics in 1998 [2]. In other words, despite it being named in 1998 cheminformatics isn’t exactly a new field.

Chen, in his article in a special issue of the Journal of Chemical Information Modeling dedicated to Professor Johann Gasteiger, provides an interesting review of the progress of cheminformatics decade by decade from the 1940s to the 2000s [4]. Dr. Peter Willet provides another perspective of how cheminformatics has developed since the 1950s [25]. Lavine et al have produced a set of 19 nearly annual reviews covering chemometrics specific topics of pattern recognition, multivariate curve resolution, and multivariate calibration [26]. These resources, as well as the introductory chapters in many of the texts on cheminformatics [3], provide a thorough and annotated history of the cheminformatics field. We will not reproduce the entirety of these works or as much depth material as they provide. Instead, we wish to provide a simple understanding of what has been challenging throughout the decades and led to the emergence of cheminformatics. Most notable papers are provided for the interested reader to explore.
3.1 Cheminformatics in the 1940s and 1950s
The 1940s and 1950s saw the birth of the computers and the first applications of computers in chemistry. The Chemical Abstract Service (CAS) of the American Chemical Society (ACS) was one of the early adopters of computers in chemistry and produced some of the first chemical information databases. Today they exist with a sole purpose of finding, collecting, and organizing substance information with over 16 million organic and inorganic substances organized [27]. In 1957, Ray and Kirsch provided the first algorithm in cheminformatics: a substructure algorithm using graph theory [28].

3.2 Cheminformatics in the 1960s
The 1960s saw the creation of the Journal of Chemical Documentation which changed its name to Journal of Chemical Information and Modeling (JCIM). JCIM is one of the core journals for cheminformatics alongside the Journal of Cheminformatics.

The interest in categorizing molecules from the 1950s spurred an interest in categorizing reactions in the 1960s. Vleduts suggested classifying reactions by their reaction center in 1963. This work started computer-aided synthesis design where possible reaction schemes are considered to synthesize particular targets from a set of reactants (Vleduts 1963). In a similar manner, computer-aided structure elucidation was started in 1965 to determine the molecule that has a set of particular spectral data (Sasaki, DENDRAL project). The 1960s also saw the seminal paper in QSAR. In 1962 Hansch et al used multiple regression to regress biological activity on molecular properties [29].

3.3 Cheminformatics in the 1970s
The work by Hansch et al was limited to studies where the dataset shared similar properties and most applicable in lead optimization studies. Lead discovery, on the other hand, uses structurally diverse datasets and in 1974 Cramer et al published the first QSAR study using substructural analysis [30]. The 1970s saw improvements in visualization as Wipke et al developed a 3D visualization computer program from 2D data called PRXBLD [31].

3.4 Cheminformatics in the 1980s
In his history, Chen explains that the 1980s saw the development of rule-based 3D structure builders and database-based 3D structure builders [4]. The first focuses on the use of logical rules to determine 3D structure where CORINA [32] is the most widely used, current version with more than 60 installations worldwide. The second, database-based 3D structure builders, focus on locating analogies of a structure to produce the particular 3D structure of interest. This work in 3D structure elucidation was matched by the first methods for 3D-QSAR in what is now called structure-based drug design [25]. In addition to 3D advancements in cheminformatics, the 1980s saw the creation of Simplified Molecular Input Line Entry System (SMILES) by Weininger [33]. Gasteiger provides an in-depth history of linear notations including their roots from the invention of the 1949 Wiswesser line notation (WLN) [3], [34].

3.5 Cheminformatics in the 1990s
The 1990s saw many new file formats that are still widely used today. Some examples are the SDfile, RGfile, RXNfile, and XDfile. The SDfile, which stands for structure-data file, is likely the most widely used of these formats and is a wrapper for a MDL Molfile. The SDfile contains information about a header, atoms, bonds, molecular names, a connection table, and (sometimes) properties. Using these files it has become possible to quickly relay all important information about a molecules 2D and 3D properties. Many search engines were developed to quickly sort through libraries of compounds and return applicable data in HTML or other file formats.

3.6 Cheminformatics in the 2000s and Today
2000s and current research has seen a move towards algorithms on increasing large datasets. As mentioned before the PubChem database has over 34 million compounds and is freely available. Tools and toolkits have become a mainstream way to share research and develop custom computer applications. The last 15 years have seen development of quite a few toolkits for cheminformatics. For example, the Chemistry Development Kit (CDK) is an open-source Java library that has been incorporated into several applications such as R, Taverna, Bioclipse, Cinfony, and Excel. The aim of CDK is to package common functions for cheminformatic tasks such as diagram generation and QSAR descriptor calculations for research and application development.
To recap, a significant amount of research has been done in cheminformatics since the first principles were developed in the 1940s and 1950s. The area has gained some momentum and is attracting many new researchers and interests from a wide array of other fields. Figure 4 below shows the number of journal articles published in cheminformatics since the 1940s by decade. This information was taken from Web of Science©. Additionally, GoogleScholar© shows 713 articles and patents when searching for titles that contain “cheminformatics” or “chemoinformatics”.

![Figure 4: Number of published articles with “cheminformatics” or “chemoinformatics” in the title pulled from Web of Knowledge©.](image)

4. Aspirin as an Example
There are many examples of cheminformatics in research. Some snippets are presented to show how various topics are used and a longer research example is discussed below. The first example has to deal with searching for molecules. Let’s assume we are interested in finding everything there is to know about Aspirin. There are web services like ChemSpider™ that allow the searching of chemical data. When querying Aspirin we see that the molecular formula is C₉H₈O₄ with a SMILES of CC(=O)Oc1ccccc1C(=O)O and a 2D picture given below. When looking at the SMILES string and the figure you can see the two are related: the SMILES is a left-to-right description of the 2D figure.

![Figure 5: 2D representation and infrared spectra of Aspirin](image)

From this search ChemSpider provides the ability to find other compounds with the same skeleton or conduct a structure search. Also, you can easily browse through chemical properties, spectra, CIFs, related articles, vendors, data sources, related patents, RSC databases, video based description from YouTube™, medical subject headings classification, pharmacological links, SimBioSys LASSO values, and images. This is a massive amount of information reaching the limits that one would ever want to know about Aspirin as a particular chemical compound (c.f. CSID:2157, http://www.chemspider.com/Chemical-Structure.2157.html).

<table>
<thead>
<tr>
<th>Coordinates</th>
<th>Element</th>
<th>Element in Bond First</th>
<th>Bond Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>Y</td>
<td>Z</td>
<td></td>
</tr>
<tr>
<td>3.7320</td>
<td>-0.0600</td>
<td>0</td>
<td>O</td>
</tr>
<tr>
<td>6.3301</td>
<td>1.4400</td>
<td>0</td>
<td>O</td>
</tr>
<tr>
<td>4.5981</td>
<td>1.4400</td>
<td>0</td>
<td>O</td>
</tr>
<tr>
<td>2.8660</td>
<td>-1.5600</td>
<td>0</td>
<td>O</td>
</tr>
<tr>
<td>4.5981</td>
<td>-0.5600</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>5.4641</td>
<td>-0.0600</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>4.5981</td>
<td>-1.5600</td>
<td>0</td>
<td>C</td>
</tr>
</tbody>
</table>
The SDFiles for Aspirin can be found in many databases. Here a glimpse of the 2D SDFile from PubChem is shown in Table 6. It shows only the atom and bond information and does not show molecular data such as charges or mass.

5. Conclusions and Future Predictions

This introduction has presented the history and background of cheminformatics for Industrial Engineers. Several grand challenges were presented; some challenges were addressed throughout the development of cheminformatics and some remain pressing issues to be researched. This work has shown there is considerable overlap between Industrial Engineering skills and the problems in cheminformatics that are challenging. As research and technology continue to flourish we predict the following research areas will gain additional momentum:

- **Real-time data analysis**: The capabilities of real-time sensor technologies has been steadily advancing. New pattern recognition software for both LIMS and ELN for chemistry manufacturing processes may allow for further preventative maintenance and tighter control of systems or processes. In order of this to occur there must be technologies developed that can process large amounts of chemical data and link process data to possible issues. Additionally, data storage has developed considerably since the 1950s and leads to the question: with the advancements in ELN and LIMS and their adoption by companies is it possible to develop specific machine learning/pattern recognition techniques for this setting?

- **Improved design methodologies**: The number of alternatives considered in a design problem is vast leading to a seemingly infinite decision theory problem. Current Edisonian-esque, test-and-fail methodologies of design in many manufacturing settings are suboptimal. Enhanced data manipulation and understanding of cheminformatic techniques can greatly reduce the R&D costs for chemical (and pharmaceutical) plants.

- **Generalization of cheminformatic principles to various products**: Cheminformatics has a research history that focuses greatly on drug design and development. As the field continues to grow and prosper there is a need for applications of cheminformatic algorithms to additional fields than drug design. This need is most evident in chemical manufacturing that make products that have to pass numerous physics and chemistry tests in order to be sold to clients.

- **Analysis of “Big Data”**: Single databases such as PubChem have more than 34 million chemical-structure records and this is just a minutia of the 10⁶th molecules that are feasibly synthesized right now. Some interesting results exist for designing new drugs such. For example, the Lipinski’s Rule of Five states that an orally active drug is unlikely to have more than one violation of five simple, independent, and easily verified rules. Considering this an open question remains: Is it possible that the Lipinski Rule of Five can be developed for other types of molecules than drug molecules?

- **Advanced artificial compound evolvers**: Structure elucidation and synthesis have a long history of artificial intelligence as pointed out by Chen [4]. An open grand challenge in chemistry to identify new artificial intelligence techniques in compound evolution using some of the latest advanced techniques from the field such as agent based modeling.

References


