Chemoinformatics—An Introduction for Computer Scientists

NATHAN BROWN

Novartis Institutes for BioMedical Research

Chemoinformatics is an interface science aimed primarily at discovering novel chemical entities that will ultimately result in the development of novel treatments for unmet medical needs, although these same methods are also applied in other fields that ultimately design new molecules. The field combines expertise from, among others, chemistry, biology, physics, biochemistry, statistics, mathematics, and computer science. In this general review of chemoinformatics the emphasis is placed on describing the general methods that are routinely applied in molecular discovery and in a context that provides for an easily accessible article for computer scientists as well as scientists from other numerate disciplines.

1. INTRODUCTION

Chemistry research has only in recent years had available the technology that allows chemists to regularly synthesize and test thousands, if not hundreds of thousands, of novel molecules for new applications, whereas before these technologies existed, a typical chemist would consider only one to two molecules a week. However, information management and retrieval systems have not developed sufficiently in pace with these technologies to allow for the generated information to be collated and analyzed in a standard and efficient manner thereby making best use of our knowledge base.

Author's address: The Institute of Cancer Research, 15 Cotswold Road, Sutton, SM2 5NG, Surrey, U.K.; email: nathan.brown@icr.ac.uk.

Indeed, there are many significant challenges still open not only in chemical databases, but also in the methods to analyze properly the molecular structures to transform the data into information that can be interpreted. This is one of the most important aspects of current chemistry research, since it can be predicted that making available computational tools to analyze molecules will reduce the numbers of experiments that chemists must perform. It has even been speculated that the vast majority of the discovery process for novel chemical entities (NCEs) will one day be performed *in silico* rather than *in vitro* or *in vivo*.

An example of the historical interface between chemistry and computer science is provided in the story surrounding the development of a fragment screening system in chemical systems that used fragment codes. The fragment screening systems were initially developed to increase the speed at which large databases of molecules could be prefiltered for the presence or absence of a particular chemical substructure, before proceeding on to a more intensive graph-matching algorithm. These methods were later adapted to text searching by Michael Lynch—a pioneer in chemical informatics—and colleagues in Sheffield and later for text compression by Howard Petrie, also in Sheffield. Lynch discussed this concept with the Israeli scientists Jacob Ziv and Abraham Lempel in the early 1970s, who went on to generalize the concept and adapt it in their Ziv-Lempel (LZ77 and LZ78) algorithms, which went on to become the Lempel-Ziv-Welch (LZW) algorithm, which is now used widely in compression of data. Therefore, we can see that the cross-fertilization of ideas from two essentially similar fields in theory, yet different in application, has led to paradigm shifts that were completely unanticipated [Lynch 2004].

This article is intended as a general introduction to the current standard methods applied in the field of chemoinformatics that is accessible to computer scientists. In this aim, it is intended to provide a useful and extensive starting point for computer scientists, which will also be of general interest to those in any numerate discipline including the field of chemistry itself. The article covers historical aspects of chemistry and informatics to set the field in context, while also covering many of the current challenges in chemoinformatics and popular techniques and methods that are routinely applied in academia and industry.

Many of the activities performed in chemoinformatics can be seen as types of information retrieval in a particular domain [Willett 2000]. In document-based information retrieval we can apply transforms to describe a document of interest that then permits an objective determination of similarity to additional documents to locate those documents that are likely to be of most interest. In chemoinformatics this is similar to searching for new molecules of interest when a single molecule has been identified as being relevant.

Although perhaps less known than the sister field of bioinformatics [Cohen 2004], chemoinformatics has a considerable history both in research and in years. Whereas bioinformatics focuses on sequence data, chemoinformatics focuses on structure information of small molecules. A great deal of chemoinformatics research has been conducted in a relatively small number of world-class academic laboratories. However, due to the applied nature of the field, a considerable amount has also been achieved in large chemical companies, including pharmaceuticals, petrochemicals, fine chemicals, and the food sciences. Although many areas of research within chemoinformatics will be touched upon in this introduction for computer scientists, it is not intended as a full and comparative critique of existing technologies, but as a brief overview of the main endeavors that a typical scientist conducting research in this field would know instinctively. References have been provided to review articles on particular topics to allow the interested reader ready access to articles that detail the most salient points regarding each of the methods and applications discussed herein. Further details
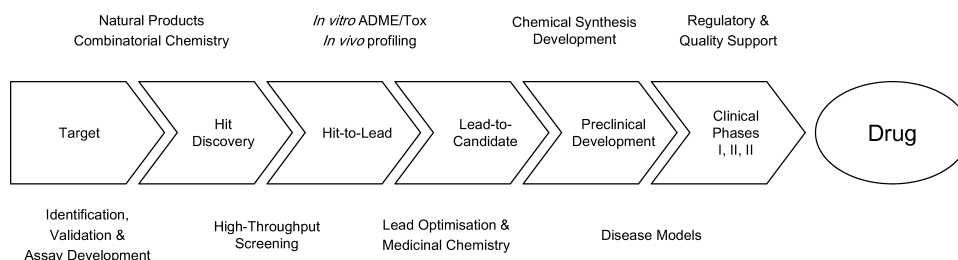
Fig. 1. An illustration of the typical workflow of a drug discovery endeavor.

on chemoinformatics are available in two excellent textbooks [Leach and Gillet 2003; Gasteiger and Engel 2003], and in more general surveys of the field [Bajorath 2004; Gasteiger 2003; Oprea 2005a].

1.1. Chemoinformatics and Drug Discovery

The discovery of new medical treatments to meet unmet medical needs is one of the most important endeavors in humanity. The process is time consuming, expensive, and fraught with many challenges. Drug discovery is perhaps the area of research in which chemoinformatics has found the greatest application. In this article we focus on the research phase of drug discovery where NCEs are designed that bring about a desired biological effect that will help the patient.

Drug discovery generally follows a set of common stages (Figure 1). First, a biological target is identified that is screened against many thousands of molecules in parallel. The results from these screens are referred to as *hits*. A number of the hits will be followed up as *leads* with various profiling analyses to determine whether any of these molecules are suitable for the target of interest. The leads can then be converted to *candidates* by optimizing on the biological activity and other objectives of interest, such as the number of synthetic steps. Once a suitable candidate has been designed, the candidate enters preclinical development. Chemoinformatics is involved from initially designing screening libraries, to determining hits to take forward to lead optimization and the determination of candidates by application of a variety of tools.

The lifecycle of drug discovery can take many years and hundreds of millions of dollars to achieve. Recent estimates are leaning toward an average drug discovery that lasts at least a decade (10–15 years) and costs almost a billion US dollars ($897 million) in bringing a single drug to market, with many avenues necessarily being explored throughout the process [DiMasi et al. 2003]. However these figures are only estimates, and there is no doubt that the discovery of new medical treatments is a very complex process and the field of chemoinformatics is assisting in allowing us to explore new areas that we have hitherto been unable to access.

1.2. Chemistry Space

Chemistry space is the term given to the space that contains all of the theoretically possible molecules and is therefore theoretically infinite. However, when considering druglike chemistry space, the space becomes bounded according to known conditions that would make these molecules unlikely to be druglike molecules—such as the Lipinski Rule-of-5 (Ro5) [Lipinski et al. 2001] where a set of empirically derived rules is used to define molecules that are more likely to be orally available as drugs. However, even this reduced druglike chemistry space is estimated to contain anything from 10^{12} to 10^{180} molecules [Gorse 2006]. To put this in context, the Chemical Abstracts Service

(CAS) currently contains fewer than 33 million (as of September 19, 2007) molecules. Therefore, the theoretical druglike chemistry space contains anything from 3×10^5 to 10^{173} times more druglike molecules than we have currently registered in CAS. Moreover, CAS also contains molecules that are not necessarily druglike.

To be able to consider this vast druglike chemistry space, it is necessary to deploy computer systems using many diverse methods which allow a rational exploration of the space without evaluating every molecule while also capitalizing on the extant molecules we have stored in our compound archives.

In drug discovery the search for NCEs is a filtering and transformation process. Initially, we can consider the potential druglike space; however, this can be difficult and leads to issues in synthetic accessibility—that is can we actually make these molecules in the laboratory that have not yet been solved for practical application? What is more normal is to consider the lists of available molecules from compound vendors and also molecules that are natural products. However, these lists still run into the many millions.

At this point we can introduce chemoinformatics techniques to assist in filtering the space of available molecules to something more manageable while also maximizing our chances of covering the molecules with the most potential to enter the clinic and maintaining some degree of structural diversity to avoid prospective redundancies or premature convergence. Each of these objectives must be balanced, which is no easy task. However, once this stage has been finalized to the satisfaction of the objectives under consideration, it typically contains only a few millions of molecules. These compound archives require vast and complex automation systems to maintain them and facilitate the usage of the compounds contained therein.

From this library screening sets are selected where each of the molecules in the set is tested against a biological target of interest using initial high-throughput screening (HTS). A hit in this context is a molecule that is indicated to have bound to our protein of interest. Typically, these hitlists are filtered using chemoinformatics methods to select only those molecules in which we are most interested; this process is often referred to as *HTS triaging*. From this comes a smaller triaged hitlist with which it is possible to cope in a higher-quality validation screen that often tests molecules in replicate to avoid potential technological artifacts. A summary of this filtering process in drug discovery is given in Figure 2.

1.3. High-Throughput Screening

As the name suggests, HTS is a method of testing many molecules rapidly *in vitro* against a particular protein target. There exist many alternative HTS technologies and these are outside the scope of this article. High-throughput experimentation (HTE) may be of interest to general readers due to the levels of automation and robotics involved; an example of the development of an HTS screening laboratory is given in Cechetto et al. [2004].

In general, HTS technologies apply the following workflow. HTS technologies use well plates with varying numbers of wells per plate, typically containing 96, 384, or 1536 wells. Each of these wells contains a compound and some biological matter of experimental interest, such as a protein, cells, or an animal embryo. If a desired response is observed then the molecules that were tested are referred to as *hits*.

In the context of chemoinformatics, the challenges in HTS are manifold: deciding what to contain in our corporate and screening libraries, analyzing primary hitlists for various artifacts and signals, prioritizing hits to be taken further into validation screens. HTS is arguably the area in which chemoinformatics is of most importance in the drug discovery process.
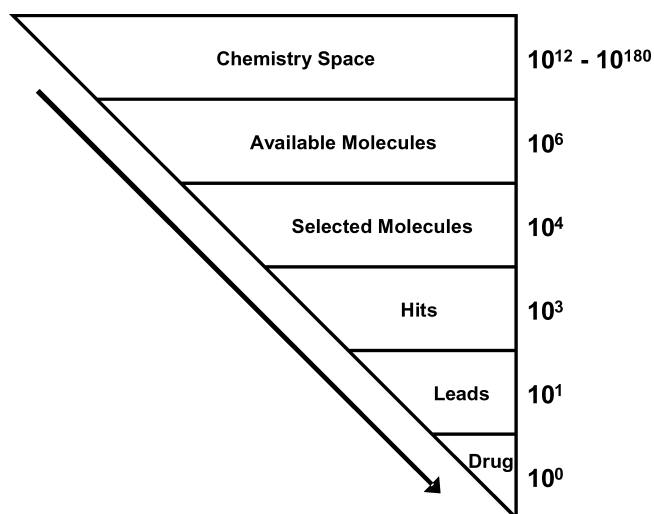
Fig. 2.   The filtering processes in drug discovery that are applied to our molecular collections and ultimately result in a drug that can be brought to the clinic. The hits are those molecules returned from an initial screening program, while the leads or lead candidates are those that are followed up in medicinal chemistry.

The first step in HTS is to decide which particular compounds are to be screened. Although druglike chemistry space is vast, only a relatively tiny subset of this space is available. However, even this space runs into the many millions of compounds. Therefore, it is important to determine which molecules should be included in our screening libraries to be tested with HTS.

Two extremes are evident in many HTS programs: diverse and focused screening libraries. The diverse libraries are often used for exploratory research, while focused sets are used to exploit knowledge from related targets in maximizing our hits in new targets. Corporate collections, historically, tend to be skewed to the particular areas of endeavor for which each of the companies are known. Recently, however, efforts have been made to increase the coverage of the available chemistry space to maximize the possibility of covering our biological activity space, typically using diversity selection methods, as discussed in Section 6 [Schuffenhauer et al. 2006].

Once HTS has been completed on a particular library, it is frequently necessary to further rationalize the set of hits for reasons of pragmatism or the particular HTS technology applied since HTS can return many hits dependent on the particular assay. Here again similar methods can be applied as is in the compound acquisition phase, although undesirable compounds will have already been filtered by this stage. Many alternative approaches exist in the literature for this HTS triaging phase, and typically the aim is once again to maximize our set of hits using extant knowledge, while also exploring new and potentially interesting regions of our explored chemistry space. This is often achieved using *in silico* tools that permit case-by-case alterations to our workflow.

1.4. Origins of Chemoinformatics

Chemoinformatics, in all but name, has existed for many decades. It could be argued that research has been conducted in this area since the advent of computers in the 1940s. However, the term *chemoinformatics* has only been in existence for the past decade [Brown 1998] yet it has quickly become a popular term with a number of books published that provide excellent overviews of the field. However, there is no true

definition of chemoinformatics, most likely due to its highly interdisciplinary characteristics, and this has been a source of debate in recent years. Here, a few quotes from leading industrial and academic scientists in chemoinformatics are provided to indicate the breadth of definitions [Russo 2002]. The differences in definitions are largely a result of the types of analyses that particular scientists practice and no single definition is intended to be all encompassing.

> 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.
>
> Frank K. Brown [1998] (cited in Russo [2000], page 4)

> [Chemoinformatics involves]. . . the computer manipulation of two- or three-dimensional chemical structures and excludes textual information. This distinguishes the term from chemical information, largely a discipline of chemical librarians and does not include the development of computational methods.
>
> Peter Willett, 2002 (cited in Russo [2001], page 4)

> . . . the application of informatics to solve chemical problems . . . [and] chemoinformatics makes the point that you're using one scientific discipline to understand another scientific discipline.
>
> Johann Gasteiger, 2002 (cited in Russo [2002], page 5)

From these quotations from leading scientists in the field of chemoinformatics, it is clear that there is still some dispute as to the "true" sphere of influence of chemoinformatics. Indeed, even the spelling of chemoinformatics is hotly debated, and the field is also referred to as *cheminformatics*, *chemical informatics*, *chemi-informatics*, and *molecular informatics*; and more recently our group at The Institute of Cancer Research is called *In Silico Medicinal Chemistry*.

1.5. The Similar-Structure, Similar-Property Principle

Much of chemoinformatics is essentially based on the fundamental assertion that similar molecules will also tend to exhibit similar properties; this is known as the *similar-structure*, *similar-property principle*, often simply referred to as the *similar-property principle*, and described for molecules by Rouvray [Johnson and Maggiora 1990, page 18], thus: ". . . the so-called principle of similitude, which states that systems constructed similarly on different scales will possess similar properties."

In large part, this is true; however, it must be emphasized that the similar-property principle is a heuristic and therefore will break down in certain circumstances. Another important caveat is that there are many ways of defining similarity (see Section 6).

2. CHEMISTRY AND GRAPH THEORY

The fields of chemistry and *graph theory* have a long-standing partnership from the mid-eighteenth century onwards. It is well known to many computer scientists that graph theory was first formalized as a mathematical abstraction system by Leonhard Euler in 1736. However, what is perhaps less well known is that the name for the field came directly from efforts in chemistry to define a pictorial representation of molecules after the development of atomistic theory.

2.1. Chemistry and the Origins of Graph Theory

A form of graph theory was first applied by William Cullen (1710–1790) and Joseph Black (1728–1799) in Edinburgh in the 1750s. The nodes of these graphs were molecular substances while the edges defined an affinity value between each of the compounds. Over a century later efforts were made to represent molecular structures using the new
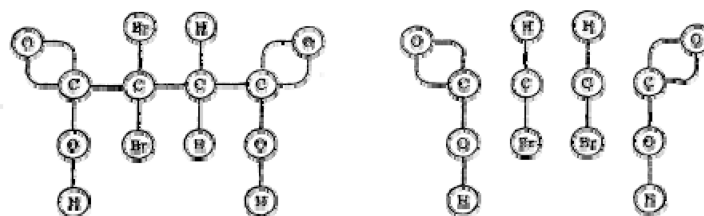
Fig. 3. Two examples of the molecular graphic notations by Alexander Crum Brown; note that the hydrogen atoms are explicit in these diagrams.

atomistic theory and this was where the most substantial interest in the mathematical abstractions from Euler became apparent. Numerous chemists worked to formalize representation systems of molecules. Arthur Cayley (1821–1895) used graph-like structures in his studies on enumerating the isomers of alkanes. However, two scientists in particular provided the field of graph theory with the beginnings of a formal name: Alexander Crum Brown (1838–1922) and James Joseph Sylvester (1814–1897). Crum Brown developed the constitutional formulae in 1864 representing atoms as nodes and the bonds as edges [Crum Brown 1864]. Crum Brown was at pains to state that these graphs represented abstractions of molecules in that they were not intended to be accurate representations of real molecules, but merely to illustrate the relationships between the atoms (Figure 3). Sylvester also developed his very similar representation (Figure 4) around the same time as Crum Brown. Crum Brown referred to his structures as *molecular graphic notations* while Sylvester called his *chemicographs*. While it is difficult to say with any certainty which scientist applied the graph term first, it is incontrovertible that both Crum Brown and Sylvester, both chemists, at least assisted in the development of a new name for the field: graph theory.

The field of graph theory continued in its own right as a field of mathematics resulting in the very active field we know today. However, chemistry had also only just begun its foray into graph theory. Almost a century after the research from Crum Brown and Cayley, the broader field of mathematical chemistry emerged in its own right, with this field applying mathematics in an effort to understand chemical systems and make predictions of molecular structure.

2.2. Graph Theory and Chemoinformatics

Graph theoretic techniques are widely applied in computer science; however, it is prudent here to provide a brief overview of graph theory and the terms and standards used in this article [Diestel 2000]. A graph G is a collection of objects $V(G)$ and the relationships between those objects $E(G)$ called *nodes* (or *vertices*) and *edges* (or *arcs*), respectively. In the context of chemoinformatics, the nodes are the atoms of a molecule and the edges are the bonds. The nodes in G are connected if there exists an edge ($\mathbf{v}_i$, $\mathbf{v}_j$) $\in E(G)$ such that $\mathbf{v}_i \in V(G)$ and $\mathbf{v}_j \in V(G)$. The order of a graph G is given by the size of $|V(G)|$. A node $\mathbf{v}_i$ is incident with an edge if that edge is connected to the node, while two nodes, $\mathbf{v}_i$ and $\mathbf{v}_j$, are said to be adjacent if they are connected by the edge ($\mathbf{v}_i$, $\mathbf{v}_j$) $\in E(G)$. Two edges are said to be incident if they have a node in common. A complete graph is where every node is connected to every other node in the graph. The edge density of a graph can then be calculated as the number of edges in a particular graph normalized between the number of edges in a connected graph ($|V(G)| - 1$), and the number of edges in the complete graph ($|V(G)| \cdot (|V(G)| - 1) / 2$), with the given
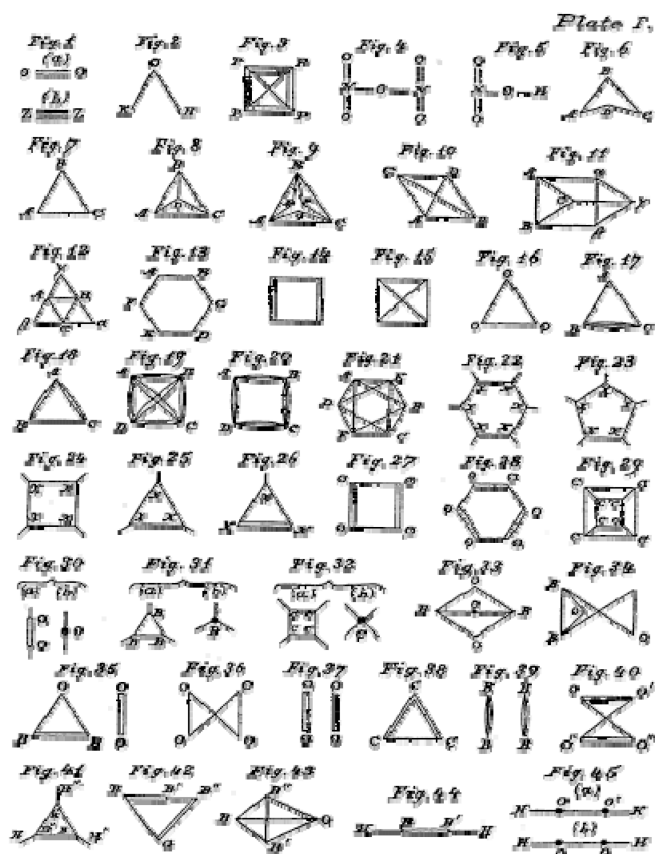
Fig. 4.   Examples of the chemicograph representations of molecules by James Joseph Sylvester.

number of nodes, $|V(G)|$:

$$\frac{2 \cdot |E(G)|}{|V(G)| \cdot (|V(G)| - 1)}. \tag{1}$$

One of the most important applications of graph theory to chemoinformatics is that of graph-matching problems.

It is often desirable in chemoinformatics to determine differing types of structural similarity between two molecules, or a larger set of molecules. Two graphs are said to be *isomorphic* when they are structurally identical. Subgraph isomorphism of G_1, G_2 holds if G_1 is isomorphic to some subgraph in G_2. On the other hand, the identification of the maximum common subgraph between two graphs is the determination of the largest connected subgraph in common between the two. Last, the maximum overlap set is the set of the, possibly disconnected, largest subgraphs in common between two graphs. In chemoinformatics, the term *structure* is often used in place of graph. These graph-matching problems are NP-complete and therefore numerous methods have been applied to prune the search tree. These methods use alternative representations. Graph-matching algorithms also have application in pharmacophore searching and docking; see Sections 4.5 and 7.3 respectively.
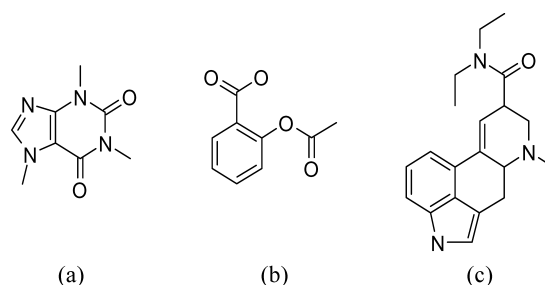
Fig. 5. The hydrogen-depleted molecular graphs of (a) caffeine, (b) aspirin, and (c) D-lysergic acid diethylamide.

The molecular graph is a type of graph that is undirected and where the nodes are colored and edges are weighted. The individual nodes are colored according to the particular atom type they represent carbon (C), oxygen (O), nitrogen (N), chlorine (Cl), etc., while the edges are assigned weights according to the bond order single, double, triple, and aromatic. Aromaticity is an especially important concept in chemistry. An aromatic system, such as the benzene ring, involves a delocalized electron system where the bonding system can be described as somewhere between single and double bonds, as in molecular orbital (MO) theory [Bauerschmidt and Gasteiger 1997]. In the case of the benzene ring—a six-member carbon ring—six π electrons are delocalized over the entire ring. A common approach to representing an aromatic system in a computer is to use resonant structures, where the molecule adopts one of two bonding configurations using alternating single and double bonds. However, this is an inadequate model for the representation of aromaticity and therefore the use of an aromatic bond type is also used. Molecular graphs also tend to be hydrogen depleted, that is, the hydrogens are implicitly represented in the graph since they are assumed to fill the unused valences of each of the atoms in the molecule. Each atom is ascribed a particular valence that is deemed at least to be indicative of the typical valence of the molecule: carbon has a valence of 4, oxygen has 2, and hydrogen has 1. The molecular graph representations of (a) caffeine, (b) aspirin, and (c) D-lysergic acid diethylamide are provided in Figure 5.

3. MOLECULAR REPRESENTATIONS

Molecules are complicated real-world objects; however, their representation in the computer is subject to a wide range of pragmatic decisions based largely on the domain of interest to which the data structures are to be applied, but also decisions that were made according to the availability of computational resources. There is a hierarchy of molecular representations with each point in the hierarchy having its own domain of applicability.

In chemoinformatics, the most popular representation is the two-dimensional (2D) chemical structure (*topology*) with no explicit geometric information. However, even these objects necessitate the application of highly complex computational algorithms to perform comparisons and transforms.

As a molecule of significant importance to computer science, various representations of caffeine have been provided in Figure 6. The representations range from the general name itself (*caffeine*), through an arbitrary number assigned to the molecule (*CAS Registry Number*), a typical database record entry in *Simplified Molecular Input Line*

Representation	Name
Caffeine	Common Name
trimethylxanthine, theine, mateine, guaranine, methyltheobromine	Synonyms
$C_8H_{10}N_4O_2$	Empirical Formula
1,3,7-trimethylpurine-2,6-dione	IUPAC Name
58-08-2	CAS Registry Number
T56 BN DN FNVNVJ B F H	WLN
CN1C=NC2=C1C(=O)N(C(=O)N2C)C	SMILES
CN1C(=O)N(C)c2ncn(C)c2C1=O	SMILES (Aromatic)
1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3	InChI

```
       C 0 0 0 1 0 0 0 0 0 0 1 0 2
       0 C 0 0 0 0 0 0 0 0 1 1 2 0
       0 0 C 0 2 0 0 0 0 1 1 0 0 0
       0 0 0 C 0 0 0 0 1 2 0 0 0 0
       1 0 2 0 C 0 0 0 1 0 0 0 0 0
       0 0 0 0 0 C 0 0 0 0 1 0 0 0
       0 0 0 0 0 0 C 0 0 0 0 1 0 0
       0 0 0 0 0 0 0 C 1 0 0 0 0 0
       0 0 0 1 1 0 0 1 N 0 0 0 0 0
       0 0 1 2 0 0 0 0 0 N 0 0 0 0
       0 1 1 0 0 1 0 0 0 0 N 0 0 0
       1 1 0 0 0 0 1 0 0 0 0 N 0 0
       0 2 0 0 0 0 0 0 0 0 0 0 O 0
       2 0 0 0 0 0 0 0 0 0 0 0 0 O
```

Adjacency Matrix

```
Caffeine
  Comment Line

 14 15  0  0  0  0              999 V2000
    3.0312   -2.1688    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.7457   -0.9312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3168   -0.9312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0473   -1.3437    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.3168   -1.7563    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0312    0.3063    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.4602   -2.1688    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2773   -2.7958    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.5322   -2.0112    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    1.5322   -0.6763    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.0312   -0.5187    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.7457   -1.7563    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    4.4602   -0.5187    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    3.0312   -2.9938    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
 11  3  1  0  0  0  0
 11  2  1  0  0  0  0
  5  1  1  0  0  0  0
```

Connection Table (SDF)

Fig. 6. Some of the many ways in which molecules can be represented from simple names, empirical formulae, and line notations, through to computational models of their structure.

*Entry Specification* (SMILES) and *Structure Data Format* (SDF), and on to graph-based and geometric-based models of the molecule itself.

However, between the explicit and implicit hydrogen models of molecules, there exists the polar hydrogen model that includes those hydrogens that are likely to be in long-range hydrogen bonds, and is more frequently used in molecular mechanics (MM) applications.
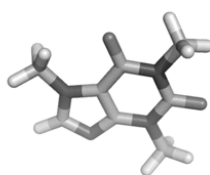
```
  1 12  1  0  0  0  0
 12  2  1  0  0  0  0
  3 10  1  0  0  0  0
  4  9  1  0  0  0  0
  4 10  2  0  0  0  0
  9  5  1  0  0  0  0
  5  3  2  0  0  0  0
 11  6  1  0  0  0  0
  2 13  2  0  0  0  0
 12  7  1  0  0  0  0
  1 14  2  0  0  0  0
  9  8  1  0  0  0  0
M  END
$$$$
```
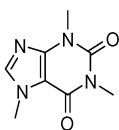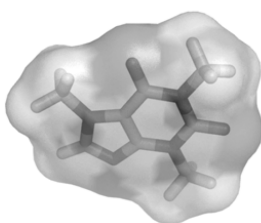
Topology
(2D Structure)

Topography
(3D Structure)

Topography
(Surface Model)

Fig. 6. Continued.

3.1. Adjacency Matrix

A common method of molecular representation is the adjacency matrix (AM), which is primarily used as a representation in structure processing, but may also be used for structure representation in files. The AM is a square matrix with dimensions that are equal to the number of atoms in the given molecule, N. Each position in the AM specifies the presence or absence of a bond connecting those two atoms, i and j, respectively. Since molecular graphs do not contain recursive edges, nodes that are self-connected, the diagonal elements of an AM are always, zero denoting no connection. Typical connections that are represented in the AM are: 1, 2, and 3 for single, double, and triple bonds, respectively. Additional bond types are represented in different systems often with the extension to 4 and 5, for amide and aromatic bonds, respectively. The atom identifiers tend to be stored in a separate vector, thereby allowing access to the complete topology of the molecule. Additional information may also be stored in this manner, recording spatial locations and charge information.

The AM itself is symmetrical and therefore redundant, that is, $\mathbf{AM}_{ij} = \mathbf{AM}_{ji}$. However, the use of a more complicated data structure avoids storing this redundancy. The AM for caffeine is given in the tenth example of Figure 6.

molecule name

number of atoms

```
                     Caffeine
                     Comment Line                        number of bonds
                    14 15  0  0  0                     999 V2000
                        3.0312    -2.1688    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        3.7457    -0.9312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        2.3168    -0.9312    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        1.0473    -1.3437    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        2.3168    -1.7563    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        3.0312     0.3063    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        4.4602    -2.1688    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        1.2773    -2.7958    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
                        1.5322    -2.0112    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
                        1.5322    -0.6763    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
                        3.0312    -0.5187    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
                        3.7457    -1.7563    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
                        4.4602    -0.5187    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
                        3.0312    -2.9938    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
                    11   3  1  0  0  0  0
                    11   2  1  0  0  0  0
                     5   1  1  0  0  0  0
                     1  12  1  0  0  0  0          atom information: first 3 real numbers
                    12   2  1  0  0  0  0          give x, y, and z, co-ordinates,
                     3  10  1  0  0  0  0          respectively, followed by the atom
                     4   9  1  0  0  0  0          type
                     4  10  2  0  0  0  0
                     9   5  1  0  0  0  0
                     5   3  2  0  0  0  0          bonding information: first two numbers
                    11   6  1  0  0  0  0          give the source and target atoms,
                     2  13  2  0  0  0  0          while the third number is the bond
                    12   7  1  0  0  0  0          order
                     1  14  2  0  0  0  0
                     9   8  1  0  0  0  0
                    M   END
                    $$$$
```

Fig. 7. The connection table representation of caffeine, with the most pertinent information regarding the structure highlighted.
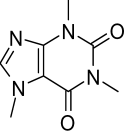
3.2. Connection Table

An alternative representation of a molecular graph is the connection table (CT), which is the preferred representation for the most common file formats such as SDF, MOL2, and, more recently, CML (Chemical Mark-up Language). The CT separates the atom and bond information into two blocks. The atom block details the atomic type together with additional information. This additional information can define *chiral centre* information for molecules that are structurally identical but can adopt geometries that are mirror images of each other, 2D or 3D coordinate data, and *charge* information. Each of the atoms must be assigned a unique identifier for the correct referencing of the edges in the graph, and this is quite often based simply on the line number in the atom block. The bond block contains the typical information regarding the particular bond type, but also the unique node identifiers are used to specify the atoms that are connected by that particular bond.

For typical usage the most popular molecular file formats that use connection tables as their data representation are SDF, from Symyx Technologies,[1] formerly Molecular Design Limited (MDL), and SYBYL MOL2, from Tripos Inc.[2] However, many alternatives exist and tend to arise due to proprietary software systems from chemoinformatics software providers. One exception that is rising in popularity is the CML format based on XML (eXtensible Mark-up Language). However, for this article, we will consider only the SDF format, although this is very similar to most of the other file formats. The SDF of the caffeine molecule is given in Figure 7.

---

[1] www.symyx.com.
[2] www.tripos.com.

Name	Topology	SMILES
Methane	C	C
Ethanol		CCO
Acetic Acid		CCC(=O)O
Cyclohexane		C1CCCCC1
Benzene		C1=CC=CC=C1 c1ccccc1
Cubane		C12C3C4C1C5C4C3C25
Caffeine		CN1C=NC2=C1C(=O)N(C(=O)N2C)C CN1C(=O)N(C)c2ncn(C)c2C1=O

Fig. 8. Some examples of simple molecules, their topologies, and the corresponding SMILES strings that explain connectivity, branching, and ring systems. In the fifth and seventh instances, alternative SMILES representations are given for representing aromatic systems.

3.3. Line Notations

A number of approaches have been suggested over the years to encode molecular structures as simple line notations, mainly for reasons of data storage capacity. The first such line notation was the Wiswesser Line Notation (WLN). A number of alternative approaches have been proposed over time such as Representation of Organic Structures Description Arranged Linearly (ROSDAL), and SYBYL Line Notation (SLN). However, perhaps the most elegant molecular line notation is the SMILES language [Weininger and Smiles 1988] and this is the line notation that we will explore in more detail here due to its simplicity and widespread adoption. Some examples of SMILES codes for some simple molecules are given in Figure 8. The simple topological encoding system relies on several simple rules:

(1) Atoms are represented by their atomic symbols, usually in upper-case.
(2) Bonds are represented by the characters "−," "=," "#" and ":" for single, double, triple, and aromatics bonds, respectively. Single bonds may also, and generally, remain implicit.
(3) Branches are encoded by round parentheses surrounding the branching fragment and these may be nested or stacked.
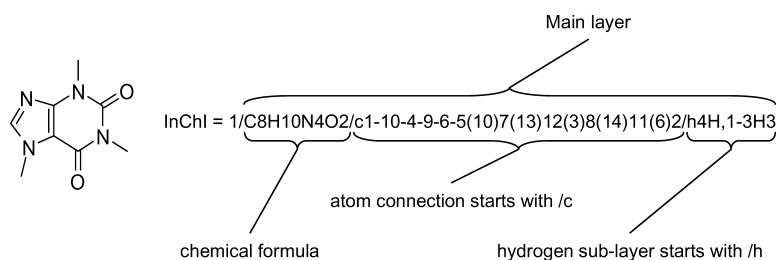
Fig. 9. Example of the InChI code for caffeine with the chemical formula, atom connection, and hydrogen sublayers of the main layer.

(4) Cycles are represented by a common number following the atoms that connect to form the cycle.

(5) In aromatic systems the atomic characters that are part of the aromatic systems are written in lower-case.

(6) Last, since single bonds are implicit, it is necessary for a disconnection to be encoded explicitly with the full-stop or period (".") character.

One issue that quickly arose with the SMILES notation was the lack of a unique representation since a molecule can be encoded beginning anywhere in addition to other limitations with regard to which path to take in encoding the molecule. From Figure 8, ethanol could be encoded as the following four SMILES strings, each of which is valid: CCO, OCC, C(C)O, and C(O)C. This limited the application of SMILES as a unique identifier in database systems. Therefore, a method of encoding a molecule was quickly developed that provided an invariant SMILES representation. The Morgan algorithm [Morgan 1965] was proposed in the 1960s to provide a canonical ordering of atoms in a molecule for just such an application: indexing in database systems. The Morgan algorithm proceeds by assigning values to each of the atoms of a molecule iteratively, based on their extended connectivities; initially assigned values are based on the node degree of each of the atoms, excluding hydrogens. The node coding and partitioning approaches in the Morgan algorithm are analogous to the methods applied in most graph and subgraph isomorphism algorithms.

Recent developments in line notations are the InChI (International Chemical Identifier) codes, supported by the International Union of Pure and Applied Chemistry (IUPAC), which can uniquely describe a molecule in a very compact form (Figure 9), but is not intended for readability [Adam 2002; Coles et al. 2005]. The InChI code of a molecule encodes that molecule in a series of six layers: main, charge, stereochemical, isotopic, fixed-H, and reconnected. Each of these layers can be further split into sublayers, but no further splitting of layers is permitted. The main layer can be split further into chemical formula, atom connections, and hydrogen atom sublayers; the main layer and the chemical formula sublayer are the only two mandatory components of every InChI code. The InChI code system was developed to be an open identifier for chemical structures primarily for printed and electronic publications to allow the data to be captured for subsequent indexing and searching. The sublayers are used to discriminate unique molecules that are otherwise not distinct when using alternative representation methods such as can be the case with SMILES.

3.4. Generation of Molecular Geometries

The automated computational generation of a geometric structure of a molecule, called a *conformer* or *conformation*, is desired in many applications in chemoinformatics. Since

molecules are 3D objects, it would seem that one of the standard 3D graph-layout algorithms would suffice; however these layouts do not necessarily take into account chemical knowledge. Therefore, a number of programs have been developed that can generate a single conformer, or multiple possible conformers (since molecules take different conformers depending on their environment), that represents a *minimum-energy conformation*—a conformation in which the geometric arrangement of atoms leads to a global minimum in the internal energy of the system. Two of the most popular programs for this purpose are Concord [Pearlman 1987], and CORINA (COoRdINAtes) [Gasteiger et al. 1990]. Both of these programs operate in a similar way through the application of chemical knowledge.

The CORINA program, for example, has a number of rules for bond angles and lengths based on the atom type involved and its particular *hybridization* state. Rings are considered individually, with particular conformations being selected from ring conformation libraries that have been generated from mining the information in crystallographic databases. Pseudo-force-field calculations and the removal of overlapping atoms are then performed to clean the resultant conformers.

The resultant low-energy conformations returned by CORINA have been demonstrated to have low root-mean-square deviation (RMSD) values when compared with X-ray crystal structures from the Cambridge Structural Database (CSD) from the Cambridge Crystallographic Data Centre (CCDC).

Many of the molecular geometry generator programs also permit the generation of multiple low-energy conformations of a single molecule, allowing the user to select the preferred conformation for their particular application, or in some way combining all the conformations and then applying this combined description. However, although intuitively it is expected that 3D will be superior to 2D or topological representations, this has not been shown to be the case in many instances.

3.5. 2D Versus 3D Molecular Representations

Much of the initial work conducted in the field focused on topological methods and analyses. This was considered to be a pragmatic solution due to the paucity of computational resources at the time of development. It was anticipated that, as processing speeds increased, over time methods would be developed using molecular geometry that were far superior to the topological approaches. However, this has been demonstrated not to be the case, with topological methods still out-performing geometric methods in standard cases [Willet 1991].

The main reason for the reduction in efficacy of geometric methods is generally accepted to be the conformer problem. A conformer or conformation of a molecule is a single geometric arrangement of atoms in a molecule. However, a molecule may adopt infinite conformations and without sufficient coverage of this space, or knowledge of the region of conformer space of most interest, the resulting analyses will be flawed. Topological methods, however, implicitly encapsulate this information in the 2D structure of the molecules being considered. In general, when considering only a single conformer of a given molecule in analyses, its use can often introduce only noise into the system [Brown and Martin 1997].

However, this is not to say that 3D structure is not important, only that we have still to develop computational methods to deal with these structures in a way that properly takes into account the *conformational flexibility* of molecules in a way that is most effective. Conformational flexibility can be described in a trivial way by a simple count of the number of *rotatable bonds* in a molecule, where a rotatable bond is a single, nonring bond between two nonterminal heavy atoms (i.e., not hydrogen). Indeed, methods such as Comparative Molecular Field Analysis (CoMFA) [Cramer et al. 1988]

have been demonstrated to be effective for the predictive modeling of congeneric series of molecules based on molecular alignments—whether these are aligned by hand or by an automated process. A congeneric series of molecules is one in which each molecule contains a significant amount of structural similarity to permit meaningful alignments.

4. MOLECULAR DESCRIPTORS

The generation of informative data from molecular structures is of high importance in chemoinformatics since it is often the precursor to permitting statistical analyses of the molecules. However, there are many possible approaches to calculating informative molecular descriptors [Karelson 2000; Todeschini and Consonni 2000].

When discussing molecular descriptors, one is reminded of the parable of the blind men and the elephant by John Godfrey Saxe. In this poem, six blind men endeavor to describe an elephant and variously determine that the elephant reminds them of a rope (tail), tree (leg), snake (trunk), spear (tusk), fan (ear), and wall (body). This emphasizes the local context information to which the blind men are privy. Essentially, in only considering selected aspects of the elephant, the overall description of the elephant is not forthcoming. In this way, molecules are similar to elephants since they contain many features that in themselves are not particularly informative, but considered in combination provide a rich characterization of the object under study.

4.1. Information Versus Knowledge

Molecular descriptors are descriptions of molecules that aim to capture the salient aspects of molecules for application with statistical methods. Many different molecular descriptors are available and come from a range of distinct descriptor classes. The two main classes of these are knowledge-based and information-based descriptors. Knowledge-based descriptors tend to describe what we expect, whereas information-based descriptors describe what we have. Examples of knowledge-based descriptors are those that calculate molecular properties based on extant knowledge or models based on such data. Information-based descriptors, however, tend to encapsulate as much information as is possible within a particular molecular representation. In this section particular classes of molecular descriptors are discussed together with examples of these types of descriptors. In addition, the situations in which these different descriptors are best applied are discussed, with particular regard given to levels of interpretability and efficacy.

4.2. Topological Indices

The class of molecular descriptors referred to as *topological indices* (TIs) calculate a scalar variable based on the topology of a hydrogen-depleted molecular graph [Balaban 1985]. This type of descriptor is essentially an information-based approach since only the structural information is used in generating the description. This reduces the interpretability of the descriptor, but can lend to the efficacy in application. Many variants of TIs have been published in the literature; here two of the more well known are discussed.

4.2.1. Wiener Index. The Wiener index W is calculated as the sum of the number of bonds between all nodes in a molecular graph, G. The shortest path is used to determine the number of edges between the nodes and therefore the Floyd or Dijkstra algorithms to calculate shortest paths between nodes are typically applied. The W index is calculated

thus:

$$W = \sum_{i=2}^{N} \sum_{j=1}^{i} \mathbf{D}_{ij}, \tag{2}$$

where N is the size of the molecule in atoms and $\mathbf{D}_{ij}$ is the shortest path distance between atoms i and j.

4.2.2. Randić Index. This topological index was developed by Randić and characterizes the branching of a given molecule. It is also referred to as the *connectivity* or *branching index*. The index is calculated by multiplying the product of the node degrees, δ, of the nodes that are incident with each of the edges in the graph:

$$R = \sum_{b=1}^{B} \frac{1}{\sqrt{(\delta_i \cdot \delta_j)_b}}. \tag{3}$$

4.3. Physicochemical Properties

This class of descriptors attempts to provide estimations of physical properties of molecules using predictive modeling techniques. Physicochemical properties provide values that are intuitive to chemists and thereby permit informed decisions to be made. Perhaps the most widespread use of physicochemical descriptors is through their application as heuristics using the so-called Lipinski Ro5 where molecules can be filtered according to their satisfaction of these rules. The physicochemical descriptors considered by the Ro5 are molecular weight (MW), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), and ClogP [Raevsky 2004]. Based on historical data, it was deemed that oral drugs would tend to have a MW of less than 500 daltons (1 Da $\approx$ 1.661 $\times$ 10^{-27} kg or one-twelfth of the mass of one atom of carbon-12), less than 10 HBAs, less than five HBDs, and a logP of less than 5. However, these are heuristics and rigorous adherence to rules such as these should not be encouraged.

One can calculate each of these parameters *in silico*. MW is simply a summation function according to the numbers and types of atoms that are present in the molecule under consideration. A trivial way of calculating the numbers of HBAs and HBDs is by the numbers of nitrogen (N) and oxygen (O) atoms and NH and OH groups, respectively. However, more complicated methods also exist such as the sum of F, N, and O atoms in a molecule—excluding N atoms with formal positive charges, in higher oxidation states, and in the pyrrolyl form of N for HBA—and the sum of hydrogens attached to all the N and O atoms in a molecule for HBD as coded in the Dragon software.[3] The calculation of logP is somewhat more complicated, as described below.

The computer-generated or calculated logP (ClogP) descriptor is one of the most frequently used physicochemical descriptors and predicts the logarithm of the partition coefficient between octanol and water. The logP value provides an indication of solubility of a molecule, but does not account for ionizable species. It is an important parameter in the design of NCEs since it indicates the general lipophilicity (or hydrophobicity). The most common method of calculating logP is through a combination of logP values that have been measured for individual molecular fragments. In early work this was effective only for series that were *congeneric* (part of the same chemical series or class), but would tend to give greater prediction errors when considered across chemical series [Fujita et al. 1964]. Alternatively, atom-additive schemes are also used where each atom

---

[3]http://www.talete.it.

is given a contribution factor with the ClogP calculated as the sum of the products of the number of each particular atom and its contribution [Ghose and Crippen 1986].

Other common physicochemical descriptors that can be calculated *in silico* to varying degrees of accuracy are pK_a (acid dissociation constant), logD (octanol-water distribution, as opposed to partition, coefficient), logS (aqueous solubility, or logW $logS_w$ (water solubility)), and the PSA (polar surface area). A review of these physicochemical properties has been given by Raevsky [2004].

### 4.4. Vector Representations (Molecular Fingerprints)

Subgraph isomorphism (called *substructure searching* in chemoinformatics) in large molecular databases is quite often time consuming to perform on large numbers of structures given that it is an NP-complete (nondeterministic polynomial time) problem. It was for this reason that substructure screening was developed as a rapid method of filtering out those molecules that definitely do not contain the substructure of interest.

4.4.1. Structure-Key Fingerprints. The structure-key (or dictionary-based) fingerprint uses a dictionary of defined substructures to generate a binary string where each bit in the string equates to a one-to-one mapping between the molecule and a substructure in the dictionary. Therefore, these types of fingerprints can be interpreted quite easily by using the dictionary as a lookup table. Essentially, a bit is set in the fingerprint if the substructural key it represents is present in the molecule that is being encoded. Augmented atoms are those that are seeded from a central atom together with its neighboring atoms. Atom sequences are linear paths of atoms either with or without specific bond type information. Bond sequences are similar to atom sequences, but with atom typing either present or not. In addition, substructural fragments may be included as keys, such as ring systems and functional groups. A number of alternative substructure dictionaries are available; the set of 166 keys used by Elsevier MDL, Inc., were recently published [Durant et al. 2002].

Structure-key fingerprints are a form of knowledge-based descriptor since the dictionaries are designed according to knowledge of extant chemical entities and in particular what is expected to be of interest to the domain for which the dictionary was designed. In the design of structure-key dictionaries it is important to ensure that the substructures that are included are somewhat orthogonal or independent of each other while also being fairly equifrequent in their occurrence. This last aim is to ensure that structure keys are included such that they neither occur too rarely (so as to be an irrelevance in general) nor too frequently (such that their presence is invariant). A schematic example of the generation of a structure-key fingerprint is provided in Figure 10.

One aspect that can sometimes be significant with structure-key fingerprints is that certain molecules may result in a fingerprint that contains very little information, or theoretically none at all, since none of their substructural elements occurs in the dictionary. This degree of brittleness can lead to problems when applying these descriptors to novel chemical classes in that their efficacy is reduced.

4.4.2. Hash-Key Fingerprints. Hash-key fingerprints are generated in a manner that is quite different from structure-key fingerprints since there is no predefined database of chemically interesting molecular substructures; rather the keys are generated from the molecule itself. The typical hash-key fingerprint enumerates multiple atom-bond paths from each molecule, generally between a path length in bonds of 0 to 7. Each of these paths is then used as the input to a hash function—such as a Cyclic Redundancy Check (CRC) algorithm—to generate a large integer value. This integer then may be folded using modulo arithmetic such that it conforms to the length of the binary string
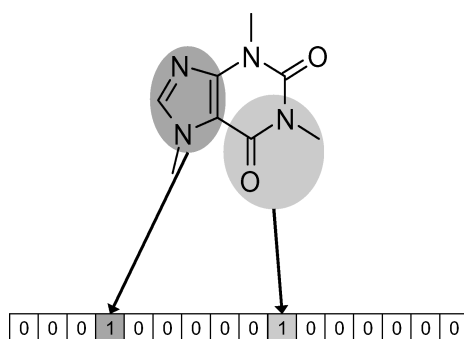
Fig. 10. An example of the encoding of a simple molecule as a structure-key fingerprint using a defined substructure or fragment dictionary. A defined fragment is assigned a single bit position on the string to which it, and no other fragment, is mapped.

being used to represent the molecule. Alternatively, the output from the CRC can be used as a seed for a random number generator (RNG) and a number of indices (typical 4 or 5) being taken from the RNG, again using modulo arithmetic, to be mapped to the fingerprint being generated. The rationale for the use of the RNG is to reduce the effect of different molecular paths colliding at the same index in the fingerprint. Since each path is now represented by four or five indices, the probability of another molecular path exhibiting precisely the same bit pattern is vastly reduced. A schematic example of the generation of a hash-key fingerprint is provided in Figure 11.

A recent advance in hash-key fingerprints has been provided by SciTegic in their PipelinePilot workflow software [Brown et al. 2005]. In this configuration, circular atom environments are enumerated, rather than atom paths, with these being canonicalized using an algorithm such as that from Morgan as described previously, providing a unique representation that acts as a key. Although circular substructures, or atom environments, were first used in the 1970s for screening systems, this recent innovation has provided a new type of molecular fingerprint descriptor that has been demonstrated to be of great application in similarity searching [Hert et al. 2004; Rogers et al. 2005]. An example of the enumeration of atom environments for a single atom is provided in Figure 12 for bond radii from the root atom of 0, 1, 2, and 3, respectively.

Although apparently quite simplistic in design, molecular hash-key fingerprint algorithms have been demonstrated to be highly effective in encapsulating molecular information, which is evident in the widespread application to many challenges in chemoinformatics. The Fingerprinting Algorithm (Fingal) descriptor [Brown et al. 2005] is one such recent example of a hash-key fingerprint descriptor and is based substantially on the fingerprints from Daylight Chemical Information Systems, Inc.,[4] while also being extended to encapsulate additional structural information, such as molecular geometries in an alignment-free way. Fingal is additionally capable of encoding the Euclidean distance between the current atom in the path and all of the previous atoms in the path thus far.

Hash-key fingerprints are very rapid to calculate and encapsulate a great deal of the information necessary for them to be effective in many applications in chemoinformatics. However, they do have significant limitations and are not universally applicable. The resultant descriptors are highly redundant and, more significantly, they are not readily interpretable. However, the latter issue can be overcome to a large extent with a

---

[4]http://www.daylight.com.
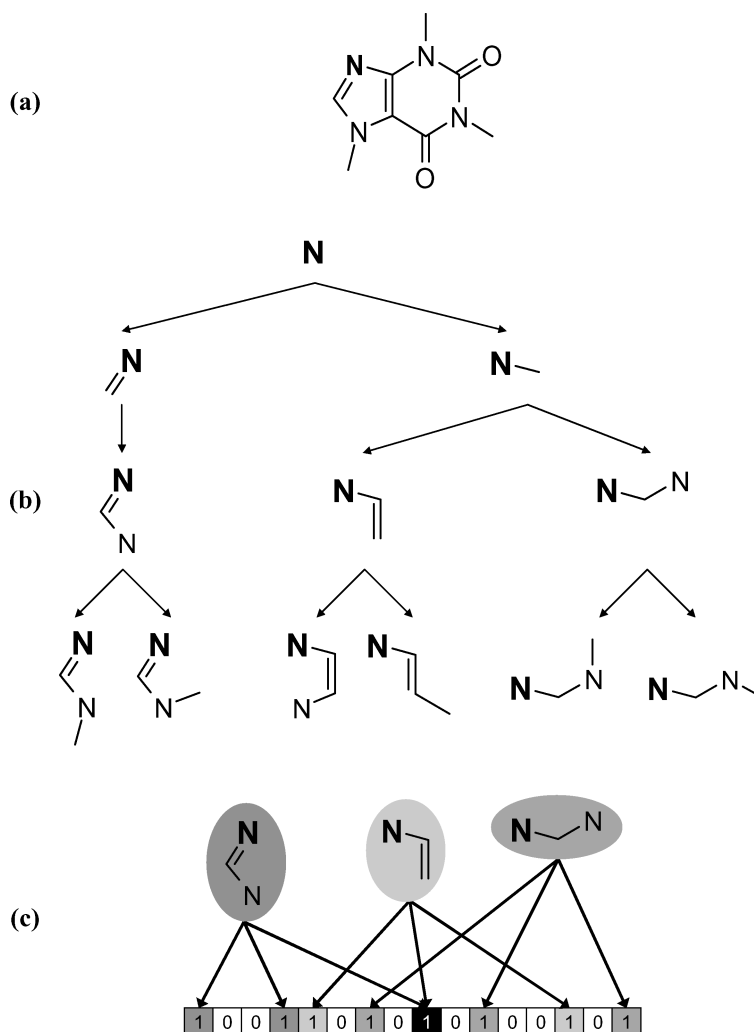
Fig. 11. A partial example of the encoding of caffeine as a hash-key fingerprint. The original structure caffeine (a) with the root nitrogen atom for this particular path enumeration highlighted in bold. The enumeration (b) of paths from the root atom up to three bonds away represented as a tree. Each of the paths in all levels of the enumerated path tree is then converted into one or more integer values using a hashing algorithm and a pseudorandom number generator to give n bit positions (3 in this case) that are set for each of the paths, here shown only for level 3, of the tree in (c); an instance of a "bit collision" is highlighted.

certain memory overhead required to store the molecular fragments subject to a degree of fuzziness brought on the hashing approach.

4.4.3. On the Reversibility of Molecular Descriptors.   Molecular descriptors, by their nature, characterize aspects of the molecules being described such that pertinent information is encapsulated. Therefore, it follows that, given a single molecule, the representation of that molecule in a particular descriptor space could be used to reverse engineer the descriptor to that specific structure [Oprea 2005b]. However, this all depends on the particular descriptors used. One would expect that a small set of physicochemical descriptors would provide a greater number of collisions in descriptor space such that it
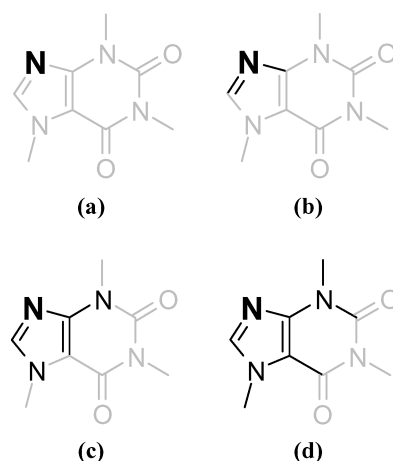
Fig. 12. An example of the enumeration of atom environments (augmented atom or circular substructure) of caffeine at bond radii (a) 0, (b) 1, (c) 2, and (d) 3, respectively, highlighted in black, with the remaining portion of the molecule in gray.

is difficult to map back to a single structure, although the particular ligands (specific molecules that bind to a protein to evoke a response) that are interesting would most likely still be apparent. The use of an information-rich descriptor such as hash-key fingerprints that do their best to encode as much information as provided will tend to map back only to a single molecular entity; the lack of frequency of occurrence information in binary fingerprints would limit this ability, but it is still generally possible to map back to the region of chemistry space. Alternatively, the use of integer fingerprints permits a more accurate mapping back from the descriptor to the structure space and has been used for this purpose in *de novo* design [Brown et al. 2004]; see Section 5 for more information on *de novo* design.
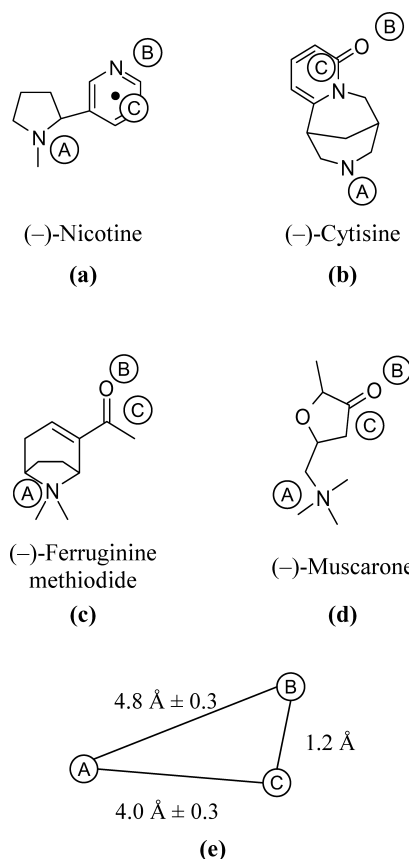
4.5. Pharmacophore Models

A *pharmacophore* is a pharmaceutically relevant arrangement of molecular features that are potential interaction points. These features could be particular molecular substructures of interest or merely a desired interaction potential such as an *electrostatic potential*. The term *pharmacophore* was first used by Paul Ehrlich (1854–1915) in 1909 to refer to a molecular framework that carries (*phoros*) the essential features responsible for the biological activity of a drug (*pharmacon*) [Güner 2005]. This was later refined by Gund [1979], page 299, as follows:

> . . . a set of structural features in a molecule that is recognized at a receptor site and is responsible for the molecule's biological activity.

There are two distinct types of pharmacophores: structure based and ligand based. The structure-based pharmacophores use information from a protein binding site or a bound conformation or crystallized structure of a ligand to derive a pharmacophore model, whereas ligand-based pharmacophores consider only a given topology of a molecule of interest and attempt to derive the salient information from that using various molecular characterization methods including non-target-specific conformations.

4.5.1. Structure-Based Pharmacophores. When information is present regarding the particular geometric constraints of a biological target (a protein target site to which

(−)-Nicotine

(a)

(−)-Cytisine

(b)

(−)-Ferruginine
methiodide

(c)

(−)-Muscarone

(d)

4.8 Å ± 0.3

1.2 Å

4.0 Å ± 0.3

(e)

Fig. 13. Four molecules: (a) (−)-nicotine, (b) (−)-cytisine, (c) (−)-ferruginine methiodide, and (d) (−)-muscarone used to derive the nicotinic pharmacophore by distance geometry and (e) the pharmacophore model obtained. The pharmacophoric points of each of the molecules are given by the letters A, B, and C, respectively, and equate to the features in the pharmacophore model in (e). Adapted from Leach [2001].

a ligand is sought to bind), this information can be used to develop a spatial arrangement of desirable molecular properties to describe the desired molecule. This pharmacophore model can then be used to search against a 3D library of molecules with the aim of retrieving molecules that will exhibit a similar pharmacophoric arrangement of features and therefore be more likely to also exhibit the same biological activity, which is concomitant with the similar-property principle.

Typically, 3-point pharmacophore descriptions are used—although 4-point pharmacophores have been published—and can be designed by hand using knowledge, or derived automatically from given information. The distances between each of the feature points are specified in Ångströms (Å, or 10^{-10} m) but also permitted to be within some tolerance to permit a fuzzy matching with given molecules. An example of a pharmacophore for the nicotinic receptor is provided in Figure 13.

This approach is related to the concept of *bioisosterism*, where molecules or fragments are said to be bioisosteric if they present a biologically important and similar arrangement of features to the target. Bioisosterism is of great interest since replacing substituents of molecules with bioisosteres can assist in improving potency against the target, but also assist in designing out off-target effects that can cause undesired

0010-4-1100-6-0100-6-
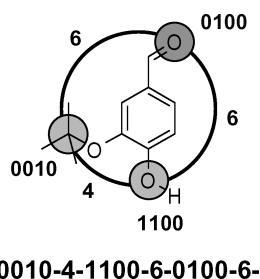
Fig. 14.   Example of a single Similog key determination. Adapted from Schuffenhauer et al. [2003].

responses such as adverse cardiac events common to potential drugs interacting with hERG (human Ether-a-go-go Related Gene) liabilities [Ertl 2007].

4.5.2. Ligand-Based Pharmacophores.   An alternative form of pharmacophoric description is one where explicit information is not known regarding the geometry of the target or a bound conformation. Given only the topology, it is still possible to characterize molecules in a way that is a topological analog of geometric pharmacophores. These include enumerating conformations of a particular molecule and using these as descriptions of the molecules for pharmacophores. Alternatively, the topology can be used by disconnecting the description from the underlying structure (in a departure from the connection-based descriptions of the fingerprints described previously). Such a disconnection is achieved by either using *though-graph* distances or *though-space* distances from generated conformers. The through-graph distance is the length of the shortest edge path between two atoms through the molecular graph, while the through-space distance is the geometric distance in Ångströms between two atoms.

Similog pharmacophoric keys are one type of topological pharmacophore where the nodes of a molecular graph are first generalized by four features with these being represented as 4 b/atom. Similog keys then represent triplets of these atom descriptions [Schuffenhauer et al. 2003]. In this way they can be seen as topological analogs of traditional 3-point geometric pharmacophore models. A summary of a Similog key construction is given in Figure 14.

The keys are based on a DABE atom typing scheme in which the constituents are potential hydrogen bond *D*onor, potential hydrogen bond *A*cceptor, *B*ulkiness, and *E*lectropositivity. Therefore, theoretically there are 24 potential atom types; however, six of these are not possible with neutral organic species and the key 0000 is uninformative, leaving nine possible keys. All possible triplet combinations of these nine keys give a theoretical limit of 8031 triplets; however, only 5989 had been found at the time the original article was written [Schuffenhauer et al. 2003]. The sorted triplet keys are each then encoded into a binary vector of length 5989, to denote presence or absence of a particular key.

4.6. Molecular Scaffolds and Scaffold Hopping

The term *scaffold* (or *chemotype*) is used extensively to describe the core structure of a molecule and molecular scaffold representations are a very active and important area of current research in chemoinformatics. Taken literally, the core structure is the central component of a molecule: the substantial substructure that contains the molecular material necessary to ensure that the *functional groups* are in a desired geometric arrangement and therefore bioisosteric. However, the core structure can also simply refer to the key component or components of the molecule that a particular scientist

adenosine   caffeine   KW-6002 (Kyowa-Hakko)
(a)   **(b)**   **(c)**
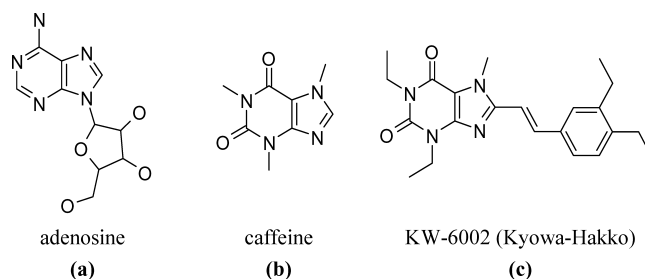
Fig. 15.   Illustration of known inhibitors of known adenosine A_{2A}-antagonists, important for treatment of Parkinson's disease: from the two natural products (a) adenosine (an agonist) and (b) caffeine (a subtype-unselective antagonist) to the designed ligand (c) an instance of an A_{2A}-antagonist. Adapted from Böhm et al. [2004].

defines, and not necessarily a scaffold in the literal sense. The definition of a scaffold is important since it is possible to determine whether an instance of *scaffold hopping* (*leapfrogging*, *lead-hopping*, *chemotype switching*, and *scaffold searching*) has occurred: that is, a complete change in the region of chemistry space being investigated, yet one that elicits a similar biological response [Böhm et al. 2004; Brown and Jacoby 2006]. An example of scaffold hopping is given in Figure 15 for A_{2A} antagonists. It is important to note that, although the molecule in Figure 15(c) wholly contains the caffeine scaffold, this is still classified as a scaffold hop since the core structure has been modified with an additional six-member ring system (benzene).

Experts with different backgrounds and knowledge will tend to define a scaffold differently depending on their particular domains of interest. For instance, a synthetic chemist may define a molecular scaffold based on the diversity not of the core structure, but on the relative diversity of the synthetic routes to the molecules themselves, whereas, patent lawyers would typically consider only the general similarity of the internal structure of the molecule to determine whether or not that particular region of scaffold chemistry space has prior art in terms of its Markush structure representation (see below) for the particular application domain of interest. Chemoinformaticians, however, will always favor an objective and invariant algorithm that will provide a solution rapidly and without ambiguity. In this case, a scaffold definition is provided by a graph transformation algorithm that, given a molecular topological graph, ensures that the scaffold can be realized deterministically. However, there are also significant limitations in the scaffold determination algorithm that maintains current favor in chemoinformatics.

4.6.1. Early Definitions of Scaffolds.   One of the earliest scaffold descriptions was that introduced by Eugene A. Markush of the Pharma-Chemical Corporation in a landmark patent that was granted on August 26, 1924 [Markush 1924]—although this was not the first patent to include such a generic definition. Markush's patent covered an entire family of pyrazolone dye molecules:

> The process for the manufacture of dyes which comprises coupling with a halogen-substituted pyrazolone, a diazotized unsulphonated material selected from the group consisting of aniline, homologues of aniline and halogen substitution products of aniline [Markush 1924, page 2].

In making this claim, Markush was able to claim rights not to just an individual compound of interest, but also a large number of molecules of only potential interest in the chemistry space surrounding the actual molecule synthesized at the center of the claim. Markush structures are now used extensively to protect chemical series of
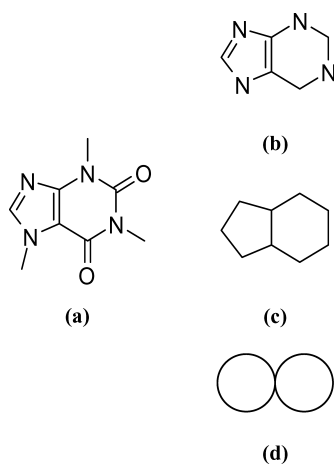
Fig. 16. The (b) molecular, (c) graph, (d) reduced scaffold framework representations for the caffeine molecule (a), respectively.

interest in patents in any industry that develops NCEs. The Markush generic structure is more concerned with intellectual property rights rather than a scientific basis and it is therefore not absolutely necessary that all of the molecules covered by a Markush representation can be synthesized.

The previous scaffold definition was designed specifically for intellectual property applications, but the scientific definition is also important to describe classes of related molecules accurately and invariantly. However, the definition of a scaffold is deceptively trivial to state, but incredibly difficult—if at all possible—to reduce to a set of generic rules that do not consider how the definition will be applied. For an early reference for an acceptable definition of a scaffold, as we tend to mean it today, we can go to the American Chemical Society (ACS) literature database and an article by Reich and Cram [1969] that describes it thus:

> The ring system is highly rigid, and can act as a scaffold for placing functional groups in set geometric relationships to one another for systematic studies of transannular and multiple functional group effects on physical and chemical properties (page 3527).

Although the definition is explanatory, it does not provide the practitioner with a rigorous and invariant description which would allow the determination of the scaffold component of any given molecule. Typically, the scaffold definitions given by Bemis and Murcko [1996] are now used widely in chemoinformatics to determine the "scaffold" of a molecule in an approach that is invariant and unbiased. These abstracted graph descriptions can then be applied in classification problems to enable the practitioner to group molecules by "scaffold" as another approach to diversity selection (discussed in more detail in Section 6.4).

From a single molecule, it is possible to generate the Bemis and Murcko [1996] scaffold or molecular framework, as well as the graph framework, as required. The former prunes side chains of the molecule, but maintains the original atom typing and bond orders used. The latter takes the same molecular framework and then proceeds to further abstract the atoms and bonds to uncolored and unweighted nodes and edges, respectively, thus giving an indication of the general framework of each molecule considered. The graph frameworks can be further abstracted by representing the ring systems as nodes of the graph. An example of a molecule, caffeine (Figure 16(a)) with its molecular,

graph, and reduced scaffold frameworks is provided in Figures 16(b), 16(c), and 16(d), respectively.

4.7. Overview: Molecular Descriptors

It is important as a practitioner to determine the preferred descriptors based on requirements in the problem domain. Essentially this represents a tradeoff surface between the necessity for interpretation and the efficacy of the descriptors to the problem being considered. Physicochemical descriptors are highly interpretable yet can lead to inferior application in the predictive sense, whereas information-based descriptors are very effective at describing the structure and permit the generation of highly predictive models; however, the typical representations are highly unintuitive for interpretation and require significant additional analysis to interpret what is salient.

Scaffolds are of significant importance in the discovery of NCEs since it is possible to patent a class of chemical compounds as opposed to just a single molecule, thereby affording increased intellectual property (IP) protection. As a result, scaffold hopping has emerged as a technique that attempts to circumnavigate the protected patent space by discovering molecules that still exhibit the desired biological response although not protected by extant IP.

5. *IN SILICO DE NOVO* MOLECULAR DESIGN

A very active area of research in chemoinformatics is that of the design of NCEs entirely within the computer that exhibit desired properties and molecular diversity and are designed such that they can be synthesized in the laboratory. A number of alternative methods have been proposed to design novel molecules *in silico* and each has its own particular challenge. Some of the more popular approaches are described in this section.

5.1. *De Novo* Design

Numerous *de novo* design programs have been developed over recent years to optimize molecules *in silico*. Typically, there are two general approaches to *de novo* design: those that optimize the molecule directly based on it satisfying binding site constraints of a particular known target of interest, and those that optimize based on a desired biological activity or other property profile. The former is 3D based, while the latter uses only topological information. Since the literature is vast on this subject, only some aspects will be discussed in this article; the interested reader is referred to the recent and extensive review by Schneider and Fechner [2005].

If the receptor site for which a ligand is required is known, then novel molecules may be generated *in silico* by either positioning fragments with desirable interactions into feature points in the site and then connecting those fragments with more molecular material, or iteratively growing the molecular from a seed atom or fragment.

If, however, only a ligand, or ligands, of interest is known, then multiple approaches are available that permit the exploration of the chemistry space around the extant ligand. One approach can use quantitative structure-activity relationships (QSARs) to optimize molecules that satisfy the desired biological activity target. Although extrapolation of the model can be troublesome, a range of additional safeguards have been advised to overcome this challenge [Eriksson et al. 2003; Brown and Lewis 2006]. An alternative is to optimize molecules that are similar in some defined regard (whether structural or having a property similarity) and explore the potential molecules around these structures in a controlled manner using multiobjective optimization techniques [Nicolaou et al. 2007].

5.2. Virtual Combinatorial Synthesis

Combinatorial synthesis *in vitro* uses sets of molecular fragments to synthesize all possible combinations of designed fragments as molecular products. Virtual combinatorial synthesis operates in the same way, but is performed *in silico*, permitting exploration of the combinatorial space prior to committing to the more expensive *in vitro* combinatorial enumeration. The products of the reagents are enumerated *in silico*, and then a subset design can be performed on these products, referred to as *cherry-picking*, by investigating their predicted properties and the overall molecular diversity of the products (see Section 6.4). One may then backtrack and select only those reagents, or a further subset thereof, thereby reducing the size of the *in vitro* combinatorial synthesis to generate the final set of molecules. These methods often employ a genetic algorithm (GA) to perform the subset selection given the combinatorial nature of the search space [Gillet et al. 1999].

5.3. Synthesis Design

One of the most significant challenges in *de novo* molecular design is the issue as to whether a medicinal chemist can easily realize the molecule through synthesis [Corey and Cheng 1995]. Many methods and tools have been proposed to achieve this, but this is still largely considered to be an unsolved problem.

One method of synthesis design is embodied in the Workbench for the Organization of Data for Chemical Applications (WODCA) system that uses a knowledge-based approach to determine synthesis routes. However, WODCA was developed with a primary intent of reducing the complexity of the *synthesis trees* that were output by extant synthesis design programs. A synthesis tree defines the potential routes to synthesis for a particular product and allows for the design of a synthetic route that efficient in both the number of synthetic steps and also the chemical reactions that are necessary. To this end, WODCA is an interactive system that provides the necessary information to the synthetic chemistry at each stage in the synthesis of new molecules [Gasteiger et al. 2000].

The Retrosynthetic Combinatorial Analysis Procedure (RECAP) [Lewell et al. 1998] approach defines a set of graph fragmentation (or bond cleavage, retrosynthetic) rules for molecular graphs that mimic common synthetic transforms performed by chemists. These rules can then be applied to define molecular fragments that may be recombined into novel molecules that have a greater likelihood of being synthetically accessible by a chemist.

6. SIMILARITY AND DIVERSITY

One of the most pressing issues in the application domain of chemoinformatics involves the rationalization of a large number of compounds such that we only retain what is desirable. Unfortunately, it is often impossible to define a priori what is desirable, and therefore various statistical learning methods are applied in an attempt to enrich subsets with molecules that are more likely to be interesting for further study based on prior knowledge.

The twin challenges of defining molecular similarity and diversity are uppermost in the minds of the majority of practitioners in chemoinformatics. More often than not, the preferred method of calculating molecular similarity or diversity is to use a coefficient that compares two molecular fingerprints. The Tanimoto (or Jaccard) association coefficient has been found to be the most useful when comparing binary molecular descriptors such as fingerprints and is the one most frequently used, although many others are used; the most common coefficients are provided in Table I [Willett et al.

Table I.   The Similarity and Distance Coefficients That Are Used Most Frequently in chemoinformatics (For the dichotomous variants, the variables are defined as a = number of bits set in the first binary vector, b = number of bits set in the second binary vector, and c = the number of bits set in common between both binary vectors.)

Name	Dichotomous	Continuous
Cosine	$\dfrac{c}{\sqrt{a \cdot b}}$	$\dfrac{\sum_{k=1}^{K} x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^{K}(x_{ik})^2 \sum_{k=1}^{K}(x_{jk})^2}}$
Dice	$\dfrac{2c}{a+b}$	$\dfrac{2\sum_{k=1}^{K} x_{jk}x_{jk}}{\sum_{k=1}^{K}(x_{ik})^2 + \sum_{k=1}^{K}(x_{jk})^2}$
Euclidean	$\sqrt{a+b-2c}$	$\sqrt{\sum_{k=1}^{K}\left(x_{ik}-x_{jk}\right)^2}$
Hamming	$a+b-2c$	$\sum_{k=1}^{K}|x_{ik}-x_{jk}|$
Tanimoto (Jaccard)	$\dfrac{c}{a+b-c}$	$\dfrac{\sum_{k=1}^{K} x_{ik}x_{jk}}{\sum_{k=1}^{k}(x_{ik})^2 + \sum_{k=1}^{k}(x_{jk})^2 - \sum_{k=1}^{k} x_{ik}x_{jk}}$

1998]. This method of calculating the similarity of or distance between molecules is used largely since the fingerprint representations can be generated and compared very rapidly when compared with graph-based similarity methods, although these methods are also used.

6.1. Substructure Screening and Searching

It is frequently necessary to search many molecules as to the presence or absence of a particular substructural component. This can be a particularly interesting functional group, or potentially toxic or reactive fragment. Therefore, early in the development of chemical information retrieval systems, methods were developed that could screen rapidly a large collection of compounds typically using a two-stage filtering process. In addition, much effort has been expended in 3D substructure searching.

As we mentioned earlier, subgraph isomorphism is computationally expensive and therefore an additional screening step was introduced that uses chemical fingerprints. The process proceeds by characterizing both the substructure of interest and the molecule being investigated as fingerprints (whether structure-key or hash-key fingerprints, described in Section 4.4). The substructure is then said to be in the molecule if all of the bits set in the substructure are also set in the molecule. Although this does not provide absolute certainty of the presence of the substructure in the molecule, the process reliably screens out the vast majority of molecules that do not contain the substructure (true negatives) and never screens out molecules with the substructure (false negatives). The more time-consuming substructure search algorithm is then applied to the remaining screened molecules.

Substructure searching has also been extended to the 3D problem, but tends to involve searching for a particular geometric arrangement of substructures or pharmacophoric features. This challenge is particularly fraught when considering multiple conformers of interest. Typically, in flexible structure searching, conformational sampling is employed to explore the available conformational space. However, there is an obvious tradeoff in exploration of this space with the increased runtime involved in exploring space, where the aim is to improve the search results while also remaining pragmatic in that the search process is rapid enough to be practical in general.

6.2. Clustering Methods

From the statistical learning literature, there are many extant methods for partitioning a set of N objects into k nonoverlapping cells—an approach that is referred to as *(crisp) clustering*. Largely due to pragmatic reasons in application to challenges in drug discovery, a few methods have become the most popular in chemoinformatics [Barnard and Downs 1992].

While an extensive review of the many statistical learning methods that are available and are applied in chemoinformatics is outside the scope of this article, the interested reader is referred to Hastie et al. [2001] for an excellent introduction to many of the methods that are typically applied in this field. Among the most common clustering methods are sequential agglomerative hierarchical nonoverlapping (SAHN) clustering, K-means clustering, self-organizing maps (SOMs, also known as *Kohonen maps* after their inventor), principal components analysis (PCA), multidimensional scaling (MDS, also known as *Sammon mapping*).

6.2.1. Stirling Numbers of the Second Kind. The space of possible cluster assignments for a given set of N objects into k unlabeled and disjoint cells or clusters is truly astronomical for all but the smallest of values of N and k. These values are referred to as *Stirling numbers of the second kind*, after the Scottish mathematician James Stirling [Goldberg et al. 1972]. The calculation of the Stirling number of the second kind is through the recurrence relation

$$\mathrm{S}(N, k) = k \cdot \mathrm{S}(N - 1, k) + \mathrm{S}(N - 1, k - 1), \tag{4}$$

where $1 \leq k < N$ given that the following initial conditions, or base cases, are met:

$$\mathrm{S}(N, N) = \mathrm{S}(N, 1) = 1. \tag{5}$$

The Stirling number of the first kind is different to the first kind in that the clusters or cells are labeled and therefore all permutations of cluster assignments are considered as distinct. The Bell number is the summation of the Stirling number of the second kind for all values of k from 1 to $N - 1$ and therefore provides the total number of possible clustering partitions available.

As an indicator, according to the Stirling number of the second kind, to partition 1006 objects into 196 nonempty cells, there are approximately 6.294×10^{1939} possible unique partitioning schemes. Therefore, it is readily apparent that the space of potential partitioning schemes is vast and it would therefore be very easy to arrive at a grouping that is even slightly incorrect, whatever incorrect means in this context.

6.3. Similarity Searching

A common approach to assisting in the decision of which molecules to test is to generate a ranked list of the structures according to their similarity to a known structure of interest [Willett et al. 1998; Bender and Glen 2004]. This allows a prioritization of molecules based on their likelihood of exhibiting a similar activity to our reference compound. Similarity searching is the embodiment of the similar-property principle in that it is expected that molecules which are similar to a molecule of interest will tend to exhibit similar properties and they are therefore more desirable to be tested.

Similarity searching proceeds by first calculating a similarity score between the reference (or query) molecule against a set of database compounds. The list of database molecules may then be sorted in order of decreasing similarity to provide a ranked list

ranging from the most to the least similar. The assumption here is that if we have restricted resources, such as only being able to test $n\%$ of the dataset, we can prioritize compounds using similarity searching such that the top $n\%$ of our ranked list will be more likely to exhibit our desired properties.

The typical way to objectively evaluate the quality of a particular similarity searching campaign is to test the recall of a similarity search using a particular search query molecule that is known to be active against a dataset that has been seeded with active compounds with the same activity, but not necessarily similar. This may then be plotted as an enrichment curve, where the axes are the percentage or number of database compounds screened, against the number of actives recalled at that screening level. This provides a very intuitive method of evaluating the quality of one particular method over another and one can readily determine the quality of enrichment relevant to one's particular screening level (Figure 17(a)).

Recently, an approach known as *data fusion* (or *consensus scoring* in the docking community) has gained widespread acceptance in combining values from multiple sources for a single object to further enrich our similarity searches [Hert et al. 2004]. Essentially, there are two particular approaches to data fusion that are currently applied in similarity searching: similarity fusion and group fusion. In similarity fusion, a single query or reference molecule is used to search against a structure database using a variety of methods such as alternate descriptors or similarity coefficients. Group fusion, however, uses multiple reference molecules with a single descriptor and coefficient.

6.4. Diverse Subset Selection

As we have seen already, the potential druglike chemistry space is vast and only a small fraction can realistically be considered, even *in silico*. However, how do we decide best which molecules should be selected from this very large space? To address this problem, a number of molecular diversity methods have been proposed over the years that offer many different ways of selecting subsets of the potential space while also maintaining levels of molecular diversity, where the definition of diversity is not necessarily a constant [Schuffenhauer and Brown 2006; Schuffenhauer et al. 2007].

Early studies into molecular diversity attempted to select subsets of a dataset that bound the dataset by covering the extremities of the space. However, it quickly became apparent that this level of diversity was not what was desired since the molecules toward the center of this hyperdimensional space were, as a result of the design of the process, being omitted. Therefore, diversity became coupled with the objective of maintaining the representivity of the space, or covering the space in the designed subset. However, considering this for a moment, what is essentially required of this process is a subset of the space that mirrors the data point density of the local regions in the chemistry space. Essentially, this is what is returned when a random selection is performed, when large enough subsets are required. This therefore appears to obviate any need for intelligently designed representative subsets, since random subsets are, by their nature, representative of the space being considered. An example of diversity selection, using a sphere-exclusion style approach is provided in Figure 17(b).

6.4.1. Dissimilarity-Based Compound Selection. An iterative method that is particularly popular is the class of algorithms called *dissimilarity-based compound selection* (DBCS) [Snarey et al. 1997]. Initially, the algorithms select a seed molecule according to the particular implementation of the algorithm such as picking a molecule at random or picking the most distal molecule in the dataset according to the descriptor space under consideration. The algorithms then proceed to select new molecules based on those selected before.
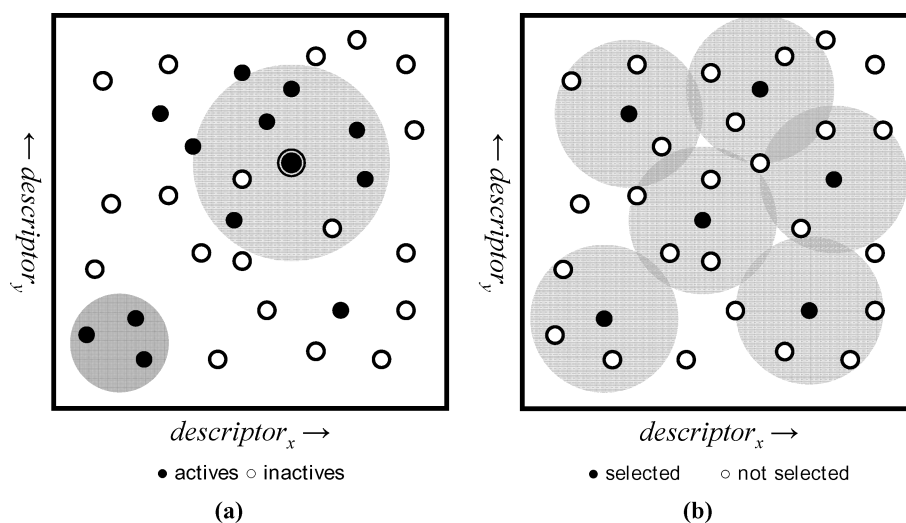
Fig. 17. An illustration of the related tasks of (a) similarity searching and (b) diversity (or subset) selection. In the case of similarity searching (a), the aim is to locate the nearest neighbors of a molecule of interest. However, it can be seen that potentially interesting "islands" of activity can be missed using this approach. Diversity selection (b), on the other hand, seeks to select a subset of compounds from a larger set such that the space is covered—here, in an approach referred to as *sphere exclusion*.

6.4.2. Cell-Based Compound Selection. Cell-based methods are one of the simplest approaches to compound selection. The algorithm proceeds by partitioning the, most likely, high-dimensional descriptor space into equidistant or varidistant partitions with the aim of arriving at a number of cells in the space that is closest to the desired number of data points. Then a single point is selected from each of the cells according to some rule. Again, this may simply be random, or by picking the most central point, or centroid.

6.4.3. Cluster Analysis. The clustering methods mentioned previously can also be applied for diverse subset selection [Schuffenhauer et al. 2006]. One proceeds by performing the cluster analysis such that the number of clusters, equals the number of data points required in the diverse subset. A single object may then be selected from each of the clusters, preferably be selecting the object that is nearest the center of the cluster (or centroid).

6.4.4. Onion Design. The final method to be considered here is the use of a method called *D-optimal onion design* (DOOD) by extension of D-optimal design as a space-filling design [Eriksson et al. 2004], which in isolation gives a shell design. The application of D-optimal design on a given space will provide a designed set of points that covers the surface of the space covered by the given objects, which is limited in applicability as mentioned earlier in this section. However, by iteratively performing a D-optimal design, and removing the surface points from the space, this will develop a designed subset that covers the volume of the space.

7. PREDICTIVE MODELING

As we have seen already, mathematics and chemistry have long been related disciplines. As early as the mid-19th century, Alexander Crum Brown (1838–1922) and Thomas Richard Fraser (1841–1920) suggested that a mathematical relationship can

be defined between the physiological action, Φ, of a molecule as a function of its chemical constitution, C:

$$\Phi = \mathrm{f}(C). \tag{6}$$

In doing so, they provided the following justification [Crum Brown and Fraser 1869]:

> There can be no reasonable doubt that a relation exists between the physiological action of a substance and its chemical composition and constitution, understanding by the latter term the mutual relations of the atoms in the substance (page 15).

The challenge in defining the function mapping was seen as largely due to the accuracy in defining C, and Φ, respectively [Livingstone 2000]. These definitions remain a challenge today, but the general issue of relating structure to property is of considerable importance in modern drug discovery and is used widely to guide decision making, even though our models are not as accurate as we would wish them to be.

Statistical models are of great importance in chemoinformatics since they allow the correlation of a measured response (dependent variable) such as biological activity with calculated molecular descriptors (independent variables). These models can then be applied to the forward problem of predicting the responses for unseen data points entirely *in silico*.

Two particular types of supervised learning methods are applied widely in chemoinformatics: *classification* and *regression*. Classification methods assign new objects, in our case molecules, to two or more classes—most frequently either biologically active or inactive. Regression methods, however, attempt to use continuous data, such as a measured biological response variable, to correlate molecules with that data so as to predict a continuous numeric value for new and unseen molecules using the generated model. The most-often used methods for classification are partial least squares—discriminant analysis (PLS-DA), naïve Bayesian classifier (NBC), recursive partitioning (RP), and support vector machines (SVM), whereas, for regression modeling, other methods are used like multiple linear regression (MLR), partial least squares (PLS), and artificial neural networks (ANNS).

The clustering methods discussed in section 6.2 and classification methods described here fall into two particular types of statistical learning methods: *unsupervised* and *supervised*, respectively. Unsupervised learning is used to determine natural groupings of objects based solely on their independent variables, as in SAHN, K-means clustering, SOM, PCA, and MDS, whereas, supervised statistical learning uses a priori information regarding the classes to which the objects in a training set belong, as in PLS-DA, NBC, RP, and SVM. This model is then used to classify new objects. In like vein, regression modeling is also a type of supervised learning method.

Predictive models and the modelers who generate them generally fall into one of two domains of application: the predictive and the interpretive modes. There is generally a tradeoff between prediction quality and interpretation quality, with the modeler determining which is preferred for the given application (Figure 18). Interpretable models are generally desired in situations where the model is expected to provide information about the problem domain and how best to navigate through chemistry space allowing the medicinal chemist to make informed decisions. However, these models tend to suffer in terms of prediction quality as they become more interpretable.

The reverse is true with predictive models in that their interpretation suffers as they become more predictive. Models that are highly predictive tend to use molecular descriptors that are not readily interpretable by the chemist such as information-based descriptors. However, predictive models are generally not intended to provide transparency, but predictions that are more reliable and can therefore be used as high-throughput models for filtering tasks or similar approaches.
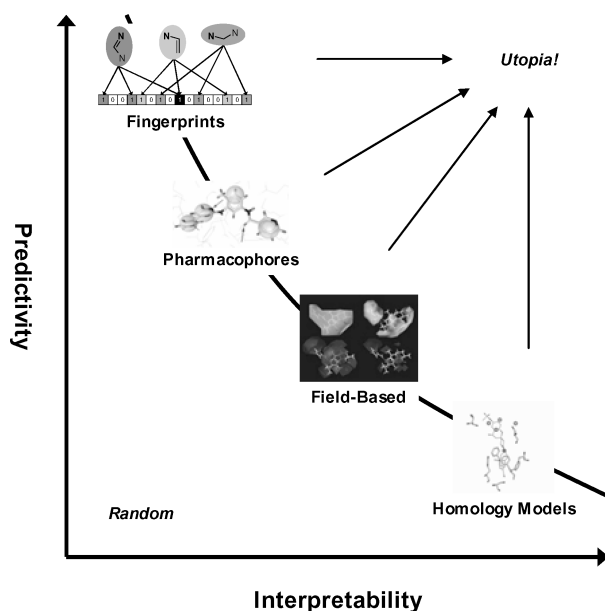
Fig. 18. The tradeoff surface between the predictive and interpretive modes of predictive models and examples of the types of descriptors or methods that are used. This figure has been adapted from an original diagram by Lewis, published in Brown and Lewis [2006].

7.1. Variable Selection and Model Parsimony

An additional question that is fraught with issues is the concept of variable selection. Occam's (or Ockham's) razor dictates that our models should be as complex as needed but no more so; this is also called the *principle of parsimony*. Therefore, by following parsimony, we should include only those descriptors that are necessary to model the response required. Two approaches are applied with this aim: the *build-up* and *build-down* variable selection methods.

The build-up method is performed by gradually adding descriptor variables from our available descriptors one at a time, checking the model quality at each step; someone skilled in the art of determining which descriptors are most likely to be important for this particular property.

The build-down method uses all the molecular descriptors that are available, which now run into the thousands, to generate a model and then using statistics from the model to select variables to remove if they are deemed to be of low importance for the response.

Both methods can be controversial, but each has a place. Chance correlations may arise if the statistical modeling method is prone to these, as is the case in multiple linear regression (MLR). However, partial least squares (PLS) does not suffer from this problem since a low-dimensional linear projection of the descriptors is determined [Migliavacca 2003].

Whatever method is chosen to develop predictive models, it is important to take heed of the model quality statistics and ensure a correct modeling methodology is used such as testing the model against an external and unseen test set to ensure it is not overfitting to the training set. Model extrapolation is another concern that frequently occurs when models are applied outside the space from which the models were generated. Again,

numerous model statistics are available that can indicate if new data points, from which responses are to be predicted, can be applied to the model [Brown and Lewis 2006].

7.2. Quantitative Structure-Activity Relationships

Many of the methods that have been described previously in this article have considered qualitative information: a molecule is either active or not active against a particular target. However, we frequently have quantitative data that we can also apply using computational and statistical learning techniques. Perhaps the most common continuous measure of biological activity is the $\log(IC_{50})$ (inhibitory concentration), which measures the concentration of a particular compound necessary to induce a 50% inhibition of the biological activity under investigation. Similarly, ED_{50} (effective dose), and LD_{50} (lethal dose), also provide measures of the compound required to exhibit 50% effectiveness and lethality, respectively.

From quantitative data, we can build a quantitative structure-activity relationship model that seeks to correlate our particular response variable of interest with molecular descriptors that have been calculated *in silico* or even measured from the molecules themselves. As we have seen already, Crum Brown and Fraser [1869] in the mid-19th century proposed that response is a function of constitution, but what we today refer to as *QSAR methods* were first pioneered by Corwin Hansch and colleagues in the 1940s. Today there is a family of methods related to QSAR: QSPR for property prediction; QSSR, for selectivity; QSTR, for toxicity. Each of these approaches essentially performs the same task: correlating a dependent variable (a measured response) with a set of independent variables (measured responses or, more often, calculated properties).

7.3. Molecular Docking

This method of predictive modeling is an approach to simulate the physical processes involved in a ligand binding to a protein binding site or the prediction of the structure of receptor-ligand complexes [Brooijmans and Kuntz 2003; Kitchen et al. 2004]. However, more recently an additional intention of docking has been toward scoring molecules *in silico* as a method of prioritizing those to be screened *in vitro* [Ferrara et al. 2006].

A number of approaches are applied in docking. In rigid docking, the ligand molecule is docked into a binding pocket using only translations and rotations, but not by exploring the potential conformations of the ligand itself. Even in this more limited search space, the degree of freedom is six dimensional.

In scoring the binding of a particular ligand to a protein, both the steric (or shape) complementarity as well as the interaction complementarity, respectively, between the ligand and the protein is considered to provide a predicted binding energy that is expected to be more accurate.

The first program that was published for docking was the DOCK program [Kuntz et al. 1982]. In this program, the receptor site is used to determine the inverse of the site using overlapping spheres of varying diameters each of which touches the surface of the receptor at two points only. These spheres then define the potential interaction points for a given ligand. Each ligand that is docked into the receptor can then be scored based on the calculation of an interaction (or binding) energy.

A more recent system is the genetic optimization for ligand docking (GOLD) program developed at the University of Sheffield, in collaboration with the CCDC and Glaxo-SmithKline (GSK) [Jones et al. 1997], which remains a very popular flexible docking program. The program uses a GA in which separate groups of potential ligand conformations and orientations are maintained as individual populations of valid solutions. In this chromosome encoding strategy, the conformational, rotational, and translational

spaces are therefore explored as the generations of the GA are iterated while the migration of solutions between the individual populations has the effect of improving the efficiency of the algorithm.

8. OVERVIEW

Chemoinformatics is now an essential component of chemical discovery, and nowhere is this more apparent than in the development of new pharmaceutical treatments to treat unmet medical needs. The field has a long and varied heritage, exhibiting influences from chemistry, of course, but also from mathematics, statistics, biology, computer science, and more besides. Now, the field of chemoinformatics is truly an interface science requiring skilled scientists from all necessary fields to be able to direct our research endeavors in the right direction and lead to an enriching and rewarding area of research that will only increase in its importance to drug discovery and other fields of chemistry in the coming years.

In conclusion, it is fitting to reflect on how rapidly chemoinformatics has become a mainstay of chemical research with an "Irishism" from an early pioneer of the field, Michael Lynch: "Here we sit side by side with those on whose shoulders we stand."

REFERENCES

ADAM, D. 2002. Chemists synthesize a single naming system. *Nature 417*, 369.

BAJORATH, J., ED. 2004. *Chemoinformatics: Concepts, Methods and Tools for Drug Discovery*. Humana Press, Totowa, NJ.

BALABAN, A. T. 1985. Applications of graph theory in chemistry. *J. Chem. Inf. Comput. Sci. 25*, 334–343.

BARNARD, J. M. AND DOWNS, G. M. 1992. Clustering of chemical structures on the basis of two-dimensional similarity measures. *J. Chem. Inf. Comput. Sci. 32*, 644–649.

BAUERSCHMIDT, S. AND GASTEIGER, J. 1997. Overcoming the limitations of a connection table description: A universal representation of chemical species. *J. Chem. Inf. Comput. Sci. 37*, 705–714.

BEMIS, G. W. AND MURCKO, M. A. 1996. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem. 39*, 2887–2893.

BENDER, A. AND GLEN, R. C. 2004. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem. 2*, 3204–3218.

BÖHM, H.-J., FLOHR, A., AND STAHL, M. 2004. Scaffold hopping. *Drug Discov. Today: Tech. 1*, 217–224.

BROOIJMANS, N. AND KUNTZ, I. D. 2003. Molecular recognition and docking algorithms. *Ann. Rev. Biophys. Biomol. Struct. 32*, 335–373.

BROWN, F. K. 1998. Chemoinformatics: What is it and how does it impact drug discovery? *Ann. Rep. Med. Chem. 33*, 375–384.

BROWN, N. AND JACOBY, E. 2006. On scaffolds and hopping in medicinal chemistry. *Mini Rev. Med. Chem. 6*, 1217–1229.

BROWN, N. AND LEWIS, R. A. 2006. Exploiting QSAR methods in lead optimization. *Curr. Opin. Drug Discov. Devel. 9*, 419–424.

BROWN, N., MCKAY, B., AND GASTEIGER, J. 2005. Fingal: A novel approach to geometric fingerprinting and a comparative study of its application to 3D QSAR modelling. *QSAR Comb. Sci. 24*, 480–484.

BROWN, N., MCKAY, B., GILARDONI, F., AND GASTEIGER, J. 2004. A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *J. Chem. Inf. Comput. Sci. 44*, 1079–1087.

BROWN, R. D. AND MARTIN, Y. C. 1997. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci. 37*, 1–9.

CECHETTO, J. D., ELOWE, N. H., BLANCHARD, J. E., AND BROWN, E. D. 2004. High-throughput screening at McMaster University: Automation in academe. *J. Assoc. Lab. Auto. 9*, 307–311.

COHEN, J. 2004. Bioinformatics—an introduction for computer scientists. *ACM Comput. Surv. 36*, 122–158.

COLES, S. J., DAY, N. E., MURRAY-RUST, P., RZEPA, H. S., AND ZHANG, Y. 2005. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem. 3*, 1832–1834.

COREY, E. J. AND CHENG, X.-M. 1995. *The Logic of Chemical Synthesis.* Wiley, New York, NY.

CRAMER, R. D., III., PATTERSON, D. E., AND BUNCE, J. D. 1988. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carried proteins. *J. Amer. Chem. Soc. 110*, 5959–5967.

CRUM BROWN, A. 1864. On the theory of isomeric compounds. *Trans. Roy. Soc. Edinb. 23*, 707–719.

CRUM BROWN, A. AND FRASER, T. R. 1869. V.—On the connection between chemical constitution and physiological action. Part. I.—On the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Trans. Roy. Soc. Edinb. 25*, 151–203.

DIESTEL, R. 2000. *Graph Theory,* 2nd Ed. Springer-Verlag, New York, NY.

DIMASI, J. A., HANSEN, R. W., AND GRABOWSKI, H. G. 2003. The price of innovation: New estimates of drug development costs. *J. Health Econ. 22*, 151–185.

DURANT, J. L., LELAND, B. A., HENRY, D. R., AND NOURSE, J. G. 2002. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci. 42*, 1273–1280.

ERIKSSON, L., ARNHOLD, T., BECK, B., FOX, T., JOHANSSON, E., AND KRIEGL, J. M. 2004. Onion design and its application to a pharmaceutical QSAR problem. *J. Chemomet. 18*, 188–202.

ERIKSSON, L., JAWORSKA, J., WORTH, A. P., CRONIN, M. T. D., AND MCDOWELL, R. M. 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ. Health Perspect. 111*, 1361–1375.

ERTL, P. 2007. *In silico* identification of bioisosteric functional groups. *Curr. Opin. Drug Discov. Devel. 10*, 281–288.

FERRARA, P., PRIESTLE, J. P., VANGREVELINGHE, E., AND JACOBY, E. 2006. New developments and applications of docking and high-throughput docking for drug design and *in silico* screening. *Curr. Comp.-Aided Drug Des. 2*, 83–91.

FUJITA, T., IWASA, J., AND HANSCH, C. 1964. A new substituent constant, π, derived from partition coefficients. *J. Amer. Chem. Soc. 86*, 5175–5180.

GASTEIGER, J., ED. 2003. *The Handbook of Chemoinformatics*. Wiley-VCH, Weinheim, Germany.

GASTEIGER, J. AND ENGEL, T., EDS. 2003. *Chemoinformatics: A Textbook*. Wiley-VCH, Weinheim, Germany.

GASTEIGER, J., PFÖRTNER, M., SITZMANN, M., HÖLLERING, R., SACHER, O., KOSTKA, T., AND KARG, N. 2000. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Persp. Drug Discov. Des. 20*, 1–21.

GASTEIGER, J., RUDOLPH, C., AND SADOWSKI, J. 1990. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahed. Comput. Methodol. 3*, 537–547.

GHOSE, A. K. AND CRIPPEN, G. M. 1986. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J. Comp. Chem. 7*, 565–577.

GILLET, V. J., WILLETT, P., BRADSHAW, J., AND GREEN, D. V. S. 1999. Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci. 39*, 169–177.

GOLDBERG, K., NEWMAN, M., AND HAYNSWORTH, E. 1972. Combinatorial Analysis. In *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, 10th ed. Abramowitz, M., Stegun, I. A. Eds. U.S. Government Printing Office: Washington, DC, 824–825.

GORSE, A.-D. 2006. Diversity in medicinal chemistry space. *Curr. Top. Med. Chem. 6*, 3–18.

GUND, P. 1979. Pharmacophoric pattern searching and receptor mapping. *Ann. Rep. Med. Chem. 14*, 299–308.

GÜNER, O. F. 2005. The impact of pharmacophore modeling in drug design. *IDrugs 8*, 567–572.

HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, NY.

HERT, J., WILLETT, P., WILTON, D. J., ACKLIN, P., AZZAOUI, K., JACOBY, E., AND SCHUFFENHAUER, A. 2004. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci. 44*, 1177–1185.

JOHNSON, M. A. AND MAGGIORA, G. M. Eds. 1990. *Concepts and Applications of Molecular Similarity*. Wiley Inter-Science, New York, NY.

JONES, G., WILLETT, P., GLEN, R. C., LEACH, A. R., AND TAYLOR, R. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol. 267*, 727–748.

KARELSON, M. 2000. *Molecular Descriptors in QSAR/QSPR*. Wiley-VCH, Weinheim, Germany.

KITCHEN, D. B., DECORNEZ, H., FURR, J. R., AND BAJORATH, J. 2004. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nature Rev. Drug Discov. 3*, 935–949.

KUNTZ, I. D., BLANEY, J. M., OATLEY, S. J., LANGRIDGE, R., AND FERRIN, T. E. 1982. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol. 161*, 269–288.

LEACH, A. R. 2001. *Molecular Modelling: Principles and Applications,* 2nd ed. Prentice Hall, Harlow, U.K.

LEACH, A. R. AND GILLET, V. J. 2003. *An Introduction to Chemoinformatics*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

LEWELL, X. Q., JUDD, D. B., WATSON, S. P., AND HANN, M. M. 1998. RECAP—retrosynthetic analysis procedure: A powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci. 38*, 511–522.

LIPINSKI, C. A., LOMBARDO, F., DOMINY, B. W., AND FEENEY, P. J. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev. 46*, 3–26.

LIVINGSTONE, D. J. 2000. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci. 40*, 195–209.

LYNCH, M. F. 2004. Introduction of computers in chemical structure information systems, or what is not recorded in the annals. In *The History and Heritage of Scientific and Technological Information Systems: Proceedings of the 2002 Conference*, W. B. Rayward and M. E. Bowden, Eds. Information Today, Inc., Medford, NJ, 137–148.

MARKUSH, E. A. 1924. Pyrazolone dye and process of making the same. U.S. Patent No. 1,506,316, August 26.

MIGLIAVACCA, E. 2003. Applied introduction to multivariate methods used in drug discovery. *Mini Rev. Med. Chem. 3*, 831–843.

MORGAN, H. L. 1965. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *J. Chem. Doc. 5*, 107–113.

NICOLAOU, C. A., BROWN, N., AND PATTICHIS, C. S. 2007. Molecular optimization using multi-objective methods. *Curr. Opin. Drug Discov. Devel. 10*, 316–324.

OPREA, T. Ed. 2005a. *Chemoinformatics in Drug Discovery*. Wiley-VCH, Weinheim, Germany.

OPREA, T. 2005b. Is safe exchange of data possible? *Chem. Eng. News 83*, 24–29.

PEARLMAN, R. S. 1987. Rapid generation of high quality approximate 3D molecular structures. *Chem. Des. Automa. News 2*, 5–7.

RAEVSKY, O. A. 2004. Physicochemical descriptors in property-based drug design. *Mini Rev. Med. Chem. 4*, 1041–1052.

REICH, H. J. AND CRAM, D. J. 1969. Macro rings. XXXVII. Multiple electrophilic substitution reactions of [2,2]paracyclophanes and interconversions of polysubstituted derivatives. *J. Am. Chem. Soc. 91*, 3527–3533.

ROGERS, D., BROWN, R. D., AND HAHN, M. 2005. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up. *J. Biomol. Screen. 10*, 682–686.

RUSSO, E. 2002. Chemistry plans a structural overhaul. *Nature Jobs 419*, 4–7.

SCHNEIDER, G. AND FECHNER, U. 2005. Computer-based *de novo* design of drug-like molecules. *Nature Rev. Drug Discov. 4*, 649–663.

SCHUFFENHAUER, A. AND BROWN, N. 2006. Chemical diversity and biological activity. *Drug Discov. Today: Technol. 3*, 387–395.

SCHUFFENHAUER, A., BROWN, N., SELZER, P., ERTL, P., AND JACOBY, E. 2006. Relationships between molecular complexity, biological activity, and structural activity. *J. Chem. Inf. Mod. 46*, 525–535.

SCHUFFENHAUER, A., FLOERSHEIM, P., ACKLIN, P., AND JACOBY, E. 2003. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci. 43*, 391–405.

SCHUFFENHAUER, A., BROWN, N., ERTL, P., JENKINS, J. L., SELZER, P., AND HAMON, J. 2007. Clustering and rule-based classifications of chemical structures evaluated in the biological activity space. *J. Chem. Inf. Mod. 47*, 325–336.

SNAREY, M., TERRETT, N. K., WILLETT, P., AND WILTON, D. J. 1997. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Mod. 15*, 372–385.

TODESCHINI, R. AND CONSONNI, V. 2000. *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, Germany.

WEININGER, D. 1988. Smiles a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci. 28*, 31–36.

WILLETT, P. 1991. *Three-Dimensional Chemical Structure Handling*. Research Studies Press, Balclock, Hertfordshine, U.K.

WILLETT, P. 2000. Textual and chemical information processing: Different domains but similar algorithms. *Inform. Res. 5*, `http://informationr.net/ir/5-2/paper69.html`.

WILLETT, P., BARNARD, J. M., AND DOWNS, G. M. 1998. Chemical similarity searching. *J. Chem. Inf. Comput. Sci. 38*, 983–996.