PART I

DATA MINING IN THE PHARMACEUTICAL INDUSTRY: A GENERAL OVERVIEW

1

A HISTORY OF THE DEVELOPMENT OF DATA MINING IN PHARMACEUTICAL RESEARCH

David J. Livingstone and John Bradshaw

Table of Contents

Pharmaceutical Data Mining: Approaches and Applications for Drug Discovery,
Edited by Konstantin V. Balakin

1.1 INTRODUCTION

From the earliest times, chemistry has been a classification science. For example, even in the days when it was emerging from alchemy, substances were put into classes such as "metals." This "metal" class contained things such as iron, copper, silver, and gold but also mercury, which, even though it was liquid, still had enough properties in common with the other members of its class to be included. In other words, scientists were grouping together things that were related or similar but were not necessarily identical, all important elements of the subject of this book: data mining. In today's terminology, there was an underlying data model that allowed data about the substances to be recorded, stored, analyzed, and conclusions drawn. What is remarkable in chemistry is that not only have the data survived more than two centuries in a usable way but that the data have continued to leverage contemporary technologies for its storage and analysis.

In the early 19th century, Berzelius was successful in persuading chemists to use alphabetic symbols for the elements: "The chemical signs ought to be letters, for the greater facility of writing, and not to disfigure a printed book" [1]. This Berzelian system [2] was appropriate for the contemporary storage and communication medium, i.e., paper, and the related recording technology, i.e., manuscript or print.

One other thing that sets chemical data apart from other data is the need to store and to search the compound structure. These structural formulas are much more than just pictures; they have the power such that "the structural formula of, say, p-rosaniline represents the same substance to Robert B. Woodward say, in 1979 as it did to Emil Fischer in 1879" [3]. As with the element symbols, the methods and conventions for drawing chemical structures were agreed at an international level. This meant that chemists could record and communicate accurately with each other, the nature of their work.

As technologies moved on and volumes of data grew, chemists would need to borrow methodology from other disciplines. Initially, systematic naming of compounds allowed indexing methods, which had been developed for text handling and were appropriate for punch card sorting, to deal with the explosion of known structures. Later, graph theory was used to be able to handle structures directly in computers. Without these basic methodologies to store the data, data mining would be impossible.

The rest of this chapter represents the authors' personal experiences in the development of chemistry data mining technologies since the early 1970s.

1.2 TECHNOLOGY

When we began our careers in pharmaceutical research, there were no computers in the laboratories. Indeed, there was only one computer in the company and that was dedicated to calculating the payroll! Well, this is perhaps a slight exaggeration. A Digital Equipment Corporation (DEC) PDP-8 running in-house regression software was available to one of us and the corporate mainframes were accessible via teleprinter terminals, although there was little useful scientific software running on them.

This was a very different world to the situation we have today. Documents were typed by a secretary using a typewriter, perhaps one of the new electric golf ball typewriters. There was no e-mail; communication was delivered by post, and there was certainly no World Wide Web. Data were stored on sheets of paper or, perhaps, punched cards (see later), and molecular models were constructed by hand from kits of plastic balls. Compounds were characterized for quantitative structure–activity relationship (QSAR) studies by using lookup tables of substituent constants, and if an entry was missing, it could only be replaced by measurement. Mathematical modeling consisted almost entirely of multiple linear regression (MLR) analysis, often using self-written software as already mentioned.

So, how did we get to where we are today? Some of the necessary elements were already in existence but were simply employed in a different environment; statistical software such as BMDP, for example, was widely used by academics. Other functionalities, however, had to be created. This chapter traces the development of some of the more important components of the systems that are necessary in order for data mining to be carried out at all.

1.3 COMPUTERS

The major piece of technology underlying data mining is, of course, the computer. Other items of technology, both hardware and software, are of course important and are covered in their appropriate sections, but the huge advances in our ability to mine data have gone hand in hand with the development of computers. These machines can be split into four main types: mainframes, general-purpose computers, graphic workstations, and personal computers (PCs).

1.3.1 Mainframes

These machines are characterized by a computer room or a suite of rooms with a staff of specialists who serve the needs of the machine. Mainframe computers were expensive, involving considerable investment in resource, and there was thus a requirement for a computing department or even division within the organizational structure of the company. As computing became available within the laboratories, a conflict of interest was perceived between the computing specialists and the research departments with competition for budgets, human resources, space, and so on. As is inevitable in such situations, there were sometimes "political" difficulties involved in the acquisition of both hardware and software by the research functions.

Mainframe computers served some useful functions in the early days of data mining. At that time, computing power was limited compared with the requirements of programs such as ab initio and even semi-empirical quantum chemistry packages, and thus the company mainframe was often employed for these calculations, which could often run for weeks. As corporate databases began to be built, the mainframe was an ideal home for them since this machine was accessible company-wide, a useful feature when the organization had multiple sites, and was professionally maintained with scheduled backups, and so on.

1.3.2 General-Purpose Computers

DEC produced the first retail computers in the 1960s. The PDP-1 (PDP stood for programmable data processor) sold for $120,000 when other computers cost over a million. The PDP-8 was the least expensive general-purpose computer on the market [4] in the mid-1960s, and this was at a time when all the other computer manufacturers leased their machines. The PDP-8 was also a desktop machine so it did not require a dedicated computing facility with support staff and so on. Thus, it was the ideal laboratory computer. The PDP range was superseded by DEC's VAX machines and these were also very important, but the next major step was the development of PCs.

1.3.3 Graphic Workstations

The early molecular modeling programs required some form of graphic display for their output. An example of this is the DEC GT40, which was a monochrome display incorporating some local processing power, actually a PDP-11 minicomputer. A GT40 could only display static images and was usually connected to a more powerful computer, or at least one with more memory, on which the modeling programs ran. An alternative lower-cost approach was the development of "dumb" graphic displays such as the Tektronix range of devices. These were initially also monochrome displays, but color terminals such as the Tek 4015 were soon developed and with their relatively low cost allowed much wider access to molecular modeling systems. Where molecular modeling was made generally available within a company, usually using in-house software, this was most often achieved with such terminals.

These devices were unsuitable, however, for displaying complicated systems such as portions of proteins or for animations. Dedicated graphic workstations, such as the Evans and Sutherland (E&S) picture systems, were the first workstations used to display the results of modeling macromolecules. These were expensive devices and thus were limited to the slowly evolving computational chemistry groups within the companies. E&S workstations soon faced competition from other companies such as Sun and, in particular, Silicon Graphics International Corporation (SGI). As prices came down and computing performance went up, following Moore's law, the SGI workstation became the industry standard for molecular modeling and found its way into the chemistry departments where medicinal chemists could then do their own molecular modeling. These days, of course, modeling is increasingly being carried out using PCs.

1.3.4 PCs

IBM PCs or Apple Macintoshes gradually began to replace dumb terminals in the laboratories. These would usually run some terminal emulation software so that they could still be used to communicate with the large corporate computers but would also have some local processing capability and, perhaps, an attached printer. At first, the local processing would be very limited, but this soon changed with both the increasing sophistication of "office" suites and the usual increasing performance/decreasing price evolution of computers in general. Word processing on a PC was a particularly desirable feature as there was a word processing program running on a DEC VAX (MASS-11), which was nearly a WYSIWYG (what you see is what you get) word processor, but not quite! These days, the PC allows almost any kind of computing job to be carried out.

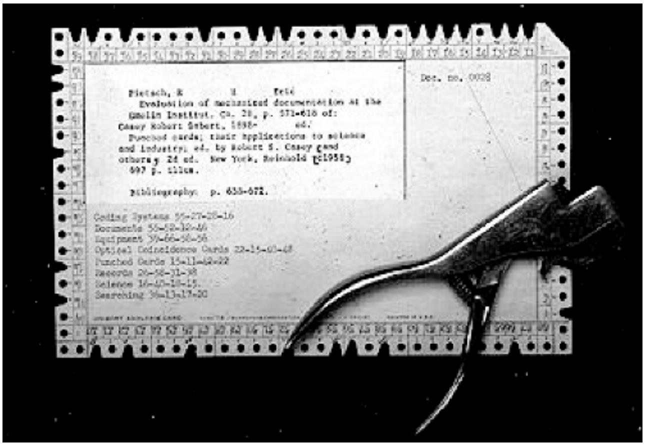
This has necessarily been a very incomplete and sketchy description of the application of computers in pharmaceutical research. For a detailed discussion, see the chapter by Boyd and Marsh [5].

1.4 DATA STORAGE AND MANIPULATION

Information on compounds such as structure, salt, melting point, molecular weight, and so on, was filed on paper sheets. These were labeled numerically and were often sorted by year of first synthesis and would be stored as a complete collection in a number of locations. The data sheets were also microfilmed as a backup, and this provided a relatively faster way of searching the corporate compound collection for molecules with specific structural features or for analogues of compounds of interest. Another piece of information entered on the data sheets was an alphanumeric code called the Wiswesser line notation (WLN), which provided a means of encoding the structure of the compound in a short and simple string, which later, of course, could be used to represent the compound in a computer record. WLN is discussed further in a later section.

Experimental data, such as the results of compound screening, were stored in laboratory notebooks and then were collated into data tables and eventually reports. Individual projects sometimes used a system of edge-notched cards to store both compound and experimental information. Figure 1.1 shows one of these edge-notched cards.

Edge-notched cards were sets of printed cards with usually handwritten information. Along the edge were a series of holes, which could be clipped to



Doc. no. 0028

Pietsch, E H Erik
Evaluation of mechanized documentation at the Gmelin Institut. Ch. 28, p. 571-618 of:
Casey Robert Sebert, 1898- ed.
Punched cards; their applications to science and industry, ed. by Robert S. Casey [and others] 2d ed. New York, Reinhold [c1958]
697 p. illus.

Bibliography: p. 655-672.

Coding Systems 55-27-28-16
Documents 55-52-32-46
Equipment 39-66-50-56
Optical Coincidence Cards 22-15-03-48
Punched Cards 15-11-42-22
Records 26-58-31-38
Science 16-45-18-15.
Searching 36-13-17-20

Figure 1.1 Edge-notched card and card punch.

form a notch. Each of these notches corresponded to some property of the item on the card. Which property corresponded to which notch did not matter, as long as all cards in a project used the same system. Then, by threading a long needle or rod through the hole corresponding to a desired property and by lifting the needle, all the cards that did *not* have that property were retained on the needle and were removed. (Note this is a principle applied to much searching of chemical data—first remove all items that could not possibly match the query.) The cards with a notch rather than a complete hole fall from the stack. Repeating the process with a single needle allows a Boolean "and" search on multiple properties as does using multiple needles. Boolean "or" search was achieved by combining the results of separate searches [6]. This method is the mechanical equivalent of the bit screening techniques used in substructure searching [7]. The limitations of storing and searching chemical information in this way are essentially physical. The length of the needle and the dexterity of the operator gave an upper limit to the number of records that could be addressed in a single search, although decks of cards could be accessed sequentially. There was no way, though, that all of the company compound database could be searched, and the results of screening molecules in separate projects were effectively unavailable. This capability would have to wait until the adoption of electronic databases.

1.5 MOLECULAR MODELING

Hofmann was one of the earliest chemists to use physical models to represent molecules. In a lecture at the Royal Society in 1865, he employed croquet balls as the atoms and steel rods as the bonds. To this day, modeling kits tend to use the same colors as croquet balls for the atoms. In the 1970s, models of

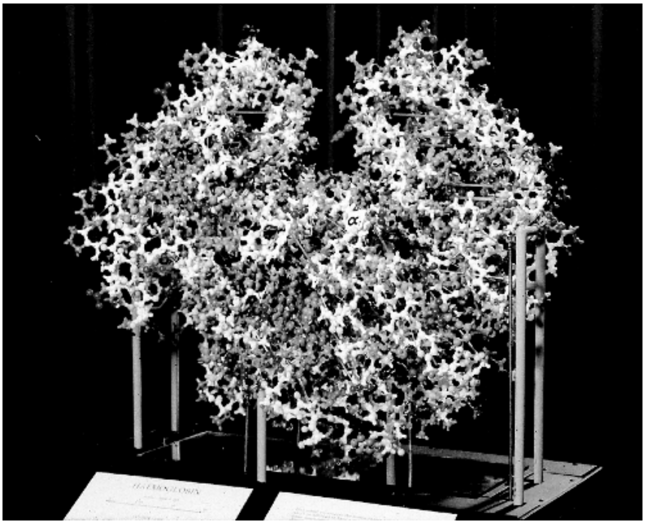


Figure 1.2 Physical model of hemoglobin in the deoxy conformation. The binding site for the natural effector (2,3-bisphosphoglycerate) is shown as a cleft at the top.

small molecules or portions of proteins used in the research laboratories were physical models since computer modeling of chemistry was in its infancy. An extreme example of this is shown in Figure 1.2, which is a photograph of a physical model of human hemoglobin built at the Wellcome research laboratories at Beckenham in Kent. This ingenious model was constructed so that the two α and β subunits were supported on a Meccano framework, allowing the overall conformation to be changed from oxy- to deoxy- by turning a handle on the base of the model. To give an idea of the scale of the task involved in producing this model, the entire system was enclosed in a perspex box of about a meter cube.

Gradually, as computers became faster and cheaper and as appropriate display devices were developed (see Graphic Workstations above), so molecular modeling software began to be developed. This happened, as would be expected, in a small number of academic institutions but was also taking place in the research departments of pharmaceutical companies. ICI, Merck, SKF, and Wellcome, among others, all produced in-house molecular modeling systems. Other companies relied on academic programs at first to do their molecular modeling, although these were soon replaced by commercial systems. Even when a third party program was used for molecular modeling, it was usually necessary to interface this with other systems, for molecular orbital calculations, for example, or for molecular dynamics, so most of the computational chemistry groups would be involved in writing code. One of the great advantages of having an in-house system is that it was possible to add any new technique as required without having to wait for its implementation by a software company. A disadvantage, of course, is that it was necessary to maintain the system as changes to hardware were made or as the operating systems evolved through new versions. The chapter by Boyd gives a nice history of the development of computational chemistry in the pharmaceutical industry [8].

The late 1970s/early 1980s saw the beginning of the development of the molecular modeling software industry. Tripos, the producer of the SYBYL modeling package, was formed in 1979 and Chemical Design (Chem-X) and Hypercube (Hyperchem) in 1983. Biosym (Insight/Discover) and Polygen (QUANTA/CHARMm) were founded in 1984. Since then, the software market grew and the software products evolved to encompass data handling and analysis, 3-D QSAR approaches, bioinformatics, and so on. In recent times, there has been considerable consolidation within the industry with companies merging, folding, and even being taken into private hands. The article by Allen Richon gives a summary of the field [9], and the network science web site is a useful source of information [10].

1.6 CHARACTERIZING MOLECULES AND QSAR

In the 1970s, QSAR was generally created using tabulated substituent constants to characterize molecules and MLR to create the mathematical models. Substituent constants had proved very successful in describing simple chemical reactivity, but their application to complex druglike molecules was more problematic for a number of different reasons:

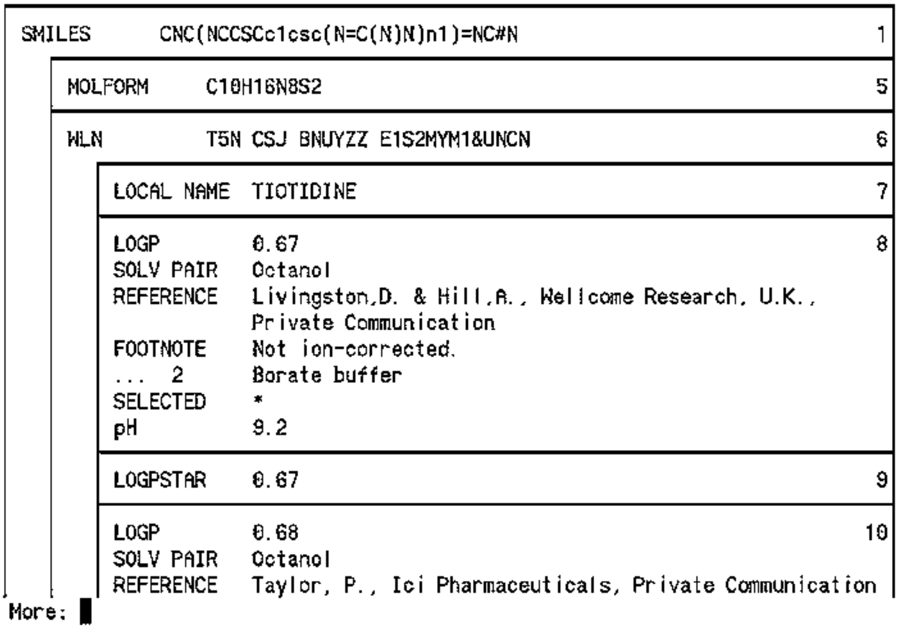
- It was often difficult to assign the correct positional substituent constant for compounds containing multiple, sometimes fused, aromatic rings.
- Missing values presented a problem that could only be resolved by experimental measurement, sometimes impossible if the required compound was unstable. Estimation was possible but was fraught with dangers.
- Substituent constants cannot be used to describe noncongeneric series.

An alternative to substituent constants, which was available at that time, was the topological descriptors first described by Randic [11] and introduced to the QSAR literature by Kier and Hall [12]. These descriptors could be rapidly calculated from a 2-D representation of any structure, thus eliminating the problem of missing values and the positional dependence of some substituent constants. The need for a congeneric series was also removed, and thus it would seem that these parameters were well suited for the generation of QSARs. There was, however, some resistance to their use.

One of the perceived problems was the fact that so many different kinds of topological descriptors could be calculated and thus there was suspicion that relationships might be observed simply due to chance effects [13]. Another objection, perhaps more serious, was the difficulty of chemical interpretation. This, of course, is a problem if the main aim of the construction of a QSAR is the understanding of some biological process or mechanism. If all that is required, however, is some predictive model, then QSARs constructed using topological descriptors may be very useful, particularly when calculations are needed for large data sets such as virtual libraries [14,15].

One major exception to the use of substituent constants was measured, whole-molecule, partition coefficient (log *P*) values. The hydrophobic substituent constant, π, introduced by Hansch et al. [16], had already been shown to be very useful in the construction of QSARs. The first series for which this parameter was derived was a set of monosubstituted phenoxyacetic acids, but it soon became clear that π values were not strictly additive across different parent series, due principally to electronic interactions, and it became necessary to measure π values in other series such as substituted phenols, benzoic acids, anilines, and so on [17]. In the light of this and other anomalies in the hydrophobic behavior of molecules, experimental measurements of log *P* were made in most pharmaceutical companies. An important resource was set up at Pomona College in the early 1970s in the form of a database of measured partition coefficients, and this was distributed as a microfiche and computer tape (usually printed out for access) at first, followed later by a computerized database. Figure 1.3 shows a screen shot from this database of some measured values for the histamine H2 antagonist tiotidine.

The screen shot shows the Simplified Molecular Input Line Entry System (SMILES) and WLN strings, which were used to encode the molecular struc-



```
SMILES      CNC(NCCSCc1csc(N=C(N)N)n1)=NC#N                              1
   MOLFORM     C10H16N8S2                                                 5
   WLN         T5N CSJ BNUYZZ E1S2MYM1&UNCN                               6
      LOCAL NAME  TIOTIDINE                                               7
      LOGP        0.67                                                    8
      SOLV PAIR   Octanol
      REFERENCE   Livingston,D. & Hill,A., Wellcome Research, U.K.,
                  Private Communication
      FOOTNOTE    Not ion-corrected.
      ...  2      Borate buffer
      SELECTED    *
      pH          9.2
      LOGPSTAR    0.67                                                    9
      LOGP        0.68                                                   10
      SOLV PAIR   Octanol
      REFERENCE   Taylor, P., Ici Pharmaceuticals, Private Communication
More: █
```

Figure 1.3 Entry from the Pomona College log *P* database for tiotidine.

ture (see later) and two measured log *P* values. One of these has been selected as a log *P* "star" value. The "starlist" was a set of log *P* values that were considered by the curators of the database to be reliable values, often measured in their own laboratories. This database was very useful in understanding the structural features that affected hydrophobicity and proved vitally important in the development of the earliest expert systems used in drug research—log *P* prediction programs. The two earliest approaches were the fragmental system of Nys and Rekker [18], which was based on a statistical analysis of a large number of log *P* values and thus was called reductionist, and the alternative (constructionist) method due to Hansch and Leo, based on a small number of measured fragments [19]. At first, calculations using these systems had to be carried out by hand, and not only was this time-consuming but for complicated molecules, it was sometimes difficult to identify the correct fragments to use. Computer programs were soon devised to carry out these tasks and quite a large number of systems have since been developed [20,21], often making use of the starlist database.

Theoretical properties were an alternative way of describing molecules, and there are some early examples of the use of quantities such as superdelocalizability [22] and Ehomo [23,24]. It was not until the late 1980s, however, that theoretical properties began to be employed routinely in the creation of QSARs [25]. This was partly due to the increasing availability of relatively easy-to-use molecular orbital programs, but mostly due to the recognition of the utility of these descriptors. Another driver of this process was the fact that many pharmaceutical companies had their own in-house software and thus were able to produce their own modules to carry out this task. Wellcome, for example, developed a system called PROFILES [26] and SmithKline Beecham added a similar module to COSMIC [27]. Table 1.1 shows an early example of the types of descriptors that could be calculated using these systems.

Since then, the development of all kinds of descriptors has mushroomed until the situation we have today where there are thousands of molecular properties to choose from [29,30], and there is even a web site that allows their calculation [31].

The other component of the creation of QSARs was the tool used to establish the mathematical models that linked chemical structure to activity. As already mentioned, in the 1970s, this was almost exclusively MLR but there were some exceptions to this [32,33]. MLR has a number of advantages in that the models are easy to interpret and, within certain limitations, it is possible to assess the statistical significance of the models. It also suffers from some limitations, particularly when there are a large number of descriptors to choose from where the models may arise by chance [13] and where selection bias may inflate the values of the statistics used to judge them [34,35]. Thus, with the increase in the number of available molecular descriptors, other statistical and mathematical methods of data analysis began to be employed [36]. At first, these were the "regular" multivariate methods that had been developed and

TABLE 1.1 An Example of a Set of Calculated Properties (Reproduced with Permission from Hyde and Livingstone [28])

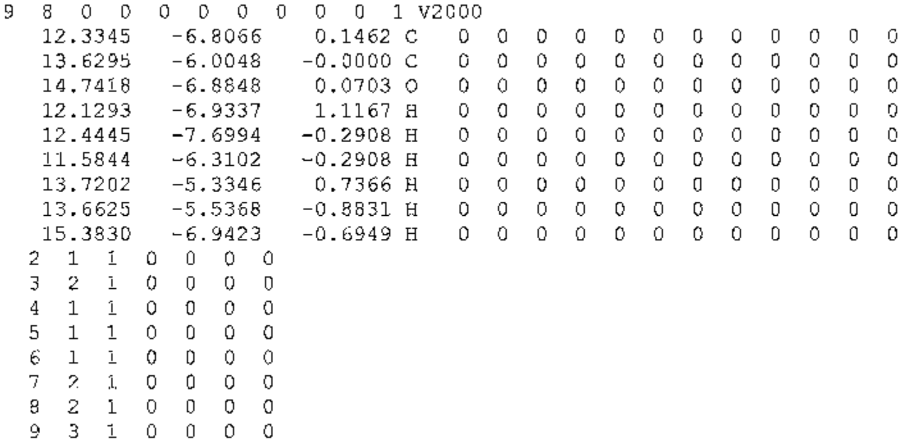
Calculated Property Set (81 Parameters, 79 Compounds)	
Whole-molecule properties	
"Bulk" descriptors	M.Wt., van der Waals' volume, dead space volume, collision diameter, approach diameter, surface area, molar refraction
"Shape" descriptors	Moment of inertia in *x*-, *y*-, and *z*-axes; principal ellipsoid axes in *x, y*, and *z* directions
Electronic and energy descriptors	Dipole moment; *x, y*, and *z* components of dipole moment; energies (total, core–core repulsion and electronic)
Hydrophobicity descriptors	Log *P*
Substituent properties	
For two substituents	Coordinates (*x, y*, and *z*) of the center, ellipsoid axes (*x, y*, and *z*) of the substituent
Atom-centered properties	
Electronic	Atom charges and nucleophilic and electrophilic superdelocalizability for atom numbers 1–14
Shape	Interatomic distances between six pairs of heteroatoms

applied in other fields such as psychology, but soon other newer techniques such as artificial neural networks found their way into the molecular design field [37]. As with any new technique, there were some problems with their early applications [38], but they soon found a useful role in the construction of QSAR models [39,40].

This section has talked about the construction of QSAR models, but of course this was an early form of data mining. The extraction of knowledge from information [41] can be said to be the ultimate aim of data mining. (See edge-notched cards above.)

1.7 DRAWING AND STORING CHEMICAL STRUCTURES

Chemical drawing packages are now widely available, even for free from the web, but this was not always the case. In the 1970s, chemical structures would be drawn by hand or perhaps by using a fine drawing pen and a stencil. The first chemical drawing software package was also a chemical storage system called MACCS (Molecular ACCess System) produced by the software company MDL, which was set up in 1978. MDL was originally intended to offer consultancy in computer-aided drug design, but the founders soon realized that their customers were more interested in the tools that they had developed



```
9  8  0  0  0  0  0  0  0  0  1 V2000
   12.3345   -6.8066    0.1462 C   0  0  0  0  0  0  0  0  0  0  0  0
   13.6295   -6.0048   -0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   14.7418   -6.8848    0.0703 O   0  0  0  0  0  0  0  0  0  0  0  0
   12.1293   -6.9337    1.1167 H   0  0  0  0  0  0  0  0  0  0  0  0
   12.4445   -7.6994   -0.2908 H   0  0  0  0  0  0  0  0  0  0  0  0
   11.5844   -6.3102   -0.2908 H   0  0  0  0  0  0  0  0  0  0  0  0
   13.7202   -5.3346    0.7366 H   0  0  0  0  0  0  0  0  0  0  0  0
   13.6625   -5.5368   -0.8831 H   0  0  0  0  0  0  0  0  0  0  0  0
   15.3830   -6.9423   -0.6949 H   0  0  0  0  0  0  0  0  0  0  0  0
  2  1  1  0  0  0  0
  3  2  1  0  0  0  0
  4  1  1  0  0  0  0
  5  1  1  0  0  0  0
  6  1  1  0  0  0  0
  7  2  1  0  0  0  0
  8  2  1  0  0  0  0
  9  3  1  0  0  0  0
```

Figure 1.4 Connection table for ethanol in the MDL mol file format.

for handling chemical information and so MACCS was marketed in 1979. MDL may justly be regarded as the first of the cheminformatics software companies.

MACCS allowed chemists to sketch molecules using a suitable graphics terminal equipped with a mouse or a light pen [42] and then to store the compound in a computer using a file containing the information in a format called a connection table. An example of a simple connection table for ethanol is shown in Figure 1.4. The connection table shows the atoms, preceded in this case by their 3-D coordinates, followed by a list of the connections between the atoms, hence the name. The MACCS system stored extra information known as keys, which allowed a database of structures to be searched rapidly for compounds containing a specific structural feature or a set of features such as rings, functional groups, and so on. One of the problems with the use of connection tables to store structures is the space they occupy as they require a dozen or more bytes of data to represent every atom and bond. An alternative to connection tables is the use of line notation as discussed below.

1.7.1 Line Notations

Even though Berzelius had introduced a system that allowed chemical elements to be expressed within a body of text, there was still a need to show the structure of a polyatomic molecule. Structural formulas became more common, and the conventions used to express them were enforced by international committees, scientific publications, and organizations, such as Beilstein and Chemical Abstracts. However, there were two areas where the contemporary technology restricted the value of structural formula.

First, in published articles, printing techniques often separated illustrative pictures from the text so authors attempted to put the formula in the body of the text in a line format. This gave it authority, as well as relevance to the surrounding text. Once you move away from linear formulas constrained to read left to right by the text in which they are embedded, you need to provide a whole lot of information like numbering the atoms to ensure that all the readers get the same starting point for the eye movement, which recognizes the structure. So linear representations continued, certainly as late as 1903, for structures as complicated as indigo [43]. Even today we may write C_6H_5OH. It has the advantage of being compact and internationally understood and to uniquely represent a compound, which may be known as phenol or carbolic acid in different contexts.

Second, organizations such as Beilstein and Chemical Abstracts needed to be able to curate and search the data they were holding about chemicals. Therefore, attempts were made to introduce systematic naming. So addressing the numbering issues alluded to above. Unfortunately, different organizations had different systematic names (Chemical Abstracts, Beilstein, IUPAC), which also varied with time so you needed to know, for instance, which Collective Index of Chemical Abstracts you were accessing to know what the name of a particular chemical was (see Reference 44 for details). The upside for the organization was that the chemical names, *within the organization*, were standard so they could use the indexing and sorting techniques already available for text to handle chemical structures. With the advent of punched cards and mechanical sorting, the names needed to be more streamlined and less dependent on an arbitrary parent structure, and thus there was a need for a linear notation system that could be used to encode any complex molecule.

Just such a system of nomenclature, known as WLN, had been invented by William Wisswesser in 1949 [45]. WLN used a complex set of rules to determine how a molecule was coded. A decision had to be made about what was the parent ring system, for example, and the "prime path" through the molecule had to be recognized. WLN had the advantage that there was only one valid WLN for a compound, but coding a complex molecule might not be clear even to experienced people, and disputes were settled by a committee. Even occasional users of WLN needed to attend a training course lasting several days, and most companies employed one or more WLN "experts." An example of WLN coding is shown below:

6-dimethylamino-*4-phenylamino*-naphthalene-2-sulfonic acid;

the WLN is

L66J *BMR&* DSWQ **INI&1**.

Here the four sections of the WLN have been separated by spaces (which does not happen in a regular WLN string) to show how the four sections of the sulfonic acid, indicated by regular text, italic, underline, and bold, have been coded into WLN.

Beilstein, too, made a foray into line notations with ROSDAL, which required even more skill to ensure you had the correct structure. The corresponding ROSDAL code for the sulfonic acid above is

1=-5-=10=5,10-1,1-11N-12-=17=12,3-18S-19O,18=20O,18=21O, 8-22N-23,22-24.

Despite the complexity of the system and other problems [46], WLN became heavily used by the pharmaceutical industry and by Chemical Abstracts and was the basis for CROSSBOW (Computerized Retrieval Of StructureS Based On Wiswesser), a chemical database system that allowed susbstructure searching, developed by ICI pharmaceuticals in the late 1960s.

A different approach was taken by Dave Weininger, who developed SMILES in the 1980s [47,48]. This system, which required only five rules to specify atoms, bonds, branches, ring closures, and disconnections, was remarkable easy to learn compared to any other line notation system. In fact it was so easy to learn that "SMILES" was the reaction from anyone accustomed to using a line notation system such as WLN when told that they could learn to code in SMILES in about 10 minutes since it only had five rules. One of the reasons for the simplicity of SMILES is that coding can begin at any part of the structure and thus it is not necessary to determine a parent or any particular path through the molecule. This means that there can be many valid SMILES strings for a given structure, but a SMILES interpreter will produce the same molecule from any of these strings.

This advantage is also a disadvantage if the SMILES line notation is to be used in a database system because a database needs to have only a single entry for a given chemical structure, something that a system such as WLN provides since there is only one valid WLN string for a molecule. The solution to this problem was to devise a means by which a unique SMILES could be derived from any SMILES string [49]. Table 1.2 shows some different valid SMILES strings for three different molecules with the corresponding unique SMILES.

Thus, the design aims of the SMILES line notation system had been achieved, namely, to encode the connection table using printable characters but allowing the same flexibility the chemist had when drawing the structure and reserving the standardization, so the SMILES could be used in a database system, to a computer algorithm. This process of canonicalization was exactly analogous to the conventions that the publishing houses had instigated for structural diagrams. Thus, for the sulfonic acid shown earlier, a valid SMILES is **c1ccccc1Nc2cc(S(=O)(=O)O)cc3c2cc(N(C)C)cc3** and the unique or canonical SMILES is **CN(C)c1ccc2cc(cc(Nc3ccccc3)c2c1)S(=O)(=O)O**.

It was of concern to some that the SMILES canonicalizer was a proprietary algorithm, and this has led to attempts to create another linear representation,

TABLE 1.2 Examples of Unique SMILES

CH_3CH_2OH (1), $CH_2{=}CHCH_2CH{=}CHCH_2OH$ (2), 4-Cl-3Br-Phenol (3)

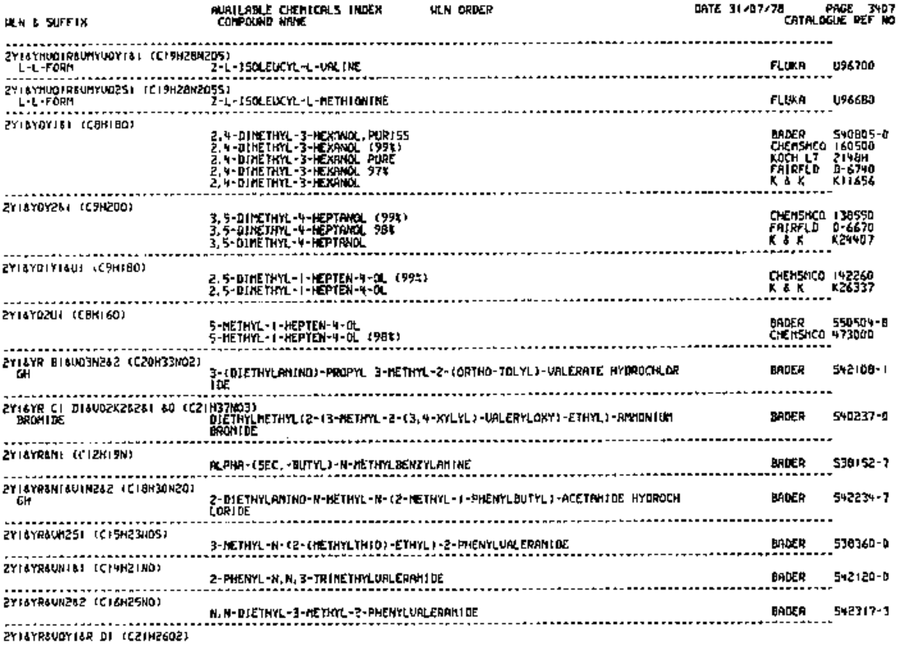
Compound	SMILES	Unique SMILES
1	OCC	CCO
1	CC(O)	CCO
1	C(O)C	CCO
2	C=CCC=CCO	OCC=CCC=C
2	C(C=C)C=CCO	OCC=CCC=C
2	OCC=CCC=C	OCC=CCC=C
3	OC1C=CC(Cl)=C(Br)C=1	Oc1ccc(Cl)c(Br)c1
3	Oc1cc(Br)c(Cl)cc1	Oc1ccc(Cl)c(Br)c1
3	c(cc1O)c(Cl)c(Br)c1	Oc1ccc(Cl)c(Br)c1

International Chemical Identifier (InChI), initially driven by IUPAC and NIH (for details, see Reference 50).

1.8 DATABASES

Nowadays, we take databases for granted. All kinds of databases are available containing protein sequences and structures, DNA sequences, commercially available chemicals, receptor sequences, small molecule crystal structures, and so on. This was not always the case, although the protein data bank was established in 1971 so it is quite an ancient resource. Other databases had to be created as the need for them arose. One such need was a list of chemicals that could be purchased from commercial suppliers. Devising a synthesis of new chemical entities was enough of a time-consuming task in its own right without the added complication of having to trawl through a set of supplier catalogs to locate the starting materials. Thus, the Commercially Available Organic Chemical Intermediates (CAOCI) was developed. Figure 1.5 shows an example of a page from a microfiche copy of the CAOCI from 1978 [51]. The CAOCI developed into the Fine Chemicals Directory, which, in turn, was developed into the Available Chemicals Directory (ACD) provided commercially by MDL.

The very early databases were simply flat computer files of information. These could be searched using text searching tools, but the ability to do complex searches depended on the way that the data file had been constructed in the first place, and it was unusual to be able to search more than one file at a time. This, of course, was a great improvement on paper- or card-based systems, but these early databases were often printed out for access. The MACCS chemical database system was an advance over flat file systems since this allowed structure and substructure searching of chemicals. The original MACCS system stored little information other than chemical structures, but a combined data and chemical information handling system (MACCS-II) was soon developed.



```
                              AVAILABLE CHEMICALS INDEX      WLN ORDER                   DATE 31/07/78      PAGE  3407
WLN & SUFFIX                   COMPOUND NAME                                                          CATALOGUE REF NO
--------------------------------------------------------------------------------------------------------------
2Y1&YMUQ1R&UMYUOY1&1 (C19H28N2O5)
  L-L-FORM                    2-L-ISOLEUCYL-L-VALINE                                      FLUKA     U96700
--------------------------------------------------------------------------------------------------------------
2Y1&YMUQ1R&UMYU02S1 (C19H28N2O5S)
  L-L-FORM                    2-L-ISOLEUCYL-L-METHIONINE                                  FLUKA     U966B0
--------------------------------------------------------------------------------------------------------------
2Y1&YQY1&1 (C8H18O)
                              2,4-DIMETHYL-3-HEXANOL,PURISS                               BADER     540805-0
                              2,4-DIMETHYL-3-HEXANOL (99%)                                CHEMSMCO  160500
                              2,4-DIMETHYL-3-HEXANOL PURE                                 KOCH LT   2148H
                              2,4-DIMETHYL-3-HEXANOL 97%                                  FAIRFLD   D-6740
                              2,4-DIMETHYL-3-HEXANOL                                      K & K     K11656
--------------------------------------------------------------------------------------------------------------
2Y1&YQY2&1 (C9H20O)
                              3,5-DIMETHYL-4-HEPTANOL (99%)                               CHEMSMCO  138550
                              3,5-DIMETHYL-4-HEPTANOL 98%                                 FAIRFLD   D-6670
                              3,5-DIMETHYL-4-HEPTANOL                                     K & K     K24407
--------------------------------------------------------------------------------------------------------------
2Y1&YQ1Y1&U1 (C9H18O)
                              2,5-DIMETHYL-1-HEPTEN-4-OL (99%)                            CHEMSMCO  142260
                              2,5-DIMETHYL-1-HEPTEN-4-OL                                  K & K     K26337
--------------------------------------------------------------------------------------------------------------
2Y1&YQ2U1 (C8H16O)
                              5-METHYL-1-HEPTEN-4-OL                                      BADER     550504-8
                              5-METHYL-1-HEPTEN-4-OL (98%)                                CHEMSMCO  473000
--------------------------------------------------------------------------------------------------------------
2Y1&YR B1&VO3N2&2 (C20H33NO2)
  GH                          3-(DIETHYLAMINO)-PROPYL 3-METHYL-2-(ORTHO-TOLYL)-VALERATE HYDROCHLOR   BADER     542108-1
                              IDE
--------------------------------------------------------------------------------------------------------------
2Y1&YR C1 D1&VO2K2&2&1 &Q (C21H37NO3)
  BROMIDE                     DIETHYLMETHYL(2-(3-METHYL-2-(3,4-XYLYL)-VALERYLOXY)-ETHYL)-AMMONIUM    BADER     540237-0
                              BROMIDE
--------------------------------------------------------------------------------------------------------------
2Y1&YR&M1 (C12H19N)
                              ALPHA-(SEC.-BUTYL)-N-METHYLBENZYLAMINE                      BADER     538152-7
--------------------------------------------------------------------------------------------------------------
2Y1&YR&N1&V1N2&2 (C18H30N2O)
  GH                          2-DIETHYLAMINO-N-METHYL-N-(2-METHYL-1-PHENYLBUTYL)-ACETAMIDE HYDROCH   BADER     542234-7
                              LORIDE
--------------------------------------------------------------------------------------------------------------
2Y1&YR&VM2S1 (C15H23NOS)
                              3-METHYL-N-(2-(METHYLTHIO)-ETHYL)-2-PHENYLVALERAMIDE        BADER     538360-0
--------------------------------------------------------------------------------------------------------------
2Y1&YR&VN1&1 (C14H21NO)
                              2-PHENYL-N,N,3-TRIMETHYLVALERAMIDE                          BADER     542120-0
--------------------------------------------------------------------------------------------------------------
2Y1&YR&VN2&2 (C16H25NO)
                              N,N-DIETHYL-3-METHYL-2-PHENYLVALERAMIDE                     BADER     542317-3
--------------------------------------------------------------------------------------------------------------
2Y1&YR&VOY1&R D1 (C21H26O2)
```

Figure 1.5 Entry (p. 3407) from the available chemical index of July 1978.

The great advance in database construction was the concept of relational databases as proposed by E.F. Codd, an IBM researcher, in 1970 [52]. At first, this idea was thought to be impractical because the computer hardware of the day was not powerful enough to cope with the computing overhead involved. This soon changed as computers became more powerful. Relational databases are based on tables where the rows of the table correspond to an individual entry and the columns are the data fields containing an individual data item for that entry. The tables are searched (related) using common data fields. Searching requires the specification of how the data fields should be matched, and this led to the development, by IBM, of a query "language" called Structured Query Language (SQL).

One of the major suppliers of relational database management software is Oracle Corporation. This company was established in 1977 as a consulting company, and one of their first contracts was to build a database program for the CIA code named "oracle." The adoption of a relational database concept and the use of SQL ensured their success and as a reminder of how they got started, the company is now named after that first project.

About 10 years ago, Oracle through its cartridges [53], along with other relational database providers such as Informix with its DataBlades [54], allowed users to add domain-specific data and search capability to a relational database. This is a key step forward as it allows chemical queries to be truly integrated with searches on related data. So, for instance, one can ask for "all compounds which are substructures of morphine which have activity in test1 > 20 and log P < 3 but have not been screened for mutagenicity, and there is >0.01 mg available." The databasing software optimizes the query and returns the results. These technologies, while having clear advantages, have not been taken up wholesale by the pharmaceutical industry. Some of this is for economic reasons, but also there has been a shift in the industry from a hypothesis-testing approach, which required a set of compounds to be preselected to test the hypothesis [55], to a "discovery"-based approach driven by the ability to screen large numbers of compounds first and to put the intellectual effort into analyzing the results.

1.9 LIBRARIES AND INFORMATION

In the 1970s, each company would have an information (science) department whose function was to provide access to internal and external information. This broad description of their purpose encompassed such diverse sources as internal company reports and documents, the internal compound collection, external literature, patents both in-house and external, supplier's collections, and so on. Part of their function included a library that would organize the circulation of new issues of the journals that the company subscribed to, the storage and indexing of the journal collection and the access, through interlibrary loans, of other scientific journals, books, and information. Company libraries have now all but disappeared since the information is usually delivered directly to the scientist's desk, but the other functions of the information science departments still exist, although perhaps under different names or in different parts of the organization. The potential downside to this move of chemical information from responsibility of the specialists is that there is a loss of focus in the curation of pharmaceutical company archives. Advances in data handling in other disciplines no longer have a channel to be adapted to the specialist world of chemical structures. The scientist at his/her desk is not likely to be able to influence a major change in company policy on compound structure handling and so will settle for the familiar and will keep the status quo. This could effectively prevent major advances in chemical information handling in the future.

1.10 SUMMARY

From the pen and paper of the 19th century to the super-fast desktop PCs of today, the representation of chemical structure and its association with data has kept pace with evolving technologies. It was driven initially by a need to communicate information about chemicals and then to provide archives, which could be searched or in today's terminology "mined." Chemistry has always been a classification science based on experiment and observation, so a tradition has built up of searching for and finding relationships between structures based on their properties. In the pharmaceutical industry particularly, these relationships were quantified, which allowed the possibility of predicting the properties of a yet unmade compound, totally analogous to the prediction of elements by Mendeleev through the periodic table. Data representation, no matter what the medium, has always been "backward compatible." For instance, as we have described, for many pharmaceutical companies, it was necessary to be able to convert legacy WLN files into connection tables to be stored in the more modern databases. This rigor has ensured that there is a vast wealth of data available to be mined, as subsequent chapters in this book will reveal.

REFERENCES

1. Berzelius JJ. Essay on the cause of chemical proportions and some circumstances relating to them: Together with a short and easy method of expressing them. *Ann Philos* 1813;2:443–454.
2. Klein U. Berzelian formulas as paper tools in early nineteenth century chemistry. *Found Chem* 2001;3:7–32.
3. Laszlo P. *Tools and Modes of Representation in the Laboratory Sciences*, p. 52. London: Kluwer Academic Publishers, 2001.
4. Web page of Douglas Jones of the University of Iowa. Available at http://www.cs.uiowa.edu/~jones/pdp8/.
5. Boyd DB, Marsh MM. *Computer Applications in Pharmaceutical Research and Development*, pp. 1–50. New York: Wiley, 2006.
6. Wikipedia contributors. Edge-notched card. *Wikipedia, The Free Encyclopedia*. Available at http://en.wikipedia.org/w/index.php?title=Edge-notched_card&oldid=210269872 (accessed May 12, 2008).
7. Weininger D, Delany JJ, Bradshaw J. *A Brief History of Screening Large Databases*. Available at http://www.daylight.com/dayhtml/doc/theory/theory.finger.html#RTFToC77 (accessed May 12, 2008).
8. Boyd DB. *Reviews in Computational Chemistry*, Vol. 23, pp. 401–451. New York: Wiley-VCH, 2007.
9. Richon AB. An early history of the molecular modeling industry. *Drug Discov Today* 2008;13:659–664.
10. Network Science web site and links thereon. Available at http://www.netsci.org/Science/Compchem/.
11. Randic M. On characterization of molecular branching. *J Am Chem Soc* 1975;97:6609–6615.
12. Kier LB, Hall LH. *Molecular Connectivity in Chemistry and Drug Research*. New York: Academic Press, 1976.
13. Topliss JG, Edwards RP. Chance factors in studies of quantitative structure-activity relationships. *J Med Chem* 1979;22:1238–1244.

14. Huuskonen JJ, Rantanen J, Livingstone DJ. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotopological state indices. *Eur J Med Chem* 2000;35:1081–1088.

15. Livingstone DJ, Ford MG, Huuskonen JJ, Salt DW. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J Comput Aided Mol Des* 2001;15:741–752.

16. Hansch C, Maloney PP, Fujita T, Muir RM. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 1962;194:178–180.

17. Fujita T, Iwasa J, Hansch C. A new substituent constant, π, derived from partition coefficients. *J Am Chem Soc* 1964;86:5175–5180.

18. Nys GC, Rekker RF. Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. Introduction of hydrophobic fragmental constants (f values). *Eur J Med Chem* 1964;8:521–535.

19. Hansch C, Leo AJ. *Substituent Constants for Correlation Analysis in Chemistry and Biology*, pp. 18–43. New York: Wiley, 1979.

20. Livingstone DJ. Theoretical property predictions. *Curr Top Med Chem* 2003;3: 1171–1192.

21. Tetko IV, Livingstone DJ. *Comprehensive Medicinal Chemistry II: In Silico Tools in ADMET*, Vol. 5, pp. 649–668. Elsevier, 2006.

22. Yoneda F, Nitta Y. Electronic structure and antibacterial activity of nitrofuran derivatives. *Chem Pharm Bull Jpn* 1964;12:1264–1268.

23. Snyder SH, Merril CR. A relationship between the hallucinogenic activity of drugs and their electronic configuration. *Proc Nat Acad Sci USA* 1965;54:258–266.

24. Neely WB, White HC, Rudzik A. Structure-activity relations in an imidazoline series prepared for their analgesic properties. *J Pharm Sci* 1968;57:1176–1179.

25. Saunders MR, Livingstone DJ. *Advances in Quantitative Structure-Property Relationships*, pp. 53–79. Greenwich, CT: JAI Press, 1996.

26. Glen RC, Rose VS. Computer program suite for the calculation, storage and manipulation of molecular property and activity descriptors. *J Mol Graph* 1987; 5:79–86.

27. Livingstone DJ, Evans DA, Saunders MR. Investigation of a charge-transfer substituent constant using computer chemistry and pattern recognition techniques. *J Chem Soc Perkin 2* 1992;1545–1550.

28. Hyde RM, Livingstone DJ. Perspectives in QSAR: Computer chemistry and pattern recognition. *J Comput Aided Mol Des* 1988;2:145–155.

29. Todeschini R, Consonni V. *Handbook of Molecular Descriptors*. Mannheim: Wiley-VCH, 2000.

30. Livingstone DJ. The characterisation of chemical structures using molecular properties—A survey. *J Chem Inf Comput Sci* 2000;40:195–209.

31. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone DJ, Ertl P, Palyulin VA, Radchenko EV, Makarenko AS, Tanchuk VY, Prokopenko R. Virtual Computational Chemistry Laboratory. Design and description. *J Comput Aided Mol Des* 2005;19:453–463. Available at http://www.vcclab.org/.

32. Hansch C, Unger SH, Forsythe AB. Strategy in drug design. Cluster analysis as an aid in the selection of substituents. *J Med Chem* 1973;16:1217–1222.
33. Martin YC, Holland JB, Jarboe CH, Plotnikoff N. Discriminant analysis of the relationship between physical properties and the inhibition of monoamine oxidase by aminotetralins and aminoindans. *J Med Chem* 1974;17:409–413.
34. Livingstone DJ, Salt DW. Judging the significance of multiple linear regression models. *J Med Chem* 2005;48:661–663.
35. Salt DW, Ajmani S, Crichton R, Livingstone DJ. An Improved Approximation to the estimation of the critical F values in best subset regression. *J Chem Inf Model* 2007;47:143–149.
36. Livingstone DJ. Molecular design and modeling: Concepts and applications. In: *Methods in Enzymology*, Vol. 203, pp. 613–638. San Diego, CA: Academic Press, 1991.
37. Aoyama T, Suzuki Y, Ichikawa H. Neural networks applied to structure-activity relationships. *J Med Chem* 1990;33:905–908.
38. Manallack DT, Livingstone DJ. Artificial neural networks: Application and chance effects for QSAR data analysis. *Med Chem Res* 1992;2:181–190.
39. Manallack DT, Ellis DD, Livingstone DJ. Analysis of linear and non-linear QSAR data using neural networks. *J Med Chem* 1994;37:3758–3767.
40. Livingstone DJ, Manallack DT, Tetko IV. Data modelling with neural networks—Advantages and limitations. *J Comput Aided Mol Des* 1997;11:135–142.
41. Applications of artificial neural networks to biology and chemistry, artificial neural networks. In: *Methods and Applications Series: Methods in Molecular Biology*, Vol. 458. Humana, 2009.
42. http://depth-first.com/articles/2007/4 (accessed May 20, 2008).
43. Bamberger E, Elger F. Über die Reduction des Orthonitroacetophenons- ein Betrag zur Kenntis der ersten Indigosynthese. *Ber Dtsch Chem Ges* 1903;36: 1611–1625.
44. Fox RB, Powell WH. *Nomenclature of Organic Compounds: Principle and Practice*. Oxford: Oxford University Press, 2001.
45. Wisswesser WJ. How the WLN began in 1949 and how it might be in 1999. *J Chem Inf Comput Sci* 1982;22:88–93.
46. Bradshaw J. Introduction to Chemical Info Systems. Available at http://www.daylight.com/meetings/emug02/Bradshaw/Training/ (accessed May 12, 2008).
47. Weininger D. Smiles 1. Introduction and encoding rules. *J Chem Inf Comput Sci* 1988;28:31–36.
48. SMILES—A Simplified Chemical Language. Available at http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html (accessed May 25, 2008).
49. Weininger D, Weininger A, Weininger JL. SMILES 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* 1989;29:97–101.
50. http://www.InChI.info/ (accessed May 25, 2008).
51. Walker SB. Development of CAOCI and its use in ICI plant protection division. *J Chem Inf Comput Sci* 1983;23:3–5.
52. Codd EF. A relational model of data for large shared data banks. *Commun ACM* 1970;13:377–387.

53. De Fazio S. Oracle8 Object, Extensibility, and Data Cartridge Technology. Available at http://www.daylight.com/meetings/mug98/DeFazio/cartridges.html (accessed June 5, 2008).

54. Anderson J. Taking Advantage of Informix DataBlade Technology. Available at http://www.daylight.com/meetings/mug98/Anderson/datablades.html (accessed June 5, 2008).

55. Bradshaw J. *Chronicles of Drug Discovery*, Vol. 3, pp. 45–81. Washington DC: ACS, 1993.