three-dimensional (3D) models became possible using the program CONCORD, introduced by Pearlman in 1987 (34). This enabled the introduction of 3D structural databases with the ability to generate, store, and search 3D molecular models on a large scale. These 3D database systems included ALADDIN by Daylight Chemical Information Systems, UNITY3D by Tripos, CHEMDBS3D by Chemical Design Ltd., and MACCS3D by MDL (35).

1.3.3 The 1990s—Relational Data Storage. This period saw the decline of single-computer mainframe chemical management programs and the rise of server-based systems and distributed computing. By far, the most significant influences on chemical information management were the Internet, the introduction of relational database technology, and the shift to high-throughput combinatorial chemistry. In a relational database, information that formerly was kept in a single large table is stored in numerous smaller tables, indexed by "keys." This is a much more flexible architecture, and combining different fields from several tables into a "view" of the data gives the user the impression of a single large table, as before. At the end of the decade, chemical and pharmaceutical firms could obtain chemical structure, reaction, and 3D model databases from a variety of vendors. These databases were even somewhat integrated with molecular modeling, quantum mechanics, and docking programs, and to literature, spectra, and biological databases. The largest database of known chemical structures, the Chemical Abstracts Registry, grew to about 20 million structures, whereas a typical corporate inventory increased to between 100,000 and 1,000,000 structures. A database of billions of virtual chemical structures was constructed and made available for drug-design purposes by Tripos, Inc. (36).

1.3.4 The 2000s. Like the customization and distributed computing of the 1980s that followed the introduction of mini-mainframe systems, the 2000s are witnessing the customization and further distribution of relational and integrated database systems. Chemical structure-specific and reaction-specific search types can be integrated into relational databases, to take maximal advantage of the scale and performance of these systems. We see the increasing use of web-based clients, also known as "thin" clients, because they need little software other than a web browser. Former single databases are turning into distributed and replicated database systems, and we see increasing use of data marts and data warehouses, more fully integrated structure, reaction, data, and citation searching, and increasingly "intelligent" database systems.

2 CHEMICAL REPRESENTATION

Chemical structures and reactions can be represented in many ways. At the most fundamental level, the parameters of the time-dependent Schrödinger equation—the atomic and molecular orbitals—do a more or less complete job of characterizing a chemical compound. Storing and representing structures as mathematical wave functions is obviously not suitable for thousands or millions of structures; nor is such a representation useful for drug discovery, except perhaps to a molecular modeler. Synthetic chemists still function in a mostly 2D chemical structure space. Intuition, training, and experience allow a chemist to extrapolate from a flat representation with a few stereochemical hints—dashed and wedged bonds or Z/E double bonds—to a higher-dimensional mental representation of a structure. Chemical representation systems are a compromise of several factors, including the needs of the chemist, the storage and performance characteristics of the chemical database system, and the ultimate 3D reality of chemical structures.

2.1 Types of Chemical Entities

There are several ways to look at chemical representation. One approach is to classify according to the type of chemical data that is stored. The most basic types of chemical structure data are shown in Fig. 9.4, including the following.

2.1.1 Sequences. For linear chemical systems, such as DNA, RNA, and proteins, the sequence of subunits (nucleotide bases or amino acids) provides most of the information