Representation of chemical structures

Wendy A. Warr*

At the root of applications for substructure and similarity searching, reaction retrieval, synthesis planning, drug discovery, and physicochemical property prediction is the need for a machine-readable representation of a structure. Systematic nomenclature is unsuitable, and notations and fragment codes have been superseded, except in certain specific applications. Connection tables are widely used, but there is no formal standard. Recently the International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChI) has started to attract interest. This review also summarizes the representation of chemical reactions and three-dimensional structures. © 2011 John Wiley & Sons, Ltd. *WIREs Comput Mol Sci* 2011 1 557–579 DOI: 10.1002/wcms.36

INTRODUCTION

Computers have been used since the 1960s for storing chemical structures in databases and for making use of chemical structural information in applications such as similarity searching, reaction retrieval, synthesis planning, drug discovery, and physicochemical property prediction. At the root of all these applications is the need for a machine-readable representation of a structure. Although there are two well-established ways of naming compounds, overseen by the International Union of Pure and Applied Chemistry (IUPAC)[1–3] and Chemical Abstracts Service (CAS),[4] systematic chemical nomenclature is not suitable for chemical structure handling systems because names are often long and complex, as are the rules used to generate them, whereas the use of trivial names and nonunique names further complicates the issue. It is worth noting, though, that several programs have been written which successfully convert a high proportion of names into machine-readable structures,[5–14] and there are also programs that can assign systematic names for input structures.[5,15–19] ACD/Labs (Toronto, Canada), CambridgeSoft (Cambridge, MA), ChemAxon (Budapest, Hungary), InfoChem (Munich, Germany), OpenEye Scientific Software (Santa Fe, NM, USA), and the University of Cambridge (Cambridge, UK) have all worked on converting names to structures.

To the practicing chemist, the language of chemistry is the two-dimensional (2D) structure diagram and most chemical information systems feature graphical input and output of chemical structures; the machine-held representation need not be meaningful to the synthetic chemist. In the ideal (unique) representation there is only one 'code' for a given structure and any one code can be interpreted to give only one structure. A unique representation is essential for chemical registration systems in which the novelty of a structure is determined before it is recorded in a database. Some representations, for example, molecular formulas, are not unique; one molecular formula will generate more than one full structure. Some nonunique representations (e.g., molecular formulas and fragment codes) do, however, play a part in certain chemical information systems, even though they do not represent the full topology of a structure.

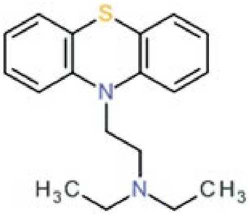
NOTATIONS

Line notations represent structures as a linear string of alphanumeric symbols. Their compactness was an advantage in the early days of cheminformatics when storage space was at a premium, and even nowadays, it can be faster to enter a structure as a notation instead of using a chemical structure drawing program. Several notations[20–22] were proposed in the 1950s and 1960s, but only one, the Wiswesser Line-Formula Notation (WLN; see Figure 1)[23–28] became widely used,[29–47] despite the fact that the Dyson notation was formally adopted by IUPAC.[20,21] WLN started to fall out of use in the early 1980s. The range

*Correspondence to: wendy@warr.com

Wendy Warr & Associates, Holmes Chapel, Cheshire, England

DOI: 10.1002/wcms.36



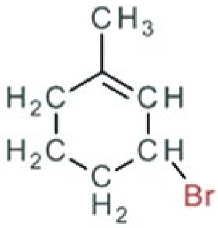
T C666 BN ISJ B2N2&2

FIGURE 1 | A Wiswesser Line Notation.

of structures it could handle was limited, it was unable to encode the finer points of structures such as stereochemistry, and its arcane rules were unacceptable in the new era of user-friendly systems for use by chemists without the need for an intermediary.

Later in the 1980s the Simplified Molecular Input Line Entry System (SMILES; see Figure 2)[48,49] was developed at Pomona College (Claremont, CA) and implemented by Daylight Chemical Information Systems (Santa Fe, NM). SMILES is still widely used today. Daylight uses an extension of SMILES called SMARTS to describe structure queries for searching chemical databases. Sybyl Line Notation (SLN)[50,51] is also still in use with software from Tripos (St. Louis, MO). Another notation, called representation of structure diagram arranged linearly (ROSDAL),[52,53] was written to transfer structures quickly in a compact form over a network to enable searching of the Beilstein database online.[54] ROSDAL is still supported by InfoChem, and by Elsevier (Amsterdam, The Netherlands) in Reaxys (*vide infra*) and the Beilstein CrossFire structure editor.

A given chemical structure can have many valid and unambiguous representations (e.g., it is possible to start with any atom to derive a SMILES string) but for comparison purposes it is desirable to have a unique representation known as the 'canonical' one. WLNs were 'canonicalized' (or 'canonized') by choosing the one that occurred first alphanumerically. More efficient methods have been devised for deriving a unique SMILES for any structure,[49] but nowadays the usual way of storing structures in a computer is a canonicalized connection table. A connection table is a listing of atoms and bonds, and other data, in tabular form.



CC1=CC(Br)CCC1

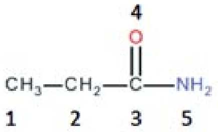
FIGURE 2 | A Simplified Molecular Input Line Entry System notation.

CONNECTION TABLES

A redundant connection table is shown in Figure 3. It is termed redundant because each connection is described twice. The redundancy is removed when a unique version of the table is stored. Hydrogen atoms are not necessarily included explicitly in a connection table: they may be implicit. Canonicalization usually involves renumbering the atoms in a unique and reproducible way, for example, using the Morgan algorithm initially developed by Gluck at DuPont (Wilmington, DE, USA) and adapted by CAS.[55] When the database is constructed, issues such as aromaticity, tautomerism, and stereochemistry are addressed before canonical numbering. The 2D coordinates needed to display a structure retrieved from the database may be stored in the connection table to make depiction easy or consistent. If they are not stored, depiction software will be needed. 'Laying out' a pleasing structure is no simple task[7,56–58]; this is a key task, for example, in algorithms that generate chemical structures from systematic names.

Once structures are stored in connection tables in a database, they can be searched by substructure, that is, all the molecules in the database that contain a specified substructure can be identified.[59,60] Also, full structures that match exactly can be retrieved. Substructure searches are carried out by treating the structure as a graph and then applying graph theoretical algorithms to carry out the match. Topological graph theory is a branch of mathematics particularly useful in cheminformatics. The atoms of a structure are treated as nodes in a graph and the bonds as edges joining the nodes. The nodes and edges can be 'colored' to distinguish them (e.g., oxygen atoms or double bonds). Of course, chemical structures and topological graphs are not entirely equivalent: a connection table is akin to a description of a single valence bond structure and does not take account, for example, of delocalized bonds.

Alternative approaches have been suggested. Dietz[61] has proposed a 'molecular multigraph': a connected, labeled, and undirected multigraph whose



Atom number	Atomic symbol	Bond order	Attached atom number	Bond order	Attached atom number	Bond order	Attached atom number
1	C	1	2				
2	C	1	1	1	3		
3	C	1	2	2	4	1	5
4	O	2	3				
5	N	1	3				

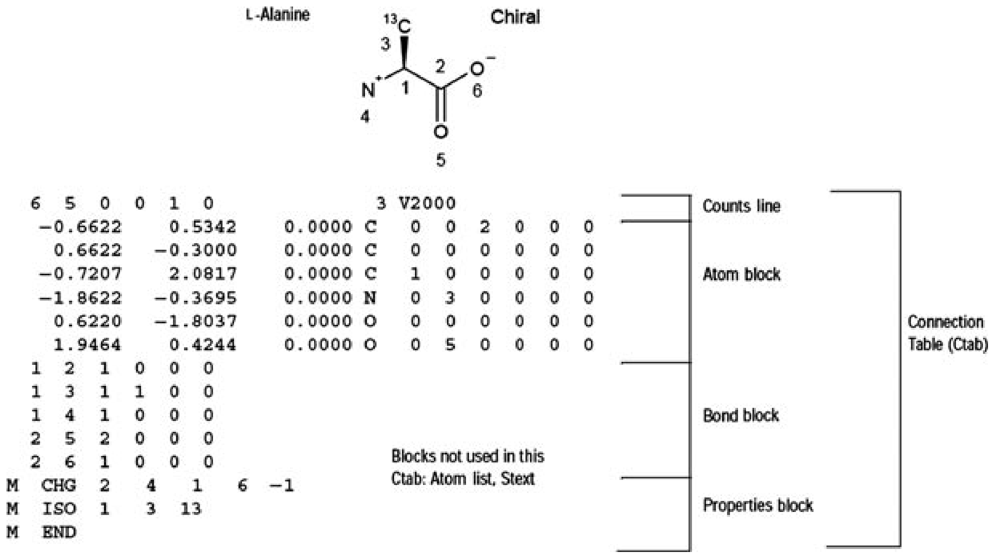
FIGURE 3 | A redundant connection table.

vertices are atoms and whose edges are bonding relations. A multigraph may have several edges between the same two vertices, so it is no longer possible to represent a structure in a simple matrix form. Bauerschmidt and Gasteiger[62] have also recognized the limits of a connection table description and its unsuitability for handling delocalized π-systems, inorganics, coordination compounds, and reaction intermediates. Their system separates σ- and π-electrons into two bond types, σ- and π-electron systems, and introduces a third bond type for coordination compounds. Electrons may be delocalized between more than two atoms in all the three bond types. The Molecular Structure Encoding System, MOSES, from Molecular Networks (Erlangen, Germany) is a later development from Gasteiger's team.

The Morgan algorithm identifies atoms based on an extended connectivity value. The atom with the highest value becomes the first atom in the name, and its neighbors are then listed in descending order. Ties are resolved based on additional parameters, for example, bond order and atomic number. The original Morgan algorithm did not handle stereochemistry; the stereochemically unique naming algorithm [stereochemical extension of Morgan algorithm (SEMA)] was developed to handle stereoisomers.[63] SEMA was adopted by MDL Information Systems (now Symyx Software, San Ramon, CA). Symyx's newly enhanced Morgan algorithm (NEMA)[64] produces a unique name and key for a wider range of structures than SEMA (More will be said about keys in a later section.). The work of Wipke et al.[65] identified the value of a constitutional key and a stereo key. This approach has been incorporated into NEMA that extends perception to nontetrahedral stereogenic centers, and supports both 2D and three-dimensional (3D) stereochemistry perception.

The MDL (now Symyx) connection table (see Figure 4),[66] or CTfile, has become the *de facto* standard for exchange of datasets. It separates atoms and bonds into separate blocks. There are various versions of this connection table. A molecule file, or 'molfile,' describes a single molecular structure that can contain disjoint fragments. A molfile consists of a header block and a connection table. The header block identifies the file by molecule name, user's name, program, date, and so on. An Rgroup file (RGfile) describes a single molecular query with Rgroups (features of a generic structure, which will be discussed later). A reaction file (rxnfile) contains the structural information for the reactants and products of a single reaction. Structure–data files (SDfiles) contain structures and data for any number of molecules. Reaction–data files (**RDfiles) are s**imilar to SDfiles in concept, but the RDfile is a more general format that can include reactions as well as molecules, together with their associated data. XML data files (XDfiles) are a data format based on Extensible Markup Language (XML) for transferring record sets of structure or reaction information with associated data. An XDfile can contain structures or reactions that use any of the CTfile formats, Chime strings, or SMILES strings (Chime is an encrypted format that is used to render structures and reactions on a web page). A white paper detailing the latest version of the formats is available on the Symyx web site.

Different vendors have developed proprietary connection table formats. Efforts have been made to establish an agreed standard format but they have not been generally unsuccessful. The standard



```
  6  5  0  0  1  0            3 V2000
   -0.6622    0.5342    0.0000 C   0  0  2  0  0  0
    0.6622   -0.3000    0.0000 C   0  0  0  0  0  0
   -0.7207    2.0817    0.0000 C   1  0  0  0  0  0
   -1.8622   -0.3695    0.0000 N   0  3  0  0  0  0
    0.6220   -1.8037    0.0000 O   0  0  0  0  0  0
    1.9464    0.4244    0.0000 O   0  5  0  0  0  0
  1  2  1  0  0  0
  1  3  1  1  0  0
  1  4  1  0  0  0
  2  5  2  0  0  0
  2  6  1  0  0  0
M  CHG  2   4   1   6  -1
M  ISO  1   3  13
M  END
```

FIGURE 4 | MDL connection table organization.

molecular data (SMD) format[67–69] never gained wide acceptance. Standard file formats have, however, been established for crystallographic information and the IUPAC International Chemical Identifier (InChI) is now used fairly widely. These formats will be discussed later. Chemical Markup Language (CML)[70–80] uses the XML protocol for data exchange using the Internet.

A substructure search query can be matched against a connection table atom-by-atom but the so-called subgraph isomorphism algorithm that is used in substructure search to compare one graph against another is slow and complex and it is likely that there may be many mismatches before a hit is found. A substructure search can be carried out faster if an initial screening stage is carried out to filter out quickly structures that could not possibly be matches. A common method is to use substructure fragments as the filter.

Hyperstructures have been suggested as a way of representing the structural information in a set of connection tables in a nonredundant form, to reduce storage and processing costs.[81–83] A hyperstructure is a pseudomolecule formed by the superimposition of sets of molecules in such a way that areas of structural commonality are stored only once. The use of a fragment screen, however, is the more usual approach in improving the efficiency of substructure search.

FRAGMENT CODES

A fragment coding system is based on a collection of small substructures or features in a closed list (a controlled 'dictionary' of structural features) or an open-ended list, for example, all linear paths of up to a defined number of atoms, typically seven (paths of length zero, paths of length one, and so on). Historically, each fragment was represented by a hole in a punch card and the occurrence of any of these fragments in a given structure was recorded by punching the appropriate holes in a card. One code, Ringcode, developed by a group of companies called the Pharma Documentation Ring, was used by Derwent (London, England) (now Thomson Reuters) for the Chemical Reactions Documentation Service, and for literature searching.[32,84–87] Fragment codes have been particularly useful for storing and retrieving structures from patents. They can still be used for searching the Derwent World Patents Index (a Thomson Reuters database), although a software product called TOPFRAG can now be used to generate the codes automatically from a structure input graphically.

The Derwent World Patent Index Chemical Code[85,88] is a closed code with about 1000 terms. It can be searched online on Questel (Paris, France). The IFI/Plenum Code is an open-ended code used in the 'CLAIMS' database of US patents. It is searchable online on STN International (Columbus, OH, USA). A group of mainly German companies used a code called GREMAS (originally generic retrieval by magnetic tape storage) for many years for retrieving both patent and chemical reaction information,[85,88–97] but its use was discontinued in the 1990s. The GREMAS code was a very sophisticated, open-ended code with good retrieval performance.

Sub structural 'keys' from a fragment dictionary are usually recorded as a binary bitstring, or fingerprint: the fragments present in a structure can be represented as a sequence of 0s and 1s, where 0 means that the fragment is not present in the structure and 1 means that it is present in the structure. Each 0 or 1 can be represented as a single bit: the *i*th substructure in the dictionary corresponds to the *i*th bit in the bitstring. These bitstrings are often called structure 'fingerprints.'

Comparing fingerprint bitstrings is very fast[98] and is well suited to the screening stage of a substructure search. Much work was carried out in the 1970s to determine the most effective screening system, typically based on the frequency of occurrence of fragments.[99–104] The very large screen set dictionary devised by the BASIC group of chemical companies (Basel, Switzerland) is used in the online system for searching the CAS database.[94,105,106] CAS uses 12 different types of screens, including augmented atoms (a central atom with its neighboring atoms and bonds), atom sequences (linear sequences of connected atoms), bond sequences (atom sequences with the bonds differentiated but not the atoms), and screens associated with ring systems, such as the number of rings and the ring type.

The CAS screen set was specifically designed for use in substructure searching, as were the dictionary fingerprints used by Symyx, in the so-called MDL, ISIS, or MACCS keys[107–112] of 166 and 960 bits. On the contrary customized dictionaries from Digital Chemistry (Sheffield, UK; formerly Barnard Chemical Information) were designed for use in applications such as clustering of chemical structures to analyze their similarity or diversity[113–116] and distinguishing drugs from nondrugs.[117] Fingerprints originally designed for use in substructure searching have also been used as 'descriptors' in studies of the chemical diversity of collections of compounds and in quantitative structure–activity relationship (QSAR) analyses.[107–122]

The alternative to structural keys is a 'hashed fingerprint.' Each of the fragment paths in an open-ended set is submitted to a hashing procedure that sets a small number of bits (usually four or five) to 1 in the fingerprint bitstring. Hashed fingerprints are typically used in software from Daylight Chemical Information Systems. Tripos uses a combination of structural keys and hashed fingerprints. The French system Description, Acquisition, Recherche et Corrélation (DARC) uses a different sort of fragments called Fragments Reduced to an Environment which is Limited (FRELs).[94,123–125] FRELs describe two concentric layers of atoms around a focus, which is an atom with at least three (or in some cases two) neighbors. Around 1990, Questel offered online access to certain CAS databases online under DARC, but nowadays substructure search of CAS REGISTRY is possible only with systems supplied by CAS.

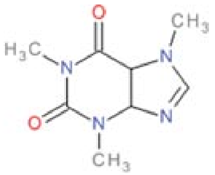
The DARC FRELs are a type of 'circular' fingerprint, as are SciTegic's (San Diego, CA, USA) (now Accelrys') extended connectivity fingerprints or 'ECFPs.'[126] ECFPs were developed specifically for QSAR. They have been widely used but details have only recently been published. ECFPs were developed specifically for QSAR, whereas circular fingerprints described much earlier by Willett[104] were developed as substructure search screens.

REGISTRY SYSTEMS

CAS REGISTRY, produced by CAS, is an authoritative collection of disclosed chemical substance information, containing more than 53 million organic and inorganic substances and more than 61 million sequences, abstracted from the patent and journal literature.[127–140] Each substance is identified by a CAS Registry Number (CAS RN) and there are links to the document where data on the molecule were published. In the Beilstein Registry file,[141,142] which also covers the scientific literature, data are stored with the compound. Beilstein is online on STN International[86,143] and Dialog (Morrisville, NC, USA);[54,143] the version on DIALOG was mentioned earlier in connection with ROSDAL. Software from InfoChem is now used for substructure searching on DIALOG. A very recent service, Reaxys, from Elsevier merges Elsevier's CrossFire Beilstein[144–147] and CrossFire Gmelin (inorganic chemistry) databases with the company's Patent Chemistry Database in a new workflow solution for synthetic chemists.

Chemical and pharmaceutical companies also have registration systems for their corporate compound collections. A key feature of a registry system is checking for the novelty of a chemical structure before registering in the database and assigning it a registry number. If the structure is not found to be novel, any new data related to the structure can be recorded alongside the structure already on file. A molecular formula is usually recorded with a structure, and consistency between the two items will be checked.

Registry numbers are often meaningless in themselves, that is, they contain no chemical information. A new addition to the CAS REGISTRY, for example, will be assigned the next highest number in a sequence. A CAS RN includes up to 10 digits that are separated into three groups by hyphens. The first part of the number, starting from the left, has up to seven digits; the second part has two digits; and the final part consists of a single check digit. CAS RNs are used in many other public and private databases as well as chemical inventory listings and are included in all CAS-produced databases. Proprietary registry numbers in the chemical and pharmaceutical industries often have a hierarchical structure, for example, parent compound, stereoisomer, salt, and batch.



InChI = 1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
InChIKey = RYYVLZVUVIJVGH-UHFFFAOYSA-N

FIGURE 5 | Example of an IUPAC International Chemical Identifier (InChI) and InChIKey.

IUPAC INTERNATIONAL CHEMICAL IDENTIFIER

IUPAC developed the InChI as a freely available, nonproprietary identifier for chemical substances that can be used in printed and electronic data sources, thus enabling easier linking of diverse data compilations and unambiguous identification of chemical substances. IUPAC decided to tackle this problem because the increasing complexity of molecular structures was making conventional naming procedures inconvenient, and because there was no suitable, openly available electronic format for exchanging chemical structure information over the Internet. The goal of InChI is to provide a unique string representing a chemical substance of known structure, independent of specific depiction, derived from a conventional connection table.[148] InChI is freely available and extensible. The InChI project was initially undertaken by IUPAC with the cooperation of the US National Institute for Standards and Technology (NIST).

An InChI is created from an input connection table in three steps: normalization, canonicalization, and serialization. In the normalization step, electron density is ignored; salts and metal atoms in organometallic compounds are disconnected; and mobile hydrogens, variable protonation, and charge are normalized. The step is needed, for example, to remove variations in the ways of representing a nitro group. NIST wrote the canonical numbering algorithm by modifying a more recent version[149] of the Morgan algorithm. In the final step, the labeled structure is serialized and the InChI character string is the output.

The identifier is hierarchically 'layered'; each layer holds a distinct and separable class of structural information, with the layers ordered to provide successive structural refinement. There are currently six InChI layer types, each representing a different class of structural information: the main layer, a charge layer, a stereochemical layer, an isotopic layer, a fixed-H layer, and a reconnected layer. Except for the main layer (atoms and their bonds), the presence of a layer is not required and appears only when corresponding input information has been provided. Layers and sublayers are separated by the forward slash (/) delimiter. Except in the case of the chemical formula layer, each layer starts (after the slash mark) with a lower-case letter to indicate the type of information held. An example is given in Figure 5; here, the connectivity layer begins with 'c' and the hydrogen layer with 'h.'

InChIKey is a condensed digital representation of the identifier. This key facilitates Web searching, previously complicated by unpredictable breaking of InChI character strings by search engines. It also allows development of Web-based InChI lookup services, permits an InChI representation to be stored in fixed length fields, and makes chemical structure database indexing easier.

The first part of the key is 14 characters long and encodes the molecular skeleton (connectivity). After a hyphen, there is a second string of 10 characters, the first eight of which encode stereochemistry and isotopes. The 10-character block ends with a flag character indicating that this is a standard InChIKey (produced out of standard InChI) and a version character indicating the version number of InChI. The key ends with a hyphen followed by a character indicating (de)protonation state. Both parts of the InChIKey are based on a truncated SHA-256 hash (secure hash standard)[150] of the corresponding InChI layers. There is a finite, but extremely small probability of finding two structures with the same InChIKey.

Figure 5 shows the standard InChI and InChIKey for caffeine. In the key, the first block of 14 letters (RYYVLZVUVIJVGH) encodes the molecular skeleton (connectivity). The first eight letters of the second block (UHFFFAOY) encode stereochemistry and isotopes. After that, 'S' indicates that the key was produced from standard InChI and 'A' indicates that version 1 of InChI was used. The final character, 'N,' means 'neutral.' Use of InChIKey allows searches based solely on atom connectivity (the first 14 characters).

Currently, the InChI algorithm can handle neutral and ionic organic molecules, radicals, and inorganic, organometallic, and coordination compounds. Because InChI is composed of hierarchical layers, new layers could be added to extend the scope of the identifier. Work is currently underway to extend InChI to include polymer representation and reactions.

It is worth noting how InChI differs from SMILES. Like InChI, the SMILES language allows a canonical serialization of molecular structure. However, SMILES is proprietary and unlike InChI, it is not an open project. This has led to the use of different generation algorithms, and thus, different SMILES versions of the same compound have been found. InChI is not a registry system as that of CAS; it does not depend on the existence of a database of unique substance records to establish the next available sequence number for any new chemical substance being assigned an InChI.

InChI has been used in chemical enhancement of the semantic Web[151–154] and in annotation of 3D structures.[155] The Royal Society of Chemistry (RSC) uses InChI in Project Prospect,[154] the aim of which is to make the science within RSC journal articles machine-readable through semantic enrichment and the integration of metadata into text. Text mining is used to attach structural information (InChI, SMILES, and CML) to chemical names.[9,13] A significant number of publishers and database producers are now using InChI. The Internet-based ChemSpider database uses InChI to register chemical structures,[154,156] and so does a system for chemical structure indexing of toxicity data on the Internet.[157]

OTHER IDENTIFIERS

The National Cancer Institute Computer Aided Drug Design (NCI/CADD) identifiers are calculated for the Chemical Structure Lookup Service (CSLS) on the Internet.[158,159] They are based on hashcodes calculated by the cheminformatics toolkit Chemical Algorithms Construction Threading and Verification System (CACTVS).[158,160,161] The National Institutes of Health's PubChem substructure search system is also based on CACTVS.[156] CACTVS hashcodes[162] (see Figure 6) represent a chemical structure uniquely as a 16-digit hexadecimal number, have a high sensitivity to structural features of a compound, and change if the connectivity changes. Structure normalization is performed for any incoming structure set to be registered, or searched by, in CSLS. Each parent structure is then subjected to a hashcode calculation to generate the NCI/CADD identifier.

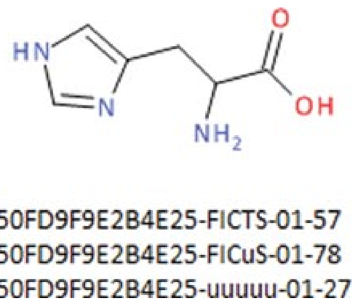

HN
N
O
OH
NH2
9850FD9F9E2B4E25-FICTS-01-57
9850FD9F9E2B4E25-FICuS-01-78
9850FD9F9E2B4E25-uuuuu-01-27

FIGURE 6 | National Cancer Institute Computer Aided Drug Design identifiers.

The normalization has adjustable levels of sensitivity. The Fragment Isotope Charge Tautomer Stereo (FICTS) identifier is a representation of the exact structure drawing, sensitive to all the five features. The FICuS identifier is not sensitive to tautomers ('u' stands for 'unsensitive'), and comes close to how chemists perceive a chemical. The uuuuu identifier links closely related forms. Currently, there are eight identifier variants defined for a structure: FICTS, FICTu, FICuS, FICuu, uuuTS, uuuTu, uuuuS, and uuuuu. Three of them, FICTS, FICuS, and uuuuu, are searchable for all the structure records in CSLS. Similar principles are used in Symyx's Flexmatch search.[98] Examples of Symyx's NEMA keys, compared with InChIKeys, are shown in Figures 7 and 8.

INTERCONVERSION

Broadly speaking, the more the information contained in the representation of a structure, the more likely it is that it will be faithfully convertible into another representation. Interconversion of connection tables is often possible. There is a difference between 'file format' connection tables and those used internally for algorithmic processing. The SMD format was not intended to be universal database format, still less an internal format for different systems. It was intended as an interchange format allowing different programs to exchange data with minimal need for conversion routines.[68]

Connection tables can, as we have seen, be converted to identifiers, but because there will be loss of information it is not possible; for example, to convert an InChIKey or a fingerprint into a connection table. There are resolvers on the Internet that allow a structure to be displayed for an InChI or InChIKey and from that a connection table could be made.

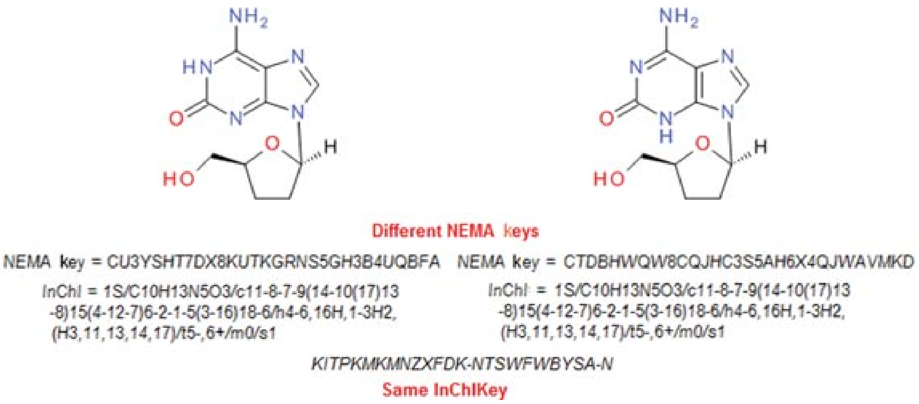

NH2
HN
N
N
O
N
O
H
HO
NH2
N
N
O
N
H
N
O
H
HO
Different NEMA keys
NEMA key = CU3YSHT7DX8KUTKGRNS5GH3B4UQBFA
NEMA key = CTDBHWQW8CQJHC3S5AH6X4QJWAVMKD
InChI = 1S/C10H13N5O3/c11-8-7-9(14-10(17)13-8)15(4-12-7)6-2-1-5(3-16)18-6/h4-6,16H,1-3H2,(H3,11,13,14,17)/t5-,6+/m0/s1
InChI = 1S/C10H13N5O3/c11-8-7-9(14-10(17)13-8)15(4-12-7)6-2-1-5(3-16)18-6/h4-6,16H,1-3H2,(H3,11,13,14,17)/t5-,6+/m0/s1
KITPKMKMNZXFDK-NTSWFWBYSA-N
Same InChIKey

FIGURE 7 | Tautomers with different newly enhanced Morgan algorithm (NEMA) keys but the same InChIKeys. (Reprinted with permission from Symyx Technologies, unpublished work)

Tools such as OpenBabel are also freely available for converting between file formats. Digital Chemistry's MOLSMART can convert MDL (Symyx) structure and reaction query formats to Daylight's SMARTS and SMIRKS strings, and MDL molfiles to Daylight's SMILES strings. A number of tools intended for viewing and editing molecular structures are also able to read files in a number of formats and write them out in other formats.

SPECIFIC CHALLENGES

Constructing a connection table, and identifier, for many organic compounds is fairly straightforward but certain structures present special problems.[62] Features such as aromaticity and tautomerism[161,163] need to be perceived. The CAS REGISTRY system uses a 'normalized' bond type for all rings with alternating single and double bonds; this includes some systems that are not aromatic and omits some that are. The process of normalization has already been mentioned: a structure can be simplified down to its 'core connectivity.' Structure conventions, sometimes called 'business rules' are applied to handle the different representations of substructures such as nitro groups. Whether tautomers are ultimately recorded as the same or different compounds[163] will depend on the application in question; for example, the representation of a chemical substance in a corporate database might need to be independent of specific tautomeric form, whereas spectral properties often require the distinction between specific forms.

Many different systems are in use for handling stereochemistry.[164] Symyx numbers atoms around a tetrahedral carbon atom with 1, 2, 3, and 4, in order of increasing connection table atom number and views the stereocenter so that the bond with atom 4 projects behind the plane formed by atoms 1, 2, and 3. If the numbers increase clockwise, the parity value is 1; if they increase counterclockwise, the parity is 2. The parity value is stored at the node for the stereocenter atom. A parity of zero is used for no stereochemistry and a parity of 3 means unknown stereochemistry. A bond type code is used to indicate double bond stereochemistry.

Isomeric SMILES, which covers stereochemistry and isotopes, has further increased the utility of

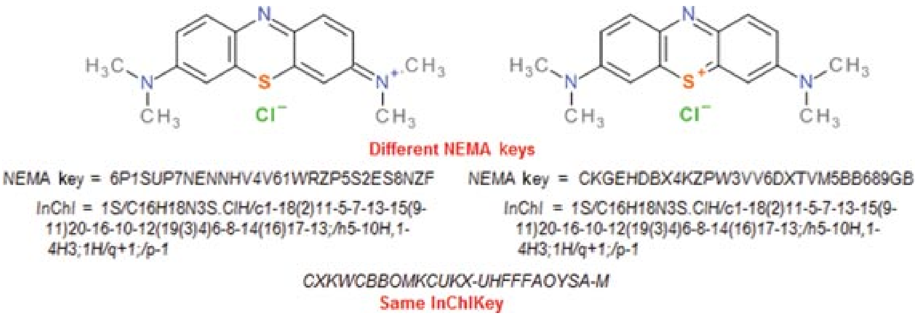

N
H3C
N
CH3
S
Cl⁻
N+
CH3
CH3
N
H3C
N
CH3
S+
Cl⁻
N
CH3
CH3
Different NEMA keys
NEMA key = 6P1SUP7NENNHV4V61WRZP5S2ES8NZF
NEMA key = CKGEHDBX4KZPW3VV6DXTVM5BB689GB
InChI = 1S/C16H18N3S.ClH/c1-18(2)11-5-7-13-15(9-11)20-16-10-12(19(3)4)6-8-14(16)17-13;/h5-10H,1-4H3;1H/q+1;/p-1
InChI = 1S/C16H18N3S.ClH/c1-18(2)11-5-7-13-15(9-11)20-16-10-12(19(3)4)6-8-14(16)17-13;/h5-10H,1-4H3;1H/q+1;/p-1
CXKWCBBOMKCUKX-UHFFFAOYSA-M
Same InChIKey

FIGURE 8 | Mesomers with the same InChIKey but different newly enhanced Morgan algorithm (NEMA) keys. (Reprinted with permission from Symyx Technologies, unpublished work)

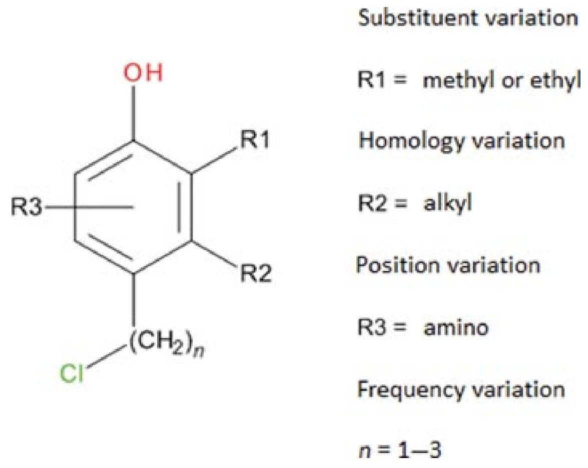


FIGURE 9 | Types of variation in Markush structures. (Reprinted with permission from Digital Chemistry, unpublished work)

canonical SMILES but it should be noted that that OpenEye Scientific Software's canonical SMILES, Daylight's canonical SMILES, ChemAxon's canonical SMILES, and so on, are all independent unique descriptors. None of them can be used as interchangeable indices in cheminformatics.

Multicenter bonds (as in ferrocene), coordination compounds, inorganic compounds, macromolecules and polymers, and incompletely defined substances all present special problems. Some systems use 'shortcuts' or 'superatoms' for subunits (e.g., amino acids) of macromolecules to reduce the complexity of the representation but, in principle, all the atoms could be represented in the traditional manner. There are two common approaches to polymer structures: monomer representation, in which the original monomers are stored and additional information is given textually, and structural repeating unit representation, which stores the repeating units as shortcuts, with details of their length, and so on.[165–172]

Generic Structures

Generic structures (also known as Markush structures) are important in chemical patents in which the inventor claims a whole class of related compounds. They can also be used to describe combinatorial libraries[173–175] (combinatorial chemistry allows very large numbers of chemical entities to be synthesized by condensing a small number of reagents together in all possible combinations. A 'chemical library' is a set of mixtures or discrete compounds made by one combinatorial reaction). A number of variants are possible in patents (see Figure 9), although not all of them are common in combinatorial libraries.

Early systems for storing and retrieving generic structures used fragmentation code systems but these were later supplemented (and to some extent replaced) by topological systems. In a compact representation for a typical set of molecules, the common parts are shown only once. The representation can be considered as a formal 'grammar' for generating valid molecules, but enumeration of the coverage of a patent is usually impractical. In some cases, it is impossible; some patents represent an infinite number of structures. Thus, suitable algorithms take advantage of a Markush representation and avoid enumeration; they compare finite grammars rather than infinite sets of valid sentences.

Sheffield University ran an extended research project on Markush structure storage and retrieval[176–193] from 1979 to 1994. This influenced the development of commercial systems, although independent work was also done at CAS, Derwent (now Thomson Reuters), and Questel.[96,194–196] At Sheffield, two storage formats were designed: an external generic structure language, GENSAL,[177] and an internal extended connection table representation.[179] The system involved a formalized version of the language used in patent specifications, in a design analogous to a programming language. The GENSAL Interpreter program[181] generated the internal representation based on partial connection tables with links between them.

Reduced graphs were also applied at Sheffield in generic chemical structure retrieval.[183,188] Graph reduction involves the generalization of certain features of chemical structures, resulting in a simpler graph. For example, every ring might be replaced by a node 'R.' Reduced graphs can be searched more rapidly (using a query whose graph is also reduced) because the number of nodes is smaller. Reduced graphs have since been successfully applied in similarity searching and other drug discovery applications.[197–200]

The Sheffield generic structures system was never implemented commercially but some of its concepts were incorporated into two commercial systems: Markush DARC[201–206] and MARPAT.[203–208] Markush DARC was developed jointly by Questel (the online host and software developer), Derwent Information (now Thomson Reuters, producer of the World Patent Index Markush database), and the French Patent Office, which offered the PHARMSEARCH database. An integrated database, the Merged Markush File is now available. MARPAT is a software and database combination from CAS, available online on STN International. It is integrated with the CAS REGISTRY database of specific compounds. Commercial chemical information management systems such as that from Symyx are capable of handling 'R-group queries' but are not true Markush structure search systems. Digital Chemistry and ChemAxon,[209] and more recently InfoChem, have expertise in handling Markush structures. The possibility of in-house systems for patent searching[210] is now being discussed.

CHEMICAL REACTIONS

Representing chemical reactions presents a much greater degree of complexity than searching structures alone. Questions that need to be answered in a reaction search include: 'Which reactions convert compound A to compound C'? 'What happens when compound A reacts with compound B'? 'How do I make compound C'? More complex still is a query representing the substructural transformation in which only the reacting substructures in the reactant and product are specified.[211,212] In addition, there might be questions about the reaction conditions, and about other functionality in the reactant that might be affected by the reaction conditions. Even defining the 'novelty' of a reaction is not straightforward: how much do the reagents, conditions, and yield need to differ before the reaction becomes 'different'?

The reaction center of a reaction is the collection of atoms and bonds that are changed during the reaction. Identifying the reaction center is a fundamental feature of a reaction storage and retrieval system. Early reaction retrieval methods using fragments (*vide supra*) and WLN[39,40,211,213] are now of only historical interest; reaction searching is based on connection tables (the Symyx RDfile was mentioned earlier). It is possible to store a structure (as a connection table) and label it as a reactant and to store another structure and label it as a product, but searching for a reaction that converts a ketone into an alcohol is not the same as searching for reactions in which there is a keto group in the starting material and an alcohol in the product. The latter will produce unwanted hits where there is a keto group in the starting material but it is unchanged by the reaction. Atom-to-atom mapping[214–216] ensures that the keto and alcohol groups are both in the reaction site (see Figure 10): atoms on each side of the reaction can be numbered to show which corresponds to which and similar mappings can be used in the reaction query. For large databases, the mapping is usually done automatically when the reaction is registered in a database, but mapping algorithms will be inaccurate in some cases. It is also important to remember that the mapping takes no account of the true reaction mechanism.

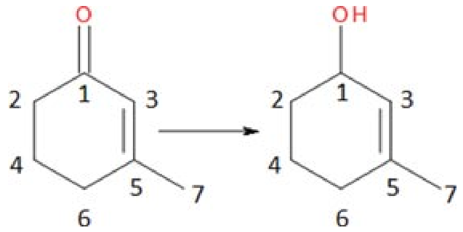


FIGURE 10 | Atom-to-atom mapping.

A method for detecting structural similarities in pairs of reactant and product molecules known as the maximum common subgraph (MCS) approach[217] presents formidable computational challenges, but an approximate MCS method and a method using approximate reaction sites as input to an exact MCS routine were developed at Sheffield University.[218,219] This work led to a number of commercial reaction database systems,[211] some of which are still in use: REACCS (now part of Isentris and ISIS from Symyx),[212] CASREACT from CAS,[220,221] and Beilstein CrossFire (*vide supra*) which has been updated as Reaxys by Elsevier. Work is currently in progress on a 'reaction InChI' or RInChI. The aim of the RInChI project is to create a unique data string to describe a reaction, using the InChI software.

Another reaction identifier has been used for some years in reaction classification, to increase the efficiency of reaction information retrieval (it can be used, e.g., to cluster similar reactions if the number of reactions retrieved is very large) and to provide a chemically meaningful link between different reaction databases. The algorithm CLASSIFY from InfoChem is widely used for these reasons. It is based on InfoChem's mapping algorithm. Hashcodes are calculated for all reaction centers, taking into account atom properties. The sum of all reaction center hashcodes of all reactants and one product of a reaction provides the unique reaction classification code, the ClassCode. Atoms in the immediate environment of the reaction center (spheres) may be included for a broad, medium, or narrow search: only reaction centers will give a large-sized cluster or hit list; reaction centers plus alpha atoms, excluding hydrogens, will give a medium-sized cluster or hit list; and reaction centers plus beta atoms, excluding consecutive sp^3 atoms, will give a small-sized cluster or hit list (see Figure 11). Three hash-coded numbers are thus assigned to each reaction. CLASSIFY has also been used to produce subsets[222] of the SPRESIweb database.[223]

Retrosynthesis

Thus far, reaction retrieval systems have been discussed but there are also synthetic analysis programs,

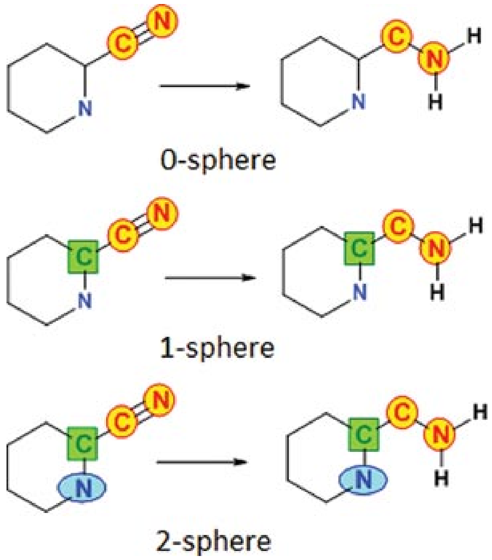


FIGURE 11 | InfoChem spheres around reaction centers.

which in turn can be divided into synthesis design (also called synthesis planning) programs,[224–229] reaction prediction programs,[226,228,230–235] and mechanism elucidation programs.[236] Retrosynthetic analysis (synthesis design) programs take an input molecule and step by step deduce possible precursors, using specialized techniques to address the challenge of combinatorial explosion. Logic and Heuristics Applied to Synthetic Analysis (LHASA)[224,229] is the oldest of these synthetic planning programs. It is an expert system, relying on a knowledge base of reactions that is manually constructed from reaction transforms coded in a language developed specifically for LHASA.

Another program, Workbench for the Organization of Data for Chemical Applications (WODCA),[227,228] performs retrosynthesis in a logic-oriented fashion, looking for promising available starting materials by substructure and similarity searches in catalogs of chemical suppliers, and searching for strategic bonds to break in the target molecule by means of calculated physicochemical effects. WODCA has been superseded by THERESA that is sold by Molecular Networks. Two other programs use automated methods to generate the transforms needed in a retrosynthesis system, and to address the problem of combinatorial explosion: ARChem Route Designer[237] from SimBioSys (Toronto, Canada) and InfoChem's IC*SYNTH*.

3D STRUCTURES

The 2D representations discussed so far may be the natural language of organic chemistry but in reality molecules are 3D: the atoms can be positioned in space in multiple conformations, more than one of which may be a low energy form. Approaches such as quantum mechanics that more accurately reflect a molecule's properties are too complex for large numbers of structures, so other methods had to be developed for representing conformers efficiently. Gasteiger and coworkers[238–240] at Erlangen have worked for many years on representing 3D space by means of physicochemical properties, in order to predict reactivity (*vide supra*), spectra, and biological properties.

In the 1990s, development of 3D structure methods was spurred on by programs for the generation of 3D structures from 2D structures.[241] The two most widely used structure generation programs are CONCORD[242–244] and CORINA.[214,244–246] Such programs were particularly important in the early 1990s because at that time only a limited number of experimentally determined 3D structures were available in databases. They are still used frequently for deterministic 3D structures and in ligand preparation prior to searching in pharmacophore or docking algorithms.

Experimental 3D Databases

The history of crystallographic databases goes back to the early 1970s but it has taken many years for them to grow. The Cambridge Structural Database[247–250] is the world repository of small molecule crystal structures. By December 2009, it contained more than 500,000 structures. The Protein Data Bank (PDB) began as a grassroots effort in 1971. It has grown from a small archive containing a dozen structures to a major international resource for structural biology containing more than 40,000 entries.[251–253] It contains information about experimentally determined structures of proteins, nucleic acids, and complex assemblies.

The Crystallographic Information File (CIF)[254,255] was adopted by the International Union of Crystallography for the archiving and electronic transmission of crystallographic data. The macromolecular CIF (mmCIF)[256,257] is an extension of CIF, replacing the historical PDB file format. The Molecular Information File was developed from SMD and is compatible with CIF.[258]

3D Structure Representation

3D searches of databases such as the CSD, or of in-house compound databases in the pharmaceutical industry aim to identify conformations that match the query. The most common 3D query is a pharmacophore. According to IUPAC,[259] 'a pharmacophore is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response. A pharmacophore does not represent a real molecule or a real association of functional groups but a purely abstract concept that accounts for the common molecular interaction capacities of a group of compounds toward their target structure. The pharmacophore can be considered as the largest common denominator shared by a set of active molecules.'

In the case of 3D database, searching the pharmacophore is defined as a set of features, such as hydrogen bond donors and acceptors, positively and negatively charged groups, and hydrophobic regions and aromatic rings, together with their relative spatial orientation. The spatial relationships can be specified as distances or distance ranges or by defining the locations (coordinates) of the features together with some distance tolerance. 3D substructure search is carried out by a procedure analogous to 2D substructure search, but in this case the query can be defined as a group of atoms, with specified interatomic distances, and both query and database structures are topological graphs in which the nodes are atoms, but the edges are interatomic distances.[260,261] In early programs, only one conformer was considered[262–267] but in the extension from 'rigid' to 'flexible' 3D searching, multiple conformers were included. The exploration of multiple conformations can be tackled by generating and storing multiple representative conformations or by exploring conformational space 'on the fly.'[268–276]

More recently, there has been increased use of field- or shape-based approaches. The shape comparison program Rapid Overlay of Chemical Structures[277] is used to perceive similarity between molecules based on their 3D shape. The objective of this method is to find and quantify the maximal overlap of the volume of two molecules. Matches are based only on volume overlap of optimally aligned molecules; therefore, they are virtually independent of the atom types and bonding patterns present in the query and search molecules. The goal of the approach is to identify molecules that can adopt shapes very similar to the query and in doing so increase the chance of 'scaffold hopping' or 'lead hopping.'

Cresset BioMolecular Discovery's (Welwyn Garden City, Herts, United Kingdom) approach[278] uses molecular fields that provide a way of analyzing the surface properties of molecules that in turn allows an understanding of how the atomic structure of a compound can be translated into biologically relevant binding properties. The field point pattern is a sophisticated 'pharmacophore' that can be used to define a template for binding. Molecules can be overlaid using their fields, rather than structure, and the field similarity between two molecules can be quantified and converted to a similarity value. The expectation is that modeling in 'field' rather than 'structural' space will facilitate more innovative discoveries.

CONCLUSION

All the basic work on structure representation was complete by the 1990s; research since then has largely been into applications. There have, however, been some significant developments over the past 20 years. Efforts to standardize connection table formats were not successful (partly because the molfile format had become a *de facto* standard and CAS had another proprietary standard), but the development of InChI has renewed interest in a nonproprietary standard. InChI is being actively developed by the InChI Trust, but it does not yet cover as many 'unusual' structures as some other systems do.

Contemporaneously, there has been much research into text mining,[8–13,279] much of it as yet unpublished or available only in conference proceedings. Text mining recognizes chemical entities in journal articles or patents and extracts them for conversion to connection tables (and InChIs in some cases); related work converts images of chemical structures into connection tables.[280–285] Research continues into the special problems of structure representation: more unusual forms of stereochemistry, polymers, and generic structures, for example. There has also been renewed interest in retrosynthesis of late.

The successful machine representation of chemical structures has had an enormous impact on progress in other fields, as is evidenced by the large number of Wiley Interdisciplinary Reviews articles related to this one. Using cheminformatics, chemical and pharmaceutical companies have been able to prevent duplication of synthetic effort and have been better able to handle patent information. In particular, the pharmaceutical industry has benefited from systems that enable it to keep track of its inventories and proprietary compounds, and from software that makes drug discovery significantly faster and more efficient.

REFERENCES

1. Panico R, Powell WH, Richter J-C. *A Guide to IUPAC Nomenclature of Organic Compounds Recommendations 1993*. Oxford: Blackwell Science; 1993.
2. Favre HA, Hellwich K-H, Moss GP, Powell WH, Traynham JG. Corrections to a guide to IUPAC nomenclature of organic compounds (IUPAC recommendations 1993). *Pure Appl Chem* 1999, 71:1327–1330.
3. Leigh GJ, Favre HA, Metanomski WV. *Principles of Organic Nomenclature*. Oxford: Blackwell Science; 1998.
4. Anonymous. *Naming and Indexing of Chemical Substances for Chemical Abstracts*. Columbus, OH: Chemical Abstracts Service; 2007.
5. Wisniewski JL. Nomenclature: automatic generation and conversion. In: Gasteiger J, ed. *Handbook of Cheminformatics: from Data to Knowledge*. Vol 1. Weinheim: Wiley-VCH; 2003, 51–79.
6. Cooke-Fox DI, Kirby GH, Lord MR, Rayner JD. Computer translation of IUPAC systematic organic chemical nomenclature. 4. Concise connection tables to structure diagrams. *J Chem Inf Comput Sci* 1990, 30:122–127.
7. Brecher J. Name = Struct: a practical approach to the sorry state of real-life chemical nomenclature. *J Chem Inf Comput Sci* 1999, 39:943–950.
8. Banville DL. Mining chemical structural information from the drug literature. *Drug Discov Today* 2006, 11:35–42.
9. Corbett P, Murray-Rust P. High-throughput identification of chemistry in life science texts. In: Berthold MR, Glen RC, Fischer I, eds. *Computational Life Sciences II*. Berlin, Heidelberg: Springer; 2006, 107–118.
10. Kolarik C, Hofmann-Apitius M, Zimmermann M, Fluck J. Identification of new drug classification terms in textual resources. *Bioinformatics* 2007, 23:i264–i272.
11. Klinger R, Kolarik C, Fluck J, Hofmann-Apitius M, Friedrich CM. Detection of IUPAC and IUPAC-like chemical names. *Bioinformatics* 2008, 24:i268–i276.
12. Banville DL. Mining chemical and biological information from the drug literature. *Curr Opin Drug Discov Devel* 2009, 12:376–387.
13. Banville DL, ed. *Chemical Information Mining: Facilitating Literature-Based Discovery*. Boca Raton, FL: CRC Press; 2009.
14. Sayle R. Foreign language translation of chemical nomenclature by computer. *J Chem Inf Model* 2009, 49:519–530.
15. Wisniewski JL. AUTONOM: system for computer translation of structural diagrams into IUPAC-compatible names. 1. General design. *J Chem Inf Comput Sci* 1990, 30:324–332.
16. Goebels L, Lawson AJ, Wisniewski JL. AUTONOM: system for computer translation of structural diagrams into IUPAC-compatible names. 2. Nomenclature of chains and rings. *J Chem Inf Comput Sci* 1991, 31:216–225.
17. Williams A, Yerin A. The need for systematic naming software tools for exchange of chemical information. *Molecules* 1999, 4:255–263.
18. Engel T. Chemical nomenclature by mouse click. *Nachr Chem* 2005, 53:428–431.
19. Eller GA. Improving the quality of published chemical names with nomenclature software. *Molecules* 2006, 11:915–928.
20. Dyson GM, Lynch MF, Morgan HL. A modified IUPAC-Dyson notation system for chemical structures. *Inform Storage Retrieval* 1968, 4:27–83.
21. Dyson GM. Dyson-IUPAC notation. In: Ash JE, Hyde E, eds. *Communication: Storage and Retrieval of Chemical Information*. Chichester: Ellis Horwood; 1975, 130–155.
22. Skolnik H, Clow A. A notation system for indexing pesticides. *J Chem Doc* 1964, 4:221–227.
23. Smith EG. *The Wiswesser Line-Formula Chemical Notation*. New York, NY: McGraw-Hill; 1968.
24. Smith EG, Barker PA. *The Wiswesser Line-Formula Chemical Notation (WLN)*. 3rd ed. Cherry Hill, NJ: Chemical Information Management; 1976.
25. Palmer G. Wiswesser Line-Formula Notation. *Chem Brit* 1970, 6:422–426.
26. Gibson GW, Granito CE. Wiswesser chemical line-notation. *Am Lab* 1972, 4:27–34, 36–37.
27. Baker PA, Palmer G, Nichols PWL. Wiswesser line-formula notation. In: Ash JE, Hyde E, eds. *Communication: Storage and Retrieval of Chemical Information*. Chichester: Ellis Horwood; 1975, 97–129.
28. Vollmer JJ. Wiswesser line notation: an introduction. *J Chem Educ* 1983, 60:192–196.
29. Hyde E, Matthews FW, Thomson LH, Wiswesser WJ. Conversion of Wiswesser notation to a connectivity matrix for organic compounds. *J Chem Doc* 1967, 7:200–204.
30. Thomson LH, Hyde E, Matthews FW. Organic search and display using a connectivity matrix derived from Wiswesser notation. *J Chem Doc* 1967, 7:204–209.
31. Garfield E, Revesz GS, Granito CE, Dorr HA, Calderon MM, Warner A. Index Chemicus Registry System: pragmatic approach to substructure chemical retrieval. *J Chem Doc* 1970, 10:54–58.

32. Sasamoto M. Qualitative comparison of Wiswesser line notation with Ringdoc. *J Chem Doc* 1973, 13:206.
33. Eakin DR, Hyde E, Parker G. Use of computers with chemical structural information. ICI CROSSBOW system. *Pestic Sci* 1974, 5:319–326.
34. Osinga M, Verrijn SAA. Documentation of chemical reactions. II. Analysis of the Wiswesser Line Notation. *J Chem Doc* 1974, 14:194–198.
35. Sheng A, Lupi L, Ronayne M, Sprules A, Zornetzer S. Hoffmann-La Roche's on-line/batch interactive chemical information system. *J Chem Doc* 1974, 14:179–185.
36. Eakin DR. ICI CROSSBOW system. In: Ash JE, Hyde E, eds. *Communication: Storage and Retrieval of Chemical Information*. Chichester: Ellis Horwood; 1975, 227–242.
37. Garfield E, Sim M. The Index Chemicus Registry System—past, present and future. *Pure Appl Chem* 1977, 49:1803–1805.
38. Lynch MF, Nunn PR, Radcliffe J. Production of printed indexes of chemical reactions using Wiswesser line notations. *J Chem Inf Comput Sci* 1978, 18:94–96.
39. Lynch MF, Willett P. The production of machine-readable descriptions of chemical reactions using Wiswesser Line Notations. *J Chem Inf Comput Sci* 1978, 18:149–154.
40. Lynch MF, Willett P. The automatic detection of chemical reaction sites. *J Chem Inf Comput Sci* 1978, 18:154–159.
41. Bond VB, Bowman CM, Davison LC, Roush PF, Young LF. Applications of the Wiswesser line notation at the Dow Chemical Company. *J Chem Inf Comput Sci* 1982, 22:103–105.
42. Eakin DR. Graphics challenge WLN. Can WLN hold fast? *J Chem Inf Comput Sci* 1982, 22:101–103.
43. Fritts LE, Schwind MM. Using the Wiswesser line notation (WLN) for online, interactive searching of chemical structures. *J Chem Inf Comput Sci* 1982, 22:106–109.
44. Johns TM, Clare M. Wiswesser line notation as a structural summary medium. *J Chem Inf Comput Sci* 1982, 22:109–113.
45. Rosenberg MD, DeBardeleben MZ, DeBardeleben JF. Chemical supply catalog indexing: now and the future. An ideal place for use of the Wiswesser line notation. *J Chem Inf Comput Sci* 1982, 22:93–98.
46. Warr WA. Diverse uses and future prospects for Wiswesser line-formula notation. *J Chem Inf Comput Sci* 1982, 22:98–101.
47. Walker SB. Development of CAOCI and its use in ICI plant protection division. *J Chem Inf Comput Sci* 1983, 23:3–5.
48. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988, 28:31–36.
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. Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL Line Notation (SLN): a versatile language for chemical structure representation. *J Chem Inf Comput Sci* 1997, 37:71–79.
51. Homer RW, Swanson J, Jilek RJ, Hurst T, Clark RD. SYBYL line notation (SLN): a single notation to represent chemical structures, queries, reactions, and virtual libraries. *J Chem Inf Model* 2008, 48:2294–2307.
52. Barnard JM, Jochum CJ, Welford SM. A universal structure/substructure representation for PC-host communication. In: Warr WA, ed. *Chemical Structure Information Systems*. Vol 400. Washington DC: ACS Symposium Series; 1989, 76–81.
53. Rohbeck HG. Representation of structure description arranged linearly. In: Gmehling J, ed. *Software Development in Chemistry 5, Proceedings of the 5th Workshop 'Computational Chemistry.'* Berlin, Heidelberg: Springer; 1991, 49–58.
54. Hartwell IO, Haglund KA. An overview of DIALOG. In: Heller SR, ed. *The Beilstein Online Database*. Vol 436. Washington DC: ACS Symposium Series; 1990, 42–63.
55. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at Chemical Abstracts Service. *J Chem Doc* 1965, 5:107–113.
56. Brecher J. Graphical representation of stereochemical configuration. *Pure Appl Chem* 2006, 78:1897–1970.
57. Brecher J. Graphical representation standards for chemical structure diagrams (IUPAC recommendations 2008). *Pure Appl Chem* 2008, 80:277–410.
58. Clark AM, Labute P, Santavy M. 2D structure depiction. *J Chem Inf Model* 2006, 46:1107–1123.
59. Barnard JM. Substructure searching methods: old and new. *J Chem Inf Comput Sci* 1993, 33:532–538.
60. Barnard JM. Structure representation and search. In: Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR, eds. *Encyclopedia of Computational Chemistry*. Vol 4. Chichester: John Wiley & Sons; 1998, 2818–2826.
61. Dietz A. Yet another representation of molecular structure. *J Chem Inf Comput Sci* 1995, 35:787–802.
62. Bauerschmidt S, Gasteiger J. Overcoming the limitations of a connection table description: a universal representation of chemical species. *J Chem Inf Comput Sci* 1997, 37:705–714.

63. Wipke WT, Dyott TM. Stereochemically unique naming algorithm. *J Am Chem Soc* 1974, 96:4834–4842.

64. Hillard R, Taylor KT. InChI keys as standard global identifiers in chemistry web services (abstract #CINF-015). Salt Lake City, UT: 237th ACS National Meeting; March 22–26, 2009.

65. Wipke WT, Krishnan S, Ouchi GI. Hash functions for rapid storage and retrieval of chemical structures. *J Chem Inf Comput Sci* 1978, 18:32–37.

66. Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J Chem Inf Comput Sci* 1992, 32:244–255.

67. Bebak H, Buse C, Donner WT, Hoever P, Jacob H, Klaus H, Pesch J, Roemelt J, Schilling P, Woost B, et al. The standard molecular data format (SMD format) as an integration tool in computer chemistry. *J Chem Inf Comput Sci* 1989, 29:1–5.

68. Barnard JM. Draft specification for revised version of the Standard Molecular Data (SMD) format. *J Chem Inf Comput Sci* 1990, 30:81–96.

69. Barnard JM. The standard molecular data (SMD) format. In: Warr WA, ed. *Chemical Structures* 2. Berlin, Heidelberg: Springer; 1993, 185–193.

70. Murray-Rust P, Rzepa HS. Chemical Markup, XML, and the Worldwide Web. 1. Basic principles. *J Chem Inf Comput Sci* 1999, 39:928–942.

71. Murray-Rust P, Rzepa HS. Chemical Markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *J Chem Inf Comput Sci* 2001, 41:1113–1123.

72. Gkoutos GV, Murray-Rust P, Rzepa HS, Wright M. Chemical Markup, XML, and the World-Wide Web. 3. Toward a signed semantic chemical web of trust. *J Chem Inf Comput Sci* 2001, 41:1124–1130.

73. Murray-Rust P, Rzepa HS. Chemical Markup, XML, and the World Wide Web. 4. CML schema. *J Chem Inf Comput Sci* 2003, 43:757–772.

74. Murray-Rust P, Rzepa HS, Williamson MJ, Willighagen EL. Chemical Markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. *J Chem Inf Comput Sci* 2004, 44:462–469.

75. Holliday GL, Murray-Rust P, Rzepa HS. Chemical Markup, XML, and the World Wide Web. 6. CMLReact, an XML vocabulary for chemical reactions. *J Chem Inf Model* 2006, 46:145–157.

76. Kuhn S, Helmus T, Lancashire RJ, Murray-Rust P, Rzepa HS, Steinbeck C, Willighagen EL. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML vocabulary for spectral data. *J Chem Inf Model* 2007, 47:2015–2034.

77. Adams N, Winter J, Murray-Rust P, Rzepa HS. Chemical markup, XML and the World-Wide Web. 8. Polymer Markup Language. *J Chem Inf Model* 2008, 48:2118–2128.

78. Murray-Rust P, Rzepa HS, Wright M, Zara S. A universal approach to web-based chemistry using XML and CML. *Chem Commun* 2000, 1471–1472.

79. Murray-Rust P, Rzepa HS, Wright M. Development of chemical markup language (CML) as a system for handling complex chemical content. *New J Chem* 2001, 25:618–634.

80. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL. The Blue Obelisk-interoperability in chemical informatics. *J Chem Inf Model* 2006, 46:991–998.

81. Brown RD, Downs GM, Willett P, Cook APF. Hyperstructure model for chemical structure handling: generation and atom-by-atom searching of hyperstructures. *J Chem Inf Comput Sci* 1992, 32:522–531.

82. Brown RD, Downs GM, Jones G, Willett P. Hyperstructure model for chemical structure handling: techniques for substructure searching. *J Chem Inf Comput Sci* 1994, 34:47–53.

83. Downs GM, Gill GS, Willett P, Walsh P. Automated descriptor selection and hyperstructure generation to assist SAR studies. *SAR QSAR Environ Res* 1995, 3:253–264.

84. Bawden D, Devon TK, Jackson FT, Wood SI, Lynch MF, Willett P. A qualitative comparison of Wiswesser line notation descriptors of reactions and the Derwent Chemical Reaction Documentation Service. *J Chem Inf Comput Sci* 1979, 19:90–93.

85. Suhr C, Von HE, Dethlefsen W. Derwent's CPI and IDC's GREMAS: remarks on their relative retrieval power with regard to Markush structures. In Barnard JM, ed. *Computer Handling of Generic Chemical Structures*. Aldershot: Gower; 1984, 96–105.

86. Zass E. A user's view of chemical reaction information sources. *J Chem Inf Comput Sci* 1990, 30:360–372.

87. Bador P, Surrel MN. Computer systems for searching of chemical reaction databases and systems for computer-aided design of organic synthesis. *New J Chem* 1992, 16:413–423.

88. Franzreb KH, Hornbach P, Pahde C, Ploss G, Sander J. Structure searches in patent literature: a comparison study between IDC GREMAS and Derwent Chemical Code. *J Chem Inf Comput Sci* 1991, 31:284–289.

89. Meyer E. Topological brief description of chemical structure formulas for the documentation with computers. *Inf Storage Retrieval* 1965, 2:205–215.

90. Roessler S, Kolb A. The GREMAS system, an intergral part of the IDC system for chemical documentation. *J Chem Doc* 1970, 10:128–134.

91. Fugmann R, Bitterlich W. Reaction documentation using the GREMAS system. *Chem-Ztg* 1972, 96:323–330.

92. Fugmann R, Kusemann G, Winter JH. The supply of information on chemical reactions in the IDC system. *Inf Process Manage* 1979, 15:303–323.

93. Mullen A. Means of information in chemistry. Part II. New trends. *Prax Naturwiss Chem* 1980, 29:243–247.

94. Jordis U, Oberhauser O. Status of computer search of the chemical literature: (partial) structural research with GREMAS, DARC, and CAS ONLINE. *Oesterr Chem Z* 1982, 83:311–314.

95. Fricke C, Nickelsen I, Fugmann R, Sander J. GREDIA: a new access to GREMAS databases. *Tetrahedron Comput Methodol* 1989, 2:167–175.

96. Barnard JM. A comparison of different approaches to Markush structure handling. *J Chem Inf Comput Sci* 1991, 31:64–68.

97. Stiegler G, Maier B, Lenz H. Automatic translation of GENSAL representations of Markush structures into GREMAS fragment codes at IDC. In: Warr WA, ed. *Chemical Structures* 2. Berlin, Heidelberg: Springer; 1993, 105–114.

98. Christie BD, Leland BA, Nourse JG. Structure searching in chemical databases by direct lookup methods. *J Chem Inf Comput Sci* 1993, 33:545–547.

99. Lynch MF, Orton J, Town WG. Organization of large collections of chemical structures for computer searching. *J Chem Soc C* 1969, 1732–1736.

100. Adamson GW, Cowell J, Lynch MF, McLure AHW, Town WG, Yapp AM. Strategic considerations in the design of a screening system for substructure of chemical structure files. *J Chem Doc* 1973, 13:153–157.

101. Adamson GW, Clinch VA, Creasey SE, Lynch MF. Distributions of fragment representations in a chemical substructure search screening system. *J Chem Doc* 1974, 14:72–74.

102. Hodes L. Selection of descriptors according to discrimination and redundancy. Application to chemical structure searching. *J Chem Inf Comput Sci* 1976, 16:88–93.

103. Willett P. The effect of screen set size on retrieval from chemical substructure search systems. *J Chem Inf Comput Sci* 1979, 19:253–255.

104. Willett P. A screen set generation algorithm. *J Chem Inf Comput Sci* 1979, 19:159–162.

105. Graf W, Kaindl HK, Kniess H, Warszawski R. The third BASIC fragment search dictionary. *J Chem Inf Comput Sci* 1982, 22:177–181.

106. Dittmar PG, Farmer NA, Fisanick W, Haines RC, Mockus J. The CAS ONLINE search system. 1. General system design and selection, generation, and use of search screens. *J Chem Inf Comput Sci* 1983, 23:93–102.

107. McGregor MJ, Pallai PV. Clustering of large databases of compounds: using the MDL "keys" as structural descriptors. *J Chem Inf Comput Sci* 1997, 37:443–448.

108. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002, 42:1273–1280.

109. Olah M, Bologa C, Oprea TI. An automated PLS search for biologically relevant QSAR descriptors. *J Comput Aided Mol Des* 2004, 18:437–449.

110. Henry DR, Durant JL Jr. Optimization of MDL substructure search keys for the prediction of activity and toxicity. In: Lavine BK, ed. *Chemometrics and Chemoinformatics*. Vol 894. Washington DC: ACS Symposium Series; 2005, 145–156.

111. Joergensen AMM, Langgaard M, Gundertofte K, Pedersen JT. A fragment-weighted key-based similarity measure for use in structural clustering and virtual screening. *QSAR Comb Sci* 2006, 25:221–234.

112. Renner S, Schneider G. Scaffold-hopping potential of ligand-based similarity concepts. *ChemMedChem* 2006, 1:181–185.

113. Bayada DM, Hamersma H, van Geerestein VJ. Molecular diversity and representativity in chemical databases. *J Chem Inf Comput Sci* 1999, 39:1–10.

114. Wild DJ, Blankley CJ. Comparison of 2D fingerprint types and hierarchy level selection methods for structural grouping using Ward's clustering. *J Chem Inf Comput Sci* 2000, 40:155–162.

115. Barnard JM, Downs GM. Chemical fragment generation and clustering software. *J Chem Inf Comput Sci* 1997, 37:141–142.

116. Warr WA. Commercial software systems for diversity analysis. *Perspect Drug Discovery Des* 1997, 7/8:115–130.

117. Wagener M, van Geerestein VJ. Potential drugs and nondrugs: prediction and identification of important structural features. *J Chem Inf Comput Sci* 2000, 40:280–292.

118. Brown RD, Martin YC. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J Chem Inf Comput Sci* 1996, 36:572–584.

119. Brown RD, Martin YC. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J Chem Inf Comput Sci* 1997, 37:1–9.

120. Willett P, Barnard JM, Downs GM. Chemical similarity searching. *J Chem Inf Comput Sci* 1998, 38:983–996.

121. Willett P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov Today* 2006, 11:1046–1053.

122. Willett P. Similarity methods in chemoinformatics. *Ann Rev Inf Sci Technol* 2009, 43:3–71.

123. Attias R, Dubois JE. Substructure systems: concepts and classifications. *J Chem Inf Comput Sci* 1990, 30:2–7.

124. Dubois J-E. Chemical complexity and molecular topology. The DARC concepts and applications. *l'Actual Chim* 2008, 320–321:37–42.

125. Holliday JD, Willett P. The influence of the DARC project on chemoinformatics research at the University of Sheffield. *l'Actual Chim* 2008, 320–321: 45–50.

126. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model* 2010, 50:742–754.

127. Dittmar PG, Stobaugh RE, Watson CE. The Chemical Abstracts Service Chemical Registry System. I. General design. *J Chem Inf Comput Sci* 1976, 16:111–121.

128. Freeland RG, Funk SA, O'Korn LJ, Wilson GA. The Chemical Abstracts Service Chemical Registry System. II. Augmented connectivity molecular formula. *J Chem Inf Comput Sci* 1979, 19:94–98.

129. Blackwood JE, Elliott PM, Stobaugh RE, Watson CE. The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry. *J Chem Inf Comput Sci* 1977, 17:3–8.

130. vanderStouw GG, Gustafson C, Rule JD, Watson CE. The Chemical Abstracts Service Chemical Registry System. IV. Use of the Registry System to support the preparation of index nomenclature. *J Chem Inf Comput Sci* 1976, 164:213–218.

131. Zamora A, Dayton DL. The Chemical Abstracts Service Chemical Registry System. V. Structure input and editing. *J Chem Inf Comput Sci* 1976, 16:219–222.

132. Stobaugh RE. The Chemical Abstracts Service Chemical Registry System. VI. Substance-related statistics. *J Chem Inf Comput Sci* 1980, 20:76–82.

133. Mockus J, Stobaugh RE. The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and alternating bonds. *J Chem Inf Comput Sci* 1980, 20:18–22.

134. Moosemiller JP, Ryan AW, Stobaugh RE. The Chemical Abstracts Service Chemical Registry System. VIII. Manual registration. *J Chem Inf Comput Sci* 1980, 20:83–88.

135. Ryan AW, Stobaugh RE. Chemical Abstracts Service Chemical Registry System. 9. Input structure conventions. *J Chem Inf Comput Sci* 1982, 22:22–28.

136. Hamill KA, Nelson RD, Stouw GGV, Stobaugh RE. Chemical Abstracts Service Chemical Registry System. 10. Registration of substances from pre-1965 indexes of Chemical Abstracts. *J Chem Inf Comput Sci* 1988, 28:175–179.

137. Stobaugh RE. Chemical Abstracts Service Chemical Registry System. 11. Substance-related statistics: update and additions. *J Chem Inf Comput Sci* 1988, 28:180–187.

138. Blackwood JE, Blower PE Jr, Layten SW, Lillie DH, Lipkus AH, Peer JP, Qian C, Staggenborg LM, Watson CE. Chemical Abstracts Service Chemical Registry System. 13. Enhanced handling of stereochemistry. *J Chem Inf Comput Sci* 1991, 31:204–212.

139. Weisgerber DW. Chemical Abstracts Service Chemical Registry System: history, scope, and impacts. *J Am Soc Inf Sci* 1997, 48:349–360.

140. Lipkus AH, Yuan Q, Lucas KA, Funk SA, Bartelt WF, Schenck RJ, Trippe AJ. Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry. *J Org Chem* 2008, 73:4443–4451.

141. Heller SR, ed. *The Beilstein System: Strategies for Effective Searching*. Washington DC: American Chemical Society; 1998.

142. Jochum C. The Beilstein Information System is not a reaction database, or is it? *J Chem Inf Comput Sci* 1994, 34:71–73.

143. Barth A, Westermann U, Pasucha B. Messenger and S4: a comparison of structure search systems. *J Chem Inf Comput Sci* 1994, 34:714–722.

144. Parkar FA, Parkin D. Comparison of Beilstein CrossFirePlusReactions and the selective reaction databases under ISIS. *J Chem Inf Comput Sci* 1999, 39:281–288.

145. Coste J, Gien O, Dietz A, Laurenco C. Use of reaction databases. *l'Actual Chim* 1999, 7:27–32.

146. Meehan P, Schofield H. Crossfire: a structural revolution for chemists. *Online Inf Rev* 2001, 25:241–249.

147. Cooke F, Schofield H. A framework for the evaluation of chemical structure databases. *J Chem Inf Comput Sci* 2001, 41:1131–1140.

148. Heller SR, McNaught AD. The IUPAC international chemical identifier (InChI). *Chem Int* 2009, 31:7–9.

149. McKay BD. Practical graph isomorphism. *Cong Numer* 1981, 30:45–87.

150. Anonymous. *Secure hash signature standard. In: Federal Information Processing Standards Publication* 180-2. Washington DC: NIST; 2002.

151. Murray-Rust P, Rzepa HS, Stewart JJP, Zhang Y. A global resource for computational chemistry. *J Mol Model* 2005, 11:532–541.

152. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y. Enhancement of the chemical semantic web through the use of InChI identifiers. *Org Biomol Chem* 2005, 3:1832–1834.

153. Casher O, Rzepa HS. SemanticEye: a semantic web application to rationalize and enhance chemical electronic publishing. *J Chem Inf Model* 2006, 46:2396–2411.

154. Williams AJ. Internet-based tools for communication and collaboration in chemistry. *Drug Discov Today* 2008, 13:502–506.

155. Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN. Application of InChI to curate, index, and query 3-D structures. *Proteins* 2005, 60:1–4.

156. Williams AJ. A perspective of publicly accessible/open-access chemistry databases. *Drug Discov Today* 2008, 13:495–501.
157. Richard AM, Gold LS, Nicklaus MC. Chemical structure indexing of toxicity data on the Internet: moving toward a flat world. *Curr Opin Drug Discov Devel* 2006, 9:314–325.
158. Ihlenfeldt W-D, Voigt JH, Bienfait B, Oellien F, Nicklaus MC. Enhanced CACTVS browser of the open NCI database. *J Chem Inf Comput Sci* 2002, 42:46–57.
159. Sitzmann M, Filippov IV, Nicklaus MC. Internet resources integrating many small-molecules databases. *SAR QSAR Environ Res* 2008, 19:1–9.
160. Ihlenfeldt WD, Takahashi Y, Abe H, Sasaki S. Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility. *J Chem Inf Comput Sci* 1994, 34:109–116.
161. Oellien F, Cramer J, Beyer C, Ihlenfeldt W-D, Selzer PM. The impact of tautomer forms on pharmacophore-based virtual screening. *J Chem Inf Model* 2006, 46:2342–2354.
162. Ihlenfeldt WD, Gasteiger J. Hash codes for the identification and classification of molecular structure elements. *J Comput Chem* 1994, 15:793–813.
163. Warr WA. Tautomerism in chemical information management systems. *J Comput Aided Mol Des* 2010, 24:497–520.
164. Rohde B. Representation and manipulation of stereochemistry. In: Gasteiger J, ed. *Handbook of Chemoinformatics: From Data to Knowledge*. Vol 1. Weinheim: Wiley-VCH; 2003, 206–230.
165. Gushurst AJ, Nourse JG, Hounshell WD, Leland BA, Raich DG. The substance module: the representation, storage, and searching of complex structures. *J Chem Inf Comput Sci* 1991, 31:447–454.
166. Nourse JG, Hounshell WD, Leland BA, Gushurst AJ, Raich DG. Computer representation and searching of chemical substances. In: Warr WA, ed. *Chemical Structures 2*. Berlin, Heidelberg: Springer; 1993, 221–233.
167. Schultz JL, Wilks ES. A nomenclature and structural representation system for asymmetrical "I"-shaped hyperbranched polymers. *J Chem Inf Comput Sci* 1996, 36:1109–1117.
168. Schultz JL, Wilks ES. Nomenclature and structural representation for linear, single-strand polymers aftertreated to hyperconnected networks. *J Chem Inf Comput Sci* 1996, 36:955–966.
169. Wilks ES. Polymer nomenclature and structure: a comparison of systems used by CAS, IUPAC, MDL, and DuPont. 4. Stereochemistry, inorganic, coordination, double-strand, polysiloxanes, oligomers, telomers. *J Chem Inf Comput Sci* 1997, 37:224–235.
170. Wilks ES. Polymer nomenclature and structure: a comparison of systems used by CAS, IUPAC, MDL, and DuPont. 3. Comb/graft, cross-linked, and dendritic/hyperconnected/star polymers. *J Chem Inf Comput Sci* 1997, 37:209–223.
171. Wilks ES. Polymer nomenclature and structure: a comparison of systems used by CAS, IUPAC, MDL, and DuPont. 2. Aftertreated (post-treated), alternating/periodic, and block polymers. *J Chem Inf Comput Sci* 1997, 37:193–208.
172. Wilks ES. Polymer nomenclature and structure: a comparison of systems used by CAS, IUPAC, MDL, and DuPont. 1. Regular single-strand organic polymers. *J Chem Inf Comput Sci* 1997, 37:171–192.
173. Downs GM, Barnard JM. Techniques for generating descriptive fingerprints in combinatorial libraries. *J Chem Inf Comput Sci* 1997, 37:59–61.
174. Barnard JM, Downs GM. Computer representation and manipulation of combinatorial libraries. *Perspect Drug Discovery Des* 1997, 7/8:13–30.
175. Barnard JM, Downs GM, von Scholley-Pfab A, Brown RD. Use of Markush structure analysis techniques for descriptor generation and clustering of large combinatorial libraries. *J Mol Graphics Modell* 2000, 18:452–463.
176. Lynch MF, Barnard JM, Welford SM. Computer storage and retrieval of generic chemical structures in patents. 1. Introduction and general strategy. *J Chem Inf Comput Sci* 1981, 21:148–150.
177. Barnard JM, Lynch MF, Welford SM. Computer storage and retrieval of generic chemical structures in patents. 2. GENSAL, a formal language for the description of generic chemical structures. *J Chem Inf Comput Sci* 1981, 21:151–161.
178. Welford SM, Lynch MF, Barnard JM. Computer storage and retrieval of generic chemical structures in patents. 3. Chemical grammars and their role in the manipulation of chemical structures. *J Chem Inf Comput Sci* 1981, 21:161–168.
179. Barnard JM, Lynch MF, Welford SM. Computer storage and retrieval of generic structures in chemical patents. 4. An extended connection table representation for generic structures. *J Chem Inf Comput Sci* 1982, 22:160–164.
180. Welford SM, Lynch MF, Barnard JM. Computer storage and retrieval of generic chemical structures in patents. 5. Algorithmic generation of fragment descriptors for generic structure screening. *J Chem Inf Comput Sci* 1984, 24:57–66.
181. Barnard JM, Lynch MF, Welford SM. Computer storage and retrieval of generic chemical structures in patents. 6. An interpreter program for the generic structure description language GENSAL. *J Chem Inf Comput Sci* 1984, 24:66–71.
182. Gillet VJ, Welford SM, Lynch MF, Willett P, Barnard JM, Downs GM, Manson G, Thompson J. Computer

storage and retrieval of generic chemical structures in patents. 7. Parallel simulation of a relaxation algorithm for chemical substructure search. *J Chem Inf Comput Sci* 1986, 26:118–126.

183. Gillet VJ, Downs GM, Ling A, Lynch MF, Venkataram P, Wood JV, Dethlefsen W. Computer storage and retrieval of generic chemical structures in patents. 8. Reduced chemical graphs and their applications in generic chemical structure retrieval. *J Chem Inf Comput Sci* 1987, 27:126–137.
184. Downs GM, Gillet VJ, Holliday JD, Lynch MF. Computer storage and retrieval of generic chemical structures in patents. 9. An algorithm to find the extended set of smallest rings in structurally explicit generics. *J Chem Inf Comput Sci* 1989, 29:207–214.
185. Downs GM, Gillet VJ, Holliday JD, Lynch MF. Computer storage and retrieval of generic chemical structures in patents. 10. Assignment and logical bubble-up of ring screens for structurally explicit generics. *J Chem Inf Comput Sci* 1989, 29:215–224.
186. Dethlefsen W, Lynch MF, Gillet VJ, Downs GM, Holliday JD, Barnard JM. Computer storage and retrieval of generic chemical structures in patents. 11. Theoretical aspects of the use of structure languages in a retrieval system. *J Chem Inf Comput Sci* 1991, 31:233–253.
187. Dethlefsen W, Lynch MF, Gillet VJ, Downs GM, Holliday JD, Barnard JM. Computer storage and retrieval of generic chemical structures in patents. 12. Principles of search operations involving parameter lists: matching-relations, user-defined match levels, and transition from the reduced graph search to the refined search. *J Chem Inf Comput Sci* 1991, 31:253–260.
188. Gillet VJ, Downs GM, Holliday JD, Lynch MF, Dethlefsen W. Computer storage and retrieval of generic chemical structures in patents. 13. Reduced graph generation. *J Chem Inf Comput Sci* 1991, 31:260–270.
189. Holliday JD, Downs GM, Gillet VJ, Lynch MF. Computer storage and retrieval of generic chemical structures in patents. 14. Fragment generation from generic structures. *J Chem Inf Comput Sci* 1992, 32:453–462.
190. Holliday JD, Downs GM, Gillet VJ, Lynch MF. Computer storage and retrieval of generic chemical structures in patents. 15. Generation of topological fragment descriptors from nontopological representations of generic structure components. *J Chem Inf Comput Sci* 1993, 33:369–377.
191. Holliday JD, Lynch MF. Computer storage and retrieval of generic chemical structures in patents. 16. The refined search: an algorithm for matching components of generic chemical structures at the atom-bond level. *J Chem Inf Comput Sci* 1995, 35:1–7.
192. Holliday JD, Lynch MF. Computer storage and retrieval of generic chemical structures in patents. 17. Evaluation of the refined search. *J Chem Inf Comput Sci* 1995, 35:659–662.
193. Downs GM, Barnard JM. Chemical patents and structural information—the Sheffield research in context. *J Doc* 1998, 54:106–120.
194. Nakayama T, Fujiwara Y. Computer representation of generic chemical structures by an extended block-cutpoint tree. *J Chem Inf Comput Sci* 1983, 23:80–87.
195. Berks AH. Current state of the art of Markush topological search systems. *World Pat Inf* 2001, 23:5–13.
196. Berks AH. Current state of the art of Markush topological search systems. In: Gasteiger J, ed. *Handbook of Cheminformatics: From Data to Knowledge*. Vol 2. Weinheim: Wiley-VCH; 2003, 885–903.
197. Barker EJ, Gardiner EJ, Gillet VJ, Kitts P, Morris J. Further development of reduced graphs for identifying bioactive compounds. *J Chem Inf Comput Sci* 2003, 43:346–356.
198. Gillet VJ, Willett P, Bradshaw J. Similarity searching using reduced graphs. *J Chem Inf Comput Sci* 2003, 43:338–345.
199. Birchall K, Gillet VJ, Harper G, Pickett SD. Training similarity measures for specific activities: application to reduced graphs. *J Chem Inf Model* 2006, 46:577–586.
200. Birchall K, Gillet VJ, Willett P, Ducrot P, Luttmann C. Use of reduced graphs to encode bioisosterism for similarity-based virtual screening. *J Chem Inf Model* 2009, 49:1330–1346.
201. Bonnet JC. Going to an actual chemical patent management system with DARC. In: Barnard JM, ed. *Computer Handling of Generic Chemical Structures*. Aldershot: Gower; 1984, 162–166.
202. Benichou P, Klimczak C, Borne P. Handling genericity in chemical structures using the Markush DARC software. *J Chem Inf Comput Sci* 1997, 37:43–53.
203. Cloutier KA. A comparison of three online Markush databases. *J Chem Inf Comput Sci* 1991, 31:40–44.
204. Schmuff NR. A comparison of the MARPAT and Markush DARC software. *J Chem Inf Comput Sci* 1991, 31:53–59.
205. Tokuno H. Comparison of Markush structure databases. *J Chem Inf Comput Sci* 1993, 33:799–804.
206. Wilke RN. *Searching for simple generic structures. J Chem Inf Comput Sci* 1991, 31:36–40.
207. Fisanick W. The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts. *J Chem Inf Comput Sci* 1990, 30:145–154.
208. Ebe T, Sanderson KA, Wilson PS. The Chemical Abstracts Service generic chemical (Markush) structure storage and retrieval capability. 2. The MARPAT file. *J Chem Inf Comput Sci* 1991, 31:31–36.
209. Csepregi S, Mate N, Dorant S, Biro E, Csizmazia T, Csizmadia F. Representation, searching and

enumeration of Markush structures: from molecules toward patents (abstract #COMP-245). Philadelphia, PA: 236th ACS National Meeting; August 17–21, 2008.

210. Barnard JM, Wright PM. Towards in-house searching of Markush structures from patents. *World Pat Inf* 2009, 31:97–103.

211. Willett P, ed. *Modern Approaches to Chemical Reaction Searching: Proceedings of a Conference; University of York, July 8–11, 1985*. Aldershot: Gower; 1986.

212. Chen L, Nourse JG, Christie BD, Leland BA, Grier DL. Over 20 years of reaction access systems from MDL: a novel reaction substructure search algorithm. *J Chem Inf Comput Sci* 2002, 42:1296–1310.

213. Ash JE, Warr WA, Willett P, eds. *Chemical structure systems: computational techniques for representation, searching, and processing of structural information*. Chichester: Ellis Horwood; 1991.

214. Reitz M, Sacher O, Tarkhov A, Truembach D, Gasteiger J. Enabling the exploration of biochemical pathways. *Org Biomol Chem* 2004, 2:3226–3237.

215. Koerner R, Apostolakis J. Automatic determination of reaction mappings and reaction center information. 1. The imaginary transition state energy approach. *J Chem Inf Model* 2008, 48:1181–1189.

216. Apostolakis J, Sacher O, Koerner R, Gasteiger J. Automatic determination of reaction mappings and reaction center information. 2. Validation on a biochemical reaction database. *J Chem Inf Model* 2008, 48:1190–1198.

217. Raymond JW, Willett P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J Comput Aided Mol Des* 2002, 16:521–533.

218. Willett P. The evaluation of an automatically indexed, machine-readable chemical reactions file. *J Chem Inf Comput Sci* 1980, 20:93–96.

219. McGregor JJ, Willett P. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J Chem Inf Comput Sci* 1981, 21:137–140.

220. Barth A. Status and future developments of reaction databases and online retrieval systems. *J Chem Inf Comput Sci* 1990, 30:384–393.

221. Blake JE, Dana RC. CASREACT: more than a million reactions. *J Chem Inf Comput Sci* 1990, 30:394–399.

222. Town WG. Large chemical structure and reaction databases: challenges and solutions. In: Collier H, ed. *Proceedings of the Montreux 1992 International Chemical Information Conference*. Tetbury: Infonortics; 1992, 209–223.

223. Roth DL. SPRESIweb 2.1, a selective chemical synthesis and reaction database. *J Chem Inf Model* 2005, 45:1470–1473.

224. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. *Science* 1969, 166:178–192.

225. Fick R, Ihlenfeldt W-D, Gasteiger J. Computer-assisted design of synthesis for heterocyclic compounds. *Heterocycles* 1995, 40:993–1007.

226. Gasteiger J, Ihlenfeldt WD, Roese P, Wanke R. Computer-assisted reaction prediction and synthesis design. *Anal Chim Acta* 1990, 235:65–75.

227. Ihlenfeldt W-D, Gasteiger J. Computer-assisted planning of organic syntheses: the second generation of programs. *Angew Chem Int Ed Engl* 1996, 34:2613–2633.

228. Gasteiger J, Pfortner M, Sitzmann M, Hollering R, Sacher O, Kostka T, Karg N. Computer-assisted synthesis and reaction planning in combinatorial chemistry. *Perspect Drug Discov Des* 2000, 20:245–264.

229. Judson P. *Knowledge-Based Expert Systems in Chemistry: Not Counting on Computers*. Cambridge: Royal Society of Chemistry; 2009.

230. Dugundji J, Ugi I. Algebraic model of constitutional chemistry as a basis for chemical computer programs. *Fortschr Chem Forsch* 1973, 39:19–64.

231. Bauer J, Herges R, Fontain E, Ugi I. IGOR and computer assisted innovation in chemistry. *Chimia* 1985, 39:43–53.

232. Jorgensen WL, Laird ER, Gushurst AJ, Fleischer JM, Gothe SA, Helson HE, Paderes GD, Sinclair S. CAMEO: a program for the logical prediction of the products of organic reactions. *Pure Appl Chem* 1990, 62:1921–1932.

233. Hoellering R, Gasteiger J, Steinhauer L, Schulz K-P, Herwig A. Simulation of organic reactions: from the degradation of chemicals to combinatorial synthesis. *J Chem Inf Comput Sci* 2000, 40:482–494.

234. Kostka T, Selzer P, Gasteiger J. A combined application of reaction prediction and infrared spectra simulation for the identification of degradation products of s-triazine herbicides. *Chem Eur J* 2001, 7:2254–2260.

235. Gasteiger J, Bauerschmidt S, Burkard U, Hemmer MC, Herwig A, Von HA, Hollering R, Kleinoder T, Kostka T, Schwab C, et al. Decision support systems for chemical structure representation, reaction modeling, and spectra simulation. *SAR QSAR Environ Res* 2002, 13:89–110.

236. Ott MA. Cheminformatics and organic chemistry. Computer-assisted synthetic analysis. In: Noordik JH, ed. *Cheminformatics Developments*. Amsterdam: IOS Press; 2004, 83–109.

237. Law J, Zsoldos Z, Simon A, Reid D, Liu Y, Khew SY, Johnson AP, Major S, Wade RA, Ando HY. Route Designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J Chem Inf Model* 2009, 49:593–602.

238. Gasteiger J, Sadowski J, Schuur J, Selzer P, Steinhauer L, Steinhauer V. Chemical information in 3D-space. *J Chem Inf Comput Sci* 1996, 36:1030–1037.

239. Sadowski J, Wagener M, Gasteiger J. Assessing similarity and diversity of combinatorial libraries by spatial autocorrelation functions and neural networks. *Angew Chem Int Ed Engl* 1996, 34:2674–2677.

240. Gasteiger J. The challenge of molecular structure representation for property prediction. *Actual Chim* 2008, 320–321:51–55.

241. Sadowski J, Schwab CH, Gasteiger J. 3D structure generation and conformation searching. In: Bultinck P, ed. *Computational Medicinal Chemistry for Drug Discovery*. New York, NY: Marcel Dekker; 2004, 151–212.

242. Pearlman RS. Rapid generation of high quality approximate 3D molecular structures. *Chem Des Autom News* 1987, 2:5–7.

243. Rusinko A III, Sheridan RP, Nilakantan R, Haraki KS, Bauman N, Venkataraghavan R. Using CONCORD to construct a large database of three-dimensional coordinates from connection tables. *J Chem Inf Comput Sci* 1989, 29:251–255.

244. Sadowski J, Gasteiger J, Klebe G. Comparison of automatic three-dimensional model builders using 639 X-ray structures. *J Chem Inf Comput Sci* 1994, 34:1000–1008.

245. Hiller C, Gasteiger J. An automatic molecule builder. In: Gasteiger J, ed. *Software Development in Chemistry. 1. Proceedings of the Workshops on the Computer in Chemistry, Hochfilzen/Tirol, November 19–21*, 1986. Vol 1. Berlin: Springer; 1987, 53–66.

246. Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D atomic coordinates for organic molecules. *Tetrahedron Comput Methodol* 1990, 3:537–547.

247. Kennard O, Watson DG, Town WG. Cambridge Crystallographic Data Centre. I. Bibliographic file. *J Chem Doc* 1972, 12:14–19.

248. Allen FH, Kennard O, Motherwell WDS, Town WG, Watson DG. Cambridge Crystallographic Data Centre. II. Structural data file. *J Chem Doc* 1973, 13:119–123.

249. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* 2002, 58:380–388.

250. Allen FH, Battle G, Robertson S. The Cambridge Structural Database. In: Triggle DJ, Taylor JB, eds. *Comprehensive Medicinal Chemistry II. Vol* 3. Amsterdam: Elsevier; 2007, 389–410.

251. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucl Acids Res* 2000, 28:235–242.

252. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, et al. The Protein Data Bank. *Acta Crystallogr D* 2002, 58:899–907.

253. Berman HM. The Protein Data Bank: a historical perspective. *Acta Crystallogr A* 2008, 64:88–95.

254. Hall SR, Allen FH, Brown ID. The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Crystallogr A* 1991, 47:655–685.

255. Brown ID, McMahon B. CIF: the computer language of crystallography. *Acta Crystallogr B* 2002, 58:317–324.

256. Bourne PE, Berman HM, McMahon B, Watenpaugh KD, Westbrook JD, Fitzgerald PMD. Macromolecular crystallographic information file. *Methods Enzymol* 1997, 277:571–590.

257. Westbrook JD, Fitzgerald PMD. The PDB format, mmCIF, and other data formats. *Methods Biochem Anal* 2003, 44:161–179.

258. Allen FH, Barnard JM, Cook APF, Hall SR. The Molecular Information File (MIF): core specifications of a new standard format for chemical data. *J Chem Inf Comput Sci* 1995, 35:412–427.

259. Wermuth CG, Ganellin CR, Lindberg P, Mitscher LA. Glossary of terms used in medicinal chemistry (IUPAC recommendations 1998). *Pure Appl Chem 1998*, 70:1129–1143.

260. Willett P. *Three-dimensional Chemical Structure Handling*. Taunton: Research Studies Press; 1991.

261. Martin YC, Willett P, eds. *Designing Bioactive Molecules: Three-Dimensional Techniques and Applications*. Washington DC: American Chemical Society; 1998.

262. Jakes SE, Willett P. Pharmacophoric pattern matching in files of 3-D chemical structures: selection of interatomic distance screens. *J Mol Graphics* 1986, 4:12–20.

263. Jakes SE, Watts N, Willett P, Bawden DD Fisher J. Pharmacophoric pattern matching in files of 3D chemical structures: evaluation of search performance. *J Mol Graphics* 1987, 5:41–48.

264. Brint AT, Willett P. Pharmacophoric pattern matching in files of 3D chemical structures: comparison of geometric searching algorithms. *J Mol Graphics* 1987, 5:49–56.

265. Martin YC, Danaher EB, May CS, Weininger D. MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical, or geometric properties. *J Comput Aided Mol Des* 1988, 2:15–29.

266. Van Drie JH, Weininger D, Martin YC. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J Comput Aided Mol Des* 1989, 3:225–251.

267. Cringean JK, Pepperrell CA, Poirrette AR, Willett P. Selection of screens for three-dimensional

substructure searching. *Tetrahedron Comput Methodol* 1990, 3:37–46.

268. Hurst T. Flexible 3D searching: the directed tweak technique. *J Chem Inf Comput Sci* 1994, 34:190–196.
269. Clark DE, Jones G, Willett P, Kenny PW, Glen RC. Pharmacophoric pattern matching in files of three-dimensional chemical structures: comparison of conformational-searching algorithms for flexible searching. *J Chem Inf Comput Sci* 1994, 34:197–206.
270. Moock TE, Henry DR, Ozkabak AG, Alamgir M. Conformational searching in ISIS/3D databases. *J Chem Inf Comput Sci* 1994, 34:184–189.
271. Smellie A, Kahn SD, Teig SL. Analysis of conformational coverage. 1. Validation and estimation of coverage. *J Chem Inf Comput Sci* 1995, 35:285–294.
272. Smellie A, Kahn SD, Teig SL. Analysis of conformational coverage. 2. Applications of conformational models. *J Chem Inf Comput Sci* 1995, 35:295–304.
273. Smellie A, Teig SL, Towbin P. Poling: promoting conformational variation. *J Comput Chem* 1995, 16:171–187.
274. Barnum D, Greene J, Smellie A, Sprague P. Identification of common functional configurations among molecules. *J Chem Inf Comput Sci* 1996, 36:563–571.
275. Güner OF, ed. *Pharmacophore: Perception, Development, and Use in Drug Design*. La Jolla, CA: International University Line; 2000.
276. Leach AR, Gillet VJ, Lewis RA, Taylor R. Three-dimensional pharmacophore methods in drug discovery. *J Med Chem* 2010, 53:539–558.
277. Rush TS III, Grant JA, Mosyak L, Nicholls A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J Med Chem* 2005, 48:1489–1495.
278. Cheeseright T, Mackey M, Rose S, Vinter A. Molecular field extrema as descriptors of biological activity: definition and validation. *J Chem Inf Model* 2006, 46:665–676.
279. Zimmermann M, Fluck J, Thi LTB, Kolarik C, Kumpf K, Hofmann M. Information extraction in the life sciences: perspectives for medicinal chemistry, pharmacology and toxicology. *Curr Top Med Chem* 2005, 5:785–796.
280. McDaniel JR, Balmuth JR. Kekule: OCR-optical chemical (structure) recognition. *J Chem Inf Comput Sci* 1992, 32:373–378.
281. Simon A, Johnson AP. Recent advances in the CLiDE project: logical layout analysis of chemical documents. *J Chem Inf Comput Sci* 1997, 37:109–116.
282. Zimmermann M. The art of teaching the computer to read. *Nachr Chem* 2007, 55:997–1000.
283. Filippov IV, Nicklaus MC. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J Chem Inf Model* 2009, 49:740–743.
284. Park J, Rosania GR, Shedden KA, Nguyen M, Lyu N, Saitou K. Automated extraction of chemical structure information from digital raster images. *Chem Cent J* 2009, 3 (Online).
285. Valko AT, Johnson AP. CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. *J Chem Inf Model* 2009, 49:780–787.

FURTHER READING

Barnard JM. Representation of chemical structures—overview. In: Gasteiger J, ed. *Handbook of Cheminformatics: From Data to Knowledge*. Vol 1. Weinheim: Wiley-VCH; 2003, 27–50.

Bishop N, Gillet VJ, Holliday JD, Willett P. Chemoinformatics research at the University of Sheffield: a history and citation analysis. *J Inf Sci* 2003, 29:249–267.

Chen WL. Chemoinformatics: past, present, and future. *J Chem Inf Model* 2006, 46:2230–2255.

Engel T, Gasteiger J. Chemical structure representation for information exchange. *Online Inf Rev* 2002, 26:139–145.

Engel T. Basic overview of chemoinformatics. *J Chem Inf Model* 2006, 46:2267–2277.

Gasteiger J, ed. *Handbook of Chemoinformatics: From Data to Knowledge*. Weinheim: Wiley-VCH; 2003.

Gasteiger J, Engel T, eds. *Chemoinformatics: A Textbook*. Weinheim: Wiley-VCH; 2003.

Gasteiger J. The central role of chemoinformatics. *Chemom Intell Lab Syst* 2006, 82:200–209.

Gasteiger J. Chemoinformatics: a new field with a long tradition. *Anal Bioanal Chem* 2006, 384:57–64.

Leach AR, Gillet VJ. *An Introduction to Chemoinformatics*. Dordrecht: Kluwer; 2003.

Lynch MF, Willett P. Current research into chemical and textual information retrieval at the Department of Information Studies, University of Sheffield. *Inf Process Manage* 1987, 23:447–463.

Paris GC. Chemical structure handling by computer. *Ann Rev Inf Sci Technol* 1997, 32:271–337.

Paris GC. Structure databases. In: Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR, eds. *Encyclopedia of Computational Chemistry*. Vol 4. Chichester: John Wiley & Sons; 1998, 2771–2785.

Paris GC. Databases of chemical structures. In: Gasteiger J, ed. *Handbook of Cheminformatics: From Data to Knowledge*. Vol 2. Weinheim: Wiley-VCH; 2003, 523–555.

Suhr C. A change of paradigms: looking back to the pioneer years of patent information management (1960–1990). *World Pat Inf* 2004, 26:41–43.

Warr WA. Available systems of structure representation. In: Lees R, Smith AF, eds. *Chemical Nomenclature Usage*. Chichester: Ellis Horwood; 1983, 124–131.

Warr WA, ed. *Chemical Structure Information Systems: Interfaces Communication and Standards*. Washington DC: American Chemical Society Symposium Series 400; 1989.

Warr WA, Suhr C. *Chemical information management*. Weinheim: Wiley-VCH; 1992.

Willett P. A history of chemoinformatics. *Handbook of Chemoinformatics: From Data to Knowledge*. Vol 1. Weinheim: Wiley-VCH; 2003, 6–20.

Willett P. From chemical documentation to chemoinformatics: 50 years of chemical information science. *J Inf Sci* 2008, 34:477–499.

Willett P. Similarity methods in chemoinformatics. *Ann Rev Inf Sci Technol* 2009, 43:3–71.