Yet Another Representation of Molecular Structure[†]

Andreas Dietz

GDR 1093 of the CNRS "Traitement Informatique de la Connaissance en Chimie Organique",
CCIPE - UPR 9023, Rue de la Cardonille, 34094 Montpellier Cedex 5, France

A representation of molecular structure has been devised for application in computer-assisted synthesis design systems and chemical information storage and retrieval systems. It takes into account constitution, configuration, and conformation. Constitution is described in terms of unshared valence electrons and "bonding systems". The latter comprise all pairwise bonding relations between the atoms of a chemical structure. Bonding relations arising from both valence electron sharing and electrostatic interaction are accounted for. Stereochemistry is treated on an "approximate-geometrical" level: Instead of exact bond angles and lengths only the relative spatial arrangement of atoms is taken into consideration. Thus, our molecular structure representation can be regarded as the computer-oriented counterpart of the graphical representation via structural formulas.

1. INTRODUCTION

The importance which structural formulas have gained as the international language of chemistry proves that they provide, for many purposes, the adequate level of abstraction for the representation of molecular structure. For example, when a chemist develops a synthetic plan, he must account for both constitutional and stereochemical features of the involved molecules. He graphically represents the molecules using structural formulas and perceives (relationships between) certain structural elements. Associating knowledge with these structural units is a crucial step in his reasoning process. For instance, retrosynthetic analysis essentially relies on the association of synthetic methods with structural units of the target molecule.

The representation of molecular structure decisively determines the scope of a chemical computer program. Our goal is to provide a versatile computer-oriented molecular structure representation for chemical information storage and retrieval as well as for computer-assisted synthesis design. Structural formulas describe molecular structure on the proper level of abstraction for these applications. Chemical information storage and retrieval systems should allow to specify both the constitutional and the stereochemical features of the chemical structures for which associated information is to be entered or sought. The same is true for entry and look-up of information associated with a given chemical structure in the knowledge base of a computer-assisted synthesis design system. It is therefore desirable that the computer-oriented representation of molecular structure be as expressive as the structural formulas.

Which information does a structural formula provide about the molecule that it represents?

First, it tells us about its chemical constitution: Which atoms are present? Which pairwise bonding relations do exist between them? How are the valence electrons distributed among and between the atoms? A bonding relation between two atoms may result from shared electron interaction, for example, in a covalent single bond. It may also result from unshared electron interaction, for example, in ionic bonding or hydrogen bonding. The same electrons may give rise to more than one bonding relation. For example, the delocalized π electrons in benzene give rise to six bonding relations. On the other hand, two atoms may be connected by several bonding relations. A chemical bond is the combined effect of all bonding relations that exist between two atoms. For example, two localized σ electrons as well as six delocalized π electrons contribute to the bond between two neighbor carbon atoms in a benzene ring.

Second, a structural formula conveys information about stereochemical features of the represented molecule, that is, information about its configuration and conformation. It is important to note that the underlying model of stereochemistry is an "approximate-geometrical" one: A structural formula does not provide any information about exact bond angles and lengths, rather it informs us about the relative spatial arrangement of the atoms.

We may draw different structural formulas to graphically represent a given molecule, according to its orientation in three-dimensional Euclidean space. Structural formulas are therefore unambiguous but not unique representations of molecular structure. A valuable computer-oriented representation of molecular structure has to be both unique and unambiguous. It has therefore to be built on the perspective independent features of structural formulas, that is, the pairwise bonding relations between atoms, the valence electron distribution, and the relative spatial arrangement of atoms.

We shall propose a computer-oriented molecular structure representation such that any chemical structure which can be represented by a structural formula will have an adequate computer representation.

Although structural formulas have been in use for about a century, to our knowledge no attempts have yet been made to formally define the underlying approximate-geometrical model of stereochemistry. The concepts of configuration and conformation are therefore being used by chemists without ever having been assigned a clearcut meaning. The IUPAC recommendations[1] do not remedy this problem. First, it is stated that there exist several definitions for these concepts, without giving preference to any of them. Second,

---

the meaning of the proposed definitions depends on what is meant by "arrangement of its atoms in space".[1] We suppose that the *relative* arrangement in space is meant, as the examples of configurationally and conformationally different molecules are represented by structural formulas without indication of bond lengths and angles.

We shall therefore formally define configuration and conformation in terms of the relative spatial arrangement of atoms.

The structure of this article reflects the factorization of molecular structure into constitution and stereochemistry (configuration and conformation).

2. ATOMS, ELECTRONS, AND CHEMICAL STRUCTURES

In order to describe molecular structure, it is common practice to conceptually dissect an atom into an atomic core and a set of valence electrons. The atomic core comprises the atomic nucleus together with the inner-shell electrons. Only the valence electrons are regarded to be relevant for chemical bonding. Which electrons are to be considered as valence electrons depends on the position of the element in the periodic table. Usually, ns and np electrons are considered as valence electrons of atoms belonging to the main group elements of the nth row of the periodic table. Transition-metal atoms of the nth row additionally have the $(n-1)d$ electrons as valence electrons.[2] In the following, the term atom will, somewhat inaccurately, also be used as a synonym for atomic core.

Let A be a set of atoms. We can give a chemical semantics to A by imposing structural constraints onto A. Constitutional constraints concern the distribution of valence electrons among and between the atomic cores as well as bonding relations between the latter; stereochemical constraints concern the relative spatial arrangement of the atomic cores. A "chemical structure" is an atom set with structural constraints; it may be any set of contiguously bonded atoms with a certain relative spatial arrangement, for example, an atom, a molecule, or a molecular fragment.

3. CONSTITUTION

In this section we shall concern ourselves with constitution, the most basic feature of molecular structure. We shall first show the limitations of existing approaches to define and represent constitution. Then we shall propose a novel definition and representation which overcome these problems. To this end, we shall introduce the concept of "bonding system". The latter will enable us to cope with chemical bonds of any type occurring in organic and organometallic chemistry. Examples will illustrate the versatility of our approach.

3.1. Traditional Definition and Representation of Constitution. Usually, constitution is defined in terms of chemical bonds between atoms and it is represented within the mathematical framework of graph theory.[3,4] A "molecular graph" represents the constitution of a chemical structure. It is a connected, labeled, and undirected graph. The vertices are atoms, and the edges are bonds. A vertex is labeled with an atom type (chemical element) and possibly the number of unshared electrons; an edge is labeled with a bond order or a bond type (single, double, aromatic, etc.).

In the following, we shall investigate scope and limitations of molecular graphs in the modeling of chemical constitution. The be-matrices of the Dugundji-Ugi model[5] and their successors will serve as examples. Connection tables are another popular notation for molecular graphs.[6] Clearly, scope and limitations of the graph-theoretical approach do not depend on which notation is being used.

A be-matrix represents an "ensemble of molecules".[5] Therefore, a be-matrix is equivalent to a set of molecular graphs; a vertex of such a graph is labeled with its atom type and the number of unshared valence electrons; an edge is labeled with the formal bond order.

The be-matrices are adequate to represent the constitution of chemical structures exclusively having localized bonds. In this case, the off-diagonal entries (formal bond orders) in the corresponding be-matrices are all integers. Problems arise when chemical structures with delocalized bonds or multicenter bonds are encountered. It has been argued that be-matrices would not be suitable for the representation of "molecular systems having delocalized electrons or multicenter bonds [...], because fractional bond orders are encountered".[7] However, rational bond orders by themselves do not prevent the be-matrices from an elegant and uniform representation of delocalized electron systems and multicenter bonds. As we shall see below, it is rather a matter of how the rational bond orders are interpreted.

In the original publication[5] on the Dugundji-Ugi model, the authors state the following: "The strongly contributing valence bond structures of resonance systems with delocalized π-electrons are represented by a class of *be*-matrices which differ only in bond orders and diagonal entries but not in connectivities [...]". A later publication[8] explicitly suggests that such classes of be-matrices could serve to represent the constitution of chemical structures with delocalized bonds. However, such a representation would be rather clumsy, especially if one considers the large number of "strongly contributing valence bond structures" that would have to be taken into account in the case of fused aromatic systems.

On the other hand, Dugundji and Ugi have proposed that rational bond orders in be-matrices should be used to represent multicenter bonds.[5] The rational off-diagonal matrix entries were of the form x_{ij}/y_{ij}; the numerator x_{ij} was to indicate the number of valence electrons that belong to the multicenter bond in which atoms i and j participate; the denominator y_{ij} indicated the number of atomic cores that share these electrons.

We believe that the distinction between multicenter bonds and delocalized bonds is somewhat artificial. Both distinguish themselves from localized bonds by the fact that bonding electrons are shared by more than two atoms. Consequently, they should be treated in a uniform manner.

The following minor modification of the Dugundji-Ugi model should have been sufficient to integrate delocalized electron systems and multicenter bonds uniformly into the be-matrix representation of chemical constitution: Both delocalized electron systems and multicenter bonds would be considered as delocalized bonding. Each rational off-diagonal matrix entry would carry information about the contributions of both localized and delocalized bonding to a bond between two atoms. Rational bond orders would have the general form $i^n/_d$. The integer part i, the "localized bond order", would be equal to half the number of electrons
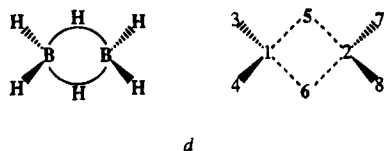
YET ANOTHER REPRESENTATION OF MOLECULAR STRUCTURE

J. Chem. Inf. Comput. Sci., Vol. 35, No. 5, 1995 **789**

d

Figure 1. Electron deficiency bonding in diborane.

localized between the corresponding atoms. The fractional part $^n/_d$, the "delocalized bond order", would be equal to half the ratio $^e/_p$, where e is the number of delocalized electrons, and p is the number of neighbor atom pairs in the delocalized electron system.

The representation of reactions involving molecules with delocalized bonding would require to modify the r-matrices[5] analogously to the be-matrices. A rational off-diagonal r-matrix entry of the general form $i^n/_d$ would convey at the same time information about changes in localized bond order (integer part i) and about changes in delocalized bond order (fractional part $^n/_d$).

If the Dugundji-Ugi model was modified in the above way, the operation "+" (elementwise matrix addition) in the matrix equation $B + R = E^5$ would be preserved. (The integer and the fractional parts of rational entries would have to be added separately, though.) Also, the sum over all entries of a r-matrix would still be zero, and the sum over all entries in a be-matrix would still correspond to the number of valence electrons in the ensemble of molecules.

Ugi and co-workers have not pursued the original idea of rational bond orders in be-matrices[5] any further. Instead, the be-matrices have been limited to integer entries, and xbe-matrices ("extended be-matrices") have been introduced.[7,9] An xbe-matrix is an integer be-matrix which includes one additional row/column for each delocalized electron system. Such a row/column specifies which atoms participate in a delocalized electron system and how many electrons are delocalized. Localized bonding is represented by the integer be-matrix.

The transition from be-matrices to xbe-matrices corresponds to a transition from a set of molecular graphs to a set of "molecular hypergraphs". A molecular hypergraph is a connected and labeled hypergraph.[10] The vertices of a molecular hypergraph are atoms, and the edges ("hyperedges") are sets of at least two atoms. If an edge contains two elements, it represents localized bonding; if it contains more than two elements, it represents a delocalized electron system. Vertices are labeled with atom types and the number of unshared electrons; edges are labeled with the number of electrons which are localized between or delocalized over the corresponding atoms.

However, representing delocalized bonding via hyperedges has a serious disadvantage: A hyperedge containing more than two atoms gives us no information about the binary neighborhood relationships between them. That means we have no information at our disposal concerning the way in which the electrons are delocalized over these atoms.

In order to overcome this problem, one could assume that the delocalized electrons follow a framework of localized bonds. For example, the delocalized π electrons of a benzene molecule follow the σ bond framework of the ring. However, this assumption is not generally valid.

For example, in the diborane molecule d (Figure 1) there is no framework of localized bonds in the region occupied by the delocalized electrons. The electron deficiency two
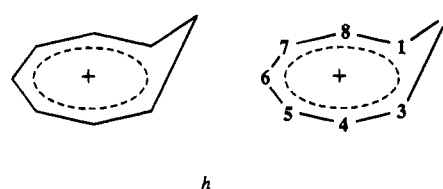

h

Figure 2. The atom adjacencies in the π electron system differ from those in the σ bond framework.

electron/three center B$-$H$-$B bonds would give rise to hyperedges {1,2,5} and {1,2,6}. (Atoms are represented by natural numbers.) However, these hyperedges could represent electron deficiency bonds linearly connecting two boron atoms via a hydrogen bridge as well as cyclic electron deficiency systems.

Diborane is not the only molecule where the atom adjacencies in a delocalized electron system do not coincide with those of a σ bond framework. Other examples are transition-metal complexes or the homotropylium ion h (in Figure 2, hydrogen atoms have been omitted for convenience). The latter is an example of a homoaromatic molecule;[11] it has a cyclic delocalized π electron system. This is another example of a chemical structure that cannot be adequately represented using an xbe-matrix. An xbe-matrix can only inform us about the presence of a delocalized electron system containing six electrons and involving all carbon atoms but the sp^3 carbon atom 2. Information whether the delocalization is cyclic or linear (following the σ bond framework along atoms 3, 4, 5, 6, 7, 8, and 1) is beyond the scope of xbe-matrices.

In order to describe conjugated and aromatic systems, Gasteiger[12] separates σ, π, and unshared electrons. His representation of chemical constitution corresponds to a molecular graph where each atom is labeled (among other things) with the number of electrons that it contributes to a delocalized π electron system. Clearly, the scope of this approach is limited by the underlying assumption that the delocalized electrons follow a framework of localized bonds (see above).

In the so-called "symbolic xbe-matrices"[13] (sxbe-matrices, formerly "modified xbe-matrices"[7]), bond types replace the integer localized bond orders of the xbe-matrices. Thus, an sxbe-matrix indicates all pairwise connections of atoms, including the delocalized electron systems.[13] The additional rows/columns of an sxbe-matrix only serve to indicate the number of electrons in the corresponding delocalized electron systems.

Therefore, unlike the xbe-matrices, the sxbe-matrices do not correspond to a set of molecular hypergraphs but to a set of molecular graphs. A vertex of such a graph is labeled with its atom type and the number of unshared valence electrons; an edge is labeled with a bond type and, if the bond type is one of the three types reserved for bonds belonging to delocalized electron systems, with the number of delocalized electrons.[13,14] (The type "edsys" characterizes bonds belonging to electron deficiency systems, "pisys" is used for bonds belonging to aromatic systems or conjugated π systems, and "coord" for coordination bonds between a metal atom and an organic ligand.)

Any constitutional representation which labels the edges of a molecular graph with bond types has the following drawback: Its expressiveness is limited by the set of predefined bond types. The set of bond types may be
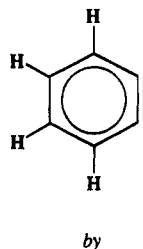
790 *J. Chem. Inf. Comput. Sci., Vol. 35, No. 5, 1995*

DIETZ

by

Figure 3. Benzyne.

extensible. However, there is a tradeoff between the expressiveness and simplicity of the constitutional representation. Hence, the set of available bond types in the original sxbe-matrix approach[7] had to be modified and reduced.[13,14] As we shall see below, it is easy to find examples of molecules containing bonds for which no adequate type is available in a given set. Enhancing the expressiveness by including a new bond type for every exceptional case is certainly not a very elegant solution.

For example, consider the benzyne molecule *by* in Figure 3. The electronic situation is similar to that of benzene, that is, the aromatic sextet is intact;[11] two C—H σ bonds in benzene are replaced by a reactive C—C π bond in benzyne. The corresponding carbon atoms are bonded by six delocalized π electrons as well as by four localized σ and π electrons. Due to the lack of a corresponding bond type in the proposed bond type set,[13,14] the benzyne molecule *by* cannot be represented by an sxbe-matrix.

A bond type has a chemical interpretation, that is, an interpretation which goes beyond the simple assignment of a number of electrons to a bond. For example, if a bond in an sxbe-matrix is labeled with the type "edsys", this means that it is involved in an electron deficient multicenter bond, where the formal bond order of each participating bond lies between zero and one.[13,14]

The homotropylium ion *h* (Figure 2) is another example of a chemical structure that cannot be represented by an sxbe-matrix because no adequate bond type is available. Six electrons are delocalized over a cyclic region defined by seven pairs of adjacent atoms. Thus, the formal order of the bond between carbon atoms 1 and 3 is $^3/_7$. However, we cannot label this bond with the type "edsys", because it does not belong to an electron deficient multicenter bond. (The remaining bonds between the carbon atoms involved in the delocalized electron system have formal order $1^3/_7$.) Neither can this bond be labeled with any other bond type belonging to the set proposed in refs 13 and 14.

Bonding based on unshared electron interaction cannot be represented by be-matrices, because a formal bond order of zero is interpreted as the absence of bonding between the corresponding atoms. In the xbe-matrices, this is not necessarily the case, as these atoms might be part of delocalized electron systems. (See, for example, the diborane molecule.) Unshared electron interaction could therefore be represented by xbe-matrices, if one declares the corresponding atoms to take part in delocalized electron systems without electrons. However, no information about the binary neighborhood relationships between these atoms would be available. Symbolic xbe-matrices could, in principle, cope with unshared electron interaction bonding. However, the necessary bond types are missing in the proposed bond type set.[13,14] (The bond type "coord" is interpreted in terms of unshared electron interaction bonding. However, coordina-

tion bonding stems from overlap of filled ligand orbitals with unfilled metal orbitals and from overlap of filled metal orbitals with unfilled ligand orbitals.[2] Therefore, coordination bonding is based on shared electron interaction, and the semantics of the bond type "coord" should be defined accordingly.)

Molecular graphs are based on the concepts of atom and bond, where a bond is a binary relationship between two atoms. However, delocalized bonds stretch over more than two atoms. This is why molecular graphs cannot adequately represent the constitution of chemical structures with delocalized bonds. These must be considered as a whole as well as in detail: Information about which atoms are involved in a delocalized bond as well as information about the binary neighborhood relationships between those atoms must be available. Hypergraphs do not meet the second requirement. In the following, we shall present a novel approach to model chemical constitution which avoids the problems encountered with molecular (hyper)graphs.

3.2. A Novel Definition and Representation of Constitution. Constitution. The constitution of a chemical structure is made up by the distribution of (possibly empty) sets of unshared valence electrons among the atomic cores and by the set of all "bonding systems" that are present. A bonding system consists of a (possibly empty) set of electrons which are spread over a region of pairwise adjacent atoms. A bonding relation exists between any two adjacent atoms ("neighbor atoms") of a bonding system. It either stems from shared electron interaction or, if the electron set is empty, from unshared electron interaction.

The bonding system concept is very general. It allows us to uniformly treat any kind of chemical bond in organic and organometallic chemistry. A chemical bond between atoms x and y is the combined effect of all bonding relations between x and y. Its type is therefore determined by the different bonding systems in which x and y are neighbor atoms.

Constitution as defined above can be modeled on the basis of "molecular multigraphs". A molecular multigraph represents the constitution of a chemical structure. It is a connected, labeled, and undirected multigraph[15] whose vertices are atoms and whose edges are bonding relations. A vertex is labeled with its atom type and the number of unshared electrons; an edge is labeled with the identifier of the corresponding bonding system and the number of electrons in that bonding system. (We are not interested in a distinction of individual electrons, so we represent electron sets simply by their cardinal numbers.)

A multigraph may have several edges between the same two vertices. Therefore, matrix notation is not adequate to represent molecular multigraphs. In the following, we shall describe a notation which is based on tuples and sets.

The order and multiplicity of elements is characteristic for a tuple or a list but not for a set. An n-tuple is an element of the Cartesian product of n sets, thus having fixed cardinality, whereas the cardinality of lists and sets is arbitrary. We use parentheses to denote lists as well as tuples, and we use braces to denote sets. For example, (x,y) is a pair and $\{x,y\}$ is a pair set; $()$ denotes the empty list and $\{\}$ denotes the empty set.

The constitution of a chemical structure is represented by a pair consisting of two sets: The first set represents the
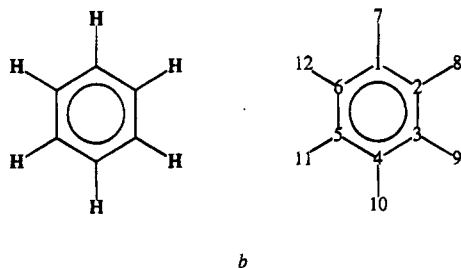
b

Figure 4. The benzene molecule comprises thirteen bonding systems.

distribution of unshared valence electrons among the atoms; for each atom, this set includes a triple consisting of the number of its unshared valence electrons, the atom itself, and the chemical element to which it belongs. The second set contains the bonding systems; a bonding system is represented by a pair consisting of the cardinal number of its electron set and a nonempty set of atom pair sets, each of them representing two neighbor atoms.

The bonding system identifiers which are required for the edge labeling of the molecular multigraph are implicit in this notation. The set of atom pair sets belonging to a bonding system may itself be considered as the identifier of the latter.

Let *Constit* be a function generating the constitutional representation of a chemical structure. The constitution of the hydrogen molecule *hy* is represented by

$$Constit(hy) = (\{(0,1,H),(0,2,H)\},\{(2,\{\{1,2\}\})\})$$

Neither hydrogen atom 1 nor hydrogen atom 2 has unshared valence electrons. This fact is represented by the triples (0,1,H) and (0,2,H).

The bonding system (2,{{1,2}}) is the only one present in *hy*. The cardinality of its electron set is 2. These electrons occupy a region confined to neighbor atoms 1 and 2, because {1,2} is the only atom pair set involved in this bonding system. On the other hand, because {1,2} is involved in no other bonding system, these two electrons are the only ones shared by atoms 1 and 2. Chemists refer to bonds such as the one between atoms 1 and 2 as covalent single bonds.

The prototype of aromatic compounds, benzene, may serve to illustrate the representation of aromatic bonds. The constitution of the benzene molecule *b* (Figure 4) is represented by

$Constit(b) =$
  $(\{(0,1,C),(0,2,C),(0,3,C),(0,4,C),(0,5,C),(0,6,C),$
  $(0,7,H),(0,8,H),(0,9,H),(0,10,H),(0,11,H),(0,12,H)\},$
  $\{(2,\{\{1,7\}\}),(2,\{\{2,8\}\}),(2,\{\{3,9\}\}),$
  $(2,\{\{4,10\}\}),(2,\{\{5,11\}\}),(2,\{\{6,12\}\}),$
  $(2,\{\{1,2\}\}),(2,\{\{2,3\}\}),(2,\{\{3,4\}\}),$
  $(2,\{\{4,5\}\}),(2,\{\{5,6\}\}),(2,\{\{6,1\}\}),$
  $(6,\{\{1,2\},\{2,3\},\{3,4\},\{4,5\},\{5,6\},\{6,1\}\})\})$

Each one of the first six bonding systems (2,{{1,7}}),..., (2,{{6,12}}) contains a set of two electrons which are responsible for a single bonding relation between a carbon and a hydrogen atom. The corresponding neighbor atom pair sets are not involved in any of the remaining bonding systems. Therefore, these bonding systems correspond to C—H single bonds.

Bonding systems seven to 12 represent the σ parts of the C—C ring bonds. Each bonding system includes a set of two electrons establishing a bonding relation between two

carbon atoms. But the corresponding carbon atom pair sets also figure together in the last bonding system. There, a set of six electrons occupies a cyclic region defined by the set of these six atom pair sets. This is the π electron system, whose electrons are at the same time responsible for six bonding relations between each two carbon neighbor atoms. Affiliation of the carbon atom pair set {x,y} both to one of the bonding systems seven to 12 and to the last bonding system determines the type of chemical bond between x and y. Chemists refer to it as aromatic.

In the benzene molecule, σ electrons and π electrons belong to distinct bonding systems. This is because the σ electrons are spread over a region given by one atom pair set, whereas the π electrons cover a region of six atom pair sets. However, in a double or a triple bond both σ and π electrons occupy the region between two atoms. Their union belongs to a single bonding system. A triple bond between atoms x and y would be represented by the bonding system (6,{{x,y}}).

Diborane is an example of a molecule with electron deficiency bonds. The constitution of the diborane molecule *d* (Figure 1) is represented by

$Constit(d) =$
  $(\{(0,1,B),(0,2,B),(0,3,H),(0,4,H),$
  $(0,5,H),(0,6,H),(0,7,H),(0,8,H)\},$
  $\{(2,\{\{1,3\}\}),(2,\{\{1,4\}\}),$
  $(2,\{\{2,7\}\}),(2,\{\{2,8\}\}),$
  $(2,\{\{1,5\},\{2,5\}\}),(2,\{\{1,6\},\{2,6\}\})\})$

The first four bonding systems correspond to covalent B—H single bonds, whereas the electron deficiency two electron/ three center B—H—B bonds are described by the last two bonding systems. For example, (2,{{1,5},{2,5}}) represents the distribution of two electrons over a region stretching from boron atom 1 over hydrogen atom 5 to boron atom 2. Note that, in our terminology, a chemical bond always is a binary relationship between two atoms, the type of the chemical bond being determined by the participation of the corresponding neighbor atom pair set in one or more bonding systems. Therefore, instead of one electron deficiency two electron/three center B—H—B bond we rather speak of two (electron deficiency) B—H bonds sharing the same set of two electrons.

Just like the delocalized electrons of the diborane molecule *d*, the cyclically delocalized electrons of the homotropylium ion *h* (Figure 2) do not follow a framework of σ bonds. If we ignore hydrogen atoms, the constitution of *h* is represented by

$Constit(h) =$
  $(\{(0,1,C),(0,2,C),(0,3,C),(0,4,C),$
  $(0,5,C),(0,6,C),(0,7,C),(0,8,C)\},$
  $\{(2,\{\{1,2\}\}),(2,\{\{2,3\}\}),(2,\{\{3,4\}\}),(2,\{\{4,5\}\}),$
  $(2,\{\{5,6\}\}),(2,\{\{6,7\}\}),(2,\{\{7,8\}\}),(2,\{\{8,1\}\}),$
  $(6,\{\{1,3\},\{3,4\},\{4,5\},\{5,6\},\{6,7\},\{7,8\},\{8,1\}\})\})$

We may also represent coordination bonding in transition-metal complexes, for example, in the ferrocene molecule *fc* depicted in Figure 5. (Hydrogen atoms are omitted for the sake of simplicity.) We imaginarily dissect *fc* into Fe^{2+} and two Cp^- units, attributing six valence electrons to each of these units.[2] Coordination bonding may then be described by three bonding systems *i*, *ii*, and *iii*, represented by the dotted lines in Figure 5:
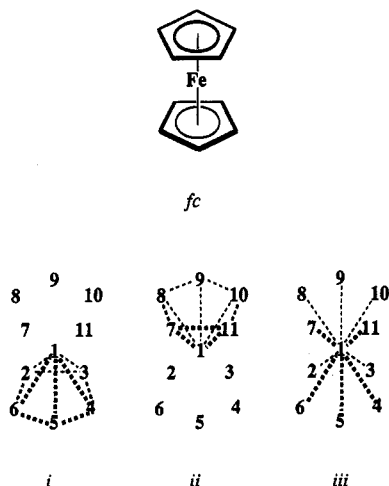
fc

i   *ii*   *iii*

Figure 5. An example for coordination bonding.

i : $(6, \{\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\},$
         $\{2,3\}, \{3,4\}, \{4,5\}, \{5,6\}, \{6,2\}\})$

ii : $(6, \{\{1,7\}, \{1,8\}, \{1,9\}, \{1,10\}, \{1,11\},$
         $\{7,8\}, \{8,9\}, \{9,10\}, \{10,11\}, \{11,7\}\})$

iii : $(6, \{\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{1,6\},$
         $\{1,7\}, \{1,8\}, \{1,9\}, \{1,10\}, \{1,11\}\})$

In bonding system i a set of six electrons is responsible for ten bonding relations. They include the five bonding relations between the iron atom and the carbon atoms of one cyclopentadienide ring. These bonding relations model the contribution of the ligand π electron system to the coordination bonds ("donation"). The remaining five bonding relations between ring carbon atoms model the contribution of the π electron system to the aromatic ring bonds. Bonding system ii is analogous to bonding system i but involves the other cyclopentadienide ring. Finally, the contribution of the electrons of the iron atom to the coordination bonds is modeled by bonding system iii. The latter represents "backdonation" from the iron atom toward both ligands by ten Fe—C bonding relations which are due to the presence of a set of six valence electrons.

The bonding systems corresponding to the C—H single bonds and to the σ parts of the aromatic ring bonds are represented analogously to those of the benzene molecule b (Figure 4).

Note that a molecular structure representation cannot free the user from the task to decide *how* a chemical structure should be represented. However, a lack of versatility might force the user to represent a chemical structure in a certain manner, even if he actually would prefer to represent it differently.

Our constitutional representation is rather flexible. If one wishes to represent an ionic ferrocene, one may do so. If one wishes to represent coordination bonding in ferrocene with a single bonding system containing 18 electrons, one may do so. We have chosen the above representation of ferrocene in order to reflect the synergetic bonding concept of coordination bonding.[2]

All types of chemical bonding presented so far were based on shared electron interaction. Now we will show how to describe bonding based on unshared electron interaction, that is, hydrogen bonding and ionic bonding. Because these stem from electrostatic interaction,[16] the respective bonding systems have an empty set of electrons. We may distinguish ionic bonding from hydrogen bonding by calculating the
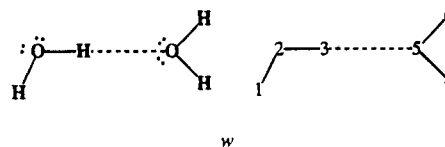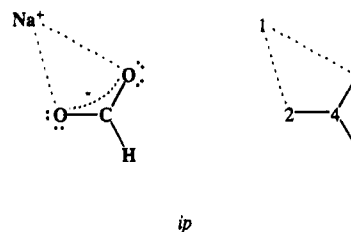

w

Figure 6. Hydrogen bonding in water.

ip

Figure 7. Ionic bonding in Na formiate.
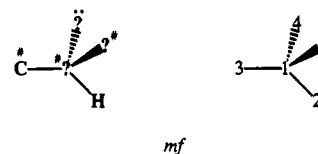

mf

Figure 8. Incompletely specified atoms in a molecular fragment.

formal charges (section 3.3) of the atoms between which an electrostatic bonding relation exists.

Figure 6 shows two water molecules which are associated via a hydrogen bond to give a single chemical structure w. Its constitution is represented by

$Constit(w) =$
$(\{(0,1,H), (4,2,O), (0,3,H), (0,4,H), (4,5,O), (0,6,H)\},$
$\{(2, \{\{1,2\}\}), (2, \{\{2,3\}\}), (2, \{\{4,5\}\}), (2, \{\{5,6\}\}),$
$(0, \{\{3,5\}\})\})$

Figure 7 shows a Na^+HCOO^- ion pair ip. The constitution of this chemical structure is represented by

$Constit(ip) =$
$(\{(0,1,Na), (4,2,O), (4,3,O), (0,4,C), (0,5,H)\},$
$\{(0, \{\{1,2\}, \{1,3\}\}), (2, \{\{2,4\}, \{3,4\}\}),$
$(2, \{\{2,4\}\}), (2, \{\{3,4\}\}), (2, \{\{4,5\}\})\})$

A chemical structure, for example, a molecular fragment, may contain incompletely specified atoms. Either their type or the number of their unshared valence electrons or both may be unspecified. Furthermore, not every bonding system in which an atom participates is necessarily specified. We call an atom of unspecified type a "wildcard atom" and consider it as belonging to the hypothetical chemical element "UNDEFINED"; a wildcard atom has an undefined number of valence electrons.

Consider, for example, molecular fragment mf in Figure 8. Only one bonding system involving carbon atom 3 is specified; the cardinal number of its set of unshared valence electrons is undefined, as represented by the symbol "#". Central atom 1 and its two neighbor atoms 4 and 5 are of unspecified type; they are wildcard atoms with the element symbol "?". Two unshared valence electrons are located at wildcard atom 4, but the number of unshared electrons of wildcard atoms 1 and 5 is undefined. The constitution of molecular fragment mf is represented by

$Constit(mf) =$
$(\{(\#,1,?), (0,2,H), (\#,3,C), (2,4,?), (\#,5,?)\},$
$\{(2, \{\{1,2\}\}), (2, \{\{1,3\}\}), (2, \{\{1,4\}\}), (2, \{\{1,5\}\})\})$

YET ANOTHER REPRESENTATION OF MOLECULAR STRUCTURE

J. Chem. Inf. Comput. Sci., Vol. 35, No. 5, 1995 **793**

3.3. Formal Atomic Charges and Bond Orders. In this section we shall describe how to calculate formal atomic charges and formal bond orders for chemical structures which are modeled by molecular multigraphs as defined above.

Formal Atomic Charges. Let $v_s(x)$ be the number of valence electrons belonging to a (nonwildcard) atom x within a chemical structure s and let $v(x)$ be the number of valence electrons of the isolated atom x. The formal charge $c_s(x)$ of x in s is then given by

$$c_s(x) = v(x) - v_s(x)$$

Calculation of $v_s(x)$ is possible only if s meets the following requirements: $u_s(x)$, the number of unshared electrons of x in s, and all bonding systems in which x participates must be specified. Note that a molecular fragment does not necessarily meet these conditions.

Let B_x be the subset of bonding systems belonging to chemical structure s and involving atom x. Choose $b \in B_x$ and let $n(b)$ be the cardinality of the electron set belonging to b; let furthermore $p(b)$ be the total number of atom pair sets involved in b and $p_x(b)$ the number of those atom pair sets in b which contain x. Then, b contributes $v_s^b(x) = (n(b) \times p_x(b))/(2 \times p(b))$ electrons to $v_s(x)$. The latter is given by the equation

$$v_s(x) = u_s(x) + \sum_{b \in B_x} v_s^b(x)$$

For example, consider the ferrocene molecule *fc* (Figure 5). Carbon atom 5 is involved in five bonding systems: two bonding systems which correspond to the σ parts of two ring bonds and a bonding system which corresponds to a covalent C–H single bond as well as bonding systems *i* and *iii* (Figure 5) which are modeling coordination bonds.

In each of the former three bonding systems, two electrons are responsible for a single bonding relation between carbon atom 5 and a carbon or a hydrogen atom; therefore each of them contributes $(2 \times 1)/(2 \times 1) = 1$ electron to the valence electrons of atom 5 in *fc*.

In bonding system *i*, we have a set of six electrons that is responsible for ten bonding relations, and carbon atom 5 is involved in three of them. Therefore, bonding system *i* contributes $(6 \times 3)/(2 \times 10) = 9/10$ electrons to the valence electrons of carbon atom 5.

The cardinality of the electron set belonging to bonding system *iii* is also six, but carbon atom 5 is involved in only one of the ten bonding relations. Thus, bonding system *iii* contributes $(6 \times 1)/(2 \times 10) = 3/10$ electrons.

Carbon atom 5 has no unshared electrons. Altogether this makes $4^1/_5$ valence electrons for carbon atom 5 in *fc*. An isolated carbon atom has four valence electrons, and we obtain a negative formal charge of $-^1/_5$.

Formal Bond Orders. In order to calculate the formal order of a bond between two atoms, we have to take into account all bonding relations between them.

Let $B_{\{x,y\}}$ be the set of bonding systems in which x and y are neighbor atoms. Furthermore, let $p(b)$ denote the total number of neighbor atom pair sets involved in bonding system b and let $n(b)$ be the cardinality of the electron set belonging to b. Then, the formal order of the bond between atoms x and y is

$$\frac{1}{2} \sum_{b \in B_{\{x,y\}}} \frac{n(b)}{p(b)}$$

For example, the formal order of the bond between atoms 1 and 3 in the homotropylium ion h (Figure 2) is $^1/_2 \times ^6/_7 = ^3/_7$ whereas the bond between atoms 1 and 8 has formal order $^1/_2 \times (^6/_7 + ^2/_1) = 1^3/_7$.

4. STEREOCHEMISTRY: CONFIGURATION AND CONFORMATION

We now turn our attention to the configuration and conformation of chemical structures. There exist different definitions for these concepts,[1,17] so we shall first give informal definitions which we shall use throughout the text. These definitions are based on definitions E-1.4.(c) and E-1.7.(c) of the IUPAC recommendations.[1] Then, we shall briefly discuss the drawbacks of existing computer-oriented representations of configuration and conformation. In order to avoid these problems, we shall formally define configuration and conformation of a chemical structure in terms of the relative spatial arrangement of its atoms. One flexible formalism will allow us to uniquely and unambiguously represent arbitrary relative spatial arrangements of atoms.

4.1. Informal Definitions. Stereochemistry deals with those properties of chemical structures that are related to the (relative) arrangement of their atoms in three-dimensional Euclidean space. Accordingly, configuration and conformation should be defined exclusively in (approximate-) geometric terms. Energy barriers separating from each other different (relative) spatial arrangements of atoms should play no role in these definitions. (See also Appendix 1 on p 486 of the IUPAC recommendations.[1])

For example, there is no reason to consider enantiomeric amines as conformationally different while regarding enantiomeric phosphines as configurationally different. Analogous geometrical operations "invert" the stereocenters of both enantiomeric amines and phosphines.

Although enantiomeric phosphines may be separated from each other by a higher energy barrier than enantiomeric amines, the former will also interconvert spontaneously under appropriate conditions. If the definitions of configuration and conformation would depend on the surmountability of energy barriers, we would consequently have to consider the enantiomeric phosphines sometimes as conformationally different and sometimes as configurationally different, depending on the observation conditions.

For our purposes, only a definition of configuration and conformation that is independent from observation conditions is acceptable. If different configurations or different conformations interconvert spontaneously under certain observation conditions, then we call them unstable under these conditions, stable otherwise.

Configuration. Two chemical structures with the same constitution differ with respect to their configuration, if they possess different relative spatial arrangements of the atoms that may not be interconverted by means of rotations about bonds.

For example, interconversion of the relative spatial arrangements of the atoms of the pentafluorophosphorane molecules m and n (Figure 9) requires a distortion of $\angle$FPF bond angles. Thus, m and n have different configurations.
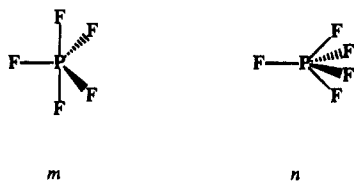
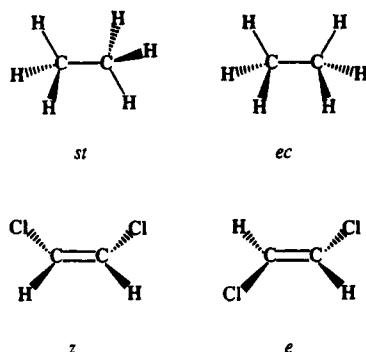**Figure 9.** Configurationally different pentafluorophosphorane molecules.

Figure 10. Pairs of conformationally different molecules.

Conformation. Two chemical structures with the same configuration differ with respect to their conformation, if they possess different relative spatial arrangements of their atoms.

For a given configuration, different conformations may be interconverted merely by means of rotations about bonds. Remember that it does not matter if the required rotations really occur under certain observation conditions. For example, the staggered and eclipsed ethane molecules *st* and *ec* (Figure 10) have different conformations. They interconvert spontaneously at ambient temperature because of the low energy barrier of the rotation about the C−C bond. The molecules *z* and *e* of *Z*- and *E*-1,2-dichloroethene also differ conformationally. But in this case no rotation about the C=C double bond is observed at ambient temperature because the corresponding energy barrier is too high.

4.2. Traditional Representation of Stereochemistry. The existing nongeometric representations of stereochemistry fall within two categories: (i) In the nomenclature-oriented approach, stereogenic units of the molecular graph are labeled with stereodescriptors, such as *R/S*, *Z/E*, cis/trans, exo/endo, etc.[18-20] and (ii) in the template-based approach, configurational and conformational "reference templates" (spatial arrangements of bonding positions) are associated with stereogenic units, and substituents are distributed on them.[21-26]

Ugi emphasizes that chemical structures are nonrigid.[27] Most template-based approaches, however, interpret configurational and conformational templates as rigid geometric figures. In the case of configurational representation, substituents are distributed on the corners of regular polyhedra;[21,22] grids corresponding to Newman projections serve as templates for conformational representation.[24,26]

An obvious drawback of the nomenclature-oriented approach is that one has to live with all the limitations of the nomenclature system that is used. For example, the CIP nomenclature system[28,29] has rather been devised for assigning stereodescriptors to stereogenic units within molecules than within molecular fragments. Generally, one cannot assign CIP stereodescriptors to stereogenic units within molecular fragments, because the CIP priorities of missing substituents are unknown, and it is not always possible to
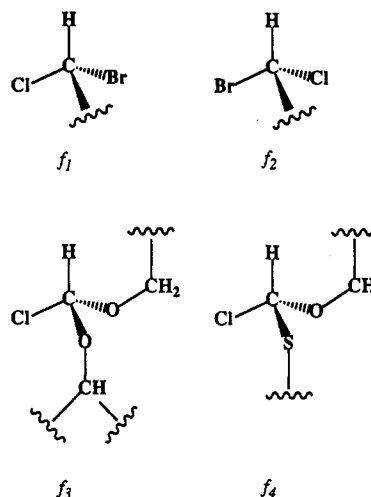

Figure 11. A CIP stereodescriptor for the carbon atom may only be determined in the case of molecular fragment f_4.

determine the CIP priorities of incompletely specified substituents.

Let us consider Figure 11. Clearly, molecular fragments f_1 and f_2 differ configurationally. However, we do not know the nature of the missing substituents and thus cannot determine CIP stereodescriptors in order to express this difference.

Arbitrarily giving the lowest CIP priority to unspecified substituents does not constitute a solution to this problem: The stereodescriptors which could be assigned making this assumption would allow, for example, for the distinguishing of molecular fragments f_1 and f_2, but they would not represent the "absolute configuration" of the carbon atoms in these fragments. Therefore, these stereodescriptors could not be used, for example, in substructure search.

Both fragments f_3 and f_4 contain a carbon atom having two incompletely specified substituents. Only in the case of fragment f_4, the known parts of the latter allow the assignment of CIP priorities.

Furthermore, nomenclature systems tend to be incomplete. If there is no stereodescriptor available for a certain kind of stereogenic unit, then one cannot completely specify the stereochemistry of the corresponding chemical structure.

Whether or not one interprets configurational and conformational templates as rigid geometric figures, the template-based approach to stereochemistry has the following shortcomings: Distributing substituents on a reference template does not result in a unique description of the corresponding stereogenic unit. For example, there are 12 different but equivalent ways to distribute four substituents on a tetrahedral template; they all represent the same atomic configuration as they correspond to 12 different spatial orientations of the same stereogenic unit. Because stereochemical features of a chemical structure are not perspective dependent, its orientation in space should not affect its structural representation. Another drawback is that the set of configurational and conformational templates has to be extended every time that a new type of stereogenic unit is encountered. Note that the same kind of problem occurs with bond-type-based representations of chemical constitution.

In our approach, we shall avoid the problems mentioned above by really describing the relative spatial arrangement of the atoms in a chemical structure. Thus, we shall not be limited to a predefined set of configurational and confor-
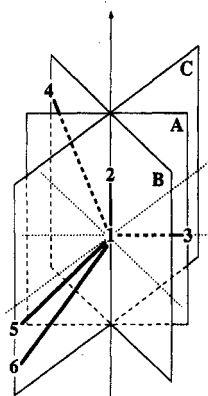
Figure 12. The configurational paddle wheel of atom 1 that is associated with neighbor atom 2.

mational templates or nomenclature-based stereodescriptors. We will be able to describe any conceivable configuration and conformation. Furthermore, for a chemical structure of given constitution, our representation of stereochemistry will be unique and unambiguous.

The reader should recall what we have said in the context of chemical constitution about the representation of chemical structures (p 792): "Note that a molecular structure representation cannot free the user from the task to decide *how* a chemical structure should be represented." Of course, this is also true in the context of stereochemistry. It means that the user has to decide if he needs to specify a certain configuration or conformation of a chemical structure. He may decide that configurational or conformational features are not relevant to his problem or he may not specify a configuration or conformation because it is unstable under the given observation conditions.

In the following, we shall give representations of configurations and conformations which may be unstable under certain observation conditions (for example, the amines in Figure 15). These representations only serve illustration purposes, and we do not mean their specification to be mandatory.

4.3. Formal Definition and Representation of Configuration. In the following, atoms are considered as points in three-dimensional Euclidean space.

(Non-)terminal Atom. An atom is called terminal if it has at most one neighbor atom, otherwise it is called nonterminal.

Configurational Paddle Wheel. Let Ax be the axis which is directed from nonterminal atom x toward its neighbor atom y. The configurational paddle wheel of x that is associated with y is the relative spatial arrangement of x and its neighbor atoms on the pencil of planes which intersect along Ax and contain x and its neighbor atoms.

Configuration of a Nonterminal Atom. The atomic configuration of nonterminal atom x is the set of its configurational paddle wheels.

Configuration of a Chemical Structure. The configuration of a chemical structure is the set of configurations of its nonterminal atoms.

We shall now formalize the concept of configurational paddle wheel, that is, we shall describe a procedure which generates a configurational paddle wheel from an arbitrary geometrically specified three-dimensional arrangement of an atom and its neighbor atoms. For example, consider the

unusual three-dimensional arrangement of pentacoordinate atom 1 and its neighbor atoms in Figure 12.

In order to generate the configurational paddle wheel that is associated with neighbor atom y of nonterminal atom x, we first must establish the pencil of planes which contain x and its neighbor atoms and which intersect along axis Ax directed from x toward y.

Ax subdivides each plane in two half planes. (Note that Ax belongs to neither half plane.) Let n be the cardinality of the pencil of planes. Let furthermore H_1 and H_2 denote two different half planes. We define a function *cdiff* such that $cdiff(H_1,H_2) = i$ modulo $2n$, if we encounter H_2 after i steps, when we look in the direction opposite to that of Ax and if we proceed clockwise from half plane H_1. (Read: "The clockwise difference from H_1 to H_2 is i modulo $2n$".) The clockwise difference is not symmetric. Only if H_1 and H_2 belong to the same plane, we have $cdiff(H_1,H_2) = cdiff(H_2,H_1) = n$. Let $n \geq 3$ be the cardinality of the pencil of planes and let H, H_1, and H_2 be three pairwise different half planes. If either $cdiff(H,H_1) < n$ and $cdiff(H,H_2) < n$ or $cdiff(H,H_1) > n$ and $cdiff(H,H_2) > n$, then H_1 and H_2 are lying in the same half space with respect to plane P to which H belongs. (Note that P belongs to neither half space.) If $cdiff(H,H_1) < n$ and $cdiff(H,H_2) > n$, then H_1 and H_2 lie in different half spaces with respect to P.

The relative spatial arrangement of atom x and its neighbor atoms on the pencil of planes can be expressed in terms of the clockwise arrangement of the corresponding half planes around Ax and the relative spatial arrangement of the atoms within each half plane.

The relative spatial arrangement of the atoms within a half plane corresponds to a sequential order of these atoms according to the following scheme (remember that Ax belongs to neither half plane): A plane intersecting Ax perpendicularly in atom x subdivides each half plane into an intersection line (dotted lines in Figure 12), a region below the intersection line and a region above the intersection line. (Ax is oriented from "below" the plane to "above" the plane.) We now order the atoms within the lower and upper region of a half plane. (On the intersection line, there is at most one atom, because the angle $\angle uxv = 0° \leftrightarrow u = v$ for any two neighbor atoms u and v of atom x.) Let N be the set of neighbor atoms of x lying in the lower (upper) region of a half plane. Then, a neighbor atom $m \in N$ precedes another neighbor atom $n \in N$ within this region if the angle $\angle mxy$ is bigger than the angle $\angle nxy$. Finally, the atoms in the lower part of a half plane precede the atom on the intersection line which in turn precedes the atoms in the upper region of that half plane.

Note that, although Ax does not belong to any of the half planes of a configurational paddle wheel, we do not lose information about the relative spatial arrangement of x and its neighbor atoms on the pencil of planes intersecting along Ax, because (i) we know that atom x is situated at the intersection of Ax with the perpendicular plane; (ii) we also know that atom y is lying in the upper part of Ax, because Ax is oriented from x toward y and Ax points from "below" the perpendicular plane to "above"; (iii) besides atom y, atom x can have at most a second neighbor atom z lying on Ax, and if it has, then z must be situated in the lower part of Ax. This is because y already occupies the upper part of Ax and because $\angle uxv = 0° \leftrightarrow u = v$ for any two neighbor atoms u

and v of atom x. The presence of a neighbor atom z on the lower part of Ax is implied by the absence of z from all half planes.

Figure 12 shows the configurational paddle wheel that is associated with neighbor atom 2 of nonterminal atom 1. Its axis Ax is directed from atom 1 to atom 2. The pencil of planes which intersect along Ax and contain 1 and its neighbor atoms comprises three planes: plane A contains atoms 1, 2, and 3; plane B contains atoms 1, 2, and 4; and plane C contains atoms 1, 2, 5, and 6.

Ax subdivides each of these three planes in an empty and a nonempty half plane. Let P_i denote a half plane, where $P \in \{A,B,C\}$ indicates the corresponding plane, and the index $i \in \{e,ne\}$ specifies if the half plane is empty or nonempty. If we look in the direction opposite to that of Ax, we obtain the following clockwise arrangement of half planes: A_{ne} precedes B_e; the latter is followed by C_{ne}, which precedes A_e; that half plane precedes B_{ne} which precedes C_e; finally, the latter is followed by A_{ne}. Note that $\forall P \in \{A,B,C\}$: $cdiff(P_e,P_{ne}) = cdiff(P_{ne},P_e) = 3$. Furthermore, $cdiff(A_{ne},B_e) < 3$ and $cdiff(A_{ne},C_{ne}) < 3$, that is, B_e and C_{ne} lie in the same half space with respect to plane A.

A_{ne} contains atom 3 on the intersection line, but no atoms in either the lower or the upper region. In B_{ne}, no atom is lying on the intersection line; the order of atoms in the upper and lower region, respectively, is trivial, because the lower region is empty and the upper region contains only atom 4. C_{ne} contains atoms 5 and 6 which are both lying in the lower region; atom 6 precedes atom 5 because $\angle 612 > \angle 512$.

A configurational paddle wheel is represented by a pair whose first element is the neighbor atom with which the configurational paddle wheel is associated; the second element is a *cyclic* list of half planes which represents the clockwise arrangement of half planes around the axis of the configurational paddle wheel. Note that, for example, the four lists (w,x,y,z), (x,y,z,w), (y,z,w,x), and (z,w,x,y) are equivalent notations for the same cyclic list, in which element w precedes x which is followed by y; y precedes z which in turn is followed by w. The representation of a half plane is a triple: The first element is a (possibly empty) list of the ordered atoms in the lower region. If there is an atom on the intersection line, then this atom is the second element of the triple; if not, then the second element is ϵ, the "empty atom". The third element of the triple is a (possibly empty) list containing the ordered atoms of the upper region of the half plane. For convenience, $((), \epsilon, ())$ will be abbreviated as $\mathcal{E}$.

Thus, the configurational paddle wheel which is depicted in Figure 12 is represented by

$$(2, (((), 3, ()), \mathcal{E}, ((6,5), \epsilon, ()), \mathcal{E}, ((), \epsilon, (4)), \mathcal{E}))$$

Let us summarize the characteristics of the configurational paddle wheel that is associated with neighbor atom y of a nonterminal atom x:

1. Direction of the axis Ax from x to y.
2. Pencil of planes intersecting along Ax and containing x and its neighbor atoms.
3. Arrangement of the corresponding half planes relative to each other.
4. Relative spatial arrangement of x and its neighbor atoms within each half plane: below/on/above the intersection line of the half plane with a plane intersecting Ax perpendicularly in atom x; order of atoms within the lower/
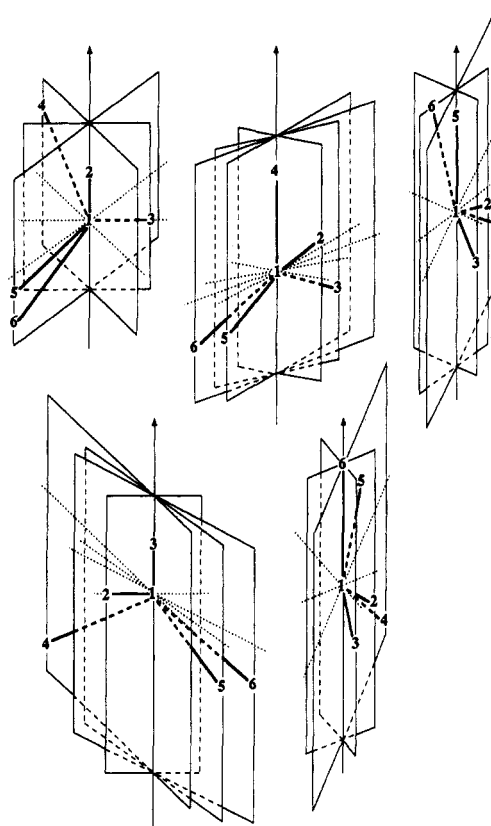

Figure 13. The configurational paddle wheels belonging to the configuration of atom 1.

upper region according to increasing size of the angle $\angle zxy$ (z is a neighbor atom in the lower/upper region).

A configurational paddle wheel imposes constraints onto the possible three-dimensional atomic coordinates without explicitly specifying exact bond angles and lengths. These are rather allowed to vary within certain limits.

The atomic configuration, that is, the relative spatial arrangement of an atom and its neighbor atoms, is simply a superposition of all configurational paddle wheels belonging to that atom. This amounts to a combination of constraints that further limits the range of atomic coordinates that may be adopted by these atoms. Bond angle ranges and bond length ranges are interdependent. Unlike the latter, the former do not depend on the scale of the approximate-geometrical model of a chemical structure.

For example, the configuration of atom 1 is the superposition of the configurational paddle wheels depicted in Figure 13.

An atomic configuration is represented by a pair consisting of the atom whose configuration is to be described and the set of its configurational paddle wheels.

Let *AtConfig* be a function generating the representation of an atomic configuration. The following is the representation of the configuration of atom 1 in Figure 13.

$AtConfig(1) =$
$(1, \{(2, (((), 3, ()), \mathcal{E}, ((6,5), \epsilon, ()), \mathcal{E}, ((), \epsilon, (4)), \mathcal{E})),$
$(3, (((), 2, ()), \mathcal{E}, \mathcal{E}, ((4), \epsilon, ()), \mathcal{E}, ((6), \epsilon, ()), ((5), \epsilon, ()), \mathcal{E})),$
$(4, (((), \epsilon, (2)), ((5), \epsilon, ()), ((6), \epsilon, ()), \mathcal{E}, \mathcal{E}, \mathcal{E}, \mathcal{E}, ((3), \epsilon, ()))),$
$(5, (((2), \epsilon, ()), ((3), \epsilon, ()), \mathcal{E}, ((), \epsilon, (6)), \mathcal{E}, ((4), \epsilon, ()))),$
$(6, (((2), \epsilon, (5)), ((3), \epsilon, ()), \mathcal{E}, \mathcal{E}, \mathcal{E}, ((4), \epsilon, ())))\})$

We shall now illustrate with a few examples that our formal definition of configuration has been adequately

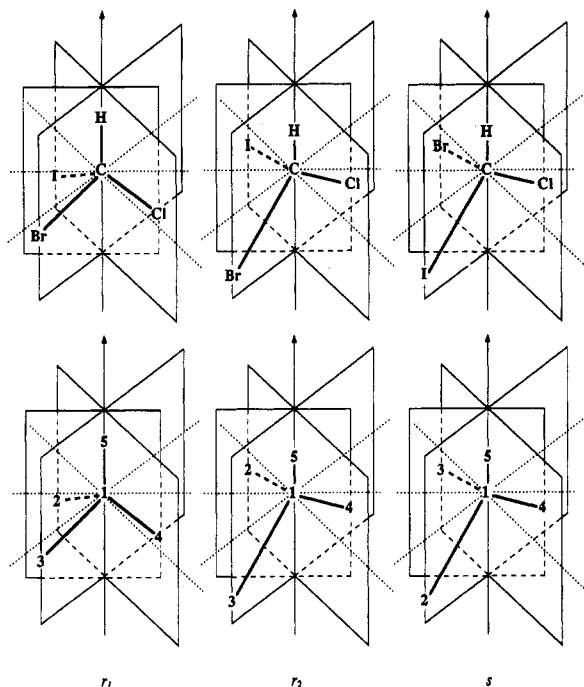YET ANOTHER REPRESENTATION OF MOLECULAR STRUCTURE

J. Chem. Inf. Comput. Sci., Vol. 35, No. 5, 1995 **797**

Figure 14. R and S configurations of an asymmetric carbon atom.

Figure 15. Spontaneously interconverting amine configurations.

Figure 16. Staggered and eclipsed configurations of ferrocene.

chosen. Arbitrary relative spatial arrangements of an atom and its neighbor atoms may be described and distinguished from each other. Chemically relevant examples range from the simple R- and S-configurations of asymmetric carbon atoms to the complex atomic configurations of the iron atom in staggered and eclipsed ferrocene molecules.

There are different three-dimensional arrangements of an atom and its neighbor atoms which should nevertheless be considered to have the same *relative* spatial arrangement, that is, the same atomic configuration.

For example, the three-dimensional arrangements r_1 and r_2 of a carbon atom and its four neighbor atoms in Figure 14 should be considered to have the same configuration, because both are molecules of (R)-bromochloroiodomethane. As can easily be seen, the configurational paddle wheel of the carbon atom which is associated with the hydrogen atom is the same for r_1 and r_2. The same is true for the remaining three configurational paddle wheels (not depicted in Figure 14).

Consequently, the superposition of the configurational paddle wheels is the same, and the carbon atoms in r_1 and r_2 have the same atomic configuration, represented by

$AtConfig(1) =$
$(1, \{(2, (((3), \epsilon, ()), \mathcal{E}, ((4), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E})),$
$(3, (((2), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E}, ((4), \epsilon, ()), \mathcal{E})),$
$(4, (((2), \epsilon, ()), \mathcal{E}, ((3), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E})),$
$(5, (((2), \epsilon, ()), \mathcal{E}, ((4), \epsilon, ()), \mathcal{E}, ((3), \epsilon, ()), \mathcal{E}))\})$

On the other hand, we must be able to distinguish molecule $r_{1,2}$ from molecule s (Figure 14), because the latter belongs to (S)-bromochloroiodomethane. Figure 14 illustrates the difference between the configurational paddle wheel of the carbon atom which is associated with the hydrogen atom in molecule $r_{1,2}$ and the corresponding configurational paddle wheel in molecule s. Molecule $r_{1,2}$ also differs from molecule s with respect to the remaining configurational paddle wheels of the carbon atom which are associated with Br, Cl, and I. Thus, the atomic configuration of the carbon
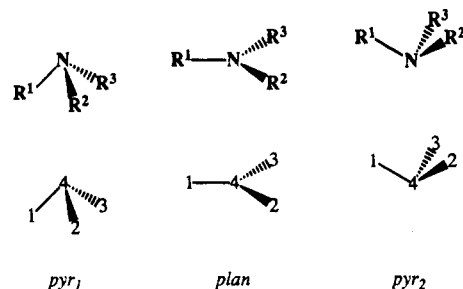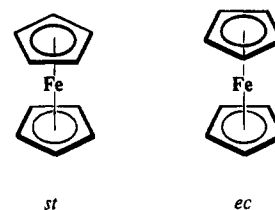
atom in $r_{1,2}$ is different from that of the carbon atom in s. The latter configuration is represented by

$AtConfig(1) =$
$(1, \{(3, (((2), \epsilon, ()), \mathcal{E}, ((4), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E})),$
$(2, (((3), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E}, ((4), \epsilon, ()), \mathcal{E})),$
$(4, (((3), \epsilon, ()), \mathcal{E}, ((2), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E})),$
$(5, (((3), \epsilon, ()), \mathcal{E}, ((4), \epsilon, ()), \mathcal{E}, ((2), \epsilon, ()), \mathcal{E}))\})$

Tertiary amines are configurationally flexible; the pyramidal configurations pyr_1 and pyr_2 spontaneously interconvert via the planar configuration *plan* (Figure 15). Just like the configurations of the carbon atom in molecule $r_{1,2}$ and molecule s (Figure 14), the configurations of the nitrogen atom in pyr_1 and pyr_2 differ by an odd permutation of neighbor atoms. In contrast to this, the nitrogen configuration in *plan* is "completely different" from those in pyr_1 and pyr_2, that is, the structures of the configurational paddle wheels are different. This can easily be seen from the following representations:

$AtConfig(4_{pyr_1}) =$
$(4, \{(1, (((2), \epsilon, ()), ((3), \epsilon, ()), \mathcal{E}, \mathcal{E})),$
$(2, (((1), \epsilon, ()), \mathcal{E}, \mathcal{E}, ((3), \epsilon, ()))),$
$(3, (((1), \epsilon, ()), ((2), \epsilon, ()), \mathcal{E}, \mathcal{E}))\})$

$AtConfig(4_{plan}) =$
$(4, \{(1, (((2), \epsilon, ()), ((3), \epsilon, ()))),$
$(2, (((1), \epsilon, ()), ((3), \epsilon, ()))),$
$(3, (((1), \epsilon, ()), ((2), \epsilon, ())))\})$

$AtConfig(4_{pyr_2}) =$
$(4, \{(1, (((2), \epsilon, ()), \mathcal{E}, \mathcal{E}, ((3), \epsilon, ()))),$
$(3, (((1), \epsilon, ()), \mathcal{E}, \mathcal{E}, ((2), \epsilon, ()))),$
$(2, (((1), \epsilon, ()), ((3), \epsilon, ()), \mathcal{E}, \mathcal{E}))\})$

The staggered and eclipsed configurations *st* and *ec* of ferrocene (Figure 16) provide examples of configurations of a highly coordinate atom. Note that chemists generally speak of staggered and eclipsed *conformations* of ferrocene; according to our informal definitions of configuration and conformation, *st* and *ec* differ *configurationally*, because the relative spatial arrangement of the atoms in molecules *st* and *ec* cannot be interconverted by rotation about bonds. We consider a bond as a binary relationship between two atoms. Thus, rotation of the ligands about the vertical lines con-
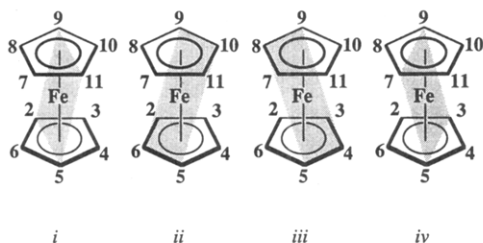
Figure 17. Planes belonging to the configurational paddle wheels associated with atoms 5 and 9 in a staggered ferrocene molecule.

necting them with the iron atom is no rotation about bonds, because each of these lines is a shorthand notation for five Fe−C bonds.

Let us first consider the staggered ferrocene molecule *st*. Figure 17 shows the planes belonging to the configurational paddle wheels that are associated with carbon atoms 5 and 9. Note that the planes belonging to the remaining configurational paddle wheels may be obtained by application of atom permutations which correspond to rotations of the molecule as a whole. (The spatial orientation of the planes remains the same.)

A permutation of a set S is an automorphism of S, that is, a bijective mapping from S to S. Let $\{s_1, s_2, ..., s_s\} \subseteq S$. Then, $\lambda = (s_1 \; s_2 \; ... \; s_s)$ denotes the cyclic permutation

$$\lambda : \quad S \to S;$$
$$s_i \mapsto s_{i+1} \quad (s_i \in \{s_1, s_2, \ldots, s_{s-1}\}),$$
$$s_s \mapsto s_1,$$
$$s_i \mapsto s_i \quad (s_i \in S \setminus \{s_1, s_2, \ldots, s_s\}).$$

Every permutation can be written as a combination of cyclic permutations.

For example, the planes belonging to the configurational paddle wheels that are associated with carbon atoms 4 and 8 can be obtained by applying the permutation (5 4 3 2 6) (11 10 9 8 7) to *i*, *ii*, *iii*, and *iv* (Figure 17).

The configuration of the iron atom 1 in molecule *st* is represented by

$AtConfig(1) =$
$(1, \{(5, (((), \epsilon, (2)), ((), \epsilon, (6)), ((8), \epsilon, ()), ((7), \epsilon, ()),$
$\quad ((11), \epsilon, ()), ((10), \epsilon, ()), ((), \epsilon, (4)), ((), \epsilon, (3)))),$
$(4, (((), \epsilon, (6)), ((), \epsilon, (5)), ((7), \epsilon, ()), ((11), \epsilon, ()),$
$\quad ((10), \epsilon, ()), ((9), \epsilon, ()), ((), \epsilon, (3)), ((), \epsilon, (2)))),$
$(3, (((), \epsilon, (5)), ((), \epsilon, (4)), ((11), \epsilon, ()), ((10), \epsilon, ()),$
$\quad ((9), \epsilon, ()), ((8), \epsilon, ()), ((), \epsilon, (2)), ((), \epsilon, (6)))),$
$(2, (((), \epsilon, (4)), ((), \epsilon, (3)), ((10), \epsilon, ()), ((9), \epsilon, ()),$
$\quad ((8), \epsilon, ()), ((7), \epsilon, ()), ((), \epsilon, (6)), ((), \epsilon, (5)))),$
$(6, (((), \epsilon, (3)), ((), \epsilon, (2)), ((9), \epsilon, ()), ((8), \epsilon, ()),$
$\quad ((7), \epsilon, ()), ((11), \epsilon, ()), ((), \epsilon, (5)), ((), \epsilon, (4)))),$
$(9, (((2), \epsilon, ()), ((3), \epsilon, ()), ((4), \epsilon, ()), ((), \epsilon, (10)),$
$\quad ((), \epsilon, (11)), ((), \epsilon, (7)), ((), \epsilon, (8)), ((6), \epsilon, ()))),$
$(10, (((3), \epsilon, ()), ((4), \epsilon, ()), ((5), \epsilon, ()), ((), \epsilon, (11)),$
$\quad ((), \epsilon, (7)), ((), \epsilon, (8)), ((), \epsilon, (9)), ((2), \epsilon, ()))),$
$(11, (((4), \epsilon, ()), ((5), \epsilon, ()), ((6), \epsilon, ()), ((), \epsilon, (7)),$
$\quad ((), \epsilon, (8)), ((), \epsilon, (9)), ((), \epsilon, (10)), ((3), \epsilon, ()))),$
$(7, (((5), \epsilon, ()), ((6), \epsilon, ()), ((2), \epsilon, ()), ((), \epsilon, (8)),$
$\quad ((), \epsilon, (9)), ((), \epsilon, (10)), ((), \epsilon, (11)), ((4), \epsilon, ()))),$
$(8, (((6), \epsilon, ()), ((2), \epsilon, ()), ((3), \epsilon, ()), ((), \epsilon, (9)),$
$\quad ((), \epsilon, (10)), ((), \epsilon, (11)), ((), \epsilon, (7)), ((5), \epsilon, ()))))\})$

In the eclipsed configuration of ferrocene, each configurational paddle wheel consists of nine planes. Figure 18 shows the planes belonging to the configurational paddle wheel associated with the carbon atom 5. (Plane *ix* is vertical
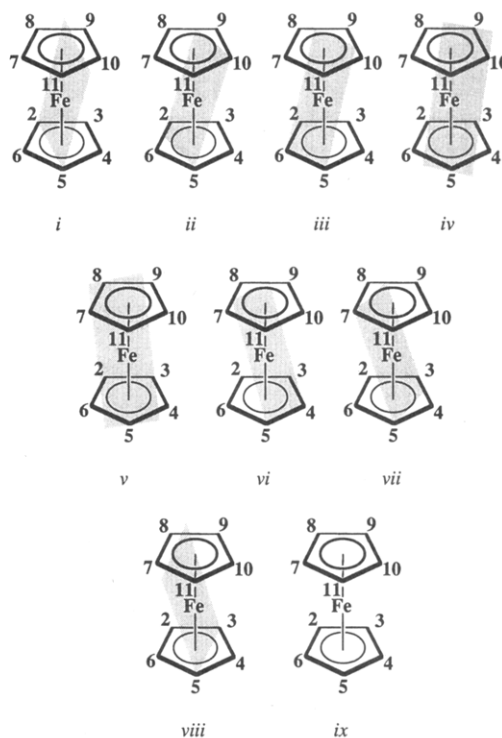

Figure 18. Planes belonging to the configurational paddle wheel associated with atom 5 in an eclipsed ferrocene molecule.

to the paper plane and passes through the iron atom and the carbon atoms 5 and 11.)

$$(5, (((), \epsilon, (2)), \mathcal{E}, ((), \epsilon, (6)), \mathcal{E}, ((8), \epsilon, ()), \mathcal{E},$$
$$((7), \epsilon, ()), \mathcal{E}, ((11), \epsilon, ()), \mathcal{E}, ((10), \epsilon, ()), \mathcal{E},$$
$$((9), \epsilon, ()), \mathcal{E}, ((), \epsilon, (4)), \mathcal{E}, ((), \epsilon, (3)), \mathcal{E}))$$

represents this configurational paddle wheel. Application of the permutation (5 4 3 2 6)(11 10 9 8 7) to this configurational paddle wheel yields the one that is associated with atom 4; further applications of this permutation yield the configurational paddle wheels associated with atoms 3, 2, and 6, respectively.

Application of the permutation (5 11)(6 10)(4 7)(2 9) (3 8) to the representation of the configurational paddle wheel associated with atom 5 yields

$$(11, (((), \epsilon, (9)), \mathcal{E}, ((), \epsilon, (10)), \mathcal{E}, ((3), \epsilon, ()), \mathcal{E},$$
$$((4), \epsilon, ()), \mathcal{E}, ((5), \epsilon, ()), \mathcal{E}, ((6), \epsilon, ()), \mathcal{E},$$
$$((2), \epsilon, ()), \mathcal{E}, ((), \epsilon, (7)), \mathcal{E}, ((), \epsilon, (8)), \mathcal{E}))$$

This represents the configurational paddle wheel associated with atom 11, from which the configurational paddle wheels associated with atoms 7, 8, 9, and 10 can be obtained by one, two, three, and four applications of the permutation (11 7 8 9 10)(5 6 2 3 4), respectively.

4.4. Formal Definition and Representation of Conformation. Directed Bond. A directed bond from atom x to atom y is a bond between x and y which is labeled with the direction from x to y.

Directed Bond Sequence. A directed bond sequence from atom x to atom y is either a directed bond from x to y or the concatenation of a directed bond from atom x to atom z and a directed bond sequence from atom z to atom y.

γ Sequence. Let A be the set of atoms that we encounter when traversing the directed bond sequence from atom x to atom y. This directed bond sequence is called a γ sequence,
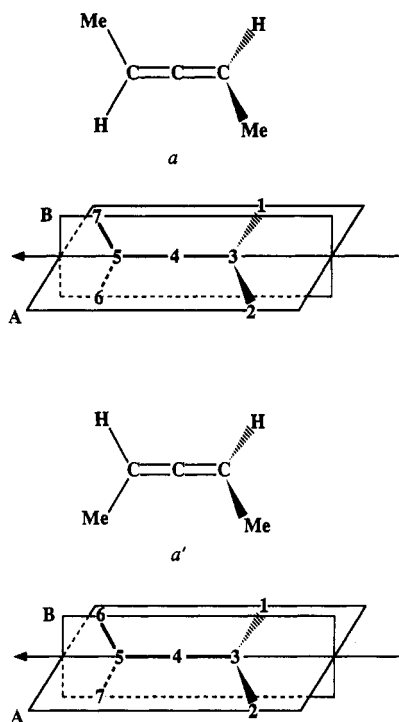
Figure 19. Enantiomeric dimethylallene conformations.

if each of the atoms in $A\backslash\{x,y\}$ has exactly two neighbor atoms and if all directed bonds in the sequence are collinear.

Maximal γ Sequence. Let $B(X)$ denote the set of directed bonds that are part of directed bond sequence X. A γ sequence S is called maximal, if there exists no other γ sequence S' such that $B(S) \subset B(S')$.

Nonterminal Maximal γ Sequence. A maximal γ sequence from atom x to atom y is called nonterminal if neither x nor y is a terminal atom.

Conformational Paddle Wheel. Let S be a nonterminal maximal γ sequence from atom x to atom y; let furthermore Ax be the axis directed from x toward y and let N be the set containing those neighbor atoms of x and y which are not lying on Ax. The conformational paddle wheel that is associated with S is the relative spatial arrangement of the atoms of N on the pencil of planes which intersect along Ax and which contain the atoms of N.

Note that the neighbor atoms of x and y that are lying on Ax are conformationally irrelevant with respect to this axis, because their positions are invariant under rotation about Ax.

Conformation of a Chemical Structure. The conformation of a chemical structure is the set of its conformational paddle wheels.

We shall now formalize the concept of conformational paddle wheel, that is, we shall describe a procedure which generates the conformational paddle wheel associated with a nonterminal maximal γ sequence in an arbitrary geometrically specified chemical structure. For example, consider the conformational paddle wheels of the enantiomeric dimethylallene molecules a and a' depicted in Figure 19. (Methyl groups are considered as atomic units.)

The first step in generating the conformational paddle wheel associated with the nonterminal maximal γ sequence S from atom x to atom y is to establish the pencil of planes which intersect along the axis Ax (directed from x to y) and which contain the atoms of N (the set of neighbor atoms of x and y that are not lying on Ax).

Just like in the context of atomic configuration, Ax subdivides each plane in two half planes. (Remember that Ax belongs to neither half plane.) We may therefore describe the relative spatial arrangement of the atoms of N on the pencil of planes in terms of the clockwise arrangement of the corresponding half planes around Ax and the relative spatial arrangement of the atoms within each half plane.

Let us now define the relative spatial arrangement of the neighbor atoms of x and y within an individual half plane. Let X and Y be the sets of neighbor atoms in that half plane belonging to x and y, respectively. (Note that X and Y need not necessarily be disjoint; for example, the oxygen atom of an epoxide is at the same time neighbor atom of two carbon atoms.) The relative spatial arrangement of the atoms of $X \cup Y$ within the half plane corresponds to a sequential order of these atoms according to the following scheme: First, we impose an order onto the atoms within the sets X and Y:

$$\forall m, n \in X (m \neq n) : m \sqsubset n \Leftrightarrow \angle mxy > \angle nxy$$
$$\forall m, n \in Y (m \neq n) : m \sqsubset n \Leftrightarrow \angle myx < \angle nyx$$

The order relation $m \sqsubset n$ reads: "Atom m precedes atom n." Finally, the atoms of X precede those of Y.

Let us consider molecule a in Figure 19. There are two nonterminal maximal γ sequences: $S_{3\bar{5}}$, which is directed from atom 3 to atom 5, and $S_{5\bar{3}}$, which is directed from atom 5 to atom 3. The axis Ax of the conformational paddle wheel that is associated with $S_{3\bar{5}}$ is directed from atom 3 to atom 5. There are four neighbor atoms of 3 and 5 which are not lying on Ax: 1, 2, 6, and 7. Plane A contains atoms 1 and 2; plane B contains atoms 6 and 7.

Let P_i denote a half plane, where $P \in \{A,B\}$ indicates the corresponding plane and the index $i \in \{1,2,6,7\}$ specifies the atom which the half plane contains. If we look in the direction opposite to that of Ax, we obtain the following clockwise arrangement of half planes: A_1 precedes B_7; the latter is followed by A_2, which precedes B_6; that half plane finally precedes A_1. Note that for either $P = A$ and $i = 1$ or $P = B$ and $i = 6$ we have $cdiff(P_i, P_{i+1}) = cdiff(P_{i+1}, P_i) = 2$, the cardinality of the pencil of planes.

As each half plane contains only a single atom, the order corresponding to the relative spatial arrangement of the atoms within an individual half plane is trivial.

A conformational paddle wheel is represented by a pair consisting of an atom pair and a cyclic list of half planes. The atom pair (x,y) represents the nonterminal maximal γ sequence from atom x toward atom y with which the conformational paddle wheel is associated. The cyclic list of half planes represents the clockwise arrangement of half planes around the axis Ax, when looking in the direction opposite to that of Ax. A half plane is represented by a pair which consists of the (possibly empty) list of ordered atoms belonging to X and the (possibly empty) list of ordered atoms belonging to Y. (X and Y are the sets of neighbor atoms of x and y, respectively, in that half plane. Remember that Ax belongs to no half plane.)

The conformational paddle wheel associated with the nonterminal maximal γ sequence from atom 3 to atom 5 in molecule a (Figure 19) is represented by

$$((3,5),(((1),()),((),(7)),((2),()),((),(6))))$$

Let us summarize the characteristics of the conformational paddle wheel that is associated with the nonterminal maximal γ sequence from atom x to atom y:

1. Direction of the axis Ax from x to y.

2. Pencil of planes intersecting along Ax and containing the neighbor atoms of x and y which are not lying on Ax.

3. Arrangement of the corresponding half planes relative to each other.

4. Relative spatial arrangement of the neighbor atoms of x and y within each half plane: the neighbor atoms of x in a half plane are ordered according to decreasing size of the angle $\angle zxy$, whereas the neighbor atoms of y in a half plane are ordered according to increasing size of the angle $\angle zyx$ (z is a neighbor atom of x and y, respectively); the neighbor atoms of x precede those of y.

Just like a configurational paddle wheel, a conformational paddle wheel imposes constraints onto the possible three-dimensional atomic coordinates without explicitly specifying exact bond angles and lengths. Configurational and conformational constraints combine and determine together the range of possible atomic coordinates.

For each nonterminal maximal γ sequence from atom x to atom y there exists an oppositely directed one. The associated conformational paddle wheels convey the same information, because a change to the opposite direction of the axis is coupled with an inversion of the cyclic sequence of half planes around the axis and an inversion of the ordering of the atoms within each half plane.

The conformational paddle wheels associated with the nonterminal maximal γ sequences from atom 3 to atom 5 and from atom 5 to atom 3 in molecule a (Figure 19) are represented by

$$((3,5),(((1),()),((),(7)),((2),()),((),(6))))$$
$$((5,3),(((),(1)),((6),()),((),(2)),((7),())))$$

The conformation of a chemical structure has been defined as the set of conformational paddle wheels, each of which is associated with one of its nonterminal maximal γ sequences. Let *Conform* be a function generating the conformational representation of a chemical structure. Because of the redundant information conveyed by paddle wheels associated with oppositely directed nonterminal maximal γ sequences, only one of them is included in the conformational representation of a chemical structure. Choosing to represent the conformational paddle wheel associated with the bond sequence which is directed from the numerically smaller atom to the numerically greater atom guarantees the uniqueness of the conformational representation of chemical structures.

The conformations of the enantiomeric dimethylallene molecules a and a' (Figure 19) are represented by

$$Conform(a) = \{((3,5),(((1),()),((),(7)),((2),()),((),(6))))\}$$

$$Conform(a') = \{((3,5),(((1),()),((),(6)),((2),()),((),(7))))\}$$

Thus, molecules a and a' may be distinguished.

Further examples of pairs of conformationally different molecules are molecules st and ec of ethane and molecules z and e of 1,2-dichloroethene (Figure 10). The relevant conformational paddle wheels are depicted in Figure 20, and their conformational representations are

$$Conform(st) = \{((1,2),(((3),()),((),(6)),((4),()),$$
$$((),(8)),((5),()),((),(7))))\}$$
$$Conform(ec) = \{((1,2),(((3),(6)),((4),(8)),((5),(7))))\}$$
$$Conform(z) = \{((1,2),(((3),(5)),((4),(6))))\}$$
$$Conform(e) = \{((1,2),(((4),(5)),((3),(6))))\}$$

4.5. Stereochemistry of Rings. Of course, our stereochemical representation is not limited to acyclic chemical structures. In this section we shall see that, for example, boat/chair conformations and cis/trans fused ring systems may be distinguished.

Figure 21 shows the energetically preferred conformations of cyclohexane, the chair and the boat. Our conformational representation enables us to pairwise distinguish the conformations $chair_1$, $chair_2$, and $boat$.

$$Conform(chair_1) =$$
$$\{((1,2),(((),(3)),((7),()),((),(10)),$$
$$((8),()),((),(9)),((6),()))),$$
$$((2,3),(((),(4)),((1),()),((),(11)),$$
$$((10),()),((),(12)),((9),()))),$$
$$((3,4),(((),(5)),((11),()),((),(14)),$$
$$((12),()),((),(13)),((2),()))),$$
$$((4,5),(((),(6)),((3),()),((),(15)),$$
$$((14),()),((),(16)),((13),()))),$$
$$((5,6),(((),(1)),((15),()),((),(18)),$$
$$((),(16)),((17),()),((),(4)))),$$
$$((1,6),(((),(5)),((2),()),((),(17)),$$
$$((8),()),((),(18)),((7),()))))\}$$

$$Conform(chair_2) =$$
$$\{((1,2),(((),(3)),((6),()),((),(10)),$$
$$((7),()),((),(9)),((8),()))),$$
$$((2,3),(((),(4)),((10),()),((),(11)),$$
$$((9),()),((),(12)),((1),()))),$$
$$((3,4),(((),(5)),((2),()),((),(14)),$$
$$((11),()),((),(13)),((12),()))),$$
$$((4,5),(((),(6)),((14),()),((),(15)),$$
$$((13),()),((),(16)),((3),()))),$$
$$((5,6),(((),(1)),((4),()),((),(18)),$$
$$((15),()),((),(17)),((16),()))),$$
$$((1,6),(((),(5)),((8),()),((),(17)),$$
$$((7),()),((),(18)),((2),()))))\}$$

$$Conform(boat) =$$
$$\{((1,2),(((),(3)),((7),()),((),(10)),$$
$$((8),()),((),(9)),((6),()))),$$
$$((2,3),(((1),(4)),((10),(11)),((9),(12)))),$$
$$((3,4),(((),(5)),((2),()),((),(14)),$$
$$((11),()),((),(13)),((12),()))),$$
$$((4,5),(((),(6)),((14),()),((),(15)),$$
$$((13),()),((),(16)),((3),()))),$$
$$((5,6),(((4),(1)),((15),(18)),((16),(17)))),$$
$$((1,6),(((),(5)),((2),()),((),(17)),$$
$$((8),()),((),(18)),((7),()))))\}$$

In order to distinguish *cis*- and *trans*-decaline molecules c and t (Figure 22), let us consider the conformational paddle wheels that are associated with the ring fusion bond.

$$((1,2),(((3),()),((),(8)),((5),()),$$
$$((),(6)),((4),()),((),(7)))) \in Conform(c)$$
$$((1,2),(((3),()),((),(8)),((5),()),$$
$$((),(7)),((4),()),((),(6)))) \in Conform(t)$$

The conformational paddle wheels reflect the cis and trans relations of hydrogen atoms 5 and 6 in molecules c and t, respectively.

Yet Another Representation of Molecular Structure

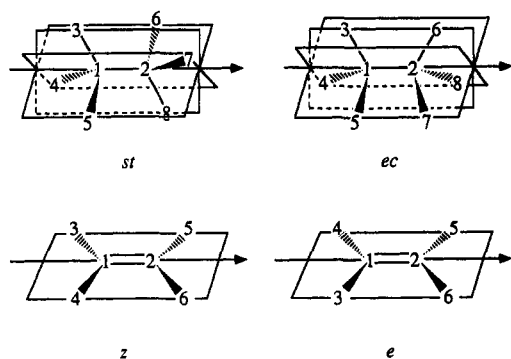*J. Chem. Inf. Comput. Sci., Vol. 35, No. 5, 1995* **801**

Figure 20. Conformational paddle wheels of molecules *st*, *ec*, *z*, and *e* in Figure 10.
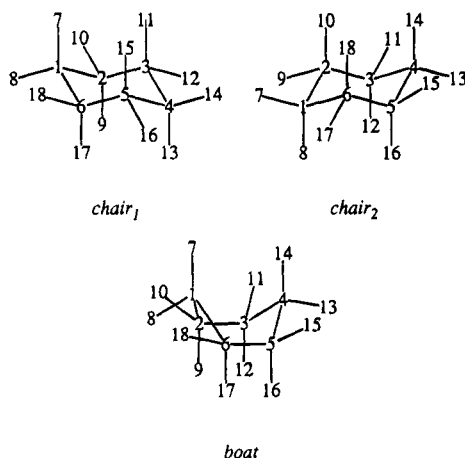

Figure 21. Some conformations of cyclohexane.

In each of the two conformational paddle wheels, the cyclic list of half planes has six elements. Consequently, the cardinality of the pencil of planes is three.

Let us now consider molecule *c*. We abbreviate "the half plane containing atom *x*" as H_x. Half planes H_8 and H_4 belong to the same plane *P*, because $cdiff(H_8,H_4) = cdiff(H_4,H_8) = 3$ (p 795). Furthermore, $cdiff(H_8,H_5) < 3$ and $cdiff(H_8,H_6) < 3$, that is, H_5 and H_6 are lying in the same half space with respect to *P* (p 795). Atoms 5 and 6 are therefore in cis relation. In molecule *t* we have $cdiff(H_8,H_5) < 3$ and $cdiff(H_8,H_6) > 3$. Atoms 5 and 6 are in trans relation, as H_5 and H_6 lie in different half spaces with respect to *P* (p 795).

It is important to note that the difference of the conformational paddle wheels associated with the ring fusion bond is *not* a conformational one; it is the consequence of different configurations of carbon atom 2 in molecules *c* and *t*. Therefore, we can as well distinguish molecules *c* and *t* by their atomic configurations of carbon atom 2.

5. CONCLUSIONS

We propose a computer-oriented representation of molecular structure that comprises constitution and stereochemistry, that is, configuration and conformation. It has been devised for chemical information storage and retrieval systems as well as for computer-assisted synthesis design systems. The drawbacks of existing molecular structure representations are discussed (sections 3.1 and 4.2) and a novel approach which avoids these shortcomings is presented (sections 3.2, 4.3, and 4.4). The goal was to reach the expressiveness of the language of structural formulas whose importance as a graphical representation of molecular structure is indisputable.
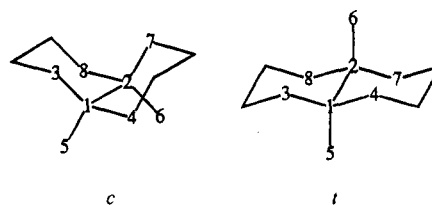

Figure 22. *Cis*- and *trans*-decaline.

Because structural features are perspective independent, our representation does not reflect the orientation of a chemical structure in three-dimensional Euclidean space, in contrast to structural formulas. We exclusively account for the perspective independent features of structural formulas: the valence electron distribution among and between the atoms, the pairwise bonding relations between them, and their relative spatial arrangement. Thus, the representation is unique and unambiguous; it is canonical if the atoms are canonically numbered.[30-32]

We have based our representation of chemical constitution on so-called bonding systems rather than on bonds. The bonding systems of a chemical structure indicate all pairwise bonding relations between its atoms. It does not matter if electron sharing or electrostatic interaction is at their origin (shared electron interaction bonding or unshared electron interaction bonding). A chemical bond between two atoms is the combined effect of all bonding relations between them. The bonding system concept allows us to uniformly represent any conceivable bond type in the fields of organic and organometallic chemistry.

The fundamental idea behind our representation of stereochemistry is to really describe the relative spatial arrangement of the atoms of a chemical structure. For a given constitution, we obtain a unique and unambiguous stereochemical representation. No limitation to predefined types of stereogenic units exists; any conceivable relative spatial arrangement of atoms may be uniformly represented by one universally applicable formalism. Furthermore, this idea led us to a formal definition of the approximate-geometrical model of stereochemistry underlying the graphical representation of chemical structures via structural formulas.

We can still improve our representation of molecular structure. In its present form, it does not account for the topological properties of chemical structures which arise from their embedding in three-dimensional space. For example, we cannot adequately represent molecules having the shape of a topological knot or a helix. Work to include these topological properties in our molecular structure representation is in progress.

Our representation of molecular structure has been designed for the description of discrete chemical structures. Neither lattice structures, such as ion crystals or metals, nor polymers have been accounted for. These types of compounds can be regarded to consist of repeatedly occurring basic structural units. The latter can be represented by our approach, but a method has to be devised in order to describe the way these structural units combine.

ACKNOWLEDGMENT

802 *J. Chem. Inf. Comput. Sci., Vol. 35, No. 5, 1995*

DIETZ

REFERENCES AND NOTES

(1) IUPAC. *Nomenclature of Organic Chemistry, Sections A, B, C, D, E, F and H*, 1979 ed.; Pergamon Press: Oxford, 1979.

(2) Elschenbroich, C.; Salzer, A. *Organometallchemie*, 2nd ed.; Teubner: Stuttgart, 1988.

(3) Balaban, A. T., Ed. *Chemical Applications of Graph Theory*; Academic Press: London, 1976.

(4) Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, 1992.

(5) Dugundji, J.; Ugi, I. An Algebraic Model of Constitutional Chemistry as a Basis for Chemical Computer Programs. *Top. Curr. Chem.* **1973**, *39*, 19−64.

(6) Barnard, J. M. In *Chemical Structure Systems: Computational Techniques for Representation, Searching and Processing of Structural Information*; Ash, J. E., Warr, W. A., Willet, P., Eds.; Ellis Horwood: Chichester, 1991; Chapter 1, pp 9−56.

(7) Ugi, I.; Stein, N.; Knauer, M.; Gruber, B.; Bley, K.; Weidinger, R. New Elements in the Representation of the Logical Structure of Chemistry by Qualitative Mathematical Models and Corresponding Data Structures. *Top. Curr. Chem.* **1993**, *166*, 199−233.

(8) Ugi, I.; Bauer, J.; Brandt, J.; Friedrich, J.; Gasteiger, J.; Jochum, C.; Schubert, W. New Applications of Computers in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 111−123.

(9) Weidinger, R. Ein stereochemischer Netzwerkgenerator. Master's thesis, Universität Passau, 1991.

(10) Berge, C. *Graphes et Hypergraphes*; Dunod: Paris, 1970.

(11) March, J. *Advanced Organic Chemistry*, 4th ed.; John Wiley & Sons: New York, 1992.

(12) Gasteiger, J. A. Representation of π Systems for Efficient Computer Manipulation. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111−115.

(13) Stein, N. Das sXBE- und sXR-Modell der konstitutionellen Chemie. Doctoral thesis, Technische Universität München, 1994.

(14) Ugi, I.; Bauer, J.; Blomberger, C.; Brandt, J.; Dietz, A.; Fontain, E.; Gruber, B.; von Scholley-Pfab, A.; Senff, A.; Stein, N. Models, Concepts, Theories, and Formal Languages in Chemistry and Their Use as a Basis for Computer Assistance in Chemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 3−16.

(15) Hartsfield, N.; Ringel, G. *Pearls in Graph Theory*; Academic Press: San Diego, 1990.

(16) Pauling, L. *Grundlagen der Chemie*; Verlag Chemie: Weinheim, 1973.

(17) Eliel, E. L.; Wilen, S. H.; Mander, L. N. *Stereochemistry of Organic Compounds*; John Wiley & Sons: New York, 1994.

(18) Blackwood, J. E.; Elliott, P. M.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. III. Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 3−8.

(19) Hanessian, S.; Franco, J.; Gagnon, G.; Laramée, D.; Larouche, B. Computer-Assisted Analysis and Perception of Stereochemical Features in Organic Molecules Using the CHIRON Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 413−425.

(20) Mata, P.; Lobo, A. M.; Marshall, C.; Johnson, A. P. Implementation of the Cahn-Ingold-Prelog System for Stereochemical Perception in the LHASA Program. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 491−504.

(21) McDonnell, P. M.; Pasternack, R. F. A Line-Formula Notation System for Coordination Compounds. *J. Chem. Doc.* **1965**, *5*, 56−60.

(22) Petrarca, A. E.; Rush, J. E. Methods for Computer Generation of Unique Configurational Descriptors for Stereoisomeric Square-Planar and Octahedral Complexes. *J. Chem. Doc.* **1969**, *9*, 32−37.

(23) Ugi, I.; Marquarding, D.; Klusacek, H.; Gokel, G.; Gillespie, P. Chemistry and Logical Structures. *Angew. Chem., Int. Ed. Engl.* **1970**, *9*, 703−730.

(24) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.

(25) Dugundji, J.; Showell, J.; Kopp, R.; Marquarding, D.; Ugi, I. A Group Theoretical Analysis of Conformational Flexibility. *Isr. J. Chem.* **1980**, *20*, 20−35.

(26) Fella, A. L.; Nourse, J. G.; Smith, D. H. Conformation Specification of Chemical Structures in Computer Programs. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 43−47.

(27) Ugi, I.; Dugundji, J.; Kopp, R.; Marquarding, D. *Perspectives in Theoretical Stereochemistry*; Lecture Notes in Chemistry 36, Springer-Verlag: Berlin, 1984.

(28) Cahn, R. S.; Ingold, C. K.; Prelog, V. Specification of Molecular Chirality. *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 385−415.

(29) Prelog, V.; Helmchen, G. Basic Principles of the CIP-System and Proposals for a Revision. *Angew. Chem., Int. Ed. Engl.* **1982**, *21*, 567−583.

(30) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(31) Schubert, W.; Ugi, I. Constitutional Symmetry and Unique Descriptors of Molecules. *J. Am. Chem. Soc.* **1978**, *100*, 37−41.

(32) Balaban, A. T.; Mekenyan, O.; Bonchev, D. Unique Description of Chemical Structures Based on Hierarchically Ordered Extended Connectivities (HOC Procedures). I. Algorithms for Finding Graph Orbits and Canonical Numbering of Atoms. *J. Comput. Chem.* **1985**, *6*, 538−551.