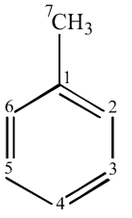
`_display_id` numbers at level 1. The connection object is specified with a `_display_conn_symbol` code, which must be a standard value in the dictionary definition, as is the colour of the icon if specified by `_display_conn_colour`.

2.4.6. Bonding conventions

Chemical information systems use a variety of conventions to specify attributes such as aromaticity, bond-order alternation, tautomerism *etc.* These system-dependent conventions decide the values that are permitted for quantities such as bond order, electronic charge and hydrogen-atom count. Most systems also provide for redundancy between chemical attributes. For example, the valency, the number of connected non-hydrogen atoms, the number of terminal hydrogen atoms and the bond types associated with a given atom are clearly related. Operational systems make use of these relationships to perform internal checks and to provide flexibility in substructure search processes.

The MIF data definitions provide for three bonding conventions. These are the data items `_bond_type_mif`, `_bond_type_casreg3` and `_bond_type_ccdc`. The 'mif' convention defines only single, double, triple and other bonds, while the 'casreg3' convention (Mockus & Stobaugh, 1980) extends these to include aromaticity in terms of 'ring alternating normalized bonds' and tautomerism *via* a 'tautomer normalized bond'. The 'ccdc' convention is that employed in the Cambridge Structural Database System (Allen *et al.*, 1991; Allen, 2002) to categorize bond types encountered in both organic and metal-organic molecules.

An important advantage of the MIF approach is that a molecule can be represented using all three bonding conventions within the same data block. An example of alternative bonding conventions encoded for toluene is shown in Fig. 2.4.6.1.




```
data_toluene

_molecule_name_common          toluene

loop_
    _atom_id
    _atom_type
    _atom_attach_h
          1  C  0    2  C  1    3  C  1    4  C  1
          5  C  1    6  C  1    7  C  3

loop_
    _bond_id_1
    _bond_id_2
    _bond_type_mif
    _bond_type_casreg3
    _bond_type_ccdc
          1  2  D  A  A    2  3  S  A  A
          3  4  D  A  A    4  5  S  A  A
          5  6  D  A  A    1  6  S  A  A
          1  7  S  S  S
```

Fig. 2.4.6.1. Three alternative bonding conventions for toluene stored in the same MIF data block.

2.4.7. Structural templates

In many chemical information systems, it is standard practice to build complete 2D molecular representations through the use of a library of commonly referenced structural templates, *e.g.* ligands, functional groups, amino-acid units *etc.*

In a MIF, molecular templates can be encapsulated as save frames, either within a data block for a specific molecule, or within a global block that is accessible to many data blocks. A simple application of a MIF template is shown in Fig. 2.4.7.1, where a 4-methylcyclohexyl ligand is used to encode the molecule tris(methylcyclohexyl)phosphine. In this example a molecular fragment is constructed in the save frame mechex, where the 'atom' sites and 'bond' connections appear in `_atom_*` and `_bond_*` loops. The molecule (2-methylcyclohexyl)(3-methylcyclohexyl)(4-methylcyclohexyl)phosphine is encoded by referencing the template fragment as the save frame $mechex. In the 'atom' loop, the item `_atom_environment` identifies the components of the target molecule as an 'atom' or 'frag' (fragment). If the component is a fragment, the items `_atom_frag_key` and `_atom_frag_id` are used to specify the frame code and the ID of the attached atom in the fragment, respectively. In the 'bond' loop, the connections from the atom P(1) to the template are encoded simply in terms of the `_atom_id` values. The necessary redefinition of the hydrogen and non-hydrogen counts of the template atoms is accomplished using the `_atom_attach_h` and `_atom_attach_nh` items, respectively. The external values override any values that are contained in, or derived from, the data in the template.

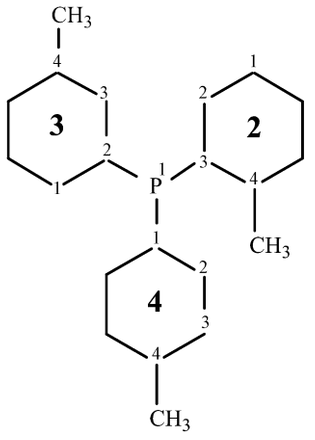
The same approach is used to construct the dipeptide alanylserine in Fig. 2.4.7.2. This employs the template peptide units described by the atoms and bonds in the save frames $alanyl and $seryl. The complete dipeptide is specified in its 'atom' list as the template peptides (identified by their save-frame names) and an additional carboxylate O atom. Note that only the atom sites affected by molecule formation are identified explicitly in this list, which gives the values of `_atom_attach_nh`, `_atom_attach_h` and `_atom_charge` for the modified sites in the zwitterionic form of alanylserine.

2.4.8. Stereochemistry and geometry at stereogenic centres

The Cahn–Ingold–Prelog (CIP) notation (Cahn *et al.*, 1966; Prelog & Helmchen, 1982) is available in the MIF definitions to specify the stereochemistry of a molecule. The CIP notation is restricted to tetrahedral atomic centres and to olefinic type stereogenic bonds, and the CIF approach is unsuitable for describing molecules with partially known stereochemistry, molecules containing more complex geometries or substructural queries. The MIF data items representing stereochemical quantities are as follows:

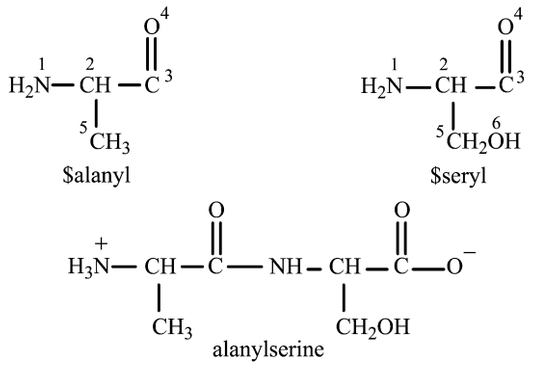
```
_define_stereo_relationship
_atom_cip
_bond_cip
_stereo_atom_id
_stereo_bond_id_1
_stereo_bond_id_2
_stereo_geometry
_stereo_vertex_id
```

The CIP stereochemical designators (R, S, E, Z, r, s, e, z *etc.*) are specified with the MIF data items `_atom_cip` and `_bond_cip`. The MIF atom-property data for the molecule (+)-3-bromocamphor are shown in Fig. 2.4.8.1. In this the absolute configuration is expressed as the atom CIP values *R*, *R* and *S* for nodes 1, 3 and 4. The period in this example is used to indicate a null field.



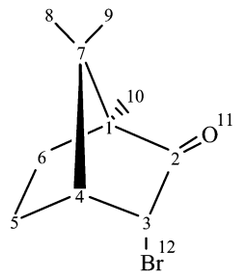
```
global_

save_mechex
     loop_
          _atom_id
          _atom_type
          _atom_attach_h
             1  C  2    2  C  2    3  C  2    4  C  1
             5  C  2    6  C  2    7  C  3
     loop_
          _bond_id_1
          _bond_id_2
          _bond_type_mif
             1  2  S    2  3  S    3  4  S    4  5  S
             5  6  S    1  6  S    4  7  S
save_

data_tris(methylcyclohexyl)phosphine

_molecule_name_common
; (2-methylcyclohexyl)(3-methylcyclohexyl)(4-
   methylcyclohexyl)phosphine
;

     loop_
          _atom_id
          _atom_environment
          _atom_frag_key
          _atom_frag_id
          _atom_type
          _atom_attach_nh
          _atom_attach_h
             1   atom   .         .  P  3  0
             2   frag   $mechex  3  C  3  1
             3   frag   $mechex  2  C  3  1
             4   frag   $mechex  1  C  3  1
     loop_
          _bond_id_1
          _bond_id_2
          _bond_type_mif
             1  2  S      1  3  S      1  4  S
```

Fig. 2.4.7.1. MIF representation of (2-methylcyclohexyl)(3-methylcyclohexyl)-(4-methylcyclohexyl)phosphine using a single global save frame that encapsulates the structure of methylcyclohexane, together with 'external' referencing of save-frame atoms in **_atom_** and **_bond_** loops.



```
global_

save_alanyl
     loop_
          _atom_id
          _atom_type
          _atom_attach_h
                1  N  2    2  C  1    3  C  0
                4  O  0    5  C  3
     loop_
          _bond_id_1
          _bond_id_2
          _bond_type_mif
                1  2  S    2  3  S    3  4  D
                2  5  S
save_
save_seryl
     loop_
          _atom_id
          _atom_type
          _atom_attach_h
                1  N  2    2  C  1    3  C  0
                4  O  0    5  C  2    6  O  1
     loop_
          _bond_id_1
          _bond_id_2
          _bond_type_mif
                1  2  S    2  3  S    3  4  D
                2  5  S    5  6  S
save_

data_alanylserine

_molecule_name_common     alanylserine
     loop_
          _atom_id
          _atom_environment
          _atom_frag_key
          _atom_frag_id
          _atom_type
          _atom_attach_nh
          _atom_attach_h
          _atom_charge
                1  frag   $alanyl   1   N   1   3  +1
                2  frag   $alanyl   3   C   3   0   0
                3  frag   $seryl    1   N   2   1   0
                4  frag   $seryl    3   C   3   0   0
                5  atom   .         .   O   1   0  -1
     loop_
          _bond_id_1
          _bond_id_2
          _bond_type_mif
                2 3 S    4 5 S
```

Fig. 2.4.7.2. MIF representation of the dipeptide alanylserine constructed using alanyl and seryl templates encapsulated in global save frames.




```
data_bromocamphor_2

_molecule_name_common       (+)-3-bromocamphor

_molecule_name_iupac
; 3R-bromo-1R,7,7-trimethyl-4S-bicyclo[2.2.1]heptan-
  2-one
;
    loop_
        _atom_id
        _atom_type
        _atom_attach_h
        _atom_cip
              1  C  0  R    2  C  0  .    3  C  1  R
              4  C  1  S    5  C  2  .    6  C  2  .
              7  C  0  .    8  C  3  .    9  C  3  .
             10  C  3  .   11  O  0  .   12  Br 0  .
```

Fig. 2.4.8.1. CIP stereochemical descriptors for (+)-3-bromocamphor.

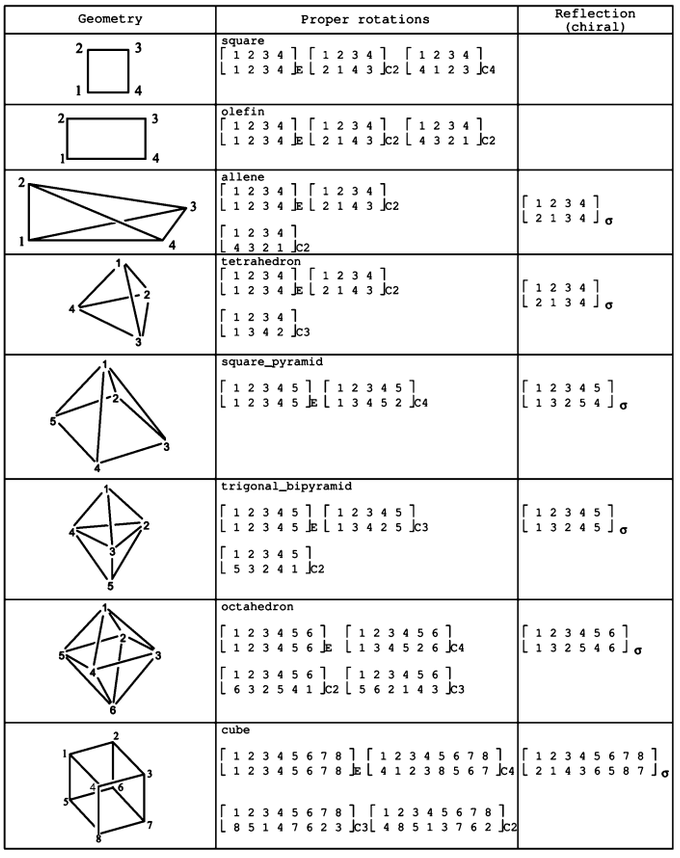


Fig. 2.4.8.2. Archetypal coordination geometries used in stereochemical definition of the MIF data item **_stereo_geometry**.

The stereogenic centre of a stereo group in a molecule has a relationship within that group that is specified by **_define_stereo_relationship**. Descriptions of the standard codes for **_define_stereo_relationship** are as follows.

absolute: The configuration of all stereogenic centres is exactly as described. This represents an enantiomerically pure compound with a known absolute configuration.

relative: The configuration of the stereogenic centres is only relative and the mirror reflection of the centres will also describe the same molecule. Only the configuration described in the MIF, or its mirror image, will be present in the molecule. This represents an enantiomerically pure compound with the described relative configuration.

racemic: The configuration of the stereogenic centres is only relative and the mirror reflection of the centres will also describe the same molecule. Both this configuration and its mirror image are present in a 1:1 ratio. This represents a racemic mixture of the molecule with the described relative configuration.

absolute_excess: The configuration of the stereogenic centres describes the absolute configuration of the excess component of a mixture of this configuration and its mirror reflection. This describes an enantiomeric excess in which the excess component has the described absolute configuration.

relative_excess: The configuration of the stereogenic centres is only relative. A mixture of this configuration and its mirror image is present, with one or other of the components in excess. This describes an enantiomeric excess mixture.

unknown: The configurational relationship between the stereogenic centres is not known.

The geometry of each stereogenic centre is described individually in terms of a prototype geometrical model. The basic principles of this approach have been described elsewhere (Barnard *et al.*, 1990). The eight geometries currently defined for the MIF data item **_stereo_geometry** are given in Fig. 2.4.8.2. They include the organic stereogenic geometries (the tetrahedron, the rectangular description of olefin-related compounds and the anti-rectangle used to describe allene-related systems) and the common archetypal metal coordination geometries (square planar, tetrahedral, trigonal bipyramidal, square pyramidal, octahedral and cubic). This list is non-exclusive and can be extended as required in later versions of the MIF dictionary.

The vertex site of the geometrical model must be occupied by either an atom, an explicit or implicit hydrogen atom, or by an explicitly declared electron pair. In each case, there exist permutations of the enumerated vertices that, if applied, do not change the meaning of the description of the relevant stereo element. Thus, the MIF does not define a canonical ordering for citing geometric vertices and the comparison of two geometries requires the use of the permutation operators. These permutations are also indicated in Fig. 2.4.8.2.

For each stereogenic centre (defined by a **_stereo_atom_id**, or by **_stereo_bond_id_1** and ***_2**), the atom sites forming the stereochemical element specified by a **_stereo_geometry** code are stored as a sequence of **_stereo_vertex_id** values. An example of the specification of absolute stereochemistry, including the ordered enumeration of the tetrahedral vertices for the four stereogenic centres, is given in Fig. 2.4.8.3. In this example, the null symbol (a period) is used to indicate an implicit hydrogen atom or an unshared electron pair.

2.4.9. MIF query applications

A MIF is suitable for interrogating databases because data items are permitted to have a single value, or a 'sequence' of alternative values. This latter option is designated by the dictionary attribute **_type_conditions** which, for MIF applications, is set to