

## 3.6. CLASSIFICATION AND USE OF MACROMOLECULAR DATA

ATOM\_SITES\_ALT\_GEN category explicitly tie together the alternative conformations that contribute to each ensemble. Finally, the atoms in each alternative conformation are identified in the ATOM\_SITE category by the data item `_atom_site.label_alt_id`.

The current version of the mmCIF dictionary cannot be used to describe an NMR structure determination completely. However, an mmCIF can be used to store the multiple models usually used to describe a structure determined by NMR using the data items in these categories.

Example 3.6.7.3 is a simplified version of the example given in the mmCIF dictionary (see Fig. 3.6.7.2).

## 3.6.7.2. Molecular chemistry

The categories describing molecular chemistry are as follows:

*Molecular chemistry in the core CIF dictionary* (§3.6.7.2.1)

CHEMICAL group

CHEMICAL  
CHEMICAL\_CONN\_ATOM  
CHEMICAL\_CONN\_BOND  
CHEMICAL\_FORMULA

*Chemical components* (§3.6.7.2.2)

CHEM\_COMP group

CHEM\_COMP  
CHEM\_COMP\_ANGLE  
CHEM\_COMP\_ATOM  
CHEM\_COMP\_BOND  
CHEM\_COMP\_CHIR  
CHEM\_COMP\_CHIR\_ATOM  
CHEM\_COMP\_PLANE  
CHEM\_COMP\_PLANE\_ATOM  
CHEM\_COMP\_TOR  
CHEM\_COMP\_TOR\_VALUE

*Chemical links* (§3.6.7.2.3)

CHEM\_LINK group

CHEM\_COMP\_LINK  
CHEM\_LINK  
CHEM\_LINK\_ANGLE  
CHEM\_LINK\_BOND  
CHEM\_LINK\_CHIR  
CHEM\_LINK\_CHIR\_ATOM  
CHEM\_LINK\_PLANE  
CHEM\_LINK\_PLANE\_ATOM  
CHEM\_LINK\_TOR  
CHEM\_LINK\_TOR\_VALUE  
ENTITY\_LINK

The detailed chemistry of the components of a macromolecular structure can be described using data items in the CHEM\_COMP and CHEM\_LINK category groups. These mmCIF categories are used in preference to those in the CHEMICAL category group in the core CIF dictionary, as macromolecules are in most cases linked assemblies of a limited number of monomers and so they are most efficiently described by defining the monomers and the links between them, rather than by a formal definition of every bond and angle.

All the categories relevant to molecular chemistry are listed in the summary above; note in particular the presence of the category ENTITY\_LINK within the formal CHEM\_LINK category group.

## 3.6.7.2.1. Molecular chemistry in the core CIF dictionary

The data items in these categories are as follows:

(a) CHEMICAL

- `_chemical.entry_id`  
→ `_entry.id`

- `_chemical.absolute_configuration`
- `_chemical.compound_source`
- `_chemical.melting_point`
- `_chemical.melting_point_gt`
- `_chemical.melting_point_lt`
- `_chemical.name_common`
- `_chemical.name_mineral`
- `_chemical.name_structure_type`
- `_chemical.name_systematic`
- `_chemical.optical_rotation`
- `_chemical.properties_biological`
- `_chemical.properties_physical`
- + `_chemical.temperature_decomposition`
- `_chemical.temperature_decomposition_gt`
- `_chemical.temperature_decomposition_lt`
- + `_chemical.temperature_sublimation`
- `_chemical.temperature_sublimation_gt`
- `_chemical.temperature_sublimation_lt`

(b) CHEMICAL\_CONN\_ATOM

- `_chemical_conn_atom.number`
- `_chemical_conn_atom.charge`
- `_chemical_conn_atom.display_x`
- `_chemical_conn_atom.display_y`
- `_chemical_conn_atom.NCA`
- `_chemical_conn_atom.NH`
- `_chemical_conn_atom.type_symbol`

(c) CHEMICAL\_CONN\_BOND

- `_chemical_conn_bond.atom_1`
- `_chemical_conn_bond.atom_2`
- `_chemical_conn_bond.type`

(d) CHEMICAL\_FORMULA

- `_chemical_formula.entry_id`  
→ `_entry.id`
- `_chemical_formula.analytical`
- `_chemical_formula.iupac`
- `_chemical_formula.moiety`
- `_chemical_formula.structural`
- `_chemical_formula.sum`
- `_chemical_formula.weight`
- `_chemical_formula.weight_meas`

The bullet (•) indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow (→) is a reference to a parent data item. Items in italics have aliases in the core CIF dictionary formed by changing the full stop (.) to an underscore (\_). Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string `_esd` to the data name listed.

Descriptions of molecular chemistry in an mmCIF are normally made using data items in the CHEM\_COMP and CHEM\_LINK category groups. The CHEMICAL category group is retained in the mmCIF dictionary solely for consistency with the core CIF dictionary and Section 3.2.4.2 may be consulted for details.

Two of the categories in this group, CHEMICAL\_CONN\_ATOM and CHEMICAL\_CONN\_BOND, have existing category keys in the core dictionary. The formal keys `_chemical.entry_id` and `_chemical_formula.entry_id` have been added to CHEMICAL and CHEMICAL\_FORMULA, respectively, to provide the category keys required by the DDL2 data model.

It is emphasized that these items will not appear in the description of a macromolecular structure, but they are retained to allow the representation of small-molecule or inorganic structures in the DDL2 formalism of mmCIF.

## 3.6.7.2.2. Chemical components

Data items in these categories are as follows:

(a) CHEM\_COMP

- `_chem_comp.id`
- `_chem_comp.formula`
- `_chem_comp.formula_weight`
- `_chem_comp.model_details`
- `_chem_comp.model_eref`

### 3. CIF DATA DEFINITION AND CLASSIFICATION

```

_chem_comp.model_source
_chem_comp.mon_nstd_class
_chem_comp.mon_nstd_details
_chem_comp.mon_nstd_flag
_chem_comp.mon_nstd_parent
_chem_comp.mon_nstd_parent_comp_id
  → _chem_comp.id
_chem_comp.name
_chem_comp.number_atoms_all
_chem_comp.number_atoms_nh
_chem_comp.one_letter_code
_chem_comp.three_letter_code
_chem_comp.type

```

#### (b) CHEM\_COMP\_ANGLE

```

• _chem_comp_angle.atom_id_1
  → _chem_comp_atom.atom_id
• _chem_comp_angle.atom_id_2
  → _chem_comp_atom.atom_id
• _chem_comp_angle.atom_id_3
  → _chem_comp_atom.atom_id
• _chem_comp_angle.comp_id
  → _chem_comp.id
+ _chem_comp_angle.value_angle
+ _chem_comp_angle.value_dist

```

#### (c) CHEM\_COMP\_ATOM

```

• _chem_comp_atom.atom_id
• _chem_comp_atom.comp_id
  → _chem_comp.id
  _chem_comp_atom.alt_atom_id
  _chem_comp_atom.charge
+ _chem_comp_atom.model_Cartn_x
+ _chem_comp_atom.model_Cartn_y
+ _chem_comp_atom.model_Cartn_z
  _chem_comp_atom.partial_charge
  _chem_comp_atom.substruct_code
  _chem_comp_atom.type_symbol
  → _atom_type.symbol

```

#### (d) CHEM\_COMP\_BOND

```

• _chem_comp_bond.atom_id_1
  → _chem_comp_atom.atom_id
• _chem_comp_bond.atom_id_2
  → _chem_comp_atom.atom_id
• _chem_comp_bond.comp_id
  → _chem_comp.id
  _chem_comp_bond.value_order
+ _chem_comp_bond.value_dist

```

#### (e) CHEM\_COMP\_CHIR

```

• _chem_comp_chir.id
• _chem_comp_chir.comp_id
  _chem_comp_chir.atom_id
  → _chem_comp_atom.atom_id
  _chem_comp_chir.atom_config
  → _chem_comp.id
  _chem_comp_chir.number_atoms_all
  _chem_comp_chir.number_atoms_nh
  _chem_comp_chir.volume_flag
+ _chem_comp_chir.volume_three

```

#### (f) CHEM\_COMP\_CHIR\_ATOM

```

• _chem_comp_chir_atom.atom_id
  → _chem_comp_atom.atom_id
• _chem_comp_chir_atom.chir_id
  → _chem_comp_chir.id
• _chem_comp_chir_atom.comp_id
  → _chem_comp.id
  _chem_comp_chir_atom.dev

```

#### (g) CHEM\_COMP\_LINK

```

• _chem_comp_link.link_id
  → _chem_link.id
  _chem_comp_link.details
  _chem_comp_link.type_comp_1
  → _chem_comp.type
  _chem_comp_link.type_comp_2
  → _chem_comp.type

```

#### (h) CHEM\_COMP\_PLANE

```

• _chem_comp_plane.id
• _chem_comp_plane.comp_id
  → _chem_comp.id
  _chem_comp_plane.number_atoms_all
  _chem_comp_plane.number_atoms_nh

```

#### (i) CHEM\_COMP\_PLANE\_ATOM

```

• _chem_comp_plane_atom.atom_id
  → _chem_comp_atom.atom_id
• _chem_comp_plane_atom.comp_id
  → _chem_comp.id
• _chem_comp_plane_atom.plane_id
  → _chem_comp_plane.id
+ _chem_comp_plane_atom.dist

```

#### (j) CHEM\_COMP\_TOR

```

• _chem_comp_tor.id
• _chem_comp_tor.comp_id
  → _chem_comp.id
  _chem_comp_tor.atom_id_1
  → _chem_comp_atom.atom_id
  _chem_comp_tor.atom_id_2
  → _chem_comp_atom.atom_id
  _chem_comp_tor.atom_id_3
  → _chem_comp_atom.atom_id
  _chem_comp_tor.atom_id_4
  → _chem_comp_atom.atom_id

```

#### (k) CHEM\_COMP\_TOR\_VALUE

```

• _chem_comp_tor_value.comp_id
• _chem_comp_tor_value.tor_id
+ _chem_comp_tor_value.angle
  → _chem_comp_atom.comp_id
+ _chem_comp_tor_value.dist
  → _chem_comp_tor.id

```

The bullet (•) indicates a category key. Where multiple items within a category are marked with a bullet, they must be taken together to form a compound key. The arrow (→) is a reference to a parent data item. Data items marked with a plus (+) have companion data names for the standard uncertainty in the reported value, formed by appending the string *\_esd* to the data name listed.

Data items in the CHEM\_COMP and related categories allow the covalent geometry, stereochemistry and Cartesian coordinates for the chemical components of the structure to be specified. These components may be monomers, *e.g.* the amino acids that form proteins, the nucleotides that form nucleic acids or the sugars that form oligosaccharides, or they may be the small-molecule compounds, ions or water molecules that co-crystallize with the macro-molecule(s).

In a small-molecule structure determination, the chemistry is often deduced from the electron density distribution. In contrast, in macromolecular crystallography, the chemistry of the monomers that form a polymeric macromolecule is usually known in advance and is used to interpret the electron density. In many cases, the chemistry of the monomers is so well determined that it is not worth storing a copy of the geometric restraints used in every mmCIF that uses the same set of data for the monomers. In these cases, the data item *\_chem\_comp.model\_eref* can be used to identify an external reference file (e.r.f.) that contains standard chemical data for these monomers. Although the present version of the mmCIF dictionary does not specify the form that the file identifier might take, it is likely that users will specify the location of the file in their local file system or the URL of files of reference data accessible over the Internet. In the long term, it would be helpful to have a standard repository of reference data for monomers with a stable identifier that is independent of file names or access protocols.

The relationships between the categories used to describe chemical components are shown in Fig. 3.6.7.3.

The CHEM\_COMP category provides data items for the chemical formula and formula weight of each component, the total number

### 3.6. CLASSIFICATION AND USE OF MACROMOLECULAR DATA

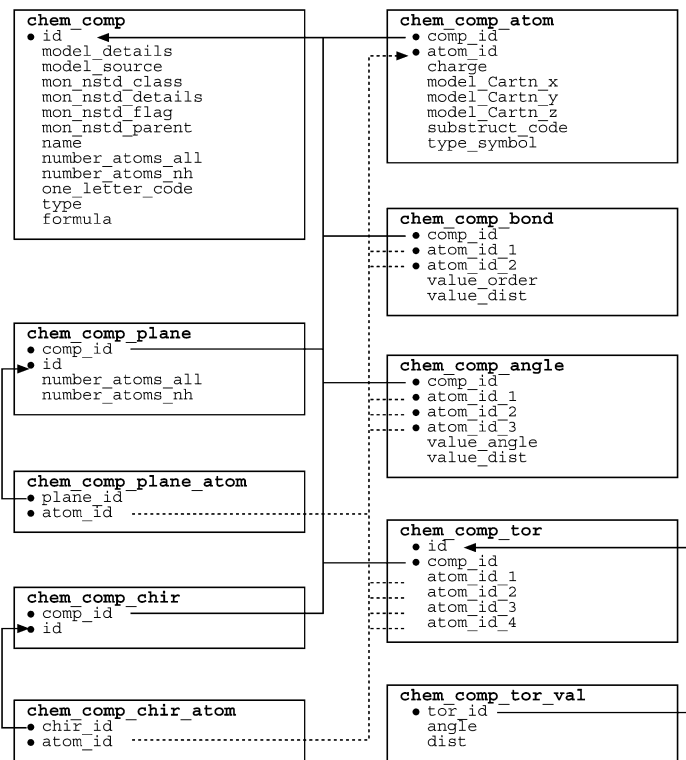


Fig. 3.6.7.3. The family of categories used to describe the chemical and structural features of the monomers and small molecules used to build a model of a structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

of atoms, the number of non-hydrogen atoms, and the name of the component. The name of the component will typically be a common name such as ‘alanine’ or ‘valine’; it is recommended that the IUPAC name is used for components that are not among the usual monomers that make up proteins, nucleic acids or sugars.

The one-letter or three-letter code for a standard component may be given (using `_chem_comp.one_letter_code` and `_chem_comp.three_letter_code`, respectively). Values of `x` for the one-letter code or `UNK` for the three-letter code are used to indicate components that do not have a standard abbreviation. A component that has been formed by modification of a standard component can be indicated by prefixing the code with a plus sign. A value of ‘.’, which means ‘not applicable’, should be used for components that are not monomers from which a polymeric macromolecule is built, for example co-crystallized small molecules, ions or water.

The data item `_chem_comp.type` can be used to describe the structural role of a monomer within a polymeric molecule. The types that are recognized are classified as linking monomers (for proteins, nucleic acids and sugars), monomers with an N-terminal or C-terminal cap (for proteins), and monomers with a 5’ or 3’ terminal cap (for nucleic acids). The specification of types for sugars is less complete than for proteins and nucleic acids and no types of terminal groups are currently specified for sugars. The values `non-polymer` and `other` are provided for types that have not been defined explicitly.

Information about the source of the model for the chemical component can be given using `_chem_comp.model_source` and `_chem_comp.model_details`. `_chem_comp.model_source` is a text field where the user might, for example, supply a reference to the Cambridge Structural Database or another small-molecule crystallographic database, or describe a molecular-modelling process. `_chem_comp.model_details` can be used to discuss any modification made to the model given in `_chem_comp.model_source`.

As mentioned previously, `_chem_comp.model_errf` can be used to specify the location of an external reference file if the model is not described within the current data block.

Macromolecules often contain modifications of standard monomers, such as phosphorylated serines and threonines. In the mmCIF data model, a nonstandard monomer should be treated as a separate `CHEM_COMP` entry and described in full. However, it may be useful to refer to the standard monomer from which it was derived using the `_chem_comp.mon_nstd_*` data items. There are no fixed rules for what constitutes a ‘standard’ or ‘nonstandard’ monomer in this context, but any covalent modification of a standard amino acid or nucleotide would generally be considered nonstandard. Sometimes it is difficult to decide whether a monomer is standard or nonstandard: selenomethionine is not one of the standard 20 amino acids, but it is so commonly used that geometric restraints for it are included in many standard packages for protein structure refinement.

Data items in the `CHEM_COMP_ATOM` category can be used to describe the atoms in a component. The position of each atom is given in orthogonal ångström coordinates. These coordinates correspond to the atom positions in the model of the component used in the refinement, not to the final set of refined atom positions recorded in the `ATOM_SITE` list.

Other `CHEM_COMP_ATOM` data items can be used to specify what element the atom is and its formal electronic charge, or partial charge. A code may also be assigned to the atom to indicate its role within a substructural classification of the component. The allowed codes are `main` and `side` for the main-chain and side-chain parts of amino acids, and `base`, `phos` and `sugar` for the base, phosphate and sugar parts of nucleotides. Atoms that do not belong to a substructure may be assigned the code `none`.

Data items in the `CHEM_COMP_BOND` category can be used to describe the intramolecular bonds between atoms in a component. Bond restraints may be described by the distance between the bonded atoms, the bond order, or both. The recognized bond types are the same as those for the core CIF dictionary data item `_chemical_conn_bond.type`, and they fulfil the same role: to characterize a model that could be used for database substructure searching, rather than to give a detailed description of unusual bond types.

In the `CHEM_COMP_ANGLE` category, atom 2 defines the vertex of the angle involving atoms 1, 2 and 3. The angle may be described as either an angle at the vertex atom or as a distance between atoms 1 and 3.

Data items in the `CHEM_COMP_CHIR` category can be used to describe the conformation of chiral centres within the component. The absolute configuration and the chiral volume may be specified, as well as the total number of atoms and the number of non-hydrogen atoms bonded to the chiral centre. There is also a flag to indicate whether a restrained chiral volume should match the target value in sign as well as in magnitude. Because chiral centres can involve a variable number of atoms, a separate list of the atoms should be given in `CHEM_COMP_CHIR_ATOM`.

Data items in the `CHEM_COMP_PLANE` category can be used to define planes within a component. The number of non-hydrogen atoms and the total number of atoms in each plane can be recorded. The atoms defining each plane should be listed separately in `CHEM_COMP_PLANE_ATOM`.

Data items in the `CHEM_COMP_TOR` category can be used to give details about the torsion angles in a component. A torsion angle may be described either as an angle or as a distance between the first and last atoms. (A torsion angle cannot be completely described by a distance, but sometimes a distance

Example 3.6.7.4. *The description of a component (adriamycin) of a macromolecule with data items in the CHEM\_COMP, CHEM\_COMP\_ATOM, CHEM\_COMP\_BOND, CHEM\_COMP\_TOR and CHEM\_COMP\_TOR\_VALUE categories (Leonard et al., 1993).*

```

_chem_comp.id          'DM2'
_chem_comp.name        'adriamycin'
_chem_comp.type        non-polymer
_chem_comp.formula     'C27 H29 N1 O11'
_chem_comp.number_atoms_all 68
_chem_comp.number_atoms_nh 39
_chem_comp.formula_weight 543.51

loop
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.model_Cartn_x
_chem_comp_atom.model_Cartn_y
_chem_comp_atom.model_Cartn_z
  DM2 'C1'  C  12.996  0.476  12.694
  DM2 'C2'  C  13.982 -0.225  13.183
  DM2 'C3'  C  12.482  0.165  11.515
# - - - abbreviated - - -

loop
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id 1
_chem_comp_bond.atom_id 2
_chem_comp_bond.value_order
_chem_comp_bond.value_dist
_chem_comp_bond.value_dist_esd
  DM2 'C1' 'C2' sing 1.517 0.0210
  DM2 'C2' 'C3' sing 1.445 0.0040
# - - - abbreviated - - -

loop
_chem_comp_tor.comp_id
_chem_comp_tor.id
_chem_comp_tor.atom_id 1
_chem_comp_tor.atom_id 2
_chem_comp_tor.atom_id 3
_chem_comp_tor.atom_id 4
  phe phe_chi1  N   CA   CB   CG
  phe phe_chi2  CA   CB   CG   CD1
  phe phe_ring1 CB   CG   CD1  CE1
  phe phe_ring2 CB   CG   CD2  CE2
  phe phe_ring3 CG   CD1  CE1  CZ
  phe phe_ring4 CD1  CE1  CZ   CE2
  phe phe_ring5 CE1  CZ   CE2  CD2

loop
_chem_comp_tor_value.tor_id
_chem_comp_tor_value.comp_id
_chem_comp_tor_value.angle
_chem_comp_tor_value.dist
  phe_chi1  phe -60.0 2.88
  phe_chi1  phe 180.0 3.72
  phe_chi1  phe 60.0 2.88
  phe_chi2  phe 90.0 3.34
  phe_chi2  phe -90.0 3.34
  phe_ring1 phe 180.0 3.75
  phe_ring2 phe 180.0 3.75
  phe_ring3 phe 0.0 2.80
  phe_ring4 phe 0.0 2.80
  phe_ring5 phe 0.0 2.80

```

restraint is used in refinement, where the value of the angle is assumed to be close to the target value.) As torsion angles can have more than one target value, the target values are specified in the CHEM\_COMP\_TOR\_VALUE category.

Data items in the CHEM\_COMP\_LINK category can be used to provide a table of links between the components of the structure. Each link is assigned an identifier (`_chem_comp_link.link_id`) and the types of monomer at each end of the link are stated. The types are those allowed for the parent data item `_chem_comp.type`.

The use of many of these data items to describe a typical component is shown in Example 3.6.7.4.

### 3.6.7.2.3. Chemical links

The data items in these categories are as follows:

#### (a) CHEM\_LINK

- `_chem_link.id`
- `_chem_link.details`

#### (b) CHEM\_LINK\_ANGLE

- `_chem_link_angle.atom_id_1`
- `_chem_link_angle.atom_id_2`
- `_chem_link_angle.atom_id_3`
- `_chem_link_angle.link_id`
- `_chem_link.id`
- `_chem_link_angle.atom_1_comp_id`
- `_chem_link_angle.atom_2_comp_id`
- `_chem_link_angle.atom_3_comp_id`
- + `_chem_link_angle.value_angle`
- + `_chem_link_angle.value_dist`

#### (c) CHEM\_LINK\_BOND

- `_chem_link_bond.atom_id_1`
- `_chem_link_bond.atom_id_2`
- `_chem_link_bond.link_id`
- `_chem_link.id`
- `_chem_link_bond.atom_1_comp_id`
- `_chem_link_bond.atom_2_comp_id`
- + `_chem_link_bond.value_dist`
- `_chem_link_bond.value_order`

#### (d) CHEM\_LINK\_CHIR

- `_chem_link_chir.id`
- `_chem_link_chir.link_id`
- `_chem_link.id`
- `_chem_link_chir.atom_comp_id`
- `_chem_link_chir.atom_id`
- `_chem_link_chir.atom_config`
- `_chem_link_chir.number_atoms_all`
- `_chem_link_chir.number_atoms_nh`
- `_chem_link_chir.volume_flag`
- + `_chem_link_chir.volume_three`

#### (e) CHEM\_LINK\_CHIR\_ATOM

- `_chem_link_chir_atom.atom_id`
- `_chem_link_chir_atom.chir_id`
- `_chem_link_chir.id`
- `_chem_link_chir_atom.atom_comp_id`
- `_chem_link_chir_atom.dev`

#### (f) CHEM\_LINK\_PLANE

- `_chem_link_plane.id`
- `_chem_link_plane.link_id`
- `_chem_link.id`
- `_chem_link_plane.number_atoms_all`
- `_chem_link_plane.number_atoms_nh`

#### (g) CHEM\_LINK\_PLANE\_ATOM

- `_chem_link_plane_atom.atom_id`
- `_chem_link_plane_atom.plane_id`
- `_chem_link_plane.id`
- `_chem_link_plane_atom.atom_comp_id`

#### (h) CHEM\_LINK\_TOR

- `_chem_link_tor.id`
- `_chem_link_tor.link_id`
- `_chem_link.id`
- `_chem_link_tor.atom_1_comp_id`
- `_chem_link_tor.atom_2_comp_id`
- `_chem_link_tor.atom_3_comp_id`
- `_chem_link_tor.atom_4_comp_id`
- `_chem_link_tor.atom_id_1`
- `_chem_link_tor.atom_id_2`
- `_chem_link_tor.atom_id_3`
- `_chem_link_tor.atom_id_4`

#### (i) CHEM\_LINK\_TOR\_VALUE

- `_chem_link_tor_value.tor_id`
- `_chem_link_tor.id`
- + `_chem_link_tor_value.angle`
- + `_chem_link_tor_value.dist`