International Tables for Crystallography (2006). Vol. G, Section 3.6.7, pp. 164–190.

Example 3.6.6.13. *The data set used in the refinement of an HIV-1 protease structure (PDB 5HVP) described using data items in the* REFLNS *and* REFLNS_SHELL *categories.*

```
_reflns.entry_id                        '5HVP'
_reflns.data_reduction_method
; Xengen program scalei. Anomalous pairs were merged.
  Scaling proceeded in several passes, beginning with
  1-parameter fit and ending with 3-parameter fit.
;
_reflns.data_reduction_details
; Merging and scaling based on only those reflections
  with I > sigma(I).
;

_reflns.d_resolution_high               2.00
_reflns.d_resolution_low                8.00

_reflns.limit_h_max                     22
_reflns.limit_h_min                     0
_reflns.limit_k_max                     46
_reflns.limit_k_min                     0
_reflns.limit_l_max                     57
_reflns.limit_l_min                     0

_reflns.number_obs                      7228
_reflns.observed_criterion_sigma_I      1.0
_reflns.details                         none

loop_
_reflns_shell.d_res_high
_reflns_shell.d_res_low
_reflns_shell.meanI_over_sigI_obs
_reflns_shell.number_measured_obs
_reflns_shell.number_unique_obs
_reflns_shell.percent_possible_obs
_reflns_shell.Rmerge_F_obs
   31.38   3.82   69.8   9024   2540   96.8    1.98
    3.82   3.03   26.1   7413   2364   95.1    3.85
    3.03   2.65   10.5   5640   2123   86.2    6.37
    2.65   2.41    6.4   4322   1882   76.8    8.01
    2.41   2.23    4.3   3247   1714   70.4    9.86
    2.23   2.10    3.1   1140    812   33.3   13.99
```

a reflection as being observed are given using the data item `_reflns.observed_criterion`. This is a free-text field so is not automatically parsable. Therefore it is supplemented in the mmCIF dictionary by data items that can be used to stipulate the criterion in terms of the values of F, I or the uncertainties in these quantities (Example 3.6.6.13). The percentage of the total number of reflections that meet the criterion can be recorded.

Data items are also provided for describing the selection of the reflections used to calculate the free R factor, and for giving the $R_{\rm merge}$ values for all reflections and for the subset of 'observed' reflections. Data items in the REFLNS_SCALE and REFLNS_SHELL categories are used in the same way in the mmCIF and core CIF dictionaries, and Section 3.2.3.2.2 can be consulted for details.

As with the related categories DIFFRN_REFLNS_CLASS and REFINE_LS_CLASS, the core dictionary category REFLNS_CLASS was introduced after the release of the first version of the mmCIF dictionary. It provides a more general way of describing the treatment of particular subsets of the observations, but it is not expected to be used in macromolecular structural studies, where partition by shells of resolution is traditional.

3.6.7. Atomicity, chemistry and structure

The basic concepts of the mmCIF model for describing a macromolecular structure were outlined in Section 3.6.3. The present section describes the components of the model in more detail. The category groups used to describe the molecular chemistry and structure are: the ATOM group describing atom positions (Section 3.6.7.1); the CHEMICAL, CHEM_COMP and CHEM_LINK groups describing molecular chemistry (Section 3.6.7.2); the ENTITY group describing distinct chemical species (Section 3.6.7.3); the GEOM group describing molecular or packing geometry (Section 3.6.7.4); the STRUCT group describing the large-scale features of molecular structure (Section 3.6.7.5); and the SYMMETRY group describing the symmetry and space group (Section 3.6.7.6).

The CHEMICAL category group itself is not generally used in an mmCIF. The purpose of this category group in the core CIF dictionary is to specify the chemical identity and connectivity of the relatively simple molecular or ionic species in a small-molecule or inorganic crystal. In principle, a macromolecular structure determined to atomic resolution could be represented as a coherent chemical entity with a complete connectivity graph. However, in practice, biological macromolecules are built from units from a library of models of standard amino acids, nucleotides and sugars. Data items in the CHEM_COMP and CHEM_LINK category groups of the mmCIF dictionary describe the internal connectivity and standard bonding processes between these units.

Molecular or packing geometry is also rarely tabulated for large macromolecular complexes, so the GEOM category group is rarely used in an mmCIF.

3.6.7.1. Atom sites

The categories describing atom sites are as follows:

ATOM group
Individual atom sites (§3.6.7.1.1)
ATOM_SITE
ATOM_SITE_ANISOTROP
Collections of atom sites (§3.6.7.1.2)
ATOM_SITES
ATOM_SITES_FOOTNOTE
Atom types (§3.6.7.1.3)
ATOM_TYPE
Alternative conformations (§3.6.7.1.4)
ATOM_SITES_ALT
ATOM_SITES_ALT_ENS
ATOM_SITES_ALT_GEN

The ATOM category group represents a compromise between the representation of a small-molecule structure as an annotated list of atomic coordinates and the need in macromolecular crystallography to present a more structured view organized around residues, chains, sheets, turns, helices *etc.* The locations of individual atoms and other information about the atom sites are given using data items in this category group. The categories within the group may be classified as shown in the summary above.

The ATOM_SITE, ATOM_SITES and ATOM_TYPE categories have many data items that are aliases of equivalent data items in the same categories in the core CIF dictionary, but the conventions for the labelling of the atom sites are different.

The ATOM_SITE_ANISOTROP and ATOM_SITES_FOOTNOTE categories are new to the mmCIF dictionary, as are the categories related to alternative conformations: ATOM_SITES_ALT, ATOM_SITES_ALT_ENS and ATOM_SITES_ALT_GEN.

3.6.7.1.1. *Individual atom sites*

The data items in these categories are as follows:

(*a*) ATOM_SITE
• **_atom_site.id** (∼ ***_atom_site_label***)
_atom_site.adp_type
+ _atom_site.aniso_B[1][1]
⇌ ***_atom_site_anisotrop.B[1][1]***

+ _atom_site.aniso_B[1][2]
⇌ *_atom_site_anisotrop.B[1][2]*
+ _atom_site.aniso_B[1][3]
⇌ *_atom_site_anisotrop.B[1][3]*
+ _atom_site.aniso_B[2][2]
⇌ *_atom_site_anisotrop.B[2][2]*
+ _atom_site.aniso_B[2][3]
⇌ *_atom_site_anisotrop.B[2][3]*
+ _atom_site.aniso_B[3][3]
⇌ *_atom_site_anisotrop.B[3][3]*
_atom_site.aniso_ratio
⇌ *_atom_site_anisotrop.ratio*
+ _atom_site.aniso_U[1][1]
⇌ *_atom_site_anisotrop.U[1][1]*
+ _atom_site.aniso_U[1][2]
⇌ *_atom_site_anisotrop.U[1][2]*
+ _atom_site.aniso_U[1][3]
⇌ *_atom_site_anisotrop.U[1][3]*
+ _atom_site.aniso_U[2][2]
⇌ *_atom_site_anisotrop.U[2][2]*
+ _atom_site.aniso_U[2][3]
⇌ *_atom_site_anisotrop.U[2][3]*
+ _atom_site.aniso_U[3][3]
⇌ *_atom_site_anisotrop.U[3][3]*
_atom_site.attached_hydrogens
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.auth_comp_id
_atom_site.auth_seq_id
+ *_atom_site.B_equiv_geom_mean*
+ *_atom_site.B_iso_or_equiv*
_atom_site.calc_attached_atom
_atom_site.calc_flag
+ *_atom_site.Cartn_x*
+ *_atom_site.Cartn_y*
+ *_atom_site.Cartn_z*
_atom_site.chemical_conn_number
→ _chemical_conn_atom.number
_atom_site.constraints
_atom_site.details (∼ *_atom_site_description*)
_atom_site.disorder_assembly
_atom_site.disorder_group
_atom_site.footnote_id
+ *_atom_site.fract_x*
+ *_atom_site.fract_y*
+ *_atom_site.fract_z*
_atom_site.group_PDB
_atom_site.label_alt_id
→ _atom_sites_alt.id
_atom_site.label_asym_id
→ _struct_asym.id
_atom_site.label_atom_id
→ _chem_comp_atom.atom_id
_atom_site.label_comp_id
→ _chem_comp.id
_atom_site.label_entity_id
→ _entity.id
_atom_site.label_seq_id
→ _entity_poly_seq.num
+ *_atom_site.occupancy*
_atom_site.refinement_flags
_atom_site.refinement_flags_adp
_atom_site.refinement_flags_occupancy
_atom_site.refinement_flags_posn
_atom_site.restraints
_atom_site.symmetry_multiplicity
_atom_site.thermal_displace_type
_atom_site.type_symbol
→ _atom_type.symbol
+ *_atom_site.U_equiv_geom_mean*
+ *_atom_site.U_iso_or_equiv*
_atom_site.Wyckoff_symbol

(*b*) ATOM_SITE_ANISOTROP
● _atom_site_anisotrop.id
+ *_atom_site_anisotrop.B[1][1]* (∼ *_atom_site_aniso_B_11*)
+ *_atom_site_anisotrop.B[1][2]* (∼ *_atom_site_aniso_B_12*)
+ *_atom_site_anisotrop.B[1][3]* (∼ *_atom_site_aniso_B_13*)
+ *_atom_site_anisotrop.B[2][2]* (∼ *_atom_site_aniso_B_22*)
+ *_atom_site_anisotrop.B[2][3]* (∼ *_atom_site_aniso_B_23*)
+ *_atom_site_anisotrop.B[3][3]* (∼ *_atom_site_aniso_B_33*)
_atom_site_anisotrop.ratio (∼ *_atom_site_aniso_ratio*)
→ _atom_site.id
_atom_site_anisotrop.type_symbol
(∼ *_atom_site_aniso_type_symbol*)
→ _atom_type.symbol
+ *_atom_site_anisotrop.U[1][1]* (∼ *_atom_site_aniso_U_11*)
+ *_atom_site_anisotrop.U[1][2]* (∼ *_atom_site_aniso_U_12*)
+ *_atom_site_anisotrop.U[1][3]* (∼ *_atom_site_aniso_U_13*)
+ *_atom_site_anisotrop.U[2][2]* (∼ *_atom_site_aniso_U_22*)
+ *_atom_site_anisotrop.U[2][3]* (∼ *_atom_site_aniso_U_23*)
+ *_atom_site_anisotrop.U[3][3]* (∼ *_atom_site_aniso_U_33*)

The bullet (●) indicates a category 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 (_) except where indicated by the ∼ symbol. 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. The double arrow (⇌) indicates alternative names in a distinct category.*

The refined coordinates of the atoms in the crystallographic asymmetric unit are stored in the ATOM_SITE category. Atom positions and their associated uncertainties may be given using either Cartesian or fractional coordinates, and anisotropic displacement factors and occupancies may be given for each position.

The relationships between categories describing atom sites are shown in Fig. 3.6.7.1.

Several of the mmCIF data names arise from the need to associate atom sites with residues and chains. As in the core CIF dictionary, the identifier for the atom site is the data item **_atom_site_label**. To accommodate standard practice in macromolecular crystallography, the mmCIF atom identifier is the aggregate of **_atom_site.label_alt_id**, ***.label_asym_id**, ***.label_atom_id**, ***.label_comp_id** and ***.label_seq_id**. For the two types of files to be compatible, the data item **_atom_site.id**, which is independent of the different modes of identifying atoms (discussed below), was introduced. The mmCIF identifier **_atom_site.id** is aliased to the core CIF identifier **_atom_site_label**.

Since the identifier does not need to be a number, it is quite possible (although it is not recommended) to use a complex label with an internal structure corresponding to the label components that the mmCIF dictionary provides as separate data items. This scheme is described in Section 3.2.4.1.1. However, normal practice in mmCIFs should be to label sites with the functional components available and to assign a simple numeric sequence to the values of **_atom_site.id** (see Example 3.6.7.1).

In addition to labelling information, each entry in the ATOM_SITE list must contain a value for the data item **_atom_site.type_symbol**, which is a pointer to the table of element symbols in the ATOM_TYPE category. All other data items in the ATOM_SITE category are optional, but it is normal practice to

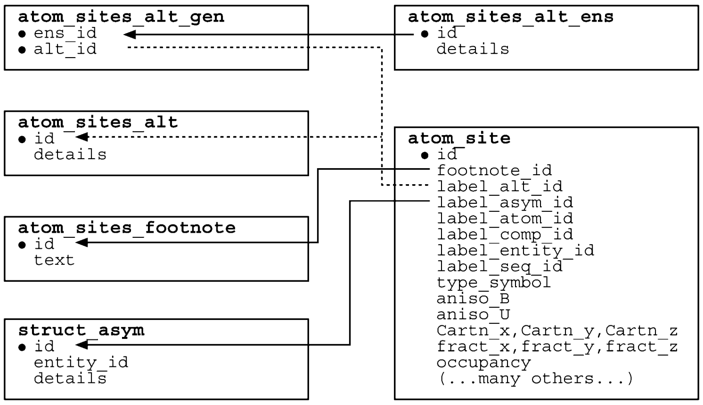

Fig. 3.6.7.1. The family of categories used to describe atom sites. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.1. *Part of the coordinate list for an HIV-1 protease structure (PDB 5HVP) described with data items in the* ATOM_SITE *category. Atoms are given for both polymer and non-polymer regions of the structure, and atoms in the side chain of residue 12 adopt alternative conformations.*

```
loop_
_atom_site.group_PDB
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_seq_id
_atom_site.label_alt_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.footnote_id
_atom_site.auth_seq_id
_atom_site.id
 ATOM N N   THR  A  12  .  26.095  32.930  14.590
      1.00  18.97  4  12 8
 ATOM C CA  THR  A  12  .  25.734  32.995  16.032
      1.00  19.80  4  12 9
 ATOM C C   THR  A  12  .  24.695  34.106  16.113
      1.00  20.92  4  12 10
 ATOM O O   THR  A  12  .  24.869  35.118  15.421
      1.00  21.84  4  12 11
 ATOM C CB  THR  A  12  .  26.911  33.346  17.018
      1.00  20.51  4  12 12
 ATOM O OG1 THR  A  12  3  27.946  33.921  16.183
      0.50  20.29  4  12 13
 ATOM O OG1 THR  A  12  4  27.769  32.142  17.103
      0.50  20.59  4  12 14
 ATOM C CG2 THR  A  12  3  27.418  32.181  17.878
      0.50  20.47  4  12 15
 ATOM C CG2 THR  A  12  4  26.489  33.778  18.426
      0.50  20.00  4  12 16
# - - - abbreviated - - -
 HETATM C C1 APS  C  .  1  4.171  29.012   7.116
      0.58  17.27  1 300 101
 HETATM C C2 APS  C  .  1  4.949  27.758   6.793
      0.58  16.95  1 300 102
 HETATM O O3 APS  C  .  1  4.800  26.678   7.393
      0.58  16.85  1 300 103
 HETATM N N4 APS  C  .  1  5.930  27.841   5.869
      0.58  16.43  1 300 104
# - - - abbreviated - - -
```

give either the Cartesian or fractional coordinates. Most macromolecular structures use Cartesian coordinates. Isotropic displacement factors are normally placed directly in the ATOM_SITE category, using `_atom_site.B_iso_or_equiv`. Anisotropic displacement factors may be placed directly in the ATOM_SITE category *or* in the ATOM_SITE_ANISOTROP category. U's may be used instead of B's. It is not acceptable to use both U's and B's, nor is it acceptable to have anisotropic displacement factors in both the ATOM_SITE category and the ATOM_SITE_ANISOTROP category.

Each atom within each chemical component is uniquely identified using the data item `_atom_site.label_atom_id`, which is a reference to the data item `_chem_comp_atom.atom_id` in the CHEM_COMP_ATOM category.

The specific object in the asymmetric unit to which the atom belongs is indicated using the data item `_atom_site.label_asym_id`, which is a reference to the data item `_struct_asym.id` in the STRUCT_ASYM category. For macromolecules, it is useful to think of this identifier as a chain ID.

The chemical component to which the atom belongs is indicated using the data item `_atom_site.label_comp_id`, which is a reference to the data item `_chem_comp.id` in the CHEM_COMP category. The chemical component that is referenced in this way may be either a non-polymer or a monomer in a polymer; if it is a monomer in a polymer, it is useful to think of this identifier as the residue name.

The correspondence between the sequence of an entity in a polymer and the sequence information in the coordinate list (and in the STRUCT categories) is established using the data item `_atom_site.label_seq_id`, which is a reference to the data item `_entity_poly_seq.num` in the ENTITY_POLY_SEQ category. This identifier has no meaning for entities that are not part of a polymer; in a polymer it is useful to think of this identifier as the residue number. Note that this is strictly a number. If the combination of a number with an insertion code is needed, `_atom_site.auth_seq_id` should be used (see below).

An alternative set of identifiers can be used for the `*_asym_id`, `*_atom_id`, `*_comp_id` and `*_seq_id` identifiers, but not for `*_alt_id`. The `_atom_site.label_*` data names are standard; there are rules for these identifiers such as the requirement that residue numbers are sequential integers. Different databases may also have their own rules. However, the author of an mmCIF may wish to use a nonstandard labelling scheme, *e.g.* to reflect the residue numbering scheme of a structure to which the present structure is homologous, apart from insertions and gaps. Another situation in which a nonstandard labelling scheme might be used is to follow a local convention for atom names in a non-polymer, such as a haem, that conflicts with the scheme required by a database in which the structure is to be deposited. In these situations, alternative identifiers can be given using the data names (`_atom_site.auth_*`).

In regions of the structure with alternative conformations, the specific conformation to which an atom belongs can be indicated using the data item `_atom_site.label_alt_id`, which is a reference to the data item `_atom_sites_alt.id` in the ATOM_SITES_ALT category.

The chemically distinct part of the structure (*e.g.* polymer chain, ligand, solvent) to which an atom belongs can be indicated using the data item `_atom_site.label_entity_id`, which is a reference to the data item `_entity.id` in the ENTITY category.

Most of the information that needs to be associated with an atom site is conveyed by the values of specific data names in mmCIF. However, for historical reasons, a pointer to additional free-text information about an atom site or about a group of atom sites can be given using the data item `_atom_site.footnote_id`, which is a reference to the data item `_atom_sites_footnote.id` in the ATOM_SITES_FOOTNOTE category.

The data item `_atom_site.group_PDB` is a place holder for the tags used by the PDB to identify types of coordinate records. It allows interconversion between mmCIFs and PDB format files. The only permitted values are ATOM and HETATM.

As in the core CIF dictionary, anisotropic displacement parameters in an mmCIF can be given in the same list as the atom positions and occupancies, or can be given in a separate list. However, DDL2 does not permit the same data names to be used for both constructs. Therefore, in mmCIF, anisotropic displacement parameters presented in a separate list are handled in a separate category with its own key, `_atom_site_anisotrop.id`, which must match a corresponding label in the atom-site list, `_atom_site.id`.

The individual elements of the anisotropic displacement matrix are labelled slightly differently in the mmCIF dictionary than in the core CIF dictionary in order to emphasize their matrix character. However, the definitions of the corresponding data items are identical in the two dictionaries.

3.6.7.1.2. *Collections of atom sites*

The data items in these categories are as follows:

(*a*) ATOM_SITES

```
• _atom_sites.entry_id
       → _entry.id
  _atom_sites.Cartn_transf_matrix[1][1]
       (~ _atom_sites_Cartn_tran_matrix_11)
  _atom_sites.Cartn_transf_matrix[1][2]
       (~ _atom_sites_Cartn_tran_matrix_12)
  _atom_sites.Cartn_transf_matrix[1][3]
       (~ _atom_sites_Cartn_tran_matrix_13)
  _atom_sites.Cartn_transf_matrix[2][1]
       (~ _atom_sites_Cartn_tran_matrix_21)
  _atom_sites.Cartn_transf_matrix[2][2]
       (~ _atom_sites_Cartn_tran_matrix_22)
  _atom_sites.Cartn_transf_matrix[2][3]
       (~ _atom_sites_Cartn_tran_matrix_23)
  _atom_sites.Cartn_transf_matrix[3][1]
       (~ _atom_sites_Cartn_tran_matrix_31)
  _atom_sites.Cartn_transf_matrix[3][2]
       (~ _atom_sites_Cartn_tran_matrix_32)
  _atom_sites.Cartn_transf_matrix[3][3]
       (~ _atom_sites_Cartn_tran_matrix_33)
  _atom_sites.Cartn_transf_vector[1]
       (~ _atom_sites_Cartn_tran_vector_1)
  _atom_sites.Cartn_transf_vector[2]
       (~ _atom_sites_Cartn_tran_vector_2)
  _atom_sites.Cartn_transf_vector[3]
       (~ _atom_sites_Cartn_tran_vector_3)
  _atom_sites.Cartn_transform_axes
  _atom_sites.fract_transf_matrix[1][1]
       (~ _atom_sites_fract_tran_matrix_11)
  _atom_sites.fract_transf_matrix[1][2]
       (~ _atom_sites_fract_tran_matrix_12)
  _atom_sites.fract_transf_matrix[1][3]
       (~ _atom_sites_fract_tran_matrix_13)
  _atom_sites.fract_transf_matrix[2][1]
       (~ _atom_sites_fract_tran_matrix_21)
  _atom_sites.fract_transf_matrix[2][2]
       (~ _atom_sites_fract_tran_matrix_22)
  _atom_sites.fract_transf_matrix[2][3]
       (~ _atom_sites_fract_tran_matrix_23)
  _atom_sites.fract_transf_matrix[3][1]
       (~ _atom_sites_fract_tran_matrix_31)
  _atom_sites.fract_transf_matrix[3][2]
       (~ _atom_sites_fract_tran_matrix_32)
  _atom_sites.fract_transf_matrix[3][3]
       (~ _atom_sites_fract_tran_matrix_33)
  _atom_sites.fract_transf_vector[1]
       (~ _atom_sites_fract_tran_vector_1)
  _atom_sites.fract_transf_vector[2]
       (~ _atom_sites_fract_tran_vector_2)
  _atom_sites.fract_transf_vector[3]
       (~ _atom_sites_fract_tran_vector_3)
  _atom_sites.solution_hydrogens
  _atom_sites.solution_primary
  _atom_sites.solution_secondary
  _atom_sites.special_details
```

(*b*) ATOM_SITES_FOOTNOTE

```
• _atom_sites_footnote.id
  _atom_sites_footnote.text
```

The bullet (•) indicates a category 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 (_) except where indicated by the ~ symbol.

The ATOM_SITES category of the core dictionary, which is used to record information that applies collectively to all the atom sites in the model of the structure, is incorporated without change into the mmCIF dictionary, and Section 3.2.4.1.2 can be consulted for details.

In practice, the data names in the PHASING categories are preferred to the aliases to the core CIF data items `_atom_sites.solution_primary`, `*_secondary` and `*_hydrogens`. The data items in the mmCIF PHASING categories are designed to allow a much more detailed description of how a macromolecular structure was solved.

Example 3.6.7.2. *Footnotes for particular groups of atom sites in an HIV-1 protease structure (PDB 5HVP) using data items in the* ATOM_SITES_FOOTNOTE *category.*

```
loop_
  _atom_sites_footnote.id
  _atom_sites_footnote.text
    3
; The positions of these water molecules correlate
  with the alternative orientations of the
  inhibitor. Water molecules with alternative ID
  "1" and occupancy 0.58 correlate with
  inhibitor orientation "1". Water molecules with
  alternative ID "2" and occupancy 0.42 correlate
  with inhibitor orientation "2".
;
    4
; Side chains of these residues adopt alternative
  orientations that do not correlate with the
  alternative orientation of the inhibitor.
;
```

The data item `_atom_sites.entry_id` has been added to the ATOM_SITES category to provide the formal category key required by the DDL2 data model.

The ATOM_SITES_FOOTNOTE category can be used to note something about a group of sites in the ATOM_SITE coordinate list, each of which is flagged with the same value of `_atom_site.footnote_id`. For example, an author may wish to note atoms for which the electron density is very weak, or atoms for which static disorder has been modelled. Example 3.6.7.2 shows how an author has used these data items to describe alternative orientations in part of a structure. However, the very large number of data names describing specific structural characteristics in the mmCIF dictionary mean that these rather general data names are rarely needed.

3.6.7.1.3. *Atom types*

The data items in this category are as follows:

ATOM_TYPE

```
• _atom_type.symbol
  _atom_type.analytical_mass_percent
     (~ _atom_type_analytical_mass_%)
  _atom_type.description
  _atom_type.number_in_cell
  _atom_type.oxidation_number
  _atom_type.radius_bond
  _atom_type.radius_contact
  _atom_type.scat_Cromer_Mann_a1
  _atom_type.scat_Cromer_Mann_a2
  _atom_type.scat_Cromer_Mann_a3
  _atom_type.scat_Cromer_Mann_a4
  _atom_type.scat_Cromer_Mann_b1
  _atom_type.scat_Cromer_Mann_b2
  _atom_type.scat_Cromer_Mann_b3
  _atom_type.scat_Cromer_Mann_b4
  _atom_type.scat_Cromer_Mann_c
  _atom_type.scat_dispersion_imag
  _atom_type.scat_dispersion_real
  _atom_type.scat_dispersion_source
  _atom_type.scat_length_neutron
  _atom_type.scat_source
  _atom_type.scat_versus_stol_list
```

The bullet (•) indicates a category key. Items in italics have aliases in the core CIF dictionary formed by changing the full stop (.) to an underscore (_) except where indicated by the ~ symbol.

The ATOM_TYPE category, which provides information about the atomic species associated with each atom site in the model of the structure, is used in the same way in the mmCIF dictionary as in the core CIF dictionary. See Section 3.2.4.1.3 for details.

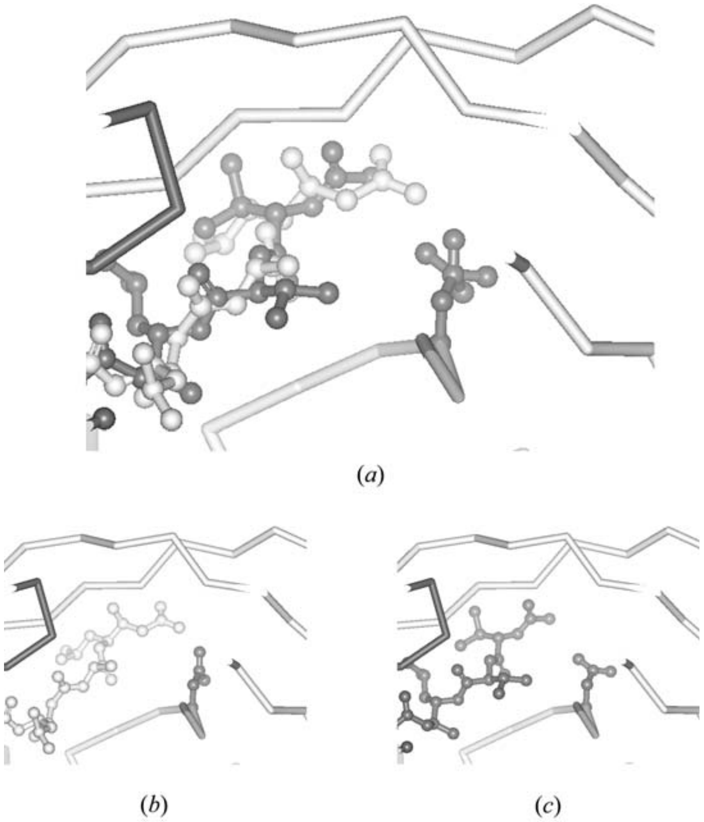

Fig. 3.6.7.2. Alternative conformations in an HIV-1 protease structure (PDB 5HVP) to be described with data items in the ATOM_SITES_ALT, ATOM_SITES_ALT_ENS and ATOM_SITES_ALT_GEN categories. (*a*) Complete structure, (*b*) ensemble 1, (*c*) ensemble 2.

3.6.7.1.4. *Alternative conformations*

The data items in these categories are as follows:

(*a*) ATOM_SITES_ALT

```
• _atom_sites_alt.id
  _atom_sites_alt.details
```

(*b*) ATOM_SITES_ALT_ENS

```
• _atom_sites_alt_ens.id
  _atom_sites_alt_ens.details
```

(*c*) ATOM_SITES_ALT_GEN

```
• _atom_sites_alt_gen.alt_id
        → _atom_sites_alt.id
• _atom_sites_alt_gen.ens_id
        → _atom_sites_alt_gen.ens_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.

Biological macromolecules are often very flexible, and as the resolution of a structure determination increases, it becomes increasingly possible to model reliably the alternative conformations that the structure adopts. Typically, partial occupancies are assigned to atom sites within the alternative conformations to indicate the relative frequency of occurrence of each conformation. It can, however, be difficult to deduce the possible different conformations of the whole structure from inspection of the atom-site occupancies alone. For instance, a segment of protein main chain might adopt one of three slightly different conformations, and within each conformation a particular side chain might adopt one of two possible conformations, one of which sterically distorts an adjacent residue sequence, while the other does not. The data model in the mmCIF dictionary allows these kinds of correlations in positions to be described.

The relationships between the categories used to describe alternative conformations are shown in Fig. 3.6.7.1.

In the core CIF dictionary, alternative conformations are indicated by using the `_atom_site.disorder_assembly` and `*.disorder_group` data items. Aliases to these data items are present in the mmCIF dictionary, but it is not intended that they should be used to describe disorder in a macromolecular structure.

The model for describing alternative conformations in mmCIF uses the ATOM_SITES_ALT family of categories. Ensembles of correlated alternative conformations can be identified using the category ATOM_SITES_ALT_ENS. Each ensemble is generated from one or more of the alternative conformations given in the list of alternative sites in the ATOM_SITES_ALT category. Data items in the

Example 3.6.7.3. *Alternative conformations in an HIV-1 protease structure (PDB 5HVP) described with data items in the ATOM_SITES_ALT, ATOM_SITES_ALT_ENS and ATOM_SITES_ALT_GEN categories.*

```
loop_
_atom_sites_alt.id
_atom_sites_alt.details
  .
; Atom sites with the alternative ID set to null are
  not modelled in alternative conformations
;
  1
; Atom sites with the alternative ID set to 1 have
  been modelled in alternative conformations with
  respect to atom sites marked with alternative
  ID 2. The conformations of amino-acid side chains
  with alternative ID set to 1 correlate with the
  conformation of the inhibitor marked with
  alternative ID 1. Atoms in these side chains have
  been given an occupancy of 0.58 to match the
  occupancy assigned to the inhibitor.
;
  2
; Atom sites with the alternative ID set to 2 have
  been modelled in alternative conformations with
  respect to atom sites marked with alternative
  ID 1. The conformations of amino-acid side chains
  with alternative ID set to 2 correlate with the
  conformation of the inhibitor marked with
  alternative ID 2. Atoms in these side chains have
  been given an occupancy of 0.42 to match the
  occupancy assigned to the inhibitor.
;

loop_
_atom_sites_alt_ens.id
_atom_sites_alt_ens.details
  'Ensemble 1'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  This conformational ensemble includes the more-
  populated conformation of the inhibitor (ID=1) and
  the amino-acid side chains that correlate with this
  inhibitor conformation.
;
  'Ensemble 2'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  This conformational ensemble includes the less-
  populated conformation of the inhibitor (ID=2) and
  the amino-acid side chains that correlate with this
  inhibitor conformation.
;

loop_
_atom_sites_alt_gen.ens_id
_atom_sites_alt_gen.alt_id
   'Ensemble 1'  .
   'Ensemble 1'  1
   'Ensemble 2'  .
   'Ensemble 2'  2
```

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_erf
  _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 macromolecule(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_erf` 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

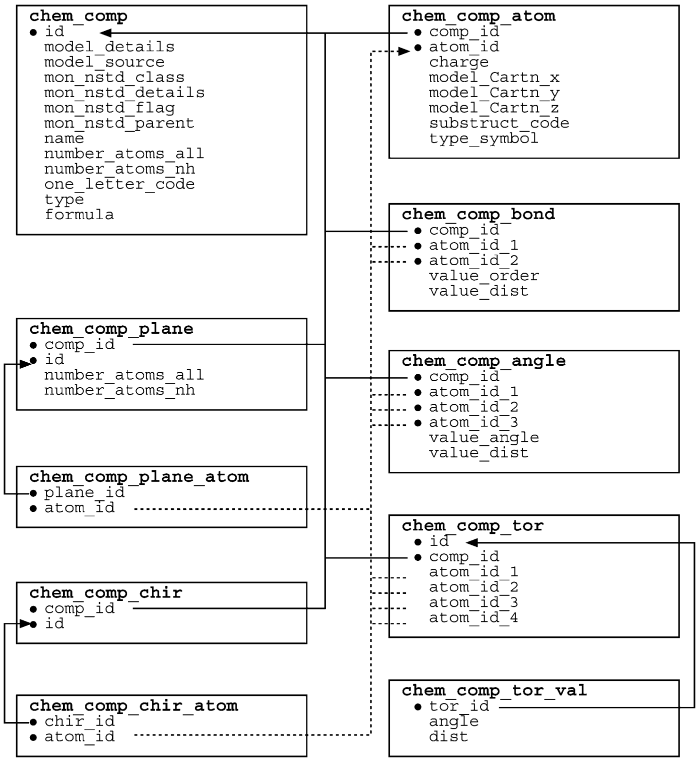

Fig. 3.6.7.3. The family of categories used to describe the chemical and structural features of the monomers and small molecules used to build a model of a structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships 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_erf** 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 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
```

(*j*) ENTITY_LINK

```
• _entity_link.link_id
        → _chem_link.id
  _entity_link.details
  _entity_link.entity_id_1
        → _entity.id
  _entity_link.entity_id_2
        → _entity.id
  _entity_link.entity_seq_num_1
        → _entity_poly_seq.num
  _entity_link.entity_seq_num_2
        → _entity_poly_seq.num
```

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.*

The geometry of the links between chemical components or entities can be described in the CHEM_LINK group of categories. Chemical components may be linked together according to the type of the component; defining the linking according to the type of the component rather than by each component in turn allows a type of polymer link for all the monomers in a polymer to be specified (*e.g.* L-peptide linking). The geometry of the links can be specified in the remaining CHEM_LINK categories. The relationships between categories used to describe links between chemical components are shown in Fig. 3.6.7.4, which also shows how information about the links is passed to the CHEM_COMP and CHEM_LINK categories. For simplicity, the categories CHEM_COMP_PLANE, CHEM_COMP_PLANE_ATOM, CHEM_COMP_CHIR, CHEM_COMP_CHIR_ATOM and ENTITY_LINK are not included in Fig. 3.6.7.4.

Note that this category group can be used to describe the links that connect the monomers within a macromolecular polymer (using the CHEM_LINK categories) and also the intramolecular links between separate molecules in the whole complex (using the ENTITY_LINK category). Intramolecular links, for example a covalent bond formed between a bound ligand and an amino-acid side chain, are usually discovered as a result of the structure determination, and it would therefore seem more appropriate to describe them in the STRUCT_CONN category. However, since one of the roles of the CHEM_LINK category group is to record target values used for restraints or constraints during the refinement of the model of the structure, ideal values for the geometry of any entity-to-entity links should be given here.

Data items in the CHEM_LINK category are used to assign a unique identifier to each link and allow the author to record any unusual aspects of each link. The other categories in the CHEM_LINK category group describe the geometric model of each link, and are closely analogous to the similarly named categories in the CHEM_COMP group.

The relationships among these categories are complex (see Fig. 3.6.7.4). Each atom that participates in an aspect of the link (for example, a bond, an angle, a chiral centre, a torsion angle or a plane) must be identified and it must also be specified whether the atom is in the first or second of the components that form the link.

Data items in the CHEM_LINK_BOND category describe the bonds between atoms participating in an intermolecular link between chemical components. Bond restraints may be described by the distance between the bonded atoms, the bond order or both.

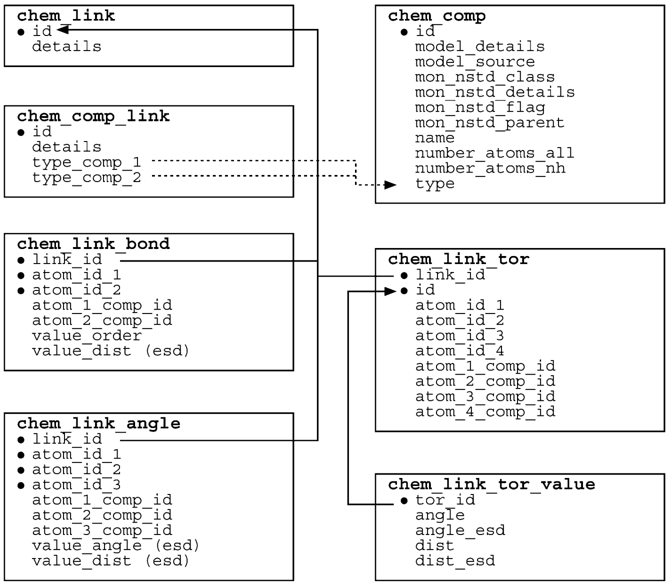

Fig. 3.6.7.4. The family of categories used to describe the links between chemical components. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

An angle at a link may be described in the CHEM_LINK_ANGLE category as either an angle at the vertex atom or as a distance between the atoms attached to the vertex. For data items in both the CHEM_LINK_BOND and CHEM_LINK_ANGLE categories, a target value and its associated standard uncertainty may be specified (Example 3.6.7.5).

Data items in the CHEM_LINK_CHIR category can be used to describe the conformation of chiral centres in a link between two chemical components. 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_LINK_CHIR_ATOM.

Data items in the CHEM_LINK_PLANE category can be used to list planes defined across a link between two chemical components. Because planes can involve a variable number of atoms, a separate list of the atoms should be given in CHEM_LINK_PLANE_ATOM.

Data items in the CHEM_LINK_TOR category can be used to give details of the torsion angles across a link between two chemical

Example 3.6.7.5. *A peptide bond described with data items in the CHEM_LINK_BOND and CHEM_LINK_ATOM categories.*

```
loop_
  _chem_link_bond.link_id
  _chem_link_bond.value_dist
  _chem_link_bond.value_dist_esd
  _chem_link_bond.atom_id_1
  _chem_link_bond.atom_1_comp_id
  _chem_link_bond.atom_id_2
  _chem_link_bond.atom_2_comp_id
    PEPTIDE  1.329  0.014  C  1  N  2

loop_
  _chem_link_angle.link_id
  _chem_link_angle.value_angle
  _chem_link_angle.value_angle_esd
  _chem_link_angle.atom_id_1
  _chem_link_angle.atom_1_comp_id
  _chem_link_angle.atom_id_2
  _chem_link_angle.atom_2_comp_id
  _chem_link_angle.atom_id_3
  _chem_link_angle.atom_3_comp_id
    PEPTIDE  116.2  2.0  CA 1  C  1  N  2
    PEPTIDE  123.0  1.6  O  1  C  1  N  2
    PEPTIDE  121.7  1.8  C  1  N  2  CA 2
```

components. The torsion angle may be described either as an angle or as a distance between the first and last atoms. As torsion angles can have more than one target value, the target values are specified in the CHEM_LINK_TOR_VALUE category.

The ENTITY_LINK category is used to identify the participants in links between distinct molecular entities. A pointer to the details of the link is given in `_entity_link.link_id`, which matches a value of `_chem_link.id` in the CHEM_LINK category.

3.6.7.3. Distinct chemical species

The categories describing distinct chemical entities are as follows:

ENTITY group
Entities (§3.6.7.3.1)
ENTITY
ENTITY_KEYWORDS
ENTITY_NAME_COM
ENTITY_NAME_SYS
ENTITY_SRC_GEN
ENTITY_SRC_NAT
Polymer entities (§3.6.7.3.2)
ENTITY_POLY
ENTITY_POLY_SEQ

The ENTITY categories of the mmCIF dictionary should be used in preference to the CHEMICAL categories of the core CIF dictionary. In a typical small-molecule structure determination, for which the core CIF dictionary was designed, the substance being studied can be thought of as a single chemical species, even if it contains distinct ions or ligands. In a macromolecular structure, it is more often the case that separate descriptions are appropriate for each of the distinct chemical species that comprise the structural complex. The ENTITY categories allow the species present and their basic chemical properties to be specified. Their structures and connectivity are described in other categories.

It is important, therefore, to remember that the ENTITY data do not represent the result of the crystallographic experiment; those results are given using the ATOM_SITE data items and are discussed and described using data items in the STRUCT family of categories. The ENTITY categories describe the chemistry of the molecules under investigation and are most usefully considered as the ideal groups to which the structure is restrained or constrained during refinement.

It is also important to remember that entities do not correspond directly to the total contents of the asymmetric unit. Entities are described only once, even in structures in which the entity occurs several times. The STRUCT_ASYM data items, which reference the list of entities, describe and label the contents of the asymmetric unit.

The following discussion treats the data items used for entities in general (Section 3.6.7.3.1) and those used more specifically to describe polymeric entities (Section 3.6.7.3.2) separately.

3.6.7.3.1. *Description of entities*

The data items in these categories are as follows:

(*a*) ENTITY
```
• _entity.id
  _entity.details
  _entity.formula_weight
  _entity.src_method
  _entity.type
```

(*b*) ENTITY_KEYWORDS
```
• _entity_keywords.entity_id
        → _entity.id
• _entity_keywords.text
```

(*c*) ENTITY_NAME_COM
```
• _entity_name_com.entity_id
        → _entity.id
• _entity_name_com.name
```

(*d*) ENTITY_NAME_SYS
```
• _entity_name_sys.entity_id
        → _entity.id
• _entity_name_sys.name
  _entity_name_sys.system
```

(*e*) ENTITY_SRC_GEN
```
• _entity_src_gen.entity_id
        → _entity.id
  _entity_src_gen.gene_src_common_name
  _entity_src_gen.gene_src_details
  _entity_src_gen.gene_src_genus
  _entity_src_gen.gene_src_species
  _entity_src_gen.gene_src_strain
  _entity_src_gen.gene_src_tissue
  _entity_src_gen.gene_src_tissue_fraction
  _entity_src_gen.host_org_common_name
  _entity_src_gen.host_org_details
  _entity_src_gen.host_org_genus
  _entity_src_gen.host_org_species
  _entity_src_gen.host_org_strain
  _entity_src_gen.plasmid_details
  _entity_src_gen.plasmid_name
```

(*f*) ENTITY_SRC_NAT
```
• _entity_src_nat.entity_id
        → _entity.id
  _entity_src_nat.common_name
  _entity_src_nat.details
  _entity_src_nat.genus
  _entity_src_nat.species
  _entity_src_nat.strain
  _entity_src_nat.tissue
  _entity_src_nat.tissue_fraction
```

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.

An entity in mmCIF is a chemically distinct molecular component of the structural complex described in the mmCIF. The three possible types of molecular entities are polymer, non-polymer and water. Note that the 'water' entity is water, and only water. Any other well ordered solvent molecules or ions should be treated as non-polymer entities. The relationships between categories used to describe the features of entities are shown in Fig. 3.6.7.5, which also shows how the information describing the entity is linked to the coordinate list in the ATOM_SITE category.

Data items in the ENTITY category are used to label each distinct chemical molecule with a reference code (`_entity.id`), to give the formula weight in daltons (if available) and to define the type of the entity as one of polymer, non-polymer or water. The method by which the entity was produced may be indicated using the item `_entity.src_method`, whose allowed values are `nat` (indicating that the sample was isolated from a natural source), `man` (indicating a genetically manipulated source) or `syn` (indicating a chemical synthesis). A value of `nat` indicates that additional details should be given in the ENTITY_SRC_NAT category and a value of `man` indicates that additional details should be given in the ENTITY_SRC_GEN category. As these flags are only relevant to the macromolecular entities of a structural complex, a value of '`.`', indicating 'inapplicable', should be given to `_entity.src_method` for solvent or water molecules. The `_entity.details` field can be used for a free-text description of any special features of the entity.

Keywords characterizing the individual molecular species may be given using data items in the ENTITY_KEYWORD category. These keywords should only be used to record information that does not depend on knowledge of the molecular structure. Thus a polypeptide could be described as a polypeptide, or an enzyme, or

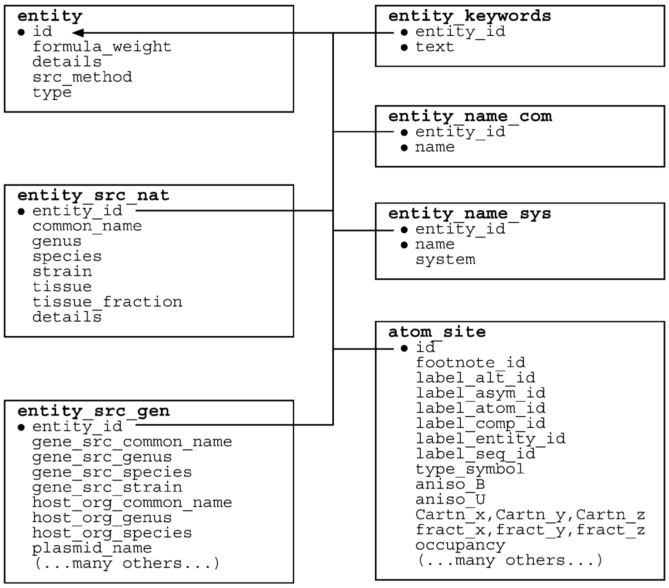

Fig. 3.6.7.5. The family of categories used to describe chemical entities. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the parent data item.

a protease, but it should not be described as an $\alpha\beta$-barrel; a number of categories within the STRUCT family allow keywords specific to the structure of the macromolecule to be given.

Data items in the ENTITY_NAME_COM category may be used to give any common names for an entity. Several different names can be recorded for each entity if appropriate.

Similarly, data items in the ENTITY_NAME_SYS category may be used to give systematic names for each entity. Again, several different names can be recorded for each entity if appropriate. The data item **_entity_name_sys.system** can be used to record the system according to which the systematic name was generated.

Example 3.6.7.6. *An example of the description of the entities in an HIV-1 protease structure (PDB 5HVP), described using data items in the* ENTITY, ENTITY_NAME_COM, ENTITY_NAME_SYS *and* ENTITY_SRC_GEN *categories.*

```
loop_
_entity.id
_entity.type
_entity.formula_weight
_entity.details
   1  polymer       10916
; The enzymatically competent form of HIV protease is
  a dimer. This entity corresponds to one monomer of
  an active dimer.
;
   2  non-polymer   647.2    .
   3  water            18    .

loop_
_entity_name_com.entity_id
_entity_name_com.name
   1  'HIV-1 protease monomer'
   1  'HIV-1 PR monomer'
   2  'acetyl-pepstatin'
   2  'acetyl-Ile-Val-Asp-Statine-Ala-Ile-Statine'
   3  'water'

_entity_name_sys.entity_id       1
_entity_name_sys.name            'EC 2.1.1.1'
_entity_name_sys.system          'Enzyme convention'

loop_
_entity_src_gen.entity_id
_entity_src_gen.gene_src_common_name
_entity_src_gen.gene_src_strain
_entity_src_gen.host_org_common_name
_entity_src_gen.host_org_genus
_entity_src_gen.host_org_species
_entity_src_gen.plasmid_name
1 'HIV-1' 'NY-5' 'bacteria' 'Escherichia' 'coli'
'pB322'
```

The ENTITY_SRC_GEN category allows a description of the source of entities produced by genetic manipulation to be given. There are data items for describing the tissue from which the gene was obtained, the plasmid into which it was incorporated for expression, and the host organism in which the macromolecule was expressed (Example 3.6.7.6).

The ENTITY_SRC_NAT category allows a description of the source of entities obtained from a natural tissue to be given. Data items are provided for the common and systematic name (by genus, species and, where relevant, strain) of the organism from which the material was obtained. Other data items can be used to describe the tissue (and if necessary the subcellular fraction of the tissue) from which the entity was isolated.

3.6.7.3.2. *Polymer entities*

The data items in these categories are as follows:

(*a*) ENTITY_POLY

```
● _entity_poly.entity_id
      → _entity.id
  _entity_poly.nstd_chirality
  _entity_poly.nstd_linkage
  _entity_poly.nstd_monomer
  _entity_poly.number_of_monomers
  _entity_poly.type
  _entity_poly.type_details
```

(*b*) ENTITY_POLY_SEQ

```
● _entity_poly_seq.entity_id
      → _entity.id
● _entity_poly_seq.mon_id
      → _chem_comp.id
● _entity_poly_seq.num
  _entity_poly_seq.hetero
```

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.

The polymer type, sequence length and information about any nonstandard features of the polymer may be specified using data items in the ENTITY_POLY category. The sequence of monomers in each polymer entity is given using data items in the ENTITY_POLY_SEQ category. The relationships between categories describing polymer entities are shown in Fig. 3.6.7.6, which also shows how the information describing the polymer is linked to the coordinate list in the ATOM_SITE category and to the full chemical description of each monomer or nonstandard monomer in the CHEM_COMP category.

Non-polymer entities are treated as individual chemical components, in the same way in which monomers within a polymer are treated as individual chemical components. They may be fully described in the CHEM_COMP group of categories (Example 3.6.7.7).

Data items in the ENTITY_POLY category can be used to give the number of monomers in the polymer and to assign the type of the polymer as one of the set of types `polypeptide(D)`, `polypeptide(L)`, `polydeoxyribonucleotide`, `polyribonucleotide`, `polysaccharide(D)`, `polysaccharide(L)` or `other`. Details of deviations from a standard type may be given in **_entity_poly.type_details**.

In some cases, the polymer is best described as one of the standard types even if it contains some nonstandard features. Flags are provided to indicate the presence of three types of nonstandard features. The presence of chiral centres other than those implied

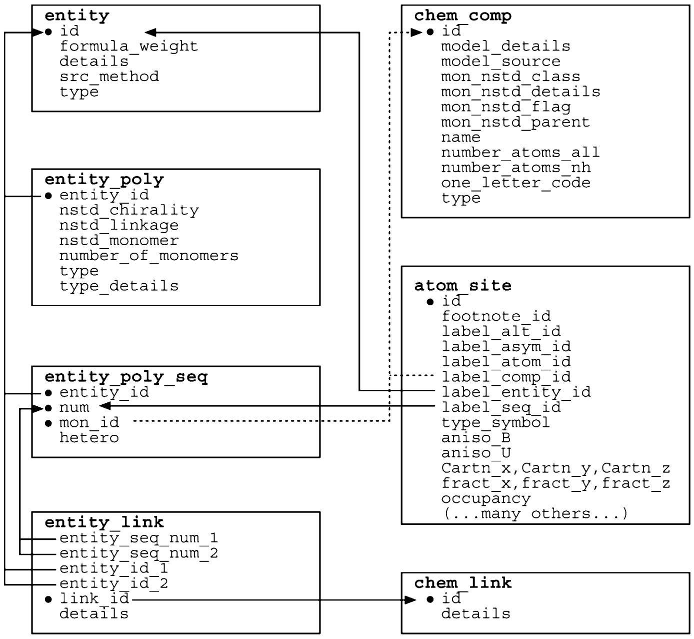

Fig. 3.6.7.6. The family of categories used to describe polymer chemical entities. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.7. *An example of both polymer and non-polymer entities in a drug–DNA complex (NDB DDF040) described with data items in the* ENTITY, ENTITY_KEYWORDS, ENTITY_NAME_COM, ENTITY_POLY *and* ENTITY_POLY_SEQ *categories (Narayana et al., 1991).*

```
loop_
_entity.id
_entity.type
_entity.src_method
   1  polymer     man
   2  non-polymer man
   3  water        .

loop_
_entity_keywords.entity_id
_entity_keywords.text
   1  'nucleic acid'
   2  'drug'

loop_
_entity_name_com.entity_id
_entity_name_com.name
   2  adriamycin
   3  water

loop_
_entity_poly.entity_id
_entity_poly.number_of_monomers
_entity_poly.type
   1  8  'polydeoxyribonucleotide'

loop_
_entity_poly_seq.entity_id
_entity_poly_seq.mon_id
_entity_poly_seq.num
   1  T  1
   1  G  2
   1  G  3
   1  C  4
   1  C  5
   1  A  6
# - - - abbreviated - - -
```

by the assigned type is indicated by assigning a value of `yes` to the data item **`_entity_poly.nstd_chirality`**. A value of `yes` for **`_entity_poly.nstd_linkage`** indicates the presence of monomer-to-monomer links different from those implied by the assigned type and a value of `yes` for **`_entity_poly.nstd_monomer`** indicates the presence of one or more nonstandard monomer components.

Data items in the ENTITY_POLY_SEQ category describe the sequence of monomers in a polymer. By including **`_entity_poly_seq.mon_id`** in the category key, it is possible to allow for sequence heterogeneity by allowing a given sequence number to be correlated with more than one monomer ID. Sequence heterogeneity is shown in the example of crambin in Section 3.6.3.

3.6.7.4. Molecular or packing geometry

The categories describing geometry are as follows:

GEOM group
 GEOM
 GEOM_ANGLE
 GEOM_BOND
 GEOM_CONTACT
 GEOM_HBOND
 GEOM_TORSION

The categories within the GEOM group are used in the core CIF dictionary to describe the geometry of the model that results from the structure determination, and can be used to select values that will be published in a report describing the structure. The complexity of macromolecular structures means that a different approach to presenting the results of a structure determination is needed. The STRUCT family of categories was created to meet this need. The GEOM categories are retained in the mmCIF dictionary, but only for consistency with the core CIF dictionary.

The data items in the categories in the GEOM group are:

(*a*) GEOM

```
● _geom.entry_id
        → _entry.id
  _geom.details (~ _geom_special_details)
```

(*b*) GEOM_ANGLE

```
● _geom_angle.atom_site_id_1
        (~ _geom_angle_atom_site_label_1)
● _geom_angle.atom_site_id_2
        (~ _geom_angle_atom_site_label_2)
● _geom_angle.atom_site_id_3
        (~ _geom_angle_atom_site_label_3)
● _geom_angle.site_symmetry_1
● _geom_angle.site_symmetry_2
● _geom_angle.site_symmetry_3
  _geom_angle.atom_site_auth_asym_id_1
        → _atom_site.auth_asym_id
  _geom_angle.atom_site_auth_atom_id_1
        → _atom_site.auth_atom_id
  _geom_angle.atom_site_auth_comp_id_1
        → _atom_site.auth_comp_id
  _geom_angle.atom_site_auth_seq_id_1
        → _atom_site.auth_seq_id
  _geom_angle.atom_site_auth_asym_id_2
        → _atom_site.auth_asym_id
  _geom_angle.atom_site_auth_atom_id_2
        → _atom_site.auth_atom_id
  _geom_angle.atom_site_auth_comp_id_2
        → _atom_site.auth_comp_id
  _geom_angle.atom_site_auth_seq_id_2
        → _atom_site.auth_seq_id
  _geom_angle.atom_site_auth_asym_id_3
        → _atom_site.auth_asym_id
  _geom_angle.atom_site_auth_atom_id_3
        → _atom_site.auth_atom_id
  _geom_angle.atom_site_auth_comp_id_3
        → _atom_site.auth_comp_id
  _geom_angle.atom_site_auth_seq_id_3
        → _atom_site.auth_seq_id
        → _atom_site.id
  _geom_angle.atom_site_label_alt_id_1
        → _atom_site.label_alt_id
  _geom_angle.atom_site_label_asym_id_1
        → _atom_site.label_asym_id
  _geom_angle.atom_site_label_atom_id_1
        → _atom_site.label_atom_id
```

_geom_angle.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_angle.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_angle.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_angle.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_angle.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_angle.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_angle.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_angle.atom_site_label_alt_id_3
→ _atom_site.label_alt_id
_geom_angle.atom_site_label_asym_id_3
→ _atom_site.label_asym_id
_geom_angle.atom_site_label_atom_id_3
→ _atom_site.label_atom_id
_geom_angle.atom_site_label_comp_id_3
→ _atom_site.label_comp_id
_geom_angle.atom_site_label_seq_id_3
→ _atom_site.label_seq_id
_geom_angle.publ_flag
+ *_geom_angle.value* (∼ *_geom_angle*)

(*c*) GEOM_BOND
• *_geom_bond.atom_site_id_1*
(∼ *_geom_bond_atom_site_label_1*)
• *_geom_bond.atom_site_id_2*
(∼ *_geom_bond_atom_site_label_2*)
• *_geom_bond.site_symmetry_1*
• *_geom_bond.site_symmetry_2*
_geom_bond.atom_site_auth_asym_id_1
→ _atom_site.auth_asym_id
_geom_bond.atom_site_auth_atom_id_1
→ _atom_site.auth_atom_id
_geom_bond.atom_site_auth_comp_id_1
→ _atom_site.auth_comp_id
_geom_bond.atom_site_auth_seq_id_1
→ _atom_site.auth_seq_id
_geom_bond.atom_site_auth_asym_id_2
→ _atom_site.auth_asym_id
_geom_bond.atom_site_auth_atom_id_2
→ _atom_site.auth_atom_id
_geom_bond.atom_site_auth_comp_id_2
→ _atom_site.auth_comp_id
_geom_bond.atom_site_auth_seq_id_2
→ _atom_site.auth_seq_id
→ _atom_site.id
_geom_bond.atom_site_label_alt_id_1
→ _atom_site.label_alt_id
_geom_bond.atom_site_label_asym_id_1
→ _atom_site.label_asym_id
_geom_bond.atom_site_label_atom_id_1
→ _atom_site.label_atom_id
_geom_bond.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_bond.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_bond.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_bond.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_bond.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_bond.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_bond.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
+ *_geom_bond.dist* (∼ *_geom_bond_distance*)
_geom_bond.publ_flag
_geom_bond.valence

(*d*) GEOM_CONTACT
• *_geom_contact.atom_site_id_1*
(∼ *_geom_contact_atom_site_label_1*)
• *_geom_contact.atom_site_id_2*
(∼ *_geom_contact_atom_site_label_2*)
• *_geom_contact.site_symmetry_1*
• *_geom_contact.site_symmetry_2*
_geom_contact.atom_site_auth_asym_id_1
→ _atom_site.auth_asym_id
_geom_contact.atom_site_auth_atom_id_1
→ _atom_site.auth_atom_id
_geom_contact.atom_site_auth_comp_id_1
→ _atom_site.auth_comp_id
_geom_contact.atom_site_auth_seq_id_1
→ _atom_site.auth_seq_id
_geom_contact.atom_site_auth_asym_id_2
→ _atom_site.auth_asym_id
_geom_contact.atom_site_auth_atom_id_2
→ _atom_site.auth_atom_id
_geom_contact.atom_site_auth_comp_id_2
→ _atom_site.auth_comp_id
_geom_contact.atom_site_auth_seq_id_2
→ _atom_site.auth_seq_id
→ _atom_site.id
_geom_contact.atom_site_label_alt_id_1
→ _atom_site.label_alt_id
_geom_contact.atom_site_label_asym_id_1
→ _atom_site.label_asym_id
_geom_contact.atom_site_label_atom_id_1
→ _atom_site.label_atom_id
_geom_contact.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_contact.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_contact.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_contact.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_contact.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_contact.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_contact.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
+ *_geom_contact.dist* (∼ *_geom_contact_distance*)
_geom_contact.publ_flag

(*e*) GEOM_HBOND
• *_geom_hbond.atom_site_id_A*
→ _atom_site.id
• *_geom_hbond.atom_site_id_D*
→ _atom_site.id
• *_geom_hbond.atom_site_id_H*
→ _atom_site.id
• *_geom_hbond.site_symmetry_A*
• *_geom_hbond.site_symmetry_D*
• *_geom_hbond.site_symmetry_H*
+ *_geom_hbond.angle_DHA*
_geom_hbond.atom_site_auth_asym_id_A
→ _atom_site.auth_asym_id
_geom_hbond.atom_site_auth_atom_id_A
→ _atom_site.auth_atom_id
_geom_hbond.atom_site_auth_comp_id_A
→ _atom_site.auth_comp_id
_geom_hbond.atom_site_auth_seq_id_A
→ _atom_site.auth_seq_id
_geom_hbond.atom_site_auth_asym_id_D
→ _atom_site.auth_asym_id
_geom_hbond.atom_site_auth_atom_id_D
→ _atom_site.auth_atom_id
_geom_hbond.atom_site_auth_comp_id_D
→ _atom_site.auth_comp_id
_geom_hbond.atom_site_auth_seq_id_D
→ _atom_site.auth_seq_id
_geom_hbond.atom_site_auth_asym_id_H
→ _atom_site.auth_asym_id
_geom_hbond.atom_site_auth_atom_id_H
→ _atom_site.auth_atom_id
_geom_hbond.atom_site_auth_comp_id_H
→ _atom_site.auth_comp_id
_geom_hbond.atom_site_auth_seq_id_H
→ _atom_site.auth_seq_id
_geom_hbond.atom_site_label_alt_id_A
→ _atom_site.label_alt_id
_geom_hbond.atom_site_label_asym_id_A
→ _atom_site.label_asym_id

_geom_hbond.atom_site_label_atom_id_A
→ _atom_site.label_atom_id
_geom_hbond.atom_site_label_comp_id_A
→ _atom_site.label_comp_id
_geom_hbond.atom_site_label_seq_id_A
→ _atom_site.label_seq_id
_geom_hbond.atom_site_label_alt_id_D
→ _atom_site.label_alt_id
_geom_hbond.atom_site_label_asym_id_D
→ _atom_site.label_asym_id
_geom_hbond.atom_site_label_atom_id_D
→ _atom_site.label_atom_id
_geom_hbond.atom_site_label_comp_id_D
→ _atom_site.label_comp_id
_geom_hbond.atom_site_label_seq_id_D
→ _atom_site.label_seq_id
_geom_hbond.atom_site_label_alt_id_H
→ _atom_site.label_alt_id
_geom_hbond.atom_site_label_asym_id_H
→ _atom_site.label_asym_id
_geom_hbond.atom_site_label_atom_id_H
→ _atom_site.label_atom_id
_geom_hbond.atom_site_label_comp_id_H
→ _atom_site.label_comp_id
_geom_hbond.atom_site_label_seq_id_H
→ _atom_site.label_seq_id
+ *_geom_hbond.dist_DA* (∼ *_geom_hbond_distance_DA*)
+ *_geom_hbond.dist_DH* (∼ *_geom_hbond_distance_DH*)
+ *_geom_hbond.dist_HA* (∼ *_geom_hbond_distance_HA*)
_geom_hbond.publ_flag

(*f*) GEOM_TORSION
● *_geom_torsion.atom_site_id_1*
(∼ *_geom_torsion_atom_site_label_1*)
● *_geom_torsion.atom_site_id_2*
(∼ *_geom_torsion_atom_site_label_2*)
● *_geom_torsion.atom_site_id_3*
(∼ *_geom_torsion_atom_site_label_3*)
● *_geom_torsion.atom_site_id_4*
(∼ *_geom_torsion_atom_site_label_4*)
● *_geom_torsion.site_symmetry_1*
● *_geom_torsion.site_symmetry_2*
● *_geom_torsion.site_symmetry_3*
● *_geom_torsion.site_symmetry_4*
_geom_torsion.atom_site_auth_asym_id_1
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_1
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_1
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_1
→ _atom_site.auth_seq_id
_geom_torsion.atom_site_auth_asym_id_2
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_2
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_2
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_2
→ _atom_site.auth_seq_id
_geom_torsion.atom_site_auth_asym_id_3
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_3
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_3
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_3
→ _atom_site.auth_seq_id
_geom_torsion.atom_site_auth_asym_id_4
→ _atom_site.auth_asym_id
_geom_torsion.atom_site_auth_atom_id_4
→ _atom_site.auth_atom_id
_geom_torsion.atom_site_auth_comp_id_4
→ _atom_site.auth_comp_id
_geom_torsion.atom_site_auth_seq_id_4
→ _atom_site.auth_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_1
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_1
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_1
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_1
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_1
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_2
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_2
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_2
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_2
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_2
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_3
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_3
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_3
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_3
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_3
→ _atom_site.label_seq_id
→ _atom_site.id
_geom_torsion.atom_site_label_alt_id_4
→ _atom_site.label_alt_id
_geom_torsion.atom_site_label_asym_id_4
→ _atom_site.label_asym_id
_geom_torsion.atom_site_label_atom_id_4
→ _atom_site.label_atom_id
_geom_torsion.atom_site_label_comp_id_4
→ _atom_site.label_comp_id
_geom_torsion.atom_site_label_seq_id_4
→ _atom_site.label_seq_id
_geom_torsion.publ_flag
+ *_geom_torsion.value* (∼ *_geom_torsion*)

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 (_) except where indicated by the ∼ symbol. 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.

3.6.7.5. Molecular structure

The categories describing molecular structure are as follows:
STRUCT group
Higher-level macromolecular structure (§3.6.7.5.1)
STRUCT
STRUCT_ASYM
STRUCT_BIOL
STRUCT_BIOL_GEN
STRUCT_BIOL_KEYWORDS
STRUCT_BIOL_VIEW
Secondary structure (§3.6.7.5.2)
STRUCT_CONF
STRUCT_CONF_TYPE
Structural interactions (§3.6.7.5.3)
STRUCT_CONN
STRUCT_CONN_TYPE
Structural features of monomers (§3.6.7.5.4)
STRUCT_MON_DETAILS
STRUCT_MON_NUCL
STRUCT_MON_PROT
STRUCT_MON_PROT_CIS
Noncrystallographic symmetry (§3.6.7.5.5)
STRUCT_NCS_DOM
STRUCT_NCS_DOM_LIM
STRUCT_NCS_ENS

STRUCT_NCS_ENS_GEN
STRUCT_NCS_OPER

External databases (§3.6.7.5.6)
STRUCT_REF
STRUCT_REF_SEQ
STRUCT_REF_SEQ_DIF

β-sheets (§3.6.7.5.7)
STRUCT_SHEET
STRUCT_SHEET_TOPOLOGY
STRUCT_SHEET_ORDER
STRUCT_SHEET_RANGE
STRUCT_SHEET_HBOND

Molecular sites (§3.6.7.5.8)
STRUCT_SITE_GEN
STRUCT_SITE_KEYWORDS
STRUCT_SITE_VIEW

The results of the determination of a structure can be described in mmCIF using data items in the categories contained in the STRUCT category group. This is a very large group of categories and it has been divided into eight groups of related categories for the discussions that follow: (1) those that describe the structure at the level of biologically relevant assemblies; (2) those that describe the secondary structure of the macromolecules present; (3) those that describe the structural interactions that determine the conformation of the macromolecules; (4) those that describe properties of the structure at the monomer level; (5) those that describe ensembles of identical domains related by noncrystallographic symmetry; (6) those that provide references to related entities in external databases; (7) those that describe the β-sheets present in the structure; and (8) those that provide detailed descriptions of the structure of biologically interesting molecular sites.

3.6.7.5.1. *Higher-level macromolecular structure*

The data items in these categories are as follows:

(*a*) STRUCT
```
• _struct.entry_id
        → _entry.id
  _struct.title
```

(*b*) STRUCT_ASYM
```
• _struct_asym.id
  _struct_asym.details
  _struct_asym.entity_id
        → _entity.id
```

(*c*) STRUCT_BIOL
```
• _struct_biol.id
  _struct_biol.details
```

(*d*) STRUCT_BIOL_GEN
```
• _struct_biol_gen.asym_id
        → _struct_asym.id
• _struct_biol_gen.biol_id
        → _struct_biol.id
• _struct_biol_gen.symmetry
  _struct_biol_gen.details
```

(*e*) STRUCT_BIOL_KEYWORDS
```
• _struct_biol_keywords.biol_id
        → _struct_biol.id
• _struct_biol_keywords.text
```

(*f*) STRUCT_BIOL_VIEW
```
• _struct_biol_view.biol_id
        → _struct_biol.id
• _struct_biol_view.id
  _struct_biol_view.details
  _struct_biol_view.rot_matrix[1][1]
  _struct_biol_view.rot_matrix[1][2]
  _struct_biol_view.rot_matrix[1][3]
  _struct_biol_view.rot_matrix[2][1]
  _struct_biol_view.rot_matrix[2][2]
  _struct_biol_view.rot_matrix[2][3]
  _struct_biol_view.rot_matrix[3][1]
  _struct_biol_view.rot_matrix[3][2]
  _struct_biol_view.rot_matrix[3][3]
```

(*g*) STRUCT_KEYWORDS
```
• _struct_keywords.entry_id
        → _entry.id
• _struct_keywords.text
```

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.

The data items in these categories serve two related but distinct purposes.

The first purpose is to label each of the entities in the asymmetric unit, using data items in the STRUCT_ASYM category. These labels become part of the category key that identifies each coordinate record and they are used extensively throughout the STRUCT family of categories, so care must be taken to select a labelling scheme that is concise and informative.

The second function is descriptive. The categories descending from STRUCT_BIOL allow the author of the mmCIF to identify and annotate the biologically relevant structural units found by the structure determination. What constitutes a biological unit can depend on the context. Take the case of a structure with two polymers related by noncrystallographic symmetry, each of which binds a small-molecule cofactor. If the author wishes to describe the dimer interface, the biological unit could be taken to be the two protein molecules. If the author wishes to highlight the cofactor binding mode, the biological unit could be taken to be one protein molecule and its bound cofactor. In this second case, there could be an additional biological unit of the second protein molecule and its bound cofactor, which may or may not be identical in conformation to the first.

The relationships between categories used to describe higher-level structure are illustrated in Fig. 3.6.7.7.

The STRUCT category serves to link the structure to the overall identifier for the data block, using `_struct.entry_id`, and to supply a title that describes the entire structure. The importance of this title as a succinct description of the structure should not be underestimated, and the author should express concisely but clearly in `_struct.title` the components of interest and the importance of this particular study. It is useful to think of this title as describing

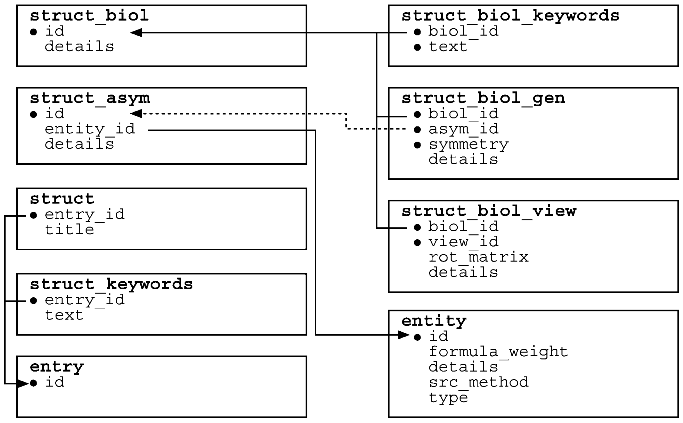

Fig. 3.6.7.7. The family of categories used to describe the higher-level macromolecular structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.8. *The higher-level structure of the complex of HIV-1 protease with an inhibitor (PDB 5HVP) described with data items in the STRUCT_ASYM, STRUCT_BIOL, STRUCT_BIOL_KEYWORDS and STRUCT_BIOL_GEN categories.*

```
loop_
_struct_asym.id
_struct_asym.entity_id
_struct_asym.details
   A  1  ’one monomer of the dimeric enzyme’
   B  1  ’one monomer of the dimeric enzyme’
   C  2
 ’one partially occupied position for the inhibitor’
   D  2
 ’one partially occupied position for the inhibitor’

loop_
_struct_biol.id
_struct_biol.details
   1
; significant deviations from twofold symmetry exist
  in this dimeric enzyme
;
   2
; The drug binds to this enzyme in two roughly
  twofold symmetric modes.

  Hence this biological unit (2) is roughly twofold
  symmetric to biological unit (3). Disorder in the
  protein chain indicated with alternative ID 1
  should be used with this biological unit.
;
   3
; The drug binds to this enzyme in two roughly
  twofold symmetric modes.

  Hence this biological unit (3) is roughly twofold
  symmetric to biological unit (2). Disorder in the
  protein chain indicated with alternative ID 2
  should be used with this biological unit.
;

loop_
_struct_biol_gen.biol_id
_struct_biol_gen.asym_id
_struct_biol_gen.symmetry
   1  A  1_555     1  B  1_555
   2  A  1_555     2  B  1_555     2  C  1_555
   3  A  1_555     3  B  1_555     3  D  1_555
```

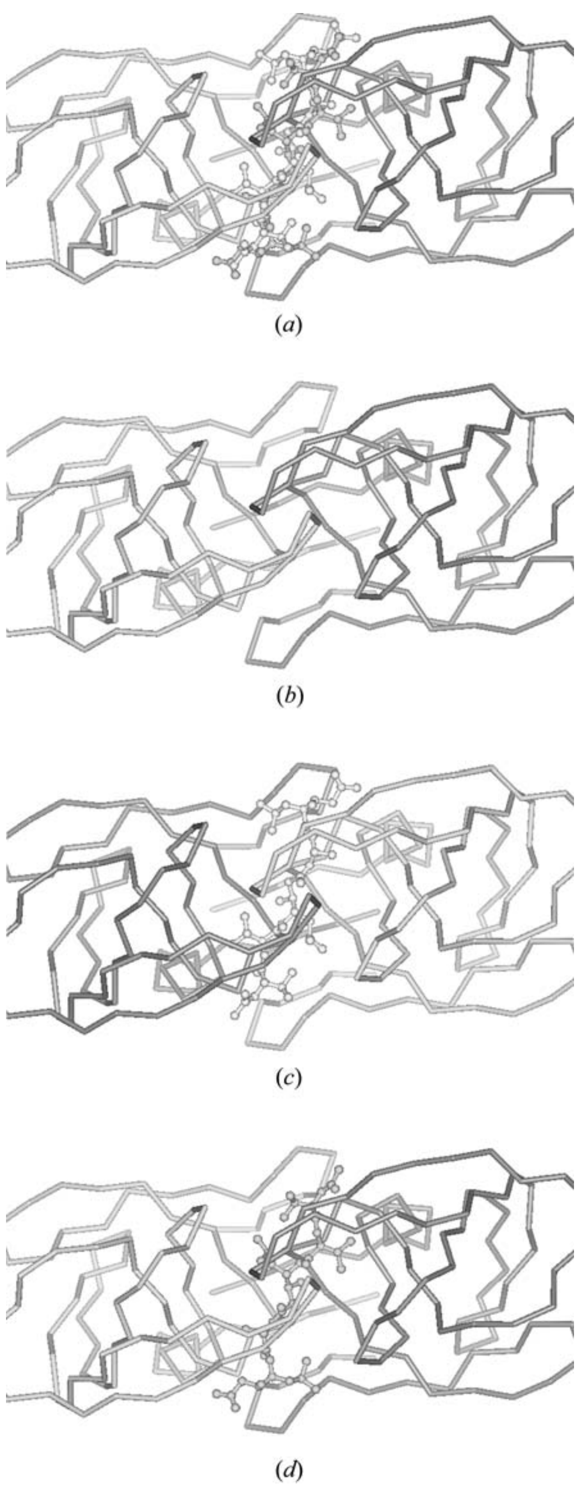

Fig. 3.6.7.8. The higher-level structure of the complex of HIV-1 protease with an inhibitor (PDB 5HVP) to be described with data items in the STRUCT_ASYM, STRUCT_BIOL, STRUCT_BIOL_KEYWORDS and STRUCT_BIOL_GEN categories. (*a*) Complete structure; (*b*), (*c*), (*d*) three different biological units.

the motivation for the structure determination, rather than the result. For instance, if the goal of the study was to determine the structure of enzyme A at pH 7.2 as part of a study of the mechanism of the reaction catalysed by the enzyme, an appropriate value for **_struct.title** would be ‘Enzyme A at pH 7.2’, even if the structure was found to contain two molecules per asymmetric unit, a bound calcium ion and a disordered loop between residues 47 and 52.

The STRUCT_KEYWORDS category allows an author to include keywords for the structure that has been determined. Other categories, such as STRUCT_BIOL_KEYWORDS and STRUCT_SITE_KEYWORDS, allow more specific keywords to be given, but the STRUCT_KEYWORDS category is the most likely category to be searched by simple information retrieval applications, so the author of an mmCIF might want to duplicate any keywords given elsewhere in the mmCIF in STRUCT_KEYWORDS as well.

The chemical entities that form the contents of the asymmetric unit are identified using data items in the ENTITY categories. The data items in the STRUCT_ASYM category link these entities to the structure itself. A unique identifier is attached to each occurrence of each entity in the asymmetric unit using **_struct_asym.id**. This identifier forms a part of the atom label in the ATOM_SITE category, which is used throughout the many categories in the STRUCT group in describing the structure. The identifier is also used in generating biological assemblies.

The usual reason for determining the structure of a biological macromolecule is to get information about the biologically relevant assemblies of the entities in the crystal structure. These assemblies take many forms and could encompass the complete contents of the asymmetric unit, a fraction of the contents of the asymmetric unit or the contents of more than one asymmetric unit. Each assembly, or ‘biological unit’, is given an identifier in the STRUCT_BIOL category and the author may annotate each biological unit using the data item **_struct_biol.details**. Keywords for each biological unit can be given using data items in the STRUCT_BIOL_KEYWORD category.

The entities that comprise the biological unit are specified using data items in the STRUCT_BIOL_GEN category by reference to the appropriate values of **`_struct_asym.id`** and by specifying any symmetry transformation that must be applied to the entities to generate the biological unit.

Data items in the STRUCT_BIOL_VIEW category allow the author to specify an orientation of the biological unit that provides a useful view of the structure. The comments given in **`_struct_biol_view.details`** may be used as a figure caption if the view is intended to be a figure in a report describing the structure.

The example of crambin in Section 3.6.3 shows the relations between the categories defining higher-level structure for the straightforward case of a single protein molecule (with a small co-crystallization molecule and solvent) in the asymmetric unit. The structure of HIV-1 protease with a bound inhibitor (PDB 5HVP), shown in Example 3.6.7.8, is considerably more complex. There are two entities: the monomeric form of the enzyme and the small-molecule inhibitor. The asymmetric unit contains two copies of the enzyme monomer (both fully occupied) and two copies of the inhibitor (each of which is partially occupied) (Fig. 3.6.7.8). Three biological assemblies are constructed for this system. One biological unit contains only the dimeric enzyme (Fig. 3.6.7.8*b*), the second contains the dimeric enzyme with one partially occupied conformation of the inhibitor (Fig. 3.6.7.8*c*) and the third contains the dimeric enzyme with the second partially occupied conformation of the inhibitor (Fig. 3.6.7.8*d*). There are alternative conformations of the side chains in the enzyme that correlate with the binding mode of the inhibitor.

3.6.7.5.2. *Secondary structure*

The data items in these categories are as follows:

(*a*) STRUCT_CONF_TYPE

```
• _struct_conf_type.id
  _struct_conf_type.criteria
  _struct_conf_type.reference
```

(*b*) STRUCT_CONF

```
• _struct_conf.id
  _struct_conf.beg_label_asym_id
       → _atom_site.label_asym_id
  _struct_conf.beg_label_comp_id
       → _atom_site.label_comp_id
  _struct_conf.beg_label_seq_id
       → _atom_site.label_seq_id
  _struct_conf.beg_auth_asym_id
       → _atom_site.auth_asym_id
  _struct_conf.beg_auth_comp_id
       → _atom_site.auth_comp_id
  _struct_conf.beg_auth_seq_id
       → _atom_site.auth_seq_id
  _struct_conf.conf_type_id
       → _struct_conf_type.id
  _struct_conf.details
  _struct_conf.end_label_asym_id
       → _atom_site.label_asym_id
  _struct_conf.end_label_comp_id
       → _atom_site.label_comp_id
  _struct_conf.end_label_seq_id
       → _atom_site.label_seq_id
  _struct_conf.end_auth_asym_id
       → _atom_site.auth_asym_id
  _struct_conf.end_auth_comp_id
       → _atom_site.auth_comp_id
  _struct_conf.end_auth_seq_id
       → _atom_site.auth_seq_id
```

The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.

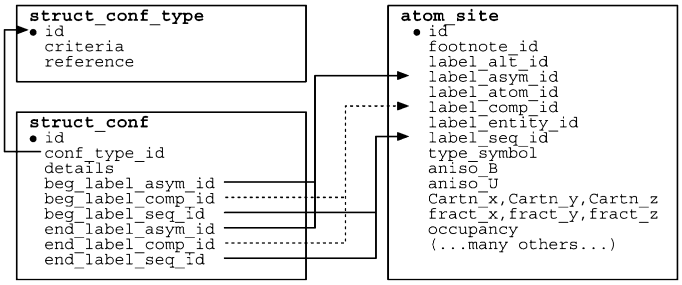

Fig. 3.6.7.9. The family of categories used to describe secondary structure. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.9. *Secondary structure in an HIV-1 protease structure (PDB 5HVP) described with data items in the STRUCT_CONF_TYPE and STRUCT_CONF categories.*

```
loop_
_struct_conf_type.id
_struct_conf_type.criteria
   HELX_RH_AL_P  'author judgement'
   STRN          'author judgement'
   TURN_TY1_P    'author judgement'
   TURN_TY1P_P   'author judgement'
   TURN_TY2_P    'author judgement'
   TURN_TY2P_P   'author judgement'

loop_
_struct_conf.id
_struct_conf.conf_type_id
_struct_conf.beg_label_comp_id
_struct_conf.beg_label_asym_id
_struct_conf.beg_label_seq_id
_struct_conf.end_label_comp_id
_struct_conf.end_label_asym_id
_struct_conf.end_label_seq_id
   HELX1  HELX_RH_AL_P  ARG  A   87  GLN  A   92
   HELX2  HELX_RH_AL_P  ARG  B  287  GLN  B  292
   STRN1  STRN          PRO  A    1  LEU  A    5
   STRN2  STRN          CYS  B  295  PHE  B  299
   STRN3  STRN          CYS  A   95  PHE  A  299
   STRN4  STRN          PRO  B  201  LEU  B  205
   TURN1  TURN_TY1P_P   ILE  A   15  GLN  A   18
   TURN2  TURN_TY2_P    GLY  A   49  GLY  A   52
   TURN3  TURN_TY1P_P   ILE  A   55  HIS  A   69
   TURN4  TURN_TY1_P    THR  A   91  GLY  A   94
```

The primary structure of a macromolecule is defined by the sequence of the components (amino acids, nucleic acids or sugars) in the polymer chain. The polymer chains assume conformations based on the torsion angles adopted by the rotatable bonds in the polymer backbone; the resulting conformations are referred to as the secondary structure of the polymer. Several patterns of values of backbone torsion angles have been described and given names, such as α-helix, β-strand, turn and coil for proteins, and A-, B- and Z-helix for nucleic acids.

In the mmCIF dictionary, these secondary structures are described in the STRUCT_CONF and STRUCT_CONF_TYPE categories. Note that the data items in these categories describe only the secondary structure; the tertiary organization of β-strands into β-sheets is described in the STRUCT_SHEET_* categories. There are no data items for describing the tertiary organization of α-helices or nucleic acids in the current version of the mmCIF dictionary.

The relationships between categories used to describe secondary structure are shown in Fig. 3.6.7.9.

The type of the secondary structure is specified in the STRUCT_CONF_TYPE category, along with the criteria used to identify it. The range of monomers assigned to each secondary-structure element is given in the STRUCT_CONF category.

The allowed values for the data item **_struct_conf_type.id** cover most types of protein and nucleic acid secondary structure (Example 3.6.7.9). The criteria that define the secondary structure may be given using the data item **_struct_conf_type.criteria**. **_struct_conf_type.reference** can be used to specify a reference to the literature in which the criteria are explained in more detail.

The residues that define the beginning and end of each region of secondary structure are identified with the appropriate ***_asym**, ***_comp** and ***_seq** identifiers. The standard labelling system or the author's alternative labelling system may be used. The identification of the residues assigned to each region of secondary structure is linked to the labelling information in the ATOM_SITE category. Unusual features of a conformation may be described using **_struct_conf.details**.

3.6.7.5.3. *Structural interactions*

The data items in these categories are as follows:

(*a*) STRUCT_CONN_TYPE

```
• _struct_conn_type.id
  _struct_conn_type.criteria
  _struct_conn_type.reference
```

(*b*) STRUCT_CONN

```
• _struct_conn.id
  _struct_conn.conn_type_id
        → _struct_conn_type.id
  _struct_conn.details
  _struct_conn.ptnr1_label_alt_id
        → _atom_sites_alt.id
  _struct_conn.ptnr1_label_asym_id
        → _atom_site.label_asym_id
  _struct_conn.ptnr1_label_atom_id
        → _chem_comp_atom.atom_id
  _struct_conn.ptnr1_label_comp_id
        → _atom_site.label_comp_id
  _struct_conn.ptnr1_label_seq_id
        → _atom_site.label_seq_id
  _struct_conn.ptnr1_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_conn.ptnr1_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_conn.ptnr1_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_conn.ptnr1_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_conn.ptnr1_role
  _struct_conn.ptnr1_symmetry
  _struct_conn.ptnr2_label_alt_id
        → _atom_sites_alt.id
  _struct_conn.ptnr2_label_asym_id
        → _atom_site.label_asym_id
  _struct_conn.ptnr2_label_atom_id
        → _chem_comp_atom.atom_id
  _struct_conn.ptnr2_label_comp_id
        → _atom_site.label_comp_id
  _struct_conn.ptnr2_label_seq_id
        → _atom_site.label_seq_id
  _struct_conn.ptnr2_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_conn.ptnr2_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_conn.ptnr2_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_conn.ptnr2_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_conn.ptnr2_role
  _struct_conn.ptnr2_symmetry
```

The bullet (•) indicates a category key. The arrow (→) is a reference to a parent data item.

The structural interactions that are described with data items in the STRUCT_CONN family of categories are the tertiary result of a structure determination, not the chemical connectivity of the components of the structure. In general, the interactions described using the STRUCT_CONN data items are noncovalent, such as hydrogen bonds, salt bridges and metal coordination.

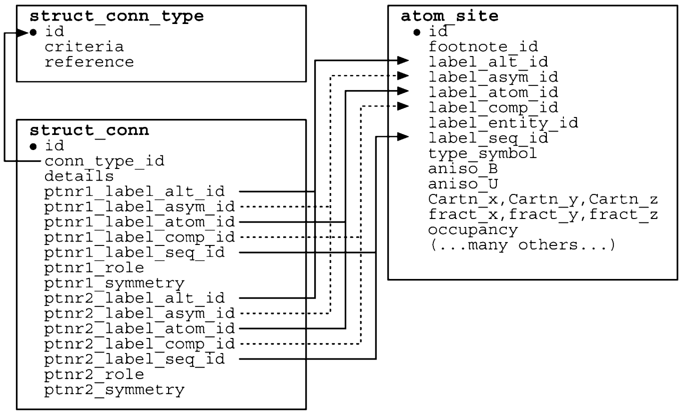

Fig. 3.6.7.10. The family of categories used to describe structural interactions such as hydrogen bonding, salt bridges and disulfide bridges. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

It is useful to think of the structure interactions given in CHEM_COMP_BOND, CHEM_LINK and ENTITY_LINK as the covalent interactions that are known in advance of the structure determination because the chemistry of the components is well defined. Literature or calculated values for these interactions are often used as restraints during the refinement. In contrast, the structural interactions described in the STRUCT_CONN family of categories are not known in advance and are part of the results of the structure determination.

This distinction only holds approximately, as there are clearly bonds, such as disulfide links, that are covalent and usually restrained during the refinement but that are also a result of the folding of the protein revealed by the structure determination, and thus should be described using STRUCT_CONN data items.

In general, the STRUCT_CONN data items would not be used to list all the structure interactions. Instead, the author of the mmCIF would use the STRUCT_CONN data items to identify and annotate only the structural interactions worthy of discussion. The relationships between categories used to describe structural interactions are shown in Fig. 3.6.7.10.

Structural interactions such as hydrogen bonds, salt bridges and disulfide bridges can be described in the STRUCT_CONN category. The type of each interaction and the criteria used to identify the interaction can be specified in the STRUCT_CONN_TYPE category (Example 3.6.7.10).

The atoms participating in each interaction are arbitrarily labelled as 'partner 1' and 'partner 2'. Each is identified by the ***_alt**, ***_asym**, ***_atom**, ***_comp** and ***_seq** constituents of the corresponding atom-site label. The role of each partner in the interaction (*e.g.* donor, acceptor) may be specified, and any crystallographic symmetry operation needed to transform the atom from the position given in the ATOM_SITE list to the position where the interaction occurs can be given. The atoms participating in the interaction may also be identified using an alternative labelling scheme if the author has supplied one.

Unusual aspects of the interaction may be discussed in **_struct_conn.details**. The general type of an interaction can be indicated using **_struct_conn.conn_type_id**, which references one of the standard types described using data items in the STRUCT_CONN_TYPE category.

The specific types of structural connection that may be recorded are those allowed for **_struct_conn_type.id**, namely covalent and hydrogen bonds, ionic (salt-bridge) interactions, disulfide

Example 3.6.7.10. *A hypothetical salt bridge and hydrogen bond described with data items in the* STRUCT_CONN_TYPE *and* STRUCT_CONN *categories.*

```
loop_
_struct_conn_type.id
_struct_conn_type.criteria
   saltbr
; negative to positive distance > 2.5 Angstroms,
  < 3.2 Angstroms
;
   hydrog
; N-O distance > 2.5 Angstroms, < 3.5 Angstroms,
  N-O-C  angle < 120 degrees
;

loop_
_struct_conn.id
_struct_conn.conn_type_id
_struct_conn.ptnr1_label_comp_id
_struct_conn.ptnr1_label_asym_id
_struct_conn.ptnr1_label_seq_id
_struct_conn.ptnr1_label_atom_id
_struct_conn.ptnr1_role
_struct_conn.ptnr1_symmetry
_struct_conn.ptnr2_label_comp_id
_struct_conn.ptnr2_label_asym_id
_struct_conn.ptnr2_label_seq_id
_struct_conn.ptnr2_label_atom_id
_struct_conn.ptnr2_role
_struct_conn.ptnr2_symmetry
  C1 saltbr ARG  A  87 NZ1 positive 1_555
            GLU  A  92 OE1 negative 1_555
  C2 hydrog ARG  B 287 N   donor    1_555
            GLY  B 292 O   acceptor 1_555
```

links, metal coordination, mismatched base pairs, covalent residue modifications and covalent modifications of nucleotide bases, sugars or phosphates. The criteria used to define each interaction may be described in detail using `_struct_conn_type.criteria` or a literature reference to the criteria can be given in `_struct_conn_type.reference`.

3.6.7.5.4. *Structural features of monomers*

The data items in these categories are as follows:

(*a*) STRUCT_MON_DETAILS

```
• _struct_mon_details.entry_id
        → _entry.id
  _struct_mon_details.prot_cis
  _struct_mon_details.RSCC
  _struct_mon_details.RSR
```

(*b*) STRUCT_MON_NUCL

```
• _struct_mon_nucl.label_alt_id
        → _atom_sites_alt.id
• _struct_mon_nucl.label_asym_id
        → _atom_site.label_asym_id
• _struct_mon_nucl.label_comp_id
        → _atom_site.label_comp_id
• _struct_mon_nucl.label_seq_id
        → _atom_site.label_seq_id
  _struct_mon_nucl.alpha
  _struct_mon_nucl.auth_asym_id
        → _atom_site.auth_asym_id
  _struct_mon_nucl.auth_comp_id
        → _atom_site.auth_comp_id
  _struct_mon_nucl.auth_seq_id
        → _atom_site.auth_seq_id
  _struct_mon_nucl.beta
  _struct_mon_nucl.chi1
  _struct_mon_nucl.chi2
  _struct_mon_nucl.delta
  _struct_mon_nucl.details
  _struct_mon_nucl.epsilon
  _struct_mon_nucl.gamma
  _struct_mon_nucl.mean_B_all
  _struct_mon_nucl.mean_B_base
  _struct_mon_nucl.mean_B_phos
  _struct_mon_nucl.mean_B_sugar
  _struct_mon_nucl.nu0
  _struct_mon_nucl.nu1
  _struct_mon_nucl.nu2
  _struct_mon_nucl.nu3
  _struct_mon_nucl.nu4
  _struct_mon_nucl.P
  _struct_mon_nucl.RSCC_all
  _struct_mon_nucl.RSCC_base
  _struct_mon_nucl.RSCC_phos
  _struct_mon_nucl.RSCC_sugar
  _struct_mon_nucl.RSR_all
  _struct_mon_nucl.RSR_base
  _struct_mon_nucl.RSR_phos
  _struct_mon_nucl.RSR_sugar
  _struct_mon_nucl.tau0
  _struct_mon_nucl.tau1
  _struct_mon_nucl.tau2
  _struct_mon_nucl.tau3
  _struct_mon_nucl.tau4
  _struct_mon_nucl.taum
  _struct_mon_nucl.zeta
```

(*c*) STRUCT_MON_PROT

```
• _struct_mon_prot.label_alt_id
        → _atom_sites_alt.id
• _struct_mon_prot.label_asym_id
        → _atom_site.label_asym_id
• _struct_mon_prot.label_comp_id
        → _atom_site.label_comp_id
• _struct_mon_prot.label_seq_id
        → _atom_site.label_seq_id
  _struct_mon_prot.auth_asym_id
        → _atom_site.auth_asym_id
  _struct_mon_prot.auth_comp_id
        → _atom_site.auth_comp_id
  _struct_mon_prot.auth_seq_id
        → _atom_site.auth_seq_id
  _struct_mon_prot.chi1
  _struct_mon_prot.chi2
  _struct_mon_prot.chi3
  _struct_mon_prot.chi4
  _struct_mon_prot.chi5
  _struct_mon_prot.details
  _struct_mon_prot.RSCC_all
  _struct_mon_prot.RSCC_main
  _struct_mon_prot.RSCC_side
  _struct_mon_prot.RSR_all
  _struct_mon_prot.RSR_main
  _struct_mon_prot.RSR_side
  _struct_mon_prot.mean_B_all
  _struct_mon_prot.mean_B_main
  _struct_mon_prot.mean_B_side
  _struct_mon_prot.omega
  _struct_mon_prot.phi
  _struct_mon_prot.psi
```

(*d*) STRUCT_MON_PROT_CIS

```
• _struct_mon_prot_cis.label_alt_id
        → _atom_sites_alt.id
• _struct_mon_prot_cis.label_asym_id
        → _atom_site.label_asym_id
• _struct_mon_prot_cis.label_comp_id
        → _atom_site.label_comp_id
• _struct_mon_prot_cis.label_seq_id
        → _atom_site.label_seq_id
  _struct_mon_prot_cis.auth_asym_id
        → _atom_site.auth_asym_id
  _struct_mon_prot_cis.auth_comp_id
        → _atom_site.auth_comp_id
  _struct_mon_prot_cis.auth_seq_id
        → _atom_site.auth_seq_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.

Most macromolecules have complex structures which contain regions of well defined structure and flexible regions that are difficult to model accurately. Overall measures of the quality of a model, such as the standard crystallographic *R* factors, do not represent the local quality of the model. During the development of

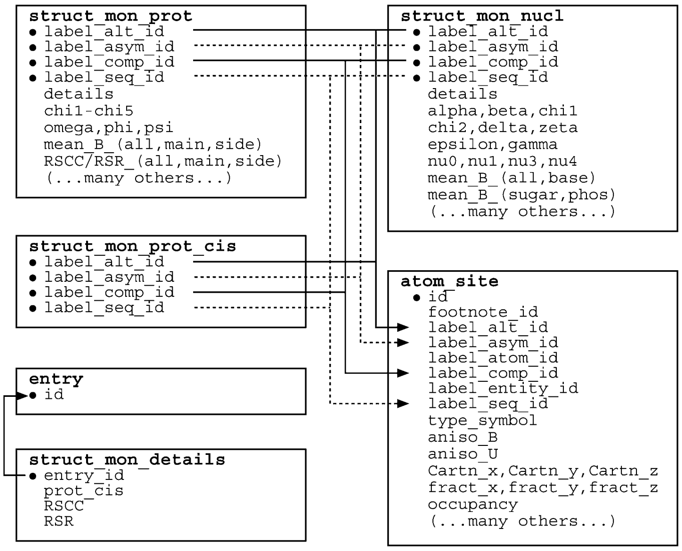

Fig. 3.6.7.11. The family of categories used to describe the structural features of monomers. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

the mmCIF dictionary, it was found that the biological crystallography community felt that mmCIF should contain data items that allowed the local quality of the model to be recorded: these data items are found in the categories STRUCT_MON_DETAILS, STRUCT_MON_NUCL (for nucleotides), and STRUCT_MON_PROT and STRUCT_MON_PROT_CIS (for proteins). Using these categories, quantities that reflect the local quality of the structure, such as isotropic displacement factors, real-space R factors and real-space correlation coefficients, can be given at the monomer and submonomer levels.

In addition, these categories can be used to record the conformation of the structure at the monomer level by listing side-chain torsion angles. These values can be derived from the atom coordinate list, so it would not be common practice to include them in an mmCIF for archiving a structure unless it was to highlight conformations that deviate significantly from expected values (Engh & Huber, 1991). However, there are applications, such as comparative studies across a number of independent determinations of the same structure, where it would be useful to store torsion-angle information without having to recalculate it each time it is needed.

The relationships between the categories used to describe the structural features of monomers are shown in Fig. 3.6.7.11.

Three indicators of the quality of a structure at the local level are included in this version of the dictionary: the mean displacement (B) factor, the real-space correlation coefficient (Jones *et al.*, 1991) and the real-space R factor (Brändén & Jones, 1990). Other indicators are likely to be added as they become available. In the current version of the dictionary, these metrics can be given at the monomer level, or at the levels of main- and side-chain for proteins, or base, phosphate and sugar for nucleic acids (Altona & Sundaralingam, 1972).

The variables used when calculating real-space correlation coefficients and real-space R factors, such as the coefficients used to calculate the map being evaluated or the radii used for including points in a calculation, can be recorded using the data items **_struct_mon_details.RSC** and **_struct_mon_details.RSR**.

Example 3.6.7.11. *A hypothetical example of the structural features of a single protein residue described with data items in the STRUCT_MON_PROT category.*

```
_struct_mon_prot.label_comp_id         ARG
_struct_mon_prot.label_seq_id          35
_struct_mon_prot.label_asym_id         A
_struct_mon_prot.label_alt_id          .
_struct_mon_prot.chi1               -67.9
_struct_mon_prot.chi2              -174.7
_struct_mon_prot.chi3               -67.7
_struct_mon_prot.chi4               -86.3
_struct_mon_prot.chi5                 4.2
_struct_mon_prot.RSCC_all             0.90
_struct_mon_prot.RSR_all              0.18
_struct_mon_prot.mean_B_all          30.0
_struct_mon_prot.mean_B_main         25.0
_struct_mon_prot.mean_B_side         35.1
_struct_mon_prot.omega              180.1
_struct_mon_prot.phi                -60.3
_struct_mon_prot.psi                -46.0
```

These data items are also provided for recording the full conformation of the macromolecule, using a full set of data items for the torsion angles of both proteins and nucleic acids. Although one could use these data items to describe the whole macromolecule, it is more likely that they would be used to highlight regions of the structure that deviate from expected values (Example 3.6.7.11). Deviations from expected values could imply inaccuracies in the model in poorly defined parts of the structure, but in some cases nonstandard torsion angles are found in very well defined regions and are essential to the proper configurations of active sites or ligand binding pockets.

A special case of nonstandard conformation is the occurrence of *cis* peptides in proteins. As the *cis* conformation occurs quite often, the category STRUCT_MON_PROT_CIS is provided so that an explicit list can be made of *cis* peptides. The related data item **_struct_mon_details.prot_cis** allows an author to specify how far a peptide torsion angle can deviate from the expected value of 0.0 and still be considered to be *cis*.

In these categories, properties are listed by residue rather than by individual atom. The only label components needed to identify the residue are ***_alt**, ***_asym**, ***_comp** and ***_seq**. If the author has provided an alternative labelling system, this can also be used. Since the analysis is by individual residue, there is no need to specify symmetry operations that might be needed to move one residue so that it is next to another.

3.6.7.5.5. *Noncrystallographic symmetry*

Data items in these categories are as follows:

(*a*) STRUCT_NCS_ENS

```
• _struct_ncs_ens.id
  _struct_ncs_ens.details
  _struct_ncs_ens.point_group
```

(*b*) STRUCT_NCS_ENS_GEN

```
• _struct_ncs_ens_gen.dom_id_1
        → _struct_ncs_dom.id
• _struct_ncs_ens_gen.dom_id_2
        → _struct_ncs_dom.id
• _struct_ncs_ens_gen.ens_id
        → _struct_ncs_ens.id
• _struct_ncs_ens_gen.oper_id
        → _struct_ncs_oper.id
```

(*c*) STRUCT_NCS_DOM

```
• _struct_ncs_dom.id
  _struct_ncs_dom.details
```

(*d*) STRUCT_NCS_DOM_LIM

```
• _struct_ncs_dom_lim.beg_label_alt_id
        → _atom_sites_alt.id
• _struct_ncs_dom_lim.beg_label_asym_id
        → _atom_site.label_asym_id
• _struct_ncs_dom_lim.beg_label_comp_id
        → _atom_site.label_comp_id
```

```
• _struct_ncs_dom_lim.beg_label_seq_id
        → _atom_site.label_seq_id
• _struct_ncs_dom_lim.dom_id
• _struct_ncs_dom_lim.end_label_alt_id
        → _atom_sites_alt.id
• _struct_ncs_dom_lim.end_label_asym_id
        → _atom_site.label_asym_id
• _struct_ncs_dom_lim.end_label_comp_id
        → _atom_site.label_comp_id
• _struct_ncs_dom_lim.end_label_seq_id
        → _atom_site.label_seq_id
  _struct_ncs_dom_lim.beg_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_ncs_dom_lim.beg_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_ncs_dom_lim.beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_ncs_dom_lim.end_auth_asym_id
        → _atom_site.auth_asym_id
  _struct_ncs_dom_lim.end_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_ncs_dom_lim.end_auth_seq_id
        → _atom_site.auth_seq_id
```

(*e*) STRUCT_NCS_OPER

```
• _struct_ncs_oper.id
  _struct_ncs_oper.code
  _struct_ncs_oper.details
  _struct_ncs_oper.matrix[1][1]
  _struct_ncs_oper.matrix[1][2]
  _struct_ncs_oper.matrix[1][3]
  _struct_ncs_oper.matrix[2][1]
  _struct_ncs_oper.matrix[2][2]
  _struct_ncs_oper.matrix[2][3]
  _struct_ncs_oper.matrix[3][1]
  _struct_ncs_oper.matrix[3][2]
  _struct_ncs_oper.matrix[3][3]
  _struct_ncs_oper.vector[1]
  _struct_ncs_oper.vector[2]
  _struct_ncs_oper.vector[3]
```

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.

Biological macromolecular complexes may be built from domains related by symmetry transformations other than those arising from the crystal lattice symmetry. These domains are not necessarily discrete molecular entities: they may be composed of one or more segments of a single polypeptide or nucleic acid chain, of segments from more than one chain, or of small-molecule components of the structure. The categories above allow the distinct domains that participate in ensembles of structural elements related by noncrystallographic symmetry to be listed and described in detail. The relationships between categories used to describe noncrystallographic symmetry are shown in Fig. 3.6.7.12.

In the mmCIF model of noncrystallographic symmetry, the highest level of organization is the ensemble, which corresponds to the complete symmetry-related aggregate (*e.g.* tetramer, icosahedron). An identifier is given to the ensemble using the data item **`_struct_ncs_ens.id`**.

The symmetry-related elements within the ensemble are referred to as domains. The elements of structure that are to be considered part of the domain are specified using the data items in the STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories. By using the STRUCT_NCS_DOM_LIM data items appropriately, domains can be defined to include ranges of polypeptide chain or nucleic acid strand, bound ligands or cofactors, or even bound solvent molecules. Note that the category keys for STRUCT_NCS_DOM_LIM include the domain ID and the range specifiers. Thus a single domain may be composed of any number of ranges of elements.

Finally, the ensemble is generated from the domains using the rotation matrix and translation vector specified by data items in the STRUCT_NCS_OPER category, which are referenced by the data items in the STRUCT_NCS_ENS_GEN category. There are data items appropriate for two common methods of describing noncrystallographic symmetry:

(1) In the first method, the coordinate list includes all copies of domains related by noncrystallographic symmetry and the aim is to describe the relationships between domains in the ensemble; in this case the data items in STRUCT_NCS_ENS_GEN specify a pair of domains and reference the appropriate operator in STRUCT_NCS_OPER. This method is indicated by giving the data item **`_struct_ncs_oper.code`** the value `given`.

(2) In the second method, the coordinate list contains only one copy of the domain and the aim is to generate the entire ensemble; in this case the data items in STRUCT_NCS_ENS_GEN

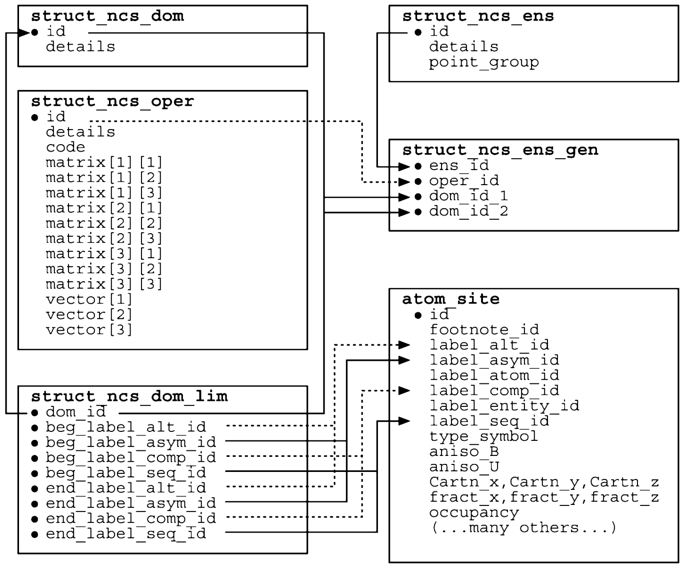

Fig. 3.6.7.12. The family of categories used to describe noncrystallographic symmetry. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

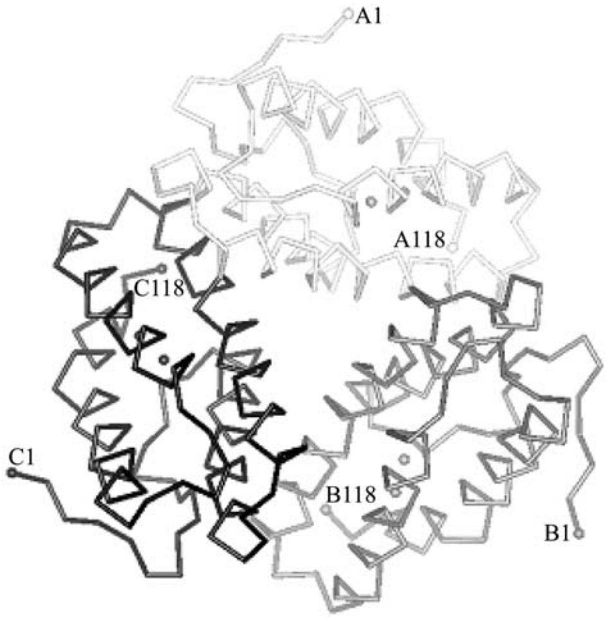

Fig. 3.6.7.13. Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) to be described with data items in the STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM and STRUCT_NCS_DOM_LIM categories.

Example 3.6.7.12. *Noncrystallographic symmetry in the structure of trimeric haemerythrin (PDB 1HR3) described with data items in the* STRUCT_NCS_ENS, STRUCT_NCS_ENS_GEN, STRUCT_NCS_DOM *and* STRUCT_NCS_DOM_LIM *categories. For brevity, the data items in the* STRUCT_NCS_OPER *category are not shown.*

```
_struct_ncs_ens.id                    trimer
_struct_ncs_ens.point_group           3

loop_
_struct_ncs_ens_gen.ens_id
_struct_ncs_ens_gen.dom_id_1
_struct_ncs_ens_gen.dom_id_2
_struct_ncs_ens_gen.oper_id
 trimer  chain_A  chain_B  1
 trimer  chain_A  chain_C  2

loop_
_struct_ncs_dom.id
 chain_A    chain_B    chain_C

loop_
_struct_ncs_dom_lim.dom_id
_struct_ncs_dom_lim.beg_label_asym_id
_struct_ncs_dom_lim.beg_label_comp_id
_struct_ncs_dom_lim.beg_label_seq_id
_struct_ncs_dom_lim.beg_label_alt_id
_struct_ncs_dom_lim.end_label_asym_id
_struct_ncs_dom_lim.end_label_comp_id
_struct_ncs_dom_lim.end_label_seq_id
_struct_ncs_dom_lim.end_label_alt_id
 chain_A  A  ala  1  .  A  ala  118  .
 chain_B  B  ala  1  .  B  ala  118  .
 chain_C  C  ala  1  .  C  ala  118  .
```

specify a pair of domains and reference the appropriate operator in STRUCT_NCS_OPER, but now the data item `_struct_ncs_oper.code` is given the value `generate`.

Noncrystallographic symmetry in a trimeric molecule is shown in Fig. 3.6.7.13 and described in Example 3.6.7.12.

3.6.7.5.6. *External databases*

The data items in these categories are as follows:

(*a*) STRUCT_REF

```
• _struct_ref.id
  _struct_ref.biol_id
        → _struct_biol.id
  _struct_ref.db_code
  _struct_ref.db_name
  _struct_ref.details
  _struct_ref.entity_id
        → _entity.id
  _struct_ref.seq_align
  _struct_ref.seq_dif
```

(*b*) STRUCT_REF_SEQ

```
• _struct_ref_seq.align_id
  _struct_ref_seq.db_align_beg
  _struct_ref_seq.db_align_end
  _struct_ref_seq.details
  _struct_ref_seq.ref_id
        → _struct_ref.id
  _struct_ref_seq.seq_align_beg
        → _entity_poly_seq.num
  _struct_ref_seq.seq_align_end
        → _entity_poly_seq.num
```

(*c*) STRUCT_REF_SEQ_DIF

```
• _struct_ref_seq_dif.align_id
        → _struct_ref_seq.align_id
• _struct_ref_seq_dif.seq_num
        → _entity_poly_seq.num
  _struct_ref_seq_dif.db_mon_id
        → _chem_comp.id
  _struct_ref_seq_dif.details
  _struct_ref_seq_dif.mon_id
        → _chem_comp.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 in the STRUCT_REF category allow the author of an mmCIF to provide references to information in external databases that is relevant to the entities or biological units described in the mmCIF. For example, the database entry for a protein or nucleic acid sequence could be referenced and any differences between the sequence of the macromolecule whose structure is reported in the mmCIF and the sequence of the related entry in the external database can be recorded. Alternatively, references to external database entries can be used to record the relationship of the structure reported in the mmCIF to structures already reported in the literature, for example by referring to previously determined structures of the same or a similar protein, or to a small-molecule structure determination of a bound inhibitor or cofactor. STRUCT_REF data items are not intended to be used to reference a database entry for the structure in the mmCIF itself (this would be the role of data items in the DATABASE_2 category), but it would not be formally incorrect to do so.

When the data items in these categories are used to provide references to external database entries describing the sequence of a polymer, data items from all three categories could be used. The value of the data item `_struct_ref.seq_align` is used to indicate whether the correspondence between the sequence of the entity or biological unit in the mmCIF and the sequence in the related external database entry is `complete` or `partial`. If the value is `partial`, the region (or regions) of the alignment may be identified using data items in the STRUCT_REF_SEQ category. Comments on the alignment may be given in `_struct_ref_seq.details` (Example 3.6.7.13).

The value of the data item `_struct_ref.seq_dif` is used to indicate whether the two sequences contain point differences. If the value is `yes`, the differences may be identified and annotated using data items in the STRUCT_REF_SEQ_DIF category. Comments on specific point differences may be recorded in `_struct_ref_seq_dif.details`.

Example 3.6.7.13. *The relationship of the sequence of the protein PDB 5HVP to a sequence in an external database described with data items in the* STRUCT_REF *and* STRUCT_REF_SEQ *categories.*

```
loop_
  _struct_ref.id
  _struct_ref.biol_id
  _struct_ref.entity_id
  _struct_ref.db_name
  _struct_ref.db_code
  _struct_ref.seq_align
  _struct_ref.seq_dif
 seq_pdb     1  .  PDB      5HVP      .         .
 seq_genbank .  1  GenBank  AAG30358  complete  yes

loop_
  _struct_ref_seq.align_id
  _struct_ref_seq.ref_id
  _struct_ref_seq.seq_align_beg
  _struct_ref_seq.seq_align_end
  _struct_ref_seq.db_align_beg
  _struct_ref_seq.db_align_end
  _struct_ref_seq.details
    align_seq_pdb_genbank seq_genbank 1 99 24 122
; The genbank reference is to the sequence of
  residues 1-376 of the viral pol 1 polypeptide;
  the protease is proteolytically released from
  this precursor during viral maturation.
;
```

References do not have to be to entries in databases of sequences: any external database can be referenced. For other kinds of databases, only the data items in the STRUCT_REF category would usually be used. The element of the structure that is referenced could be either an entity or a biological unit, that is, either a building block of the structure or a structurally meaningful assembly of those building blocks. Since the identification of the part of the structure being linked to an entry in an external database can be made using either `_struct_ref.biol_id` or `_struct_ref.entity_id`, and since any part of the structure could be linked to any number of entries in external databases, the data item `_struct_ref.id` was introduced as the category key.

3.6.7.5.7. *β-sheets*

Data items in these categories are as follows:

(*a*) STRUCT_SHEET

```
• _struct_sheet.id
  _struct_sheet.details
  _struct_sheet.number_strands
  _struct_sheet.type
```

(*b*) STRUCT_SHEET_TOPOLOGY

```
• _struct_sheet_topology.range_id_1
        → _struct_sheet_range.id
• _struct_sheet_topology.range_id_2
        → _struct_sheet_range.id
• _struct_sheet_topology.sheet_id
        → _struct_sheet.id
  _struct_sheet_topology.offset
  _struct_sheet_topology.sense
```

(*c*) STRUCT_SHEET_RANGE

```
• _struct_sheet_range.id
• _struct_sheet_range.sheet_id
        → _struct_sheet.id
  _struct_sheet_range.beg_label_asym_id
        → _struct_asym.id
  _struct_sheet_range.beg_label_comp_id
        → _chem_comp.id
  _struct_sheet_range.beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_range.end_label_asym_id
        → _struct_asym.id
  _struct_sheet_range.end_label_comp_id
        → _chem_comp.id
  _struct_sheet_range.end_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_range.beg_auth_asym_id
        → _atom_site.auth_atom_id
  _struct_sheet_range.beg_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_sheet_range.beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_range.end_auth_asym_id
        → _atom_site.auth_atom_id
  _struct_sheet_range.end_auth_comp_id
        → _atom_site.auth_comp_id
  _struct_sheet_range.end_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_range.symmetry
```

(*d*) STRUCT_SHEET_ORDER

```
• _struct_sheet_order.range_id_1
        → _struct_sheet_range.id
• _struct_sheet_order.range_id_2
        → _struct_sheet_range.id
• _struct_sheet_order.sheet_id
        → _struct_sheet.id
  _struct_sheet_order.offset
  _struct_sheet_order.sense
```

(*e*) STRUCT_SHEET_HBOND

```
• _struct_sheet_hbond.range_id_1
        → _struct_sheet_range.id
• _struct_sheet_hbond.range_id_2
        → _struct_sheet_range.id
• _struct_sheet_hbond.sheet_id
        → _struct_sheet.id
  _struct_sheet_hbond.range_1_beg_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_1_beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_1_end_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_1_end_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_2_beg_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_2_beg_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_2_end_label_atom_id
        → _atom_site.label_atom_id
  _struct_sheet_hbond.range_2_end_label_seq_id
        → _atom_site.label_seq_id
  _struct_sheet_hbond.range_1_beg_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_1_beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_hbond.range_1_end_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_1_end_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_hbond.range_2_beg_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_2_beg_auth_seq_id
        → _atom_site.auth_seq_id
  _struct_sheet_hbond.range_2_end_auth_atom_id
        → _atom_site.auth_atom_id
  _struct_sheet_hbond.range_2_end_auth_seq_id
        → _atom_site.auth_seq_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.

Different methods of describing β-sheets are in widespread use. The mmCIF dictionary provides data items for two methods and it is anticipated that future versions of the dictionary could cover others. The model used in the STRUCT_SHEET_TOPOLOGY category is the simpler of the two. It is a convenient shorthand for describing the topology, but it does not provide details about strand registration and it is not suitable for describing sheets that contain strands from more than one polypeptide. A more general model is provided by the linked data items in the STRUCT_SHEET_RANGE, STRUCT_SHEET_ORDER and STRUCT_SHEET_HBOND categories. For both methods of representing β-sheets, data items in the parent category STRUCT_SHEET can be used to provide an identifier for each sheet, a free-text description of its type, the number of participating strands and a free-text description of any peculiar aspects of the sheet. The relationships between categories used to describe β-sheets are shown in Fig. 3.6.7.14.

In the description of β-sheet topology based on the STRUCT_SHEET_TOPOLOGY category, the strand that occurs first in the polypeptide chain is numbered 1. Subsequent strands are described by their position in the sheet relative to the previous strand (+1, −3 *etc.*) and by their orientation relative to the previous strand (parallel or antiparallel).

While writing this chapter, a few errors in the mmCIF dictionary were discovered. The use of `_struct_sheet_topology.range_id_1` and `*_2` as pointers to the residues participating in β-sheets is one; the correct data items should be `_struct_sheet_topology.comp_id_1` and `*_2`, and these data items should be pointers to `_atom_site.label_comp_id`. This error will be corrected in future versions of the dictionary. As the data model encoded in the current version of the dictionary is incorrect, no example of its use is given.

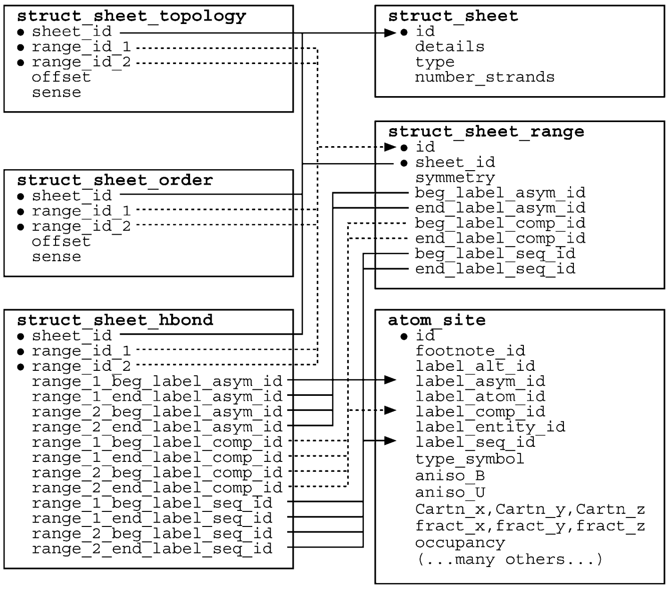

Fig. 3.6.7.14. The family of categories used to describe β-sheets. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

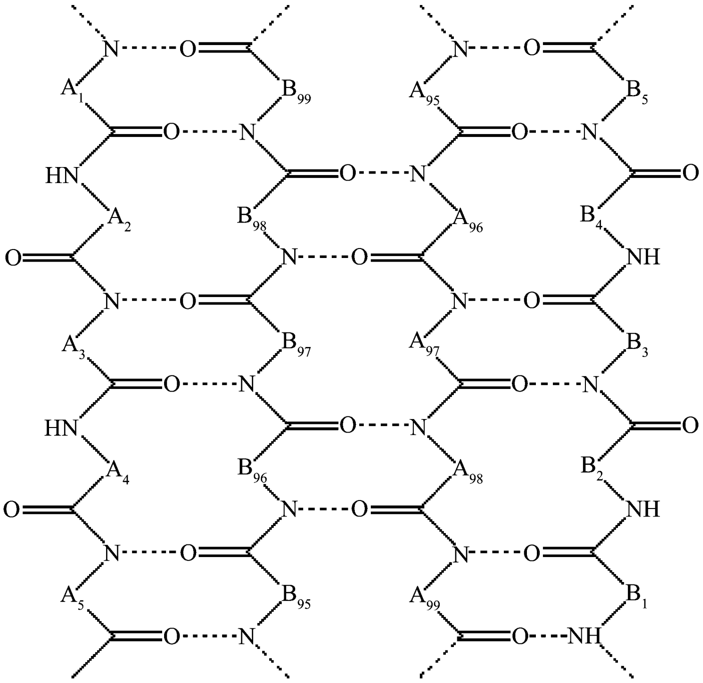

Fig. 3.6.7.15. A hypothetical β-sheet to be described with data items in the STRUCT_SHEET, STRUCT_SHEET_ORDER, STRUCT_SHEET_RANGE and STRUCT_SHEET_HBOND categories. Note that the strands come from two different polypeptides, labelled A and B.

In the more detailed and more general method for describing β-sheets, data items in the STRUCT_SHEET_RANGE category specify the range of residues that form strands in the sheet, data items in the STRUCT_SHEET_ORDER category specify the relative pairwise orientation of strands and data items in the STRUCT_SHEET_HBOND category provide details of specific hydrogen-bonding interactions between strands (see Fig. 3.6.7.15 and Example 3.6.7.14). Note that the specifiers for the strand ranges include the amino acid (**`*_comp_id`** and **`*_seq_id`**), the chain (**`*_asym_id`**) and a symmetry code (**`_struct_sheet_range.symmetry`**). Thus sheets that are composed of strands from more than one polypeptide chain or from polypeptides in more than one asymmetric unit can be described.

Example 3.6.7.14. *A hypothetical β-sheet described with data items in the STRUCT_SHEET, STRUCT_SHEET_ORDER, STRUCT_SHEET_RANGE and STRUCT_SHEET_HBOND categories.*

```
loop_
  _struct_sheet.id
  _struct_sheet.number_strands
    S1   4

loop_
  _struct_sheet_order.sheet_id
  _struct_sheet_order.range_id_1
  _struct_sheet_order.range_id_2
  _struct_sheet_order.sense
    S1 1 2 anti-parallel
    S1 2 3 anti-parallel
    S1 3 4 anti-parallel
    S2 1 2 anti-parallel

loop_
  _struct_sheet_range.sheet_id
  _struct_sheet_range.id
  _struct_sheet_range.beg_label_comp_id
  _struct_sheet_range.beg_label_asym_id
  _struct_sheet_range.beg_label_seq_id
  _struct_sheet_range.end_label_comp_id
  _struct_sheet_range.end_label_asym_id
  _struct_sheet_range.end_label_seq_id
    S1 1 PRO A 1  LEU A 5
    S1 2 CYS B 95 PHE B 99
    S1 3 CYS A 95 PHE A 99
    S1 4 PRO B 1  LEU B 5

loop_
  _struct_sheet_hbond.sheet_id
  _struct_sheet_hbond.range_id_1
  _struct_sheet_hbond.range_id_2
  _struct_sheet_hbond.range_1_beg_label_atom_id
  _struct_sheet_hbond.range_1_beg_label_seq_id
  _struct_sheet_hbond.range_2_beg_label_atom_id
  _struct_sheet_hbond.range_2_beg_label_seq_id
    S1 1 2 A 3  O 97
    S1 2 3 B 98 O 96
    S1 3 4 A 97 O 3
```

It is conventional to assign the number 1 to an outermost strand. The choice of which outermost strand to number as 1 is arbitrary, but would usually be the strand encountered first in the amino-acid sequence. The remaining strands are then numbered sequentially across the sheet.

In some simple cases, the complete hydrogen bonding of the sheet could be inferred from the strand-range pairings and the relationship between the strands (parallel or antiparallel). However, in most cases it is necessary to specify at least one hydrogen bond between adjacent strands in order to establish the registration. The data items in the STRUCT_SHEET_HBOND category can be used to do this. Hydrogen bonds also need to be specified precisely when a sheet contains a nonstandard feature such as a β-bulge. This is a case where it is sufficient to specify a single hydrogen-bonding interaction to establish the registration; here only the **`*_beg_*`** or **`*_end_*`** data items need to be used to reference the atom-label components. However, it is preferable, wherever possible, to specify the initial and final atoms of the two ranges participating in the hydrogen bonding.

3.6.7.5.8. *Molecular sites*

The data items in these categories are as follows:

(*a*) STRUCT_SITE
● **`_struct_site.id`**
  `_struct_site.details`

(*b*) STRUCT_SITE_KEYWORDS

```
• _struct_site_keywords.site_id
        → _struct_site.id
• _struct_site_keywords.text
```

(*c*) STRUCT_SITE_GEN

```
• _struct_site_gen.id
• _struct_site_gen.site_id
        → _struct_site.id
  _struct_site_gen.details
  _struct_site_gen.label_alt_id
        → _atom_sites_alt.id
  _struct_site_gen.label_asym_id
        → _atom_site.label_asym_id
  _struct_site_gen.label_atom_id
        → _chem_comp_atom.atom_id
  _struct_site_gen.label_comp_id
        → _atom_site.label_atom_id
  _struct_site_gen.label_seq_id
        → _atom_site.label_seq_id
  _struct_site_gen.auth_asym_id
        → _atom_site.auth_asym_id
  _struct_site_gen.auth_atom_id
        → _atom_site.auth_atom_id
  _struct_site_gen.auth_comp_id
        → _atom_site.auth_comp_id
  _struct_site_gen.auth_seq_id
        → _atom_site.auth_seq_id
  _struct_site_gen.symmetry
```

(*d*) STRUCT_SITE_VIEW

```
• _struct_site_view.id
  _struct_site_view.details
  _struct_site_view.rot_matrix[1][1]
  _struct_site_view.rot_matrix[1][2]
  _struct_site_view.rot_matrix[1][3]
  _struct_site_view.rot_matrix[2][1]
  _struct_site_view.rot_matrix[2][2]
  _struct_site_view.rot_matrix[2][3]
  _struct_site_view.rot_matrix[3][1]
  _struct_site_view.rot_matrix[3][2]
  _struct_site_view.rot_matrix[3][3]
  _struct_site_view.site_id
        → _struct_site.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.

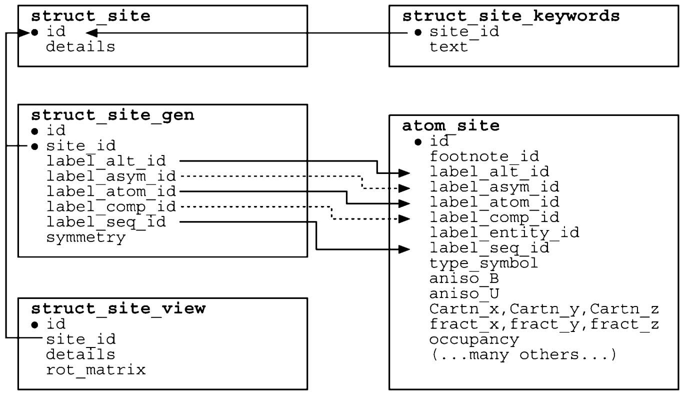

Fig. 3.6.7.16. The family of categories used to describe molecular sites. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (•). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

Example 3.6.7.15. *A DNA binding site with an intercalated drug (NDB DDF040) described with data items in the STRUCT_SITE, STRUCT_SITE_KEYWORDS, STRUCT_SITE_GEN and STRUCT_SITE_VIEW categories.*

```
loop_
_struct_site.id
_struct_site.details
   B1  'Binding at TG/AC Step 1'

loop_
_struct_site_keywords.site_id
_struct_site_keywords.text
   B1  'Intercalation complex'

loop_
_struct_site_gen.id
_struct_site_gen.site_id
_struct_site_gen.label_asym_id
_struct_site_gen.label_comp_id
_struct_site_gen.label_seq_id
_struct_site_gen.symmetry
   1  B1  A  T    1  1_555
   2  B1  A  G    2  1_555
   3  B1  A  C    5  8_555
   4  B1  A  A    6  8_555
   5  B1  D  DM2  .  8_555

loop_
_struct_site_view.id
_struct_site_view.site_id
_struct_site_view.details
_struct_site_view.rot_matrix[1][1]
_struct_site_view.rot_matrix[1][2]
# - - - abbreviated - - -
_struct_site_view.rot_matrix[3][3]
   View1  B1
  'View along the base-pair plane'
   0.133  0.922  . . . . . . -0.172
```

Substrate-binding sites, active sites, metal coordination sites and any other sites of interest may be described using data items in a collection of categories descending from STRUCT_SITE. These categories are intended to enable the author to generate views of molecular sites that could be used as figures in a report describing the structure or to enable a database to store standard views of common molecular sites (*e.g.* ATP-binding sites or the coordination of a calcium atom). The relationships between categories used to describe structural sites are shown in Fig. 3.6.7.16.

An identifier for each site that an author wishes to describe is given using **_struct_site.id** and the site can be described using **_struct_site.details**.

Keywords can be given for each site using data items in the STRUCT_SITE_KEYWORD category. Because keywords can be given at many levels of the mmCIF description of a structure, it may be worth duplicating the most significant higher-level keywords at this level to ensure that the site is detected in all search strategies.

The structural elements that generate each molecular site can be specified using data items in the STRUCT_SITE_GEN category. 'Structural elements' in this sense may be at any level of detail in the structure: single atoms, complete amino acids or nucleotides, or elements of secondary, tertiary or quaternary structure. Therefore the labels for each element may include, as required, the relevant ***_alt**, ***_asym**, ***_atom**, ***_comp** or ***_seq** parts of atom or residue identifiers. If the author has used an alternative labelling scheme, this can also be used. Noteworthy features of a structural element that forms part of the site can be described using the data item **_struct_site_gen.details**. Any crystallographic symmetry operations that are needed to form the site can be given using **_struct_site_gen.symmetry**.

Data items in the STRUCT_SITE_VIEW category allow the author to specify an orientation of the molecular site that gives a useful view of the components. The comments given in **_struct_site_view.details** could be used as a figure caption if the view is intended for use as a figure in a report.

Example 3.6.7.15 illustrates the use of these categories for describing a DNA binding site.

3.6.7.6. Crystal symmetry

The categories describing symmetry are as follows:

SYMMETRY group
SYMMETRY
SYMMETRY_EQUIV
SPACE_GROUP
SPACE_GROUP_SYMOP

Data items in the SYMMETRY category are used to give details about the crystallographic symmetry. The equivalent positions for the space group are listed using data items in the SYMMETRY_EQUIV category. These categories are used in the same way in the core CIF and mmCIF dictionaries, and Section 3.2.4.4 can be consulted for details.

The current version of the mmCIF dictionary includes the SPACE_GROUP categories that were derived from the symmetry CIF dictionary (Chapter 3.8) and included in version 2.3 of the core CIF dictionary. At the time of writing, macromolecular applications have not yet begun to make use of these new categories.

Data items in these categories are as follows:

(*a*) SYMMETRY
• *_symmetry.entry_id*
→ _entry.id
_symmetry.cell_setting
_symmetry.Int_Tables_number
_symmetry.space_group_name_Hall
_symmetry.space_group_name_H-M

(*b*) SYMMETRY_EQUIV
• *_symmetry_equiv.id* (∼ *_symmetry_equiv_pos_site_id*)
_symmetry_equiv.pos_as_xyz

(*c*) SPACE_GROUP
• *_space_group.id*
_space_group.crystal_system
_space_group.IT_number
_space_group.name_H-M_alt
_space_group.name_Hall

(*d*) SPACE_GROUP_SYMOP
• *_space_group_symop.id*
_space_group_symop.operation_xyz
_space_group_symop.sg_id

The bullet (•) indicates a category 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 (_) except where indicated by the ∼ symbol.*

The data item `_symmetry.entry_id` has been added to the SYMMETRY category to provide the formal category key required by the DDL2 data model.

3.6.7.7. Bond-valence information

The categories describing bond valences are as follows:

VALENCE group
VALENCE_PARAM
VALENCE_REF

These categories were introduced into version 2.2 of the core CIF dictionary to provide the information about bond valences required in inorganic crystallography. They appear in the mmCIF dictionary only for full compatibility with the core dictionary.

Data items in these categories are as follows:

(*a*) VALENCE_PARAM
• *_valence_param.atom_1*
• *_valence_param.atom_1_valence*
• *_valence_param.atom_2*
• *_valence_param.atom_2_valence*
_valence_param.B
_valence_param.details
_valence_param.id
_valence_param.ref_id
→ _valence_ref.id
_valence_param.Ro

(*b*) VALENCE_REF
• *_valence_ref.id*
_valence_ref.reference

The bullet (•) indicates a category 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 (_).*

Information about the use of these data items in the core CIF dictionary is given in Section 3.2.4.5.

3.6.8. Publication

The results of the determination of the crystal structure of a biological macromolecule might be published in an academic journal and/or deposited in a structural database. The data items in the core CIF dictionary cover most of the requirements for constructing an article for publication from an mmCIF and the many well defined data fields in mmCIF allow an extensively annotated record of the structure to be deposited in a database. However, the formalism of two of the core CIF categories for publication did not fit the relational database model of mmCIF, so new categories were required. The core CIF category COMPUTING, which is used to list the programs used to determine the structure, is replaced by the mmCIF category SOFTWARE, and the core CIF category DATABASE, which is used to identify the records associated with the structure in various databases, is replaced by the mmCIF category DATABASE_2.

The category groups discussed here are: the CITATION group, which is used to give citations to the literature (Section 3.6.8.1); the COMPUTING group, which is used to cite software (Section 3.6.8.2); the DATABASE group for citing related database entries (Section 3.6.8.3), which includes a group of categories used to ensure compatibility with specific database records in the Protein Data Bank (Section 3.6.8.3.2); journal administration categories that might be used by a publisher (Section 3.6.8.4.1); and the PUBL family of categories used to store the text of an article for publication (Section 3.6.8.4.2).

3.6.8.1. Literature citations

The categories describing literature citations are as follows:

CITATION group
CITATION
CITATION_AUTHOR
CITATION_EDITOR

Data items in these categories are as follows:

(*a*) CITATION
• *_citation.id*
_citation.abstract
_citation.abstract_id_CAS
_citation.book_id_ISBN
_citation.book_publisher
_citation.book_publisher_city
_citation.book_title
_citation.coordinate_linkage
_citation.country
_citation.database_id_CSD
_citation.database_id_Medline
_citation.journal_abbrev
_citation.journal_full
_citation.journal_id_ASTM
_citation.journal_id_CSD
_citation.journal_id_ISSN
_citation.journal_issue
_citation.journal_volume
_citation.language
_citation.page_first