3. CIF DATA DEFINITION AND CLASSIFICATION


```
struct_mon_prot
● label_alt_id
● label_asym_id
● label_comp_id
● label_seq_id
  details
  chi1-chi5
  omega,phi,psi
  mean_B_(all,main,side)
  RSCC/RSR_(all,main,side)
  (...many others...)

struct_mon_prot_cis
● label_alt_id
● label_asym_id
● label_comp_id
● label_seq_id

entry
● id

struct_mon_details
● entry_id
  prot_cis
  RSCC
  RSR

struct_mon_nucl
● label_alt_id
● label_asym_id
● label_comp_id
● label_seq_id
  details
  alpha,beta,chi1
  chi2,delta,zeta
  epsilon,gamma
  nu0,nu1,nu3,nu4
  mean_B_(all,base)
  mean_B_(sugar,phos)
  (...many others...)

atom_site
● id
  footnote_id
  label_alt_id
  label_asym_id
  label_atom_id
  label_comp_id
  label_entity_id
  label_seq_id
  type_symbol
  aniso_B
  aniso_U
  Cartn_x,Cartn_y,Cartn_z
  fract_x,fract_y,fract_z
  occupancy
  (...many others...)
```
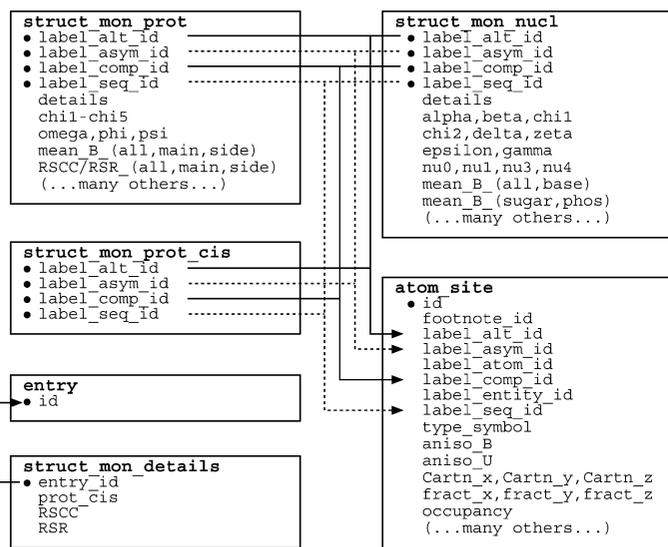
Fig. 3.6.7.11. The family of categories used to describe the structural features of monomers. Boxes surround categories of related data items. Data items that serve as category keys are preceded by a bullet (●). Lines show relationships between linked data items in different categories with arrows pointing at the parent data items.

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
```

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.

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

references