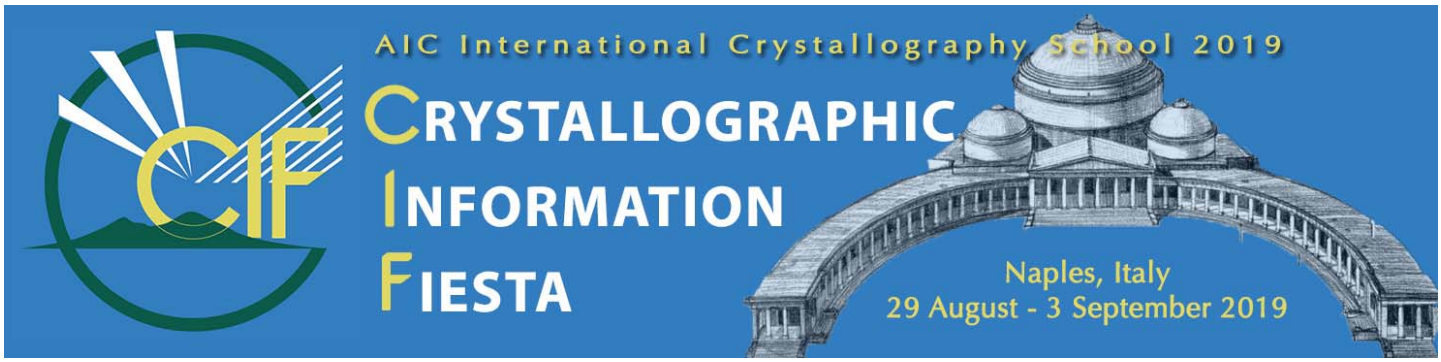


The core CIF dictionary and the mmCIF dictionary

Brian McMahon

International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK

bm@iucr.org

An overview is given of the core and mmCIF dictionaries and how they are arranged in categories related to different aspects of the crystallographic experiment. The core categories will be described in detail. mmCIF categories describing protein structure will be introduced in a general way for a plenary audience (more specific macromolecular structure considerations will be addressed in the presentation by John Berrisford).

Core CIF dictionary

- First published as integral part of the CIF standard (1991)
- Includes the data items needed to describe any crystal structure
- Includes discrete experimental data from single-crystal diffraction (structure factors)
- Includes essential experimental metadata
- Permits some chemical characterisation of the crystal
- Allows documentation in the form of a journal article
- Defines a general data model, extensible through other dictionaries

The first two slides give a brief overview of the material that will be covered in this lecture. The first part, on the core CIF dictionary, will give a historical background to the development of a standard describing essential data items in a machine-readable format. We will aim to explain something of the design requirements and choices (pointing out that the approach was designed by people who had also been involved in writing crystallographic structure solution packages, and so had an excellent working knowledge of the input/output requirements and capabilities of such packages). The core dictionary aims to cover a large part of the information needed to describe any reasonably well-ordered crystal structure, and is particularly well tuned to structure determination of small-unit-cell structures by single-crystal diffraction (especially of X-rays). Students will be exposed to some of the detail of assigned categories and data item definitions, but the main objective is to instill an understanding of why an orderly classification of data items is beneficial.

Macromolecular CIF (mmCIF) dictionary

- Commissioned soon after publication of core dictionary (1991)
- Long development period (published 1997)
- Attempt to capture all relevant information from an MX experiment
- Describe secondary structure of proteins and nucleic acids
- Accommodate all the data and metadata in a PDB entry
- Required more complex data model (DDL2)
- Became basis for the PDB database schema, subsequently expanded

This slide is intended to parallel the first, showing how the same principles were applied in the field of protein (and nucleic acid) crystallography. As this is a plenary lecture, it will not go too deeply into the complexities of protein structure determination; nevertheless, it is felt to be useful to expose non-macromolecular crystallographers to some sense of what mmCIF tries to do that is different from the core and other small-unit-cell extension dictionaries. A particular complication in the early adoption of mmCIF was that it sought to improve the description of structures beyond what was then available in the PDB format. As such, it was poorly tuned to existing software, which hindered its early adoption. Because of a relatively wide spread of computation approaches and a more diverse user community, it needed to be developed in a more flexible and rapidly growing way than the core dictionary.

Core CIF dictionary – history

- Many different software authors in the 1960s/70s/80s
- Each program had its own format
- Each diffractometer produced data files with their own formats
- Need for programs to pass data from one to the next ('pipeline')
- Attempts to settle on a common standard
- Standard Crystallographic File Structure (SCFS)
  - Brown, I. D. (1988). *Standard Crystallographic File Structure-87. Acta Cryst.* A**44**, 232.
- Crystallographic Information File (CIF)
  - Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The crystallographic information file (CIF): a new standard archive file for crystallography*. *Acta Cryst*. (1991). A**47**, 655-685

In this slide we do not give a detailed history of the evolution of the CIF standard (which may be covered in other course lectures), but we want to emphasise a few things. There were already many ways to describe the specific pieces of information ("data items") that each program required. Inevitability there was some similarity (there is really only one way to calculate and represent a unit-cell volume) but also much divergence (different programs took different approaches as to how to describe the site occupancy of an atom located on a crystallographic special position). There was also a growing need to handle diverse types of data as program systems were developed to address all the challenges along the structure solution and refinement workflow. CIF stood on the shoulders of its predecessors – of particular importance were the classification and categorisation of the data items that were needed by the SCFS project, backed up by the requirements for publication of *Acta Crystallographica* and the requirements for database deposition of the CCDC; and the data storage approach within the *Xtal* suite of programs developed and managed by Syd Hall and others.

Organisation of the core dictionary

Topic	*Category group*	*Subject covered*
(*a*) Experimental measurements	CELL	Unit cell
	DIFFRN	Diffraction experiment
	EXPTL	Experimental conditions
(*b*) Analysis	REFINE	Refinement procedures
	REFLN	Reflection measurements
(*c*) Atomicity, chemistry and structure	ATOM	Atom sites
	CHEMICAL	Chemical properties and nomenclature
	GEOM	Geometry of atom sites
	SYMMETRY	Symmetry information
	VALENCE	Bond-valence information
(*d*) Publication	CITATION	Bibliographic references
	COMPUTING	Computational details of the experiment
	DATABASE	Database information
	JOURNAL	Journal housekeeping
	PUBL	Contents of a published article
(*e*) File metadata	AUDIT	Dictionary maintenance and identification

We discuss the organisation of the core dictionary thematically, using the approach to classification described in *International Tables* G (Chapter 3.2). In practice the dictionary is organised strictly alphabetically, but this approach may help to appreciate its overall design. “Category groups” are not formally defined in DDL1, the original formalism used for the core dictionary, but they are formal structures in the mmCIF category and can usefully be adopted informally to better understand the core dictionary structure.

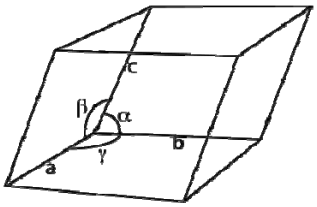
Experimental measurements – the CELL group

CELL category

```
_cell_angle_alpha
_cell_angle_beta
_cell_angle_gamma
_cell_formula_units_Z
_cell_length_a
_cell_length_b
_cell_length_c
_cell_measurement_pressure
_cell_measurement_radiation
_cell_measurement_reflns_used
_cell_measurement_temperature
_cell_measurement_theta_max
_cell_measurement_theta_min
_cell_measurement_wavelength
_cell_reciprocal_angle_alpha
_cell_reciprocal_angle_beta
_cell_reciprocal_angle_gamma
_cell_reciprocal_length_a
_cell_reciprocal_length_b
_cell_reciprocal_length_c
_cell_special_details
_cell_volume
```

CELL_MEASUREMENT_REFLN category

```
_cell_measurement_refln_index_h
_cell_measurement_refln_index_k
_cell_measurement_refln_index_l
```



EXAMPLE

```
_cell_length_a      20.572(3)
_cell_length_b      3.9052(5)
_cell_length_c      14.811(3)
_cell_angle_alpha      90
_cell_angle_beta     110.95(2)
_cell_angle_gamma      90
_cell_volume     1111.2(4)
_cell_formula_units_Z     8
_cell_measurement_reflns_used      2315
_cell_measurement_theta_min     2.9395
_cell_measurement_theta_max     30.5514
_cell_measurement_temperature      298(2)
```

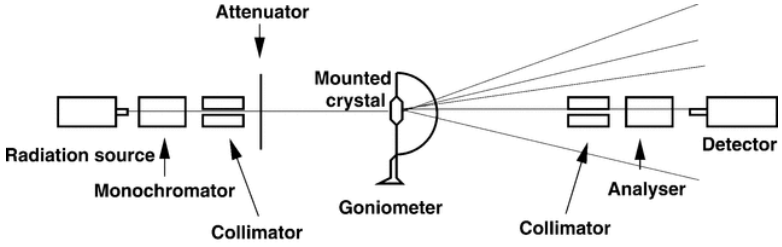
Portalone, G. (2019). 6-Methyluracil: a redetermination of polymorph (II). *IUCrData*, **4**, x190861.

There are two categories in this group, describing the unit cell and its measurement. CELL_MEASUREMENT_REFLN is a category on its own, because it loops the reflections used in the determination of the unit cell on a diffractometer. In practice these do not seem to be reported much. The CELL category combines both aspects of the experimental conditions and the refined cell parameters, which is arguably a property of the derived structure model. Perhaps the thinking is that determining the cell parameters establishes input data for the structure determination on a par with other experimental conditions. In any case, the values required by structure determination software are well characterised by their individual definitions in the dictionary.

Experimental measurements – DIFFRN group

General

DIFFRN



Before the crystal

DIFFRN_ATTENUATOR
DIFFRN_RADIATION

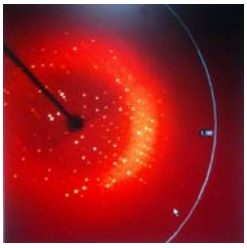
DIFFRN_RADIATION_WAVELENGTH
DIFFRN_SOURCE

At the crystal

DIFFRN_MEASUREMENT
DIFFRN_ORIENT_MATRIX
DIFFRN_ORIENT_REFLN

After the crystal

DIFFRN_DETECTOR



Intensity measurements

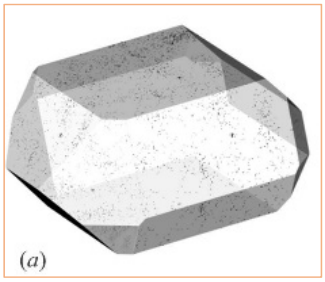
DIFFRN_REFLN
DIFFRN_REFLNS
DIFFRN_REFLNS_CLASS
DIFFRN_SCALE_GROUP
DIFFRN_STANDARD_REFLN
DIFFRN_STANDARDS

The DIFFRN family of categories relate to the diffraction experiment, and in principle cover any instrument, technique or methodology. They are grouped roughly according to the scheme in the slide. The DIFFRN category describes the ambient conditions, crystal treatment and any noteworthy aspects of diffraction point symmetry, systematic absences, inferred space group etc. The categories “before the crystal” characterise the radiation probe (not always X-rays) and its source. The categories “at the crystal” describe the goniometer or other mounting device and the relationship between the crystal-centric reference frame and coordinates given in the frame of reference of the instrumentation. “After the crystal” comes some information about the radiation detector, while the intensity measurements themselves are listed using the DIFFRN_REFLN and related categories. These categories were developed to record peak intensities from point diffractometers; nowadays raw diffraction images can be described *in toto* within the imgCIF (“image CIF”) framework.

Experimental measurements – EXPTL group

EXPTL

```
_exptl_absorpt_coefficient_mu
_exptl_absorpt_correction_T_max
_exptl_absorpt_correction_T_min
_exptl_absorpt_correction_type
_exptl_absorpt_process_details
_exptl_crystals_number
_exptl_special_details
```

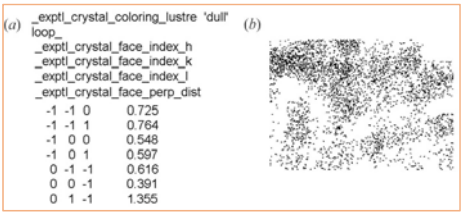


EXPTL_CRYSTAL

```
_exptl_crystal_id
_exptl_crystal_colour
_exptl_crystal_colour_lustre
_exptl_crystal_colour_modifier
_exptl_crystal_colour_primary
_exptl_crystal_density_diffrn
_exptl_crystal_density_meas
_exptl_crystal_density_meas_gt
_exptl_crystal_density_meas_lt
_exptl_crystal_density_meas_temp
_exptl_crystal_density_meas_temp_gt
_exptl_crystal_density_meas_temp_lt
_exptl_crystal_density_method
_exptl_crystal_description
_exptl_crystal_F_000
_exptl_crystal_preparation
_exptl_crystal_pressure_history
_exptl_crystal_recrystallization_method
_exptl_crystal_size_length
_exptl_crystal_size_max
_exptl_crystal_size_mid
_exptl_crystal_size_min
_exptl_crystal_size_rad
_exptl_crystal_thermal_history
```

EXPTL_CRYSTAL_FACE

```
_exptl_crystal_face_index_h
_exptl_crystal_face_index_k
_exptl_crystal_face_index_l
_exptl_crystal_face_diffr_chi
_exptl_crystal_face_diffr_kappa
_exptl_crystal_face_diffr_phi
_exptl_crystal_face_diffr_psi
_exptl_crystal_face_perp_dist
```



Figures from Kaminsky, W. (2007). From CIF to virtual morphology using the *WinXMorph* program. *J. Appl. Cryst.* **40**, 382-385

Although very many of the CIF data categories describe aspects of the experimental setup and conduct, the categories in the "EXPTL" group itself are largely concerned with the crystal itself – its habit, size, density, preparation and treatment. Clearly they are concerned with single crystals (powders are described in separate (PD_SPEC) categories in the pdCIF dictionary, for example).

Analysis – structure refinement

REFINE

_refine_diff_density_max
_refine_diff_density_min
_refine_diff_density_rms
_refine_ls_abs_structure_details
_refine_ls_abs_structure_Flack
_refine_ls_abs_structure_Rogers
_refine_ls_d_res_high
_refine_ls_d_res_low
_refine_ls_extinction_coef
_refine_ls_extinction_expression
_refine_ls_extinction_method
_refine_ls_goodness_of_fit_all
_refine_ls_goodness_of_fit_gt
_refine_ls_goodness_of_fit_obs
_refine_ls_goodness_of_fit_ref
_refine_ls_hydrogen_treatment
_refine_ls_matrix_type
_refine_ls_number_constraints
_refine_ls_number_parameters
_refine_ls_number_reflns
_refine_ls_number_restraints
_refine_ls_R_factor_all
_refine_ls_R_factor_gt
_refine_ls_R_factor_obs
_refine_ls_R_Fsqd_factor
_refine_ls_R_I_factor
_refine_ls_restrained_S_all
_refine_ls_restrained_S_gt
_refine_ls_restrained_S_obs
_refine_ls_shift/esd_max
_refine_ls_shift/esd_mean
_refine_ls_shift/su_max
_refine_ls_shift/su_max_lt
_refine_ls_shift/su_mean
_refine_ls_shift/su_mean_lt
_refine_ls_structure_factor_coef
_refine_ls_weighting_details
_refine_ls_weighting_scheme
_refine_ls_wR_factor_all
_refine_ls_wR_factor_gt
_refine_ls_wR_factor_obs
_refine_ls_wR_factor_ref
_refine_special_details

REFINE_LS_CLASS

_refine_ls_class_code
_reflns_class_code
_refine_ls_class_d_res_high
_refine_ls_class_d_res_low
_refine_ls_class_R_factor_all
_refine_ls_class_R_factor_gt
_refine_ls_class_R_Fsqd_factor
_refine_ls_class_R_I_factor
_refine_ls_class_wR_factor_all

There are a large number of data items relating to the structure refinement, as the metrics produced during least-squares refinement provide some indication of the likely quality of the resulting structure model. The “REFINE_LS_CLASS” items allow for handling particular groups of intensities (“reflection classes”) in different ways. This can be useful, for example, in identifying the reflections from different phases of a composite material. Note in this list the presence of italicised data names. Their use is deprecated (in most of these examples the name has been changed to reflect preferred terminology); but as CIF is intended as an archive mechanism, the deprecated items are retained in the dictionary and attributes are added to indicate their replacements.

Analysis – reflections used in refinement

Groups of reflections

REFLNS

```
_reflns_d_resolution_high
_reflns_d_resolution_low
_reflns_Friedel_coverage
_reflns_limit_h_max
_reflns_limit_h_min
_reflns_limit_k_max
_reflns_limit_k_min
_reflns_limit_l_max
_reflns_limit_l_min
_reflns_number_gt
_reflns_number_total
_reflns_special_details
_reflns_threshold_expression
```

REFLNS_CLASS

```
_reflns_class_code
_reflns_class_d_res_high
_reflns_class_d_res_low
_reflns_class_description
_reflns_class_number_gt
_reflns_class_number_total
_reflns_class_R_factor_all
_reflns_class_R_factor_gt
_reflns_class_R_Fsqd_factor
_reflns_class_R_I_factor
_reflns_class_wR_factor_all
```

REFLNS_SCALE

```
_reflns_scale_group_code
_reflns_scale_meas_F
_reflns_scale_meas_F_squared
_reflns_scale_meas_intensity
```

REFLNS_SHELL

```
_reflns_shell_d_res_high
_reflns_shell_d_res_low
_reflns_shell_meanI_over_uI_all
_reflns_shell_meanI_over_uI_gt
_reflns_shell_number_measured_all
_reflns_shell_number_measured_gt
_reflns_shell_number_possible
_reflns_shell_number_unique_all
_reflns_shell_number_unique_gt
_reflns_shell_percent_possible_all
_reflns_shell_percent_possible_gt
_reflns_shell_percent_possible_obs
_reflns_shell_Rmerge_F_all
_reflns_shell_Rmerge_F_gt
_reflns_shell_Rmerge_I_all
_reflns_shell_Rmerge_I_gt
```

The reflections actually used for the (final) refinement are stored in a CIF as a loop in the REFLN category. The various “REFLNS_” categories provide coarse-grained summaries of the reflections, either in total or as split up by resolution shell (REFLNS_SHELL), by scale group (REFLNS_SCALE) or by some other arbitrary category (REFLNS_SCALE). Note that, for conciseness of display, deprecated data names have been dropped from this and succeeding lists.

Analysis – reflections used in refinement

Individual reflections – structure factors

REFLN

```
_refln_index_h
_refln_index_k
_refln_index_l
_refln_A_calc
_refln_A_meas
_refln_B_calc
_refln_B_meas
_refln_class_code
_refln_crystal_id
_refln_d_spacing
_refln_F_calc
_refln_F_meas
_refln_F_sigma
_refln_F_squared_calc
_refln_F_squared_meas
_refln_F_squared_sigma
_refln_include_status
_refln_intensity_calc
_refln_intensity_meas
_refln_intensity_sigma
_refln_mean_path_length_tbar
_refln_phase_calc
_refln_phase_meas
_refln_refinement_status
_refln_scale_group_code
_refln_sint/lambda
_refln_symmetry_epsilon
_refln_symmetry_multiplicity
_refln_wavelength
_refln_wavelength_id
```

Example

```
loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   4   0   0       3307.04       2953.75        43.53 o
   6   0   0      11608.51      11255.44       106.34 o
   8   0   0       7328.16       7688.11        62.41 o
  10   0   0         72.13         76.83         7.67 o
  12   0   0       1309.75       1424.38        30.84 o
  14   0   0         92.59         64.57         7.43 o
  16   0   0          0.64          3.80         2.56 o
  18   0   0        493.14        464.68        28.68 o
   1   1   0        271.31        338.59         6.55 o
   3   1   0       3736.66       4174.68        29.94 o
   5   1   0         10.36         10.77         0.61 o
   7   1   0        562.07        658.90         9.17 o
```

Portalone, G. (2019). 6-Methyluracil: a redetermination of polymorph (II). *IUCrData*, **4**, x190861.

Structure factor listings generally comprise the intensities (or *F* squared values), the phases being generally unknown. The CIF dictionary does have provision for listing calculated and measured phases, for listing the *A* and *B* structure-factor components, wavelength associated with individual reflections (in the case of Laue, energy-dispersive or polychromatic methods), etc. In practice, most single-crystal monochromatic structure determinations have listings such as those shown in the example, which is from *SHELXL*-2014. Note that this still uses the deprecated data name **`_refln_observed_status`** instead of the preferred **`_refln_include_status.`**

Structure – the ATOM group

ATOM_SITES

```
_atom_sites_fract_tran_matrix_11
_atom_sites_fract_tran_matrix_12
_atom_sites_fract_tran_matrix_13
_atom_sites_fract_tran_matrix_21
_atom_sites_fract_tran_matrix_22
_atom_sites_fract_tran_matrix_23
_atom_sites_fract_tran_matrix_31
_atom_sites_fract_tran_matrix_32
_atom_sites_fract_tran_matrix_33
_atom_sites_Cartn_tran_matrix_11
_atom_sites_Cartn_tran_matrix_12
_atom_sites_Cartn_tran_matrix_13
_atom_sites_Cartn_tran_matrix_21
_atom_sites_Cartn_tran_matrix_22
_atom_sites_Cartn_tran_matrix_23
_atom_sites_Cartn_tran_matrix_31
_atom_sites_Cartn_tran_matrix_32
_atom_sites_Cartn_tran_matrix_33
_atom_sites_Cartn_tran_vector_1
_atom_sites_Cartn_tran_vector_2
_atom_sites_Cartn_tran_vector_3
_atom_sites_Cartn_transform_axes
```

```
_atom_sites_fract_tran_vector_1
_atom_sites_fract_tran_vector_2
_atom_sites_fract_tran_vector_3
_atom_sites_solution_hydrogens
_atom_sites_solution_primary
_atom_sites_solution_secondary
_atom_sites_special_details
```

ATOM_TYPE

```
_atom_type_symbol
_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
```

Most end-users are most interested in the atomic positional coordinates and anisotropic displacement parameters, which effectively describe the three-dimensional molecule or coordinated structure. Most of this information is in a single category (or looped list of data items), which we shall consider in a moment. For a full understanding of the structure described in the ATOM_SITE loop, one also needs to know the transformation matrix between Cartesian and fractional cell coordinates (described in the ATOM_SITES loop), as well as the elemental identity of the atom or atoms occupying each site. The latter can be inferred from the information in the ATOM_TYPE list, which gives scattering coefficients and other properties of each atom type. It is worth remembering that (averaged throughout the whole crystal) the “atoms” at each site may appear as an average of different elements or may have in some sense “unusual” scattering coefficients. All such factors can be recorded in the ATOM_TYPE data.

Structure – the ATOM group

ATOM_SITE

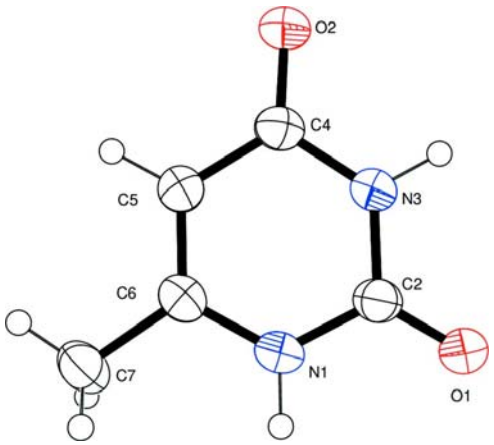
```
_atom_site_label
_atom_site_adp_type
_atom_site_aniso_B_11
_atom_site_aniso_B_12
_atom_site_aniso_B_13
_atom_site_aniso_B_22
_atom_site_aniso_B_23
_atom_site_aniso_B_33
_atom_site_aniso_label
_atom_site_aniso_ratio
_atom_site_aniso_type_symbol
_atom_site_aniso_U_11
_atom_site_aniso_U_12
_atom_site_aniso_U_13
_atom_site_aniso_U_22
_atom_site_aniso_U_23
_atom_site_aniso_U_33
_atom_site_attached_hydrogens
_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
_atom_site_constraints
_atom_site_description
_atom_site_disorder_assembly
_atom_site_disorder_group
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_label_component_0
_atom_site_label_component_1
_atom_site_label_component_2
_atom_site_label_component_3
_atom_site_label_component_4
_atom_site_label_component_5
_atom_site_label_component_6
_atom_site_occupancy
_atom_site_refinement_flags_posn
_atom_site_refinement_flags_adp
_atom_site_refinement_flags_occupancy
_atom_site_restraints
_atom_site_symmetry_multiplicity
_atom_site_type_symbol
_atom_site_U_equiv_geom_mean
_atom_site_U_iso_or_equiv
_atom_site_Wyckoff_symbol
```

The ATOM_SITE data are, as mentioned before, the things that most CIF end-users are most interested in. When these data were published in print journals, it was often the case that the anisotropic displacement parameters were listed in a separate table from the coordinates, purely as a matter of formatting convenience. This convention is still found in many CIFs, but it is quite possible (and from a computational point of view, possibly more efficient) to include all of these items in a single table (loop).

Structure – the ATOM group

EXAMPLE

```
loop_
    _atom_site_type_symbol
    _atom_site_label
    _atom_site_fract_x
    _atom_site_fract_y
    _atom_site_fract_z
    _atom_site_U_iso_or_equiv
    _atom_site_adp_type
    _atom_site_calc_flag
    O O1 0.06097(6) 0.5147(3) -0.06247(9) 0.0486(4) Uani d
    O O2 0.26885(6) 0.0108(3) 0.10454(9) 0.0468(3) Uani d
    N N1 0.07366(7) 0.2401(3) 0.07840(9) 0.0380(3) Uani d
    H H1 0.0293(10) 0.304(5) 0.0715(13) 0.049(5) Uiso d
    C C2 0.09750(8) 0.3466(4) 0.00759(11) 0.0366(4) Uani d
    N N3 0.16426(6) 0.2545(3) 0.02038(9) 0.0361(3) Uani d
    H H3 0.1820(10) 0.333(5) -0.0258(15) 0.057(5) Uiso d
    C C4 0.20846(7) 0.0710(4) 0.09840(11) 0.0357(4) Uani d
    C C5 0.17907(7) -0.0298(4) 0.16844(11) 0.0372(4) Uani d
    H H5 0.2066 -0.1641 0.2240 0.045 Uiso calc
    C C6 0.11357(8) 0.0595(4) 0.15811(11) 0.0359(4) Uani d
    C C7 0.08038(9) -0.0227(5) 0.23014(13) 0.0488(5) Uani d
    H H7A 0.1100 -0.1790 0.2783 0.073 Uiso calc
    H H7B 0.0354 -0.1289 0.1975 0.073 Uiso calc
    H H7C 0.0742 0.1864 0.2615 0.073 Uiso calc
```



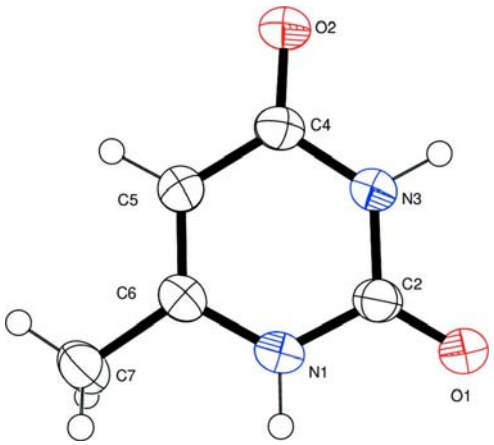
Portalone, G. (2019). 6-Methyluracil: a redetermination of polymorph (II). *IUCrData*, **4**, x190861.

In this (slightly edited) example, the atom coordinates are presented separately from the anisotropic displacement parameters. It is quite clear by inspection where atoms have been constrained – the positions of the hydrogens attached to the N atoms are refined, those attached to carbons fixed; but the explicit flags (“calc”) would permit an automated analysis based on such properties.

Structure – the ATOM group

EXAMPLE

```
loop_
    _atom_site_aniso_label
    _atom_site_aniso_U_11
    _atom_site_aniso_U_22
    _atom_site_aniso_U_33
    _atom_site_aniso_U_12
    _atom_site_aniso_U_13
    _atom_site_aniso_U_23
    O1 0.0372(6) 0.0662(8) 0.0451(7) 0.0128(5) 0.0178(5) 0.0152(5)
    O2 0.0319(6) 0.0639(8) 0.0475(7) 0.0091(5) 0.0178(5) 0.0105(5)
    N1 0.0300(7) 0.0451(7) 0.0423(7) 0.0023(5) 0.0171(5) 0.0025(6)
    C2 0.0312(7) 0.0414(8) 0.0387(8) 0.0008(6) 0.0142(6) -0.0007(6)
    N3 0.0308(6) 0.0435(7) 0.0365(7) 0.0021(5) 0.0150(5) 0.0031(5)
    C4 0.0304(7) 0.0389(7) 0.0383(8) -0.0006(6) 0.0130(6) -0.0022(6)
    C5 0.0344(8) 0.0416(8) 0.0359(8) 0.0005(6) 0.0129(6) 0.0038(6)
    C6 0.0369(8) 0.0356(7) 0.0385(8) -0.0035(6) 0.0174(6) -0.0016(6)
    C7 0.0496(10) 0.0552(10) 0.0519(11) -0.0006(7) 0.0308(8) 0.0042(7)
```



Portalone, G. (2019). 6-Methyluracil: a redetermination of polymorph (II). *IUCrData*, **4**, x190861.

The anisotropic displacement parameters are presented in a separate loop, but formally this and the coordinate table on the previous slide are part of the same category, and so could be presented together in a composite loop. The arrangement here is historic, because of constraints of publishing in print.

Structure – the CHEMICAL group

CHEMICAL

```
_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
```

CHEMICAL_FORMULA

```
_chemical_formula_analytical
_chemical_formula_iupac
_chemical_formula_moiety
_chemical_formula_structural
_chemical_formula_sum
_chemical_formula_weight
_chemical_formula_weight_meas
```

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
```

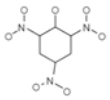
CHEMICAL_CONN_BOND

```
_chemical_conn_bond_atom_1
_chemical_conn_bond_atom_2
_chemical_conn_bond_type
```

This group of categories characterises the chemistry of the material whose structure is being determined. The CHEMICAL and CHEMICAL_FORMULA items relate to the gross properties of the material; the CHEMICAL_CONN_* items provide a very simple way to describe chemical connectivity (*i.e.* provide a two-dimensional chemical structure outline, but without indications of bond types, charge distribution, etc. In principle this could allow simple substructure searching (at least of organic molecules). These categories exist in the core dictionary because the typical single-crystal structure determination is interested in determining the two- and three-dimensional structure of a single chemical species. For mmCIF, a different approach was followed, because biological macromolecules are often bound to a variety of other molecules and ions. Since the focus of interest is usually the protein itself, the various ligand species are often handled as idealised structures (and are perhaps not refined). We will see in the mmCIF that they are described by the CHEM_COMP (meaning “chemical component”) categories.

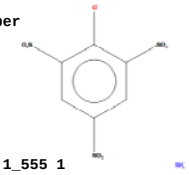
Structure – the CHEMICAL group

EXAMPLE (CIF)



```
loop_
    _chemical_conn_atom_number
    _chemical_conn_atom_type_symbol
    _chemical_conn_atom_display_x
    _chemical_conn_atom_display_y
    _chemical_conn_atom_NCA
    _chemical_conn_atom_NH
    1  C  4.8497  2.9504  2  1
    2  C  4.8497  4.3504  3  0
    3  C  3.6373  5.0504  3  0
    4  C  2.4249  4.3504  3  0
    5  C  2.4249  2.9504  2  1
    6  C  3.6373  2.2504  3  0
    7  O  3.6373  6.4504  1  0
    8  N  1.2124  5.0504  3  0
    9  O  1.2124  6.4504  1  0
    10 O  0.0000  4.3504  1  0
    11 N  3.4393  0.8644  3  0
    12 O  2.1401  0.3429  1  0
    13 O  4.5406  0.0000  1  0
    14 N  6.0622  5.0504  3  0
    15 O  7.2746  4.3504  1  0
    16 O  6.0622  6.4504  1  0
```

EXAMPLE (MIF)



```
loop_
_ccdc_atom_site_atom_id_number
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_ccdc_atom_site_symmetry
_ccdc_atom_site_base
1 C1 C 0.2500 0.1280(1) 0.0 1_555 1
2 O1 O 0.2500 0.06477(7) 0.0 1_555 2
3 C2 C 0.1668(1) 0.16901(8) -0.0608(2) 1_555 3
4 N1 N 0.0776(1) 0.13664(7) -0.1342(2) 1_555 4
5 O2 O 0.05679(9) 0.07942(7) -0.0825(2) 1_555 5
6 O3 O 0.02663(9) 0.16864(7) -0.2449(2) 1_555 6
7 C3 C 0.1672(1) 0.23847(8) -0.0632(2) 1_555 7
8 C4 C 0.2500 0.2725(1) 0.0 1_555 8
9 N2 N 0.2500 0.3458(1) 0.0 1_555 9
10 O4 O 0.1773(1) 0.37478(7) -0.0604(2) 1_555 10
11 N3 N 0.1165(2) 0.0 0.2500 1_555 11
12 H1 H 0.111(1) 0.261(1) -0.112(3) 1_555 12
13 H2 H 0.153(2) 0.017(1) 0.148(3) 1_555 13
14 H3 H 0.076(2) 0.030(1) 0.301(3) 1_555 14
15 C3G C 0.3328 0.23847 0.0632 8_555 7
16 C2G C 0.3332 0.16901 0.0608 8_555 3
17 N1G N 0.4224 0.13664 0.1342 8_555 4
18 O2G O 0.44321 0.07942 0.0825 8_555 5
19 O3G O 0.47337 0.16864 0.2449 8_555 6
20 H1G H 0.389 0.261 0.112 8_555 12
21 O4G O 0.3227 0.37478 0.0604 8_555 10
22 H2B H 0.153 -0.017 0.352 3_555 13
23 H3B H 0.076 -0.030 0.199 3_555 14

loop_
_atom_type_symbol
_atom_type_radius_bond
C 0.68
H 0.23
N 0.68
O 0.68

loop_
_atom_id
_atom_type
_atom_attach_nh
_atom_attach_h
_atom_charge
1 C 3 0 0
2 O 1 0 -1
3 C 3 0 0
4 N 3 0 0
5 O 1 0 0
6 O 1 0 0
7 C 2 1 0
8 C 3 0 0
9 N 3 0 0
10 O 1 0 0
11 N 0 4 1
15 C 2 1 0
16 C 3 0 0
17 N 3 0 0
18 O 1 0 0
19 O 1 0 0
21 O 1 0 0

loop_
_bond_id_1
_bond_id_2
_bond_type_ccdc
_bond_environment
1 2 S chain
3 1 A ring
4 3 S chain
5 4 D chain
6 4 D chain
7 3 A ring
8 7 A ring
9 8 S chain
10 9 D chain
11 13 S chain
12 7 S chain
14 11 S chain
15 8 A ring
16 1 A ring
17 16 S chain
18 17 D chain
19 17 D chain
20 15 S chain
21 9 D chain
22 11 S chain
23 11 S chain
15 16 A ring
```

The idea behind this slide is to demonstrate the very rudimentary abilities of the CHEMICAL_CONN_* groups with the slightly more expressive MIF (molecular information file) approach. Even MIF does not accommodate the full range of expression required for chemical structure representation, and it accommodates different conventions for describing bond orders (i.e. is not an unequivocal standard). The CIF content was not fully developed and is rarely, if ever, used. It was expected that a more complex formalism such as MIF (which used more of the STAR File feature set than CIF) would be necessary to describe chemical structures properly, but MIF never gained much attraction in the chemistry community. The Cambridge Structural database can output CIF/MIF files that include three- and two-dimensional structure descriptions. As these conform to the CIF syntax, they are restricted to a subset of MIF features. Their usefulness appears to be limited.

Structure – the GEOM group

GEOM
GEOM_ANGLE
GEOM_BOND
GEOM_CONTACT
GEOM_HBOND
GEOM_TORSION

EXAMPLE

```
loop_
    _geom_bond_atom_site_label_1
    _geom_bond_atom_site_label_2
    _geom_bond_site_symmetry_2
    _geom_bond_distance
    _geom_bond_publ_flag
    Zn1 N12 . 2.1454(18) y
    Zn1 N12 2_656 2.1454(18) ?
    Zn1 N14 . 2.1704(18) y
    Zn1 N14 2_656 2.1705(18) ?
    Zn1 N1 . 2.2101(19) y
    Zn1 N1 2_656 2.2101(19) ?
    N1 C1 . 1.323(3) ?
    N1 C5 . 1.350(3) ?
    C1 C2 . 1.394(4) ?
    C1 H1 . 0.9300 ?
    C2 C3 . 1.347(5) ?
    C2 H2 . 0.9300 ?
    C3 C4 . 1.399(5) ?
    C3 H3 . 0.9300 ?
    C4 C5 . 1.402(3) ?
    C4 C6 . 1.426(5) ?
```

Table 1
Selected geometric parameters (Å, °)

Zn1—N12	2.1454 (18)	Zn1—N1	2.2101 (19)
Zn1—N14	2.1704 (18)		
N12—Zn1—N12^{i}	91.24 (10)	N14—Zn1—N1	89.34 (7)
N12—Zn1—N14	90.43 (7)	N12—Zn1—N1^{i}	171.59 (7)
N12—Zn1—N14^{i}	88.55 (7)	N14—Zn1—N1^{i}	91.83 (7)
N14—Zn1—N14^{i}	178.54 (11)	N1—Zn1—N1^{i}	75.01 (11)
N12—Zn1—N1	96.92 (7)		

Symmetry code: (i) $-x+1, y, -z+\frac{3}{2}$.

Ossinger, S., Näther, C. & Tuczek, F. (2019). Crystal structure of bis[dihydrobis(pyrazol-1-yl)borato-$\kappa^2N^2,N^{2'}$](1,10-phenanthroline- κ^2N,N')zinc(II). *Acta Cryst*. E**75**, 1112-1116.

The various geometry categories require little detailed explanation: they are a straightforward way to list geometric values derived from the structural coordinates. As purely derived quantities, their inclusion in a CIF is in one sense redundant. However, they serve as a useful validation cross check, and they are convenient for identifying specific values that the author wishes to appear in a publication, as demonstrated in this example. It is perhaps worth pointing out here that in IUCr journals, the “supporting crystallographic data” which are available for every published structure allow visualisation of any geometric parameter (bond distance, angle, torsion, hydrogen bond) simply by clicking on the table entry.

Structure – SYMMETRY and SPACE_GROUP

SYMMETRY

```
_symmetry_Int_Tables_number
_symmetry_cell_setting
_symmetry_space_group_name_H-M
_symmetry_space_group_name_Hall
```

SYMMETRY_EQUIV

```
_symmetry_equiv_pos_site_id
_symmetry_equiv_pos_as_xyz
```

SPACE_GROUP

```
_space_group_crystal_system
_space_group_id
_space_group_IT_number
_space_group_name_Hall
_space_group_name_H-M_alt
```

SPACE_GROUP_SYMOP

```
_space_group_symop_operation_xyz
_space_group_symop_sg_id
```

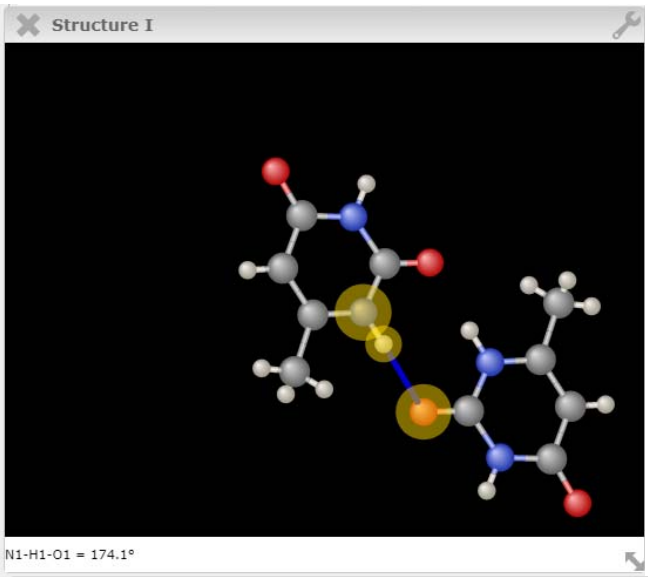
EXAMPLE

```
_space_group_crystal_system     monoclinic
_space_group_name_H-M_alt     'C 2/c'
_space_group_name_Hall     '-C 2yc'

loop_
    _space_group_symop_operation_xyz
    'x, y, z'
    '-x, y, -z+1/2'
    'x+1/2, y+1/2, z'
    '-x+1/2, y+1/2, -z+1/2'
    '-x, -y, -z'
    'x, -y, z-1/2'
    '-x+1/2, -y+1/2, -z'
    'x+1/2, -y+1/2, z-1/2'
```

Portalone, G. (2019). 6-Methyluracil: a redetermination of polymorph (II). *IUCrData*, **4**, x190861.

This slide shows the deprecated SYMMETRY and replacement SPACE_GROUP categories, with an example from a recent *SHELXL*-2014 CIF demonstrating that real-world CIFs are still missing the formal identifiers in the symmetry operations list, compromising the position-independent nature of CIF data structures.




```
loop_
  _space_group_symop_id
  _space_group_symop_operation_xyz
  1  'x,y,z'               2   '-x, y,-z+1/2'
  3  'x+1/2,y+1/2,z'       4   '-x+1/2,y+1/2,-z+1/2'
  5  '-x,-y,-z'            6   'x,-y,z-1/2'
  7  '-x+1/2,-y+1/2,-z'    8   'x+1/2,-y+1/2,z-1/2'

loop_
  _geom_hbond_atom_site_label_D
  _geom_hbond_atom_site_label_H
  _geom_hbond_atom_site_label_A
  _geom_hbond_site_symmetry_A
  _geom_hbond_distance_DH
  _geom_hbond_distance_HA
  _geom_hbond_distance_DA
  _geom_hbond_angle_DHA
  N1 H1 O1 5_565 0.91(2) 1.95(2) 2.8594(17) 174.0(16)
  N3 H3 O2 7     0.93(2) 1.90(2) 2.8246(18) 171.5(18)
```

Portalone, G. (2019). 6-Methyluracil: a redetermination of polymorph (II). *IUCrData*, **4**, x190861.

Hydrogen-bond geometry (Å, º)

D—H···*A*	*D*—H	H···*A*	*D*···*A*	*D*—H···*A*
N1—H1···O1^{i}	0.91 (2)	1.95 (2)	2.8594 (17)	174.0 (16)
N3—H3···O2^{ii}	0.93 (2)	1.90 (2)	2.8246 (18)	171.5 (18)

Symmetry codes: (i) −*x*, −*y*+1, −*z*; (ii) −*x*+1/2, −*y*+1/2, −*z*.

This example links the GEOM and SPACE_GROUP categories. The **…_site_symmetry_A** entry is a pointer to the symmetry operator with a matching value of **_space_group_symop_id**. Here I am still keeping the real-world data names (*i.e.* in DDL1 all-underscore formalism), but I have properly added in the **_space_group_symop_id** entries.

Structure – the VALENCE group

VALENCE_PARAM
VALENCE_REF

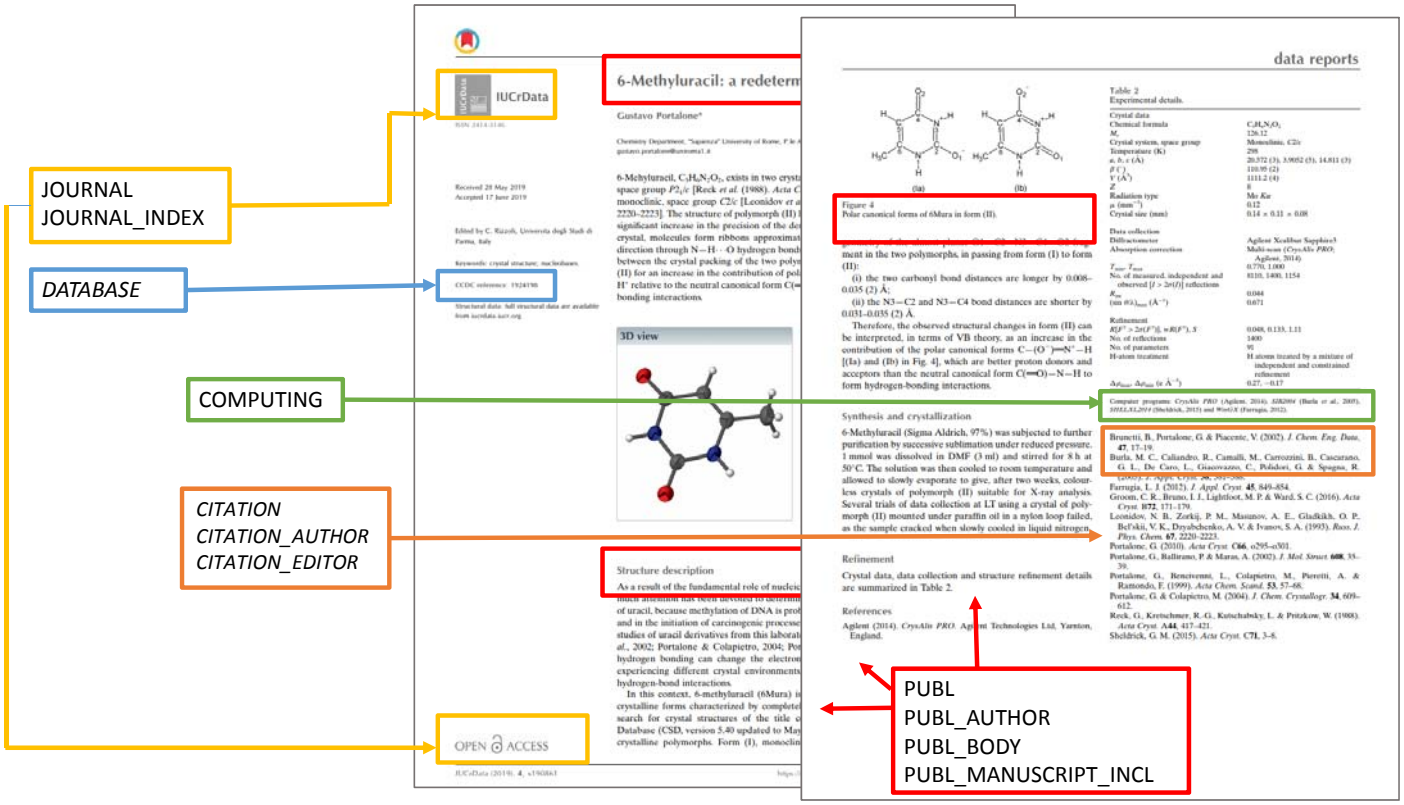
EXAMPLE

```
loop_
    _valence_param_atom_1
    _valence_param_atom_1_valence
    _valence_param_atom_2
    _valence_param_atom_2_valence
    _valence_param_Ro
    _valence_param_B
    _valence_param_ref_id
    _valence_param_details
      Cu 2 O -2 1.679 0.37 a .
      Cu 2 O -2 1.649 0.37 j .
      Cu 2 N -3 1.64  0.37 m '2-coordinate N'
      Cu 2 N -3 1.76  0.37 m '3-coordinate N'

loop_
    _valence_ref_id
    _valence_ref_reference
       a
            'Brown & Altermatt (1985), Acta Cryst. B41, 244-247'
       j
            'Liu & Thorp (1993), Inorg. Chem. 32, 4102-4205'
       m
            'See Krause & Strub (1998), Inorg. Chem. 37, 5369-5375'
```

Bond valences are calculated in inorganic structures and provide another opportunity for validating the chemical reasonableness of the structure.

Publication



This is a rather busy slide, but in practice will be shown in stages. The various category groups associated with each part of an article are keyed visually to an example paper. The JOURNAL categories provide bibliographic metadata about the publication. The PUBL categories include the textual content of the article. The COMPUTING, DATABASE and CITATION categories link to software, database entries and literature citations respectively. The latter two category groups are not generally used by IUCr journals (linking information is expected to be embedded in the article text), but were imported from the original mmCIF dictionary which considered it important to have granular linking between entries in structural and bibliographic databases. The detailed content of these categories will be explored by example in the publication tutorials.

File metadata – the AUDIT group

AUDIT
- _audit_block_code
- _audit_creation_date
- _audit_creation_method
- _audit_update_record

AUDIT_AUTHOR
- _audit_author_address
- _audit_author_name

AUDIT_CONFORM
- _audit_conform_dict_location
- _audit_conform_dict_name
- _audit_conform_dict_version

AUDIT_CONTACT_AUTHOR
- _audit_contact_author_address
- _audit_contact_author_email
- _audit_contact_author_fax
- _audit_contact_author_name
- _audit_contact_author_phone

AUDIT_LINK
- _audit_link_block_code
- _audit_link_block_description

EXAMPLE

```
data_example

_audit_block_code                 xyzzy_2002-04-05
_audit_creation_date              2002-04-05
_audit_creation_method            'SHELXL97'

_audit_update_record
; 2002-04-09  discussion added                      BM
  2002-04-17  coeditor number XY1234 assigned      BM
  2002-04-18  revised comment after referee report BM
;

loop_
  _audit_conform_dict_name
  _audit_conform_dict_version
  _audit_conform_dict_location
    cif_core.dic      2.3   .
    cif_pd.dic        1.0   .
```

The AUDIT family of categories supply a variety of types of metadata concerning the file itself. For crystal structure determinations, AUDIT_AUTHOR can be used to identify the crystallographers involved in data collection, who are not always listed amongst the publication authors. AUDIT_CONFORM provides a mechanism for identifying an appropriate level of validation based on the detailed content of the contemporary dictionaries when the file was generated. AUDIT_LINK may describe the relationships between data blocks (*e.g.* the separate phases or substructures identified in a composite material).

mmCIF dictionary – history

- Protein Data Bank established 1971
- Biological macromolecular structures “deposit first, then publish”
- PDB file format well established during 1970s, 1980s
- By 1990s, shortcomings becoming apparent
- Seemed a natural extension to the new CIF standard
- Long and difficult development (1991-1997)

As with the history of the core dictionary, we do not want to go into too much detail. Unlike the case for small-unit-cell structures, there was already a de facto standard for data requirements in the form of the PDB file. However, developed initially in the Fortran fixed-format style, the PDB format was already showing limitations in coping with the size of large molecules, and its annotations describing detailed structural features were free-text and therefore unstructured. Despite some “hacks” to improve matters, it was reasonably clear that a radically new approach was needed.

mmCIF dictionary – philosophy

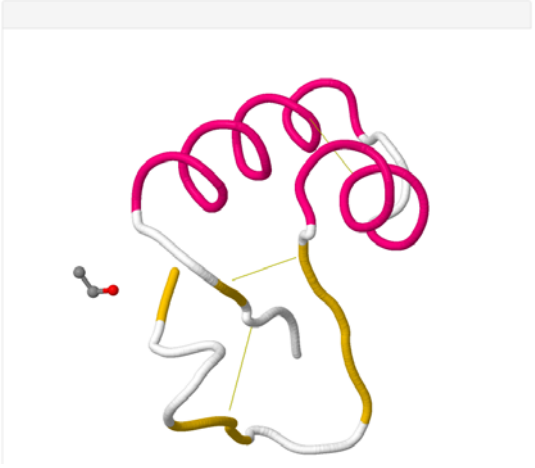
- Provide a full description of the macromolecular experiment and its results
- Specify relationships between different components of a complex macromolecular structure
- Required richer dictionary definition language (DDL2)
- Dictionary concepts translated well to relational database model
- mmCIF category structure underpins wwPDB database schema
- However, only fully accepted as data exchange format in 2010s

In practice, capturing all the information needed to give a complete description of the experiment and the structure was a major undertaking, both in terms of specifying the necessary data items, and subsequently in persuading the community to provide them. The technical challenge was first met by the mmCIF dictionary in 1997, subsequently greatly enlarged and refined with the PDBx family of extensions. Persuading the community to provide all the information that could be of use for future scientific re-use and development remains something of a challenge.

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION



Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Specification of three distinct components of the crambin structure

```
loop_
  _struct_asym.id
  _struct_asym.entity_id
  _struct_asym.details
   chain_a A       'single polypeptide chain'
   ethanol ethanol 'cocrystallized ethanol molecule'
   water   HOH     .
```

Identification of the biological function of the structural components of crambin

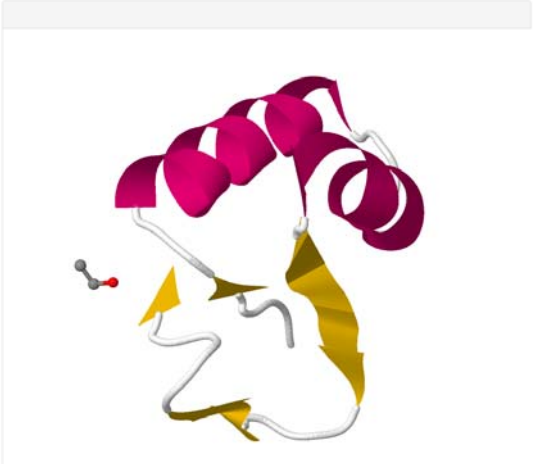
```
_struct_biol.id                 crambin_1
_struct_biol.details
; The function of this protein is unknown and
  therefore the biological unit is assumed to be
  the single polypeptide chain without
  co-crystallization factors i.e. ethanol.
;
_struct_biol_gen.biol_id        crambin_1
_struct_biol_gen.asym.id        chain_a
_struct_biol_gen.symmetry       1_555
```

Structural biologists are very interested not only in the structure of a protein or nucleic acid molecule, but also in its biological function. The mmCIF dictionary was designed to allow detailed annotation of many of these ancillary features of interest. We illustrate some of these with a very simple example: crambin, a small seed storage protein from the Abyssinian cabbage. There is a hierarchy of levels of structure description. At the topmost level, the STRUCT category group describes the gross structure – the components of the crystallography asymmetric unit, the biological function of the different components of a protein assembly, the secondary structure of the proteins, intramolecular interactions.

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION



Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Description of the secondary structure of the crambin protein

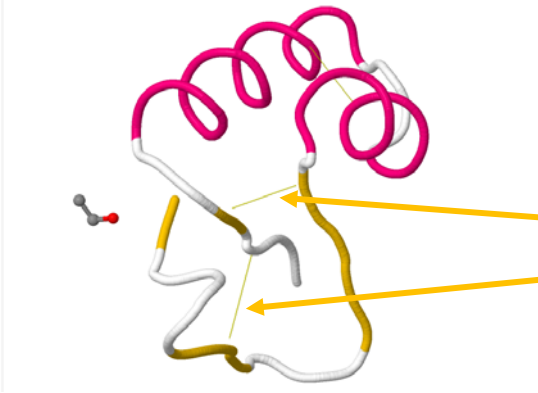
```
loop_
  _struct_conf.id
  _struct_conf.conf_type.id
  _struct_conf.beg_label_comp_id
  _struct_conf.beg_label_asym_id
  _struct_conf.beg_label_seq_id
  _struct_conf.end_label_comp_id
  _struct_conf.end_label_asym_id
  _struct_conf.end_label_seq_id
  _struct_conf.details
    H1 HELX_RH_AL_P ILE chain_a  7 PRO chain_a 19
                                  'HELX-RH3T 17-19'
    H2 HELX_RH_AL_P GLU chain_a 23 THR chain_a 30
                                  'Alpha-N start'
    S1 STRN_P       CYS chain_a 32 ILE chain_a 35 .
    S2 STRN_P       THR chain_a  1 CYS chain_a  4 .
    S3 STRN_P       ASN chain_a 46 ASN chain_a 46 .
    S4 STRN_P       THR chain_a 39 PRO chain_a 41 .
    T1 TURN-TY1_P   ARG chain_a 17 GLY chain_a 20 .
    T2 TURN-TY1_P   PRO chain_a 41 TYR chain_a 44 .
```

The secondary structure (STRUCT_CONF refers to the backbone conformation) is described in terms of structural motifs beginning and ending at certain residues in the polypeptide (or nucleic acid). Here we see a catalogue of right-handed alpha helices, beta strands and type I turns.

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION



Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Interactions between portions of the crambin structure

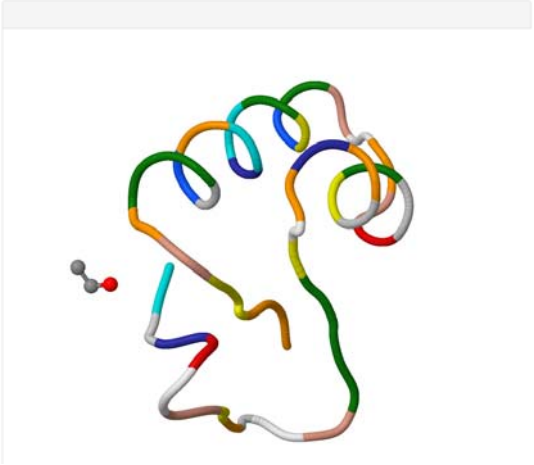
```
loop_
  _struct_conn.id
  _struct_conn.conn_type_id
  _struct_conn.ptnr1_label_comp_id
  _struct_conn.ptnr1_label_asym_id
  _struct_conn.ptnr1_label_seq_id
  _struct_conn.ptnr1_label_atom_id
  _struct_conn.ptnr1_role
  _struct_conn.ptnr1_symmetry
  _struct_conn.ptnr2_label_comp_id
  _struct_conn.ptnr2_label_asym_id
  _struct_conn.ptnr2_label_seq_id
  _struct_conn.ptnr2_label_atom_id
  _struct_conn.ptnr2_role
  _struct_conn.ptnr2_symmetry
  _struct_conn.details
    SS2 disulf CYS chain_a  4 S  .        1_555
               CYS chain_a 32 S  .        1_555 .
    SS1 disulf CYS chain_a  3 S  .        1_555
               CYS chain_a 40 S  .        1_555 .
    HB1 hydrog SER chain_a  6 OG positive 1_555
               LEU chain_a  8 O  negative 1_556 .
    HB2 hydrog ARG chain_a 17 N  positive 1_555
               ASP chain_a 43 O  negative 1_554 .
```

The STRUCT_CONN loop describes two of the three disulfide bonds shown in the *Jmol* figure, as well as two hydrogen bonds (not shown).

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION



Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Description of the crambin polypeptide

```
_entity_name_com.entity_id       A
_entity_name_com.name            crambin
_entity_src_nat.entity_id        A
_entity_src_nat.common_name      'Abyssinian Cabbage'
_entity_src_nat.genus            Crambe
_entity_src_nat.species          abyssinica
_entity_src_nat.details          ?
_entity_poly.entity_id           A
_entity_poly.type                polypeptide(L)
_entity_poly.nstd_chirality      no
_entity_poly.nstd_linkage        no
_entity_poly.nstd_monomers       no
_entity_poly.type_details
 'Sequence heterogeneity at residues 22 and 25'
loop_
  _entity_poly_seq.entity_id
  _entity_poly_seq.num
  _entity_poly_seq.mon_id
    A    1   THR     A    2   THR
\# - - abbreviated - - -
    A   22   PRO     A   22   SER
    A   23   GLU     A   24   ALA
    A   25   LEU     A   25   ILE
\# - - abbreviated - - -
    A   47   ALA     A   48   ASN
```

The distinct chemical entities in the assembly can be annotated in significant detail. In this example the amino acid sequence is listed, together with some characteristics of the protein molecule itself (as distinct from ligands or other complexed molecules).

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION

Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Separate chemical components forming the crambin polypeptide

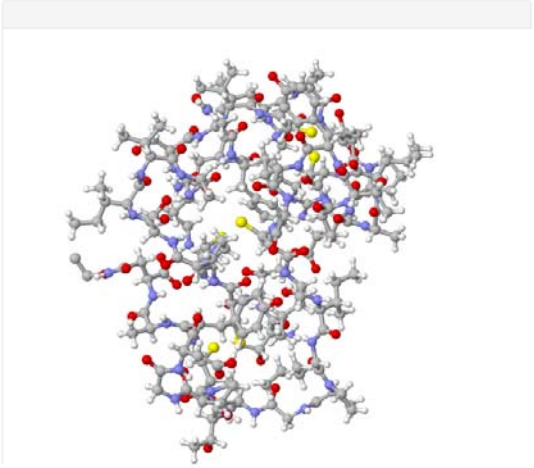
```
loop_
_chem_comp.id
_chem_comp.mon_nstd_flag
_chem_comp.formula
_chem_comp.name
  ethanol .   'C2 H6 O1'        "ethanol"
  ALA     yes 'C3 H7 N1 O2'     "alanine"
  ARG     yes 'C6 H14 N4 O2'    "arginine"
  ASN     yes 'C4 H8 N2 O3'     "asparagine"
  ASP     yes 'C4 H7 N1 O4'     "aspartic acid"
  CYS     yes 'C3 H7 N1 O2 S1'  "cysteine"
  GLU     yes 'C5 H9 N1 O4'     "glutamic acid"
  GLY     yes 'C2 H5 N1 O2'     "glycine"
  ILE     yes 'C6 H13 N1 O2'    "isoleucine"
  LEU     yes 'C6 H13 N1 O2'    "leucine"
  PHE     yes 'C9 H11 N1 O2'    "phenylalanine"
  PRO     yes 'C5 H9 N1 O2'     "proline"
  SER     yes 'C3 H7 N1 O3'     "serine"
  THR     yes 'C4 H9 N1 O3'     "threonine"
  TYR     yes 'C9 H11 N1 O3'    "tyrosine"
  VAL     yes 'C5 H11 N1 O2'    "valine"
```

The CHEM_COMP category provides the chemical identity of each monomeric residue in the polypeptide assembly. The (idealised) structure of each such molecule (and small molecules that appear as ligands) is described using other categories within the CHEM_COMP group.

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION



Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.

Atomic positional coordinates of the crambin model

```
loop_
_atom_site.label_seq_id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_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.label_entity_id
_atom_site.id
1  N  N    THR chain_a A  16.864  14.059   3.442
                          0.80  6.22  .   A    1
1  N  N    THR chain_a B  17.633  14.126   4.146
                          0.20  8.40  .   A    2
1  C  CA   THR chain_a A  16.868  12.814   4.233
                          0.80  4.45  .   A    3
1  C  CA   THR chain_a B  17.282  12.671   4.355
                          0.20  7.82  .   A    4
1  C  C    THR chain_a .  15.583  12.775   4.990
                          1.00  4.39  .   A    5
1  O  O    THR chain_a .  15.112  13.824   5.431
                          1.00  7.04  .   A    6
```

At the most granular level, each atomic position can be traced upwards through residues, helices, chains and molecules to provide a complete structure description of the macromolecular complex.

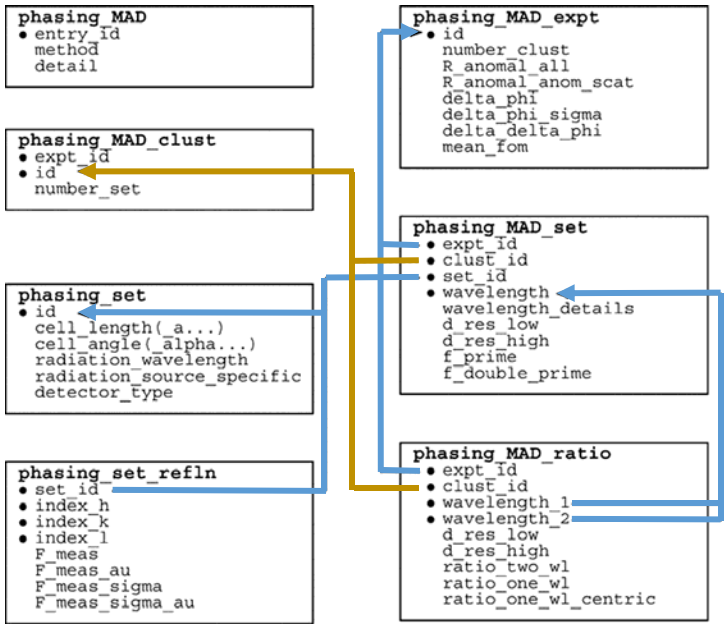
mmCIF Analysis – phasing

Overall description of phasing
PHASING
Phasing via molecular averaging
PHASING_AVERAGING
Phasing via isomorphous replacement
PHASING_ISOMORPHOUS
Phasing via multiple-wavelength anomalous dispersion
PHASING_MAD
PHASING_MAD_CLUST
PHASING_MAD_EXPT
PHASING_MAD_RATIO
PHASING_MAD_SET

Phasing via multiple isomorphous replacement
PHASING_MIR
PHASING_MIR_DER
PHASING_MIR_DER_REFLN
PHASING_MIR_DER_SHELL
PHASING_MIR_DER_SITE
PHASING_MIR_DER_SHELL
Phasing data sets
PHASING_SET
PHASING_SET_REFLN

Because of the size and complexity of biological macromolecules, a variety of phasing strategies may need to be employed in solving the structure. This list of the phasing-related categories demonstrates the different approaches involved in the molecular averaging, single and multiple isomorphous replacement, and multiple-wavelength anomalous dispersion methods. There are complex interactions between the categories; the PHASING_SET and PHASING_SET_REFLNS categories allow intensity and phase information for the data sets to be used in phasing to be stored in the same data block as the information for the refined structure. This approach helps to encourage validation and reproducibility of the derived structural model.

mmCIF Analysis – phasing



Because of the size and complexity of biological macromolecules, a variety of phasing strategies may need to be employed in solving the structure. This list of the phasing-related categories demonstrates the different approaches involved in the molecular averaging, single and multiple isomorphous replacement, and multiple-wavelength anomalous dispersion methods. There are complex interactions between the categories; the PHASING_SET and PHASING_SET_REFLNS categories allow intensity and phase information for the data sets to be used in phasing to be stored in the same data block as the information for the refined structure. This approach helps to encourage validation and reproducibility of the derived structural model.

mmCIF Analysis – phasing

EXAMPLE

```
_phasing_MAD.entry_id        'NCAD'

loop_
_phasing_MAD_expt.expt_id
_phasing_MAD_expt.number_clust
_phasing_MAD_expt.R_normal_all
_phasing_MAD_expt.R_normal_anom_scat
_phasing_MAD_expt.delta_delta_phi
_phasing_MAD_expt.delta_phi_sigma
_phasing_MAD_expt.mean_fom
   1  2  0.063  0.451  58.5  20.3  0.88
   2  1  0.051  0.419  36.8  18.2  0.93

loop_
_phasing_MAD_ratio.expt_id
_phasing_MAD_ratio.clust_id
_phasing_MAD_ratio.wavelength_1
_phasing_MAD_ratio.wavelength_2
_phasing_MAD_ratio.d_res_low
_phasing_MAD_ratio.d_res_high
_phasing_MAD_ratio.ratio_two_wl
_phasing_MAD_ratio.ratio_one_wl
_phasing_MAD_ratio.ratio_one_wl_centric
   1  'four wavelength'  1.4013 1.4013  20.00  4.00
        .  0.084 0.076
   1  'four wavelength'  1.4013 1.3857  20.00  4.00
     0.067      .     .
   1  'four wavelength'  1.4013 1.3852  20.00  4.00
     0.051      .     .
```

```
loop_
_phasing_MAD_clust.id
_phasing_MAD_clust.expt_id
_phasing_MAD_clust.number_set
  'four wavelength'  1  4
  'five wavelength'  1  5
  'five wavelength'  2  5


loop_
_phasing_MAD_set.expt_id
_phasing_MAD_set.clust_id
_phasing_MAD_set.set_id
_phasing_MAD_set.wavelength
_phasing_MAD_set.wavelength_details
_phasing_MAD_set.d_res_low
_phasing_MAD_set.d_res_high
_phasing_MAD_set.f_prime
_phasing_MAD_set.f_double_prime
   1 'four wavelength' aa 1.4013 'pre-edge'  20.00
      3.00   -12.48   3.80
   1 'four wavelength' bb 1.3857 'peak'      20.00
      3.00.  -31.22  17.20
```

MAD phasing of the structure of N-cadherin (Shapiro *et al.*, 1995) described using data items in the PHASING_MAD and related categories.

This is a small portion of an example of MAD phasing using data from two experiments, the first of which had two clusters, using respectively four and five wavelengths. The second experiment was a single five-wavelength cluster.
Reference: Shapiro, L., Fannon, A. M., Kwong, P. D., Thompson, A., Lehmann, M. S., Grubel, G., Legrand, J. F., Als-Nielsen, J., Colman, D. R. & Hendrickson, W. A. (1995). Structural basis of cell-cell adhesion by cadherins. *Nature* (*London*) **374**, 327-337.

mmCIF structure – alternative conformations

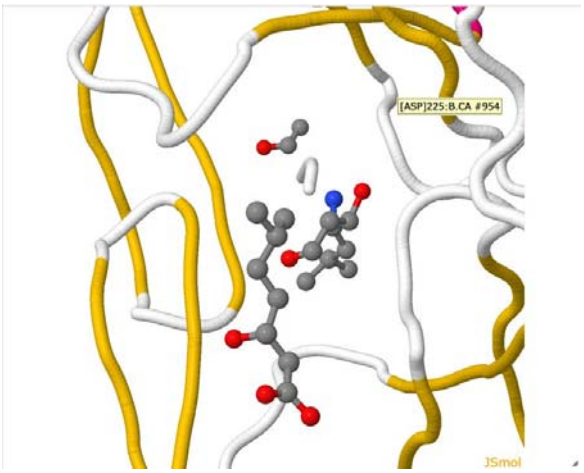
```
loop_
_atom_sites_alt_ens.id
_atom_sites_alt_ens.details
  'Ensemble 1'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  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
```

EXAMPLE



Alternative conformations in an HIV-1 protease structure (Fitzgerald *et al.*, 1990) described with data items in the ATOM_SITES_ALT, ATOM_SITES_ENS and ATOM_SITES_GEN categories.

Portion of the description of two conformations in which the inhibitor binds to the enzyme in a HIV-a protease structure. Reference: Fitzgerald, P.M., McKeever, B.M., VanMiddlesworth, J.F., Springer, J.P., Heimbach, J.C., Leu, C.T., Herber, W.K., Dixon, R.A., Darke, P.L. (1990). Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0-Å resolution. *J. Biol. Chem.* **265**, 14209-14219.

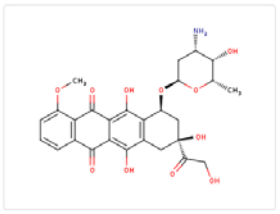
mmCIF structure – CHEM_COMP

EXAMPLE

```
_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 - - -
```



Rotate | Hydrogens | Labels

Chemical Component Summary

Name	DOXORUBICIN
Identifiers	(7S,9S)-7-[(2R,4S,5S,6S)-4-amino-5-hydroxy-6-methyl-oxan-2-yl]oxy-6,9,11-trihydroxy-9-(2-hydroxyethanoyl)-4-methoxy-8,10-dihydro-7H-tetracene-5,12-dione
Formula	$C_{27}H_{29}NO_{11}$
Molecular Weight	543.52
Type	NON-POLYMER
Isomeric SMILES	COc1cccc2C(=O)c3c(O)c4C[C@](O)(C[C@H](O[C@H]5C[C@H](N)[C@H](O)[C@H](C)O5)c4c(O)c3C(=O)c12)C(=O)CO
InChI	InChI=1S/C27H29NO11/c1-10-22(31)13(28)6-17(38-10)39-15-8-27(36,16(30)9-29)7-12-19(15)26(35)21-20(24(12)33)23(32)11-4-3-5-14(37-2)18(11)25(21)34/h3-5,10,13,15,17,22,29,31,33,35-36H,6-9,28H2,1-2H3/t10-,13-,15-,17-,22+,27-/m0/s1
InChIKey	AOJJSUZBOXZQNB-TZSSRYMLSA-N

The CHEM_COMP category provides a mechanism for describing the (idealised) chemical structure of components of a macromolecular complex, which could be the individual amino acids in a polypeptide chain or ligands which bind to an enzyme. The Protein Data Bank maintains a ligand database which characterises all of these chemical species. It provides a significant extension to the scope of the CHEMICAL category of the CIF dictionary, and in conjunction with the mmCIF CHEM_LINK group is better suited to the description of linked assemblies of repeating monomers that characterises many biological macromolecular structures.

mmCIF structure – the STRUCT categories

Higher-level macromolecular structure
STRUCT
STRUCT_ASYM
STRUCT_BIOL
STRUCT_BIOL_GEN
STRUCT_BIOL_KEYWORDS
STRUCT_BIOL_VIEW

Secondary structure
STRUCT_CONF
STRUCT_CONF_TYPE

Structural interactions
STRUCT_CONN
STRUCT_CONN_TYPE

Structural features of monomers
STRUCT_MON_DETAILS
STRUCT_MON_NUCL
STRUCT_MON_PROT
STRUCT_MON_PROT_CIS

Noncrystallographic symmetry
STRUCT_NCS_DOM
STRUCT_NCS_DOM_LIM
STRUCT_NCS_ENS
STRUCT_NCS_ENS_GEN
STRUCT_NCS_OPER

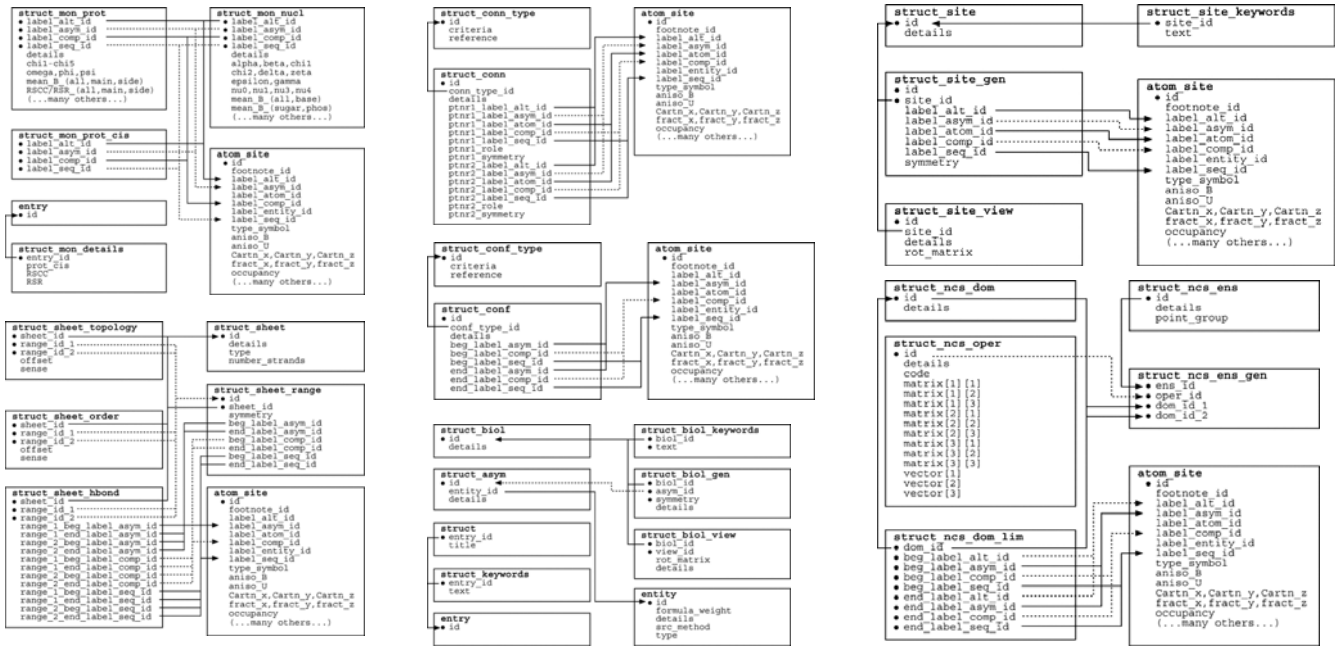
External databases
STRUCT_REF
STRUCT_REF_SEQ
STRUCT_REF_SEQ_DIF

β-sheets
STRUCT_SHEET
STRUCT_SHEET_TOPOLOGY
STRUCT_SHEET_ORDER
STRUCT_SHEET_RANGE
STRUCT_SHEET_HBOND

Molecular sites
STRUCT_SITE_GEN
STRUCT_SITE_KEYWORDS
STRUCT_SITE_VIEW

We are all familiar with the often beautiful cartoon representations of protein structures (the example is BgaR, a lactose sensor [Newman, J., Caron, K., Nebl, T. & Peat, T. S. (2019). *Acta Cryst.* D**75**, 639-646]). There is a considerable challenge translating this into machine-readable database records. The mmCIF dictionary approaches this by describing structural features at various levels of granularity, each of which has its own group of categories, and then relating these categories to the chemical characterisations in the various ENTITY and CHEM_COMP categories, and ultimately to the ATOM_SITE list of individual atomic coordinates.

mmCIF structure – STRUCT relationships



This slide is simply to illustrate the number and complexity of relationships between the different structure-related categories and the other levels of description of the structure model. Full details are given in *International Tables for Crystallography* Volume G 1st ed., Chapter 3.6.