dently for the free and working sets of reflections. The remaining new data items have more specialized roles, which are discussed below.

The data item **_refine.entry_id** has been added to the REFINE category to provide the formal category key required by the DDL2 data model.

Many macromolecular structure refinements now use the statistical cross-validation technique of monitoring a 'free' R factor (Brünger, 1997). $R_{\rm free}$ is calculated the same way as the conventional least-squares R factor, but using a small subset of reflections that are not used in the refinement of the structural model. Thus $R_{\rm free}$ tests how well the model predicts experimental observations that are not themselves used to fit the model.

The mmCIF dictionary provides data names for $R_{\rm free}$ and for the complementary $R_{\rm work}$ values for the 'working' set of reflections, which are the reflections that are used in the refinement. Separate data items are provided for unweighted and weighted versions of each R factor. A fixed percentage of the total number of reflections is usually assigned to the free group, and this percentage can be specified. Further details about the method used for selecting the free reflections can be given using **_reflns.R_free_details**. The estimated error in the $R_{\rm free}$ value may also be given, along with the method used for determining its value.

The purposes of having a set of reflections that are not used in the refinement are to monitor the progress of the refinement and to ensure that the R factor is not being artificially reduced by the introduction of too many parameters. However, as the refinement converges, the working and free R factors both approach stable values. It is common practice, particularly in structures at high resolution, to stop monitoring $R_{\rm free}$ at this point and to include all the reflections in the final rounds of refinement. It is thus worth noting a distinction between **_refine.ls_R_factor_obs** and **_refine.ls_R_factor_R_work**: **_refine.ls_R_factor_obs** relates to a refinement in which all reflections more intense than a specified threshold were used, while **_refine.ls_R_factor_R_work** relates to a refinement in which a subset of the observed reflections were excluded from the refinement and were used to calculate the free R factor. The dictionary allows the use of both values if a free R factor were calculated for most of the refinement, but all of the observed reflections were used in the final rounds of refinement; the protocol for this may be explained in **_refine.details**. When a full history of the refinement is provided using data items in the REFINE_HIST category, it is preferable to specify a change in protocol using data items in this category.

Other data items help to provide an assessment of the quality of the refinement. The scale-independent correlation coefficient between the observed and calculated structure factors may be recorded for the reflections included in the refinement using the data item **_refine.correlation_coeff_Fo_to_Fc**. There is a similar data item for the reflections that were not included in the refinement.

Overall standard uncertainties for positional and displacement parameters can be recorded according to a number of conventions. A maximum-likelihood residual for the positional parameters can be given using **_refine.overall_SU_ML** and the corresponding value for the displacement parameters can be given using **_refine.overall_SU_B**. Diffraction-component precision indexes for the displacement parameters based on the crystallographic R factor (the Cruickshank DPI; Cruickshank, 1999) can be given using **_refine.overall_SU_R_Cruickshank_DPI**. The corresponding value for $R_{\rm free}$ can be given using **_refine.overall_SU_R_free**.

Example 3.6.6.7. *Results of the overall refinement of an HIV-1 protease structure (PDB 5HVP) described using data items in the* REFINE *and* REFINE_FUNCT_MINIMIZED *categories.*

```
_refine.entry_id                          '5HVP'
_refine.ls_number_reflns_obs              12901
_refine.ls_number_restraints              6609
_refine.ls_number_parameters              7032
_refine.ls_R_factor_obs                   0.176
_refine.ls_weighting_scheme               calc
_refine.ls_weighting_details
;  Sigdel model of Konnert-Hendrickson:
   Sigdel: Afsig +  Bfsig*(sin(theta)/lambda-1/6)
   Afsig = 22.0, Bfsig = -150.0 at the beginning
      of refinement.
   Afsig = 15.5, Bfsig =  -50.0 at the end of
      refinement.
;
loop_
  _refine_funct_minimized.type
  _refine_funct_minimized.number_terms
  _refine_funct_minimized.residual
    'sum(W*Delta(Amplitude)^2^'       3009   1621.3
    'sum(W*Delta(Plane+Rigid)^2^'       85     56.68
    'sum(W*Delta(Distance)^2^'        1219    163.59
    'sum(W*Delta(U-tempfactors)^2^'   1192     69.338
```

The quality of a data set used for the refinement of a macromolecular structure is often given not only in terms of the scaling residuals, but also in terms of the data redundancy (the ratio of the number of reflections measured to the number of crystallographically unique reflections). Data items are provided to express the redundancy of all reflections, as well as those that have been marked as 'observed' (*i.e.* exceeding the threshold for inclusion in the refinement). The percentage of the total number of reflections that are considered observed is another metric of the quality of the data set, and a data item is provided for this (**_refine.ls_percent_reflns_obs**).

The limited resolution of many macromolecular data sets makes it inappropriate to refine anisotropic displacement factors for each atom. For these low- to medium-resolution studies, an overall anisotropic displacement model may be refined. The data items **_refine.aniso_B*** are provided for recording the unique elements of the matrix that describes the refined anisotropy.

The two-parameter method for modelling the contribution of the bulk solvent to the scattering proposed by Tronrud is used in several refinement programs. The data items **_refine.solvent_model_*** can be used to record the scale and displacement factors of this model, and any special aspects of its application to the refinement.

The average phasing figure of merit can be given for the working and free reflections. Unusually high or low values of displacement factors or occupancies can be a sign of problems with the refinement, so data items are provided to record the high, low and mean values of each. Further indicators of the quality of the refinement are found in the REFINE_ANALYZE category (Section 3.6.6.2.2).

The data items in the REFINE_FUNCT_MINIMIZED category allow a brief description of the function minimized during refinement to be given (Example 3.6.6.7). It is not possible to reconstruct the functioned minimized during the refinement by automatic parsing of the values of these data items, but the details given in them may still be helpful to someone reading the mmCIF.

3.6.6.2.2. *Analysis of the refined structure*

The data items in this category are as follows:

REFINE_ANALYZE

- **_refine_analyze.entry_id**
  → **_entry.id**

```
_refine_analyze.Luzzati_coordinate_error_free
_refine_analyze.Luzzati_coordinate_error_obs
_refine_analyze.Luzzati_d_res_low_free
_refine_analyze.Luzzati_d_res_low_obs
_refine_analyze.Luzzati_sigma_a_free
_refine_analyze.Luzzati_sigma_a_free_details
_refine_analyze.Luzzati_sigma_a_obs
_refine_analyze.Luzzati_sigma_a_obs_details
_refine_analyze.number_disordered_residues
_refine_analyze.occupancy_sum_hydrogen
_refine_analyze.occupancy_sum_non_hydrogen
_refine_analyze.RG_d_res_high
_refine_analyze.RG_d_res_low
_refine_analyze.RG_free
_refine_analyze.RG_free_work_ratio
_refine_analyze.RG_work
```

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

In small-molecule crystallography, there is general agreement on the metrics that should be used to assess the quality of a structure determination, and data items in the REFINE category of the core CIF dictionary can be used to record them. For macromolecular structure determinations, no such agreement has been achieved yet and new metrics are frequently suggested as the field evolves. The REFINE_ANALYZE category can be used to record the metrics that were in common use at the time that the mmCIF dictionary was constructed; it is anticipated that new metrics will be added in future versions of the dictionary, and that some of the current metrics may fall into disuse.

Luzzati (1952) devised a method for estimating the average positional shift that would be needed in an idealized refinement to reach an R factor of zero by using a plot of R factors against resolution. For some time, macromolecular crystallographers have used a modification of this approach to assess the average positional error. Recent practice has used Luzzati plots based on the free R values to yield a cross-validated error estimate. Data items are provided for recording these coordinate-error estimates and the range of resolution included in the plot (Example 3.6.6.8). Related data names allow the specification of the value of σ_a used in constructing the Luzzati plot.

A general feature of introducing more parameters in the model of the structure is a reduction in the R factor, but the statistical significance of this is often obscured by the simultaneous reduction in the ratio of observations to parameters. Attempts to extend Hamilton's (1965) test to macromolecular structures are usually confounded by the use of restraints. Tickle *et al.* (1998) proposed the use of a Hamilton generalized R factor analyzed separately for reflections in the working set (those used in the refinement) and for reflections in the free set (those set aside for cross validation), and these metrics are often reported in the literature. Data items are provided for recording the Hamilton generalized R factor for the working and free set of reflections, and for the ratio of the two.

Other indicators of a successful refinement involve the relative order of the model. Data items are provided for recording the sum of the occupancies of the hydrogen and non-hydrogen atoms in the model. The number of disordered residues may also be recorded.

3.6.6.2.3. *Restraints and refinement by shells of resolution*

The data items in these categories are as follows:

(*a*) REFINE_LS_RESTR

```
• _refine_ls_restr.type
  _refine_ls_restr.criterion
  _refine_ls_restr.dev_ideal
  _refine_ls_restr.dev_ideal_target
  _refine_ls_restr.number
  _refine_ls_restr.rejects
  _refine_ls_restr.weight
```

Example 3.6.6.8. *Aspects of the refinement of an HIV-1 protease structure (PDB 5HVP) described with data items in the REFINE_ANALYZE category.*

```
loop_
_refine_analyze.entry_id                          '5HVP'
_refine_analyze.Luzzati_coordinate_error_obs       0.32
_refine_analyze.Luzzati_d_res_low_obs              5.0
```

(*b*) REFINE_LS_RESTR_NCS

```
• _refine_ls_restr_ncs.dom_id
        → _struct_ncs_dom.id
  _refine_ls_restr_ncs.ncs_model_details
  _refine_ls_restr_ncs.rms_dev_B_iso
  _refine_ls_restr_ncs.rms_dev_position
  _refine_ls_restr_ncs.weight_B_iso
  _refine_ls_restr_ncs.weight_position
```

(*c*) REFINE_LS_RESTR_TYPE

```
• _refine_ls_restr_type.type
        → _refine_ls_restr.type
  _refine_ls_restr_type.distance_cutoff_high
  _refine_ls_restr_type.distance_cutoff_low
```

(*d*) REFINE_LS_SHELL

```
• _refine_ls_shell.d_res_high
• _refine_ls_shell.d_res_low
  _refine_ls_shell.number_reflns_all
  _refine_ls_shell.number_reflns_obs
  _refine_ls_shell.number_reflns_R_free
  _refine_ls_shell.number_reflns_R_work
  _refine_ls_shell.percent_reflns_obs
  _refine_ls_shell.percent_reflns_R_free
  _refine_ls_shell.R_factor_all
  _refine_ls_shell.R_factor_obs
  _refine_ls_shell.R_factor_R_free
  _refine_ls_shell.R_factor_R_free_error
  _refine_ls_shell.R_factor_R_work
  _refine_ls_shell.redundancy_reflns_all
  _refine_ls_shell.redundancy_reflns_obs
  _refine_ls_shell.wR_factor_all
  _refine_ls_shell.wR_factor_obs
  _refine_ls_shell.wR_factor_R_free
  _refine_ls_shell.wR_factor_R_work
```

(*e*) REFINE_LS_CLASS

```
• _refine_ls_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
```

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.

These categories were introduced in the mmCIF dictionary to allow a detailed description of several aspects of structure refinement to be given. Data items in the REFINE_LS_RESTR category allow geometric restraints to be specified and the deviations of restrained parameters from ideal values in the final model to be given. The type of the geometric restraints can be described in more detail using data items in the REFINE_LS_RESTR_TYPE category. Data items in the REFINE_LS_RESTR_NCS category can be used to give information about any restraints on noncrystallographic symmetry used in the refinement and the category REFINE_LS_SHELL contains data items that allow the results of refinement to be given by shells of resolution.

Data items in the REFINE_LS_RESTR category can be used to record details about the restraints applied to various classes of parameters during least-squares refinement (Example 3.6.6.9). It is clearly useful to tabulate the various classes of restraint, their deviation from ideal target values and the criteria used to reject