

Protein Data Bank Chemical Components Dictionary Content Description

INTRODUCTION

In addition to biopolymers (proteins and nucleic acids), the PDB archive contains more than 10,000 unique non-biopolymer entities which are collectively called *Chemical Components* (Components). They are cataloged in the *Chemical Components Dictionary* (CCD). These components are very diverse in nature and include ions, solvents, natural and modified amino nucleic and acids, ligands such as drugs, cofactors, metal clusters, surfactants, and others. More of this diversity will become obvious in the remainder of this document.

The CCD provides a systematic, standard and common point of reference for the Components. Along with other information, it contains chemical and atom nomenclature, connectivity, bond orders, a representative set of deposited PDB coordinates and idealized coordinates. Each of these unique Components might occur in many PDB entries (an individual PDB ID). Each such occurrence is referred to as an "Instance", and all instances are standardized to the template that CCD provides. This provides consistency between PDB entries. For example, all Instances of ATP will have the same atom and chemical nomenclature. New Components are added as new unique Chemical Components are found in newly deposited PDB entries. However, they are released to the public only when the corresponding PDB entry is released.

The CCD can be searched through on-line web services operated by the wwPDB partners. It can also be downloaded in bulk in several formats. The list of all of the entries that contain a given Component is also available through individual search. For example, all entries containing ATP can be rapidly identified. They can also be downloaded in bulk.

CHEMICAL COMPONENT DICTIONARY CONTENT

The CCD contents include but are not limited to the following:

Component Identifier: Unique alphanumeric 3 letter code (legacy Components may have 2 or 1 letter code)

Chemical and administrative data: Molecular weight, empirical formula, atom and bond counts, formal charge, dates of creation and modification, release status, processing site.

Atom names, bond connectivity and bond order: Atom pairs forming the bond and type of bond, e.g.: single, double, triple, etc.

Internal geometry: Coordinates are provided of both an experimental representative of the instances of this component in one of the PDB entries and also that of an idealized, computed model. This includes atomic Cartesian coordinates and internal geometry such as bond lengths, bond angles, and torsion angles.

Nomenclature and chemical identifiers (both molecular and atomic): SMILES strings, InChI descriptors, IUPAC and common names, synonyms

Other information:

IMPORTANT INFORMATION ABOUT CHEMICAL COMPONENTS

Chemical Components are unique, independent and free-standing, and no Component contains another. For example, hemoglobin (Component code HEM) contains an iron atom which is denoted as its common symbol, Fe, but this does not refer to the Component FE which is specific code for the ferric state of iron, FE(III). In other words, the HEM component stands alone.

A [Variants Dictionary](#) provides protonation states of the 20 essential amino acids and 9 nucleic acids and will be soon extended to other molecules.

AVAILABILITY OF INFORMATION ON THE CHEMICAL COMPONENTS:

The released Components are available for public search and download at wwPDB member sites (PubChem and [Ligand Expo](#)). The mmCIF Components file data fields are described at http://mmcif.pdb.org/dictionaries/mmcif_pdbx.dic/Categories/chem_comp.html.

UPDATES TO THE CHEMICAL COMPONENTS DICTIONARY

The CCD is constantly evolving. Additions and modifications are made as needed, and errors are corrected on an individual basis and during periodic remediations. Questions and observations of errors should be [reported](http://www.rcsb.org/pdb/home/contactUs.do) at <http://www.rcsb.org/pdb/home/contactUs.do>.

GOALS OF THE COMPONENT STANDARDIZATION

The wwPDB staff strives to process ligands in the PDB entries with the intention of where possible to create physically and chemically reasonable neutral stable entities wherein all valences are satisfied, bond orders are properly defined, and, where possible and reasonable, formal charges are reported. This is done to aid search capabilities and to allow assignment of nomenclature.

In many cases, if the Component represents an isolated, “free standing” chemical entity in the protein entries, the Component structure is obvious and the definition is straight-forward. However, the user should be aware that the diverse nature of this collection of non-polymer components means that these goals in fact *cannot* be met in all cases. Sometime the stability of a Component depends on its interactions with a Biopolymer. Consequently, there are special cases whose Component description cannot fit the ideal, which are described below. It should be noted that the depositor of a particular entry might request that a particular “grouping” of atoms be considered a component. The PDB respects their wishes whenever possible.

ADDITIONAL INFORMATION AND SPECIAL CASES

Flags and tokens

When possible, additional information about the Chemical Components and special cases are noted within the data files via [flags](#) (mmCIF format) or REMARK statements (for particular Instances, PDB format).

Ambiguous flag

The ambiguous flag, `_chem_comp.pdbx_ambiguous_flag`, indicates that some aspect of the component could not be completely defined. For example, atomic valence might not be satisfied or formal charge might not be defined.

The major complication for making systematic programmatic use of the PDB ligand set is where valence is not satisfied. In these cases the flag `_chem_comp.pdbx_ambiguous_flag` is set to "Y" in the Dictionary. In these cases the SMILES and InChI codes are incomplete, there are unlikely to be idealized coordinates and the systematic names will not be correct.

Obsolete or Replaced Component token

Some components are no longer used and have become obsolete based on later consideration. These can be identified via the `_chem_comp.pdbx_release_status` token which will have the value OBS, see [CRY](#) for example. In case Component is superseded by another definition, the mmCIF token `_chem_comp.pdbx_replaced_by` will give the code of the corrected, current definition. The superseding Component mmCIF will, in its turn, contain the code of the replaced Component in token `_chem_comp.pdbx_replaces`. See [GOL](#) for example.

Leaving atom flag

Leaving atoms are added to complete the Component when otherwise it would be a fragment, such as a Component that is covalently bound to another Component or a Biopolymer. Normally these Components represent the original molecules used for crystallization or molecule synthesis. In case these are not known, the leaving atoms are added making every attempt to make them of compatible chemistry, often duplicating the atom making the covalent bond. Each leaving atom has the flag `_chem_comp_atom.pdbx_leaving_atom_flag`. The best examples of Components with leaving atoms are the standard amino acids. The OXT, HXT and H2 atoms are lost as a result of peptide bond formation. See [SER](#) for example.

DESCRIPTION AND EXAMPLES OF SPECIAL CASES

Charge

The PDB policy is that the total charge of a Component is assigned only when it is known and is adjusted to 0 charge, if possible. For ionizable species this is done even if this is not in agreement with any particular pKa/pH considerations. For example GLU is listed as having uncharged amino and carboxylic groups.

Substituents and fragments

Some Components do not achieve the goals described above, but are used due to prior convention or for convenience of search and representation. Examples of such not free-standing molecules include [NH2](#) (a terminal NH₂ group, used to represent amidation), [OQE](#) (chloromethyl group, used to represent chloromethylketone inhibitor peptides), [CF0](#) (fluoromethyl group, used to represent fluoromethylketone inhibitor peptides), etc. Also, there are a few cases of single atoms which are not normally free-standing, but used as individual Components such as isolated [O](#), [SE](#), and [ARS](#) atoms.

Components covalently bound to biopolymers resulting in bond order change or ring opening

For Components covalently bound to a Biopolymer or another Component, the parent (pre-bound) compound will be normally used as the Component with the corresponding atom or atom group flagged as leaving. However, if the bond order changes as a result of the reaction, the final product will be built as a Component. Similarly, when a ring opening occurs, the final product is built as the Component. See [167](#) for example.

Valence tautomerism

Tautomers are not enumerated and only that one most relevant is built as a Component, although different tautomers of the same molecule might be present as different Chemical Components.

Dependent (non-free standing) components

Many metal-containing cofactors and prosthetic groups have the metal coordination completed by either amino acid, nucleic acid, or other Chemical Components. These components are clearly distinct chemically from the biopolymer but difficult to represent as free-standing moieties. Examples include clusters such as the Sirohaem-Fe₄S₄ enzymes, the molybdopterin-containing proteins, and the specific components [F3S](#) (Fe₃S₄) or [CLF](#) (Fe₈S₇) which are stabilized by interactions between the iron atoms and cysteine sulfhydryl groups. These components will not have valence fulfilled.

Racemates

The stereochemistry of a chemical component is absolute and is that seen in the crystal structure as reported by the depositor. Different stereoisomers of the same molecule are considered as different Components, such as [ALA](#) (L-alanine) and [DAL](#) (D-alanine).