This dataset has been obtained from the protein data bank (<https://www.rcsb.org/>) where you can search for collate data on a selection or all structures within the database. The first dataset contains selected information from all structures within the database and primarily summarises the data quality and some basic information about each structure. More data can be added. Gaining insight into this can allow us to gain an idea of how data has changed.

**File;** datacollection\_keep\_data.csv

**Glossary of terms found in the proteins data set.**

**Entry ID**

This is the unique code assigned to a structure, sometimes called the PDB id. This is searchable at the Protein Data Bank (https://www.rcsb.org/) and can be used to link the structure to other data sources.

**Experimental Method**

This is the method to obtain the data that resulted in the structure.

**Resolution (Å)**

This can be used to indicate the quality of the data (from X-ray and cryo-EM). The smaller the value the “higher” or better the Resolution. The higher the resolution the atoms in the structure are more defined. It is given in angstroms (Å) equal to 0.1 nm.

**PubMed ID**

In cases where the structure has formed part of a publication this is an integer code that links to the publication. It is searchable at <https://pubmed.ncbi.nlm.nih.gov/>

**Release Date**

This is the date on which the structure was made publicly available.

**Number of Non-Hydrogen atoms per deposited model**

This is the total number of atoms found in the structure. Hydrogens are rarely seen/added and so are not counted here.

**Total number of Polymer Instances (Chains)**

A protein or single strand of DNA or RNA is a polymer. These are classed as chains. Therefore, this is the total number of chains found in the structure.

**Total number of Polymer Residues per Deposited Model.**

Proteins, DNA, RNA polymers are made up of Amino acids and Nucleotide bases. This is the total count of Amino acids and/or Nucleotide bases that form part of the polymers in the structure.

**Number of Water molecules per Deposited Model**

Biological structures interact with their surroundings, including water. This is the total number of water molecules found in the structure.

**Disulfide Bond Count per Deposited Model**

Proteins that contain the amino acid cysteine may contain disulfide bonds. These form between two cysteine side chains. This is the total number of disulfides found in the structure.

**Number of Distinct Molecular Entities**

A protein chain, DNA chain, RNA chain or other molecule or atom (such as a metal) is a distinct molecular entity. This is the total count of these, it does not include waters.

**Molecular Weight per Deposited model**

This is the total molecular weight of all atoms in the structure. It is given in kilodaltons (kDa).

**Number of Distinct Non-polymer Entities**

Any molecule or atom (not including water) that does not form part of a polymer.

**Entry Polymer composition**

The types of polymer molecules that make up the structure. Is it Protein, DNA, RNA, oligosaccharide or some form of mix.

**Number of Distinct DNA Entities**

The number of distinct (unique) DNA polymers found in the structure.

**Number of Distinct NA Hybrid Entities**

The number of distinct (unique) nucleic acid hybrid polymers found in the structure. ie. RNA/DNA hybrid polymer.

**Number of Distinct Protein Entities**

The number of distinct (unique) protein polymers found in the structure.

**Entry Polymer Types**

What types of polymer make up the structure.

**Journal Name (Abbrev)**

If the structure formed part of a publication the abbreviated name of the journal in which it was published.

**R Free**

A statistical value to assess the fit of structures solved using X-Ray crystallography to the original data.

**R Work**

Related to R Free a further method to asses the fit of structures solved using X-Ray crystallography

**File;** combined\_metals\_keep\_data.csv

**PDB code**

This is the unique code assigned to a structure, called Entry ID in the previous data set.

**Metal**

The ion type found in the structure.

**Chains**

The chain that the ion interacts with. A alone means the ion interacts with only chain A. B\_A means the ion interacts with chains B and A, ie. It is in an interface.

**Number of hydrogen bonds**

Total number of direct bonds (such as coordination bonds) the ion forms to the polymer structure. This is called hydrogen bonds because of the nomenclature used by pdbsum.

**Number of non-bonded contacts**

Total number of non-bonded contacts, these tend to be longer than hydrogen bonds.

This is called non-bonded contacts because of the nomenclature used by pdbsum.

**Average Distance HB**

Average distance of closer (coordination bonds).

**Std Distance HB**

Standard deviation for the average distance.

**Max Distance HB**

Maximum distance classed as a hydrogen bond.

**Min Distance HB**

Minimum distance classed as a hydrogen bond.

**Average Distance NB**

As for HB but for longer nonbonded contacts

**Std Distance NB**

As for HB but for longer nonbonded contacts

**Max Distance NB**

As for HB but for longer nonbonded contacts

**Min Distance NB**

As for HB but for longer nonbonded contacts

**Unique ALA count, etc**

This is the number of unique alanines found interacting with the ion. For example if an ion interacts with Ala1, Ala2 and His3 (the number being the position in the polymer chain) this would give 2 in the Ala column and 1 in the his column. It does not count the number of interactions to a single amino acid, which may be more than 1.

This is given for each of the 20 amino acids.