**The crystal structure dataset**

**Data preprocessing**

Starting from the original crystal antibody data, we only kept the antibodies that have:

* Complete alignment information for heavy *and* light chains
* Successful structure predictions using ABodyBuilder *and* IgFold for heavy and light chains separately *and* as paired antibodies.

**Antibody count**

**859** unique antibodies (PDBs)

**Enclosing directory**

The main “[crystal\_dataset](https://drive.google.com/drive/folders/12j9N74vVpqtoVv4uAc9xDIcKplqXtKlx)” directory contains:

1. The file “**aligned\_crystal\_dataset**.csv” contains **859** unique antibodies with complete (heavy+light) sequence information + germline and region definition (FRs and CDRs). Thus, the “PDB” field is repeated twice, however the “identifier” field is unique for each row [1718 rows].
2. The file “**AbChain\_crystal\_dataset**.csv” contains the same information as the file in (1) as well as the sequence-based DPs measured for the unpaired chains [1718 rows].

*Notes:*

* The two immunogenicity parameters (AbChain\_immunopeptide\_start\_region and AbChain\_immunopeptide\_end\_region) are qualitative/descriptive measures. Better to be avoided in quantitative (correlation) studies.

1. The two sub directories (AbStruc\_paired) and (AbStruc\_unpaired) that contain the structure-based DP calculations as measured on for the paired and unpaired antibodies of this dataset, respectively.

*Notes:*

* The DP “AbStruc\_clashes” was removed because it mostly 0 for this dataset.
* The DP “AbStruc\_pi\_helices” is included to be treated as a control/comparison DP (with the native repertoire), but not to be used in correlation studies due to limited variability in its values.
* The DP “AbStruc\_weak\_hbonds” shows variability (in contrast to the native repertoire.