**清洗ASD-pro-lig数据：**

**预处理**

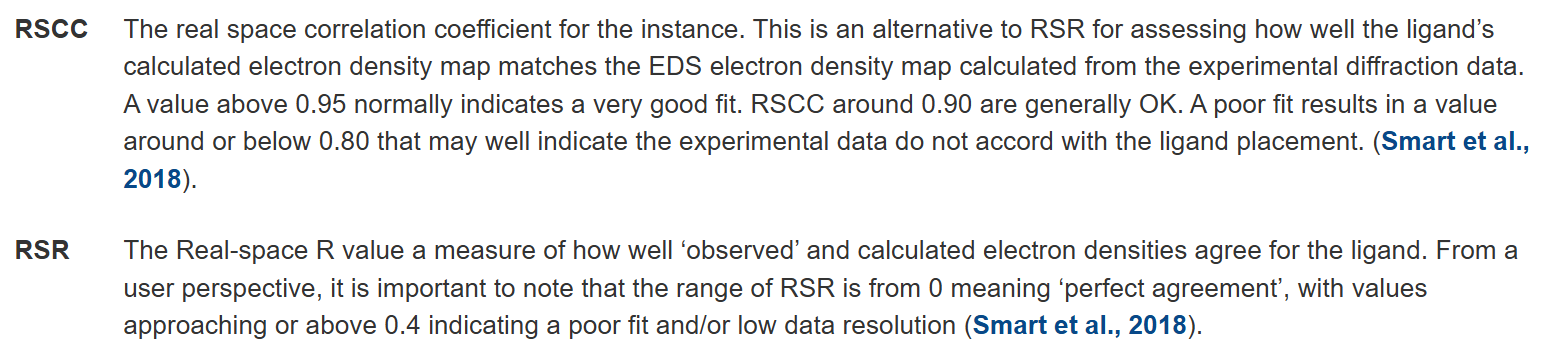
1. 原始数据source from xinyi & qiuhan **(3,117+96=3,213)**
2. 排除共价+非常规小分子**（2,630+96=2,726）**
3. 目视核查并清除多配体互作或螯合情况**（2,661）**
4. 清除重复记录**（2,557）**
5. QBioLip提示ligand name记录错误或是共结晶分子**（2,516）**
6. 因为QBio没有ligand resi的信息，使用BioLip cross-maping to correct recording error and unique **(2,381)**

**从这里开始，我们有了初步预处理好的数据集。**都是dirty action，只有共价+非常规小分子和多配体螯合的情况是需要在正文提及的

汇报的话，**3,213** (原始数据) **→ 2,557** (共价+非常规) **→ 2,516** (共结晶分子) **→ 2,381** (记录错误 & 删去重复)

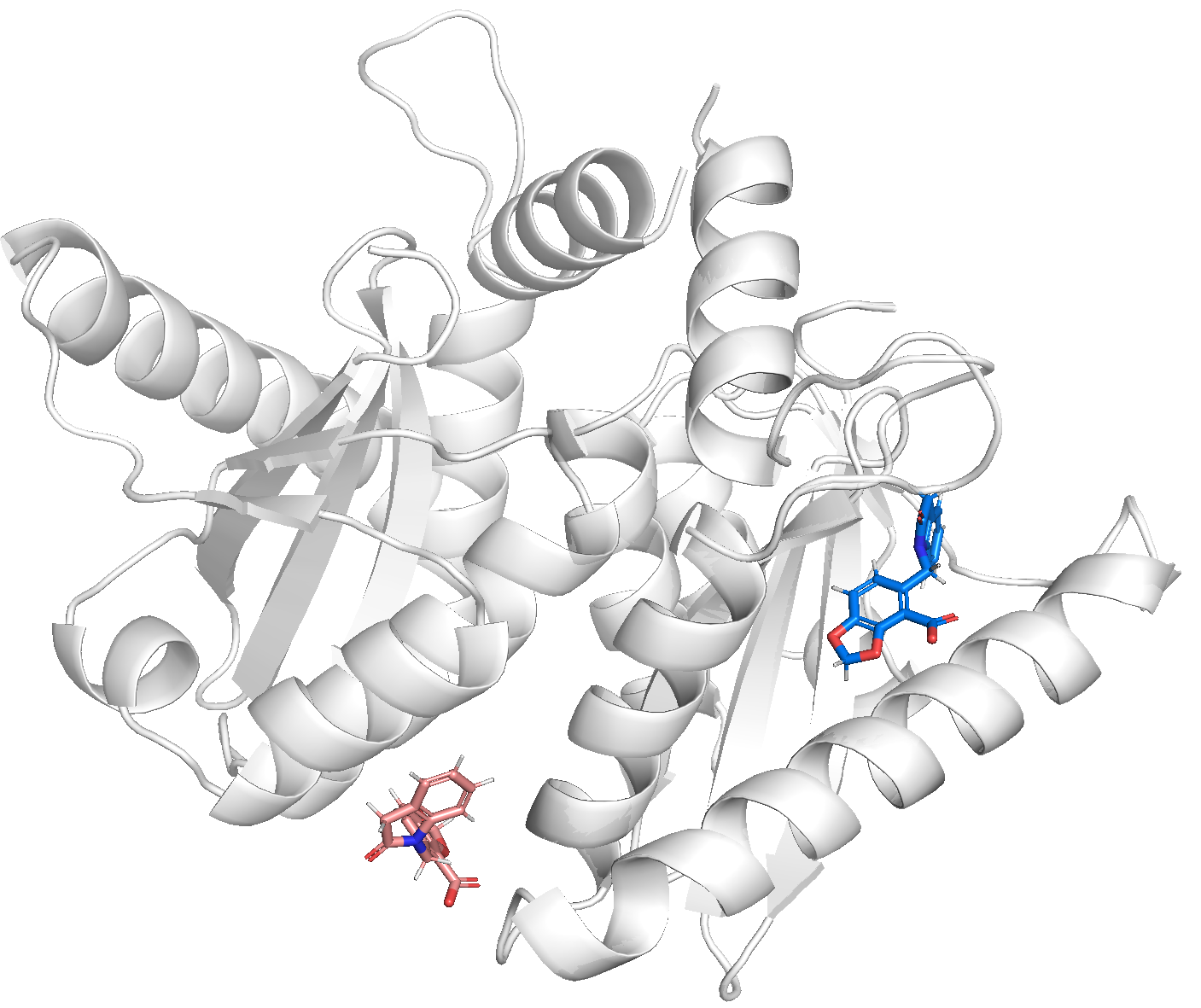
**正式处理**

1. Resolution (3Å, 2,173; **2.5Å, 1,666**; 2Å, 855;)
2. Ligand quality (rsr <= 0.2 & rsrc >=0.95: 607; rsr <= 0.4 & rsrc >=0.8: 1,317; 607; **rsr <= 0.4 & rsrc >=0.9: 1,030**)



1. PDBBind-opt (Ligands containing only H, C, O, N, P, S, F, Cl atoms & Ligands that are not covalently bound to protein & Steric clash 2Å & other HETATM not in hetatm\_cutoff 8 Å: **852**)
   1. 7 covalent
   2. 3 steric clashes
   3. 6 rare elements
   4. 157 hetatm\_cutoff 8 Å
2. Ligands properties: (**840**)
   1. < 6 heavy atoms (4)
   2. 100 < nm< 900 (8)
3. sequence similarity by CD-hit (0.9, 228 class)
4. ligand fp similarity by rdkit (0.9, **799 + 1**, 见下图)
5. different from pdbbind timesplit
   1. AlloBind pdbids: 700
   2. AlloBind in train: 243
   3. AlloBind in val: 18
   4. AlloBind in test: 0
   5. AlloBind unique: 439

3nf6\_IMV\_A\_230 4chp\_IMV\_A\_1216



**From PDBBind Refine Set**

We anticipated that our PDBbind database will be of major interest to researchers working on structure-based drug design, particularly in the docking/scoring field. Not every 1622 PLEXBAS identified through the previous steps, however, **is the right material for docking/scoring studies**. Consequently, we have applied a number of criteria to filter out the unqualified entries from them, and the remaining entries formed a “refined set”. The following criteria were applied to the selection of the refined set:

(1) Only crystal structures with an overall resolution better than or equal to 2.5 Å were accepted into the refined set. This cutoff of 2.5 Å was chosen to select the structures with relatively high resolution. Structures resolved by NMR techniques were neglected in our work.

(2) Only noncovalently bound complexes were accepted into the refined set. Our program detected any covalent connection between the protein and the ligand by analyzing the “LINK” and “CONECT” records in the given PDB file. It further checked the distance between the protein and the ligand using the available coordinate information. If the distance between any two non-hydrogen atoms on the protein and the ligand was found to be shorter than 2.0 Å, this complex was rejected, even if it was not identified as covalently bound.

(3) Only binary complexes were accepted into the refined set. In other words, the complex must be a structure formed unambiguously between one protein molecule and one ligand molecule. In some cases, two ligand molecules bind together inside the same binding site on the protein molecule to form a ternary complex. In such cases, experimental measurement of the binding affinity of each individual ligand molecule becomes much more complicated, since the existence of the other ligand molecule must be taken into account, and none of the known docking/scoring tools is really capable of handling ternary protein−ligand complexes. Thus, we decided not to include them in the refined set. If there were more than one ligand molecule in the given complex structure, our program would check the distance between them. **If any non-hydrogen atom on one ligand molecule was within 8.0Å of any non-hydrogen atom on another ligand molecule**, the given complex was considered to be a ternary complex and, accordingly, was rejected.

(4) Only complexes with known Kd or Ki values were accepted into the refined set. Kd and Ki are equilibrium constants, which are in principle associated with the thermodynamic properties of a protein−ligand binding process. They are also independent of the concentration levels of the protein and the ligand used in binding assay and thus can be considered as a measurement of absolute binding affinities. In contrast, IC50 values depend on the design of the binding assay and reflect only the relative binding affinities of the inhibitor molecules evaluated in the same binding assay. Therefore, we decided not to mix IC50 values with Kd and Ki values.

(5) Some specific restrictions on the ligand and the protein were also applied. To be included in the refined set, the ligand molecule in the given complex must consist of only common organic elements, i.e., **C, N, O, P, S, F, Cl, Br, I, and H**. Otherwise, the complex was rejected. This is a practical concern, since not all of today's molecular modeling software has the necessary parameters to handle organic molecules with uncommon elements, such as Be, B, Si, or metals. Similarly, if the protein molecule in the given complex had nonstandard amino acid residues close (i.e. within 8.0 Å) to where the ligand molecule was bound, the complex was rejected. In addition, to control the size of the ligand molecules, the **molecular weight of the ligand molecule was restricted to 1000 or less.**

