Findsitemetal notes:

* First had to run LOMETS to get template lists for every protein
* Then ran findsitemetal to predict metal binding sites based on structures and template lists
* Results
  + Don’t need data from “\*.findsitemetal.alignments.dat” right now
  + “\*.findsitemetal.templates.pdb” all-atom coordinates of template structures aligned and superimposed on the target protein structure (same coordinate system as target)
  + “\*.findsitemetal.metals.pdb” has Cartesian coordinates of metal atoms from template structures aligned and superimposed on target protein structure
  + “\*.findsitemetal.sites.pdb” has Cartesian coordinates of predicted metal atoms in their binding sites, but since I’m taking my own approach to this based on centroid of metal binding residues, I don’t need this file
  + “\*.findsitemetal.sites.dat” is probably the only file really needed. From here we can parse out:
    - The number of the binding site: “SITE 1 7 1.0000 2.2058 0.6435 0.2115 4” 🡨 second column
    - The number of metal binding residues in the site: “SITE 1 7 1.0000 2.2058 0.6435 0.2115 4” 🡨 last column
    - Which metal(s) bind to each site: “METAL MG 1.000000 0.764762” 🡨 third column equal to 1 means the site does bind this metal, otherwise 0
    - Which residues make up the binding site: “RESIDUE 25 T \* 7 4.205 0.35060 0.74038 0.85714 0.63671 0.05634” 🡨 if asterisk follows third column, it is a binding residue, else not
    - Residue position(s) making up the binding site: “RESIDUE 25 T \* 7 4.205 0.35060 0.74038 0.85714 0.63671 0.05634” 🡨 second column
    - Amino acid type making up the binding site: “RESIDUE 25 T \* 7 4.205 0.35060 0.74038 0.85714 0.63671 0.05634” 🡨 third column
    - Parsed result columns look like: col1 = site#, col2 = METAL1,METAL2…, col3 = AA#,AA#...
* What features to extract
  + Total number of metal-specific binding sites on the protein (this should have a column for each type of metal, and their sum as well, and will consist of integers)
  + Local metal ion contacts (should have a column for each type of metal, their sum, and will be integers)
    - Approximate the locations of bound metal atoms as the centroid of all atoms (or if possible, just sidechain atoms) of all metal binding residues for each site; it is not known which specific amino acid atoms metals may form bonds with because it is difficult to know the partial charges on the protein, protonation states, etc., so an approximation is probably appropriate; this method leads to very slightly different centroids than those output by findsitemetal itself but is also extensible to metal binding sites from UniProt
    - Use Bubble method (5-angstrom radius) and count the number of metal atoms for each type of metal within this radius (distance from centroid to potential CS atom < [5 angstroms + atomic radius of metal]) for a given RKPT residue
    - Metal atomic radii (Chimera default ionic radii, coordination number = 6; from R. D. Shannon (1976). "Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides". Acta Crystallogr A. 32: 751–767.):
      * CA2+ = 1.00 angstroms
      * CO2+ = 0.65 angstroms
      * CU2+ = 0.73 angstroms
      * CU1+ = 0.77 angstroms
      * FE2+ = 0.61 angstroms
      * K1+ = 1.38 angstroms
      * MG2+ = 0.72 angstroms
      * MN2+ = 0.83 angstroms
      * MO4+ = 0.65 angstroms
      * NA1+ = 1.02 angstroms
      * NI2+ = 0.69 angstroms
      * ZN2+ = 0.74 angstroms

Decide whether to say findsitemetal “FE” can stand-in for UniProt “Iron sulfur” clusters

* “TP” = 468 = Count number of overlapping residues that findsitemetal predicts bind FE and UniProt says bind an iron-sulfur cluster
* “FN” = 249 = Count number of residues that UniProt says bind an iron-sulfur cluster but findsite metal does not predict bind FE
* If TP >> FN, then when merging UniProt and findsitemetal sites, treat iron-sulfur clusters like they are just free FE, otherwise, these can’t be merged and probably should toss out iron-sulfur cluster data from UniProt; TP >> FN can be evaluated by seeing if FE-ironSulfur(TP/FN) >= FE-FE(TP/FN)
  + FE-FE(TP) = 76
  + FE-FE(FN) = 32+41 = 73
* FE-ironSulfur(TP/FN) = 468/249 = 1.88
* FE-FE(TP/FN) = 76/73 = 1.04
* This result suggests that if FE binding prediction by findsitemetal is considered to be predictive of iron-sulfur cluster binding sites, it performs substantially better at predicting FE binding overall, but since findsitemetal does not provide for the orientation of iron-sulfur clusters, to be conservative, I’ll treat those binding sites as binding free FE. This means that I’ll use the ionic radius of free FE2+ instead of the substantially larger size of the iron-sulfur cluster, and the position of the FE should be approximately the centroid of the position of the iron-sulfur cluster. RKPT will have to be closer to these FEs in order to count towards their Bubble feature for proximal metal binding than they would if the binding position of the iron-sulfur cluster was known.

Merging UniProt metal binding site data with findsitemetal predictions

* If sites between UniProt and findsitemetal overlap at least partially in both residues and type of metal, merge them into one site by taking the union of residues and metals for both sites
* If sites between UniProt and findsitemetal overlap only in partially in residues and do not overlap at all in metals, consider the two as distinct sites that bind different metals; very few (76) such cases in my protein structures
* If sites between UniProt and findsitemetal overlap only in metals and not at all in residues, consider the two as distinct sites that bind the same type of metal but not the same actual ion simultaneously
* After merging in this fashion, make sure to include all findsitemetal results for proteins not annotated with metal binding sites in UniProt and all UniProt binding sites that either don’t overlap with any findsitemetal results or for proteins with no predicted findsitemetal results
* De-duplicating
  + If there are entirely identical final binding sites because of multiple findsitemetal sites that use the same binding residues, remove the higher-numbered site (or one with more metals associated; or combine metals for otherwise identical sites)
  + If there are entirely identical final binding sites because of merging 2 or more uniprot sites with findsitemetal sites, de-merge them and keep only the uniprot sites as they were in uniprot; this is especially important because not doing so will reduce the count of real metal binding sites to be used in the Bubble method