**Membrane topological features to be annotated:**

* Are these files for the asymmetric unit or the biological unit?
  + Probably asymmetric unit, per PDB conventions
* Information about the membrane itself? This seems difficult to ascertain
* Accession number
* XYZ coordinates from PDB? Transformation (like OPM/PDBTM)?
* Chains (if multiple)
* N-terminus: cytoplasmic or extracellular
* C-terminus: cytoplasmic or extracellular
* Transmembrane spanning regions (maybe something like spanfiles for MPDock?)
  + Hydrophobicity information
  + Core hydrophobic regions
  + Full transmembrane domains
* Structure motifs (alpha-helix, beta-barrel, etc.)
* Features that aren’t loops – are there standard definitions for these?
  + Head
  + Neck
  + Stalk
  + Stem
* Functional domains and motifs
  + Ligand binding
  + Protein-protein interactions
  + Signal sequences
  + Specific functions (e.g. S4 voltage sensor in voltage-gated ion channels)
  + Check these against a database? Pfam or Interpro, etc?
    - Pfam may have some limitations –didn’t work for my example protein (labeled all of the residues as the same domain)
    - Interpro seems a bit better
    - GO terms?
* Post-translational modification (might not be necessary, a lot of these aren’t relevant to amino acids in the membrane):
  + Histone marks (acetylation, methylation, etc.)
  + Phosphorylation sites (Ser, Thr, Tyr, His)
  + Sugars: N-glycosylation (Asn-X-Ser/Thr), O-glycosylation (Ser, Thr)
  + Lipidation: Myristoylation, palmitoylation, etc.
  + Addition of peptides: SUMOylation, ubiquitination, etc.
  + Amino acid modification: Citrullination, deamidation, etc.
  + Nitrosylation (Cys)
  + Disulfide bridges?
* Lipid rafts/caveolin interaction?
* Hydrophobic thickness/depth
* Tilt angle
* Delta G transfer (kcal/mol)
* Resolution (if applicable)

**How exactly will this filetype relate to PDBs?**

* OPM, PDBTM directly modify PDB files
* How much information will these files have to supplement the PDB?
  + At least transmembrane spans, domains, and motifs?
  + Could follow something like an OCTOPUS file or a FASTQ file
    - Have primary structure with specific information
    - Somewhere in between the complexity of a FASTA and a PDB file
* How will domain/motif information be standardized?
  + Pulled from an existing database would be easiest, but has limitations
    - Pfam labeled my entire protein sequence as a single domain in my SARS-CoV-2 example
    - Interpro properly showed extracellular, transmembrane, and cytoplasmic domains
  + Some set of internally consistent keywords or something similar?
* How will MFTA files be compared?
  + Multiple sequence alignment equivalence?
  + This is probably separate from the database, as it will require computation
* How human-readable should files be?
  + PDBs are somewhat daunting
  + FASTA are easy but lacking in depth
  + FASTQ have more information but aren’t very human readable
* How will these files be generated?
  + Direct relationship to how human readable/accessible they are

**MFTA mockfile annotation:**

I wasn’t sure exactly how complicated this should be, so I took a simple approach at first that can be added to later. I think we should have some discussion moving forward about what all should be included and how to best represent it.

* **Metadata** at the top (name, accession numbers, etc.)
* **Seqnum**: amino acid residue number
* **Chain**: each chain of a PDB file is separated here as well
* **Res**: amino acid residue, single letter format
* **Topo**: is the residue located outside (O), in the membrane (M), or inside (I)?
  + This could be extended further into smaller bins (head/stalk/neck/etc.)
* **Domain**: domain or motif information
  + I mocked this with some shorthand, it would need to be standardized
  + Pfam or InterPro seem like natural places to start, but I’m open to other ideas