- Search UniProtKB for all keywords associated with toxins
    - KW-0800 (toxins in general)
    - KW-0008 (acetylcholine receptor inhibitor)
    - KW-1204 (blood coagulation cascade activating toxin)
    - KW-1203 (blood coagulation cascade inhibiting toxin)
    - KW-1222 (bradykinin receptor impairing toxin)
    - KW-0108 (calcium channel impairing toxin)
    - KW-1221 (calcium activated potassium channel impairing toxin)
    - KW-0123 (cardiotoxin)
    - KW-1217 (cell adhesion impairing toxin)
    - KW-1265 (chloride channel impairing toxin)
    - KW-1216 (complement system impairing toxin)
    - KW-0204 (cytolysin, exotoxin)
    - KW-1061 (dermonecrotic toxin)
    - KW-0260 (enterotoxin)
    - KW-1206 (fibrinogenolytic toxin)
    - KW-1205 (fibrinolytic toxin)
    - KW-1214 (g protein coupled acetylcholine receptor impairing toxin)
    - KW-1213 (g protein coupled receptor impairing toxin)
    - KW-1200 (hemorrhagic toxin)
    - KW-1199 (hemostasis impairing toxin)
    - KW-0872 (ion channel impairing toxin)
    - KW-1028 (ionotropic glutamate receptor inhibitor)
    - KW-0959 (myotoxin)
    - KW-0528 (neurotoxin)
    - KW-1202 (platelet aggregation activating toxin)
    - KW-1201 (platelet aggregation inhibiting toxin)
    - KW-0629 (postsynaptic neurotoxin)
    - KW-0631 (potassium channel impairing toxin)
- Split dataset into manually reviewed (SwissProt) vs unreviewed (TREMBL)
    - In manually reviewed, add data from other manually reviewed datasets
        - Add all toxic protein sequences from conoserver for conch peptides
        - Add all toxic protein sequences from arachnoserver for spider peptides
        - Add all toxic protein sequences from DBETH for bacterial toxins
        - Add all toxic protein sequences from ATDB for animal toxins
- Keep only sequences <50 AA in both manually reviewed and unreviewed
- Keep only unique sequences in both manually reviewed and unreviewed