A diagram of a diagram

Description automatically generated with medium confidence

**Supplementary Figure 1: The BulkProt pipeline.** A CSV-format file of UniProt queries is used as input to the BulkProt pipeline. Each search is queried against UniProt via the API, and a seed table is generated per search term, containing the entries associated with that term. For each seed table, the protein and gene names are extracted and used to construct a second query, which is again queried against UniProt via the API. The resulting main search table will contain all of the proteins and genes associated with the initial seed table search term, but will also contain irrelevant entries as a result of the unsupervised query construction. To remove these, entries are dropped if they do not have a gene name that is present in the initial seed table. Dropped and filtered entries are written to separate tables. The final output of BulkProt is the seed table, the main table, the filtered table, and the dropped table.