# Probe Ligand Selection Algorithm

The probe ligand selection algorithm is designed to efficiently identify a minimal set of ligands (probes) that collectively maximize the coverage of true binding pocket residues across a diverse set of protein targets. The process operates in a batch-wise manner, iteratively selecting ligands that contribute the most new coverage of pocket residues for each batch of proteins. For each batch, the algorithm first determines the set of true pocket residues for the proteins in the batch. It then initializes the current coverage using predictions from the already selected probe set. If the current coverage already exceeds a predefined threshold (e.g., 80%), the batch is skipped. Otherwise, the algorithm evaluates each candidate ligand by predicting its pocket coverage across the batch and quantifies how many new, previously uncovered, true pocket residues it identifies. Ligands that contribute new coverage are added to the probe set, and the process continues until the coverage threshold is met or all ligands are exhausted. This greedy, coverage-driven approach ensures that the selected probe set is both efficient and effective in representing the diversity of binding pockets in the dataset.

## Algorithm: Probe Ligand Selection

Input: batch\_size, max\_batches  
Output: probe\_set, coverage\_history  
  
1. Initialize probe\_set ← ∅, coverage\_history ← []  
2. Shuffle all proteins and partition into batches of size batch\_size  
3. For each batch (up to max\_batches):  
 a. Get true pocket residues for proteins in batch  
 b. Initialize current\_predictions using probe\_set (if not empty)  
 c. If current coverage ≥ threshold (e.g., 80%), record and skip batch  
 d. For each ligand associated with batch proteins:  
 i. Predict pocket residues for all proteins in batch  
 ii. Calculate number of new true pocket residues covered  
 iii. If new residues > 0:  
 - Add ligand to probe\_set  
 - Update current\_predictions  
 - Update coverage  
 - If coverage ≥ threshold, break  
 e. Record batch statistics in coverage\_history  
 f. If consecutive skipped batches ≥ 3, terminate  
4. Save probe\_set to database  
5. Return probe\_set, coverage\_history