# Sequence-Structure-Function Relationships In The Microbial Protein Universe

- **The traditional paradigm**: similar sequences = similar structures = similar function
  - This paper shows that similar function can be achieved with different sequences and structures as well as to explore neglected parts of the "protein universe"
- **Definitions**:
  - **UMAP**: Algorithm for projecting multi-dimensional data into a 2d graph
  - **Structural Motifs**: specific arrangement of secondary structures (like alpha helices and beta sheets) that recur in various proteins and are associated with particular functions
  - **TM Score**: Metric used to assess the structural similarity between two protein structures.
- For this paper, they created a gene catalog using the following steps:
  1. Extracted sequences not found in other databases (since they are looking for novel folds)
  2. Align these extracted sequences to identify similarities and differences.
  3. Of those with deep enough alignments, they used **Rosetta** and **DMPFold** to predict the numerous 3d structures of those proteins.
     - **Rosetta** had fewer "coil" residues compared to those made with **DMPFold**
     - **DMPFold** models were higher quality for larger proteins
  4. Using these predicted structures, they then curated the dataset to keep only the highest quality predictions; discarding the 25% lowest-quality ones
     - Too many coil resides were considered low quality (filter out)
     - If the models produced from both **Rosetta** and **DMPFold** were similar, they were considered to be high quality
     - **AlphaFold2** was used to double check the predicted structures
  5. Finally, they took this curated dataset of predicted structures and ran it through **DeepFRI**; a model for predicting functions of those sequences
- **Takeaways**:
  - Identified 148 novel folds
  - Protein structure evolves very gradually, slowly resulting in new folds. Fold space is continuous rather than discrete
  - *Confirms* existing knowledge: (on average) Similar sequences = similar structures = similar function

- However, there are *some* examples where this is not true where dissimilar sequences lead to similar structures and intern similar functions and vice versa