Improved global protein homolog detection with major gains in function identification

- Although we recently now have access to massive amounts of protein sequences, we still struggle to understand the relationships between them
    - Existing models struggle to detect homologs when the sequence identity is low and are computationally expensive
    - Existing models also struggle with homolog detection when protein evolution increases (the structure evolves rapidly)
    - This paper attempts to solves these using a LLM (PRotein Ortholog Search Tool aka "PROST")
- **PROST**:
    - Applies IDCT to embeddings from the ESM1-b model
    - This is done to compress the embeddings to retain only the information essential for homolog detection
    - ESM1-b embeddings are in a 34 x N x 1280 matrix
        - Of the 34 output laters, each layer has 20 **attention heads** that learn different relevance of the input sequence
        - What each attention head learns is unknown
        - *To solve this and determine the most relevant layers, we compress each layer with 2d-iDCT and then test that layer's accuracy at predicting a sequence*?
            - **QUESTION**: Are we figuring out layer or attention head accuracy? Attention heads are part of a layer?

Vocab:

- **Homolog**: protein sequences that form similar structures
- **Twilight-Zone Proteins**: Proteins with low sequence identities (25-30%)
- **Quantization**: Method of reducing high-dimensional data to a low dimensional representation
- **Inverse Direct Cosine Transform (iDCT)**: Algorithm for compressing data