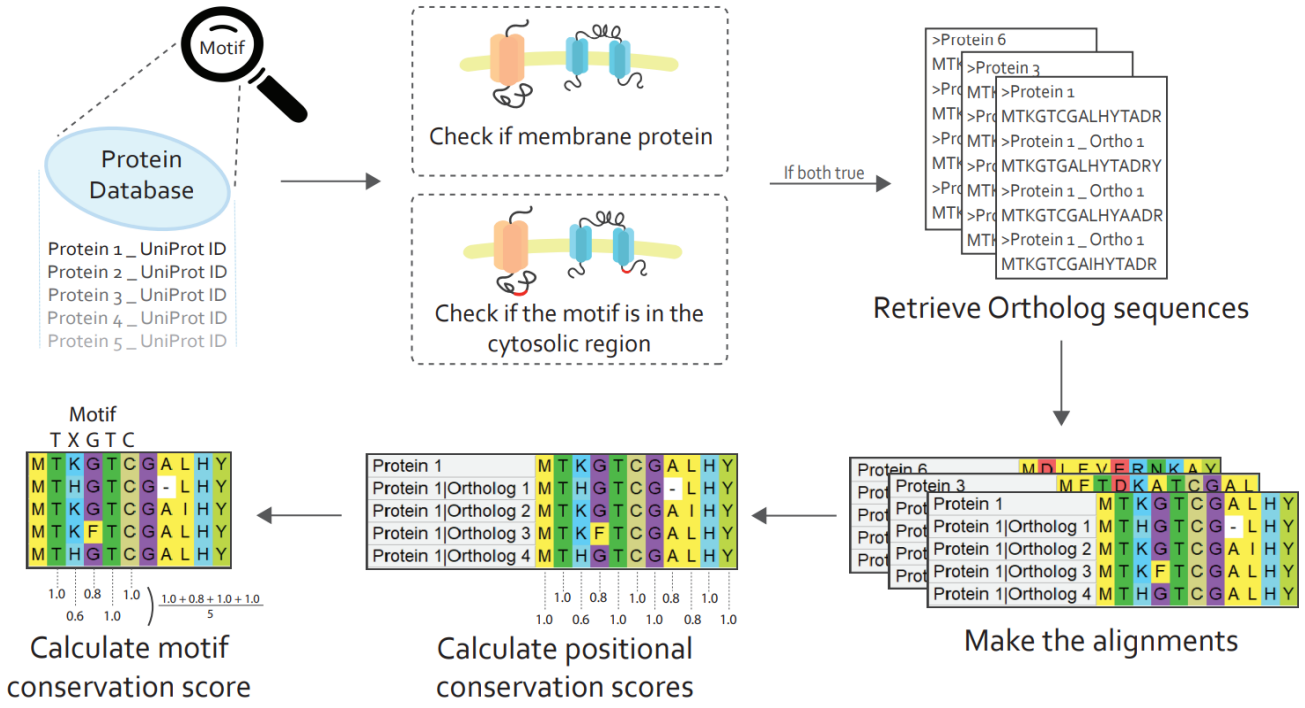


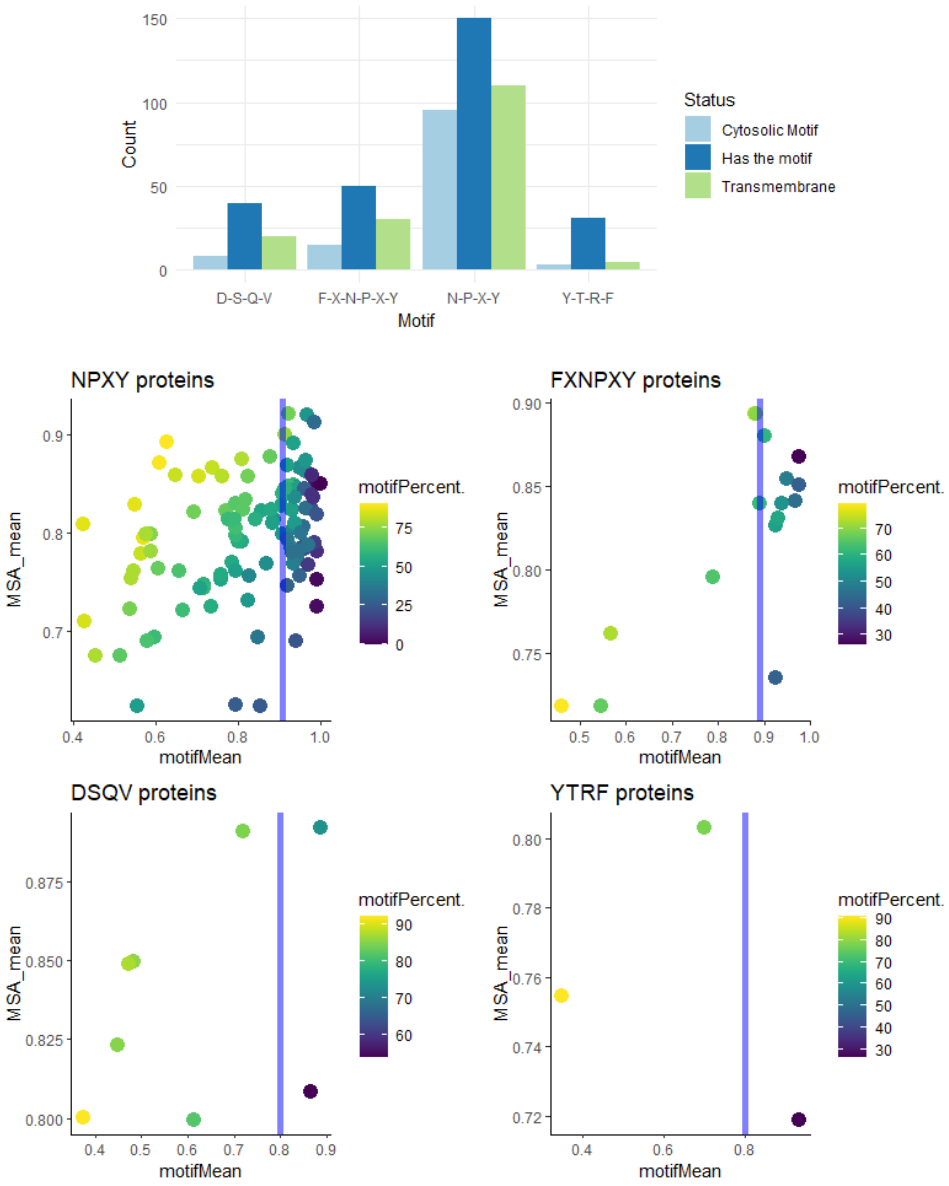
Method Figure

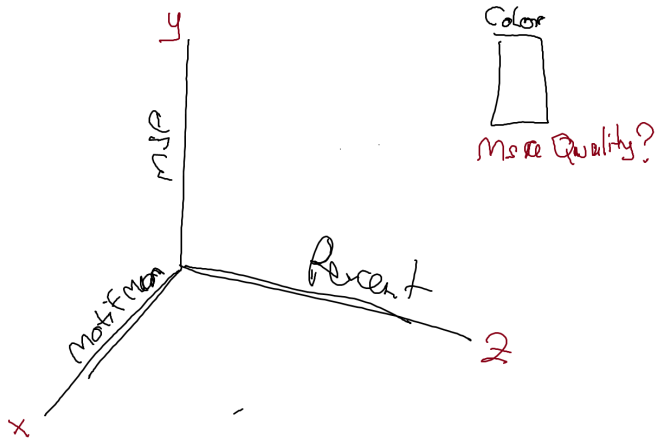
Identification of candidate transcytosis proteins through sequence analysis.

A set of motifs has been identified from literature which have been demonstrated to play a crucial role in the transcytosis process (as detailed in Table 1). To further investigate these motifs, the REST (Representational State Transfer) web service of the PrositeScan tool (de Castro et al., 2006) provided by the ExPASY portal was employed to search for proteom sequences that possess these motifs.The UniProt proteom database was searched for each motif, with isoforms filtered out and the search restricted to the Homo Sapiens organism. Metadata is generated using UniProt REST API, including the location of the motifs on the sequence, topology information, and database reference IDs of each protein. Through the analysis of topology information, proteins without transmembrane domains are filtered out. The proteins that contain the motif in their cytoplasmic domains were selected from the remaining protein sequences. These proteins have been designated as preliminary candidates. The orthologous protein sequences of the preliminary candidate proteins are obtained from the OMA (Orthologous Matrix) (Altenhoff vd., 2021) database. Using the orthologous sequences, multiple sequence alignments (MSAs) were built for each of the pre-candidate protein using the Clustal Omega (Sievers vd., 2011) alignment method.

The conservation level at each position in the multiple sequence alignments was determined by calculating the proportion of the most common amino acid at that position to the total number of sequences in the alignment. In order to take into account the effects of gaps in the multiple

sequence alignment, the conservation scores were calculated both with and without the positions that were mostly composed of gaps.




y
MRP
Most Frequent
Recent
z
x
Color
More Quality?