# 2. Methods

## 2.1 Dataset

In this study, we utilized the Protein Folding Kinetics Database, PFDB dataset (http://lee.kias.re.kr/%7ebala/PFDB), which comprises 141 single-domain proteins without covalently bound prosthetic groups or disulfide linkages [1]. Among these proteins, 89 were classified as two-state (2S) and 52 as non-two-state (N2S) globular proteins. PFDB, being the most recently updated database with accurate ln(kf) values and folding types, serves as the gold standard for developing prediction models. Additionally, we collected protein sequences from the Protein Data Bank (PDB) website (https://www.rcsb.org/), using the PDB\_IDs listed in the PFDB dataset. Initially, we downloaded files in .fasta format, each corresponding to a PDB\_ID containing the protein sequence. Subsequently, we developed a Python program to read these files and extract the protein sequences based on the specified chains and residues in the PFDB dataset. Finally, we integrated these sequences into the existing PFDB dataset, leaving the other columns unchanged.

References

1. B. Manavalan, K. Kuwajima, J. Lee, PFDB: a standardized protein folding database with temperature correction, Sci Rep 9 (2019) 1588