# A Protein Representation Model for Low-Data Protein Function Prediction

Serbulent Unsal[1,2*], Sinem Özdemir[1], Işık Özdinç[1], Amine Bayraklı[1], Muammer Albayrak[1], Kemal Turhan[1], Tunca Dogan[3*], Aybar C. Acar [2*]

*[1] Karadeniz Technical University, [2] Middle East Technical University, [3] Hacettepe University*

Proteins are macromolecules that are both building blocks and essential machinery of life. However, the knowledge about proteins is still limited. In particular, the UniProt knowledgebase, the largest protein information hub for protein science, has approximately 210 million protein records but only 0.5% of them are manually annotated or reviewed. Manual annotation of protein functions needs wet-lab experiments and interpretation of results by human experts. This is a slow and high-cost process. Especially in the last decade, rigorous efforts have been made to annotate proteins with automated systems such as machine learning algorithms. Machine learning-based protein annotation (a.k.a protein function prediction - PFP) has two major issues. The first one is the need for manual efforts in the data preprocessing process. Classical machine learning methods highly rely on manually extracted features (e.g. physical and chemical properties of proteins). The second problem is about the performance of these methods. The CAFA experiment is a well-known benchmark study in the PFP domain and is repeated periodically, where the aim is to predict gene ontology (GO) annotations in terms of molecular function (MF), biological process (BP) and cellular component (CC) categories. The results of CAFA showed that the development of successful PFP methods is still an open problem (e.g., the best performing method in the BP category could achieve Fmax: 0.42).

In this study, we aimed to develop a holistic protein representation using a multimodal learning model to predict protein functions even in the cases of low training data. We created representation vectors using protein sequence, protein describing text and protein-protein interaction data to achieve this goal. In theory sequence data includes all the knowledge about a protein. However, it currently is not possible to infer all of this knowledge directly from the sequence data. This is the reason behind our choice of using a multimodal approach. We took advantage of our previous benchmark study to choose the best sequence-based protein representation method, which is a protein language model. We incorporated protein-protein interaction based representations to our holistic protein representation model. The rationale is the assumption that interacted proteins are likely to act in the same function or biological process. Also these proteins are probably located at the same location in the cell. We use pre-trained natural language processing models to calculate text-based protein representations. This is the most meaningful data in the semantic context. Most of the time text data is directly refined from experimental results or literature. We aimed to increase low-data prediction performance using these 3 data types together. Moreover we showed that, once the sequence data is enriched with text and ppi data, this associative knowledge can be used to develop high performing sequence based models, which can be applied to low-data protein function prediction problems even when text and ppi data does not exist. For each data type, a selection of representation methods are applied on a carefully created benchmarking dataset. These representations are integrated together by using different approaches which are concatenation, autoencoder models and multimodal autoencoders. We also used transfer learning to create sequence based multimodal protein representations (SMPRs). SMPRs were intended to learn the relationship between sequence and other modalities and use them when only protein sequence data exists. The integrated representations were tested on our PFP benchmark (using GO categories of MF, BP, CC). This benchmark is specifically designed to evaluate protein representation learning methods under different training sample sizes (i.e., low, middle, and high) and function specificities (i.e., specific, normal, and shallow) to explore strengths and weaknesses of the benchmarked methods. We also tested our holistic protein representations for discovery of new immune-escape proteins in lung adenocarcinoma. The schematic representation of the stıdy is given in Figure 1.

We observed that multimodal learning approaches couldn't perform better than the best-performing representation (i.e., text-based). However, text data is scarce, so we also trained sequence-based multimodal models. These models are trained with multiple modalities. Then, we transferred this information to a sequence-based model. We observed better results compared to solely sequence-based models, on average, using this approach. Moreover, we showed that sequence-based multimodal models could perform notably better in predicting functions in low data scenarios.

A rigorous effort has been made for PFP with deep learning in the last decade. Our study contributes to this field with novel approaches, especially considering the subjects of data availability and annotation specificity. Although our carefully curated dataset is needed to be expanded with diverse approaches (e.g., train/test

splits considering annotation times), we believe our contributions will trigger the development of novel holistic approaches to be used in the PFP domain.

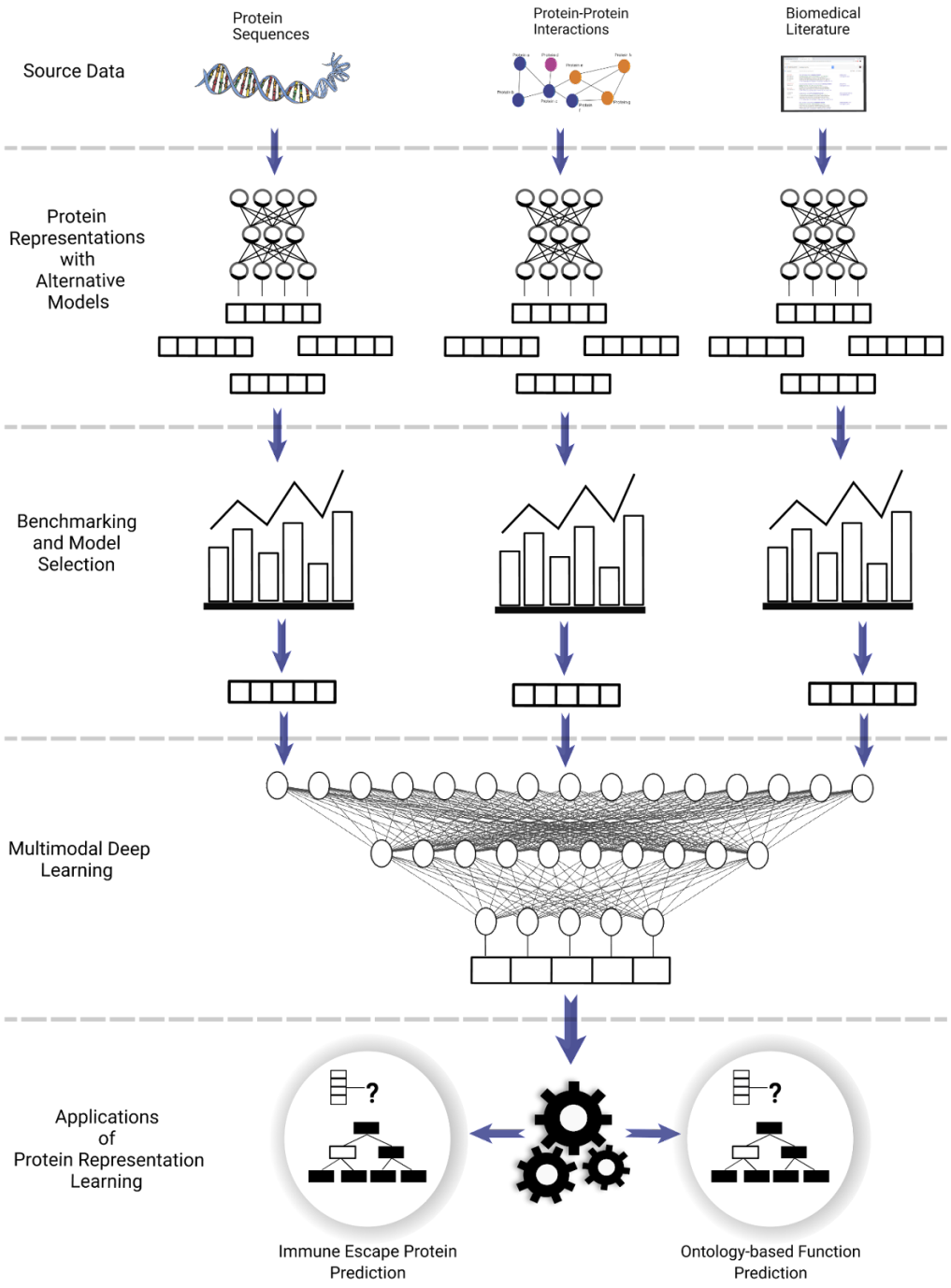


**Figure 1.** The overview of the study. We first created protein representations using protein sequence, protein-protein interactions, and literature-based text data. Then we benchmark them to find the best performing representation model for each protein function prediction task, especially in the cases of low training data. Finally, we tested our model to find new tumor immune-escape proteins and evaluate the results.