BERT-HemoPep60: A Deep Learning Method for Quantitative Hemolytic Activity Prediction of Peptides Based on Transformers and Domain Adaptive Pretraining

ProtBERT finetuned peptide toxicity regressor: Peptide Hemolytic Activity Prediction to human red blood cells (RBCs) for sequences with 5 to 60 residues.

we proposed BERT-HemoPep60, a deep learning method based on the transformer architecture and domain-adaptive pretraining (DAPT) to predict the hemolytic activity. By leveraging DAPT, our model can align peptide hemolytic data with Protein Language Models (PLMs), maximizing the power of large language models (LLMs). To enhance prediction using scarce and heterogenous experimental hemolysis data, we developed an innovative prefix-prompt approach. Experimental hemolysis data for six mammalian species and multiple hemolytic concentrations (HC5, HC10, and HC50) were integrated to train a prediction model for the human RBCs. Through five-fold cross validation, BERT-HemoPep60 achieved Pearson Correlation Coefficients (PCC) of 0.7431, 0.8088, and 0.7606 for HC5, HC10, and HC50 predictions, respectively.

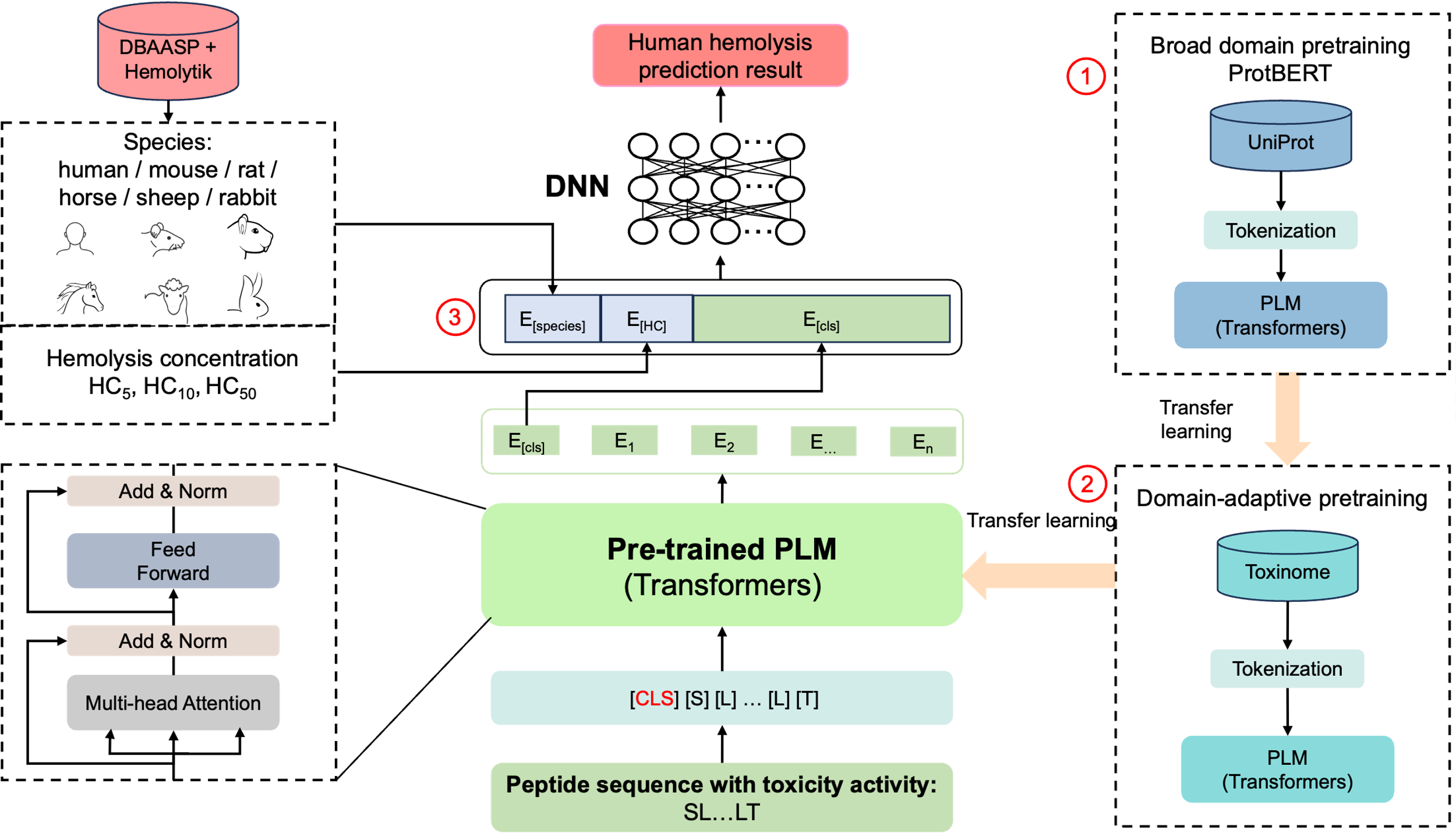


Figure 1: Model architecture of BERT-HemoPep60. ① A protein language model (PLM), ProtBERT proposed by Elnagga, is trained in a broad-domain corpus UniProt. ② The PLM is trained in an in-domain corpus Toxinome. For an input sequence, it is first tokenized for data representation, and the encoder layer derived from the pre-trained PLM are task fine-tuned for the downstream hemolytic regression task. ③ The [CLS] token embedding representing the aggregated features of the input sequence is fed into a deep neural network, along with a prefix prompt containing the target species and lysis concentration.