**Toxin Data Preparation**

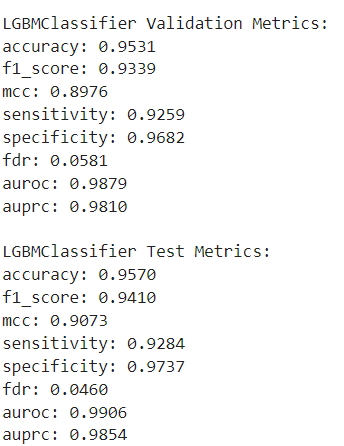
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | **Dataset** | **Number of Toxic(positive)** | **Number of Non-toxic(negative)** | | --- | --- | --- | --- | | Protein | Training set(60%) | 2053 | 3380 | | Testing set(20%) | 670 | 1142 | | Validation set(20%) | 648 | 1163 | | Peptide | Training set(60%) | 1812 | 1560 | | Testing set(20%) | 617 | 508 | | Validation set(20%) | 604 | 520 | |

Total Protein-9056, Total Peptides-5621

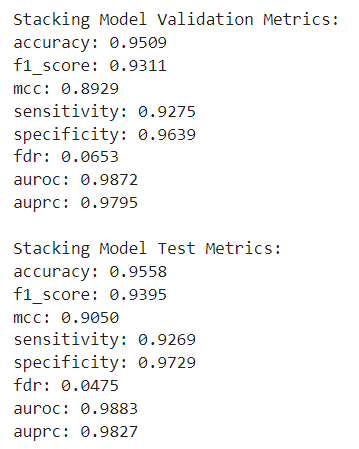
Experiment 1:(AAC + DPC)

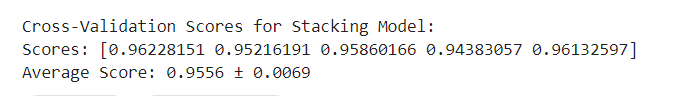
Here we process a dataset containing protein and peptide sequences to extract amino acid composition (AAC) and dipeptide composition (DPC) features. We then calculated AAC and DPC for each sequence, constructs a comprehensive feature set.

1.Protein results

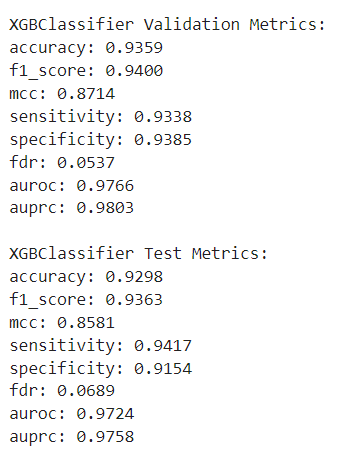


Stack model:

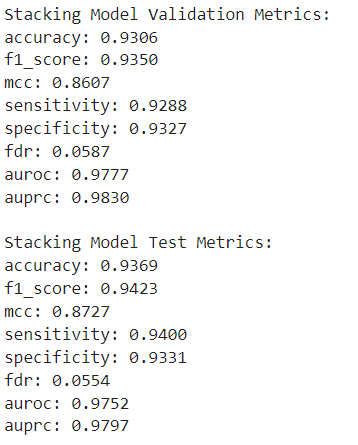


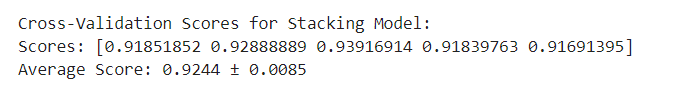


2.Peptide results:



Stack model:

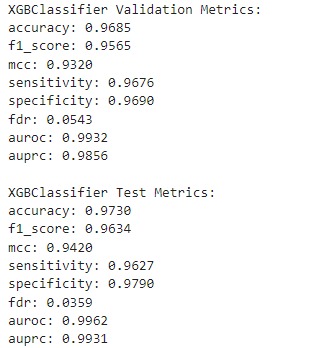




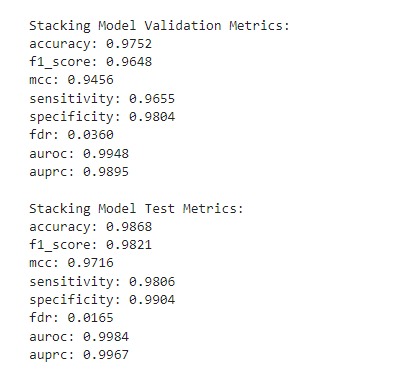
Experiment 2:(ProtBert)

In experiment 2, we have used ProtBERT to extract features of peptides and protein. ProtBERT is a protein language model inspired by the BERT architecture, designed to understand and generate protein sequences by learning the patterns and relationships between amino acids. These models, trained on extensive protein sequence datasets, can extract features that encapsulate the biological and chemical properties of the sequences.

1.Protein results:

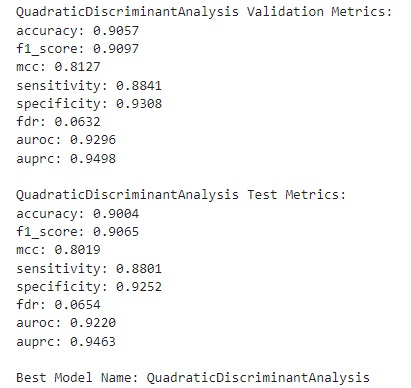


Stack model

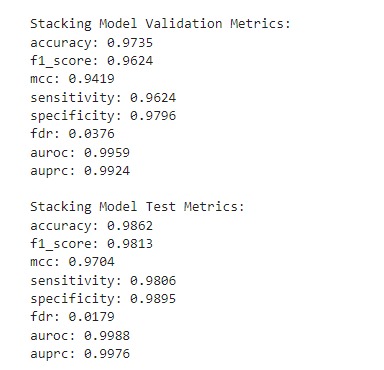


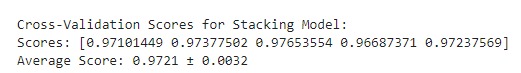


2.Peptide results:



Stack model

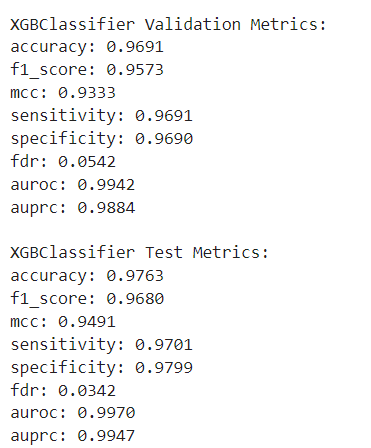




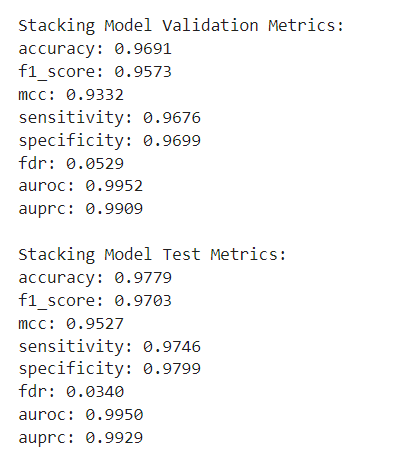
EXPT3-ESM

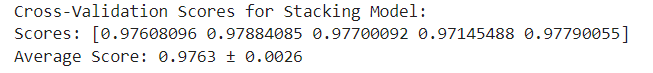
In experiment 3 , we have used Evolutionary scale modelling to extract feature of peptide. ESM is a technique based on protein language models, which are designed to understand and generate protein sequences by learning patterns and relationships between amino acids. These models, like those based on transformer architectures, are trained on large-scale protein sequence datasets. They can extract features that encapsulate the biological and chemical properties of the sequences.

1.Protein results:

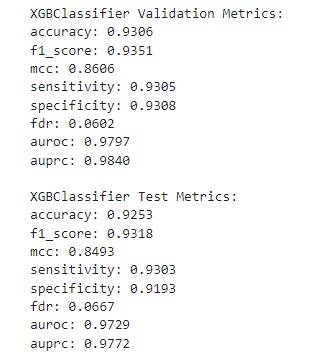


Stack model

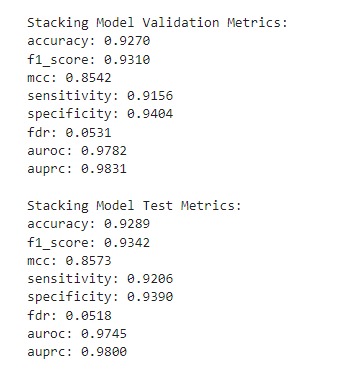




2.Peptide results:



Stack Model





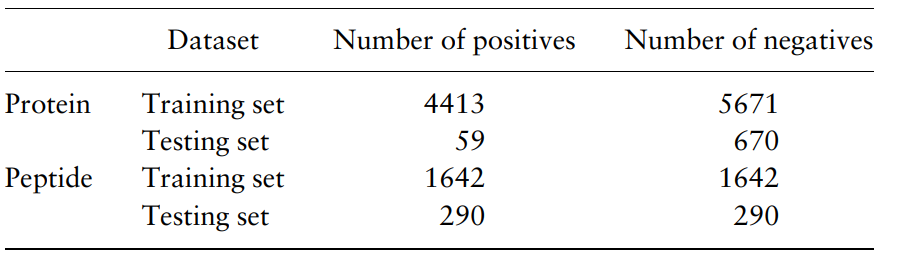
References:

ToxIBTL: prediction of peptide toxicity based on information bottleneck and transfer learning

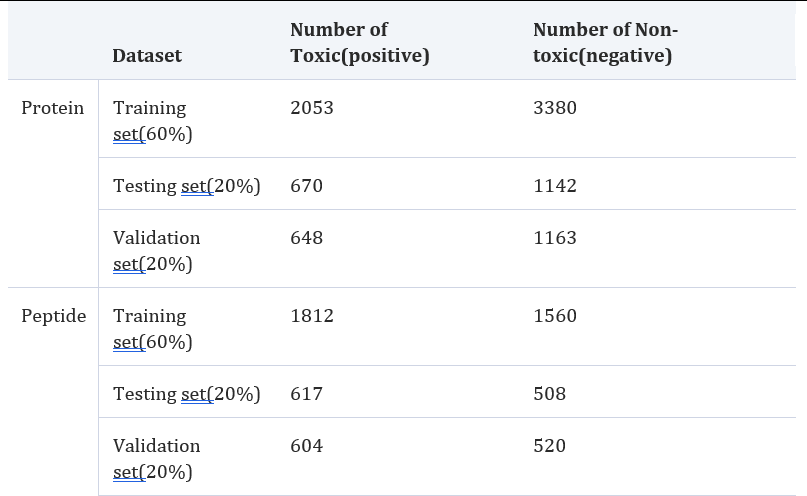
Comparision with the TOXIBTL:

The paper presents ToxIBTL, a deep learning model developed to predict the toxicity of peptides and proteins.

Data Used(TOXIBTL):



Data Used(By us):



Performance:

Predictive performance of various methods on Protein dataset:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | *F1\_score* | *MCC* | *auROC* | *auPRC* | *SN* | *SP* | *FDR* | *ACC* |
| ToxIBTL | *0.830* | *0.816* | *0.953* | *0.847* |  |  |  |  |
| Expt1 | *0.941* | *0.907* | *0.991* | *0.985* | *0.928* | *0.974* | *0.581* | *0.957* |
| Expt1(stack) | *0.940* | *0.905* | *0.988* | *0.983* | *0.928* | *0.973* | *0.065* | *0.956* |
| Expt2 | *0.963* | *0.942* | *0.996* | *0.993* | *0.968* | *0.979* | *0.054* | *0.973* |
| Expt2(stack) | ***0.982*** | ***0.972*** | ***0.998*** | ***0.997*** | *0.981* | *0.990* | *0.036* | *0.987* |
| Expt3 | *0.968* | *0.949* | *0.997* | *0.995* | *0.970* | *0.980* | *0.054* | *0.976* |
| Expt3(stack) | *0.970* | *0.953* | *0.995* | *0.993* | *0.975* | *0.980* | *0.053* | *0.978* |

Predictive performance of various methods on Peptide dataset:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Method | *F1\_score* | *MCC* | *auROC* | *auPRC* | *SN* | *SP* | *FDR* | *ACC* |
| ToxIBTL |  | *0.921* |  |  | *0.963* | *0.954* | *0.046* | *0.960* |
| Expt1 | *0.940* | *0.871* | *0.977* | *0.980* | *0.942* | *0.939* | *0.068* | *0.936* |
| Expt1(stack) | *0.942* | *0.873* | *0.978* | *0.983* | *0.940* | *0.933* | *0.059* | *0.937* |
| Expt2 | *0.910* | *0.813* | *0.930* | *0.950* | *0.884* | *0.931* | ***0.065*** | *0.906* |
| Expt2(stack) | *0.981* | ***0.970*** | *0.999* | *0.998* | ***0.981*** | ***0.990*** | *0.038* | ***0.986*** |
| Expt3 | *0.935* | *0.861* | *0.980* | *0.984* | *0.931* | *0.931* | *0.067* | *0.931* |
| Expt3(stack) | *0.934* | *0.857* | *0.978* | *0.983* | *0.921* | *0.940* | *0.053* | *0.929* |