Cyspresso

-Shreyasi Ghosh (423563)

Overview of the Dataset

- The original dataset used in the CysPresso model contained a list of 1249 CDPs(Cysteine Dense Peptides), including
  - ❏ UniProt accession number
  - ❏ primary sequence
  - ❏ source organism
  - ❏ expressibility
- After removing duplicate entries, 1227 CDPs remained for analysis. The dataset was further processed to identify CDPs as knottins and not knottins based on the presence of inhibitor cysteine knot peptides

- In this dataset, CDPs of 30 to 50 amino acids in length were tested.

- Protein representations were generated from this dataset using pre-trained models such as SeqVec and ProteInfer, resulting in numeric embeddings compatible with machine learning operations .

Note: The authors have used the same dataset, after preprocessing, to train a regression model and in classification (Random Forest).

Evaluation of Cyspresso

CysPresso was assessed using a variety of performance metrics to evaluate its predictive capabilities. The model's performance was evaluated through **Leave-one-out cross-validation**, which is a technique used to assess the model's ability to generalize to new data.

The assessment included measures such as **area under the curve (AUC)**, **sensitivity, specificity, precision, accuracy, and F1 score for both non-knottin and knottin CDP models.**

These metrics provided a comprehensive evaluation of CysPresso's performance in predicting the expressibility of cysteine-dense peptides.

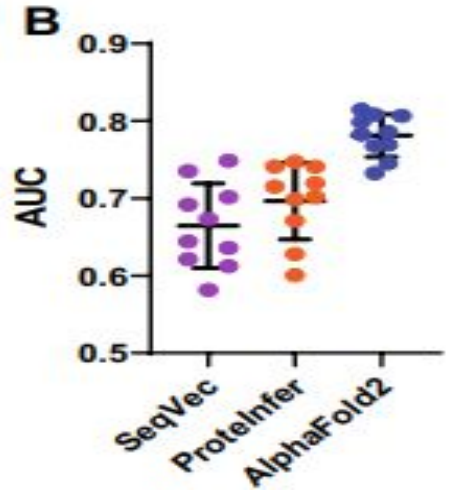
Table 2 Classifier performance metrics for CysPresso

Measure	Non-knottin CDP model	Knottin model
AUC	0.798	0.852
Sensitivity	0.853	0.764
Specificity	0.584	0.812
Precision	0.743	0.783
Accuracy	0.742	0.789
F1 Score	0.794	0.773

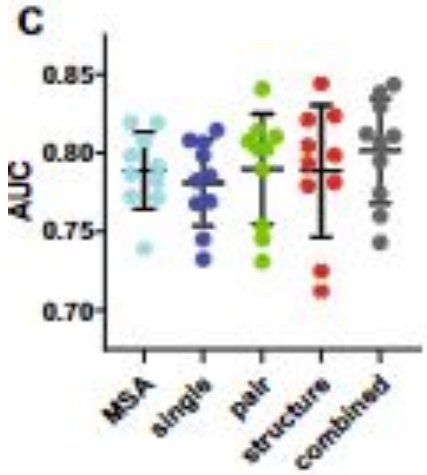
The performance of the non-knottin and knottin CysPresso models were evaluated by leave-one-out cross-validation

Comparison of Model Performance

- The models (SeqVec,ProteInferAlphafold2) were evaluated with a permutation test over 50 random permutations of the stratified dataset with a 90-10 split for training and validation sets. The different models were ranked in order of AUC (Area Under the Curve) and plotted.
- Out of the different protein representations tested, **SeqVec** has the least predictive power and the best performance was obtained with the embeddings from **AlphaFold2**.

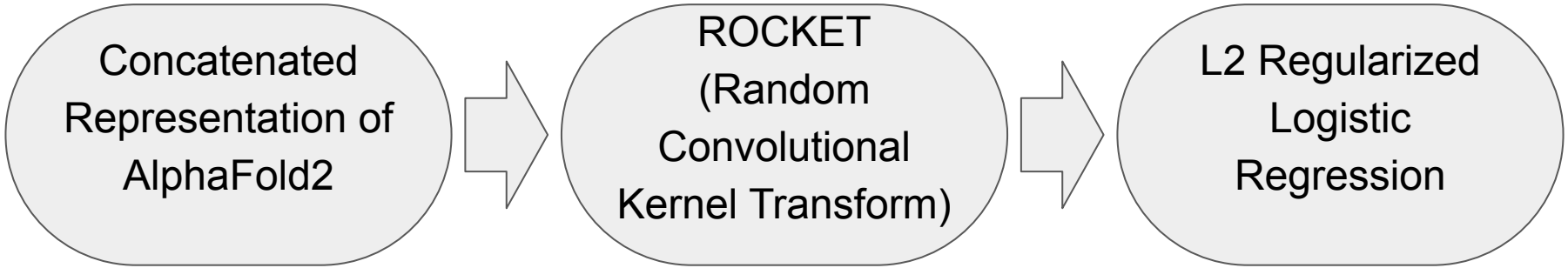


- AlphaFold2 uses four different types representations for structure prediction - MSA, Pair,single and structured. The researchers examined whether any of them provided superior model performance at predicting CDP expressibility and found out their AUC scores are similar. Then they **concatenated four representations** and that improved the performance.



Pipeline for the different Sequence Representation Learning

Pipeline 1(Cyspresso):



Architecture Diagram of Cyspresso

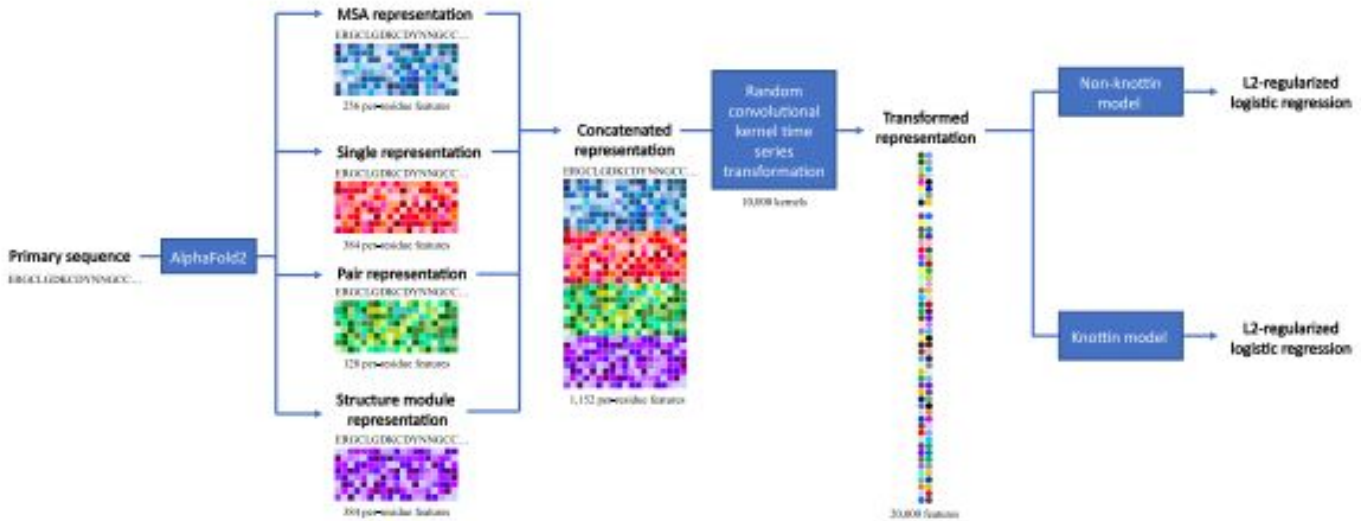
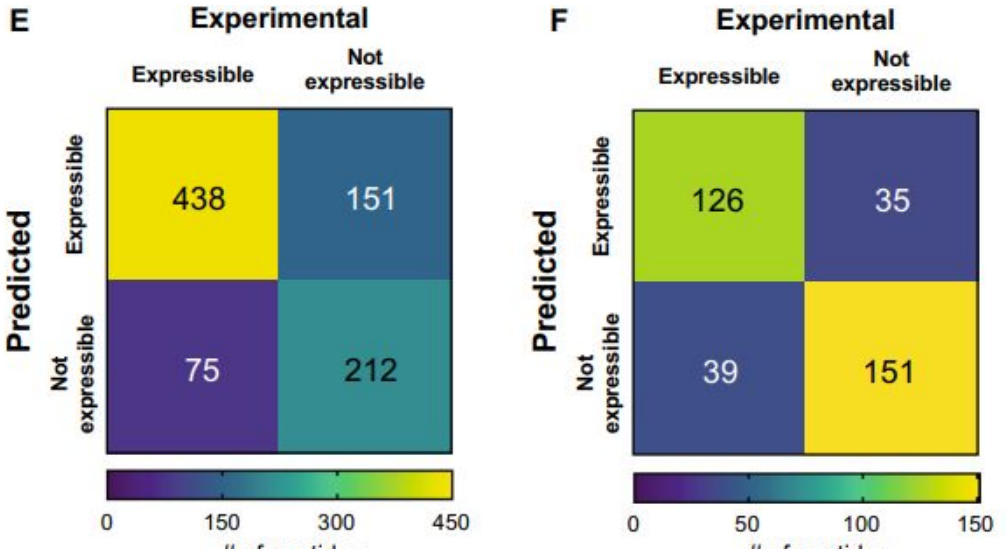


Fig. 3 Machine learning architecture diagram of CysPresso. The primary sequence of the CDP is used to generate MSA, single, pair, and structure module AlphaFold2 representations. The four representations are concatenated, and a time series transformation utilizing random convolutional kernels is carried out on the concatenated representation. The transformed representation is then used to predict expressibility using L2-regularized logistic regression machine learning models for knottin and non-knottin CDPs

- Various protein representations are tested to identify the ideal model to utilize with machine learning algorithms to predict expressibility from primary sequences. The tested representations include SeqVec, ProteInfer, and AlphaFold2. AlphaFold2 representations are found to perform best at predicting expressibility, and combining the four AlphaFold2 representations enhances model performance.

- The dataset is then stratified into knottin and non-knottin CDPs to improve predictions of expressibility for knottin peptides.

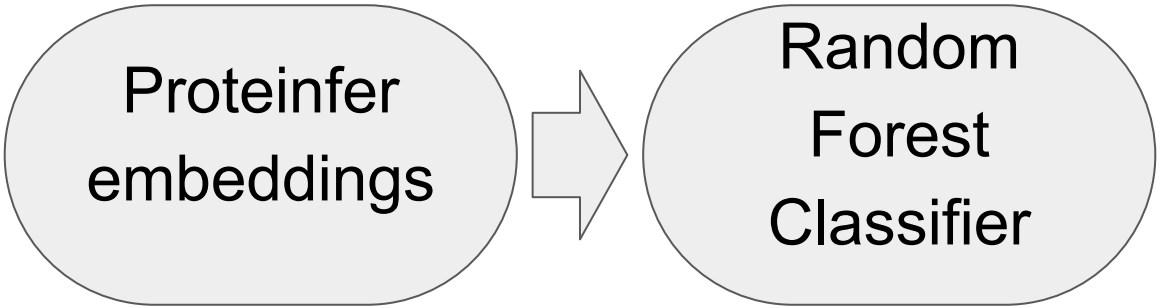
- A time series classification method implementing random convolutional kernels is applied to further enhance model performance for knottin peptide expressibility. (ROCKET uses random convolutional kernels to transform sequential features and is useful for capturing patterns in ordered data)
- The ROCKET representations were then classified using a L2-regularized logistic regression model.
- To evaluate the quality of the classification in terms of class-specific accuracies, we computed confusion matrices for both knottin and non-knottin models using ROCKET time-series classification with the combined AlphaFold2 representations. A leave-one-out cross validation scheme was followed, training as many model instances as samples per set and aggregating the predicted labels over all samples.



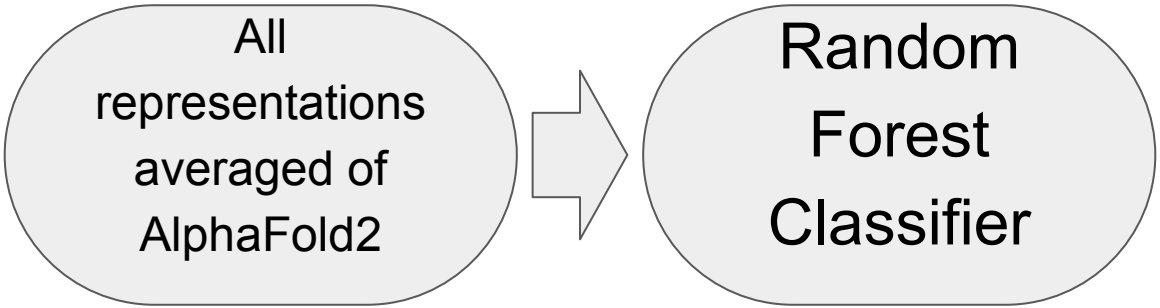
E :Confusion matrix of the fnal machine learning model for non-knottin CDPs evaluated by leave-one-out cross validation.

F: Confusion matrix of the fnal machine learning model for knottins evaluated by leave-one-out cross validation

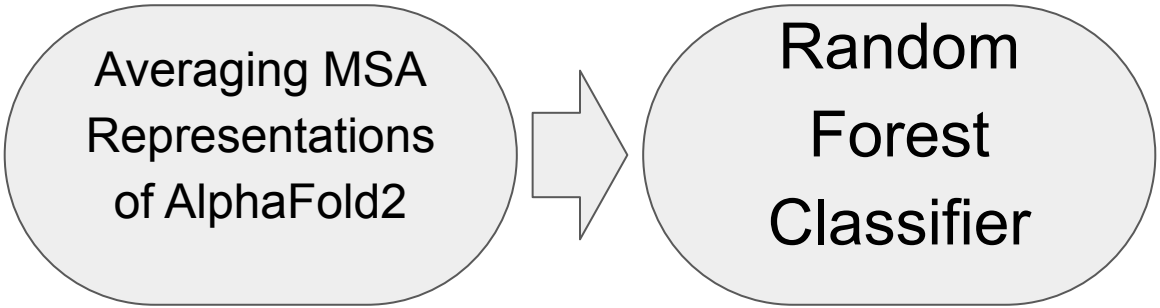
Pipeline 2



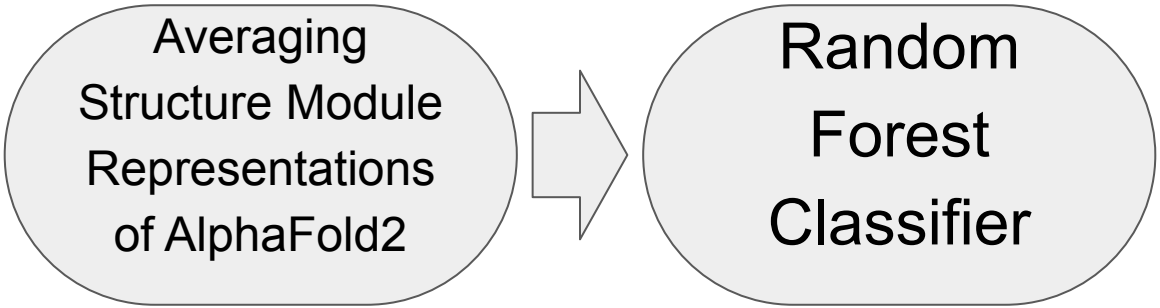
Pipeline 3:


All representations averaged of AlphaFold2
Random Forest Classifier

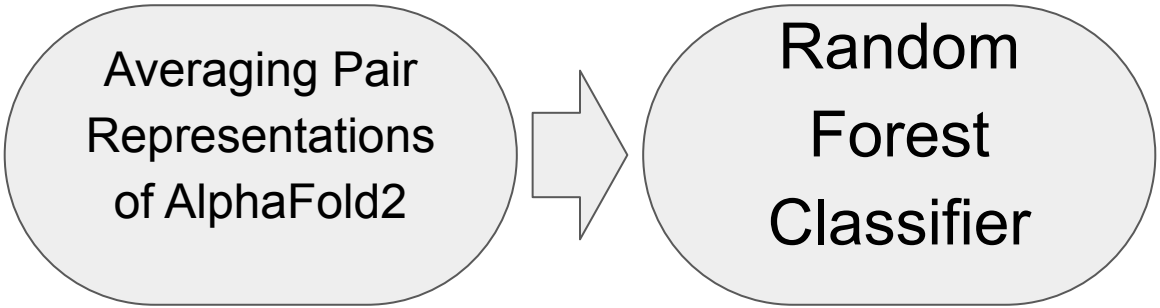
Pipeline 4:


Averaging MSA Representations of AlphaFold2
Random Forest Classifier

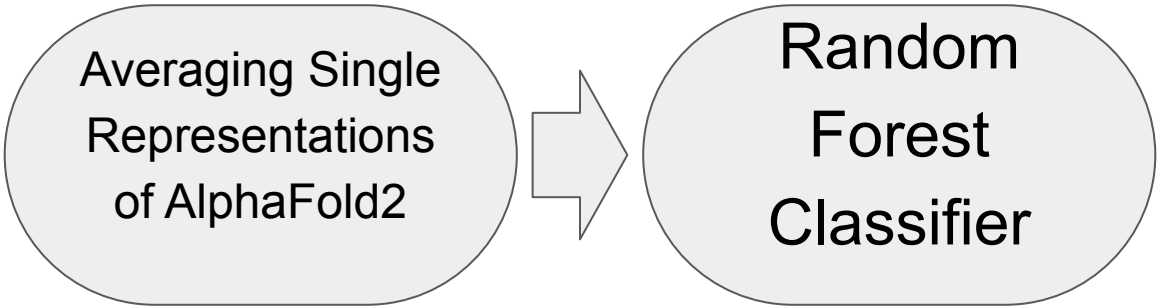
Pipeline 5


Averaging
Structure Module
Representations
of AlphaFold2
Random
Forest
Classifier

Pipeline 6


Averaging Pair Representations of AlphaFold2
Random Forest Classifier

Pipeline 7:


Averaging Single Representations of AlphaFold2
Random Forest Classifier

Some Pointers:

- The models from Pipeline 1 to Pipeline 7 is tested with the help of **ten fold cross validation.**
- Pipeline 1 to 7 has been only applied to the knottin dataset.
- Similarly, the researchers have applied the same pipeline on the not knottin dataset.