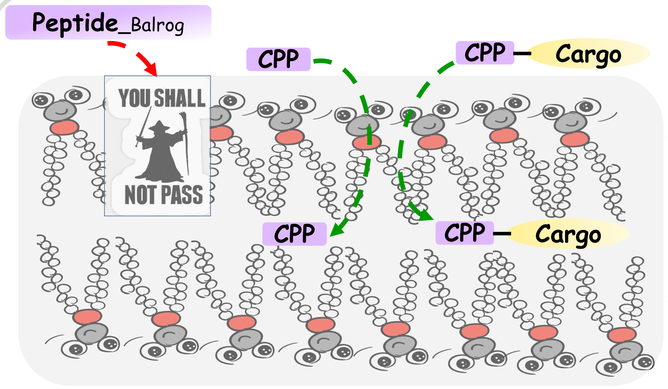
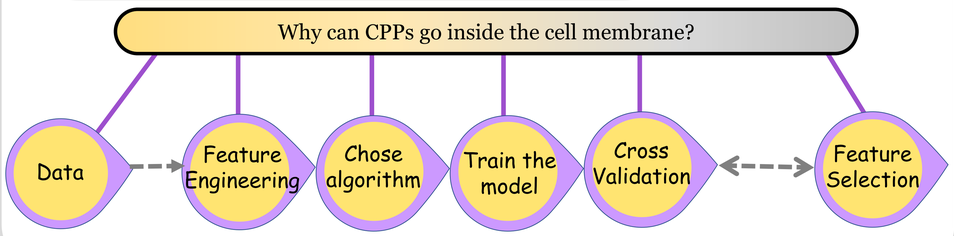
Building and Interpreting Random Forest based Cell Penetrating Peptide Prediction Model

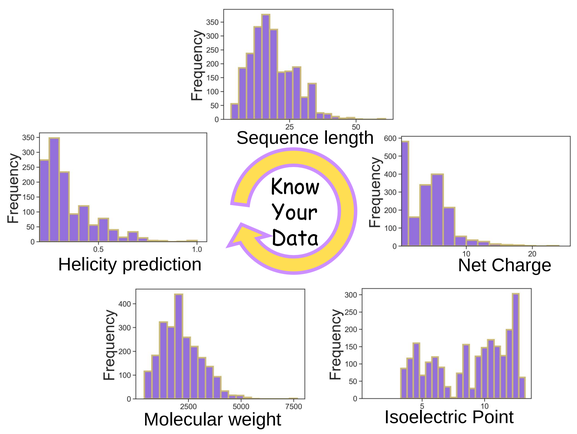
Shilpa YADAHALLI, Chandra S. VERMA

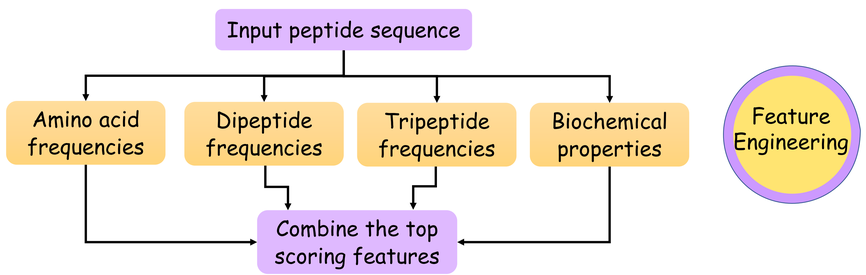
Bioinformatics Institute, A*STAR (Agency for Science, Technology and Research), Singapore

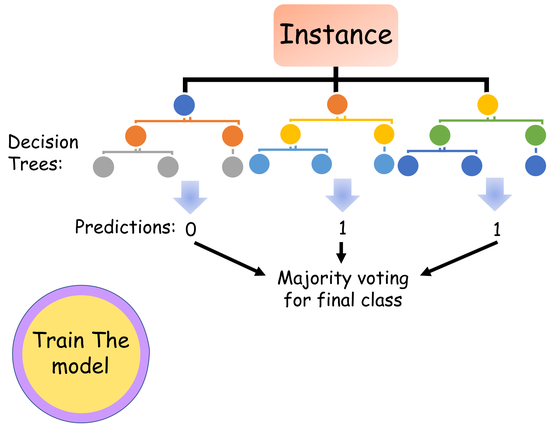


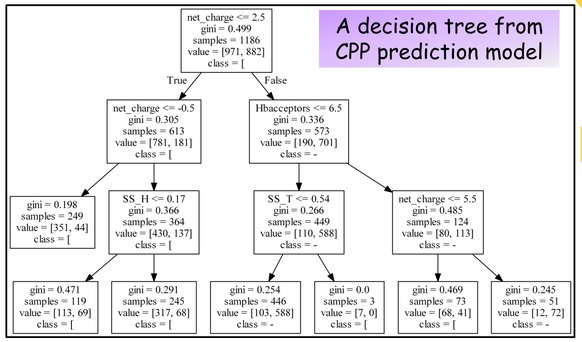
- The process of introducing drugs into cells has always proved a major challenge
- The cell membrane is selectively permeable and supports 'no generic mechanism' for their uptake
- It is important to know what makes a peptide permeable

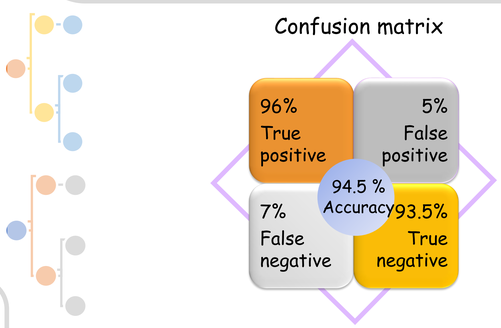


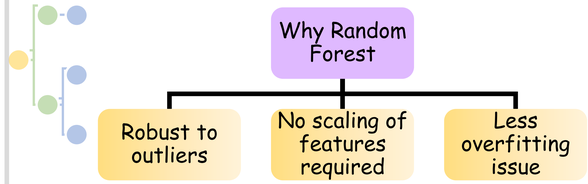




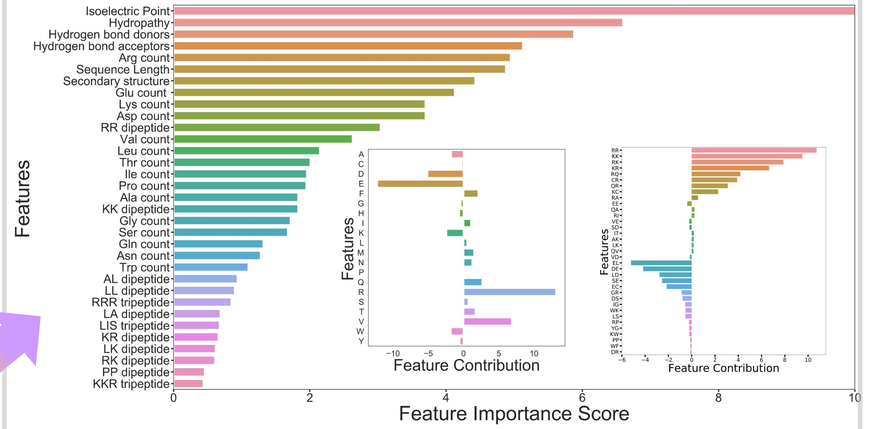


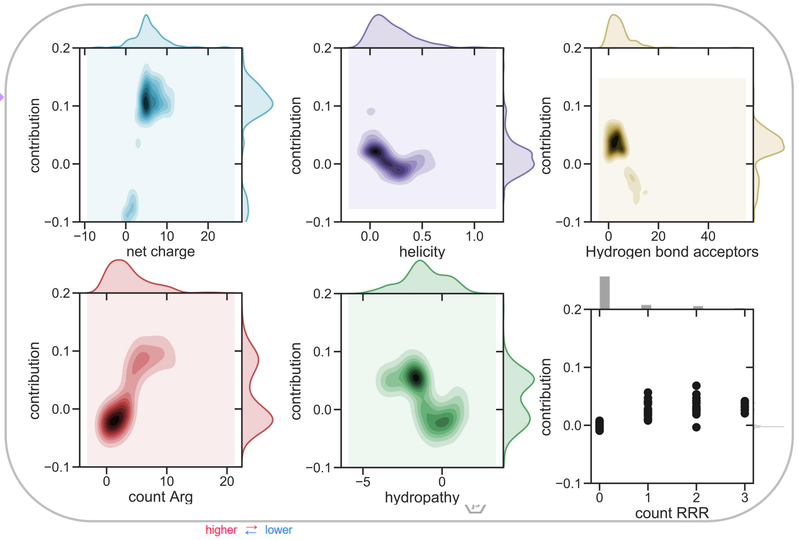


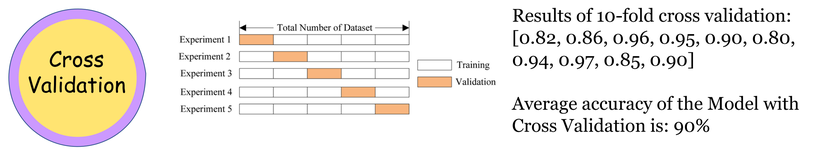




Interpreting the model: Light of Earendil for a Dark Forest







Results of 10-fold cross validation: [0.82, 0.86, 0.96, 0.95, 0.90, 0.80, 0.94, 0.97, 0.85, 0.90]

Average accuracy of the Model with Cross Validation is: 90%

Summary: Targeting intracellular pathways with peptides is increasingly becoming popular. However, cellular permeability of the peptides remains a major challenge with no apparent sequence-structure-activity relationship. One class of peptides called Cell Penetrating Peptides (CPPs), have an ability to get across multiple cell membranes without irreversibly damaging the membrane. They are also capable of delivering biologically active cargo into cells. While there exist various models of peptide uptake by cells (ranging from passive permeation to receptor-mediated uptake), there is no consensus view on the relationship between a peptide's sequence and its ability to enter cells. In the current study, we explore the application of machine learning methods to a dataset of CPPs. We annotate the peptides with a range of properties based on their amino acid sequences and find that Random Forest Classifiers can predict cellular uptake with an accuracy of ~94.5%. The proposed model is a simple and fast sequence-based predictor which shows higher accuracy than most of the state-of-the-art predictors available. We have also done in depth analysis of the features, their contribution to the prediction values and probability of the prediction. The contribution values are studied for each class separately, which tells us if they affect the permeability in positive/negative/neutral manner.

There appears to be a delicate balance of amino acid composition between peptides which enter cells without disrupting the membrane and peptides that destroy the membrane. We plan to extend our study to probe a dataset of antimicrobial peptides (AMPs) to understand this further. These models can hopefully guide the design of functional peptides. We also discuss issues related to the robustness of experimental data.