With the advancement of sequencing technology, there have been a growing number of uncharacterized or unannotated protein sequences. Experimentally determining their functionality is both costly and time consuming. Computational methods that are capable of automatically predicting the annotations from protein sequences are the best alternative. Machine Learning based supervised classification algorithms have evolved as the go-to method to solve this problem because of their ability to discern generalizable patterns from data. These methods require a batch of characterized (labelled) protein sequences from which they learn a set of rules or patterns employed to functionally annotate new uncharacterized sequences. However, the characterized dataset needs to meet certain rigid specifications such as balanced dataset, enough instances per category of labels, feature set independence among others, to ensure generalizability of the ML model. Most common protein classification datasets fail to meet at least one if not all of these specifications which leads to suboptimal performance of ML models on blinded test set.

In this work, we presented a computational base carefully designed to address all the issues prevalent in protein classification pertinent datasets. To test the effectiveness of our developed computational base, we applied it on a specific protein classification task, categorization of plant acyl-ACP TEs into their respective substrate specificity groups. Acyl-ACP TEs have been effectively demonstrated to confer substrate specificity in transgenic hosts engineered for production of free fatty acids and related derivatives. This strategy has enabled enrichment of product distributions for medium-chain chemical species in bacteria, yeast, and plant systems [1], [2], [3]. While the medium-chain activity is most desirable due to the relative scarcity of 8 to 12 carbon oleochemicals, most acyl-ACP TE exhibit a preference for C14 and above.

Thus, several efforts have been made to bioprospect genomes of plants with high fractions of the medium-chain oils to identify and implement the thioesterase gene responsible for the narrow substrate specificity [4], [1]. While progress has been demonstrated in identifying the features which dictate specificity in acyl-ACP thioesterases among plants [5], the throughput for bioprospecting, characterizing, and in some cases, engineering the acyl-ACP thioesterase is largely inhibited by the testing pipeline, which requires derivatization of the free fatty acids into fatty acid methyl esters prior to analysis with gas chromatography [6]. A method for inferring substrate specificity from thioesterase gene sequence would therefore expedite this process, removing the necessity of expressing each homolog in a host to gain insight to its selectivity profile.

\*\*Importance of plant acyl TE classification into medium chain – big picture, small picture (lack of any existing computational model) \*\*. As part of this study, we also compiled a characterized TE dataset consisting of the TE sequences and their corresponding substrate specificity category, long chain, medium chain or mixed. The assembled dataset is the perfect representation of the most frequently occurring datasets in protein classification domain. It exhibits all the attributes of a typical protein classification dataset including but not limited to small size, class imbalance, high dimensionality, correlated features, which makes a conventional ML algorithm falter on unseen test sets. Our method on the other hand displayed consistent performance when assessed through a rigorous validation strategy, devised so as to capture any trace of overfitting. We achieved classification accuracies (measured on the basis of correctly predicting the substrate specificity category of TEs) in the range of 0.59 to as high as 1 with mean accuracy of 0.83 and standard deviation of 0.06 across 10000 simulations of our model validation scheme (discussed in Methods section). The model also attained a mean precision score of 0.89 and mean recall score of 0.91 on medium chain TE class prediction, the TE substrate specificity category that we are primarily interested in. In contrast an existing sequence similarity based modeling approach achieved a mean accuracy of 0.37, mean precision and mean recall scores of 0.005 and 0.002 respectively on medium chain TE class prediction (detailed study given in Results section).

Although we have attained reasonably high accuracy on TE substrate specificity classification task where the TEs were grouped into three different bins each representing a range of substrate specific chain lengths, we acknowledge that our model is currently unable to accomplish a deeper level of TE classification across all chain lengths. It must be noted that the referred limitation to achieve a higher resolution of TE classification can be ascribed to a lack of characterized TE dataset and is not solely the fault of the model per se. It might be also be more desirable to obtain continuous valued substrate predictions across the chain lengths. Our method provides the flexibility to adapt to such requirements by simply modifying the base learners from classifiers to regressors. Moreover, it is worth mentioning that the three bins/categories of TE substrate specificity are not equally well predicted; prediction of the medium chained TE bin obtained the highest precision score of 0.89 followed by the long chained bin and the bin of mixed specificity which achieved mean precision scores of 0.79 and 0.69 respectively. Even though at face value it can be viewed as a modeling limitation, the SVM base model hyperparameter “class weights” can be easily readjusted to impose further emphasis on the most poorly predicted category or the one that represents the subject of interest, thereby providing another technique to deal with class imbalance. In the TE substrate specificity prediction problem, the model was already biased towards the subject of our interest, medium chained TEs. Hence, we decided against tuning the “class weight” hyperparameter. Furthermore, it has been previously recognized that primary sequence alone may be insufficient to perfectly classify protein sequences and addition of protein structural features might boost prediction accuracy \cite. Taking that argument into account, a Convolutional Neural Network (CNN) base model might be conveniently incorporated into the ensemble framework where the CNN extracts structural features from a 3-D voxel representation of proteins, given that protein structural information is readily available, and passes on its predictions to the meta learner.

Our method can be effortlessly extrapolated to other application areas in the protein classification domain ranging from applications as general as protein structural class prediction \cite or protein-protein interactions \cite to as specific as TE substrate specificity prediction or protein glycosylation sites prediction \cite. While general applications such as protein-protein interactions suffer from dataset imbalance \cite, more specific tasks, for instance glycosylation sites prediction may encounter yet another set of difficulties relating to small sized datasets. Issues related to high dimensionality and correlated feature set are ubiquitous in the protein classification domain. Our framework helps to resolve these omnipresent drawbacks in protein classification datasets while maintaining the computational efficiency required for swift functional characterization.