Classification of proteins into their respective functional categories remains a long-standing key challenge in computational biology. Machine Learning (ML) based discriminative algorithms have been used extensively to solve this problem. Typical ML algorithms perform accurately on an ideal dataset that contains appropriate number of sequences for each class of proteins, required to learn generalizable patterns embedded in the dataset. However, it is extremely common in computational biology to encounter small to medium-sized noisy, imbalanced characterized datasets with very high sequence similarity between proteins that belong to varying functional groups. Traditional ML based supervised classification methods when trained on such datasets produce suboptimal performance on blinded test sets since they suffer from overfitting, class imbalance and the curse of dimensionality. Although these problems have been acknowledged in previous literature, efforts to tackle them remain sparse. Herein we present a model that is specifically designed to address the issues of high dimensionality, small size, class imbalance and high sequence similarity in the dataset at every stage of its pipeline. We used our model to classify plant acyl Thioesterases into their respective substrate specificity category. [One or two lines about plant acyl TEs, their uses and how identifying medium chained TEs can help on a grander scheme of things]. The characterized TE dataset generated as a part of this study is an ideal representation of the majority of computational biology datasets that exhibit the issues described above. We followed a rigorous model evaluation scheme to measure model generalizability and ensure model robustness. Our method can be effortlessly extrapolated to other application areas in protein classification domain.

Our tool is freely available in GitHub and can be accessed using the following link: https://github.com/deeprob/ThioesteraseEnzymeSpecificity