

**Invitation to M.Tech. Thesis Defense of Srijanee Gupta: July 20, 2023 (Thursday): 02:00 PM – 03:00 PM IST**

In Partial Fulfillment of the Requirements for the Degree of

**M.Tech. CB**

**Srijanee Gupta (MT21231)**

Will defend her thesis

**Title: “ThpPred: An ML-based tool for predicting therapeutic proteins/peptides”**

IIIT-D Faculty and Students are invited

**Date: July 20, 2023 (Thursday)
Time:** **02:00 PM – 03:00 PM IST**

**Venue Details:** **A-320, 3rd Floor, R&D Building**

**Examiner: Internal:   Gaurav Ahuja**

**External/~~Internal~~: Jaspreet Kaur Dhanjal**

**Advisor: G P S Raghava**

**Co-Advisor NA**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Abstract**

ThpPred is a web-based tool, developed for predicting druggable proteins/peptides. The main dataset used in this study contained 356 therapeutic proteins/peptides and 356 random proteins/peptides, curated from DrugBank, Uniprot and other sources. In order to provide a fair assessment, we did internal validation on 80% of the data and external validation on the remaining 20%. In this study, we have implemented the following methods for predicting druggability of proteins/peptides; i) machine learning models on features chosen using SVC-L1, Variance Threshold, and correlation coefficient; ii) machine learning models on single feature (AAC, DPC & TPC); and iii) MERCI-based motif search. The goal was to construct the best model and install it on a web server by training it on protein sequences of already existing medications. When compared to other models, the XGB-based model performed the best on AAC features and obtained maximum AUCs of 0.91 and 0.91 on the training and validation datasets, respectively for the alternate dataset consisting of 356 positive sequences and 3560 negative sequences. On the other hand, the RF-based model performed admirably on DPC features and obtained maximum AUCs of 0.91 and 0.89 on the training and validation datasets for the main dataset. The AUC score and accuracy for both datasets improved when motif labels were added to ML predicted labels. ThpPred was created to determine if a protein is therapeutic or not by combining motif search with RF and XGB models. The platform helps the scientific community create more effective protein-based medicines by providing a free web server and a standalone package. Overall, the results of the study indicate that ThpPred has the potential to improve the development of pharmaceuticals and protein-based treatments for the treatment of numerous diseases.