SL-Pred

A multi-view subcellular localization prediction tool for multi-location human proteins

राष्ट्रीय प्रौद्योगिकी संस्थान पटना

NIT·PATNA

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

Authors:
Go¨khan O¨zsarı, Ahmet Sureyya Rifaioglu,
Ahmet Atakan, Tunca Dogan,
Maria Jesus Martin, Rengu¨l Cetin Atalay a
nd Volkan Atalay

November 28, 2022

Members

- B Rahul – 2006028
- G A Venkata Harish – 2006175
- M Karthik - 2006171

Department of Computer Science and Engineering
National Institute of Technology Patna

Overview

1.Introduction

2.Problem Statement

3.Novelty

4.Datasets and APIs

5.Methods and Technique used

6.Predictions Obtained

Introduction

- Proteins have evolved to function optimally in a specific subcellular localization. Hence, the correct transport of a protein to its final destination is crucial to its function.
- SL-Pred is a tool to predict multi-view subcellular localization for human proteins.
- This tool consists of nine independently developed models for proteins which have annotation with nine subcellular localizations.
- It also exploits the features of 40 different protein descriptors from the publicly available API tools: POSSUM, SPMAP and iFeature.
- A Support Vector Machine (SVM) is used to construct prediction models, which produces probabilistic scores indicating the probability for localization for every protein sequence.
- Finally by applying a threshold on the weighted score, we get a binary prediction for the localization of that particular protein sequence.

Problem Statement

- ✓ To analyze and quantify the uncertainties and achieve good performance under noisy data.

- ✓ Computational approaches have been used for the prediction of protein attributes, such as functions protein and small molecule interactions, implications and high-level heterogeneous relationships with the aim of aiding experimental studies by reducing costs and required times.

- ✓ With the aim of preparing both comprehensive and reliable training / test datasets for SL prediction, the author introduced a new SL hierarchy (Supplementary Section S1) that combines UniProt Knowledgebase (UniProtKB) SLs (UniProt-SL) and Gene Ontology (GO) Cellular Component (CC).

INPUT: Protein Sequences.

OUTPUT: Binary predictions of subcellular localizations.

Novelty

The main contributions of this work are as follows:

- The author perceive prediction by probabilistic models through binary classification by applying a threshold on the weighted model.
- The performance of this SL-Pred is evaluated and compared with six state-of-the-art methods, namely Multiloc2, LocTree2, Cello2.5, SubCons, DeepLoc and YLoc+.
- All the four benchmarking datasets are used evaluate all the methods and their performances.

Featurization and APIs

- **POSSUM** is a bioinformatics toolkit for generating numerical sequence feature descriptors based on PSSM (Position-Specific Scoring Matrix) profiles. It is a versatile with an online web-server that can generate 21 types of PSSM-based feature descriptors which addresses a crucial need for bioinformatics.

- **SPMap** (Subsequence-profile map) is used for functional classification of protein sequences which is a system based feature space mapping. It considers all the subsequences as a distribution over a quantized space by discretizing and reducing the dimension of huge space of all possible subsequences

- **iFeature** is a python-based web server toolkit which is used for calculation of a wide range of structural feature descriptors from protein and peptide sequences as well as other macromolecules.

Datasets

In the SL-Pred, for the training and evaluation purposes, four datasets have been used.

Four datasets are used for the training and evaluation of SL-Pred:

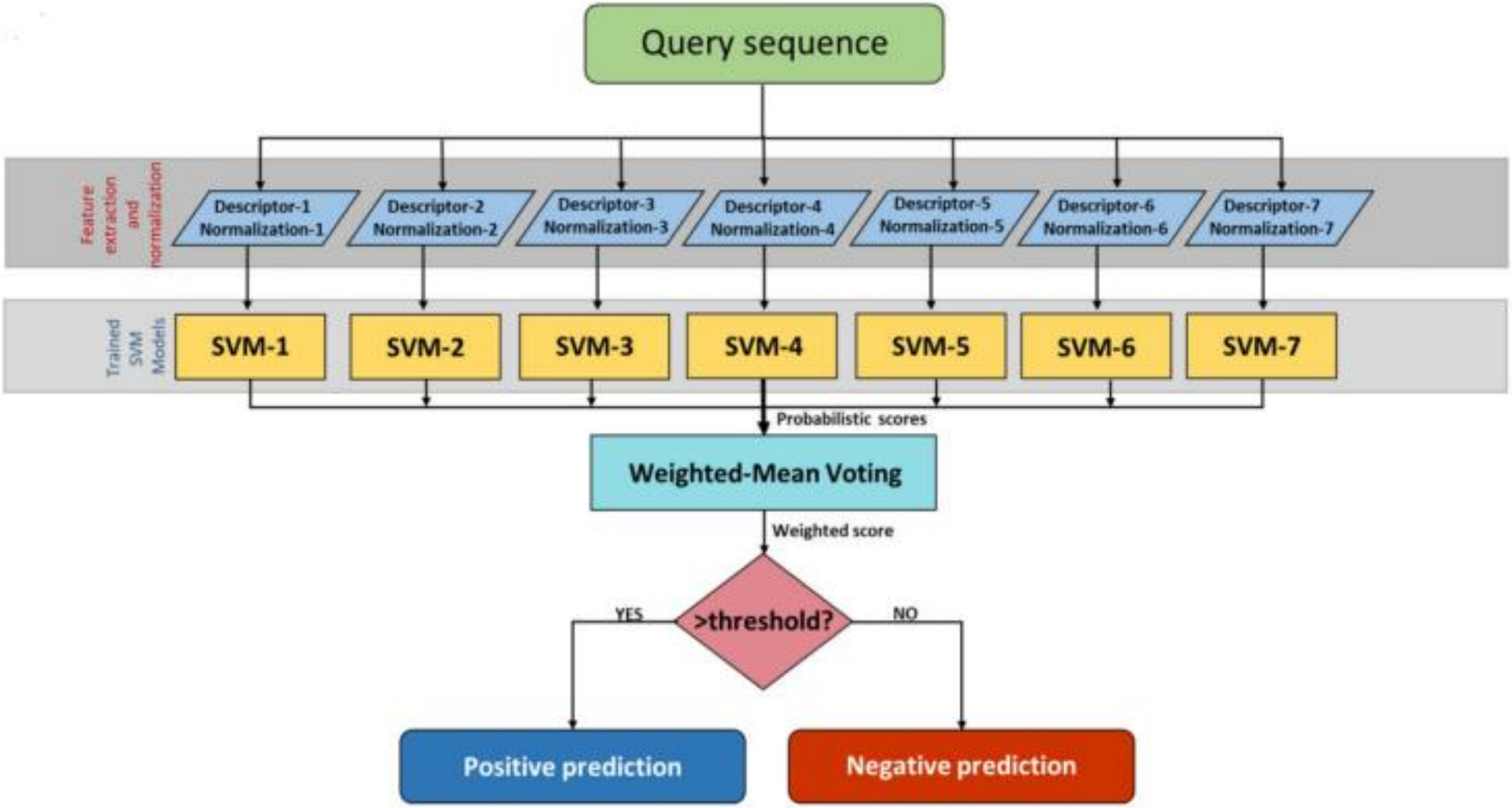
- Trust dataset is our in-house dataset, composed of 4431 human proteins, 35 unique SL terms and 8098 SL annotations, that is employed for the training, validation (10-fold cross -validation on 'trust-train') and testing of our method.
- The multi-labeled-test dataset comprises 559 subcellular localization annotations for 363 proteins, where the proteins may have one or multiple location annotations. Golden data set was constructed as a benchmark dataset by the developers of the SubCons tool and composed of 1226 human proteins and 3306 annotations.
- Golden-trust dataset is the refined version of the golden dataset, which is composed of 57 2 human proteins and 1810 annotations. Here, the golden dataset is modified according to the procedure applied in constructing the trust-dataset on top of removing proteins that are in the trust-dataset.
- The multi-labeled, golden- and golden-trust datasets are used for the independent evaluation of SL-Pred and comparison with the state-of-the-art methods.

Method and Technique Used

- First, we map the UniProt-SL terms to their equivalents in GO-CC.
- Second, we introduce additional UniProt-SL term relationships based on semantic relationships of the corresponding terms in GO-CC.
- This way, all disconnected terms in UniProt-SL gets connected to each other, constituting a connected component of 521 SL terms.
- Using the newly constructed hierarchy, we propagate SL annotations of UniProtKB/Swiss-Prot human proteins all the way to the root and then select the annotations of nine main SLs to be used in the model construction.

Method and Technique Used [Contd...]

The walkthrough pictorial representation of the SL-Pred from the protein sequence to the binary prediction is as follows


Query sequence
Feature extraction and normalization
Descriptor-1 Normalization-1
Descriptor-2 Normalization-2
Descriptor-3 Normalization-3
Descriptor-4 Normalization-4
Descriptor-5 Normalization-5
Descriptor-6 Normalization-6
Descriptor-7 Normalization-7
Trained SVM Models
SVM-1
SVM-2
SVM-3
SVM-4
SVM-5
SVM-6
SVM-7
Probabilistic scores
Weighted-Mean Voting
Weighted score
>threshold?
YES
NO
Positive prediction
Negative prediction

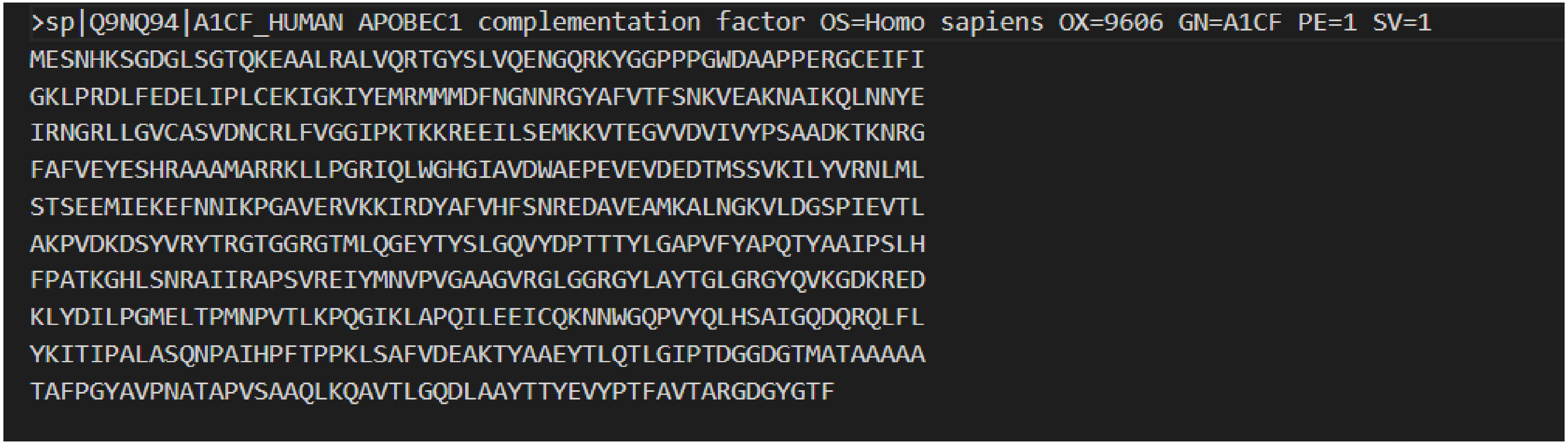
Results Obtained

- The SL-Pred and its entire functionalities along with the required datasets, features and APIs are bundled into a directory, mapping all required files properly.
- This entire directory is uploaded to Google Drive (-) then imported and mounted to Google Colaboratory and the driver file of the entire directory (run_SL-Pred.py) will be executed.
- The required output will be saved in the 'input_predictions.csv' file.

Input Type:

Inputs are passed in .fasta format,

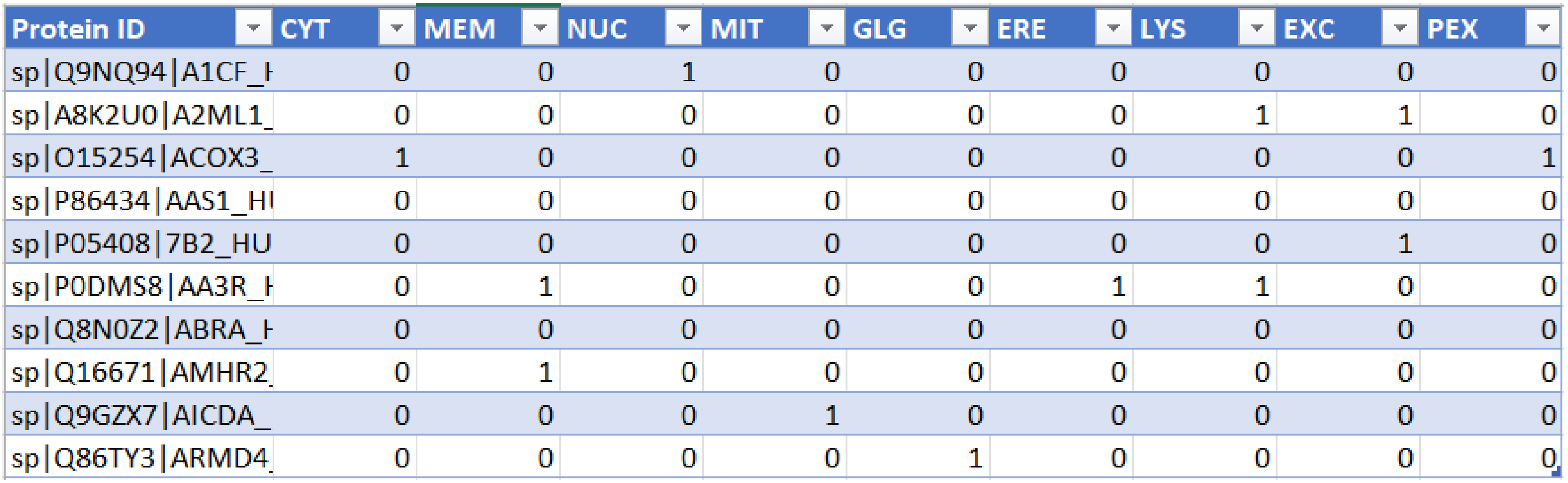
.fasta format saves data of nucleotide sequence or amino acid(protein).



```
>sp|Q9NQ94|A1CF_HUMAN APOBEC1 complementation factor OS=Homo sapiens OX=9606 GN=A1CF PE=1 SV=1
MESNHKSGDGLSGTQKEAALRALVQRTGYSLVQENGQRKYGGPPPGWDAAPPERGCEIFI
GKLPRDLFEDELIPLCEKIGKIYEMRMMMDFNGNNRGYAFVTFSNKVEAKNAIKQLNNYE
IRNGRLLGVCASVDNCRLFVGGIPKTKKREEILSEMKKVTEGVVDVIVYPSAADKTKNRG
FAFVEYESHRAAAMARRKLLPGRIQLWGHGIAVDWAEPEVEVDEDTMSSVKILYVRNLML
STSEEMIEKEFNNIKPGAVERVKKIRDYAFVHFSNREDAVEAMKALNGKVLDGSPIEVTL
AKPVDKDSYVRYTRGTGGRGTMLQGEYTYSLGQVYDPTTTYLGAPVFYAPQTYAAIPSLH
FPATKGHLSNRAIIRAPSVREIYMNVPVGAAGVRGLGGRGYLAYTGLGRGYQVKGDKRED
KLYDILPGMELTPMNPVTLKPQGIKLAPQILEEICQKNNWGQPVYQLHSAIGQDQRQLFL
YKITIPALASQNPAIHPFTPPKLSAFVDEAKTYAAEYTLQTLGIPTDGGDGTMATAAAAA
TAFPGYAVPNATAPVSAAQLKQAVTLGQDLAAYTTYEVYPTFAVTARGDGYGTF
```

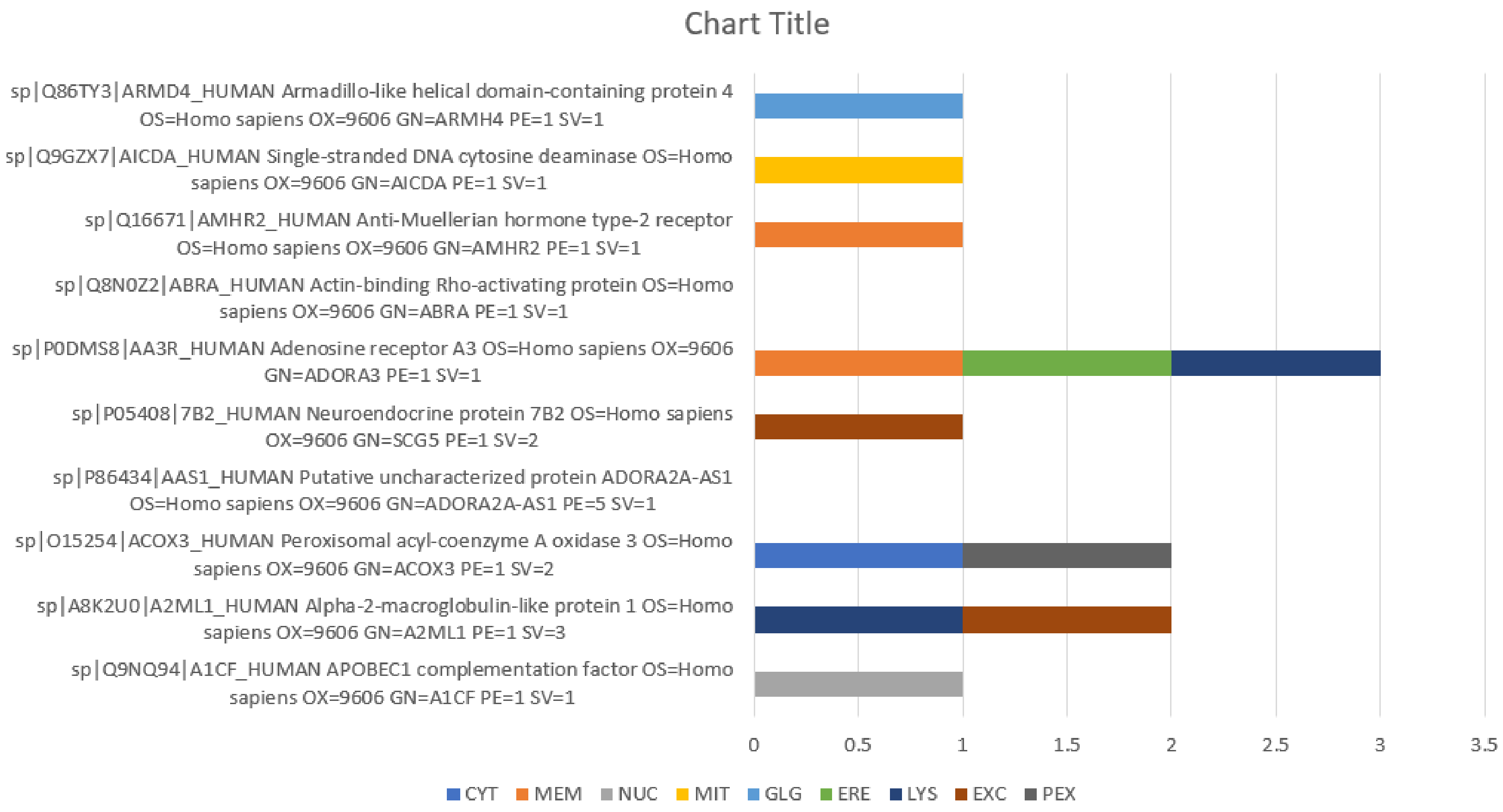
Output Predictions

Output/Predictions has been saved in csv format with name 'input_predictions' After running the code:



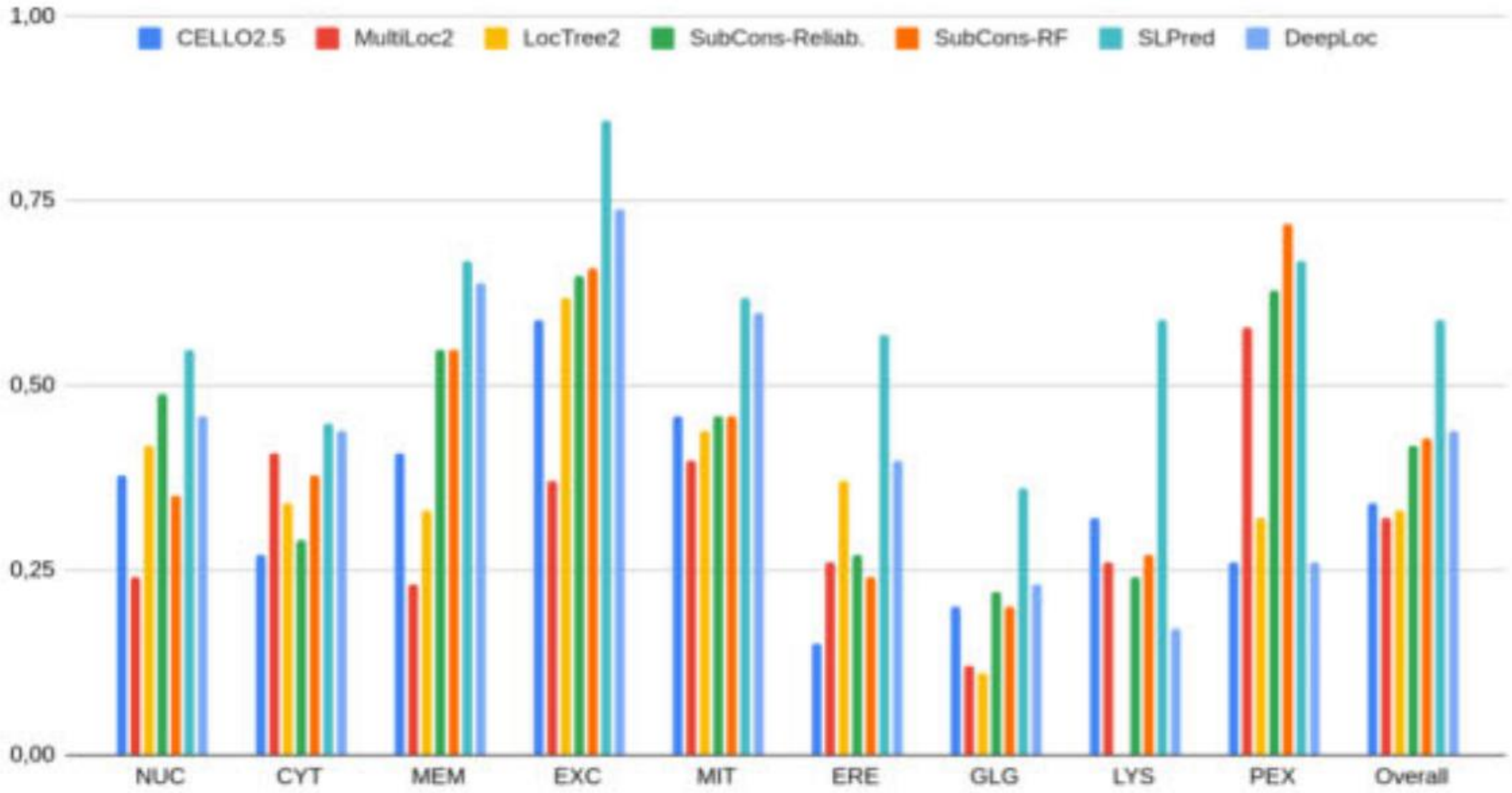
Protein ID	CYT	MEM	NUC	MIT	GLG	ERE	LYS	EXC	PEX
sp|Q9NQ94|A1CF_H	0	0	1	0	0	0	0	0	0
sp|A8K2U0|A2ML1_	0	0	0	0	0	0	1	1	0
sp|O15254|ACOX3_	1	0	0	0	0	0	0	0	1
sp|P86434|AAS1_HI	0	0	0	0	0	0	0	0	0
sp|P05408|7B2_HU	0	0	0	0	0	0	0	1	0
sp|P0DMS8|AA3R_H	0	1	0	0	0	1	1	0	0
sp|Q8N0Z2|ABRA_H	0	0	0	0	0	0	0	0	0
sp|Q16671|AMHR2_	0	1	0	0	0	0	0	0	0
sp|Q9GZX7|AICDA_	0	0	0	1	0	0	0	0	0
sp|Q86TY3|ARMD4_	0	0	0	0	1	0	0	0	0

Output Predictions



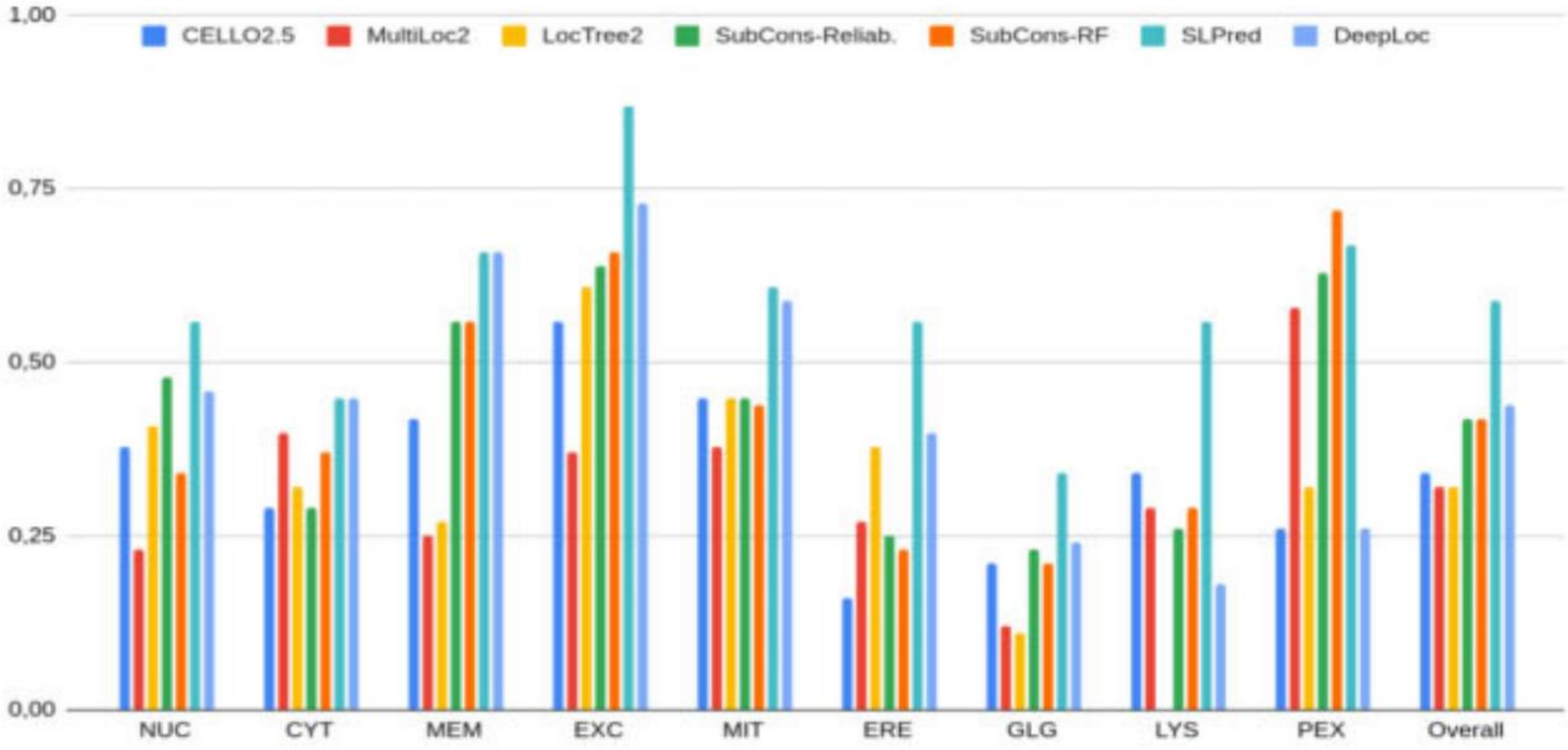
Results Obtained [Contd...]

a) Trust – set dataset
MCC score 0.59



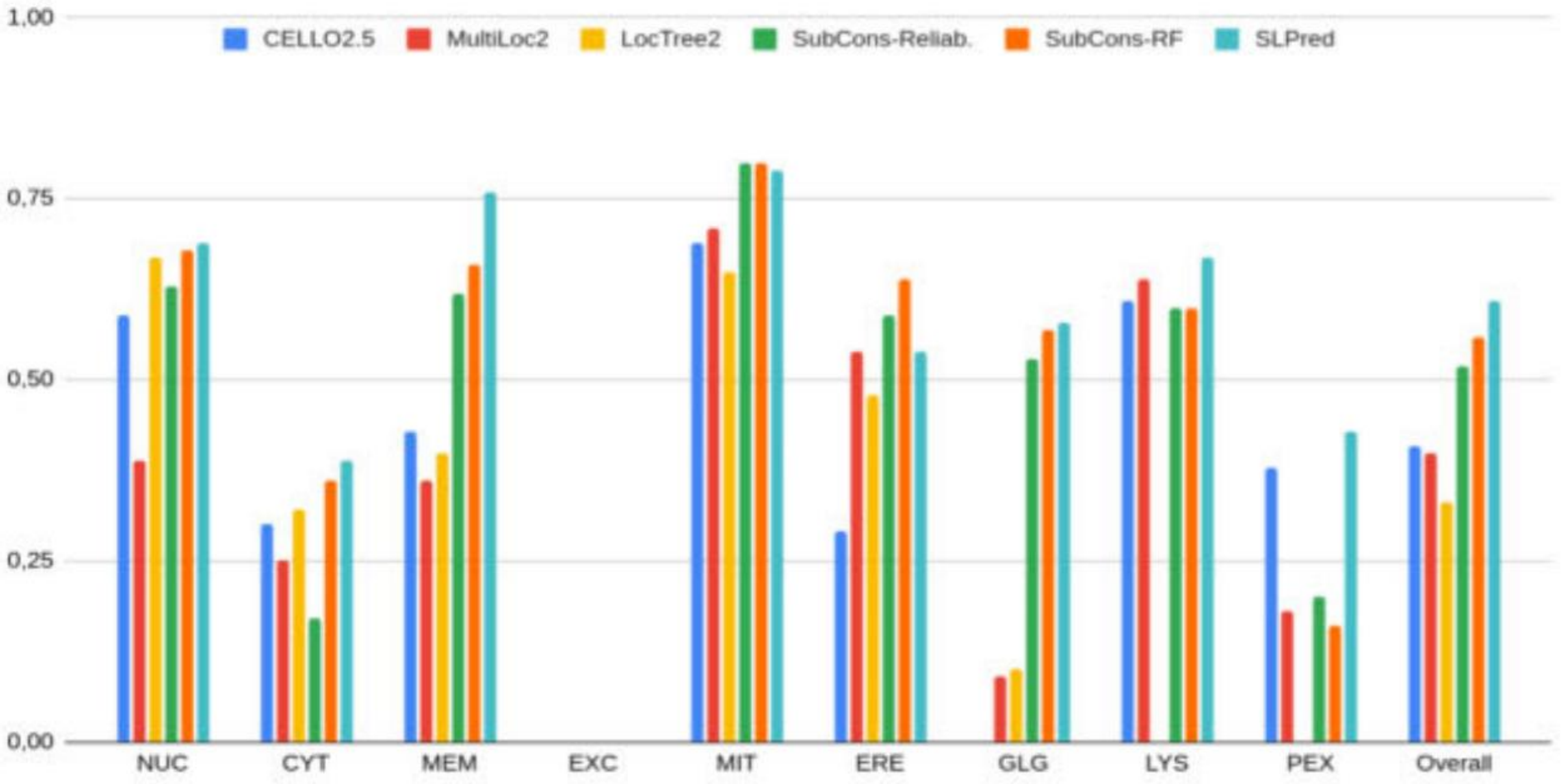
Results Obtained [Contd...]

b) Multi-labeled trust dataset
MCC score: 0.45



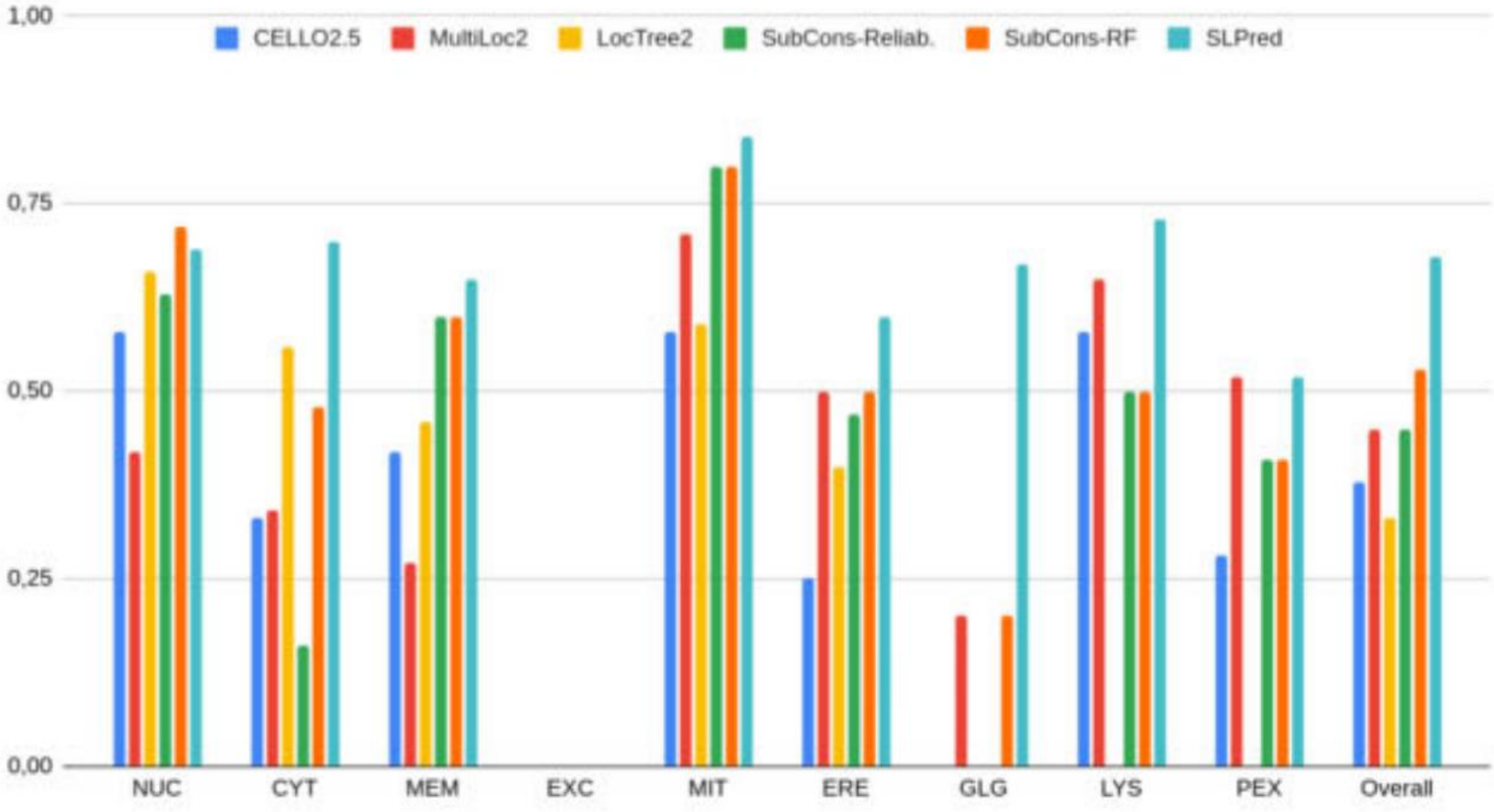
Results Obtained [Contd...]

c) Golden Dataset
MCC score:061



Results Obtained [Contd...]

d) Golden-trust Dataset
MCC Score: 0.68



Summary

- Accurately predict the subcellular locations (SLs) of proteins is a critical topic in protein science.

- For a query protein sequence, SL-Pred provides predictions for nine main SLs using independent machine-learning models trained for each location.

- They used UniProtKB/Swiss-Prot human protein entries and their curated SL annotations as our source data.

- We tested SL-Pred on multiple benchmarking datasets including the dataset set by authors and compared its performance against six methods.

THANK YOU!