Prediction of Ligand Activity at Subcellular Location

Manikya Varshney[1,2], Srijan Verma[1,3], Govinda KC[4], Giovanni Bocci[5], Tudor I Oprea[5] and Suman Sirimulla[4,6,7]

[1] Department of Pharmaceutical Sciences, The University of Texas at El Paso, Texas 79968, USA
[2] Department of Biological Sciences, BITS Pilani, Rajasthan, India.
[3] Department of Pharmacy, BITS Pilani, Rajasthan, India.
[4] Computational Science Program, The University of Texas at El Paso, Texas 79968, USA.
[5] Division of Translational Informatics, University of New Mexico Health Sciences Center, Albuquerque, NM 87131, USA.
[6] Department of Pharmaceutical Sciences, School of Pharmacy, The University of Texas at El Paso, Texas 79902, USA.
[7] Department of Computer Science, The University of Texas at El Paso, Texas 79968, USA.

Abstract

Understanding subcellular distribution and the mechanism of xenobiotics can help in modulating subcellular dysfunction mediated diseases. Therefore, with improved knowledge of how xenobiotics are distributed across subcellular locations and the mechanism for a specific molecule can play a crucial role in assessing drug efficacy and toxicity. Such knowledge would widen therapeutic windows by allowing specific receptors to be targeted efficiently. Based on datasets that provide information on the subcellular locations of proteins and their ligands, we developed machine learning models for 40 subcellular locations. Such models were trained and validated based on the grid search method and best models based on Cohen's Kappa scores were selected.

Introduction

Overview :

1) This work focuses precise subcellular drug delivery which holds great promise in treating diseases effectively.
2) Herein we report the first in-silico study for predicting drug localization for 40 different subcellular regions.
3) We manually curated the subcellular locations and collected ligand data. Forty different classification models were developed by exploring various RDKit features and trained using different machine learning (ML) classifiers from scikit-learn.

Outcome :

1) Based on the extensive evaluation of the performance of different features and ML algorithms, best models were selected. These models can predict the binding site for the given drug.
2) Models were tested across five evaluation metrics, and the best models, after performing grid search, were implemented.

References

(1) The Portal is available at : https://drugdiscovery.utep.edu/subcell/
(2) Zheng, N. Cheminformatic and Mechanistic Study of Drug Subcellular Transport/Distribution, deepblue.lib.umich.edu, 2011.
(3) RDKit: Cheminformatics and Machine Learning Software http://www.rdkit.org.
(4) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Others. Scikit-Learn: Machine Learning in Python. The Journal of machine Learning research 2011, 12, 2825–2830.

Acknowledgement

We thank the High-Performance Computing Center at UTEP for assistance in using the Chanti cluster and the Texas Advanced Computing Center (TACC) at The UT Austin for providing HPC resources that have contributed to the research results reported within this paper.

Methodology

1) Data was collected from PDSP Ki and Pharos databases.

2) PDSP and Pharos datasets were filtered separately, after which they were merged.

3) A total of 21 features of five different kinds (Circular , Path-based, Substructure keys, RDKit descriptors , and VolSurf+) were generated for each of the 40 subcellular locations.

4) An Evaluation metrics is made and the best models were selected based on Cohen's Kappa scores.

5) Based on the evaluation metrics, the machine learning models built consist of 19 different fingerprints-based features for 40 different subcellular locations using 28 different ML classifiers.

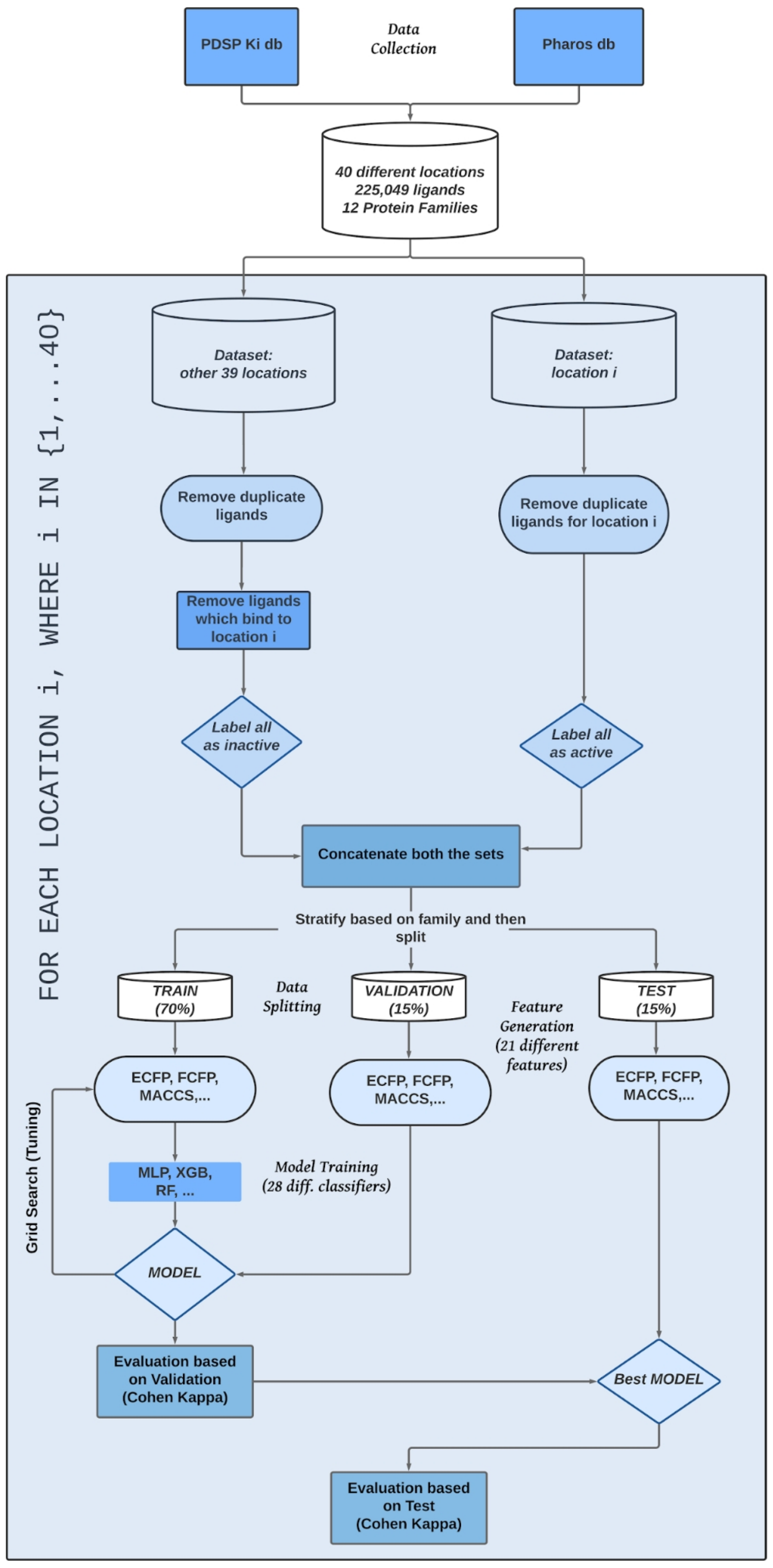
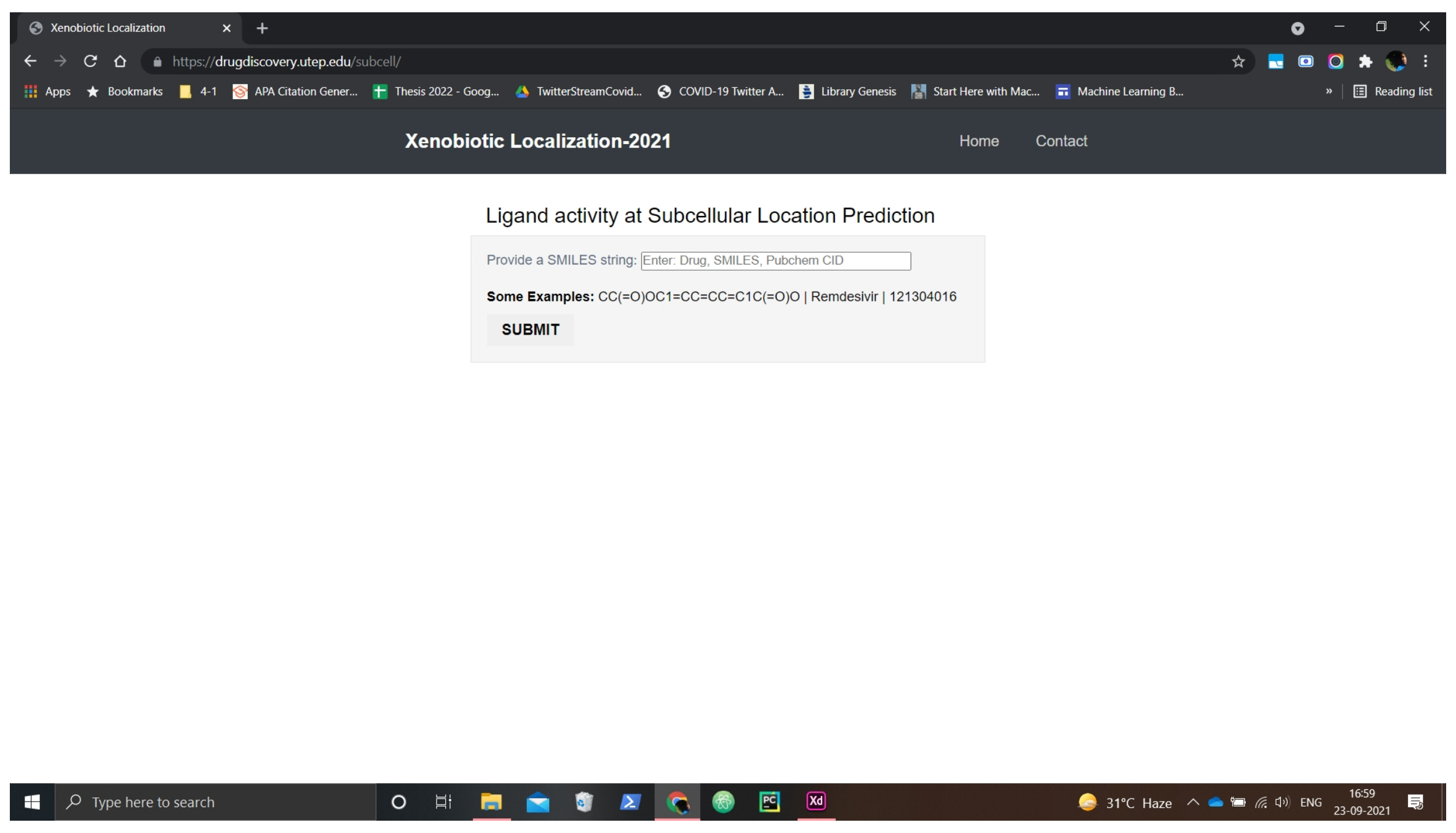

PDSP Ki db
Data Collection
Pharos db
40 different locations
225,049 ligands
12 Protein Families
FOR EACH LOCATION i, WHERE i IN {1,...40}
Dataset: other 39 locations
Dataset: location i
Remove duplicate ligands
Remove duplicate ligands for location i
Remove ligands which bind to location i
Label all as inactive
Label all as active
Concatenate both the sets
Stratify based on family and then split
TRAIN (70%)
Data Splitting
VALIDATION (15%)
TEST (15%)
Feature Generation (21 different features)
ECFP, FCFP, MACCS,...
ECFP, FCFP, MACCS,...
ECFP, FCFP, MACCS,...
Grid Search (Tuning)
MLP, XGB, RF, ...
Model Training (28 diff. classifiers)
MODEL
Evaluation based on Validation (Cohen Kappa)
Best MODEL
Evaluation based on Test (Cohen Kappa)

Figure 1: A schematic of the workflow for model development.

Results



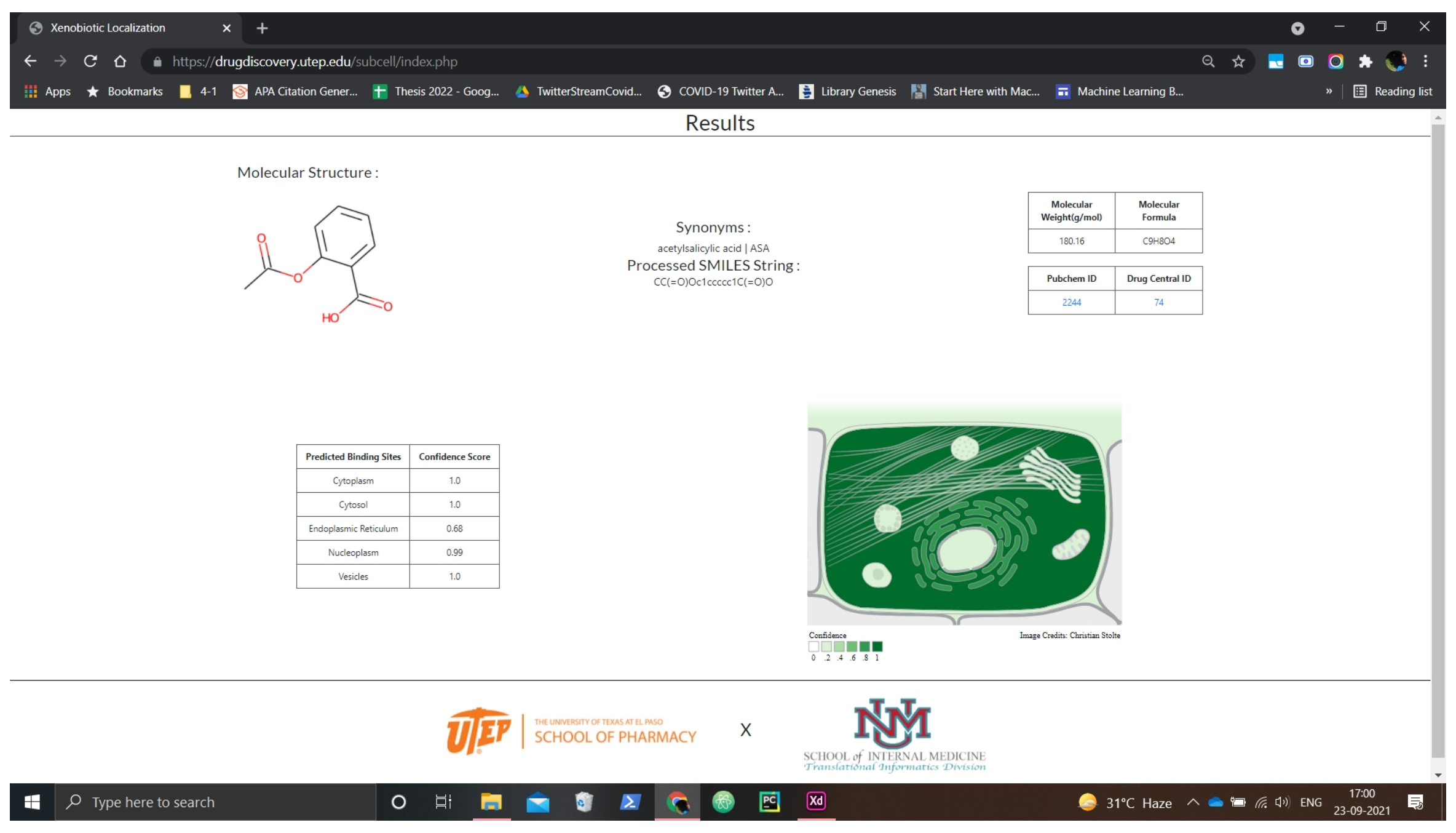
Xenobiotic Localization-2021

Ligand activity at Subcellular Location Prediction

Provide a SMILES string: Enter: Drug, SMILES, Pubchem CID

Some Examples: CC(=O)OC1=CC=CC=C1C(=O)O | Remdesivir | 121304016

Figure 2: The portal predicts the binding site for Drug Name, SMILE or Pubchem CID



Results

Molecular Structure :

Synonyms :
acetylsalicylic acid | ASA
Processed SMILES String :
CC(=O)Oc1ccccc1C(=O)O

Molecular Weight(g/mol)	Molecular Formula
180.16	C9H8O4

Pubchem ID	Drug Central ID
2244	74

Predicted Binding Sites	Confidence Score
Cytoplasm	1.0
Cytosol	1.0
Endoplasmic Reticulum	0.68
Nucleoplasm	0.99
Vesicles	1.0

Figure 3: The predicted binding site, image of the predicted site along with the ligand structure and other information is shown.

In summary, we have developed some reliable predictive models for localization for 40 different subcellular locations. Manually curated data on protein localization along with data obtained from Pharos and PDSP Ki databases were used. Several features from RDkit and different machine learning algorithms from scikit-learn were used in training the models. The best models were selected based on model's Cohen Kappa on its validation set. MLP showed best results in most cases. Among 21 different molecular features mainly based on RDKit, LECFP6 outperformed others. However, a scaffold analysis revealed that very few scaffolds are likely to be in positive and negative molecules simultaneously. This can potentially bias fingerprint-based strategies. Hence, using other methodologies such as VolSurf+ might be preferable, despite the lower performances in prediction.

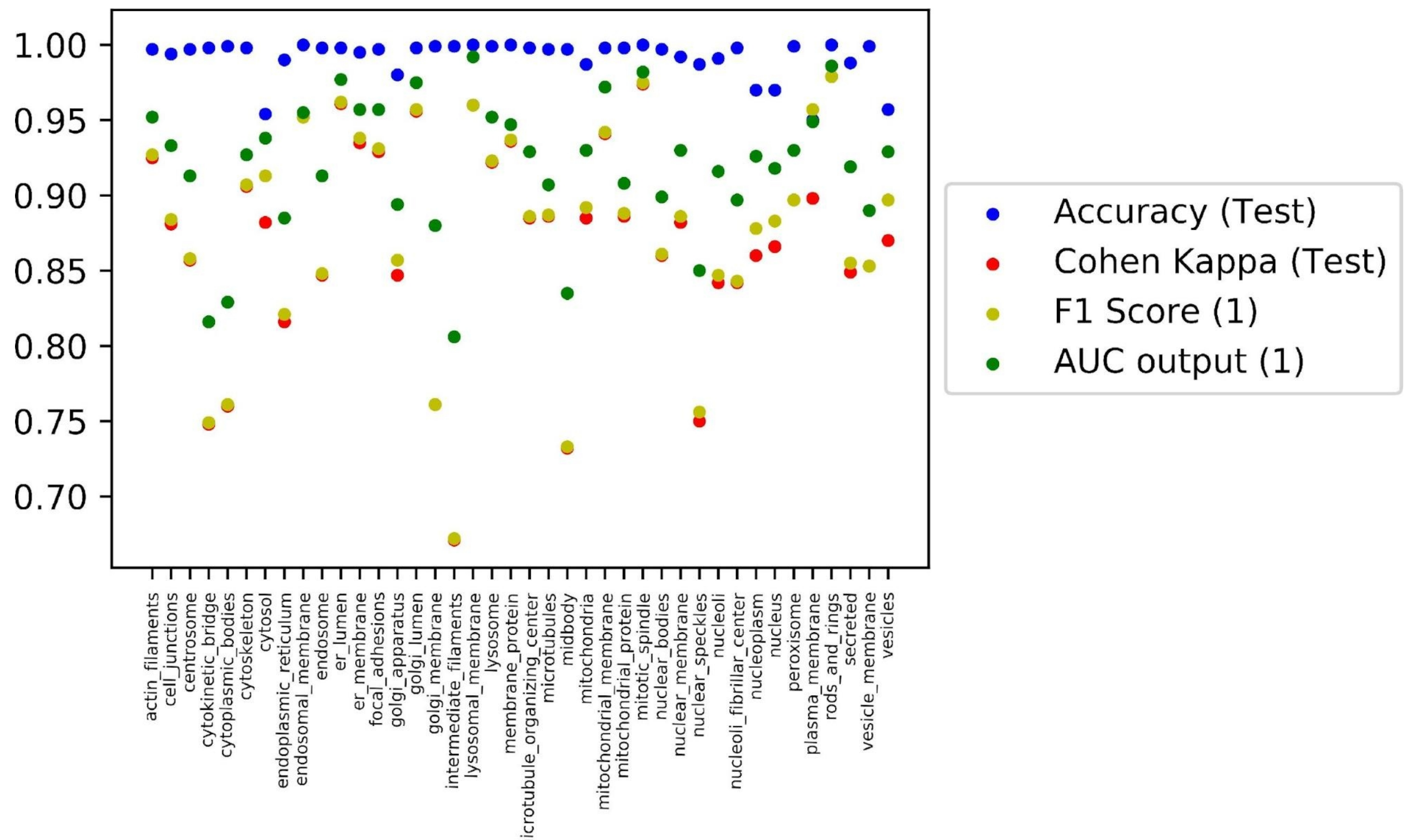


Figure 4: Plot showing the performance of 40 models on the test set.