Software Requirements Specification (SRS)

Project

Smiler

AI-Powered Protein-Ligand Interaction Predictor

Date: 30/12/2024

Group Members

Hina Ali (FA21-BSI-020)

Syeda Javeria Zahid (FA21-BSI-045)

Introduction

Purpose

The purpose of this document is to outline the requirements for the development of *Smiler*, an AI-powered tool designed to predict protein-ligand interactions. The tool will allow researchers to input molecular structures (SMILES strings) and determine their likelihood of binding to a specific target protein.

Scope

Smiler will be a desktop application for:

- Predicting bioactivity of molecules for drug discovery.
- Processing molecular data efficiently.
- Providing an easy-to-use interface for researchers.

Definitions, Acronyms, and Abbreviations

SMILES: Simplified Molecular Input Line Entry System – a chemical notation for molecules.

ChEMBL: A database for bioactive molecules and their properties.

GUI: Graphical User Interface – a visual tool for interacting with the application.

ML: Machine Learning.

General Description

Product Perspective

Smiler is a standalone tool integrating molecular data processing, machine learning, and visualization capabilities for bioactivity predictions. It will use ChEMBL data for training predictive models.

Product Features

Data Input: Upload SMILES strings directly or through CSV file.

Feature Generation: Compute molecular properties for the molecules such as LogP.

Bioactivity Prediction: Predict activity labels (active/inactive) for target proteins.

Visualization: Display molecular structures and prediction confidence scores.

User Characteristics

Primary Users: Researchers in pharmaceuticals, bioinformatics, and academia.

Technical Requirements: Basic understanding of molecular structures; no programming skills needed.

Constraints

- Dependence on external data quality and availability.
- Limited accuracy with small or biased datasets.
- Must be completed within one month.

System Requirements

Hardware Requirements

Processor: Intel Core i5 or equivalent

Memory (RAM): 8 GB (minimum), 16 GB (recommended) for handling large datasets

Storage:

- Minimum: 10 GB of free disk space
- Recommended: 20 GB for additional libraries and temporary files

Operating System: Windows 10/11

Software Requirements

Programming Environment:

- Python 3.8 or later

Libraries and Tools:

- RDKit (for molecular feature extraction)
- scikit-learn (for machine learning models)
- PyTorch or TensorFlow (if deep learning models are used)
- pandas, numpy, matplotlib, and seaborn (for data handling and visualization)

API Tools:

- ChEMBL API or downloaded datasets for molecular and bioactivity data

GUI/Visualization Libraries:

- tkinter (for building the interface) .

Network Requirements

Internet access is required for accessing ChEMBL API or downloading datasets and Installing/updating Python libraries

Functional Requirements

Data Handling

- The system should retrieve SMILES strings and activity data from ChEMBL or user uploads.
- It must clean data by removing duplicates and invalid SMILES strings.
- It should make sure that the data uploaded is valid and provide specific error messages.

Molecular Feature Extraction

- Compute molecular properties (e.g., molecular weight, LogP).
- Generate molecular fingerprints using RDKit or similar tools.

Model Training and Prediction

- Use machine learning models (Random Forest) for prediction.
- Support classification (active/inactive) and regression (IC50, Ki values).

Visualization and GUI

- Display molecular structures alongside prediction results.
- Allow users to interactively explore confidence scores and features.

Interface Requirements

User Interface

- Input panel for typing or uploading SMILES strings.
- Output panel showing predicted results, confidence scores, and molecular structures.
- Option to download prediction results as CSV.

System Interface

- Integration with the ChEMBL API for retrieving molecular data.
- Utilize RDKit for descriptor calculations and structure visualizations.

Performance Requirements

- Predictions for a single molecule must be processed in a short time.
- Support batch processing for upto 500 molecules.
- Achieve over 80% accuracy for classification models on test datasets.

Schedule

Phase	Timeline
Requirement Gathering	2 Days
Data Collection and Cleaning	5 Days
Model Development	10 Days
GUI Development	5 Days
Testing and Deployment	5 Days

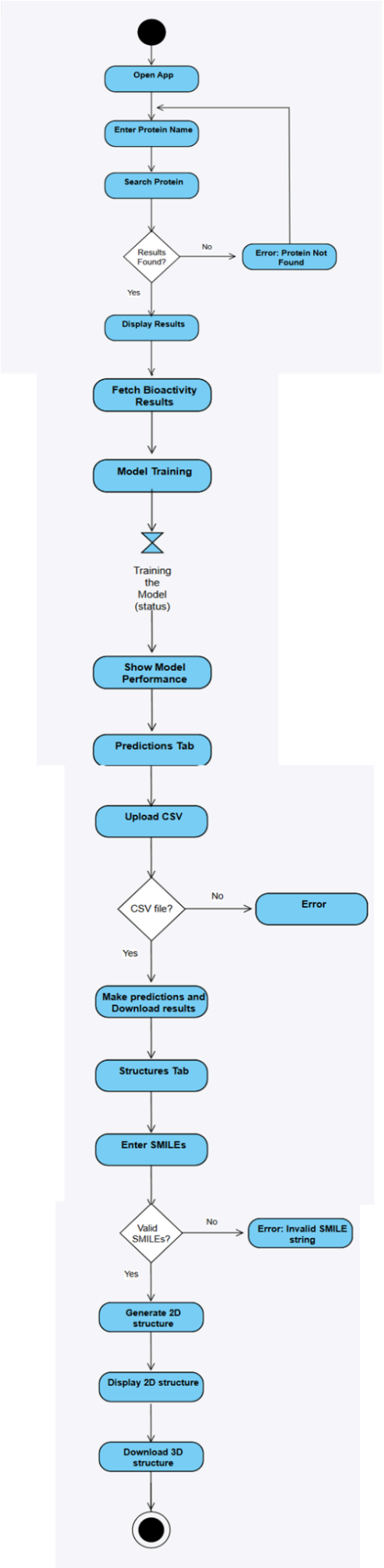
Conclusion

This SRS outlines the plan for creating *Smiler*, a fast, free, and easy-to-use tool for predicting protein-ligand interactions. The tool is designed to be simple and efficient, making it ideal for researchers with limited resources.

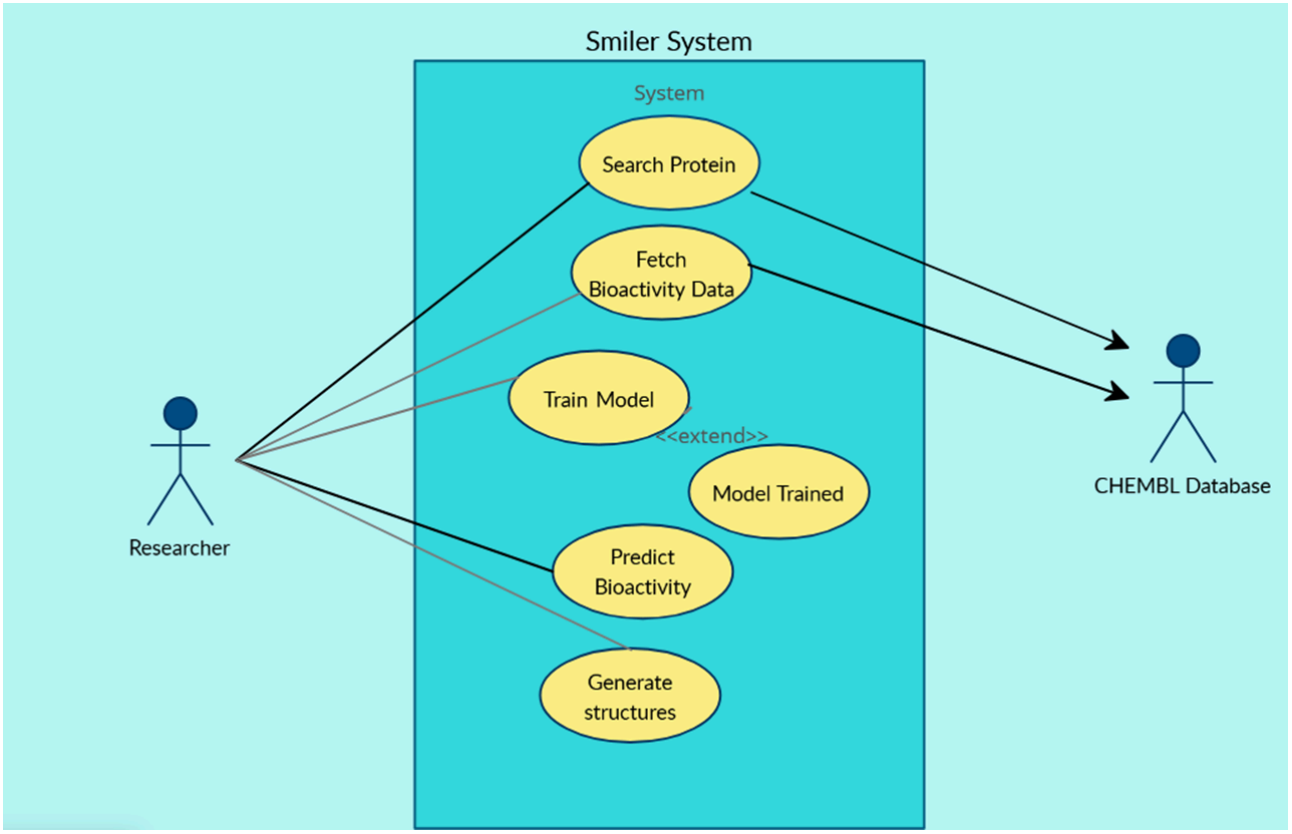
Thanks to built-in error handling and data checks, the system will work smoothly with very few errors. Even with large datasets, it will aim to provide accurate and reliable results as quickly as possible. Using advanced machine learning and tools like RDKit, *Smiler* ensures dependable predictions.

In summary, *Smiler* will help researchers save time and effort, make bioactivity predictions easier, and contribute to advancements in drug discovery and molecular research.

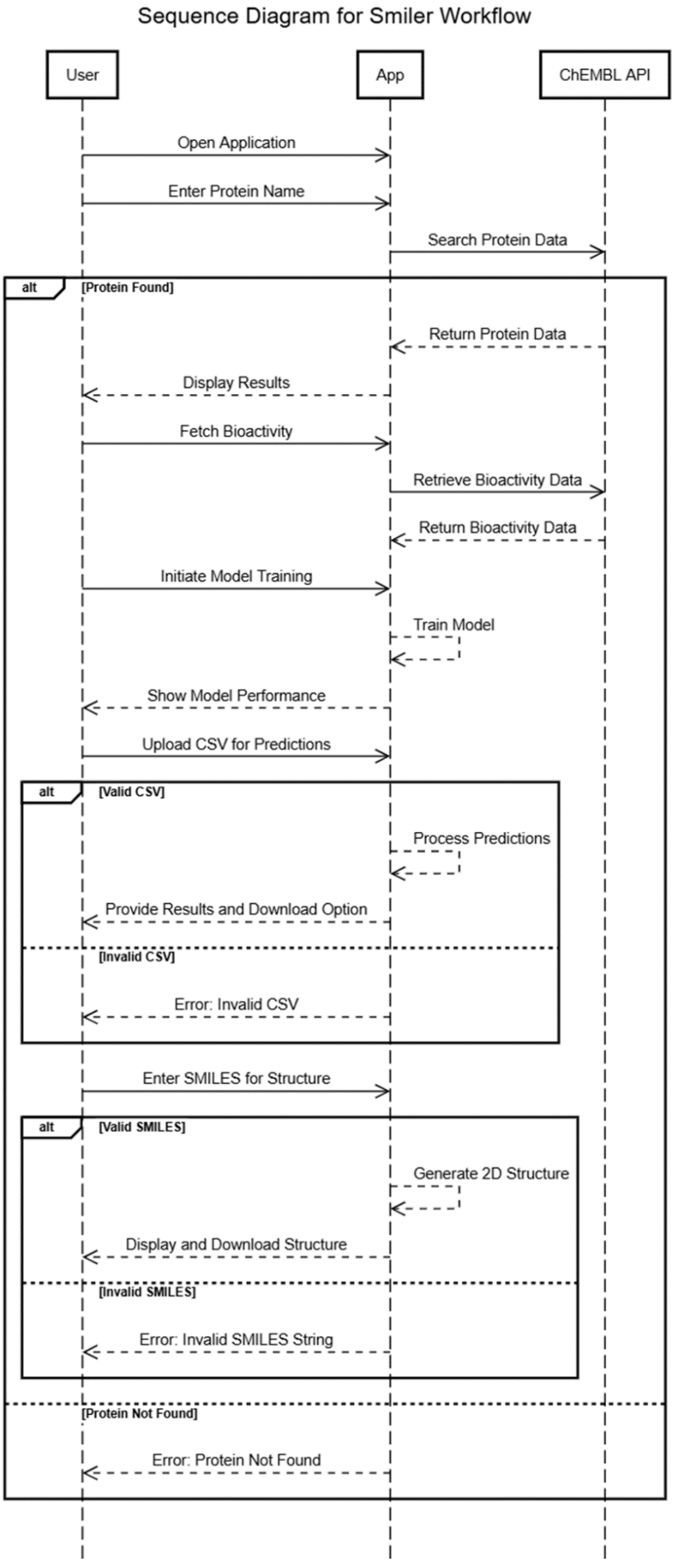
ACTIVITY DIAGRAM



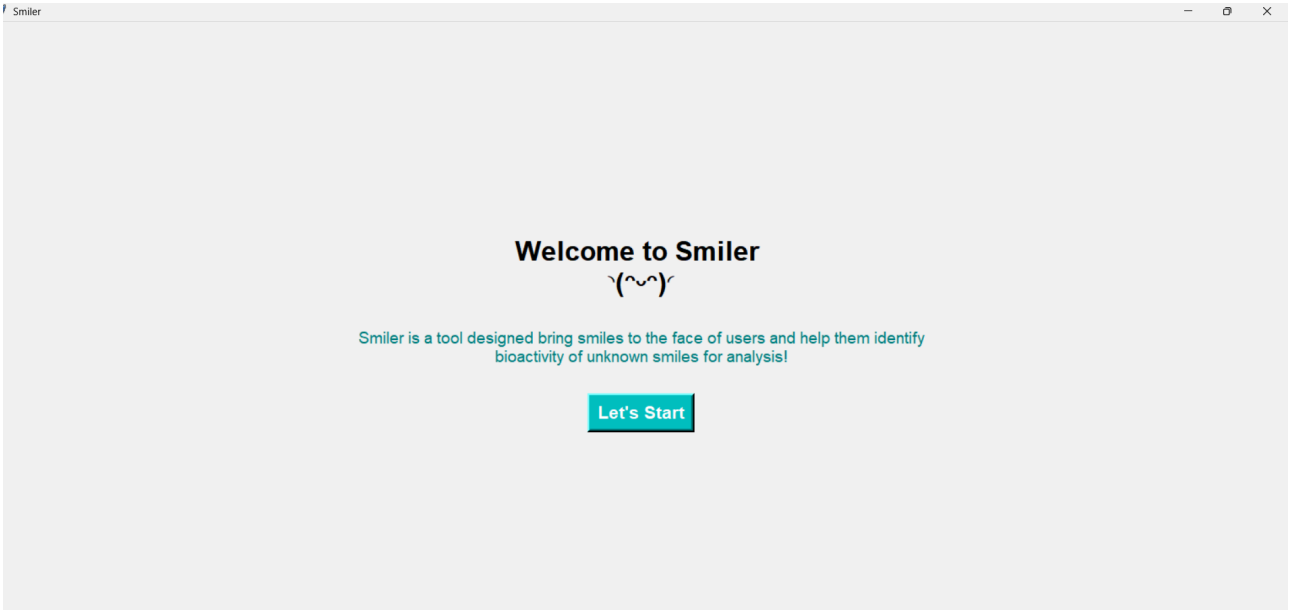
USE CASE DIAGRAM

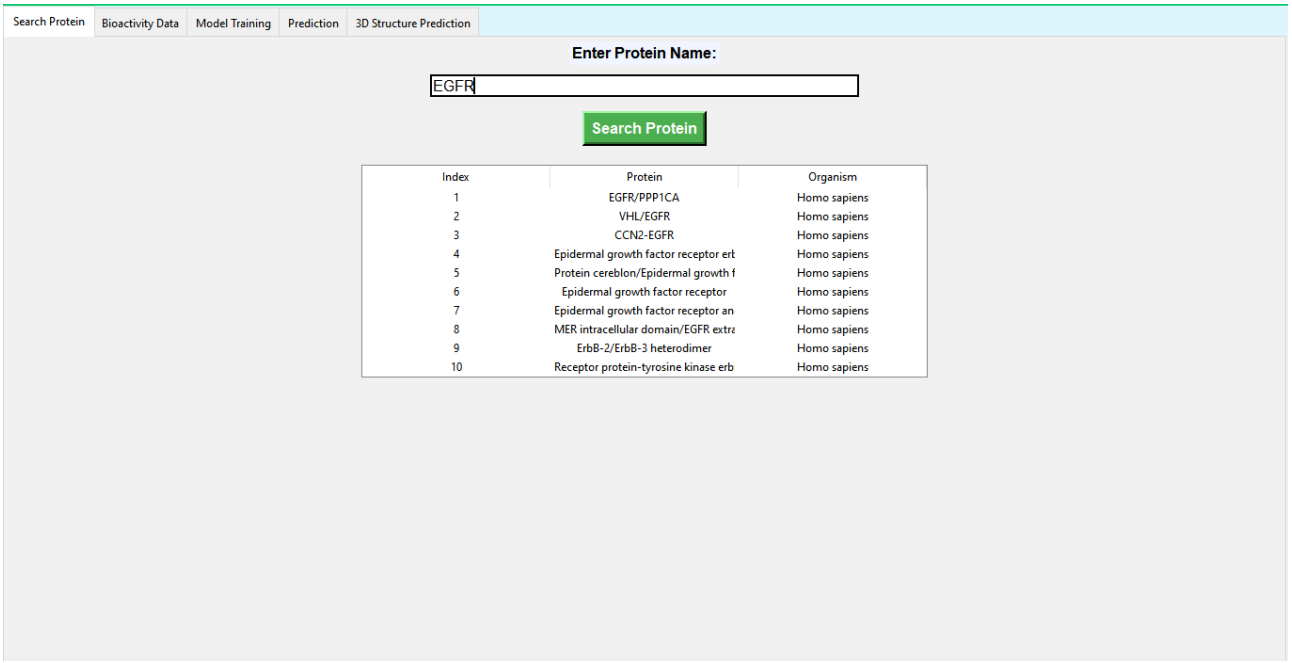


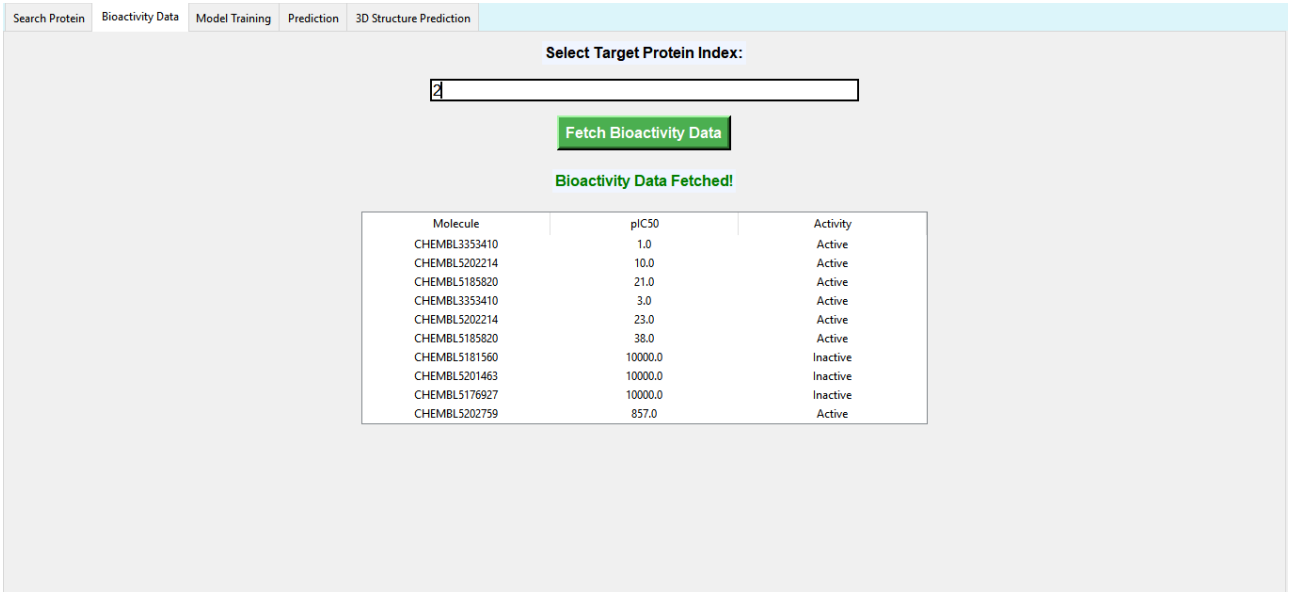
SEQUENCE DIAGRAM

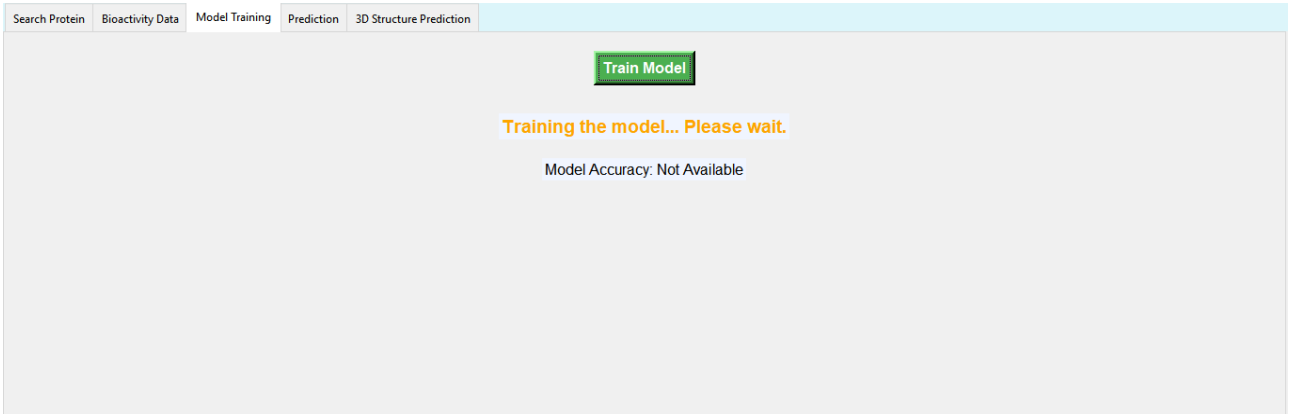


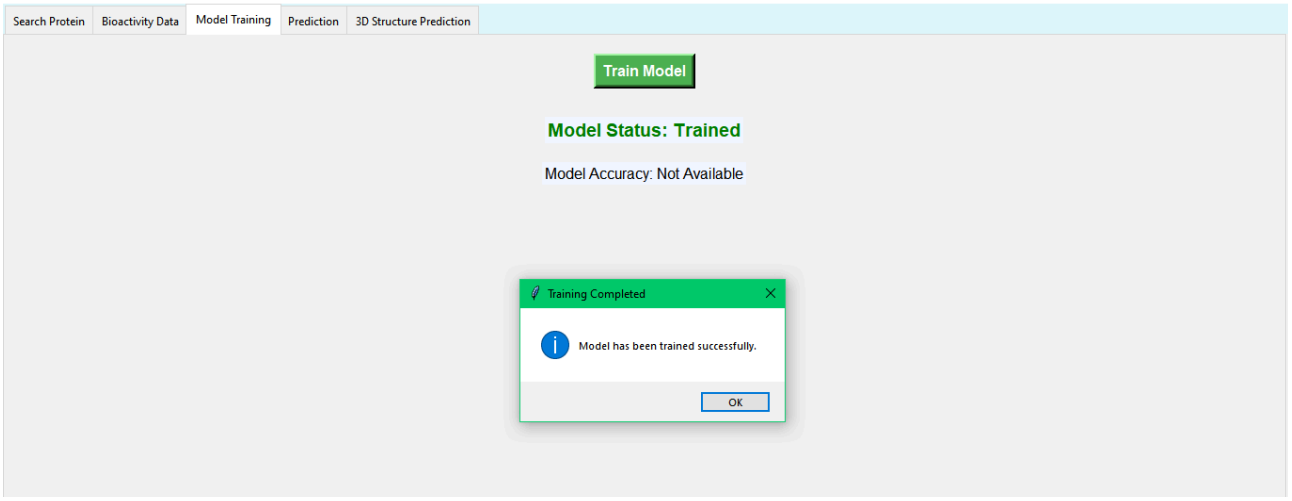
Smiler tool output:

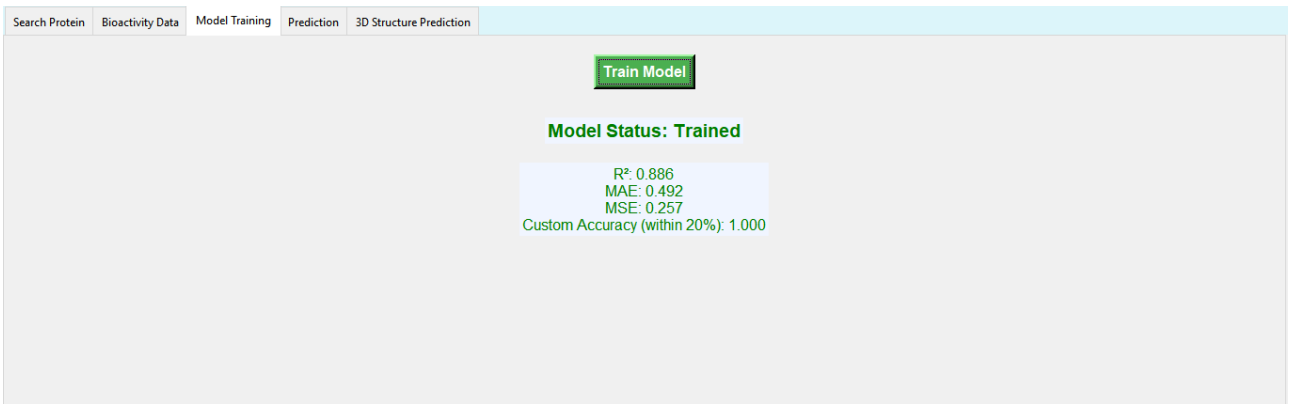


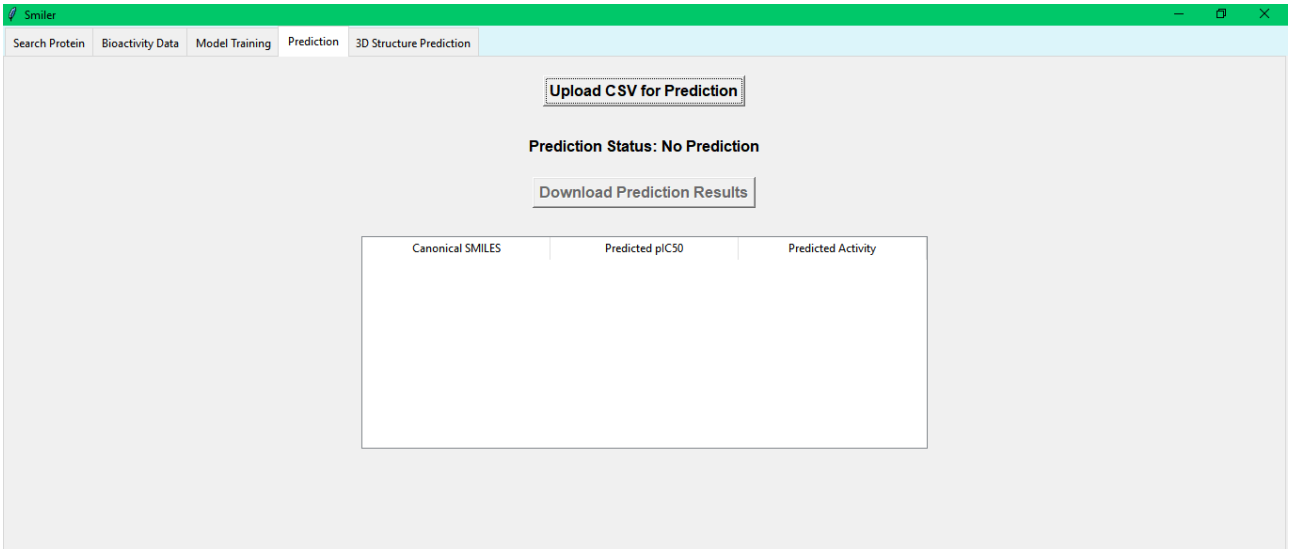


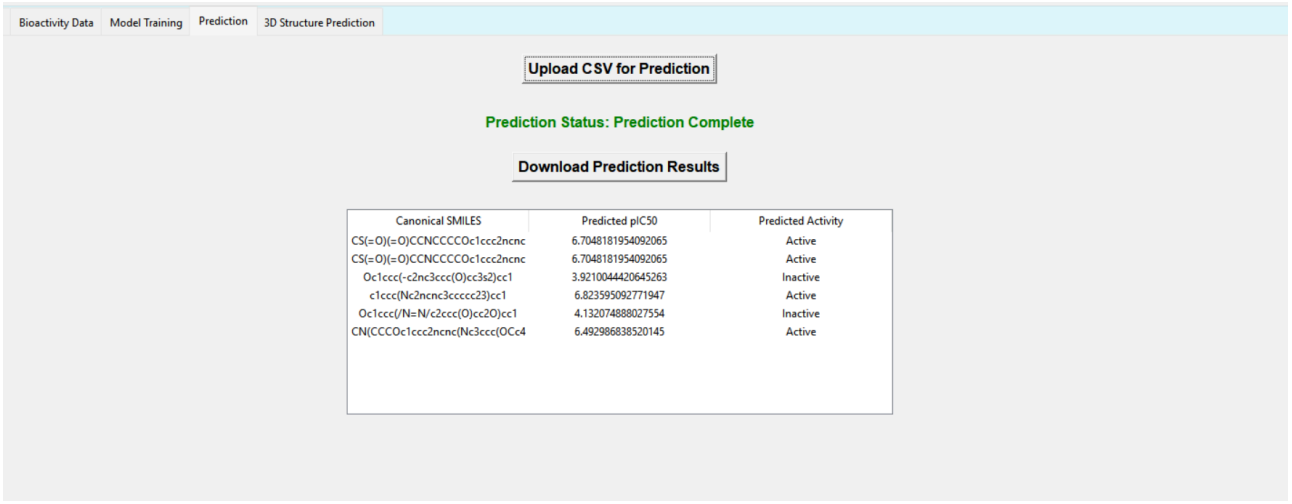

Search Protein
Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Select Target Protein Index:
2
Fetch Bioactivity Data
Bioactivity Data Fetched!
Molecule
pIC50
Activity
CHEMBL3353410
1.0
Active
CHEMBL5202214
10.0
Active
CHEMBL5185820
21.0
Active
CHEMBL3353410
3.0
Active
CHEMBL5202214
23.0
Active
CHEMBL5185820
38.0
Active
CHEMBL5181560
10000.0
Inactive
CHEMBL5201463
10000.0
Inactive
CHEMBL5176927
10000.0
Inactive
CHEMBL5202759
857.0
Active


Search Protein
Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Train Model
Training the model... Please wait.
Model Accuracy: Not Available

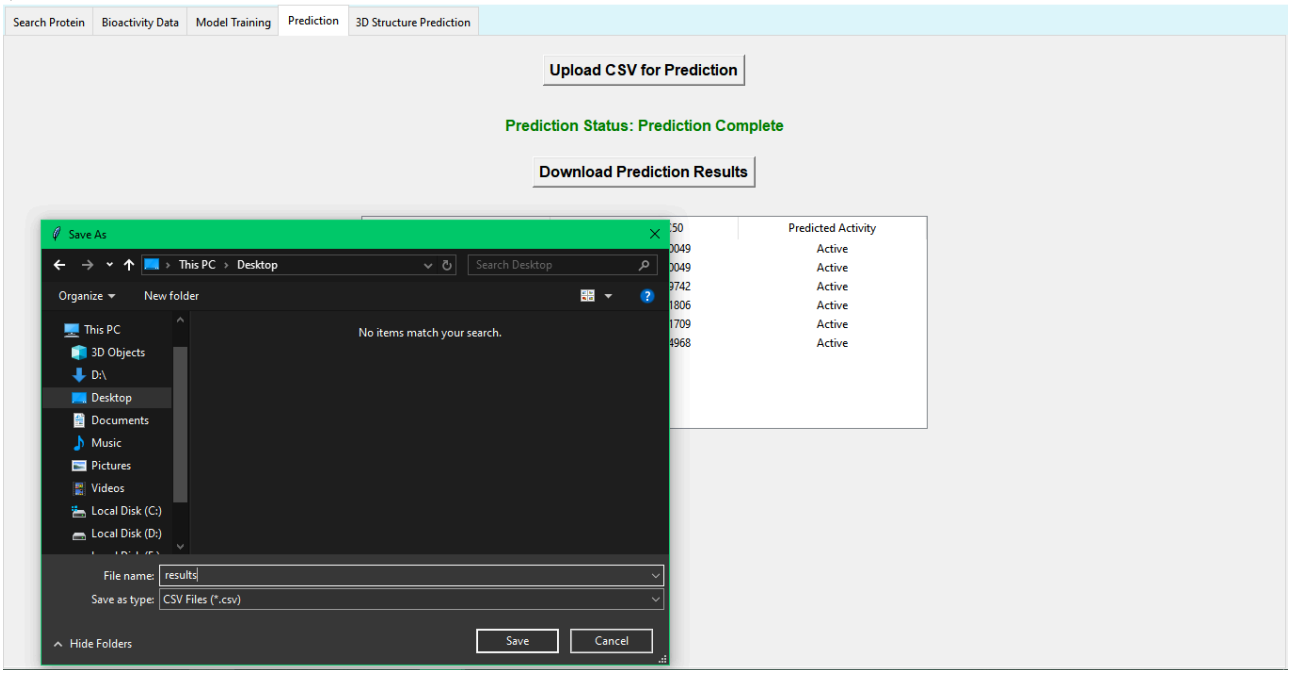

Search Protein
Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Train Model
Model Status: Trained
Model Accuracy: Not Available
Training Completed
Model has been trained successfully.
OK

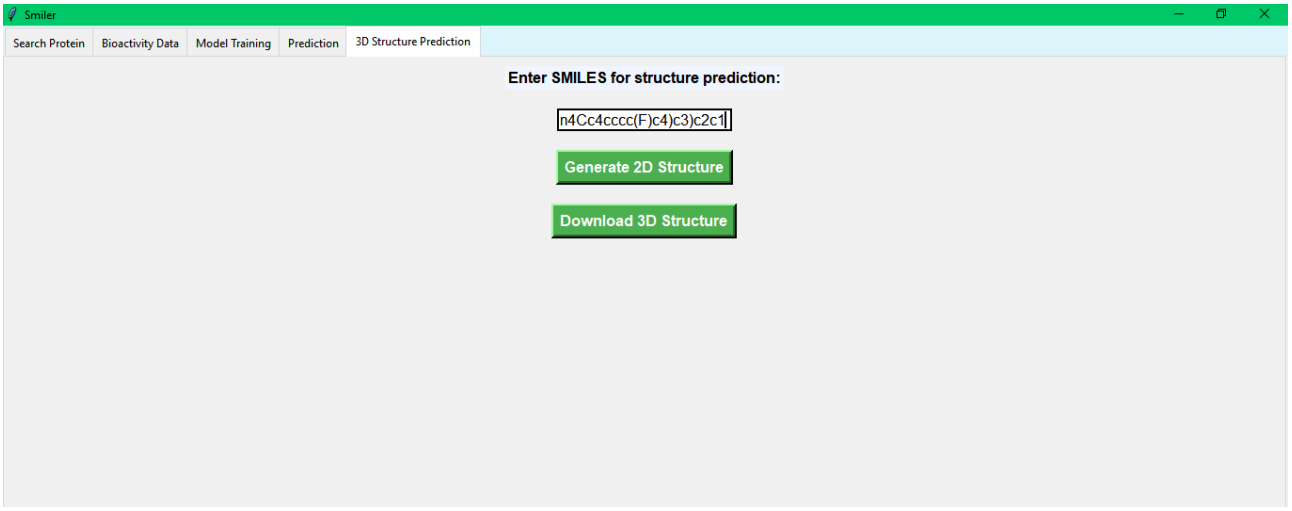

Search Protein
Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Train Model
Model Status: Trained
R²: 0.886
MAE: 0.492
MSE: 0.257
Custom Accuracy (within 20%): 1.000

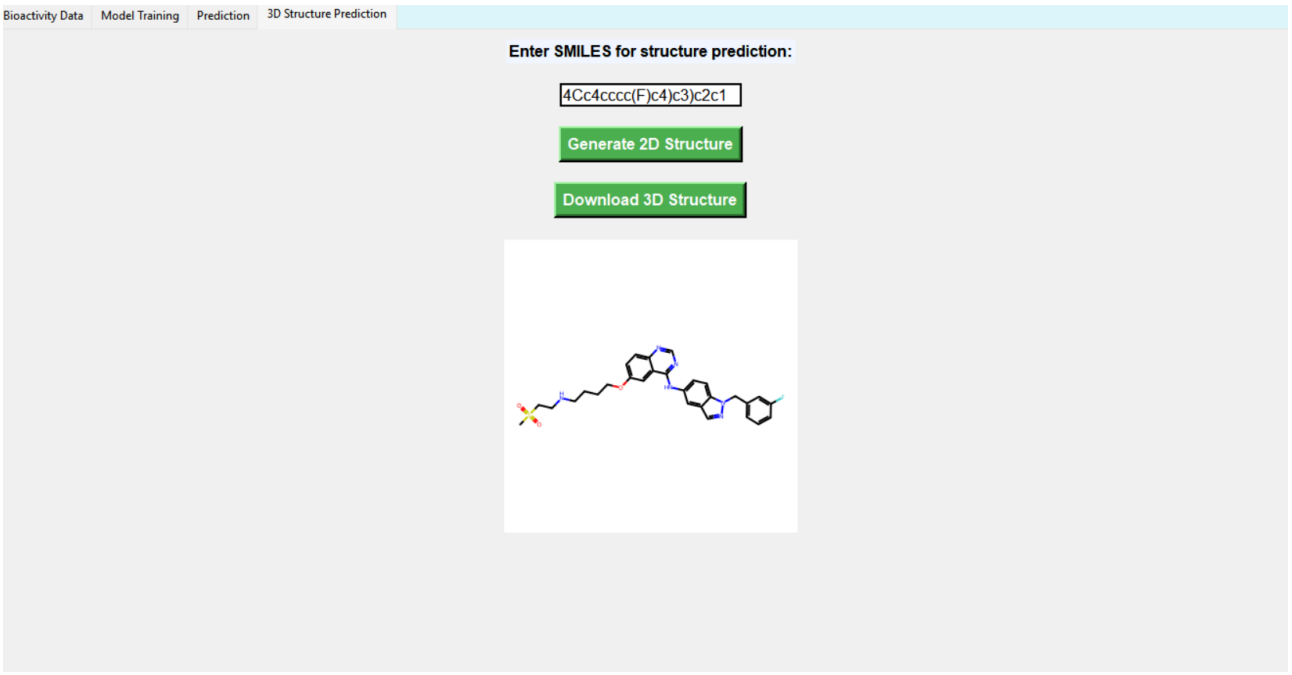


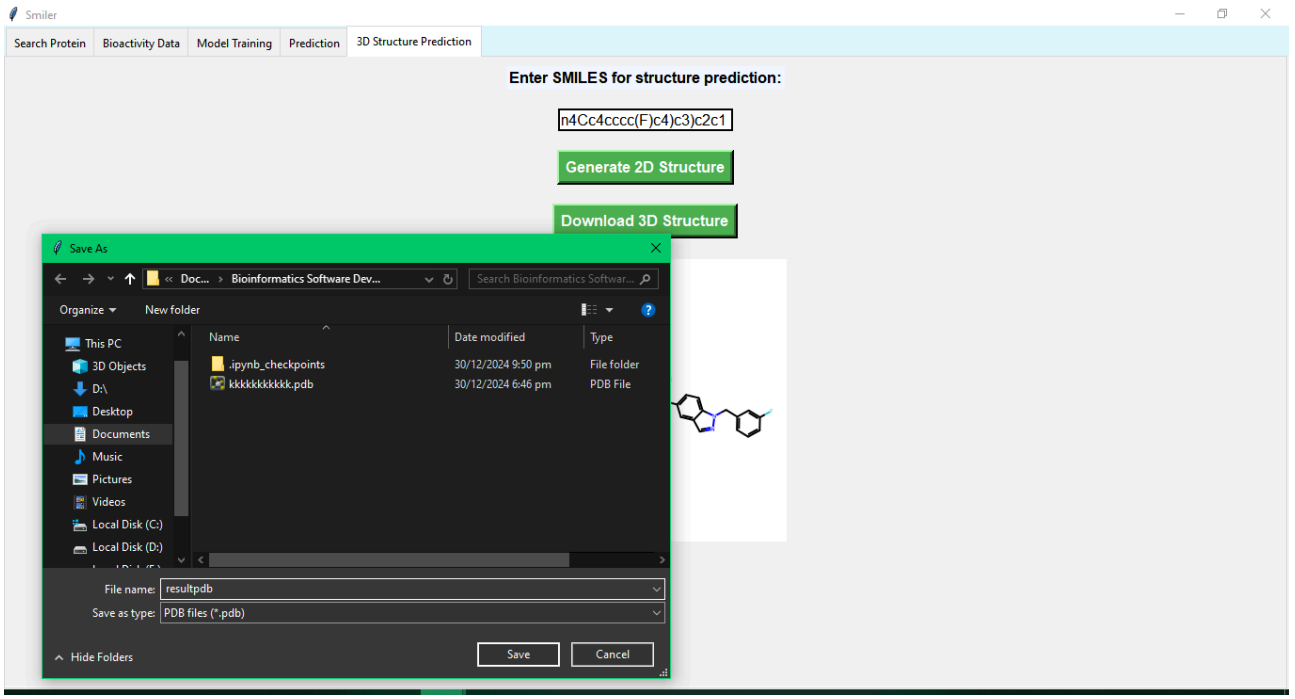


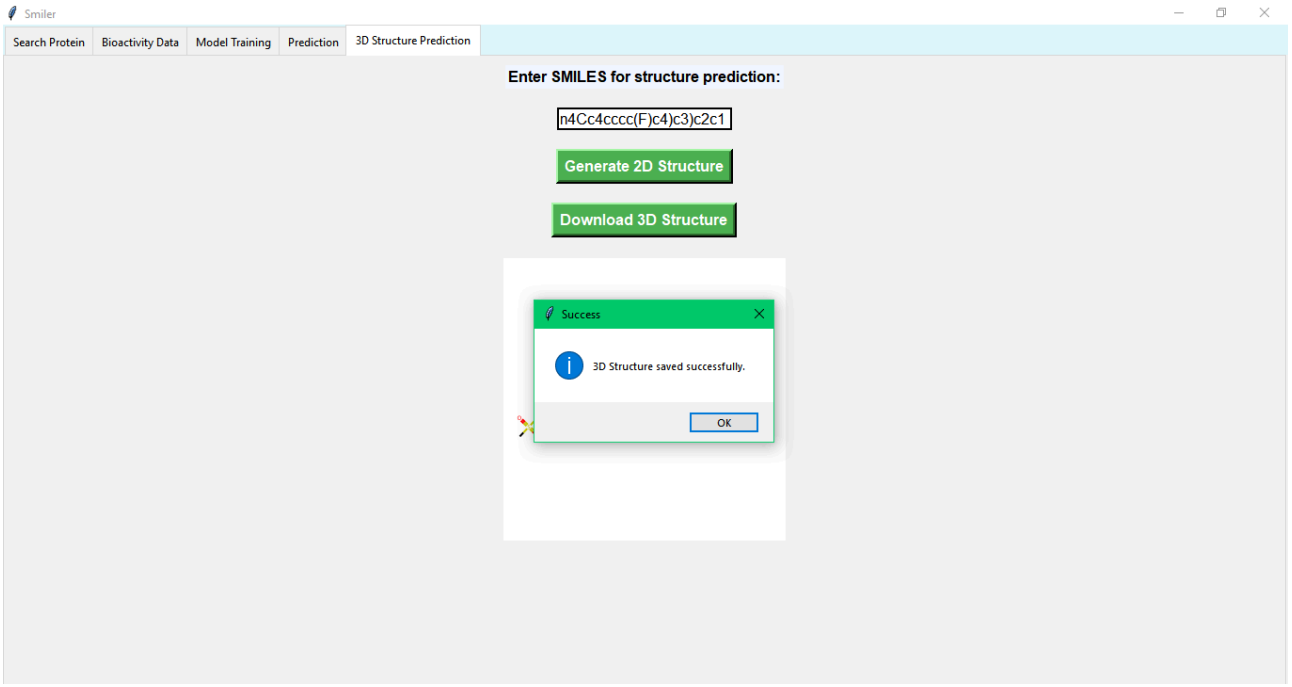
Download the results



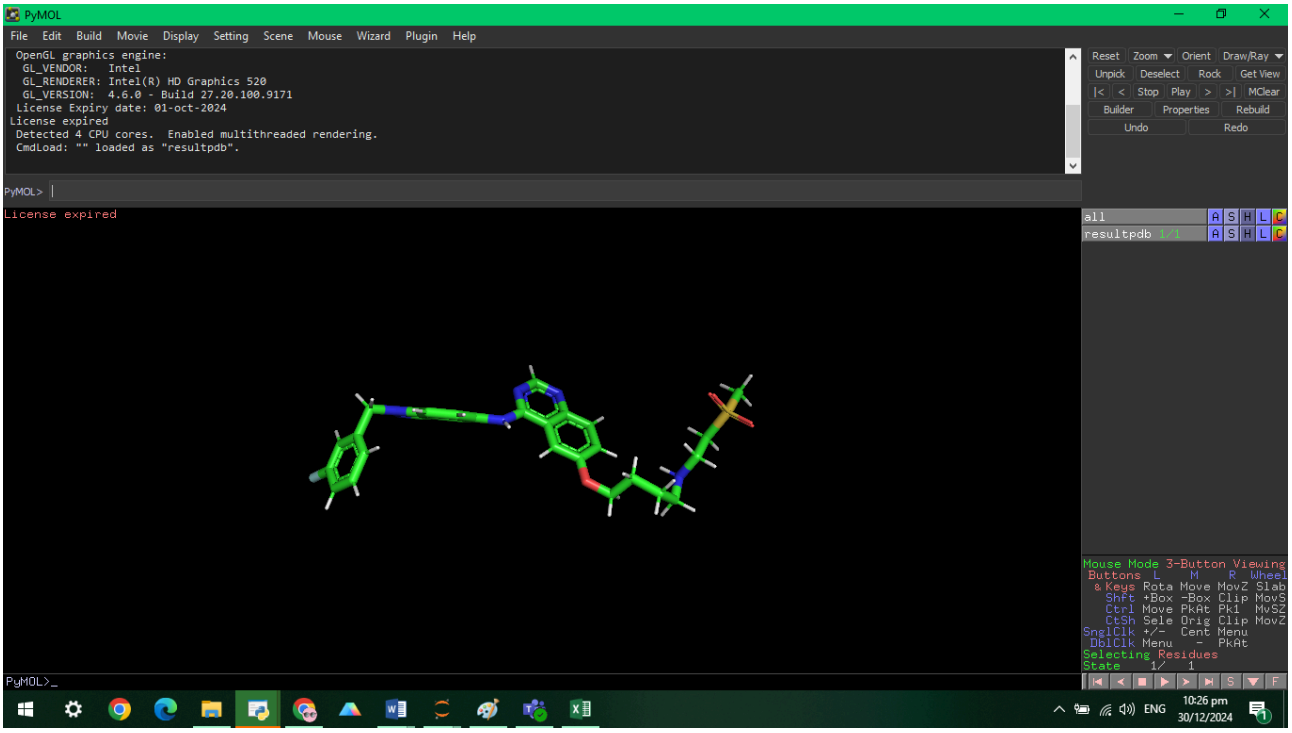

Smiler
Search Protein
Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Enter SMILES for structure prediction:
n4Cc4cccc(F)c4)c3)c2c1
Generate 2D Structure
Download 3D Structure


Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Enter SMILES for structure prediction:
4Cc4cccc(F)c4)c3)c2c1
Generate 2D Structure
Download 3D Structure


Smiler
Search Protein
Bioactivity Data
Model Training
Prediction
3D Structure Prediction
Enter SMILES for structure prediction:
n4Cc4cccc(F)c4)c3)c2c1
Generate 2D Structure
Download 3D Structure
Save As
Doc... › Bioinformatics Software Dev...
Search Bioinformatics Softwar...
Organize
New folder
This PC
3D Objects
D:\
Desktop
Documents
Music
Pictures
Videos
Local Disk (C:)
Local Disk (D:)
Name
Date modified
Type
.ipynb_checkpoints
30/12/2024 9:50 pm
File folder
kkkkkkkkkkk.pdb
30/12/2024 6:46 pm
PDB File
File name:
resultpdb
Save as type:
PDB files (*.pdb)
Hide Folders
Save
Cancel

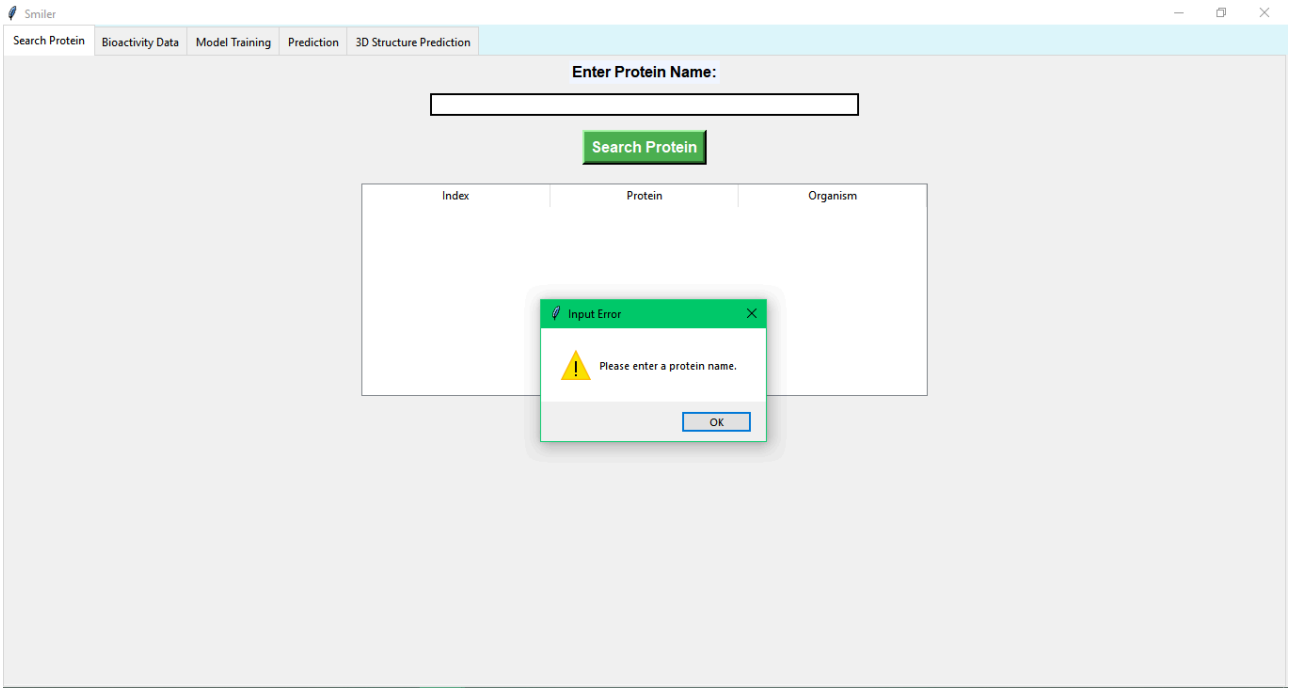


View structure in pymol

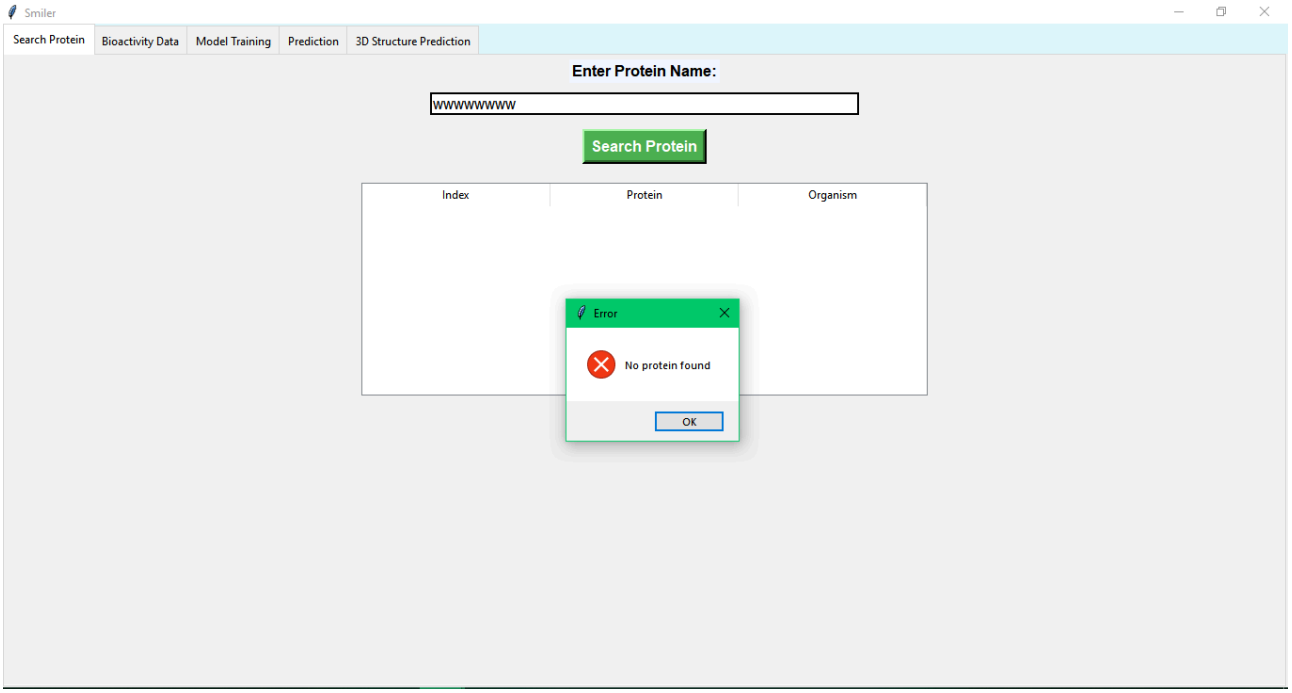


ERROR HANDLING:

In case of empty



Incase of incorrect protein



In case training has an issue error message will appear