# PREDICTIVE STUDY ON DRUG BIOACTIVITY IN INHIBITING ACETYLCHOLINESTERASE FOR ALZHEIMER'S DISEASE

## PB PROJECT


**PRESENTED BY:** GROUP 5

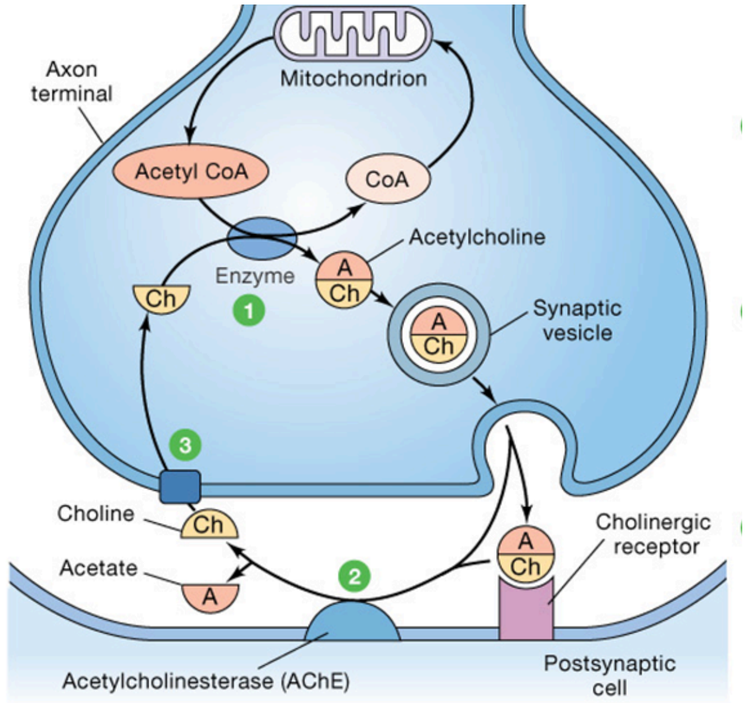
**GROUP MEMBERS:** ANISHA BOPARDIKAR, GRISHMA BELLANI, RIYA GUPTA, SHREYANSH SRIVASTAV, SHUBHAM DWIVEDI, VIMANSH MAHAJAN

# BACKGROUND OF THE STUDY

- Alzheimer's disease involves cognitive decline from reduced acetylcholine levels.
- One of the key factors contributing to cognitive decline in AD is the deficiency of the neurotransmitter acetylcholine, which plays a crucial role in memory and cognitive function.

# PROBLEM STATEMENT

**Despite advancements, Alzheimer's disease lacks effective treatments. Our project addresses this by assessing various drugs' efficacy in modulating Acetylcholinesterase (AChE) levels, a critical target for Alzheimer's treatment, aiming to identify potent therapeutic candidates.**



# SOLUTION

**By modulating AChE activity, the levels of acetylcholine in the brain can be increased, potentially improving cognitive function and alleviating some symptoms of Alzheimer's disease.**

- **Quantitative Structure-Activity Relationship (QSAR)**
  - **Develop a QSAR model to predict the bioactivity of compounds towards inhibiting AChE based on their molecular descriptors.**

- **Machine Learning Algorithms**
  - **Utilize machine learning algorithms, such as Random Forest to build predictive models for AChE inhibition.**

# FEASIBILITY OF OUR SOLUTION

LINK TO THE STUDY: NCBI WEBSITE

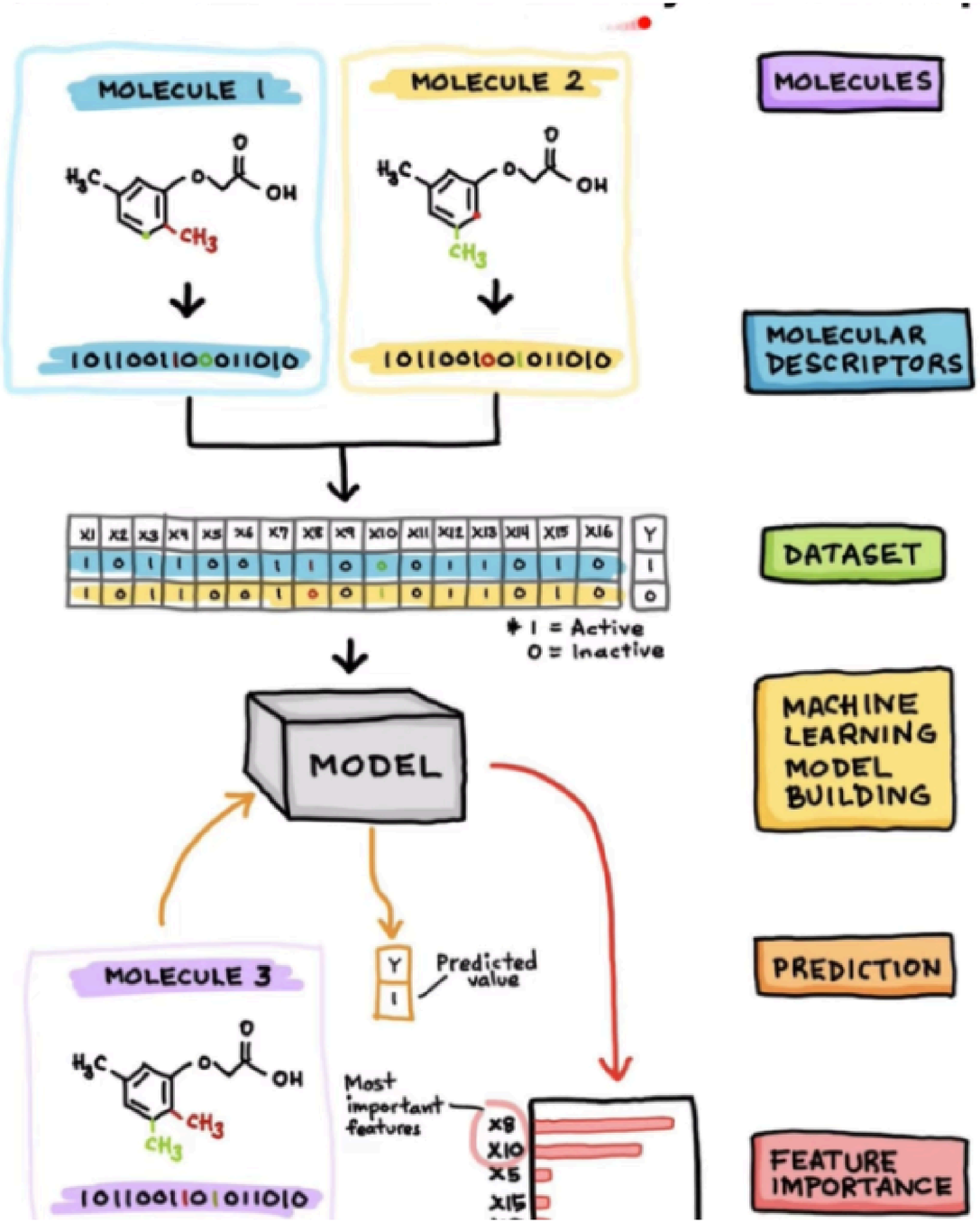
**Several researches including the one we have mentioned below by NCBI indicate that our solution will indeed be correct and feasible in finding the best drug for Alzheimer's out of all those available in the market today.**

Cholinesterase inhibitors

The cholinergic hypothesis of AD concludes that cholinergic systems in the basal forebrain are affected early in the disease process, including loss of acetylcholine neurons, loss of enzymatic function for acetylcholine synthesis and degradation, resulting in memory loss and deterioration of other cognitive and noncognitive functions such as neuropsychiatric symptoms [Bartus *et al.* 1982; Cummings and Back, 1998]. A strategy to enhance the cholinergic transmission by using CIs to delay the degradation of acetylcholine between the synaptic cleft has been proposed. To date, three CIs are approved for the treatment of mild to moderate AD: donepezil (Pfizer, New York, NY, USA), rivastigmine (Novartis, Basel, Switzerland) and galantamine (Janssen, Beerse, Belgium) [Farlow,

# QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP

**Utilize computational methods, such as QSAR modeling, to predict the bioactivity of compounds towards inhibiting AChE. This approach can help identify potential drug candidates and reduce the number of compounds that need to be experimentally tested.**



# WORKFLOW

## DATA COLLECTION

We have sourced raw data from the **ChEMBL website**, filtering compounds based on specific parameters.
Through classification, we will categorize compounds as either active or intermediate, streamlining the dataset for targeted drug discovery efforts.

## EDA

We will utilize **Lipinski's Rule**, which assesses drug-likeness based on molecular weight, LogP, hydrogen bond donors, and hydrogen bond acceptors, to sort compounds for drug candidacy.

## CALCULATION AND DATASET PREPARATION

We will compute molecular descriptors, providing quantitative characterizations of compounds in the dataset, to prepare data for subsequent model construction.

## BUILDING A REGRESSION MODEL

We will construct a **regression model** to predict the activity of acetylcholinesterase inhibitors based on molecular descriptors.

# HOW FAR WE'VE COME IN DATA COLLECTION..

## A. DATA RETREIVAL

**We obtained the raw data from the ChEMBL database for the acetylcholinesterase enzyme, specifically retrieving bioactivity data related to Human Acetylcholinesterase.**

| | activity_comment | activity_id | activity_properties | assay_chembl_id | assay_description | assay_type | bao_endpoint | bao_format |
|---|---|---|---|---|---|---|---|---|
| **0** | None | 33969 | [] | CHEMBL643384 | Inhibitory concentration against acetylcholine... | B | BAO_0000190 | BAO_0000357 |
| **1** | None | 37563 | [] | CHEMBL643384 | Inhibitory concentration against acetylcholine... | B | BAO_0000190 | BAO_0000357 |
| **2** | None | 37565 | [] | CHEMBL643384 | Inhibitory concentration against acetylcholine... | B | BAO_0000190 | BAO_0000357 |
| **3** | None | 38902 | [] | CHEMBL643384 | Inhibitory concentration against acetylcholine... | B | BAO_0000190 | BAO_0000357 |
| **4** | None | 41170 | [] | CHEMBL643384 | Inhibitory concentration against acetylcholine... | B | BAO_0000190 | BAO_0000357 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **7021** | None | 18798886 | [] | CHEMBL4274263 | Inhibition of human erythrocyte AChE using ace... | B | BAO_0000190 | BAO_0000357 |
| **7022** | None | 18798887 | [] | CHEMBL4274263 | Inhibition of human | B | BAO_0000190 | BAO_0000357 |

# HOW FAR WE'VE COME IN DATA COLLECTION..

## B. DATA FILTERING

**Upon obtaining the raw data, we filtered it based on specific parameters such as standard_value, standard_type (IC50), and canonical_smiles. This ensured that the dataset contained relevant and usable information for our analysis.**

| | molecule_chembl_id | canonical_smiles | standard_value |
|---|---|---|---|
| **0** | CHEMBL133897 | CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1 | 750.0 |
| **1** | CHEMBL336398 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1 | 100.0 |
| **2** | CHEMBL131588 | CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1 | 50000.0 |
| **3** | CHEMBL130628 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F | 300.0 |
| **4** | CHEMBL130478 | CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C | 800.0 |
| **...** | ... | ... | ... |
| **7019** | CHEMBL4293155 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccccc3Cl)CC... | 2440.0 |
| **7020** | CHEMBL4282558 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3cccc(Cl)c3)... | 2540.0 |
| **7021** | CHEMBL4281727 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(Br)cc3)... | 3810.0 |
| **7022** | CHEMBL4292349 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3cccc([N+](=... | 3460.0 |
| **7023** | CHEMBL4278260 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(C#N)cc3... | 2780.0 |

# HOW FAR WE'VE COME IN DATA COLLECTION..

## C. BIOACTIVITY CLASSIFICATION

**Compounds were categorized based on IC50 values:**

- **<1000 nM as active, indicating strong inhibitory activity;**
- **>10,000 nM as inactive, suggesting minimal inhibition;**
- **1,000-10,000 nM as intermediate, showing moderate inhibitory potential.**

| | molecule_chembl_id | canonical_smiles | standard_value | class |
|---|---|---|---|---|
| **0** | CHEMBL133897 | CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1 | 750.0 | active |
| **1** | CHEMBL336398 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1 | 100.0 | active |
| **2** | CHEMBL131588 | CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1 | 50000.0 | inactive |
| **3** | CHEMBL130628 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F | 300.0 | active |
| **4** | CHEMBL130478 | CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C | 800.0 | active |
| **...** | ... | ... | ... | ... |
| **4690** | CHEMBL4293155 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccccc3Cl)CC... | 2440.0 | intermediate |
| **4691** | CHEMBL4282558 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3cccc(Cl)c3)... | 2540.0 | intermediate |
| **4692** | CHEMBL4281727 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(Br)cc3)... | 3810.0 | intermediate |
| **4693** | CHEMBL4292349 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3cccc([N+](=... | 3460.0 | intermediate |
| **4694** | CHEMBL4278260 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(C#N)cc3... | 2780.0 | intermediate |

# EXPLORATORY DATA ANALYSIS

## A. CALCULATION OF LIPINSKI DESCRIPTORS

**Lipinski's Rule, also known as the Rule of Five, evaluates a compound's drug-likeness based on four key factors:**

- **Molecular weight < 500 Dalton**
- **Octanol-water partition coefficient (LogP) < 5**
- **Hydrogen bond donors < 5**
- **Hydrogen bond acceptors < 10**

## COMBINED DATASET WITH LIPINSKI'S FACTORS

| | molecule_chembl_id | canonical_smiles | standard_value | class | MW | LogP | NumHDonors | NumHAcceptors |
|---|---|---|---|---|---|---|---|---|
| **0** | CHEMBL133897 | CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1 | 750.0 | active | 312.325 | 2.80320 | 0.0 | 6.0 |
| **1** | CHEMBL336398 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1 | 100.0 | active | 376.913 | 4.55460 | 0.0 | 5.0 |
| **2** | CHEMBL131588 | CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1 | 50000.0 | inactive | 426.851 | 5.35740 | 0.0 | 5.0 |
| **3** | CHEMBL130628 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F | 300.0 | active | 404.845 | 4.70690 | 0.0 | 5.0 |
| **4** | CHEMBL130478 | CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C | 800.0 | active | 346.334 | 3.09530 | 0.0 | 6.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **4690** | CHEMBL4293155 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccccc3Cl)CC... | 2440.0 | intermediate | 511.150 | 7.07230 | 2.0 | 3.0 |
| **4691** | CHEMBL4282558 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3cccc(Cl)c3)... | 2540.0 | intermediate | 511.150 | 7.07230 | 2.0 | 3.0 |
| **4692** | CHEMBL4281727 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(Br)cc3)... | 3810.0 | intermediate | 555.601 | 7.18140 | 2.0 | 3.0 |
| **4693** | CHEMBL4292349 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3cccc([N+](=... | 3460.0 | intermediate | 521.702 | 6.32710 | 2.0 | 5.0 |
| **4694** | CHEMBL4278260 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(C#N)cc3... | 2780.0 | intermediate | 501.715 | 6.29058 | 2.0 | 4.0 |

4695 rows × 8 columns

# EXPLORATORY DATA ANALYSIS

## B. CONVERT IC50 TO PIC50

- To allow IC50 data to be more uniformly distributed, we will convert IC50 to the negative logarithmic scale which is essentially -log10(IC50).

- Steps for the procedure:

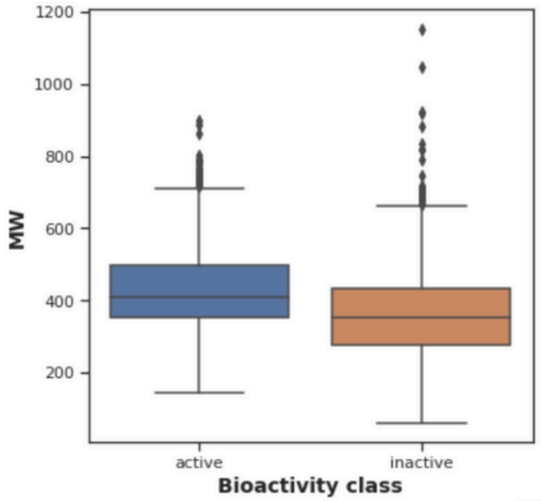
- Take the IC50 values from the standard_value column and converts it from nM to M by multiplying the value by 10 -9
- Take the molar value and apply -log10
- Delete the standard_value column and create a new pIC50 column and delete rows of “intermediate class”.

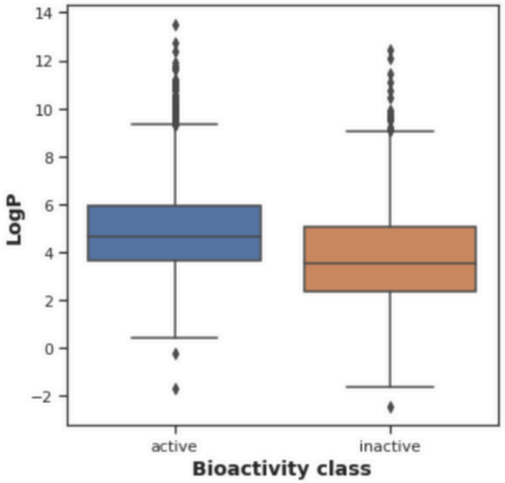
## MODIFIED DATASET WITH PIC50 VALUES

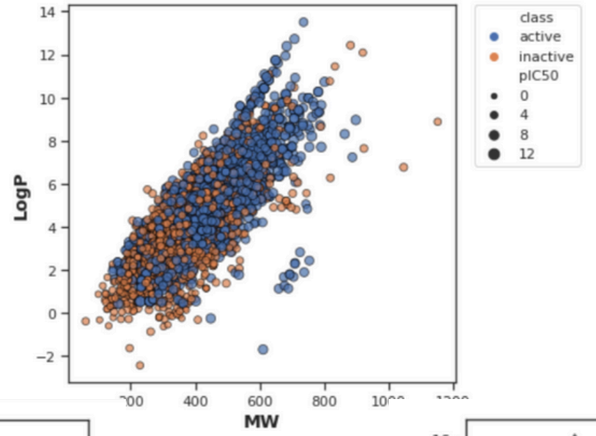
| | molecule_chembl_id | canonical_smiles | class | MW | LogP | NumHDonors | NumHAcceptors | pIC50 |
|---|---|---|---|---|---|---|---|---|
| 0 | CHEMBL133897 | CCOc1nn(-c2cccc(OCc3ccccc3)c2)c(=O)o1 | active | 312.325 | 2.8032 | 0.0 | 6.0 | 6.124939 |
| 1 | CHEMBL336398 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC1CC1 | active | 376.913 | 4.5546 | 0.0 | 5.0 | 7.000000 |
| 2 | CHEMBL131588 | CN(C(=O)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F)c1ccccc1 | inactive | 426.851 | 5.3574 | 0.0 | 5.0 | 4.301030 |
| 3 | CHEMBL130628 | O=C(N1CCCCC1)n1nc(-c2ccc(Cl)cc2)nc1SCC(F)(F)F | active | 404.845 | 4.7069 | 0.0 | 5.0 | 6.522879 |
| 4 | CHEMBL130478 | CSc1nc(-c2ccc(OC(F)(F)F)cc2)nn1C(=O)N(C)C | active | 346.334 | 3.0953 | 0.0 | 6.0 | 6.096910 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 4675 | CHEMBL4284261 | CCN(C)Cc1cc(N)ccc1O.Cl.Cl | inactive | 180.251 | 1.4261 | 2.0 | 3.0 | 3.015428 |
| 4676 | CHEMBL4276921 | CN(C)Cc1cc(N)ccc1O.Cl.Cl | inactive | 166.224 | 1.0360 | 2.0 | 3.0 | 2.813467 |
| 4677 | CHEMBL4292574 | CNCc1cc(N)ccc1O.Cl.Cl | inactive | 152.197 | 0.6938 | 3.0 | 3.0 | 3.476904 |
| 4685 | CHEMBL4292766 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccccc3F)CC2... | active | 494.695 | 6.5580 | 2.0 | 3.0 | 6.124939 |
| 4687 | CHEMBL4284475 | CC(C)(C)c1cc(/C=C/C(=O)NCCC2CCN(Cc3ccc(F)cc3)C... | active | 494.695 | 6.5580 | 2.0 | 3.0 | 6.008774 |

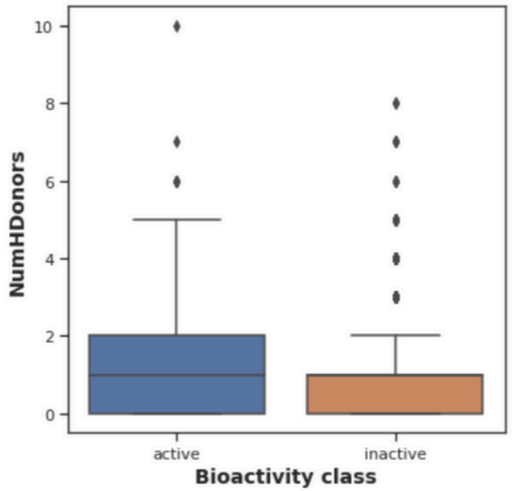
# EXPLORATORY DATA ANALYSIS

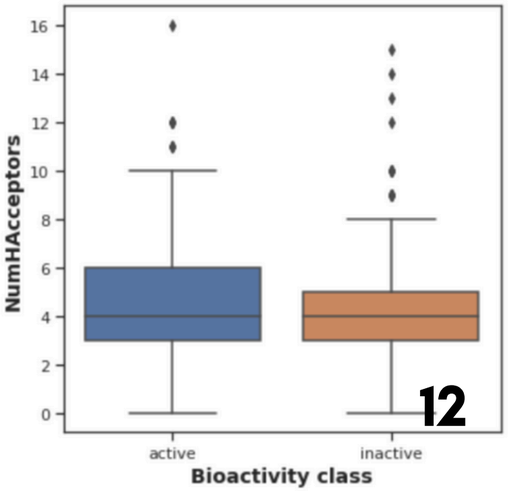
## C. PLOT OF LIPINSKI'S FACTOR VS BIOACTIVITY CLASS IS OBTAINED AS RESULT



# CALCULATION AND DATASET PREPARATION

## A. CALCULATE FINGERPRINT DESCRIPTORS USING PADEL DESCRIPTORS

- **Individually calculate the X (fingerprints matrix) and the Y (PIC50 data matrix) and then combine both the matrices into a single table.**

| | PubchemFP0 | PubchemFP1 | PubchemFP2 | PubchemFP3 | PubchemFP4 | PubchemFP5 | PubchemFP6 | PubchemFP7 | PubchemFP8 | PubchemFP9 | PubchemFP10 | PubchemFP11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **1** | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **2** | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **3** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **4** | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **4690** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **4691** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **4692** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **4693** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| **4694** | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |

4695 rows × 882 columns

# REGRESSION MODELS WITH RANDOM FOREST

1

**IDENTIFY**

The Acetylcholinesterase data set contains 881 input features and 1 output variable (pIC50 values).

2

**OPTIMISE**

Remove low variance features

3

**DATA SPLIT**

Split the data in 80/20 ratio (training: testing)

4

**BUILD**

Build a regression model using Random Forest

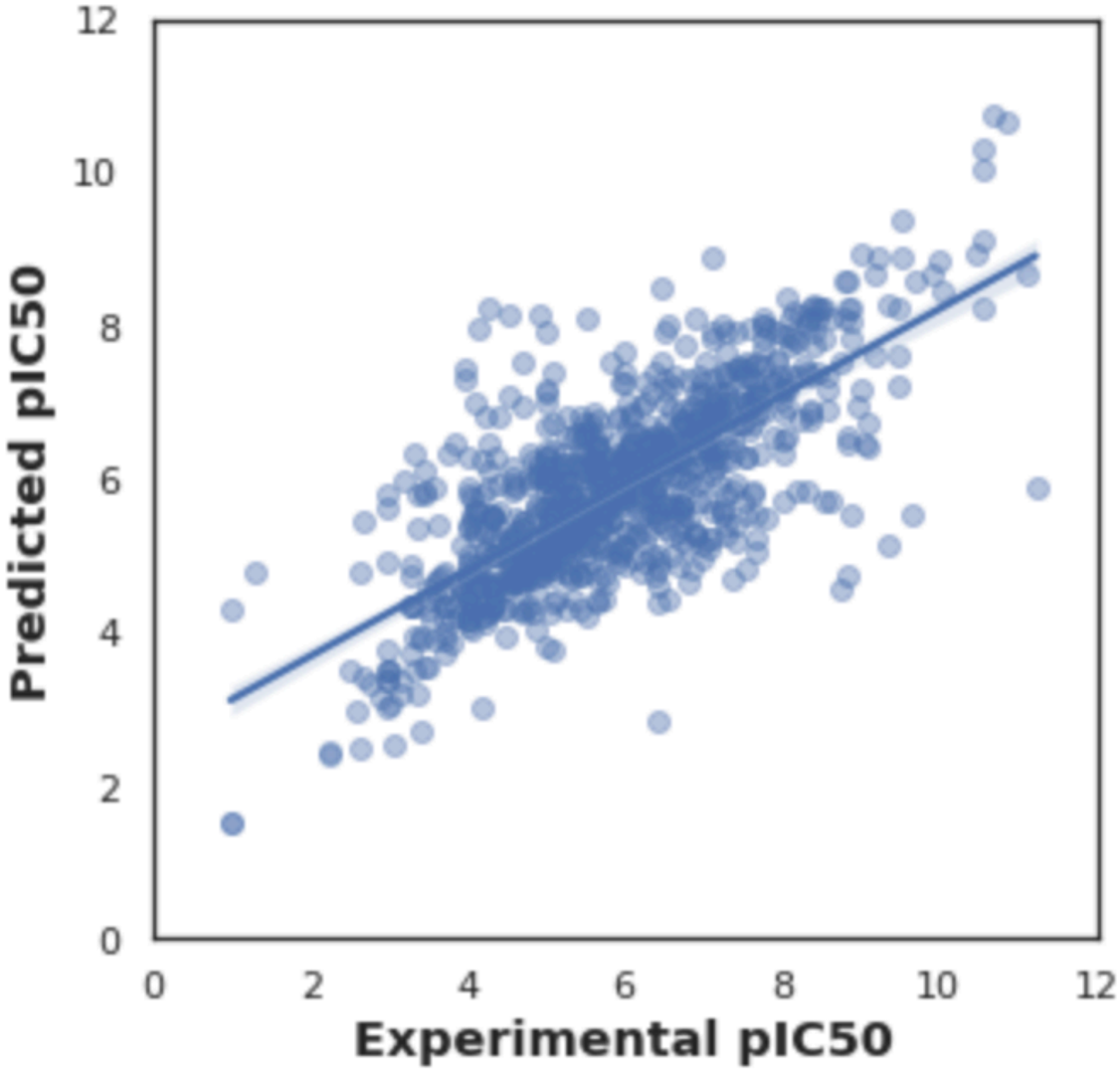
5

**ANALYZE**

Obtain Scatter Plot of Experimental vs Predicted pIC50 Values

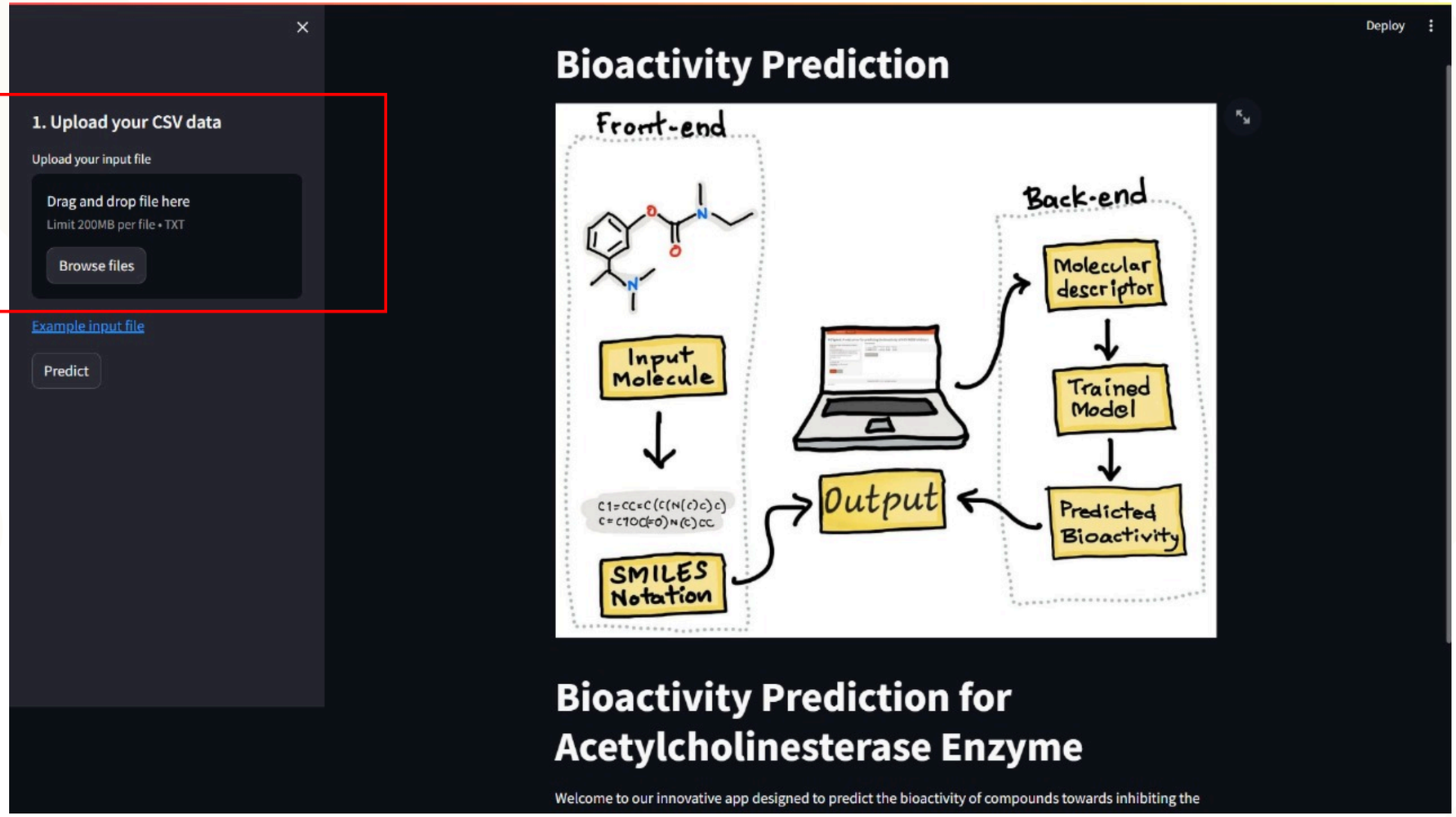
# SCATTER PLOT OF EXPERIMENTAL VS PREDICTED PIC50 VALUES

## Performance Percentage= 86%

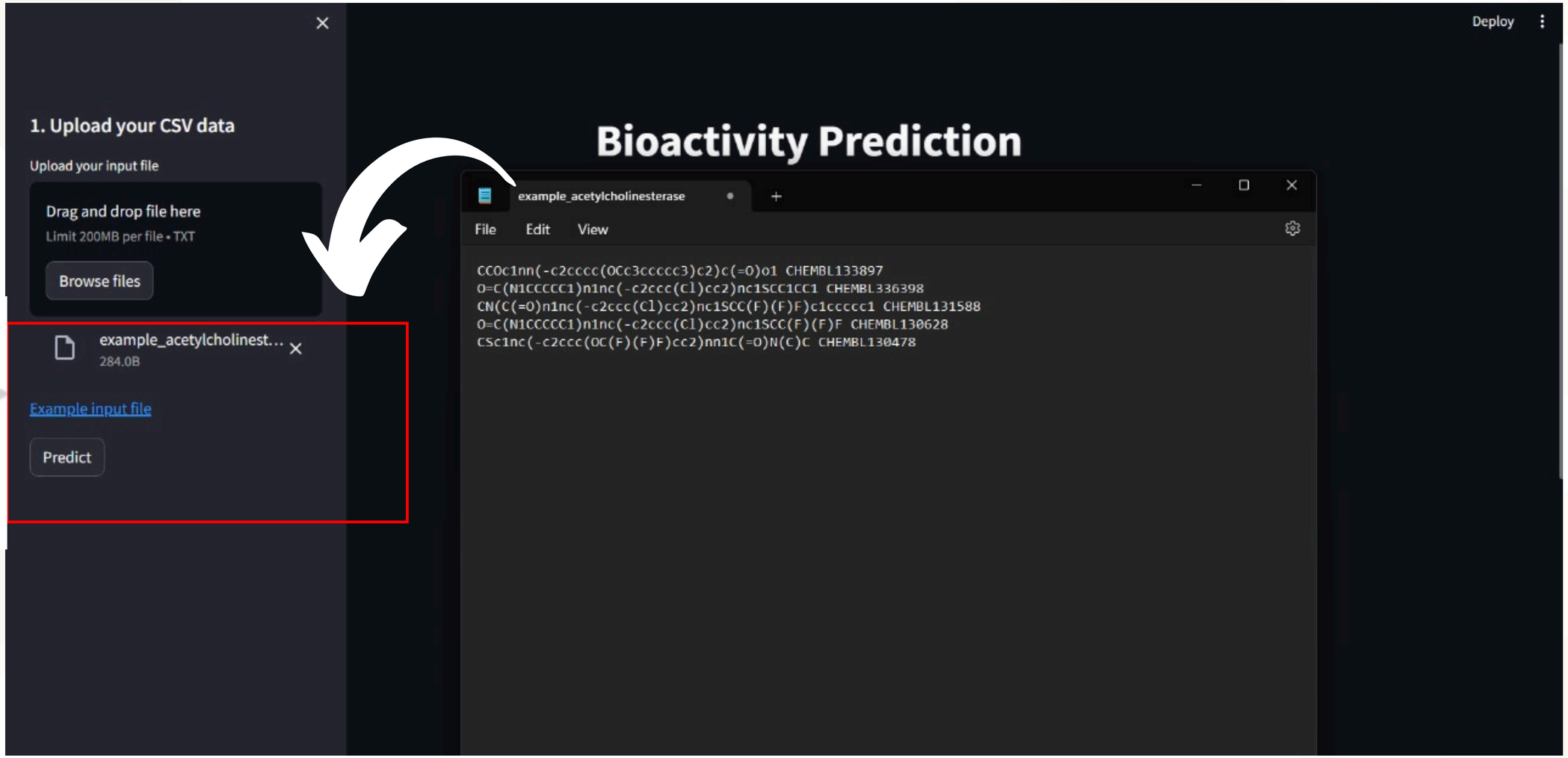


# OUR WEBPAGES WITH WORKING ML MODEL

STEP 1:
UPLOAD CSV
FILE HERE

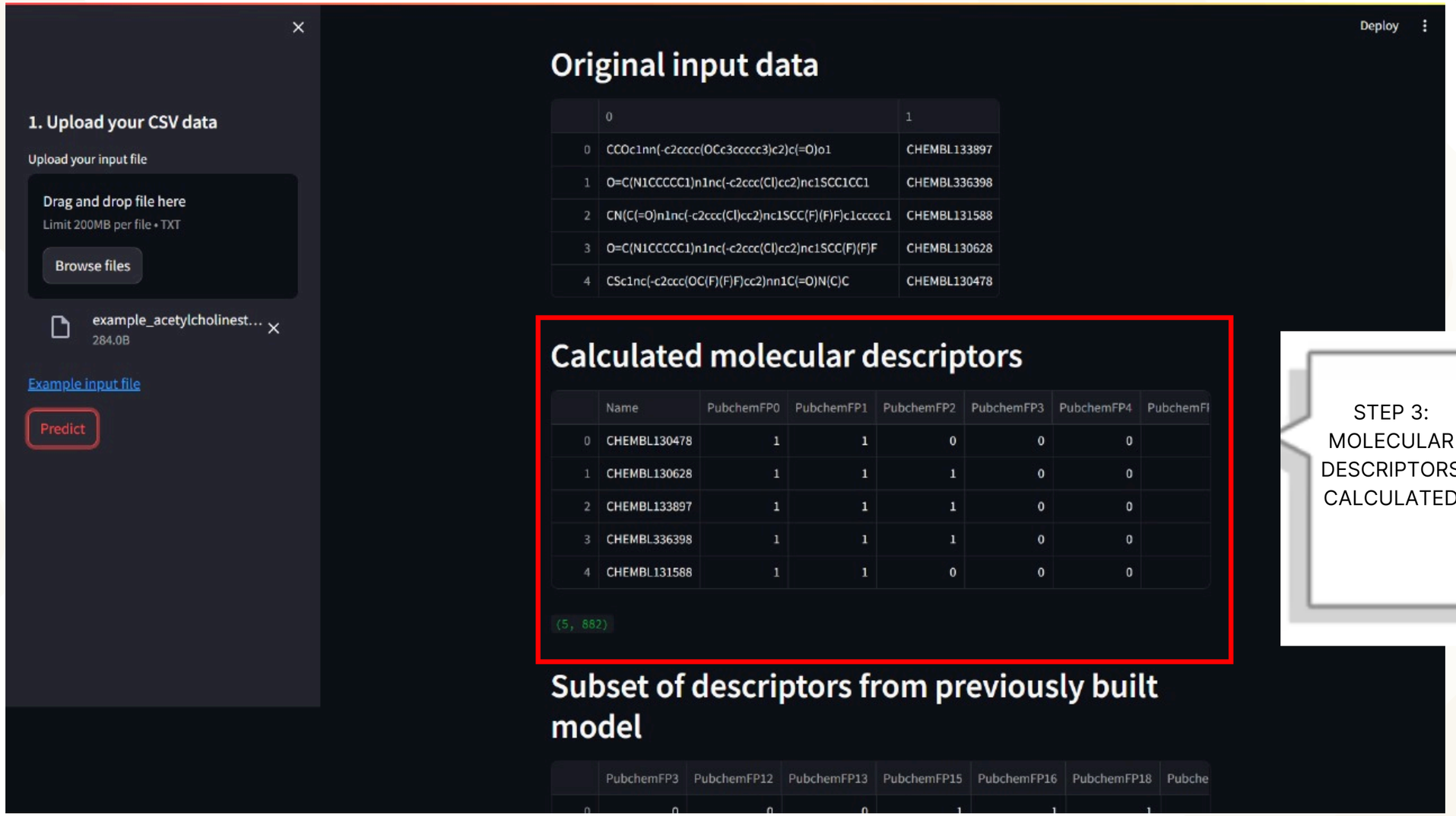


# OUR WEBPAGES WITH WORKING ML MODEL



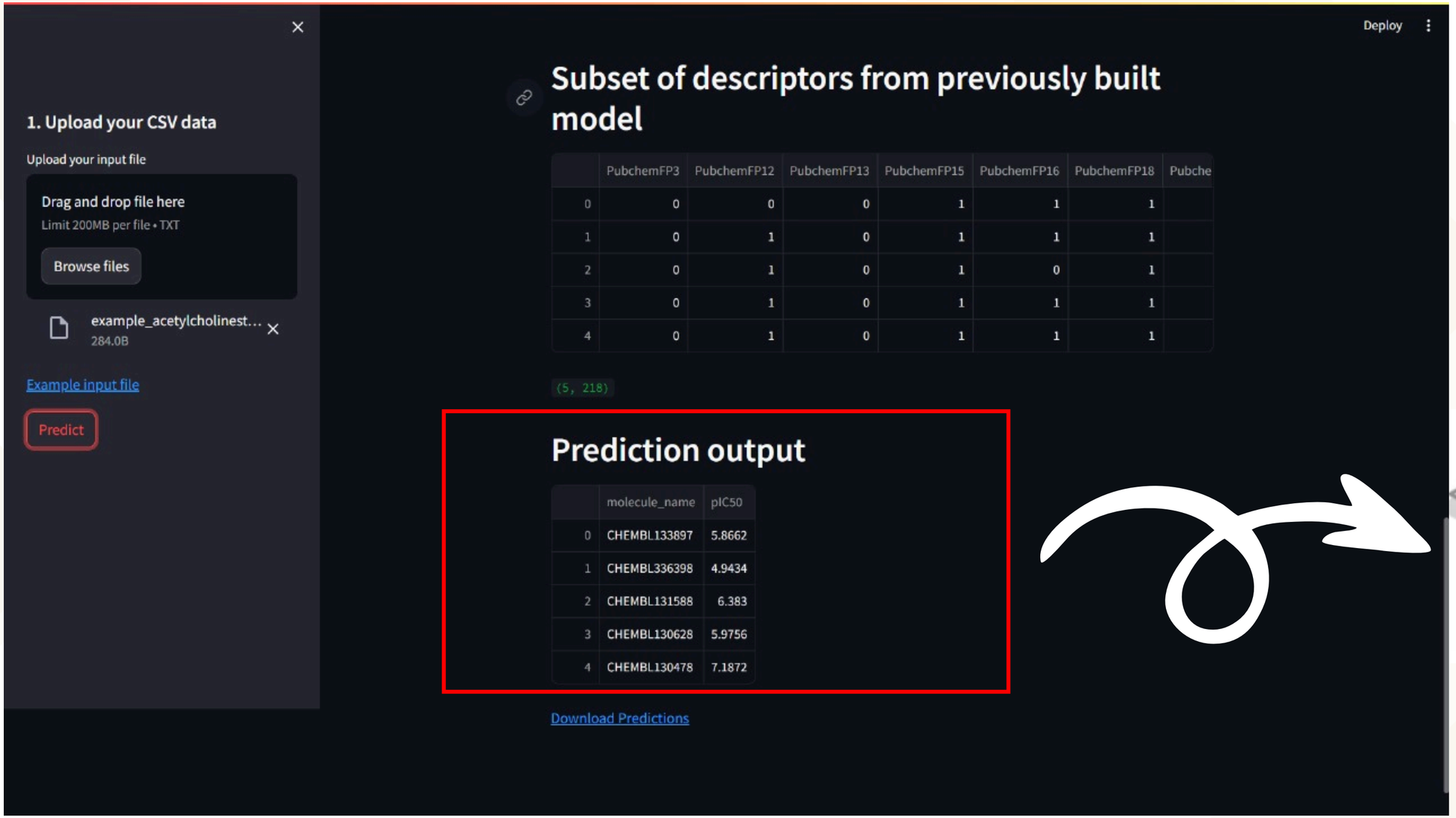

STEP 2: FILE UPLOADED

# OUR WEBPAGES WITH WORKING ML MODEL




STEP 3: MOLECULAR DESCRIPTORS CALCULATED

# OUR WEBPAGES WITH WORKING ML MODEL



# THANK YOU!

## REFERENCES


- NCBI
- PUBMED (NCBI)
- ChEMBL
- PaDEL


## CONTRIBUTIONS:

- Data collection, Website - Grishma, Shubham, Shreyansh
- Data set preparation, PPT- Anisha, Riya, Vimansh
- QSAR and Regression Model- Anisha, Grishma, Riya, Shubham, Shreyansh, Vimansh