Problem Statement Title: Drug Quality Test Using AI Simulation
Team Name: Zombies

Team members details

Team Name	Zombies		
Institute Name/Names	Indraprastha Institute of Information Technology		
Team Members >	**1 (Leader)**	**2**	**3**
Name	Tanishq Tiwari	Om Garg	Aditya Arya
Batch	2025	2025	2025

Programming Modules and Libraries:

Frontend:

1. HTML , CSS – Tailwind CSS
2. Java script

Backend:

1. Python
2. Django Framework

Libraries of Data gathering:

1. Mordred : Generating descriptors
2. RDkit : Generating 2d and 3D descriptors
3. Pubchempy : Converting drug ids into smiles
4. Joblib : Storing the trained model and predicting the output
5. Chembl_WebResource_Client : Converting drugs names to drugs ids
6. Scikit : Importing regression and classification models
7. Sklearn : Importing machine learning models
8. Ipython : Displaying molecular structures

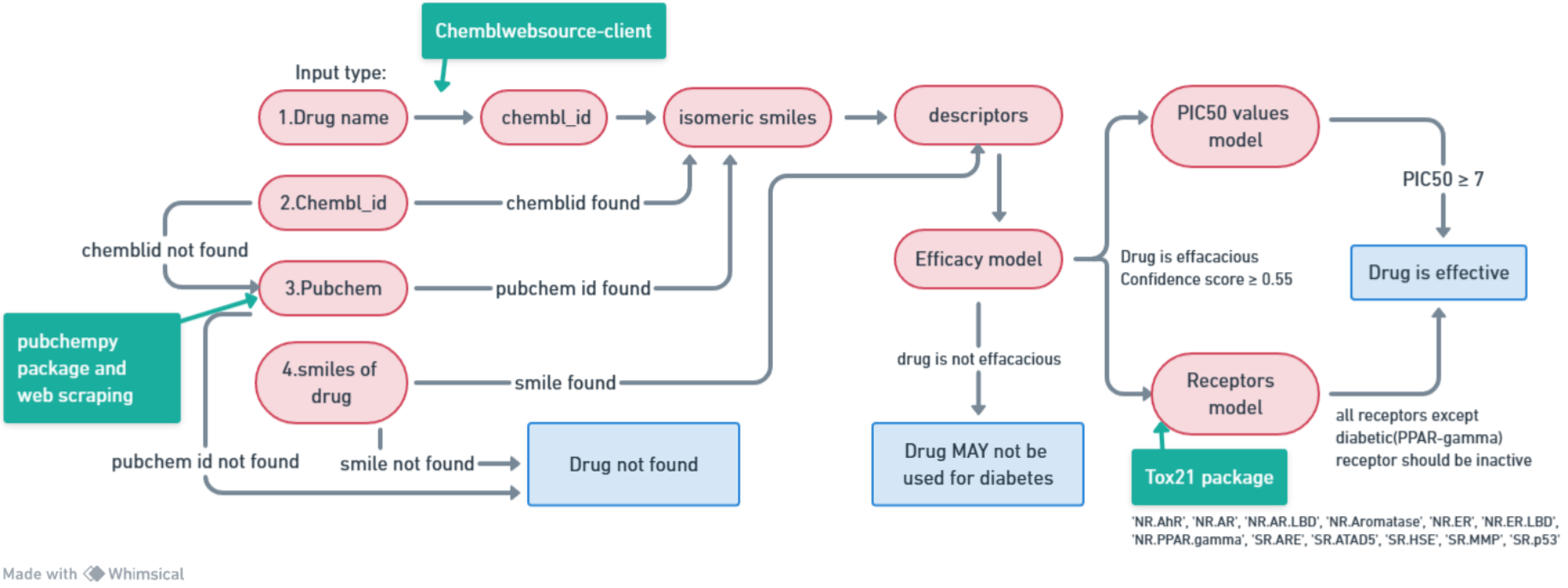
Software we Built:

- *Website* : Our responsive website allows users to input the name, chembl_id, pubchem_id, or smiles of a drug. Using this information, our platform predicts the drug's efficacy and effectiveness. Additionally, users can download a report that provides insights into the drug's quality assessment.

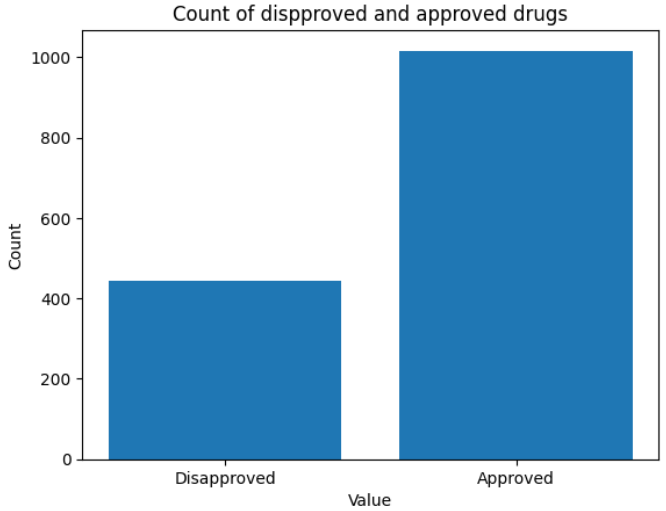
Use-cases

1. **Reducing Adverse Effects:** The website could forecast the risk of negative effects or drug interactions for particular patients by examining patient data. Doctors might use this information to suggest alternate medications or change dosages to reduce hazards.
2. **Preclinical Research:** The predictions can be used by researchers to rank drug candidates for preclinical testing. Choosing to concentrate on the most promising prospects can help you save time and resources.
3. **Educational Tool:** Researchers, doctors, and students can use the website as a learning resource to become familiar with drug efficacy prediction models and their limitations.

Approach for predicting Efficacy and Effectiveness



Data Gathering:



Sources:

1. Chembl_WebResource
2. PubChempy
3. Tdcommons.ai
4. Web scraping
5. DrugBank

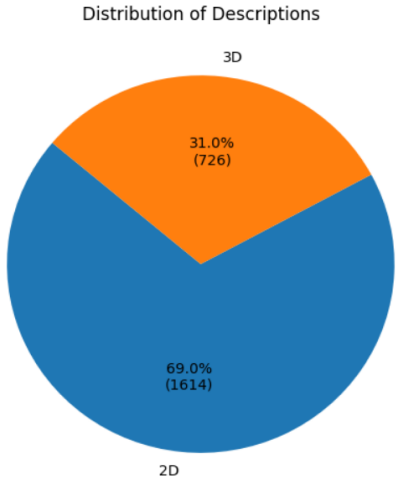
- Collected information from TDcommons.ai's clinical trial section and divided the medications into two groups based on whether they had received approval or not.
- Approved drugs- Drugs passed in phase 3 trials.
- Disapproved drugs- Drugs failed in phase1, phase2, phase3 trials.
- Collected information from a number of research papers**, compared the drugs to PubChempy's clinical trials section, and then took the medications that had passed phase 3, and phase 4 studies.
- We took the authorized and rejected medications from DrugBank and placed the former in "Approved Drugs" and the latter in "Disapproved Drugs."

Descriptor's Data:

Libraries used:

1. Mordred
2. Rdkit : 3D descriptors
3. Rmol Descriptors

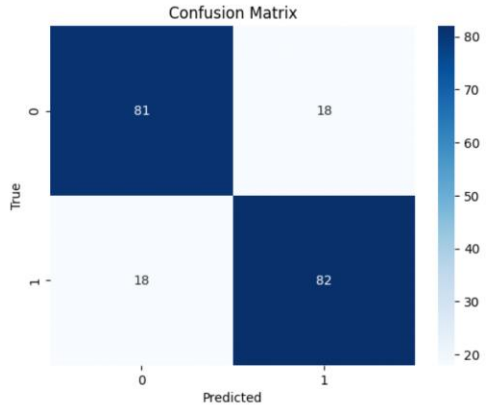
- To generate 2D descriptors for the medication, we used Mordred descriptors.
- Both some 2D and some 3D descriptors were created using Rdkit.
- The remaining Rdkit 3D descriptors were created using Rmol descriptors.
- 2340 descriptors were produced in total.

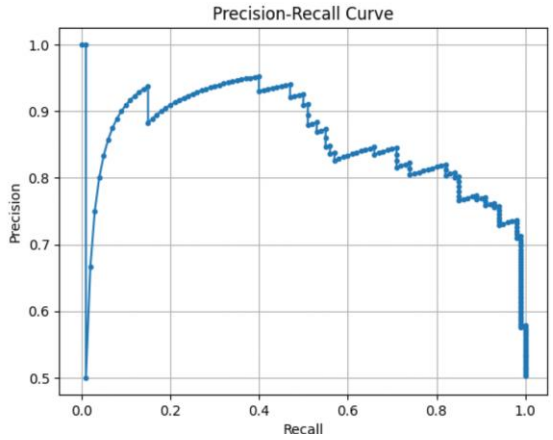


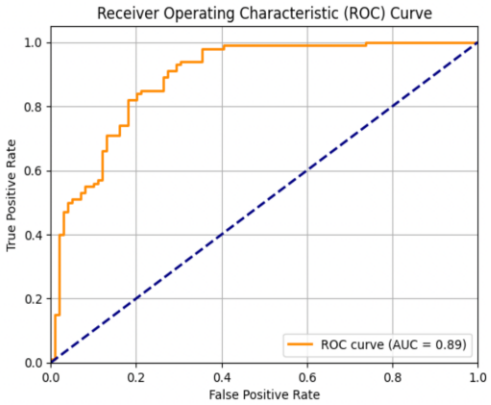
Question:
Why use descriptors as input features to predict efficacy and effectiveness?

Answer: Complex chemical structures can be quantitatively and uniformly represented using molecular descriptors. They reduce a molecule's many structural, physicochemical, and topological details into numerical values that machine learning algorithms can absorb and comprehend.

Predictions from Efficacy Model







Some Inferences:

- Ensemble stacking model is giving the best result
- Sensitivity: 0.81
- Specificity: 0.82
- Precision: 0.81
- F1Score: 0.81

Predictions from Other Model

Testing Accuracies of different Model:

1. SR.ATAD5 : 93.2%
2. SR.HSE :91.2%
3. SR.MMP :92.4%
4. SR.p53 :91.1%
5. NR.PPAR.gamma :93.3%
6. SR.ARE :93.4%
7. NR.ER.LBD : 96.3%
8. NR.AR :93.2%
9. NR.ER :94.1%
10. NR.AhR :91.2%
11. NR.Aromatase :93.3%
12. NR.AR.LBD :94.4%
13. PIC50 R2 score: 81%

Project Scopes

In Scope:

1. Data collection and preprocessing:
   Collected and preprocessed molecular data, clinical trials, descriptors(biomarkers) associated with the drug.
2. Machine learning model development:
   Built predictive models such as classification for efficacy and receptors and regression for PIC50 values.
3. Model Validation:
   validated predictive model using 5 fold cross validation technique.

Advantages:

1. As different qualities are available on several websites, one can anticipate the quality of any diabetes-related drugs at one platform, which can save a lot of time and boost productivity.
2. Accuracies are good enough to rely on our model.

Future Scope

- To compare the structures of diabetic treatments, we will create 3D structural models using GNN. By doing so, we will be better able to determine how similar a drug is to a diabetic drug and forecast its efficacy.
- We can apply machine learning algorithms to predict product quality during manufacture using placebo and medication API data.
- To forecast a drug's unfavourability and side effects, we can also model its ADMET features.
- For bigger smiles, we will mail the result to user afterwards since the computation takes longer time.
- We can forecast the therapeutic effects of the medicine using SIDERDB.
- Instead of just concentrating on diabetes, we can eventually broaden the scope to include other medications.
- We can validate the toxicity data from CLINTOX and LD50 values databases.
- By considering protein characteristics, we may model our findings for some drug-protein interactions.
- We can use the concept of Personalized Medicine for the diagnosis of diseases that rely on individual's genetic profile, we can employ PharmaKinetics and Pharma Dynamics for the same.

Limitations

- For the time being, we are unable to predict the outcome for bigger grins, but we will be able to mail the results to the user shortly.
- We are only forecasting the PIC50 values for targets related to diabetes, some of targets are following:

i. CHEMBL1914266
ii. CHEMBL235
iii. CHEMBL2459
iv. CHEMBL4797
v. CHEMBL1932903
vi. CHEMBL2095161
vii. CHEMBL2095162
viii. CHEMBL2095163
ix. CHEMBL2096976
x. CHEMBL2111325
xi. CHEMBL2111371
xii. CHEMBL2111394
xiii. CHEMBL3559683

- Only diabetic medications are included in our model for estimating efficacy and effectiveness.
- Predicting model based on structure similarity i.e. we intend to find the molecular similarity based on descriptors of drugs.

Cost for setting up test lab and model:

Ram Used in our models: (2-3)GB of Ram.
CPU Utilization - 2.4 GHz Processor.
GPU Ram was not used at all.

So Estimated Cost would be around ~ 200$
We have defined the cost by using Amazon EC2 Instance in which we have used 1v core CPU and 2-4 GB of RAM.

https://aws.amazon.com/ec2/pricing/on-demand/

We have also used gputil package for RAM and CPU Usage but total Cost was coming — 0.01$(so may not be correct). So we defined our requirements and check on Amazon AWS Services which gave us cost of around ~200$ yearly

References:

- https://go.drugbank.com/
- https://www.ebi.ac.uk/chembl/
- https://tdcommons.ai/single_pred_tasks/tox/
- https://pubchempy.readthedocs.io/en/latest/api.html
- https://www.nature.com/articles/nrd4128
- https://www.nature.com/articles/s41597-022-01203-x **
- http://bioinf.jku.at/research/DeepTox/tox21.html **
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2485237/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC400435/
- https://tox-new.charite.de/protox_II/
-

Video
link: https://drive.google.com/file/d/1dJZqurlOxU1ixXeNirnDwTcVo8pnGxkY/view?usp=sharing


Flipkart
GRiD
5.0
Thank You