REPORT

Gagan Neeli

Data collection:

Ligand activity data for Acetylcholinesterase (AChE) was retrieved from the ChEMBL database using its UniProt ID (P22303). The dataset includes bioactivity values (IC50, Ki, Kd, pChEMBL) for molecules tested against AChE. Only ligands with relevant activity data were retained.

Molecular structures were extracted in SMILES format using ChEMBL's molecule endpoint. The final dataset, AChE_ligands.csv, contains ChEMBL IDs, activity values, and SMILES for further analysis.

Cleaning and preprocessing data:

The dataset was cleaned and preprocessed to ensure quality and consistency. Unnecessary columns (type, units, value) were removed. Numeric columns (pChEMBL value, standard value) were converted to numerical format for analysis.

A normalization process was applied to scale activity values between 0 and 1, and an Activity Score was computed using a weighted formula (70% pChEMBL, 30% standard value). Molecules were then classified as Active or Inactive based on a threshold of 0.5. The cleaned and labeled dataset was saved for further analysis.

Feature extraction:

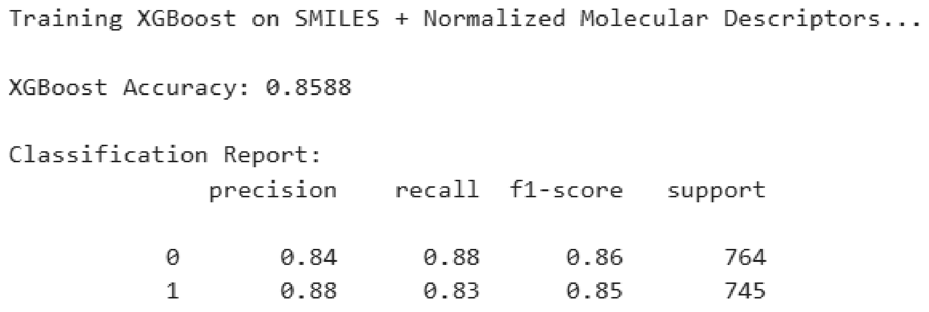
Key molecular descriptors (MolWt, LogP, TPSA, HBA, HBD, RotBonds) were extracted from SMILES using RDKit to analyze molecular properties. Invalid SMILES were removed, and the processed dataset was saved as dataset_with_descriptors.csv.

Training and testing different models for best performing model for classification of active or inactive ligands:

Multiple machine learning models (Random Forest, XGBoost, SVM, etc.) and deep learning models (ChemBERTa) were trained to classify molecules as Active or Inactive.

- Feature Engineering: Molecular descriptors and fingerprints were used as input features.
- Training: Models were trained on an 80-20 train-test split, with standardization applied where needed.
- Evaluation: Performance was assessed using accuracy, precision, recall, and F1-score to select the best-performing model.

XGBoost with SMILES-based fingerprints + molecular descriptors showed the highest accuracy, making it the best model for predicting ligand activity.



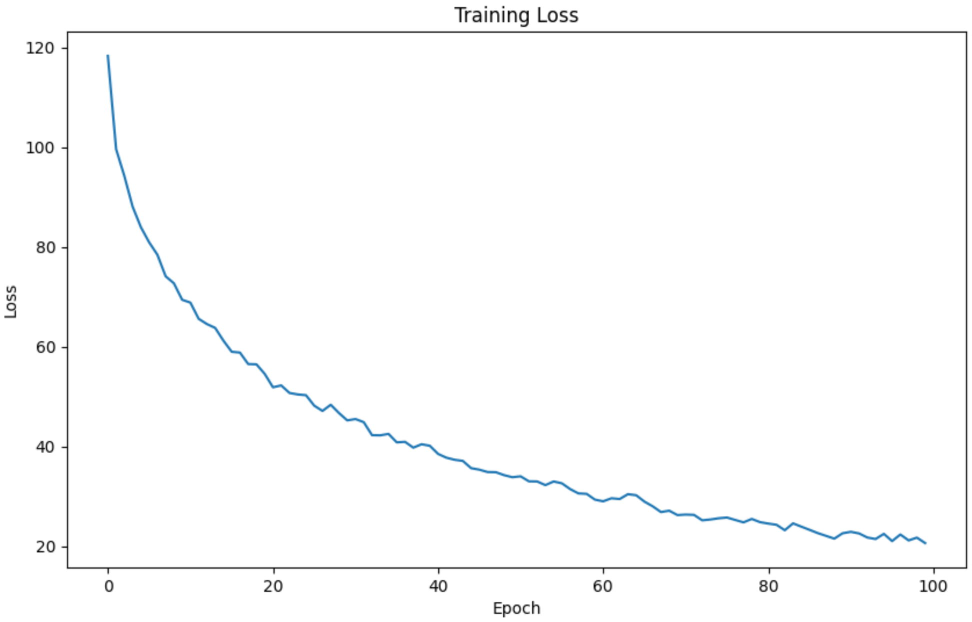
```
Training XGBoost on SMILES + Normalized Molecular Descriptors...

XGBoost Accuracy: 0.8588

Classification Report:
             precision    recall  f1-score   support

           0       0.84      0.88      0.86       764
           1       0.88      0.83      0.85       745
```

Smiles generating model:

A Variational Autoencoder (VAE) was used to generate novel molecules based on active SMILES. The model was trained on character-level SMILES sequences, encoding molecules into a latent space and decoding them to generate new structures.

- **Preprocessing:** Active SMILES were tokenized, and character mappings were created.
- **Model Training:** A bidirectional **GRU-based VAE** was trained for **100 epochs** to learn molecular representations.
- **Molecule Generation:** The trained model sampled from the latent space to create new SMILES.
- **Validation:** Generated molecules were checked for **chemical validity** using RDKit.

The final valid generated **molecules** were saved for further evaluation.



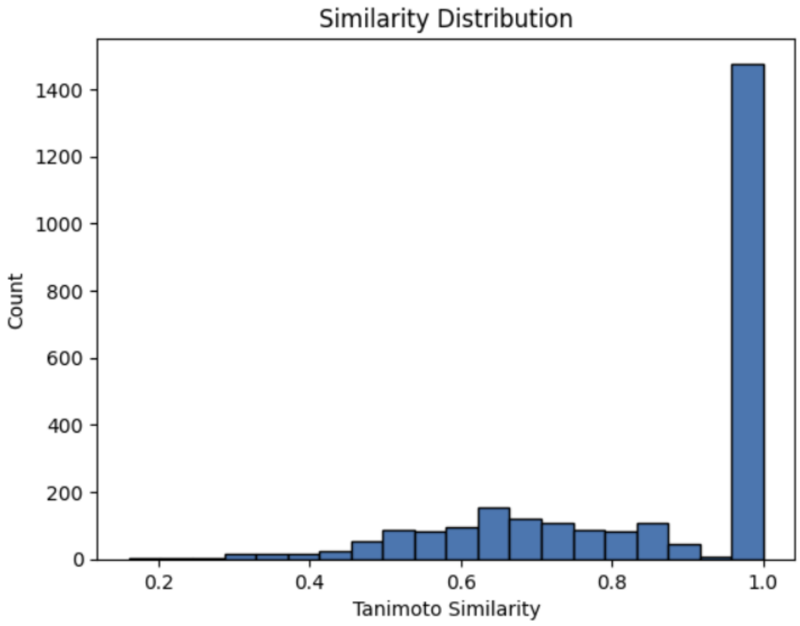
Checking if generated molecules are active:

To assess the activity of generated molecules, pre-trained XGBoost model was used to predict whether each molecule is Active or Inactive based on its molecular fingerprint. The generated SMILES were converted into Morgan fingerprints (ECFP4, 2048-bit) and fed into the model for classification. Each prediction included an activity label (Active/Inactive) along with a confidence score. The final results were displayed and saved for further analysis.

Similarity checks for newly generated ligands:

To evaluate the structural similarity between the newly generated ligands and known active ligands, the Tanimoto Similarity was calculated using Morgan fingerprints (ECFP4, 2048-bit). Each generated molecule was compared to the most similar active ligand, and the highest similarity score was recorded.

An average similarity score of 0.8556 indicates that the generated ligands share significant structural features with known active molecules. The results were saved in Similarity_Scores.csv for further analysis.



Strengths of this approach:

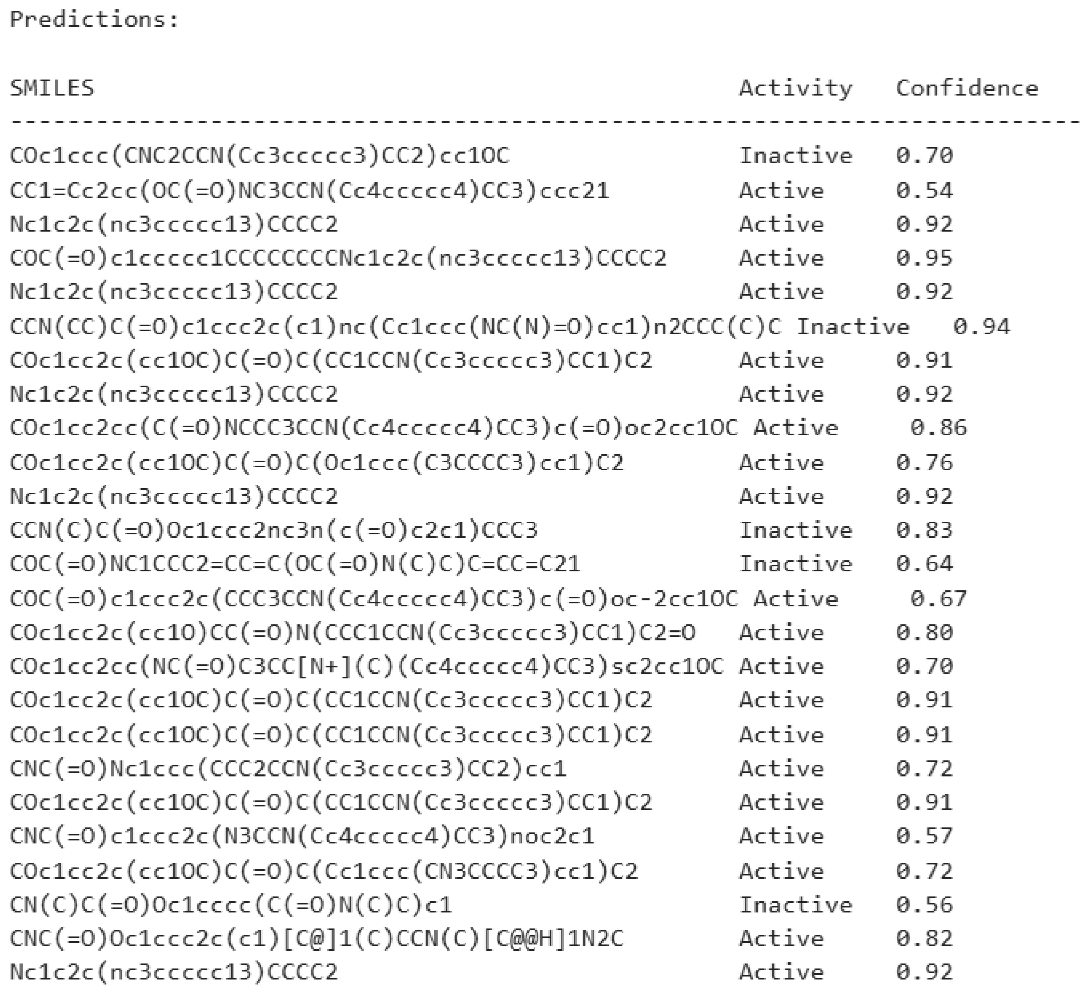
- **High Success Rate in Generating Active Molecules** – Most generated SMILES are classified as active, ensuring relevance for drug discovery.
- **Best Accuracy Among Trained Models** – The classification model outperforms others, leading to reliable activity predictions.
- **Controlled Vocabulary for SMILES Generation** – The model only generates molecules using known chemical substructures, reducing the likelihood of unrealistic SMILES.
- **Reliable Data Retrieval from ChEMBL** – Ensures clean, structured data without corruption, improving training quality.
- **Robust Activity Classification** – The **weighted activity score** improves accuracy in distinguishing active and inactive compounds.

Limitations:

- **Limited Dataset Size** – With only **7,500 data points** from a single protein target, the model's learning capacity is constrained.
- **Generation of Invalid SMILES** – Some generated molecules are invalid, though the model effectively filters them out.
- **Similarity to Known Ligands** – About 45**% of** valid SMILES closely resemble existing active molecules, likely due to the small dataset size limiting structural diversity.

Result:

Few generated ligands and their activity state:



```
Predictions:

SMILES                                             Activity   Confidence
---------------------------------------------------------------------------
COc1ccc(CNC2CCN(Cc3ccccc3)CC2)cc1OC                Inactive   0.70
CC1=Cc2cc(OC(=O)NC3CCN(Cc4ccccc4)CC3)ccc21         Active     0.54
Nc1c2c(nc3ccccc13)CCCC2                            Active     0.92
COC(=O)c1ccccc1CCCCCCCCNc1c2c(nc3ccccc13)CCCC2     Active     0.95
Nc1c2c(nc3ccccc13)CCCC2                            Active     0.92
CCN(CC)C(=O)c1ccc2c(c1)nc(Cc1ccc(NC(N)=O)cc1)n2CCC(C)C Inactive   0.94
COc1cc2c(cc1OC)C(=O)C(CC1CCN(Cc3ccccc3)CC1)C2      Active     0.91
Nc1c2c(nc3ccccc13)CCCC2                            Active     0.92
COc1cc2cc(C(=O)NCCC3CCN(Cc4ccccc4)CC3)c(=O)oc2cc1OC Active     0.86
COc1cc2c(cc1OC)C(=O)C(Oc1ccc(C3CCC3)cc1)C2         Active     0.76
Nc1c2c(nc3ccccc13)CCCC2                            Active     0.92
CCN(C)C(=O)Oc1ccc2nc3n(c(=O)c2c1)CCC3              Inactive   0.83
COC(=O)NC1CCC2=CC=C(OC(=O)N(C)C)C=CC=C21           Inactive   0.64
COC(=O)c1ccc2c(CCC3CCN(Cc4ccccc4)CC3)c(=O)oc-2cc1OC Active     0.67
COc1cc2c(cc1O)CC(=O)N(CCC1CCN(Cc3ccccc3)CC1)C2=O   Active     0.80
COc1cc2cc(NC(=O)C3CC[N+](C)(Cc4ccccc4)CC3)sc2cc1OC Active     0.70
COc1cc2c(cc1OC)C(=O)C(CC1CCN(Cc3ccccc3)CC1)C2      Active     0.91
COc1cc2c(cc1OC)C(=O)C(CC1CCN(Cc3ccccc3)CC1)C2      Active     0.91
CNC(=O)Nc1ccc(CCC2CCN(Cc3ccccc3)CC2)cc1            Active     0.72
COc1cc2c(cc1OC)C(=O)C(CC1CCN(Cc3ccccc3)CC1)C2      Active     0.91
CNC(=O)c1ccc2c(N3CCN(Cc4ccccc4)CC3)noc2c1          Active     0.57
COc1cc2c(cc1OC)C(=O)C(Cc1ccc(CN3CCCC3)cc1)C2       Active     0.72
CN(C)C(=O)Oc1cccc(C(=O)N(C)C)c1                    Inactive   0.56
CNC(=O)Oc1ccc2c(c1)[C@]1(C)CCN(C)[C@@H]1N2C        Active     0.82
Nc1c2c(nc3ccccc13)CCCC2                            Active     0.92
```