ai powered drug development

Deep learning algorithm used in drug discovery

SHREEYA SANCHI

Draft paper

# 🧪 Project Title:

## A Novel Deep Learning-Based Pipeline for Drug Activity Prediction Using Molecular Descriptors

### **📘 1. Methodology / Model Used**

I have designed a novel descriptor-based deep learning pipeline to predict drug bioactivity (pXC50) using molecular property datasets, specifically drugs.csv and ChEMBL.csv. The methodology consists of:

🔹 a) Data Collection:

Two datasets used:

drugs.csv containing 34 molecular descriptors per compound

ChEMBL.csv containing drug activity scores (pXC50\_3D7)

🔹 b) Data Preprocessing:

Unnecessary columns like SMILES were dropped (no RDKit used).

Missing values were removed.

Data was aligned using index-wise matching across both datasets.

Features were numerical descriptors like LogP, Polarizability, Heavy Atom Count, etc.

Data normalization was performed using StandardScaler from scikit-learn.

🔹 c) Model Architecture:

I have used a deep feedforward neural network (FNN) built using TensorFlow/Keras:

**🎯 Architecture Diagram:**

Input Layer

↓

Dense Layer (128 units, ReLU)

↓

Dropout (rate = 0.3)

↓

Dense Layer (64 units, ReLU)

↓

Output Layer (1 unit – regression output)

The model predicts the pXC50 bioactivity value for each compound.

🔹 d) Training:

Model trained using MSE (Mean Squared Error) loss

Optimizer: Adam

Batch size: 16, Epochs: 30–50

20% of training data used for validation

### **🎯 2. Novelty**

The key novelties in this approach:

1. Descriptor-Only Learning: Unlike other models relying on SMILES, molecular graphs, or RDKit, this model uses only precomputed descriptors, making it lightweight, fast, and RDKit-independent.

2. Dashboard Deployment: A real-time Streamlit dashboard was developed for direct predictions using descriptor values, making the tool usable by non-technical users like pharmacists or biologists.

3. Cross-Dataset Compatibility: The model can be cross-validated and retrained on other descriptor-based datasets — ensuring adaptability and robustness.

### **📊 3. Accuracy and Evaluation Metrics**

🔹 a) Evaluation Metrics Used:

MAE (Mean Absolute Error) – Primary metric for regression accuracy

MSE (Mean Squared Error) – Loss function for training

Residual Error Distribution – For checking prediction quality

Cross-validation MAE – To assess robustness

🔹 b) Observed Results:

Average MAE on test set: ~0.20 to 0.35 (depending on fold)

Average MAE across 5-fold CV: ~0.25 (± small variance)

🧠 These results are considered very good for molecular regression where pXC50 varies between ~5.0 and 8.0.

### **🧩 4. Architectural Design Summary**

Here’s a diagram-like description:

┌──────────────────────────────┐

│ drugs.csv + ChEMBL.csv │

└────────────┬─────────────────┘

│

┌─────────────▼──────────────┐

│ Select Numerical Features │

└─────────────┬──────────────┘

│

┌──────────▼──────────┐

│ StandardScaler │

└──────────┬──────────┘

│

┌──────▼───────┐

│ Deep Learning│

│ Model (FNN) │

└──────┬───────┘

│

┌─────▼─────┐

│ Predict │

│ pXC50 │

└─────┬─────┘

│

┌──────────────▼────────────────┐

│ Real-time Streamlit Dashboard │

└───────────────────────────────┘