Protein-Ligand Binding Prediction: Siamese Neural Network and DeepDTA Approaches

1. Introduction

The accurate prediction of protein-ligand binding is a critical step in drug discovery. This project aimed to develop machine learning-based approaches for predicting binding interactions between proteins and small molecules, leveraging two distinct models:

1. **Siamese Neural Network (SNN)** for binary classification (binding/non-binding).
2. **DeepDTA** for predicting continuous binding affinity scores (KIBA scores).

Due to time constraints, the initial focus was on building a binary classification model with high-quality data (directly measured KIBA scores). Later, the project expanded to leverage sequence and molecular data using DeepDTA to predict binding affinities with more granularity.

Why Two Models?

- **Siamese Neural Network (SNN)**: Suitable for similarity-based classification tasks. Ideal when focusing on binary predictions (binding vs non-binding) and when feature-based numerical data is available.
- **DeepDTA**: A regression-based approach that leverages sequence data and SMILES strings. It excels in predicting continuous KIBA scores and capturing intricate patterns in protein and ligand sequences.

Both models were chosen after reviewing relevant literature and considering the time constraints for the project.

2. Data Preparation and Feature Engineering

2.1 Dataset Overview

The provided dataset contained:

- **UniProt IDs** for proteins.
- **PubChem IDs** for small molecules (ligands).
- **KIBA scores** (binding affinity).

- **`kiba_score_estimated`** flag to indicate whether the score was estimated (`True`) or directly measured (`False`).

To enhance the dataset, additional features were sourced from the **BindingDB dataset**, including:

1. **Molecular Weight**
2. **LogP** (Partition coefficient)
3. **TPSA** (Topological Polar Surface Area)
4. **Number of Rotatable Bonds**

2.2 Data Curation Strategy

1. **High-Quality Data**:
   - Focused on rows where **`kiba_score_estimated = False`** to ensure the use of reliable, directly measured binding scores.
   - Only **4.11%** of the dataset had `kiba_score_estimated = False` (~53,000 rows).
   - For DeepDTA I first pre-trained it on **`kiba_score_estimated = True`** and then trained and tested on the False rows.
2. **Merging with BindingDB**:
   - Used **`preprocess_merge_data.py`** to merge the provided dataset with BindingDB data to enrich the features.
   - Dropped rows with missing critical numerical features due to time constraints.
3. **Final Dataset**:
   - Approximately **26,897 rows** with `kiba_score_estimated = False` after merging and cleaning.
   - Created a balanced dataset of **positive** and **negative** pairs for training.

2.3 Synthetic Negative Pair Generation

- Initially used **random pairing** to generate negative examples.
- Improved the negative pair generation by using **K-Nearest Neighbors (KNN)** to select ligands with similar molecular properties, making the task more challenging and realistic.

---

3. Model Design and Implementation

3.1 Siamese Neural Network (SNN)

Model Overview

The Siamese Neural Network was designed for binary classification (binding/non-binding) based on the numerical features derived from the dataset.

Architecture

- **Three Fully Connected Layers** with Batch Normalization and Dropout (0.4) for regularization.
- **Contrastive Loss** to compute the distance between embeddings of protein-ligand pairs.

Hyperparameters

- **Learning Rate**: 0.0001
- **Batch Size**: 32
- **Epochs**: 20
- **Early Stopping**: Patience of 10 epochs.

Results

Metric	Value
Accuracy	0.9523
Precision	0.9582
Recall	1.000
F1 Score	0.9553

Confusion Matrix:

Actual \ Predicted	No Bind	Bind
No Bind	3902	133
Bind	0	4035

Inference Example

Selected Pair	UniProt_ID	PubChem_CID
Pair 1	A0A0B4J268	7428.0
Pair 2	A0A0B4J268	65303.0

- **Feature Vector 1**: [1.5031, -1.1973, -1.3804, -0.1029, -0.8692]
- **Feature Vector 2**: [1.5016, -0.7970, -1.1081, 0.0249, -0.7091]

- **Model Output (Distance)**: 0.4798
- **Prediction**: **Binding**

3.2 DeepDTA

Model Overview

The **DeepDTA** model is a deep learning-based approach designed for predicting continuous binding affinity scores (KIBA scores). It leverages:

- **Convolutional Neural Networks (CNNs)** for encoding:
  - **SMILES Strings** for ligands.
  - **Protein Sequences** for proteins.
- **Multi-Head Attention Mechanisms** for capturing interactions between drug molecules and proteins.

Architecture

- **SMILES Encoder**: CNN module for drug molecules.
- **Protein Encoder**: CNN module for protein sequences.
- **Multi-Head Attention**: Enhances the interaction learning between drugs and proteins.
- **Fully Connected Layer**: Predicts the binding affinity score.

Results

Metric	Value
Loss	0.99
RMSE	0.98
R²	0.5775
CI	0.7569

---

4. Comparison of Siamese Neural Network and DeepDTA

Criteria	Siamese Neural Network (SNN)	DeepDTA
Task	Binary Classification (Binding/Non-Binding)	Regression (Continuous KIBA Scores)
Input Features	Numerical Features (Molecular Descriptors)	SMILES Strings & Protein Sequences

Architecture	Fully Connected Layers	CNN + Multi-Head Attention
Performance	High Accuracy and Recall	Good RMSE and Concordance Index (CI)
Best Use Case	When high-quality numerical features are available	When sequence data is available and detailed affinity scores are needed
Scalability	Faster training, suitable for smaller datasets	More complex, suitable for larger datasets

---

5. Challenges and Future Improvements

Challenges

1. **Limited Time**:
    - Focused only on `kiba_score_estimated = False` rows.
    - Dropped rows with missing data rather than imputing or estimating them.
2. **Negative Pair Generation**:
    - Used KNN for more realistic negative pairs, but further improvements could be achieved using advanced clustering techniques.
3. **Threshold Selection**:
    - Explored a fixed threshold (0.5) for the SNN. Conducting A/B tests with thresholds like [0.4, 0.45, 0.5, 0.55] could further optimize predictions.

Future Work

1. **Incorporate Estimated Scores**:
    - Balance the dataset by including rows where `kiba_score_estimated = True`.
2. **Advanced Feature Representation**:
    - **Protein Sequence Embeddings**: Use models like **ProtBERT** or **ESM**.
    - **Ligand SMILES Embeddings**: Use Transformer-based models like **ChemBERTa**.
3. **Model Enhancements**:
    - **Graph Neural Networks (GNNs)** for richer drug representations.
    - **Transformer-based Models** to improve interaction learning.
4. **A/B Testing**:
    - Compare different negative pair generation methods and feature sets.

---

6. Conclusion

This project demonstrates the potential of using both **Siamese Neural Networks** and **DeepDTA** for protein-ligand binding prediction. Each approach has its strengths, and combining numerical features with sequence data can lead to robust and accurate predictions. Future work can focus on integrating richer features, optimizing model architectures, and systematically improving model performance through A/B testing.

Appendix

- **`main.py`**: Contains the main execution code for training and testing the models.
- **`data_utils.py`**: Handles data loading, preprocessing, and pair generation.
- **`siamese_model.py`**: Defines the Siamese Neural Network architecture.
- **`preprocess_merge_data.py`**: Merges the Deloitte dataset with BindingDB.
- **`add_seq.py`**: Adds protein sequences to the merged dataset.
- **`test_inference_true_rows.py`**: Runs inference on the test set using the trained models.

Datasets:

- **Deloitte_DrugDiscovery_dataset.csv**
- **BindingDB_All_2D.sdf**

Merged Dataset: `merged_dataset.csv`