FINAL PROJECT:
DRUG–PROTEIN BINDING AFFINITY PREDICTION WITH TEFDTA

ALEJANDRO SILVA RODRÍGUEZ
MARTA CUEVAS RODRÍGUEZ

APRENDIZAGE COMPUTACIONAL
UNIVERSIDAD DE MÁLAGA

DICIEMBRE 2024

INDEX

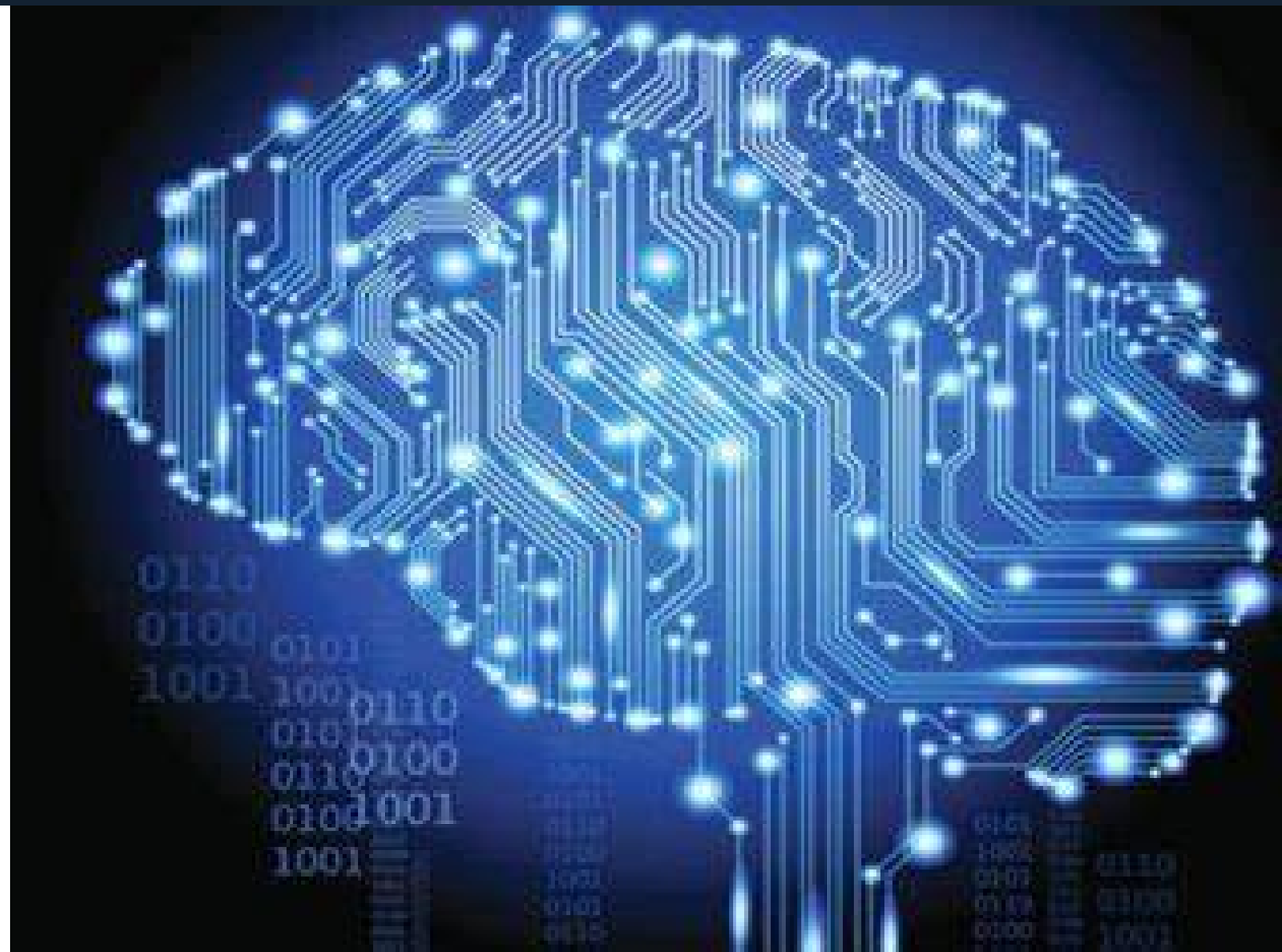
- INTRODUCTION
- OBJETIVES
- METHODOLOGY AND RESULTS
  - DATASET DESCRIPTION
  - MODEL ARCHITECTURE
  - PERFORMANCE ANALYSIS
- DISCUSION
- ACCESS TO THE REPOSITORY

1_ INTRODUCTION

INRODUCTION

Challenge in drug discovery: Predicting drug–target interactions (DTI) is crucial for identifying binding affinities between small molecules and proteins, with covalent interactions being particularly challenging.

The model proposed by Zongquan Li et al., combines transformers and fingerprint-based molecular representations demonstrating robustness in predicting covalent and non-covalent interactions.



2 – OBJECTIVES

OBJECTIVES

REPRODUCE TEFDTA EXPERIMENTS

IMPLEMENT AND VALIDATE THE EXPERIMENTS USING THE PROVIDED DATASETS AND CODE.

EVALUATE MODEL PERFORMANCE

ASSESS THE TEFDTA MODEL'S ACCURACY FOR PREDICTING COVALENT AND NON-COVALENT DRUG-TARGET INTERACTIONS.

COMPARE RESULTS

ANALYZE DIFFERENCES BETWEEN REPRODUCED FINDINGS AND THOSE IN THE ORIGINAL PAPER TO CONFIRM REPRODUCIBILITY.

3 – METHODOLOGY AND RESULTS

DATASETS DESCRIPTION

THE MODEL HAS BEEN EVALUATED IN THREE DIFFERENT DATASETS

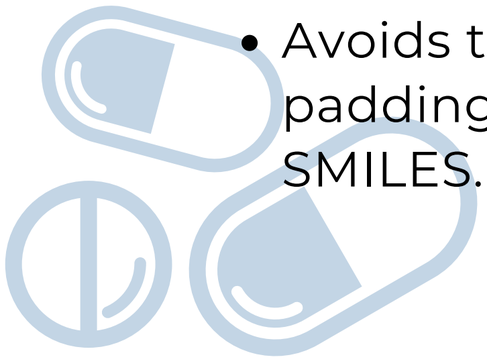
- **DAVIS**: THIS DATASET COMPRISES 442 PROTEINS AND 68 DRUGS, RESULTING IN 30, 056 BINDING AFFINITY VALUES.
- **KIBA**: CONSISTING OF 2, 111 DRUGS, 229 TARGETS, AND 118, 254 BIOACTIVITY SCORES.
- **COVALENTINDB**: A SPECIALIZED DATASET OF COVALENT DRUG-TARGET INTERACTIONS, CURATED TO FINE-TUNE THE MODEL FOR PREDICTING BONDED INTERACTIONS.

MODEL ARCHITECTURE:

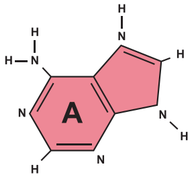
INPUT REPRESENTATION

DRUGS

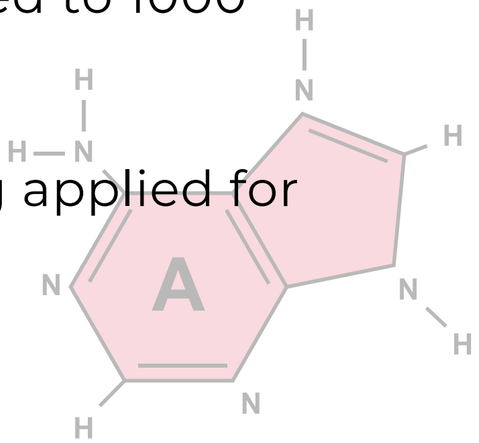
- Represented as MACCS fingerprints (166-bit binary vectors).
- Avoids the sequence padding required by SMILES.



PROTEINS



- Encoded as integers (amino acids mapped to integers).
- Sequences truncated to 1000 residues.
- Positional encoding applied for spatial awareness.



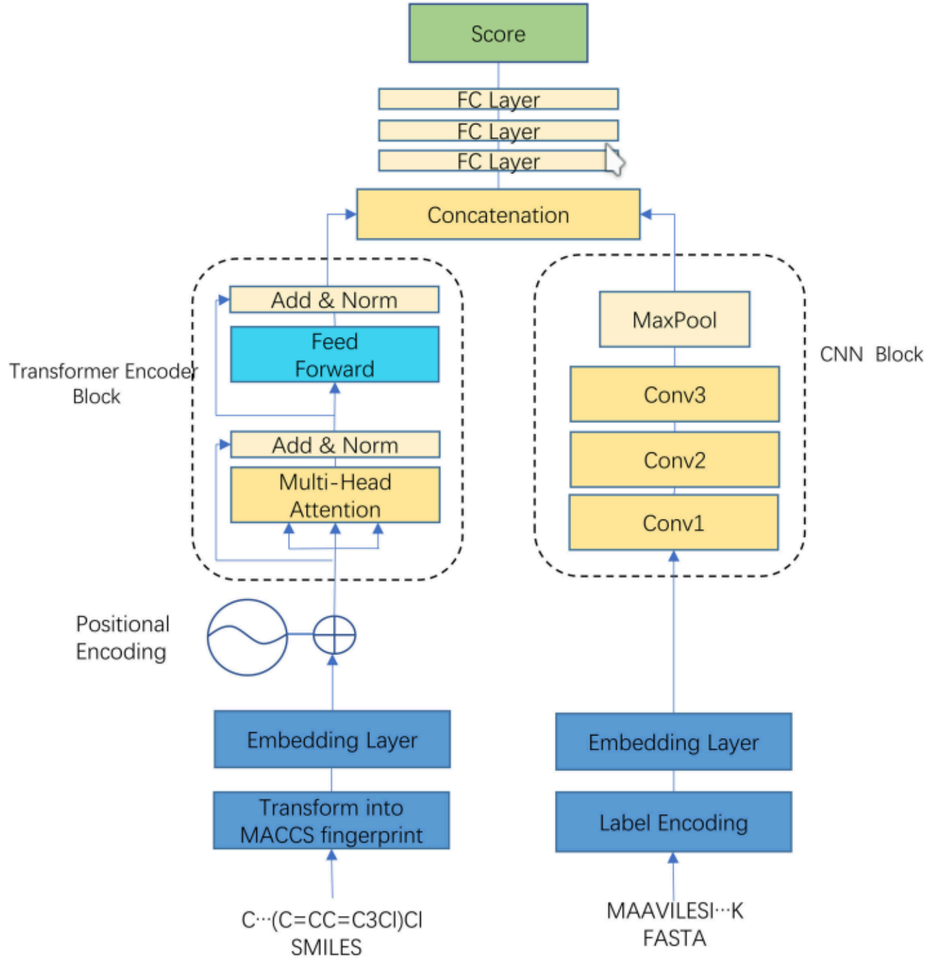
MODEL ARCHITECTURE:
FEATURE EXTRACCTION AND PREDICTION

- **Drug Features:** Extracted using Transformer Encoder with multi-head attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- **Protein Features:** Extracted using 1D-CNN, capturing local amino acid patterns through convolution and pooling layers.
- **Feature Fusion:**

  - Combined drug and protein features into a single representation vector.
  - Passed through fully connected layers for binding affinity prediction.



PERFORMANCE ANALYSIS

- **CI (Correlation Index)**: A measure of the correlation between predicted and true values, with its standard deviation (SD) in parentheses.
- **rm^2(Coefficient of Determination)**: A measure of how well the predicted values explain the variance in the true values, with its SD in parentheses.
- **MSE (Mean Squared Error)**: A measure of the average squared difference between predicted and true values.

Table 1: Performance comparison of different models on Davis dataset.

Model	CI (SD)	MSE	r_m^2 (SD)
KronRLS	0.871 (0.001)	0.379	0.407 (0.005)
SimBoost	0.872 (0.002)	0.282	0.644 (0.006)
DeepDTA	0.878 (0.004)	0.261	0.630 (0.017)
DeepCDA	0.891 (0.003)	0.248	0.649 (0.009)
TEFDTA	**0.890 (0.002)**	**0.199**	**0.756 (0.008)**
TEFDTA Replication	**0.877**	**0.217**	**0.723**

Table 2: Performance comparison of different models on KIBA dataset.

Model	CI (SD)	MSE	r_m^2 (SD)
KronRLS	0.782 (0.001)	0.411	0.342 (0.001)
SimBoost	0.836 (0.001)	0.222	0.629 (0.007)
DeepDTA	0.863 (0.002)	0.194	0.673 (0.009)
DeepCDA	0.889 (0.002)	**0.176**	0.682 (0.008)
TEFDTA	0.860 (0.001)	0.184	**0.731 (0.006)**
TEFDTA Replication	**0.850**	**0.206**	**0.727**

3– DISCUSION

DISCUSION

VALIDATION OF TEFDTA MODEL:

- RESULTS CLOSELY MATCH THE ORIGINAL PAPER, CONFIRMING THE REPRODUCIBILITY AND ROBUSTNESS OF THE MODEL.
- TEFDTA EFFECTIVELY PREDICTS BOTH COVALENT AND NON-COVALENT INTERACTIONS.

MINOR DEVIATIONS:

SLIGHT DIFFERENCES IN METRICS DUE TO:

- RANDOM INITIALIZATION OF WEIGHTS.
- HARDWARE/SOFTWARE VARIABILITY.
- FEWER TRAINING EPOCHS.

IMPACT ON DRUG DISCOVERY:

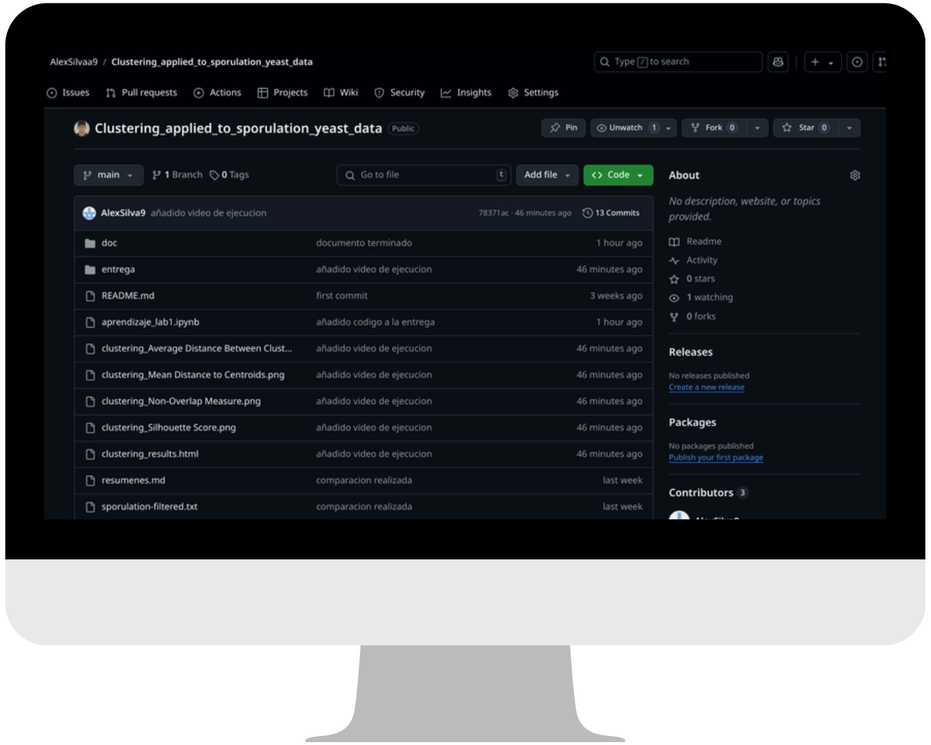
- DEMONSTRATED SENSITIVITY TO COVALENT BINDING INTERACTIONS, A CRUCIAL AREA IN MODERN DRUG DESIGN.
- ENHANCED GENERALIZABILITY ACROSS DATASETS, HIGHLIGHTING ITS UTILITY FOR BROADER APPLICATIONS.

3 – REPOSITORY ACCESS

REPOSITORY ACCESS

- ALL ADDITIONAL INFORMATION, INCLUDING SOURCE CODE AND FULL DOCUMENTATION, IS AVAILABLE IN THE GITHUB REPOSITORY:

https://github.com/AlexSilvaa9/artificial_intelligence_final_project




THANK
YOU