

1. Project Information

Project Title: Artificial Intelligence based approach for Predicting Drug-Target Binding Affinity

Team Members:

Mohamed Ayman Elsayed Abdelghany (240103514)

AbdulAzim Alaa Abdul Azim Ali (230200005)

Abdallah Mohamed Saad Mahmoud (240100184)

Mohamed Ashraf Abdelmoneem Abdelaal (240103481)

Abdallah Ahmed Saad Mohamed Soliman (240103457)

Omar Osama Saeid Abo Al kaseem (230105621)

Team Leader: Abdallah Mohamed Saad Mahmoud

Supervisor: Dr. Dalia Ezzat

Email: dalia.ezzat@sut.edu.eg



2. Project Overview

This project proposes the development of an artificial intelligence to predict drug-target binding affinity (DTA), a critical factor in drug discovery. The high cost and lengthy timeline of traditional drug development are largely due to the inefficient experimental screening of millions of compounds. Our solution aims to address this by creating a deep learning model that can accurately predict the binding strength between a drug molecule and a protein target computationally. The model will leverage advanced neural network architectures, such as Graph Neural Networks (GNNs) for encoding molecular structures and Convolutional Neural Networks (CNNs) or Transformers for processing protein sequences. By learning from large-scale biochemical datasets, the model will learn to identify complex patterns that determine binding. The outcome will be a predictive tool that can prioritize the most promising drug candidates for further laboratory testing, thereby accelerating the early stages of drug discovery.

3. Problem Statement & Market Need & Industry collaboration

The process of discovering new drugs is notoriously inefficient, often exceeding 10 years and \$2 billion in costs. A significant portion of this effort is spent on experimentally testing thousands of compounds for their ability to bind to a specific disease-related protein target, with a high failure rate. This represents a major gap in the pharmaceutical R&D pipeline.

There is an urgent market need for computational tools that can accurately predict binding affinity to serve as a high-throughput pre-screening filter. This directly addresses the industry's need to reduce costs, accelerate timelines, and improve the probability of success. An accurate AI-based DTA prediction tool has immense industrial relevance globally and can be a catalyst for innovation in local biotechnology and pharmaceutical sectors, potentially leading to research collaborations and positioning the team at the forefront of AI-driven healthcare technology. Explain the industrial or societal relevance such as relevance to Egypt's energy, technology, or manufacturing sector.

4. Objectives

- Select appropriate public-domain datasets such as BindingDB, and KIBA for model training and validation.
- Design and implement a deep learning architecture that effectively represents and integrates features from both drug compounds (as graphs or SMILES strings) and protein targets (as sequences).
- Train, validate, and optimize the model to achieve competitive performance against established benchmarks, using metrics such as Mean Squared Error (MSE).



- Develop a basic functional prototype, such as a Python script or a simple web application demo, that can accept input data and return a predicted affinity value.

5. Innovation & Expected Impact

The proposed approach innovates by leveraging modern deep learning architectures (like GNNs) that are particularly well-suited for capturing the structural intricacies of molecules and the functional domains of proteins, potentially leading to more accurate and generalizable predictions than traditional quantitative structure-activity relationship (QSAR) models. The proposed approach has a scientific impact and societal impact by accelerating drug discovery, the technology can contribute to bringing new treatments for diseases to patients faster.

6. Methodology & Technical Approach

The setup of the proposed approach will follow a standard machine learning workflow:

- (A) Data Collection & Preprocessing: Acquire datasets from public databases. Clean the data, handle missing values, and standardize affinity measurements (e.g., Kd, Ki, IC50). Split the data into training, validation, and test sets.

(B) Feature Engineering:

Drug Representation: Represent drug molecules as molecular graphs (using GNNs) or as tokenized SMILES strings (using NLP techniques).

Protein Representation: Represent protein targets as amino acid sequences, encoded and processed by 1D-CNNs or Recurrent Neural Networks (RNNs)/Transformers to capture sequential patterns.

- (C) Model Architecture: A dual-input neural network will be designed. One branch processes the drug representation, the other processes the protein representation. The outputs of these branches will be concatenated and fed into fully connected layers to perform the final regression prediction for affinity.

- (D) Tools & Technologies: Python, TensorFlow, deep learning libraries for graphs such as scikit-learn, pandas, NumPy, and Colab Notebooks for development.



7. Project Deliverables

- (A) A trained and validated deep learning model for DTA prediction.
- (B) A comprehensive final report detailing the literature review, methodology, experiments, and results.
- (C) A clean, well-documented, and version-controlled code repository (such as on GitHub).
- (D) A presentation summarizing the project's motivation, approach, key findings, and demonstration of the prototype.
- (E) A simple software prototype demonstrates the model's functionality.

8. Timeline

