Exploring the Effectiveness of Graph Neural Networks in Predicting Candidate Drugs

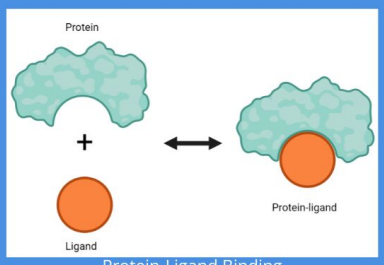
By: Ayam Babu, Supervisor: Dr. John Mitchell

University of St Andrews

LAIDLAW FOUNDATION

Introduction:

- It costs approximately $1 Billion to bring a new drug to market; this needs to be reduced
- Genetic diseases often lead to the production of proteins that can't perform their usual functions
- Drugs (also called a "ligand") bind to sites on malformed proteins and regulates its functions
- I created an AI tool that:
  - Takes: proposed drug molecules and a disease protein
  - Gives: whether the proposed drug molecules can treat the disease
- The tool helps scientists identify drug molecules to investigate further



Protein-Ligand Binding

Literature Review:

There are two main categories of methods:

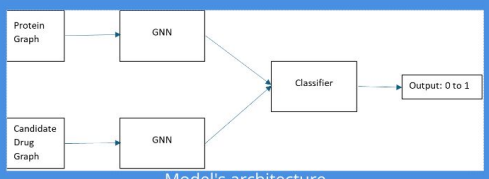
- Knowledge-based: specifically define how to calculate the likelihood of the protein and candidate drug binding. Example: DLIGAND2
- GNN-based: use AI to analyze and process the chemical structures of the candidate drug and protein and predict binding. Examples: Neural Fingerprinting (NF) and ParaVS-ND

Data:

- The Database of Useful (Docking) Decoys – Enhanced dataset (DUD-E) is used
- This dataset contains details of disease proteins and many associated candidate drug molecules, including those that won't treat the disease
- Protein and candidate drug molecules are represented as atoms with details for each atom, forming a graph
- The protein and candidate drug molecule graphs are paired together to create the base dataset for the model

Methodology:

- The DUD-E dataset is processed, and the protein and candidate drug molecule graphs are paired
- These pairs are split into a training dataset (for learning) and testing dataset (for evaluation)
- Paired graphs from the training dataset are fed into two different Graph Neural Networks (GNNs) that embed the graphs as vectors (more readable format for the classifier)
- These vectors are fed into a classifier, which outputs a value between 0 and 1, indicating whether the drug molecule should be investigated further
- The model's output is compared with the expected result and parameters are adjusted with feedback
- The model, once fully trained, is evaluated on the test dataset



Model's architecture

Results:

	ROC AUC	EF at 1%	EF at 2%	EF at 5%	EF at 10%	EF at 20%	EF max
TwinGNN	0.82	22.36	15.97	9.67	6.43	4.08	35.99
DLIGAND2	0.77	6.67	N/A	3.31	2.55	N/A	N/A
NF	0.90	N/A	N/A	N/A	N/A	N/A	N/A
ParaVS-ND	0.98	N/A	39.7	N/A	N/A	4.9	62.6

TwinGNN refers to this project's model. ROC AUC measures a model's ability to differentiate between positive and negative cases (higher is better). EF at x% measures the accuracy of the model's top x% drug candidates.

- ParaVS-ND and NF outperform TwinGNN partly because they focus on parts of protein relevant to candidate drugs
- TwinGNN outperforms DLIGAND2 because DLIGAND2 simplifies factors necessary for binding, missing key information

Conclusion and Further Research:

- In the future, the model could be tested on other datasets
- The model could also be improved by using specific features of a protein and candidate drug, rather than using entire molecules

Acknowledgement:

I am deeply grateful to Lord Laidlaw and the team at the Laidlaw Foundation for the research and leadership development opportunities through the Laidlaw Scholars program. I also thank Dr. John Mitchell, my research supervisor, for his valuable time and advice. Finally, I would like to thank my family for their support and flexibility, which enabled me to pursue my Laidlaw plans.