Deep neural networking for Drug Discovery

Name: Arushi Tejpal
student ID: 28130006
Tutor: Monday 5pm Tooba Jalalidi

Table of Contents

1. PROJECT DESCRIPTION:

Introduction

From the evolution deep neural networking methodology we are able to revolutionize the complex, time consuming and often unsuccessful process of drug discovery.

A factor which decreases the accuracy of the drug discovery process is the lack of precise knowledge of the three dimensional structure of the compounds and target molecules (seen in fig 1.). The binding affinity and kinetics of the molecule is what determine effects of the drugs. Previously, many computational methods have been used for drug discovery, however, have not been successful with its accuracy.

Biological molecules in our body rely on molecular interactions and their ability to bind with one another. This important attribute of our biological system is a curtail way to understand diseases and further create drugs which help improve these interactions or limit harmful interactions. Furthermore, the binding affinity and kinetics of the molecule is what determine effects of the drugs.

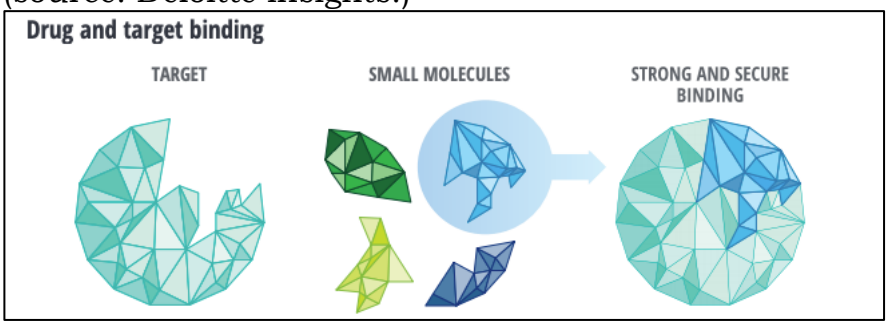
Through Deep neural networking we are able to extract relationships between data and learn independently from patterns of data. Using deep convolutional neural networking (CNN) techniques for binding affinity predictions will further provide accurate relationships with biological molecules. In this study we will look into how the CNN technique of deep learning will enable us identify potential therapeutics for any disease target.

Benefits:

- Provide accurate drugs for disease.
- Reduces cost through efficiency
- Increase accuracy of predictions on the safety and effect of drugs.
- Allow a new variation of drugs into the market
- Ability to find patterns and insights in data that medical chemists cannot.

Convolutional neural network (CNN) uses a class of deep learning applied in analysing visual imagery. We will be using this technique to identify binding sites of molecules.

Figure 1:
The drug (small molecules) and target molecule binding to target site. (source: Deloitte insights.)



1.1 Proposed Model

The Aim

The aim of this model is to implement deep neural networking approaches such as the convolutional neural network (CNN) to accurately detect therapeutic methods for disease. The CNN will enable us to identify predications of biological binding of target molecule and potential therapeutic drugs.

Data Science Roles and responsibilities

- **Data Engineer/ Data Scientist:**
  Strong database and software engineering skills, ideally in the scientific field such as cheminformatics or bioinformatics. Contributing to the design and development of datasets for machine learning, construct analysis which inform machine learning directions for drug discovery research.
- **Machine learning scientists:** have the ability to perform predictions and analysis of the safety drugs and effect of drugs.
- **Bioinformatics experts:** Needed for their expertise in the knowledge of protein structure, protein analysis along with the understanding of biological data.
- **Data business person:** Relies on decision making regarding business model and plans.

2. BUSINESS MODEL

AI and Data science is able to intelligently search vast amounts of data sets, scientific publications and clinical trial data which in turn accelerates the process and allows predictive analysis.

- Through computer aided data analysis we are able to identify spatial structure of target molecules.
- Data scientists collaboration with biochemists is essential.
- Collaboration of big pharma organisation *is* critical in order for data collection.
- Collaboration of Big pharma organisations and using sources such as Pub Med for curating data.
- organisations is curtail to understand technical structure of model.

Challenges:

- Finding the ideal lock and key complex proteins.
- Having accurate and a broad range of data to conduct accurate predictions.
- The CNN needs a powerful GPU.
- Filling gaps in existing data sets cost money and time.

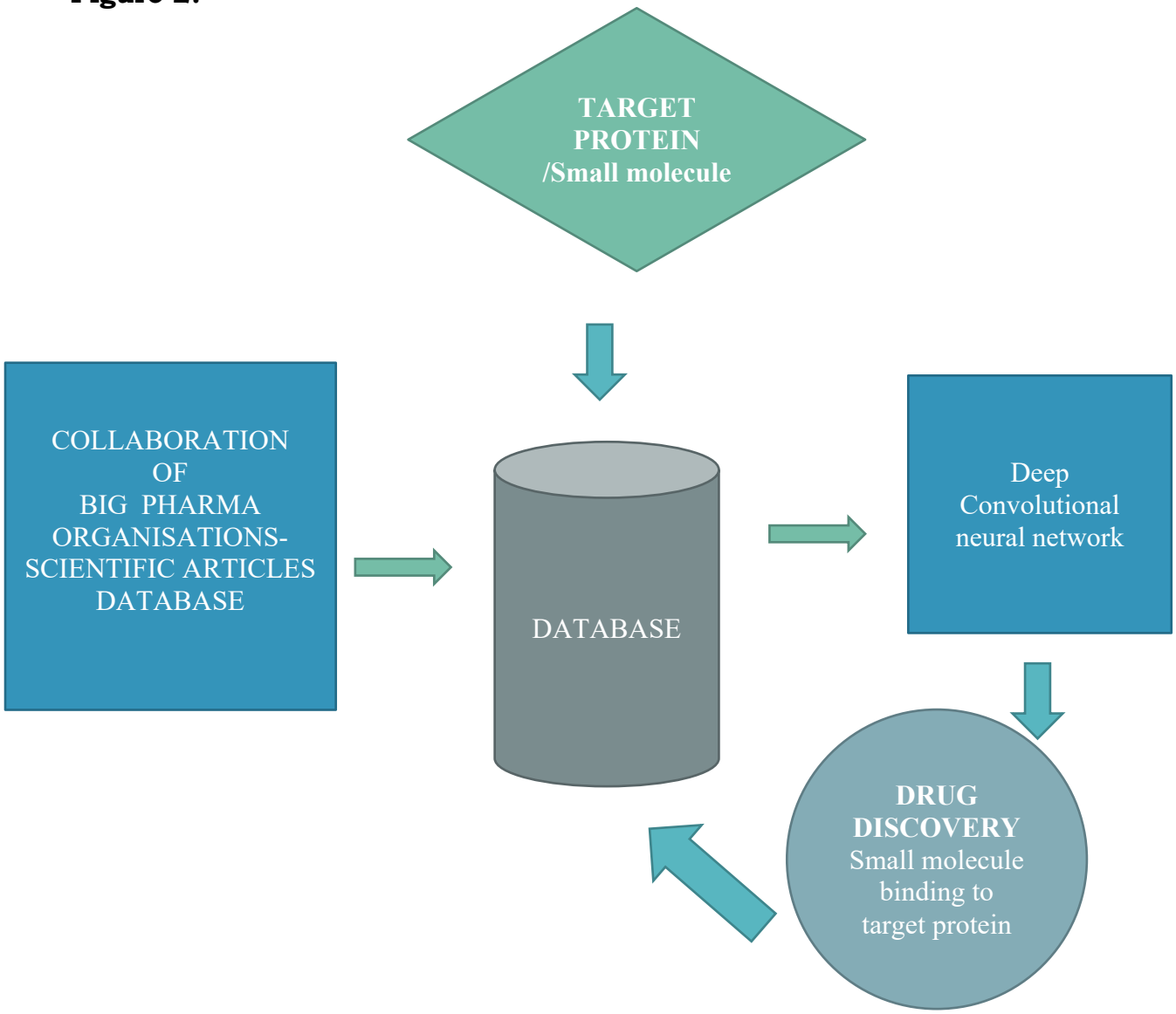
- **Data standardization:** data processes are different from company to company. To get generate and extract data we need to collaborate with different companies and organisations through a unified database platform. Furthermore , through different company collaboration the data quality and data value my vary.
- **Employee training and development:** chemists and biochemists need to be aware of biotechnologies and machine learning technologies to be able to analyse data. Organisations need to invest in training and development so they are aware of AI and data technologies.

Core Values:

- Integrity : There should be integrity within the data used for drug discovery within the business.
- Collaboration : Collaboration within the business with all working towards one goal of drug discovery.
- Innovation: Forming innovative ideas through new technologies within AI.

BASIC PRINCIPLE:

Figure 2.



3. Characterising the Data and Data Processing

- Raw Data
- Metadata

- **Variety:**
  Biggest challenge is to combine the variety of data from different scientific articles, journals, organisations to gain insight and a solution of drug discovery.
- **Volume:**
  The amount of data accumulated from the different sources; May amount to gigabytes even terabytes through deep neural networking. As seen in fig 3.
- **Veracity:**
  How accurate is the data? The biological structure of molecule and their interactions need to be accurate in order to formulate successful prediction

- **Value:** How reliable is this data? This is a curtail point in order to identify which information and source is reliable to form a drug for an illness.

  We will be using deep neural networking to identify drug discovery . Previous method of docking and using QSAR used only bytes to conduct their drug discovery. Now the advanced method of deep learning and Deep neural networking can enable us to find possible therapeutic methods for diseases in a more efficient way.

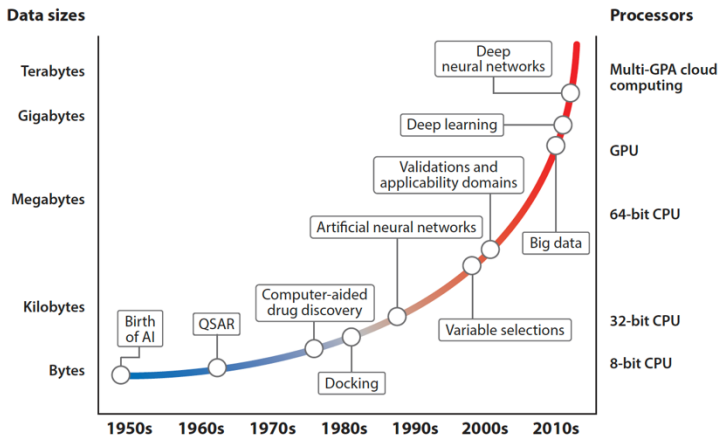


Figure 3. historical process of drug discovery through AI. (Sliowski, et al. 2015)

3.1 *Data Sets:*

1. *DUDE*

We will use DUDE (Directory of Useful Decoys Enhanced) benchmark which contains verified interactive molecules. DUDE is a structured based virtual screening method, containing diverse sets of active molecules for a set of target proteins.

- 22,886 active compounds and their 102 binding targets.
- 50 decoys for each active having similar physico-chemical properties but dissimilar 2-D topology. (Carpenter, 2018)

2. ChEMBL-20 PMD

ChEMBL: is a chemical database od bioactive molecules which have drug like properties. Chembl database contains data from scientific literature, public database, patents and depositded data sets from pharmaceutical companies. The chembl contains 18000,000 compounds as shown in the figure 3.1 (EMBL-EBI, 2020):

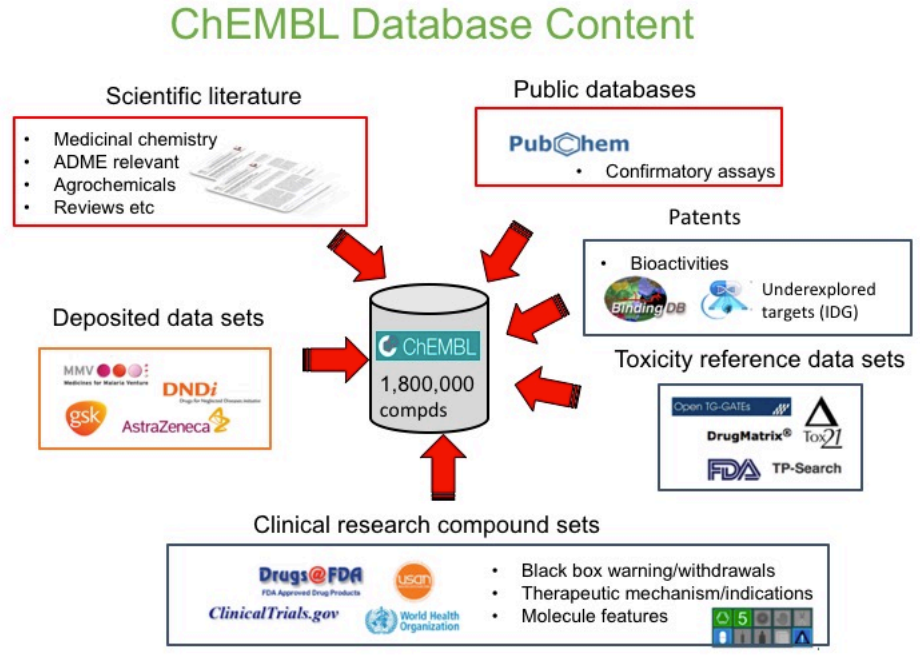
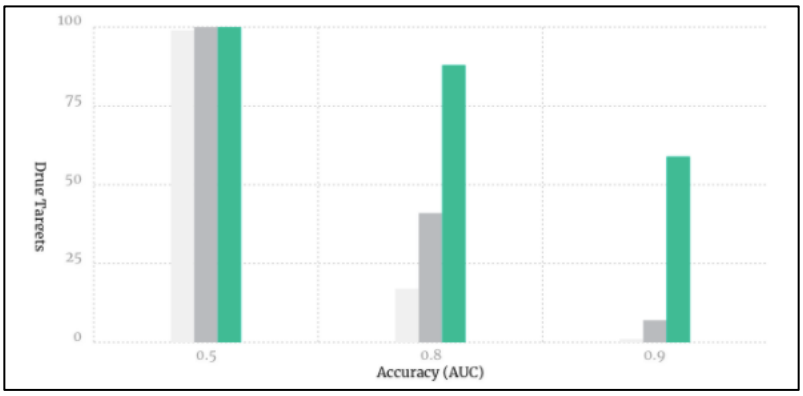


Figure 3.1 ChemBL database content (EMBL-EBI, 2020).



(AtomNet, 2015)

Figure 3.2, graph shows how the CNN model which uses the DUD-E data set to make over 1million predictions and compares the result with historical data with previous technologies used such as DOCK (light grey) and Autodock-sima (Dark grey). The CNN (green) model is far greater in terms of drug prediction compared to the previous methods.

4. Data Processing

Hardware:

- **GPU:** NVIDA V100 Tensor Core: for powerful graphic processing .

- **Apache Spark:** cluster- computing framework for large scale data processing.

Through image recognition we will be able to identify the protein molecules and their interactions this will enable us to understand and make predictions of possible drug therapies. Using a powerful and fast Graphic processing unit such as, **NVIDA V100 Tensor Core** can be used to create computer graphics for molecular dynamics, image segmentation, annotation and analysis.
The packages used for deep learning for CNN are open sourced such as:

- *TensorFlow: open source library for machine learning*
- *Caffe: deep learning framework*
- *PyTorch: open source library for machine learning*
- *Keras : open source library with python for neural netoworking.*
- *Theano: python library for mathematical expressions (Chen, 2018)*

Apache spark is a data processing mechanism and can be used for convolutional neural network modelling for images.

5. Data Analysis

Graph- structured Data:

Deep learning on graph structured data. This is used to graph data directly as input to the deep learning methodology.
CNN involving deep artificial neural networks to analyse visual imagery. Helps cluster images by similarity along with having image recognition as seen in fig 4.3 . Will help with recognising molecule structures and binding sites.
CNN is a combination of distinct layers which are stacked the layers include: convolutional layers, pooling layers and fully connected layers.

For drug target interaction we need data for:
Input: Compounds and protein / targets and drugs.
Output: number between 0-1 (if interaction affinity to target protein) scoring functions used to represent protein-ligand interactions (Ragoza, 2017)

5.1 Method

The target molecule and drug are paired randomly and the process repeats making different combination of drug and target molecule. If detection of interaction is successful 1 will appear if no detection of interaction 0 will appear.

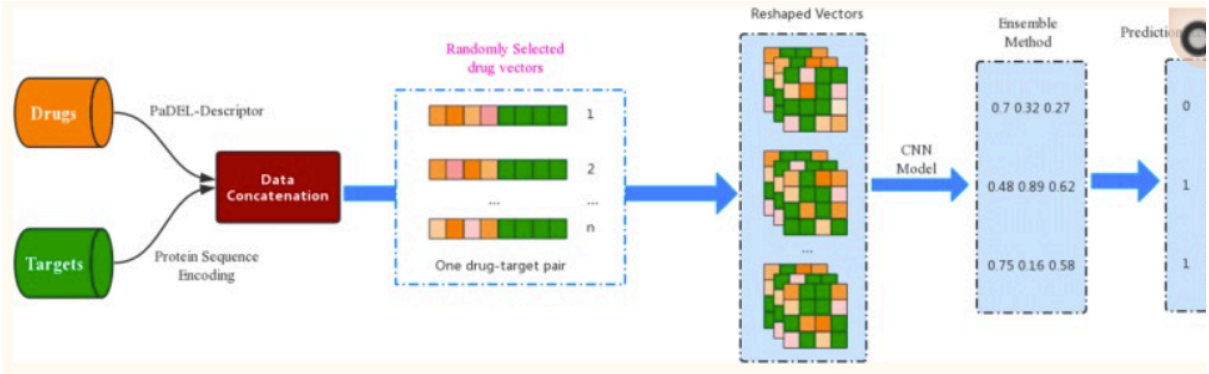
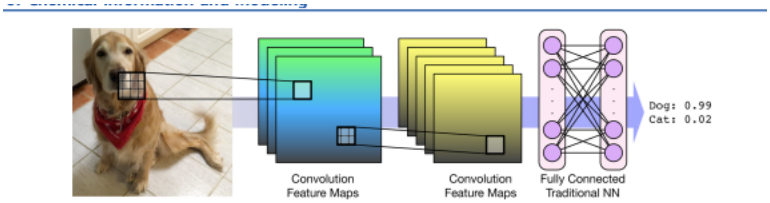


figure 4.2 : Convolutional network. (Hu, 2019)



(Figure 4.3: image recognition through CNN)

The powerful non-linear Convolutional Neural Networking is a tool used which enables image recognition within biological molecules. CNN scoring function automatically learns key features of protein ligand interactions what result in binding. The CNN learns complex features through layering on top of each other.
This method allows a new level of accuracy to drug discovery compared to previous time computing methods. The learning of accurate patterns is depended on how well structured the data is. The CNN can provide better outcomes and accuracy compared to machine learning algorithms.

6. DATA MANAGEMENT:

- **Policy for access and sharing:** There is intense competition between major pharmaceutical companies and thus results in limitation of data sharing. This causes a problem within the topic of drug discovery as acquiring more data will allow more drugs to be discovery. We will keep our data policy transparent for sharing data with pharmaceutical companies, this will allow the formation of partnerships with other pharmaceutical companies and thus allow more data to be accessible.

- **Data privacy:** The use of patient data is highly sensitive, the studies and experiments conducted need to monitor patient data

- **Data credibility:** The main challenges in the scientific community is the credibility of the scientific work. The data manager along with the bioinformatic experts will be responsible for the management of the credibility of biological data provided

7.Conclusion:

Drug -target interaction is time consuming and laborious from the CNN-based model we are able to eliminate this challenge and provide a more efficient method for drug discovery. CNN- based model allows image recognition and a diverse range of databases to allow for accurate predictions of interactions to occur. This model will enable cost efficiency, potential to find new drugs and allow accurate predictions. Deep learning techniques is the future for the pharmaceutical industry and is needed in order to cure diseases and find new potential therapeutic methods.

8. YouTube Video:

https://youtu.be/lpo_HQtRbaQ

9. REFERENCES:

- Assignment 2.Proposal. Arushi Tejpal.2020.

- Accelerating drug discovery. Deloitte Insights. Retrieved from:
- https://www2.deloitte.com/content/dam/insights/us/articles/32961_intelligent-drug-discovery/DI_Intelligent-Drug-Discovery.pdf

- Wallach, I., Dzamba, M., Abraham, H. 2015. AtomNew: deep convolutional neural netowrki for bioactivity prediction.
- https://arxiv.org/pdf/1510.02855.pdf

- Fooladi, H. 2020.Deep learning in drug discovery. Review: deep learning in drug discovery. https://towardsdatascience.com/review-deep-learning-in-drug-discovery-f4c89e3321e1

- Sliowski, G., Kothiwale, S., Meiler, J. 2015. Computational methods in drug discovery.
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3880464/

- Ragoza, M., Hochili, J., Idrobo, E. 2017. Protein-ligand scoring with convolutional neural netoworks.
- https://pubs.acs.org/doi/10.1021/acs.jcim.6b00740

- EML-EBI. What is chemBL?.2020. Retrived from:
- https://www.ebi.ac.uk/training-beta/online/courses/chembl-quick-tour/what-is-chembl/

- Atomwise.2020. introducing AtomNet.https://blog.atomwise.com/introducing-atomnet-drug-design

- Carpenter, K., Cohen, D., Jarrell, J. 2018 Deep Learning and virtual drug screening. https://www.atomwise.com/wp-content/uploads/2018/11/Carpenter-et-al.pdf