Machine Learning
+
Drug-Protein Interaction Prediction

Team - Superwise

SAMHAR COVID-19

The Problem

A Lead molecule is a small drug-like molecule that is expected to interact with a specific protein (target) in a specific way.

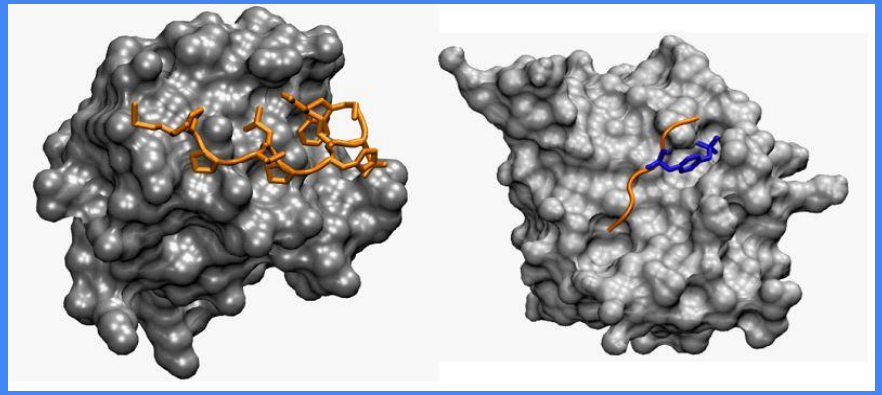
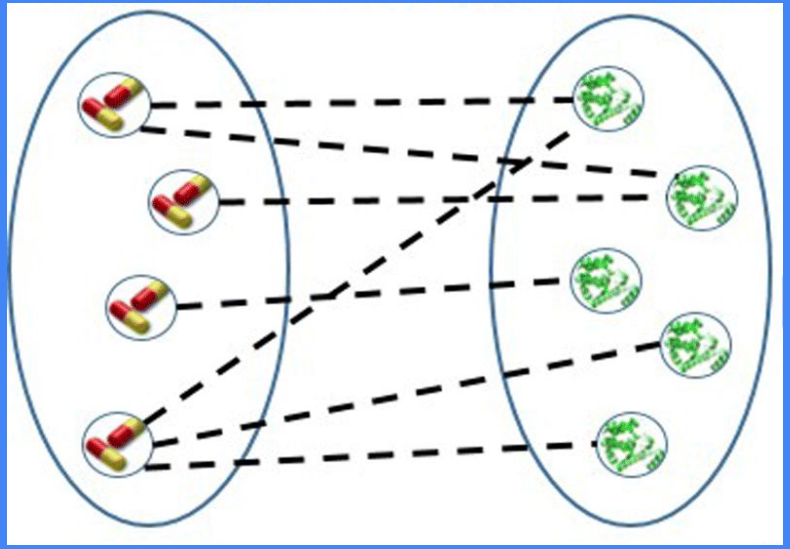


Image:https://www.intechopen.com/books/binding-protein/protein-peptide-interactions-revolutionize-drug-development

These Lead molecules can have various effects ranging from inhibiting viral replication, to stopping vital processes of pathogenic organisms.

The Problem

Large Databases for Molecules (Drugs), Targets, and their Interaction Affinities are freely available.



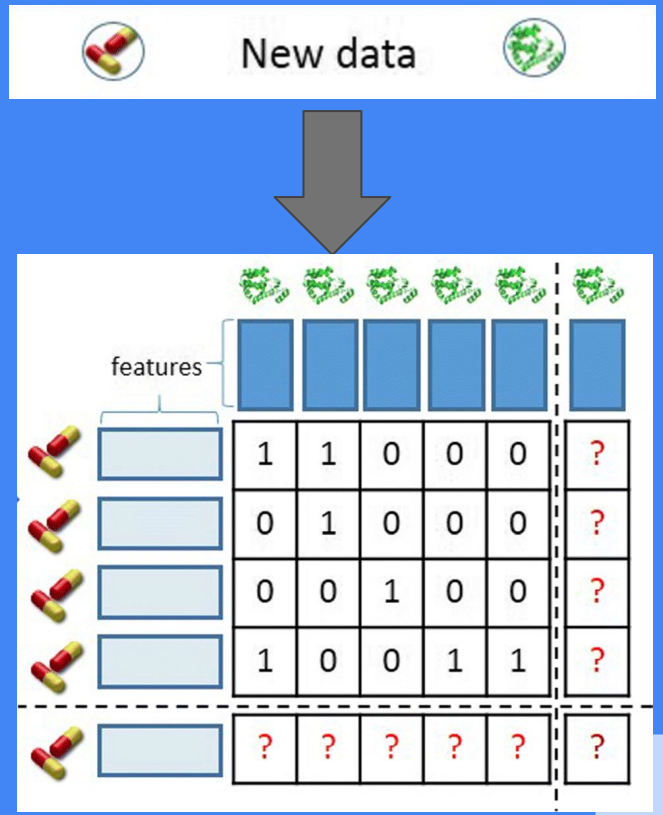
A Drug-Target Interaction Network. A Bipartite Graph, with one set of Nodes as Drug Molecules, the other as the Target Protein.

The Problem

What If we want to see if a NEW Target has affinity to any existing drug? Such as SARS-CoV-2 Enzymes?

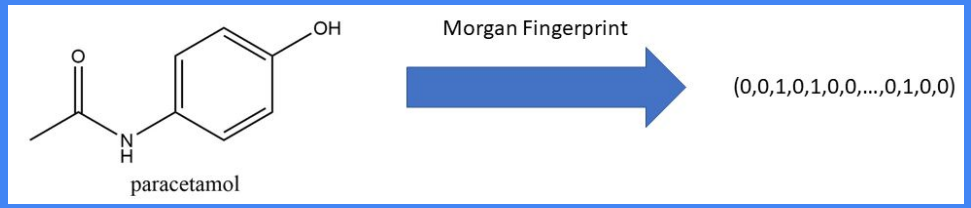
Or If we want to see if a NEW Drug has an affinity to any existing Target?

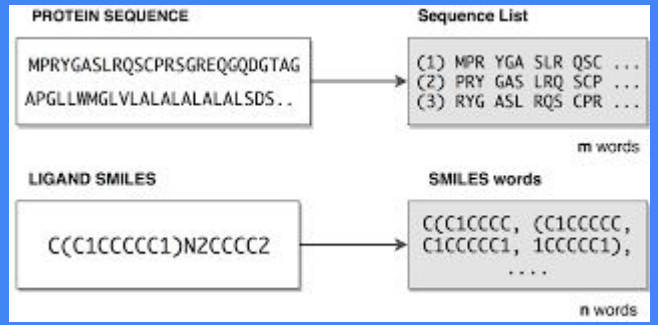
Or If we want to see if a new Drug has an affinity to a NEW Target? Such as SARS-CoV-2 Enzymes?



The Formulation

- Drug Molecules and Protein Sequences can be represented digitally as either in Text or as a Graph.
- The Inputs:
  - Molecule Fingerprint
  - Protein Sequence
- The Output:
  - Interaction Affinity (0/1)
- Hence this problem can be seen as a Binary Classification Problem.
- There can be multiple Outputs, to capture additional information such as Side effects.





The Data

The Binding Database

BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of protein considered to be drug-targets with small, drug-like molecules. BindingDB contains 1,854,767 binding data, for 7,493 protein targets and 820,433 small molecules.

From here, We extract Positive pairs, the Drug-Target Pair which have some interactions.
And then assume all other possible pairs to be Negative pairs, i.e. have no affinity.

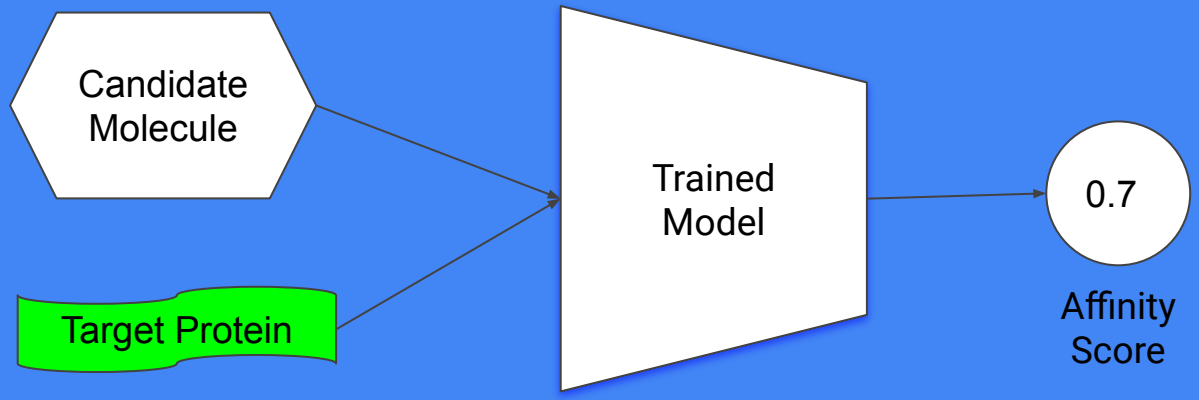
We divide this dataset into three parts randomly into

- Training set
- Validation set
- And Testing set

The Data

MOSES Dataset

MOSES is a Database of over 2 Million molecules. Some of the molecules are drugs, but majority of them are generated. This Dataset will be used to find possible Lead molecules for COVID-19.



The Expected Outcomes

Once a Model has been trained and evaluated:

- Drug Repurposing
  - Existing Drugs can be tested for their affinity towards new Targets (Proteins).
  - Drugs that are already in the market could be used for treating COVID-19.
  - Remdesiver and HCQ are few of the already existing drugs that are being tested for COVID-19 treatment.

- Drug Discovery
  - This Model can be paired up with another Generative Model that can generate new Molecules that could potentially be used as a Drug.
  - This can also be used with Human designed molecules. But this allows for extremely fast prototyping.

- Bringing down time & cost of experimentation of Candidate Drugs.