Reverse Drug Discovery Web Portal (RDD)

Presented by:

Brian Loh

bloh@swinburne.edu.my

About Us

- Created as part of final year project
  - Focus on addressing real world problems through software solutions
  - Interact with industry client to solve project requirements
- Project team consists of:
  - Chin Jing Jie
  - Liew Woun Kai
  - Nicholas Lim Tun Yang
- Stakeholders
  - Dr. Xavier Chee
  - Professor Ts. Dr. Lau Bee Theng
  - Dr. Brian Loh
  - Professor Dr. Patrick Then

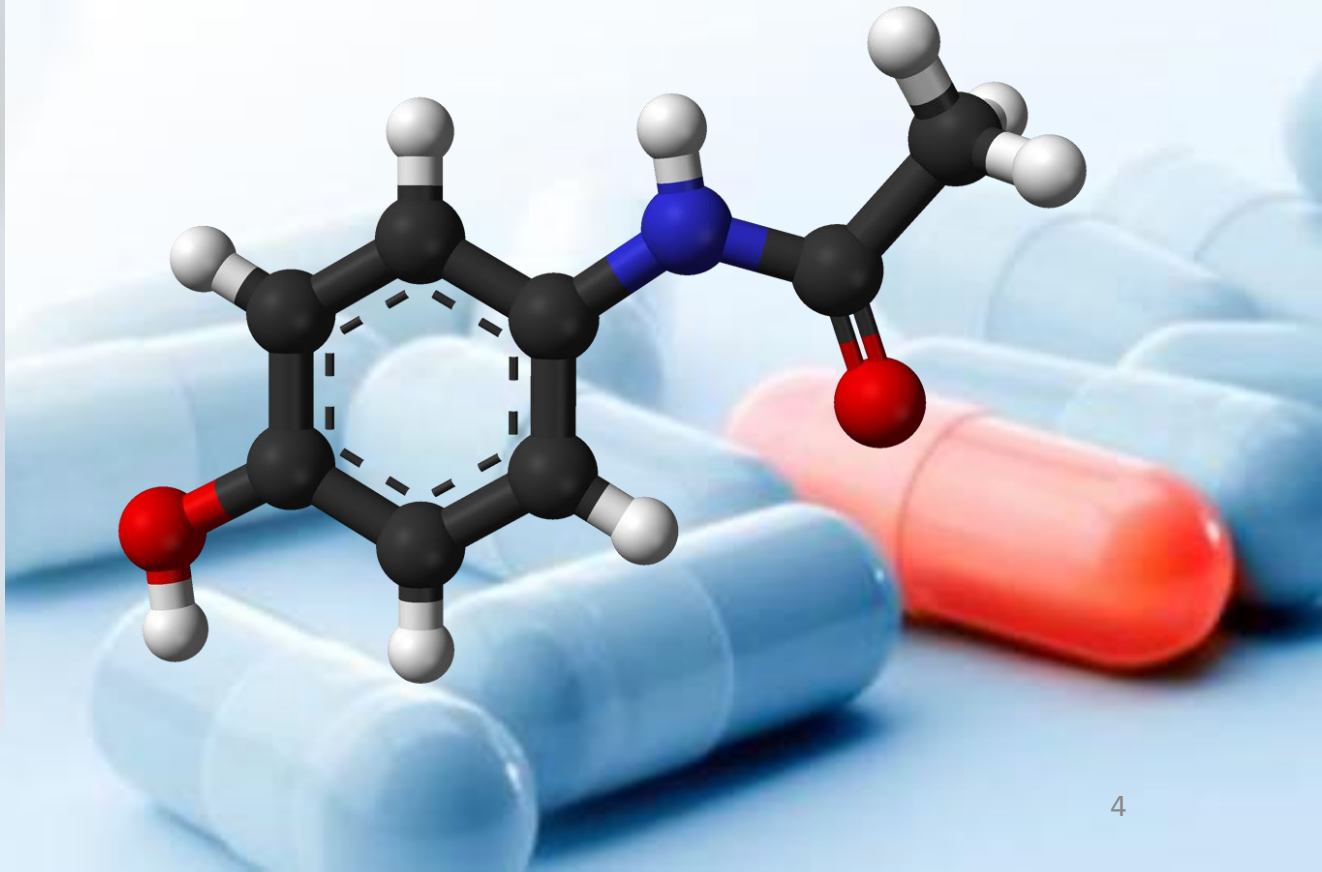
Overview

- Background
- Proposed Solution
- Experiments & results
- Conclusion

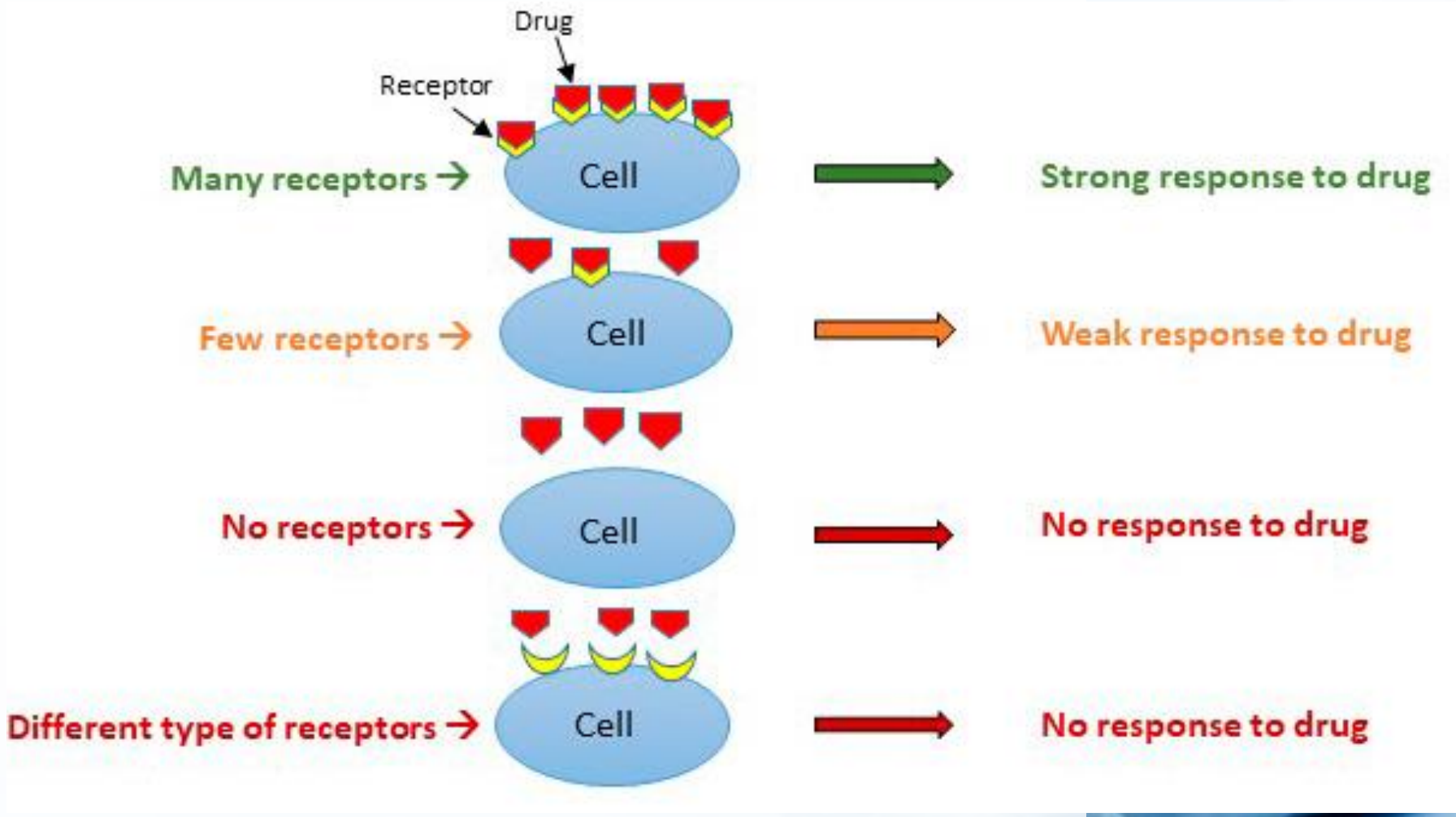
How Do Drugs Work?

Paracetamol Drug

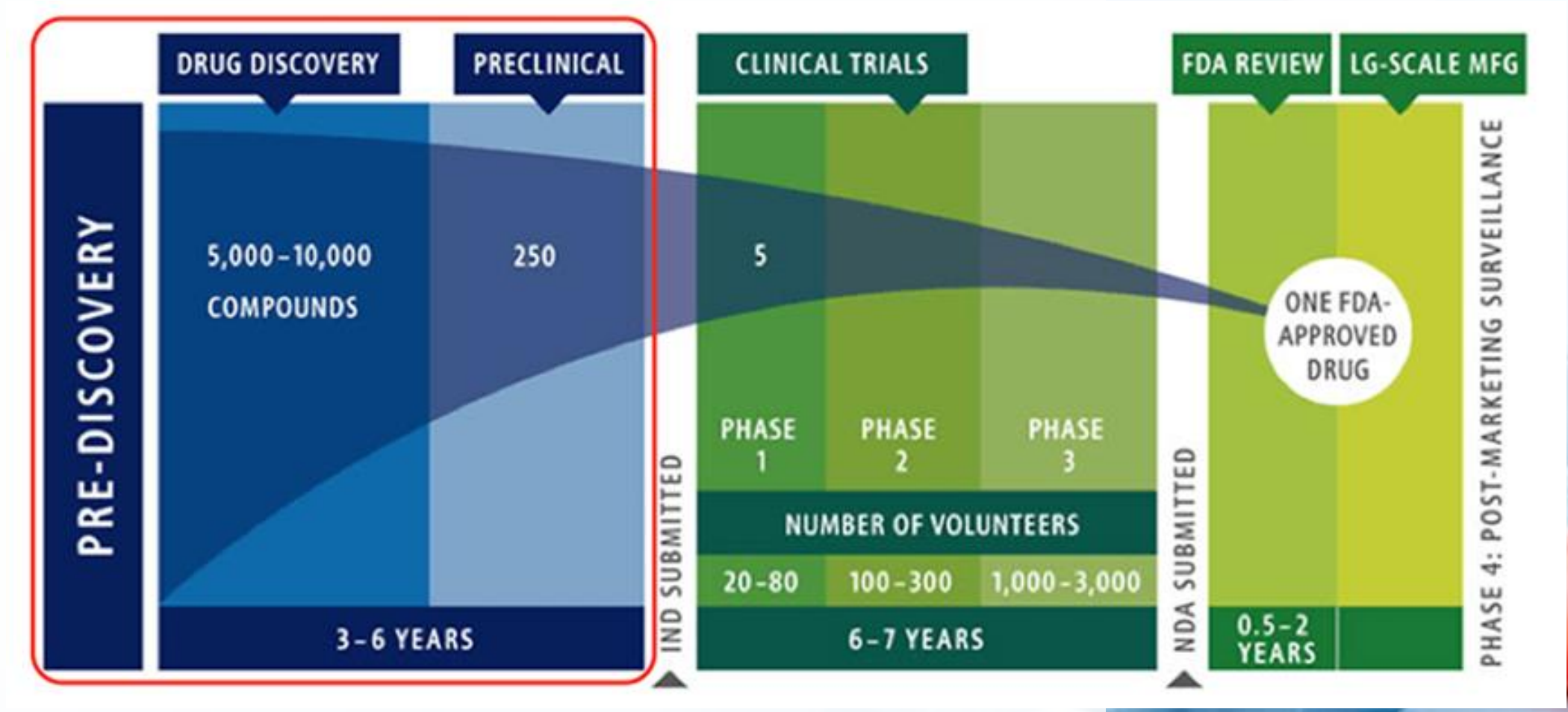
Paracetamol Molecule (C8H9NO2)



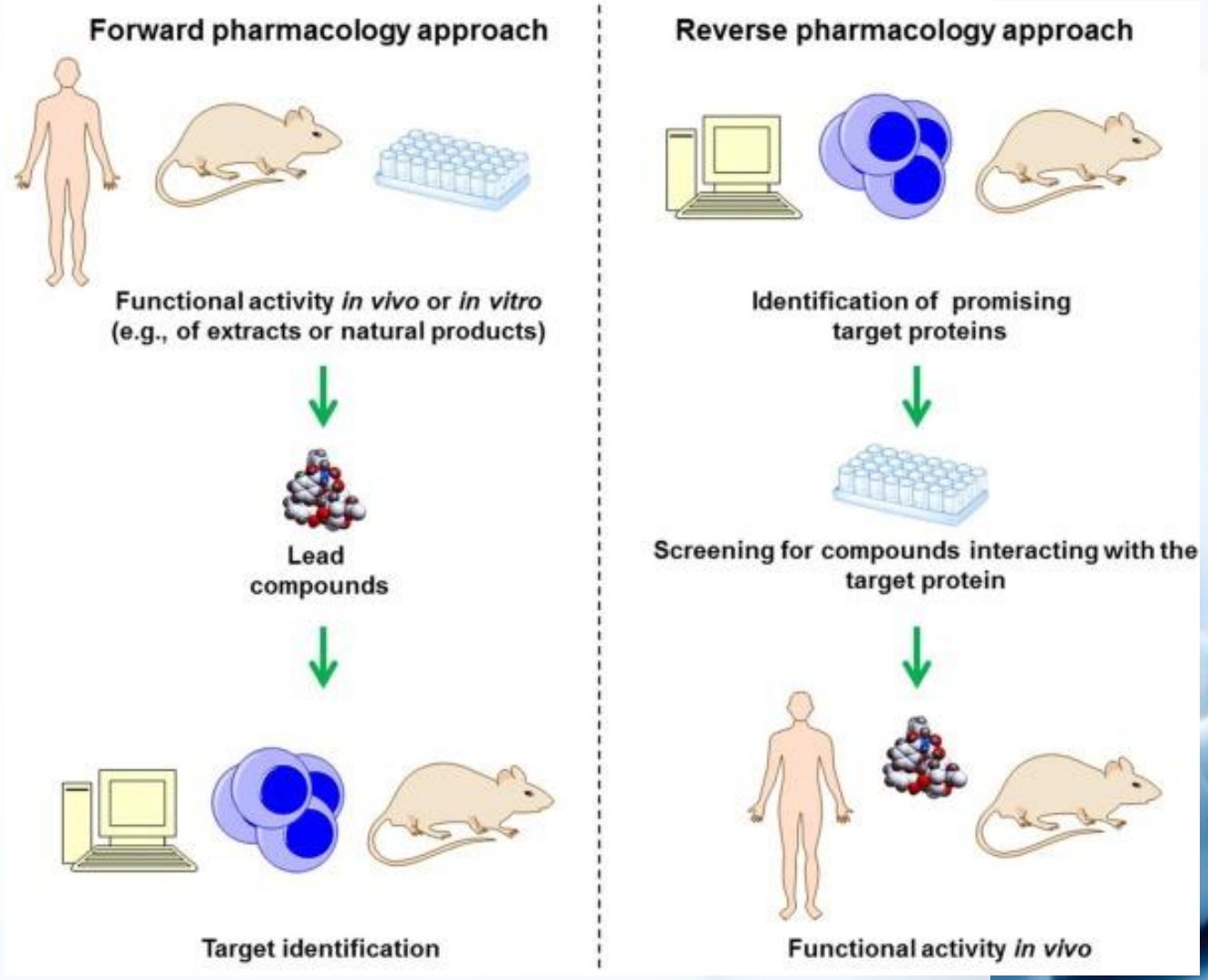
How Do Drugs Work?



Drug Discovery Pipeline

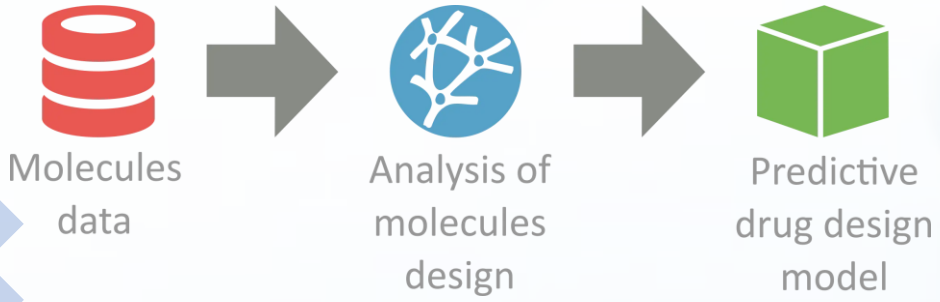


Reverse Drug Discovery



Drug Discovery with Machine Learning

- Simplified molecular-input line-entry system (SMILES)
  - Describes chemical compound structure using short ASCII strings
    - Paracetamol (C8H9NO2), SMILES: *CC(=O)Nc1ccc(O)cc1*
- Quantitative structure-activity relationship (QSAR)
  - Computational modelling method for revealing relationships between structural properties of chemical compounds and biological activities



SMILES	Class
Clc1cccc(N2CCNCC2)c1	0
COc1cccc([N+](=O)[O-])c1	0
CCC(=O)CC(C)CC	1

ABC	ABCGG	SpAbs_A	SpMax_A
6.54	6.23	11.18	2.19
2.44	2.68	3.46	1.73
3.93	4.24	12.09	2.24

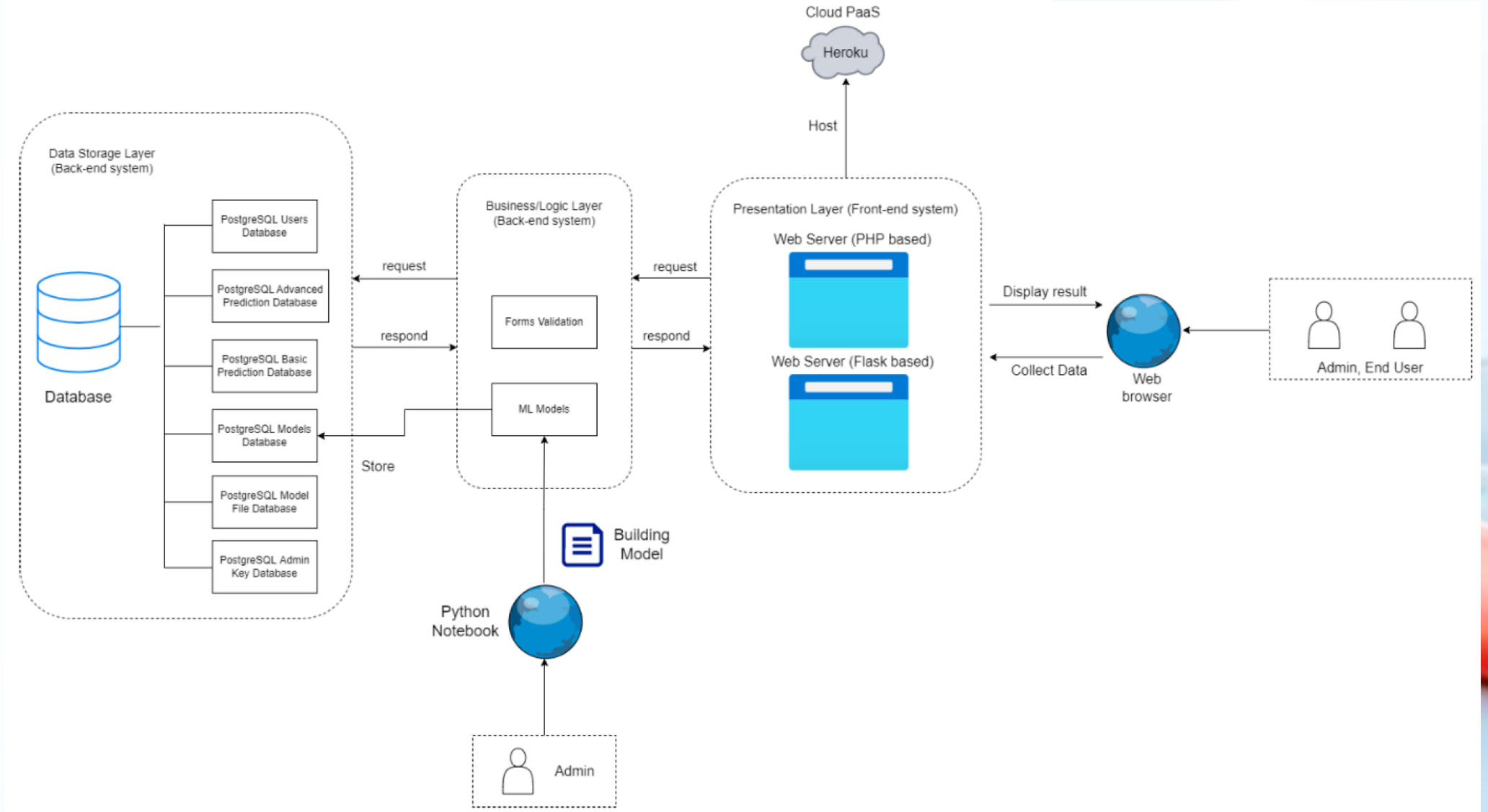
Fingerprint
1101000101
1010011101
1110010110

Proposed Solution

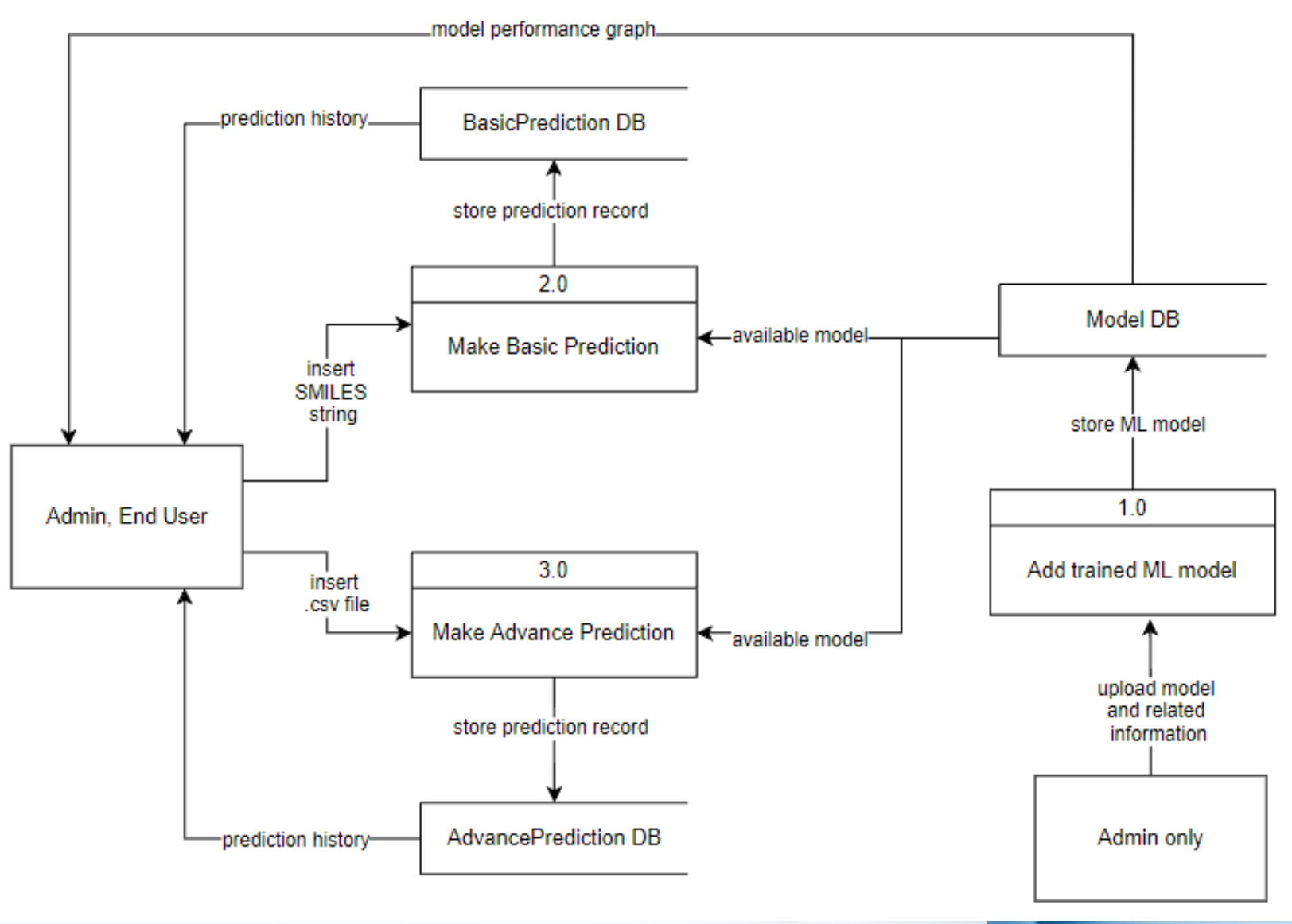
- Drug discovery is a difficult and time-consuming process
- Machine learning approaches can be difficult to implement

- Reverse Drug Discovery Web Portal (RDD)
  - Low-cost cloud-based system
  - Enables drug discovery with machine learning

System Architecture



System Process Flow



Reverse Drug Discovery

Prediction History Profile Logout

Machine Learning Reverse Drug Discovery

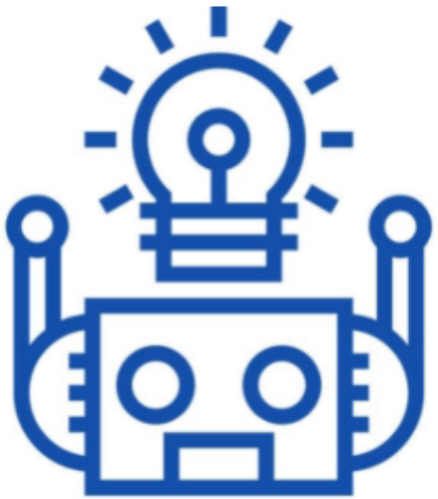
Basic Prediction

Enter smiles string

Select Target Disease

Select Specific Model

Predict



Experiments

- Datasets
  - HIV and COVID
- SMILES are described in 3 forms
  - Molecular descriptor (MD)
  - Chemical Fingerprints (FP)
  - MD + FP
- Data preprocessing
  - Remove records with missing information
  - Perform under-sampling / oversampling
  - Data normalization
  - Dimensionality reduction
- 11 different classification models

Results

HIV	Classification Models	Accuracy (Training)	Accuracy (Testing)
1.	XGB Classifier model	0.99	0.84
2.	Gradient Boosting Classifier model	0.91	0.83
3.	Ada Boost Classifier	0.81	0.79
4.	Logistic Regression	0.73	0.77
5.	Gaussian NB	0.72	0.77
6.	Ridge Classifier	0.73	0.78
7.	K-Nearest Neighbors Classifier	0.93	0.84
8.	SVC	0.85	0.84
9.	Bagging Classifier	1.00	0.85
10.	**Random Forest Classifier**	**1.00**	**0.86**
11.	Stacking Classifier model	1.00	0.85

COVID	Classification Models	Accuracy (Training)	Accuracy (Testing)
1.	**XGB Classifier model**	**1.00**	**0.95**
2.	Gradient Boosting Classifier model	0.94	0.84
3.	Ada Boost Classifier	0.68	0.66
4.	Logistic Regression	0.55	0.56
5.	Gaussian NB	0.56	0.56
6.	Ridge Classifier	0.55	0.56
7.	K-Nearest Neighbors Classifier	0.91	0.84
8.	SVC	0.84	0.82
9.	Bagging Classifier	1.00	0.94
10.	**Random Forest Classifier**	**1.00**	**0.95**

Conclusion

- Reverse Drug Discovery Web Portal (RDD)
  - Assists in the drug discovery process by automating target identification
  - Provides a simple interface for managing processes
- Achieved by incorporating
  - Cloud technologies
  - Machine learning algorithms

Thank you

Q & A