TEKNOFEST

AEROSPACE AND TECHNOLOGY FESTIVAL

BIOTECHNOLOGY INNOVATION COMPETITION

PROJECT DETAIL REPORT

PROJECT CATEGORY FOR THE BACHELOR DEGREE AND UPPER-LEVEL

PROJECT NAME: A New Computational Tool for Drug Discovery : Case Study Alzheimer's Disease

TEAM NAME: Biolegends.AI

APPLICATION ID: 316795

MEMBERS:

Cyrille Mesue NJUME (Captain) - *Yildiz Technical University*

Atakan ÜNLÜ - *Yildiz Technical University*

CONSULTANT:

Asst.Prof. Alper YILMAZ - Yildiz Technical University

Table of Contents

1. Project Summary

Drug discovery is a challenging and expensive process that often takes years to complete. Though Computer-Aided-Drug-Design has reduced the discovery time, it is also met with various challenges such as low competency in data handling, small data size, large correlation in data, missing data and the inability or difficulty in explaining computational models. This project aims to develop a computational tool that employs data mining, preprocessing, and cleaning, feature engineering, machine learning for predicting the bioactivity of hit compounds, and molecular docking between the hit and target compound. While the molecular docking part is meant to confirm the predictions of our machine learning models through visualization, we plan to incorporate drug repurposing into our product to ease the process of lead compound discovery. This tool will help scientists researching Alzheimer's drugs, drugs for different carcinomas and other diseases to save time and money by running a check on the target to find bioactive drug candidates. The algorithm will also be made to provide automatic data mining from python supported client databases such as ChEMBL and NCBI, exploratory analysis and visualization, as well as building, training and optimizing machine learning models from scratch. Most importantly, our algorithms will be packaged into a user-friendly web application that can be accessed through the net. In short, we built a python API to mine and clean the dataset, build, train and optimize machine learning models, as well as build a web application where users can use our pretrained models to predict the bioactivity of certain compounds on their target compounds of interest, validate the drug using molecular docking

2. Problem

In a normal drug discovery pipeline, scientists may spend billions of dollars for upto 15 years and still end up not discovering a drug for a given disease. Efforts using computational approaches such as making use of drug databases , drug repurposing models, molecular docking and machine learning, has made it obvious that computational drug discovery will go a long way to saving humanity by reducing the drug discovery time and cost. Although these approaches show great promise in drug discovery, they have some major challenges. For instance, in a structure-based drug discovery paradigm (Batool et al., 2019) such as that employed in our research, molecular descriptors are often used to fit machine learning models which are used to predict the bioactivity of the compound on a given target. One major challenge is that there are hundreds of molecular descriptors, usually greater than the number of observable samples. A situation like this often makes it very hard to apply machine learning to predict because the models often overfit. Another problem is that most of the computational methods require expertise knowledge in computing to implement. Therefore, researchers who do not have coding experience nor experience in interpreting machine learning models may not be able to use these models. A common solution to solve the overfitting problem is to use more observable samples to train the models. However, biological data is often small limited in the number of samples (Zhang & Rajapakse, 2009, 1-43). Another challenge is that some of the data entries contain some missing values, and the performance of models may be poorly evaluated. Moreso, predicting a drug as bioactive is not enough to validate it for clinical trials. Our research aims to solve or contribute to a solution to most of these problems by building a web application with the possibility to predict the

bioactivity of compounds using machine learning, and validate compounds by molecular docking.

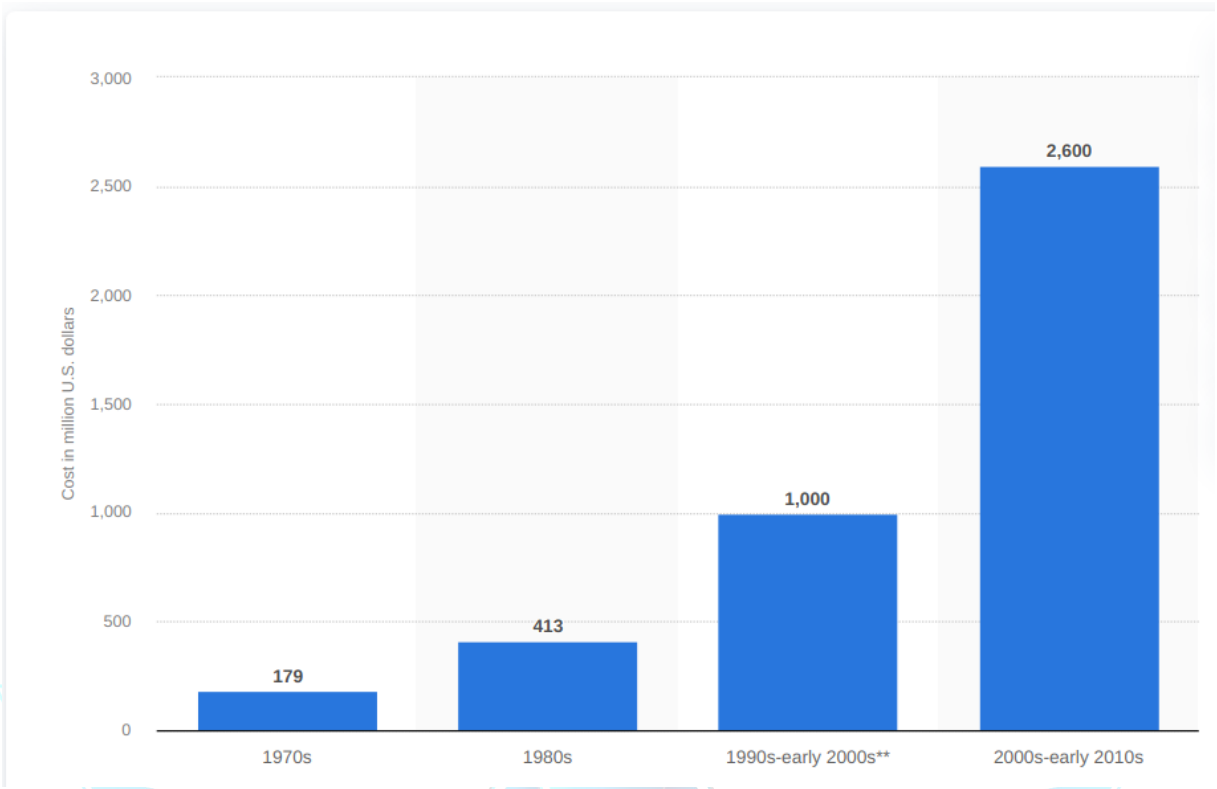

Figure 1. *Cost of developing a drug from the 1970s until 2010s ("Drug development cost United States 1970-2016 | Statista", 2020).*

3. Solution

The unceasing development of technology and the widespread use of artificial intelligence throughout the world has led to important developments in medicine as well as in other fields. In recent years, with the effective use of computational approaches in the field of medicine, diagnosis, treatment and follow-up of diseases are performed much faster and more efficiently compared to previous times (Asai, Konno, Taniguchi, Vecchione & Ishii, 2021). One of the aims of the use of artificial intelligence in healthcare is to reduce the cost and time of drug discovery and also to contribute to the production of much more effective drugs. The use of computational approaches is known to have great potential for drug discovery, but still has challenges that need to be developed. For this purpose, in our study, we will develop a web application to be used in drug discovery (especially for Alzheimer's disease) by combining the necessary computational approaches such as use of drug databases (ChEMBL), machine learning, drug repurposing and molecular docking. First, the data is obtained from the Chembl database and cleaned as explained in the data mining part. Next, with the cleaned data, machine learning models are used for determining the best type of molecular descriptor calculator and best feature selection methods. And, the best descriptors are used to train several classifiers which are evaluated using a stratified k-fold strategy (Kamber et al., 2011;

Manorathna, 2020). The best model is selected and optimized. Finally, to check the model's sensitivity, molecular docking is performed between active compounds and the target and based on the root mean square deviation (RMSD) score, the compound is validated as active or not. Our project pipeline can be summarized in figure 2 below.

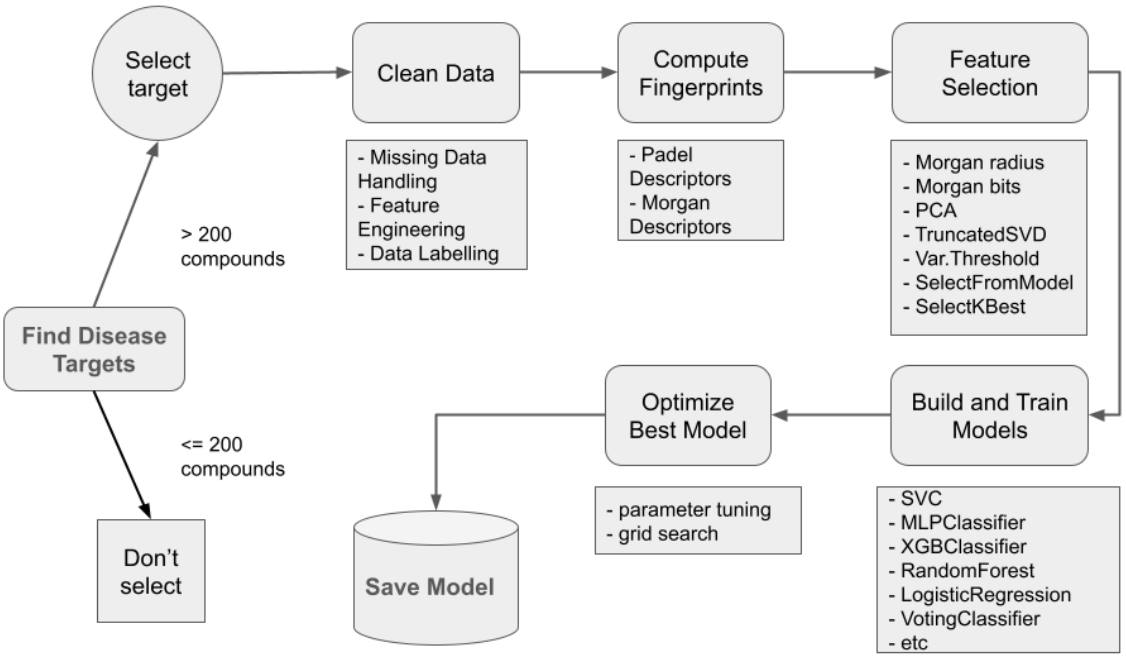

Figure 2 (a). *Pipeline in service development*

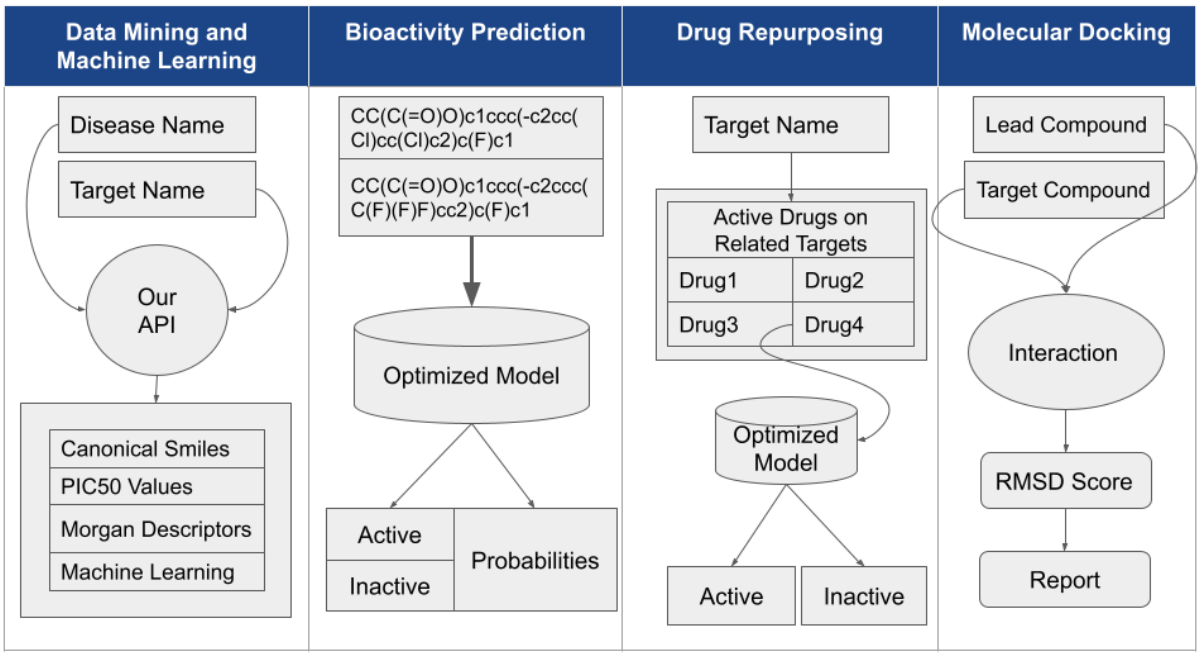

Figure 2 (b). *Pipeline in service provision*

4. Method

4.1 Data Mining

The data used for our project is obtained from the ChEMBL database (Mendez D et al, 2019), and is obtained by real-time data mining using the chembl client python API. With the chembl API, data can be obtained for diseases and targets by passing the disease name or target name to a class object in the API. The resulting data is raw, displayed as a pandas dataframe and may contain missing data entries. For machine learning purposes, the relevant columns are the columns that contain the molecule's canonical smiles and the standard bioactivity values (IC50) on the target in nano-molar. The canonical smiles can be transformed into vectors that represent the molecule's descriptors and these vectors can then be fit into machine learning models along with the bioactivity values. Therefore as a first step to cleaning the data, along with molecule IDs, only the canonical smiles and standard bioactivity columns are selected. For further data cleaning, all entries with either missing smiles or missing standard values are removed. Also, Some bioactivity values are negative, these are removed as well because they cannot be standardized to pIC50 values. Next, the standard values are normalized and then converted to pIC50 values which are stored on a new column named pIC50. Finally, standard activities greater than or equal to 1000 are considered inactive whereas values less than 1000 are considered active and these new labels are stored on a new column called bioactivity class. It is important to note that the IC50 values actually represent the concentration of drug compounds needed to cause 50% deactivation of the target (Swinney, 2011). And so the smaller the IC50 value, the more active the molecule is, and vice versa. The 1000 nM threshold is chosen.

Table 1. A Sample Clean Dataset from our Data Mining Pipeline

	molecule_chembl_id	canonical_smiles	bioactivity_class	standard_value	pIC50
390	CHEMBL3286987	COc1cc(/C=c2\sc3nc(-c4ccncc4C)cn3c2=O)cc(OC)c1O	active	32.0	7.494850
391	CHEMBL3286988	COc1cc(/C=c2\sc3nc(-c4cccc(Cl)c4)cn3c2=O)cc(OC...	active	6.0	8.221849
392	CHEMBL3286989	COc1cc(/C=c2\sc3nc(-c4cccc(C(F)(F)F)c4)cn3c2=O...	active	123.0	6.910095
393	CHEMBL3286990	COc1cc(/C=c2\sc3nc(-c4cccc(OC(F)(F)F)c4)cn3c2=...	inactive	10000.0	5.000000
394	CHEMBL191083	CN(C)c1ccc2nc3ccc(=[N+](C)C)cc-3sc2c1	active	550.0	6.259637
395	CHEMBL3330737	N#C/C(=C\c1ccc(-c2ccc(-c3ccc(N(c4ccccc4)c4cccc...	active	410.0	6.387216
396	CHEMBL3330738	N#CC(C#N)=Cc1ccc(-c2ccc(-c3ccc(N(c4ccccc4)c4cc...	inactive	10000.0	5.000000
397	CHEMBL3330739	O=Cc1ccc(-c2ccc(-c3ccc(N(c4ccccc4)c4ccccc4)cc3...	inactive	7100.0	5.148742
398	CHEMBL3330740	N#C/C(=C\c1ccc(-c2ccc(-c3ccc(N(c4ccccc4)c4cccc...	inactive	1500.0	5.823909
399	CHEMBL3330741	O=C1NC(=O)C(=Cc2ccc(-c3ccc(-c4ccc(N(c5ccccc5)c...	inactive	10000.0	5.000000

4.2 Machine Learning

From the clean data above, both regression and classification models can be built. Both types of models make use of the canonical smiles as the input features while the bioactivity class column is used as labels for classification and pIC50 column for regression. First, the dataset is upsampled incase of class imbalance. Then, regression is used to determine the best type of molecular descriptor calculator while classification is implemented more in the other steps of our project. And these experiments were all performed on the beta amyloid 4 protein target for

Alzheimer's disease.

4.2.1 Selecting Descriptor Calculator

The two descriptor calculators we worked with are the Padel Software and Morgan FingerPrinter on RDKit python API. Among the Padel descriptors were Pubchem Fingerprint, EState Fingerprint, AtomPairs2D Fingerprint, GraphOnly Fingerprint, KlekotaRoth Fingerprint, etc. Several regression models (RandomForest, DecisionTree, SVR, SGDRegressor, LinearRegression, MLPRegressor, VotingRegressor, etc) were trained on each of these descriptors including the morgan fingerprints and the average score of these models, the best fingerprints were thus determined.

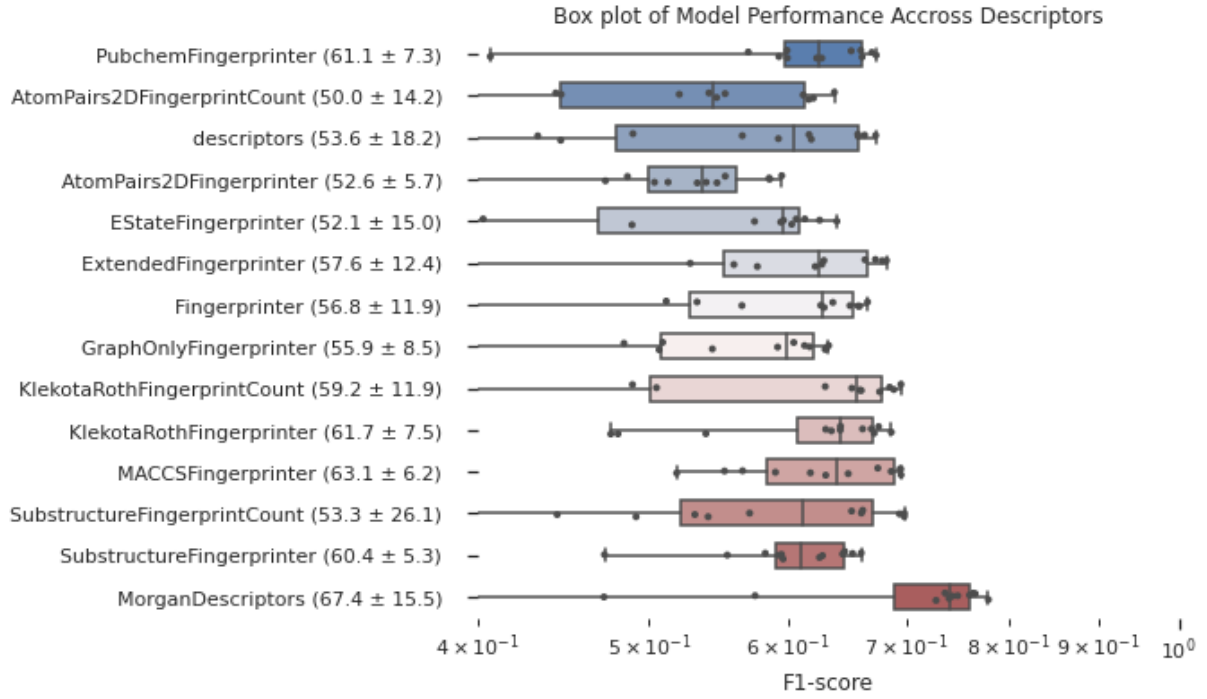

Figure 3. *Determination of Best Molecular Descriptors Using Trained Model Results*

As can be seen from figure 1 above, the majority of the models perform best with morgan descriptors compared to when other descriptors are used. For this reason, morgan descriptors were selected for further experimentation.

4.2.2 Selecting a Feature Selection Method and Best Model

Morgan descriptors calculated using RDKit can be as many as 30,000 or more with varying radii. By default, 1024 descriptors are computed with a Morgan radius of 1. Compared to the number of samples in the experimental dataset (1098), this makes a number of sample to feature ratio of approximate 1. This ratio is not so good and may result in overfitting of models. Therefore, different feature selection methods are tried. This time, classification models are used since our aim is to classify compounds into the active and inactive classes.

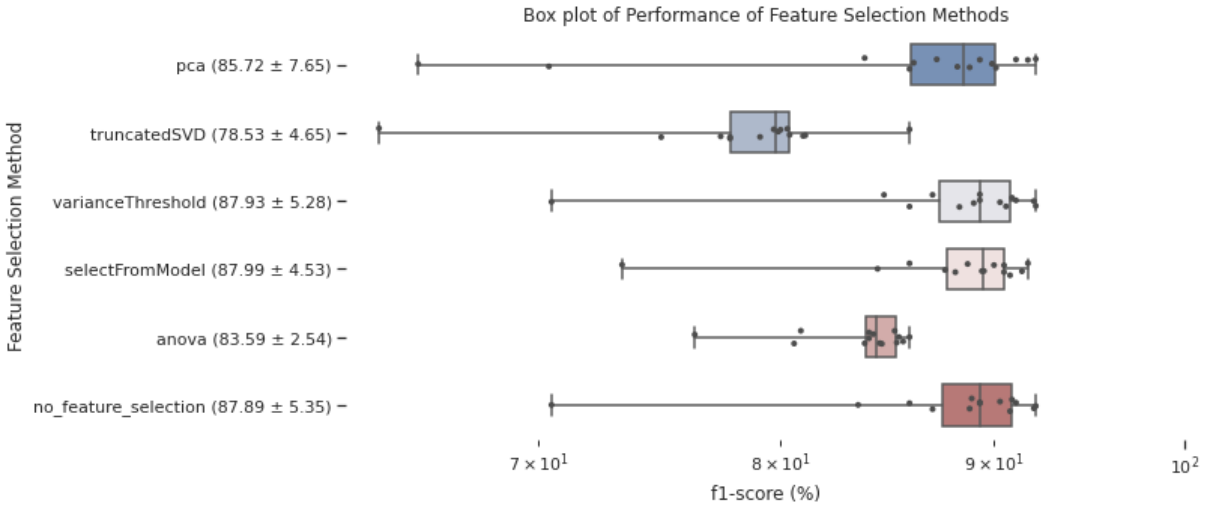

Figure 4. *Feature Selection Method Determination Using Trained Models*

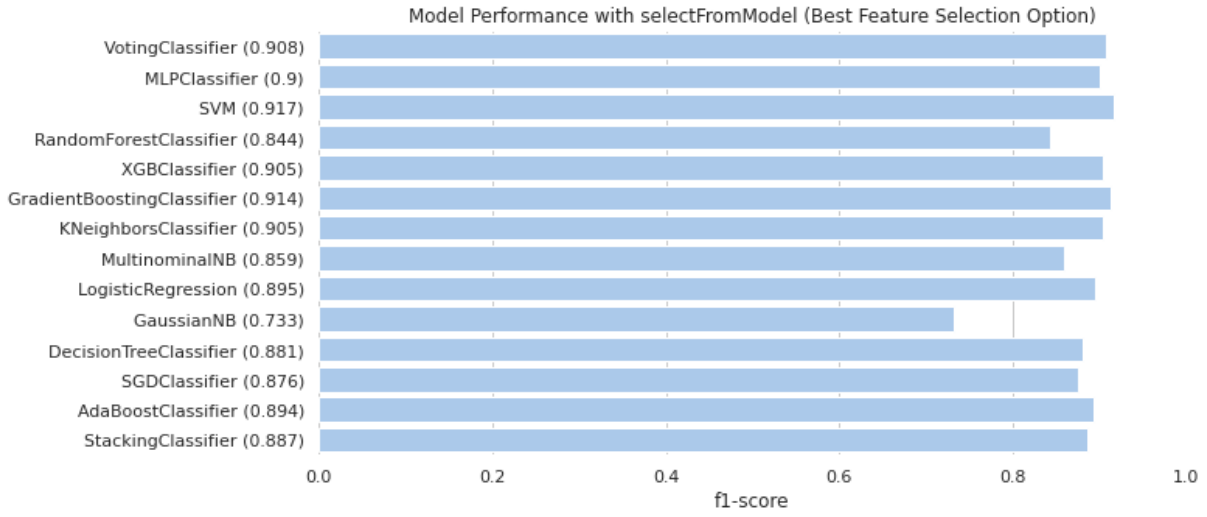

Figure 5. *Best Model Across Different Feature Selection Methods*

A stratified 10-fold cross validation strategy was used to evaluate the models on each descriptor dataset and average taken. To determine the best feature selection method, the average score for all models is taken on each feature selection method and in our case study, selectFromModel scikit-learn method was the best with mean score 87.99%. Moreso, to select the best model, the average of model performance for each model is taken across all feature selection methods, and the support vector classifier seems to be the best. Other models like the gradientboostingclassifier, voting classifier, K-nearest neighbour and neural networks are also doing well with mean scores above 90%.

To ensure that feature selection is not applied if not needed, an option with no feature selection is implemented in our pipeline.

4.2.3 Morgan Parameter Optimization

Since Morgan descriptors are computed differently when different radii and number of bits are

specified, fine tuning is performed for radius values in the range 1 to 5 and bits values from 512 to 2024 with steps of 256, to determine the best model parameters. For this determination, the best model SVC, and no feature selection are used. Thus, a best accuracy of 92.6 (± 1.6) % was obtained on Morgan radius 4 and number of bits 768. And a score of 92.5 (± 1.6) % was obtained with feature selection.

4.2.4 Model Optimization and Evaluation

For the selected best model (SVC), hyper-parameter tuning is performed on its parameters using grid search and the results are compared before and after tuning. To do this, the dataset is split into 90 % train (988 samples) and 10% test (110 samples). The model is trained on the training dataset and evaluated on the test dataset both before and after tuning and the results were obtained using sklearn classification report method as shown on table 2 below. The optimized model was then saved for later use.

Table 2. *Test Results of SVC Before and After Hyper-parameter Tuning*

Results on Normal SVC				
	Precision	**Recall**	**F1-score**	**Support**
Active	0.88	0.88	0.88	42
Inactive	0.93	0.93	0.93	68
Accuracy			0.91	110
Macro-average	0.90	0.90	0.90	110
Weighted-average	0.91	0.91	0.91	110
Results on Optimized SVC				
	Precision	**Recall**	**F1-score**	**Support**
Active	1.00	0.98	0.99	42
Inactive	0.99	1.00	0.99	68
Accuracy			0.99	110
Macro-average	0.99	0.99	0.99	110
Weighted-average	0.99	0.99	0.99	110

These results show that very accurate machine models can be built to predict the bioactivity drugs on the beta amyloid 4 protein. To make it easier to reproduce this machine learning pipeline on any other target, an API was built to handle the data mining and cleaning, as well as the machine learning model building, selection of feature selection method, selection of best model, Morgan tuning and hyper parameter tuning. Therefore in a similar way we plan to build machine learning models for as many targets as possible considering that the data is available.

4.3 Drug Repurposing

Drug repurposing, also called drug repositioning, is a method used to assign FDA approved drugs to other diseases for which the drug was not meant for (Pushpakom et al., 2018). It involves searching through drug databases and using computational approaches to determine if the drugs will work on a given target. Our approach aims at narrowing the search field. That is, instead of searching for all drugs in the drug database, the search will be limited to drug compounds that are active on other targets that are closely related to the given target. To compute the relationship between targets, 14,855 targets are downloaded from the ChEMBL database and a similarity score will be computed between all targets pair-wise according to their number of shared active compounds. Then a graph will be created after applying sklearn agglomerative clustering algorithm on the similarity scores. This graph will show the relationship between targets. That is, targets with a great number of shared drugs will likely be clustered together whereas targets with no active drugs in common will be far apart. Also, the clustering results will be stored into a structured data format for later use. Selected compounds are then screened a second time using the pretrained machine learning models on a given target and any compound predicted as active may be considered a drug candidate. Figure 6 below shows the drug repurposing pipeline.

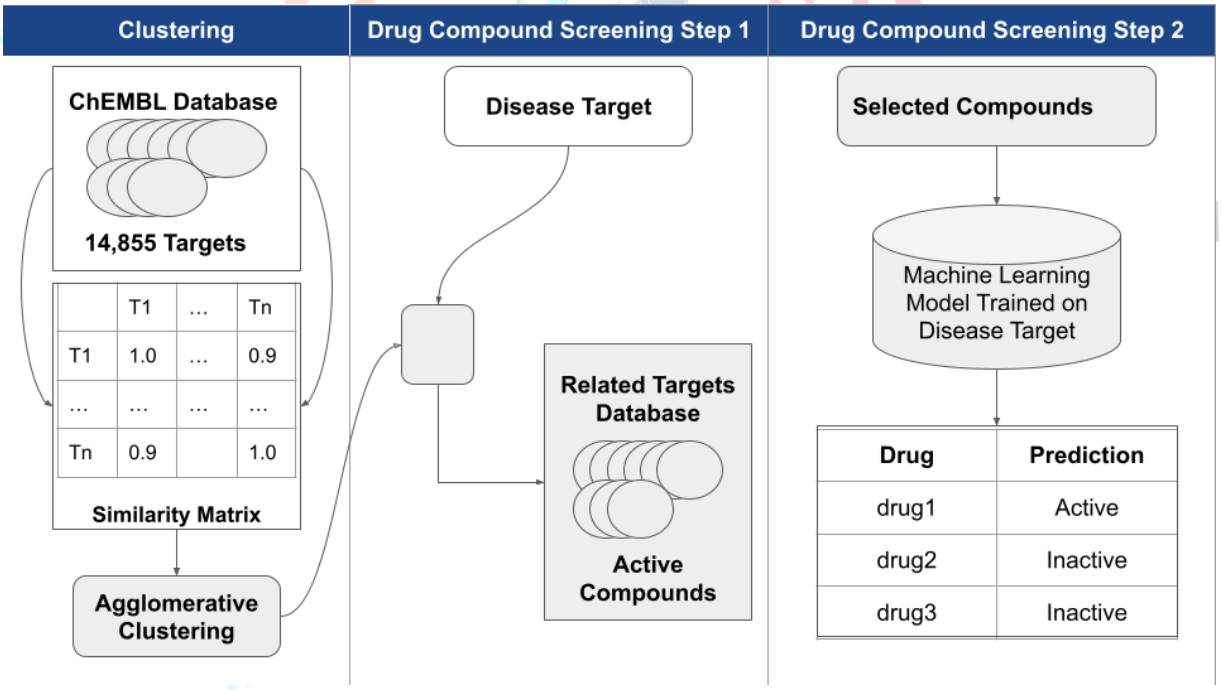

Figure 6. *Using Clustering and Supervised Machine Learning for Drug Repurposing*

4.4 Molecular Docking

Molecular docking is a computational technique used to model the interaction between a ligand (drug molecule) and a target molecule (Morris & Lim-Wilby, 2008; Meng, Zhang, Mezei & Cui, 2011). Molecular binding software aims to predict the binding energy and affinity for the ligand to the target compound. Several docking software such are pyGOLD,

CCCC, PTools (Patel, Brinkjost & Koch, 2017; Jones, Willett & Glen, 1995;Saladin, Fiorucci, Poulain, Prévost & Zacharias, 2009) have open source APIs that can be accessed using the python programming language. We intend to experiment on each of these and select the best API for our project. Molecular docking will be added to our pipeline as the final screening phase for all molecules that passed both the unsupervised and supervised screening steps. Compounds with a high RMSD score can be validated as active drug candidates and recommended for pre-clinical trials.

4.5 Web Application

The commercial product resulting from this project is an interactive web application where users can benefit from this great innovation. By incorporating three screening elements on the web app, users can reduce their work from working with millions of drugs to only a few hundreds. The components of the web application are; a home page where users will be guided on how to use the app, a page data mining, incase users want to analyze clean data themselves, a page to offer users the possibility to predict the bioactivity on any target in our database, a page to screen drugs using pretrained models, a page for further drug screening using molecular docking, and a page with our contact information through which users can reach us for support. The design of the web app is still in progress and figure 7 below shows how the interface would look like for some components.

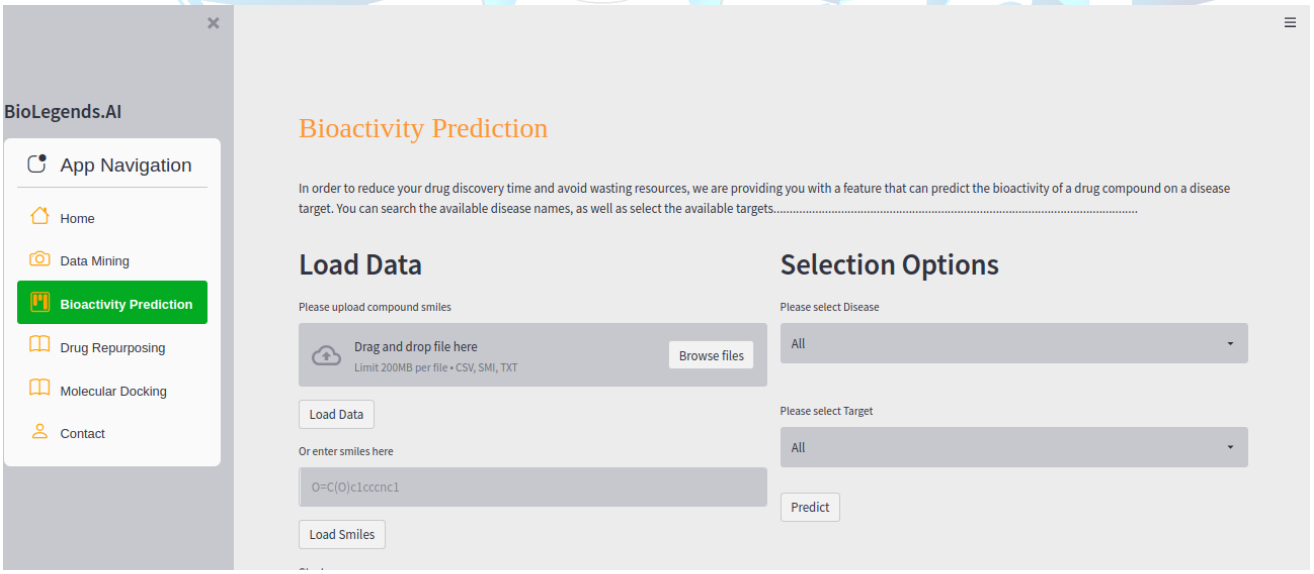

Figure 7 (a). *User Interface For Predicting Bioactivity of Drug Compounds on Targets: User Options*

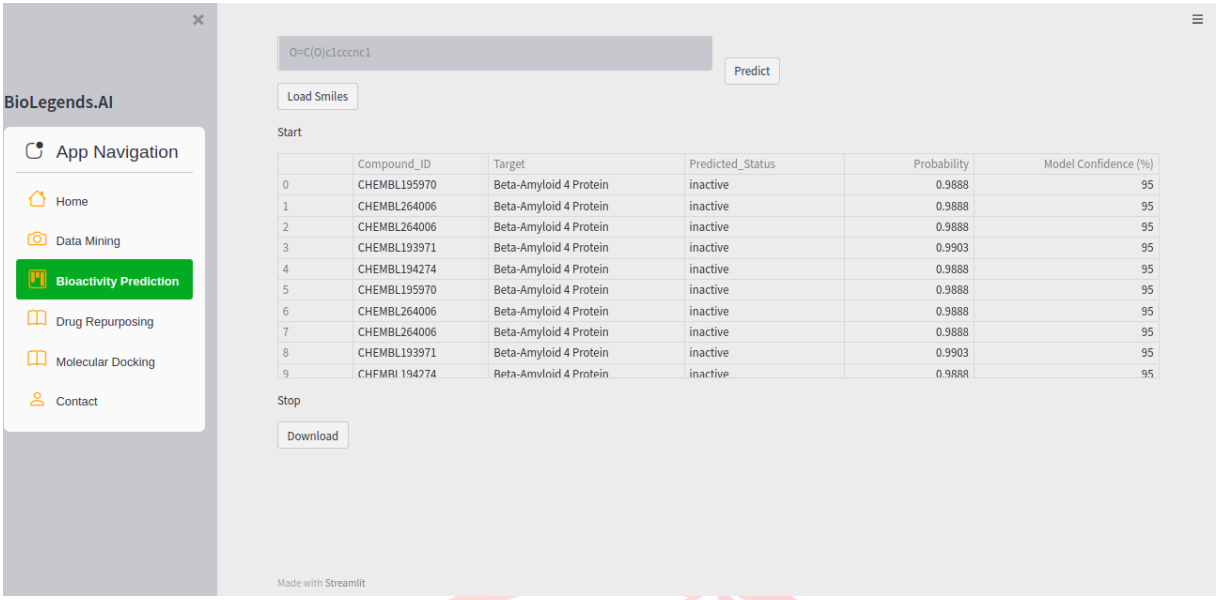

Figure 7 (b). *User Interface For Predicting Bioactivity of Drug Compounds on Targets: Results*

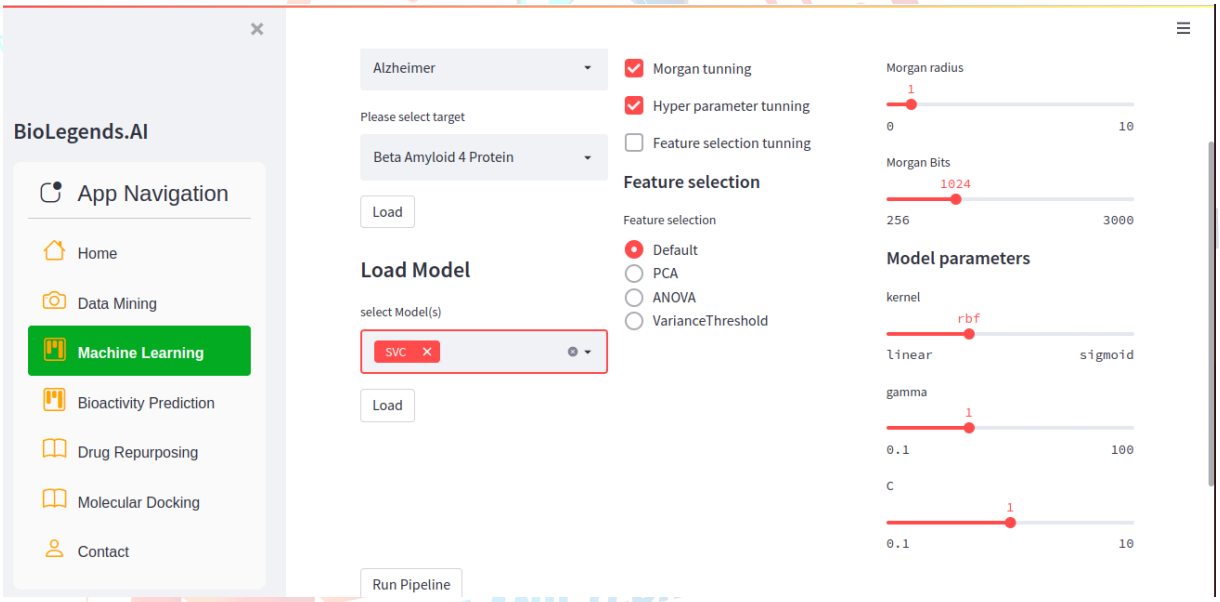

Figure 7 (c). *User Interface for Building Custom Machine Learning Models*

5. Innovative Direction

In our novel approach, we provide a pipeline for data preprocessing and analysis, as well as machine learning models that are completely dependent on the data's characteristics. This will ensure that the models are not killed by the huge variability in various dataset characteristics. One important aspect of good research is its reproducibility. Therefore, the software will be made user friendly and adaptable to all possible use cases and computational drug discovery experiments. Researchers can choose any target molecule, and based on the availability of data in the involved databases, hit compounds, based on their bioactivity, can be validated for the given target. Also, the integration of unsupervised machine for drug screening and

supervised learning for further screening in drug repositioning is a novelty on its own. As a case study, we use the available dataset of amyloid-beta protein target due to its importance in Alzheimer's disease (Yamin et al., 2008), and to contribute to the current efforts and hope to discover a new drug in treating Alzheimer's disease in the process.

6. Applicability

A common pipeline used by many drug discovery researchers is to create models that act as filters for reducing the number of drugs to work with until a lead is found. We not just aim to adapt our pipeline to this, but to build a web application that is hosted on the internet for all users to access. This will require us to purchase an online server for hosting the application. Cloud servers can be purchased from Upcloud and paid monthly with a minimal fee of 10 $ per month. Also, as lots of models will be trained, a database will be established on the cloud server for storing models and intermediate results of our experiments. It is also planned to purchase super computing services in order to improve our machine learning model training efficiency and accelerate our research. Also, to provide users with the possibility of training new models in real time is a service we wish to provide, however, this will require the purchase of a server with much computational power, considering that many users will be using the app simultaneously. This is perhaps a shortcoming to our project where we need help. Nevertheless, with the data mining, bioactivity prediction, drug repurposing and molecular docking pipelines, users will definitely have a gain making use of our application.

7. Estimated Cost and Project Scheduling

Table 3. Project Schedule

	Assignee	Mar	Apr	May	Jun	Jul	Aug
1. Data Mining	Cyrille						
2. Machine Learning	Atakan						
3. Drug Repurposing	Cyrille Atakan						
4. Molecular Docking	Cyrille Atakan						
5. Building and Hosting Web Application	Cyrille						

Table 4. *Project Cost Estimate*

Product	Cost	Time
Amazon Computer	720 $	1 year
Upcloud Server (1 core, 2 GB memory, 50 GB storage, 2TB transfer)	120 $	1 year
Total	**840 $**	

8. Target Audience of Project Idea (Users)

The target audience of our project is pharmaceutical companies and scientists. This project will enable the drug development process to take place in a shorter time, with less cost, and to be more effective. Thus, it will be easier than usual to both increase the effect of drugs for diseases that already have treatment and to develop drugs for diseases for which there is no cure yet.

9. Risks

	Important Risks	Risk Management (Plan B)
1.	Machine Learning models might perform poorly and predict inactive compounds as active	• Molecular Docking will be added to the pipeline as extra screener to make sure that all inactive compounds are removed
2.	There might not be enough storage on web server to store trained models and clustering results	• Models will be saved for only common diseases to save space • Users will be provided with an option to train custom models on real time • Purchase a server with more storage and memory.
3.	Users might experience delay while running the machine learning pipeline as it is time consuming.	• Experiments will be done to select the best models and best optimal parameters such that users don't have to perform hyper-parameter tuning nor waste time on training a weak model for their study. • Users can select any model or any group of models and their experimental optimal parameters may be applied.

10. Resources

Batool, M., Ahmad, B., & Choi, S. (2019). A Structure-Based Drug Discovery Paradigm. *New Avenues in Molecular Docking for Drug Design*, *20*(11), 2783. https://doi.org/10.3390/ijms20112783

Mendez D et al (2019). ChEMBL: towards direct deposition of bioassay data. *Database issue Publ*, *37*. ChEMBL. doi: 10.1093/nar/gky1075

Kamber, M., Pei, J., & Han, J. (2011). *Data Mining: Concepts and Techniques, 3rd edition* (3rd ed.). Elsevier Science. https://doi.org/10.1016/B978-0-12-381479-1.00008-3

Manorathna, R. (2020, December). *k-fold cross-validation explained in plain English (For evaluating a model's performance and hyperparameter tuning)*. ResearchGate.

Meng, X.-Y., Zhang, H.-X., Mezei, M., & Cui, M. (2011, June 1). Molecular Docking: A powerful approach for structure-based drug discovery. *Current Computer Aided Drug Design*, *7*(2), 146-157. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3151162/

Morris, G. M., & Lim-Wilby, M. (2008). Molecular docking. *Methods in Molecular Biology*, *443*, 365-378. https://doi.org/10.1007/978-1-59745-177-2_19

Yamin, G., Ono, K., Inayathullah, M., & Teplow, D. B. (2008). Amyloid β-Protein Assembly as a Therapeutic Target of Alzheimer's Disease. *Curr Pherm Des.*, *14*(30), 3231–3246. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5650911/

Zhang, Y., & Rajapakse, J. C. (Eds.). (2009). *Machine Learning in Bioinformatics*. Wiley. https://doi.org/10.1002/9780470397428

Pushpakom, S., Iorio, F., Eyers, P., Escott, K., Hopper, S., & Wells, A. et al. (2018). Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, *18*(1), 41-58. doi: 10.1038/nrd.2018.168

Asai, A., Konno, M., Taniguchi, M., Vecchione, A., & Ishii, H. (2021). Computational healthcare: Present and future perspectives (Review). *Experimental And Therapeutic Medicine*, *22*(6). doi: 10.3892/etm.2021.10786

Morris, G., & Lim-Wilby, M. (2008). Molecular Docking. *Methods In Molecular Biology*, 365-382. doi: 10.1007/978-1-59745-177-2_19

Meng, X., Zhang, H., Mezei, M., & Cui, M. (2011). Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer Aided-Drug Design*, *7*(2), 146-157. doi: 10.2174/157340911795677602

Jones, G., Willett, P., & Glen, R. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal Of Molecular Biology*, *245*(1), 43-53. doi: 10.1016/s0022-2836(95)80037-9

Saladin, A., Fiorucci, S., Poulain, P., Prévost, C., & Zacharias, M. (2009). PTools: an opensource molecular docking library. *BMC Structural Biology*, *9*(1). doi: 10.1186/1472-6807-927

Swinney, D. (2011). Molecular Mechanism of Action (MMoA) in Drug Discovery. *Annual Reports In Medicinal Chemistry*, 301-317. doi: 10.1016/b978-0-12-386009-5.00009-6

Drug development cost United States 1970-2016 | Statista. (2020). Retrieved 10 May 2022, from https://www.statista.com/statistics/265054/pharmaceuticals-cost-of-development-in-the-us/