VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka

Department of Computer Science Engineering

PROJECT WORK-4 REPORT

on

"Bioinformatics: Drug Discovery"

"Submitted by"

Anagha Acharya (1BM19CS224)
Ramya Ramesh(1BM19CS227)
Tasmiya Fathima (1BM19CS172)
Trisha Lakhani (1BM19CS214)

Under the Guidance of
Dr. Asha GR
Assistant Professor, BMSCE

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE AND ENGINEERING

B. M. S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019
April-2022 to July-2022

B. M. S. College of Engineering,
Bull Temple Road, Bangalore 560019
(Affiliated To Visvesvaraya Technological University, Belgaum)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project work entitled "**Bioinformatics: Drug Discovery**" carried out by **Anagha Acharya (1BM19CS224),Ramya Ramesh(1BM19CS227), Tasmiya Fathima (1BM19CS172) and Trisha Lakhani (1BM19CS214)** who are bonafide students of **B. M. S. College of Engineering**. It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visveswaraiah Technological University, Belgaum during the year 2022. The project report has been approved as it satisfies the academic requirements in respect of **Project Work-4(20CS6PWPW4)** work prescribed for the said degree.

Signature of the Guide	Signature of the HOD
Dr. Asha GR	Dr. Jyothi S Nayak
Assistant Professor	Professor and Head, Dept. of CSE
BMSCE, Bengaluru	BMSCE, Bengaluru

External Viva

Name of the examiner	Signature with date
1._____	_____
2._____	_____

B. M. S. College of Engineering

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, Anagha Acharya (1BM19CS224), Ramya Ramesh (1BM19CS227), Tasmiya Fathima (1BM19CS172) and Trisha Lakhani (1BM19CS214), students of 6th Semester, B.E, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bangalore, here by declare that, this Project Work-4 entitled "Bioinformatics : Drug Discovery" has been carried out by us under the guidance of Prof Asha GR, Assistant Professor, Department of CSE, B. M. S. College of Engineering, Bangalore during the academic semester Apr-2022-Aug-2022

We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

Anagha Acharya (1BM19CS224)

Ramya Ramesh (1BM19CS227)

Tasmiya Fathima(1BM19CS172)

Trisha Lakhani (1BM19CS214)

Contents

1 Introduction 5
1.1 Motivation 5
1.2 Scope of the Project 5
1.3 Problem Statement 5
2 Literature Survey 7
3 Design 19
3.1 High Level Design 19
3.2 Detailed Design 20
3.3 Sequence Design 23
3.4 Use Case Design 24
4 Implementation 25
4.1 Proposed methodology 25
4.2 Algorithm used for implementation 25
4.3 Tools and Technologies Used 26
4.4 Testing 26
5 Results and Discussions 27
6 Conclusion and Future Work 30

1 Introduction

Bioinformatic-analysis can not only accelerate drug target identification and drug candidate screening and refinement, but also facilitate characterization of side effects and predict drug resistance. High-throughput data such as genomic, epigenetic, transcriptomic, proteomic, and ribosome profiling data have all made significant contributions to mechanism-based drug discovery and drug repurposing. Moreover, bioinformatics has also innovated personalized medicine research thus bringing new discoveries in terms of drugs that can be personalized to someone's genetic pattern.

1.1 Motivation

Anaplastic Lymphoma Kinase, is an enzyme encoded by the ALK gene. It helps in controlling cell growth. It is part of a family of proteins called receptor tyrosine kinases (RTKs). Mutated forms of the ALK gene and protein have been found in non-small cell lung cancer, anaplastic large cell lymphoma and neuroblastoma. Alk-positive lung cancer occurs in about 5% of all lung cancer patients and is usually metastatic. Without treatment, the expectancy of patients is within 12 months. With research, improved treatments and better medicines are being found out. The ALK gene provides instructions for making a protein called ALK receptor tyrosine kinase, which . ALK inhibitors work by blocking these receptors. Various ML algorithms are used to find the best model for predicting the best chemical compound for acting against ALK. Our objective is to find the best compounds useful for this inhibition action on the basis of pIC50 values. IC50 is a quantitative measure that indicates how much of a particular inhibitory substance (e.g. drug) is needed to inhibit, in vitro, a given biological process or biological component by 50%. pIC50 is the negative log of the IC50 value when converted to molar.

1.2 Scope of the Project

Build and design a model that can effectively utilize input from the user in the form of compound ID and canonical smiles from Drug database and predict the pIC50 values of those compounds to identify the best compounds among them that require the least amount to display their inhibitory action in the most effective way.

1.3 Problem Statement

The long development pipeline faces increasing costs and additional challenges, including the lack of predictive validity of current animal models, insufficient knowledge regarding underlying mechanisms of disease, patient heterogeneity, lack of targets and biomarkers, a high rate of failed clinical trials etc. Bioinformatics analyses provide key information throughout the entire drug discovery and development process, from

aiding the identification and validation of drug targets and leads through to helping assess the outcomes of phase 1, 2 and 3 clinical trials; as well as supporting drug repurposing efforts. The aim of the project is to learn about Bioinformatics through Drug Discovery. The study on various drug-target interactions are made to make predictions on their action. This is done by searching for target enzyme and acquiring curated bioactivity data, performing exploratory data analysis and building regression models and scatter plots to compare Predicted and Actual values of drug action on target following which Model comparisons based on their performances can be made.

2 Literature Survey

1. Machine Learning Techniques and Drug Design

Introduction: The interest in the application of machine learning techniques (MLT) as drug design tools is growing in the last decades. The reason for this is related to the fact that the drug design is very complex and requires the use of hybrid techniques.
Methodology: This study provides a brief overview of certain MLT, including self-organizing maps, multilayer perceptrons, Bayesian neural networks, counter-propagation neural networks, and support vector machines. The effectiveness of the described approaches is compared to various traditional statistical techniques (such as partial least squares and multiple linear regression), which demonstrates that MLT has important advantages. These methods are now being used in a lot more medicinal chemistry studies, especially when support vector machines are involved. The use of these methods to build more trustworthy QSAR models is included in the state of the art and future trends of MLT applications. The models produced by MLT can be used as filters and in virtual screening tests to create/discover novel compounds.
Result: Therefore, this review provides a critical point of view on the main MLT and shows their potential ability as a valuable tool in drug design.

2. Performance measures in evaluating machine learning based bioinformatics predictors for classifications

Introduction: Many existing bioinformatics predictors are based on machine learning technology. When applying these predictors in practical studies, their predictive performances should be well understood.
Methodology: Different evaluation techniques and performance measures are used in diverse research. Different words, nomenclatures, or notations may emerge in different contexts even for the same performance metric. Three methods can be used in machine learning to assess a predictor. They are referred to as the independent dataset test, the re-substituting test, and cross validation, as shown in Figure. The leave-one-out cross validation and the n-fold cross validation are two further subsets of the cross validation method. In bioinformatics, cross validation approaches have emerged as the most suggested evaluation techniques due to the abundance of samples accessible. All of the data are treated as both the training dataset and the testing dataset in a cross-validation test.
Results: We carried out a review on the most commonly used performance measures and the evaluation methods for bioinformatics predictors.

3. Advancing Drug Discovery via Artificial Intelligence

Introduction: Most de novo design methodologies have been based on computational predictions of atomic and molecular characteristics. An AI subfield called machine learning can more accurately and more quickly predict the quantum mechanically-level physical

and chemical features of tiny molecules. Additionally, molecular representations and biological and toxicological activity can be compared using artificial intelligence (AI). In order to effectively investigate the paths of synthesis of potential medication candidates, AI-based algorithms are also being created. By using automated analysis of reaction feasibility in conjunction with robotic platforms, the chemical space for novel reactions can be explored.
Methodology: In primary drug screening, cell types are categorized using LS-SVM, and cells are sorted using DNN. DNNs, RF, and GBMs can all be used to predict the physical attributes and bioactivity of compounds. DeepTox can predict a compound's toxicity, whereas AlphaFold can predict the 3D structure of a target protein utilizing DNNs for structural analysis and drug-protein interactions. The 3N-MCTS approach can be used to recursively search for "backward" reaction paths until a collection of easier, readily available precursor molecules is identified, at which point the retrosynthesis pathway can be anticipated. On this basis, insights into the reaction process and reaction yield prediction can be produced. In order to digitize and standardize a synthesis operation, ChASM employs a chemical descriptive language (XDL) that openly and methodically combines all the necessary information. Additionally, the development of AI that allows for the design of medicinal compounds using generative adversarial networks is a byproduct of advancements in precision medicine (GANs).
Conclusion: For a variety of reasons, computational techniques, including AI, do not currently perform well in this domain. First off, as AI is a data-mining technique, training DNNs successfully requires a lot of training data. Second, sometimes the quality of the data is insufficient for effective AI learning. Furthermore, numerous, conflicting datasets could be present in public databases. Thirdly, while converting 3D atomic space to a 2D interpretation for AI calculations, crucial information about the 3D target structure is lost.

4. Graph Neural Networks and Their Current Applications in Bioinformatics

Introduction: As a subset of deep learning in non-Euclidean space, graph neural networks (GNNs) excel at a variety of tasks that require processing graph structure and network input.
Methodology: Commonly used GNN models: 1) Graph Convoluted Networks [Spectral-based]. GCN exploits the principle of Laplacian and Fourier transform to map the irregular structure of a graph to a regular Euclidean space for convolution operation.The convolution operation is defined by spatial-based GCN directly using the information-dissemination mechanism on the graph, and its propagation method is similar to that of the original GNN; 2) Graph Attention Networks introduces the attention mechanism into the propagation step of the graph to learn the weight between two connected nodes; 3) Graph Autoencoder Networks an autoencoder which is a form of ANN, is employed to learn effective codings for unlabeled data using AE-based graph generation models (unsupervised learning). Then, three representative tasks—node categorization, link prediction, and graph generation—are suggested based on the three tiers of structural information that GNNs can learn. In the meanwhile, we classify and discuss the relevant studies in three areas: illness prediction, drug discovery, and biomedical imaging, in accordance with the specialised applications for

various omics data.
Conclusion: Even while GNNs already perform exceptionally well in a wide range of biological activities, they still have a long way to go because of issues with poor data processing, methodology, and interpretability.

5. Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods

Introduction: The necessity for quick, precise, and effective drug discovery pipelines has been brought to light by the present global health emergency known as the Coronavirus 2019 (COVID-19) pandemic. In vitro high-throughput screening (HTS) has traditionally been used to find new drugs, but these initiatives are expensive and need complex experimental setups that are only available to large pharmaceutical corporations.
Methodology: Utilizing cutting-edge computational techniques and contemporary artificial intelligence (AI)-based algorithms for quick lead identification in the repurposable chemical space [approved drugs and natural products (NPs) with established pharmacokinetic profiles] is a potent way to conserve time and resources. One well-liked in silico medication repurposing method is structure-based drug repurposing. Structure-based drug discovery (SBDD) pipelines are examined in this review using both conventional and cutting-edge AI-based computational approaches and tools.
Conclusion: Finding targets, predicting the 3D structures of target proteins from their sequence, screening large numbers of small drug-like molecules, performing generative tasks to suggest new ligands, providing retrosynthetic pathways for synthesis, controlling robotic systems to physically synthesize compounds, processing the signal corresponding to molecule characterization based on spectra, and predicting the results of clinical trials are all possible with the help of machine learning (ML) methods.

6. Machine Learning guided early drug discovery of small molecules

Introduction: The early stages of the drug development process have seen widespread adoption of machine learning (ML) techniques, particularly when it comes to small-molecule therapeutic candidates. Despite this, the usage of ML in the pharmacokinetic/pharmacodynamic (PK/PD) application arena is still restricted. The methods now employed for predicting the ADME (absorption, distribution, metabolism, and excretion) properties of small compounds based on their structures and for predicting the structures based on the required qualities for molecular screening and optimization. The use of ML to forecast PK to rank drug candidates' capacity for achieving the right exposures and, in turn, offer significant insights into safety and efficacy, is discussed in the article's last section. **Methodology:** Machine learning models employ molecular representation techniques. Three sorts of molecular representation techniques exist: the use of descriptors, natural language, or both. **Conclusion:** These methods have significantly improved and been more widely used, especially in the context of early drug

development, thanks to advancements in molecular representations, computational tools, and computer power.

7. Utilizing graph machine learning within drug discovery and development

Introduction: Graphs can be used to visualize the interrelated nature of the biomedical data that is generated and utilized in the drug discovery process. Biomolecules that represent spatial and structural relationships are seen at the molecular level. Interactions between biomolecular species are seen at the intermediary level. At the higher level, complex relationships between drugs, side effects, diagnoses, associated treatments, and test results are seen. Graph machine learning (GML) is gaining popularity due to its capacity to integrate multi-omic data sets with other types of data, model biomolecular structures and the functional links between them. GML, a new family of ML techniques that take advantage of the structure of graphs and other irregular datasets, is the convergence of DL and NLP.
Methodology:
1. Traditional approaches: Graph statistic and Random Walks
2. Geometric approaches: knowledge graph embedding posits each relation type as a geometric transformation from source to target in the embedding.
3. Matrix/Tensor factorization: approximate a matrix X by the product of n low-dimensional latent factors, Fi, i 1, ..., n.
4. GNN: Message passing network, GCN, GAT, R-GCN, Graph pooling . **Applications:** Drug-target-indication interaction and relation prediction using knowledge graph embedding, molecular property prediction (ADME) profiles, early work in target identification to de novo molecular design, and other features are all available in GML for mining graph-structured data. Most notably, it is employed to instruct message-passing GNNs that operate on molecular structures to offer prospects for antibiotic development that have been repurposed. It is applied in the drug development process for a variety of purposes, including as target discovery, small molecule therapy design, the creation of new biological entities, and drug repurposing.
Limitations: The study of GML is still in its early phases. Feature over-smoothing and information over-squashing are problems for deeper GNNs.

8. FL-QSAR: a federated learning based QSAR prototype for collaborative drug discovery

Introduction: Quantitative structure-activity relationship (QSAR) analysis is commonly used in drug discovery. For the first time, we verified the feasibility of applying the horizontal federated learning (HFL), which is a recently developed collaborative and privacy-preserving learning framework to perform QSAR analysis.
Methodology: A prototype platform of federated-learning-based QSAR modeling for collaborative drug discovery, i.e, FL-QSAR, is presented accordingly. We first compared the

HFL framework with a classic privacy-preserving computation framework, i.e., secure multiparty computation (MPC) to indicate its difference from various perspectives. Then we compared FL-QSAR with the public collaboration in terms of QSAR modeling.
Result: Taking together, our results indicate that FL-QSAR under the HFL framework provides an efficient solution to break the barriers between pharmaceutical institutions in QSAR modeling.

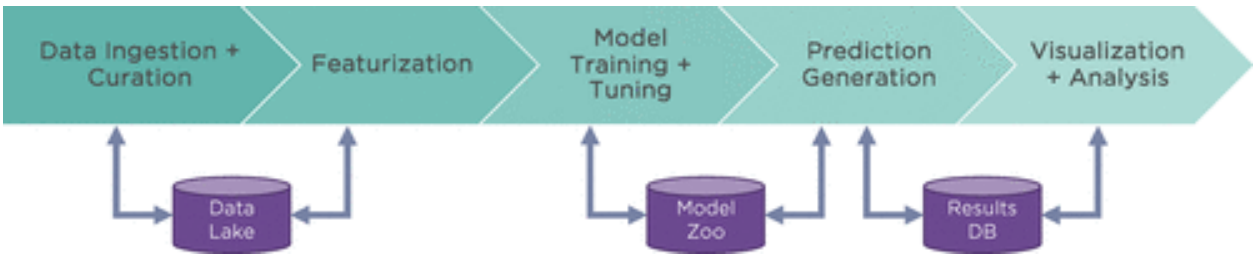
9. Understanding the Performance of Knowledge Graph Embeddings in Drug Discovery

Introduction: Knowledge Graphs (KG) and associated Knowledge Graph Embedding (KGE) models have recently begun to be explored in the context of drug discovery and have the potential to assist in key challenges such as target identification.
Methodology: In the field of drug development, KGs can be used as a step in a process that could lead to the execution of lab-based trials, have an impact on other choices, and, most significantly, eventually affect patient healthcare. A deeper comprehension of performance and the many variables affecting it is necessary for KGE models to have an influence in this field. In this study, we assess the prediction performance of five KGE models on two publicly available drug discovery-focused KGs over the course of many thousands of tests. Instead of concentrating on the most effective overall model or configuration, our objective is to examine more closely at how performance might be impacted by changes in the training setting, choice of training methods, and other factors.
Conclusion: Our results highlight that these factors have significant impact on performance and can even affect the ranking of models. Indeed these factors should be reported along with model architectures to ensure complete reproducibility and fair comparisons of future work, and we argue this is critical for the acceptance of use, and impact of KGEs in a biomedical setting.

10. AMPL: A Data-Driven Modeling Pipeline for Drug Discovery

Introduction: The total traceability and reproducibility of the model development and evaluation process is one of the essential conditions for integrating machine learning (ML) into the drug discovery process. In light of this, we have created an end-to-end modular and extendable software pipeline for creating and disseminating ML models that foretell crucial metrics vital to the pharmaceutical industry. The ATOM Modeling PipeLine, or AMPL, expands the capabilities of the open source library DeepChem and includes a number of machine learning (ML) and molecular feature-based tools that enable the user to develop models for a wide range of molecular features required for in silico drug discovery.

Methodology:



Conclusion:
1) Only on larger data sets did neural networks typically yield more accurate models.
2) For both random forests and neural networks, the proprietary MOE descriptors fared better than the open-source Mordred descriptors. ECFP was outperformed by graph convolutions in neural network representations. To describe molecular properties, physicochemical descriptors and deep learning-based graph representations perform noticeably better than conventional fingerprints.
3) It appears that many data sets prominently display little networks.
4) As data set size rose, model performance often increased, indicating the necessity for transfer learning or methodologies that integrate public data sets.
5) Hyperparameter optimization increases performance, sometimes noticeably.
6) There was little link between uncertainty quantification and error, and the effectiveness of applying UQ to filter predictions varied widely across data sets and model types.

11. An overview of neural networks for drug discovery and the inputs used

Introduction: Neural network-based artificial intelligence systems (NNs) discover drug discovery criteria based on training molecules, but first the molecules must be represented in specific.
Methodology: Artificial neural networks (ANNs) have long used molecular descriptors and fingerprints as inputs, but other methods of molecular description are only employed for storing and presenting molecular data. Researchers now have more options for drug discovery thanks to the advancement of deep learning, which allows ANN versions to employ a variety of inputs. regions covered The authors give a concise description of how NNs are used in drug discovery. The properties of various molecular characterization techniques combined with corresponding NN-based methodologies offer new options for drug discovery.
Result: As a result of DL, there are now more ways to describe molecules that are suitable for ANNs and their variations, giving researchers more options.

12. Advanced Graph and Sequence Neural Networks for Molecular Property prediction and Drug Discovery

Introduction: Introduction: One of the main issues in cheminformatics is the prediction of molecular properties because these properties are indicative of their functions and have applications in many different areas. In the literature, molecular graphs and simplified molecular-input line-entry system (SMILES) sequences are the two most often used methods for representing molecules in silico. Computational approaches for predicting chemical characteristics are gaining traction as a result of deep learning method advancements.
Methodology: First off, AdvProp is made up of a variety of comprehensive machine learning techniques that span numerous data sources and technique categories. It is predicted to deliver supplementary data for predicting molecular properties and produce improved results. Second, a brand-new graph-based deep learning technique called the multi-level message passing neural network is suggested (ML-MPNN). In molecular graphs, nodes, edges, subgraphs, and the complete graph can all be used to aggregate information. ML-MPNN can take full advantage of these extremely informational molecular graphs. Third, AdvProp includes contrastive-BERT, an unique sequence-based deep learning technique. Included are a suggested masked embedding recovery task and an unique self-supervised pre-training task using contrastive learning to use a lot of unlabeled molecules and produce competitive results on downstream tasks.Fourth, areas under curves (AUC), such as areas under receiver operating characteristic (ROC) curves and areas under precision-recall curves, can be optimized using AdvProp's efficient stochastic optimization techniques (PRC).
Conclusion: On highly skewed datasets, it is shown that they can improve prediction performance in terms of ROC-AUC and PRC-AUC considerably. Additionally, this won first place in the AI Cures Open Challenge for finding new COVID-19 drugs.

13. EDock-ML: A web server for using ensemble docking with machine learning to aid drug discovery

Introduction: In computational structure-based drug development, ensemble docking corresponds to the creation of an "ensemble" of drug target conformations. A web server called EDock-ML makes it easier to utilize ensemble docking with machine learning to determine whether a molecule is valuable enough to be taken into consideration later in the drug development process. A cost-effective method of taking receptor flexibility into account in molecular docking is ensemble docking. The use of the resulting docking scores to determine whether a drug is likely to be useful is improved by machine learning.
Methodology: To enhance predictions for the proteins in its database, EDock-ML uses a bottom-up methodology in which machine-learning models are created one protein at a time. Novice users don't need to worry about selecting the right parameters because machine learning models are designed to be used without changing the docking and model parame-

ters with which they were trained. A user can upload a file created by a chemical drawing software or enter a compound's ID from the ZINC database to submit it. They will then receive an output that will assist them determine whether the molecule is likely to be active or inactive for a medication target.
Conclusion: It was discovered that the machine-learning enhanced ensemble docking was able to eliminate the previously noted artifacts that adding more structures to an ensemble could actually have an adverse effect on performance rather than the opposite. In most instances, K-nearest neighbors and random forests outperformed logistic regression and support vector machines in terms of performance. For the proteins analyzed, the best area under the receiver operating characteristic curve ranged from 0.72 to almost 1, with the majority being more than 0.8. The performance is typically better than the outcomes found in earlier investigations using NNScore produced from ML.

14. Application of network link prediction in drug discovery

Introduction: Technological and scientific developments have generated enormous amounts of biomedical data. These data can be used to simulate the entities and interactions in biological and other complex systems when they are represented as networks (graphs). Predicting outcomes from drug-drug, drug-disease, and protein-protein interactions is of special interest to the fields of network biology and network medicine in order to speed up the process of drug development.
Methodology:
1) Medication Target prediction: Taking into account the bipartite networks of pharmaceuticals and their target proteins, the goal is to forecast which drug will affect which unknown proteins.
2) Drug-Disease Prediction: In order to detect commonalities across drug structures, it is important to understand the chemical makeup of medications as well as their intended protein targets.
3) Prediction of drug-drug interactions.
Conclusion: We can use these techniques to anticipate the interactions (links), or missing links, between medications and their targets, such as diseases, proteins, or other drugs, by using network-based techniques, and in particular link prediction methodologies. This strategy has already produced encouraging results whether it comes to finding a novel medicine to treat urine leakage problems or repurposing current drugs for the treatment of breast cancer.

15. ADMET modeling approaches in drug discovery

Introduction: A well-balanced mixture of pharmacodynamics (PD) and pharmacokinetics (PK), as well as sufficient absorption, distribution, metabolism, excretion, and tolerable tox-

icity (ADMET), are all characteristics of effective and safe medications . An essential part of pharmaceutical research and development is in silico ADMET prediction. Fully integrated ADMET prediction platforms that simultaneously target various PK parameters can quickly rule out inappropriate compounds, cutting down on the number of synthesis-evaluation cycles and costly late-stage failures.

Methodology: ML approaches that have lately excelled in the analysis of ADMET features include k-nearest neighbor (k-NN), support vector machines (SVM), random forests (RF), and artificial neural networks (ANNs). A free online tool for PK and toxicity prediction is called ATMETLab. Six algorithms—RP, DT, RF, SVM, PLS, and NB—were utilised to build the 289 000 chemicals used in the platform's QSPR models. RF, SVM, RP, and PLS were used to create the regression models, and RF, SVM, NB, and DT were used to create the classification models. The user of ADMETlab is able to do drug-likeness assessments based on the five rules.

Conclusion: SVM and RF fared better than other methods that were used concurrently.

16. Artificial intelligence in early drug discovery enabling precision medicine

Introduction: The idea behind precision medicine is to treat illnesses based on a patient's molecular profile, lifestyle, and environmental exposures. This strategy has been shown to boost clinical trial success rates and quicken drug approval processes. Only a few molecular biomarkers, however, are currently used in early drug development applications of precision medicine.

Methodology:

Table 1. Overview of AI algorithms used in precision medicine.

AI Algorithm		Advantages	Disadvantages	Applications discussed (section)
Shallow Learning	Linear/Logistic Regression	+ Interpretable	- Limited to linear trends	* drug response (2.2) * drug combinations (2.3) * MHC affinity (3.1) * T-cell specificity (3.1)
	Support Vector Machines (SVM)	+ Nonlinear function approximation + Flexible through kernel	- Less interpretable - Hard to design kernels for nonstandard data	* MHC affinity (3.1)
	Random Forests	+ Nonlinear function approximation + Automatic handling of different data types + Interpretable	- Not well equipped for regression tasks - Less interpretable	* disease subtyping (2.1) * patient stratification (2.1) * drug response (2.2) * drug combinations (2.3) * T-cell specificity (3.1)
	Gaussian Processes	+ Nonlinear function approximation + Flexible through kernel + Fully Bayesian	- Does not scale well to large datasets - Hard to design kernels for nonstandard data	* MHC affinity (3.1) * T-cell specificity (3.1)
	Dimensionality reduction and feature synthesis	+ No labels needed	- Limited expressive power	* disease subtyping (2.1) * patient stratification (2.1) * drug response (2.2)
Deep Learning	Generative	+ Nonlinear + Scales well + Handles unstructured data + Can generate novel examples + Few labels are needed (if at all)	- Hard to interpret - Needs lots of data and compute resources - Novel examples can be hard to evaluate	* protein sequence (3.2) modeling * small molecule modeling (3.3)
	Discriminative	+ Nonlinear + Scales well + Handles unstructured data	- Hard to interpret - Needs lots of data and compute resources	* disease subtyping (2.1) * drug response (2.2) * drug combinations (2.3) * MHC affinity (3.1) * T-cell specificity (3.1)

Conclusion: The majority of the time, AI-aided drug design is utilized as a preprocessing step to cut down on the number of compounds or modifications that need to be tested experimentally. However, in order to fully realize the potential of machine learning models for drug development, an AI-driven experimental design and a closed feedback loop are necessary.

17. An effective self-supervised framework for learning expressive molecular global representations to drug discovery

Introduction: The production of expressive molecular representations is a critical problem in AI-driven drug development, to say the least. Graph neural network (GNN) modeling of molecular data has become a powerful approach. However, the lack of labeled data and limited generalizability of earlier supervised techniques are frequent problems.
Methodology: Molecular representations are learned from massive amounts of unlabeled molecules using the unique molecular pre-training graph-based deep learning framework known as MPG. In MPG, we developed an efficient self-supervised approach for pre-training the model at both the node and graph levels. We called this approach MaGNet, and it is a potent GNN for modeling molecular graphs. MaGNet was found to be able to develop interpretable representations after pre-training on 11 million unlabeled molecules. With just one more step, the pre-trained MaGNet can be adjusted.
Conclusion: MPG makes accurate and logical predictions about drug-drug interactions. The model may now accept two graph inputs at once thanks to this intentional design. This makes it simple to utilize MPG in some activities that require graph pair input, such the widely used DDI. When many medications are provided at once, DDI refers to how one drug may alter another's activity.

18. MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery

Introduction: For the development of novel drugs, drug repurposing, and the detection of off-target effects, it is essential to identify interactions between bioactive small molecules and target proteins. The Multi-channel Deep Proteochemometric Predictor for Binding Affinity (MDeePred) method pursues the idea of building multiple channels that represent the input proteins from different aspects in a thorough manner. It is a novel protein featureization approach to be used in deep learning-based compound target protein binding affinity prediction.
Methodology: Multiple 2D vectors containing various protein features, including sequence, structural, evolutionary, and physicochemical characteristics, are fed to cutting-edge pairwise input hybrid deep neural networks to predict the real-valued compound-target protein interactions. The procedure uses a proteochemometric strategy, in which the input level

modelling of the chemical and target protein interactions is based on the properties of both molecules. The entire system is known as MDeePred, and it is a novel approach for computational drug discovery .
Conclusion: Evaluation of MDeePred was done on well-known benchmark datasets and compared its performance with the state-of-the-art methods. A scalable technique with a good level of predictive performance is MDeePred. Other protein-related prediction tasks can also be accomplished using the featurization method suggested here.

19. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches

Introduction: It has been observed that medication molecules contain poly-pharmacological capabilities, meaning that they can interact with targets other than their main therapeutic targets. A practical method to discover new drug-target interactions (DTIs) at minimal cost with reasonable accuracy is to use computational methods. A DDR, an effective DTI prediction approach based on the usage of a heterogeneous graph that incorporates known DTIs with numerous similarities between medications and multiple similarities between target proteins, is created in an effort to further increase the accuracy of DTI prediction.
Methodology: To create an optimum combination of similarities, DDR undertakes a preprocessing stage where a selection of similarities is chosen using a heuristic procedure. Then, in order to merge various similarities, a non-linear similarity fusion method is used. A random forest model is then applied by DDR employing various graph-based features that were taken from the DTI.
Result: It can be seen that DDR significantly lowers the AUPR score error compared to the next best state-of-the-art method for predicting DTIs by 34% when the drugs are new, 23% when the targets are new, and 34% when the drugs and the targets are known but not all DTIs between them are not known using 5-repeats of 10-fold cross-validation, 3 testing setups, and AUPR scores. This shows that identifying the right DTIs can be done effectively using DDR.
Limitations: DDR's present implementation only works with binary DTI data and aims to categorize each DTI as either binding (label = 1) or non-binding (label 1=0).

20. Machine learning and image-based profiling in drug discovery

Introduction: Automated imaging has been used in a variety of preclinical drug discovery applications and has proven to enable scalable and systematic phenotypic profiling of small molecules, establishing itself as a complementary approach to target-based in-vitro screening.
Approaches: The initial strategy for image-based profiling entails screening software targeted at pre-defined, particular phenotypes in order to find medications or pharmacolog-

ical targets that alter it. The second method analyzes cells after they have been exposed to genetic, pathogenic, or chemical perturbations and is an addition to methods like transcriptional profiling. Without any assistance from humans, computer vision can extract multivariate feature vectors of cell morphology. Unsupervised machine learning techniques are used to cluster image-based profiles in order to find tiny molecules that share the MAO molecule according to the "guilt-by-association" rule. It is common practice to group small molecule profiles based on their profile similarity for big data sets using hierarchical (unsupervised) clustering techniques.

3 Design

3.1 High Level Design

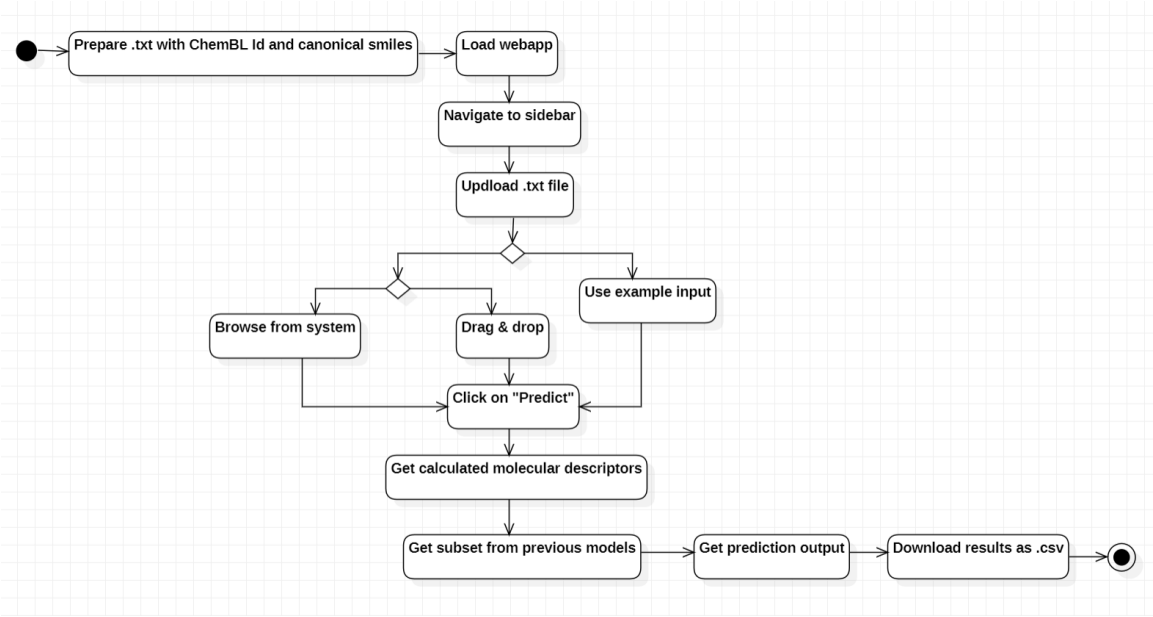


Figure 1: High Level Design

The user who can access the Bioactivity Prediction app must first input the data with ChemBL ID and canonical smiles, upload the file in the app and get outputs in the form of calculated molecular descriptors, subset from previous models built and prediction output with the molecule ChemBL ID and pIC50 values of the drugs predicted. The user has the option to save these prediction results by downloading them.

3.2 Detailed Design

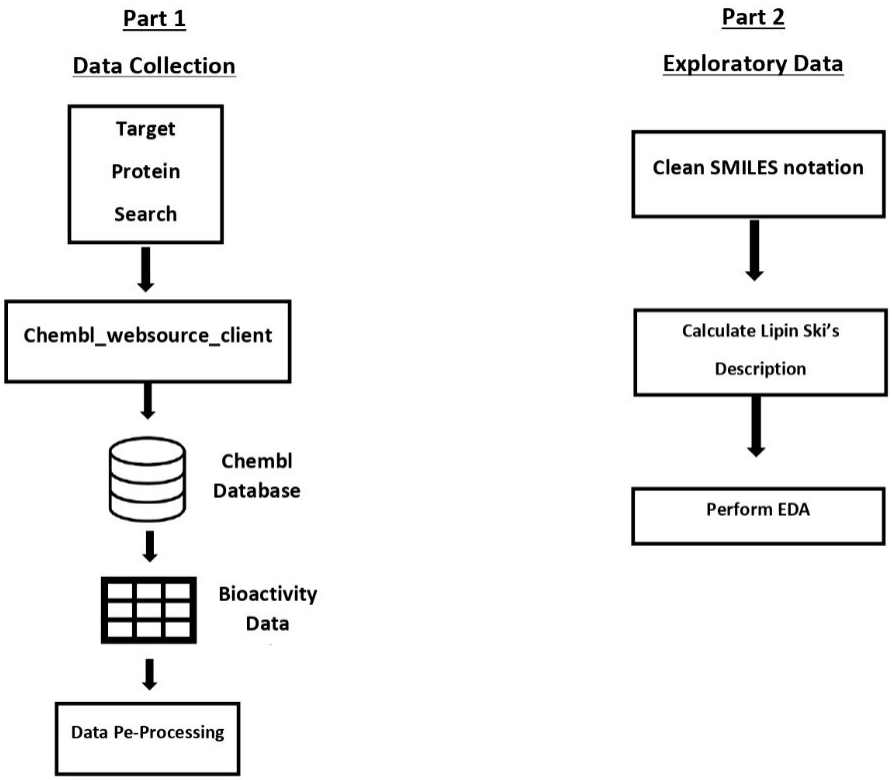


Figure 2: Detailed Design

- **PART-1 : Data Collection**
  Search for a Target Protein/enzyme of our Interest and conduct Data pre-processing (drop missing values, drop duplicates, feature selection and label compounds according to bioactivity threshold) after which we obtain the Curated Bioactivity data that can be used for our next step.
- **PART-2 : Exploratory Data Analysis**
  We use clean SMILES notation that we will require to analyze and remove small organic compounds after which using Lipinski's description we analyze other chemical properties like number of H-bond donors and acceptors and finally perform EDA using box plot, scatter plot and other statistical analysis tests.

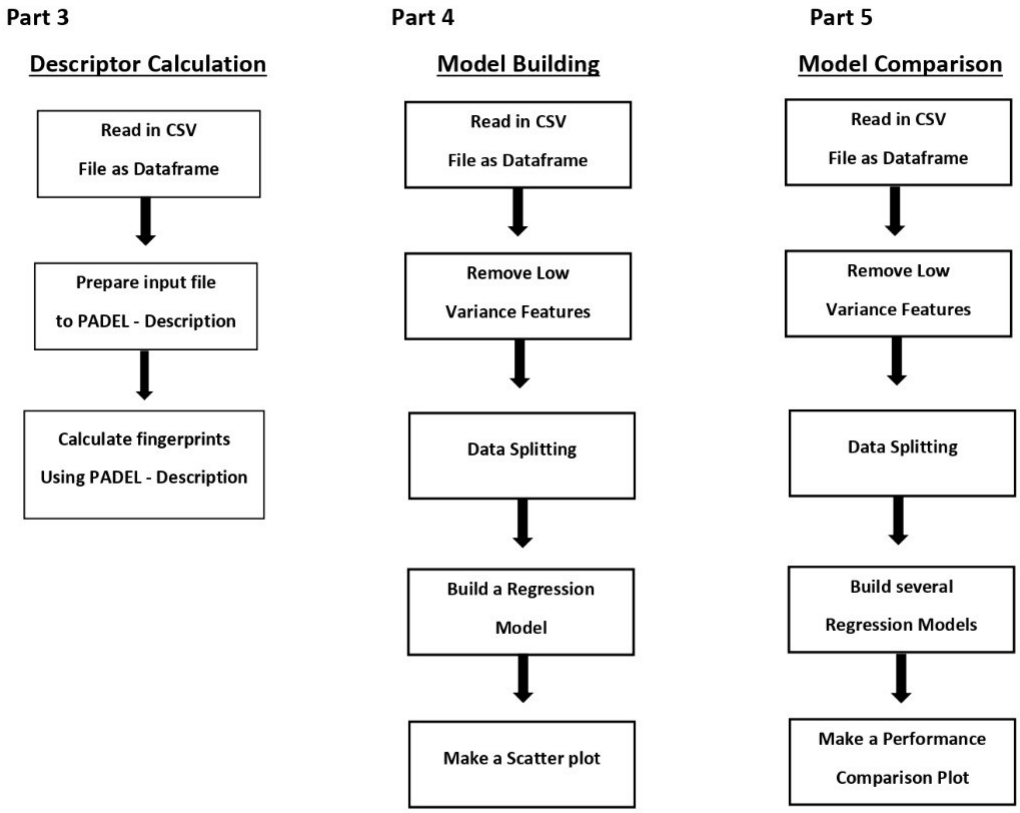


Figure 3: Detailed Design

- **PART-3 : Descriptor calculation**
  The Clean SMILES notation is exported as a CSV file and read as a Dataframe, an input file is prepared and PADEL-Descriptor to calculate molecular descriptors and fingerprints.
- **PART-4 : Model Building**
  The data is read from the CSV File, it is split into Training and Test set, Low variance features are removed and a Random forest Regressor Model is fit to predict pIC50 values and scatter plots are made comparing Predicted and Actual values using evaluation metrics after which hyperparameter tuning is done to search for best values of hyperparameters.
- **PART-5 : Model Comparison**
  Several Regression models are fit to the training dataset by firstly removing low variance features and then all these models are compared by making performance comparison plots w.r.t R2 scores and RMSE.

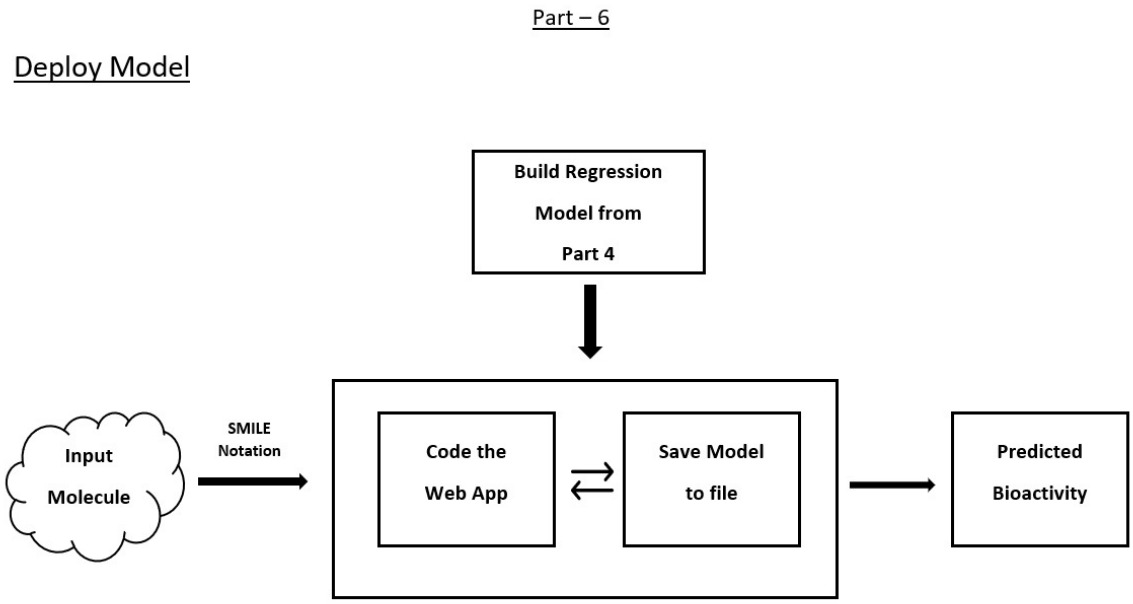


Figure 4: Detailed Design

- **PART-6 : Deploy Model as Web App**
  Web app takes in user csv input and fits the ML model to provide prediction output of ChemBL ID and corresponding pIC50 values for comparison and selection of best inhibitory compounds.

3.3 Sequence Design

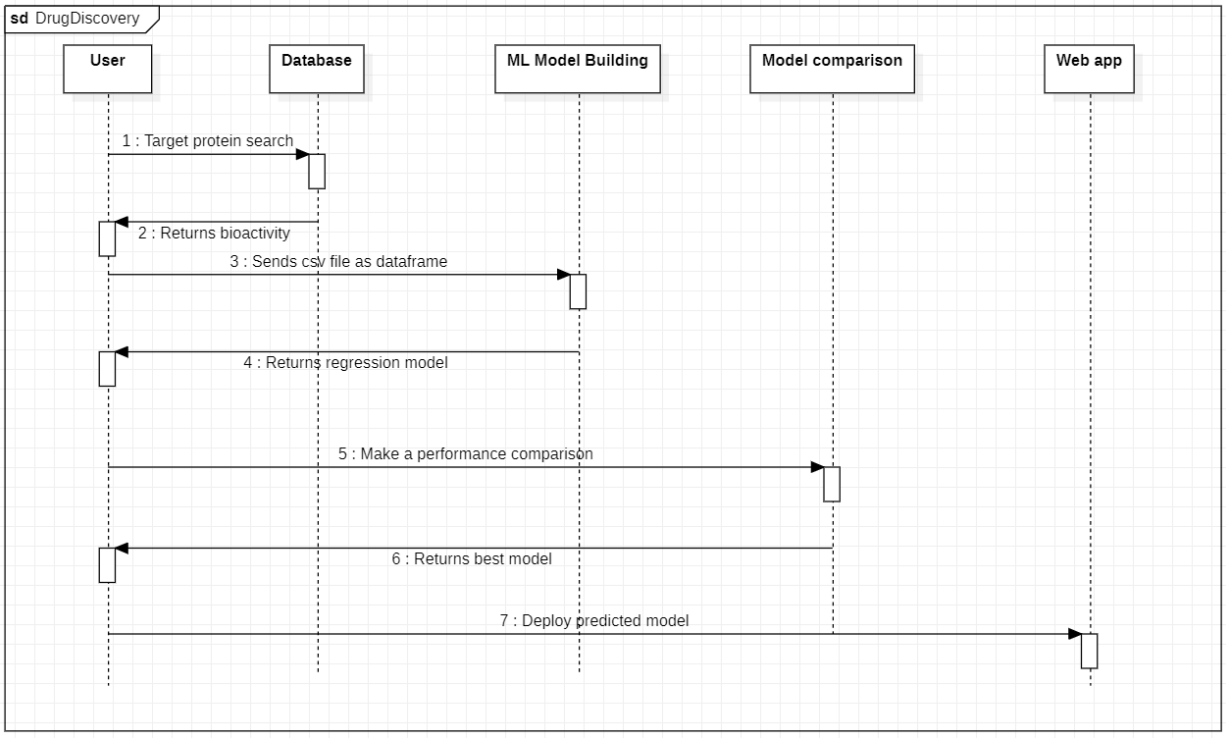


Figure 5: Sequence Design

The user searches the target protein in the database, here target protein is protein on which drug will act on. The database returns bioactivity of the protein with the drug. The data collected is used to build various ML models. Then there is a comparison based on their performance and the best model is returned to the user. The same is deployed to the web app for interactive user experience.

3.4 Use Case Design

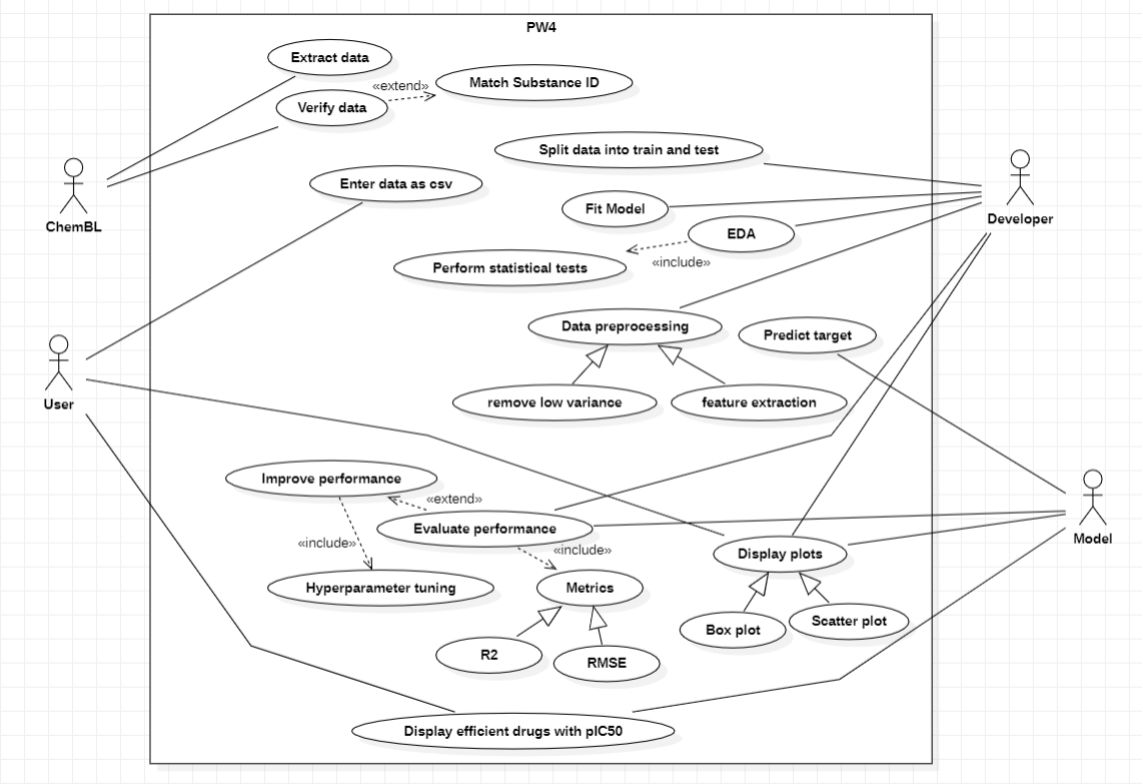


Figure 6: Use Case Design

- From ChemBL (Online database for bioactivity of drug-like compounds), relevant data can be extracted and verified by matching it with Substance ID.
- Developers are involved in model building and deployment. They get the input dataset, appropriately split it, do data-preprocessing, Exploratory data analysis using statistical tests, fit Models to derive necessary output and evaluate performance using pre-defined metrics.
- Models that are designed fit the training datasets, learn their parameters and predict outputs upon which performance comparison is done. They can generate plots to display outputs and output performance.
- Users can enter data in csv format to the Web app, view plots and save the predicted output with ChemBL ID and pIC50 values.

4 Implementation

4.1 Proposed methodology

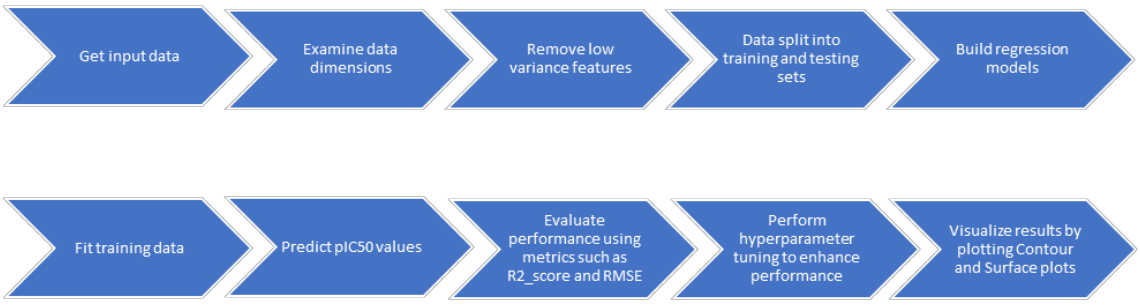


Figure 7: Proposed Methodology

4.2 Algorithm used for implementation

Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

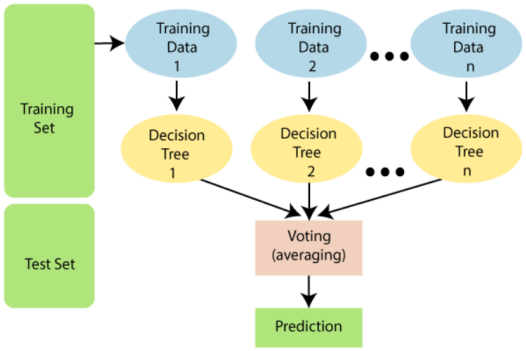


Figure 8: Random Forest

4.3 Tools and Technologies Used

- Jupyter Notebook / Google colab
- Conda
- Python libraries (numpy, pandas, matplotlib, scikit-learn)
- RDkit
- PaDEL Descriptors
- Lazy predict
- Streamlit (for Web App)

4.4 Testing

```
pd.DataFrame(data={'Actuals': Y_test, 'Predictions': Y_pred})
```

	Actuals	Predictions
23	9.301030	8.657282
29	5.080922	5.380648
30	5.031517	5.293212
32	4.744727	5.153417
44	4.301030	5.746938
...	...	...
1937	5.337242	5.409336
1946	5.267606	5.756796
1947	5.443697	5.756796
1950	5.408935	5.756796
1951	5.086186	7.682671

392 rows × 2 columns

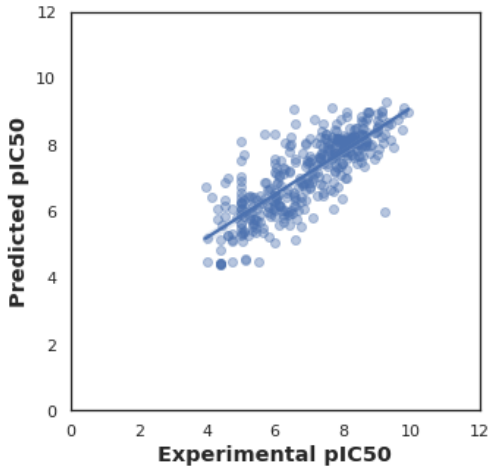


Figure 9: Testing

5 Results and Discussions

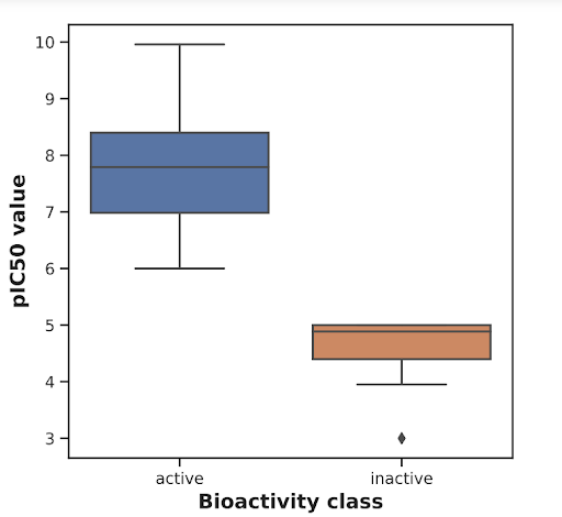


Figure 10: Bioactivity vs pIC50 plot

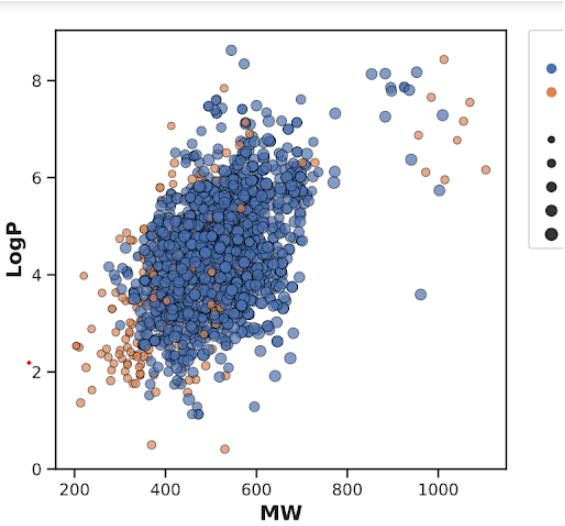


Figure 11: MW vs LogP plot

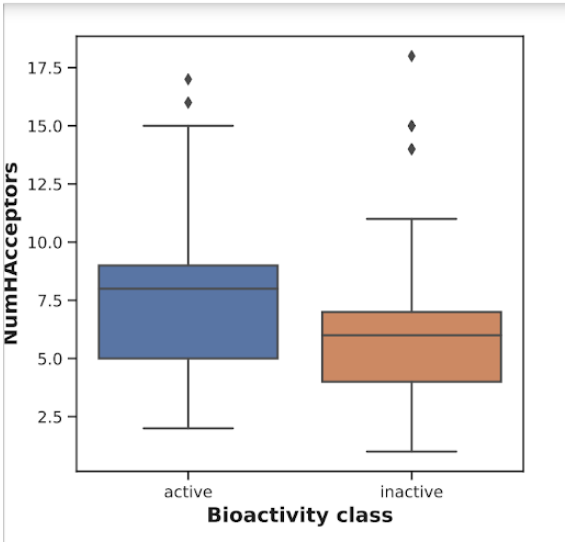


Figure 12: Bioactivity vs NumHAcceptors plot

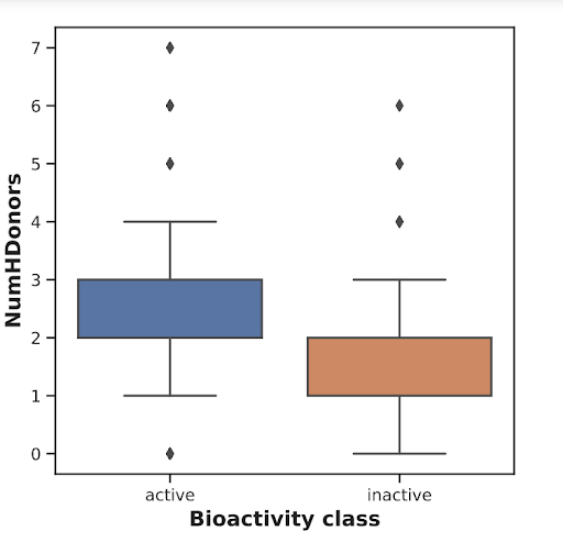


Figure 13: Bioactivity vs NumHDonors plot

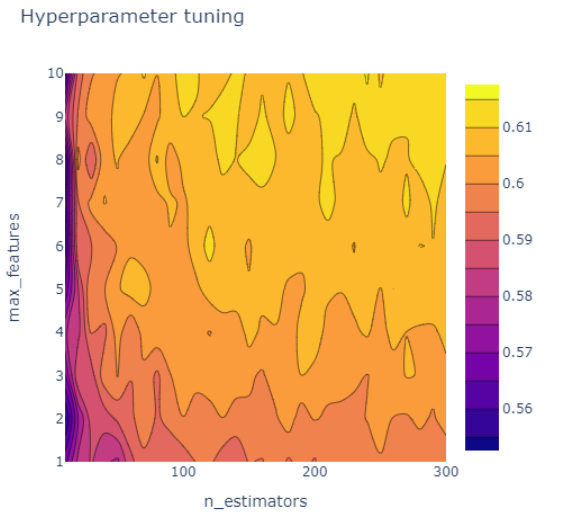


Figure 14: 2D Contour plot

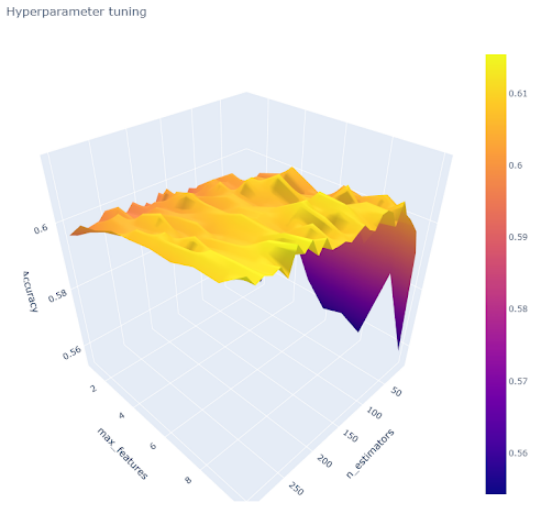


Figure 15: 3D Surface plot

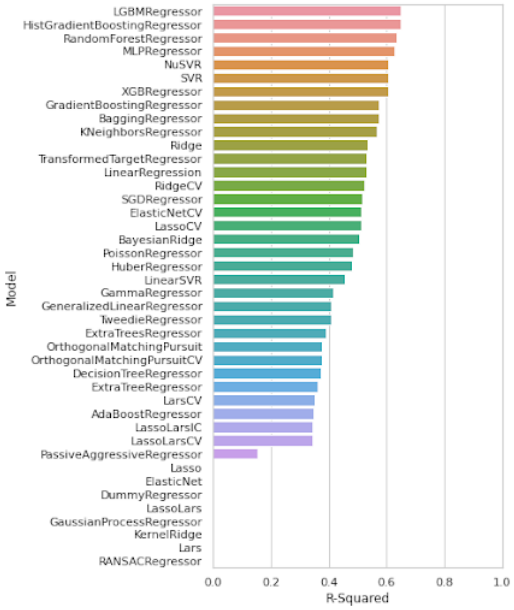


Figure 16: R2 performance plot comparing Regressors

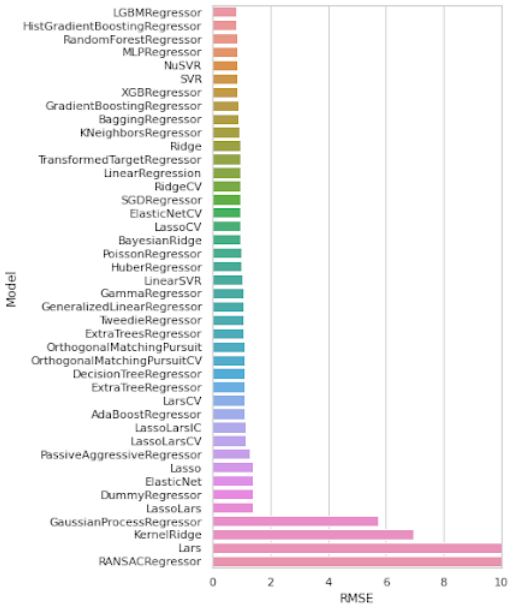


Figure 17: RMSE performance plot comparing Regressors

Prediction output

	Molecule_name	pIC50
0	CHEMBL867052	5.5006
1	CHEMBL867052	5.0719
2	CHEMBL867052	5.1835
3	CHEMBL867052	4.8830
4	CHEMBL867052	5.7715
5	CHEMBL867052	4.6718
6	CHEMBL867052	5.2923
7	CHEMBL867052	6.0569
8	CHEMBL867052	5.4651
9	CHEMBL867052	5.3369

Download Predictions

Figure 18: Predicted Output: pIC50 values of drugs

6 Conclusion and Future Work

The methodical process through which new candidate medications are found is known as drug discovery. It is a difficult, risky, time-consuming, yet potentially very profitable process. Bioinformatic analysis can speed up the identification of therapeutic targets, the screening of drug candidates, and the refinement of those candidates. It can also make it easier to characterize side effects and anticipate drug resistance.

To create a model, various ML techniques are applied to the training set of data. The best chemicals (based on pIC50) acting on ALK to inhibit cancer growth are then found using the model.

Prediction, identification, and storage of data relating to physiologically active candidates are the main focuses of current bioinformatics techniques. To locate targets for drug repurposing and identify new therapeutics, we use data analysis and machine learning.

The process can be shortened by conducting additional research on potential therapeutic targets, resulting in speedier and more effective pharmaceuticals entering the market and saving lives.

Better Deep learning methods can be used for increasing the efficiency of the model. The performance of the rf value can be enhanced by taking more parameters for hyperparameter tuning.

References

Literature survey

[1] J.C. Gertrudesa, V.G. Maltarollob, R.A. Silvaa, P.R. Oliveiraa, K.M. Honórioa,b and A.B.F. da Silva, "Machine Learning Techniques and Drug Design", 2012 Bentham Science Publishers, Current Medicinal Chemistry, 2012, 19, 4289-4297.

[2] Yasen Jiao and Pufeng Du, "Performance measures in evaluating machine learning based bioinformatics predictors for classifications", Higher Education Press and Springer-Verlag Berlin Heidelberg 2016,Quantitative Biology 2016, 4(4): 320–330.

[3] H.C. Stephen Chan, Hanbin Shan, Thamani Dahoun, Horst Vogel, Shuguang Yuan, "Advancing Drug Discovery via Artificial Intelligence", Trends in Pharmacological Sciences, Volume 40, Issue 8, August 2019, Pages 592-604.

[4] Xiao-Meng Zhang, Li Liang , Lin Liu, and Ming-Jing Tang, "Graph Neural Networks and their current applications in Bioinformatics", frontiers in Genetics, METHODS article, 29 July 2021.

[5] Choudhury C, Arul Murugan N, Priyakumar UD, "Structure-based drug repurposing: Traditional and advanced AI/ML-aided methods", Drug Discovery Today. 2022 Mar 14:S1359-6446(22)00112-X.

[6] Pillai N, Dasgupta A, Sudsakorn S, Fretland J, Mavroudis PD, "Machine Learning guided early drug discovery of small molecules" Drug Discovery Today. 2022 Mar 29:S1359-6446(22)00127-1.

[7] Gaudelet, Thomas, et al. "Utilizing graph machine learning within drug discovery and development." Briefings in bioinformatics 22.6 (2021): bbab159.

[8] Shaoqi Chen1, Dongyu Xue1, Guohui Chuai1, Qiang Yang2,3, Qi Liu1,"FL-QSAR: a federated learning based QSAR prototype for collaborative drug discovery", Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, 200092, China.

[9] Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Charles Tapley Hoyt, William L Hamilton, "Understanding the Performance of Knowledge Graph Embeddings in Drug Discovery", arXiv:2105.10488v3, 9 March 2022.

[10] Amanda J. Minnich, Kevin McLoughlin, Margaret Tse, Jason Deng, Andrew Weber, Neha Murad, Benjamin D. Madej, Bharath Ramsundar, Tom Rush, Stacie Calad-Thomson, Jim Brase, and Jonathan E. Allen, "AMPL: A Data-Driven Modeling Pipeline for Drug

Discovery", Journal of Chemical Information and Modeling 2020 60 (4), 1955-1968.

[11] Yinqiu Xu, Hequan Yao Kejiang Lin,"An overview of neural networks for drug discovery and the inputs used ",Expert opinion on drug discovery 2018, Vol. 13, No. 12, 1091-1102.

[12] Zhengyang Wang, Meng Liu, Youzhi Luo, Zhao Xu, Yaochen Xie, Limei Wang, Lei Cai, Qi Qi, Zhuoning Yuan, Tianbao Yang, Shuiwang Ji, "Advanced Graph and Sequence Neural Networks for molecular property prediction and drug discovery", Bioinformatics, Volume 38, Issue 9, 1 May 2022, Pages 2579–2586.

[13] Chandak T, Wong CF, "EDock-ML: A web server for using ensemble docking with machine learning to aid drug discovery", Protein Science, 2021, 30 :1087–1097.

[14] Abbas K, Abbasi A, Dong S, Niu L, Yu L, Chen B, Cai SM, Hasan Q, "Application of network link prediction in drug discovery", BMC Bioinformatics 2021 Apr 12; 22(1): 187.

[15] Leonardo L.G.Ferreira, Adriano D.Andricopulo, "ADMET modeling approaches in drug discovery", Drug Discovery Today, Volume 24, Issue 5, May 2019, Pages 1157-1165.

[16] Fabio Boniolo, Emilio Dorigatti, Alexander J. Ohnmacht, Dieter Saur, Benjamin Schubert Michael P. Menden (2021), "Artificial intelligence in early drug discovery enabling precision medicine", Expert Opinion on Drug Discovery, 2021 Sep; 16(9): 991-1007.

[17] Li P, Wang J, Qiao Y, Chen H, Yu Y, Yao X, Gao P, Xie G, Song S, "An effective self-supervised framework for learning expressive molecular global representations to drug discovery", Brief Bioinform. 2021 Nov 5; 22(6): bbab109.

[18] Rifaioglu, Ahmet Süreyya, et al. "MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery." Bioinformatics 37.5 (2021): 693-704.

[19] Olayan, Rawan S., Haitham Ashoor, and Vladimir B. Bajic. "DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches." Bioinformatics 34.7 (2018): 1164-1173.

[20] Christian Scheeder, Florian Heigwer, Michael Boutros, "Machine learning and image-based profiling in drug discovery", Current Opinion in Systems Biology, Volume 10, August 2018, Pages 43-52.

Others

1. https://www.datacamp.com/tutorial/streamlit

2. Saw Simeon, Nuttapat Anuwongcharoen, Watshara Shoombuatong1 Aijaz Ahmad Malik, Virapong Prachayasittikul, Jarl E.S. Wikberg, Chanin Nantasenamat, "Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking", PeerJ, August 9, 2016.

3. Sullivan, Ivana, and David Planchard. "ALK inhibitors in non-small cell lung cancer: the latest evidence and developments." Therapeutic advances in medical oncology 8.1 (2016): 32-47.

4. https://medlineplus.gov/genetics/gene/alk/