Visvesvaraya Technological University, Belagavi – 590018

PROJECT WORK PHASE - 2 REPORT
ON

Predicting Bioactivity Using Chemical Fingerprint Models and Cell Morphology with Deep Neural Networks

Submitted in partial fulfillment of the requirements for the degree

BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE & ENGINEERING

Submitted by

EADOIN LOBO	**4SO20CS045**
JEROME JOSEPH	**4SO20CS065**
SAMSON EIOAN DSOUZA	**4SO20CS135**
SHAUN MATHEW CRASTA	**4SO20CS144**

Under the Guidance of

Dr Shrisha H S
Associate Professor, Department of CSE

DEPT. OF COMPUTER SCIENCE AND ENGINEERING
ST JOSEPH ENGINEERING COLLEGE
An Autonomous Institution
(Affiliated to VTU Belagavi, Recognized by AICTE, Accredited by NBA)
Vamanjoor, Mangaluru - 575028, Karnataka

2023-24

ST JOSEPH ENGINEERING COLLEGE

An Autonomous Institution

(Affiliated to VTU Belagavi, Recognized by AICTE, Accredited by NBA)

Vamanjoor, Mangaluru - 575028, Karnataka

DEPT. OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"Predicting Bioactivity Using Chemical Fingerprint Models and Cell Morphology with Deep Neural Networks"** carried out by

EADOIN LOBO	**4SO20CS045**
JEROME JOSEPH	**4SO20CS065**
SAMSON EIOAN DSOUZA	**4SO20CS135**
SHAUN MATHEW CRASTA	**4SO20CS144**

the bonafide students of VIII semester Computer Science & Engineering in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi during the year 2023-2024. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the said degree.

Dr Shrisha H S	**Dr Sridevi Saralaya**	**Dr Rio D'Souza**
Project Guide	HOD-CSE	Principal

External Viva:

Examiner's Name	Signature with Date
1.	.....
2.	.....

ST JOSEPH ENGINEERING COLLEGE

An Autonomous Institution

(Affiliated to VTU Belagavi, Recognized by AICTE, Accredited by NBA)

Vamanjoor, Mangaluru - 575028, Karnataka

DEPT. OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We hereby declare that the entire work embodied in this Project Report Titled **"Predicting Bioactivity Using Chemical Fingerprint Models and Cell Morphology with Deep Neural Networks"** has been carried out by us at St Joseph Engineering College, Mangaluru under the supervision of **Dr Shrisha H S,** for the award of **Bachelor of Engineering** in **Computer Science & Engineering**. This report has not been submitted to this or any other University for the award of any other degree.

EADOIN LOBO	**4SO20CS045**
JEROME JOSEPH	**4SO20CS065**
SAMSON EIOAN DSOUZA	**4SO20CS135**
SHAUN MATHEW CRASTA	**4SO20CS144**

Acknowledgement

We dedicate this page to acknowledge and thank those responsible for the shaping of the project. Without their guidance and help, the experience while constructing the dissertation would not have been so smooth and efficient.

We sincerely thank our Project Guide **Dr Shrisha H S**, Associate Professor, Computer Science and Engineering for his guidance and valuable suggestions which helped us to complete this project. We also thank our Project coordinator **Ms Supreetha D R**, Dept of CSE, for her consistant encouragement.

We owe a profound gratitude to **Dr Sridevi Saralaya**, Head of the Department, Computer Science and Engineering, whose kind support and guidance helped us to complete this work successfully.

We are extremely thankful to our Principal, **Dr Rio D'Souza**, for his valuable guidance and encouragement throughout the project.

We express my deep gratitude to our Director, **Rev. Fr Wilfred Prakash D'Souza**, and Assistant Director, **Rev. Fr Kenneth Rayner Crasta** for their support and encouragement.

We would like to thank all faculty and staff of the Deprtment of Computer Science and Engineering who have always been with us extending their support, precious suggestions, guidance, and encouragement through the project.

We also extend our gratitude to our friends and family members for their continuous support.

Abstract

This project is dedicated to advancing predictions for potential pharmaceutical compounds. Through the integration of chemical fingerprints and cell morphology data using sophisticated neural networks, our objective is to elevate the precision of bioactivity predictions. A critical review of existing literature reveals a gap in the comprehensive exploration of this integration, prompting our research.

Our methodology involves a meticulously planned series of steps, including data collection, preprocessing, model development, training, evaluation, and integration. The feasibility of this approach is substantiated by the availability of relevant datasets and the incorporation of established practices in deep learning. In conclusion, our project modestly contributes to the enhancement of bioactivity predictions, thereby bolstering the efficiency of drug discovery processes. The potential benefits extend to resource savings in terms of cost, labor, space, and energy. As we look ahead, future work may explore broader applications and refine predictive models to meet evolving demands in pharmaceutical research.

Table of Contents

Chapter 1

Introduction

1.1 Background

The landscape of drug discovery has witnessed a transformation with the integration of computational methods and artificial intelligence. Conventional approaches to predicting bioactivity in drug candidates often rely on chemical fingerprint models or the analysis of cell morphology data independently. However, the limitations of these singular modalities have prompted a surge in interest in developing integrative solutions. This project is situated against the backdrop of the growing need for more accurate and nuanced bioactivity predictions. The integration of chemical and cell morphology data through deep neural networks presents an innovative approach to addressing the challenges faced by traditional methods.

Deep learning techniques, particularly deep convolutional neural networks, have demonstrated remarkable effectiveness in diverse domains such as image and speech recognition. In the realm of drug discovery, the hierarchical and compositional nature of deep learning architectures remains underexplored. The proposed project, inspired by recent advancements, aims to apply these principles to structure-based drug discovery. AtomNet, a deep convolutional neural network introduced in a seminal paper, serves as a pioneering example of utilizing the spatial and temporal structure of molecular information for bioactivity prediction. This project seeks to build upon such foundations, pushing the boundaries of integration between chemical and cellular data for enhanced predictive modeling.

Accurate bioactivity prediction holds paramount importance in the context of drug development, significantly impacting the efficiency of screening processes and allowing researchers to prioritize the most promising drug candidates. With the surge in available chemical and biological data, the proposed project takes a forward-looking approach by leveraging the wealth of information available on the internet. By combining chemical fingerprints and cell morphology data in a unified deep learning framework, the project aims to contribute to the ongoing evolution of drug discovery methodologies.

1.2 Problem statement

The goal of our project is to enhance the prediction accuracy of chemical compound bioactivity for pharmaceutical applications by developing an innovative deep learning-based model. This model integrates chemical fingerprint data and cell morphology information, addressing the limitations of existing predictive approaches that rely on singular data types. Our research aims to streamline the drug discovery process, significantly reducing the time and resources spent on experimental validation of potential drug candidates by leveraging a comprehensive, multi-modal dataset. Through this integration, we aim to unveil intricate patterns and interactions that singular approaches might overlook, thus pushing the boundaries of current methodologies in bioactivity prediction and offering a more efficient pathway to identifying promising drug candidates.

1.3 Scope

The primary aim of this project is to develop an advanced predictive model for accurately assessing the bioactivity of chemical compounds. By integrating chemical fingerprints and cell morphology data through deep neural networks, our goal is to enhance the efficiency of identifying potential drugs. This approach addresses the limitations of traditional drug discovery methods, offering a more accurate and streamlined process. In summary, our project seeks to create a robust and effective tool that saves time and resources in drug discovery by providing researchers with a comprehensive understanding of bioactivity. The successful implementation of our predictive model has the potential to significantly improve the precision and effectiveness of identifying promising drug candidates.

Chapter 2

Literature Survey

This chapter gives the details of the various works which were carried out for Bioactivity Prediction.

2.1 Paper 1 - Predicting Compound Activity from Phenotypic Profiles and Chemical Structures

This project systematically evaluates the relative strength of three highthroughput data sources for predicting compound bioactivity. By synergizing chemical structures, imaging, and gene-expression profiles, the study enhances prediction accuracy, potentially expediting the early stages of drug discovery. The research underscores the potential of unbiased phenotypic profiling to advance compound bioactivity prediction, contributing valuable insights to the field.

Identified Problem: The study focuses on addressing the high cost and slow pace associated with drug discovery, underscoring the critical need for more efficient methods.

Methodology & Implementation: The research methodology integrates the evaluation of three high-throughput data sources—chemical structures, imaging, and gene-expression profiles—to determine their collective impact on the prediction accuracy of compound bioactivity. To ensure the robustness and reliability of predictions, Bemis-Murcko clustering techniques are employed, facilitating the analysis of molecular frameworks and identification of core chemical structures pivotal in biological interactions. For practical implementation, a suite of software tools including CellProfiler and the Bemis-Murcko clustering algorithm are utilized. CellProfiler analyzes and processes imaging data, extracting and quantifying cell morphology features from high-throughput microscopy images, crucial for understanding compound effects on cellular structures and functions. Bemis-Murcko clustering categorizes compounds based on skeletal frameworks, exploring the relationship between molecular architecture and bioactivity. This compre-

hensive approach leverages computational techniques to uncover patterns and relationships across diverse data types, enhancing the efficiency of drug discovery processes by providing a more nuanced prediction model for compound bioactivity.

Results: The study finds that combining details about chemical structures, imaging, and gene-expression profiles improves predicting how compounds will act. This helps speed up drug discovery. Results show that using various types of information gives a clearer picture. It's like solving a puzzle with different pieces. Overall, the study deepens our understanding of how different factors influence drug effectiveness, which is crucial for developing new medicines quickly and efficiently.

Limitations/Future Scope: While the study acknowledges the potential for lower accuracy in practical applications and considers factors influencing assay predictability, it highlights the need for future research to explore additional high-throughput data sources and methodologies. This exploration aims to further enhance the efficiency and effectiveness of drug discovery processes.

2.2 Paper 2 - AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery

AtomNet pioneers a paradigm shift in drug discovery by introducing a deep convolutional neural network tailored for predicting the bioactivity of smallmolecules. In contrast to conventional docking approaches, AtomNet harnesses the spatial and temporal structure of the domain, yielding unprecedented predictive performance. Its superiority is evident as it outperforms existing methods, positioning itself as a formidable tool in structure-based drug discovery.

Identified Problem: The research addresses limitations in existing ligand-based techniques, such as their inability to predict for novel targets, dependence on known ligands, and constraints in pre-specified molecular structures during fingerprinting. The blind model to the target restricts guidance for medicinal chemists during molecule optimization.

Methodology & Implementation: AtomNet addresses limitations by integrating ligand and target structure information, requiring atom locations in the target binding site to discover molecular features and interactions. The construction process involves experimental benchmarks, data encoding, and the design of a deep convolutional network. Implementation employs deep convolutional neural networks to merge ligand and target

structure information, facilitating the discovery of molecular features and interactions. Tested on realistic benchmarks, AtomNet demonstrates efficacy in structure-based bioactivity prediction experiments, showcasing its potential for advancing drug discovery by enhancing the understanding of molecular interactions and facilitating more accurate predictions of compound bioactivity.

Results: AtomNet showcases remarkable proficiency in forecasting new active molecules for targets lacking known modulators, exhibiting superior performance over conventional docking methods. Its successful application to challenging benchmarks underscores its promising role in structure-based drug discovery, offering a robust platform for identifying molecular features and interactions crucial for drug development. The results imply that AtomNet's efficacy stems from its unique ability to integrate both ligand and target structure information, enabling the discovery of previously unrecognized active molecules. By surpassing traditional docking approaches, AtomNet demonstrates its capacity to enhance drug discovery by providing more accurate predictions of compound activity. The inference drawn from these findings is clear: the synergy between ligand and target structure information is pivotal for AtomNet's success. This integration empowers AtomNet to uncover molecular intricacies and predict new active molecules with greater precision and reliability. Moreover, AtomNet's superiority over conventional methods suggests a paradigm shift in structure-based drug discovery, emphasizing the importance of holistic approaches that consider both ligand and target characteristics. Overall, these results herald AtomNet as a promising tool for accelerating drug discovery efforts, offering new avenues for identifying potent compounds and advancing therapeutic development.

Limitations/Future Scope: While AtomNet is a significant advancement, considerations include the model's dependence on atom locations, which may pose challenges in practical implementation. Further research is needed for scalability and applicability across diverse targets and compounds. Future enhancements can focus on addressing these limitations and expanding AtomNet's capabilities.

2.3 Paper 3 - De Novo Design and Bioactivity Prediction of SARS-CoV-2 Main Protease Inhibitors using Recurrent Neural Network-based Transfer Learning

In response to the global urgency for antivirals against SARS-CoV-2, this project introduces a sophisticated deep learning platform for de novo design of potential inhibitors. Employing transfer learning, the model not only outshines existing counterparts but also

generates novel structures and identifies promising hits. This research significantly contributes to ongoing efforts against COVID-19, providing an innovative approach to drug discovery.

Identified Problem: The study addresses the need for drug candidates against SARS-CoV-2, emphasizing the urgency in the COVID-19 pandemic.

Methodology & Implementation: A recurrent neural network-based transfer learning approach is adopted for both de novo design and bioactivity prediction of potential SARS-CoV-2 main protease inhibitors. Implementation involves training a chemical model using ULMFit for de novo design, followed by fine-tuning a classifier for bioactivity prediction. Notably, the chemical space of generated molecules overlaps significantly with the target chemical space of Mpro inhibitors, ensuring relevance and potential efficacy in inhibiting the SARS-CoV-2 main protease.

Results: The platform demonstrates its capability by successfully generating molecules with properties resembling those of known SARS-CoV-1 Mpro inhibitors, suggesting potential efficacy in inhibiting the virus. The inference drawn from these results is significant: the utilization of deep learning techniques accelerates the discovery process of potential antiviral compounds against SARS-CoV-2. This highlights the effectiveness of advanced computational approaches in rapidly identifying promising candidates for combating viral infections, paving the way for expedited drug discovery efforts in the fight against emerging pathogens like SARS-CoV-2.

Limitations/Future Scope: Considerations include careful molecule selection and experimental confirmation. Future work may explore modifications' impact and intellectual property issues.

2.4 Paper 4 - Learning Machine Reasoning for Bioactivity Prediction of Chemicals

Responding to the demand for higher-level vector representations in bioactivity prediction, this project introduces "reason vectors." In contrast to high-dimensional chemical fingerprints, these vectors—comprising only about five dimensions—enable abstract reasoning for bioactivity. Facilitating potent similarity searches and handling novel feature combinations, the methodology offers a straightforward approach to activity prediction, applicable to both binary and continuous outcomes.

Identified Problem: The study addresses limitations in traditional QSAR models, emphasizing challenges in handling novel feature combinations.

Methodology & Implementation: The proposed method systematically learns abstract representations by reconstructing molecules stepwise, guided by reason vectors that consider the distributed, continuous fingerprints of building blocks. In implementation, the method undergoes testing across diverse datasets including AMES, AHR, SKINSENS, and LD50. Remarkably, it outperforms traditional QSARs in terms of accuracy, showcasing its efficacy in generating more precise predictions across various chemical datasets. This approach offers a systematic framework for learning molecular structures and their properties, leveraging reason vectors to guide predictions and reconstruct molecules. Its successful application across multiple datasets underscores its robustness and potential for advancing computational chemistry methods, providing valuable insights into chemical interactions and toxicity prediction. Ultimately, this methodology presents a promising avenue for enhancing predictive modeling in drug discovery and chemical risk assessment, offering improved accuracy and efficiency in compound evaluation and design.

Results: The proposed method attains superior sensitivity, specificity, and accuracy, showcasing its stability in predicting chemical bioactivity. These findings imply a dependable approach for bioactivity prediction based solely on chemical structure, facilitating abstract reasoning. The method's robust performance suggests its reliability and convenience in assessing compound effectiveness and potential interactions. Overall, these results underscore the method's efficacy in providing accurate predictions of chemical bioactivity, offering a promising avenue for enhancing drug discovery processes and understanding molecular interactions.

Limitations/Future Scope: Considerations include database size and diversity limitations. Future work may involve expanding the reference database and exploring advanced machine learning techniques.

2.5 Paper 5 - Bio-activity Prediction of Drug Candidate Compounds Targeting SARS-Cov-2 using Machine Learning Approaches

Addressing the pressing need for swift bioactivity predictions against SARSCoV-2, this project meticulously evaluates 27 machine learning classifiers. A neural network model is introduced, showcasing remarkable accuracy on ChEMBL and PubChem datasets. This model emerges as a cost-effective and timely solution for screening potential drug can-

didates against the novel coronavirus, holding significant implications for combatting COVID19.

Identified Problem: The research addresses the urgent need for effective drug candidates targeting SARS-CoV-2 during the COVID-19 pandemic.

Methodology & Implementation: Machine learning techniques are employed to predict the bioactivity of drug candidate compounds targeting SARS-CoV-2. Essential structural features are captured using various molecular representations, including fingerprints, descriptors, graphs, and SMILES strings. Diverse machine learning models, such as deep neural networks, support vector machines, random forests, and decision trees, are utilized. In implementation, data from ChEMBL and BioAssay databases are curated and used to train the model. Subsequently, the trained model evaluates candidate compounds for activity against SARS-CoV-2 using REDIAL, providing a comprehensive and effective framework for drug discovery and evaluation against the virus.

Results: The model achieves impressive results with 93% accuracy and a 0.94 f-1 score utilizing a neural network design. Active inhibitors identified in the study undergo further evaluation for potential activities against SARS-CoV-2 using REDIAL. From these findings, it can be inferred that machine learning effectively predicts the bioactivity of drug candidates against SARS-CoV-2, offering insights into crucial fingerprints, substructures, and their effects. This underscores the utility of machine learning approaches in drug discovery, particularly in rapidly assessing potential candidates for combating viral infections like SARS-CoV-2, thereby aiding in the development of effective therapeutic interventions.

Limitations/Future Scope: Limitations include the need for in-vitro and in-vivo studies to confirm efficacy and considerations of applicability domain issues. Future research can explore additional molecular descriptors and conduct in-depth studies for optimal drug creation.

2.6 Paper 6 - Bioactivity Prediction Using Convolutional Neural Network

Identified Problem: The study focuses on the challenge of predicting the bioactivity of compounds based on their chemical structure. This is crucial for drug discovery, as structurally similar compounds often exhibit similar properties and biological activities. The existing computational systems have limitations in accurately predicting molecular binding, necessitating physical experiments for confirmation. The paper identifies the

need for a more reliable predictive system that can enhance the efficiency and accuracy of bioactivity prediction, reducing the reliance on costly and time-consuming physical experiments.

Methodology & Implementation: The research introduces a novel predictive system that uses a CNN to predict molecular bioactivities using a unique molecular matrix representation. This approach involves comparing the performance of the proposed system with three classical machine learning algorithms across various datasets. The methodology includes the development of a new molecular representation technique, Mol2fgs, which characterizes compounds within an n x n matrix to capture their molecular properties. The datasets used for validation include the MDDR and ChEMBL datasets, aimed at validating the molecule classification based on structure-activity relationships. The study designs a deep convolutional network architecture and evaluates its performance in comparison to other machine learning algorithms implemented in the WEKA Workbench.

Results: The CNN model demonstrated superior prediction accuracy, with a rate of 90.21%, significantly outperforming the classical machine learning algorithms it was compared against. The study finds that molecules with atom numbers between 9 to 11 offer the best representation for bioactivity prediction using the proposed model. This showcases the CNN model as a potentially stable and efficient approach for the activity prediction of unknown chemical compounds.

Limitations/Future Scope: While the CNN model shows promising results, the study acknowledges the need for continued exploration in this area to enhance predictive accuracy further and to apply the model across a wider array of datasets and compound classes. The paper suggests a significant opportunity for future research to improve upon the findings and extend the applicability of the model in the field of drug discovery.

2.7 Paper 7 - Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data

Identified Problem: The research identifies the limitation of machine learning models trained on structural fingerprints in predicting biological endpoints due to the lack of diversity in the chemical space of the training data. It highlights the challenge of extending the applicability domain of structural models to improve the reliability of predictions for new compounds.

Methodology & Implementation: The study introduces similarity-based merger

models that combine predictions from individual models trained on cell morphology (using Cell Painting) and chemical structure (using chemical fingerprints). The merger employs logistic regression models, utilizing both predictions and similarities as features, to predict assay hit calls for 177 assays from databases like ChEMBL, PubChem, and the Broad Institute. This approach seeks to expand the applicability domain by better extrapolating to new structural and morphological spaces.

Results: The similarity-based merger models demonstrated improved performance over individual models and other baseline models, with an additional 20% of assays (79 out of 177 assays) achieving an AUC ¿ 0.70. This indicates that merging structure and cell morphology models can more accurately predict a wide range of biological assay outcomes and expand the applicability domain.

Limitations/Future Scope: While the models showed promising results, the study acknowledges the need for further exploration to enhance predictive accuracy further and extend the model's applicability across different types of bioactivity data and datasets. Future research could explore improving the predictive models and their application in identifying active compounds, predicting selectivity profiles, and selecting compounds for progression.

2.8 Paper 8 - Imputation of Assay Bioactivity Data using Deep Learning

Concentrating on the imputation of assay pIC50 values, this project unveils a novel deep learning neural network method. Trained on sparse bioactivity data typical of public and commercial databases, the method outperforms traditional quantitative structure-activity relationship (QSAR) models. By augmenting accuracy and precision, the project presents a promising methodology for imputing assay values, holding potential implications for drug discovery.

Identified Problem: The paper addresses the challenge of sparse experimental data availability in drug discovery, which hinders the efficient identification of new hits and the progression of compounds through various stages of development. The sparse nature of public and commercial bioactivity databases limits the potential insights that could be derived if the missing data were accurately imputed. Traditional QSAR models and machine learning approaches struggle with sparse data, which motivates the development of a new method that can effectively utilize incomplete bioactivity information.

Methodology & Implementation: The methodology revolves around a deep learning neural network that can learn from sparse bioactivity data and molecular descriptors to predict assay bioactivity values. This neural network is capable of exploiting correlations between different bioactivities and between molecular descriptors and bioactivities. A key feature of this approach is its ability to handle missing data, enabling it to learn from incomplete datasets typical of those found in drug discovery. The paper also introduces a method for estimating the uncertainty of each prediction, allowing for a focus on the most confident results. The performance of this neural network was compared with traditional QSAR models, random forests, multi-target deep neural networks, and matrix factorization approaches using public domain datasets.

Results: The neural network method showed superior performance in imputing assay bioactivity data compared to traditional QSAR models and other machine learning approaches. Specifically, the method achieved a significant improvement in prediction accuracy, with the ability to increase the accuracy to $R^2 > 0.9$ for the most confident predictions compared to lower accuracies when reporting all predictions. This demonstrates the method's effectiveness in accurately imputing missing bioactivity data from sparse datasets.

Limitations/Future Scope: While the neural network approach demonstrates promising results, the paper acknowledges the need for further validation and exploration of its applicability across different types of bioactivity data and datasets. The ability of the method to estimate prediction uncertainty and focus on the most confident results offers a valuable tool for drug discovery, suggesting its potential for broader application in identifying active compounds, predicting selectivity profiles, and selecting compounds for progression. Future research could explore the method's effectiveness in various contexts and its integration into drug discovery workflows.

2.9 Comparison of Existing Methods:

Paper Name	Problem	Method	Results	Limitations
Predicting Compound Activity from Phenotypic Profiles and Chemical Structures	High cost and slow drug discovery process.	Evaluation of three data sources (chemical structures, imaging, gene-expression profiles).	Improved prediction accuracy by combining data sources.	Lower accuracy in practical applications; need for further research.
AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery	Limitations of existing ligand-based techniques.	Deep convolutional neural network that integrates ligand and target structure information.	Superior performance in predicting bioactivity.	Model's dependence on atom locations; needs scalability improvement.
De Novo Design and Bioactivity Prediction of SARS-CoV-2 Main Protease Inhibitors using Recurrent Neural Network-based Transfer Learning	Need for drug candidates against SARS-CoV-2.	Recurrent neural network-based transfer learning for de novo design and bioactivity prediction.	Successful generation of molecules with potential efficacy against SARS-CoV-2.	Need for experimental confirmation; intellectual property issues.
Learning Machine Reasoning for Bioactivity Prediction of Chemicals	Challenges in handling novel feature combinations in QSAR models.	Method learning abstract representations through "reason vectors".	Superior sensitivity, specificity, and accuracy.	Database size and diversity limitations; exploring advanced techniques.
Bio-activity Prediction of Drug Candidate Compounds Targeting SARS-Cov-2 using Machine Learning Approaches	Urgent need for effective drug candidates against SARS-CoV-2.	Evaluation of 27 machine learning classifiers for bioactivity prediction.	High accuracy and f-1 score with a neural network model.	Need for in-vitro and in-vivo studies; applicability domain issues.
Bioactivity Prediction Using Convolutional Neural Network	Predicting bioactivity of compounds is challenging due to structural similarities.	Convolutional Neural Network (CNN) for molecular bioactivities prediction.	CNN model demonstrated superior prediction accuracy.	Need for enhanced predictive accuracy and broader applicability.
Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data	Limitation of machine learning models trained on structural fingerprints.	Similarity-based merger models combining cell morphology and chemical structure predictions.	Improved performance over individual models and other baselines.	Need for further exploration to enhance predictive accuracy.
Imputation of Assay Bioactivity Data using Deep Learning	Sparse experimental data availability in drug discovery.	Deep learning neural network for imputing assay bioactivity data.	Superior performance in imputing assay bioactivity data.	Need for further validation and exploration across different datasets.

Figure 2.1: Comparison

2.10 Proposed System:

Importance of the Chosen Problem/Project

The chosen problem of predicting bioactivity is crucial in the field of drug discovery. Accurate predictions help identify potential drug candidates more efficiently, reducing the time and resources required for experimental testing. This is particularly significant given the vast number of compounds to be screened in drug discovery, making computational methods a valuable tool for prioritization.

Novelty in Proposed Project Work

The primary novelty in the proposed project lies in leveraging deep neural networks for bioactivity prediction. While existing works often focus on either chemical features or cellular responses, the proposed project not only integrates chemical fingerprint models and cell morphology data but does so through advanced neural networks. This unique approach introduces a more holistic perspective, considering not only the molecular structure and cellular morphology but also harnessing the power of neural networks. The application of deep learning techniques has the potential to uncover hidden relationships and significantly improve the overall accuracy of bioactivity predictions.

Advancement in the State-of-the-Art

The proposed project advances the state-of-the-art by providing a more comprehensive and accurate bioactivity prediction model. By utilizing deep neural networks to analyze integrated chemical and cell morphology data, the project aims to outperform existing methods. This advancement contributes to the efficiency of drug discovery processes, enabling researchers to identify promising drug candidates with greater confidence.

Differences from Existing Works ?

Unlike existing works that often focus on single-modal data or traditional machine learning approaches, the proposed project distinguishes itself by combining chemical and cell morphology features through deep neural networks. This unique integration, coupled with the application of neural networks, sets the project apart, exploring a more diverse set of data sources and employing advanced learning techniques. This approach holds the potential to offer superior predictive capabilities compared to the surveyed works.

Chapter 3

Methodology

3.1 Introduction

The development of an advanced neural network-based system for predicting compound bioactivity incorporates diverse data sources, including chemical structures and cell morphology data. This methodology section outlines the architecture, design, and specific neural network types employed for each component within this comprehensive approach.

3.2 Architecture

The system is designed to process heterogeneous data sources through specialized neural networks, culminating in an aggregation model that synthesizes predictions and additional data to provide a nuanced bioactivity prediction. The architecture includes:

3.2.1 Chemical Structure Analysis Network (ChemNet)

1. Neural Network Type
    - Dense Neural Network (DNN)
2. Purpose
    - Processes Morgan Fingerprint data to estimate the probability of bioactivity based on the compound's chemical structure.
3. Output
    - Probability score of bioactivity considering chemical structure (FP-Prob).

3.2.2 Cellular Morphology Analysis Network (CellNet)

1. Neural Network Type
    - Dense Neural Network (DNN)
2. Purpose
    - Analyzes Cell Painting feature data (numerical) to predict bioactivity probability based on cell morphology insights.
3. Output
    - Probability of bioactivity considering cellular morphology (CP-Prob).

3.2.3 Bioactivity Aggregation Network (BioNet)

1. Neural Network Type
    - Dense Neural Network Dense Neural Network (DNN) with Attention Mechanisms
2. Purpose
    - Integrates FP-Prob and CP-Prob, alongside additional similarity scores and fields, to finalize the compound's bioactivity prediction.
3. Output
    - Final prediction of bioactivity status.

3.3 Design

3.3.1 ChemNet Design (Dense Neural Network)

Designed to understand and predict based on chemical structures:

1. Input Layer
    - Scaled to accommodate Morgan Fingerprints, typically 2048 bits.
2. Hidden Layers
    - Features several dense layers with ReLU activation functions to capture complex chemical structure patterns. Includes dropout layers for overfitting prevention.
3. Output Layer
    - Utilizes a sigmoid activation to deliver the FP-Prob.

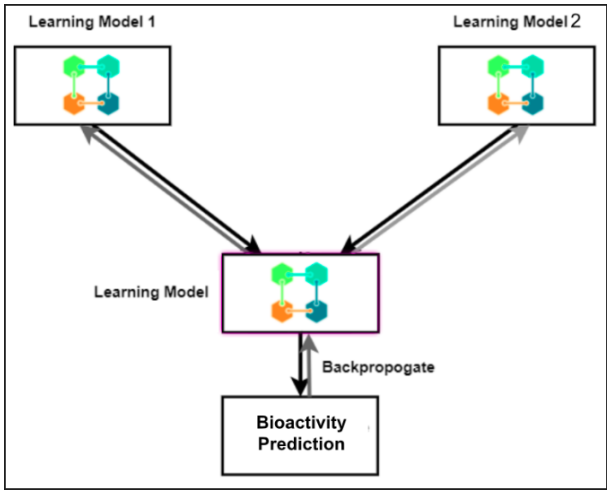

Figure 3.1: Design Model

3.3.2 CellNet Design (Dense Neural Network)

Optimized for numerical data analysis from Cell Painting features:

1. Input Layer
    - Configured to accept the dimensionality of the preprocessed Cell Painting features.
2. Hidden Layers
    - Mirrors the structure of ChemNet with dense layers and ReLU activation, tailored to learn from cellular morphology data efficiently. Dropout layers are interspersed to mitigate overfitting risks.
3. Output Layer
    - A sigmoid-activated neuron outputs the CP-Prob.

3.3.3 BioNet Design (Dense Neural Network with Attention Mechanisms)

Sophisticated integration of predictive scores and additional data for bioactivity assessment:

1. Input Layer
    - Receives an integrated vector containing FP-Prob, CP-Prob, and additional data such as similarity scores.
2. Attention Mechanism

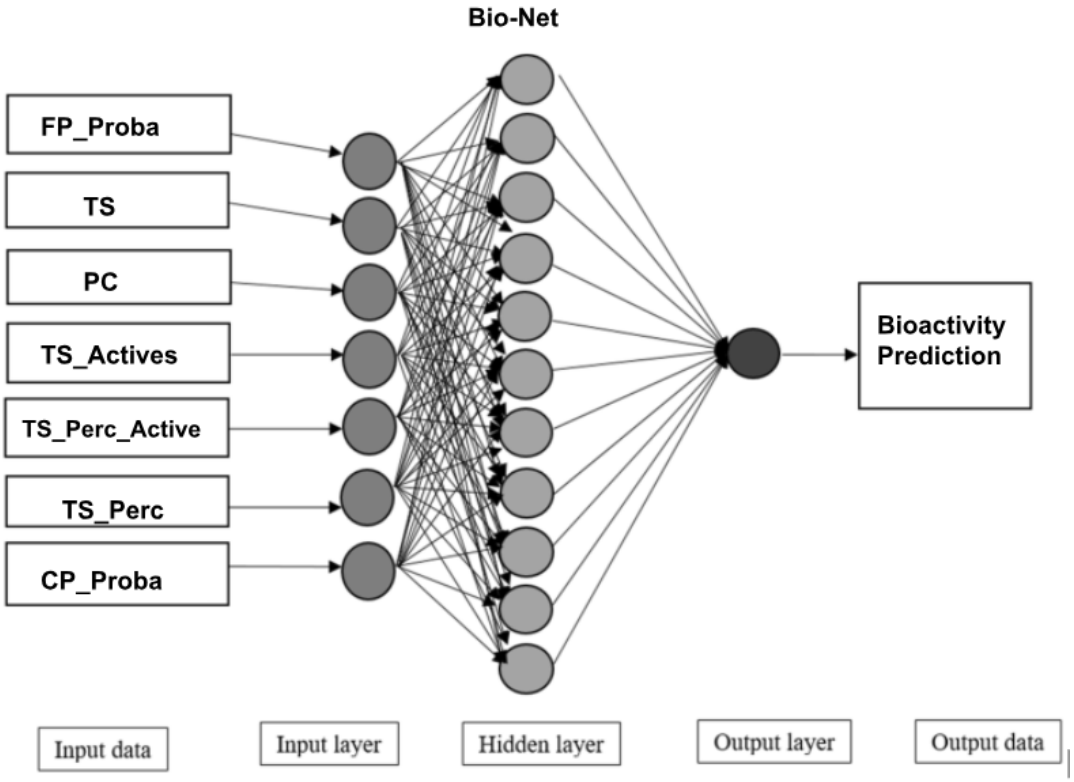

Figure 3.2: BioNet

- Allocates focus to the most pertinent features by calculating attention scores, enabling the network to prioritize information crucial for accurate bioactivity prediction.

3. Hidden Layers

   - Comprises multiple dense layers with ReLU activation, processing the attention-refined inputs with regularization strategies like dropout in place.

4. Output Layer

   - Finalizes the prediction through a sigmoid activation neuron, indicating the bioactivity status.

3.3.4 Algorithm

1. Data Preparation

   - Standardization and preprocessing are applied to Morgan Fingerprints and Cell Painting features for neural network compatibility.
   - Additional fields, including similarity scores, are computed and normalized for subsequent integration.

2. Model Training

    - Independently train ChemNet and CellNet with designated datasets, fine-tuning parameters for optimal performance.
    - Utilize these networks to generate FP-Prob and CP-Prob for further processing.

3. BioNet Training

    - Employ the comprehensive dataset, including FP-Prob, CP-Prob, and additional inputs, to train BioNet, aiming for precise bioactivity predictions.

4. Evaluation

    - System performance is evaluated using metrics such as AUC-ROC, precision, recall, and F1-score, ensuring robustness through cross-validation techniques.

5. Optimization

    - Based on performance evaluations, the system undergoes iterative refinements to enhance predictive accuracy and reliability, including adjustments to neural network architectures.

3.4 Implementation

3.4.1 Dataset Description

The exploration of three pivotal domains in chemical and biological research—Standardized International Chemical Identifier (InChI) codes, Simplified Molecular-Input Line-Entry System (SMILES) representations, and Cell Morphology Descriptors—unveils their profound significance and versatile applications. These domains serve as pillars, bolstering the consistency, interoperability, and analytical depth essential for managing and dissecting chemical and biological data effectively.

InChI codes stand as beacons of identification, providing a standardized means to unequivocally label chemical compounds. Their adoption fosters uniform recognition across diverse databases, facilitating seamless data exchange and integration. This standardization not only ensures the reproducibility of findings but also underpins computational endeavors such as virtual screening and Quantitative Structure-Activity Relationship (QSAR) studies, thereby propelling scientific inquiry forward.

Complementing the InChI framework, SMILES notation offers a succinct, human-readable representation of molecular structures. Its accessibility promotes ease of use and interoperability across an array of computational chemistry tools, enriching the analytical toolkit available to researchers. By accommodating diverse molecular features

and applications, SMILES augments the versatility and adaptability of chemical data analysis, further amplifying research capabilities.

Meanwhile, Cell Morphology Descriptors delve into the intricate changes observed in cellular structures in response to chemical perturbations. By capturing these nuanced variations, they furnish valuable insights into the cellular effects of compounds, shedding light on fundamental biological processes and mechanisms. Their integration into research endeavors fosters a holistic understanding of chemical-biological interactions, laying the groundwork for targeted interventions and therapeutic advancements.

Collectively, these domains converge to form a robust framework for chemical and biological research, characterized by enhanced data management, analytical prowess, and collaborative potential. Their synergy drives innovation and discovery, propelling the boundaries of scientific knowledge ever outward and empowering researchers to confront complex challenges with clarity and confidence.

3.4.2 Public Assay Filtered Dataset

Introduction: The dataset comprises InChI (International Chemical Identifier) standardized codes associated with 88 distinct assays, offering a standardized means of labeling chemical compounds. These codes facilitate interoperability and exchange of chemical information across platforms. This resource finds application in chemistry, biochemistry, pharmacology, and related fields. Each InChI code corresponds to a specific assay, representing a unique chemical compound evaluated for its biological activity, efficacy, or toxicity.

Dataset Utilization:

1. **Chemical Informatics:** Perform structure-activity relationship (SAR) analysis, virtual screening, and compound prioritization based on assay data.

2. **Drug Discovery:** Identify lead compounds, optimize drug candidates, and predict biological activity profiles for novel molecules.

3.4.3 Cell Painting Filtered Dataset

Introduction: This dataset merges InChI standardized codes with Cell Painting features extracted from biological assays. It encapsulates a diverse array of chemical compounds represented by their molecular structures (InChI codes) and cellular responses captured through Cell Painting assays. This integration offers insights into the interactions between chemical compounds and biological systems, with applications in drug discovery, toxicology studies, and chemical informatics.

Dataset Overview: The dataset comprises InChI standardized codes and Cell Painting features. InChI codes uniquely identify chemical compounds, while Cell Painting fea-

	InChICode_standardised	588458	588334	2642	2156	2330	2216	743015	504444	894	...	720579	720533	720542	smiles_r
0	InChI=1S/C14H13N5O5S2/c1-2-5-3-25-12-8(11(21)1...	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	C=CC1=C(C(=O)O)N2C(=O)[C@@H](N=C(O)/C(=N\O)c3c...
1	InChI=1S/C22H21NO2S/c23-20(21(24)25)16-26-22(1...	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NC(CSC(c1ccccc1)(c1ccccc1)c1ccccc1)C(=O)O
2	InChI=1S/C15H8O7/c16-6-3-8-12(10(18)4-6)14(20)...	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	...	NaN	NaN	NaN	O=C(O)c1cc(O)c2c(c1)C(=O)c1cc(O)cc(O)c1C2=O
3	InChI=1S/C8H10O4/c1-5(2)8(10)6(11-3)4-7(9)12-8...	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0	...	NaN	NaN	NaN	C=C(C)C1(O)OC(=O)C=C1OC
4	InChI=1S/C14H14O4S2/c15-5-7-19-13-11(17)9-3-1-...	1.0	NaN	NaN	NaN	1.0	NaN	NaN	1.0	1.0	...	NaN	NaN	NaN	O=C1C(SCCO)=C(SCCO)C(=O)c2ccccc21
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9871	InChI=1S/C11H8O3/c1-7(12)9-6-8-4-2-3-5-10(8)14...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	CC(=O)c1cc2ccccc2oc1=O
9872	InChI=1S/C15H18N2/c1-2-7-14-12(5-1)13-6-3-4-11...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	c1cc2c3c(c1)c1c(n3CCNC2)CCCC1
9873	InChI=1S/C33H44N4O4/c1-22-18-37(23(2)20-38)32(...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	C[C@H]1CN([C@H](C)CO)C(=O)c2c(c3ccccc3n2C)-c2c...
9874	InChI=1S/C18H16NO/c1-14-11-12-19(17-10-6-5-9-1...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	Cc1cc[n+](CC(=O)c2ccccc2)c2ccccc12
9875	InChI=1S/C13H11Cl2N3O/c1-8-5-6-9(7-16-8)17-13(...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	Cc1ccc(N=C(O)Nc2c(Cl)cccc2Cl)cn1

Figure 3.3: Public Assay Filtered Dataset

	InChICode_standardised	Nuclei_Texture_Variance_DNA_5_0	Nuclei_Texture_Variance_ER_10_0	Nuclei_Texture_Variance_ER_3_0	Nuclei_Texture_Variance_ER_5_
0	InChI=1S/C14H13N5O5S2/c1-2-5-3-25-12-8(11(21)1...	-0.126439	-0.158874	-0.111242	-0.1116
1	InChI=1S/C22H21NO2S/c23-20(21(24)25)16-26-22(1...	-0.143813	0.057710	0.098878	0.0421
2	InChI=1S/C15H8O7/c16-6-3-8-12(10(18)4-6)14(20)...	0.029446	0.055282	0.089878	0.0841
3	InChI=1S/C8H10O4/c1-5(2)8(10)6(11-3)4-7(9)12-8...	-0.356086	-0.123535	-0.165911	-0.1410
4	InChI=1S/C14H14O4S2/c15-5-7-19-13-11(17)9-3-1-...	0.431928	0.727071	1.850584	1.9420
...	...	...	...	...	
9871	InChI=1S/C11H8O3/c1-7(12)9-6-8-4-2-3-5-10(8)14...	-0.086964	-0.050364	-0.058893	-0.0311
9872	InChI=1S/C15H18N2/c1-2-7-14-12(5-1)13-6-3-4-11...	-0.173884	-0.053152	-0.028996	-0.0390
9873	InChI=1S/C33H44N4O4/c1-22-18-37(23(2)20-38)32(...	0.457065	0.102048	0.166616	0.1322
9874	InChI=1S/C18H16NO/c1-14-11-12-19(17-10-6-5-9-1...	-0.098800	-0.080274	-0.114318	-0.1274
9875	InChI=1S/C13H11Cl2N3O/c1-8-5-6-9(7-16-8)17-13(...	-0.023186	-0.005595	-0.047118	-0.0337

Figure 3.4: Cell Painting Filtered Dataset

tures represent cellular morphology and response characteristics observed through high-content imaging assays.

Dataset Utilization:

1. **Chemical Biology:** Explore relationships between chemical structures and cellular phenotypes, elucidate structure-activity relationships (SAR), and uncover potential drug targets.

2. **Drug Discovery:** Prioritize lead compounds, optimize drug candidates, and screen for compounds with desirable biological profiles based on Cell Painting features.

3.4.4 Merged Dataset for Biological Prediction

Introduction: This merged dataset combines a public assay dataset with Cell Painting features and Molecular Fingerprints (MFP) features to facilitate biological predictions based on comprehensive molecular and cellular characteristics. By integrating diverse data modalities, it enables robust predictions of biological responses to chemical compounds, offering insights into drug discovery, toxicity assessment, and chemical biology.

Dataset Overview: The dataset integrates a public assay dataset containing InChI standardized codes and biological activity measurements with Cell Painting features extracted from high-content imaging assays. Molecular Fingerprints (MFP) features are also incorporated to capture molecular structural information.

Dataset Utilization:

1. **Biological Prediction:** Train predictive models using combined features to predict biological activity or toxicity profiles of chemical compounds.
2. **Feature Engineering and Selection:** Explore feature engineering techniques and identify relevant features for predicting biological activity, improving model interpretability and performance.
3. **Model Evaluation and Validation:** Perform rigorous evaluation and validation of predictive models using cross-validation techniques and performance metrics to assess robustness and generalization ability.

3.5 Software Requirements

- **Programming Language:** Python 3.7 or newer.
- **Development Environments:**
  - Google Colab for cloud-based Jupyter notebook environments with pre-configured Python environments. Includes free access to GPUs and TPUs.
  - Local IDEs like PyCharm, Visual Studio Code, or Jupyter Notebooks for development on local machines.
- **Primary Libraries:**
  - TensorFlow 2.4+ for model development, training, and inference.
  - Pandas & NumPy for data manipulation and numerical computations.
  - Scikit-learn for data preprocessing and machine learning tasks.
  - Matplotlib & Seaborn for data visualization.
- **Version Control:** Git, with GitHub for repository hosting.

3.6 Hardware Requirements

- **CPU:** Intel i5/i7/i9 or equivalent AMD processor with at least 4 cores.
- **Memory (RAM):** Minimum 16 GB, 32 GB or more recommended.
- **Storage:** SSD with at least 256 GB (1 TB recommended).
- **GPU (Optional but Recommended):** NVIDIA GPU with CUDA and CuDNN support. Models like NVIDIA GTX 1060 or better are suitable.
- **Internet Connection:** Required for accessing Google Colab, dataset downloads, and online collaboration.

3.6.1 Utilizing Google Colab

- Google Colab provides various GPU options for free tiers, with better access in Pro or Pro+ tiers.
- The free tier offers around 12 GB of RAM and 100 GB of disk space, suitable for small to medium datasets.
- Use Google Drive for storing and accessing large datasets directly from Colab notebooks.

3.6.2 Additional Considerations

- **Data Storage and Backup:** External HDDs or cloud storage solutions for data backup. Google Drive integration with Colab for dataset access.
- **Power Supply & Cooling:** Adequate power supply and cooling system for desktop setups with powerful GPUs.

Bibliography

[1] **Predicting Compound Activity from Phenotypic Profiles and Chemical Structures** - Suman K. Chakravarti.

[2] **AtomNet: A Deep Convolutional Neural Network for Bioactivity in Structure based Drug Discovery** - Izhar Wallach, Michael Dzamba, Abraham Heifets.

[3] **De Novo Design and Bioactivity Prediction of SARS-CoV-2 Main Protease Inhibitors using Recurrent Neural Network-based Transfer Learning.** - Marcos V. S. Santana, Floriano P. Silva-Jr.

[4] **Learning Machine Reasoning for Bioactivity Prediction of Chemicals.** - Suman K. Chakravarti.

[5] **Bio-activity Prediction of Drug Candidate Compounds Targeting SARS-Cov-2 using Machine Learning Approaches.** - Faisal Bin Ashraf, Sanjida Akter, Sumona Hoque Mumu, Muhammad Usama Islam, Jasim Uddin.

[6] **Bioactivity Prediction Using Convolutional Neural Network.** - Hentabli Hamza, Maged Nasser, Naomie Salim, Faisal Saeed.

[7] **Merging bioactivity predictions from cell morphology and chemical fingerprint models using similarity to training data** - Srijit Seal1 , Hongbin Yang1 , Maria-Anna Trapotsi1 , Satvik Singh2 , Jordi Carreras-Puigvert3 , Ola Spjuth3 and Andreas Bender.

[8] **Imputation of Assay Bioactivity Data using Deep Learning.** - T.M. Whitehead, B.W.J. Irwin, P. Hunt, M.D. Segall, G.J. Conduit.