COMPOUND SIMILARITY PREDICTION

# A PROJECT REPORT

Submitted in partial fulfillment of the requirements for the award of the degree of

## Bachelor of Technology

*in*

COMPUTER SCIENCE AND ENGINEERING

### BY

**M.JYOTHSNA K. ASHOK KUMAR**

**Roll No: 18331A05A2 Roll No: 19335A0509**

**P. SAI JAHNAVI S.NIKHILA**

**Roll No: 18331A05B5 Roll No: 18331A05B3**

**Under the Supervision of**

**Dr. P. SRINIVASA RAO**

**Associate Professor**



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING MVGR COLLEGE OF ENGINEERING (Autonomous) VIZIANAGARAM-535005, AP (INDIA)**

**(Accredited by NBA, NAAC, and Permanently Affiliated to Jawaharlal Nehru Technological University Kakinada)**

# CERTIFICATE



This is to certify that the project report entitled **“Compound Similarity Prediction”** being submitted by **M.JYOTHSNA, K.ASHOK KUMAR, P. SAI JAHNAVI, S.NIKHILA** bearing registered numbers 18331A05A2, 19335A0509, 18331A05B5, 18331A05B3 respectively, in partial fulfillment for the award of the degree of “**Bachelor of Technology” in Computer Science and Engineering** is a record of Bonafide work done by them under my supervision during the academic year 2021-2022.

**HOD CSE Supervisor**

**Dr. P. Ravi Kiran Varma Dr. P. Srinivasa Rao**

**Associate Professor Associate Professor**

**Department of CSE Department of CSE**

**MVGR College of Engineering MVGR College of Engineering**

# ABSTRACT

The drug industry is one of the major players guiding the development of the medicines, biotechnology and pharmacology field. Drug discovery is the process by which drugs are discovered and designed. It is a process which aims at identifying a compound therapeutically useful in curing and treating disease. Drug discovery and development pipelines are long, complex and depend on numerous factors. Drug designers mine chemical data from large databases to extract chemical compound that becomes the lead compound in drug discovery. The drug target designing combine machine learning and deep learning algorithm that improve the quality of drugs discovered. Drug discovery involves seven step process that includes disease selection, target hypothesis, lead identification, lead optimization, pre- clinical trial, clinical trial, pharmacogenic identification. Machine learning can be applied to identify the drug targets and in optimization of lead compound. Both supervised and unsupervised algorithms when applied to the databases increases the efficacy of identifying a new target and optimize the lead compound. The process failure rate when machine learning is applied will be very low compared to the traditional drug discovery process. Here we present a Compound Similarity Prediction by using k-means clustering and using dbscan.

Through this compound similarity prediction, the compounds are clustered according to their similarity. So, if the reaction of the new compound is unknown, we could know its reaction through this compound similarity prediction. The main advantage of compound similarity prediction is to know the reaction of the new compound. If the reaction of the compound is known then we can speed up the process of drug discovery.

## PROJECT AREA:

## Machine Learning

## TOPIC SELECTED:

## Drug Discovery (Compound Similarity Prediction).

## INPUTS:

## MOLV3000 sdf file

## EXPECTED OUTPUT:

## We can analyse the similar compounds and can estimate the amout of dose required by patient.

## Hardware and Software requirements

**Hardware requirements :**

* Processor: Base Frequency – 1.1 GHz
* RAM: 4 GB
* Hard Disk: 1 TB

**Software requirements :**

* Operating system: Windows
* Language: JSP,python
* Tool: Anaconda Navigator, Jupyter Notebook,Gephi
* Database: mysql
* Browser: Chrome, Mozilla, Opera etc.
* Technologies: Image Processing, Convolutional Neural Networks, Deep Learning, Machine Learning.

# LITERATURE SURVEY

1. Hakime Ozt¨urk ¨ a , Arzucan Ozg¨ur ¨ a , Philippe Schwaller b , Teodoro Lainob, ∗ , Elif Ozkirimlic,d, “*Exploring Chemical Space using Natural Language Processing Methodologies for Drug Discovery.”,arXiv peprint ,May 2020.*

Hakime Ozturk *et al*[1], explained an approach on offensive content clustering in compound similarity in Drug Discovery using Machine learning. the Information about the design and discovery of drugs with the column called SMILES. SMILES(Simplified Molecular Input Entry Specification) consists of all the chemical formula of each and every compound. In this agglomative clustering, is used to combine the similar compounds. With this review, they have summarized the impact of NLP on bio/cheminformatics to encourage this already interdisciplinary field to take advantage of recent advances. Future Perspectives are to increase in the biochemical data available in public databases combined

with the advances in computational power and NLP methodologies have given rise to a rapid growth in the publication rate in bio/cheminformatics, especially through pre-print servers. As this interdisciplinary field grows, novel opportunities come hand in hand with novel challenges. The challenges are Bench marking, Reproducibility, Bias in data.

[2] Hakime Ozt¨urk ¨ a , Arzucan Ozg¨ur ¨ a , Philippe Schwaller b , Teodoro Lainob, ∗ , Elif Ozkirimlic,d, “*Exploring Chemical Space using Natural Language Processing Methodologies for Drug Discovery.”,arXiv peprint ,May 2020.*

Hakime Ozturk *et al*[2], Explained the study of SMILES-based Compund similarity functions

for drug-target interaction prediction. Molecular structures can be represented as strings of

special characters using SMILES. Since each molecule is represented as a string, the similarity

between compounds can be computed using SMILES-based string similarity functions. Most

previous studies on drug-target interaction prediction use 2D-based compound similarity kernels

such as SIMCOMP. In this study, Hakime Ozturk *et al*[2] adapt and evaluate various SMILES-

based similarity methods for drug-target interaction prediction. In addition, inspired by the

vector space model of Information Retrieval we propose cosine similarity based SMILES

kernels that make use of the Term Frequency (TF) and Term Frequency-Inverse Document

Frequency (TF-IDF) weighting approaches. We also investigate generating composite kernels

by combining our best SMILES-based similarity functions with the SIMCOMP kernel. With

this study, Hakime Ozturk *et al*[2] provided a comparison of 13 different ligand similarity

functions, each of which utilizes the SMILES string of molecule representation. Hakime Ozturk

*et al*[2] conclude the more efficient SMILES-based similarity functions performed similarly to

the more complex 2D-based SIMCOMP kernel in terms of AUC-ROC scores. The TF-IDF

based cosine similarity obtained a better AUC-PR score than the SIMCOMP kernel on the

GPCR benchmark data set. The composite kernel of TF-IDF based cosine similarity and

SIMCOMP achieved the best AUC-PR scores for all data sets.

Martin Vogt *et al*[3] Described Python package for modelling Tanimoto similarity value

distributions”. It described the information about how to use the ccbmlib Python package which

is a collection of modules for modelling similarity value distributions. From this paper i invoked

that this is used to assess the statistical significance of tanimoto coefficients and evaluate how

molecular similarity is reflected when different fingerprint representations are used and

conditional significance score to estimate where a test compound would be ranked in a search

and the resulting models have been evaluated for RDKit fingerprints, taking a collection of

ChEMBL compounds as a reference data set. I conclude that the Highly accurate models were

obtained, with difference of 1% or less for indicating high similarity. Martin Vogt *et al*[3]

concluded that the tools provided make it possible to evaluate the significance of Tc values for a variety

of fingerprints from RDKit. Users can generate distribution models for different fingerprints with respect

to reference data sets. Accurate models are obtained for most RDKIT fingerprints including the popular

MACCS and Morgan fingerprints. Based on these models, it can be assessed to what extent molecular

similarity is accounted for by fingerprints of different design and to what extent similarity between

compounds sharing the same activity is reflected by similarity scores calculated on the basis of different

fingerprint representations. Furthermore, the conditional models can be used to predict the suitability of

fingerprints for similarity searching and ligand-based virtual screening.

[4] B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen and J. Tang

*“ Comparative analysis of molecular fingerprints in prediction of drug combination effects.“,brefings in bioinformatics,22(6),2021*

B. Zagidullin *et al*[4] describes the Application of machine and deep learning methods

in drug discovery and cancer research has gained a considerable amount of attention in the past

years. As the field grows, it becomes crucial to systematically evaluate the performance of novel

computational solutions in relation to established techniques. To this end, we compare rule-

based and data-driven molecular representations in prediction of drug combination sensitivity

and drug synergy scores using standardized results of 14 high-throughput screening studies,

comprising 64 200 unique combinations of 4153 molecules tested in 112 cancer cell lines. We

evaluate the clustering performance of molecular representations and quantify their similarity by

adapting the Centered Kernel Alignment metric. Our work demonstrates that to identify an

optimal molecular representation type, it is necessary to supplement quantitative benchmark

results with qualitative considerations, such as model interpretability and robustness, which may

vary between and throughout preclinical drug development projects. B. Zagidullin *et al*[4]

compare rule-based and data-driven molecular representations in prediction of

drug combination sensitivity and finding drug synergy scores using standardized results and

later on to evaluate the clustering performance of molecular representations and quantify

their similarity by adapting the Centred Kernel Alignment metric (identify an optimal molecular

representation type (CKA)). The included Data-driven fingerprints, namely VAE 256 bits long

trained on SMILES and Infomax 300 bits long-trained molecular graphs are well-suited for

regression tasks. 1024 bits long 2D and 3D circular fingerprints are flexible and well-performant

rule-based models fit for clustering tasks.

***[5]* .** Daniel Probst and Jean‑Louis Reymond

**“***One molecular fngerprint to rule them all: drugs, biomolecules, and the metabolome Alice Capecchi.”, jouneral of cheminformatics,2020*

Daniel Probst *et al*[5] explained that MAP4 is a new molecular fingerprint suitable for drugs

and can be adopted as a universal fingerprint to describe and search chemical space so here it

takes MinHasing. MinHasing is used for MAP4 allows the construction of k-NN trees and the

creation of high-resolution chemical space tree-maps (TMAPs) for databased as diverse as

DrugBank.

Daniel Probst *et al*[5] set out to design a new fingerprint suitable for both small and large

molecules by combining substructure and atom-pair concepts. Our quest resulted in a new

fingerprint called Min Hashed atom-pair finger print up to a diameter of four bonds (MAP4). In

this finger print the circular substructures with radii of r=1 and r=2 bonds around each atom in

an atom-pair are written as two pairs of SMILES, each pair being combined with the topological

distance separating the two central atoms. These so-called atom-pair molecular shingles are

hashed, and the resulting set of hashes is Min Hashed to form the MAP4 finger print. MAP4

significantly out performs all other finger prints on an extended benchmark that combines the

Riniker and Landrum small molecule benchmark with a peptide benchmark recovering BLAST

analogs from either scrambled or point mutation analogs. MAP4 furthermore produces well-

organized chemical space tree-maps (TMAPs) for databases as diverse as Drug Bank,

ChEMBL, Swiss Prot and the Human Metabolome Database (HMBD), and differentiates

between all metabolites in HMBD, over 70% of which are indistinguishable from their nearest

neighbor using substructure finger prints.

**[6].** Ravi Manne, ” *Machine Learning Techniques in Drug Discovery and Development “,all research, 2021.*

Ravi Manne *et al*[6] examined machine learning and deep learning techniques which help

pharma industry in all stages of drug discovery which includes target validation, prognostic

bookmarks, technical traits, in which these are totally seven phases in the process of drug

discovery.

**[7].** Suresh Dara1  · Swetha Dhamercherla1  · Surender Singh Jadav2  · CH Madhu Babu1  · Mohamed Jawed Ahsan3 , “*Machine Learning in Drug Discovery: A Review”, springer Nature B.V. 2021*

Suresh Dara *et al*[7] Described the feasible literature on drug discovery through ML tools and

techniques that are enforced in every phase of drug development to accelerate the research

process. Target validation, prognostic biomarkers, digital pathology are considered as the problem

statements where In clinical trials, absolute and methodological data must be generated to tackle

many puzzles in validating ML techniques, improving decision-making, promoting awareness in

ML approaches, and deducing risk failures in drug discovery.

***[8].*** *Geeks for geeks and tutorial point, Anaconda navigator installation jupyter notebook environment setup*

I used “Anaconda and Jupyter Note Setup” to install anaconda for my execution where,the

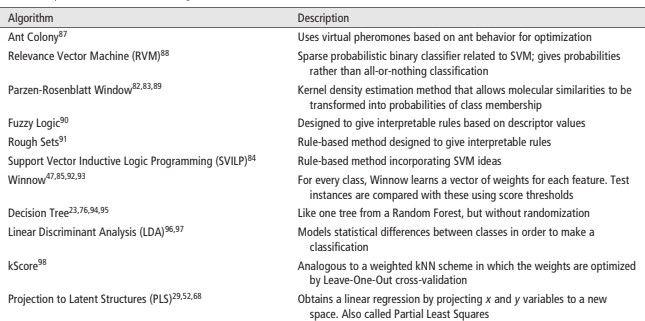
Anaconda distribution includes the Conda package manager in addition to the preconfigured

Python packages and other tools. Anaconda Navigator is a GUI tool that is included in the

Anaconda distribution and makes it easy to configure, install, and launch tools such as Jupyter

Notebook.

**ALGORITHMS USED:**

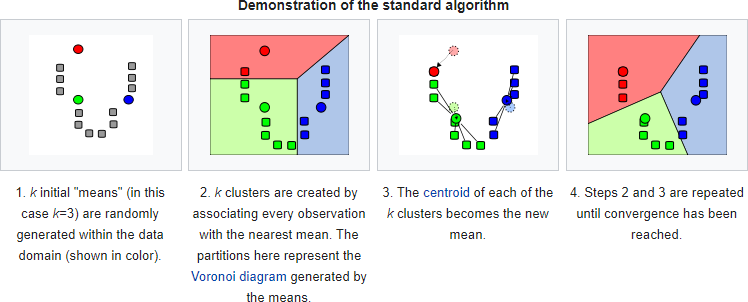


**figure 2.1 Algorithms**

## Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, such as finding the probability density function. Though unsupervised learning encompasses other domains involving summarizing and explaining data features.

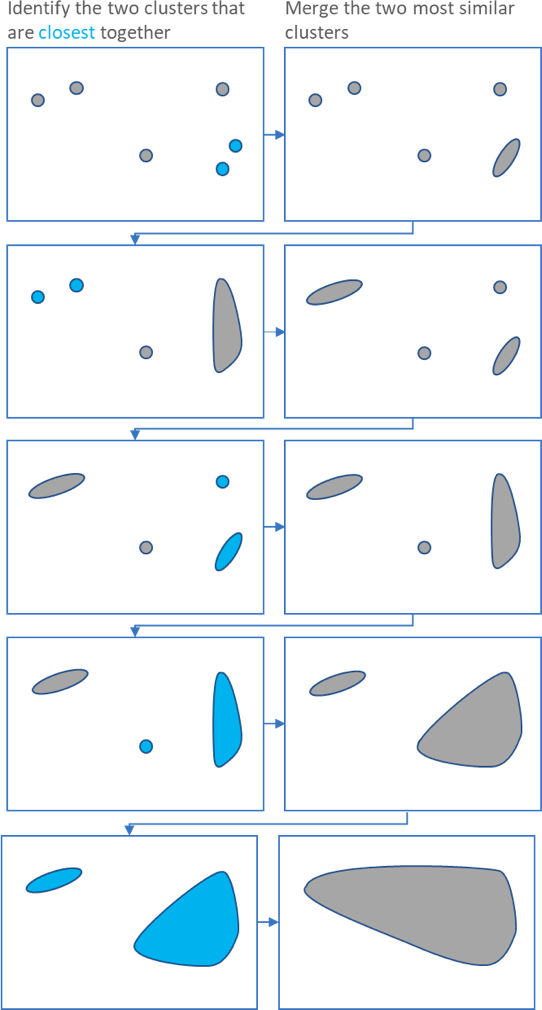
1. **k-means clustering**-A classification method that classifies data into k groups by minimizing within-group distances to the centroid. ***k*-means clustering** is a method of vector quantization, originally from signal processing, that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. *k*- means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficult Weber problem: the mean optimizes squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, better Euclidean solutions can be found using k-medians and k-medoids.



**Figure 3.8 k-means clustering**

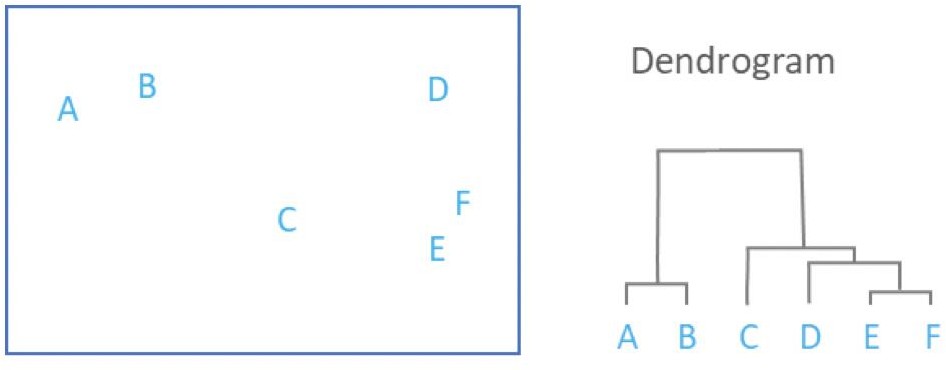
1. **Hierarchical clustering**-A classification method that builds a hierarchy of clusters by agglomerative clustering e.g., merging smaller clusters or divisive clustering e.g., splitting a large cluster to smaller ones. ***Hierarchical clustering****,* also known as *hierarchical cluster analysis,* is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters*,* where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together. This is illustrated in the diagrams below



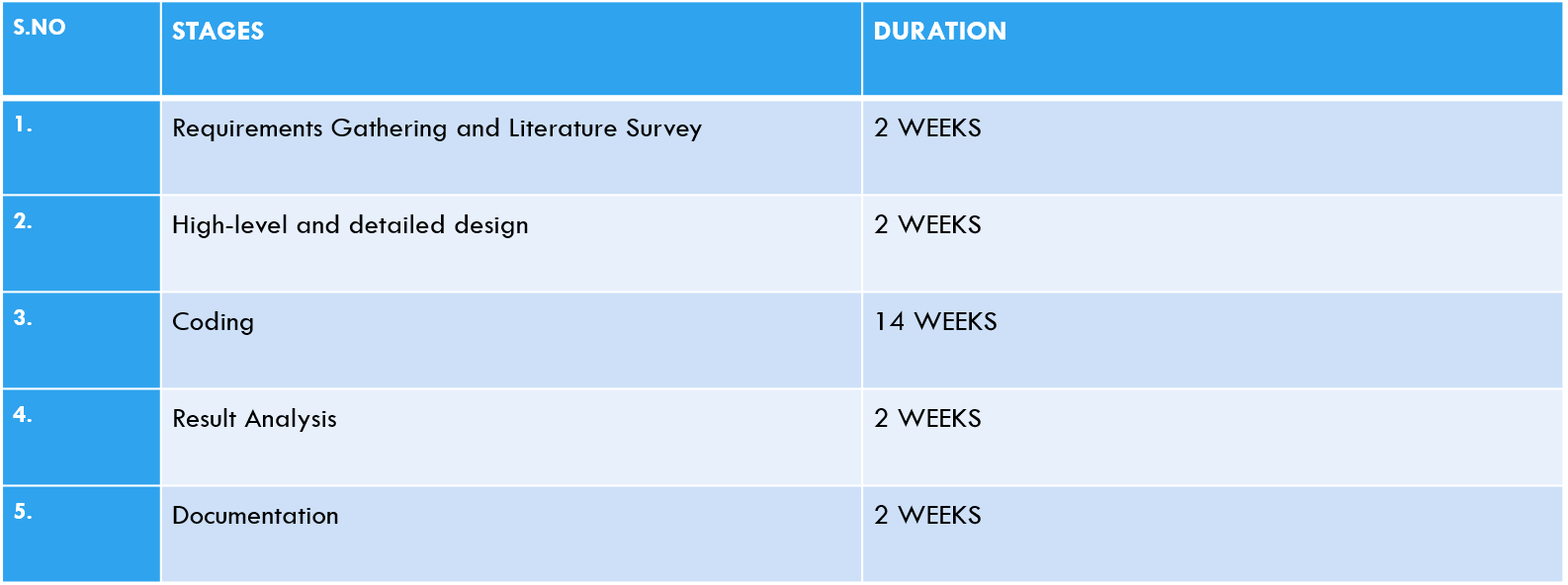
**Figure 3.9 Hierarchical clustering**

The main output of Hierarchical Clustering is a *dendrogram,* which shows the hierarchical relationship between the clusters:



**Figure 3.10 Dendrogram**

**DURATION:**



**REFERENCES:**

**[1].** Hakime Ozt¨urk ¨ a , Arzucan Ozg¨ur ¨ a , Philippe Schwaller b , Teodoro Lainob, ∗ , Elif Ozkirimlic,d, “*Exploring Chemical Space using Natural Language Processing Methodologies for Drug Discovery.”,arXiv peprint 2020.*

**[2].** Hakime Öztürk\* , Elif Ozkirimli\* and Arzucan Özgür\*

“ *A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction.” arXiv peprint 2020.*

**[3].** Martin Vogt , Jürgen Bajorath Department of Life Science Informatics, B-IT, University of Bonn, Endenicher Allee

*ccbmlib – ” a Python package for modeling Tanimoto similarity value distributions [version 1; peer review: 2 approved] “ ,* 19c, Bonn, NRW, 53115, Germany,

**[4].** B. Zagidullin, Z. Wang, Y. Guan, E. Pitkänen and J. Tang

*“ Comparative analysis of molecular fingerprints in prediction of drug combination effects.“,brefings in bioinformatics,22(6),2021*

**[5].** Daniel Probst and Jean‑Louis Reymond

**“***One molecular fngerprint to rule them all: drugs, biomolecules, and the metabolome Alice Capecchi.”, jouneral of cheminformatics,2020*

**[6].** Ravi Manne, ” *Machine Learning Techniques in Drug Discovery and Development “,all research, 2021*

**[7].** Suresh Dara1  · Swetha Dhamercherla1  · Surender Singh Jadav2  · CH Madhu Babu1  · Mohamed Jawed Ahsan3 , “*Machine Learning in Drug Discovery: A Review”, springer Nature B.V. 2021*

***[8].*** *Geeks for geeks and tutorial point, Anaconda navigator installation jupyter notebook environment setup*

***[9].*** John B. O. Mitchell∗ “*Machine learning methods in chemoinformatics”, WIREs Comput Mol Sci 2014*

***[10].*** N. Priya1 and G. Shobana2 “*Application of Machine Learning Models in Drug Discovery: A Review”, Research Trend , 2019*

***[11]. Dataset link: https://drugcentral.org/download***

**CONCLUSION:**

Drug discovery is a long process and takes years to discover a new drug. So in order to reduce time we need machine learning algorithms to speed up the process of drug discovery. We also find many drugs but we don’t know the reaction of each drug. If we know the reaction of drug then we could save some time in discovering new medicine.so we have done compound similarity prediction and until now we have read many papers regarding classification, supervised and unsupervised clustering techniques and many algorithms. We used k-means clustering for compound similarity prediction and also, we used density-based clustering algorithm. Finally, we clustered the compounds based on their similarity. We have successfully implemented compound similarity prediction and finally we can know the reaction of the new drug.

**Project Team:**

M. JYOTHSNA

K. ASHOK KUMAR

P. SAI JAHNAVI

S. NIKHILA

**Supervisor:**

**Dr P. SRINIVASA RAO**