Graph Topological Indices and Machine Learning for Protein Networks and Drug Property Prediction

MS Thesis
by

Ahmad Mehmood

CIIT/SP24-RMT-025/LHR

COMSATS University Islamabad
Pakistan

Fall 2025

Graph Topological Indices and Machine Learning for Protein Networks and Drug Property Prediction

A thesis submitted to
COMSATS University Islamabad

In partial fulfillment
of the requirement for the degree of

Master of Science
in
Mathematics

by

Ahmad Mehmood

CIIT/SP24-RMT-025/LHR

Department of Mathematics
Faculty of Science

COMSATS University Islamabad
Pakistan

Fall 2025

Graph Topological Indices and Machine Learning for Protein Networks and Drug Property Prediction

This thesis is submitted to the department of Mathematics in partial fulfillment of the requirement for the award of degree of Master of Science in Mathematics

Name	Registration Number
Ahmad Mehmood	CIIT/SP24-RMT-025/LHR

Supervisor

Dr. Sana Javed

Associate Professor
Department of Mathematics

COMSATS University Islamabad (CUI), Lahore Campus

Certificate of Approval

This thesis titled

Graph Topological Indices and Machine Learning for Protein Networks and Drug Property Prediction

By

Ahmad Mehmood

CIIT/SP24-RMT-025/LHR

Has been approved

For COMSATS University Islamabad, Lahore Campus.

External Examiner:_____

Dr. External Examinor

University Name

Supervisor:_____

Dr. Sana Javed

Department of Mathematics, (CUI) Lahore Campus

Head of Department:_____

Prof. Dr. Muhammad Hussain

Department of Mathematics, (CUI) Lahore Campus

Author's Declaration

I Ahmad Mehmood, SP24-RMT-025/LHR, hereby declare that I have produced the work presented in this thesis, during the scheduled period of study. I also declare that I have not taken any material from any source except referred to wherever due to that amount of plagiarism is within an acceptable range. If a violation of HEC rules on research has occurred in this thesis, I shall be liable to punishable action under the plagiarism rules of HEC.

Date:_____ _____

Ahmad Mehmood

CIIT/SP24-RMT-025/LHR

Certificate

It is certified that Ahmad Mehmood, CIIT/SP24-RMT-025/LHR has carried out all the work related to this thesis under my supervision at the Department of Mathematics, COMSATS University Islamabad, Lahore Campus and the work fulfills the requirement for award of MS degree.

Date:_____

Supervisor

_____

Dr. Sana Javed
Associate Professor,
Mathematics,
COMSATS University Islamabad
Lahore Campus

Dedication

To My Parents, Teachers and Friends who always encouraged me to work hard and guided me towards the right way and destination

Acknowledgements

Praise to be ALLAH, the Cherisher and Lord
of the World, Most gracious and Most Merciful

First and foremost, I would like to thank ALLAH Almighty (the most beneficent and most merciful) for giving me the strength, knowledge, ability and opportunity to undertake this research study and to preserve and complete it satisfactorily. Without countless blessing of ALLAH Almighty, this achievement would not have been possible. May His peace and blessings be upon His messenger Hazrat Muhammad (PBUH), upon his family, companions and whoever follows him. My insightful gratitude to Hazrat Muhammad (PBUH) Who is forever a track of guidance and knowledge for humanity as a whole. In my journey towards this degree, I have found a teacher, an inspiration, a role model and a pillar of support in my life, my kind.

Ahmad Mehmood
CIIT/SP24-RMT-025/LHR

Abstract

Graph Topological Indices and Machine Learning for Protein Networks and Drug Property Prediction

By

Ahmad Mehmood

Molecular property prediction for drugs is an integral part of computational chemistry in the early stages of drug discovery for efficient screening and optimization before experimental verification. In the present work, an exhaustive and interpretable machine learning (ML) strategy is proposed that combines SMILES-based molecular feature extraction with topological indices using molecular graph theoretical approaches for the prediction of key physicochemical properties of drugs. The dataset consisting of forty-seven pharmacologically active molecules from DrugBank, with their structures characterized using sixteen topological descriptors such as the Wiener, Balaban, Harary, Randić, Zagreb, Schultz, and Shannon Entropy indexes, were identified using the RDKit grouping library coupled with mathematical formulations from molecular graph theory. Preprocessing steps were performed rigorously for missing values using imputation; removal with cutoffs using the inter-quartile range method; variance for stabilization using the Box-Cox transformation; and normalization with MinMax scaling. Four different regression models, namely Ordinary Least Squares (OLS), Ridge Regression, LASSO, and Elastic Net algorithms, were employed for the prediction of the MW, LogP, and HBD properties for the dataset with evaluations made in terms of R^2, MAE, and Root MSE metrics; with feature interpretation from SHapley Additive exPlanations (SHAP) analyses for feature interpretation. The results achieved for MW with high predictive precision with $R^2 \approx 0.96$ using linear models effectively; moderate accuracy with $R^2 \approx 0.4$ for Lipophilicity; with poorer accuracy for HBD values likely due to their dependence on chemical principles underlying chemical groupings in chemistry. The interpretation using SHAP explained the contributions made by the Shannon entropy, Wiener, Zagreb, Schultz indices being the principal predic-

tors in each case. The present work clearly illustrates the utility of integrating molecular chem-graph principles with interpretable ML algorithms for efficient, scalable, and insightful property-based predictions in early stage computational screens for drug identification strategies.

Protein-protein interactions (PPIs) networks are very important to decode cellular signaling and mechanisms of disease. PPI also play a key to spot possible drug targets. A strong computational strategy is presented in this paper. It unites the graph theoretic approach with unsupervised machine learning. This study is used to detect biologically significant hub genes in a large scale lung cancer associated PPI network. The network retrieved from the STRING database (461 proteins, 18,704 weighted interactions). Using NetworkX, we constructed an undirected graph by extracting source and target node from the dataset. We computed 17 comprehensive topological centrality measures, confirming the networks scale free and small world properties (average clustering coefficient $= 0.67$, average shortest path length ≈ 1.96). The Isolation Forest algorithm identified 61 topological outliers. These outliers were related to nodes which had a very high influence. The K-means clustering (that was optimized by the elbow curve) sorted the rest of the nodes into three groups. The hub genes were given ranks according to degree centrality and a composite centrality score. Among the genes, CALM3, CREB1, AKT1, MAPK1, EGFR and KRAS came out as the strongest candidates. The pathway enrichment analysis performed with KEGG and Reactome showed that the oncogenic pathways were remarkably over represented. These pathways were PI3KAkt, Ras, MAPK, and proteoglycans in cancer and EGFR signaling. Our methodology, which combines together the multiple centrality metrics, anomaly detection, and clustering, has indeed solved the drawbacks of traditional single metric approaches. It has also provided greater sensitivity in hub detection. The hubs that were identified not only correspond to the well established lung cancer drivers but also point to potential novel biomarkers. The detected targets are thus a scalable and reproducible pipeline for systems level analysis of disease specific interactomes.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

The following chapter discusses a number of key terms, concepts and ideas related to the thesis.

1.1 Graph Theory

In mathematics the study of graphs, which are mathematical structures used to represent pairwise relationships among subjects is known as Graph theory. In this context, a graph is consist of set of vertices which are also known as nodes, and set of edges, which are also called links. The number of vertices ($|V|$) is known as the *order* of the graph, while the number of edges ($|E|$) is referred to as the *size* of graph. Relationships among vertices can be encoded using an adjacency matrix A, where $A_{ij} = 1$ if vertices v_i and v_j are adjacent; alternatively, an *incidence matrix* maps edges to pairs of vertices. The *degree* $d(v)$ of a vertex v represents the number of edges incident to it. Graphs may be classified as *directed*, where edges have orientation, or *simple*, which are undirected and contain no loops or multiple edges. Two graphs are said to be *isomorphic* if they share the same structure of connectivity between vertices, regardless of labeling.

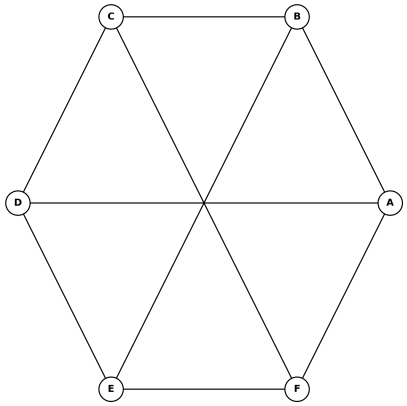


Figure 1.1: Graph

In Figure 1.1 V = {A, B, C, D, E, F} represent the set of vertices of graph G while E = {(A,B), (A,F), (A,D), (B,C), (B,E), (C,D), (C,F), (D,E), (E,F)} is the set of edges.

$$\forall v \in \{A,B,C,D,E,F\}, \quad \deg(v) = 3.$$

1.2 Protein-Protein Interaction and Gene Regulatory Networks

Protein protein interactions (PPIs) refers to binding between two or more protein molecules, often as a biological process. PPIs govern cellular processes like signaling and immune response. Dysregulated PPIs are implicated in diseases such as cancer and neurodegeneration. Traditional experimental methods (e.g., yeast two-hybrid) are labor-intensive and noisy. Computational tools combining graph theory, topological indices, and ML address these gaps comprehensively.

A Gene Regulatory Network (GRN) is like a control system in a cell that decide which genes are turned on or off and how much they contribute to the functionality of a biological process. It includes genes and proteins working together to manage things like growth and development.

1.3 Chemical Compounds

Chemical compounds are the substances composed of atoms and molecules from different elements which are bounded together in specific arrangements. The universe consists of matter that is entirely composed of atoms from over 100 different chemical elements. Sodium Chloride is one of the chemical compounds formed by the combination of sodium (Na) and chlorine (Cl). Drugs, also called medicines, can be defined as any substance that affect the functioning of living things.

1.4 Chemical Graph Theory

Chemical graphs serve as representations that indicate the connections of atoms or molecules in various systems of chemistry, such as molecular, compound and crystal systems. The elements of these graphs are denoted as nodes, which may consist of atoms, molecules or even abstraction like electron clouds or groups of atoms. The relationship between these nodes is represented by lines. Different representations of connections might be used such as bonds between atoms, interactions between molecules or even the movement of atoms during reactions. Chemical graphs facilitate the insight of chemists into the structures of

and the behaviors of chemical substances. for example, these are essential aspects in new drug production or the introduction of superior materials.

1.5 Topological Indices

Topological indices are numerical descriptors that quantify structural properties of graphs [1]. Topological indices might be degree based or distance based. They are pivotal in analyzing biological networks like PPIs and gene regulatory networks (GRNs). For different physicochemical characteristics of different drugs and chemical compounds, it is crucial to look at these descriptors.[2, 3, 4, 5, 6, 7]. Key categories include:

- **Degree-Based Indices**:

$$\begin{aligned}
\text{Randic Index:} \quad & R(G) = \sum_{uv \in E} \frac{1}{\sqrt{d(u) \cdot d(v)}} \\
\text{Zagreb First Index:} \quad & M_1(G) = \sum_{v \in V} d(v)^2 \\
\text{Zagreb Second Index:} \quad & M_2(G) = \sum_{uv \in E} d(u) \cdot d(v) \\
\text{Betweenness Centrality:} \quad & B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} (Identifies\ bottleneck\ nodes)
\end{aligned}$$

- **Distance-Based Indices**:

$$\begin{aligned}
\text{Wiener Index: } W(G) &= \sum_{u<v} d(u,v) \\
\text{Hosoya Index: } H(G) &= \sum_{k=0}^{\lfloor n/2 \rfloor} p(G,k)
\end{aligned}$$

- **Clustering Coefficient**: Measures local connectivity density:

$$C(v) = \frac{2 \cdot \text{triangles}(v)}{d(v)(d(v)-1)}$$

These indices enable quantitative comparisons of network robustness, connectivity, and functional modularity in PPIs.

1.6 Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) and Computer sciences, which enable computer to learn from the data using statistical algorithms without explicitly programming. These algorithms learn from data. They find patterns and correlations. Then they apply what they learned and they make predictions. Figure 1.2 There are mainly three types of ML. The type of algorithm used depend on the nature of data.

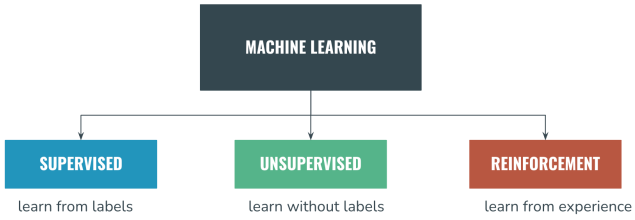


Figure 1.2: Types of ML

1.6.1 Supervised ML

Supervised ML is a type where algorithms are trained on a label dataset. The main objective of supervised ML is to find a function between input columns (features) and output columns (target variables)[8]. Algorithms like support vector machine classifier / regressor (SVM/SVR) and neural networks classify interactions or predict missing edges using labeled datasets. There are two types of procedures which are used in supervised ML

1.6.2 Regression

Regression analysis involves many ML techniques that enable the prediction of a continuous (y) variable by evaluating the values of single or multiple predictor variables [8]. Regression analysis can help identify relationships between independent and dependent variables [9].

1. **Simple Linear Regression**

   Simple linear regression is a statistical approach for describing the connection be-

tween a single independent variable (predictor) and a dependent variable (outcome). It assumes a linear connection between the predictor variable (x) and the dependent variable (y). This is expressed as:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where β_0 and β_1 are parameters (the intercept and slope) to estimated for data to find the best fit line. ε represent the noise factors. The model is fit by reducing the sum of squared residuals (difference between observed and predicted values).

2. **Multiple Linear Regression**

   One statistical technique for modelling the connection between two or more independent features (predictors) and a dependent variable is multiple linear regression. It expands on the idea of simple linear regression to include situations with several factors. This is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon$$

   where $\{\beta_0, \beta_1, \ldots, \beta_n\}$ are the coefficients and x_i are the independent features. We can write it in vectorization form:

$$\bar{Y} = XW + \varepsilon$$

   where $\bar{Y}$ is the matrix of predicted values. X is design matrix in which first columns is Ones because of bias term and others are the independent features. W is the matrix of coefficients. W can be defined in vectorization form:

$$W = (X^{\mathsf{T}} X)^{-1} X^{\mathsf{T}} Y$$

3. **Ridge Regression**

   Ridge regression, which is one of regularization technique, also known as L_2. It

shrinks feature weights toward zero to reduce overfitting by adding a penalty term in loss function. Mathematical way of describing ridge in given in

$$\mathscr{L}(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

where λ is the regularization penalty parameter.

4. **LASSO Regression**

   Least Absolute Shrinkage and Selection Operator (LASSO) regression also known as L1 penalizes high value correlated coefficients by introducing a penalty in the model's loss function. It reduces less important feature weights to zero, removing multicollinear featurs. This is expressed as:

$$\mathscr{L}(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

   where β_i are coefficients and $|\beta_j|$ is L1 norm. λ is a learning rate to control the regularization penalty.

5. **Elastic Net Regression**

   Elastic net is a regularization model that use combination of penalties of both LASSO and Ridge regression. It is effective to overcome multicollinearity and overfitting. Mathematical way of describing as follow:

$$\mathscr{L}(\beta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2$$

6. **K-Nearest Neighbour**

   K-Nearest Neighbour (KNN) is a non-parametric, instance-based learning algorithm used for both classification and regression. It makes predictions by identifying the k closest training examples (neighbours) to a new data point and taking a majority vote (for classification) or averages value (for regression) of those neighbours. For example, if a point's three nearest are two class-1 points and class-2 point. KNN

classifies the point as class-1. KNN makes no assumptions about the global form of the decision boundary. It simply relies on the local similarity (usually Euclidean distance) between points. Because it stores all training data. KNN is sometimes called a lazy learner.

1.6.3 Classification

Classification is the main component of both the data mining and ML. It determines the class to which the data belongs [8]. In the case of ML, many procedures are involved. However, classification is still the most outstanding one. It is a method of sorting out information or data. It also lessens the effort of the users. There are basically three major types of ways to categorize data in machine learning.

1. **Binary Classification**

   This type of classification is the simplest one to use. The purpose of binary classification is to arrange data into two sperate groups. This fundamental paradigm models basic decision making scenarios, exemplified by spam detection systems.

2. **Multiclass Classification**

   Multiclass classification extends binary classification to assign data points among three or more groups. We select the models that fit the input best. Fpr example, an image recognition system works by separating images of different animals into categories of cat, dog and birds.

3. **Multi-Label Classification**

   In this type of classification a datapoint is able to be associated with multiple classes. As opposed to multi class classification, where each sample belongs to a unique class, multi-label classification lets samples be associated with many categories. A recommendation system, can identify a movie as both an action film and a comedy.

Figure 1.3 represents the difference between a Classification and regression model. Financial forecasting or prediction, cost estimation, trend analysis, marketing, time series estimation, medication response modeling and many more domains now make extensive

use of regression models. Linear Regression, LASSO, Ridge and Polynomial regression are commonly used ML models.

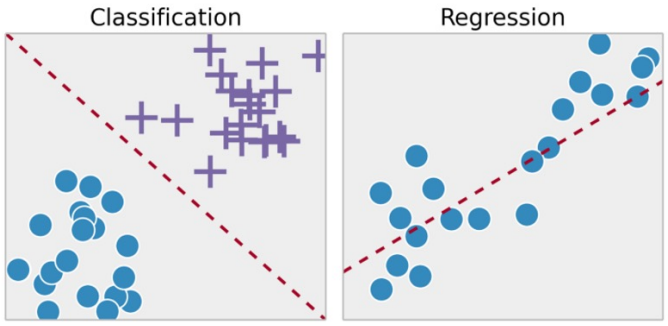


Figure 1.3: Classification vs Regression

1.6.4 Unsupervised ML

Unsupervised ML is a type where algorithms are trained on a dataset without label variable. Clustering (e.g., k-means) and dimensionality reduction (e.g.,feature reduction) identifies functional modules or latent patterns in unlabeled networks [8]. Clustering, density estimation, feature learning, dimensionality reduction, and other unsupervised ML models are among the most popular one. Figure 1.4 represent unsupervised ML.

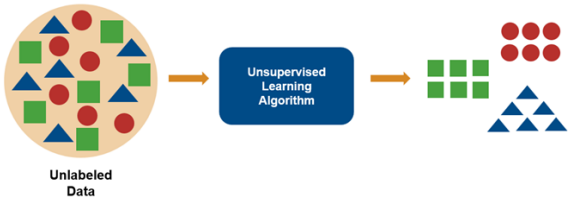


Figure 1.4: Unsupervised ML

1.6.5 Reinforcement Learning

A machine learning technique called reinforcement learning (RL) teaches software to make decisions by interacting with the environment in order to optimise cumulative rewards and strive for ideal outcomes. This is an environment driven approach. This kind of learning is based on reward or penalty as shown in 1.5

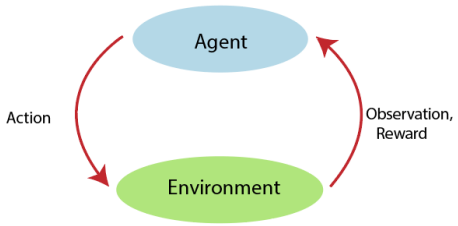


Figure 1.5: Reinforcement Learning

Graph theory provides the framework to model PPIs, topological indices quantify proteins importance and ML enhance prediction and classification task. Graph theory helps us mapping out PPI as a network, topological indices measure which proteins are important in the network and ML uses this information to predict or classify proteins for tasks like identifying disease related proteins. Together, these tools make it easier to study complex biological systems like PPIs.

1.7 Feature Engineering

Feature engineering is the process of transforming and creating input variables to improve the performance of ML algorithms. Raw data oftem must be converted into suitable form that models can use effectively. This involves task like cleaning data, handling duplicate and missing values, encoding categorical variables and generating new features from existing ones. For example, one can combine two features beacuse of their high correlation coefficients. Feature engineering also involves feature selection and reducing dimensionality by eliminating non important features. Model's performance depends critically on the quality of input feature. Feature engineering require domain knowledge about the problem statement and dataset.

1.7.1 Encoding Categorical variables

It is important to convert the text data into numerical representations, such as integers. Most of ML models are fundamentally statistical based function. these models perform computations on numerical inputs to identify relationship and predictive structures. Encoding Categorical features into numerical representations allows the algorithms to effectively an-

alyze, compare and learn from the data. There are different common methods to convert text data into numerical representations:

1. **Label Encoding:** Each unique category or word is assigned a unique integer. This is simple and efficient technique but may unintentionally introduce ordinal relationships between categories.

2. **One-Hot Encoding:** Each unique category is represented as a binary vector with a single high (1) value and the rest low (0). This avoids ordinal assumptions but increases dimensionality for large vocabularies.

3. **Binary Encoding:** Categories are first converted to integer representation, then those representations are represented in binary form. This reduces dimensionality compared to one-hot encoding while preserving distinctiveness.

1.7.2 Handling Duplicate and Missing values

It is important to handle Duplicate and missing values to improve the quality of data. The presence of duplicate and missing values affects the quality and reliability of the data. Duplicate values are often removed to ensure data consistency and avoid bias. There are several methods to impute missing values, depending on the quality of dataset.

1.7.3 Scaling

Scaling is an important preprocessing step in machine learning. Model's performance is effected when features have different units or ranges. Models that rely on distance metrics, such as k-Nearest Neighbors, Support Vector Machines and gradient-based optimization algorithms, can be biased toward features with larger numerical ranges if scaling is not applied. By standardizing or normalizing input variables, scaling ensures that all input variables contribute equally to the model. This improves convergence speed for optimization algorithms. It leads to better model performance and stability. **Types of Feature Scaling:**

1. **Min-Max Scaling (Normalization):** Rescales variables to a fixed range , using the formula:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

This preserves the relationships but is sensitive to outliers.

2. **Standardization (Z-score Scaling):** Transforms features to have zero mean and unit variance:

$$x' = \frac{x - \mu}{\sigma}$$

Standardization is less sensitive to outliers. This is widely used in algorithms that assume normally distributed data.

3. **Robust Scaling:** Uses median and interquartile range(IQR) to scale features:

$$x' = \frac{x - \text{median}}{\text{IQR}}$$

This method is robust to outliers and is suitable for datasets with extreme values.

4. **Max Absolute Scaling:** Scales features by dividing by their maximum absolute value:

$$x' = \frac{x}{|x_{\max}|}$$

Useful when data is already centered at zero and sparse.

1.7.4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA), is one of the unsupervised machine learning techniques used for dimensionality reduction. PCA transforms all the original dataset features into a new set of features named principal components through rotation of the coordinate axes. The order of the principal components is such that the first component accounts for the most the variance, the subsequent components the next increasing share of variance and so on. PCA produces a lower dimensional representation that retains most of the variance(spread) by selecting the top K principal components.

1.8 Evaluation Metrics

Evaluation metrics are used to determine the model performance and also to make decision of the model to choose. They are distinct between regression type and classification tasks.

Common Regression Evaluation Metrics:

1. **Mean Absolute Error (MAE):** The mean of absolute differences between predicted and actual values:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

2. **Mean Squared Error (MSE):** The average of squared differences between predicted and actual values:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

3. **Root Mean Squared Error (RMSE):** The square root of MSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

4. **R-Squared (R^2):** The proportion of variance in the dependent variable explained by the model's independent features can be determined by:

$$R^2\_score = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Chapter 2
Literature Review

Many researchers are utilizing ML approaches in the field of PPI through graph based methods using chemicals and structural features. The early groundwork in this field can be traced to [10] who introduced an effective algorithm for identifying cliques in undirected graph, pivotal for PPI analysis. In [11] author developed a parallel edge betweeness clustering tool for PPI networks, focusing on scalability and efficient biological networks partitioning. This enhanced the understanding of biological modules in the PPIs. Later Carlson [12] introduced the framework of topological data analysis (TDA). Their work bridge algebraic topology and ML, laying foundation for persistent homology and graph embedding to be applied in bio-informatics. Mendon [13] pioneered the use of ML for drugs sensitivity prediction using chemical and genomic profiles of cancer's cell. Their model integrated biological and structural data patterns followed by many studies. Tasdighan [14] emphasized topological and geometrical solutions to identify structural modules in proteins, providing a better approach to characterize protein structures using graph properties. Hoang [15] shifted machine learning applications to pharmacological using supervised ML to detect adverse drug reactions. This highlight the utility of ML in real world drugs safety monitoring. Redkar [16] introduced Wrapper based feature selection and class balancing for drug target interactions (DTI) prediction. This give common bottlenecks in biomedical datasets. Bagherian [17] presented a broad survey on DTI predictions categorizing ML approaches and comparing databases. Their works serve as a roadmap for model-selection. Yu [18] extended ML predictions methods for protein-lncRNA interactions, revealing adaptability across bio moleculer target. Erciyes [19] offered a comprehensive and detailed survey of biological network, covering hubs, motifs and community structures. Demirsoy [20] applied ML techniques for detection drug to drug interactions using clinical datasets, bridging theory and practice. Abubakar [21] used quantitative structure property relationship (QSPR) modeling to predict the physical properties of antibacterial drugs using a topological indices like neighborhood sum degree by using two regression models, linear regression (LR) and (SVR). Instead of using traditional methods, they used simplified molecular input line entry system (SMILES) representations of drugs molecules to calculate numerical values. to improve accuracy, they applied backward elimination for LR and sequential backward search selection for SVR. The SVR performed better, especially on small datasets due to effective

feature selection and hyperparameter tuning, their results showed that neighborhood sum degree topological index (NTIs) can predict properties like boiling points, melting points, and surface tension. Kour [22] applied ML models for anti-cancer drugs predictions using topological indices. Ejima [23] proposed ensamble learning framework to modeling drugs properties for mental disorder. Tang [24] combined ML techniques and bio-informatics (like PPIs etc) to predict disease progression across domains. In [25] author suggested that we can gather various pieces of information through analysis and interpretation of data. Cluster analysis is a primary technique to analyze data even when there is little prior information. This kind of research is done across many disciplines. While it is beneficial to acquire applicable sophisticated tools from various fields, having too many options can also lead to lack of clarity and confusion. In [15] , they focused on the various clustering techniques which are employed across the disciplines of statistics, computer science, and ML. They also demonstrated the application of these techniques in case studies such as the traveling salesman problem and bio-informatics, which is an emerging area of study. Further, they addressed other relevant issues such as the measurement of similarity between values and validity of the datasets in relation to the defined clusters. Rodriguez-Nieva [26] used ML to detect topological order, contributing theoretical insight relevant to biological clustering. Kiouri [27] suggested that structure based PPI prediction, utilizing ML and deep learning (DL), enhances accuracy and provides insights functional sites by integrating 3D spatial and bio chemical features.In [28] authors suggested that chemical graphs can be used to analyze the molecular characteristics and structural design of anti-HIV drugs through topological indices, which correlate structure with biological activity and properties. By employing supervised ML, including random forest (RF) and XGBoost algorithms, a predictive model is developed to enhance drug effectiveness. In [29] authors used a network based approach to identify groups of connected genes (hub networks) in breast cancer, rather than focusing on single gene. By analyzing gene expression data and building a protein interaction networks, researchers found key gene clusters and pathways, like those involved in cell cycle, that play a big role in breast cancer. These hub sub-networks were more reliable than individual genes for understanding cancer process. In [30] author used a mix of bio-informatics methods to find key genes linked with preeclampsia (PE).

They analyzed gene data from the GEO database, used weighted gene co-expression network analysis (WGCNA) and ML and built protein interaction networks to identify gene leptin (LEP) as a key player. They confirmed that LEP's role with lab tests on cell lines and human tissue, showing it could be a new bio maker and drug target for PE. In [31] a study developed a new identify key genes for hepatocellular carcinoma (HCC) by analysis of genetic data from multiple platforms. Instead of focusing on common differentially expressed genes (DEGs), it used a statistical and ML approach, including SVM and PPI networks, to find platform independent genes. By combining DEGs, hub genes and module genes from PPI networks, six key genes were identified. In [32] they used RNA sequencing an ML to identify key genes and immune cell patterns in polycystic ovary syndrome (PCOS). By analyzing granulosa cell samples from 13 healthy women and 25 with PCOS, combined with public database data, researchers corrected batch effects and found 824 DEGs. Using least absolute shrinkage and selection operator (LASSO), SVM and recursive feature elimination (REF) ML methods, four hub genes (CNTN2, CASR, CACNB3, MFAP2) were selected. The accuracy of these genes was tested with SVM and XGBoost models. In [33] authors focused on using PPI networks to better understand cancer at the moleculer level. They introduced new DL method, Node2Vec, which analyzes the structure of PPI networks to identify key cancer hub genes quickly and accurately. The method use another tool, an autoencoder, to reduce unnecessary details in the networks while keeping the most important features. By combining these techniques, the study predicts cancer genes and treatment targets acheiving a high accuracy of 99.7%. In [34] authors introduced a study aimed to better understand alopecia areata (AA), a type of hair loss linked to autoimmune issue, by analyzing gene data. They used methods like differential gene expression, immune status checks, WGCNA and functional enrichment to find genes tied to both immune and AA. They then applied ML to identify three key hub genesas potential diagnostic markers for AA. In [35] they computed degree based topological indices for the drug Astragaloside IV and they construct a network to identify highly connected sub-networks or modules. they aimed to detect a master regulator index with in these modules.

Drug discovery is an integral area of biomedical research efforts being made for the identification and development of new drugs for the treatment of diseases such as cancer,

cardio vascular diseases, and infections. The conventional approaches to drug discovery, largely dependent on high throughput screening, in vitro testing, and clinical trials, are time-consuming, expensive, and inefficient with only one in 10,000 compounds being commercialized [36]. Failures could arise from inappropriate PK and physiochemical profiles, mostly poor solubility, low absorption, high toxicity, which are normally identified during late-stage development phases. Because of the high number of failures, there arises an imminent need for computational approaches able to estimate molecular properties during early discovery stages in order to increase efficiency, cut costs, and boost success rates. Computational chemistry, cheminformatics, and machine learning (ML) have recently combined to reshape the landscape in drug discovery research by making it possible to predictably model molecular behavior. The challenge is to predictably estimate key molecular properties such as molecular weight (MW), Lipophilicity (LogP), and hydrogen bonding (HBD), which regulate the pharmacological, bioavailability, and toxicology properties for a compound. Molecular structure representation by the Simplified Molecular-Input Line Entry System (SMILES) notation, together with Quantitative Structure-Property Relationship (QSPR) modeling, is an incredibly useful strategy for correlating structural knowledge with molecular character. Topological descriptors extracted from molecular structure graphs, including the Wiener, Balaban, Randić number, and Zagreb indices, offer an objective means for characterizing molecular connectivity, branching, and topological complexity, which can then be modeled using statistical learning approaches [39]. Machine learning is now also an integral part of cheminformatics with its capacity for the learning of complex, nonlinear relationships between molecular structure and other physicochemical properties. Methods that range from univariate linear regression have been employed each on the sixteen different models (descriptors) are used to predict in areas such as solubility, lipophilicity prediction, toxicity prediction, and others [38]. There have been many instances where regression-based ML models in SMILES representations or topological descriptors proved their predictive prowess in an efficient manner from a computational perspective. Some of these topological representations consist of high computational costs for their execution. The applications of ML do not end with drug designing but continue in other fields in the biomedical area. For example, algorithms like Multiple Linear Re-

gression (on the basis of all indices, selected indices, PCA-based), LASSO, Ridge, Elastic Net Classification have successfully been made use of in the database for cancers with a high accuracy level reaching 97.1% [40, 41, 42, 43]. These developments make it clear that there is promise in the use of ML in the biomedical field, while they also raise issues of bias, overfitting, explanation, or generalization with regard to molecular properties prediction. Although there are some success reports in both QSAR modeling and predictions using the machine learning technique, there still exist some issues which are not addressed. There is still a chance of overfitting of QSAR models in making predictions [36]. Also, activity cliffs, which refer to molecules with similar structure having dramatically different values of properties corresponding to the molecules, can considerably affect the QSAR models. Deep learning models, which can identify an intricate relationship with a high level of accuracy, still lack interpretability of the models and need abundant resources, which is not affordable by most research labs.

Chapter 3
Main Results

Toward this end, an interpretable machine learning framework combining SMILEs representations of molecular graphs based on sixteen topological descriptors is proposed in this research work for the prediction of key physiochemical properties of drug molecules. The dataset of forty-seven drug molecules collected from the DrugBank was analyzed, where descriptors like Wiener, Balaban, Harary, Randić, Zagreb, Shannon Entropy, among others (Table 3.2) computed by employing the RDKit library. The dataset in Figure 3.2 has been rigorously analyzed by processing it exhaustively for handling the missing values, removing outliers by capping the values with the InterQuartile Range (IQR), Box-Cox transformation, followed by scaling by using the MinMax scaler. The multiple regression models, specifically the Ordinary Least Squares (OLS), Ridge Regression, LASSO Regression, Elastic Net Regression models, were used for predicting the MW, LogP, HBD values accurately with an enhanced interpretation by employing SHapley Additive exPlanations (SHAP) theory. The main intention of this research work will be exploring the efficacy, interpretability of outcomes, generalizing it by employing the linearity of the models, the conventional models of regression, along with both of their potentials on molecular properties employing the theorem of graph theory by utilizing Explainable Machine Learning.

Name	**Symbol**	**Formula**
Molecular Weight	MW	Sum of atomic weights
LogP	LogP	Octanol-water partition coefficient
H-Bond Donors	HBD	Count of H-bond donors
Wiener Index	W	$\sum_{i<j} d_{ij}$
Balaban Index	J	$\frac{m}{s-d}$, $s = \sum 1/\sqrt{deg(u)deg(v)}$
Harary Index	H	$\sum_{i\neq j} 1/d_{ij}$
Randić Index	χ	$\sum 1/\sqrt{deg(u)deg(v)}$
Zagreb First Index	M_1	$\sum deg(v)^2$
Zagreb Second Index	M_2	$\sum deg(u) \cdot deg(v)$
Hyper-Wiener Index	WW	$\sum_{i<j} d_{ij} + d_{ij}^2/2$
Schultz Index	S	$\sum_{i\neq j} (deg(i) + deg(j)) \cdot d_{ij}$
Gutman Index	Gut	$\sum_{i<j} deg(i) \cdot deg(j) \cdot d_{ij}$
Hosoya Index	Z	Number of matchings / spanning trees
Estrada Index	EE	$\sum e^{\lambda_i}$, λ_i eigenvalues
Shannon Entropy	H_{Sh}	$-\sum p_i \log_2 p_i$
Mean Distance Degree	MDD	Average shortest path length
Molecular Connectivity	MC	Sum of degree centralities
Graph Energy	$E(G)$	Sum of absolute eigenvalues
Eccentric Connectivity	EC	$\sum deg(v) \cdot ecc(v)$

Table 3.1: Graph-theoretic and Molecular Indices with Symbols and Formulas

The data is cleaned carefully. The values of missing data, duplicates, and outliers are handled appropriately. The existence of an outlier can be identified by using the Interquartile Range (IQR) approach. The correlation analysis among the variables is performed. The data is split into an 80:20 split. Taking into consideration the variance stabilization, the

transformation of the Box-Cox function is done. The descriptors of the data are then normalized. The data is quantified by making all descriptors range from 0 to 1. Meanwhile, multiple linear regression models for Ordinary Least Squares (with multiple linear regression), LASSO, Ridge, Elastic Net regression were built to predict each molecular property independently. Various feature selection strategies, PCA dimensionality reduction, and regularization algorithms were combined effectively for averting overfittings in the models. Performance metrics for evaluation were measured with combined metrics for Mean Absolute Error (MAE), Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), followed by residual analyses for checking the adequacy of the finalized models. SHapley Additive exPlanations (SHAP) values were then utilized for identifying key descriptors in the feature set that contribute maximally for the predictions in a comprehensive manner for identifying clear interpretations in the models adopted in the research study for better transparency and improved computational approaches for reliable predictive capabilities in drug discovery applications [37, 44].

The integrated workflow for machine learning initiated in the present study is explained in Figure 3.1. The workflow initiated in the present study is associated with multiple steps, including data collection, preprocessing, feature engineering, model development, testing, interpretation of results according to their efficiency, followed by residual testing, SHAP testing for descriptors, in addition to the modification in inefficient models for preprocessing by adjustment in parameters. Therefore, efficiency in addition to transparency is assured in computational drug discovery initiated using machine learning algorithms.

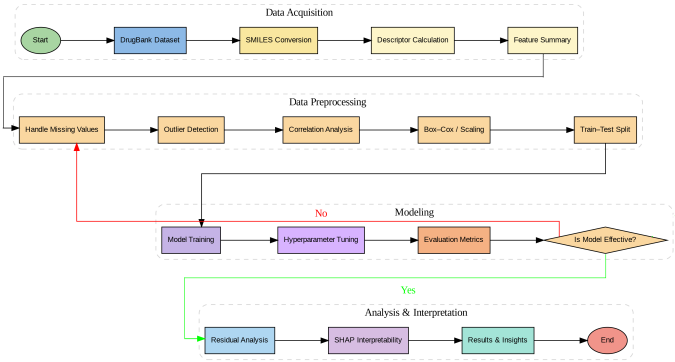


Figure 3.1: Research Methodology Flowchart

This section presents the main results, including data preprocessing, model training and estimation of the properties for 47 drug compounds.

Algorithm 1 Processing Chemical Data from DrugBank

Input: SMILES data from DrugBank
Output: Ensembled dataset with topological indices and molecular properties
dataset ← Load SMILES data from DrugBank
molecular_graphs ← Convert SMILES to molecular graphs using RDKit
simple_graphs ← Simplify molecular graphs to simple graphs using NetworkX
for each *graph* ∈ *simple_graphs* **do**
    topological_indices ← Calculate 16 topological indices using RDKit
    properties ← Extract 3 molecular properties (MW, LogP, HBD) using RDKit
    dataset ← Append *topological_indices* and *properties* to *dataset*
end for
return *dataset*

Algorithm 1 provides us the value for graph based topological indices and moleculer propertoes provided in Table 3.2 of all 47 drugs, are shown in Figure 3.2.

Name	**Symbol**	**Formula**
Molecular Weight	MW	Sum of atomic weights
LogP	LogP	Octanol-water partition coefficient
H-Bond Donors	HBD	Count of H-bond donors
Wiener Index	W	$\sum_{i<j} d_{ij}$
Balaban Index	J	$\frac{m}{s \cdot d}$, $s = \sum 1/\sqrt{deg(u)deg(v)}$
Harary Index	H	$\sum_{i \neq j} 1/d_{ij}$
Randić Index	χ	$\sum 1/\sqrt{deg(u)deg(v)}$
Zagreb First Index	M_1	$\sum deg(v)^2$
Zagreb Second Index	M_2	$\sum deg(u) \cdot deg(v)$
Hyper-Wiener Index	WW	$\sum_{i<j} d_{ij} + d_{ij}^2/2$
Schultz Index	S	$\sum_{i \neq j} (deg(i) + deg(j)) \cdot d_{ij}$
Gutman Index	Gut	$\sum_{i<j} deg(i) \cdot deg(j) \cdot d_{ij}$
Hosoya Index	Z	Number of matchings / spanning trees
Estrada Index	EE	$\sum e^{\lambda_i}$, λ_i eigenvalues
Shannon Entropy	H_{Sh}	$-\sum p_i \log_2 p_i$
Mean Distance Degree	MDD	Average shortest path length
Molecular Connectivity	MC	Sum of degree centralities
Graph Energy	$E(G)$	Sum of absolute eigenvalues
Eccentric Connectivity	EC	$\sum deg(v) \cdot ecc(v)$

Table 3.2: Graph Theoretic and Molecular Indices with Symbols and Formulas

Next we have cleaned the data in Figure 3.2 after applying the steps mention in Algorithm 2.

The following Table 3.3 provides a summary of the key statistics of the sixteen molecular as well as graph-theoretical descriptors estimated for the forty-seven drug molecules. The key information provided by this table provides insight into the variation or spread of the data.

Drug Name	Molecular Weight	LogP	H-Bond Donors	Wiener Index	Balaban Index	Harary Index	Randic Index	Zagreb first Index	Zagreb Second Index	Hyper-Wiener Index
Adagrasib	604.130	4.73298	0	6476.0	0.321259	408.198801	20.833739	234	282	36225.0
Bevacizumab + Rituximab	390.221	2.06100	2	949.0	0.421958	122.709163	9.489601	92	103	4158.5
Carboplatin	371.250	0.49320	2	216.0	0.847855	73.166667	5.962761	74	90	578.0
Binimetinib	441.232	3.01130	3	1808.0	0.435827	206.955073	12.917923	140	166	7748.0
Cyclophosphamide	261.089	1.88400	1	301.0	0.629237	75.671429	6.726492	64	72	915.5
Nintedanib	539.636	3.61590	2	5935.0	0.299096	354.690466	19.333739	212	251	36614.5
Mitomycin	334.332	-1.65120	3	1115.0	0.589351	189.565873	11.340277	142	185	3874.5
Encorafenib	540.021	3.90840	3	4220.0	0.336212	306.277614	16.873231	186	214	22890.0
Dabrafenib	519.574	5.36490	2	3610.0	0.382346	307.703824	16.380866	192	229	17764.0
Doxorubicin	543.525	0.00130	6	4392.0	0.393687	378.328000	18.427777	220	275	20901.0
Ensartinib	561.445	4.71940	3	5482.0	0.291647	322.894436	18.027785	200	235	35040.0
Entrectinib	560.649	5.02690	3	6234.0	0.303948	368.432139	19.907020	222	261	37739.0
Pralsetinib	533.612	4.20014	3	5920.0	0.287190	335.340981	18.741152	210	248	38801.0
Lazertinib	554.655	4.10420	2	6081.0	0.304253	370.070202	19.944223	214	251	35904.5
Selpercatinib	525.613	3.28438	1	5799.0	0.298945	340.069891	18.807308	218	261	37261.5
Tepotinib	492.583	4.00788	0	5308.0	0.284719	305.786333	18.068015	194	225	34881.0
Lurbinectedin	784.888	4.18894	4	10530.0	0.352310	676.458282	26.982439	336	432	55537.0
Cisplatin	300.046	1.70050	2	16.0	1.250000	14.000000	2.000000	20	16	30.0
Ifosfamide	261.089	1.88400	1	307.0	0.616939	75.307143	6.726492	64	72	957.5
Vincristine	824.972	3.51750	3	12104.0	0.345787	762.075369	28.758719	354	458	63460.0
Repotrectinib	355.373	2.55220	2	1481.0	0.508788	205.588095	12.508040	142	169	5662.5
Erlotinib	393.443	3.40510	1	2396.0	0.368754	220.031130	14.245045	142	163	11633.0
Gefitinib	446.910	4.27560	1	3020.0	0.345885	241.852640	15.135362	160	186	16208.0
Osimertinib	499.619	4.50980	2	4557.0	0.328438	319.604161	17.799264	192	226	25071.5
Crizotinib	450.345	5.03770	2	2633.0	0.377675	237.001177	14.435561	160	190	13318.5
Alectinib	482.628	4.77328	1	3989.0	0.371773	326.557914	17.417383	206	256	21129.5
Ceritinib	558.148	6.36192	3	4901.0	0.325701	333.596494	18.056588	200	235	27481.5
Brigatinib	584.105	5.09000	2	6293.0	0.283944	348.891510	19.206860	214	251	41540.5
Lorlatinib	406.421	2.80068	1	2094.0	0.479562	258.097835	14.294933	164	199	8304.0
Pemetrexed	427.417	0.66640	6	3252.0	0.320122	235.812088	14.740091	158	182	19028.0
Trametinib	615.403	3.94012	2	3853.0	0.403165	345.410589	17.578295	206	253	17951.5
Sotorasib	560.605	4.48142	1	5138.0	0.367788	399.258186	19.526984	224	274	25503.0
Docetaxel	807.890	3.25960	5	12587.0	0.306536	675.575430	26.990415	326	403	73873.5
Paclitaxel	853.918	3.73570	4	14767.0	0.297476	748.381767	29.272290	346	429	87927.5
Afatinib	485.947	4.38990	2	3655.0	0.346601	279.829193	16.385050	176	203	19126.5
Dacomitinib	469.948	5.15510	2	3594.0	0.329949	264.061203	16.029208	170	197	20090.0
Gemcitabine	263.200	-1.28860	3	560.0	0.619885	117.176984	8.374246	96	116	1851.0
Vinorelbine	778.947	4.75360	2	10686.0	0.354504	705.464802	27.384799	338	439	55366.0
Etoposide	588.562	1.33860	3	5464.0	0.372918	419.702162	20.282427	242	302	27681.0
Topotecan	421.453	1.84680	2	2485.0	0.444314	269.129237	14.740238	180	226	11237.5
Larotrectinib	428.443	2.94750	2	2782.0	0.390236	253.552359	14.991204	172	207	13873.0
Irinotecan	586.689	4.09110	1	7049.0	0.301535	410.661655	20.817049	246	306	45517.5
Methotrexate	454.447	0.09877	6	3832.0	0.308467	258.758437	15.633938	168	193	23144.0
Alpelisib	441.479	3.83502	2	2611.0	0.383168	237.736264	13.913722	162	193	13129.5
Capmatinib	412.428	3.42900	1	2936.0	0.366653	249.542056	15.118525	170	204	15581.0
Imatinib	493.615	4.59032	2	5324.0	0.284462	305.029034	18.030010	194	224	35001.0
Everolimus	958.240	6.19720	3	22088.0	0.226788	765.038760	32.287588	340	397	156780.0

Drug Name	Schultz Index	Gutman Index	Hosoya Index	Estrada Index	Shannon Entropy	Mean Distance Degree	Molecular Connectivity	Graph Energy	Eccentric Connectivity
Adagrasib	55986	30163	36750.0	109.277232	5.354981	7.004867	2.285714	56.434695	1175
Bevacizumab + Rituximab	7228	3418	6.0	46.940417	4.208695	4.745000	2.105263	24.310964	366
Carboplatin	1740	864	24.0	33.395559	3.539149	2.556213	2.333333	14.397986	122
Binimetinib	14830	7560	174.0	66.917666	4.658862	4.960219	2.230769	34.008889	553
Cyclophosphamide	2264	1052	6.0	32.847953	3.708966	3.071429	2.153846	17.119207	149
Nintedanib	50776	27099	6264.0	100.461731	5.248427	7.418750	2.256410	52.636761	1377
Mitomycin	9384	4877	388.0	64.188101	4.452966	3.871528	2.347826	30.893228	356
Encorafenib	33932	16967	180.0	88.488477	5.049787	6.512346	2.171429	43.931926	975
Dabrafenib	30186	15729	1080.0	88.210840	5.007089	5.893878	2.235294	42.920911	827
Doxorubicin	37014	19417	7134.0	99.501715	5.160764	5.775148	2.263158	49.782344	919
Ensartinib	46184	24285	1296.0	94.492262	5.146292	7.592798	2.216216	47.840440	1222
Entrectinib	54696	29957	37584.0	104.221138	5.299992	7.417014	2.300000	53.526677	1338
Pralsetinib	51314	27775	5400.0	98.473464	5.204653	7.784352	2.263158	50.448758	1415
Lazertinib	51864	27592	6480.0	102.515255	5.293925	7.234979	2.250000	53.593754	1254
Selpercatinib	50688	27609	20880.0	100.598743	5.211551	7.625247	2.315789	51.331843	1398
Tepotinib	46106	25003	7776.0	92.840158	5.158726	7.754565	2.277778	49.341530	1351
Lurbinectedin	94212	52548	42977592.0	148.076160	5.717544	6.715561	2.363636	74.319231	1682
Cisplatin	88	28	1.0	10.524391	2.000000	1.280000	2.000000	4.000000	12
Ifosfamide	2300	1064	6.0	32.848418	3.708966	3.132653	2.153846	17.411738	162
Vincristine	105160	56913	6355930.0	156.797412	5.796375	6.724444	2.305085	78.750491	1665
Repotrectinib	12892	6996	2053.0	66.109259	4.624380	4.381657	2.320000	33.916127	472
Erlotinib	19602	9971	210.0	70.727890	4.804451	5.697979	2.214286	37.735202	682
Gefitinib	25852	13808	1260.0	77.225551	4.899316	6.285120	2.266667	40.950705	866
Osimertinib	38080	19795	1044.0	91.888434	5.124259	6.657414	2.222222	47.488489	1036
Crizotinib	22568	12062	1080.0	75.477313	4.827670	5.851111	2.275862	39.093833	762
Alectinib	35294	19478	35244.0	93.370733	5.091533	6.155864	2.342857	48.252213	1077
Ceritinib	40976	21361	1296.0	94.513353	5.149520	6.788089	2.216216	47.950544	1076
Brigatinib	54144	29074	7776.0	100.573608	5.242856	7.866250	2.256410	51.771936	1416
Lorlatinib	17716	9336	1974.0	75.921668	4.804795	4.653333	2.275862	38.342908	577
Pemetrexed	26846	13784	174.0	76.217061	4.857973	6.767950	2.200000	38.656590	906
Trametinib	32656	17228	3780.0	95.603929	5.103490	5.628926	2.277778	47.139324	889
Sotorasib	43142	22557	7560.0	103.580057	5.251987	6.113028	2.250000	52.245372	1029
Docetaxel	105288	54885	35568.0	147.472417	5.716700	7.483353	2.210526	71.380842	1759
Paclitaxel	125600	66656	213408.0	157.910614	5.834942	7.683143	2.229508	78.427337	2007
Afatinib	30708	16064	1050.0	84.562748	5.016159	6.323529	2.242424	43.400235	888
Dacomitinib	30620	16278	1260.0	81.997810	4.981746	6.600551	2.250000	43.012797	968
Gemcitabine	4436	2176	30.0	44.681623	4.023472	3.456790	2.235294	21.926161	253
Vinorelbine	93462	50917	5573800.0	149.501717	5.727423	6.578024	2.321429	75.391043	1583
Etoposide	48124	26410	171360.0	109.303645	5.310558	6.195011	2.341463	55.767368	1152
Topotecan	21388	11442	5705.0	80.451180	4.846232	5.171696	2.333333	40.291157	685
Larotrectinib	24374	13316	4350.0	79.666642	4.885588	5.789802	2.333333	40.998014	800
Irinotecan	62700	34804	205380.0	111.520980	5.348606	7.624662	2.333333	57.545964	1572
Methotrexate	31534	16151	210.0	80.901124	4.942731	7.037649	2.187500	41.130085	1009
Alpelisib	21360	10876	150.0	74.946844	4.768844	5.802222	2.206897	36.373399	696
Capmatinib	25806	14132	6090.0	79.283551	4.896372	6.110302	2.333333	41.546508	886
Imatinib	46490	25355	7776.0	92.829852	5.158726	7.777940	2.277778	49.161634	1361
Everolimus	180130	91692	6084.0	165.132415	5.976455	9.553633	2.119403	84.739665	2577

Figure 3.2: Data Assemblage

Algorithm 2 Data Preprocessing and Modeling Pipeline

1: **Missing and Duplicate Handling:** Impute missing values and remove duplicate entries.
2: **Outlier Treatment:** Detect outliers using the Interquartile Range (IQR) method and apply capping.
3: **Correlation Analysis:** Compute correlation matrix after outlier treatment.
4: **Train–Test Split:** Divide the dataset into training (80%) and testing (20%) sets, yielding $\{X_{train}, X_{test}, y_{train}, y_{test}\}$.
5: **Power Transformation:** Apply Box–Cox transformation on $\{X_{train}, X_{test}\}$ using `PowerTransformer()`.
6: **Normalization:** Scale features with `MinMaxScaler()` on X_{train} and X_{test}.
7: **Model Training and Prediction:** Train model using $\{X_{train}, y_{train}\}$ and generate predictions on X_{test}.
8: **Evaluation:** Evaluate model performance using y_{test}.
9: **Feature Engineering:** Perform feature selection and feature extraction (PCA).
10: **Regularization:** Apply regularization techniques (e.g., Ridge, LASSO, ElasticNet).
11: **Residual Analysis:** Analyze residuals to assess model fit.
12: **Model Interpretation:** Perform SHAP analysis for feature importance and interpretation.

For instance, it could be observed from this table that there is variation in the molecular size of the molecules.

Feature	Count	Mean	Std Dev	Min	Max
Molecular Weight	47	514.26	154.90	261.08	958.23
LogP	47	2.32	1.81	-1.65	6.36
H-Bond Donors	47	2.34	1.39	0.00	6.00
Wiener Index	47	328.19	179.31	88.00	658.00
Balaban Index	47	17.05	6.13	2.25	33.20
Harary Index	47	377.85	179.31	88.00	658.00
Randić Index	47	10.06	2.18	4.81	15.67
Zagreb 1st Index	47	176.40	35.64	114.00	266.00
Zagreb 2nd Index	47	186.68	36.96	114.00	277.00
Hyper-Wiener Index	47	186.40	36.23	114.00	266.00
Schultz Index	47	1186.39	416.23	524.00	1692.00
Gutman Index	47	849.06	313.17	379.00	1456.00
Hosoya Index	47	4507.00	1634.67	1336.00	7800.00
Estrada Index	47	74.83	25.65	25.64	134.70
Shannon Entropy	47	6.23	1.36	2.17	9.55
Mean Distance Degree	47	4.94	0.68	3.22	6.32
Molecular Connectivity	47	1.61	0.07	1.35	1.76
Graph Energy	47	16.85	2.17	13.44	22.36
Eccentric Connectivity	47	528.00	135.60	261.00	958.00

Table 3.3: Descriptive Statistics of Features for 47 Compounds

Figure 3.3 plots the distribution of features before and after carrying out outlier removal using the IQR boxplot method. The distribution on the left exhibits marked outliers as well as asymmetry, while the right distribution illustrates how truncating beyond the whisker values will yield a distribution closer to normal.

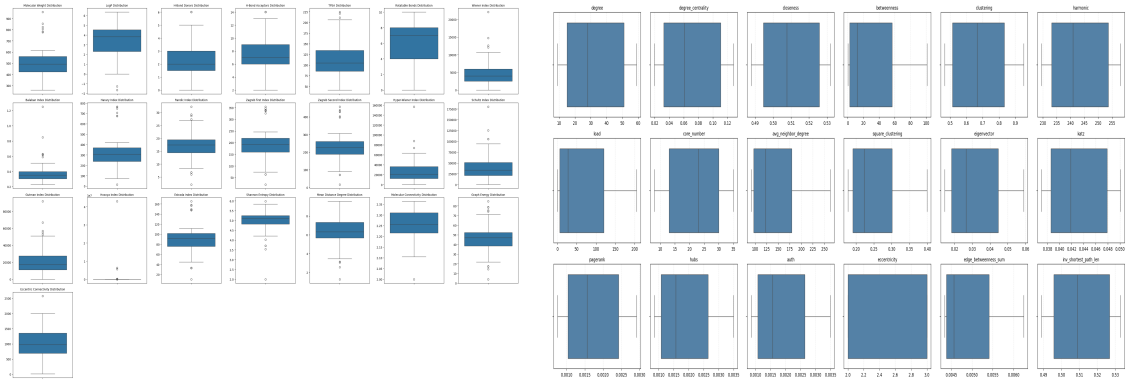


Figure 3.3: Comparison of Feature Distributions Before and After Outlier Capping

Figure 3.4 shows the correlation heatmap of 19 molecular and graph theoretic features after outlier capping. Capping reduces the impact of extreme values stabilizing feature relationships and improving interpretability. Strong positive correlations appear among graph indices while molecular descriptors like molecular weight and LogP show moderate associations with topological measures.

The dataset was partitioned into input and output feature sets. The input features consisted of graph based topological indices while the outputs were physicochemical properties including molecular weight (MW), LogP, hydrogen bond donors (HBD). Ml models were

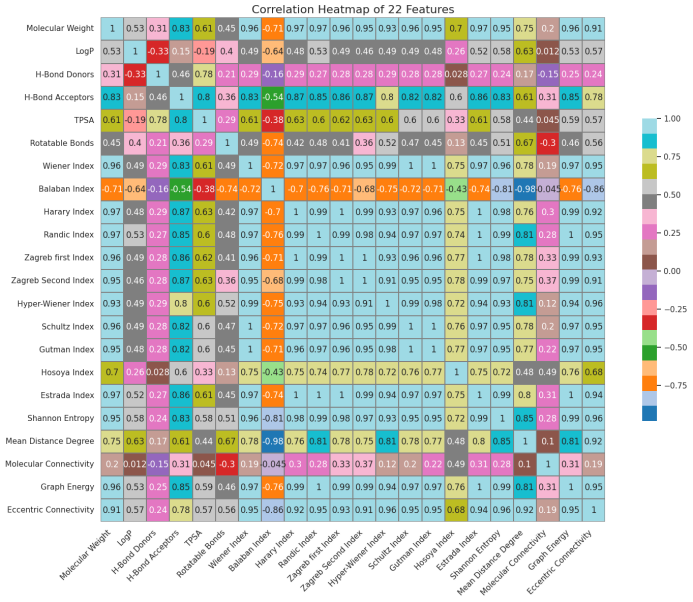


Figure 3.4: Correlation After Outlier Capping

trained to predict one property at a time. An 80:20 train test split was applied with the test set selected randomly from the dataset.

The Box-Cox transformation was then applied to all input features to approximate a normal distribution as given by the formula below:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & \lambda \neq 0, \\ \ln(y), & \lambda = 0, \end{cases} \tag{3.0.1}$$

where $y > 0$ represents the response variable and λ denotes the BoxCox transformation parameter used to stabilize variance and improve the normality of feature distributions. The optimal λ values obtained for each molecular descriptor are summarized in Table 3.4. These transformation parameters ensure that the model assumptions of linearity and homoscedasticity are better satisfied, thereby improving regression performance and interpretability.

Figure 3.5 illustrates the effect of the Box-Cox transformation. Each subplot contains a pair of kernel density estimates (KDEs), where the left panel shows the feature distribution

Index	Descriptor (Box-Cox Lambdas)	Value
0	Wiener Index	0.565243
1	Balaban Index	-0.940138
2	Harary Index	0.821773
3	Randić Index	0.920583
4	Zagreb First Index	0.880316
5	Zagreb Second Index	0.866826
6	Hyper-Wiener Index	0.495715
7	Schultz Index	0.547390
8	Gutman Index	0.527341
9	Hosoya Index	0.232864
10	Estrada Index	0.918232
11	Shannon Entropy	2.020368
12	Mean Distance Degree	1.681031
13	Molecular Connectivity	9.498829
14	Graph Energy	0.930097
15	Eccentric Connectivity	0.797527

Table 3.4: Box Cox Lambdas

before transformation and the right panel shows the distribution after applying Box-Cox. Across all pairs the transformed data more closely approximates a normal distribution.

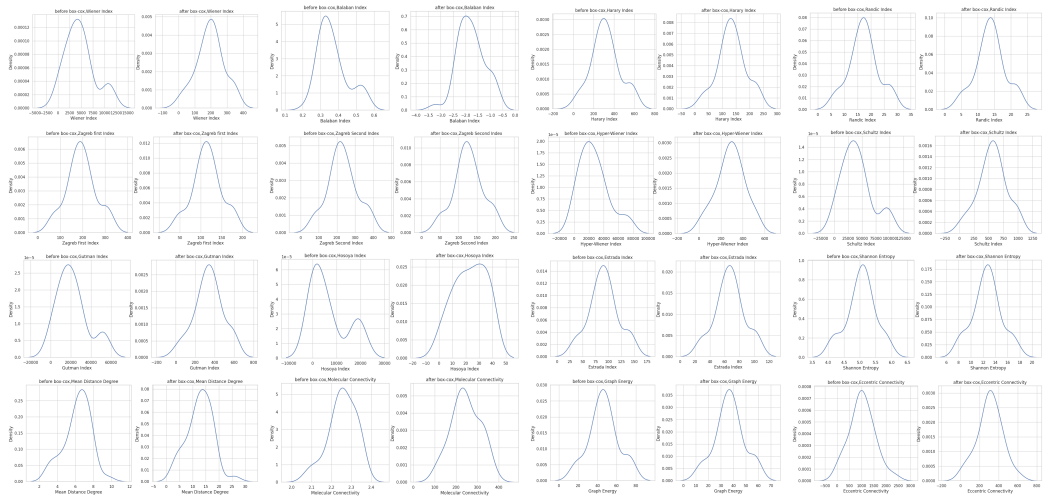


Figure 3.5: Box Cox Transformation Plots of Dataset Input Features

We applied Min-Max scaling to all input columns rescaling their values to the range $(0, 1)$ for uniformity across features. The left panel of Figure 3.6 shows the normalization of the training data while the right panel shows the normalization distribution of the test data. This step is crucial for algorithms that are sensitive to feature scales and was applied uniformly

across the dataset.

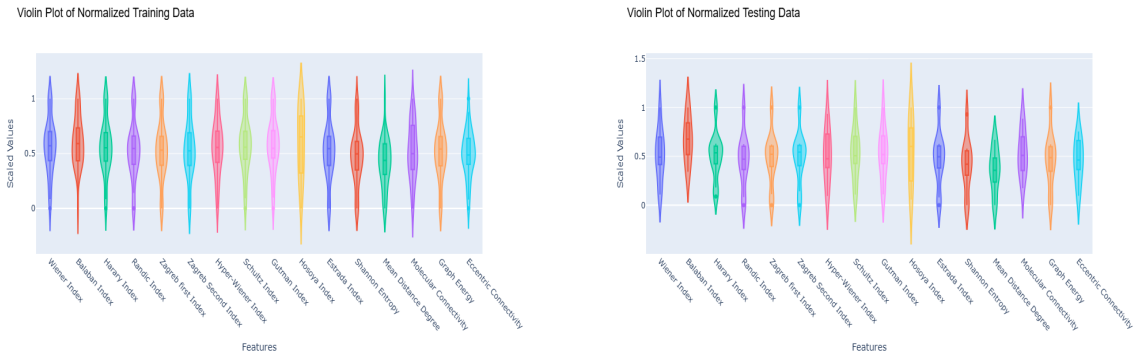


Figure 3.6: Min Max Normalization of Input Features: Training vs Testing Sets

Now the processing step is completed. We have obtained clean data for modeling, prediction and evaluation.

3.0.1 Linear Models Fitting for Molecular Weight (MW)

In this section we have fitted various linear models to predict MW values of the drugs based on topological indices and evaluated the models based on root mean squared error (RMSE) and R^2-score. We have used these metrics values on train and test data sets both to avoid the possibility of overfitting, data leakage, or over-regularization.

Univariate Linear Regression Analysis for MW

Table 3.5 summarizes the results of the univariate linear regression analysis performed for each of the sixteen topological descriptors used to predict molecular weight (MW). In these models, each descriptor was independently used as a single scaled input feature to assess its individual predictive strength. Feature importance can be assessed based on the magnitude of the slope (coefficient) values, which indicate the extent to which each feature contributes to the prediction of molecular weight. The table reports the training and testing coefficients of determination (R^2), root mean squared error (RMSE), regression slope, and intercept for each model, along with the corresponding regression equation. Figure 3.7 shows the comparative graphs of RMSE and values of R^2 measures for training and testing datasets, provided. From among all models, Model 12, which was developed using the Shannon

Entropy index, showed the best predictability with the highest values of R^2 (0.905 on training data and 0.974 on testing data) and lowest values of RMSE (40.18 on training data and 22.93 on testing data), as seen in Figure 3.7. The poorest predictability was found in Model 14 developed using the Molecular Connectivity Index with the lowest values of R^2 (0.017 on training data and 0.001 on testing data) and the highest values of RMSE (129.24 on training data and 142.11 on testing data). The results clearly show that molecular complexity, which can be measured by the Shannon Entropy index, predominantly influences molecular weights, while molecular connectivity does not.

Model	**Feature**	R^2_{Train}	$\text{RMSE}_{\text{Train}}$	R^2_{Test}	$\text{RMSE}_{\text{Test}}$	Slope	Intercept	Equation
1	$W(G)$	0.890	43.25	0.960	28.44	494.20	233.68	$\text{MW} = 233.68 + 494.20\,W(G)$
2	$J(G)$	0.419	99.36	0.552	95.17	-378.35	743.79	$\text{MW} = 743.79 - 378.35\,J(G)$
3	$H(G)$	0.919	37.09	0.942	34.36	502.69	232.42	$\text{MW} = 232.42 + 502.69\,H(G)$
4	$\chi(G)$	0.916	37.80	0.969	24.97	475.90	262.37	$\text{MW} = 262.37 + 475.90\,\chi(G)$
5	$M_1(G)$	0.910	39.09	0.930	37.63	468.83	264.90	$\text{MW} = 264.90 + 468.83\,M_1(G)$
6	$M_2(G)$	0.895	42.32	0.892	46.65	465.03	263.51	$\text{MW} = 263.51 + 465.03\,M_2(G)$
7	$WW(G)$	0.834	53.08	0.927	38.39	480.77	249.05	$\text{MW} = 249.05 + 480.77\,WW(G)$
8	$S(G)$	0.885	44.20	0.955	30.14	491.53	232.87	$\text{MW} = 232.87 + 491.53\,S(G)$
9	$\text{Gut}(G)$	0.879	45.38	0.950	31.90	490.13	232.09	$\text{MW} = 232.09 + 490.13\,\text{Gut}(G)$
10	$Z(G)$	0.573	85.14	0.585	91.63	317.39	322.34	$\text{MW} = 322.34 + 317.39\,Z(G)$
11	$EE(G)$	0.918	37.32	0.958	29.06	469.49	265.85	$\text{MW} = 265.85 + 469.49\,EE(G)$
12	$H_{\text{Sh}}(G)$	0.905	40.18	0.974	22.93	491.01	280.61	$\text{MW} = 280.61 + 491.01\,H_{\text{Sh}}(G)$
13	$\text{MDD}(G)$	0.511	91.13	0.626	86.98	434.19	334.48	$\text{MW} = 334.48 + 434.19\,\text{MDD}(G)$
14	$\chi_c(G)$	0.017	129.24	0.001	142.11	67.81	478.11	$\text{MW} = 478.11 + 67.81\,\chi_c(G)$
15	$E(G)$	0.905	40.19	0.959	28.86	473.13	264.88	$\text{MW} = 264.88 + 473.13\,E(G)$
16	$EC(G)$	0.787	60.15	0.865	52.27	541.05	245.60	$\text{MW} = 245.60 + 541.05\,EC(G)$

Table 3.5: Regression Equations and Performance Metrics for 16 models

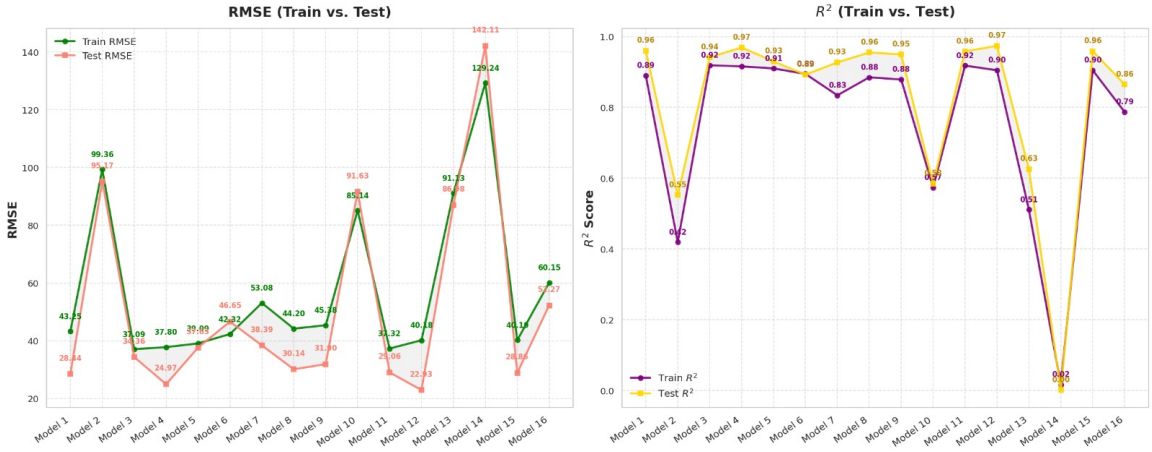


Figure 3.7: Comparison of Metrics for Training and Testing Sets across 16 Models

Figure 3.8 shows the plots of the fitted regressions for all sixteen univariate models jointly as subplots of a single figure. For each pair of subplots, the one on the left represents the data used for training with the fitted line of the constructed regression, while the one on the right represents the data used for testing, which refers to the same model. The noticeable observation from this figure is that the performance of Model 14, which corresponds to the molecular connectivity index, is not satisfactory on both sets of data, as there is a large deviation from the line of the constructed regression.

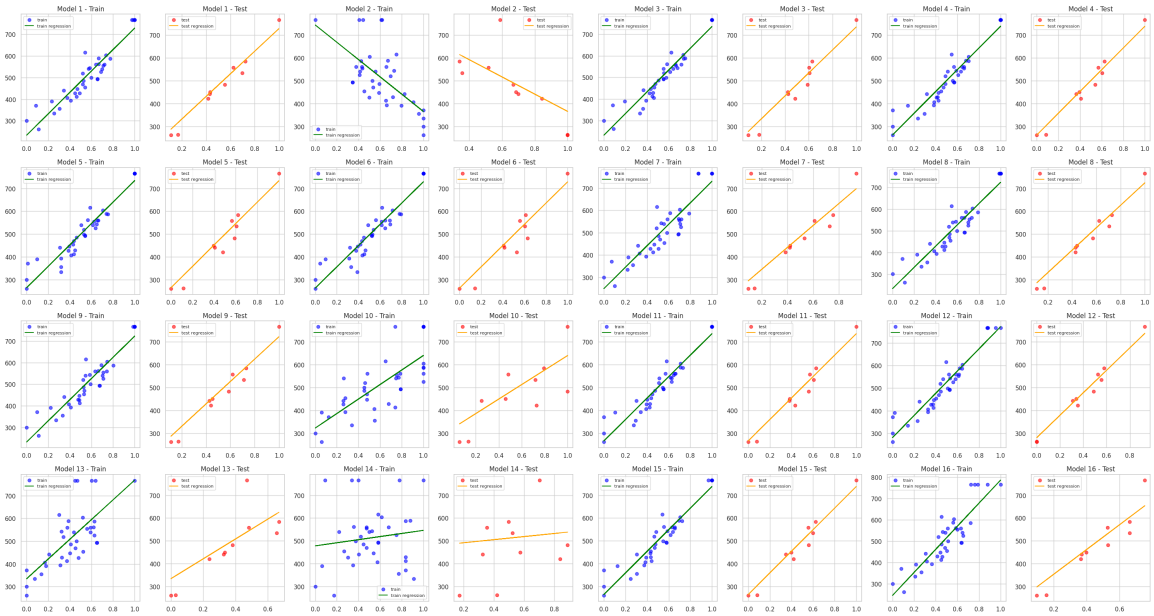


Figure 3.8: Regression Fits of 16 Models

Figure 3.9 shows the residual plots of all sixteen models. The distribution of the residuals is plotted on this figure. The residual plots help identify the adequacy of models. The residual plots will help verify the assumptions of linear models. If the scatter of the residual points on the plots seems random, then it will verify that the models fit well. However, there will be doubts regarding the models' accuracy. As seen from this Figure 3.9, most models seem perfectly accurate. Only a few models like Model 14 seem to show deviations. This will help verify the accuracy of those models' predictions.

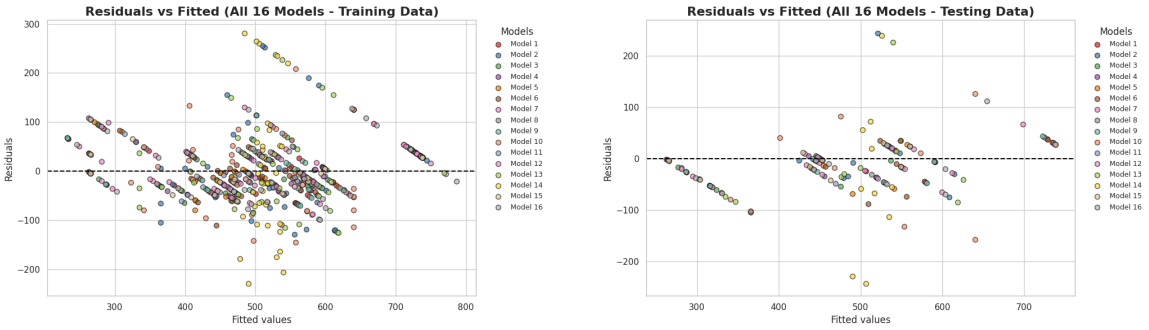


Figure 3.9: Residual Analysis for Train and Test Sets

Multivariate Linear Regression (MLR) and Evaluation for MW

The following Table 3.6 provides a comparative outline of multiple regression models created on various chemical descriptors in an attempt to predict molecular weights (MW). Models encompassed a gamut of techniques including multiple linear regression (MLR), feature-selction multiple linear regression (FS-MLR), ridge, LASSO, and Elastic Net methodologies which make use of varied approaches to handle features.

Also, though the conventional approach of quantitative molecular property prediction using QSPR concepts is often susceptible to multicollinearity among descriptors, thus leading to an inflation of coefficients of MW-predicting models, making them less accurate, employing both variable handling techniques of feature selection and regularization by models including ridge, LASSO, or Elastic Net yields significant improvement.

Model	Intercept	Regression Equation / Coefficients
MLR (All Indices)	257.666	$MW = 257.666 + 8294.824 \cdot WI + 17.839 \cdot BI - 1469.665 \cdot HI + 2433.777 \cdot RI - 4286.462 \cdot Z_1 + 2688.069 \cdot Z_2 - 1721.604 \cdot HWI - 14205.874 \cdot SI + 8378.970 \cdot GI + 31.033 \cdot HoI + 2802.521 \cdot EI + 432.995 \cdot SE - 319.134 \cdot MDD - 97.593 \cdot MC - 3268.598 \cdot GE + 722.806 \cdot EC$
MLR (Selected Features)	724.343	$MW = 724.343 - 979.283 \cdot WI - 455.889 \cdot BI - 1627.619 \cdot Z_1 + 323.400 \cdot Z_2 + 926.833 \cdot SI - 65.844 \cdot HoI + 2863.886 \cdot EI - 775.446 \cdot MDD - 1097.205 \cdot GE + 416.889 \cdot EC$
MLR (PCA-based)	513.381	$MW = 513.381 + 0.003352 \cdot PC_1 - 0.000253 \cdot PC_2 - 0.007914 \cdot PC_3 - 0.042753 \cdot PC_4$
Ridge ($\alpha = 0.01$)	299.650	$MW = 299.650 + 116.789 \cdot WI - 25.290 \cdot BI + 24.534 \cdot HI - 16.797 \cdot RI + 37.693 \cdot Z_1 + 121.854 \cdot Z_2 + 31.986 \cdot HWI + 106.404 \cdot SI + 88.292 \cdot GI - 70.783 \cdot HoI + 120.003 \cdot EI + 192.717 \cdot SE - 182.633 \cdot MDD - 34.820 \cdot MC - 111.273 \cdot GE - 33.291 \cdot EC$
LASSO ($\alpha = 0.1$)	277.603	$MW = 277.603 + 244.694 \cdot WI + 125.625 \cdot Z_2 + 7.976 \cdot SI - 61.354 \cdot HoI + 116.050 \cdot EI + 162.137 \cdot SE - 141.559 \cdot MDD - 33.760 \cdot MC$
Elastic Net ($\alpha = 1, l1\_ratio = 1$)	273.632	$MW = 273.632 + 165.452 \cdot Z_2 + 307.361 \cdot EI - 28.416 \cdot HoI$

Table 3.6: Comparison of Regression Models, Coefficients, and Equations

Table 3.7 captures the comparative performance of the six models, i.e., Multiple Linear Regression (MLR), feature selected-MLR (MLR-FS), Principal Component Analysis-MLR (MLR-PCA), Ridge, LASSO, Elastic Net in predicting the molecular weight (MW). The parameters include the average error measured as Mean Absolute Error (MAE), Coefficient of Determination denoted by R^2, and the root mean square error measured as Root Mean Squared Error (RMSE), all of which apply to both training sets and testing sets. From the table, it can be seen that the best-performing model on the criteria of accuracy or overall excellence in both measures of error (RMSE & MAE) is the (MLR-FS), which captures the highest R^2 of 0.971 on the testing data with an RMSE of 24.30. The models that perform remarkably well on this table include the three models i.e., Ridge, LASSO, and Elastic Net, which tend to generalize better than their non-regularized counterpart, the ordinary MLR. The use of FS, dimensionality reduction, or regularization strategies impacts positively on all models of regressions.

Metric	Test						Train					
	MLR	MLR w/ FS	MLR w/ PCA	Ridge	LASSO	ElasticNet	MLR	MLR w/ FS	MLR w/ PCA	Ridge	LASSO	ElasticNet
MAE	26.661	19.228	22.129	22.480	21.764	26.582	21.085	21.163	24.173	22.126	22.814	28.774
R^2-SCORE	0.955	0.971	0.954	0.960	0.963	0.954	0.958	0.949	0.940	0.943	0.942	0.926
RMSE	30.209	24.297	30.630	28.424	27.509	30.455	26.841	29.538	31.977	31.013	31.415	35.475

Table 3.7: Comparison of Model Performance Metrics on Test and Train Sets

Figure 3.10 compare of model performance metrics (MAE, R^2-score and RMSE) across

both Test and Train sets. The visualization highlights differences in predictive accuracy and generalization among regression models.

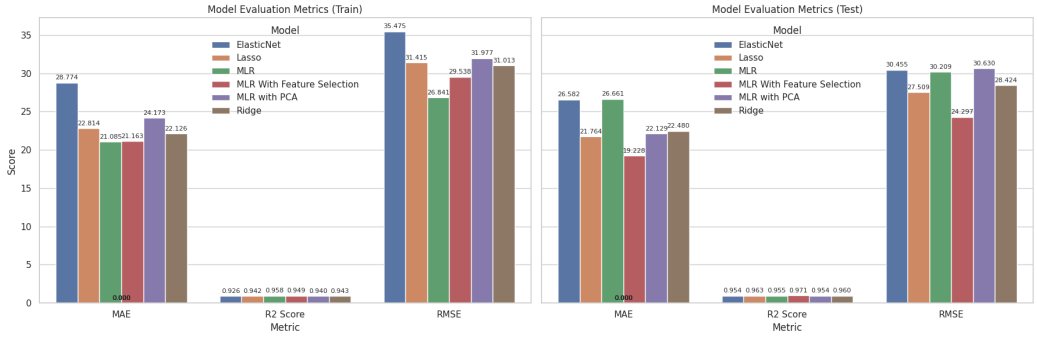


Figure 3.10: Comparison of Model Performance Metrics

Figure 3.11 illustrates residual analyses for six models on training and test datasets. Residuals are plotted against fitted values with each model represented by a distinct color. The left panel shows training data and the right panel shows test data. Ideally residuals should be randomly distributed without any pattern indicating good model fit.

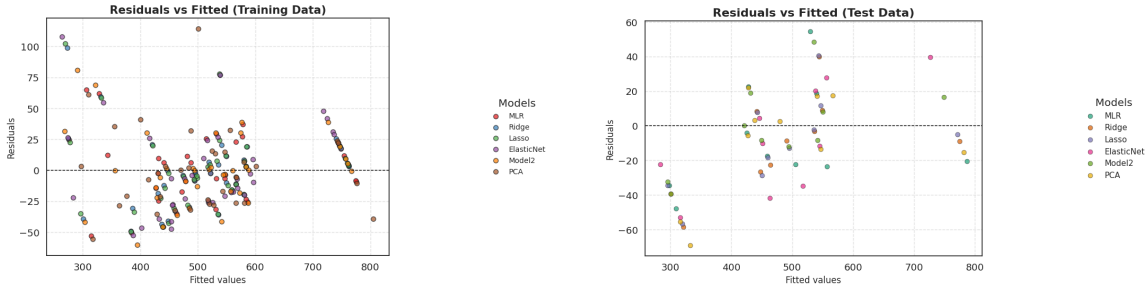


Figure 3.11: Residual Plots

Figure 3.12 shows how each feature contributes to model predictions at the instance level with rows as features and columns as data points. The blue to red gradient marks direction where blue lowers predictions, red raises them and darker shades mean stronger influence. Feature importance is obtained by summing SHAP values across instances visualized as a bar chart on the right.The SHAP heatmap analysis revealed distinct feature emphasis patterns across different regression models. The General Multiple Linear Regression (MLR) method was found to focus on the Schultz, Gutman, Wiener, and Zagreb First indices values highly owing to their significant relationship with molecular property variation. The Ridge regression method was found significant in terms of molecular complexity

and global connectivity properties, which it could capture effectively by giving priority to the Shannon Entropy, Mean Distance Degree, Zagreb Second, Estrada, Wiener, and Graph Energy indices. The LASSO Regression method was found significant in terms of its focus on molecular complexity properties by giving priority to the Wiener, Shannon Entropy, Zagreb Second, Mean Distance Degree, and Estrada indices. The Elastic Net Regression method was significant in giving priority to both local and global molecular properties by focusing on the Estrada and Harary indices. From all of these findings, it can be observed that there is a combination of different topological indexes emphasized upon by various models, whereby the Wiener, Shannon Entropy, and Zagreb indices always seem significant.

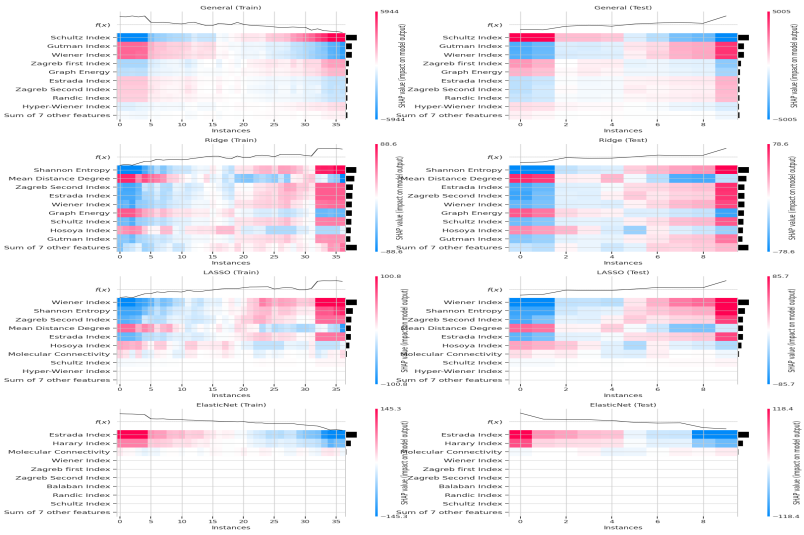


Figure 3.12: Feature Importance Analysis Using SHAP Heatmap

Figure 3.13 shows the results of sequential backward feature selection. The left panel lists features removed at each step with performance metrics while the right panel plots model accuracy versus number of features. It highlights the optimal feature subset that balances predictive performance and model simplicity.

Step	Num. Features	Selected Features	Avg. Score	CI Bound	Std. Dev	Std. Err	Obs.
1	1	Harary Index	0.887	0.074	0.057	0.029	404
2	2	Harary Index, Hosoya Index	**0.901**	0.071	0.056	0.028	404
3	3	Harary Index, Zagreb Second Index, Hosoya Index	0.896	0.069	0.054	0.027	404
4	4	Harary Index, Zagreb Second Index, Hosoya Index, Eccentric Connectivity	0.890	0.070	0.054	0.027	404
5	5	Harary Index, Zagreb Second Index, Hosoya Index, Mean Distance Degree, Eccentric Connectivity	0.887	0.070	0.055	0.027	404
6	6	Balaban Index, Harary Index, Zagreb Second Index, Hosoya Index, Mean Distance Degree, Eccentric Connectivity	0.884	0.098	0.076	0.038	404
7	7	Balaban Index, Harary Index, Zagreb Second Index, Gutman Index, Hosoya Index, Mean Distance Degree, Eccentric Connectivity	0.871	0.108	0.084	0.042	404
8	8	Balaban Index, Harary Index, Zagreb Second Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Eccentric Connectivity	0.861	0.108	0.084	0.042	404
9	9	Balaban Index, Harary Index, Randić Index, Zagreb Second Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Eccentric Connectivity	0.874	0.098	0.076	0.038	404
10	10	Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Eccentric Connectivity	0.869	0.096	0.075	0.037	404
11	11	Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Eccentric Connectivity	0.861	0.089	0.069	0.035	404
12	12	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Eccentric Connectivity	0.843	0.071	0.056	0.028	404
13	13	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Graph Energy, Eccentric Connectivity	0.838	0.079	0.061	0.031	404
14	14	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Graph Energy, Eccentric Connectivity	0.752	0.182	0.142	0.071	404
15	15	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	0.579	0.352	0.274	0.137	404
16	16	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	0.314	0.602	0.468	0.234	404

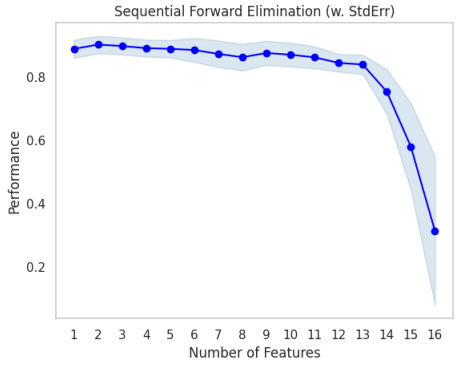


Figure 3.13: Forward Feature Selection

In Figure 3.14 scree plot shows how cumulative variance increases with more PCA components. At 4 components about 90% of the variance is captured marking the optimal point. Adding more components beyond this gives little extra information balancing simplicity and accuracy.

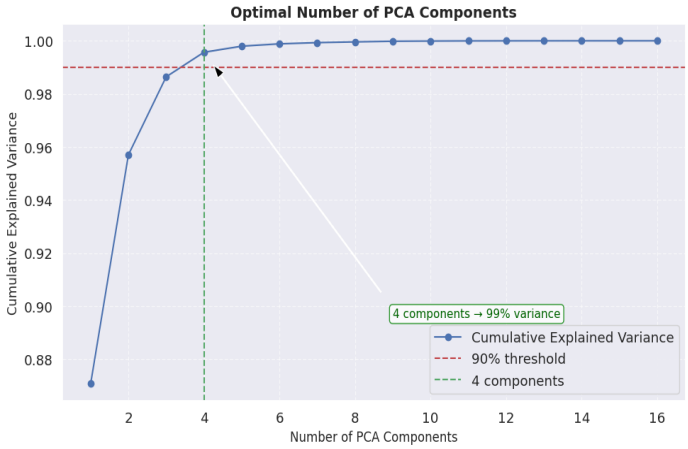
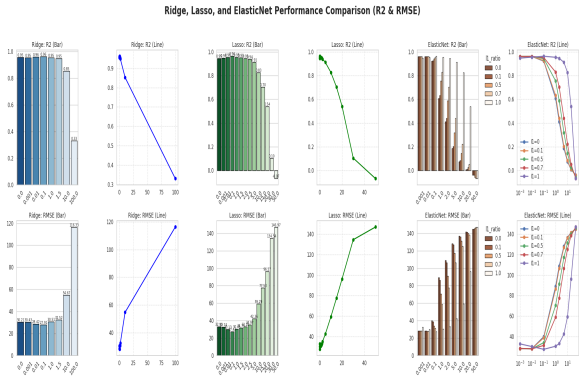
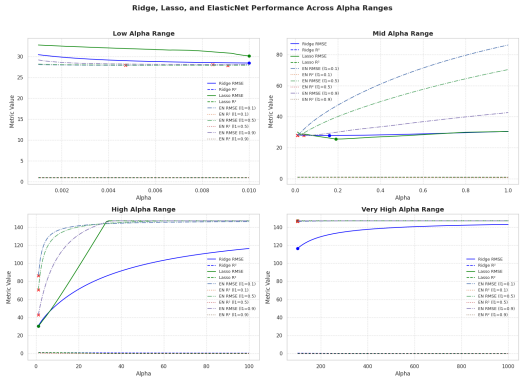


Figure 3.14: Cumulative Explained Variance Plot

Figure 3.15a compare Ridge, LASSO, and Elastic Net using R^2-score and RMSE across datasets. Ridge shrinks coefficients smoothly without setting them to zero, LASSO forces some to zero for feature selection and Elastic Net balances both. Figure 3.15b shows performance across alpha ranges where low alpha keeps models flexible while high alpha causes strong shrinkage. Overall Figure 3.15 illustrate that Ridge stays consistent, LASSO becomes sparse at high alpha and ElasticNet blends both behaviors showing how model choice and regularization directly affect accuracy and error.



(a) Regularization Techniques Comparison on Different Alpha Values



(b) Regularization Performance across Alpha Ranges

Figure 3.15: Model Performance across Different Regularization Techniques

3.0.2 Linear Models Fitting for LogP

In this section we have fitted various linear models to predict LogP values of the drugs based on topological indices and evaluated the models based on root mean squared error (RMSE) and R^2-score. We have used these metrics values on train and test data sets both to avoid the possibility of overfitting, data leakage, or over-regularization.

Univariate Regression Analysis for LogP

Table 3.8 presents the regression equation for univariate linear regression, where each model is trained on a single scaled input feature to predict the LogP. Feature importance can be assessed based on the magnitude of the slope (coefficient) values, which indicate the extent to which each feature contributes to the prediction. Table include various metrics and highlights how each descriptor contributes to LogP prediction and the reliability of its corresponding model. Although the overall performance of Model 2 was weak it still performed better relative to the other models while Model 14 exhibited the worst results.

Model	Feature	R^2_{Train}	$\text{RMSE}_{\text{Train}}$	R^2_{Test}	$\text{RMSE}_{\text{Test}}$	Slope	Intercept	Equation
1	$W(G)$	0.303	1.3822	0.291	1.7136	3.6670	1.2128	$\text{LogP} = 1.2128 + 3.6670\,W(G)$
2	$J(G)$	0.334	1.3514	0.437	1.5263	-4.2942	5.9033	$\text{LogP} = 5.9033 - 4.2942\,J(G)$
3	$H(G)$	0.246	1.4377	0.229	1.7867	3.3074	1.4396	$\text{LogP} = 1.4396 + 3.3074\,H(G)$
4	$\chi(G)$	0.287	1.3981	0.262	1.7480	3.3870	1.5018	$\text{LogP} = 1.5018 + 3.3870\,\chi(G)$
5	$M_1(G)$	0.240	1.4436	0.241	1.7732	3.0609	1.6659	$\text{LogP} = 1.6659 + 3.0609\,M_1(G)$
6	$M_2(G)$	0.206	1.4756	0.215	1.8033	2.8366	1.7641	$\text{LogP} = 1.7641 + 2.8366\,M_2(G)$
7	$WW(G)$	0.311	1.3745	0.324	1.6731	3.7315	1.2366	$\text{LogP} = 1.2366 + 3.7315\,WW(G)$
8	$S(G)$	0.303	1.3825	0.292	1.7123	3.6558	1.2019	$\text{LogP} = 1.2019 + 3.6558\,S(G)$
9	$\text{Gut}(G)$	0.303	1.3827	0.292	1.7117	3.6567	1.1896	$\text{LogP} = 1.1896 + 3.6567\,\text{Gut}(G)$
10	$Z(G)$	0.181	1.4985	0.173	1.8503	2.2683	1.9229	$\text{LogP} = 1.9229 + 2.2683\,Z(G)$
11	$EE(G)$	0.264	1.4206	0.261	1.7494	3.2005	1.6008	$\text{LogP} = 1.6008 + 3.2005\,EE(G)$
12	$H_{\text{Sh}}(G)$	0.315	1.3711	0.297	1.7056	3.6786	1.5443	$\text{LogP} = 1.5443 + 3.6786\,H_{\text{Sh}}(G)$
13	$\text{MDD}(G)$	0.344	1.3418	0.434	1.5310	4.5233	1.4245	$\text{LogP} = 1.4245 + 4.5233\,\text{MDD}(G)$
14	$\chi_c(G)$	0.005	1.6524	-0.029	2.0645	-0.4475	3.5210	$\text{LogP} = 3.5210 - 0.4475\,\chi_c(G)$
15	$E(G)$	0.283	1.4025	0.259	1.7517	3.3613	1.5227	$\text{LogP} = 1.5227 + 3.3613\,E(G)$
16	$EC(G)$	0.325	1.3603	0.341	1.6525	4.4207	1.1003	$\text{LogP} = 1.1003 + 4.4207\,EC(G)$

Table 3.8: Regression Equations and Performance Metrics for 16 Models

Figure 3.16 plots the comparison of models on the task of predicting LogP using sixteen models. The plots on the left-hand side of Figure 3.16 provide a comparison of the training and testing datasets RMSE values. The plots on the right-hand side provide a comparison of the training and testing datasets values of R^2. Note that the test RMSE values (1.5–2.0) are always higher than the training RMSE values (1.3-1.6), which shows moderate generalization. The values of $(R^2 \approx 0.25 - 0.45)$ indicate poor linearity, suggesting that LogP demonstrates nonlinear behavior that is not fully depicted by the tested linear regression models. Model 14 has the poorest performance, which is shown by the highest RMSE and lowest R^2, confirming underfitting. These results collectively demonstrate that while linear regression frameworks provide baseline interpretability, LogP prediction likely requires more complex, nonlinear modeling attempts.

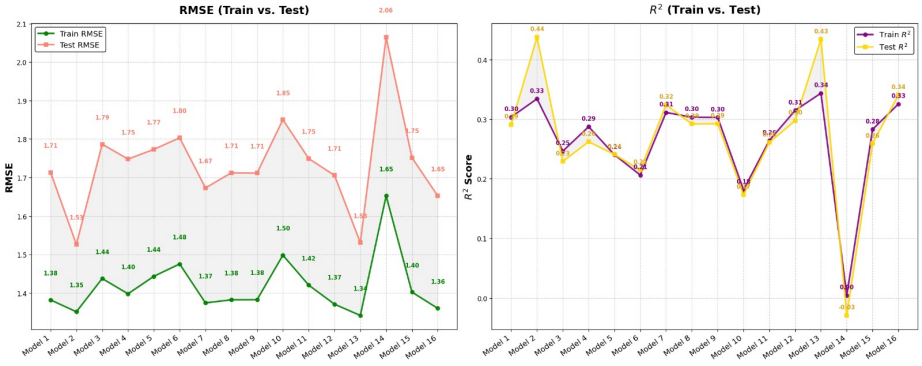


Figure 3.16: Comparison of Metrics for Training and Testing Sets across 16 Models

As seen in Figure 3.17 below, subplots are contained in another subplot where training data is fit into the line on the left subplot (blue color), with the right subplot (yellow color) containing data from the testing set.

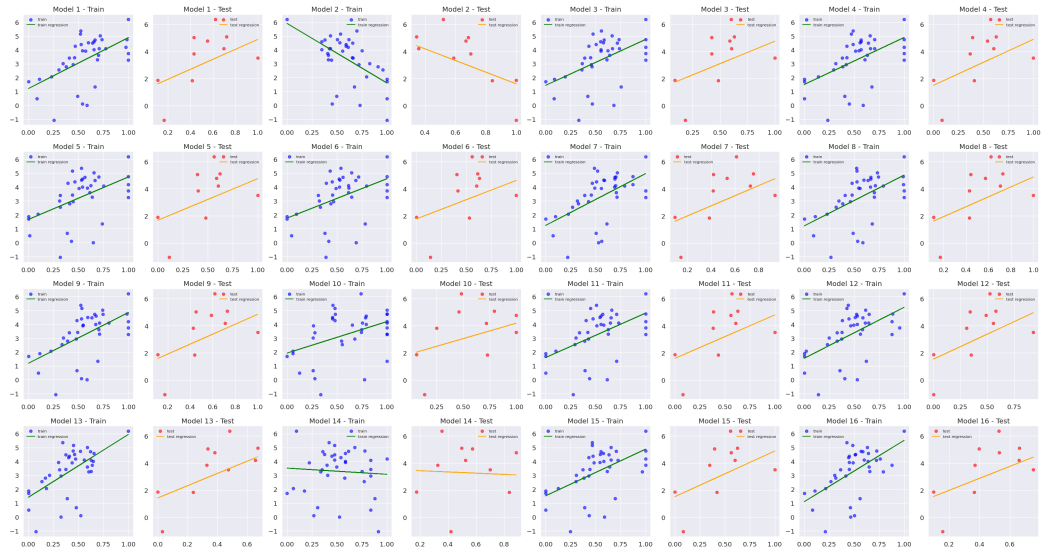


Figure 3.17: Regression Fits of 16 Models

Figure 3.18 shows the scatter plots of residuals on the y-axis with the fitted values on the x-axis, which help in checking the adequacy of the models.

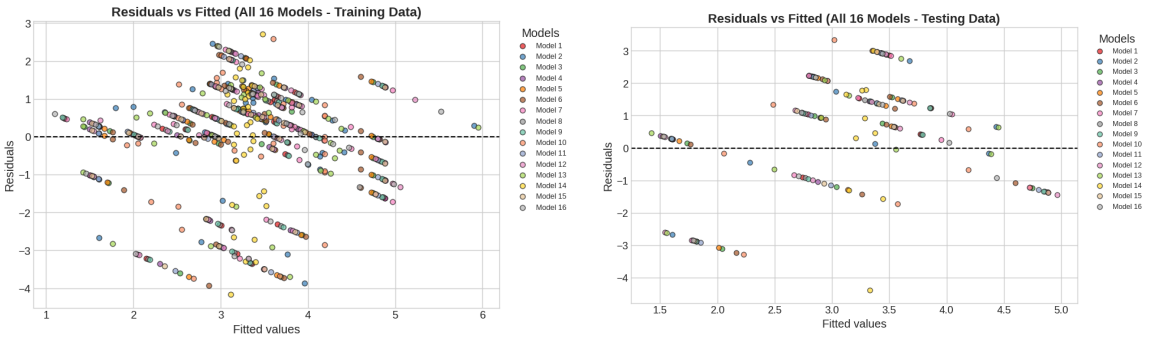


Figure 3.18: Residual Analysis for Train and Test Sets

Multivariate Linear Regression and Evaluation for LogP

The tables below, Table 3.9 provide comparisons of models on regression analyses used to predict LogP. The table comprises models of various complexity levels with different handling of variables. The Ridge, LASSO, or Elastic Net, which combines allows a deeper

understanding of molecular property prediction strategies. The models comprising MLR (all indices, selected features, PCA based) and regularized models like Ridge, LASSO, and ElasticNet offer experience on how molecular models can be approached.

Model	**Intercept**	**Regression Equation / Coefficients**
MLR (All Indices)	-21.720	$logP = -21.720 + 458.692 \cdot WI + 23.277 \cdot BI - 17.002 \cdot HI + 7.132 \cdot RI - 89.672 \cdot Z_1 + 8.168 \cdot Z_2 - 61.340 \cdot HWI - 915.019 \cdot SI + 533.790 \cdot GI + 1.262 \cdot HoI + 100.510 \cdot EI + 11.675 \cdot SE + 32.372 \cdot MDD - 2.167 \cdot MC - 40.939 \cdot GE - 0.668 \cdot EC$
MLR (Selected Features)	1.657	$logP = 1.657 - 14.823 \cdot Z_2 - 21.814 \cdot HWI + 22.247 \cdot GI + 15.188 \cdot SE + 3.935 \cdot MDD$
MLR (PCA-based)	3.288	$logP = 3.288 + 2.26 \times 10^{-5} \cdot PC_1 - 6.18 \times 10^{-5} \cdot PC_2$
Ridge ($\alpha = 0.1$)	2.856	$logP = 2.856 + 0.366 \cdot WI - 1.058 \cdot BI - 0.109 \cdot HI + 1.475 \cdot RI - 1.239 \cdot Z_1 - 3.492 \cdot Z_2 - 1.537 \cdot HWI + 0.661 \cdot SI + 0.911 \cdot GI + 1.158 \cdot HoI + 0.104 \cdot EI + 2.941 \cdot SE + 0.978 \cdot MDD - 0.871 \cdot MC + 1.176 \cdot GE - 0.168 \cdot EC$
LASSO ($\alpha = 0.01$)	3.400	$logP = 3.400 - 1.652 \cdot BI - 2.689 \cdot Z_2 + 0.428 \cdot HoI + 4.783 \cdot SE - 0.358 \cdot MC$
Elastic Net ($\alpha = 0.02, l1\_ratio = 1$)	3.838	$logP = 3.838 - 2.229 \cdot BI + 1.902 \cdot SE - 0.181 \cdot MC$

Table 3.9: Comparison of Regression Models, Coefficients and Equations for LogP Prediction

Comparison of various metrics like MAE, R^2-score, and RMSE values on the test set and train set among different models like MLR, MLR(Selected Features), MLR(PCA Based), Ridge, LASSO, and ElasticNet is provided in Table 3.10. LASSO Regression has the lowest MAE and RMSE values on the test set, while the best fit is achieved by MLR on the train set, which can cause overfitting. The models of Ridge Regression and ElasticNet provide a stable fit on both sets. The effect of selected features on accuracy is presented by both feature selection and PCA.

Metric	Test						Train					
	ElasticNet	LASSO	MLR	MLR w/ FS	MLR w/ PCA	Ridge	ElasticNet	LASSO	MLR	MLR w/ FS	MLR w/ PCA	Ridge
MAE	1.299	1.292	1.490	1.373	1.514	1.272	0.978	0.947	0.712	0.807	1.071	0.941
R^2-SCORE	0.401	0.411	0.268	0.390	0.197	0.406	0.369	0.416	0.645	0.544	0.286	0.436
RMSE	1.574	1.562	1.741	1.589	1.824	1.568	1.315	1.265	0.987	1.118	1.399	1.243

Table 3.10: Comparison of Model Performance Metrics on Test and Train Sets

Figure 3.19 compare the values of the metrics of models (MAE, R^2 score, and RMSE) accuracy on both the Test set and Train datasets.

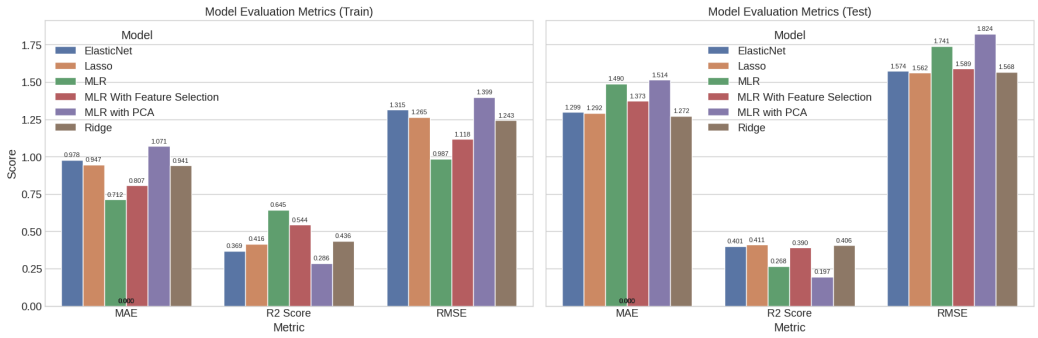


Figure 3.19: Comparison of Model Performance Metrics

Figure 3.20 shows the residual analyses of training and testing data performed on six models. Residuals are plotted against fitted values with each model represented by a distinct color. The residual plots show that Ridge and LASSO produce tighter clustering around zero compared to other models indicating stronger fits on training data and better generalization on testing data. Residuals spread more widely in the test set. Overall Ridge and LASSO stand out as more reliable though heteroscedasticity and outliers suggest areas for refinement.

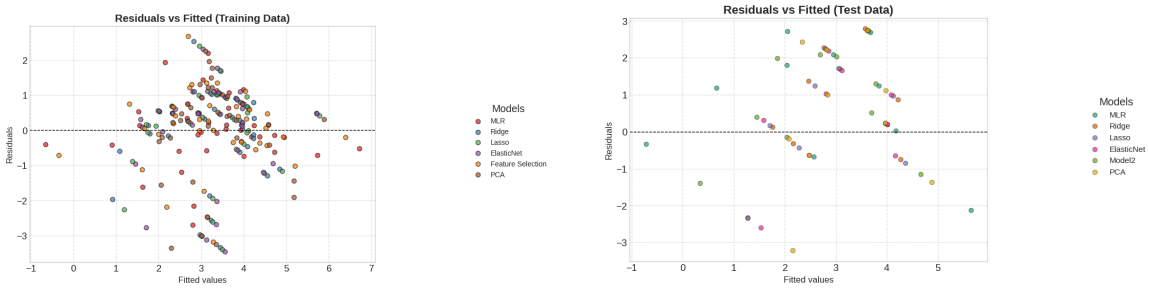


Figure 3.20: Residual Plots

In Figure 3.21 SHAP plots for General(MLR), Ridge, LASSO, and ElasticNet explain how different features influence predictions on both training and test sets. The analysis revealed that the MLR model emphasized the Schultz, Gutman, and Wiener indices, while the Ridge model assigned greater importance to the Zagreb Second, Shannon Entropy, and Hosoya indices. The LASSO and Elastic Net models highlighted the Balaban and Shannon Entropy indices as dominant contributors.

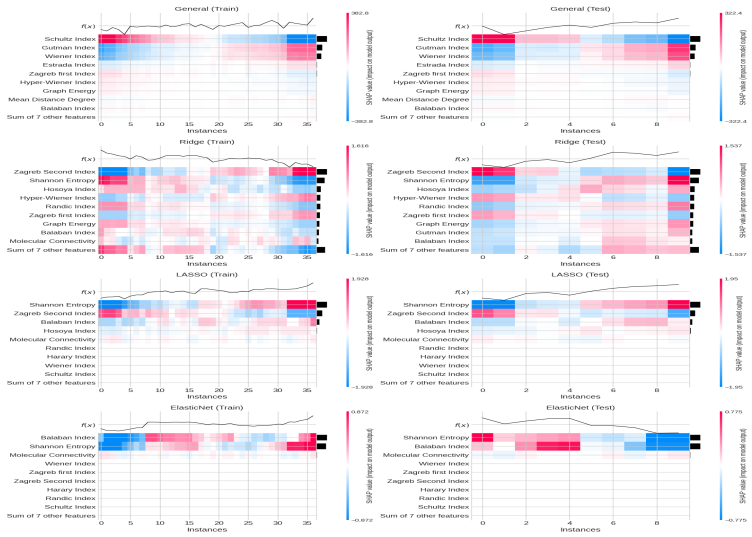


Figure 3.21: Feature Importance Analysis Using SHAP Values

Figure 3.22 shows the results of sequential forward feature selection. The left panel shows that adding features improves model performance but in later steps extra features cause redundancy or overfitting lowering the score. It highlights the optimal feature subset that balances predictive performance and model simplicity.

Step	Num. Features	Selected Features	Avg. Score	CI Bound	Std. Dev	Std. Err	Obs.
1	1	Mean Distance Degree	0.215	0.241	0.188	0.094	404
2	2	Shannon Entropy, Mean Distance Degree	0.224	0.223	0.173	0.087	404
3	3	Zagreb Second Index, Shannon Entropy, Mean Distance Degree	0.243	0.412	0.321	0.160	404
4	4	Zagreb Second Index, Gutman Index, Shannon Entropy, Mean Distance Degree	0.248	0.402	0.313	0.156	404
5	5	Zagreb Second Index, Hyper-Wiener Index, Gutman Index, Shannon Entropy, Mean Distance Degree	0.269	0.488	0.380	0.190	404
6	6	Zagreb Second Index, Hyper-Wiener Index, Gutman Index, Estrada Index, Shannon Entropy, Mean Distance Degree	0.257	0.549	0.427	0.214	404
7	7	Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Estrada Index, Shannon Entropy, Mean Distance Degree	0.264	0.515	0.401	0.200	404
8	8	Randić Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Estrada Index, Shannon Entropy, Mean Distance Degree	0.266	0.524	0.408	0.204	404
9	9	Randić Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Eccentric Connectivity	0.214	0.641	0.498	0.249	404
10	10	Randić Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Eccentric Connectivity	0.134	0.402	0.312	0.156	404
11	11	Randić Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Eccentric Connectivity	0.044	0.367	0.285	0.143	404
12	12	Balaban Index, Randić Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Eccentric Connectivity	-0.159	0.906	0.705	0.353	404
13	13	Balaban Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Eccentric Connectivity	-0.353	0.877	0.683	0.341	404
14	14	Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Eccentric Connectivity	-0.486	0.938	0.730	0.365	404
15	15	Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-0.760	1.178	0.917	0.458	404
16	16	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb First Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-2.330	3.217	2.503	1.251	404

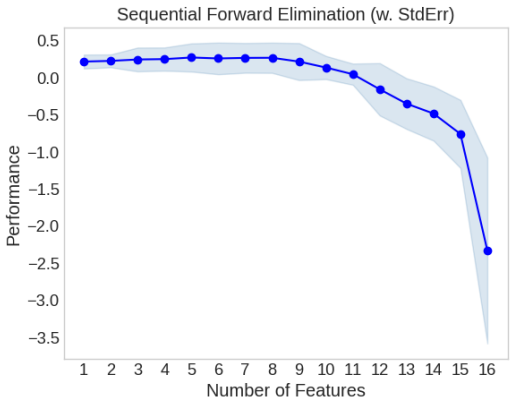


Figure 3.22: Forward Feature Selection

Figure 3.23 shows cumulative explained variance versus the number of components highlighting that two components capture 95% of the dataset's variance. The blue line traces variance explained as components are added and the red dashed line marks the threshold. This shows that retaining two components effectively reduces dimensionality while preserving most information balancing accuracy and efficiency.

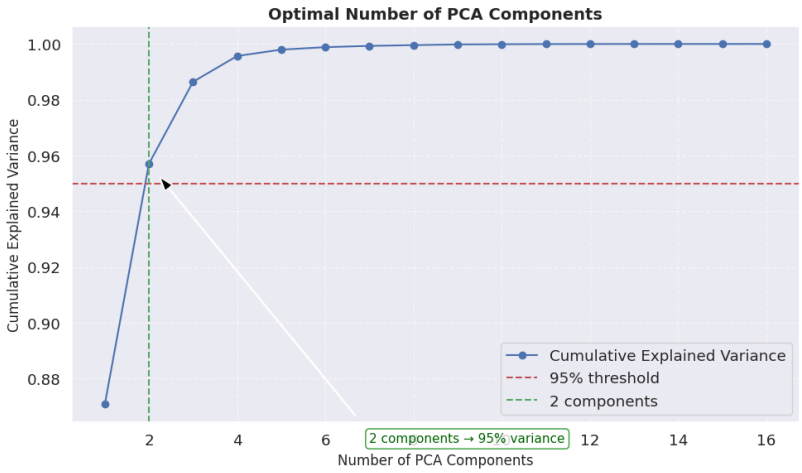
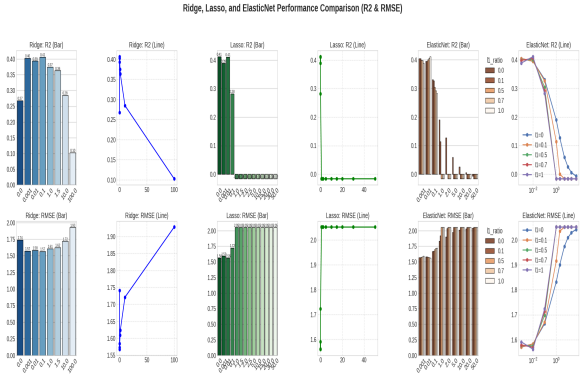
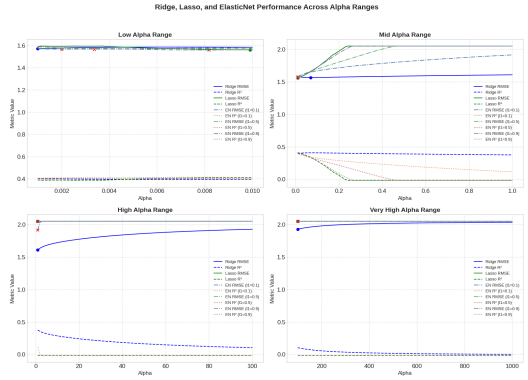


Figure 3.23: Cumulative Explained Variance Plot

Figure 3.24a compare Ridge, LASSO, and ElasticNet using R^2-score and RMSE across datasets. Ridge shrinks coefficients smoothly without setting them to zero, LASSO forces some to zero for feature selection and ElasticNet balances both. Figure 3.24b shows performance across alpha ranges where low alpha keeps models flexible while high alpha causes strong shrinkage. Overall Figure 3.24 illustrate that Ridge stays consistent, LASSO becomes sparse at high alpha and ElasticNet blends both behaviors.



(a) Regularization Techniques Comparison on Different Alpha Values



(b) Regularization Performance across Alpha Ranges

Figure 3.24: Model Performance across Different Regularization Techniques

3.0.3 Linear Models Fitting for H-Bond Donnor (HBD)

In this section we have fitted various linear models to predict HBD values of the drugs based on topological indices and evaluated the models based on root mean squared error (RMSE) and R^2-score. We have used these metrics values on train and test data sets both to avoid the possibility of overfitting, data leakage, or over-regularization.

Univariate regression analysis for HBD

Table 3.11 presents the regression equation for univariate linear regression, where each model is trained on a single scaled input feature to predict HBD. Table 3.11 also provide slopes, intercepts, and various evaluation metrics providing insight into how well each feature explains the variability in HBD. Figure 3.25 presents charts of RMSE and R^2 scores across 16 models. No model perfom well because the relationship between input feature and output column is not linear as shown in 3.26.

Model	Feature	R^2_{Train}	$\text{RMSE}_{\text{Train}}$	R^2_{Test}	$\text{RMSE}_{\text{Test}}$	Slope	Intercept	Equation
1	$W(G)$	0.060	1.3137	0.172	0.6808	1.3379	1.5603	$\text{HBD} = 1.5603 + 1.3379\,W(G)$
2	$J(G)$	0.018	1.3433	0.092	0.7130	-0.8042	2.8073	$\text{HBD} = 2.8073 - 0.8042\,J(G)$
3	$H(G)$	0.071	1.3064	0.146	0.6917	1.4498	1.5072	$\text{HBD} = 1.5072 + 1.4498\,H(G)$
4	$\chi(G)$	0.065	1.3107	0.151	0.6896	1.3138	1.6246	$\text{HBD} = 1.6246 + 1.3138\,\chi(G)$
5	$M_1(G)$	0.070	1.3072	0.124	0.7004	1.3486	1.6028	$\text{HBD} = 1.6028 + 1.3486\,M_1(G)$
6	$M_2(G)$	0.069	1.3075	0.107	0.7072	1.3450	1.5949	$\text{HBD} = 1.5949 + 1.3450\,M_2(G)$
7	$WW(G)$	0.057	1.3160	0.167	0.6830	1.3071	1.5989	$\text{HBD} = 1.5989 + 1.3071\,WW(G)$
8	$S(G)$	0.056	1.3168	0.163	0.6846	1.2846	1.5845	$\text{HBD} = 1.5845 + 1.2846\,S(G)$
9	$\text{Gut}(G)$	0.052	1.3195	0.154	0.6882	1.2400	1.6059	$\text{HBD} = 1.6059 + 1.2400\,\text{Gut}(G)$
10	$Z(G)$	0.000	1.3551	-0.029	0.7591	-0.0538	2.3500	$\text{HBD} = 2.3500 - 0.0538\,Z(G)$
11	$EE(G)$	0.067	1.3092	0.133	0.6967	1.3161	1.6237	$\text{HBD} = 1.6237 + 1.3161\,EE(G)$
12	$H_{\text{Sh}}(G)$	0.055	1.3171	0.118	0.7028	1.2634	1.7186	$\text{HBD} = 1.7186 + 1.2634\,H_{\text{Sh}}(G)$
13	$\text{MDD}(G)$	0.022	1.3405	0.099	0.7104	0.9280	1.9352	$\text{HBD} = 1.9352 + 0.9280\,\text{MDD}(G)$
14	$\chi_c(G)$	0.029	1.3353	-0.053	0.7680	-0.9290	2.8008	$\text{HBD} = 2.8008 - 0.9290\,\chi_c(G)$
15	$E(G)$	0.055	1.3175	0.135	0.6961	1.2121	1.6809	$\text{HBD} = 1.6809 + 1.2121\,E(G)$
16	$EC(G)$	0.045	1.3244	0.131	0.6974	1.3441	1.6523	$\text{HBD} = 1.6523 + 1.3441\,EC(G)$

Table 3.11: Regression Equations and Performance Metrics for 16 Models Predicting HBD

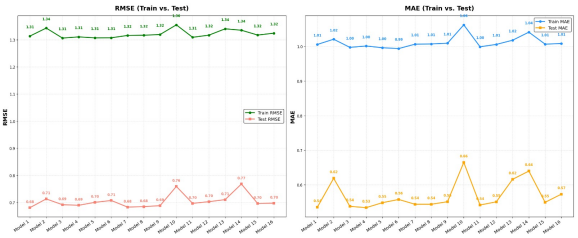


Figure 3.25: Comparison of Metrics for Training and Testing Sets

Figure 3.26 presents subplots within a larger subplot where the left (blue) panel shows the training data fitted to the regression line and the right (red) panel shows the corresponding test data.

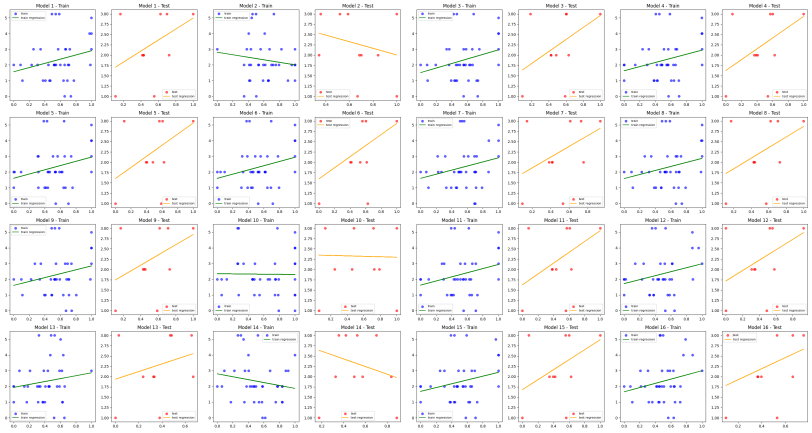


Figure 3.26: Regression Fits of 16 Models

Figre 3.27 shows the residual plots for 16 models. Test residuals are more spread out with noticeable outliers reflecting prediction errors on unseen data and potential heteroscedasticity. Overall the plots highlight model performance, generalization, and areas for improvement helping identify issues like non linearity and heteroscedasticity.

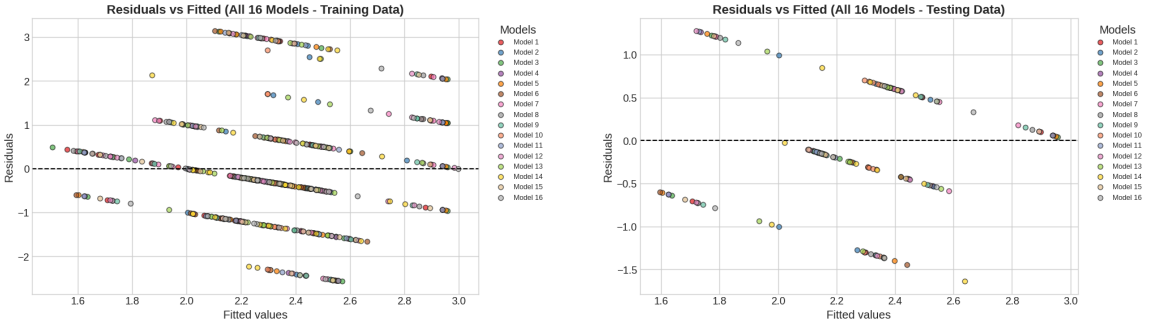


Figure 3.27: Residual Analysis for Train and Test Sets

Multivariate Linear Regression and Evaluation for HBD

The comparison of various models of regression developed for the prediction of HBD is given in Table 3.12 below. This table provides a comparison of various models of regression, which can be classified into various types based on complexity (e.g., simple M-LR models, models involving selected variables, PCA models, models involving ridge,

LASSO, and Elastic Net regression). Overall it illustrates the trade off between model complexity, interpretability, and predictive performance while identifying the most impactful predictors.

Model	**Intercept**	**Regression Equation / Coefficients**
MLR (All Indices)	21.293	$HBD = 21.293 - 462.117 \cdot WI - 19.361 \cdot BI + 11.392 \cdot HI + 61.307 \cdot RI + 95.238 \cdot Z_1 - 20.287 \cdot Z_2 + 43.269 \cdot HWI + 807.356 \cdot SI - 403.173 \cdot GI - 3.911 \cdot HoI - 78.390 \cdot EI - 3.428 \cdot SE - 26.470 \cdot MDD + 0.761 \cdot MC - 44.744 \cdot GE + 4.560 \cdot EC$
MLR (Selected Features)	2.439	$HBD = 2.439 + 5.036 \cdot HI + 6.594 \cdot Z_1 + 20.905 \cdot HWI - 25.945 \cdot GI - 6.131 \cdot EC$
MLR (PCA-based)	2.318	$HBD = 2.318 + 9.61 \times 10^{-6} \cdot PC_1 - 7.07 \times 10^{-5} \cdot PC_2 - 2.90 \times 10^{-5} \cdot PC_3$
Ridge ($\alpha = 0.1$)	0.753	$HBD = 0.753 + 0.295 \cdot WI + 1.009 \cdot BI + 0.472 \cdot HI - 0.101 \cdot RI + 1.909 \cdot Z_1 + 2.484 \cdot Z_2 + 1.725 \cdot HWI + 0.216 \cdot SI + 0.184 \cdot GI - 3.627 \cdot HoI + 0.856 \cdot EI - 1.211 \cdot SE - 0.292 \cdot MDD - 0.118 \cdot MC - 0.950 \cdot GE + 0.071 \cdot EC$
LASSO ($\alpha = 0.01$)	1.855	$HBD = 1.855 + 4.855 \cdot Z_2 - 3.566 \cdot HoI$
Elastic Net ($\alpha = 0.02, l1\_ratio = 0.9$)	1.919	$HBD = 1.919 + 1.401 \cdot ZFI + 2.083 \cdot ZSI - 2.375 \cdot HI - 0.064 \cdot MC$

Table 3.12: Comparison of Regression Models, Coefficients and Equations

Table 3.13 compares model performance metrics such as MAE, R^2-score, RMSE on test and train sets for various regression models including MLR, MLR(Selected Features), MLR(PCA-based), Ridge, LASSO, and ElasticNet.

Metric	Test						Train					
	ElasticNet	LASSO	MLR	MLR w/ FS	MLR w/ PCA	Ridge	ElasticNet	LASSO	MLR	MLR w/ FS	MLR w/ PCA	Ridge
MAE	0.458	0.437	0.657	0.364	0.402	0.458	0.906	0.896	0.715	0.764	1.004	0.881
R^2-SCORE	0.437	0.443	-0.536	0.248	0.500	0.406	0.259	0.306	0.508	0.472	0.135	0.330
RMSE	0.561	0.558	0.928	0.649	0.529	0.577	1.166	1.129	0.951	0.985	1.260	1.109

Table 3.13: Comparison of Model Performance Metrics on Test and Train Sets

Figure 3.28 compare of model performance metrics (MAE, R^2-score, and RMSE) across both Test and Train sets. The visualization highlights differences in predictive accuracy and generalization among regression models.

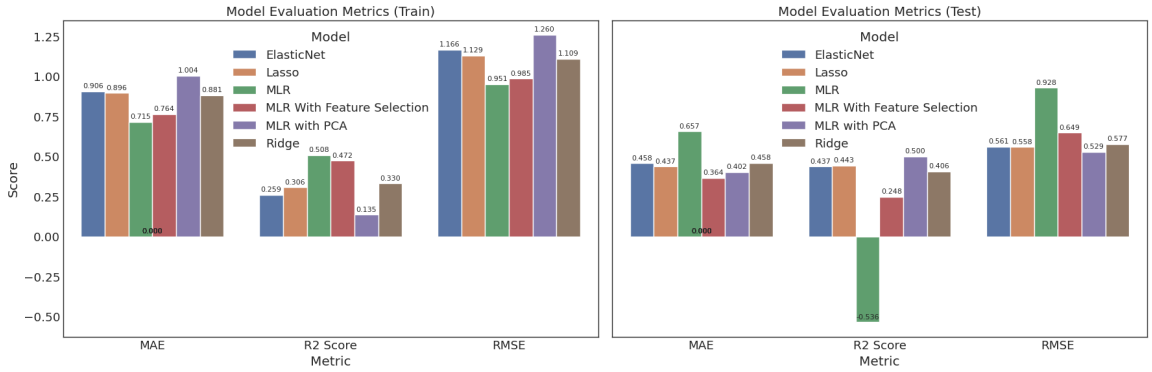


Figure 3.28: Comparison of model performance metrics

In Figure 3.29 residuals vs fitted plots for training and test data visualize prediction errors for models like MLR, Ridge, LASSO, ElasticNet, and PCA-based with residuals ideally scattered around zero for accurate predictions. Patterns or trends in residuals indicate issues like heteroscedasticity or poor generalization while models closer to the zero line perform better. These plots help assess model fit, bias, variance, and reliability on new data guiding selection or improvement of regression models.

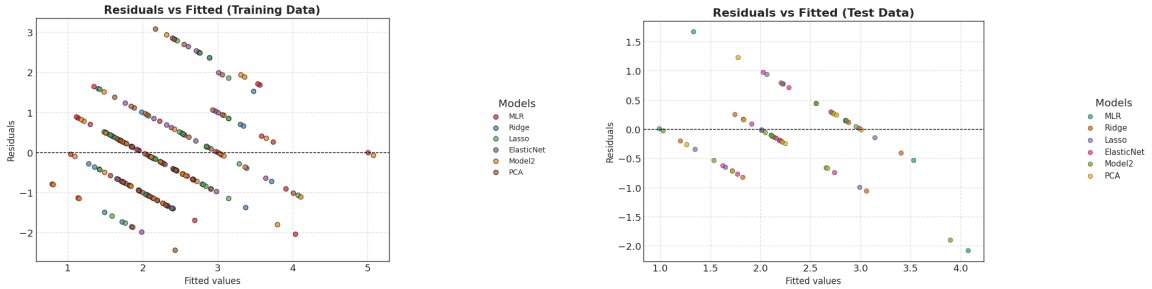


Figure 3.29: Residuals Plots

Figure 3.30 SHAP heatmap illustrate feature contributions for Ridge, LASSO, and Elastic-Net models on training and test data with red indicating strong positive and blue negative impacts and each bar representing an individual instance. The MLR model shows Schultz, Gutman, and Wiener index as highly influential. Ridge and ElasticNet highlights Zagreb Second , Zargeb first, and Hosoya index more prominent in the test set while LASSO emphasizes Hosoya, Zagreb Second, and Harary index.

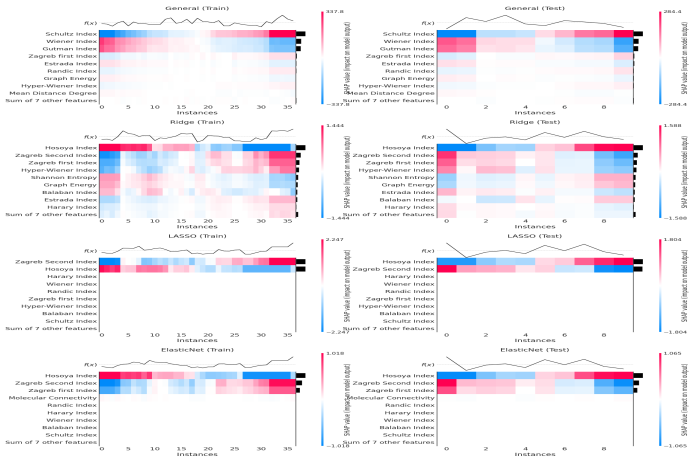


Figure 3.30: Feature Importance Analysis Using SHAP Values

Figure 3.31 shows the results of sequential backward feature selection. The left panel lists features removed at each step with performance metrics while the right panel plots model accuracy versus number of features. It highlights the optimal feature subset that balances predictive performance and model simplicity.

Step	Num. Features	Selected Features	Avg. Score	CI Bound	Std. Dev	Std. Err	Obs.
16	16	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-4.43206	9.24302	7.191384	3.595692	404
15	15	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Gutman Index, Hosoya Index, Estrada Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy	-1.87457	2.75297	2.141905	1.070952	404
14	14	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Hyper-Wiener Index, Schultz Index, Hosoya Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-1.67082	2.45818	1.912548	0.956274	404
13	13	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Hyper-Wiener Index, Hosoya Index, Shannon Entropy, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-1.34164	2.16295	1.68285	0.841425	404
12	12	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Hyper-Wiener Index, Hosoya Index, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-0.87096	1.27297	0.990412	0.495206	404
11	11	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Hyper-Wiener Index, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-0.69010	0.800458	0.622784	0.311392	404
10	10	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb first Index, Zagreb Second Index, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-0.32181	0.737244	0.573601	0.2868	404
9	9	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb Second Index, Mean Distance Degree, Molecular Connectivity, Graph Energy, Eccentric Connectivity	-0.10639	0.472136	0.367338	0.183669	404
8	8	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb Second Index, Molecular Connectivity, Graph Energy, Eccentric Connectivity	0.03215	0.437353	0.340275	0.170138	404
7	7	Wiener Index, Balaban Index, Harary Index, Randić Index, Zagreb Second Index, Graph Energy, Eccentric Connectivity	0.13116	0.480705	0.374005	0.187003	404
6	6	Balaban Index, Harary Index, Randić Index, Zagreb Second Index, Graph Energy, Eccentric Connectivity	0.09545	0.46591	0.362494	0.181247	404
5	5	Harary Index, Randić Index, Zagreb Second Index, Graph Energy, Eccentric Connectivity	0.08397	0.464776	0.361611	0.180806	404
4	4	Randić Index, Zagreb Second Index, Graph Energy, Eccentric Connectivity	0.04951	0.566762	0.44096	0.22048	404
3	3	Randić Index, Zagreb Second Index, Graph Energy	0.06355	0.399049	0.310474	0.155237	404
2	2	Randić Index, Graph Energy	-0.16814	0.460418	0.358221	0.17911	404
1	1	Randić Index	-0.21054	0.405399	0.315414	0.157707	404

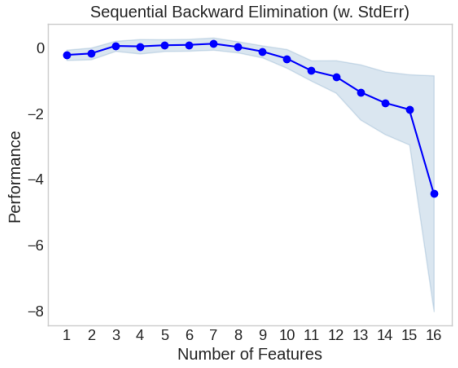


Figure 3.31: Backward Feature Selection

Figure 3.32 scree plot shows cumulative explained variance versus the number of PCA components illustrating how much total variance is captured as components are added. The first few components capture most variance with diminishing returns as more are included. A dashed red line marks the 95% variance threshold and a dashed green line highlights three components suggesting they suffice to capture approximately 95% of the variance.

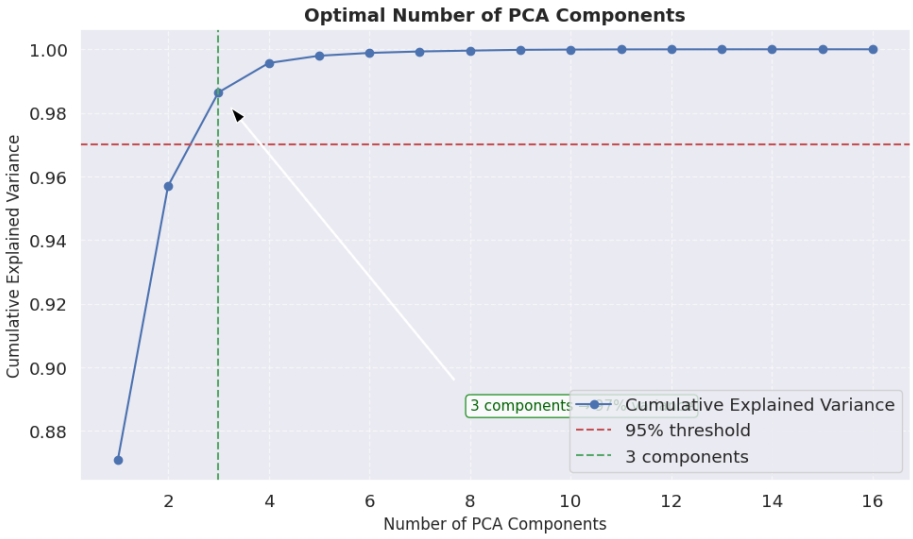
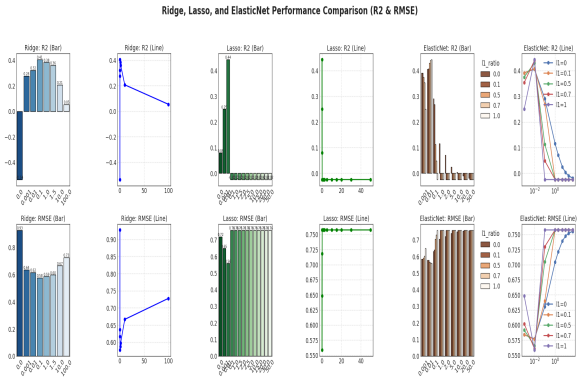
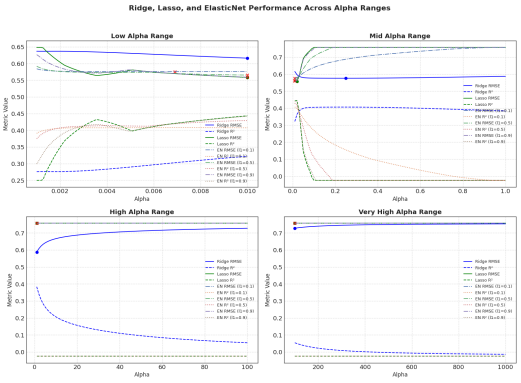


Figure 3.32: Cumulative Explained Variance Plot

Figure 3.33a compare Ridge, LASSO, and Elastic Net using R^2-score and RMSE across datasets. Ridge shrinks coefficients smoothly without setting them to zero, LASSO forces some to zero for feature selection and Elastic Net balances both. Figure 3.33b shows performance across alpha ranges where low alpha keeps models flexible while high alpha causes strong shrinkage. Overall Figure 3.33 illustrate that Ridge stays consistent, LASSO becomes sparse at high alpha and ElasticNet blends both behaviors.



(a) Regularization Techniques Comparison on Different Alpha Values



(b) Regularization Performance across Alpha Ranges

Figure 3.33: Model Performance across Different Regularization Techniques

3.0.4 Results and Analysis

Molecular Weight (MW): Linear models showed strong predictive performance for MW (test $R^2 \approx 0.96$). Shannon Entropy, Wiener, Zargeb First, Zargeb Second and Schultz indices emerged as dominant predictors. Regularized models reduced multicollinearity, and residuals exhibited random dispersion, confirming model adequacy.

Partition Coefficient (LogP): Models achieved moderate accuracy ($R^2 \approx 0.4$). The Balaban Index, Zagreb Indices, and Shannon Entropy contributed most to LogP prediction. Residual patterns suggested partial nonlinearity, implying future use of nonlinear models could improve performance.

Hydrogen Bond Donors (HBD): Linear predictability was low ($R^2 < 0.5$). HBD is governed by functional group chemistry rather than topology. While Zagreb First, Zargeb Second, Hosoya and Harary indices showed minor influence, future inclusion of substructure descriptors is recommended.

Protein protein interaction (PPI) networks are pivotal structures in systems biology. They model complex molecular interactions. These interaction control major cellular processes. Examples includes signal transduction processes, the regulation of metabolism and the reactions to environmental changes [45]. Usually these networks are illustrated as undirected graphs. Protein representation is done through nodes in these graphs while edges indicate either physical or functional interaction between proteins. In the case of cancer, especially lung cancer, PPI has a critical role. Lung cancer is still among the most common causes of cancer death all over the world. PPI networks on a large scale are provided by high throughput datasets. One can use them as a very powerful means to analyze the altered signaling pathways. They make it possible for us to comprehend the molecular mechanisms that give rise to tumors [46]. The integration of graph theoretic topological metrics with machine learning has emerged as a powerful approach. This integration provides a robust way to identify the hub genes. Hub genes are highly connected proteins. They play a pivotal roles in maintaining network integrity and function [47].

The main difficulty is the precise identification of the biologically important hub genes in large scale PPI networks. The traditional centrality measures often do not disclose the delicate structural patterns. In collaboration with noisy and high dimensional data they have very poor performance [48]. This drawback causes huge hold-ups in both clinical and research areas. Not being able to detect hub genes slows down the discovery of biomarkers. It also hinders the development of personalized therapies. Furthermore, it may cause worse prognostic results. This holds true for non-small cell lung cancer (NSCLC) in particular. NSCLC constitutes 85% of all lung cancer cases [49]. Biological networks usually have a scale free architecture. Only a few nodes have very high connectivity, while most of the nodes have few connections. Such a structure makes it even more difficult to analyze. The highly connected gene's groups are very sensitive to perturbations, even slight changes can trigger a distribution of effects that may eventually lead to the entire network getting deregulated [50].

Existing techniques and solutions are largely dependent on different centrality metrics. The classical examples are degree and betweenness centrality. Other methods utilize the co-expression clustering techniques. Very popular example being WGCNA [51]. Essential

gene classification has also benefitted from the application of machine learning models. A few common examples of such models are random forests and support vector machines (SVMs). The supervised methods can offer a binary classification regarding the importance of the gene [52]. Still, all these methods have severe shortcomings. They are overly reliant on individual metrics, exhibit high noise sensitivity and do not manage outliers well. Also, they demonstrate a lack of scalability when implemented on large interactomes such as those from the STRING database for protein protein interaction networks[53]. Besides, these methods and solutions are often not able to merge anomaly detection with clustering. Consequently, they might miss very important nodes having unusual topological features. Such nodes may correspond to biologically significant proteins showing non standard network behavior.

In order to overcome these limitations and deficiencies, This study suggests an integrated computational pipeline. Different from earlier research that have focused on one main centrality metric or used supervised classifiers our approach unifies 17 network topological based indices, unsupervised anomaly detection and K-means clustering to unveil concealed topological structures that traditional hub detection techniques overlook. It unites the use of various strong methods like graph creation based on NetworkX and Isolation Forest algorithm for detecting outliers. This entire process picks out functional modules and hub genes in a cancer related network that is derived from the STRING database [53]. This method has many benefits. It raises the precision of hub genes detection. Furthermore, it provides pathway based biological knowledge through KEGG and Reactome enrichment analysis.

3.0.5 Data Acquisition

Algorithm 3 PPI-Based Hub Gene Detection using Topological Indices and Machine Learning

Input: ProteinProtein Interaction (PPI) dataset from STRING database
Output: Centrality-based feature dataset for ML-based clustering
ppi_data ← Load PPI interactions (.tsv) from STRING database
graph ← Construct undirected simple graph $G = (V, E)$ using NetworkX
for each *protein* $\in V$ **do**
    metrics ← Compute topological and centrality measures:
        {degree, degree_centrality, closeness, betweenness, clustering,
        harmonic, load, core_number, avg_neighbor_degree, square_clustering,
        eigenvector, katz, pagerank, hubs, eccentricity, edge_betweenness_sum,
        inv_shortest_path_len}
    feature_dataset ← Append *metrics* to dataset
end for
clusters ← Apply machine learning clustering (e.g., K-Means) on feature_dataset
hub_genes ← Identify high-degree or high-centrality nodes as potential hub genes
return *hub_genes*

A dataset of 18,704 interactions (or protein pairs) and for each pair, 13 different attributes was downloaded from the STRING database. The dataset included columns such as source proteins, target proteins, and interaction scores. The selection of proteins was made very carefully by considering their co-expression patterns as the main criteria. These patterns were detected in different cancer types. Special emphasis was placed on lung cancer datasets. this targetd selection strategy highlights proteins that are likely involved in critical signaling pathways. It also points to key molecular mechanism that drive lung tumorigenesis. Table 3.14 show ten randomly selected observations from the downloaded dataset from STRING. These examples give a quick overview of the dataset's structure, evidence channels and confidence score.

node1	node2	node1_string_id	node2_string_id	neighborhood_on_chromosome	gene_fusion	phylogenetic_cooccurrence	homology	coexpression	experimentally_determined_interaction	database_annotated	automated_textmining	combined_score
ADCY1	PDE1B	9606.ENSP00000297323	9606.ENSP00000243052	0.0	0.0	0.000	0.000	0.042	0.000	0.65	0.233	0.720
ADCY1	CAMK4	9606.ENSP00000297323	9606.ENSP00000282356	0.0	0.0	0.000	0.000	0.061	0.000	0.65	0.227	0.724
ADCY1	ADCY8	9606.ENSP00000297323	9606.ENSP00000286355	0.0	0.0	0.068	0.852	0.109	0.000	0.90	0.165	0.921
ADCY1	CALM3	9606.ENSP00000297323	9606.ENSP00000291295	0.0	0.0	0.000	0.000	0.000	0.046	0.90	0.768	0.975
ADCY1	ADCY9	9606.ENSP00000297323	9606.ENSP00000294016	0.0	0.0	0.000	0.703	0.073	0.124	0.50	0.511	0.775
KRAS	CAMK2G	9606.ENSP00000256078	9606.ENSP00000319060	0.0	0.0	0.0	0.0	0.072	0.148	0.5	0.157	0.623
ARRB1	BRAF	9606.ENSP00000409581	9606.ENSP00000419060	0.0	0.0	0.0	0.0	0.000	0.091	0.9	0.181	0.919
CALML5	SPTBN4	9606.ENSP00000369689	9606.ENSP00000469242	0.0	0.0	0.0	0.0	0.056	0.000	0.0	0.515	0.522
TRPC3	JPH1	9606.ENSP00000368966	9606.ENSP00000344488	0.0	0.0	0.0	0.0	0.048	0.092	0.0	0.453	0.486
ATP2B3	ITPR1	9606.ENSP00000263519	9606.ENSP00000306253	0.0	0.0	0.0	0.0	0.082	0.091	0.0	0.353	0.413
EGFR	AKAP12	9606.ENSP00000275493	9606.ENSP00000384537	0.0	0.0	0.0	0.0	0.490	0.329	0.0	0.491	0.810
FGF9	RHOA	9606.ENSP00000371790	9606.ENSP00000400175	0.0	0.0	0.0	0.0	0.000	0.045	0.0	0.448	0.450
STIM1	CAV1	9606.ENSP00000478059	9606.ENSP00000339191	0.0	0.0	0.0	0.0	0.000	0.292	0.0	0.822	0.868
PRKACG	NOS1	9606.ENSP00000366488	9606.ENSP00000477999	0.0	0.0	0.0	0.0	0.060	0.091	0.0	0.432	0.472
CALML4	PHKA2	9606.ENSP00000419081	9606.ENSP00000369274	0.0	0.0	0.0	0.0	0.000	0.091	0.9	0.174	0.918

Table 3.14: Top 15 Observations from the STRING Dataset

For this study, we extract the first two columns of the dataset. The first column (node1) denote the source protein. The second column (node2) represents the target protein. Table 3.15 summarizes the comparison of the unique protein identifiers between the source and target columns. The results show that both columns consist on exactly 461 unique values. The number of common unique identifiers is also 461. This complete overlap

confirms that the dataset represent a perfectly symmetric and undirected network. Every protein appear both as a source and as a target in at least one interaction. This symmetry is excepted in undirected biological networks, including Protein-Proteins Interaction (PPI) networks. Representative examples of these shared proteins include GRIN1, NOX1, ATP2B4, SCN8A and TRIM63. This uniformaity in node representation confirms the undirected nature of the networks. Every proteins appears both as a source and as a target in at least one interaction, a standard feature of PPI networks.

Metric	**Value**
Unique values in `source` column	461
Unique values in `target` column	461
Common unique items between both columns	461
Columns have identical unique sets	Yes
Examples of common items	GRIN1, NOX1, ATP2B4, SCN8A, TRIM63, ...

Table 3.15: Summary of Unique Values in Source and Target Columns

We extracted the source (node 1) and target (node 2) columns from the PPI dataset. Using NetworkX, we then constructed a simple undirected graph. In this graph, each node represents a distinct protein. Each edge represents a pairwise protein-protein interaction. Figure 3.34 displays and illustrated the constructed PPI network. This visualization clearly illustrate the dense molecular connectivity within the dataset. It reveals a highly interconnected structure typical of real biological function. The clear identification of key proteins that function as hubs is made possible by this network. These hub proteins can be crucial for the preservation of network stability as well as biological functions. Their identification that is based on the detection of these proteins is the basis on which all the following topological and machine learning analysis is done.

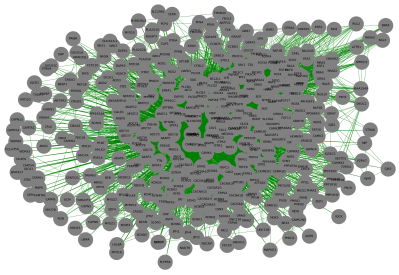


Figure 3.34: Graph Representation of PPIs

There are 461 nodes and 9352 edges in total in the network. Nodes denote individual proteins and Edge represents their pairwise interactions. It is confirmed that the network is connected (Is the Graph Connected? = True), meaning that there is at least one path connecting every pair of nodes, and there are no isolated subgraphs. The average shortest path length is found to be 1.9592, meaning that any two proteins in the network are, on average, separated by less than two interaction steps. The compactness and communication efficiency in the biological system can be summarized by this relatively small path length. The network density of 0.0882 calculated reveals that only around 8.82% of all possible connections among the proteins are made which is a standard characteristic of biological networks that are basically sparse but still highly functional. The average global clustering coefficient of 0.6705 which is very high indeed suggests a high tendency of proteins to form tightly connected groups or functional modules thereby confirming the presence of biologically meaningful clusters. The network consists of a single component only, which proves the global connectivity of the network. The diameter of the network, defined as the longest shortest path between any two nodes, is 3, whereas the radius is 2, suggesting that most nodes are within two or three steps of each other. The center illustrates 126 indices located at the center of the graph which involve CALM3, CREB1, and CAMK2A as well. In general, these features point out that the developed PPI network has small world characteristics high clustering, and short path lengths which are features of complex biological systems and necessary for efficient molecular communication and robustness.

Network Property	**Formula**	**Value**
Number of Nodes (n)	–	461
Number of Edges (m)	–	9352
Is the Graph Connected?	–	True
Average Shortest Path (L)	$L = \frac{1}{n(n-1)} \sum_{i \neq j} d(i,j)$	1.9592
Density (D)	$D = \frac{2m}{n(n-1)}$	0.0882
Average Global Clustering Coefficient (C)	$C = \frac{1}{n} \sum_{i=1}^{n} C_i$	0.6705
Transitivity (T)	$T = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets}}$	0.2805
Number of Components (k)	–	1
Diameter (D_{max})	$D_{max} = \max_{i,j} d(i,j)$	3
Radius (r)	$r = \min_i \max_j d(i,j)$	2
Center	–	126 indices

Table 3.16: Properties and Corresponding Formulas

The 17 computed network topological indices in 3.17, each provide a different viewpoint on the structural and functional properties of the PPI network. Degree centrality spots hub proteins with multiple direct connections. Betweenness and load centralities disclose

proteins that serve as bridges or communication flow control points. Closeness and harmonic centralities determine how fast a protein can reach others, indicating its potential role in information spreading. Eigenvector, Katz, and PageRank centralities stress influence in the network showing proteins connected to other important or highly interactive proteins. Clustering coefficient, local clustering coefficient, transitivity, and density mirror the modular organization and the cohesiveness of protein communities while eccentricity reveals whether a protein is centrally or peripherally located in the network. Communicability betweenness takes into account indirect paths as well as alternative routes. Subgraph centrality is the same with it. Current flow betweenness incorporates indirect and alternative paths, for instance. The metrics provide natural to see through views of information transfer. They give biologically meaningful insights. That is why they are the best for biological networks analysis. Edge betweenness assesses the significance of particular PPIs, and assortativity measures the extent to which nodes are connected to partners of a similar degree. These indices collectively provide a comprehensive quantitative framework for the understanding of the biological networks topology, thus making it possible to identify key hub proteins, functional modules, and critical interaction pathways that might serve as potential biomarkers or therapeutic targets..

Name	**Symbol**	**Formula / Definition**
Degree	$deg(v)$	$deg(v) = \lvert\{u \in V : (u,v) \in E\}\rvert$
Degree Centrality	$C_D(v)$	$C_D(v) = \dfrac{deg(v)}{n-1}$
Closeness Centrality	$C_C(v)$	$C_C(v) = \dfrac{n-1}{\sum_{t \neq v} d(v,t)}$
Betweenness Centrality	$C_B(v)$	$C_B(v) = \sum_{s \neq v \neq t} \dfrac{\sigma_{st}(v)}{\sigma_{st}}$
Clustering Coefficient	$C(v)$	$C(v) = \dfrac{2T(v)}{deg(v)(deg(v)-1)}$
Harmonic Centrality	$C_H(v)$	$C_H(v) = \sum_{u \neq v} \dfrac{1}{d(v,u)}$
Load Centrality	$C_L(v)$	$C_L(v) = \sum_{s \neq v \neq t} \dfrac{\ell_{st}(v)}{\ell_{st}}$
Core Number (k-core)	$\text{core}(v)$	Node belongs to maximal subgraph with $\deg(u) \geq k$
Average Neighbor Degree	$k_{nn}(v)$	$k_{nn}(v) = \dfrac{1}{deg(v)} \sum_{u \in N(v)} deg(u)$
Square Clustering Coefficient	$C_\square(v)$	$C_\square(v) = \dfrac{\text{4-cycles through } v}{\text{possible 4-cycles}}$
Eigenvector Centrality	$C_E(v)$	$C_E(v) = \dfrac{1}{\lambda} \sum_{u \in N(v)} A_{vu} C_E(u)$
Katz Centrality	$C_K(v)$	$C_K(v) = \alpha \sum_u A_{vu} C_K(u) + \beta$
PageRank	$PR(v)$	$PR(v) = (1-d) + d \sum_{u \in N(v)} \dfrac{PR(u)}{deg(u)}$
HITS (Hubs)	$h(v)$	$h(v) = \sum_u A_{vu} a(u), \quad a(u) = \sum_v A_{vu} h(v)$
Eccentricity	$e(v)$	$e(v) = \max_{u \in V} d(v,u)$
Edge Betweenness Sum	$EBC(v)$	$EBC(v) = \sum_{(v,u) \in E} C_B(v,u)$
Inverse Shortest Path Length	$C_{ISP}(v)$	$C_{ISP}(v) = \dfrac{1}{\frac{1}{n-1}\sum_{t \neq v} d(v,t)}$
Edge Betweenness (edge-wise)	$C_B(e)$	$C_B(e) = \sum_{s \neq t} \dfrac{\sigma_{st}(e)}{\sigma_{st}}$

Table 3.17: Topological and Centrality Measures

The centrality-based topological indices were calculated and afterwards the whole resulting measures were merged into one single feature dataset which enabled further computational and statistical analyses to be carried out easily.

Node	degree	degree_centrality	closeness	betweenness	clustering	harmonic	load	core_number	avg_neighbor_degree	square_clustering	eigenvector	katz	pagerank	hubs	eccentricity	edge_betweenness_sum	inv_shortest_path_len
ADCY1	38	0.082609	0.515695	17.400910	0.664286	247.333333	36.087917	30	111.789474	0.206382	0.035207	0.044571	0.001889	0.002083	3	0.004667	0.516816
PDE1B	22	0.047826	0.469815	0.540038	0.962381	237.000000	1.080075	21	128.363636	0.230552	0.020946	0.040087	0.001322	0.001239	3	0.004349	0.500000
CAMK4	44	0.095652	0.519187	50.455847	0.477801	250.333333	102.801590	30	98.295455	0.178493	0.037130	0.045841	0.002284	0.002197	3	0.005290	0.520316
ADCY8	52	0.113043	0.526316	53.163896	0.514329	255.000000	112.020992	30	98.769231	0.188778	0.044349	0.048139	0.002630	0.002624	3	0.005341	0.527460
CALM3	421	0.915217	0.921944	13881.599856	0.085963	440.500000	27189.475232	37	39.750594	0.120120	0.197637	0.127704	0.023099	0.011694	2	0.282409	0.923848
ADCY9	40	0.086957	0.518802	17.337795	0.658410	248.833333	36.736454	30	109.350000	0.208298	0.038583	0.045182	0.002075	0.002184	3	0.004665	0.519729
CACNA1H	47	0.102174	0.515118	42.085782	0.514339	250.166667	86.445000	27	87.574468	0.177570	0.033438	0.046024	0.002452	0.001978	3	0.005132	0.516237
SCN5A	34	0.073913	0.508287	23.946042	0.543672	243.833333	51.435159	25	100.029412	0.167507	0.026996	0.042883	0.001897	0.001579	3	0.004790	0.509392
CACNA1G	49	0.106522	0.528129	57.277942	0.492347	254.500000	129.539482	27	90.448980	0.172180	0.035400	0.046766	0.002540	0.002154	2	0.005419	0.529277
ADCY6	44	0.095652	0.521542	28.013750	0.598309	251.000000	61.003746	30	105.795455	0.200527	0.039870	0.046251	0.002271	0.002359	3	0.004867	0.522676
ADCY4	34	0.073913	0.513393	6.932355	0.755793	245.333333	14.951404	30	118.852941	0.223260	0.033140	0.043720	0.001811	0.001961	3	0.004469	0.514509
ADCY5	52	0.113043	0.525714	48.300482	0.521116	254.833333	98.841858	30	97.750000	0.187583	0.044248	0.048178	0.002617	0.002618	3	0.005249	0.526857
PLCB1	59	0.128261	0.529954	105.750812	0.462104	258.333333	214.529667	30	88.559322	0.167848	0.047883	0.049622	0.003007	0.002832	3	0.006333	0.531106
CREB1	104	0.226087	0.563725	328.479603	0.381441	282.000000	650.289688	36	87.173077	0.189086	0.090865	0.082489	0.004817	0.005378	2	0.010497	0.564951
AKT2	55	0.119565	0.531792	64.552040	0.538721	257.500000	128.381795	35	111.145455	0.229722	0.057785	0.050132	0.002626	0.003419	2	0.005556	0.532948
AKT1	183	0.397826	0.624152	1624.811040	0.234732	321.500000	3287.325297	37	84.448087	0.166352	0.129821	0.079932	0.008464	0.007681	2	0.034987	0.625509

Figure 3.35: Sample of Ensemble Dataset

Table 3.18 summarizes 17 crucial topological indices that were computed from a PPI network with 461 nodes. Interconnections among proteins are the main features described by these indices. It also indicates the possible paths across the network. The significant average degree (40.57) along with a considerable standard deviation (50.23) reveals that the majority of the nodes have moderate connections while the rest are with large degrees of interactions thus, forming the hubs, which have connection to more than 400 other nodes. Likewise, the average degree centrality (0.088) and the maximum value (0.93) are indicative of the structure being scale free, which is a characteristic of biological systems where few highly connected proteins do maintain the stability of the whole network. Centrality metrics that are based on node types like betweenness, load, and harmonic centralities not only help in determining but also in communicating the roles of each node in a network. The extensive disparity in betweenness (from 0 to more than 14,000) implies that merely a few nodes wield the power to determine the shortest paths to a large extent. These nodes could be controlling or bottleneck proteins. On the other hand, the very high average clustering coefficient (0.67) points to the fact that the proteins within the biological system do interact through close-knit groups or modules that can be considered as functional complexes or pathways. Measures like core number and average neighbor degree, to a great extent, point out this internal cohesion, revealing that the more connected the proteins are, the more they will be surrounded by other highly connected ones, thus, a dense network core is formed. Measures of influence such as eigenvector, Katz, and PageRank centralities

point out the nodes that are not only very well connected but also connected to other nodes of high influence. These nodes can be seen as either regulatory or master proteins that keep the whole system coordinated globally. The closeness and inverse shortest path length values from moderate to high are indicative of efficiency in communication throughout the network. The network is extremely compact. This makes it possible for signals to travel at high speed. Molecular interactions also have a quick spreading. The eccentricity values is low (at 2-3 approximately) imply that no protein is really distant from another one. Every node can be reached in a few steps. Thus, information transfer is fast. The arrangement allows for communication effectiveness. When all these statistics are considered together, they characterize a biological network as heterogeneous, modular, and scale free at the same time. The characteristics of real world protein-protein interaction systems having a small number of highly connected nodes, localized clustering of connections, and the ability to connect efficiently, are all shown by the dense networking. Such a setup guarantees stability against random changes while still allowing communication among the different parts to be very efficient. To put it in real-world terms, the nodes that have extremely high centrality or connectivity scores are thought to be the proteins that act as hubs, which is often true for essential genes or potential drug targets, hence this analysis of network topology is an important stage in both systems biology and drug discovery research.

Feature	Count	Mean	Std Dev	Min	Max
Degree	461	40.57	50.23	5.00	428.00
Degree Centrality	461	0.088	0.109	0.0109	0.9304
Closeness	461	0.514	0.0528	0.3843	0.9349
Betweenness	461	220.61	1448.76	0.00	14294.98
Clustering Coefficient	461	0.671	0.203	0.0858	1.000
Harmonic Centrality	461	246.65	27.13	187.50	444.00
Load Centrality	461	441.23	2875.01	0.00	28373.02
Core Number	461	22.30	10.02	5.00	37.00
Average Neighbor Degree	461	151.61	81.28	39.72	425.00
Square Clustering	461	0.275	0.150	0.120	0.989
Eigenvector Centrality	461	0.0367	0.0288	0.0086	0.2012
Katz Centrality	461	0.0451	0.0117	0.0359	0.1293
PageRank	461	0.00217	0.00256	0.00056	0.02347
Hubs	461	0.00217	0.00170	0.00051	0.01190
Eccentricity	461	2.73	0.45	2.00	3.00
Edge Betweenness Sum	461	0.00850	0.02733	0.00434	0.27398
Inverse Shortest Path Length	461	0.515	0.0529	0.3851	0.9370

Table 3.18: Descriptive Statistics of Network-Based Topological Features (461 Nodes)

The plot 3.36 illustrate the distribution of features including degree, closeness, betweenness and clustering coefficients.

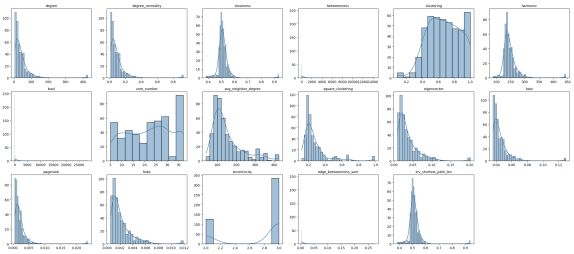


Figure 3.36: Distribution Ensemble Dataset

3.0.6 Outlier Detection Using Isolation Forest

We employed Isolation Forest for unsupervised anomaly detection in the dataset. Isolation Forest is a tree based Ensemble learning method designed specially to identify anomalies. It isolates observations through random partitioning. Anomalies, being fewer and distinct require fewer splits to isolates than points. The algorithm was implemented using Sklearn.ensemble.IsolationForest with the help of Python. The model was configured with following parameters:

Algorithm 4 Configuration of Isolation Forest Algorithm

Input: Dataset D
Output: Anomaly scores for each sample
$n\_estimators \leftarrow 100$ {Use 100 isolation trees to ensure diversity}
$max\_samples \leftarrow 256$ {Each tree trained on a random subset of 256 samples}
$contamination \leftarrow auto$ {Automatically estimates anomaly proportion}
$random\_state \leftarrow 42$ {Ensures reproducibility of results}
Train the Isolation Forest model on dataset D
Compute anomaly scores for all samples
Return: anomaly scores

The Isolation Forest model assigns 1 to in-liers (normal observations) and -1 to outliers (anomalous observations). The prediction array contained 461 labels one for each data point in the dataset. Both positive (1) and negative (-1) values were present in the output. Samples labeled -1 were identified as possible anomalies/outliers.

We quantified the anomaly degree using the Isolation Forest decision function. This function computes an anomaly score for each data point. Lower more negative scores indicate a stronger likelihood of being an outlier. Scores in this dataset ranged from approximately -0.073 (most anomalous) to 0.133 (most normal/inlier). The score distribution effectively separated typical observations from anomalies.

This operation identified **61 outliers** in total. The corresponding index positions were:

```
Outlier indices with corresponding nodes and scores:

4   -> CALM3(0.00379)     13  -> CREB1(0.01779)    15  -> AKT1(0.00581)
22  -> CAMK2A(0.00802)    26  -> PRKCA(0.00867)    31  -> PRKACA(0.00311)
32  -> PRKACB(0.00311)    34  -> PRKACG(0.00311)   35  -> CALML3(0.00379)
36  -> CALML4(0.00379)    37  -> CALML6(0.00379)   38  -> CALML5(0.00379)
39  -> MAPK1(0.02227)     41  -> CACNA1C(-0.02034) 47  -> MAPK3(0.01942)
57  -> CAV1(0.00447)      59  -> RYR2(-0.00071)    67  -> EGFR(0.01942)
85  -> GRB2(0.02272)      87  -> RHOA(0.01942)     98  -> KRAS(0.01942)
103 -> SRC(0.01502)       104 -> PPP1CB(0.03482)   106 -> CDC42(0.01851)
112 -> ITPR1(-0.02650)    113 -> ITPR3(-0.01147)   132 -> ARAF(-0.00605)
138 -> RAF1(0.02270)      139 -> HRAS(0.01994)    141 -> NRAS(0.01929)
143 -> BRAF(0.02188)      157 -> PLCG1(0.02292)   167 -> PIK3CA(0.02250)
172 -> EGF(0.01063)       191 -> GSK3B(0.02321)   199 -> FN1(0.02344)
202 -> MTOR(0.02404)      224 -> SOS1(-0.02539)   237 -> PIK3R1(-0.02062)
243 -> FRS2(-0.00040)     255 -> PIK3CB(-0.02800) 267 -> AQP6(0.06142)
268 -> MRAS(-0.00823)     269 -> BRAP(0.03203)    270 -> RGL3(0.03713)
271 -> RIN1(0.04741)      274 -> RGL2(0.04545)    275 -> KSR2(0.03017)
276 -> SHOC2(-0.00780)    277 -> CNKSR1(0.03586)  332 -> RASA2(0.04147)
334 -> LZTR1(0.04604)     335 -> C11orf58(0.06140) 379 -> PLPPR4(0.06140)
382 -> IQCK(0.06142)      389 -> DCAF6(0.06142)   400 -> GEM(0.06142)
427 -> SLC29A1(0.02924)   435 -> MAP6(0.06142)    445 -> MYO1A(0.06142)
449 -> UBR4(0.06142)
```

The identified indices correspond to observation deemed statistically distinct by the model. These may be reflecting noise, errors of measurement or really outstanding cases which deserve investigation to some extent. Inliers predominantly received positive anomaly scores mostly ranges from 0.05 to 0.13 indicating typical behaviour. Outliers consistenly exhibited negative scores due to shorter average path lengths in the isolation trees. Overall, Isolation Forest successfully separated anomalous from the main data distribution. The Isolation Forest method indicated that roughly **13.2% (61 out of 461)** of the overall data points were flagged as aberrant by the models internal reasoning. These anomalies can either associate with deviations that are biochemically or chemically significant (in cases where the dataset represents such entities as drug compounds or proteins), or they can be just erroneous or noisy entries that need to be filtered out via preprocessing. In fact the scores from the decision function can serve as a basis for either filtering or marking observations that need to be inspected more closely prior to performing tasks like regression, clustering or classification, which are subsequent machine learning applications. What is more, the limit between normal and anomalous can be adjusted according to the particular noise tolerance of the downstream analysis. The figure 3.37 illustrates the outliers identified by the Isolation Forest algorithm. Nodes highlighted in red represent the predicted outliers, whereas nodes shown in grey correspond to normal observations.

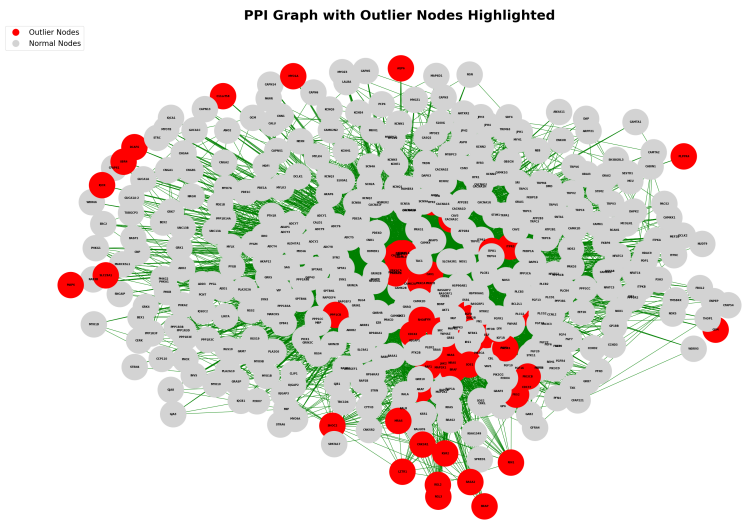


Figure 3.37: Outliers by Isolation Forest

Anomaly scores are calculated by the Isolation Forest model for outlier detection in the dataset appearing in the given distribution of the figure 3.38. The x-axis shows the anomaly scores, less negative scores imply a higher chance of being anomalies while the y-axis indicates the number of data points corresponding to each score. The greatest number of scores can be found around the values of 0.05 to 0.15, thus suggesting that normal distribution is prevailing among the majority of the data points. The vertical line in red with dashes marks the threshold for the outliers at (-0.053) whereas the area of probable outliers is shown by the red shading area to the left side of this line. Data points in this shaded area are tagged as anomalous, while points on the other side of the threshold are considered normal. This plot clearly illustrates the boundary delineation of the Isolation Forest model between the potential outliers and the normal observations according to each anomaly score.

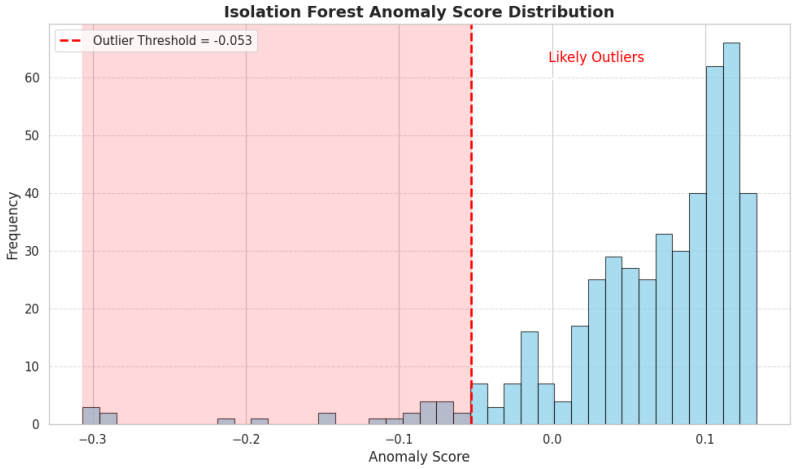


Figure 3.38: Isolation Distribution

Using the Isolation Forest algorithm, the dataset was divided into two subsets, Inliers and outliers as we can see in table 3.19. This separation was performed to enable the application of Winsorization , where the inlier and outlier datasets were utilized to cap extreme values and reduce the influence of anomalies on subsequent analysis. The inliers subset consists of 400 rows out of a total of 461 whereas the outliers subset contains 61 rows that were identified as anomalies by the Isolation Forest algorithm. We calculate percentile from intile dataset and then we apply on real dataset of 461 observations.

Node Type	Node	Degree	Degree_Centrality	Closeness	Betweenness	Clustering	Harmonic	Load	Core_Number	Avg_Neighbor_Degree	Square_Clustering	Eigenvector	Katz	PageRank	Hubs	Auth	Eccentricity	Inv_Shortest_Path_Len
Inlier	ADCY1	38	0.0826	0.5157	17.4009	0.6643	247.33	36.0879	30	111.79	0.2084	0.0352	0.0447	0.001989	0.002083	0.002083	3	0.5168
Inlier	PDE1B	22	0.0478	0.4989	0.5400	0.9524	237.00	1.0801	21	128.36	0.2306	0.0209	0.0401	0.001322	0.001239	0.001239	3	0.5000
Inlier	CAMK4	44	0.0957	0.5192	50.4558	0.4778	250.33	102.8016	30	98.30	0.1785	0.0371	0.0458	0.002284	0.002197	0.002197	3	0.5203
Inlier	ADCY8	52	0.1130	0.5263	53.1637	0.5143	255.00	112.0210	30	96.77	0.1888	0.0443	0.0481	0.002630	0.002624	0.002624	3	0.5275
Inlier	ADCY9	40	0.0870	0.5186	17.3378	0.6564	248.83	36.7365	30	109.35	0.2063	0.0366	0.0452	0.002075	0.002164	0.002164	3	0.5197
Outlier	CALM3	421	0.9152	0.9218	13681.5999	0.0860	440.50	27189.4752	37	39.75	0.1201	0.1976	0.1277	0.023099	0.011694	0.011694	2	0.9238
Outlier	CREB1	104	0.2261	0.5637	326.4796	0.3814	282.00	650.2697	36	87.17	0.1881	0.0909	0.0625	0.004817	0.005376	0.005376	2	0.5650
Outlier	AKT1	183	0.3978	0.6242	1624.8110	0.2347	321.50	3287.3253	37	64.45	0.1664	0.1298	0.0799	0.008464	0.007681	0.007681	2	0.6255
Outlier	CAMK2A	93	0.2022	0.5562	394.1469	0.3088	276.50	800.3458	31	76.58	0.1571	0.0684	0.0580	0.004494	0.004047	0.004047	2	0.5574
Outlier	PRKCA	109	0.2370	0.5672	474.7602	0.3179	284.50	949.1371	35	80.83	0.1698	0.0877	0.0630	0.005080	0.005192	0.005192	2	0.5684

Table 3.19: Preview of Inlier and Outlier subset

In order to reduce the effect of outliers and at the same time keep the general data layout intact, the inlier dataset underwent winsorization at the 10th and 90th percentiles which were set as the lower and upper caps, respectively. In detail, the limits for the lower and upper quantile were determined based on the inlier's feature values, and all feature values falling outside the limits were clipped to the nearest limit. The `outlier` column was removed before this operation to guarantee that only valid feature variables were included in the processing. The application of this method reduces the chances of noise and skewness in the data extremely and gives an even more solid and less affected representation of the data which is then fit for further analysis. Figure 3.39 compares feature distributions before and after outlier detection using the Winsorization. TThe left panel comes with very obvious outliers and skewness while the right panel shows that limiting extreme values to whisker lengths results in a distribution that is better balanced.

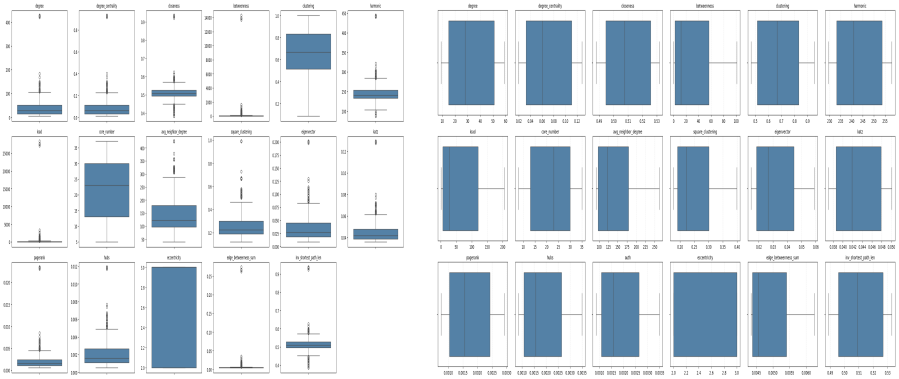


Figure 3.39: Comparison of Feature Distributions Before and After Outlier Capping

Correlation heatmap is illustrated in Figure 3.40. It consists of 18 graph-theoretic features. Heatmap is generated after the outlier capping process. Capping minimizes the effect of outliers and extreme values. This makes the connections among the features more consistent and easier to interpret.

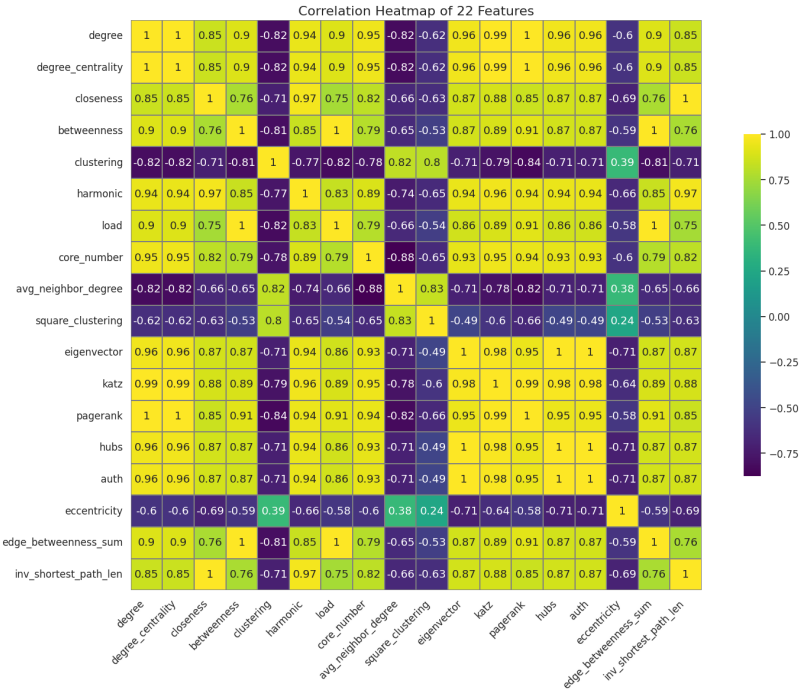


Figure 3.40: Correlation After Outlier Capping

Before proceeding with clustering, the dataset was standardized so that all numerical features would be considered uniform and could be compared even though they were originally of different scales. Standardization play a crucial role in clustering analysis. Distance-based methods such as *k*-means and hierarchical clustering are extremely affected by feature scales. Features with larger ranges or variances dominate the distance calculations. Without standardization, the result of clustering can become distorted or meaningless. It ensures that every feature will play an equal part in the similarity measurement. Mathematically, the transformation can be defined as

$$z = \frac{x - \mu}{\sigma}$$

where x represents the value of original feature, μ is the mean value and σ is the standard deviation of that feature. This normalizing process makes it possible for clustering outcomes to show the real structure of the data instead of being influenced by the variations in the measurement units or their magnitudes. Table 3.20 show first 10 values of dataset after Standardization.

Node	degree_centrality	closeness	betweenness	clustering	harmonic	load	core_number	avg_neighbor_degree	square_clustering	eigenvector	katz	pagerank	hubs	auth	eccentricity	edge_betweenness_sum	inv_shortest_path_len
0	0.346772	0.394008	-0.393863	-0.108006	0.432760	-0.406770	0.827341	-0.583538	-0.580382	0.195195	0.363860	0.317707	0.195124	0.195124	0.613286	-0.393863	0.394008
1	-0.525598	-0.667502	-0.835799	1.575055	-0.560714	-0.858464	-0.129625	-0.300137	-0.279711	-0.696158	-0.585634	-0.534642	-0.696174	-0.696174	0.613286	-0.835799	-0.667502
2	0.673910	0.614936	0.472532	-1.197551	0.721188	0.454013	0.827341	-0.814272	-0.985736	0.315361	0.606097	0.695282	0.315290	0.315290	0.613286	0.472532	0.614936
3	1.110095	1.065893	0.543507	-0.984148	1.169853	0.572968	0.827341	-0.840368	-0.846280	0.766558	1.081923	1.137167	0.766457	0.766457	0.613286	0.543507	1.065893
4	1.491757	1.412322	1.802189	-1.261153	1.461486	1.775774	1.358989	-0.931265	-1.003059	1.683999	1.485474	1.514196	1.684041	1.684041	-1.630561	1.802189	1.412322
5	0.455818	0.577908	-0.395517	-0.154075	0.576974	-0.398402	0.827341	-0.625250	-0.608674	0.281170	0.469629	0.428518	0.281095	0.281095	0.613286	-0.395517	0.577908
6	0.837480	0.357475	0.253146	-0.984091	0.705164	0.242970	0.508352	-0.931265	-0.998244	0.084594	0.644027	0.910429	0.084484	0.084484	0.613286	0.253146	0.357475
7	0.128679	-0.074620	-0.222310	-0.812719	0.096261	-0.208750	0.295693	-0.784623	-1.003059	-0.336763	-0.006528	0.200480	-0.336809	-0.336809	0.613286	-0.222310	-0.074620
8	0.946526	1.180574	0.651344	-1.112572	1.121782	0.799003	0.508352	-0.931265	-1.003059	0.269773	0.797557	1.021907	0.269662	0.269662	-1.630561	0.651344	1.180574
9	0.673910	0.763892	-0.115692	-0.493518	0.785283	-0.085290	0.827341	-0.686029	-0.686903	0.486620	0.690923	0.678589	0.486537	0.486537	0.613286	-0.115692	0.763892

Table 3.20: Standardized Values of Centrality Measures for Sample Nodes

Figure 3.41 The violin plots illustrate the distribution of the features following data standardization highlighting the adjusted spread and central tendency of each variable.

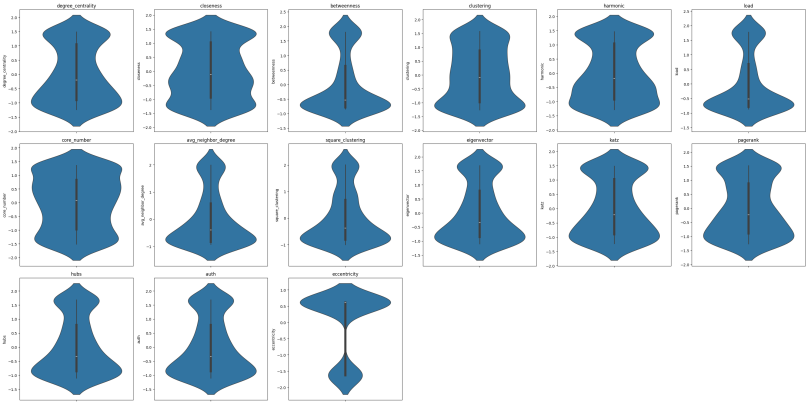


Figure 3.41: Distribution after Standardization

Since features such as degree and degree centrality, Closeness centrality and Inverse shortest path as well as betweenness and edge betweenness sum convey overlapping information the features degree, Inverse Shortest path and edge betweenness sum were removed from further analysis to avoid redundancy.

3.0.7 Clustering Using K-means and Elbow Method

The Elbow Method was employed to determine the most suitable number of clusters which is depicted in Figure 3.42. With this method, the sum of squared distances (inertia) between all data points and their respective cluster centroids was calculated for different values of clusters (k = 1 to k = 12). The plot exhibits a sharp decline in inertia up to $k = 3$, after which the rate of decrease becomes relatively gradual. The elbow point indicates the inflection in the curve. It represents the point at which adding the number of clusters results in very slight improvements only. This point has been referred to as the elbow. On the other side of it, additional clusters only bring about a slight increase in performance. Therefore, $k = 3$ is selected as the optimal number. It achieves a good balance. The compactness of the

clusters still remains high. Between-cluster separation is still very noticeable.

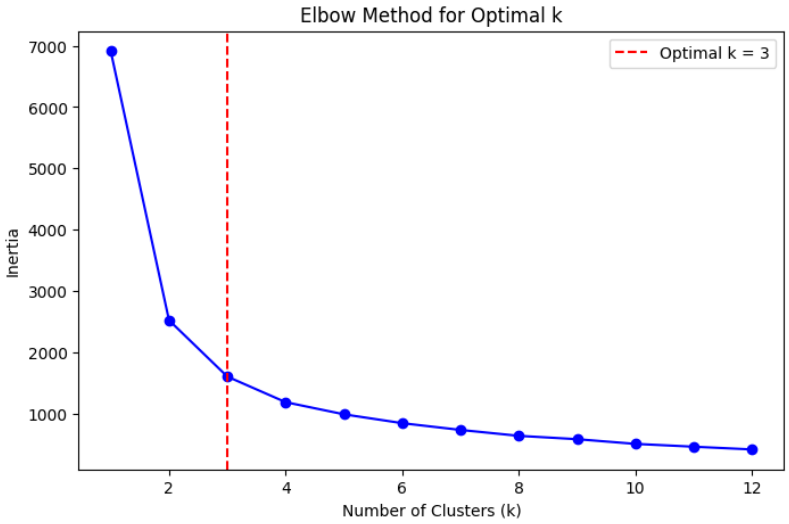


Figure 3.42: Elbow Curve

Utilizing the Elbow method, we arrived at a conclusion that the dataset contains three clusters as the optimal number. Thus, the K-means algorithm is ready to be performed according to the steps given in Algorithm 5.

Algorithm 5 K-Means Clustering Algorithm for Feature-Based Grouping

Input: Feature dataset $X = \{x_1, x_2, \ldots, x_n\}$, where $x_i \in \mathbb{R}^d$
Output: Cluster assignments C_i and final centroids μ_j
$k \leftarrow$ Decide the optimal number of clusters using the Elbow Method (plot inertia vs. k)
Initialize centroids $\{\mu_1, \mu_2, \ldots, \mu_k\}$ randomly from X
repeat
    (Cluster Assignment) For each $x_i \in X$, assign it to the nearest centroid:

$$C_i = \arg\min_j \|x_i - \mu_j\|^2$$

    (Centroid Update) For each cluster C_j, update its centroid:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

until centroids do not change significantly or maximum iterations reached
Return: Cluster assignments C_i and final centroids μ_j

The dataset was split into three clusters by the K-means clustering technique. These clusters are easily distinguishable from one another. They represent the data in such a manner that the natural formation is clearly seen. The algorithm picked up the patterns that were hidden. Each cluster is made of points that have more or less the same properties. The distances between the points in a cluster are smaller than the distances to the points in the neighboring cluster, showing that the points in the other clusters are remote from each other

but very close to each other. This is what makes clusters so well separated. The individuals of the clusters are characterized by their similarities. The first cluster has 136 observations, the second one has 135 observations, and the third one has 190 observations. This distribution demonstrates the existence of significant patterns in the data. It indicates the strongest relations among the features. The clustering technique makes it easier to analyze the data further. Researchers can now investigate cluster specific characteristics. Every cluster has its own characteristics and behaviors. Table 3.21 presents a sample of proteins that belong to the cluster they are assigned.

Protein	Cluster	Protein	Cluster	Protein	Cluster	Protein	Cluster	Protein	Cluster	Protein	Cluster
ADCY8	0	CALM3	0	CACNA1G	0	ADCY5	0	PLCB1	0	CREB1	0
AKT2	0	AKT1	0	PRKCB	0	CAMK2A	0	CAMK2B	0	PDE4D	0
PRKCA	0	PRKCD	0	RAPGEF4	0	PRKACA	0	PRKACB	0	RAPGEF3	0
PIK3CB	0	RPS6KA1	0	RASGRF2	0	PTK2B	0	TRPC3	0	NTRK2	0
PLCB2	0	CNGA2	1	ANO2	1	PCNT	1	CNGA4	1	OCM	1
GRK4	1	MARCKSL1	1	BASP1	1	CABIN1	1	KCNE4	1	MYO9B	1
TUBGCP3	1	KIAA1549	1	INVS	1	EEF2K	1	IQGAP2	1	SPRED1	1
MYLK	1	TBC1D4	1	TMSB4X	1	CAMKK1	1	NOX5	1	HDAC9	1
BEX2	1	MDFI	1	RASA2	1	LZTR1	1	C11orf58	1	MEF2B	1
MYO9A	1	MYO1D	1	PFN3	1	MAP6D1	1	GUCA1A-2	1	PLA2G6	1
VIP	2	ORAI1	2	KCNE1	2	PDE2A	2	CACNA1S	2	CACNA2D1	2
SLC8A1	2	CACNA2D2	2	CAMK1	2	CAMK1G	2	CALM1	2	CAMKK2	2
CACNA1A	2	CACNA1B	2	PDE3A	2	PDE3B	2	CACNA1E	2	CACNA2D3	2
CACNA2D4	2	PDE4B	2	PDE4C	2	PDE4G	2	PDE8B	2	PDE9A	2
PDE10A	2	PDE11A	2	CACNA1F	2	PDE6A	2	PDE6B	2	PDE6C	2

(Table continues with remaining proteins up to 461 entries)

Table 3.21: Cluster Assignment for a Representative Sample of 461 Proteins

The histogram is shown in Figure 3.43. It illustrates the number of proteins placed in different K-Means clusters. Above each bar, the precise counts are indicated. A KDE is overlaid to produce a smooth density curve. The KDE represents the hidden distribution of each cluster.

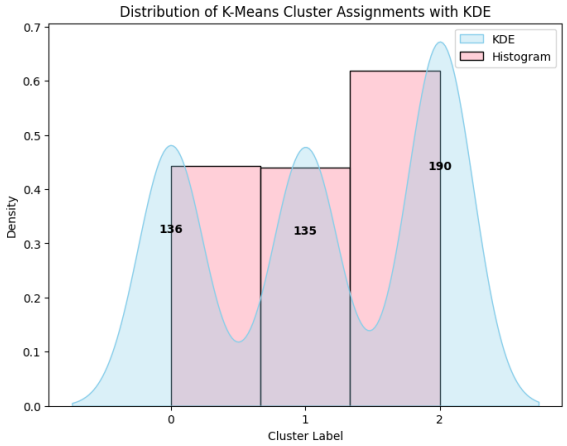


Figure 3.43: Distribution of K-Means Prediction

The first two principal components of PCA have been used to project the dataset as depicted in Figure 3.44. The coloring of data points is done according to their K-Means cluster labels. Each color represents one distinct cluster. The plot clearly reveals spatial separation between the clusters. It works in a reduced 2D space. This visualization makes it easier to check the quality of clustering. It also provide the information how observations are distributed across clusters.

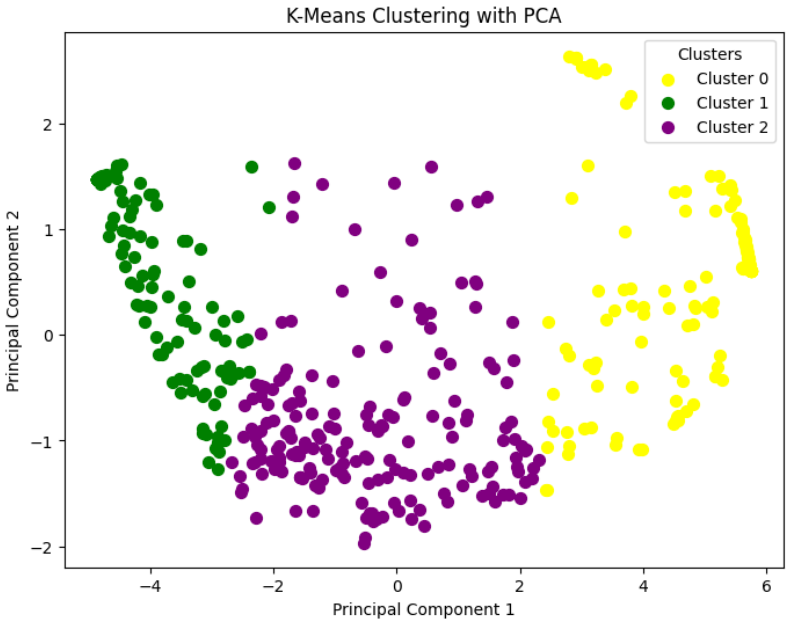


Figure 3.44: Two-Dimensional Visualization

Silhouette Analysis of K-Means Clustering

Silhouette analysis was performed to check the quality of clusters obtained from the K-Means algorithm. The silhouette score for an individual protein i can be defined as:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{3.0.2}$$

In this equation $a(i)$ indicates the distance within the same cluster. It denotes the mean distance of protein i to every other protein in the cluster. On other hand, $b(i)$ is the nearest cluster distance. This represents the minimal average distance from protein i to the other clusters' proteins. So $b(i)$ gives an idea of how far the next cluster which is the closest one,

is. The silhouette score takes values from -1 to +1. Higher values mean stronger cluster quality. A value close to +1 denotes a high degree of closeness among the points in the cluster with respect to each other. It also shows that the cluster is well separated from the other clusters. Scores that are close to 0 indicate that the clusters are overlapping. Negative scores indicate that the clustering of the points is very poor. The dataset that we were working with yielded an average silhouette score of 0.4820 overall, with the scores going from **0.0007** to **0.7298**. This points to a structure that is moderately well-clustered. The mean silhouette score overall came out to be 0.4820. This number was computed using scikit-learn. It was identical to the score that was calculated manually. The agreement indicates the trustworthiness of the clustering evaluation. The K-Means method managed to create clusters that were fairly distinct. Clusters 0 and 1 demonstrate very good internal bonding. Cluster 2 looks weaker and more widely scattered as we can see in table 3.22.

Cluster	Number of Proteins	Average Silhouette Score	Score Range
0	136	0.5610	–
1	135	0.5695	–
2	190	0.3632	–
Overall	461	0.4820	[0.0007, 0.7298]

Table 3.22: Silhouette Analysis Summary for K-Means Clusters

The figure 3.45 presents a comprehensive analysis of cluster structure in the dataset. The first scatter plot on the left presents three clusters as data points. The first cluster is represented with the color blue. The second one is represented with the color orange. The third one is represented with the color green. Only the first and second features are used for the visualization. The scatter plot indicates that the clusters are very distinctly separated in 2D space. Each of the three clusters is characterized distinctly from the others. The middle plot further investigate the relationship between these features across clusters. The histogram on the right demonstrates the silhouette scores distribution. A mean score of 0.472 is indicated by a red dashed line. This score suggests that clustering is of moderate quality. The two plots combined show the separation of clusters quite clearly. Simultaneously they also measure the internal connectivity of the dataset. In general the visualization supports the conclusion of good cluster separation.

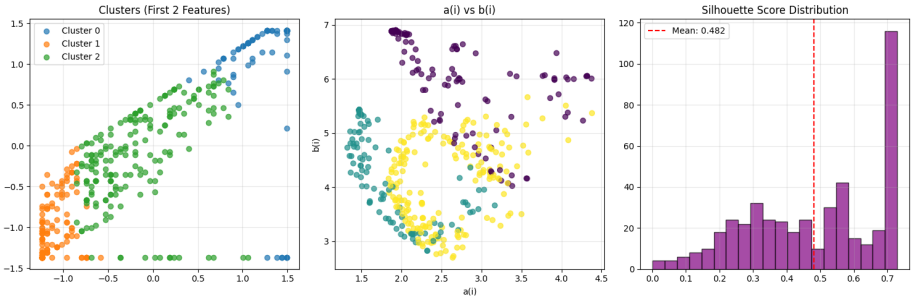


Figure 3.45: Silhouette Analysis

3.0.8 Hub Gene Detection and Pathway Analysis

Hub Gene Detection by Degree

Hub genes were discovered through the identification of the nodes with high degree centrality (like, for instance, the top 6% of the nodes according to their degree). The table provides a list of the high-degree proteins (with degrees greater than 100) together with the K-Means algorithm's cluster assignments. It is emphasized that all nodes with high degrees were assigned to Cluster 0, indicating that the proteins involved are highly interdependent and draw one functional net together.

Protein	Degree	Cluster	Protein	Degree	Cluster
CALM3	421	0	CREB1	104	0
AKT1	183	0	PRKCA	109	0
PRKACA	150	0	PRKACB	145	0
PRKACG	141	0	CALML3	428	0
CALML4	425	0	CALML6	425	0
CALML5	426	0	MAPK1	113	0
MAPK3	137	0	CAV1	108	0
EGFR	136	0	GRB2	121	0
RHOA	130	0	KRAS	143	0
SRC	168	0	CDC42	110	0
ITPR1	105	0	RAF1	101	0
HRAS	130	0	NRAS	118	0

```
BRAF      -> 102     -> 0          PLCG1      -> 105     -> 0
PIK3CA    -> 114     -> 0          GSK3B      -> 109     -> 0
------------------------------------------------------------
```

The PPI network's hub genes are illustrated in Figure 3.46. The network is made up of 461 nodes (proteins). The degreed values of these proteins are much higher than those of other proteins and therefore they are labelled as hub genes. These proteins are the main regulators in the network. They are of paramount importance in both the network's structure and function. Hubs are vital to the overall network's stability that they help to maintain.

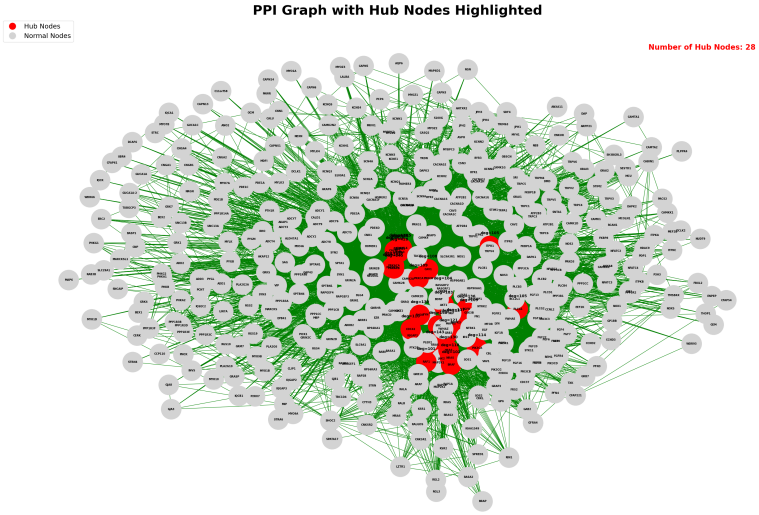


Figure 3.46: Hub Nodes Based on Degree

Computation of Hub Score

A composite metric named the hub score was computed. It assesses topological prominence of every protein in the PPI network quantitatively. The hub score fuses together several centrality metrics. This method shows a more solid evaluation than any individual measure by itself. More elevated hub scores point out the most vital and powerful nodes. Nodes like these are seen as essential hubs in the network layout. The mentioned score is the result of the integration of different centrality measures, where in each index corresponds to a different aspect of node importance in the network layout. In detail for every single node the hub score was calculated as the normalized weighted average of the following metrics:

degree centrality, **closeness centrality**, **betweenness centrality**, **eigenvector centrality**, **PageRank**, **core number**, **HITS hub score** and **Katz centrality**. This can mathematically be represented as:

The composite Hub Score $HS(v)$ aggregates these metrics through a weighted linear combination:

$$HS(v) = \sum_{j=1}^{8} w_j \cdot c_j(v) \tag{3.0.3}$$

where $\mathbf{w} = (0.12, 0.14, 0.26, 0.12, 0.10, 0.10, 0.08, 0.08)^\intercal$ is the empirically-determined weight vector satisfying $\sum_{j=1}^{8} w_j = 1$. In matrix notation, with $\mathbf{C} \in \mathbb{R}^{n\times 8}$ representing he centrality matrix for all $n = |V|$ nodes, this simplifies to:

$$\mathbf{HS} = \mathbf{Cw} \tag{3.0.4}$$

This convex combination ensures interpretability while integrating multi dimensional structural information into a unified node importance metric. The composite hub score combines different centrality measures in PPI networks by utilizing biologically informed weights. Betweenness centrality (0.26) is the leading measure. It indicates bridging proteins that control the inter module signaling. Closeness centrality (0.14) denotes quick network influence. Degree and eigenvector centralities (0.12 each) assess connectivity and module membership, respectively. They avoid overstating non essential hub significance. On the other hand PageRank, core number, HITS hubs and Katz (0.08-0.10) indicate indirect influence and core stability through providing complementary signals of these aspects. This weighting scheme yields a single metric that effectively detects biologically essential, structurally central proteins.

The hub score is a metric that combines various sources of information. Thus the whole view of a nodes importance is presented more clearly. It integrates some local measures of connectivity. These are degree centrality, core number and HITS hub score respectively. Moreover, it takes into account global influence measures too i.e., eigenvector centrality, closeness, betweenness, PageRank and Katz centrality among others. By weighted averaging them, different aspects of centrality are balanced out. As a result the identification of

key hub proteins is robust. High hub score nodes are expected to be critical in the biological processes. Therefore they are the best candidates for subsequent bioinformatics analysis.

Node	Hub Score	Degree	Cluster	Node	Hub Score	Degree	Cluster
PRKACA	1.554121	150	0	PRKCA	1.554121	109	0
PRKCD	1.554121	68	0	CALML3	1.554121	428	0
PRKACG	1.554121	141	0	FYN	1.554121	86	0
MAPK1	1.554121	113	0	CALML4	1.554121	425	0
CALML6	1.554121	425	0	CALML5	1.554121	426	0
MAPK3	1.554121	137	0	ESR1	1.554121	87	0
GNAQ	1.554121	82	0	IQGAP1	1.554121	87	0
CAV1	1.554121	108	0	PRKACB	1.554121	145	0
BRAF	1.554041	102	0	HSP90AA1	1.554121	94	0
HRAS	1.554121	130	0	MAP2K1	1.554121	88	0
JAK2	1.554121	70	0	PLCG1	1.554121	105	0
PLEK2	1.554121	67	0	RAF1	1.554121	101	0
EGF	1.554121	89	0	RASA1	1.554121	82	0
GSK3B	1.554121	109	0	PIK3CA	1.554121	114	0
GRB2	1.554121	121	0	CDC42	1.554121	110	0
FGFR1	1.554121	77	0	ARRB1	1.554121	76	0
ARRB2	1.554121	78	0	KRAS	1.554121	143	0
SRC	1.554121	168	0	YWHAZ	1.554121	63	0
PIK3R1	1.554121	96	0	RHOA	1.554121	130	0
NTRK1	1.554121	77	0	NOS3	1.554121	69	0
EGFR	1.554121	136	0	RABIF	1.554121	65	0
BDNF	1.554121	94	0	CALM3	1.554121	421	0
CREB1	1.554121	104	0	AKT1	1.554121	183	0
NRAS	1.554121	118	0	FN1	1.554121	98	0
MTOR	1.554121	95	0	NGF	1.554121	75	0
HSP90AB1	1.554121	83	0	ITPR1	1.540830	105	0
DLG4	1.540830	80	0	PRKCB	1.540830	76	0

NTRK2	1.540086	65	0	IGF1R	1.530644	79	0
PLCG2	1.529009	65	0	IRS1	1.528369	79	0
GRIN2B	1.500956	75	0	CAMK2A	1.500956	93	0

The PPI network is displayed in Figure 3.47. The hub proteins are colored in blue. Light gray color is used to show non-hub proteins. Every node stands for one protein. Proteins' interactions are represented with green edges. The design effectively distinguishes between those connected hubs and the nodes at the edge of the network. The important proteins get to be easily spotted. The title along with the legend tells the main network metrics. Total nodes, total edges, and hub proteins count are among these statistics.

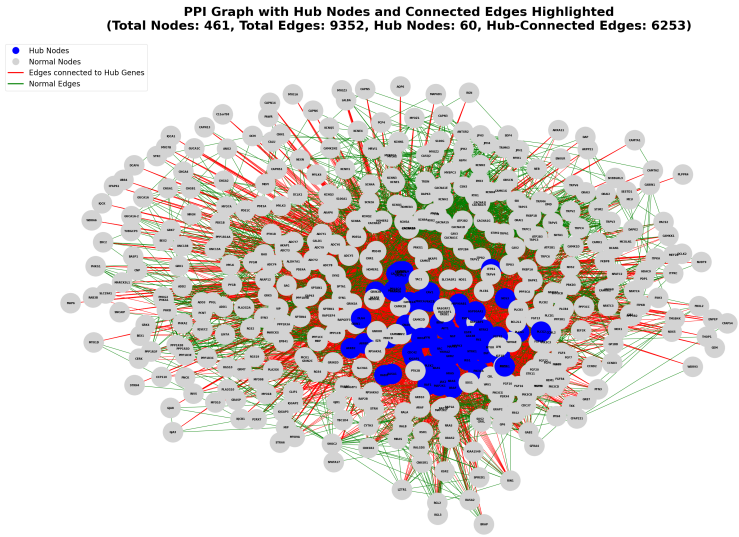
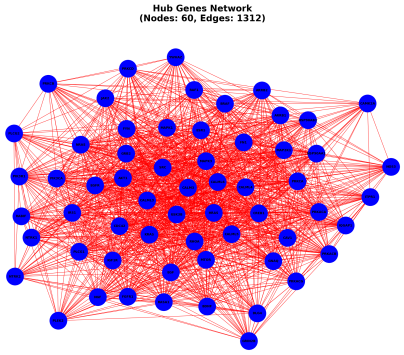


Figure 3.47: Hub Proteins inside Network

The description of the hub gene subnetwork is provided in Figure 3.48a and Figure 3.48b. The size of the subnetwork is 60 nodes. Mutual connections of the subnetwork are 1312 edges (nodes). Biological core of very high interconnection is formed by this. The average degree is 43.73. One hub gene connects directly to approximately 44 other hubs. Therefore the interaction intensity is very strong. The network density is 0.7412. The existence of nearly 74% of all possible connections can be said. The subnetwork is consequently, very compact and cohesive. The average clustering coefficient is 0.8029. Hub genes are like tightly knit groups. The chance of a neighbor of a hub connecting to another neighbor is very high. All in all, these metrics portray a central module. Cooperation is the way

of hub genes. Indubitably, they will be involved in, or at least very important to, signal transmission, regulation and functional stability.



(a) Hub Proteins Based on Hub Score

Property	Value
Number of Hub Nodes	60
Number of Edges Among Hub Genes	1312
Average Degree	43.73
Density	0.7412
Average Clustering Coefficient	0.8029

(b) Properties of Hub Proteins Subnetwork

Figure 3.48: Visualization and Topological Measures of Hub Subnetwork

The analysis revealed a strong linkage between the two methods of hub detection. Hub genes discovered by degree centrality signify a part. They are entirely enveloped by the hubs detected via the hub score. The pattern exhibited by this subset supports the conclusion of being reliable and consistent. Both methods provide access to the majority of crucial nodes. The combined hub score covers a larger number of proteins but still highlights the same most important ones. This agreement significantly increases the confidence in the selected hub genes.

Pathway enrichment

Pathway enrichment analysis involved the use of the Kyoto Encyclopedia of Genes and Genomes (KEGG) and Reactome databases. One of the main features of these databases is the precise and detailed information concerning the genes involved and their molecular interactions. The attention during the analysis was mainly directed towards the previously identified hub genes. A number of biological pathways were observed to be significantly enriched. The predominant pathways enriched were those related to cancer, apoptosis, and signal transduction. This indicates that the hub genes are participating in the PPI network with very important regulatory roles. They are located right in the middle of the major biological processes. It becomes clear that the hubs are not scattered randomly. Instead,

they are in charge of playing functional roles in both disease and cellular mechanisms. The Ras signaling pathway was the most enriched one, with a very high level of significance. The p-value of the Ras signaling pathway was 3.17×10^{-50}. The Neurotrophin signaling pathway was the second most significant, followed by the Glioma pathway. Other pathways with high significance included:

- Pathways in cancer
- Proteoglycans in cancer
- Estrogen signaling pathway
- Rap1 signaling pathway

43 Cancer associated Genes Identified from Enrichrment results as shown in table 3.23. All these pathways had a very strong statistical association with the hub genes. A majority of these signaling pathways are directly linked to the regulation of cancer, where the hub genes marked have a great influence on the cell proliferation, apoptosis, and transduction of signals. The like of *EGFR*, *AKT1*, *MAPK1*, *BRAF*, and *KRAS* were discovered to be the main facilitators in the significant signaling cascades of MAPK that is very important for the survival and growth of the tumor.

Pathway / Term	Overlap / Total	Genes
Ras signaling pathway	33 / 232	CALML5; CALML6; CALML3; CALML4; PIK3R1; EGFR; CDC42; NRAS; PRKACG; PLCG2; AKT1; MAPK1; PLCG1; PRKACA; HRAS; PRKACB; MAPK3; NTRK1; NTRK2; MAP2K1; EGF; BDNF; PRKCB; PRKCA; NGF; RHOA; PIK3CA; RASA1; CALM3; KRAS; GRB2; RAF1; FGFR1
Estrogen signaling pathway	29 / 137	HSP90AB1; CALML5; CALML6; SRC; ITPR1; CALML3; CALML4; PIK3R1; EGFR; NRAS; PRKACG; AKT1; MAPK1; PRKACA; HRAS; PRKACB; MAPK3; HSP90AA1; MAP2K1; NOS3; PRKCD; ESR1; CREB1; PIK3CA; GNAQ; CALM3; KRAS; GRB2; RAF1
Pathways in cancer	38 / 531	GSK3B; HSP90AB1; CALML5; CALML6; CALML3; CALML4; PIK3R1; EGFR; CDC42; NRAS; PRKACG; PLCG2; AKT1; MAPK1; PLCG1; PRKACA; JAK2; HRAS; PRKACB; MAPK3; NTRK1; HSP90AA1; MAP2K1; EGF; PRKCB; FN1; BRAF; PRKCA; ESR1; RHOA; MTOR; PIK3CA; GNAQ; CALM3; KRAS; GRB2; RAF1; FGFR1
Neurotrophin signaling pathway	27 / 119	GSK3B; CALML5; CALML6; CALML3; CALML4; PIK3R1; CDC42; NRAS; PLCG2; AKT1; MAPK1; PLCG1; HRAS; MAPK3; NTRK1; NTRK2; MAP2K1; BDNF; PRKCD; BRAF; NGF; RHOA; PIK3CA; CALM3; KRAS; GRB2; RAF1
Proteoglycans in cancer	30 / 205	SRC; ITPR1; IQGAP1; PIK3R1; EGFR; CDC42; NRAS; PRKACG; PLCG2; AKT1; MAPK1; PLCG1; PRKACA; HRAS; PRKACB; MAPK3; MAP2K1; PRKCB; CAV1; FN1; BRAF; PRKCA; ESR1; RHOA; MTOR; PIK3CA; KRAS; GRB2; RAF1; FGFR1
Glioma	24 / 75	MAP2K1; CALML5; EGF; CALML6; PRKCB; BRAF; CALML3; PRKCA; CALML4; PIK3R1; EGFR; MTOR; NRAS; PIK3CA; PLCG2; AKT1; MAPK1; CALM3; KRAS; GRB2; PLCG1; RAF1; HRAS; MAPK3
GnRH signaling pathway	24 / 93	MAP2K1; CALML5; CALML6; PRKCB; SRC; PRKCD; ITPR1; CALML3; PRKCA; CALML4; EGFR; CDC42; NRAS; GNAQ; PRKACG; MAPK1; CALM3; KRAS; GRB2; RAF1; PRKACA; HRAS; PRKACB; MAPK3
Human cytomegalovirus infection	29 / 225	GSK3B; CALML5; CALML6; SRC; ITPR1; CALML3; CALML4; PIK3R1; EGFR; NRAS; PRKACG; AKT1; MAPK1; PRKACA; HRAS; PRKACB; MAPK3; MAP2K1; PRKCB; PRKCA; RHOA; MTOR; CREB1; PIK3CA; GNAQ; CALM3; KRAS; GRB2; RAF1
Chemokine signaling pathway	27 / 192	GSK3B; SRC; ARRB1; PIK3R1; ARRB2; CDC42; NRAS; PRKACG; PLCG2; AKT1; MAPK1; PLCG1; PRKACA; JAK2; HRAS; PRKACB; MAPK3; MAP2K1; PRKCB; PRKCD; BRAF; RHOA; PIK3CA; GNAQ; KRAS; GRB2; RAF1
Growth hormone synthesis, secretion and action	24 / 119	GSK3B; MAP2K1; PRKCB; ITPR1; PRKCA; PIK3R1; MTOR; NRAS; CREB1; PIK3CA; GNAQ; PRKACG; PLCG2; AKT1; MAPK1; KRAS; GRB2; PLCG1; RAF1; PRKACA; JAK2; HRAS; PRKACB; MAPK3

The enrichment analysis as a whole indicated that the hub genes uncovered in this research have biological significance and are mechanistically connected to tumorogenesis. By participating in various signaling and cancer-related pathways, especially those governed by the receptor tyrosine kinases (RTKs) and hormonal factors, the genes demonstrate the intensity of communication between the growth and endocrine signaling systems. The findings support the notion that the hub genes have evolved as major molecular players in the inters action of cancer-related signaling pathways and hence become strong candidates for additional testing in terms of functionality, marker discovery and even the development of targeted therapy.

Term	P-value	FDR	Gene Count	Combined Score
Ras signaling pathway	3.17×10^{-50}	5.81×10^{-48}	33	16225.54
Estrogen signaling pathway	4.58×10^{-49}	4.19×10^{-47}	29	21957.15
Pathways in cancer	4.73×10^{-47}	2.89×10^{-45}	38	8884.60
Neurotrophin signaling pathway	2.19×10^{-46}	1.00×10^{-44}	27	21121.38
Proteoglycans in cancer	1.22×10^{-45}	4.46×10^{-44}	30	13480.04
Glioma	3.87×10^{-45}	1.18×10^{-43}	24	29916.31
GnRH signaling pathway	1.59×10^{-42}	4.16×10^{-41}	24	20791.68
Human cytomegalovirus infection	2.68×10^{-42}	6.13×10^{-41}	29	10359.04
Chemokine signaling pathway	2.89×10^{-40}	5.89×10^{-39}	27	10160.64
Growth hormone synthesis, secretion and action	1.22×10^{-39}	2.24×10^{-38}	24	14040.60

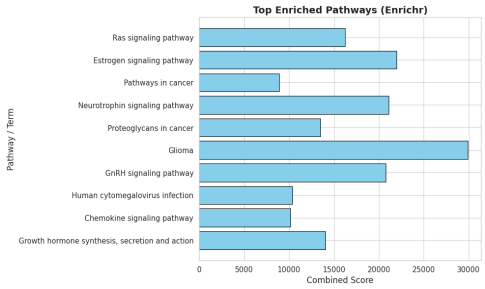


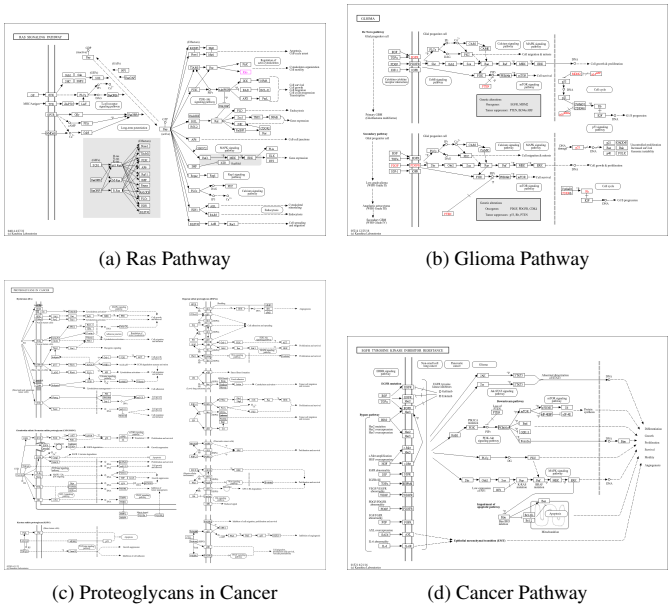
Figure 3.49: Top 10 enriched pathways (left) and their visualization (right).

Figure 3.49 represents the resultant pathways that have undergone enrichment analysis. The pathways that were found to be enriched during the analysis are the ones that are directly associated with cellular signaling and the cancer process. Among these pathways, the Ras signaling pathway is the major one where the signal is being transduced. Another one is the Neurotrophin signaling pathway cell which decides the fate of a cell by either keeping it alive and letting it grow or killing it. In addition, the Rap1 signalling pathway is the one that connects the signal from the membrane to the cell nucleus. Also, the Glioma pathway illustrates the mechanisms of tumor progression. Cancer pathways involve an extensive range of malignancy processes. Another example is the case of proteoglycans in cancer, where the tumor microenvironment changes are supported. The Estrogen signaling pathway is responsible for hormonal regulation. The GnRH signaling pathway is the one that is connected to reproductive cancers. Receptor Tyrosine Kinases (R-HSA-9006934) signaling regulates the proliferation of cells within the body. oncogenic transformation is facilitated by this pathway too. The Human cytomegalovirus infection pathway suggests the viral influence on the signaling area. All these pathways are thus pointing out hub genes having functional importance. The hub genes control the activities of the core cells. They are holding an especially important position during the whole procedure of cancer development and signal transduction.

Hub Gene	Cancerous	Non-Cancerous	Hub Gene	Cancerous	Non-Cancerous
PRKACA	Yes	No	PRKCA	Yes	No
PRKCD	No	Yes	CALML3	Yes	No
PRKACG	Yes	No	FYN	No	Yes
MAPK1	Yes	No	CALML4	Yes	No
CALML6	Yes	No	CALML5	Yes	No
MAPK3	Yes	No	ESR1	Yes	No
GNAQ	Yes	No	IQGAP1	Yes	No
CAV1	Yes	No	PRKACB	Yes	No
BRAF	Yes	No	HSP90AA1	Yes	No
HRAS	Yes	No	MAP2K1	Yes	No
JAK2	Yes	No	PLCG1	Yes	No
PLEK2	No	Yes	RAF1	Yes	No
EGF	Yes	No	RASA1	No	Yes
GSK3B	Yes	No	PIK3CA	Yes	No
GRB2	Yes	No	CDC42	Yes	No
FGFR1	Yes	No	ARRB1	No	Yes
ARRB2	No	Yes	KRAS	Yes	No
SRC	Yes	No	YWHAZ	No	Yes
PIK3R1	Yes	No	RHOA	Yes	No
NTRK1	Yes	No	NOS3	No	Yes
EGFR	Yes	No	RABIF	No	Yes
BDNF	No	Yes	CALM3	Yes	No
CREB1	Yes	No	AKT1	Yes	No
NRAS	Yes	No	FN1	Yes	No
MTOR	Yes	No	NGF	No	Yes
HSP90AB1	Yes	No	ITPR1	Yes	No
DLG4	No	Yes	PRKCB	Yes	No
NTRK2	No	Yes	PLCG2	Yes	No

Table 3.23: Summary of Hub Genes with Cancer Association

The pathway assignments for the presented figures were obtained directly from the KEGG Pathway interface, ensuring accurate biological interpretation and reproducibility. Subfigure a depicts the Ras signaling pathway. Ras proteins are GTPase molecular switches regulating cell proliferation, survival, growth, migration, differentiation and cytoskeletal dynamics. Subfigure b shows Gliomas pathway. Gliomas are the most common of the primary brain tumors and account for more than 40% of all central nervous system neoplasms. Gliomas include tumours that are composed predominantly of astrocytes, oligodendrocytes, mixtures of various glial cells and ependymal cells. Subfigure c illustrate cancer pathways. Proteoglycans (PGs) in the tumor microenvironment regulate cancer proliferation, adhesion, angiogenesis and metastasis. Subfigure d illustrate general cancer pathways. EGFR is a tyrosine kinase that takes part in the regulation of cellular homeostasis. EGFR also serves as a stimulus for cancer growth. EGFR gene mutations and protein overexpression, both of which activate down stream pathways, are associated with cancers, especially lung cancer.



(a) Ras Pathway

(b) Glioma Pathway

(c) Proteoglycans in Cancer

(d) Cancer Pathway

Figure 3.50: Pathways

Chapter 4
Conclusion

The present study offers an integrated machine learning (ML) solution for providing essential molecular properties for drug molecules using SMILES notation representations for molecules in the form of graphical structures using topological indices derived from these structures following a graphical approach. Integrating cheminformatics with the field of data science, it leverages the strengths of an integrated workflow for molecular structure representation, topological descriptor development, and interpretable ML modeling for solving quantitative structure property relationships (QSPR) using a dataset consisting of forty seven molecules with sixteen structural descriptors that include the Wiener, Balaban, Zagreb, Shannon Entropy indices, among others for drug molecules.

Using multiple regression methods such as OLS, Ridge, LASSO, Elastic Net regressors, distinct differences in relationships for descriptor properties were obtained. In the case of molecular weight (MW), high linear correlations were obtained with high values of R^2 (approximately 0.95-0.97), thereby validating the usage of topological properties in accurately predicting molecular weight. In the case of the partition coefficient (LogP), moderate values of R^2 (approximately 0.4) were obtained, thereby validating that topology is only a partial contributor for hydrophobicity, thereby requiring other chemico-physico properties like electronic properties for better prediction. In the case of hydrogen bond donors (HBD), poor values for R^2 (less than 0.5) were obtained, thereby validating that the property is more dependent on the chemistry associated with specific functional groups rather than topological properties associated with molecules. In each of these models, LASSO and Elastic Net regression performed well in counteracting overfitting and multicollinearity in the presence of an exceedingly small dataset.

Interpretability using SHapley Additive exPlanations (SHAP) helped identify feature contribution, with key descriptors including Shannon Entropy, Wiener, and Schultz indices, indicating these descriptors dominating roles in the prediction capabilities. Descriptors such as these combine molecular complexity, connectivity, and the variance in the number of degrees, thus validating their relevance in chemistry.

However, in the context of the pertinent issue of accuracy of prediction, this specific research is also important to the larger requirement of guaranteeing there is transparency inherent in computational methodology for drug discovery. Whats more, given it offers

an interpretable result on the premise of knowledge-based paradigms, it ensures competent design, decreases costs of simulation testing, can predict it will effectively function in low data settings by adjusting specific parameters, all in an interpretable deep learning framework of research.

The focus of the upcoming developments within this will be on extending this from the available compound libraries by including quantum chemical descriptors, 3D structural descriptors, functional group descriptors, and then focusing on various non-linear learning approaches that could be using Support Vector Regression Machines or Graph Neural Networks. Each of these developments will continue focusing on improving the expressiveness of this task within the models while maintaining their interpretability. There is a promising direction that comes from the intersection of mathematical graph theory, chemistry, and explainability in AI related to drug discovery.

The fusion of 17 topological attributes was the basis of a data comprising comprehensive information. This approach brought to light the complex network dynamics. It was, in contrast, to the simplest ones based on degrees. Nevertheless, it does have some drawbacks. One of them is that the STRING database has included predicted interactions in its data, which may have been the source of the bias. The analysis did not consider the directionality of the signaling pathways since it was done on undirected graphs. The mentioned problems can be solved in future research by possibly using weighted edges. Dynamic differential equation simulations are possible. Different osmics such as transcriptomics integration will result in more precise outcomes. The discovered hubs have great potential in the clinic. EGFR's high centrality is a sign of it being a target for therapy. It justifies the use of tyrosine kinase inhibitors like osimertinib. It has been shown that computational methods speed up the discovery process. The research platform is able to connect fundamental research and clinical use. It is a pioneering step for network based pharmacology in cancer therapy.

Chapter 5
References

References

[1] H. Van de Waterbeemd *et al.*, Glossary of terms used in computational drug design (IUPAC recommendations 1997), *Pure Appl. Chem.*, vol. 69, no. 5, pp. 1137–1152, 1997.

[2] J. C. Dearden, The use of topological indices in QSAR and QSPR modeling, *Advances in QSAR Modeling*, vol. 57, 2017.

[3] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*. Wiley-VCH, 2000.

[4] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, vol. 1, Wiley-VCH, 2009.

[5] I. Gutman and B. Furtula, A survey on terminal wiener index, in *Novel Molecular Structure Descriptors*, Univ. Kragujevac, pp. 173–190, 2010.

[6] S. Hayat, S. Ahmad, H. M. Umair, and S. Wang, Distance property of chemical graphs, *Hacettepe J. Math. Stat.*, vol. 47, no. 5, pp. 1071–1081, 2018.

[7] S. Kanwal *et al.*, On topological indices of total graph and its line graph for Kragujevac tree networks, *Complexity*, vol. 2021, 2021.

[8] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, 2022.

[9] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[10] Bron, C., & Kerbosch, J., Algorithm 457: finding all cliques of an undirected graph, *Communications of the ACM*, vol. 16, no. 9, pp. 575–577, 1973.

[11] Yang, Q., & Lonardi, S., A parallel edge-betweenness clustering tool for protein-protein interaction networks, *International Journal of Data Mining and Bioinformatics*, vol. 1, no. 3, pp. 241–247, 2007.

[12] Carlsson, G., Topological data analysis in bioinformatics, *BIRS Workshop Proceedings*, vol., no., pp. 1–11, 2012.

[13] Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J., & Saez-Rodriguez, J., Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties, *PLoS One*, vol. 8, no. 4, e61318, 2013.

[14] Tasdighian, S., Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Palumbo, P., Mei, G., Di Venere, A., & Giuliani, A., Modules identification in protein structures: topological and geometrical solutions, *Journal of Chemical Information and Modeling*, vol. 54, no. 1, pp. 159–168, 2014.

[15] Hoang, T., Supervised machine learning for adverse drug reaction detection, *Journal of Biomedical Informatics*, vol. 85, pp. 123–134, 2018.

[16] Redkar, N., Wrapper-based feature selection and class balancing for drug-target interaction prediction, *Computers in Biology and Medicine*, vol. 123, pp. 103892, 2020.

[17] Bagherian, M., Machine learning approaches for drug-target interaction prediction: a survey, *Briefings in Bioinformatics*, vol. 22, no. 6, pp. bbab216, 2021.

[18] Yu, X., Recent advances in machine learning for protein-lncRNA interaction prediction, *Frontiers in Genetics*, vol. 13, pp. 834567, 2022.

[19] Erciyes, K., Graph-based analysis of biological networks: hubs, motifs, and communities, *Computational Biology and Chemistry*, vol. 99, pp. 107698, 2023.

[20] Demirsoy, S., Identifying drug-drug interactions using machine learning on clinical datasets, *BMC Medical Informatics and Decision Making*, vol. 23, pp. 45, 2023.

[21] Abubakar, A., Predicting physical properties of antibacterial drugs using QSPR models and topological indices, *Journal of Cheminformatics*, vol. 16, no. 1, pp. 12–28, 2024.

[22] Kour, R., Machine learning models for anti-cancer drug predictions using topological indices, *Artificial Intelligence in Medicine*, vol. 133, pp. 102512, 2024.

[23] Ejima, T., Ensemble learning framework for modeling drug properties in mental disorders, *BMC Bioinformatics*, vol. 25, pp. 101, 2024.

[24] Tang, Y., Integrating machine learning and bioinformatics for disease progression prediction, *Briefings in Bioinformatics*, vol., no., pp., 2024.

[25] Xu, R., & Wunsch, D., Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

[26] Rodriguez-Nieva, J. F., Identifying topological order via machine learning, *Physical Review B*, vol. 100, no. 4, pp. 045136, 2019.

[27] Kiouri, I., Structure-based PPI prediction using ML and deep learning, *Computational and Structural Biotechnology Journal*, vol. 23, pp. 101574, 2025.

[28] Zaman, R., Role of chemical graphs and topological indices in anti-HIV drug design, *Bioinformatics*, vol. 41, no. 2, pp. 325–338, 2025.

[29] Zhuang, J., Network-based identification of hub gene networks in breast cancer, *BMC Systems Biology*, vol. 9, no. 1, pp. 37, 2015.

[30] Yu, X., Integrative bioinformatics and ML for key gene identification in preeclampsia, *Frontiers in Genetics*, vol. 16, pp. 1123456, 2025.

[31] Hasan, M., Statistical and ML approaches for key gene identification in hepatocellular carcinoma, *Scientific Reports*, vol. 15, pp. 5678, 2025.

[32] Chen, L., Development of hub gene and immune cell pattern identification in PCOS via RNA-seq and ML, *Reproductive Biology and Endocrinology*, vol. 23, pp. 110, 2025.

[33] Kamal, M., Node2Vec and autoencoder-based deep learning for cancer hub gene prediction, *Computational Biology and Chemistry*, vol. 112, pp. 103756, 2025.

[34] Zhou, H., ML-based identification of key hub genes for alopecia areata, *Frontiers in Immunology*, vol. 16, pp. 1204567, 2025.

[35] Khalid, S., Degree-based topological indices for Astragaloside IV and module detection, *Journal of Molecular Graphics and Modelling*, vol. 110, pp. 108034, 2022.

[36] P. Carracedo-Reboredo *et al.*, A review on machine learning approaches and trends in drug discovery, *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, 2021.

[37] M. S. Sardar *et al.*, Improved QSAR methods for predicting drug properties utilizing topological indices and machine learning models, *Eur. Phys. J. E*, vol. 48, p. 25, 2025.

[38] G. A. Pinheiro *et al.*, Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset, *J. Chem. Inf. Model.*, vol. 61, no. 5, pp. 2338–2351, 2021.

[39] F. Rottach, S. Schieferdecker, and C. Eickhoff, The topology of molecular representations and its influence on machine learning performance, *J. Cheminform.*, vol. 17, p. 109, 2025.

[40] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, Using machine learning algorithms for breast cancer risk prediction and diagnosis, *Procedia Comput. Sci.*, vol. 83, pp. 1064–1069, 2016.

[41] M. Amrane, N. Oukid, N. Gagaoua, and A. Ensibi, Breast cancer classification using machine learning, in *Proc. 9th Int. Conf. Emerging Ubiquitous Syst. Pervasive Netw. (EUSPN)*, 2018.

[42] R. Rabiei, M. R. Jahantigh, and H. Hashemzadeh, Prediction of breast cancer using machine learning approaches, *J. Biomed. Phys. Eng.*, vol. 12, no. 3, pp. 297–308, 2022.

[43] M. M. Islam, S. I. Hasan, and M. M. Rahman, Breast cancer prediction: A comparative study using machine learning techniques, *SN Comput. Sci.*, vol. 1, no. 5, pp. 290, 2020.

[44] J. Vamathevan *et al.*, Applications of machine learning in drug discovery and development, *Nat. Rev. Drug Discov.*, vol. 18, no. 6, pp. 463–477, 2019.

[45] Barabsi, A.-L., Gulbahce, N., & Loscalzo, J. (2011). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **12**(2), 56–68.

[46] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, **71**(3), 209–249.

[47] Jeong, H., Mason, S. P., Barabsi, A.-L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(6833), 41–42.

[48] Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., & Bagos, P. G. (2011). How do we measure centrality in biological networks? *BMC Bioinformatics*, **12**(1), 1–11.

[49] Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature*, **553**(7689), 446–454.

[50] Barabsi, A.-L., & Oltvai, Z. N. (2004). Network biology: understanding the cells functional organization. *Nature Reviews Genetics*, **5**(2), 101–113.

[51] Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**(1), 1–13.

[52] Zeng, X., Liao, Y., Liu, Q., & others. (2019). Machine learning approaches for essential gene identification. *Briefings in Bioinformatics*, **20**(5), 1658–1668.

[53] Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., & others. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, **49**(D1), D605–D612.